TRANSFORMACIONES A UNA LINEA RECTA,POR DR. PRIMITIVO REYES AGUILAR

A veces se detecta no linealidades a través de la prueba de falta de ajuste descrita en la sección anterior o de diagramas de dispersión y gráficas de los residuos. En algunos casos los datos se pueden transformar para que representen una relación más lineal.

Varias funciones linealizables se encuentran en la página siguiente (fig. 2.13)[1] y sus correspondientes funciones no lineales, transformaciones y formas lineales resultantes se muestran en la tabla 2.1. Dependiendo de la curvatura del comportamiento de la relación entre las variables *X* y *Y*, se puede localizar una gráfica parecida en la figura 3.13 y usar su transformación.

Tabla 2.1 Funciones linealizables y su forma lineal correspondiente.

Figura 2.13	Función	Transformación	Forma lineal
a,b	$Y = \beta_0 X^{\beta_1}$	$Y' = \log Y, X' = \log X$	$Y' = \log \beta_0 + \beta_1 X'$
c,d	$Y = \beta_0 e^{\beta_1 X}$	$Y' = \log Y$	$Y' = \ln \beta_0 + \beta_1 X$
e,f	$Y = \beta_0 + \beta_1 \log X$	$X' = \log X$	$Y' = \beta_0 + \beta_1 X'$
g,h	$Y = \frac{X}{\beta_0 X - \beta_1}$	$Y' = \frac{1}{Y}, X' = \frac{1}{X}$	$Y' = \beta_0 - \beta_1 X'$

Por ejemplo la función:

$$Y = \beta_0 e^{\beta_1 X} \varepsilon \qquad (2.19)$$

Puede ser transformada de acuerdo a la tabla 2.1 en:

$$\ln Y = \ln \beta_0 + \beta_1 X + \ln \varepsilon$$

ó

$$Y' = \beta_0' + \beta_1 X + \varepsilon'$$

Se requiere que la transformada del término de error sea normal e independientemente distribuida con media cero y varianza σ^2.

Varios tipos de transformaciones recíprocas pueden ser útiles. Por ejemplo:

$$Y = \beta_0 + \beta_1 \left(\frac{1}{X} \right) + \varepsilon$$

Puede ser linealizada usando la transformación recíproca X' = 1/X, quedando como:

$$Y = \beta_0 + \beta_1 X' + \varepsilon$$

Ejemplo 2.3 Un investigador desea determinar la relación entre la salida de Corriente Directa (Y) de un generador de molino de viento y la velocidad del viento (X), para ello colecta 25 pares de datos para ambas variables, utilizando el Minitab para su proceso. Los datos colectados son los siguientes:

[1] Montgomerey, Douglas C., *Introduction to Linear Regression Analysis,* John Wiley and Sons, Nueva York, 1992, pp. 90-91

Obs	X	Y	Fit	SE Fit	Residual	St Resid
1	5	1.582	1.3366	0.0519	0.2454	1.07
2	6	1.822	1.5778	0.0473	0.2442	1.06
3	3.4	1.057	0.9508	0.0703	0.1062	0.47
4	2.7	0.5	0.782	0.0806	-0.282	-1.27
5	10	2.236	2.5424	0.0875	-0.3064	-1.4
6	9.7	2.386	2.47	0.0828	-0.084	-0.38
7	9.6	2.294	2.4338	0.0804	-0.1398	-0.63
8	3.1	0.558	0.8664	0.0753	-0.3084	-1.38
9	8.2	2.166	2.0962	0.0609	0.0698	0.31
10	6.2	1.866	1.626	0.0472	0.24	1.04
11	2.9	0.653	0.8302	0.0776	-0.1772	-0.79
12	6.4	1.93	1.6622	0.0474	0.2678	1.16
13	4.6	1.562	1.2402	0.0555	0.3218	1.4
14	5.8	1.737	1.5295	0.0476	0.2075	0.9
15	7.4	2.088	1.9154	0.053	0.1726	0.75
16	3.6	1.137	0.999	0.0675	0.138	0.61
17	7.9	2.179	2.0239	0.0574	0.1551	0.68
18	8.8	2.112	2.253	0.0694	-0.141	-0.62
19	7	1.8	1.8189	0.05	-0.0189	-0.08
20	5.5	1.501	1.4451	0.049	0.0559	0.24
21	9.1	2.303	2.3253	0.0737	-0.0223	-0.1
22	10.2	2.31	2.5906	0.0907	-0.2806	-1.29
23	4.1	1.194	1.1196	0.0611	0.0744	0.33
24	4	1.144	1.0834	0.0629	0.0606	0.27
25	2.5	0.123	0.7217	0.0845	-0.5987	-2.72R

R denotes an observation with a large standardized residual

Durbin-Watson statistic = 1.21

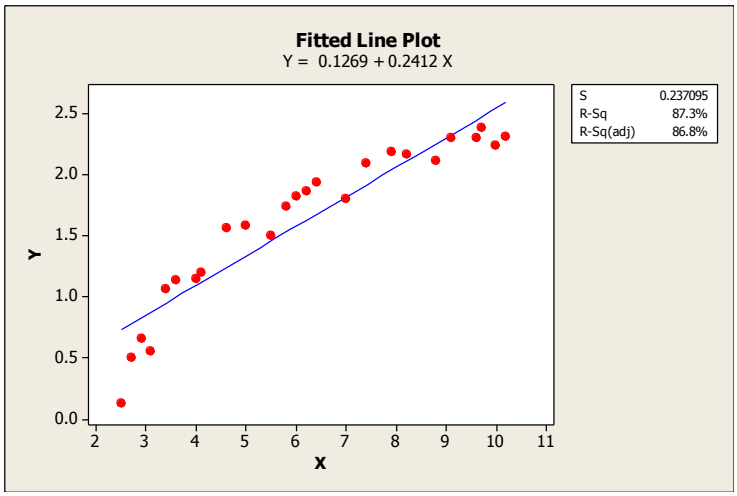
El valor del estadístico indica que no podemos llegar a conclusiones:

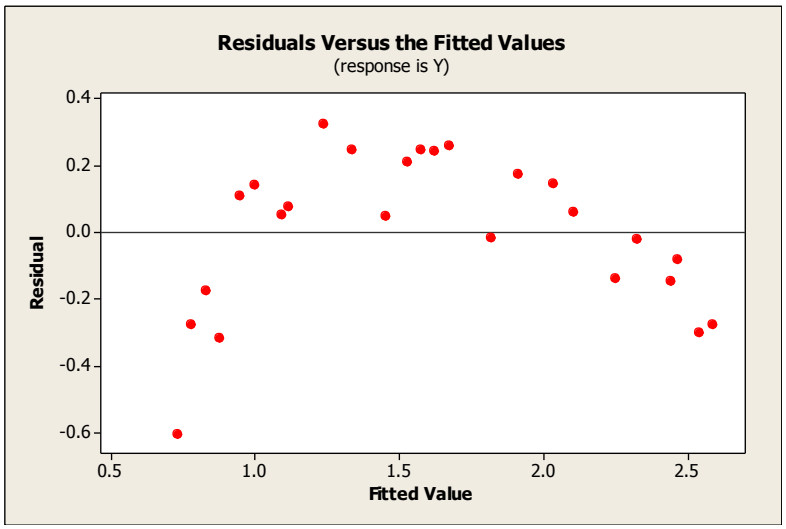
Regression Analysis: Y versus X
The regression equation is
Y = 0.131 + 0.241 X

Predictor	Coef	SE Coef	T	P
Constant	0.1309	0.1260	1.04	0.310
X	0.24115	0.01905	12.66	0.000

S = 0.2361 R-Sq = 87.4% R-Sq(adj) = 86.9%

Ajustando el modelo con una recta se tiene:



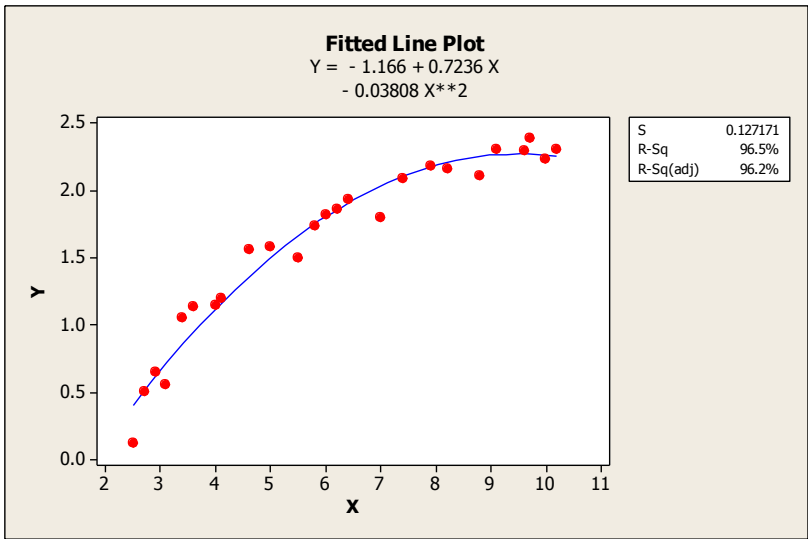


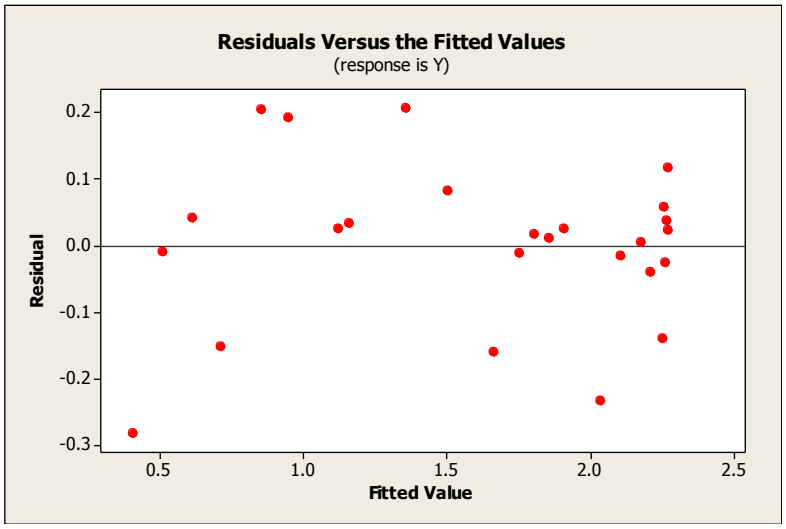
The regression equation is
Y = 0.1269 + 0.2412 X
S = 0.237095 R-Sq = 87.3% R-Sq(adj) = 86.8%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	1	8.9183	8.91827	158.65	0.000
Error	23	1.2929	0.05621		
Total	24	10.2112			

El tratar de ajustar los datos, una recta no fue la mejor opción, por lo que se intenta un modelo cuadrático, el cual se muestra a continuación.





Polynomial Regression Analysis: Y versus X

The regression equation is

Y = - 1.166 + 0.7236 X - 0.03808 X**2

S = 0.127171 R-Sq = 96.5% R-Sq(adj) = 96.2%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	2	9.8554	4.92770	304.70	0.000
Error	22	0.3558	0.01617		
Total	24	10.2112			

Sequential Analysis of Variance

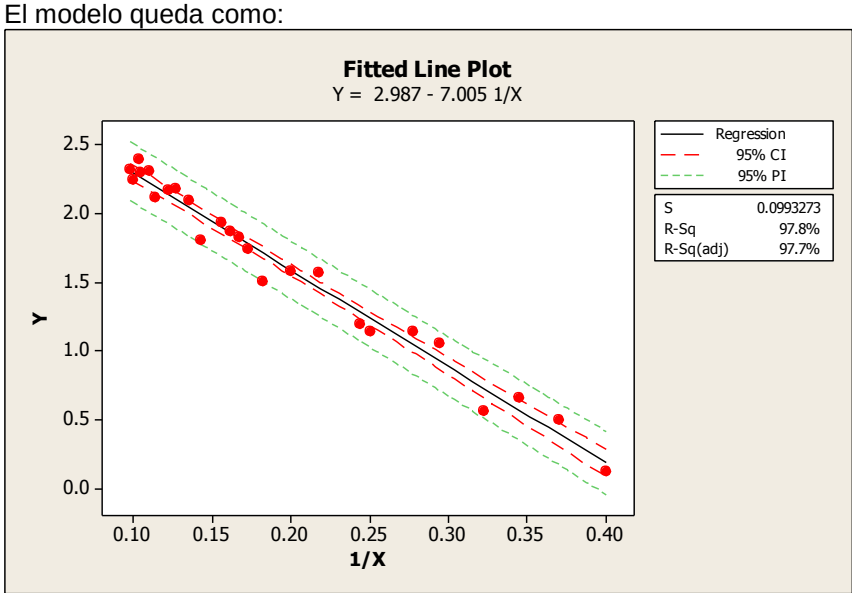
Source	DF	SS	F	P
Linear	1	8.91827	158.65	0.000
Quadratic	1	0.93713	57.95	0.000

A pesar de que la R^2 es adecuada, los residuos muestran un comportamiento anormal, por lo que será necesario transformar la variable X. Se observa que los residuos no siguen una distribución normal por lo que es necesario transformar la variable regresora:

Transformando la variable X' = 1/X se tiene, utilizando Minitab:

Obs	1/X	Y	Fit	SE Fit	Residual	St Resid
1	0.2	1.582	1.592	0.0188	-0.01	-0.11
2	0.167	1.822	1.8231	0.0199	-0.0011	-0.01
3	0.294	1.057	0.9393	0.0274	0.1177	1.31
4	0.37	0.5	0.4105	0.0404	0.0895	1.05
5	0.1	2.236	2.2854	0.0276	-0.0494	-0.55
6	0.103	2.386	2.264	0.0271	0.122	1.35
7	0.105	2.294	2.2527	0.0269	0.0413	0.46
8	0.328	0.558	0.7052	0.0329	-0.1472	-1.67
9	0.123	2.166	2.128	0.0243	0.038	0.42
10	0.161	1.866	1.8604	0.0203	0.0056	0.06
11	0.345	0.653	0.5876	0.0358	0.0654	0.75
12	0.157	1.93	1.8868	0.0206	0.0432	0.47
13	0.217	1.562	1.4713	0.0193	0.0907	0.98
14	0.172	1.737	1.7832	0.0195	-0.0462	-0.5
15	0.135	2.088	2.0418	0.0228	0.0462	0.51
16	0.278	1.137	1.0526	0.0251	0.0844	0.93
17	0.127	2.179	2.0955	0.0237	0.0835	0.92
18	0.114	2.112	2.1908	0.0256	-0.0788	-0.87
19	0.143	1.8	1.9882	0.0219	-0.1882	-2.06R
20	0.183	1.501	1.7065	0.0191	-0.2055	-2.23R
21	0.11	2.303	2.2168	0.0261	0.0862	0.95
22	0.098	2.31	2.299	0.0279	0.011	0.12
23	0.244	1.194	1.2875	0.0211	-0.0935	-1.02
24	0.253	1.144	1.2233	0.0221	-0.0793	-0.87
25	0.408	0.123	0.1484	0.0474	-0.0254	-0.31 X

El modelo queda como:



Regression Analysis: Y versus 1/X

The regression equation is

Y = 2.99 - 7.00 1/X

Predictor Coef SE Coef T P
Constant 2.98664 0.04763 62.71 0.000
1/X -7.0046 0.2202 -31.81 0.000

S = 0.0993273 R-Sq = 97.8% R-Sq(adj) = 97.7%

Analysis of Variance
Source DF SS MS F P
Regression 1 9.9843 9.9843 1012.00 0.000
Residual Error 23 0.2269 0.0099
Total 24 10.2112

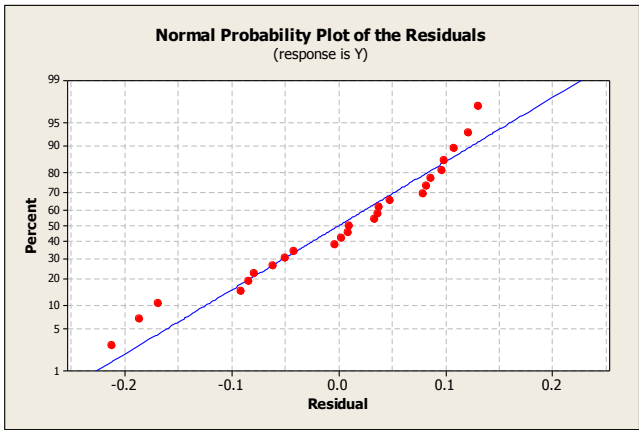
Unusual Observations

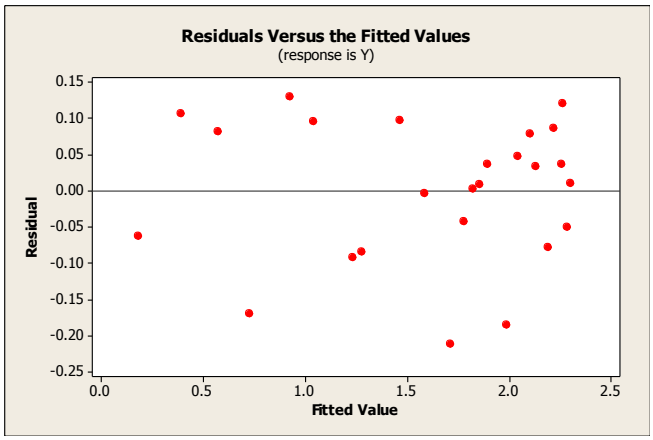
Obs 1/X Y Fit SE Fit Residual St Resid
20 0.182 1.5010 1.7131 0.0201 -0.2121 -2.18R
25 0.400 0.1230 0.1848 0.0490 -0.0618 -0.72 X

R denotes an observation with a large standardized residual.
X denotes an observation whose X value gives it large influence.

Durbin-Watson statistic = 1.52151

Como se observa ahora los residuos muestran un comportamiento normal, indicando que el modelo es adecuado.





2.5 TRANSFORMACIONES PARA ESTABILIZAR LA VARIANZA

La suposición de varianza constante es un requerimiento básico del análisis de regresión, una razón común de violación a de este supuesto es cuando la variable de respuesta Y sigue una distribución de probabilidad en la cual la varianza esta relacionada con la media. Para estos casos se utiliza transformaciones estabilizadoras de la varianza.

Si la distribución de Y es de Poisson, podemos relacionar $Y' = \sqrt{Y}$ contra X ya que la varianza de Y' es independiente de la media. Si la variable de respuesta Y es una proporción con valores entre [0,1] y la gráfica de residuos tiene el patrón de doble cresta, entonces se usa la transformación $Y' = sin^{-1}\left(\sqrt{Y}\right)$.

Otras transformaciones se muestran abajo en la tabla 2.2:

Tabla 2.2 Relaciones para transformar la varianza

Relación de σ^2 a E(Y)	Transformación	
$\sigma^2 - \alpha - constante$	$Y' = Y$	
$\sigma^2 - \alpha - E(Y)$	$Y' = \sqrt{Y}$	Datos de Poisson
$\sigma^2 - \alpha - E(Y)[1 - E(Y)]$	$Y' = sin^{-1}\sqrt{Y}$	Proporciones binomiales
$\sigma^2 - \alpha - [E(Y)]^2$	$Y' = \ln(Y)$	
$\sigma^2 - \alpha - [E(Y)]^3$	$Y' = Y^{-1/2}$	

La magnitud de la transformación, depende del grado de curvatura que induce.
La selección de la transformación se hace en base a la experiencia o de forma empírica. A continuación se presenta un ejemplo para este análisis.

Ejemplo 2.4 Se hizo un estudio entre la demanda (Y) y la energía eléctrica utilizada (X) durante un cierto periodo de tiempo, procesando los datos con Minitab se obtuvo lo siguiente:

Obs	X	Y	Fit	SE Fit	Residual	St Resid
1	679	0.79	1.649	0.351	-0.859	-0.61
2	292	0.44	0.308	0.49	0.132	0.1
3	1012	0.56	2.802	0.293	-2.242	-1.57
4	493	0.79	1.004	0.412	-0.214	-0.15
5	582	2.7	1.312	0.381	1.388	0.98
6	1156	3.64	3.301	0.297	0.339	0.24
7	997	4.73	2.75	0.294	1.98	1.38
8	2189	9.5	6.88	0.651	2.62	2.00R
9	1097	5.34	3.097	0.293	2.243	1.57
10	2078	6.85	6.495	0.6	0.355	0.27
11	1818	5.84	5.595	0.488	0.245	0.18
12	1700	5.21	5.186	0.441	0.024	0.02
13	747	3.25	1.884	0.333	1.366	0.96
14	2030	4.43	6.329	0.579	-1.899	-1.42
15	1643	3.16	4.988	0.42	-1.828	-1.31
16	414	0.5	0.73	0.441	-0.23	-0.17
17	354	0.17	0.523	0.465	-0.353	-0.25
18	1276	1.88	3.717	0.313	-1.837	-1.29
19	745	0.77	1.877	0.333	-1.107	-0.78
20	435	1.39	0.803	0.433	0.587	0.42
21	540	0.56	1.167	0.395	-0.607	-0.43
22	874	1.56	2.324	0.307	-0.764	-0.53

23	1543	5.28	4.642	0.384	0.638	0.45
24	1029	0.64	2.861	0.293	-2.221	-1.55
25	710	4	1.756	0.343	2.244	1.58

The regression equation is

Y = - 0.7038 + 0.003464 X

S = 1.46163 R-Sq = 66.4% R-Sq(adj) = 64.9%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	1	97.094	97.0943	45.45	0.000
Error	23	49.136	2.1364		
Total	24	146.231			

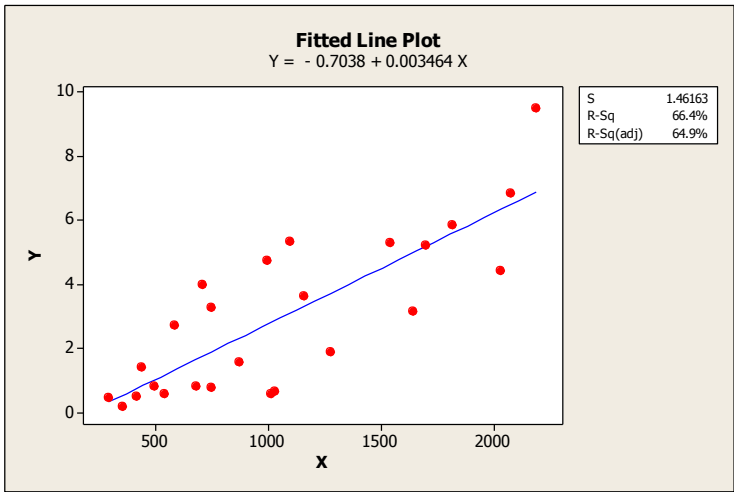
Unusual Observations

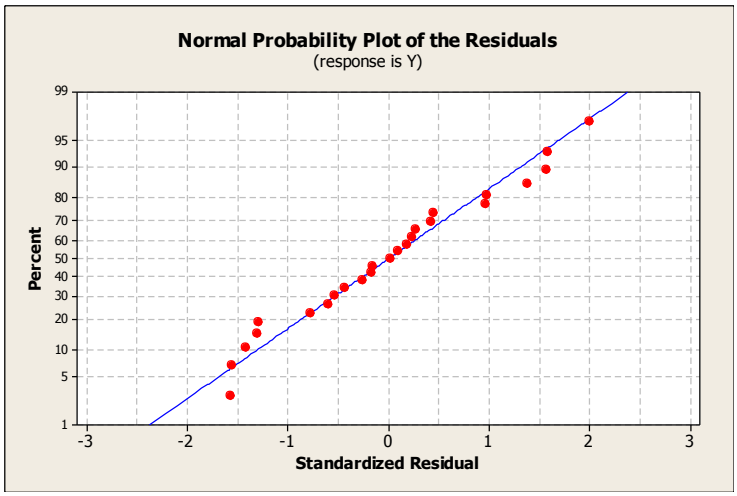
Obs	X	Y	Fit	SE Fit	Residual	St Resid
8	2189	9.500	6.880	0.651	2.620	2.00R

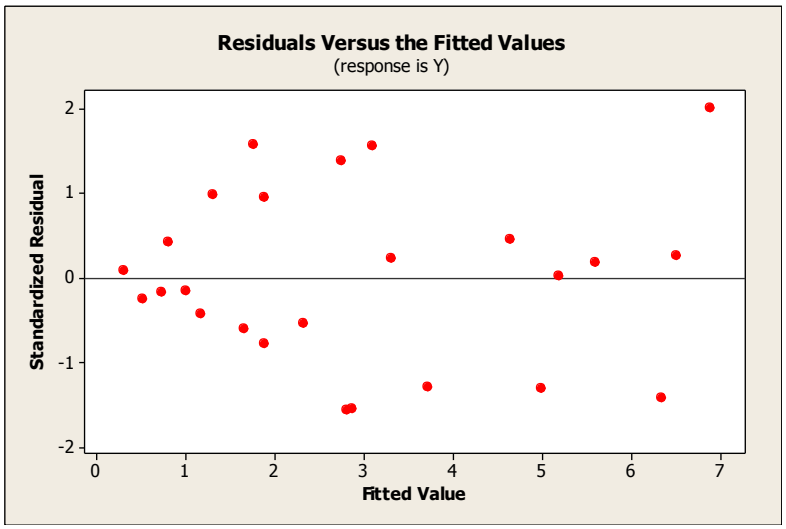
R denotes an observation with a large standardized residual.

Durbin-Watson statistic = 1.49454

Fitted Line: Y versus X







Notar que “y” es la cuenta de kilowatts utilizados por un cliente en cierta hora, se observa que la varianza aumenta conforme aumenta la media de los datos indicando que sigue el modelo de Poisson, por tanto se puede transformar con la raiz cuadrada de Y. como sigue:

Raiz(Y)	X	SRES1	TRES1	RESI1	FITS1
0.88882	679	-0.63599	-0.62755	-0.280548	1.16937
0.66333	292	-0.25322	-0.248	-0.108411	0.77174
0.74833	1012	-1.7143	-1.79523	-0.763184	1.51152
0.88882	493	-0.20513	-0.2008	-0.089439	0.97826
1.64317	582	1.30713	1.3287	0.573465	1.0697
1.90788	1156	0.55826	0.54973	0.248407	1.65947
2.17486	997	1.52481	1.57291	0.678753	1.4961
3.08221	2189	0.88812	0.88389	0.361359	2.72085
2.31084	1097	1.59927	1.65908	0.711994	1.59885
2.61725	2078	0.02523	0.02467	0.010451	2.6068
2.41661	1818	0.17965	0.17583	0.076952	2.33966
2.28254	1700	0.14802	0.14483	0.064127	2.21841
1.80278	747	1.27361	1.29201	0.563541	1.23924
2.10476	2030	-1.08504	-1.08943	-0.452723	2.55748

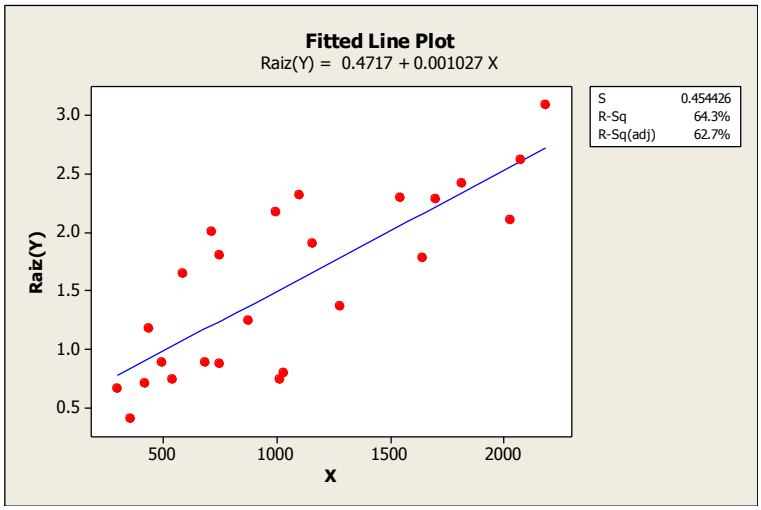
1.77764	1643	-0.87804	-0.8735	-0.38221	2.15985
0.70711	414	-0.43853	-0.4307	-0.189981	0.89709
0.41231	354	-0.98212	-0.98133	-0.423129	0.83544
1.37113	1276	-0.92738	-0.92444	-0.411636	1.78277
0.8775	745	-0.81296	-0.80676	-0.359685	1.23718
1.17898	435	0.59981	0.59127	0.260318	0.91866
0.74833	540	-0.63592	-0.62748	-0.278218	1.02655
1.249	874	-0.27173	-0.26618	-0.120724	1.36972
2.29783	1543	0.54906	0.54054	0.240723	2.0571
0.8	1029	-1.63735	-1.70373	-0.728982	1.52898
2	710	1.80812	1.90928	0.798781	1.20122

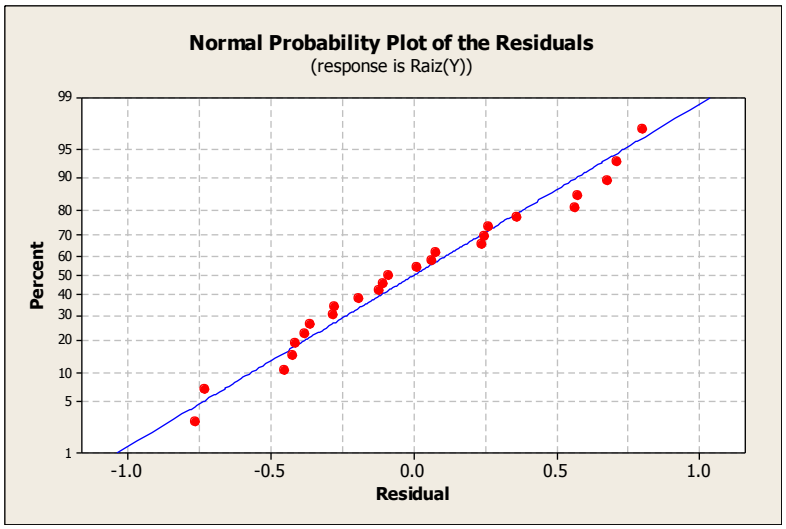
Regression Analysis: Raiz(Y) versus X

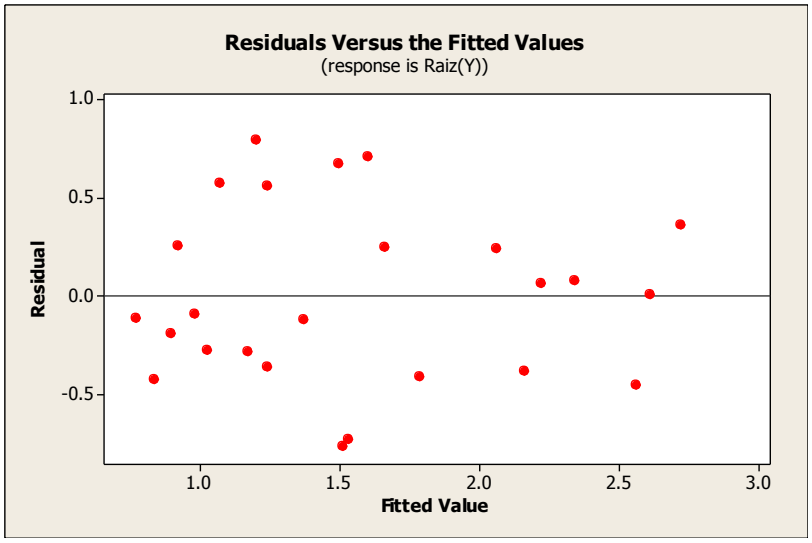
The regression equation is
Raiz(Y) = 0.4717 + 0.001027 X

S = 0.454426 R-Sq = 64.3% R-Sq(adj) = 62.7%

Durbin-Watson statistic = 1.65249







Se observa una mejor distribución normal de los residuos por lo que el modelo es adecuado. A continuación se muestra el análisis de varianza para el modelo:

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	1	8.5401	8.54008	41.36	0.000
Error	23	4.7496	0.20650		
Total	24	13.2897			

3. REGRESIÓN LINEAL MÚLTIPLE

3.1 Modelos de Regresión Múltiple

Asumiendo que N observaciones de la respuesta se puedan expresar por medio de un modelo de primer orden

$$Y_u = \beta_0 + \beta_1 X_{u1} + \beta_2 X_{u2} + ... + \beta_k X_{uk} + \varepsilon_u \qquad (3.1)$$

En la ecuación 3.1 Y*u* denota la respuesta observada en el intento *u;* X*ui* representa el nivel del factor *i* en el intento *u*; las betas son parámetros desconocidos y ε_u representa el error aleatorio en Y*u***.** Se asume que los errores ε_u tienen las características siguientes:

1. Tienen media cero y varianza común σ^2.
2. Son estadísticamente independientes.
3. Están distribuidos en forma normal.

3.2 Estimación de los parámetros del modelo

El método de mínimos cuadrados selecciona como estimados para los parámetros desconocidos beta, los valores b_0, b_1, ..., b_k respectivamente, los cuales minimizan la cantidad:

$$R(\beta_0, \beta_1, ..., \beta_k) = \sum_{u=1}^{N} (Y_u - \beta_0 - \beta_1 X_{u1} - \beta_2 X_{u2} - ..... - \beta_{uk})^2$$

Y son las soluciones a un conjunto de (k +1) ecuaciones normales.

Sobre N observaciones el modelo de primer orden puede expresarse en forma matricial como:

Y = X β + ε = [1 : D] β + ε (3.2)

Y es un vector N x 1.
X es una matriz de orden N x (k + 1), donde la primera columna es de 1's.
β es un vector de orden (k + 1) x 1.
ε es un vector de orden N x 1.
D es la matriz de Xij con i = 1, 2, ..., N; j = 1, 2, ..., k

Deseamos encontrar el vector de estimadores de mínimos cuadrados **b** que minimicen:

$$S(\beta) = \sum_{i=1}^{n} \varepsilon_i^2 = \varepsilon'\varepsilon = (Y - X\beta)'(Y - X\beta)$$

Que puede ser expresada como:

$$S(\beta) = Y'Y - \beta' X'Y - Y' X\beta + \beta' X' X\beta$$

Como $\beta' X'Y$ es una matriz 1x1 o un escalar y su transpuesta $(\beta' X'Y)' = Y' X\beta$ es el mismo escalar, se tiene:

$$S(\beta) = Y'Y - 2\beta' X'Y + \beta' X' X' \beta \quad (3.3)$$

Los estimadores de mínimos cuadrados deben satisfacer:

$$\left.\frac{\partial S}{\partial \beta}\right|_b = -2X'Y + 2X' Xb = 0$$

Que se simplifica a las ecuaciones normales de mínimos cuadrados:

X'X b = X' Y (3.4)

Los estimadores de mínimos cuadrados **b** de los elementos β son:

b = (X'X)$^{-1}$ X'Y (3.5)

El vector de valores ajustados $\hat{Y} = Xb$ se puede expresar como:

$$\hat{Y} = Xb = X(X'X)^{-1}X'Y = Hy \quad (3.5)$$

Donde la matriz **H** [n x n] se denomina la "matriz sombrero" ya que mapea el vector de valores observados dentro del vector de valores ajustados o predichos.
Como principales características de los estimadores **b** se tienen:

La matriz de varianza y covarianza de el vector de estimados **b** es:

Var**(b) = C = (X'X)$^{-1}$ σ^2** (3.6)

El elemento *(ii)* de esta matriz $c_{ii}\sigma^2 = Var(b_i)$ es la varianza del elemento *i* de **b**.

El error estándar de b_i es la raíz cuadrada positiva de la varianza de b_i o sea:

$$se.b_i = \sqrt{c_{ii}\sigma^2} \quad (3.7)$$

La covarianza del elemento b_i y b_j de **b** es $Co\,\mathrm{var}(c_{ij}) = c_{ij}\sigma^2$. (3.8)

Si los errores están normalmente distribuidos, entonces b se dice que está distribuido como:

$$b \approx N(\beta, (X'X)^{-1}\sigma^2)$$

Sea x'p un vector (1 x p) vector cuyos elementos corresponden a una fila de la matriz X, p = k + 1, entonces en la región experimental el valor de predicción de la respuesta es:

$$\hat{Y}(x) = x'_p\, b \quad (3.9)$$

Una medida de la precisión de la predicción $\hat{Y}(X)$ se puede expresar como:

$$Var(\hat{Y}(x)) = Var(x_p b) = x'_p (X'X)^{-1} x_p \sigma^2 \quad (3.10)$$

RESIDUOS
Los residuos se definen como la diferencia entre los valores reales observados y los valores predichos para estos valores de respuesta usando el modelo de ajuste y predicción, o sea:

$$r_u = Y_u - \hat{Y}(x_u), u = 1,2,...,N \quad (3.11)$$

Si se obtienen valores para los N intentos entonces en forma matricial:

$$r = \hat{Y} - Xb = Y - HY = (1-H)Y \quad (3.12)$$

los residuos tienen las propiedades siguientes:

1. 1'r = 0, donde 1' es un vector (1 x n) de 1's.
2. $\hat{Y}(X)'r = 0$
3. X'r = 0

ESTIMACIÓN DE σ
Para un modelo con p parámetros y teniendo N observaciones (N > p), la varianza se estima como sigue:

La suma de cuadros de los residuos es:

$$SSE = \sum (Y_i - \hat{Y})^2 = \sum_{i=1}^{n} e_i^2 = e'e$$

Como e = Y – X b, se tiene:

$$SSE = (Y - Xb)'(Y - Xb) = Y'Y - b'X'Y - Y'Xb + b'X'Xb = Y'Y - 2b'X'Y + b'X'Xb \quad (3.13)$$

Como X'Xb = X'Y, se transforma en:

$$SSE = Y'Y - b'X'Y \quad (3.14)$$

La suma residual de cuadrados tiene *n-p* grados de libertad asociado con el ya que se estiman *p* parámetros en el modelo de regresión. El cuadrado medio de los residuos es:

$$s^2 = MSE = \frac{SSE}{N - p} \quad (3.15)$$

3.3 Intervalos de confianza para los coeficientes de la regresión

Asumiendo que los errores son independientes y distribuidos normalmente con media cero y desviación estándar σ^2 , por tanto las observaciones Y_i también son independientes y normalmente distribuidas. Cada uno de los estadísticos:

$$\frac{b_j - \beta_j}{\sqrt{S^2 C_{jj}}}, \ldots j = 0,1,\ldots,k \quad (3.16)$$

Se distribuye con una distribución *t* con *n-p* grados de libertad, donde S^2 es la varianza del error de la ecuación (3.15). Por tanto un intervalo de confianza 100(1 - α)% para el coeficiente de regresión β_j, para *j = 0, 1, ..., k* es:

$$b_j - t_{\alpha/2,n-p} se(b_j) \le \beta_j \le b_j + t_{\alpha/2,n-p} se(b_j) \quad (3.17)$$

Donde $se(b_j)$ es el *error estándar* del coeficiente de regresión b_j.

$$se(b_j) = \sqrt{S^2 C_{jj}} \quad (3.18)$$

Siendo C_{jj} el j-ésimo elemento de la matriz $(X'X)^{-1}$.

3.3.1 Intervalos de confianza para la respuesta media en un punto en particular

Se puede construir un intervalo de confianza en la respuesta media de un punto en particular, tal como X_{01}, X_{02}, X_{03},....., X_{0K}. Definiendo el vector X_0 como:

$$X_0 = \begin{bmatrix} 1 \\ X_{01} \\ X_{02} \\ ..... \\ X_{0K} \end{bmatrix}$$

El valor ajustado en este punto es:

$$\hat{Y}_0 = X'_0 b \tag{3.19}$$

Con varianza:

$$Var(\ddot{Y}_0) = S^2 X'_0 (X'X)^{-1} X_0 \tag{3.20}$$

Por tanto el intervalo de confianza para el 100(1 - α) % es:

$$\hat{Y}_0 - t_{\alpha/2,n-p}\sqrt{S^2 X'_0 (X'X)^{-1} X_0} \le Y_0 \le \hat{Y}_0 + t_{\alpha/2,n-p}\sqrt{S^2 X'_0 (X'X)^{-1} X_0} \tag{3.21}$$

3.4 Prueba de Hipótesis en Regresión múltiple

Entre las pruebas importantes a realizar se encuentra la prueba de siginificancia de la regresión, la prueba de coeficientes individuales de la regresión y otras pruebas especiales. A continuación se analiza cada una de ellas.

3.6.1 Prueba de significancia para la regresión

La prueba de significancia de la regresión es probar para determinar si hay una relación lineal entre la respuesta *Y* y cualquiera de las variables regresoras *Xi's*, la hipótesis apropiada es:

$$\left.\begin{aligned} H_0 &: \beta_1 = \beta_2 = ... = \beta_k = 0 \\ H_0 &: \beta_j \neq 0 .....para.al.menos.una.j \end{aligned}\right\} \tag{3.22}$$

El rechazo de H_0 implica que al menos alguno de los regresores contribuye significativamente al modelo. El método es una generalización del utilizado en la regresión lineal. La suma total de cuadrados S_{yy} se divide en suma de cuadrados debidos a la regresión y la suma de cuadrados de los residuos, o sea:

$$S_{YY} = SST = SSR + SSE$$

Para la prueba de la hipótesis se utiliza el estadístico F_0 como sigue:

$$F_0 = \frac{SSR/k}{SSE/(n-k-1)} = \frac{MSR}{MSE} \qquad \text{con k = No. de variables regresoras} \tag{3.23}$$

La suma de cuadrados totales es:

$$SST = \sum_{u=1}^{N} (Y_u - \bar{\bar{Y}})^2 \qquad \text{con N-1 grados de libertad} \tag{3.24}$$

La suma de cuadrados debidos a la regresión es:

$$SSR = \sum_{u=1}^{N} (\hat{Y}(x_u) - \bar{\bar{Y}})^2 \quad \text{con p (parámetros) – 1 grados de libertad} \tag{3.25}$$

La suma de cuadrados del error o de los residuos es:

$$SSE = \sum_{u=1}^{N} (Y_u - \hat{Y}(x_u))^2$$ con (N-1) – (p –1) grados de libertad (3.26)

En forma matricial se tiene:

$$SST = Y'Y - \frac{(1'Y)^2}{N} \quad (3.27)$$

$$SSR = b'X'Y - \frac{(1'Y)^2}{N} \quad (3.28)$$

$$SSE = Y'Y - b'X'Y \quad (3.29)$$

La tabla de ANOVA para la significancia de la regresión queda como:

Fuente de variación	SS	df	MS	F_0
Regresión	SSR	K	MSR	MSR/MSE
Residuos	SSE	n – k - 1	MSE	
Total	SST	n – 1		

Para probar la hipótesis de existencia del modelo, se tiene:

$$H_0 : \beta_1 = \beta_2 = ... = \beta_k = 0$$

$$H_a : \forall \beta_i \neq 0, i = 1,2,...,k$$

Se calcula el estadístico F_0 como:

$$F_0 = \frac{MSR}{MSE} = \frac{SSR/(p-1)}{SSE/(N-p)} \quad (3.30)$$

Se compara el valor de F con el de tablas para $F_{\alpha,p-1,N-p}$ el cual es la parte superior de la distribución F, si F calculada excede a F de tablas se infiere que la variación explicada por el modelo es significativa.

El coeficiente de determinación R^2 mide *la proporción de la variación total de los valores Yu alrededor de la media Y explicada por el modelo de ajuste.* Se expresa en porcentaje.

$$R^2 = \frac{SSR}{SST} \quad (3.31)$$

3.4.2 Prueba de los coeficientes individuales de la regresión

Con frecuencia estamos interesados en probar hipótesis sobre los coeficientes de regresión individuales. Por ejemplo el modelo podría ser más efectivo con la inclusión de regresores adicionales o con la eliminación de una o más variables regresoras presentes en el modelo.

Al agregar una variable al modelo, siempre incrementa la suma de cuadrados de la regresión y decrementa la suma de cuadrados de los residuos, sin embargo también incrementa la varianza de los valores estimados Yest., de tal forma que se debe tener cuidado en incluir sólo los regresores que mejor expliquen la respuesta. Por otra parte, al agregar un regresor no importante puede incrementar el cuadrado medio de los residuos, lo que decrementa la utilidad del modelo.
La hipótesis para probar la significancia de cualquier coeficiente individual de la regresión β_j es:

$$\begin{aligned} H_0 &: \beta_j = 0 \\ H_1 &: \beta_j \neq 0 \end{aligned} \qquad (3.32)$$

Si no se rechaza H_0, indica que el regresor X_j puede ser excluido del modelo. El estadístico de prueba para esta hipótesis es:

$$t_0 = \frac{b_j}{se(b_j)} \qquad (3.33)$$

La hipótesis nula es rechazada si $|t_0| > t_{\alpha/2,n-k-1}$. Esta es una prueba parcial o marginal de la contribución de X_j dados los otros regresores en el modelo.

3.4.3 Caso especial de columnas ortogonales en X
Si dentro de la matriz X si las columnas de X_1 son ortogonales a las columnas en X_2, se tiene que $X_1'X_2 = X_2' X_1 = 0$. Entonces los estimadores de mínimos cuadrados b1 y b2 no dependen si está o no está en el modelo alguno de los otros regresores, cumpliéndose:

$$SSR(\beta_2) = SSR(\beta_1) + SSR(\beta_2) \qquad (3.34)$$

Un ejemplo de modelo de regresión con regresores ortogonales es el diseño factorial 2^3 siguiente:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon$$

Donde la matriz X es la siguiente:

$$X = \begin{bmatrix} +1,-1,-1,-1 \\ +1,+1,-1,-1 \\ +1,-1,+1,-1 \\ +1,-1,-1,+1 \\ +1,+1,+1,-1 \\ +1,+1,-1,+1 \\ +1,-1,+1,+1 \\ +1,+1,+1,+1 \end{bmatrix}$$

En este caso, $SSR(\beta_j)$, j = 1, 2, 3, mide la contribución del regresor X_j al modelo, independientemente de cualquier otro regresor esté incluido en el modelo de ajuste.

Ejemplos:
Ejemplo 3.1 Un embotellador está analizando las rutas de servicio de máquinas dispensadoras, está interesado en predecir la cantidad de tiempo requerida por el chofer para surtir las máquinas en el local (Y). La actividad de servicio incluye llenar la máquina con refrescos y un mantenimiento menor.

Se tienen como variables el número de envases con que llena la máquina (X1) y la distancia que tiene que caminar (X2). Se colectaron los datos siguientes, y se procesaron con el paquete Minitab:

X1_envases	X2_Distancia	Y_tiempo
7	560	16.68
3	220	11.5
3	340	12.03
4	80	14.88
6	150	13.75
7	330	18.11
2	110	8
7	210	17.83
30	1460	79.24
5	605	21.5
16	688	40.33
10	215	21
4	255	13.5
6	462	19.75
9	448	24
10	776	29
6	200	15.35
7	132	19
3	36	9.5
17	770	35.1
10	140	17.9
26	810	52.32
9	450	18.75
8	635	19.83
4	150	1075

De manera matricial:

	1's	X1	X2
	1	7	560
	1	3	220
	1	3	340
	1	4	80
	1	6	150
	1	7	330
X	1	2	110
	1	7	210
	1	30	1460
	1	5	605
	1	16	688
	1	10	215
	1	4	255
	1	6	462

1	9	448
1	10	776
1	6	200
1	7	132
1	3	36
1	17	770
1	10	140
1	26	810
1	9	450
1	8	635
1	4	150

La transpuesta de X es (Copiar con pegado especial Transponer):

X'

1's	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
X1	7	3	3	4	6	7	2	7	30	5	16	10	4	6	9	10	6	7	3	17	10	26	9	8
X2	560	220	340	80	150	330	110	210	1460	605	688	215	255	462	448	776	200	132	36	770	140	810	450	635

Con la función de Excel de multiplicación de matrices MMULT :
Seleccionar el rango de celdas de resultados y al final teclear (Ctrl-Shif-Enter). final)

X'X

25	219	10,232
219	3,055	133,899
10,232	133,899	6,725,688

X'y

560
7,375
337,072

El vector estimador de los coeficientes Betas es :

$$\hat{\beta} = (X'X)^{-1}X'y$$

Con la función de Excel MINVERSA

$(X'X)^{-1}$

0.113215186	-0.004449	-8.367E-05
-0.004448593	0.0027438	-4.786E-05
-8.36726E-05	-4.79E-05	1.229E-06

Matrix B = INV(X'X) X'Y

Betas est,
2.341231145
1.615907211
0.014384826

The regression equation is

Y-TENT = 2.34 + 1.62 X1-ENV + 0.0144 X2-DIST

Estadísticas de la regresión	
Coeficiente de	0.9795886
correlaciçon mçultiple	
Coeficiente de	0.9595937
determinación R^2	
R^2 ajustado	0.9559205
Error típico	3.2594734
Observaciones	25

ANÁLISIS DE VARIANZA

	Grados de libertad	*Suma de cuadrados*	*Promedio de cuadrados*	*F*	*Valor Critico de F*
Regresión	2	5550.81092	2775.405	261.235	4.6874E-16
Residuos	22	233.731677	10.62417		
Total	24	5784.5426			

	Coeficientes	*Error típico*	*Estad. t*	*Probab.*	*Inferior 95%*	*Superior 95%*	*Inferior 95.0%*	*Superior 95.0%*
Intercepción	2.3412311	1.09673017	2.134738	0.04417	0.066752	4.615710293	0.066752	4.61571029
X1_envases	1.6159072	0.17073492	9.464421	3.3E-09	1.26182466	1.969989758	1.26182466	1.96998976
X2_Distancia	0.0143848	0.00361309	3.981313	0.00063	0.00689174	0.021877908	0.00689174	0.02187791

Cálculo de la estimación de la varianza:

$Cov(\beta) = \sigma^2(X'X)^{-1}$

Si **$C = (X'X)^{-1}$**

La varianza de βi es $\sigma^2 Cjj$ y la covarianza entre βi y βj es $\sigma^2 Cij$.

Y'_tiempo	16.68	11.5	12.03	14.88	13.75	18.11	8	17.83	79.24	21.5	40.33	21	
	13.5	19.75	24	29	15.35	19	9.5	35.1	17.9	52.32	18.75	19.83	10.75

La matriz **y'y e**s:

y'y
18,310.63

β'		
2.3412	1.6159	0.0144

X'y
559.6
7375.44
337072

β'X'y
18,076.90

SSE =	233.73

σ^2 = =233.73/(25-3) **=10.6239**

SSE = y'y - β' X' y

σ^2 = MSE = SSE / (n-p)

Matrix Y'Y = 18310.6

Matrix b' = [2.34123 1.61591 0.01438]

Matrix b'X'Y = 18076.9

Matrix SSe = Y'Y - b'X'Y = 233.732

$$S^2 = \frac{SS_E}{N-p} = \frac{233.732}{25-3} = 10.624$$

Cálculo del error estándar de los coeficientes y del intervalo de confianza para α = 0.05

De ecuación 3.17 se tiene:

$$se(b_j) = \sqrt{S^2 C_{jj}}$$

Siendo C_{jj} el j-ésimo elemento de la matriz $(X'X)^{-1}$.

M8 = $(X'X)^{-1}$

0.113215186	-0.004449	-8.367E-05
-0.004448593	0.0027438	-4.786E-05
-8.36726E-05	-4.79E-05	1.229E-06

$$b_1 - t_{.025,22} se(b_1) \le \beta_1 \le b_1 + t_{.025,22} se(b_1)$$

$$1.61591 - (2.074)\sqrt{(10.6239)(0.00274378)} \le \beta_1 \le 1.6191 + (2.074)(0.17073)$$

Por tanto el intervalo de confianza para el 95% es:

$1.26181 \le \beta_1 \le 1.97001$

Cálculo del intervalo de confianza para la respuesta media

El embotellador desea construir un intervalo de confianza sobre el tiempo medio de entrega para un local requiriendo

X1 = 8 envases y cuya distancia es X2 = 275 pies. Por tanto:

$$X_0 = \begin{bmatrix} 1 \\ 8 \\ 275 \end{bmatrix}$$

El valor de respuesta estimada por la ecuación de ajuste es:

$$\hat{Y}_0 = X'_0 b = [1,8,275] \begin{bmatrix} 2.34123 \\ 1.61591 \\ 0.01438 \end{bmatrix} = 19.22 minutos$$

La varianza de $\hat{Y}_0$ es estimada por (tomando M8=inv(X'X) anterior):

$$Var(\hat{Y}_0) = S^2 X'_0 (X'X)^{-1} X_0 = 10.6239[1,8,275]M8\begin{bmatrix}1\\8\\275\end{bmatrix} = 10.6239(0.05346) = 0.56794$$

Por tanto el intervalo al 95% de nivel de confianza es:

$$19.22 - 2.074\sqrt{0.56794} \le Y_0 \le 19.22 + 2.074\sqrt{0.56794}$$

Que se reduce a:

$17.66 \le Y_0 \le 20.78$

Analysis of Variance
De ecuaciones 3.26 a 3.29

SST = 18,310.629 - $\frac{(559.6)^2}{25}$ = 5784.5426

SSR = 18,076.930 - $\frac{(559.6)^2}{25}$ = 5,550.8166

SSE = SST – SSR = 233.7260

$$F_0 = \frac{MSR}{MSE} = \frac{2775.4083}{10.6239} = 261.24$$

$$F_{0.05,2,22} = 3.44$$

Como la F calculada es mayor que la F de tablas, se concluye que existe el modelo con alguno de sus coeficientes diferente de cero.
Con el paquete Minitab se obtuvo lo siguiente:
Regression Analysis: Y_tiempo versus X1_envases, X2_Distancia

The regression equation is
Y_tiempo = 2.34 + 1.62 X1_envases + 0.0144 X2_Distancia

Predictor	Coef	SE Coef	T	P
Constant	2.341	1.097	2.13	0.044
X1_envases	1.6159	0.1707	9.46	0.000
X2_Distancia	0.014385	0.003613	3.98	0.001

S = 3.25947 R-Sq = 96.0% R-Sq(adj) = 95.6%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	2	5550.8	2775.4	261.24	0.000
Residual Error	22	233.7	10.6		
Total	24	5784.5			

Source DF Seq SS
X1_envases 1 5382.4
X2_Distancia 1 168.4

Unusual Observations

Obs	X1_envases	Y_tiempo	Fit	SE Fit	Residual	St Resid
9	30.0	79.240	71.820	2.301	7.420	3.21RX
22	26.0	52.320	56.007	2.040	-3.687	-1.45 X

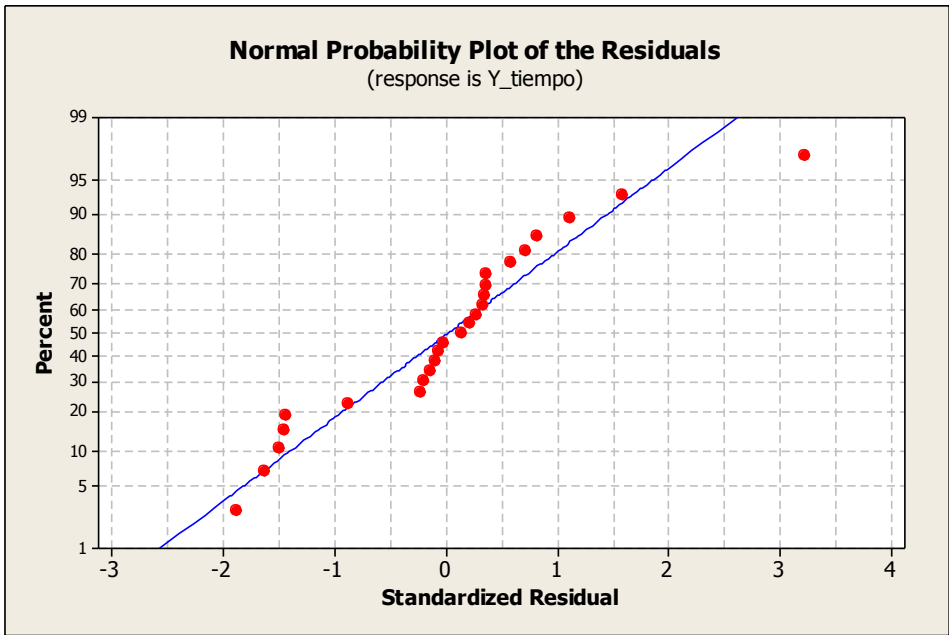
R denotes an observation with a large standardized residual.
X denotes an observation whose X value gives it large influence.

Predicted Values for New Observations

New Obs	Fit	SE Fit	95% CI	95% PI
1	19.224	0.757	(17.654, 20.795)	(12.285, 26.164)

Values of Predictors for New Observations

New Obs	X1_envases	X2_Distancia
1	8.00	275



Prueba de la significancia de los coeficientes particulares
Probando la contribución del regresor X2 (distancia) dado que la variable regresora de casos está en el modelo. Las hipótesis son:

$H_0 : \beta_2 = 0$
$H_1 : \beta_2 \neq 0$

El elemento de la diagonal principal de $(X'X)^{-1}$ correspondiente a β_2 es C22 = 0.00000123, de tal forma que el estadístico t es:

$$t_0 = \frac{b_2}{\sqrt{S^2 C_{22}}} = \frac{0.01438}{\sqrt{(10.6239)(0.00000123)}} = 3.98$$

Como $t_{0.025,22} = 2.074$, se rechaza la hipótesis H_0, concluyendo que el regresor de distancia X_2 (distancia), contribuye significativamente al modelo dado que "casos" X_1 también está en el modelo.

3.5 Predicción de nuevas observaciones

El modelo de regresión puede ser usado para predecir observaciones futuras en **y** correspondientes a valores particulares en las variables regresoras, por ejemplo X01, X02, ….., X0k. Si

$$x'_0 = [1, x_{01}, x_{02}, x_{013}, ..., x_{01k}]$$

Entonces una observación futura **y0** en este punto es:

$$\hat{y}_0 = x'_0 \hat{\beta}$$

Un intervalo de de predicción con un nivel de confianza del 100(1-alfa) porciento para una observación futura es:

$$\hat{y}_0 - t_{\alpha/2,n-p}\sqrt{\sigma^2(1+x'_0(X'X)^{-1}x_0} \le y_0 \le \hat{y}_0 + t_{\alpha/2,n-p}\sqrt{\sigma^2(1+x'_0(X'X)^{-1}x_0}$$

Es una generalización del modelo de regresión lineal simple.

Para el caso del ejemplo del embotellador:

El embotellador desea construir un intervalo de predicción sobre el tiempo de entrega para un local requiriendo

X1 = 8 envases y cuya distancia es X2 = 275 pies. Por tanto:

$$X_0 = \begin{bmatrix} 1 \\ 8 \\ 275 \end{bmatrix} \qquad Xo' = [1, 8, 275]$$

El valor de respuesta estimada por la ecuación de ajuste es:

$$\hat{Y}_0 = X'_0 b = [1,\ 8,\ 275]\begin{bmatrix} 2.34123 \\ 1.61591 \\ 0.01438 \end{bmatrix} = 19.22 \min utos$$

$$X'_0(X'X)^{-1}X_0 = 0.05346$$

Por tanto el intervalo de predicción al 95% de nivel de confianza es:

$$19.22 - 2.074\sqrt{10.6239(1+0.05346)} \le Y_0 \le 19.22 + 2.074\sqrt{10.6239(1+0.05346)}$$

Que se reduce al intervalo de predicción de:

$12.28 \le Y_0 \le 26.16$

3.6 Extrapolación oculta

AL predecir la respuesta promedio en un punto X0, se debe tener cuidado de no extrapolar más allá de la región que contiene las observaciones originales, ya que el ajuste puede no ser adecuado en esas regiones.

Para un procedimiento formal, se define el conjunto convexo más pequeño que contiene todos los n puntos originales (Xi1, Xi2, ….., Xik), i=1, 2, 3, ….,n, como la variable regresora envolvente o cáscara (*Regressor Variable Hull – RVH*). Si un punto *X0' = [X01, X02, …, X0k]* se encuentra fuera de la variable RHV entonces se requiere extrapolación. El lugar de ese punto en relación con la RVH se refleja mediante:

$h00 = X0'(X'X)^{-1}X0$

Los puntos *h00 > hmax* están fuera del elipsoide que encierra la RVH y son puntos de extrapolación.

Los elementos diagonales ***hii*** de la matriz sombrero **H = X(X'X)$^{-1}$X'** se utilizan para detectar extrapolación oculta. En general el punto que tiene el mayor valor de ***hii*** o ***hmax*** se encuentra en la frontera de la RVH. El conjunto de puntos X que satisfacen el modelo:

x'(X'X)$^{-1}$x <= hmax

es un elipsoide que engloba todos los puntos dentro de la variable RVH.

Para el caso del ejemplo del embotellador se tiene:

x'

Observación	1	1	1	1	1
X1_envases	7	3	3	4	6
X2_Distancia	560	220	340	80	150

Etc..

(X'X)$^{-1}$

0.1132152	-0.004	-8E-05
-0.0044486	0.0027	-5E-05
-8.367E-05	-5E-05	1E-06

X1'(X'X)$^{-1}$

primero

0.0352184	-0.0120421	0.0003

Segundo

0.0814614	-0.0067458	4E-05

X1'(X'X)$^{-1}$x1

Observación	**X1_envases**	**X2_Distancia**	**hii**
1	7	560	0.10180178
1	3	220	0.07070164

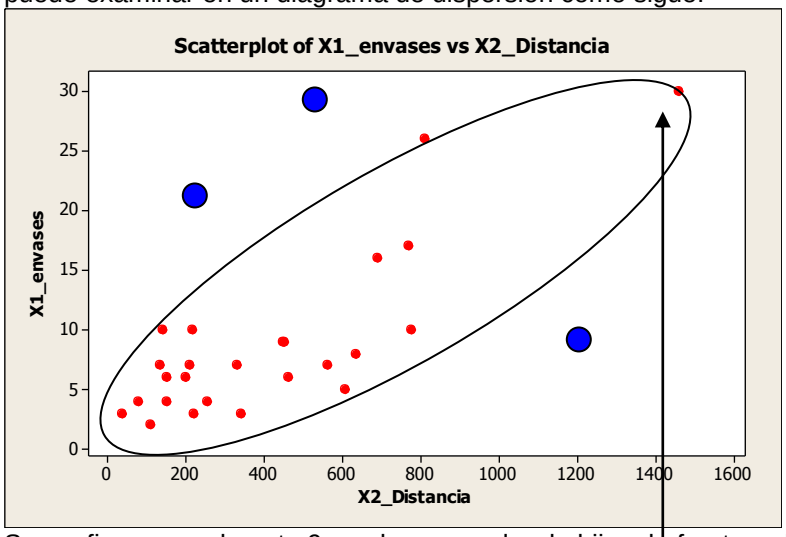
La tabla completa se muestra a continuación:

Observación	**X1_envases**	**X2_Distancia**	**hii**	
1	7	560	0.10180178	
1	3	220	0.07070164	
1	3	340	0.09874	
1	4	80	0.08538	
1	6	150	0.07501	
1	7	330	0.04287	
1	2	110	0.0818	
1	7	210	0.06373	
1	30	1460	0.49829	hmax
1	5	605	0.1963	

1	16	688	0.08613
1	10	215	0.11366
1	4	255	0.06113
1	6	462	0.07824
1	9	448	0.04111
1	10	776	0.16594
1	6	200	0.05943
1	7	132	0.09626
1	3	36	0.09645
1	17	770	0.10169
1	10	140	0.16528
1	26	810	0.39158
1	9	450	0.04126
1	8	635	0.12061
1	4	150	0.06664

Los puntos para los cuales hoo sea mayor a hmax, se encuentran fuera del elipsoide, generalmente entre menor sea el valor de hoo es más probable que se encuentre en el elipsoide.

En la tabla la observación 9 tiene el valor mayor de *hii*. Como el problema solo tiene dos regresores se puede examinar en un diagrama de dispersión como sigue:



Se confirma que el punto 9 es el mayor valor de hii en la frontera de la RHV.
Ahora supongamos que se desea considerar la predicción o estimación para los puntos siguientes:

Punto	x10	x20	h00
a	8	275	0.05346
b	20	250	0.58917
c	28	500	0.89874
d	8	1200	0.86736

Todos los puntos se encuentran dentro del rango de los regresores X1 y X2. El punto a es de interpolación puesto que *hoo <= hmax (0.05346 < 0.49829)* todos los demás son puntos de extrapolación ya que exceden a hmax, lo que se confirma en la gráfica de dispersión.

Inferencia simultanea en la regresión múltiple
Indica que se pueden hacer inferencias en forma simultanea

3.6 Evaluación de la adecuación del modelo

Como se comentó anteriormente, los residuos e_i del modelo de regresión múltiple, juegan un papel importante en la evaluación de la adecuación del modelo, de forma similar que en la regresión lineal simple. Es conveniente graficar los residuos siguientes:

1. Residuos en papel de probabilidad normal.
2. Residuos contra cada uno de los regresores X's.
3. Residuos contra cada $\hat{Y}_i, i = 1,2,...,k$
4. Residuos en secuencia de tiempo (si se conoce)

Estas gráficas se usan para identificar comportamientos anormales, outliers, varianza desigual, y la especificación funcional equivocada para un regresor. Se pueden graficar los residuos sin escalamiento o con un escalamiento apropiado.

Existen algunas técnicas adicionales de análisis de residuos útiles en el análisis de la regresión múltiple, como se describen a continuación.

Gráficas de residuos contra regresores omitidos en el modelo

Estas gráficas podrían revelar cualquier dependencia de la variable de respuesta Y contra los factores omitidos, se esta forma se puede analizar si su incorporación mejora la explicación del modelo.

Gráficas de residuos parciales

Estas gráficas están diseñadas para revelar en forma más precisa la relación entre los residuos y la variable regresora X_j. Se define el residuo parcial i-ésimo para el regresor X_j como sigue:

$$e_{ij}^{*} = e_i + b_j X_{ij}, i = 1,2,....,n \qquad (3.35)$$

La gráfica de e_{ij}^{*} contra X_{ij} se denomina *Gráfica de residuo parcial.* Esta gráfica sirve para detectar Outliers y desigualdad de varianza, dado que muestra la relación entre *Y* y el regresor X_j después de haber removido el efecto de los otros regresores X_i *(I<>j*), es el equivalente de la gráfica de *Y* contra X_j en regresión múltiple.

Gráficas de regresión parcial

Son gráficas de residuos de los cuales se ha removido la dependencia lineal de Y sobre todos los regresores diferentes de X_j, así como su dependencia lineal de otros regresores. En forma matricial se pueden escribir estas cantidades como $e_{Y|X(j)}, e_{X_j|X(j)}$ donde $\mathbf{X}_{(j)}$ es la matriz original **X** con el regresor j-ésimo removido.

del modelo general en forma matricial:

$$Y = X\beta + \varepsilon = X_{(j)}\beta + X_j\beta_j + \varepsilon \qquad (3.36)$$

Premultiplicando por [$I - H_{(j)}$] y notando que $(1 - H_{(j)})X_{(j)} = 0$ se tiene:

$$e_{Y|X(j)} = \beta_j e_{X_j|X(j)} + (1 - H_{(j)})\varepsilon \qquad (3.37)$$

Algunos programas como SAS generan gráficas de regresión parcial. Gráficas de regresores X_i *versus* X_j.

Estas gráficas pueden ser útiles para el análisis de la relación entre los regresores y la disposición de los datos en el espacio X, donde pueden descubrirse puntos remotos del resto de los datos y que

tienen influencia en el modelo. Si se encuentra que las variables regresoras están altamente correlacionadas, puede no ser necesario incluirlas ambas en el modelo. Si dos o más regresores están altamente correlacionados, se dice que hay *multicolinealidad* en los datos, esto distorsiona al modelo.

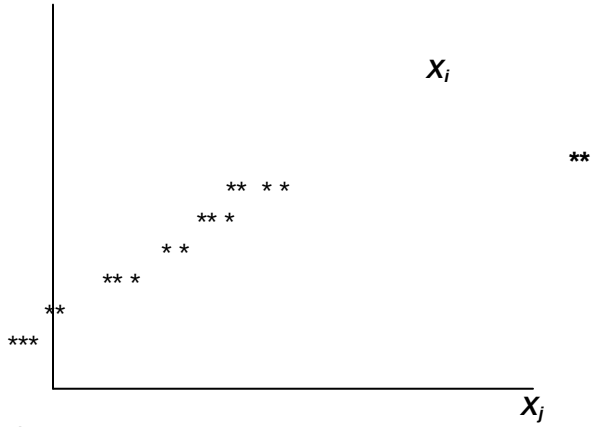


Fig. 3.1 Gráfica de X_i versus X_j

Método de escalamiento de residuos

Es difícil hacer comparaciones directas entre los coeficientes de la regresión debido a que la magnitud de b_j refleja las unidades de medición del regresor X_j. Por ejemplo:

$$\hat{Y} = 5 + X_1 + 1000X_2 \qquad (3.38)$$

Donde Y esta medida en litros, X_1 en mililitros y X_2 en litros. Note que a pesar de que b_2 es mucho mayor que b_1, su efecto en la variable de respuesta es idéntico. Por lo anterior algunas veces es importante trabajar con regresores y variables de respuesta con escala cambiada, de tal forma que produzcan coeficientes de regresión sin dimensiones.

Existen dos técnicas para esto. La primera se denomina *escala unitaria normal,*

$$Z_{ij} = \frac{X_{ij} - X_j}{S_j} \qquad \text{Con i = 1, 2, ..., n;} \quad \text{j = 1, 2, ..., k} \qquad (3.39)$$

$$Y_i^* = \frac{Y_i - Y}{S_y} \qquad \text{Con i = 1, 2, ..., n} \qquad (3.40)$$

De esta forma el modelo de regresión se transforma en:

$$Y_i^* = b_1 Z_{i1} + b_2 Z_{i2} + b_3 Z_{i3} + ... + b_k Z_{ik} + \varepsilon_i \qquad \text{i = 1, 2, ..., n} \qquad (3.41)$$

En este modelo $b_0 = 0$ y el estimador de mínimos cuadrados para $\boldsymbol{b}$ es:

$$b = (Z'Z)^{-1} Z'Y^* \qquad (3.42)$$

El otro método de escalamiento es el *escalamiento de longitud unitaria,*

$$W_{ij} = \frac{X_{ij}}{\sqrt{S_{jj}}}, \qquad \text{i = 1, 2, ..., n;} \quad \text{j = 1, 2, ..., k} \qquad (3.43)$$

$$Y_i^0 = \frac{Y_i - \bar{\bar{Y}}}{\sqrt{S_{YY}}}, \qquad i = 1, 2, \ldots\ldots\ldots, n \qquad (3.44)$$

$$S_{jj} = \sum (X_{ij} - \bar{\bar{X}}_j)^2 \qquad (3.45)$$

Esta última es la suma de cuadrados corregida para el regresor Xj. En este caso cada regresor *Wj* tiene media cero y longitud uno.

$$\left.\begin{aligned} \bar{\bar{W}}_j &= 0 \\ \sqrt{\sum_{i=1}^{n} (W_{ij} - W_j)^2} &= 1 \end{aligned}\right\} \qquad (3.46)$$

En términos de las variables de regresión, el modelo queda como:

$$Y_i^0 = b_1 W_{i1} + b_2 W_{i2} + \ldots\ldots\ldots + b_k W_{ik} + \varepsilon_i, \qquad i = 1, 2, \ldots\ldots, n \qquad (3.47)$$

El vector de mínimos cuadrados de los coeficientes es:

$$b = (W'W)^{-1} W'Y^0 \qquad (3.48)$$

La matriz de correlación W'W en la escala unitaria tiene la forma:

$$W'W = \begin{bmatrix} 1, r_{12}, r_{13} \ldots\ldots r_{1k} \\ r_{12,} 1, r_{23} \ldots\ldots r_{2k} \\ \ldots\ldots\ldots\ldots\ldots \\ r_{1k}, r_{2k}, r_{3k} \ldots . 1 \end{bmatrix}$$

Donde r_{ij} es la correlación simple entre X_i y X_j.

$$r_{ij} = \frac{\sum_{u=1}^{n} (X_{ui} - \bar{\bar{X}}_i)(X_{uj} - \bar{\bar{X}}_j)}{\sqrt{S_{ii} S_{jj}}} = \frac{S_{ij}}{\sqrt{S_{ii} S_{jj}}} \qquad (3.49)$$

De forma similar

$$W'Y^0 = \begin{bmatrix} r_{1Y} \\ r_{2Y} \\ \ldots . \\ r_{KY} \end{bmatrix}$$

Donde r_{jy} es la correlación simple entre el regresor *Xj* y la respuesta *Y*:

$$r_{jy} = \frac{\sum_{u=1}^{n}(X_{uj} - \overline{X}_j)(Y_u - \overline{Y})}{\sqrt{S_{ii}S_{YY}}} = \frac{S_{jY}}{\sqrt{S_{jj}S_{YY}}} \qquad (3.50)$$

Si se utiliza la escala normal unitaria, la matriz **Z'Z** está relacionada con **W'W** como sigue:

$$Z'Z = (n-1)\,W'W \qquad (3.51)$$

Por lo que no importa que método se utilice para escalamiento, ambos métodos producen el mismo conjunto de coeficientes de regresión sin dimensiones **b.**

La relación entre los coeficientes originales y los estandarizados es:

$$b_j = \hat{b}_j\sqrt{\frac{S_{YY}}{S_{JJ}}} \qquad j = 1, 2, ..., k \qquad (3.52)$$

y

$$b_0 = \overline{Y} - \sum_{j=1}^{k} b_j\,\overline{X}_j \qquad (3.53)$$

Si las variables originales difieren mucho en magnitud, los errores de redondeo al calcular X'X pueden ser muy grandes aún utilizando computadora, es por esto que los programas muestran tanto los valores originales como coeficientes de regresión estandarizados (coeficientes Beta). Por tanto se debe tener cuidado de usar éstos últimos para medir la importancia relativa del regresor X_j.

Ejemplo 3.5
Calculando los coeficientes de correlación entre las diferentes variables, se tiene:
Con Minitab:

Stat > Basic statistics > Correlation
Variables Y_tiempo, X1_envases, X2_Distancia
OK

Correlations: Y_tiempo, X1_envases, X2_Distancia

	Y_tiempo	X1_envases
X1_envases	0.965	
	0.000	
X2_Distancia	0.892	0.824
	0.000	0.000

$r_{12} = 0.824215$
$r_{1y} = 0.964615$
$r_{2y} = 0.891670$

La matriz de correlación para este problema W'W es:

$$W'W = \begin{bmatrix} 1.000000, & 0.824215 \\ 0.824215, & 1.000000 \end{bmatrix}$$

Las ecuaciones normales en términos de los coeficientes de la regresión estandarizados son:

$$W'Wb = \begin{bmatrix} 1.000000, & 0.824215 \\ 0.824215, & 1.000000 \end{bmatrix} \begin{bmatrix} \hat{b}_1 \\ \hat{b}_2 \end{bmatrix} = \begin{bmatrix} 0.964615 \\ 0.891670 \end{bmatrix}$$

Por tanto:

$$\begin{bmatrix} \hat{b}_1 \\ \hat{b}_2 \end{bmatrix} = \begin{bmatrix} 3.11841, & -2.57023 \\ -2.57023, & 3.11841 \end{bmatrix} \begin{bmatrix} 0.964615 \\ 0.891670 \end{bmatrix} = \begin{bmatrix} 0.716267 \\ 0.301311 \end{bmatrix}$$

El modelo ajustado es:

$$\hat{Y}^0 = 0.716267W_1 + 0.301311W_2$$

De esta forma incrementando el valor estandarizado de envases W1 en una unidad incrementa la unidad estandarizada de tiempo en 0.7162. Además incrementando el valor estandarizado de la distancia W2 en una unidad, incrementa la respuesta en 0.3013 unidades. Por lo tanto parece ser que el volumen de producto surtido es más relevante que la distancia, con ciertas precauciones dado que los coeficientes b's son sólo coeficientes parciales de regresión.

El coeficiente de determinación R^2 se calcula como sigue:

$$R^2 = \frac{SSR}{SST}$$

$$R^2 = \frac{5550.816}{5784.5426} = 0.9596$$

Por lo anterior el 96% de la variabilidad en tiempo de entrega es explicada por los dos regresores cantidad de surtimiento X_1 y distancia X_2. El índice R2 siempre se incrementa cuando se agrega una nueva variable al modelo de regresión, aunque sea innecesaria.

Un índice más real es el *índice ajustado* $\bar{R}^2$, que penaliza al analista que incluye variables innecesarias en el modelo. Se calcula como sigue:

$$\bar{R}^2 = 1 - \frac{SSE/(N-p)}{SST/(N-1)} = 1 - \frac{N-1}{N-p}(1-R^2)$$

Para el ejemplo se tiene:

$$\bar{R}^2 = 1 - \frac{25-1}{25-3}(1-0.9596) = 0.9559$$

Residuos estandarizados y estudentizados

Los residuos se estandarizan como sigue:

$$d_i = \frac{e_i}{\sqrt{MSE}}, \qquad i = 1, 2, \ldots\ldots\ldots, n \qquad (3.54)$$

Para los residuos estudentizados, utilizamos el vector de residuos:

e = (I – H) Y (3.55)

donde

H = X (X'X)$^{-1}$X' *es la matriz sombrero o "hat matriz".*

Esta matriz tiene las propiedades siguientes:
1. Es simétrica, es decir H' = H.
2. Es idempotente, es decir H H = H.
3. En forma similar la matriz I – H es simétrica e idempotente.

Por tanto se tiene:

$$e = (I - H)(X\beta + \varepsilon) = X\beta - HX\beta + (I - H)\varepsilon = X\beta - X(X'X)^{-1}X'X\beta + (I - H)\varepsilon$$

e = (I – H) ε (3.55)

De esta forma los residuos tienen la misma transformación lineal para las observaciones Y y para los errores ε.

La varianza de los residuos es:

$$Var(e) = \sigma^2(I - H) \qquad (3.56)$$

Como la matriz I – H no es diagonal, los residuos tienen diferentes varianzas y están correlacionados. La varianza del residuo i-ésimo es:

$$V(e_i) = \sigma^2(1 - h_{ii}) \qquad (3.57)$$

Donde h_{ii} es el elemento diagonal i-ésimo de H.

Tomando esta desigualdad de varianza en cuenta, varios autores recomiendan para escalamiento de los residuos, graficar los residuos "estudentizados" siguientes en lugar de e_i (o d_i):

$$r_i = \frac{e_i}{\sqrt{MSE(1 - h_{ii})}}, \qquad i = 1, 2, \ldots\ldots\ldots, n \qquad (3.58)$$

Los residuos estudentizados tienen varianza constante = 1, independientemente de la localización de *Xi*, cuando la forma del modelo es correcto. A pesar de que los residuos estandarizados y los estudentizados proporcionan casi la misma información, como cualquier punto con residuo y ***hii*** grande tiene una influencia potencial en el ajuste de mínimos cuadrados, se recomienda el análisis de los residuos estudentizados.

La covarianza entre e_i y e_j es:

$$Cov(e_i, e_j) = -\sigma^2 h_{ij} \qquad (3.59)$$

De tal forma que otra forma de escalamiento de residuos es transformar los residuos *n* dependientes en *n-p* funciones ortogonales de los errores ε.

Residuos *PRESS* – *Suma de cuadrados del error de predicción*

La suma de cuadrados del error de predicción (PRESS) propuesto por Allen (1971) proporciona un escalamiento útil para los residuos. Para calcular PRESS, seleccione una observación, por ejemplo *(i),* Ajuste el modelo de regresión a las observaciones remanentes *(N – 1),* usando la ecuación para predecir la observación retenida (Y_i). Denotando el error de predicción como:

$$e_{(i)} = Y_i - \hat{Y}_{(i)} \qquad (3.60)$$

El error de predicción es normalmente denominado el residuo i-ésimo PRESS, el procedimiento se repite para cada una de las observaciones i = 1, 2, ..., N, produciendo los residuos PRESS correspondientes. Así el estadístico PRESS se define como la suma de cuadrados de los N residuos PRESS, como:

$$PRESS = \sum_{i=1}^{N} e_{(i)}^2 = \sum \left[Y_i - \hat{Y}_{(i)} \right]^2 \qquad (3.61)$$

Así PRESS utiliza cada uno de los posibles subconjuntos de *N – 1* observaciones como el conjunto de datos de estimación, y cada observación en turno es usada para formar el conjunto de datos de predicción.

Como:

$$e_{(i)} = \frac{e_i}{1 - h_{ii}} \qquad (3.62)$$

Entonces:

$$\text{PRESS} = \sum_{i=1}^{N} \left(\frac{e_i}{1 - h_{ii}} \right)^2 \qquad (3.63)$$

De esta forma se observa que los residuos asociados con valores altos de h_{ii} serán puntos de alta influencia, donde si se excluyen mostrarán un ajuste pobre del modelo.

La varianza del residuo i-ésimo PRESS es:

$$Var(e_{(i)}) = \frac{\sigma^2}{1 - h_{ii}} \qquad (3.64)$$

Y el residuo PRESS estandarizado es:

$$\frac{e_{(i)}}{\sqrt{V(e_{(i)})}} = \frac{e_i}{\sqrt{\sigma^2 (1 - h_{ii})}} \qquad (3.65)$$

Donde si utilizamos MSE para estimar la varianza σ^2 se convierte en el *residuo estudentizado* discutido previamente.

R- STUDENT

Otro método para diagnosticar la presencia de outliers o puntos de alta influencia es el residuo estudentizado R – Student donde la estimación de la varianza se hace excluyendo la j-ésima observación, como sigue:

$$S^2_{(i)} = \frac{(N-p)MSE - e_i^2/(1-h_{ii})}{n-p-1} \qquad \text{i = 1, 2, ..., n} \qquad (3.66)$$

y el residuo estudentizado externamente R – Student, está dado por:

$$t_i = \frac{e_{(i)}}{\sqrt{S^2_{(i)}(1-h_{ii})}}, \qquad \text{i = 1, 2, ..., n} \qquad (3.67)$$

En muchas situaciones este residuo puede diferir del residuo estudentizado r_i . Si la observación i-ésima tiene influencia, entonces $S^2_{(i)} \neq MSE$ y el estadístico R-student será más sensible a este punto. También ofrece una prueba más formal de prueba de hipótesis de outliers, ya que se puede comparar todos los *n* valores de $|t_i|.versus.t_{(\alpha/2n),n-p-1}$.

El estadístico PRESS puede usarse para calcular una R^2 aproximada para predicción, o sea:

$$R^2_{\text{Predicción}} = 1 - \frac{PRESS}{S_{YY}} \qquad (3.68)$$

Para el ejemplo de las bebidas se tiene:[2]

$$R^2_{\text{Predicción}} = 1 - \frac{457.4}{5784.5426} = 0.9209$$

Por lo que esperaríamos que este modelo explicara aproximadamente el 92% de la variabilidad al predecir nuevas observaciones, que se compara con el 95.96% de la variabilidad en los datos originales explicados por el ajuste de mínimos cuadrados.

Tabla de residuos

									R Student
hii	**Y_tiempo**	**Fits =Yest**	**ei = Y - Yest**	**di=ei/Sigma**	**ri=ei/raiz(MSE(1-hii))**	**e(i)=ei/(1-hii)**	**S(i)^2**	**[ei/(1-hii)^2)**	**ti**
0.10180	16.68	21.7081	-5.0281	-1.5426	-1.6277	-5.5980	9.7897	31.3372	-1.8878
0.07070	11.5	10.3536	1.1464	0.3517	0.3648	1.2336	11.0627	1.5218	0.3847
0.09874	12.03	12.0798	-0.0498	-0.0153	-0.0161	-0.0552	11.1299	0.0031	-0.0174
0.08538	14.88	9.9556	4.9244	1.5108	1.5797	5.3840	9.8676	28.9879	1.7922
0.07501	13.75	14.1944	-0.4444	-0.1363	-0.1418	-0.4804	11.1199	0.2308	-0.1498
0.04287	18.11	18.3996	-0.2896	-0.0888	-0.0908	-0.3025	11.1259	0.0915	-0.0927
0.0818	8	7.1554	0.8446	0.2591	0.2704	0.9199	11.0931	0.8462	0.2882
0.06373	17.83	16.6734	1.1566	0.3548	0.3667	1.2353	11.0620	1.5260	0.3839
0.49829	79.24	71.8203	7.4197	2.2764	3.2138	14.7888	5.9049	218.7096	8.5921
0.1963	21.5	19.1236	2.3764	0.7291	0.8133	2.9568	10.7955	8.7429	1.0038
0.08613	40.33	38.0925	2.2375	0.6865	0.7181	2.4484	10.8692	5.9945	0.7768
0.11366	21	21.5930	-0.5930	-0.1819	-0.1933	-0.6691	11.1112	0.4477	-0.2132
0.06113	13.5	12.4730	1.0270	0.3151	0.3252	1.0939	11.0766	1.1966	0.3392

[2] Montgomery, Douglas C., Peck, Elizabeth A., *Introduction to Linear Regression Analysis,* 2º edition, John Wiley and Sons, Nueva York, 1991, p. 176

0.07824	19.75	18.6825	1.0675	0.3275	0.3411	1.1581	11.0712	1.3413	0.3625
0.04111	24	23.3288	0.6712	0.2059	0.2103	0.7000	11.1077	0.4900	0.2145
0.16594	29	29.6629	-0.6629	-0.2034	-0.2227	-0.7948	11.1050	0.6317	-0.2612
0.05943	15.35	14.9136	0.4364	0.1339	0.1380	0.4639	11.1204	0.2152	0.1434
0.09626	19	15.5514	3.4486	1.0580	1.1130	3.8159	10.5034	14.5614	1.2386
0.09645	9.5	7.7068	1.7932	0.5501	0.5788	1.9846	10.9606	3.9387	0.6306
0.10169	35.1	40.8880	-5.7880	-1.7757	-1.8736	-6.4432	9.3542	41.5145	-2.2227
0.16528	17.9	20.5142	-2.6142	-0.8020	-0.8778	-3.1318	10.7402	9.8082	-1.0460
0.39158	52.32	56.0065	-3.6865	-1.1310	-1.4500	-6.0592	10.0664	36.7137	-2.4484
0.04126	18.75	23.3576	-4.6076	-1.4136	-1.4437	-4.8059	10.0756	23.0963	-1.5463
0.12061	19.83	24.4029	-4.5729	-1.4029	-1.4961	-5.2000	9.9977	27.0403	-1.7537
0.06664	10.75	10.9626	-0.2126	-0.0652	-0.0675	-0.2278	11.1278	0.0519	-0.0707
							PRESS	**459.03907**	

3.7 Estimación del error puro a partir de vecinos cercanos

Para la regresión lineal, la suma de cuadrados del error puro SS_{PE} se calcula utilizando respuestas replicadas en el mismo nivel de X. La suma de cuadrados del error o residual se parte en un componente debido al error "puro" y un componente debido a la falta de ajuste o sea:

$SSE = SS_{PE} + SS_{LOF}$

Esto mismo podría extenderse a la regresión múltiple, donde el cálculo de SS_{PE} requiere observaciones replicadas en Y con el mismo nivel de las variables regresoras X_1, X_2, ..., X_k, o sea que algunas de las filas de la matriz X deben ser las mismas. Sin embargo estas condiciones repetidas no son comunes y este método es poco usado.

Daniel y Wood han sugerido un método para obtener un estimado del error independiente del modelo donde no hay puntos repetidos exactos. El procedimiento busca puntos en el espacio X que son "vecinos cercanos" es decir observaciones que se han tomado con niveles cercanos de X_{i1}, X_{i2}, ..., X_{ik}. Las respuestas Y_i de tales "vecinos cercanos" pueden ser consideradas como réplicas a usar para el cálculo del error puro. Como una medida de la distancia entre dos puntos X_{i1}, X_{i2}, ..., X_{ik} y X_{j1}, X_{j2}, ..., X_{jk} proponen el estadístico de suma de cuadrados ponderados de la distancia como:

$$D_{ii}^2 = \sum_{j=1}^{k}\left[\frac{b_j(X_{ij} - X_{i'j})}{\sqrt{MSE}}\right]^2 \qquad (3.69)$$

Los pares de puntos que tienen esta distancia pequeña son vecinos cercanos sobre los cuales se puede calcular el error puro, y los que generan $D_{ii}^2 >> 1$ están ampliamente separados en el espacio X.

El estimado del error puro se obtiene del rango de los residuos en el punto *i* e *i'*, como sigue:

$$E_i = |e_i - e_{i'}| \qquad (3.70)$$

Hay una relación entre el el rango de una muestra de una distribución normal y la desviación estándar de la población. Para muestras de tamaño 2, la relación es:

$$\hat{\sigma} = \frac{R}{d_2} = \frac{E}{1.128} = 0.886E$$

Esta desviación estándar corresponde al error puro.

Un algoritmo para calcular la desviación estándar estimada es como sigue:

1. Arreglar los conjuntos de datos de puntos X's en orden ascendente de Yi-est.
2. Calcular los valores de D_{ii}^2, para todos los N-1 pares de puntos con valores adyacentes de Y-est. Repetir el procedimiento para los pares de puntos separados por uno, dos o tres valores intermedios de Y-est. Lo cual producirá (4 N – 10) valores de D_{ii}^2.
4. Arreglar los (4 N –10) valores de D_{ii}^2 en orden ascendente. Sea Eu, u = 1, 2,..., 4N-10, sea el rango de los residuos en esos puntos.
5. Para los primeros m valores de Eu, calcular un estimado de la desviación estándar del error puro como:

$$\hat{\sigma} = \frac{0.886}{m}\sum_{u=1}^{m} E_u$$

No se deben incluir Eu para los cuales la suma de las distancias cuadradas ponderadas sea muy grande.

Ejemplo 3.6 La tabla 4.9 muestra el cálculo de D_{ii}^2 para pares de puntos que en términos de $\hat{Y}$ son adyacentes, en uno, dos y tres puntos. Las columnas R en la tabla identifican a los 15 valores más pequeños de D_{ii}^2.

Observ	Fits =Yest	ei = Y - Yest	X1	X2	Delta	D2ii	Ra	Delta	D2ii	Ra	Delta	D2ii	R	Delta	
7	7.1554	0.8446	2	110	0.9486	0.3524271		4.0798	1.0006243		0.3018	0.48143932		1.0572	1.0
19	7.7068	1.7932	3	36	3.1312	0.28348034	12	0.6468	0.6593958		2.0058	0.49889025		1.843	1.7
4	9.9556	4.9244	4	80	3.778	0.62751294		5.137	0.0954348	3	4.9742	1.56238413		3.8974	0.
2	10.3536	1.1464	3	220	1.359	0.34120864	15	1.1962	0.2804614	11	0.1194	0.26963257	9	1.5908	2.3
25	10.9626	-0.2126	4	150	0.1628	0.94887491		1.2396	0.2147282	6	0.2318	0.98309549		0.649	1.
3	12.0798	-0.0498	3	340	1.0768	0.38649146		0.3946	2.9150659		0.4862	2.59370393		3.4984	4.7
13	12.473	1.027	4	255	1.4714	1.19782372		0.5906	1.0420119		2.4216	2.50662458		0.1296	2.2
5	14.1944	-0.4444	6	150	0.8808	0.04869121	2	3.893	0.2520843	8	1.601	0.31588921	13	0.1548	0.8
17	14.9136	0.4364	6	200	3.0122	0.33583313	14	0.7202	0.2477215	7	0.726	0.57492644		0.6311	1.3
18	15.5514	3.4486	7	132	2.292	0.11849492	5	3.7382	0.763556		2.3811	2.36676288		1.0722	5.3
8	16.6734	1.1566	7	210	1.4462	0.28046136	10	0.0891	1.4826085		1.2198	4.02191377		3.7708	2.3
6	18.3996	-0.2896	7	330	1.3571	0.58513212		2.666	2.4560045		2.3246	2.9150659		0.3034	2.4
14	18.6825	1.0675	6	462	1.3089	0.64404848		3.6817	5.9517817		1.6605	5.12062274		6.0956	0.4
10	19.1236	2.3764	5	605	4.9906	10.3556494		2.9694	9.1067199		7.4045	1.02253537		1.7052	4.4
21	20.5142	-2.6142	10	140	2.0212	0.10955522	4	2.4139	5.6476165		3.2854	2.09339097		1.9934	2.
12	21.593	-0.593	10	215	4.4351	4.53015326		1.2642	1.3031327		4.0146	1.32136265		3.9799	4.4
1	21.7081	-5.0281	7	560	5.6993	1.2274085		0.4205	1.2187609		0.4552	0.35532909		4.3652	3.1
15	23.3288	0.6712	9	448	5.2788	7.7906E-05	1	5.2441	0.926847		1.3341	2.34113183		1.5663	13
23	23.3576	-4.6076	9	450	0.0347	0.91235651		3.9447	2.3156566		6.8451	13.1461457		1.1804	17
24	24.4029	-4.5729	8	635	3.91	1.37030746		6.8104	15.784237		1.2151	20.2626427		0.8864	80
16	29.6629	-0.6629	10	776	2.9004	8.99868534		5.1251	12.043621		3.0236	62.9406265		8.0826	10
11	38.0925	2.2375	16	688	8.0255	0.37673375		5.924	24.867275		5.1822	59.7793515			
20	40.888	-5.788	17	770	2.1015	19.9388461		13.2077	50.808538						
22	56.0065	-3.6865	26	810	11.106	12.1611961									
9	71.8203	7.4197	30	1460											

Los 15 pares de puntos se usan para estimar σ = 1.969. Sin embargo de una tabla anterior se había calcualdo $\sqrt{MSE} = \sqrt{10.6239} = 3.259.$ Por otro lado no se observa falta de ajuste y esperaríamos haber encontrado que $\hat{\sigma} = \sqrt{MSE}.$ Sin embargo en este caso $\sqrt{MSE}$ es sólo del 65% mayor que $\hat{\sigma}$, indicando una cierta falta de ajuste, lo cual puede ser debido a el efecto de regresores no presentes en el modelo o la presencia de uno o más outliers.

Determinación de la Desviación estándar

Núm.	**Observ**	**D2ii**	**Delta**	**Sigma acum**
1	15-23	7.7906E-05	5.2788	4.6770168
2	5-17	0.04869121	0.8808	2.7287028
3	4-25	0.09543477	5.137	3.336262533
4	21-12	0.10955522	2.0212	2.9498927
5	18-8	0.11849492		
6		0.21472823		
7		0.24772152		
8		0.25208425		
9		0.26963257		
10		0.28046136		
11		0.28046136		
12		0.28348034		
13		0.31588921		
14		0.33583313		
15		0.34120864		
16		0.3524271		
17		0.35532909		
18		0.37673375		
19		0.38649146		
20		0.43282602		
21		0.48143932		
22		0.49889025		
23		0.57492644		
24		0.58513212		
25		0.5964673		
26		0.62751294		
27		0.64404848		
28		0.65939581		
29		0.76355604		
30		0.87681193		
31		0.91235651		
32		0.92684701		
33		0.94887491		
34		0.98309549		
35		1.00062433		
36		1.01425787		

37		1.02253537		
38		1.0317867		
39		1.04201186	0.5907	1.983
40		1.19782372	1.4714	1.966

Desviación estándar

Diagnóstico de influyentes

A veces un pequeño grupo de puntos ejerce una influencia desproporcionada en el modelo de regresión, se deben revisar con cuidado, si son valores "mal" tomados, se deben eliminar, de otra forma se debe estudiar el porqué de su ocurrencia.

Puntos influyentes

Son observaciones remotas que tienen un apalancamiento desproporcionado potencial en los parámetros estimados, valores de predicción, y estadísticas en general.

Hoaglin y Welsch discuten el papel de la matriz sombrero **H** donde sus elementos de la diagonal principal *(h_{ij})* puede ser interpretado como la cantidad de influencia ejercida por Y_j en $\hat{Y}_i$. Así, enfocando la atención en los elementos de la diagonal de la matriz **H**, como $\sum_{i=1}^{n} h_{ii} = rango(H) = rango(X) = p$, el tamaño medio de un elemento en la diagonal principal es *p/n.* Por tanto si un elemento de la diagonal principal $h_{ii} > 2p/n$, la observación (*i*) es un punto con apalancamiento alto.

Medidas de influencia: la D de Cook

Cook sugirió un ***diagnóstico de eliminación,*** es decir, mide la infuencia de la pésima observación si se eliminara de la muestra. Sugiere medir la distancia cuadrada entre el estimado de mínimos cuadrados basado en todos los *n* puntos ***b*** y el estimado obtenido al borrar el i-ésimo punto $\boldsymbol{b_{(i)}}$, esta distancia se expresa como:

$$D_i(M,c) = \frac{(b_{(i)} - b)' M (b_{(i)} - b)}{c}, i = 1,2,.....,n \qquad (3.71)$$

Donde M = X'X y c = pMSe, obteniéndose:

$$D_i(M,c) = \frac{(b_{(i)} - b)' X' X (b_{(i)} - b)}{pMSe}, i = 1,2,.....,n \qquad (3.72)$$

Los puntos con valores grandes de *Di* tienen una influencia considerable en los estimadores de mínimos cuadrados ***b***. La magnitud de Di puede evaluarse comparándola con $F_{\alpha,p,n-p}$. Si $D_i \approx F_{.5,p,n-p}$, entonces al borrar el punto *i* moverá a ***b*** al límite del intervalo de confianza del 50% para β con base en el conjunto de datos completo. Como $F_{.5,p,n-p} \approx 1$ normalmente se considera que los puntos donde $D_i > 1$ tendrán influencia. Idealmente cada $b_{(i)}$ deberá permanecer dentro de la banda del 10 a 20% de la región de confianza.

Otra forma de escribir el estadístico Di es:

$$D_i = \frac{r_i^2}{p} \frac{V(\hat{Y}_i)}{V(e_i)} = \frac{r_i^2}{p} \frac{h_{ii}}{(1 - h_{ii})}, - - - i = 1,2,.....,n \qquad (3.73)$$

Así D_i está formado por un componente que refleja que tan bien se ajusta el modelo a la i-ésima observación Y_i y un componente que mide que tan lejos se encuentra el punto del resto de los datos. Uno o ambos componentes pueden contribuir a un valor grande de D_i .

Por ejemplo para el caso de tiempos de entrega para la primera observación se tiene:

$$D_i = \frac{r_i^2}{p}\frac{h_{ii}}{(1-h_{ii})} = \frac{-1.6277^2}{3}\frac{0.1018}{(1-0.1018)} = 0.10009$$

En la tabla mostrada abajo el valor máximo de Di = D9 = 3.41835, indicando que el punto 9 tiene una alta influencia en el estimado de los coeficientes Beta, se consideran como influyentes los puntos mayores a 1. También es la distancia euclidiana al cuadrado que se mueve el vector de los valores estimados cuando elimina la i-ésima observación.

Influencia en los valores estimados (DFFITS) y en los parámetros estimados (DFBETAS)
También se puede investigar la influencia de la observación i-ésima en la predicción de un valor. Un diagnóstico razonable es:

$$DFFITS_i = \frac{\hat{Y}_i - \hat{Y}_{(i)}}{\sqrt{S_{(i)}^2 h_{ii}}}, i = 1,2,.....,n \qquad (3.74)$$

Donde $\hat{Y}_{(i)}$ es el valor estimado de Y_i obtenido sin el uso de la iésima observación, el denominador es una estandartización, por tanto DFFITS es el número de desviaciones estándar que el valor estimado $\hat{Y}_i$ *cambia* si la observación i-ésima es removida. Computacionalmente se tiene:

$$DFFITS_i = \left(\frac{h_{ii}}{1-h_{ii}}\right)^{1/2} t_i \qquad (3.75)$$

Donde *ti* es la R-student.

Por lo general merece atención cualquier observación donde

$$\left|DFFITS_i\right| > 2\sqrt{p/n} \qquad (3.76)$$

Para el caso de DFBETAS, indica cuánto cambia el coeficiente de regresión Beta(j) en unidades de desviación estándar, si se omitiera la *i-ésima observación*.

$$DFBETAS_{j,i} = \frac{\hat{\beta}_j - \hat{\beta}_{j(i)}}{\sqrt{S_{(i)}^2 C_{jj}}}$$

C_{jj} es el j-ésimo elemento diagonal de la matriz $(X'X)^{-1}$

$\hat{\beta}_j$ es el j-ésimo coeficiente de regresión, calculado sin usar la i-ésima observación. Un valor grande de DFBETAS indica que la i-ésima observación tiene grana influencia sobre el j-ésimo coeficiente de regresión.

De R = $(X'X)^{-1}X'$, los n elementos del renglón k-ésimo de R producen el balanceo que las n observaciones de la muestra tienen sobre Beta. Si **r'**j es el j-ésimo renglón de **R**, se tiene:

$$DFBETAS_{j,i} = \frac{r_{j,i}}{\sqrt{r_j' r_j}} \frac{t_i}{\sqrt{1-h_{ii}}}$$

Ejemplo de cálculo:

Renglón	**R = (X'X)-1X'**				n elementos						
j=1	0.035217	0.081461	0.07142	0.088726	0.073971	0.054461	0.095113	0.064501	-0.14241	0.04035	-0.0155
j=2	-0.01204	-0.00675	-0.01249	0.002698	0.004835	-0.00104	-0.00423	0.004707	0.00799	-0.01968	0.00652
j=3	0.000269	4.3E-05	0.00019	-0.00018	-0.00019	-1.3E-05	-4.4E-05	-0.00016	0.000274	0.00042	-4.5E-0
	1	2	3	4	5	6	7	8	9	10	1

0.050736	0.074083	0.047866	0.03569	0.003797	0.069787	0.071028	0.096856	-0.02684	0.057011	-0.07023	0.035523	0
0.0127	-0.00568	-0.0101	-0.0012	-0.01415	0.002442	0.00844	0.00206	0.005344	0.016289	0.028124	-0.00129	-
-0.0003	3.81E-05	0.000197	3.58E-05	0.000391	-0.00013	-0.00026	-0.00018	4.84E-05	-0.00039	-0.00033	3.83E-05	0
12	13	14	15	16	17	18	19	20	21	22	23	

R'

0.03522	-0.012	0.00027
0.08146	-0.0067	4.3E-05
0.07142	-0.0125	0.00019
0.08873	0.0027	-0.0002
0.07397	0.00484	-0.0002
0.05446	-0.001	-1E-05
0.09511	-0.0042	-4E-05
0.0645	0.00471	-0.0002
-0.1424	0.00799	0.00027
0.04035	-0.0197	0.00042
-0.0155	0.00652	-5E-06
0.05074	0.0127	-0.0003
0.07408	-0.0057	3.8E-05
0.04787	-0.0101	0.0002
0.03569	-0.0012	3.6E-05
0.0038	-0.0141	0.00039
0.06979	0.00244	-0.0001
0.07103	0.00844	-0.0003
0.09686	0.00206	-0.0002
-0.0268	0.00534	4.8E-05
0.05701	0.01629	-0.0004
-0.0702	0.02812	-0.0003
0.03552	-0.0013	3.8E-05
0.02449	-0.0129	0.00031
0.08287	-0.0007	-9E-05

C

0.11322	-0.0044	-8E-05
-0.0044	0.00274	-5E-05
-8E-05	-5E-05	1.2E-06

$$D_i = \frac{r_i^2}{p}\frac{V(\hat{Y}_i)}{V(e_i)} = \frac{r_i^2}{p}\frac{h_{ii}}{(1-h_{ii})}, - - - i = 1,2,.....,n$$

Atender Di > 1

$$DFFITS_i = \left(\frac{h_{ii}}{1-h_{ii}}\right)^{1/2} t_i$$

Atender DFFITS > 2*raiz(p/n)

0.69282032

$$DFBETAS_{j,i} = \frac{r_{j,i}}{\sqrt{r_j r_j}}\frac{t_i}{\sqrt{1-h_{ii}}}$$

Atneder DFBETAS > 2/raiz(n)

0.4

-1.991908828

Calculo de Bo,i	
r(0,1) =	0.035217
raiz(Cjj)	0.3364746
t1 =	-1.8878
raiz(1-h11) =	0.9477341
-0.20848235	

r(0,2)	0.0814608
raiz(Cjj) =	0.3364746
t2 =	0.3847
raiz(1-hii) =	0.9640021
0.09661409	

r(0,3) =	0.0714204
raiz(Cjj) =	0.3364746
t3 =	-0.0174
raiz(1-hii) =	0.9493471
-0.0038904	

Tabla 6.1 Esatdísticas para detectar observaciones influyentes

MSE =

	(a)		R Student	(b)	(c)	(d)	(e)	(f)	
Observación	hii	ri=ei/raiz(MSE(1-hii))	ti	Distancia COOK Di	DFFITS	DFBETTAS (0),i	DFBETTAS (1),i	DFBETTAS (2),i	S(i)^2
1	0.1018	-1.6277	-1.8878	0.10009265	-0.63554067	**-0.208482352**			9.7897
2	0.0707	0.3648	0.3847	0.00337483	0.10610942	0.096614091			11.0627
3	0.09874	-0.0161	-0.0174	9.4662E-06	-0.00575931	-0.003890398			11.1299
4	0.08538	1.5797	1.7922	0.07765035	0.54757574				9.8676
5	0.07501	-0.1418	-0.1498	0.00054352	-0.04265823				11.1199
6	0.04287	-0.0908	-0.0927	0.00012309	-0.01961874				11.1259
7	0.0818	0.2704	0.2882	0.00217124	0.0860205				11.0931
8	0.06373	0.3667	0.3839	0.00305101	0.10015889				11.062
9	**0.49829**	3.2138	8.5921	3.41936807	8.56276509				5.9049
10	0.1963	0.8133	1.0038	0.05385259	0.49608987				10.7955
11	0.08613	0.7181	0.7768	0.01620013	0.23847575				10.8692
12	0.11366	-0.1933	-0.2132	0.00159716	-0.0763468				11.1112
13	0.06113	0.3252	0.3392	0.00229524	0.08655264				11.0766
14	0.07824	0.3411	0.3625	0.00329195	0.10561206				11.0712
15	0.04111	0.2103	0.2145	0.00063203	0.04441367				11.1077
16	0.16594	-0.2227	-0.2612	0.00328907	-0.11650648				11.105
17	0.05943	0.138	0.1434	0.0004011	0.03604595				11.1204
18	0.09626	1.113	1.2386	0.04398164	0.40423345				10.5034
19	0.09645	0.5788	0.6306	0.01192026	0.2060293				10.9606
20	0.10169	-1.8736	-2.2227	0.13245993	-0.74783684				9.3542
21	0.16528	-0.8778	-1.046	0.05085684	-0.46544828				10.7402
22	0.39158	-1.45	-2.4484	0.45105736	-1.9642234				10.0664
23	0.04126	-1.4437	-1.5463	0.0298993	-0.32078049				10.0756
24	0.12061	-1.4961	-1.7537	0.10232972	-0.64946567				9.9977
25	0.06664	-0.0675	-0.0707	0.00010844	-0.01889132				11.1278

De acuerdo a los puntos de corte de DFFITS de 0.69, los puntos 9 y 22 excend este valor por lo que se consideran influyentes.
Con base en el punto de corte de DFBETAS de 0.4, los puntos 9 y 22 tienen efectos grandes sobre los tres parámetros. La eliminación del punto 9 da como resultado que la respuesta estimada se desplace en en más de cuatro desviaciones estándar.

Medida de desempeño del modelo
Como medida escalar de la *precisión general de la estimación*, se usa el determinante de la matriz de covarianza, denominada *varianza generalizada*, para expresar el papel de la i-ésima observación en la

estimación de la precisión de la estimación, se define la relación de covarianzas (COVRATIOi) como sigue:

$$COVRATIO=\frac{(S^2_{(i)})^p}{MS^p_{\text{Res}}}\left(\frac{1}{h_{ii}}\right)$$

Notar que [1/(1-hii)] es la relación de $\left|(X'_{(i)}X_{(i)})^{-1}\right| / \left|(X'X)^{-1}\right|$, por lo que un punto de alto balanceo hará que COVRATIOi, sea grande.

Si $COVRATIO_i > 1+3p/n \;\; o$ $COVRATIO_i < 1-3p/n$ se debería considerar el i-ésimo punto como influyente.

Ejemplo:
En el caso de los refrescos: el corte para COVRATIOi es 1+-3*3/25 o sea (0.64, 1.66), se puede observar de la tabla que se salen los puntos 9 y apenas el 22.

Multicolinealidad
La multicolinealidad implica una dependencia cercana entre regresores (columnas de la matriz X), de tal forma que si hay una dependencia lineal exacta hará que la matriz X'X se singular. La presencia de dependencias cercanamente lineales impactan dramáticamente en la habilidad para estimar los coeficientes de regresión.

La varianza de los coeficientes de la regresión son inflados debido a la multicolinealidad. Esta es evidente por los valores diferentes de cero que no estan en la diagonal principal de X'X. Los cuales se denominan correlaciones simples entre los regresores. La multicolinealidad puede afectar seriamente la precisión con la cual los coeficientes de regresión son estimados.

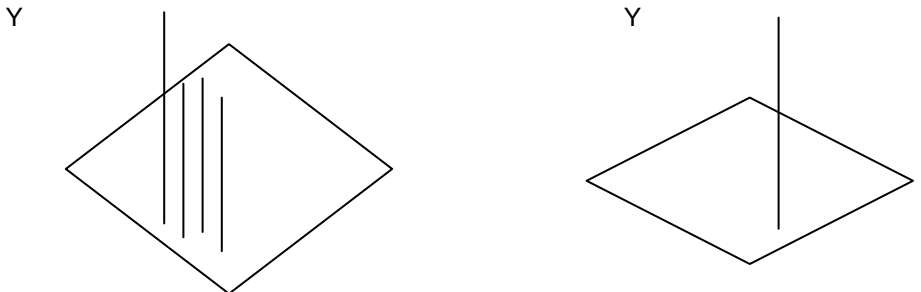
Entre las fuentes de colinealidad se encuantran:
- El método de recolección de datos empleado.
- Restricciones en el modelo o en la población.
- Especificación del modelo.
- Un modelo sobredefinido.

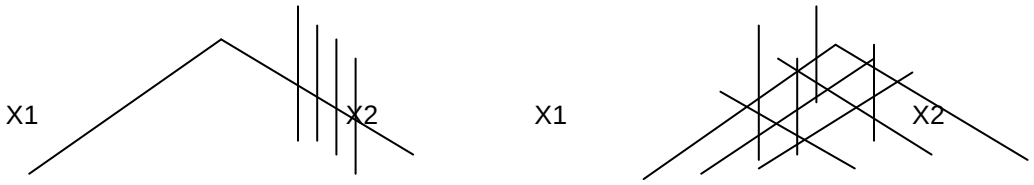
Los elementos de la diagonal principal de la matriz X'X se denominan *Factores de inflación de varianza (VIFs)* y se usan como un diagnóstico importante de multicolinealidad. El factor para el coeficiente j-ésimo coeficiente de regresión es:

$$VIF_j = \frac{1}{1-R_j^2} \qquad (3.77)$$

R_j^2 es el coeficiente de determinación múltiple obtenido al hacer una regresión de Xj con con todos los demás regresores. Si X_j es casi linealmente dependiente de algunos de los otros regresores, entonces el coeficiente de determinación R_j^2 será cercano a la unidad y el VIF_j será muy grande, de tal forma que si es mayor a 10 implica que se tienen serios problemas de multicolinealidad.

Los modelos de regresión que tienen presente multicolinealidad muestran ecuaciones de predicción pobres y los coeficientes de regresión son muy sensibles a los datos en la muestra colectada en particular. En comparación con el caso de regresores *ortogonales* que son muy estables (imaginar un plano encima).





a) Datos con multicolinealidad (muy inestable) b) Regresores ortogonales (muy estable)

Fig. 3.2 Efectos de la colinealidad en la estabilidad del sistema

En la figura anterior, un sistema ortogonal se obtiene de los datos siguientes:

X_1	X_2
5	20
10	20
5	30
10	30
5	20
10	20
5	30
10	30

Asumiendo que se utiliza el escalamiento unitario para los coeficientes de regresión, se obtiene:

$$X'X = \begin{bmatrix} 1{,}0 \\ 0{,}1 \end{bmatrix} = (X'X)^{-1}$$

Las varianzas de los coeficientes estandarizados de regresión b_1, b_2 son:

$$\frac{V(b_1)}{\sigma^2} = \frac{V(b_2)}{\sigma^2} = 1$$

Y un sistema con colinealidad es:

$$W'W = \begin{bmatrix} 1.00000{,}0.824215 \\ 0.824215{,}1.00000 \end{bmatrix} \qquad \text{donde} \quad (W'W)^{-1} = \begin{bmatrix} 3.11841{,}-2.57023 \\ -2.57023{,}3.11841 \end{bmatrix}$$

Las varianzas de los coeficientes estandarizados de regresión b_1, b_2 son:

$$\frac{V(b_1)}{\sigma^2} = \frac{V(b_2)}{\sigma^2} = 3.11841$$

Se observa que están infladas debido a la multicolinealidad.

4. MODELOS DE REGRESIÓN POLINOMIAL

4.1Introducción

El modelo de regresión lineal en forma matricial $Y = \beta X + \varepsilon$ es un modelo general para estimar cualquier relación que sea lineal en los parámetros desconocidos β. Esto incluye a los modelos de regresión polinomial de segundo orden en una variable y en dos variables. Los cuales son

ampliamante utilizados en situaciones donde la respuesta es curvilinea o muy compleja, pero que puede ser modelada por polinomios en una región con pequeños rangos en las X's.

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_{11} X_1^2 + \beta_{22} X_2^2 + \beta_{12} X_1 X_2 + \varepsilon$$

4.2. Modelos polinomiales en una variable

El modelo denominado cuadrático es el siguiente:

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \varepsilon$$

Normalmente se denomina a β_1 el parámetro del efecto lineal y β_2 el parámetro del efecto cuadrático. Como regla general el uso de polinomios de más alto orden debe evitarse a menos que no haya otra alternativa.

5. REGRESIÓN MÚLTIPLE POR PASOS (Stepwise)
Introducción
El análisis de regresión es usado para investigar y modelar las relaciones entre una variable de respuesta y uno o más predictores. Minitab proporciona mínimos cuadrados, mínimos cuadrados parciales, y procedimientos de regresión logística.

- Usar mínimos cuadrados cuando la variable de respuesta sea continua.

- Usar procedimientos de mínimos cuadrados cuando los predictores sean altamente correlacionados o excedan al número de observaciones.

- Usar regresión logística cuando la variable de respuesta sea categórica.

Tanto el método de regresión por mínimos cuadrados como la regresión logística estiman parámetros en el modelo de manera que se optimice su ajuste.

La regresión por mínimos cuadrados, minimiza la suma de cuadrados de los errores para obtener los parámetros estimados, mientras que la regresión logística obtiene estimados de los parámetros con la máxima verosimilitud.

La regresión de cuadrados parciales (PLS) extrae combinaciones lineales de los predictores para minimizar el error de predicción.

Usar...	Para...	Tipo de respuesta	Método de estimación
Regression	Realizar regression simple, multiple o regression polynomial por mínimos cuadrados.	continua	Mínimos cuadrados
Stepwise	Realizar regresión por pasos, selección de variables hacia adelante, o eliminación de variables hacia atrás para identificar un conjunto útil de predictores.	continua	Mínimos cuadrados
Best Subsets	Identificar subconjuntos de los predictores con base en el criterio R máximo.	continua	Mínimos cuadrados
Fitted Line Plot	Realizar regresión lineal y polinomial con un predictor simple y graficar una línea de regresión a través de los datos.	continua	Mínimos cuadrados

PLS	Realizar regression con datos mal condicionados (ver explicación abajo).	continua	biased, non-least squares
Binary Logistic	Realizar regresión logística sobre una respuesta que solo tiene dos valores posibles, tal como presencia o ausencia.	categórica	máxima verosimilitud
Ordinal Logistic	Realizar regresión logística en una respuesta que con tres o más valores posibles que tienen un orden natural, tal como: ninguno, medio o severo.	categórica	máxima verosimilitud
Nominal Logistic	Realizar regresión logística en una respuesta con tres o más valores posibles que no tienen un orden natural, tal como: dulce, salado, o ácido.	categórica	máxima verosimilitud

Datos mal condicionados
Los datos mal condicionados se refieren a problemas en las variables predoctoras, las cuales pueden causar dificultades computacionales y estadísticas. Se presentan dos tipos de problemas: multicolinealidad y un pequeño coeficiente de variación.

Multicolinearidad
La multicolinealidad significa que ambos predictores estan correlacionados con otros predictores. Si la correlación es alta, se pueden calcular los valores estimados y los residuos, pero el error estándar de los coeficientes será grande y su exactitud numérica puede ser afectada. Se recomienda eliminar una de las variables correlacionadas.

Para identificar los predictores que están altamente correlacionados, se puede examinar la estructura de las variables predoctoras y hacer una regresión con cada uno de los predictores sospechosos y los otros predictores. Se puede también revisar el factor de inflación VIF, que mide cuanto de la varianza de un coeficiente de regresión se incrementa, si los predictores están correlacionados. Si el VIF < 1, no hay colinealidad, pero si VIF > 1, los predictores pueden estar correlacionados. Montogomery sugiere que si se sobrepasa el límite de 5 a 10, los coeficientes tienen una estimación deficiente. Algunas soluciones al problema de multicolinealidad son:

- Eliminar los predictores del modelo, especialmente si al borrarlos tienen poco efecto en la R^2.
- Cambiar los predictores formando una combinación lineal con ellos usando la regresión parcial de mínimos cuadrados o análisis de componentes principales.
- Si se usan plinomios, restar un valor cercano a la media de un predictor antes de elevarlo al cuadrado.

Coeficientes de variación pequeños
Los predictores con coeficientes de variación pequeños (porcentaje de la desviación estándar de la media) y que casi son constantes, pueden causar problemas numéricos. Por ejemplo, la variable Año con valores de 1970 a 1975 tiene un pequeño coeficiente de variación, las diferencias numéricas se encuentran en el cuarto dígito. El problema se complica se Año es elevado al cuadrado. Se pude restar una constante de los datos, reemplazando Año con Año_desde_1970 con valores de 0 a 5.

Regresión por pasos (Stepwise regression)

Stat > Regression > Stepwise
La regresión por pasos remueve y agrega variables al modelo de regresión con el propósito de identificar un subconjunto útil de predictores. La regresión por pasos remueve y agrega variables; la selección hacia delante agrega variables y la selección hacia atrás remueve variables.

- En este método de regresión por pasos, se puede iniciar con un conjunto de variables predoctoras en **Predictors in initial model.** Estas variables se remueven si sus valores p son mayores que el valor de **Alpha to enter.** Si se quieren conservar las variables en el modelo independientemente de su valor p, seleccionarlas en **Predictors to include in every model** en la ventana principal de diálogo.

- Cuando se selecciona el método de selección por pasos o hacia delante (forward), se puede poner un valor de alfa para una nueva variables en **Alpha to enter**.
- Cuando se selecciona el método de eliminación hacia atrás, se puede establecer el valor de alfa para remover una variable del modelo en **Alpha to remove**.

Entre los problemas que se presentan con el método automático de selección se tienen los siguientes:

- Como el procedimiento automáticamente "encuentra" el mejor de muchos modelos, puede ajustar los datos demasiado bien, pero solo por azar.
- Los tres procedimientos automáticos son algoritmos heurísticos, que frecuentemente trabajan bien, pero pueden no seleccionar el modelo con la R2 más alta (para un cierto número de predictores).
- Los procedimientos automáticos no pueden tomar en cuenta el conocimiento especial que le analista puede tener sobre los datos. Por tanto, el modelo seleccionado puede no ser el mejor desde el punto de vista práctico.

Ejemplo:

Los estudiantes de un curso introductorio de estadística participan en un experimento simple. Cada estudiante registra su altura, peso, género, preferencia en fumar, nivel de actividad normal, y puso en reposo. Todos lanzan una moneda, y aquellos que les salga sol, corren durante un minuto. Después de esto el grupo coimpleto registra su pulso en reposo una vez más. Se desea encontrar los mejores predictores para la segunda tasa de pulso.

Los datos se muestran a continuación:

PULSE.MTW

Pulso1	Pulso2	Corrió	Fuma	Sexo	Estatura	Peso	Actividad
64	88	1	2	1	66	140	2
58	70	1	2	1	72	145	2
62	76	1	1	1	73.5	160	3
66	78	1	1	1	73	190	1
64	80	1	2	1	69	155	2
74	84	1	2	1	73	165	1
84	84	1	2	1	72	150	3
68	72	1	2	1	74	190	2
62	75	1	2	1	72	195	2
76	118	1	2	1	71	138	2
90	94	1	1	1	74	160	1
80	96	1	2	1	72	155	2
92	84	1	1	1	70	153	3
68	76	1	2	1	67	145	2
60	76	1	2	1	71	170	3
62	58	1	2	1	72	175	3
66	82	1	1	1	69	175	2
70	72	1	1	1	73	170	3
68	76	1	1	1	74	180	2
72	80	1	2	1	66	135	3
70	106	1	2	1	71	170	2
74	76	1	2	1	70	157	2
66	102	1	2	1	70	130	2
70	94	1	1	1	75	185	2
96	140	1	2	2	61	140	2
62	100	1	2	2	66	120	2
78	104	1	1	2	68	130	2
82	100	1	2	2	68	138	2
100	115	1	1	2	63	121	2

68	112	1	2	2	70	125	2
96	116	1	2	2	68	116	2
78	118	1	2	2	69	145	2
88	110	1	1	2	69	150	2
62	98	1	1	2	62.75	112	2
80	128	1	2	2	68	125	2
62	62	2	2	1	74	190	1
60	62	2	2	1	71	155	2
72	74	2	1	1	69	170	2
62	66	2	2	1	70	155	2
76	76	2	2	1	72	215	2
68	66	2	1	1	67	150	2
54	56	2	1	1	69	145	2
74	70	2	2	1	73	155	3
74	74	2	2	1	73	155	2
68	68	2	2	1	71	150	3
72	74	2	1	1	68	155	3
68	64	2	2	1	69.5	150	3
82	84	2	1	1	73	180	2
64	62	2	2	1	75	160	3
58	58	2	2	1	66	135	3
54	50	2	2	1	69	160	2
70	62	2	1	1	66	130	2
62	68	2	1	1	73	155	2
48	54	2	1	1	68	150	0
76	76	2	2	1	74	148	3
88	84	2	2	1	73.5	155	2
70	70	2	2	1	70	150	2
90	88	2	1	1	67	140	2
78	76	2	2	1	72	180	3
70	66	2	1	1	75	190	2
90	90	2	2	1	68	145	1
92	94	2	1	1	69	150	2
60	70	2	1	1	71.5	164	2
72	70	2	2	1	71	140	2
68	68	2	2	1	72	142	3
84	84	2	2	1	69	136	2
74	76	2	2	1	67	123	2
68	66	2	2	1	68	155	2
84	84	2	2	2	66	130	2
61	70	2	2	2	65.5	120	2
64	60	2	2	2	66	130	3
94	92	2	1	2	62	131	2
60	66	2	2	2	62	120	2
72	70	2	2	2	63	118	2
58	56	2	2	2	67	125	2
88	74	2	1	2	65	135	2
66	72	2	2	2	66	125	2

84	80	2	2	2	65	118	1
62	66	2	2	2	65	122	3
66	76	2	2	2	65	115	2
80	74	2	2	2	64	102	2
78	78	2	2	2	67	115	2
68	68	2	2	2	69	150	2
72	68	2	2	2	68	110	2
82	80	2	2	2	63	116	1
76	76	2	1	2	62	108	3
87	84	2	2	2	63	95	3
90	92	2	1	2	64	125	1
78	80	2	2	2	68	133	1
68	68	2	2	2	62	110	2
86	84	2	2	2	67	150	3
76	76	2	2	2	61.75	108	2

Corrida en Minitab:
1 Open worksheet PULSE.MTW.
2 Presionar [CTRL] + [M] para activar la session de commandos.
3 Seleccionar **Editor > Enable Commands** de forma que Minitab despliegue la sesión de comandos.
4 Ejecutar **Stat > Regression > Stepwise**.
5 En **Response**, seleccionar *Pulse2*.
6 En **Predictors**, seleccionar *Pulse1 Ran-Weight*.
7 Click **Options**.
8 **In Number of steps between pauses**, anotar *2*. Click **OK** en cada una de las ventanas de diálogo.
9 En la ventana de sesión, en el primer **More**? prompt, contestar *Yes*.
10 En la ventana de sesión, en el primer **More**? prompt, contestar *No*.

Resultados:

Results for: Pulse.MTW

```
MTB > Stepwise 'Pulso2' 'Pulso1' 'Corrió'-'Peso';
SUBC>   AEnter 0.05;
SUBC>   ARemove 0.10;
SUBC>   Best 0;
SUBC>   Steps 2;
SUBC>   Constant;
SUBC>   Press.
```

Stepwise Regression: Pulso2 versus Pulso1, Corrió, ...

Alpha-to-Enter: 0.05 Alpha-to-Remove: 0.1

Response is Pulso2 on 6 predictors, with N = 92

Step	1	2
Constant	10.28	44.48
Pulso1	0.957	0.912
T-Value	7.42	9.74
P-Value	0.000	0.000

```
Corrió              -19.1
T-Value              -9.05
P-Value              0.000

S            13.5     9.82
R-Sq         37.97    67.71
R-Sq(adj)    37.28    66.98
Mallows C-p  103.2    13.5
PRESS        17252.4  9304.69
R-Sq(pred)   35.12    65.01

More? (Yes, No, Subcommand, or Help)

SUBC> Yes

Step           3
Constant       42.62

Pulso1         0.812
T-Value        8.88
P-Value        0.000

Corrió         -20.1
T-Value        -10.09
P-Value        0.000

Sexo           7.8
T-Value        3.74
P-Value        0.000

S            9.18
R-Sq         72.14
R-Sq(adj)    71.19
Mallows C-p  1.9
PRESS        8195.99
R-Sq(pred)   69.18

More? (Yes, No, Subcommand, or Help)

SUBC> No
MTB >
```

Interpretando los resultados

Este ejemplo usa seis predictores. Se requirío a Minitab intervenir para mostrar los resultados.
La primera “página” de salida proporciona los resultados para los dos primeros pasos. En el paso 1, la variable Pulso1 entró al modelo; en el paso 2, entró la variable Corrio. No se removío ninguna variable en ninguno de los dos pasos. En cada uno de los modelos, se mostró el término constante del modelo, los coeficientes y su valor t de cada variable en el modelo, S (raíz de MSE), y R2.
Como se constestó “Yes” en “MORE?”, el procedimiento automático realizó un paso adicional, agregando la variable “Sexo”. En este punto, no más variables entraron o salieron de modo que se terminó el procedimiento automático, y otra vez preguntó por intervención, donde se indicó NO.
La salida por pasos está diseñada para presentar un resumen conciso de un número de modelos ajustados.

6. REGRESIÓN POR MEJORES SUBCONJUNTOS HALLADOS (Best Subsets)

La regresión de los mejores subconjuntos idnetifica los modelos de regresión que mejor ajusten los datos con los predictores especificados. Es una forma eficiente de identificar modelos que logreen las metas con los menores predictores que sea posible. Los modelos de subconjuntos pueden realmente estimar los coeficientes de regresión y predecir respuestas futures con varianzas más pequeñas que el modelo completo que utiliza todos los predictores.

Primero se evalúan los modelos que tienen un predictor, después los de dos predictores, etcetera. En cada caso se muestra el mejor modelo.
Ejemplo:
El flujo de calor solar se mide ocmop parte de una prueba de energía térmica solar. Se desea ver como se estima el flujo de calor con base en otras variables: aislamiento, posición de puntos focales en el este, sur, y norte, y la hora del día. (datos de D.C. Montgomery and E.A. Peck (1982). *Introduction to Linear Regression Analysis*. John Wiley & Sons. p. 486).
Los datos son los siguientes (Exh_regr.Mtw):

Flujo_de_calor	Aislamiento	Este	Sur	Norte	Hora
271.8	783.35	33.53	40.55	16.66	13.2
264	748.45	36.5	36.19	16.46	14.11
238.8	684.45	34.66	37.31	17.66	15.68
230.7	827.8	33.13	32.52	17.5	10.53
251.6	860.45	35.75	33.71	16.4	11
257.9	875.15	34.46	34.14	16.28	11.31
263.9	909.45	34.6	34.85	16.06	11.96
266.5	905.55	35.38	35.89	15.93	12.58
229.1	756	35.85	33.53	16.6	10.66
239.3	769.35	35.68	33.79	16.41	10.85
258	793.5	35.35	34.72	16.17	11.41
257.6	801.65	35.04	35.22	15.92	11.91
267.3	819.65	34.07	36.5	16.04	12.85
267	808.55	32.2	37.6	16.19	13.58
259.6	774.95	34.32	37.89	16.62	14.21
240.4	711.85	31.08	37.71	17.37	15.56
227.2	694.85	35.73	37	18.12	15.83
196	638.1	34.11	36.76	18.53	16.41
278.7	774.55	34.79	34.62	15.54	13.1
272.3	757.9	35.77	35.4	15.7	13.63
267.4	753.35	36.44	35.96	16.45	14.51
254.5	704.7	37.82	36.26	17.62	15.38
224.7	666.8	35.07	36.34	18.12	16.1
181.5	568.55	35.26	35.9	19.05	16.73
227.5	653.1	35.56	31.84	16.51	10.58
253.6	704.05	35.73	33.16	16.02	11.28
263	709.6	36.46	33.83	15.89	11.91
265.8	726.9	36.26	34.89	15.83	12.65
263.8	697.15	37.2	36.27	16.71	14.06

Instrucciones de Minitab:
1 Open worksheet EXH_REGR.MTW.
2 Seleccionar **Stat > Regression > Best Subsets**.
3 En **Response**, seleccionar **Flujo_de_Calor**.
4 En **Free Predictors**, seleccionar **Aislamiento-Hora** Click **OK**.
Los resultados se muestran a continuación:

Results for: Exh_regr.MTW

Best Subsets Regression: Flujo_de_calor versus Aislamiento, Este, ...

Response is Flujo_de_calor

Vars	R-Sq	R-Sq(adj)	Mallows C-p	S	Aislamiento	Este	Norte	Hora	Sur
1	72.1	71.0	38.5	12.328			X		
1	39.4	37.1	112.7	18.154	X				
2	85.9	84.8	9.1	8.9321			X	X	
2	82.0	80.6	17.8	10.076				X	X
3	87.4	85.9	7.6	8.5978			X	X	X
3	86.5	84.9	9.7	8.9110	X			X	X
4	89.1	87.3	5.8	8.1698	X		X	X	X
4	88.0	86.0	8.2	8.5550	X			X	X
5	89.9	87.7	6.0	8.0390	X	X	X	X	X

Interpretando los resultados
Cada línea de la salidad representa un modelo diferente. Vars es el número de variables o predictores en el modelo, R2 y R2 ajustada se convienten a porcentajes. Los predictores que están presentes se indican con una X.

En este ejemplo, no es claro que modelo ajusta mejor a los datos.

- El modelo con todas las variables tiene la mayor R2 ajustada (87.7%), un valor bajo de Cp de Mallows (6.0), y el menor valor de S (8.039).

- El modelo de cuatro predictores con todas las variables excepto la *Hora*, tiene un valor bajo de Cp (5.8), la S es ligeramente mayor (8.16) y la R2 ajustada en ligeramente menor (87.3%).

- El mejor modelo de tres predictores incluye Norte, Sur, y Este, con un valor de Cp ligeramente más alto (7.6) y un valor menor de R2 ajustado.

- El modelo con dos predictores podría ser considerado con el menor ajuste. Se puede observar que el agregar la variable Este no mejora el ajuste del modelo.

Antes de seleccionar un modelo, se debe verificar si no se viola ninguno de los supuestos de la regresión por medio de las gráficas de residuos y otras pruebas de diagnóstico, tales como las siguientes.

Verificación de la adecuación del modelo

Características de un modelo de regression adecuado	Checar usando...	Posibles soluciones
Relación lineal entre respuesta y predictores	Prueba de Lack-of-fit (falta de ajuste) Gráfica de Residuales vs variables	· Agregar terminos de mayor orden al modelo · Transformar variables.

Los Residuales tienen varianza constante.	Gráfica de Residuals vs estimados (fits)	· Transformar variables. · Mínimos cuadrados ponderados.
Los Residuales son independientes entre sí (no correlacionados).	Estadístico de Durbin-Watson Gráfica de Residualess vs orden	· Agregar un nuevo predictor. · Usar análisis de series de tiempo. · Agregar variable defasada en tiempo (lag).
Los Residuales están normalmente distribuidos.	Histograma de residuales Gráfica Normal de residuales Gráfica de Residuales vs estimados (fits) Prueba de Normalidad	· Transformar variables. · Checar puntos atípicos.
Observations No usuales, puntos atípicos o outliers.	Gráficas de Residuales Influyentes (Leverages) Distancia de Cook's DFITS	· Transformar variables. · Eliminar la observación atípica.
Datos mal condicionados (ill conditioned).	Factor de Inflación de Variance (VIF) Matriz de correlación de predictores	· Remover predictor. · Regresión de mínimos cuadrados parciales. · Transformar variables.

Si se determina que el modelo no cumple con los criterios listados en la tabla, se debe:
1. Verificar si los datos se introdujeron correctamente, especialemente identificar puntos atípicos.
2. Tratar de determinar las causas del problema. Puedes querer ver que tan sensible es el modelo al problema. Por ejemplo, si se observa un Outlier, correr el modelo sin esa observación, para ver como difieren los resultados.
3. Considerar alguna de las soluciones listadas en la tabla.

7. REGRESIÓN POR MÍNIMOS CUADRADOS PARCIALES (PLS)

Usar reegresión de mínimos cuadrados parcial (PLS) para realizar una regresión sesgada, no de mínimos cuadrados. PLS se utiliza cuando los predictores son muy colineales o se tienen más predictores que observaciones, y la regresión lineal normal falla o produce coeficientes con altos errores estándar. La PLS reduce el número de predictores a un conjunto de componentes no correlacionados y realiza la regresión de mínimos cuadrados en esos componentes.
La PLS ajusta variables de respuesta múltiple en un modelo simple. Dado que los modelos PLS tratan las respuestas como multivariadas, los resultados pueden diferir de si se tratan individualmente por separado. El modelo agrupa las respuestas múltiples sólo si estan correlacionadas.

Ejemplo:

Un productor de vino quiere saber como la composición química del vino se relaciona con las pruebas sensoriales. Se tienen 37 muestras, cada una descrita por 17 concentraciones elementales (Cd, Mo, Mn, Ni, Cu. Al, Ba, Cr, Sr, B, Mg, Si, Na, Ca, P, K) y una medida del aroma del vino de un panel de catadores. Se quiere predecir la media del aroma a partir de los 17 elementos y determinar si el modelo PLS es adecuado, dado que la relaciónde muestras a predictores es baja. Los datos son de I.E. Frank and B.R. Kowalski (1984). "Prediction of Wine Quality and Geographic Origin from Chemical Measurements by Partial Least-Squares Regression Modeling," *Analytica Chimica Acta*, 162, 241□251.
Archivo WineAroma.mtw

Cd	Mo	Mn	Ni	Cu	Al	Ba	Cr	Sr	Pb	B	Mg	Si	Na	Ca	P	K	Aroma
0.005	0.044	1.51	0.122	0.83	0.982	0.387	0.029	1.23	0.561	2.63	128	17.3	66.8	80.5	150	1130	3.3
0.055	0.16	1.16	0.149	0.066	1.02	0.312	0.038	0.975	0.697	6.21	193	19.7	53.3	75	118	1010	4.4
0.056	0.146	1.1	0.088	0.643	1.29	0.308	0.035	1.14	0.73	3.05	127	15.8	35.4	91	161	1160	3.9
0.063	0.191	0.96	0.38	0.133	1.05	0.165	0.036	0.927	0.796	2.57	112	13.4	27.5	93.6	120	924	3.9
0.011	0.363	1.38	0.16	0.051	1.32	0.38	0.059	1.13	1.73	3.07	138	16.7	76.6	84.6	164	1090	5.6
0.05	0.106	1.25	0.114	0.055	1.27	0.275	0.019	1.05	0.491	6.56	172	18.7	15.7	112	137	1290	4.6
0.025	0.479	1.07	0.168	0.753	0.715	0.164	0.062	0.823	2.06	4.57	179	17.8	98.5	122	184	1170	4.8
0.024	0.234	0.91	0.466	0.102	0.811	0.271	0.044	0.963	1.09	3.18	145	14.3	10.5	91.9	187	1020	5.3
0.009	0.058	1.84	0.042	0.17	1.8	0.225	0.022	1.13	0.048	6.13	113	13	54.4	70.2	158	1240	4.3
0.033	0.074	1.28	0.098	0.053	1.35	0.329	0.03	1.07	0.552	3.3	140	16.3	70.5	74.7	159	1100	4.3
0.039	0.071	1.19	0.043	0.163	0.971	0.105	0.028	0.491	0.31	6.56	103	9.47	45.3	67.9	133	1090	5.1
0.045	0.147	2.76	0.071	0.074	0.483	0.301	0.087	2.14	0.546	3.5	199	9.18	80.4	66.3	212	1470	3.3
0.06	0.116	1.15	0.055	0.18	0.912	0.166	0.041	0.578	0.518	6.43	111	11.1	59.7	83.8	139	1120	5.9
0.067	0.166	1.53	0.041	0.043	0.512	0.132	0.026	0.229	0.699	7.27	107	6	55.2	44.9	148	854	7.7
0.077	0.261	1.65	0.073	0.285	0.596	0.078	0.063	0.156	1.02	5.04	94.6	6.34	10.4	54.9	132	899	7.1
0.064	0.191	1.78	0.067	0.552	0.633	0.085	0.063	0.192	0.777	5.56	110	6.96	13.6	64.1	167	976	5.5
0.025	0.009	1.57	0.041	0.081	0.655	0.072	0.021	0.172	0.232	3.79	75.9	6.4	11.6	48.1	132	995	6.3
0.02	0.027	1.74	0.046	0.153	1.15	0.094	0.021	0.358	0.025	4.24	80.9	7.92	38.9	57.6	136	876	5
0.034	0.05	1.15	0.058	0.058	1.35	0.294	0.006	1.12	0.206	2.71	120	14.7	68.1	64.8	133	1050	4.6
0.043	0.268	2.32	0.066	0.314	0.627	0.099	0.045	0.36	1.28	5.68	98.4	9.11	19.5	64.3	176	945	6.4
0.061	0.245	1.61	0.07	0.172	2.07	0.071	0.053	0.186	1.19	4.42	87.6	7.62	11.6	70.6	156	820	5.5
0.047	0.161	1.47	0.154	0.082	0.546	0.181	0.06	0.898	0.747	8.11	160	19.3	12.5	82.1	218	1220	4.7
0.048	0.146	1.85	0.092	0.09	0.889	0.328	0.1	1.32	0.604	6.42	134	19.3	125	83.2	173	1810	4.1
0.049	0.155	1.73	0.051	0.158	0.653	0.081	0.037	0.164	0.767	4.91	86.5	6.46	11.5	53.9	172	1020	6
0.042	0.126	1.7	0.112	0.21	0.508	0.299	0.054	0.995	0.686	6.94	129	43.6	45	85.9	165	1330	4.3
0.058	0.184	1.28	0.095	0.058	1.3	0.346	0.037	1.17	1.28	3.29	145	16.7	65.8	72.8	175	1140	3.9
0.065	0.211	1.65	0.102	0.055	0.308	0.206	0.028	0.72	1.02	6.12	99.3	27.1	20.5	95.2	194	1260	5.1
0.065	0.129	1.56	0.166	0.151	0.373	0.281	0.034	0.889	0.638	7.28	139	22.2	13.3	84.2	164	1200	3.9
0.068	0.166	3.14	0.104	0.053	0.368	0.292	0.039	1.11	0.831	4.71	125	17.6	13.9	59.5	141	1030	4.5
0.067	0.199	1.65	0.119	0.163	0.447	0.292	0.058	0.927	1.02	6.97	131	38.3	42.9	85.9	164	1390	5.2
0.084	0.266	1.28	0.087	0.071	1.14	0.158	0.049	0.794	1.3	3.77	143	19.7	39.1	128	146	1230	4.2
0.069	0.183	1.94	0.07	0.095	0.465	0.225	0.037	1.19	0.915	2	123	4.57	7.51	69.4	123	943	3.3
0.087	0.208	1.76	0.061	0.099	0.683	0.087	0.042	0.168	1.33	5.04	92.9	6.96	12	56.3	157	949	6.8
0.074	0.142	2.44	0.051	0.052	0.737	0.408	0.022	1.16	0.745	3.94	143	6.75	36.8	67.6	81.9	1170	5
0.084	0.171	1.85	0.088	0.038	1.21	0.263	0.072	1.35	0.899	2.38	130	6.18	101	64.4	98.6	1070	3.5
0.106	0.307	1.15	0.063	0.051	0.643	0.29	0.031	0.885	1.61	4.4	151	17.4	7.25	103	177	1100	4.3
0.102	0.342	4.08	0.065	0.077	0.752	0.366	0.048	1.08	1.77	3.37	145	5.33	33.1	58.3	117	1010	5.2

Las instrucciones de Minitab son las siguientes:

1 Open worksheet WINEAROMA.MTW o tomar los datos de la tabla.
2 Seleccionar **Stat > Regression > Partial Least Squares**.
3 En **Responses**, seleccionar *Aroma*.
4 En **Predictors**, selección las variables *Cd-K*.
5 En **Maximum number of components**, indicar *17*.
6 Click **Validation**, seleccionar **Leave-one-out**. Click **OK**.
7 Click **Graphs**, luego seleccionar **Model selection plot**, **Response plot**, **Std Coefficient plot**, **Distance plot**, **Residual versus leverage plot**, y **Loading plot**. No seleccionar **Coefficient plot**.
Click **OK** en cada una de las ventanas de diálogo.

Los resultados se muestran a continuación:
PLS Regression: Aroma versus Cd, Mo, Mn, Ni, Cu, Al, Ba, Cr, ...

La primera línea, muestra el número de componentes en el modelo óptimo, el cual es definido como el modelo con la mayor R^2 Predictora (***Predicted R^2)***, en este caso de 0.46.

R^2 Predictora

Es similar a la R^2, la R^2 predictora indica que tan bien estima el modelo las respuestas a nuevas observaciones, mientras que la R^2 sólo indica que tan bien el modelo se ajusta a los datos. La R^2 predictora puede evitar el sobreajuste del modelo y es más útil que la R^2 ajustada para comparar modelos dado que es calculada con observaciones no incluidas en el cálculo del modelo.

Su valor se encuentra entre 0 y 1, y se calcula a partir del estadístico PRESS. Valores altos de R^2 Predictora sugieren modelos de mayor capacidad de predicción o estimación.

Como se tiene el mimso número de componentes que predictors (17), se pueden comparar los estadísticos de bondad de ajuste y de bondad de predicción para el modelo PLS y la solución de mínimos cuadrados.

Number of components selected by cross-validation: 2
Number of observations left out per group: 1
Number of components cross-validated: 17

El ANOVA muestra que el valor p para Aroma es 0.000 menor a 0.05, proporcionando suficiente evidencia de que el modelo es significativo.

Analysis of Variance for Aroma

Source	DF	SS	MS	F	P
Regression	2	28.8989	14.4494	39.93	0.000
Residual Error	34	12.3044	0.3619		
Total	36	41.2032			

Usar la tabla de Selección y Validación del Modelo para seleccionar el número óptimo de componentes para el modelo. Dependiendo de los datos o campo de estudio, se puede determinar que un modelo diferente del seleccionado por validación cruzada es más apropiado.

Model Selection and Validation for Aroma

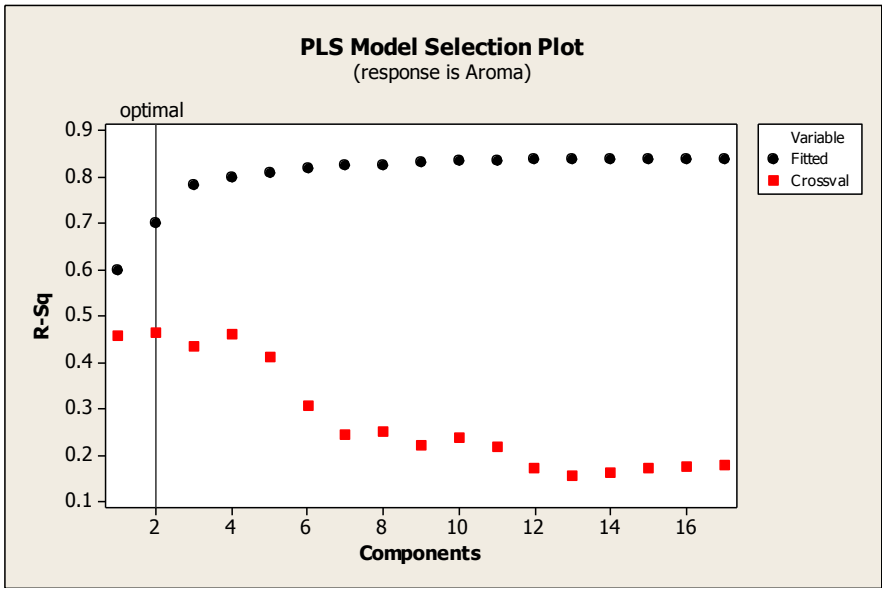
Components	X Variance	Error SS	R-Sq	PRESS	R-Sq (pred)
1	0.225149	16.5403	0.598569	22.3904	0.456585
2	0.366697	12.3044	0.701374	22.1163	0.463238
3		8.9938	0.781720	23.3055	0.434377
4		8.2761	0.799139	22.2610	0.459726
5		7.8763	0.808843	24.1976	0.412726
6		7.4542	0.819087	28.5973	0.305945
7		7.2448	0.824168	31.0924	0.245389
8		7.1581	0.826274	30.9149	0.249699
9		6.9711	0.830811	32.1611	0.219451
10		6.8324	0.834178	31.3590	0.238920
11		6.7488	0.836207	32.1908	0.218732
12		6.6955	0.837501	34.0891	0.172660
13		6.6612	0.838333	34.7985	0.155442
14		6.6435	0.838764	34.5011	0.162660
15		6.6335	0.839005	34.0829	0.172811
16		6.6296	0.839100	34.0143	0.174476
17		6.6289	0.839117	33.8365	0.178789

- El modelo con dos componentes, seleccionado por validación cruzada, tiene una R^2 de 70.1% y una R^2 de Predicción de 46.3%. El modelo de cuatro componentes tiene una R^2 predictora un poco menor, con una mayor R^2, pero también se podría utilizar.
- Comparando la R^2 predictora del modelo PLS de dos componentes con la R^2 predictora del modelo de mínimos cuadrados de 17 componentes, se puede ver que el modelo PLS predice los datos mucho

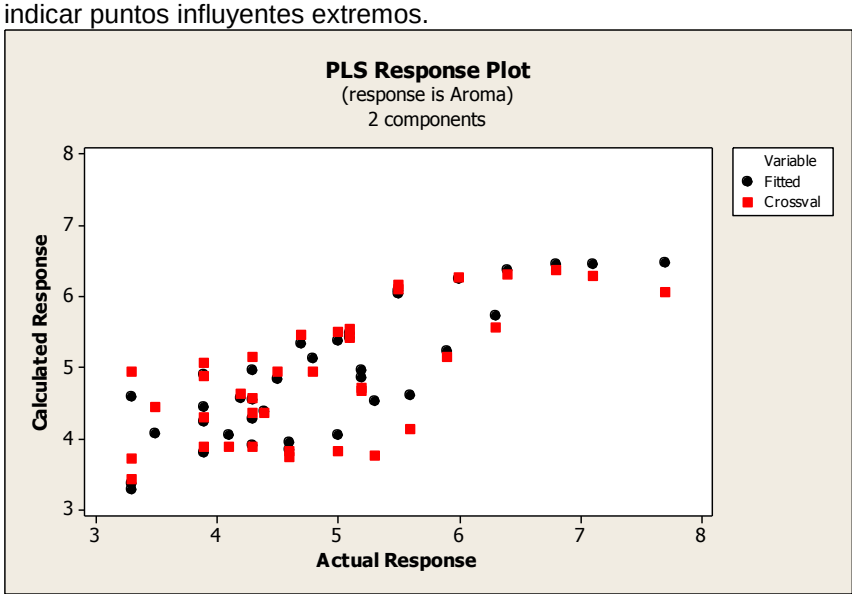
más exactamente que el modelo completo. La R^2 del modelo PLS de dos componentes es de 46%, mientreas que el de 17 componentes es de solo 18%.

- La varianza de X indica la cantidad de varianza en los predictores que es explicada por el modelo. En este ejemplo, el modelo de dos componentes explica el 36.7% de la varianza en los predictores.
-

Esta gráfica muestra la tabla de "Model Selection and Validation. La línea vertical indica que le modelo óptimo tiene dos componentes. Se puede observar que la habilidad predictiva de todos los modelos con más de cuatro componentes, se reduce significativamente, incluyendo el de 17 componententes con sólo 18%.

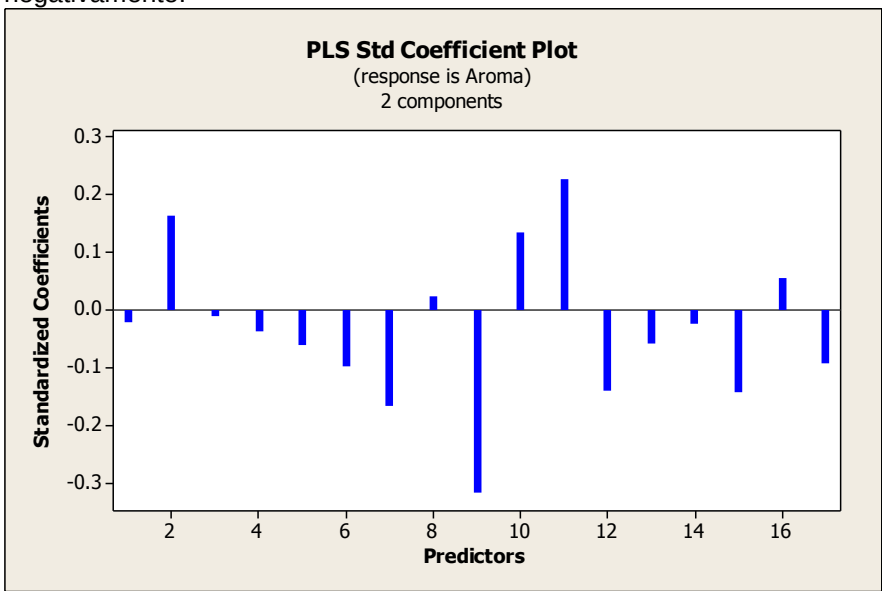


Como los puntos muestran un patrón de línea recta, de abajo hacia arriba, la gráfica de respuesta indica que el modelo ajusta los datos adecuadamente. A pesar de haber diferencias entre las respuestas estimadas (*fitted*) y las de validación cruzada (*cross-validated* indica que tan bien el modelo estima los datos, de modo que se puedan omitir), ninguno es suficientemente severo para indicar puntos influyentes extremos.

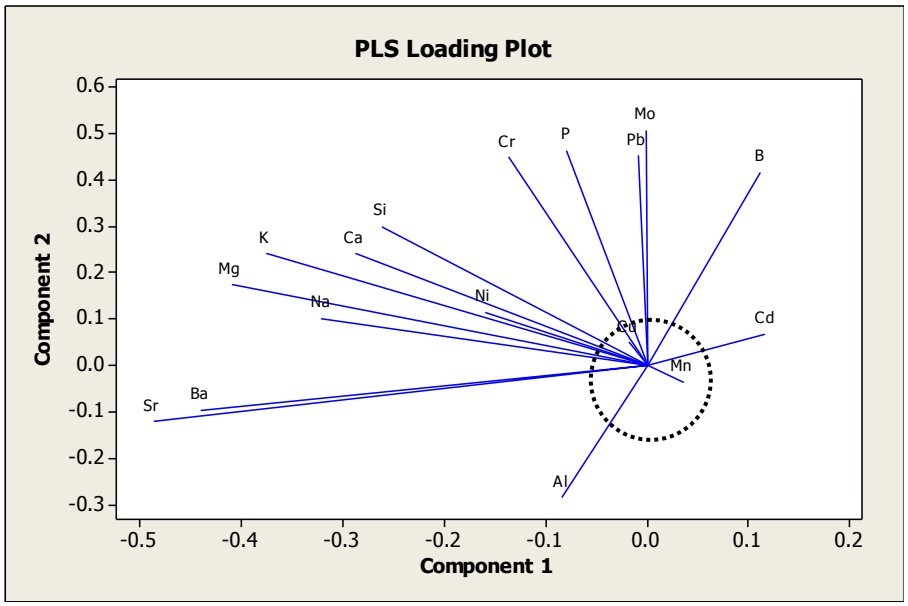


La gráfica de coeficientes muestra los coeficientes estandarizados para los predictores. Se usa para interpretar la magnitud y signo de los coeficientes. Los elementos Sr, B, Mg, Pb y Ca tienen los

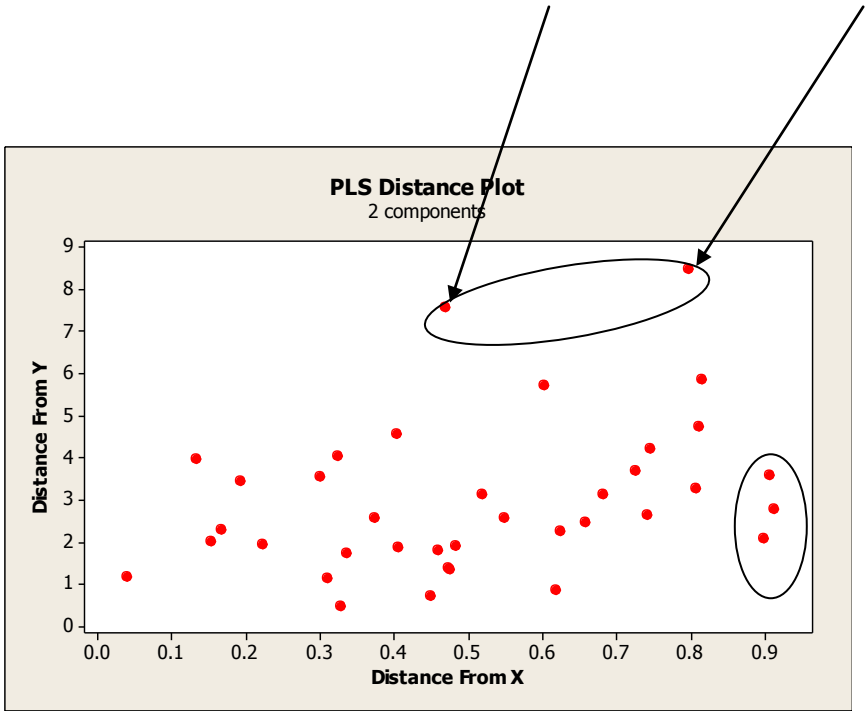
coeficientes más altos y el mayor impacto en Aroma. Los elementos Mo, Cr, Pb, y B están positivamente realcionados con Aroma, mientras que Cd, Ni, Cu, Al, BA y Sr están realcionados negativamente.



La gráfica de carga compara la influencia relativa de los predictors en la respuesta. El Cu y el Mn tienen líneas muy cortas, indicando que tienen carga baja en X y no se realcionan con Aroma. Los elementos Sr, Mg, y Ba tienen líneas largas, indicando que tienen una carga mayor y se están más relacionadas con Aroma.

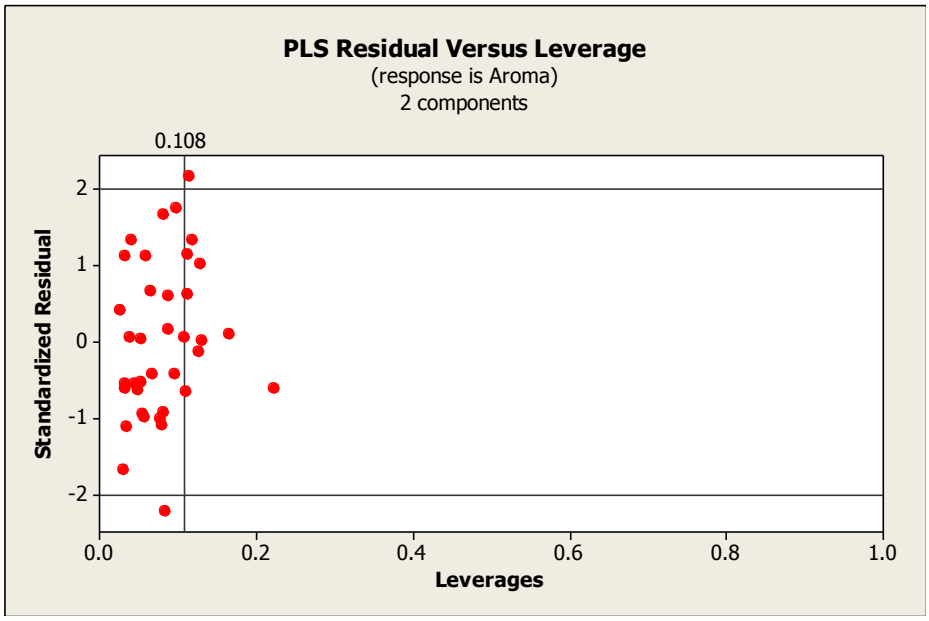


La gráfica de distancia y la gráfica de residuales versus influyentes, muestran los puntos atípicos e influyentes. Brushing la gráfica de distancia, pueden observarse comparados con el resto de datos. La observación 14 y 32 tienen una mayor distancia en el eje Y y las observaciones de los renglones 7, 12, y 23 tienen una mayor distancia en el eje X.



La gráfica de residuos versus influyentes confirma estos hallazgos, indicando que:

- Las observaciones 14 y 32 son puntos atípicos, ya que salen de las líneas de referencia horizontales.
- Las observaciones 7, 12 y 23 tienen valores influyentes extremos, dado que están a la derecha de la línea vertical de referencia.



8. REGRESIÓN LOGÍSTICA BINARIA[3]

Tanto la regression logística como la regresión por mínimos cuadrados, investigan la relación entre una variable de respuesta y uno o más predictores. Una diferencia práctica entre ellas es que las técnicas de regresión logística se utilizan con variabnles de respuesta categóricas, y las técnicas de regresión lineal son usadas con variables de respuesta comtinuas.

Hay tres procedimientos de regresión logística que se pueden utilizar para evaluar las relaciones entre uno o más vareiables predoctoras y una respuesta categórica de los tipos siguientes:

[3] Hair., Joseph Jr., *Et. Al., Multivariate Data Analysis,* Prentice Hall Internacional, Nueva Jersey, 1984, pp. 279- 325

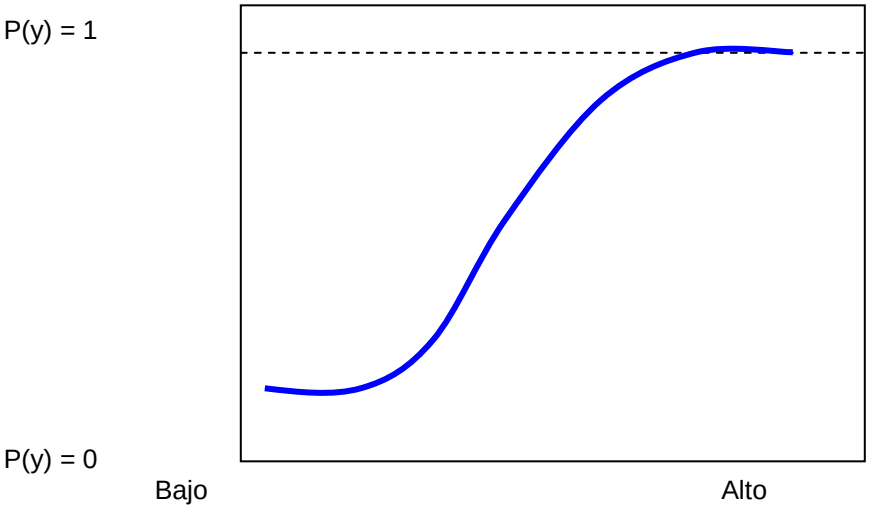
Tipo de Variable	Número de categorias	Características	Ejemplos
Binary	2	Dos niveles	Éxito, falla Si, No
Ordinal	3 o más	Orden natural de niveles	Nada, moderado, severo Fino, medio, grueso
Nominal	3 o más	Niveles sin orden natural	Azul negro, rojo, amarillo Soleado, lluvioso, nublado

Tanto los métodos de regression logísticos como los métodos de mínimos cuadrados, estiman los parámetros en el modelo de manera que el ajuste es optimizado. El de mínimos cuadrados minimiza la suma de cuadrados de los errores para estimar los parámetros, mientras que la regresión logística obtiene la máxima verosimilitud de los parámetros usando un algoritmo iterativo de mínimos cuadrados reponderados.

La regresión logística predice directamente la probabilidad de que un evento ocurra, la respuesta tiene un rango entre cero y uno con una forma de S.

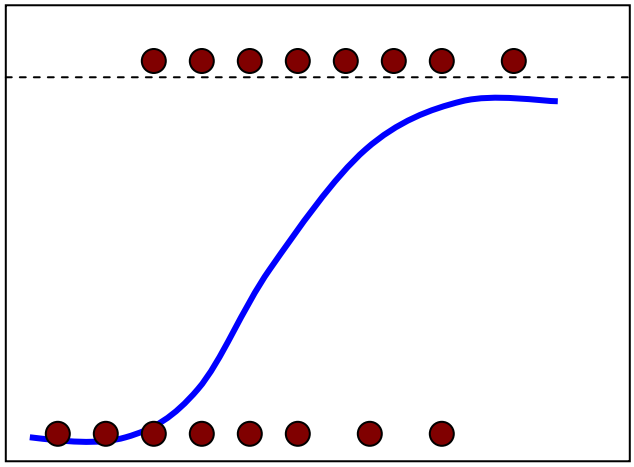
Su término de error es el de una variable discreta, que no sigue la distribución normal sino la binomial; la varianza de una variable dicotómica no es contante, creando situaciones de heteroestacidad.

Su relación única entre las variables independientes y dependiente requiere un método diferente para estimar, evaluar bondad de ajuste e interpretar los coeficientes.



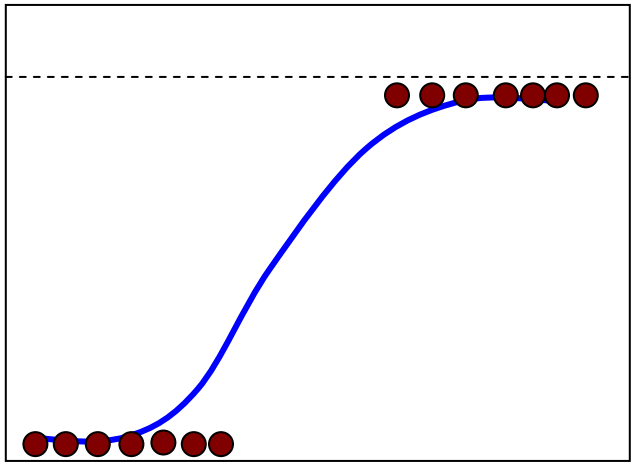
Para la estimación de sus coeficientes dada su naturaleza no lineal, se utiliza el método de máxima verosimilitud, buscando el mayor valor de verosimilitud (*likelihood value)* de que un evento ocurra, en vez de la mínima suma de cuadrados como en la regresión múltiple.

En el siguiente ejemplo se muestran ejemplos de cuando el modelo puede adecuado y cuando no.



A. Relación con ajuste pobre

Hay valores de X que tienen respuesta Y de eventos y no eventos.



B. Relación con ajuste bien definido

Los valores de X sólo tienen una respuesta en Y de eventos o no eventos.

El nombre de regresión logística deriva de la transformación utilizada en su variable dependiente. El procedimiento para calcular los coeficientes logísticos, comparan la probabilidad de que un evento ocurra con la probabilidad de que no ocurra. Esta razón de posibilidades se expresa como:

$$\frac{P(evento)}{P(no\ evento)} = e^{B0} + B_1X_1 + B_2X_2 + ... + B_nX_n$$

Los coeficientes estimados (B0, B1, … Bn) son medidas reales de las posibilidades en la relación de probabilidades. Como se expresan en logaritmos, al final se deben regresar con las funciones de antilogaritmo de modo que se pueda el efecto en las probabilidades de manera más fácil.

Cuando los coeficientes son positivos, su transformación será mayor a uno, en la razón de posibilidades se incrementa y viceversa en caso contrario.

La medición global de que tan bien ajusta el modelo, similar a la menor suma de cuadrados en la regresión múltiple, se da por el valor de verosimilitud (que es realmente menos 2 veces el logaritmo del valor de verosimilitud = -2LL). Un modelo ideal tendrá una verosimilitud de 1 y un -2LL de cero. Para determinar un “pseudos R2” de la regresión logística se puede calcular como:

$$R^2_{\text{logit}} = \frac{-2LL_{null} - (-2LL_{\text{model}})}{-2LL_{null}}$$

Para probar la significancia de los coeficientes se usa el estadístico de **Wald**, utilizado de manera similar a la regresión múltiple para probar significancia.

Ejemplo de Minitab
Un investigador está interesado en comprender el efecto de fumar y el peso en la tasa de pulso en reposo. Dado que se ha categorizado la tasa de respuesta del puso en baja y alta, el análisis de regresión logística es adecuado para comprender los efectos de fumar y peso en la tasa de pulso.

DATOS MINITAB. Exh_Regr		
Y	**X1**	**X2**
Pulso en reposo	**Fuma**	**Peso**
Bajo	No	140
Bajo	No	145
Bajo	Si	160
Bajo	Si	190
Bajo	No	155
Bajo	No	165
Alto	No	150
Bajo	No	190
Bajo	No	195
Bajo	No	138
Alto	Si	160
Bajo	No	155
Alto	Si	153
Bajo	No	145
Bajo	No	170
Bajo	No	175
Bajo	Si	175
Bajo	Si	170
Bajo	Si	180
Bajo	No	135
Bajo	No	170
Bajo	No	157
Bajo	No	130
Bajo	Si	185
Alto	No	140
Bajo	No	120
Bajo	Si	130
Alto	No	138
Alto	Si	121
Bajo	No	125
Alto	No	116

Bajo	No	145
Alto	Si	150
Bajo	Si	112
Bajo	No	125
Bajo	No	190
Bajo	No	155
Bajo	Si	170
Bajo	No	155
Bajo	No	215
Bajo	Si	150
Bajo	Si	145
Bajo	No	155
Bajo	No	155
Bajo	No	150
Bajo	Si	155
Bajo	No	150
Alto	Si	180
Bajo	No	160
Bajo	No	135
Bajo	No	160
Bajo	Si	130
Bajo	Si	155
Bajo	Si	150
Bajo	No	148
Alto	No	155
Bajo	No	150
Alto	Si	140
Bajo	No	180
Bajo	Si	190
Alto	No	145
Alto	Si	150
Bajo	Si	164
Bajo	No	140
Bajo	No	142
Alto	No	136
Bajo	No	123
Bajo	No	155
Alto	No	130
Bajo	No	120
Bajo	No	130
Alto	Si	131
Bajo	No	120
Bajo	No	118
Bajo	No	125
Alto	Si	135
Bajo	No	125
Alto	No	118
Bajo	No	122

Bajo	No	115
Bajo	No	102
Bajo	No	115
Bajo	No	150
Bajo	No	110
Alto	No	116
Bajo	Si	108
Alto	No	95
Alto	Si	125
Bajo	No	133
Bajo	No	110
Alto	No	150
Bajo	No	108

Corrida en Minitab:
1 Abrir la hoja de trabajo EXH_REGR.MTW o tomar datos de esta tabla.
2 Seleccionar **Stat > Regression > Binary Logistic Regression**.
3 En **Response**, seleccionar **RestingPulse**. En **Model**, seleccionar **Smokes Weight**. En **Factors** (**optional**), seleccionar **Smokes**.
4 Click **Graphs**. Seleccionar **Delta chi-square vs probability** y **Delta chi-square vs leverage**. Click **OK**.
5 Click **Results**. Seleccionar **In addition, list of factor level values, tests for terms with more than 1 degree of freedom, and 2 additional goodness-of-fit tests**. Click **OK** en cada uno de las ventanas de diálogo.
Model: Especificar los términos a ser incluidos en el modelo.
Factors (optional): Especificar cuales de los predictores son factores, Minitab asume que todas las variables en el modelo con covariados a menos que se especifique cuales predictors son factores. Los predoctores continuos deben ser modelados como covariados; y los predictores categóricos deben ser modelados como factores.

Los resultados se muestran a continuación:

Results for: Exh_regr.MTW

Binary Logistic Regression: RestingPulse versus Smokes, Weight

Link Function: Logit

Información de la respuesta: **-** muestra el número de valores no considerados y el número de observaciones que caen dentro de cada una de las dos categorías de respuesta. El valor de la respuesta que se ha designado como el evento de referencia es la primera entrada en Valor y se etiqueta como evento. En este caso, el evento de referencia es tasa de pulso baja.
Response Information
Variable Value Count
Pulso en reposo Bajo 70 (Event)
Alto 22
Total 92

Información de los factores: muestra todos los factores del modelo, el número de niveles para cada factor, y los valores de nivel de los factores. El nivel del factor que se ha designado como nivel de referencia es la primera entrada en Values, el sujeto no fuma.
Factor Information
Factor Levels Values
Fuma 2 No, Si

Tabla de regression logística – muestra los coeficientes estimados, error estándar de los coeficientes, su valor Z y p. Cuando se usa la función de enlace logia, se puede también obtener la tasa de posibilidades y un intervalo de confianza del 95% para esta tasa.

- De la salida, se puede ver que los coeficientes estimados para ambos Fuma (z=-2.16, p =0.031) y Peso (z= 2.04, p = 0.041), tienen valores p menores a 0.05 indicando que hay suficiente evidencia de que los coeficientes no sean cero utilizando un alfa de 0.05.
- El coeficiente estimado de -1.193 para *Fuma*, representa el cambio en el logaritmo de P(pulso bajo/P(pulso alto) cuando el sujeto fuma comparado a cuando no lo hace, con el covariado peso mantenido constante.
- El coeficiente estimado de 0.025 para *Peso* representa el cambio en el logaritmo de P(pulso bajo/P(pulso alto) con un incremento en peso de 1 libra, con el factor *Fuma* mantenido constante.
- A pesar de que hay evidencia de que el coeficiente estimado para el peso no es cero, la tasa de posibilidades es cercana a uno (1.03), indicando que un incremento de una libra en peso afecta de forma mínima a la tasa de pulso en reposo de la persona. Se puede observar una diferencia más significativa si se comparan sujetos con una diferencia más grande en peso, (por ejemplo, si la unidad de peso es de 10 libras, la tasa de posibilidades pasa a ser 1.28, indicando que las posibilidades de un sujeto para que tenga un pulso bajo se incrementan 1.28 veces con cada 10 libras de incremento en peso).
- Para *Fuma*, el coeficiente negativo de -1.193 y la tasa de posibilidades de 0.30, indica que quien fuma, tiende a tener una tasa de pulso más alta que los sujetos que no fuman. Si los sujetos tienen el mismo peso, la tasa de posibilidades se puede interpretar como las posibilidades de que los fumadores en la muestra tengan un pulso bajo sea sólo del 30% de las posibilidades de que los no fumadores tengan un pulso bajo.

Logistic Regression Table

					Odds	95% CI	
Predictor	Coef	SE Coef	Z	P	Ratio	Lower	Upper
Constant	-1.98717	1.67930	-1.18	0.237			
Fuma							
Si	-1.19297	0.552980	-2.16	0.031	0.30	0.10	0.90
Peso	0.0250226	0.0122551	2.04	0.041	1.03	1.00	1.05

Se muestra el último valor de verosimilitud logarítmica de las iteraciones de máxima verosimilitud, junto con el estadístico G. Este estadístico prueba la hipótesis nula de que todos los coeficientes asociados con los predictores son iguales a cero versus que sean diferentes de cero. En este caso, G = 7.54, con un valor P de 0.023, indica que suficiente evidencia de uno de los coeficientes es diferente de cero, para alfa de 0.05.

Log-Likelihood = -46.820
Test that all slopes are zero: G = 7.574, DF = 2, P-Value = 0.023

Las pruebas de bondad de ajuste muestran las pruebas de – Pearson, desviación, y Hosmer-Lemeshow. Como se seleccionó el enlace a la función Logia y las opciones en la ventana de resultados, además se muestran las pruebas de Brown de alternativa general y simétrica. Las pruebas de bondad de ajuste, con valor p de 0.312 y 0.724, indican que no hay suficiente evidencia para afirmar que el modelo no ajusta los datos adecuadamente, si los valores p fueran menores a alfa, el modelo no ajustaría a los datos.

Goodness-of-Fit Tests

Method	Chi-Square	DF	P
Pearson	40.8477	47	0.724
Deviance	51.2008	47	0.312
Hosmer-Lemeshow	4.7451	8	0.784
Brown:			
General Alternative	0.9051	2	0.636

Symmetric Alternative 0.4627 1 0.496

La tabla de valores observados y frecuencias esperadas – permite ver que tan bien el modelo ajusta los datos, al comparar las frecuencias observadas y esperadas. Hay evidencia insuficiente de que el modelo no ajuste a los datos bien, ya que ambas frecuencias son similares. Esto soporta las conclusiones hechas en las pruebas de bondad de ajuste.

Table of Observed and Expected Frequencies:
(See Hosmer-Lemeshow Test for the Pearson Chi-Square Statistic)

	Group										
Value	1	2	3	4	5	6	7	8	9	10	Total
Bajo											
Obs	4	6	6	8	8	6	8	12	10	2	70
Exp	4.4	6.4	6.3	6.6	6.9	7.2	8.3	12.9	9.1	1.9	
Alto											
Obs	5	4	3	1	1	3	2	3	0	0	22
Exp	4.6	3.6	2.7	2.4	2.1	1.8	1.7	2.1	0.9	0.1	
Total	9	10	9	9	9	9	10	15	10	2	92

Medidas de asociación – muestran una tabla del número y porcentaje de pares de datos concordantes, discordantes y apareados, así como las estadísticas de correlaciones comunes de rangos. Estos valores miden la asociación entre las respuestas observadas y las probabilidades estimadas.

- La tabla de pares de datos concordantes, discordantes y apareados se calcula con valores de respuesta diferentes. En este caso, se tienen 70 individuos con pulso bajo y 22 con pulso alto, resultando en 70*22 = 1540 pares con diferentes valores de respuesta. Con base en el modelo, un par es concordante si el individuo con pulso bajo tiene una probabilidad más alta de tener un pulso bajo; es discordante si ocurre lo opuesto; y pareado si las probabilidades son iguales.

- En este ejemplo, el 67.9% es concordante y 29.9% son discordantes. Se pueden usar estos valores como una medición comparativa de predicción, por ejemplo al comparar valores estimados con diferentes conjuntos de predictores o con diferentes funciones de enlace.

- Se presentan resúmenes pares concordantes y discordantes de Sommers, Goodman-Krsukal Gamma y Kendall Tau-a. Estas medidas tienden a encontrarse entre 0 y 1, donde los valores más grandes indican que le modelo tien una mejor habilidad predictiva. En este ejemplo, el rango de medición de 0.14 a 0.39 implica una predictibilidad menor a la deseable.

Measures of Association:
(Between the Response Variable and Predicted Probabilities)

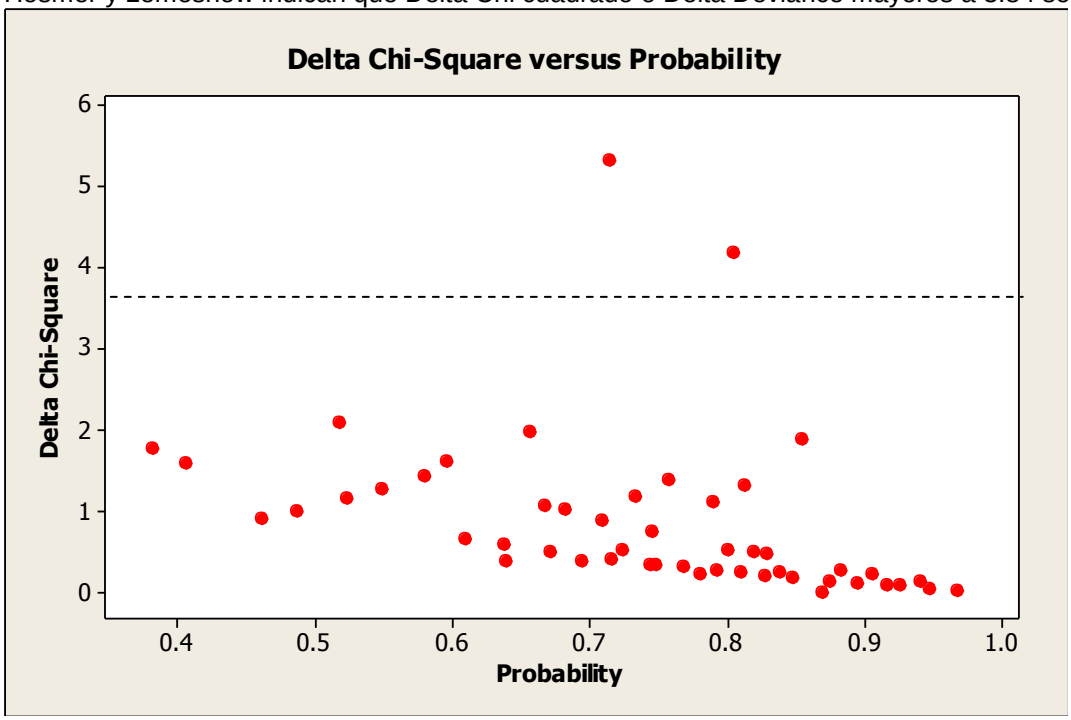
Pairs	Number	Percent	Summary Measures	
Concordant	1045	67.9	Somers' D	0.38
Discordant	461	29.9	Goodman-Kruskal Gamma	0.39
Ties	34	2.2	Kendall's Tau-a	0.14
Total	1540	100.0		

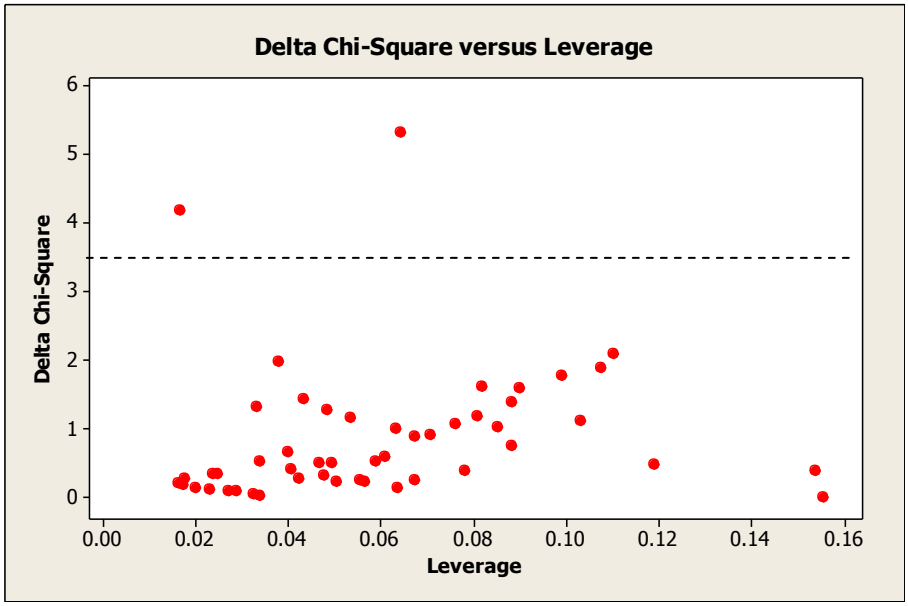
Gráficas: - En el ejemplo, se seleccionaron dos gráficas para diagnóstico, Delta Chi cuadrada de Pearson versus la probabilidad estimada del evento y Delta Pearson versus los valores influyentes.

La Delta Chi cuadrada de Pearson para el j-ésimo patrón de factor/covariado es el cambio en la Chi cuadrada de Pearson cuando se omiten todas las observaciones con ese patrón de factor/covariado.

Las gráficas indican que dos observaciones no ajustan bien en el modelo (alto Delta Chi cuadrado). Puede ser causado por un valor influeyente grande y/o un residuo alto de Pearson, que fue el caso ya que los valores influyentes fueron menores 0.1.

Hosmer y Lemeshow indican que Delta Chi cuadrado o Delta Deviance mayores a 3.84 son grandes.





Si se seleccionar Editor > Brush, se marcan los puntos, y dando clic en ellos, se identifican como valores de 31 y 66. Estos son individuos con un pulso en reposo alto, queines no fuman, y quienes tienen menos peso que el promedio (peso promedio = 116.136 libras). Se pueden hacer más investigaciones para ver por qué el modelo no se ajustó a ellos.

Corrida con SPSS
Variables

Pulsorep	String
Fuma	String
peso	Numeric

Instrucciones:
1. Analyze > Regresión > Binary Logistic
2. Seleccionar en Dependent – Pulsorep; Covariates – Fuma Peso
3. Con el botón Categorical – Fuma > Continue
4. Con boitón Options Seleccionar Calsification Plots, Hosmer Goodness of fit, CI for Exp(B) > Continue
5. OK

Exportar el reporte a Word con:
Seleccionar el reporte Output1
File > Export > seleccionar All Visible Objects y dar el nombre de archivo
OK

Cargarlo en Word y hacer comentarios:

Logistic Regression

Case Processing Summary

Unweighted Cases(a)		N	Percent
Selected Cases	Included in Analysis	92	100.0
	Missing Cases	0	.0
	Total	92	100.0
Unselected Cases		0	.0
Total		92	100.0

a If weight is in effect, see classification table for the total number of cases.

Dependent Variable Encoding

Original Value	Internal Value
Bajo	0
Alto	1

Categorical Variables Codings

		Frequency	Parameter coding (1)
FUMA	No	64	1.000
	Si	28	.000

Block 0: Beginning Block

Classification Table(a,b)					
			Predicted		
			PULSOREP		Percentage Correct
	Observed		Bajo	Alto	
Step 0	PULSOREP	Bajo	70	0	100.0
		Alto	22	0	.0
	Overall Percentage				76.1
a Constant is included in the model.					
b The cut value is .500					

Variables in the Equation							
		B	S.E.	Wald	df	Sig.	Exp(B)
Step 0	Constant	-1.157	.244	22.425	1	.000	.314

Variables not in the Equation					
			Score	df	Sig.
Step 0	Variables	FUMA(1)	3.081	1	.079
		PESO	2.721	1	.099
	Overall Statistics		7.249	2	.027

Block 1: Method = Enter

Omnibus Tests of Model Coefficients				
		Chi-square	df	Sig.
Step 1	Step	7.574	2	.023
	Block	7.574	2	.023
	Model	7.574	2	.023

Model Summary			
Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	93.640	.079	.118

Hosmer and Lemeshow Test			
Step	Chi-square	df	Sig.
1	7.561	8	.477

Contingency Table for Hosmer and Lemeshow Test						
		PULSOREP = Bajo		PULSOREP = Alto		Total
		Observed	Expected	Observed	Expected	
Step 1	1	9	8.345	0	.655	9
	2	10	9.591	1	1.409	11
	3	8	9.322	3	1.678	11
	4	7	7.379	2	1.621	9
	5	6	7.119	3	1.881	9
	6	9	6.782	0	2.218	9
	7	7	7.213	3	2.787	10
	8	6	5.419	2	2.581	8
	9	4	5.532	5	3.468	9
	10	4	3.299	3	3.701	7

Classification Table(a)					
			Predicted		
			PULSOREP		Percentage Correct
	Observed		Bajo	Alto	
Step 1	PULSOREP	Bajo	68	2	97.1
		Alto	20	2	9.1
	Overall Percentage				76.1
a The cut value is .500					

Variables in the Equation									
		B	S.E.	Wald	df	Sig.	Exp(B)	95.0% C.I.for EXP(B)	
								Lower	Upper
Step 1(a)	FUMA(1)	-1.193	.553	4.654	1	.031	.303	.103	.897

	PESO	-.025	.012	4.169	1	.041	.975	.952	.999
	Constant	3.180	1.871	2.888	1	.089	24.050		
a Variable(s) entered on step 1: FUMA, PESO.									

Step number: 1
Observed Groups and Predicted Probabilities

```
    16 ô                                                  ô
       ó                                                  ó
       ó                                                  ó
F      ó                                                  ó
R   12 ô                                                  ô
E      ó      A                                           ó
Q      ó      B                                           ó
U      ó      B                                           ó
E    8 ô      B                                           ô
N      ó      B      B                                    ó
C      ó      BA AA   B                                      ó
Y      ó      BAABA   B A B A                                  ó
     4 ô      BBBBB ABB A B A                                   ô
       ó   B B BBBBBABBB B B B      A                            ó
       ó   B B BBBBBBBBBABAB B      B                              ó
       ó B BBBBBBBBBBBBBBBBBAB BAA AB A A  B B                        ó
Predicted ÒÒÒÒÒÒÒÒÒÒÒÒÒÒÒÒÔÒÒÒÒÒÒÒÒÒÒÒÒÒÒÒÒÔÒÒÒÒÒÒÒÒÒÒÒÒÒÒÒÒÔÒÒÒÒÒÒÒÒÒÒÒÒÒÒÒÒ
 Prob:   0          .25          .5          .75          1
 Group:  BBBBBBBBBBBBBBBBBBBBBBBBBBBBBAAAAAAAAAAAAAAAAAAAAAAAAAAAAA
```

Predicted Probability is of Membership for Alto
The Cut Value is .50
Symbols: B - Bajo
A - Alto
Each Symbol Represents 1 Case.

Ejemplo con HATCO

Tomando la base de datos de HATCO donde:

X1 - Rapidez de entregas
X2 - Nivel de precios percibido
X3 - Flexibilidad en precios (para negociar)
X4 - Imagen de manufactura
X5 - Service global necesario para mantener satifacción del cliente
X6 - Imagen de la fuerza de ventas
X7 - Calidad del producto pericbida por los clientes
X8 - Tamaño de la mepresa: 1 = grande; 0 = pequeña.
X9 - Nivel de utilización, porcentaje de productos adquiridos de Hatco
X10 - Nivel de satisfacción del cliente, en las mismas unidades que las percepciones X1 a X7
X11 - Uso de especificaciones: 1 - Valor; 0-con especificaciones.
X12 - Estrutura del abastecimiento: 1-Centralizado; 0-Descentralizado
X13 - Tipo de industria: 1 - industria A; 0 - otras industrias.
X14 - Tipo de situación de compra para el cliente: 1-Nueva tarea; 2-Compra modificada; 3-Compra normal

n	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10	X11	X12	X13	X14
1	4.1	0.6	6.9	4.7	2.4	2.3	5.2	0	32	4.2	1	0	1	1

2	1.8	3	6.3	6.6	2.5	4	8.4	1	43	4.3	0	1	0	1
3	3.4	5.2	5.7	6	4.3	2.7	8.2	1	48	5.2	0	1	1	2
4	2.7	1	7.1	5.9	1.8	2.3	7.8	1	32	3.9	0	1	1	1
5	6	0.9	9.6	7.8	3.4	4.6	4.5	0	58	6.8	1	0	1	3
6	1.9	3.3	7.9	4.8	2.6	1.9	9.7	1	45	4.4	0	1	1	2
7	4.6	2.4	9.5	6.6	3.5	4.5	7.6	0	46	5.8	1	0	1	1
8	1.3	4.2	6.2	5.1	2.8	2.2	6.9	1	44	4.3	0	1	0	2
9	5.5	1.6	9.4	4.7	3.5	3	7.6	0	63	5.4	1	0	1	3
10	4	3.5	6.5	6	3.7	3.2	8.7	1	54	5.4	0	1	0	2
11	2.4	1.6	8.8	4.8	2	2.8	5.8	0	32	4.3	1	0	0	1
12	3.9	2.2	9.1	4.6	3	2.5	8.3	0	47	5	1	0	1	2
13	2.8	1.4	8.1	3.8	2.1	1.4	6.6	1	39	4.4	0	1	0	1
14	3.7	1.5	8.6	5.7	2.7	3.7	6.7	0	38	5	1	0	1	1
15	4.7	1.3	9.9	6.7	3	2.6	6.8	0	54	5.9	1	0	0	3
16	3.4	2	9.7	4.7	2.7	1.7	4.8	0	49	4.7	1	0	0	3
17	3.2	4.1	5.7	5.1	3.6	2.9	6.2	0	38	4.4	1	1	1	2
18	4.9	1.8	7.7	4.3	3.4	1.5	5.9	0	40	5.6	1	0	0	2
19	5.3	1.4	9.7	6.1	3.3	3.9	6.8	0	54	5.9	1	0	1	3
20	4.7	1.3	9.9	6.7	3	2.6	6.8	0	55	6	1	0	0	3
21	3.3	0.9	8.6	4	2.1	1.8	6.3	0	41	4.5	1	0	0	2
22	3.4	0.4	8.3	2.5	1.2	1.7	5.2	0	35	3.3	1	0	0	1
23	3	4	9.1	7.1	3.5	3.4	8.4	0	55	5.2	1	1	0	3
24	2.4	1.5	6.7	4.8	1.9	2.5	7.2	1	36	3.7	0	1	0	1
25	5.1	1.4	8.7	4.8	3.3	2.6	3.8	0	49	4.9	1	0	0	2
26	4.6	2.1	7.9	5.8	3.4	2.8	4.7	0	49	5.9	1	0	1	3
27	2.4	1.5	6.6	4.8	1.9	2.5	7.2	1	36	3.7	0	1	0	1
28	5.2	1.3	9.7	6.1	3.2	3.9	6.7	0	54	5.8	1	0	1	3
29	3.5	2.8	9.9	3.5	3.1	1.7	5.4	0	49	5.4	1	0	1	3
30	4.1	3.7	5.9	5.5	3.9	3	8.4	1	46	5.1	0	1	0	2
31	3	3.2	6	5.3	3.1	3	8	1	43	3.3	0	1	0	1
32	2.8	3.8	8.9	6.9	3.3	3.2	8.2	0	53	5	1	1	0	3
33	5.2	2	9.3	5.9	3.7	2.4	4.6	0	60	6.1	1	0	0	3
34	3.4	3.7	6.4	5.7	3.5	3.4	8.4	1	47.3	3.8	0	1	0	1
35	2.4	1	7.7	3.4	1.7	1.1	6.2	1	35	4.1	0	1	0	1
36	1.8	3.3	7.5	4.5	2.5	2.4	7.6	1	39	3.6	0	1	1	1
37	3.6	4	5.8	5.8	3.7	2.5	9.3	1	44	4.8	0	1	1	2
38	4	0.9	9.1	5.4	2.4	2.6	7.3	0	46	5.1	1	0	1	3
39	0	2.1	6.9	5.4	1.1	2.6	8.9	1	29	3.9	0	1	1	1
40	2.4	2	6.4	4.5	2.1	2.2	8.8	1	28	3.3	0	1	1	1
41	1.9	3.4	7.6	4.6	2.6	2.5	7.7	1	40	3.7	0	1	1	1
42	5.9	0.9	9.6	7.8	3.4	4.6	4.5	0	58	6.7	1	0	1	3
43	4.9	2.3	9.3	4.5	3.6	1.3	6.2	0	53	5.9	1	0	0	3
44	5	1.3	8.6	4.7	3.1	2.5	3.7	0	48	4.8	1	0	0	2
45	2	2.6	6.5	3.7	2.4	1.7	8.5	1	38	3.2	0	1	1	1
46	5	2.5	9.4	4.6	3.7	1.4	6.3	0	54	6	1	0	0	3
47	3.1	1.9	10	4.5	2.6	3.2	3.8	0	55	4.9	1	0	1	3
48	3.4	3.9	5.6	5.6	3.6	2.3	9.1	1	43	4.7	0	1	1	2
49	5.8	0.2	8.8	4.5	3	2.4	6.7	0	57	4.9	1	0	1	3

50	5.4	2.1	8	3	3.8	1.4	5.2	0	53	3.8	1	0	1	3
51	3.7	0.7	8.2	6	2.1	2.5	5.2	0	41	5	1	0	0	2
52	2.6	4.8	8.2	5	3.6	2.5	9	1	53	5.2	0	1	1	2
53	4.5	4.1	6.3	5.9	4.3	3.4	8.8	1	50	5.5	0	1	0	2
54	2.8	2.4	6.7	4.9	2.5	2.6	9.2	1	32	3.7	0	1	1	1
55	3.8	0.8	6.7	2.9	1.6	2.1	5.6	0	39	3.7	1	0	0	1
56	2.9	2.6	7.7	7	2.8	3.6	7.7	0	47	4.2	1	1	1	2
57	4.9	4.4	7.4	6.9	4.6	4	9.6	1	62	6.2	0	1	0	2
58	5.4	2.5	9.6	5.5	4	3	7.7	0	65	6	1	0	0	3
59	4.3	1.8	7.6	5.4	3.1	2.5	4.4	0	46	5.6	1	0	1	3
60	2.3	4.5	8	4.7	3.3	2.2	8.7	1	50	5	0	1	1	2
61	3.1	1.9	9.9	4.5	2.6	3.1	3.8	0	54	4.8	1	0	1	3
62	5.1	1.9	9.2	5.8	3.6	2.3	4.5	0	60	6.1	1	0	0	3
63	4.1	1.1	9.3	5.5	2.5	2.7	7.4	0	47	5.3	1	0	1	3
64	3	3.8	5.5	4.9	3.4	2.6	6	0	36	4.2	1	1	1	2
65	1.1	2	7.2	4.7	1.6	3.2	10	1	40	3.4	0	1	1	1
66	3.7	1.4	9	4.5	2.6	2.3	6.8	0	45	4.9	1	0	0	2
67	4.2	2.5	9.2	6.2	3.3	3.9	7.3	0	59	6	1	0	0	3
68	1.6	4.5	6.4	5.3	3	2.5	7.1	1	46	4.5	0	1	0	2
69	5.3	1.7	8.5	3.7	3.5	1.9	4.8	0	58	4.3	1	0	0	3
70	2.3	3.7	8.3	5.2	3	2.3	9.1	1	49	4.8	0	1	1	2
71	3.6	5.4	5.9	6.2	4.5	2.9	8.4	1	50	5.4	0	1	1	2
72	5.6	2.2	8.2	3.1	4	1.6	5.3	0	55	3.9	1	0	1	3
73	3.6	2.2	9.9	4.8	2.9	1.9	4.9	0	51	4.9	1	0	0	3
74	5.2	1.3	9.1	4.5	3.3	2.7	7.3	0	60	5.1	1	0	1	3
75	3	2	6.6	6.6	2.4	2.7	8.2	1	41	4.1	0	1	0	1
76	4.2	2.4	9.4	4.9	3.2	2.7	8.5	0	49	5.2	1	0	1	2
77	3.8	0.8	8.3	6.1	2.2	2.6	5.3	0	42	5.1	1	0	0	2
78	3.3	2.6	9.7	3.3	2.9	1.5	5.2	0	47	5.1	1	0	1	3
79	1	1.9	9.1	4.5	1.5	3.1	9.9	1	39	3.3	0	1	1	1
80	4.5	1.6	8.7	4.6	3.1	2.1	6.8	0	56	5.1	1	0	0	3
81	5.5	1.8	8.7	3.8	3.6	2.1	4.9	0	59	4.5	1	0	0	3
82	3.4	4.6	5.5	8.2	4	4.4	6.3	0	47.3	5.6	1	1	1	2
83	1.6	2.8	6.1	6.4	2.3	3.8	8.2	1	41	4.1	0	1	0	1
84	2.3	3.7	7.6	5	3	2.5	7.4	0	37	4.4	1	1	0	1
85	2.6	3	8.5	6	2.8	2.8	6.8	1	53	5.6	0	1	0	2
86	2.5	3.1	7	4.2	2.8	2.2	9	1	43	3.7	0	1	1	1
87	2.4	2.9	8.4	5.9	2.7	2.7	6.7	1	51	5.5	0	1	0	2
88	2.1	3.5	7.4	4.8	2.8	2.3	7.2	0	36	4.3	1	1	0	1
89	2.9	1.2	7.3	6.1	2	2.5	8	1	34	4	0	1	1	1
90	4.3	2.5	9.3	6.3	3.4	4	7.4	0	60	6.1	1	0	0	3
91	3	2.8	7.8	7.1	3	3.8	7.9	0	49	4.4	1	1	1	2
92	4.8	1.7	7.6	4.2	3.3	1.4	5.8	0	39	5.5	1	0	0	2
93	3.1	4.2	5.1	7.8	3.6	4	5.9	0	43	5.2	1	1	1	2
94	1.9	2.7	5	4.9	2.2	2.5	8.2	1	36	3.6	0	1	0	1
95	4	0.5	6.7	4.5	2.2	2.1	5	0	31	4	1	0	1	1
96	0.6	1.6	6.4	5	0.7	2.1	8.4	1	25	3.4	0	1	1	1
97	6.1	0.5	9.2	4.8	3.3	2.8	7.1	0	60	5.2	1	0	1	3

98	2	2.8	5.2	5	2.4	2.7	8.4	1	38	3.7	0	1	0	1
99	3.1	2.2	6.7	6.8	2.6	2.9	8.4	1	42	4.3	0	1	0	1
100	2.5	1.8	9	5	2.2	3	6	0	33	4.4	1	0	0	1

Paso 1. Obtener el comportamiento del modelo por cada variable X1 a X7:

La variable dependiente es X11:

Corrida en Minitab:
1 Abrir la hoja de trabajo HATCO.MTW o tomar datos de esta tabla.
2 Seleccionar **Stat > Regression > Binary Logistic Regression**.
3 En **Response**, seleccionar **X11** En **Model**, seleccionar **X1-X7**
4 Click **Graphs**. Seleccionar **Delta chi-square vs probability** y **Delta chi-square vs leverage**. Click **OK**.
5 Click **Results**. Seleccionar **In addition, list of factor level values, tests for terms with more than 1 degree of freedom, and 2 additional goodness-of-fit tests**. Click **OK** en cada uno de las ventanas de diálogo.
Model: Especificar los términos a ser incluidos en el modelo.

Los resultados de la corrida son los siguientes:

Binary Logistic Regression: X11 versus X1, X2, X3, X4, X5, X6, X7

Link Function: Logit
Response Information

Variable	Value	Count	
X11	1	60	(Event)
	0	40	
	Total	100	

Logistic Regression Table

						95% CI	
Predictor	Coef	SE Coef	Z	P	Odds Ratio	Lower	Upper
Constant	-1.37522	5.27926	-0.26	0.794			
X1	0.0759455	4.00067	0.02	0.985	1.08	0.00	2744.24
X2	-0.349077	4.00277	-0.09	0.931	0.71	0.00	1801.48
X3	2.21451	0.869462	2.55	0.011	9.16	1.67	50.33
X4	-2.04458	1.75315	-1.17	0.244	0.13	0.00	4.02
X5	2.63834	8.25052	0.32	0.749	13.99	0.00	1.47505E+08
X6	5.10396	2.97675	1.71	0.086	164.67	0.48	56297.08
X7	-3.39040	1.09301	-3.10	0.002	0.03	0.00	0.29

Log-Likelihood = -12.479
Test that all slopes are zero: G = 109.645, DF = 7, P-Value = 0.000

Goodness-of-Fit Tests

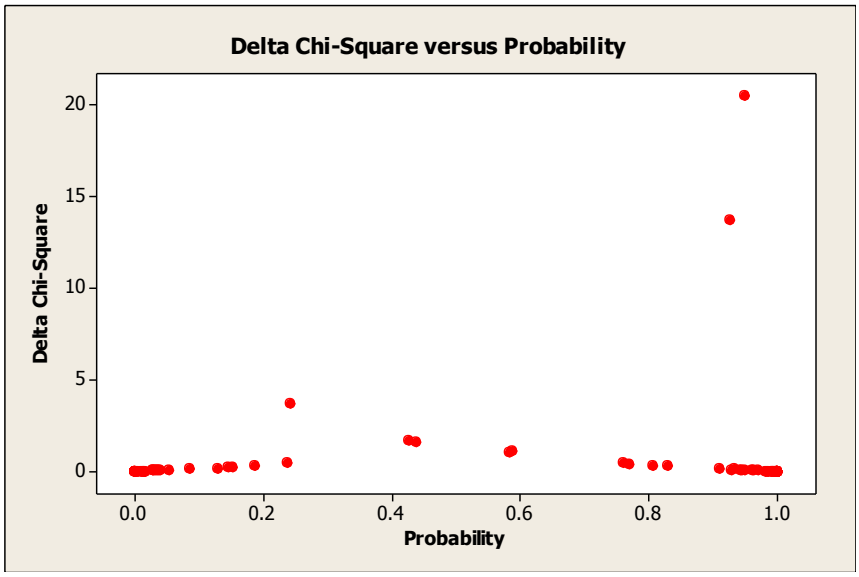
Method	Chi-Square	DF	P
Pearson	41.5472	91	1.000
Deviance	24.9571	91	1.000
Hosmer-Lemeshow	2.0928	8	0.978
Brown:			
General Alternative	2.5040	2	0.286
Symmetric Alternative	0.0018	1	0.966

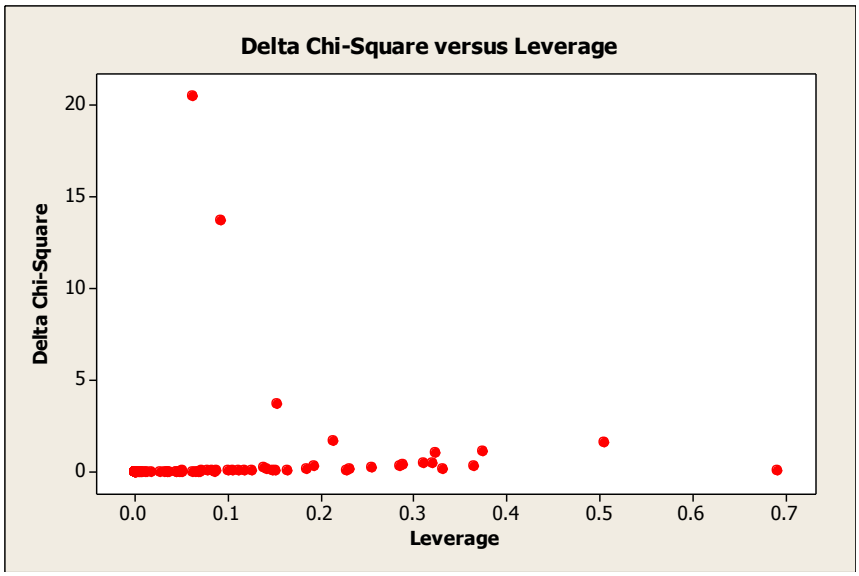
Table of Observed and Expected Frequencies:
(See Hosmer-Lemeshow Test for the Pearson Chi-Square Statistic)

		Group									
Value	1	2	3	4	5	6	7	8	9	10	Total
1											
Obs	0	0	0	2	9	9	10	10	10	10	60
Exp	0.0	0.0	0.3	2.1	8.0	9.6	9.9	10.0	10.0	10.0	
0											
Obs	10	10	10	8	1	1	0	0	0	0	40
Exp	10.0	10.0	9.7	7.9	2.0	0.4	0.1	0.0	0.0	0.0	
Total	10	10	10	10	10	10	10	10	10	10	100

Measures of Association:
(Between the Response Variable and Predicted Probabilities)

Pairs	Number	Percent	Summary Measures	
Concordant	2375	99.0	Somers' D	0.98
Discordant	25	1.0	Goodman-Kruskal Gamma	0.98
Ties	0	0.0	Kendall's Tau-a	0.47
Total	2400	100.0		





Corrida en SPSS de Hatco
Logistic Regression

Case Processing Summary

Unweighted Cases [a]		N	Percent
Selected Cases	Included in Analysis	100	100.0
	Missing Cases	0	.0
	Total	100	100.0
Unselected Cases		0	.0
Total		100	100.0

a. If weight is in effect, see classification table for the total number of cases.

Dependent Variable Encoding

Original Value	Internal Value
.00	0
1.00	1

Block 0: Beginning Block

Iteration History[a,b,c]

Iteration		-2 Log likelihood	Coefficients Constant
Step 0	1	134.603	.400
	2	134.602	.405

a. Constant is included in the model.

b. Initial -2 Log Likelihood: 134.602

c. Estimation terminated at iteration number 2 because log-likelihood decreased by less than .010 percent.

Classification Table[a,b]

Observed			Predicted		
			X11		Percentage Correct
			.00	1.00	
Step 0	X11	.00	0	40	.0
		1.00	0	60	100.0
	Overall Percentage				60.0

a. Constant is included in the model.

b. The cut value is .500

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 0	Constant	.405	.204	3.946	1	.047	1.500

Variables not in the Equation

			Score	df	Sig.
Step 0	Variables	X1	39.773	1	.000
		X2	18.312	1	.000
		X3	37.681	1	.000
		X4	.142	1	.706
		X5	4.821	1	.028
		X6	.181	1	.670
		X7	46.796	1	.000
	Overall Statistics		66.959	7	.000

Block 1: Method = Enter

Iteration History[a,b,c,d]

Iteration		-2 Log likelihood	Coefficients							
			Constant	X1	X2	X3	X4	X5	X6	X7
Step 1	1	59.008	-1.327	.842	.489	.453	-.048	-.913	.347	-.570
	2	38.779	-1.776	1.318	.850	.747	-.077	-1.409	.909	-1.126
	3	29.850	-2.073	1.594	1.054	1.109	-.251	-1.481	1.659	-1.757
	4	26.324	-1.986	1.518	.950	1.502	-.683	-.851	2.695	-2.403
	5	25.175	-1.600	.871	.356	1.887	-1.383	.811	3.969	-2.965
	6	24.965	-1.397	.216	-.226	2.149	-1.919	2.313	4.882	-3.307
	7	24.957	-1.375	.081	-.345	2.212	-2.040	2.627	5.096	-3.387
	8	24.957	-1.375	.076	-.349	2.215	-2.045	2.638	5.104	-3.390

a. Method: Enter

b. Constant is included in the model.

c. Initial -2 Log Likelihood: 134.602

d. Estimation terminated at iteration number 8 because log-likelihood decreased by less than .010 percent.

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	109.645	7	.000
	Block	109.645	7	.000
	Model	109.645	7	.000

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	24.957	.666	.900

Hosmer and Lemeshow Test

Step	Chi-square	df	Sig.
1	2.093	8	.978

Contingency Table for Hosmer and Lemeshow Test

		X11 = .00		X11 = 1.00		
		Observed	Expected	Observed	Expected	Total
Step 1	1	10	10.000	0	.000	10
	2	10	9.969	0	.031	10
	3	10	9.727	0	.273	10
	4	8	7.909	2	2.091	10
	5	1	1.965	9	8.035	10
	6	1	.368	9	9.632	10
	7	0	.059	10	9.941	10
	8	0	.002	10	9.998	10
	9	0	.000	10	10.000	10
	10	0	.000	10	10.000	10

Classification Table[a]

			Predicted		
			X11		Percentage
	Observed		.00	1.00	Correct
Step 1	X11	.00	38	2	95.0
		1.00	2	58	96.7
	Overall Percentage				96.0

a. The cut value is .500

Variables in the Equation

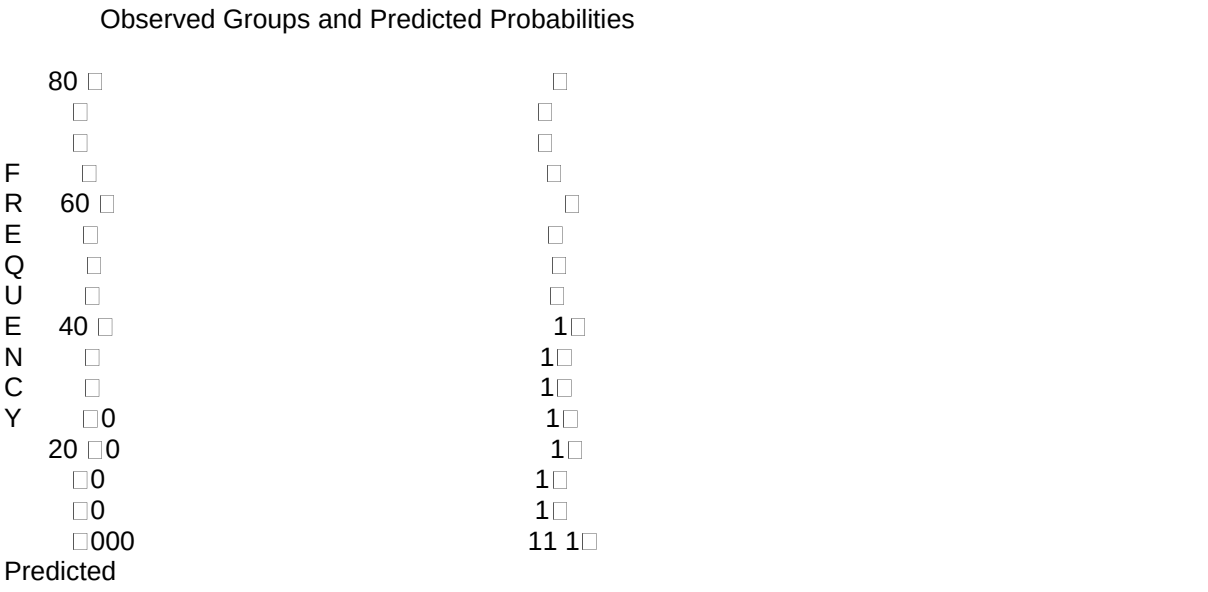
								95.0% C.I.for EXP(B)	
		B	S.E.	Wald	df	Sig.	Exp(B)	Lower	Upper
Step 1[a]	X1	.076	4.001	.000	1	.985	1.079	.000	2743.863
	X2	-.349	4.003	.008	1	.931	.705	.000	1801.224
	X3	2.215	.869	6.487	1	.011	9.157	1.666	50.331
	X4	-2.045	1.753	1.360	1	.244	.129	.004	4.021
	X5	2.638	8.251	.102	1	.749	13.990	.000	1.5E+08
	X6	5.104	2.977	2.940	1	.086	164.671	.482	56290.184
	X7	-3.390	1.093	9.622	1	.002	.034	.004	.287
	Constant	-1.375	5.279	.068	1	.794	.253		

a. Variable(s) entered on step 1: X1, X2, X3, X4, X5, X6, X7.

Correlation Matrix

		Constant	X1	X2	X3	X4	X5	X6	X7
Step 1	Constant	1.000	-.173	-.181	-.300	-.189	.146	.166	-.252
	X1	-.173	1.000	.978	-.285	.516	-.987	-.426	.235
	X2	-.181	.978	1.000	-.192	.454	-.980	-.372	.162
	X3	-.300	-.285	-.192	1.000	-.701	.309	.717	-.746
	X4	-.189	.516	.454	-.701	1.000	-.530	-.938	.631
	X5	.146	-.987	-.980	.309	-.530	1.000	.430	-.279
	X6	.166	-.426	-.372	.717	-.938	.430	1.000	-.716
	X7	-.252	.235	.162	-.746	.631	-.279	-.716	1.000

Step number: 1



Casewise List[b]

Case	Selected Status[a]	Observed X11	Predicted	Predicted Group	Temporary Variable Resid	Temporary Variable ZResid
85	S	0**	.950	1	-.950	-4.381
87	S	0**	.926	1	-.926	-3.529

a. S = Selected, U = Unselected cases, and ** = Misclassified cases.

b. Cases with studentized residuals greater than 2.000 are listed.

Ejemplo del Titanic

En 1912, se hunde el ***Titanic***, de los 2,228 pasajeros y tripulación, sólo sobrevivieron 705. Se reunió información de 1,309 pasajeros observando si sobrevivieron en función de su edad, género, tipo de boleto y número de miembros de la familia que los acompañaban.

Se investigó para tratar de determinar si había algunas variables explicativas de la supervivencia.

	pclass	survived	name	sex	age	sibsp	parch	ticket	fare	cabin	embarke
1	1	1	Allen, Miss. Elisabeth Walton	1	29	0	0	24160	211.3375	B5	1
2	1	1	Anderson, Mr. Harry	2	48	0	0	19952	26.5500	E12	1
3	1	0	Andrews, Mr. Thomas Jr	2	39	0	0	112050	.0000	A36	1
4	1	0	Artagaveytia, Mr. Ramon	2	71	0	0	PC 17609	49.5042		3
5	1	1	Aubart, Mme. Leontine Pauline	1	24	0	0	PC 17477	69.3000	B35	3
6	1	1	Barber, Miss. Ellen "Nellie"	1	26	0	0	19877	78.8500		1
7	1	1	Barkworth, Mr. Algernon Henry W	2	80	0	0	27042	30.0000	A23	1
8	1	0	Baumann, Mr. John D	2	.	0	0	PC 17318	25.9250		1
9	1	0	Baxter, Mr. Quigg Edmond	2	24	0	1	PC 17558	247.5208	B58 B60	3
10	1	1	Baxter, Mrs. James (Helene DeLa	1	50	0	1	PC 17558	247.5208	B58 B60	3
11	1	1	Bazzani, Miss. Albina	1	32	0	0	11813	76.2917	D15	3
12	1	0	Beattie, Mr. Thomson	2	36	0	0	13050	75.2417	C6	3
13	1	1	Behr, Mr. Karl Howell	2	26	0	0	111369	30.0000	C148	3
14	1	1	Bidois, Miss. Rosalie	1	42	0	0	PC 17757	227.5250		3
15	1	1	Bird, Miss. Ellen	1	29	0	0	PC 17483	221.7792	C97	1
16	1	0	Birnbaum, Mr. Jakob	2	25	0	0	13905	26.0000		3
17	1	1	Bissette, Miss. Amelia	1	35	0	0	PC 17760	135.6333	C99	1
18	1	1	Bjornstrom-Steffansson, Mr. Mau	2	28	0	0	110564	26.5500	C52	1
19	1	0	Blackwell, Mr. Stephen Weart	2	45	0	0	113784	35.5000	T	1
20	1	1	Blank, Mr. Henry	2	40	0	0	112277	31.0000	A31	3
21	1	1	Bonnell, Miss. Caroline	1	30	0	0	36928	164.8667	C7	1
22	1	1	Bonnell, Miss. Elizabeth	1	58	0	0	113783	26.5500	C103	1

Fig. 1 Características de 21 pasajeros[4]

Este problema puede ser abordado con la Regresión Logística, donde la respuesta es binaria (0,1) y no sigue una distribución normal con varianza constante.

En el modelo general:

$$E(y) = \beta_0 + \beta_1 x_1 + \ldots + \beta_q x_q$$

El valor esperado es la probabilidad de que la variable tome el valor de uno (1 = supervivencia). Para poder utilizar un modelo más general se hace una transformación logística (por ejemplo *ln(p/(1-p))*, lo que nos lleva al modelo de regresión logística:

$$\ln \frac{p}{1-p} = \beta_0 + \beta_1 x_1 + \ldots + \beta_q x_q$$

Los parámetros en la regresión logística se estiman por el método de máxima verosimilitud, en términos de p, el modelo de regresión se puede escribir como:

$$p = \frac{\exp\left(\beta_0 + \beta_1 x_1 + \ldots + \beta_q x_q\right)}{1 + \exp\left(\beta_0 + \beta_1 x_1 + \ldots + \beta_q x_q\right)}$$

En el ejemplo, “1” equivale a sobrevivió y “0” a no sobrevivió, y las cinco características de los pasajeros son:

- Pclass es la clase “1” es primera, “2” es segunda y “3” es tercera.
- Age es la edad del pasajero.
- Sex es “1” para mujeres y “1” para hombres.

[4] Landau Sabine y Everitt Brian, *Statistical Analysis USing SPSS,* Chapman & Hall/ CRC, Chicago, EEUU., 2004

- Parch, número de familiares directos padres e hijos.
- Sibsp, número de hermanos o esposa.

Las tablas de contingencia para las diferentes variables son las siguientes (*comando* ***Crosstabs…)***:

a)

Survived?* Passenger class Crosstabulation

			Passenger class			
			first class	second class	third class	Total
Survived?	no	Count	123	158	528	809
		% within Passenger class	38.1%	57.0%	74.5%	61.8%
	yes	Count	200	119	181	500
		% within Passenger class	61.9%	43.0%	25.5%	38.2%
Total		Count	323	277	709	1309
		% within Passenger class	100.0%	100.0%	100.0%	100.0%

Las proporciones de supervivencia decrecen para boletos en primera clase.

b)

Survived?* Gender Crosstabulation

			Gender		
			male	female	Total
Survived?	no	Count	682	127	809
		% within Gender	80.9%	27.3%	61.8%
	yes	Count	161	339	500
		% within Gender	19.1%	72.7%	38.2%
Total		Count	843	466	1309
		% within Gender	100.0%	100.0%	100.0%

Las proporciones de supervivencia son mayores en las mujeres que en los hombres.

d)

Survived?* Number of siblings/spouses aboard Crosstabulation

			Number of siblings/spouses aboard							
			0	1	2	3	4	5	8	Total
Survived?	no	Count	582	156	23	14	19	6	9	809
		% within Number of siblings/spouses aboard	65.3%	48.9%	54.8%	70.0%	86.4%	100.0%	100.0%	61.8%
	yes	Count	309	163	19	6	3			500
		% within Number of siblings/spouses aboard	34.7%	51.1%	45.2%	30.0%	13.6%			38.2%
Total		Count	891	319	42	20	22	6	9	1309
		% within Number of siblings/spouses aboard	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%

d)

Survived?* Number of siblings/spouses aboard Crosstabulation

			Number of siblings/spouses aboard							Total
			0	1	2	3	4	5	8	
Survived?	no	Count	582	156	23	14	19	6	9	809
		% within Number of siblings/spouses aboard	65.3%	48.9%	54.8%	70.0%	86.4%	100.0%	100.0%	61.8%
	yes	Count	309	163	19	6	3			500
		% within Number of siblings/spouses aboard	34.7%	51.1%	45.2%	30.0%	13.6%			38.2%
Total		Count	891	319	42	20	22	6	9	1309
		% within Number of siblings/spouses aboard	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%

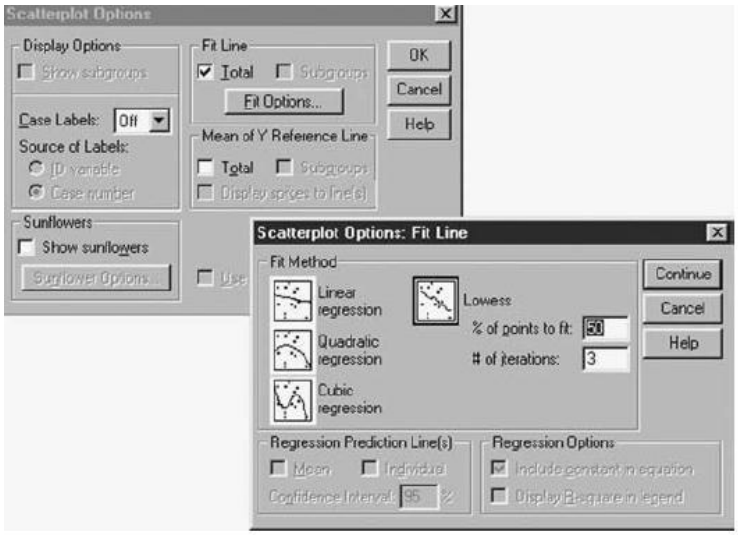
Las proporciones de supervivencia son mayores para pasajeros con un hermano o esposa o tres familiares directos (padres / hijos) con ellos.

Para examinar la asociación entre la edad y la supervivencia, se puede observar una gráfica de dispersión de dos variables, con la opción de ***Lowess curve***. La cuál proporciona una representación informal del cambio en la proporción de "1" con la edad.

_____.

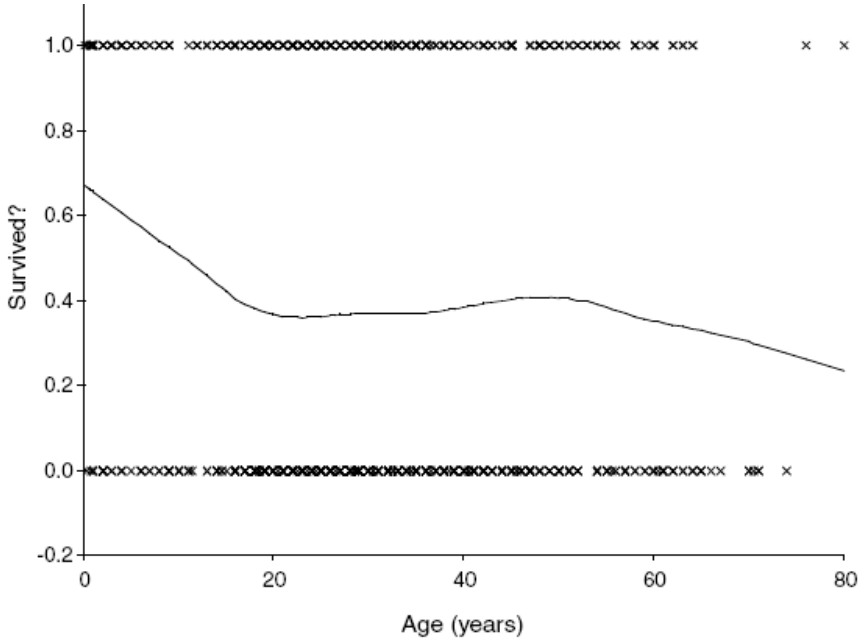
Por ejemplo al examinar las edades de las parejas que contraen matrimonio se observa que hay cierta concentración en los jóvenes, como sigue:

La curva *Lowess (locally weighted regresión fit)* permite revelar la relación entre las dos edades en vez de asumir que es lineal



2.22 Adding a Lowess curve to a scatterplot.

Para el caso que se está tratando de encontrar la relación entre edad y supervivencia se tiene:



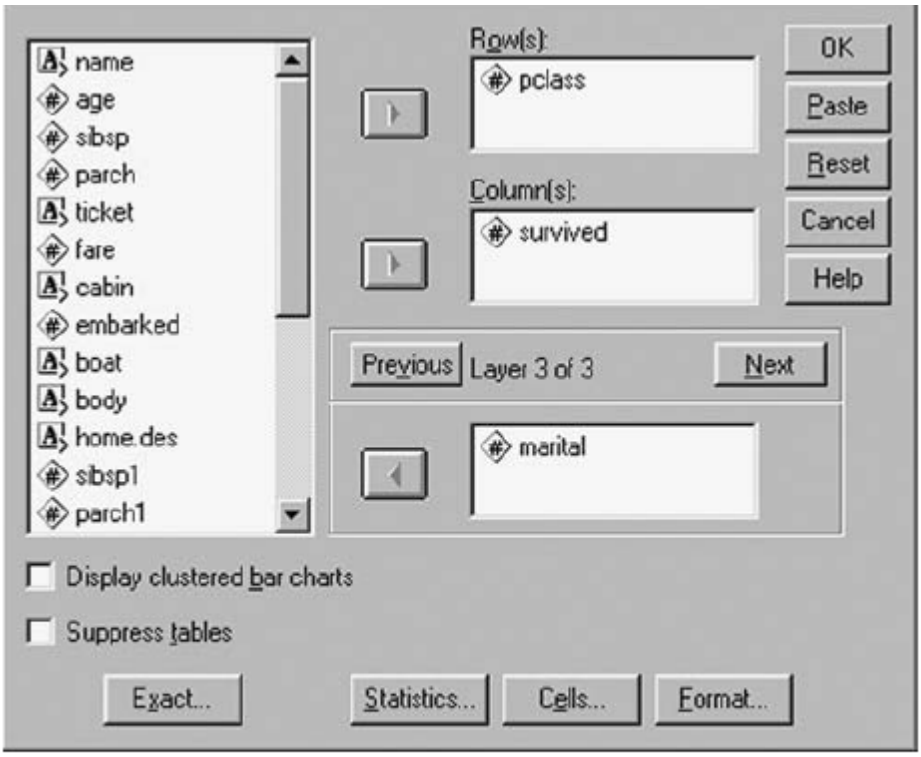
A pesar de que las tablas de contingencia y gráficas de dispersión son útiles para los análisis iniciales, no describen las posibles confusiones o interacciones entre las variables consideradas.

Haciendo un análisis de tablas de contingencia adicionales con las variables se encuentra que:

- Los hombres tienden a tener un boleto de tercera clase que las mujeres.
- Los hombres llevan menos hermanos que las mujeres.
- La mediana de edad es decreciente con la clase baja de pasajeros.
- El número de hermanos o esposa decrece con la edad.
- El número de familiares directos se incrementa con la edad.

Para clarificar la presentación de los datos, se puede hacer una clasificación múltiple de supervivencia de pasajeros dentro de estratos definidos por variables explicativas. Para lo cual se categorizar las variables edad, parch y sibsp, formando nuevas variables:

- Age_cat para categorizar a los pasajeros en niños (<21 años) y adultos (>21 años).
- Marital, para categorizar en cuatro estados civiles (1-Sin hermanos o esposa; 2-Con hermanos o esposa pero sin niños; 3- Sin hermanos o esposa pero con niños; 4- Con hermanos o esposa y además con niños). Para generar estas variables se pueden utilizar los comandos de SPSS ***Recode***, ***Compute*** e ***If Cases***. También se usa el comando ***Crosstabs*** para generar la tabla de cinco vías y ***Layer*** para indicar que forme celdas para cada combinación de las variables.



4 Defining a five-way table.

Los resultados se muestran a continuación:

Passenger class * Survived? * Gender * AGE_CAT * MARITAL Crosstabulation

% within Passenger class

MARITAL	AGE_CAT	Gender			Survived? yes
no sibs./spouse and no parents/childr.	Child	female	Passenger class	1	100.0%
				2	88.9%
				3	62.1%
		male	Passenger class	1	27.6%
				2	13.8%
				3	11.0%
	Adult	female	Passenger class	1	95.7%
				2	84.8%
				3	47.6%
		male	Passenger class	1	30.4%
				2	9.2%
				3	16.8%
sibs./spouse but no parents/childr.	Child	female	Passenger class	1	100.0%
				2	100.0%
				3	55.0%
		male	Passenger class	3	13.0%
	Adult	female	Passenger class	1	97.0%
				2	75.0%
				3	40.0%
		male	Passenger class	1	42.9%
				2	3.4%
				3	8.3%
no sibs./spouse but parents/childr.	Child	female	Passenger class	1	100.0%
				2	100.0%
				3	57.1%
		male	Passenger class	1	100.0%
				2	100.0%
				3	71.4%
	Adult	female	Passenger class	1	100.0%
				2	100.0%
				3	46.2%
		male	Passenger class	1	25.0%
				3	20.0%
sibs./spouses and parents/children	Child	female	Passenger class	1	66.7%
				2	90.9%
				3	34.2%
		male	Passenger class	1	75.0%
				2	88.9%
				3	19.2%
	Adult	female	Passenger class	1	95.2%
				2	93.8%
				3	37.5%
		male	Passenger class	1	33.3%
				2	9.1%
				3	10.0%

9.5 Tabulation of survival by passenger class, gender, age, and marital

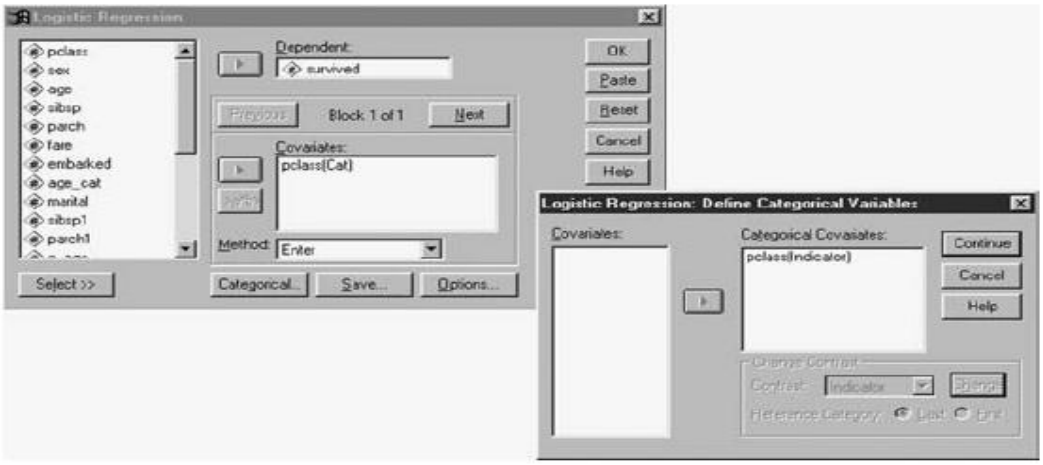
Las conclusiones del estudio indican que para los pasajeros sin hermanos o esposa o sin niños, a los cuales pertenecía el 60% de los pasajeros se observa que:

- Las mujeres con boleto de primera clase tenían una probabilidad mayor de supervivencia.

- Los hombres con boleto de tercera clase tenían menos probabilidad de sobrevivir.
- Los niños tuvieron mayor probabilidad de sobrevivir que los adultos.

Ahora se procederá a investigar las asociaciones entre la supervivencia y los cinco predictores potenciales utilizando la regresión logística con el comando:

Analyze – Regression – Binary Logistic
Se inicia incluyendo una variable a la vez para observar su efecto no ajustado, en este caso Pclass.



- La variable binaria se declara en la ventana de Dependent, y la variable explicatorio en la vantana Covariates.

- Por omisión SPSS asume que las variables explicativas se miden en una escala de intervalo. Para informar a SPSS que la variable pclass es categórica, se le indica con el botón Categorical y se incluye en la ventana Categorical Covariates. Esto hará que se generen las variables artificiales apropiadas, por omisión se generan *k-1* variables indicadoras para *k* categorías, donde el código de la categoría más alta representa la categoría de referencia, también puede cambiarse esto.

- Con el botón Options seleccionar CI for exp(B) en la ventana de diálogo, para incluir intervalos de confianza para las razones de indicadores en los resultados.

Los resultados de la codificación de la categoría de clase de boleto se muestran a continuación:

Case Processing Summary

Unweighted Cases[a]		N	Percent
Selected Cases	Included in Analysis	1309	100.0
	Missing Cases	0	.0
	Total	1309	100.0
Unselected Cases		0	.0
Total		1309	100.0

a. If weight is in effect, see classification table for the total number of cases.

Dependent Variable Encoding

Original Value	Internal Value
no	0
yes	1

Categorical Variables Codings

			Parameter coding	
		Frequency	(1)	(2)
Passenger class	first class	323	1.000	.000
	second class	277	.000	1.000
	third class	709	.000	.000

Se observa que la codificación de la variable artificial, para la variable categórica predoctora única, es (1) para primera clase, (2) para segunda clase y la tercera clase representa la categoría de referencia.

SPSS inicia con ajustar un ***null model*** vgr. Un modelo que contiene sólo un parámetro de intersección (ver Block 0: beginning block).

Classification Table [a,b]

			Predicted		
			Survived?		Percentage
	Observed		no	yes	Correct
Step 0	Survived?	no	809	0	100.0
		yes	500	0	.0
	Overall Percentage				61.8

a. Constant is included in the model.

b. The cut value is .500

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 0	Constant	-.481	.057	71.548	1	.000	.618

Variables not in the Equation

			Score	df	Sig.
Step 0	Variables	PCLASS	127.856	2	.000
		PCLASS(1)	102.220	1	.000
		PCLASS(2)	3.376	1	.066
	Overall Statistics		127.856	2	.000

9.8 Null model output from logistic regression of survival.

La primera parte de esta tabla es una “tabla de clasificación” para el modelo nulo, que compara las predicciones de supervivencia realizadas con base en el modelo ajustado con el estatus verdadero de supervivencia. Se pronostica a los pasajeros en la categoría de supervivencia si sus probabilidades son superiores a 0.05 (la cuál puede cambiarse en el diálogo Options), de manera que la proporción de no supervivencia de 0.382 está por debajo del límite de 0.5 y así el modelo calsifica a los no sobrevivientes con una exactitud del 61.8%.

A continuación la tabla de “Variables en la ecuación” proporciona la prueba de Wald para la hipótesis nula de intersección cero (o un número igual de las proporciones de supervivientes y no supervivientes). También muestra las pruebas para las variables aún no incluidas en el modelo, aquí pclass. Es claro que la supervivencia está relacionada significativamente con la clase del boleto del pasajero (Chi cuadrada = 127.9, $p < 0.001$), también se incluyen comparaciones entre las clases de pasajeros con la categoría de referencia (tercera clase).

Classification Table[a]

Observed			Predicted		
			Survived?		Percentage Correct
			no	yes	
Step 1	Survived?	no	686	123	84.8
		yes	300	200	40.0
	Overall Percentage				67.7

a. The cut value is .500

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	127.765	2	.000
	Block	127.765	2	.000
	Model	127.765	2	.000

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)	95.0% C.I.for EXP(B)	
								Lower	Upper
Step 1[a]	PCLASS			120.536	2	.000			
	PCLASS(1)	1.557	.143	117.934	1	.000	4.743	3.581	6.282
	PCLASS(2)	.787	.149	27.970	1	.000	2.197	1.641	2.941
	Constant	-1.071	.086	154.497	1	.000	.343		

a. Variable(s) entered on step 1: PCLASS.

Display 9.9 Output from logistic regression of survival on passenger class.

Los resultados anteriores muestran la “Tabla de clasificación” donde se indica que Pclass incrementa el porcentaje de clasificación correcta a 67.7%.

La tabla “Ominibus Test of Model” muestra la razón de verosimilitud (LR) o sea es una prueba para evaluar los efectos de Pclass, de nuevo se detecta un efecto significativo con Chi cuadrada = 127.8 y $p < 0.001$.

Finalmente la tabla de “Variables en la ecuación” proporciona las pruebas de Wald para todas las variables incluidas en el modelo. Consistente con las pruebas LR, el efecto de Pclass es significativo (Chi cuadrada de 120.5 con $p < 0.001$). Los parámetros estimados, son proporcionados en la columna “B” y su error estándar en “SE”. Como los efectos son difíciles de interpretar, se proporcionar en términos logarítmicos en la columna “Exp(B)”. Comparando cada clase con la tercera, se estima que las probabilidades de supervivencia fueron 4.7 veces más altas para pasajeros de primera clase (CI de 3.6 a 6.3) y 2.2 veces más altas que para la segunda clase (1.6 a 2.9). Claramente, las probabilidades de supervivencia son mayores en las dos clases superiores.

Los resultados de las otras variables categóricas explicativas consideradas individualmente se muestran a continuación, las variables sibsp y parch se recodificaron previamente en sibsp1 y parch1 dado que la supervivencia de pasajeros acompañados por muchos familiares o niños fue cero, se agruparon en una sola categoría.

Se muestra que la probabilidad de supervivencia entre pasajeros es 8.4 veces mayor para las mujeres que para los hombres.

Table 9.1 Unadjusted Effects of Categorical Predictor Variables on Survival Obtained from Logistic Regressions

Categorical Predictor	*LR Test*	*OR for Survival*	*95% CI for OR*
Passenger class (pclass)	$X^2(2) = 127.8$,		
First vs. third class	$p < 0.001$	4.743	3.581–6.282
Second vs. third class		2.197	1.641–2.941
Gender (sex)	$X^2(1) = 231.2$,		
Female vs. male	$p < 0.001$	8.396	6.278–11.229
Number of siblings/spouses aboard (sibsp1)	$X^2(3) = 14.2$, $p < 0.001$		
0 vs. 3 or more		2.831	1.371–5.846
1 vs. 3 or more		5.571	2.645–11.735
2 vs. 3 or more		4.405	1.728–11.23
Number of parents/children aboard (parch1)	$X^2(3) = 46.7$, $p < 0.001$		
0 vs. 3 or more		1.225	0.503–2.982
1 vs. 3 or more		3.467	1.366–8.802
2 vs. 3 or more		2.471	0.951–6.415

Las edades se centran en 30 años, se determinan los términos lineales, cuadráticos y cúbicos y se dividen por sus desviaciones estándar para mejor comparación.

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	16.153	3	.001
	Block	16.153	3	.001
	Model	16.153	3	.001

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)	95.0% C.I.for EXP(B)	
								Lower	Upper
Step 1[a]	C_AGE	.098	.113	.742	1	.389	1.102	.883	1.376
	C_AGE2	.235	.074	10.155	1	.001	1.265	1.095	1.461
	C_AGE3	-.339	.128	6.999	1	.008	.713	.555	.916
	Constant	-.501	.079	40.372	1	.000	.606		

[a] Variable(s) entered on step 1: C_AGE, C_AGE2, C_AGE3.

Display 9.10 LR test and odds ratios for logistic regression of survival on three ge terms.

Se observa que los términos combinados de Age tienen un efecto significativo en la supervivencia (Chi cuadrada (3) = 16.2, p = 0.001). Las pruebas de Wald indican que el modelo cuadrático y cúbico contribuyen significativamente a explicar la variabilidad en las probabilidades de supervivencia y el modelo logarítmico lineal no es suficiente.

Habiendo analizado que todos los predoctores potenciales tienen asociación con la supervivencia cuando se consideran de manera singular, el siguiente paso es estimar sus efectos simultáneamente. De esta manera, se puede estimar el efecto para cada uno, ajustado por el remanente. El modelo de regresión logística incluye en su ventana de ***Covariates***, las cuatro variables categóricas y los tres términos de edad (con el botón ***Categorical***). Los resultados se muestran a continuación:

Case Processing Summary

Unweighted Cases[a]		N	Percent
Selected Cases	Included in Analysis	1046	79.9
	Missing Cases	263	20.1
	Total	1309	100.0
Unselected Cases		0	.0
Total		1309	100.0

a. If weight is in effect, see classification table for the total number of cases.

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)	95.0% C.I.for EXP(B) Lower	Upper
Step 1[a]	PCLASS			85.200	2	.000			
	PCLASS(1)	2.159	.234	85.195	1	.000	8.661	5.476	13.699
	PCLASS(2)	.874	.206	17.950	1	.000	2.397	1.600	3.591
	SEX(1)	2.550	.176	210.422	1	.000	12.812	9.077	18.083
	C_AGE	-.413	.156	6.955	1	.008	.662	.487	.899
	C_AGE2	.195	.106	3.363	1	.067	1.215	.987	1.497
	C_AGE3	-.200	.165	1.474	1	.225	.819	.592	1.131
	SIBSP1			18.725	3	.000			
	SIBSP1(1)	2.166	.510	18.031	1	.000	8.722	3.210	23.700
	SIBSP1(2)	2.008	.516	15.152	1	.000	7.449	2.710	20.475
	SIBSP1(3)	1.541	.653	5.562	1	.018	4.668	1.297	16.798
	PARCH1			5.525	3	.137			
	PARCH1(1)	.518	.552	.880	1	.348	1.679	.569	4.955
	PARCH1(2)	1.031	.583	3.126	1	.077	2.805	.894	8.801
	PARCH1(3)	.844	.623	1.838	1	.175	2.326	.686	7.884
	Constant	-4.989	.755	43.661	1	.000	.007		

a. Variable(s) entered on step 1: PCLASS, SEX, C_AGE, C_AGE2, C_AGE3, SIBSP1, PARCH1.

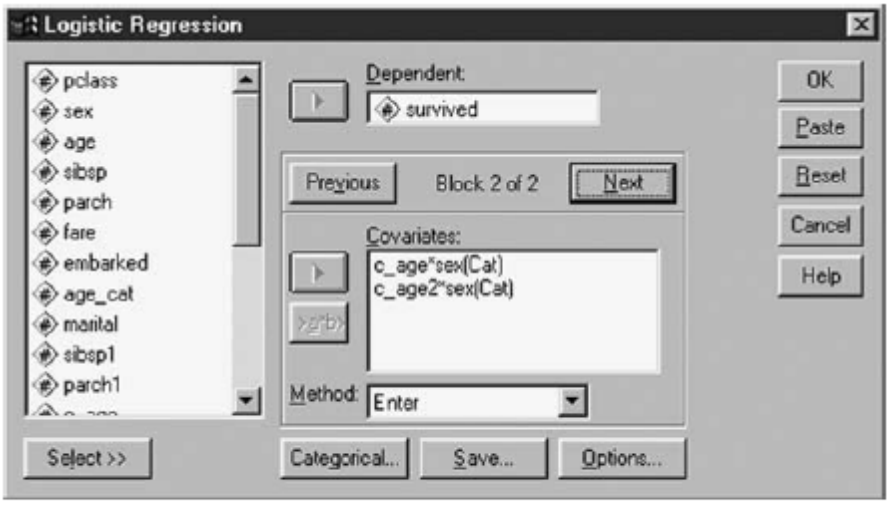
Display 9.11 Selected output from logistic regression of survival on all potential explanatory variables.

Se puede notar que de la tabla "Case Processing Summary", los casos incluidos en el análisis se reduce a 1046 dado que falta información en la variable de edad para 263 pasajeros.

La tabla "Ómnibus.." proporciona el efecto de todas las variables explicativas simultáneamente, la guía de la significancia son las pruebas de Wald. En esta corrida se observa que la variable Patch1 no contribuye a la explicación de las probabilidades de supervivencia, una vez que se introducen las otras variables, de manera que se excluye del modelo y se hace una nueva corrida, donde ahora el tercer término de la edad no es necesario.

El modelo final de efectos principales contiene términos de edad, clase del boleto, género, y número de hermanos/esposas, cada contribuye significativamente a un nivel del 5% después de ajustar los otros términos del modelo.

Ahora se prueban los términos de interacción de dos vías, una por una, por medio de la opción de bloqueo para agregar los términos de interacción de interés, a los efectos principales significativos identificados previamente. Por ejemplo para Age y Sex:



9.12 Adding interaction effects to a logistic regression model.

Un término de interacción se puede definir en la ventana de Logistic Regresión, seleccionando las variables involucradas y el botón *>a*b>* para crear términos de interacción.

Los resultados se indica como sigue:

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	23.869	2	.000
	Block	23.869	2	.000
	Model	477.239	10	.000

LR test output for age by gender interaction effect on survival

El primer término permite que el efecto del término lineal de Age varie con Sex, la segunda hace lo mismo con el término cuadrático y Age.

Se procede a analizar las otras interacciones.

De la tabla siguiente se observa que se deben incluir en el modelo las interacciones entre: género y clase de boleto; género y edad; clase de boleto y número de hermanos/esposa; y edad y número de hermanos/esposa. Si se considera el 10% también se debe incluir este último término.

Table 9.2 LR Test Results for Assessing Two-Way Interaction Effects on Survival Probabilities

Interaction Involving		*LR Test*		
Variable 1	*Variable 2*	*Deviance Change*	*DF*	*p-Value*
pclass	sex	52.8	2	<0.001
c_age, c_age2	sex	23.9	2	<0.001
sibsp2	sex	0.84	2	0.66
c_age, c_age2	pclass	7.4	4	0.12
sibsp2	pclass	16.6	4	0.002
c_age, c_age2	sibsp2	9.7	4	0.045

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)	95.0% C.I.for EXP(B)	
								Lower	Upper
Step 1[a]	PCLASS			14.394	2	.001			
	PCLASS(1)	3.974	1.102	13.003	1	.000	53.214	6.136	461.499
	PCLASS(2)	1.736	.950	3.338	1	.068	5.673	.881	36.521
	SEX(1)	1.831	.280	42.854	1	.000	6.238	3.606	10.793
	C_AGE	-.870	.128	45.865	1	.000	.419	.326	.539
	C_AGE2	.343	.105	10.769	1	.001	1.409	1.148	1.730
	SIBSP2			19.103	2	.000			
	SIBSP2(1)	1.992	.456	19.102	1	.000	7.333	3.001	17.919
	SIBSP2(2)	1.734	.485	12.782	1	.000	5.663	2.189	14.651
	PCLASS * SEX			28.309	2	.000			
	PCLASS(1) by SEX(1)	1.911	.615	9.644	1	.002	6.757	2.023	22.567
	PCLASS(2) by SEX(1)	2.390	.478	24.957	1	.000	10.915	4.273	27.878
	C_AGE by SEX(1)	.423	.214	3.900	1	.048	1.526	1.003	2.322
	C_AGE2 by SEX(1)	-.233	.199	1.369	1	.242	.792	.536	1.171
	PCLASS * SIBSP2			8.409	4	.078			
	PCLASS(1) by SIBSP2(1)	-2.500	1.110	5.070	1	.024	.082	.009	.723
	PCLASS(1) by SIBSP2(2)	-1.975	1.138	3.011	1	.083	.139	.015	1.291
	PCLASS(2) by SIBSP2(1)	-1.953	.975	4.013	1	.045	.142	.021	.959
	PCLASS(2) by SIBSP2(2)	-1.758	1.027	2.928	1	.087	.172	.023	1.291
	Constant	-3.920	.479	67.024	1	.000	.020		

[a] Variable(s) entered on step 1: PCLASS * SEX , C_AGE * SEX , C_AGE2 * SEX , PCLASS * SIBSP2 .

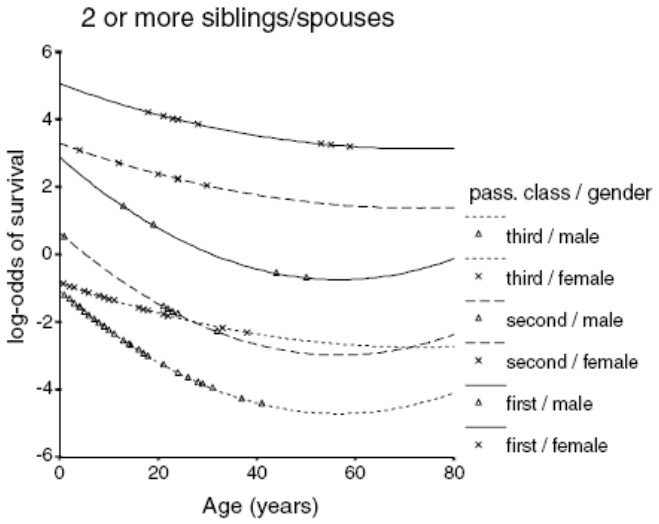
Display 9.14 Parameter estimates from final model for survival probabilities.

Como un medio alterno para interpretar el modelo logístico de ajuste, se obtienen gráficas de las probabilidades logarítmicas de la supervivencia, dado que el modelo asume efectos aditivos de las variables explicativas en esta escala.

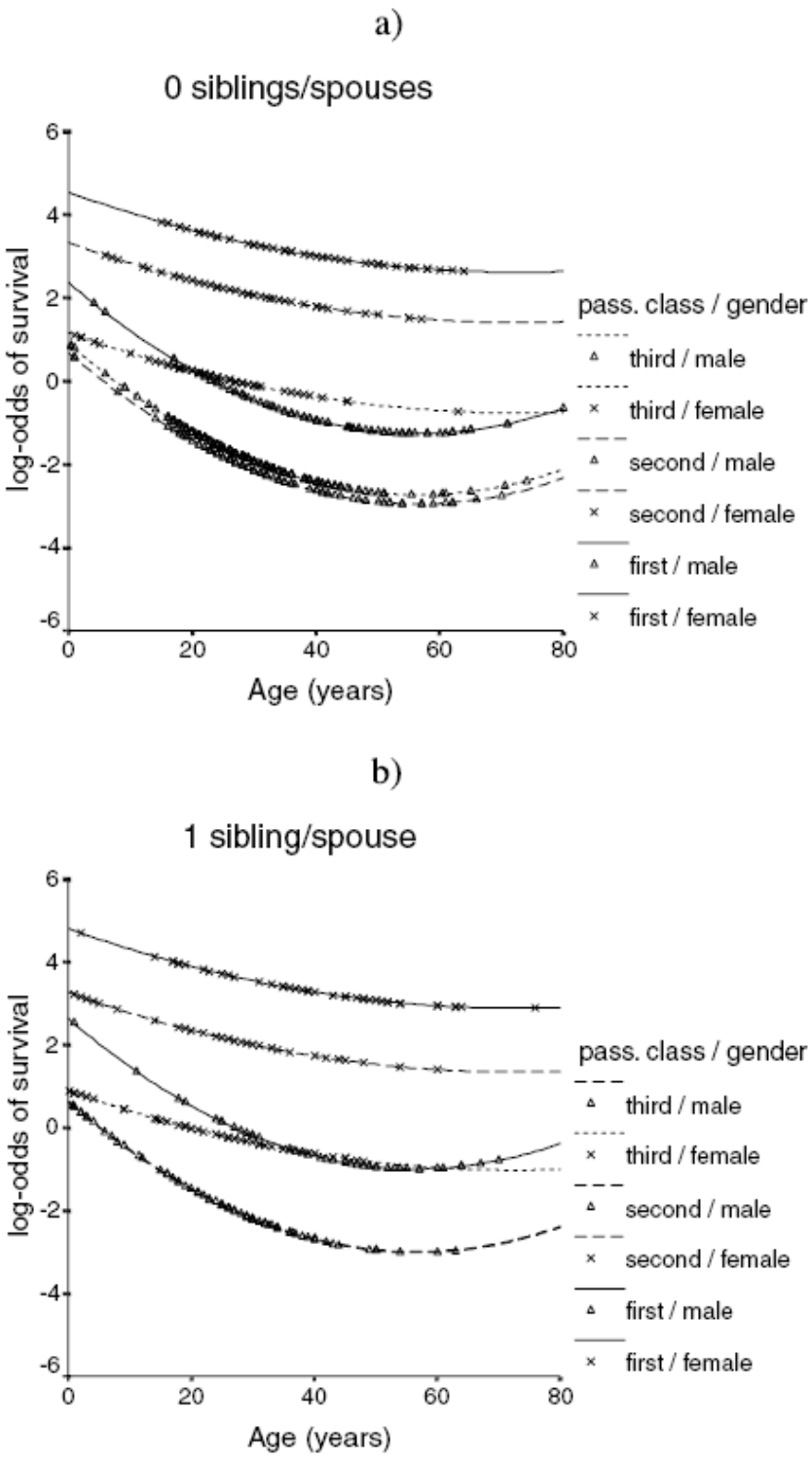
Las instrucciones son las siguientes:

- Guardar las probabilidades de supervivencia como una nueva variable pre_1, en la vista de Datos, seleccionado ***Predicted Values:Probabilities*** en la ventana ***Save New Variables*** cuando se obtenga el modelo de regresión final.
- Transformar estos valores en posibilidades usando la fórmula odds = pre_1/(1-Pre_1) y calcular la variable logarítmica con la fórmula ln_odds= ln(odds).

- Generar un factor de interacción clase y género (class.se) con ***Compute Numeric Expresión*** 100 x pclass + 1 x Sex. Resultará en un factor con 6 niveles, cada uno con tres dígitos: el primero indica la clase; el intermedio es cero; y el último indica el género.
- Usar el comando ***Split File*** para organizar la salida en grupos definidos por sibsp2.
- Usar el comando Simple Scatterplot para producir una gráfica de dispersión de ln_odds contra la edad con marcadores definidos por class.se.



Display 9.15 (continued).



Display 9.15 Log-odds of survival predicted by the final logistic model by passenger age, gender, ticket class, and number of accompanying siblings/spouses. The symbols indicate predictions for passenger characteristics observed in the sample, the curves extrapolated values.

- ***Predictores identificados:*** cada una de las variables, edad del pasajero, género, clase de boleto, y número de hermanos/esposa, hacen una contribución independiente a la predicción de las posibilidades de supervivencia. Quienes tienen mayores posibilidades son: los jóvenes (<20 años), mujeres, en primera clase. Los que tienen menos posibilidades son: los de tercera clase, adultos acompañados de dos o más hermanos/esposa.
- ***Interacción edad por género:*** Las posibilidades de supervivencia son mayores para mujeres que para hombres conforme se tiene mayor edad.
- ***Interacción de género por clase de boleto:*** Las posibilidades de supervivencia de las mujeres sobre los hombres se incrementa con la clase.

9. REGRESIÓN LOGÍSTICA ORDINAL

La regression logística ordinal realiza una regresión con una variable de respuesta ordinal. Las variables ordinales son variables categóricas que tienen tres o más niveles posibles con un orden natural, tal como fuertemente en desacuerdo, desacuerdo, de acuerdo, y fuertemente de acuerdo. Un modelo con uno o más predictores se ajusta usando un algoritmo iterativo de mínimos cuadrados reponderado, para obtener los estimados de los parámetros por máxima verosimilitud.

Se asumen líneas de regresión paralelas, y por tanto, se determina una sóla pendiente para cada covariado. En situaciones donde este supuesto no es válido, la regresión logística nominal es más apropiada, ya que genera funciones logit separadas.

Ejemplo:

Suponiendo que un biólogo cree que la población adulta de salamandras en el Norte se ha hecho más pequeña durante los últimos años. Se quiere determinar si existe alguna asociación entre el tiempo que vive una salamandra recien nacida y el nivel de toxicidad del agua, así como si hay un efecto regional. El tiempo de supervivencia se codifica como sigue: 1 si es <10 días; 2 = 10 a 30 días; 3 = 31 a 60 días.

Supervivencia	Region	NivelToxico	Supervivencia	Region	NivelToxico
1	1	62.00	2	1	40.50
1	2	46.00	2	2	60.00
2	1	48.50	3	1	57.50
3	2	32.00	2	1	48.75
2	1	63.50	2	1	44.50
1	1	41.25	1	1	49.50
2	2	40.00	2	2	33.75
3	1	34.25	2	1	43.50
2	1	34.75	2	2	48.00
1	2	46.25	3	1	34.00
2	1	43.50	1	1	50.00
2	2	46.00	3	2	35.00
2	1	42.50	1	1	49.00
1	2	53.00	2	2	43.50
1	2	43.50	3	2	37.25
1	1	56.00	3	2	39.00
2	1	40.00	3	1	34.50
1	2	48.00	2	1	47.50
2	1	46.50	1	2	42.00
2	2	72.00	2	2	45.50
2	2	31.00	2	2	38.50
1	1	48.00	2	1	36.50
2	2	36.50	2	2	37.50
2	2	43.75	3	1	38.50

2	1	34.25	2	2	47.00
2	1	41.25	2	2	39.75
2	2	41.75	1	1	60.00
2	2	45.25	2	2	41.00
2	1	43.50	2	1	41.00
2	2	53.00	3	1	30.00
3	1	38.00	2	2	45.00
2	2	59.00	2	2	51.00
2	1	52.50	2	2	35.25
2	2	42.75	1	2	40.50
2	2	31.50	2	2	39.50
2	2	43.50	3	2	36.00
2	2	40.00			

Instrucciones de Minitab
1 Open worksheet EXH_REGR.MTW.
2 Seleccionar **Stat > Regression > Ordinal Logistic Regression**.
3 En **Response**, seleccionar **Survival**. En **Model**, seleccionar **Region ToxicLevel**. En **Factors** (**optional**), seleccionar **Region**.
4 Click **Results**. Seleccionar **In addition, list of factor level values, and tests for terms with more than 1 degree of freedom**. Click **OK** en cada ventana de diálogo.
Los resultados se muestran a continuación:
Results for: Exh_regr.MTW

Ordinal Logistic Regression: Supervivencia versus Region, NivelToxico

Link Function: Logit

Información de respuesta: muestra el número de observaciones que caen dentro de cada una de las categorías de respuesta. Abajo se muestran los valores ordenados de la respuesta de menor a mayor. 1 corresponde a <10 días; 2 = 10 a 30 días; y 3 = 31 a 60 días.

Información de factores: muestra todos los factores en el modelo, el número de niveles para cada factor, y los valores de los niveles del factor. El nivel del factor que ha sido designado como el nivel de referencia, es el primer dato en Valores. En este caso *Región 1*.

Niveles de Referencia para los factores
Se requiere asignar un nivel de factor como el nivel de referencia. Los coeficientes estimados se interpretan respecto a este nivel de referencia. Minitab asigna el nivel de referencia como sigue dependiendo del tipo de datos:

- Para factores numéricos, el nivel de referencia es el valor con el menor valor numérico.
- Para fechas, el nivel de referencia es el nivel con la fecha/hora más antigua.
- Para factores de texto, el nivel de referencia es el nivel que está primero en orden alfabético.

Se puede cambiar esta configuración de Default en la ventana de diálogo de ***Options.*** Para cambiar el nivel de referencia de un factor, especificar la variable del factor seguida por el nuevo nivel de referencia en la ventana **Reference factor level.** Se puede especificar niveles de referencia para más de un factor al mismo tiempo. Si todos los niveles son texto o fecha/hora, encerrarlos entre comillas. Si ya se definió un valor de orden para un factor de texto, la regla por omisión es que se designa el primer valor en el orden definido como valor de referencia.
La regression logística crea un conjunto de variables de diseño para cada uno de los factores en el Modelo. Si hay k niveles, habrá k-1 variables de diseño y el nivel de referencia será codificado con cero. Por ejemplo:

Factor A with 4 levels (1 2 3 4)

reference level is 1		A1	A2	A3
	1	0	0	0
	2	1	0	0
	3	0	1	0
	4	0	0	1

Factor B with 3 levels (Humidity Pressure Temp)

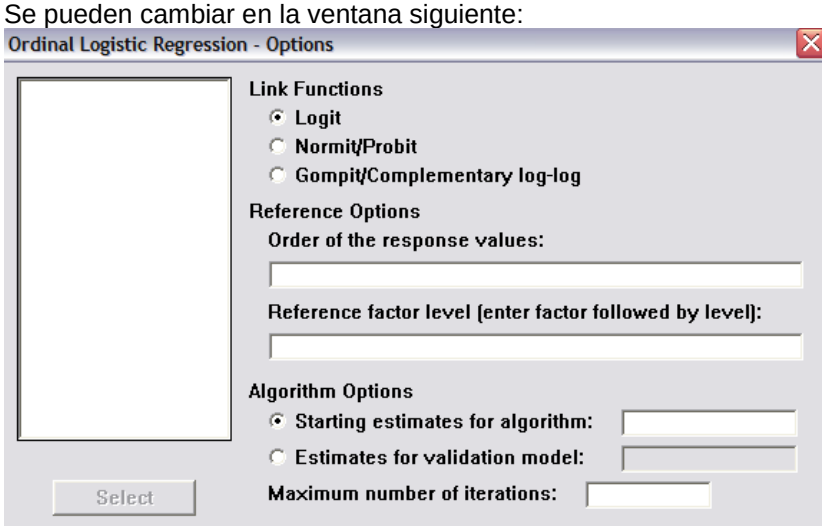
reference level is Humidity		B1	B2
	Humidity	0	0
	Pressure	1	0
	Temp	0	1

Nivel de referencia para la variable de respuesta

Minitab asigna el nivel de referencia como sigue dependiendo del tipo de datos:

- Para factores numéricos, el nivel de referencia es el valor con el mayor valor numérico.
- Para fechas, el nivel de referencia es el nivel con la fecha/hora más reciente.
- Para factores de texto, el nivel de referencia es el nivel que es último en orden alfabético.

Se pueden cambiar en la ventana siguiente:



Response Information

Variable	Value	Count
Supervivencia	1	15
	2	46
	3	12
	Total	73

Factor Information

Factor	Levels	Values
Region	2	1, 2

Tabla de regression logística: muestra los coeficientes estimados, el error estándar de los coeficientes, los valores Z, los valores p. Cuando se utiliza la función de enlace logit, se muestran las tasas de posibilidades calculadas, y un intervalo de confianza del 95% para las tasas de posibilidades.

- Los valores etiquetados Const(1) y Const(2) son intersecciones estimadas para las funciones logit de probabilidad acumuladas de supervivencia para <10 días, y para 10-30 días respectivamente.

- El coeficiente de 0.2015 para la región es el cambio estimado en la función logit acumulativa del tiempo de supervivencia cuando la región es 2 comparada con la región 1, con el covariado Nivel Toxico mantenido constante. Dado que el coeficiente estimado es 0.685, no hay suficiente evidencia de que la región tenga un efecto sobre el tiempo de supervivencia.
- Hay un coeficiente estimado para cada covariado, que da líneas paralelas para el nivel del factor. En este caso, el coeficiente estimado para un covariado simple, Nivel Toxico, es 0.121, con un valor p < 0.0005. El valor p indica que para la mayoría de niveles alfa, hay evidencia suficiente para concluir que el nivel de toxicidad afecta la supervivencia. El coeficiente positivo, y una tasa de posibilidades mayor a uno, indica que los niveles de toxicidad más altos tienden a estar asociados con menores valores de superviviencia. Específicamente, un incremento de una unidad en la toxicidad del agua resulta en un 13% de incremento en las posibilidades que la salamadra viva menos o igual a 10 días contra más de 30 días, y que la salamandra viva menos que o igual a 30 días versus más que 30 días.
- Se muestra la verosimilitud logarítmica (log Likelihood) de las iteraciones de máxima verosimilitud junto con el estadístico G. Este estadístico prueba la hipótesis que todos los coeficientes asociados con los predictores son iguales a cero versus al menos un coeficiente no es cero. En este caso G = 14.713 con un valor p de 0.001, indicando que hay suficiente evidencia para concluir que al menos uno de los coeficientes estimados es diferente de cero.

Logistic Regression Table

					Odds	95% CI	
Predictor	Coef	SE Coef	Z	P	Ratio	Lower	Upper
Const(1)	-7.04343	1.68017	-4.19	0.000			
Const(2)	-3.52273	1.47108	-2.39	0.017			
Region							
2	0.201456	0.496153	0.41	0.685	1.22	0.46	3.23
NivelToxico	0.121289	0.0340510	3.56	0.000	1.13	1.06	1.21

Log-Likelihood = -59.290
Test that all slopes are zero: G = 14.713, DF = 2, P-Value = 0.001

Prueba de bondad de ajuste: muestra tanto las pruebas de Pearson como deviance. En este ejemplo para Pearson se tiene un valor P de 0.463, y para la prueba de deviance es 0.918, indicando que no hay suficiente evidencia para afirmar que el modelo no ajusta los datos adecuadamente. Si el valor P es menor que el nivel de alfa seleccionado, la prueba rechaza la hipótesis de que el modelo ajusta los datos adecuadamente.

Goodness-of-Fit Tests

Method	Chi-Square	DF	P
Pearson	122.799	122	0.463
Deviance	100.898	122	0.918

Medidas de asociación: muestra una tabla de los números y porcentajes de parejas concordantes, discordantes y similares, y estadísticas de correlación de rango común. Estos valores miden la asociación entre las respuestas observadas y las probabilidades estimadas o pronosticadas.
- La tabla de pares concordantes, discordantes y similares, se calcula emparejando las observaciones con diferentes valores de respuestas. Si se tienen 15 1's, 46 2's, y 12 3's, resultan en 15 x 46 + 15 x 12 + 46 x 12 = 1422 pares de diferentes valores de respuesta. Para pares incluyendo los valores de respuesta codificados menores (1-2 y 1-3 pares de valores en el ejemplo), un par es concordante si la probabilidad acumualtiva hasta el valor de respuesta más bajo (aquí 1) es mayor para la observación con el valor más bajo. De manera similar para otros pares. Para pares con respuestas 2 y 3, un par es concordante si la probabilidad acumulativa hasta 2 es mayor para la observación codificada como 2. El par es discordante si ocurre lo opuesto. El par es similar si las probabilidades son iguales. En este caso, 79.3% de pares son concordantes, 20.3% son discordantes, y 0.5% son similares. Se pueden usar estos valores como medida comparativa de predicción, por ejemplo para evaluar predictores de diferentes funciones de enlace.

- Se muestran resúmenes de pares concordantes y discordantes de Somers'D, Goodman-Kruskal Gamma y la Tau-a de Kendall. Los números tienen el mismo numerador: el número de pares concordantes menos el número de pares discordantes. El denominador es el número total de pares con Somers'D, el número total de pares excepto los similares con Goodman-Kruskal Gamma, y el número de todas las posibles observaciones para la Tau-a de Kendall. Estas medidas tienden a estar entre 0 y 1 donde los valores mayores indican una mejor capacidad predictiva del modelo.

Measures of Association:
(Between the Response Variable and Predicted Probabilities)

Pairs	Number	Percent	Summary Measures	
Concordant	1127	79.3	Somers' D	0.59
Discordant	288	20.3	Goodman-Kruskal Gamma	0.59
Ties	7	0.5	Kendall's Tau-a	0.32
Total	1422	100.0		

10. REGRESIÓN LOGÍSTICA NOMINAL

Usar la regression logística nominal para realizar regresión sobre una variable de respuesta nominal, usando un algoritmo iterativo de mínimos cuadrados reponderados, para obtener la estimación de máxima verosimilitud de los parámetros.

Las variables nominales son variables categóricas que tienen tres o más niveles posibles, sin un orden natural. Por ejemplo, los niveles en un estudio de gusto por la comida, puede incluir: crujiente, fresca y firme *(crunchy, mushy, and crispy).*

Ejemplo:

Suponiendo que un director de escuela se interesa por identificar la materia favorita de los niños, como se asocia con su edad o con el método de enseñanza empleado. Se toman 30 niños, de 10 a 13 años, con clases de ciencias, matemáticas, y lenguaje, que emplean ya sea técnicas de enseñanza de exposición o discusión. Al final del año escolar, se les preguntó por su materia favorita. Se usa la regresión logística nominal porque la respuesta es categórica pero no tiene un órden implícito.
Los datos considerados son los siguientes:

Materia	MetodoEnseñanza	Edad
Matemáticas	Discusión	10
Ciencias	Discusión	10
Ciencias	Discusión	10
Matemáticas	Exposición	10
Matemáticas	Discusión	10
Ciencias	Exposición	10
Matemáticas	Discusión	10
Matemáticas	Exposición	11
Artes	Exposición	11
Ciencias	Discusión	11
Artes	Exposición	11
Matemáticas	Discusión	11
Ciencias	Exposición	11
Ciencias	Discusión	11
Artes	Exposición	11
Ciencias	Exposición	12
Ciencias	Exposición	12
Ciencias	Discusión	12
Artes	Exposición	12
Matemáticas	Discusión	12

Matemáticas	Discusión	12
Artes	Exposición	12
Artes	Discusión	13
Matemáticas	Discusión	13
Artes	Exposición	13
Artes	Exposición	13
Matemáticas	Discusión	13
Ciencias	Discusión	13
Matemáticas	Exposición	13
Artes	Exposición	13

Instrucciones de Minitab:
1 Open worksheet EXH_REGR.MTW.
2 Seleccionar **Stat > Regression > Nominal Logistic Regression**.
3 En **Response**, seleccionar **Subject**. En **Model**, seleccionar **TeachingMethod Age**. En **Factors** (**optional**), seleccionar **TeachingMethod**.
4 Click **Results**. Seleccionar **In addition, list of factor level values, and tests for terms with more than 1 degree of freedom**. Click **OK** en cada ventana de diálogo.
Los resultados se muestran a continuación:

Nominal Logistic Regression: Materia versus MetodoEnseñanza, Edad

Información de respuesta: muestra el número de observaciones que caen dentro de cada una de las categorías de respuesta (ciencias, matemáticas y artes del lenguaje).

Response Information

Variable	Value	Count	
Materia	Matemáticas	11	(Reference Event)
	Ciencias	10	
	Artes	9	
	Total	30	

Información de factores: muestra todos los factores en el modelo, el número de niveles para cada factor, y los valores de los niveles del factor. El nivel del factor que ha sido designado como el nivel de referencia, es el primer dato en Valores. Aquí, el esquema de codificación de default define el nivel de referencia como *Discusión* usando el orden alfabético.

Factor Information

Factor	Levels	Values
MetodoEnseñanza	2	Discusión, Exposición

Tabla de regression logística: muestra los coeficientes estimados, el error estándar de los coeficientes, los valores Z, los valores p. Cuando se utiliza la función de enlace logit, se muestran las tasas de posibilidades calculadas, y un intervalo de confianza del 95% para la tasa de posibilidades. El coeficiente asociado con un predictor es el cambio estimado en la función logia con el cambio de una unidad en el predictor, asumiendo que todos los otros factores y covariados permanecen constantes.
- Si hay k respuestas distintas, Minitab estima k-1 conjuntos de parámetros estimados, denominados Logia(1) y Logia (2). Estas son diferencias estimadas en logaritmo de posibilidades o logias de matemáticas y artes de lenguaje, respectivamente, comparado con la ciencia como el evento de referencia. Cada conjunto contiene una constante y coeficientes para los factores, aquí el método de enseñanza, y el covariado edad. El coeficiente del método de enseñanza es el cambio estimado en el

Logit cuando el método de enseñanza sea exposción comparado a cuando sea discusión, manteniendo la edad constante. El coeficiente de la edad es el cambio estimado en el logit con un año de incremento en edad manteniendo constante el método de enseñanza. Estos conjuntos de estimados de parámetros dan líneas no paralelas para los valores de respuesta.

- El primer conjunto de logiats estimados, etiquetados como Logia(1), son los parámetros estimados del cambio en Logias de matemáticas respecto al evento de referencia, ciencia. Como el valor p tiene valores de 0.548 y 0.756 para el método de enseñanza y edad, indica que hay insuficiente evidencia para concluir que un cambio en el método de enseñanza de discusión a exposición, o en edad afecten la selección de materia favorita cuando se compara con la ciencia.
- El segundo conjunto de logias estimados, Logia(2), son los parámetros estimados del cambio en Logias de artes del lenguaje respecto al evento de referencia ciencia. Los valores p de 0.044 y 0.083 para método de enseñanza y edad, respectivamente, indica que hay suficiente evidencia, si los valores p son menores al valor aceptable de alfa, se concluye que la selección favorece a la ciencia.
- El coeficiente positivo del método de enseñanza indica que los estudiantes que se les aplica el método de enseñanza de exposición, prefieren las artes del lenguaje sobre la ciencia comparado a estudiantes que se les da un método de enseñanza de discusión. La tasa estimada de posibilidades de 15.96 indica que las posibilidades de seleccionar el lenguaje sobre la ciencia es de alrededor de 16 veces más alto para los estudiantes, cuando el método de enseñanza cambia de discusión a lectura. El coeficiente positivo asociado con la edad indica que los estudiantes tienden a preferir las artes del lenguaje sobre las ciencias confoirme se hacen más maduros.

Logistic Regression Table

					Odds	95% CI
Predictor	Coef	SE Coef	Z	P	Ratio	Lower
Logit 1: (math/science)						
Constant	-1.12266	4.56425	-0.25	0.806		
TeachingMethod						
lecture	-0.563115	0.937591	-0.60	0.548	0.57	0.09
Age	0.124674	0.401079	0.31	0.756	1.13	0.52
Logit 2: (arts/science)						
Constant	-13.8485	7.24256	-1.91	0.056		
TeachingMethod						
lecture	2.76992	1.37209	2.02	0.044	15.96	1.08
Age	1.01354	0.584494	1.73	0.083	2.76	0.88

Predictor	Upper
Logit 1: (math/science)	
Constant	
TeachingMethod	
lecture	3.58
Age	2.49
Logit 2: (arts/science)	
Constant	
TeachingMethod	
lecture	234.91
Age	8.66

Log-Likelihood: de las iteraciones de máxima verosimilitud junto con el estadístico G. G es la diferencia en -2 log-likelihood (-2LL) para un modelo el cual sólo tiene los términos de la constante y el modelo ajustado indicado en la Tabla de la Regresión logística. G prueba la hipótesis nula que los coeficientes asociados con los predictores son iguales a cero versus que no todo son cero. G = 12.825 con un valor p de 0.012, indican que para alfa = 0.05, hay evidencia suficiente que al menos uno de los coeficientes es diferente de cero.

Log-Likelihood = -26.446
Test that all slopes are zero: G = 12.825, DF = 4, P-Value = 0.012

Prueba de bondad de ajuste: muestra tanto las pruebas de Pearson como deviance. En este ejemplo para Pearson se tiene un valor P de 0.730, y para la prueba de deviance es 0.640, indicando que no hay suficiente evidencia para afirmar que el modelo no ajusta los datos adecuadamente. Si el valor P es menor que el nivel de alfa seleccionado, la prueba rechaza la hipótesis de que el modelo ajusta los datos adecuadamente.

Goodness-of-Fit Tests

Method	Chi-Square	DF	P
Pearson	6.95295	10	0.730
Deviance	7.88622	10	0.640

BIBLIOGRAFÍA

- Montgomery, Doglas C., Peck, Elizabeth A., Introduction to Linear Regression Analysis, John Wiley and Sons, 2º edition, Inc., New York, 1992
- Chatterjee, Samprit, Price, Bertram, Regression Analysis by Example, John Wiley and Sons, Inc., 2º edition, 1991
- Draper, Norman R., Smith, Harry, Applied Regression Analysis, John Wiley and Sons, Inc., New York, 1998

TAREA NO. 1 DE ANALISIS DE REGRESIÓN
Con apoyo de Minitab

11/11/00

PROBLEMA 2.1

Calcular lo siguiente (Y vs X8):

a) La recta de regresión

The regression equation is
Y = 21.8 - 0.00703 X8

b) La tabla ANOVA y prueba de significancia

Analysis of Variance

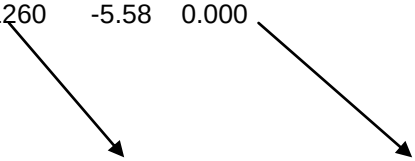
Source	DF	SS	MS	F	P
Regression	1	178.09	178.09	31.10	0.000
Residual Error	26	148.87	5.73		
Total	27	326.96			

Ftablas=F1,26,0.05=4.23

Nota: Como p = 0 equivale a Fc > F tablas y se rechaza la Ho: Beta1 = 0
quiere decir que existe la recta de regresión

c) El intervalo de confianza al 95%
de la pendiente b1

Predictor	Coef	StDev	T	P
Constant	21.788	2.696	8.08	0.000
X8	-0.007025	0.001260	-5.58	0.000



El intervalo de confianza para **β1** *se calcula como sigue:*
t0.025,26 = 2.056
*b1 ± t*std dev (Predict.X8) =-0.007025 ± 2.056* (0.00126) =*
-0.0096 <= **β1** <= *-0.004435;*

El iuntervalo de confianza para β0 es:
*b0 ± t*std dev (Constant) =21.788 ± 2.056* (2.696);*

d) % de la variabilidad explicada por la
regresión

R-Sq = 54.5%

e) El intervalo de confianza a un 95% para la media
del valor estimado de Y, cuando Xo = 2000 yardas (corresponde a CI).
Predicted Values
Fit StDev Fit 95.0%CI para media 95.0% PI p.valor futuro
7.738 0.473 *(6.766; 8.710)* (2.724; 12.752)

f) Probar la hipótesis nula de que el coeficiente de
correlación es cero. Ho: $\rho = 0$

$$t_0 = \frac{0.738234\sqrt{26}}{\sqrt{1-0.545}} = 5.58055$$ Ttablas 0.025,26 = 2.056

Cómo to > ttablas, se rechaza Ho. Es decir que ρ es diferente de cero.

g) Probar la hipótesis nula de que el coeficiente de
correlación es Ho: $\rho 0 = -0.80$
Zo = -0.76006 Z tablas = Z0.025 = 1.96
Cómo Zo < |Ztablas| no hay evidencia suficiente para rechazar Ho

h) Encontrar el intervalo de confianza del 95% para ρ.
- 0.87134<= ρ <= - 0.50396

i) Con Minitab construir las sig. gráficas de residuos
y comentar acerca de la adecuación del modelo
- Gráfica de probabilidad normal
- Gráfica de residuos contra Yi est.
- Gráfica de residuos contra Xi8..

Los residuos muestran una variación normal con varianza constante
j) Graficar los residuos contra el porcentaje de juegos
ganados X7i, ¿se mejora el modelo agregando esta
variable?.

No se mejora la distribución de los residuos

The regression equation is
Y = 17.9 - 0.00654 X8 + 0.048 X7

S = 2.432 R-Sq = 54.8% R-Sq(adj) = 51.1%

Al agregar la nueva variable X7, el modelo no mejora realmente (comparar R^2)

PROBLEMA 2.2
Si las yardas ganadas se limitan a 1800. Hallar el intervalo de predicción al 90% en el número de juegos ganados (corresponde a PI).

t(0.05,26) = 1.705616 Alfa = 0.1

Intervalo 8.1238 <= Ymedia <=10.16 4.936<=Ypuntual<=13.35

PROBLEMA 2.3
Calcular lo siguiente:

a) La recta de regresión
The regression equation is
Y1 = 607 - 21.4 X4

b) La tabla ANOVA y prueba de significancia
Analysis of Variance

Source	DF	SS	MS	F	P
Regression	1	10579	10579	69.61	0.000
Residual Error	27	4103	152		
Total	28	14682			

Ftablas=F1,27,.05=4.21

Como Fc=69.61 es mayor que Ftablas=4.21, se rechaza Ho y existe la regresión

c) El intervalo de confianza al 99%
de la pendiente β1

Predictor	Coef	StDev	T	P
Constant	607.10	42.91	14.15	0.000
X4	-21.402	2.565	-8.34	0.000

El intervalo de confianza para β1 se calcula como sigue:
t0.005,27 = 2.771 *7.1076*
*b1 ± t*std dev (Predict.X4) =-21.402 ± 2.771* (2.565) =*
-28.5096 <= **β1** <= *-14.2943*

d) % de la variabilidad explicada por la regresión R^2

R-Sq = 72.1% R-Sq(adj) = 71.0%

e) El intervalo de confianza a un 95% para la media del valor estimado de Y, cuando Xo = 16.5 (corresponde a CI).
Predicted Values

Fit	StDev Fit	95.0% CI para media	95.0% PI p.valor futuro
253.96	2.35	*(249.15; 258.78)*	(228.21; 279.71)

f) Probar la hipótesis nula de que el coeficiente de correlación es cero. Ho: ρ = 0

$$t_0 = \frac{0.84882\sqrt{27}}{\sqrt{1-0.7205}} = 8.3427$$ Ttablas 0.025,27 = 2.052

Cómo to > Ttablas, se rechaza Ho. Es decir que ρ es diferente de cero.

g) Probar la hipótesis nula de que el coeficiente de correlación es $\rho 0 = -0.80$.
Zo = 0.78172 Z tablas = Z0.025 = 1.96
Cómo Zo < |Ztablas| no hay evidencia suficiente para rechazar Ho

h) Encontrar el intervalo de confianza del 95% para ρ.
- 0.927 <= ρ <= - 0.7

i) Con Minitab construir las sig. gráficas de residuos y comentar acerca de la adecuación del modelo
- Gráfica de probabilidad normal
- Gráfica de residuos contra Yi est.
- Gráfica de residuos contra Xi4.

Unusual Observations

Obs	X4	Y1	Fit	StDev Fit	Residual	St Resid
22	17.6	254.50	229.99	3.28	24.51	2.06R
24	19.1	181.50	199.39	6.44	-17.89	-1.70 X
25	16.5	227.50	253.75	2.34	-26.25	-2.17R

R denotes an observation with a large standardized residual
X denotes an observation whose X value gives it large influence.

Los residuos no muestran una distribución aleatoria

PROBLEMA 2.7
a) Ecuación de regresión
The regression equation is
Y78 = 77.9 + 11.8 X78

b) Probar la hipótesis nula de que Ho: $\beta 1 = 0$

Analysis of Variance

Source	DF	SS	MS	F	P
Regressio	1	148.31	148.31	11.47	0.003
Residual error	18	232.83	12.94		
Total	19	381.15			

Ftablas = F0.05,1,18=4.41

Cómo Fc > F tablas se rechaza la hipótesis Ho, implicando $\beta 1 \neq 0$

c) Calcular R^2

R-Sq = 38.9%

d) Encontrar el intervalo de confianza al 95% para la pendiente:

Predictor	Coef	StDev	T	P

Constant 77.863 4.199 18.54 0.000
X78 11.801 3.485 3.39 0.003

t0.025,18 = 2.101
*b1 ± t*std dev (Predict.X78) =11.801 ± 2.101* (3.485) =*
4.47699 <= **β1** <= *19.12301*

e) Encontrar el intervalo de confianza para la pureza media si el % de hidrocarbono es de 1.00

Predicted Values

Fit StDev Fit 95.0% CI p. la media 95.0% PI p. valor futuro
89.664 1.025 *(87.510; 91.818)* (81.807; 97.521)

PROBLEMA 2.8
a) ¿Cuál es la correlación entre las dos variables?

R-Sq = 38.9% entonces r = 0.6237

b) Probar la Hipótesis nula Ho: $\rho = 0$

$$t_0 = \frac{0.6237\sqrt{18}}{\sqrt{1-0.389}} = 3.38527$$ Ttablas 0.025,18 = 2.101

Cómo to > Ttablas, se rechaza Ho. Es decir que ρ es diferente de cero.

c) Contruir un intervalo de confianza del 95% para ρ.

0.25139 <= ρ <= 0.8356

PROBLEMA 2.9
a) Ecuación de regresión

The regression equation is
Y9 = - 6.33 + 9.21 X9

b) Probar la significancia de la regresión

Analysis of Variance

Source	DF	SS	MS	F	P
Regressi	1	280590	280590	74122.78	0.000
Residual error	10	38	4		
Total	11	280627			

Como el valor de p es cero, se rechaza la hipótesis Ho: $\beta 1 = 0$, por tanto existe la regresión.

c) Si se incrementa la temperatura ambiente promedio en un grado, el consumo de vapor se incrementa en 10 unidades. ¿se soporta esta afirmación?.

Column Mean
Mean of X9 = 46.500; se incrementa en un grado

Predicted Values

Fit StDev Fit 95.0% CI 95.0% PI
421.862 0.562 (420.610; 423.113) (417.350; 426.374)
431.070 0.563 (429.816; 432.324) (426.557; 435.583)

Por los resultados observados se cumple la afirmación

d) Intervalo de predicción con un 99% de nivel de confianza para Xo = 58.

Predicted Values

Fit StDev Fit 99.0% CI 99.0% PI
527.759 0.683 (525.593; 529.925) *(521.220; 534.298)*

PROBLEMA 2.10
a) Encontrar el coeficiente de correlación r

R-Sq = 100.0% por tanto r = 1

b) Probar la Hipótesis nula Ho: $\rho = 0$

$$t_0 = \frac{0.999\sqrt{10}}{\sqrt{1-0.999}} = 272.25$$ Ttablas 0.005,10 = 1.812

Cómo to > Ttablas, se rechaza Ho. Es decir que ρ es diferente de cero.

c) Contruir un intervalo de confianza del 95% para ρ.

0.99 <= ρ <= 0.999

FÓRMULAS DE REGRESIÓN LINEAL MÚLTIPLE
Modelos de Regresión Múltiple

Asumiendo que N observaciones de la respuesta se tiene:

$$Y_u = \beta_0 + \beta_1 X_{u1} + \beta_2 X_{u2} + ... + \beta_k X_{uk} + \varepsilon_u \qquad (3.1)$$

Para N observaciones el modelo en forma matricial es:

Y = X β + ε = [1 : D] β + ε (3.2)

k es el número de variables independientes o regresores
Y es un vector N x 1.
X es una matriz de orden N x (k + 1), donde la primera columna es de 1's.
β es un vector de orden (k + 1) x 1.
ε es un vector de orden N x 1.
D es la matriz de Xij con i = 1, 2, ..., N; j = 1, 2, ..., k

Se trata de encontrar el vector de estimadores de mínimos cuadrados **b** que minimicen:

$$S(\beta) = \sum_{i=1}^{n} \varepsilon_i^2 = \varepsilon'\varepsilon = (Y - X\beta)'(Y - X\beta)$$

quedando

X'X b = X' Y (3.4)

A) VECTOR DE ESTIMADORES DE MINIMOS CUADRADOS **b** de β

$$\mathbf{b = (X'X)^{-1}\ X'Y} \qquad (3.5)$$

B) VARIANZAS Y COVARIANZAS DE **b**

$$\text{Var}\mathbf{(b) = C = (X'X)^{-1}\ \sigma^2} \qquad (3.6)$$

El elemento *(ii)* de esta matriz $c_{ii}\sigma^2 = Var(b_i)$ es la varianza del elemento **bi** .

El error estándar de b_i es la raíz cuadrada positiva de la varianza de b_i o sea:

$$se.b_i = \sqrt{c_{ii}\sigma^2} \qquad (3.7)$$

La covarianza del elemento b_i y b_j de **b** es $Co\,\text{var}(c_{ij}) = c_{ij}\sigma^2$. (3.8)

La desviación estándar se estima como sigue:

$SSE = \sum (Y_i - \hat{Y})^2 = \sum_{i=1}^{n} e_i^2 = e'e$; con p = k +1 parámetros del modelo se tiene:

$$SSE = (Y - Xb)'(Y - Xb) = Y'Y - b'X'Y - Y'Xb + b'X'Xb = Y'Y - 2b'X'Y + b'X'Xb$$

$$s^2 = MSE = \frac{SSE}{N - p} \qquad (3.15)$$

C) INTERVALO DE CONFIANZA PARA LOS COEFICIENTES β_j

Con intervalo de confianza 100(1 - α)% , para *j = 0, 1, ..., k* es:

$$b_j - t_{\alpha/2,n-p}\,se(b_j) \le \beta_j \le b_j + t_{\alpha/2,n-p}\,se(b_j) \qquad (3.17)$$

Donde *se(b_j)* es el *error estándar* del coeficiente de regresión b_j.

$$se(b_j) = \sqrt{S^2 C_{jj}} \qquad (3.18)$$

Siendo C_{jj} el j-ésimo elemento de la matriz $(X'X)^{-1}$.

D) INTERVALO DE CONFIANZA PARA LA RESPUESTA MEDIA Yo en Xo

El intervalo de confianza para el 100(1 - α) % es:

$$\hat{Y}_0 - t_{\alpha/2,n-p}\sqrt{S^2 X'_0 (X'X)^{-1} X_0} \le Y_0 \le \hat{Y}_0 + t_{\alpha/2,n-p}\sqrt{S^2 X'_0 (X'X)^{-1} X_0} \qquad (3.21)$$

E) TABLA ANOVA PARA LA REGRESIÓN

$H_0 : \beta_1 = \beta_2 = ... = \beta_k = 0$; $H_a : \forall \beta_i \ne 0, i = 1,2,...,k$

Ho se rechazará si Ft >= Fo

Fuente de variación	SS	df	MS	F_0
Regresión	SSR	k= p-1	MSR	MSR/MSE
Residuos	SSE	n–k–1= N-p	MSE	Ft=$F_{\alpha,p-1,N-p}$
Total	SST=SSR+SSE	n – 1=k+(n-k+1)		

Donde:

$SST = \sum_{u=1}^{N} (Y_u - \bar{\bar{Y}})^2$ con N-1 grados de libertad (3.24)

$$SSR = \sum_{u=1}^{N} (\hat{Y}(x_u) - \bar{\bar{Y}})^2$$ con p (parámetros) – 1 grados de libertad (3.25)

$$SSE = \sum_{u=1}^{N} (Y_u - \hat{Y}(x_u))^2$$ con (N-1) – (p –1) grados de libertad (3.26)

En forma matricial se tiene:

$$SST = Y'Y - \frac{(1'Y)^2}{N} \quad (3.27)$$

$$SSR = b' X'Y - \frac{(1'Y)^2}{N} \quad (3.28)$$

$$SSE = Y'Y - b' X'Y$$

F) PRUEBA DE LA SIGNIFICANCIA DE LOS COEFICIENTES INDIVIDUALES BETAx

$H_0 : \beta_j = 0 \qquad H_1 : \beta_j \neq 0$

Si no se rechaza Ho quiere decir que el regresor Xj puede ser excluido del modelo,

Ho es rechazada si $|t_0| > t_{\alpha/2,n-k-1}$, donde:

$$t_0 = \frac{b_j}{se(b_j)}$$

G) INTERVALO DE PREDICCIÓN PARA LA RESPUESTA Yo en Xo

El intervalo de confianza para el 100(1 - α) % es:

$$\hat{Y}_0 - t_{\alpha/2,n-p}\sqrt{S^2(1+X'_0(X'X)^{-1}X_0)} \leq Y_0 \leq \hat{Y}_0 + t_{\alpha/2,n-p}\sqrt{S^2(1+X'_0(X'X)^{-1}X_0)}$$

FORMULAS

Methods

Model fitting

Minitab uses the nonlinear iterative partial least squares (NIPALS) algorithm developed by Herman Wold [36] to solve problems associated with ill-conditioned data. PLS reduces the number of predictors by extracting uncorrelated components based on the covariance between the predictor and response variables. PLS is similar to principal components regression and ridge regression, but varies in its computational method. For a detailed comparison of these techniques, see [12].

The PLS algorithm produces a sequence of models, where each consecutive model contains one additional component. Components are calculated one at a time, starting with the standardized x- and y-matrix. Subsequent components are calculated from the x- and y-residual matrix; iterations stop upon reaching the maximum number of components or when x-residuals become the zero matrix. If the number of components equals the number of predictors, the PLS model equals the least squares regression model. Cross-validation is used to identify the number of components that minimizes prediction error.

PLS performs decomposition on both predictors and responses simultaneously. After Minitab determines the number of components and calculates the loadings, it calculates the regression coefficients for each predictor. For more detailed information on PLS and NIPALS, see [14], [20], and [23].

Cross-validation Calculates the predictive ability of potential models to help you determine the appropriate number of components to retain in your model. When the data contain multiple response variables, Minitab validates the components for all responses simultaneously. For more information, see [15].

Cross-validation procedure

For each potential model, Minitab:

1. Omits one observation or group of observations, depending on the cross-validation method you use.
2. Recalculates the model without the observation/group of observations.
3. Predicts the response, or the cross-validated fitted value, for the omitted observation/group of observations using the recalculated model and calculates the cross-validated residual value.
4. Repeats steps 1−3 until all observations have been omitted and fit.
5. Calculates the prediction sum of squares (PRESS) and predicted R^2 values.

After performing steps 1−5 for each model, Minitab selects the model with the number of components that produces the highest predicted R^2 and lowest PRESS. With multiple response variables, Minitab selects the model with the highest average predicted R^2 and lowest average PRESS.

Back to top

Model selection

R^2 (R-sq) Coefficient of determination; indicates how much variation in the response is explained by the model. The higher the R^2, the better the model fits your data. The formula is:

$$1 - \frac{\text{SS Error}}{\text{SS Total}}$$

Another presentation of the formula is:

$$\frac{\text{SS Regression}}{\text{SS Total}}$$

R^2 can also be calculated as the Correlation $(Y, \hat{Y})^2$.

Sum of squares (SS) The sum of squared distances. SS Total is the total variation in the model. SS Regression is the portion of the variation explained by the model, while SS Error is the portion not explained by the model and is attributed to error. The calculations are:

$$\text{SS Regression} = \sum (\hat{y} - \bar{y})^2$$

$$\text{SS Error} = \sum (y - \hat{y})^2$$

$$\text{SS Total} = \sum (y - \bar{y})^2$$

where y = observed response, $\hat{y}$ = fitted response, and $\bar{y}$ = mean response.

PRESS

The prediction sum of squares (PRESS) statistic assesses your model's predictive ability. PRESS, similar to the residual sum of squares, is the sum of squares of the prediction error. In PLS, Minitab only calculates PRESS if you cross-validated the model.

Minitab calculates PRESS in the following steps:

1. Minitab recalculates the model as many times as there are observations, omitting a different observation each time. For each omitted observation, Minitab calculates the fitted or predicted response using the model.
2. Minitab subtracts the predicted value from the observed response value. This is the true prediction error because the observation fit is independent of the model.
3. Once Minitab conducts this routine for all observations, Minitab calculates PRESS using the formula:

$$\sum_{i=1}^{n} (y_i - \hat{y}_{(i)})^2$$

where y_i = observed response, $\hat{y}_{(i)}$ = fitted response for the omitted observation, and n = number of observations.

In general, the smaller the PRESS value, the better the model's predictive ability. PRESS is used to calculate the predicted R^2.

Predicted R^2

Similar to R^2 in least squares regression. In PLS, Minitab only calculates predicted R^2 if you cross-validated the model.

The formula for predicted R^2 is:

$$1 - \frac{\text{PRESS}}{\text{SS Total}}$$

where PRESS = prediction sum of squares and SS Total = total sum of squares.

Predicted R^2 indicates how well the model predicts responses for new observations, whereas R^2 indicates how well the model fits your data. Predicted R^2 can prevent overfitting the model, that is, fitting the model too closely to the data in the current data set, so it is not useful for predicting new data. Predicted R^2 can be more useful than adjusted R^2 for comparing models because it is calculated with observations that are not included in model calculation.

Predicted R^2 is between 0 and 1 and is calculated from the PRESS statistic. Larger values of predicted R^2 suggest models of greater predictive ability.

Test R^2

Indicates how well the PLS model predicts your test data. The test R^2 represents the proportion of variation in the responses that is explained by the predictors in your test data set. Generally, test data is used to validate the fitted model and must include the same number of predictors as the original data set. The test R^2 can only be calculated if the test data includes response data for each observation. The test R^2 is calculated in the same way as R^2 with the formula:

$$1 - \frac{\text{SS Error}}{\text{SS Total}}$$

Back to top

Model information

Coefficients and standardized coefficients

Coefficients are the parameters in a regression equation. The estimated coefficients are used with the predictors to calculate the fitted value of the response variable and the predicted response of new observations. In contrast to least squares, the PLS coefficients are non-linear estimators. Standardized coefficients indicate the importance of each predictor in the model and correspond to the standardized x- and y-variables. In PLS, the coefficient matrix (dimension p x r, where p = number of predictors and r = number of responses) is calculated from the weights and loadings.

The formula for standardized coefficients is:

$$\beta(\text{std}) = \mathbf{W} (\mathbf{P}' \mathbf{W})^{-1} C'$$

where W = x-weight matrix, P = x-loading matrix, and C = y-loading matrix.

To calculate nonstandardized coefficients and intercept, use the formulas:

$$\beta_{(j,k)} = \beta(\text{std})_{(j,k)} * \text{StDev}(y)_{(k)} / \text{StDev}(x)_{(j)}$$

$$\beta o_k = \bar{y}_k - \Sigma (\bar{x}_j \beta_{j,k})$$

where j = predictors (1, p) and k = responses (1, r).

Leverages

In least squares regression, values that indicate how far the corresponding observations are from the center of the x-space, which is described by the x-values. In PLS, the predictors are replaced by x-scores. Observations with high leverage have x-scores far from zero and have a significant influence on the regression coefficients. Points with high leverage are outliers in the x-space, but are not necessarily outliers in the y-space.

The leverage values in PLS are calculated from the x-score matrix T. The formula to calculate leverage values is:

$h_i = \mathbf{T} (\mathbf{T}' \mathbf{T})^{-1} \mathbf{T}' + 1 / n$ where i = observations (1, n).

A leverage value greater than 2m / n, where m = number of components, is considered high and should be examined.

Distances from the x-model

A measure of how well observations are fitted in the x-space; indicates how well the x-scores describe observations. An observation with a large distance may also be a leverage point.

The formula for calculating the distance from the x-model for the i^{th} observation is:

$$\sqrt{\frac{\sum_{m=1}^{M} (t_{i,m})^2}{p - M + 1}}$$

where m = number of components (1, M), t = x-score, and p = number of predictors.

Distances from the y-model

A measure of how well observations are fitted in the y-space; indicates how well the y-scores describe observations. An observation with a large distance may also be an outlier.

The formula for calculating the distance from the y-model for the i^{th} observation is:

$$\sqrt{\frac{\sum_{m=1}^{M} (u_{i,m})^2}{r}}$$

Back to top

where m = number of components (1, M), u = y-score, and r = number of responses.

Fits and residuals

Fitted values

The predicted Y or $\hat{Y}$; the mean response value for the given predictor values using the estimated regression equation.

Cross-validated fitted values

Indicate how well your model predicts data. These values are similar to ordinary fitted values, which indicate how well your model fits the data. Cross-validated fitted values in PLS and least squares regression are conceptually similar, but their calculations differ:

- In PLS, the cross-validated fitted values are the predicted responses for the observations in your data set, calculated individually so the observation can be excluded from the model used to calculate the predicted response. The cross-validated fitted values are calculated during cross-validation and the values vary based on how many observations are omitted each time the model is recalculated.
- In least squares regression, the cross-validated fitted values are calculated directly from the ordinary fitted values. The observations are not excluded from the model used to calculate the predicted response.

Residuals

The difference (e_i) between the observed values and predicted or fitted values (data minus fits).

This part of the observation is not explained by the fitted model. The formula for the residual of an observation is:

$$e_i = (Y_i - \hat{Y}_i)$$

Cross-validated residuals

Measure the model's predictive ability and are used to calculate the PRESS statistic. Cross-validated residuals in PLS and least squares regression are conceptually similar, but their calculations differ:

- In PLS, the cross-validated residuals are the differences between the actual responses and the cross-validated fitted values. The formula is:

  $e_{i-1} = y_i - \hat{y}_{(i)}$, (i) = observation omitted from the model calculation,

  where y_i = response value and $\hat{y}_{(i)}$ = cross-validated fitted value. The cross-validated residual value varies based on how many observations are omitted each time the model is recalculated during cross-validation.

- In least squares regression, the cross-validated residuals are calculated directly from the ordinary residuals using the formula:

  $e_{(i)} = e_i / (1 - h_i)$

  where h_i = leverage value.

Standardized residuals

Also called the internally Studentized residual. The residual e_i scaled by its standard deviation. The standardized residual is helpful in identifying outliers. The formula is:

$$\frac{e_i}{\text{sqrt}[s^2 * (1 - h_i)]}$$

where s^2 = MS Error and h_i = i^{th} diagonal element of $\mathbf{X(X'X)^{-1}X'}$. The denominator is an estimator of the standard deviation of e_i.

For more information, see [28].

Standard error of fitted values (SE fit)

The standard error of the fitted value; also the estimated standard deviation of the fitted value. The formula for the standard error of the fitted value in a regression model with one predictor is:

$$\sqrt{s^2\left[\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum(x_i - \bar{x})^2}\right]}$$

where s^2 = MS Error, n= number of observations, x_0 = predictor value, $\bar{x}$ = mean of the x-values, and x_i = i^{th} x-value.

The formula for the standard error of the fitted value at a given point $x'_0 = [1, x_{1,0}, \ldots, x_{k,0}]$

$\text{sqrt}(s^2[x'_0(\mathbf{X'X})x_0]$

where s^2 = MS Error and $\mathbf{X}$ = matrix of predictors.

Confidence interval

The range in which the estimated mean response for a given set of predictor values is expected to fall. The interval is defined by lower and upper limits, which Minitab calculates from the confidence level and the standard error of the fits.

The formula is:

$$\hat{y}_0 \pm t(1 - \alpha/2; n - p) * s(\hat{y}_0)$$

where α = chosen alpha value, n = number of observations, p = number of predictors, and $s(\hat{y}_0) = \text{sqrt}[s^2(\mathbf{X'}_0(\mathbf{X'X})^{-1}\mathbf{X}_0)] = \text{sqrt}[\mathbf{X'}_0\mathbf{s}^2\{\mathbf{b}\}\mathbf{X}_0]$. s^2 = mean square error and $\mathbf{s}^2\{\mathbf{b}\}$ = variance-covariance matrix of the coefficients.

For more information, see [28].

Prediction interval

The range in which the predicted response for a new observation is expected to fall. The interval is defined by lower and upper limits, which Minitab calculates from the confidence level and the standard error of the prediction. The prediction interval is always wider than the confidence interval because of the added uncertainty involved in predicting a single response versus the mean response.

The formula is:

$$\hat{y}_0 \pm t(1 - \alpha/2; n - p) * s(\text{pred})$$

where α = chosen alpha value, n = number of observations, p = number of predictors, and $s(\text{pred}) = \text{sqrt}[s^2(1 + \mathbf{X'}_0(\mathbf{X'X})^{-1}\mathbf{X}_0)]$.

Back to top

For more information, see [28].

Component information

X-scores

Linear combinations of the predictor variables; similar to principal component scores. The x-scores form an n x m matrix of uncorrelated columns, where n = number of observations and m = number of components. Providing a window to the x-space, the x-scores are projections of the observations on the PLS components. PLS fits the x-scores, which replace the original predictors, using least squares estimation. Minitab calculates the x-score vector for the m^{th} component using the following formula:

$t_i = X * w_j$, i = observations (1, n) and j = predictors (1, p)

where X = x-residual matrix and w = weights.

X-loadings

Linear coefficients that link the predictors to the x-scores; similar to eigenvectors in principal components analysis. The loading values indicate the importance of the corresponding predictor to the m^{th} component. The x-loadings form a p x m matrix, where p = number of predictors and m = number of components.

Minitab calculates the x-loading vector for the m^{th} component using the following formula:

$l_j' = t_i' * X / t_i' t_i$ i = observations (1, n) and j = predictors (1, p)

where t = x-scores and X = predictors.

X-weights

Describe the covariance between the predictors and responses. In the algorithm, the weights ensure the x-scores are orthogonal, or unrelated to one another, and are used to calculate the x-scores. The x-weights form a p x m matrix, where p = number of predictors and m = number of components. Minitab calculates the x-weight vector for the m^{th} component using the formula:

$w_j' = u_i' X / u_i' u_i$, i = observations (1, n) and j = predictors (1, p)

where X = x-residual matrix and u = y-scores.

X-residuals

Contain the variance in the predictors not explained by the PLS model. Observations with relatively large x-residuals are outliers in the x-space, indicating that they are not well explained by the model.

The x-residuals are the differences between the actual predictor values and the x-calculated values and are on the same scale as the original predictors. The x-residual matrix, similar to the original x-matrix, is an n x p matrix, where n = number of observations and p = number of predictors.

The x-residual matrix is initialized to the standardized x-matrix. After calculating the m^{th} component and obtaining the x-score vector and the x-loading vector, Minitab calculates the x-residuals are calculated using the following formula:

$xres_{ij} = xres_{ij} - t_i * l_j'$, i = observations (1, n) and j = predictors (1, p)

where t = x-scores and l = x-loadings.

X-calculated values

Linear combinations of the x-scores; contain the variance in the predictors explained by the PLS model. Observations with relatively small x-calculated values are outliers in the x-space and are not well explained by the model.

The x-calculated matrix, similar to the original x-matrix, is an n x p matrix, where n = number of observations and p = number of predictors. The x-calculated values are on the same scale as the predictors.

The x-calculated matrix is initialized to the zero matrix. After calculating the m^{th} component and obtaining the x-score vector and the x-loading vector, Minitab calculates the x-calculated values using the following formula:

$xcal_{ij} = xcal_{ij} + t_i * l_j'$, i = observations (1, n) and j = predictors (1, p)

where t = x-scores and l = x-loadings.

If the number of components equals the number of predictors, then x-calculated value equals the original x-value.

Y-scores

Linear combinations of the response variables. The y-scores form an n x m matrix, where n = number of observations and m = number of components. Providing a window to the y-space, the y-scores are projections of the observations on the PLS components. Minitab calculates the y-score vector for the m^{th} component using the following formula:

$u_i = Y * c_k'$, i = observations (1, n) and k = responses (1, r)

where Y = y-residual matrix and c = y-loadings.

Y-loadings

Linear coefficients that link the responses to the y-scores. The loading values indicate the importance of the corresponding response to the m^{th} component. The y-loadings form an r x m matrix, where r = number of responses and m = number of components.

Minitab calculates the y-loading vector for the m^{th} component using the following formula:

$c_k' = t_i' * Y$, i = observations (1, n) and k = responses (1, r)

where t = x-scores and Y = responses.

Y-residuals

Contain the remaining variance in the responses not explained by the PLS model. Observations with relatively large y-residuals are outliers in the y-space, indicating that they are not well explained.

The y-residuals are the differences between the actual response values and the y-calculated values, and are on the same scale as the original responses. The y-residual matrix, similar to the original y-matrix, is an n x r matrix, where n = number of observations and r = number of responses.

The y-residual matrix is initially set to the standardized Y matrix. After Minitab calculates the m^{th} component and obtains the x-score and y-loading vectors, Minitab determines the standardized y-residuals using the formula:

$yres_{ik} = yres_{ik} - t_i * c_k'$, i = observations (1, n) and k = responses (1, r)

where t = x-scores and c = y-loadings. Minitab then calculates the unstandardized y-residuals by multiplying the standardized y-residuals by the standard deviation of the corresponding response values.

Y-calculated values

Linear combinations of the x-scores; contain the variance in the responses explained by the PLS model. Observations with relatively small y-calculated values are outliers in the y-space and are not well explained.

The y-calculated matrix, like the original y-matrix, is an n x r matrix, where n = number of observations and r = number of responses.

The y-calculated matrix is initially set to the zero matrix. After Minitab calculates the m^{th} component and obtains the x-score and y-loading vectors, Minitab determines the standardized y-calculated values using the formula:

$ycal_{ik} = ycal_{ik} + t_i * c_k'$, i = observations (1, n) and k = responses (1, r)

where t = x-scores and c = y-loadings. Minitab then calculates the unstandardized y-calculated values by multiplying the standardized y-calculated values by the standard deviation and adding the mean of the corresponding response values.

Back to top

Bibliografía

[1] A. Agresti (1984). *Analysis of Ordinal Categorical Data.* John Wiley & Sons, Inc.

[2] A. Agresti (1990). *Categorical Data Analysis.* John Wiley & Sons, Inc.

[3] D.A. Belsley, E. Kuh, and R.E. Welsch (1980). *Regression Diagnostics.* John Wiley & Sons, Inc.

[4] A. Bhargava (1989). "Missing Observations and the Use of the Durbin-Watson Statistic," *Biometrik,* 76, 828□831.

[5] C.C. Brown (1982). "On a Goodness of Fit Test for the Logistic Model Based on Score Statistics," *Communications in Statistics*, 11, 1087□1105.

[6] D.A. Burn and T.A. Ryan, Jr. (1983). "A Diagnostic Test for Lack of Fit in Regression Models," *ASA 1983 Proceedings of the Statistical Computing Section*, 286□290.

[7] R.D. Cook (1977). "Detection of Influential Observations in Linear Regression," *Technometrics*, 19, 15□18.

[8] R.D. Cook and S. Weisberg (1982). *Residuals and Influence in Regression.* Chapman and Hall.

[9] N.R. Draper and H. Smith (1981). *Applied Regression Analysis*, Second Edition. John Wiley & Sons, Inc.

[10] S.E. Fienberg (1987). *The Analysis of Cross-Classified Categorical Data*. The MIT Press.

[11] I.E. Frank and J.H. Friedman (1993). "A Statistical View of Some Chemometrics Regression Tool," *Technometrics*, 35, 109□135.

[12] I.E. Frank and B.R. Kowalski (1984). "Prediction of Wine Quality and Geographic Origin from Chemical Measurements by Partial Least-Squares Regression Modeling," *Analytica Chimica Acta*, 162, 241□251.

[13] M.J. Garside (1971). "Some Computational Procedures for the Best Subset Problem," *Applied Statistics*, 20, 8□15.

[14] P. Geladi and B. Kowalski (1986). "Partial Least-Squares Regression: A Tutorial," *Analytica Chimica Acta,* 185, 1□17.

[15] P. Geladi and B. Kowalski (1986). "An Example of 2-Block Predictive Partial Least-Squares Regression with Simulated Data," *Analytica Chimica Acta,* 185, 19-32.

[16] James H. Goodnight (1979). "A Tutorial on the Sweep Operator," *The American Statistician,* 33, 149□158.

[17] W.W. Hauck and A. Donner (1977). "Wald's test as applied to hypotheses in logit analysis," *Journal of the American Statistical Association*, 72, 851-853.

[18] D.C. Hoaglin and R.E. Welsch (1978). "The Hat Matrix in Regression and ANOVA," *The American Statistician*, 32, 17□22.

[19] R.R. Hocking (1976). "A Biometrics Invited Paper: The Analysis and Selection of Variables in Linear Regression," *Biometrics*, 32, 1□49.

[20] A. Hoskuldsson (1988). "PLS Regression Methods," *Journal of Chemometrics*, 2, 211□228.

[21] D.W. Hosmer and S. Lemeshow (2000). *Applied Logistic Regression.* 2nd ed. John Wiley & Sons, Inc.

[22] LINPACK (1979). Linpack User's Guide by J.J. Dongarra, J.R. Bunch, C.B. Moler, and G.W. Stewart, Society for Industrial and Applied Mathematics, Philadelphia, PA.

[23] A. Lorber, L. Wangen, and B. Kowalski (1987). "A Theoretical Foundation for the PLS Algorithm," *Journal of Chemometrics*, 1, 19□31.

[24] J.H. Maindonald (1984). *Statistical Computation.* John Wiley & Sons, Inc.

[25] P. McCullagh and J.A. Nelder (1992). *Generalized Linear Model*. Chapman & Hall.

[26] W. Miller (1978). "Performing Armchair Roundoff Analysis of Statistical Algorithms," *Communications in Statistics*, 243□255.

[27] D.C. Montgomery and E.A. Peck (1982). *Introduction to Linear Regression Analysis*. John Wiley & Sons.

[28] J. Neter, W. Wasserman, and M. Kutner (1985). *Applied Linear Statistical Models*. Richard D. Irwin, Inc.

[29] S.J. Press and S. Wilson (1978). "Choosing Between Logistic Regression and Discriminant Analysis," *Journal of the American Statistical Association,* 73, 699-705.

[30] M. Schatzoff, R. Tsao, and S. Fienberg (1968). "Efficient Calculation of All Possible Regressions," *Technometrics*, 10, 769□779.

[31] G.W. Stewart (1973). *Introduction to Matrix Computations*. Academic Press.

[32] R.A. Thisted (1988). *Elements of Statistical Computing: Numerical Computation*. Chapman & Hall.
[33] P. Velleman and R. Welsch (1981). "Efficient Computation of Regression Diagnostics," *The American Statistician*, 35, 234□242.
[34] P.F. Velleman, J. Seaman, and I.E. Allen (1977). "Evaluating Package Regression Routines," *ASA 1977 Proceedings of the Statistical Computing Section*.
[35] S. Weisberg (1980). *Applied Linear Regression*. John Wiley & Sons, Inc.
[36] H. Wold (1975). "Soft Modeling by Latent Variables; the Nonlinear Iterative Partial Least Squares Approach," in *Perspectives in Probability and Statistics, Papers in Honour of M.S. Bartlett*, ed. J. Gani, Academic Press.

TRANSFORMACIONES A UNA LINEA RECTA,POR DR. PRIMITIVO REYES AGUILAR

Enviado por:
Ing.+Lic. Yunior Andrés Castillo S.
"NO A LA CULTURA DEL SECRETO, SI A LA LIBERTAD DE INFORMACION"®
www.monografias.com/usuario/perfiles/ing_lic_yunior_andra_s_castillo_s/monografias
Página Web: yuniorandrescastillo.galeon.com
Correo: yuniorcastillo@yahoo.com
yuniorandrescastillosilverio@facebook.com
Twitter: @yuniorcastillos
Celular: 1-829-725-8571
Santiago de los Caballeros,
República Dominicana,
2015.
"DIOS, JUAN PABLO DUARTE Y JUAN BOSCH – POR SIEMPRE"®