CAPÍTULO 9

REGRESIÓN LINEAL

En este capítulo, primero se tratará la Regresión Lineal Simple, cuyos aspectos descriptivos ya fueron considerados en la Sección 3.8 del texto. La inferencia estadística en regresión simple es discutida en gran detalle. Luego se considerará el caso donde hay más de una variable predictora y se hacen las inferencias correspondientes. Finalmente se discutirá los métodos de elegir las mejores variables predictoras que produzcan un modelo confiable con el menor número de variables.

9.1 Regresión Lineal Simple

Supongamos que tenemos datos de dos variables cuantitativas continuas X e Y, las cuales se relacionan siguiendo una tendencia lineal, que puede ser detectada haciendo un diagrama de dispersion de los datos. Tendencia lineal significa que los puntos están dispuestos alrededor de una línea recta, desviándose por una cantidad aleatoria ε de la misma. Si además, asumimos que se trata de predecir el comportamiento de Y usando X, entonces el **modelo de regresión lineal simple** es de la forma:

$$Y = \alpha + \beta X + \varepsilon$$

Donde, Y es llamada la variable de respuesta o dependiente,
X es llamada la variable predictora o independiente,
α es el intercepto de la línea con el eje Y,
β es la pendiente de la línea de regresión y
ε es un error aleatorio, el cual se supone que tiene media 0 y varianza constante σ^2.

α y β son parámetros desconocidos y para estimarlos se toma una muestra de tamaño n de observaciones (x_i,y_i). La variable Y se asume que es aleatoria, pero X no necesariamente lo es.

El estimado $\hat{\alpha}$ de α y el estimado $\hat{\beta}$ de β son hallados usando el método de mínimos cuadrados, que se basa en minimizar la suma de cuadrados de los errores $Q(\alpha,\beta) = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \alpha - \beta x_i)^2$. Usando técnicas de cálculo diferencial para minimizar una función de dos variables α y β se obtienen:

$$\boxed{\hat{\beta} = \frac{S_{xy}}{S_{xx}}} \quad \text{y} \quad \boxed{\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}}$$

las cantidades Sxx y Sxy aparecen definidas en la Sección 3.8 del texto.

La ecuación $\hat{Y} = \hat{\alpha} + \hat{\beta}X$, es llamada la **línea de regresión estimada**. Para obtener esta línea en **MINITAB** se sigue la secuencia: **STAT ▸ Regression ▸ Regression**. En la salida, además de la ecuación, aparecen los valores de la prueba de *t* para probar hipótesis acerca del intercepto y la pendiente. También se muestra la tabla del Análisis de Varianza para regresión que permiten hacer inferencia estadística acerca de la pendiente de la línea de regresión poblacional.

Ejemplo 9.1. Se desea hallar una línea de regresión que permita predecir el precio de una casa (Y) basado en el área de la misma (X). Se recolectaron 15 datos:

Casa	área	precio
1	3060	179000
2	1600	126500
3	2000	134500
4	1300	125000
5	2000	142000
6	1956	164000
7	2400	146000
8	1200	129000
9	1800	135000
10	1248	118500
11	2025	160000
12	1800	152000
13	1100	122500
14	3000	220000
15	2000	141000

La ventana de diálogo para **Regression** se completará como sigue:

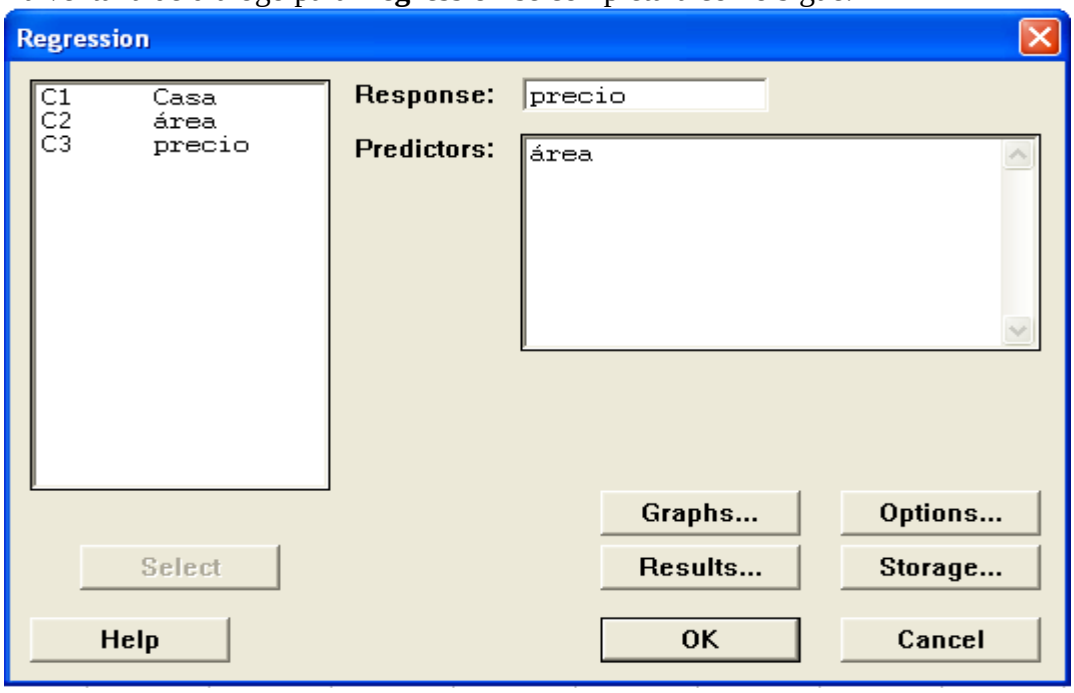

Figura 9.1. Ventana de diálogo para regresión.

En la ventana **Response** se entra la variable de respuesta Y, en la ventana de **Predictors** se entra la variable independiente X
El botón **Results** permite controlar los resultados que aparecerán en la ventana **session.** Hay 4 alternativas para controlar la salida según se muestra en la Figura 9.2.

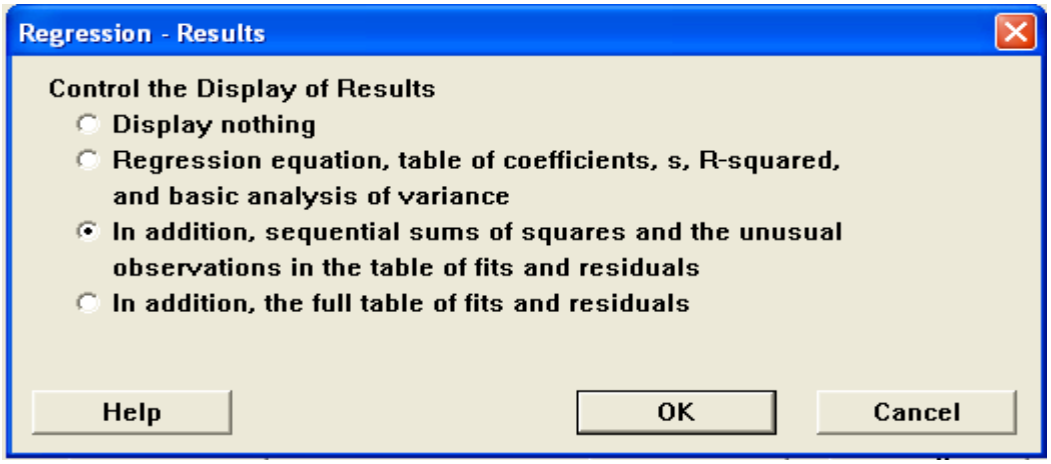

Figura 9.2. Ventana de diálogo que aparece al seleccionar el botón **results** en **regression.**

El botón **Storage** permite guardar algunas medidas importantes que aparecen en el análisis de regresión y que posteriormente se pueden usar, por ejemplo, en el análisis de residuales. La ventana de diálogo se muestra en la Figura 9.3.

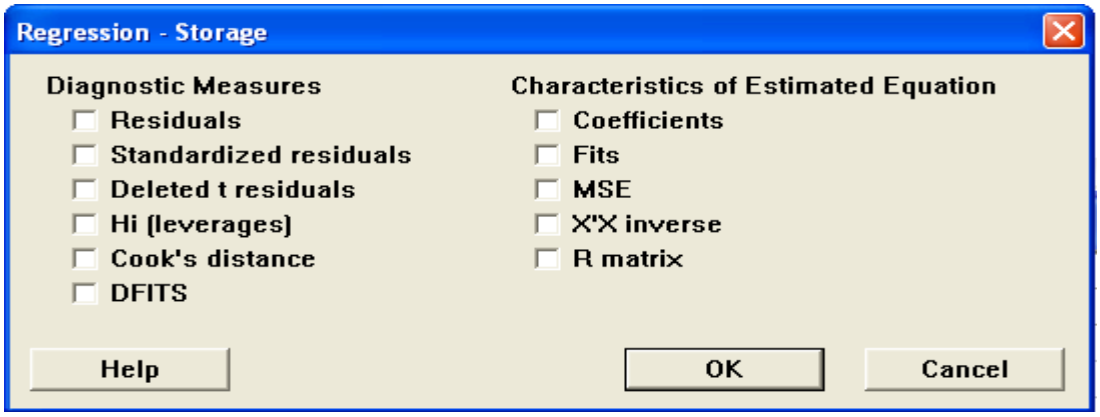

Figura 9.3. Ventana de diálogo que aparece al oprimir el botón **storage** en **regression**

El efecto de oprimir los botones **Graphs** y **Options** se explicará en las siguientes secciones. Al oprimir el botón **OK** en la ventana **regression** se obtendrán los siguientes resultados:

```
MTB > Regress 'precio' 1 'area';
SUBC>   Constant;
SUBC>   Brief 2.
```

Regression Analysis

```
The regression equation is
precio = 73168 + 38.5 area

Predictor        Coef       StDev          T        P
Constant        73168       12674       5.77    0.000
area           38.523       6.391       6.03    0.000

S = 14118       R-Sq = 73.6%     R-Sq(adj) = 71.6%

Analysis of Variance

Source            DF          SS          MS         F        P
Regression         1  7241245891  7241245891     36.33    0.000
Residual Error    13  2591087442   199314419
Total             14  9832333333

Unusual Observations
Obs       area     precio       Fit   StDev Fit    Residual   St Resid
 14       3000     220000      188737        7923       31263   2.68R

R denotes an observation with a large standardized residual
```

9.1.1. Interpretación de los Coeficientes de Regresión:

Interpretación del intercepto $\hat{\alpha}$:

Indica el valor promedio de la variable de respuesta Y cuando X es cero. Si se tiene certeza de que la variable predictora X no puede asumir el valor 0, entonces la interpretación no tiene sentido. En el ejemplo anterior, $\hat{\alpha}$ = 73,168 indicaría que si la casa no tiene área, su precio promedio será 73,158, lo cual no es muy razonable. Es más conveniente hallar una línea de regresión que no tenga intercepto.

Interpretación de la pendiente $\hat{\beta}$:

Indica el cambio promedio en la variable de respuesta Y cuando X se incrementa en una unidad. En el ejemplo anterior $\hat{\beta}$ = 38.5 indica que por cada píe cuadrado adicional de la casa su precio aumentará en promedio en 38.5 dólares.

9.2 Inferencia en Regresión Lineal

Para poder hacer inferencia en regresión hay que asumir que los errores e_i del modelo se distribuyen en forma normal con media cero y varianza constante σ^2 y además que sean idependientes entre sí. Se pueden hacer prueba de hipótesis y calcular intervalos de confianza para el intercepto α y de la pendiente β de la línea de regresión poblacional.

Asimismo se pueden establecer intervalos de confianza para el valor medio y para el valor individual de la variable de respuesta dado un valor particular de la variable predictora.

9.2.1 Inferencia acerca de los coeficientes de regresión

Con respecto a prueba de hipótesis lo más frecuente es probar H_o: $\alpha = 0$ versus H_a: $\alpha \neq 0$ y H_o: $\beta = 0$ versus H_a: $\beta \neq 0$. De aceptarse la primera hipótesis significaría que la línea de regresión pasaría por el origen, es decir, que cuando la variable predictora es cero, entonces el valor promedio de la variable de respuesta es también cero. De aceptarse la segunda hipótesis significaría que la pendiente de la línea de regresión es cero, es decir, que la variable predictora no se relaciona linealmente con la variable de respuesta. En ambos casos la prueba estadística que se usa es una prueba de *t* de Student.

Sólo discutiremos la prueba de hipótesis para la pendiente. La prueba estadística viene dada por:

$$t = \frac{\hat{\beta}}{s.e(\hat{\beta})} = \frac{\hat{\beta}}{\frac{s}{\sqrt{S_{xx}}}}$$

La cual se distribuye como una *t* con *n*-2 grados de libertad. Aquí $s = \sqrt{\frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}{n-2}}$ es la desviación estándar del error, S_{xx} es la suma de cuadrados corregida de la variable X y s.e($\hat{\beta}$) es el error estándar de $\hat{\beta}$. En el Ejemplo 9.1, s=14,118 y s.e($\hat{\beta}$)= $s/\sqrt{s_{xx}}$ =6.391.

En **MINITAB** aparece el valor de la prueba estadística y el "p-value" de la prueba, él cual se puede usar para llegar a una decisión. Un "p-value" cercano a 0, digamos menor que 0.05, lleva a la conclusión de rechazar la hipótesis nula. Si se rechaza la hipótesis nula quiere decir de que de alguna manera la variable X es importante para predecir el valor de Y usando la regresión lineal. En cambio si se acepta la hipótesis nula se llega a la conclusión de que, la variable X no es importante para predecir el comportamiento de Y usando una regresión lineal.

En el Ejemplo 9.1 el valor de la prueba estadística de *t* es 6.03 y el P-value = .0000 por lo que se rechaza la hipótesis nula. Luego hay suficiente evidencia estadística para concluir que la variable área de la casa puede ser usada para predecir el precio de la casa.

También se pueden establecer intervalos de confianza para los parámetros de regresión. Por ejemplo, un intrevalo de confianza del 100 (1-α) % para la pendiente β será de la forma:

$$\hat{\beta} \pm t_{(\alpha/2,n-2)} \frac{s}{\sqrt{S_{xx}}}$$

MINITAB no da este intervalo de confianza. Hay que calcular el percentil de la *t* de student usando la secuencia **Calc ▸ Probability Distributions ▸ t**. En el ejemplo anterior, un intervalo del 95 % para la pendiente será:

$$38.523 \pm (2.1604)6.391$$

O sea, hay una confianza del 95 % de que la pendiente de la regresión poblacional caiga en el intervalo (24.7150, 52.3301).

9.2.2 El Análisis de Varianza para Regresión Lineal Simple.

El análisis de varianza, que fue introducida por Fisher, consiste en descomponer la variación total de una variable en varias partes, cada una de las cuales es llamada una fuente de variación. En el caso de regresión, la descomposición de la variación de la variable de respuesta Y es como sigue:

VAR. TOTAL DE Y = VAR. DEBIDA A LA REGRESIÓN + VAR. DEBIDA AL ERROR

Cada variación es representada por una suma de cuadrados, definidas de la siguiente manera:

Suma de Cuadrados Total = SST = $\sum_{i=1}^{n}(y_i - \bar{y})^2$

Suma de Cuadrados de Regresión = SSR = $\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$

Suma de Cuadrados del Error = SSE = $\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$

Cada una de estas sumas de cuadrados tiene una distribución Ji-Cuadrado, SSR tiene una distribución Ji-Cuadrado no central con 1 grado de libertad, SSE tiene una distribución Ji-Cuadrado con *n*-2 grado de libertad, y SST se comporta como una Ji-Cuadrado no central con *n*-1 grados de libertad. Al dividir las sumas de cuadrados por sus grados de libertad se obtienen los Cuadrados Medios. Si la hipótesis de que la pendiente β es 0 es cierta, entonces la división del cuadrado medio de la regresión por el cuadrado medio del error se distribuye como una F con 1 grado de libertad en el numerador y *n*-2 en el denominador. Luego, la hipótesis H_o: $\beta = 0$ se rechaza si el "p-value" de la prueba de F es menor que .05. Los cálculos se resumen en la siguiente tabla llamada **tabla del análisis de varianza** para la regresión lineal simple.

Fuentes de Variación	Grados de Libertad	Suma de Cuadrados	Cuadrados Medios	F
Debido a la regresión	1	SSR	MSR=SSR/1	MSR/MSE
Debido al Error	n-2	SSE	MSE=SSE/n-2	
Total	n-1	SST		

En el ejemplo anterior la prueba de F es 36.33 y el "P-value"=.0000, por lo que se rechaza la hipótesis nula. Notar que el valor de la prueba de F = 36.33 = $(6.03)^2$ es el cuadrado de la prueba *t*.

9.2.3 El Coeficiente de Determinación

El coeficiente de determinación, denotado por R^2, es una medida de la bondad de ajuste del modelo de regresión hallado. Se calcula por:

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

donde, SSR representa la suma de cuadrados debido a la regresión, y SST representa la suma de cuadrados del total. Puede demostrarse que el coeficiente de detreminación es simplemente el cuadrado del coeficiente de correlación. El coeficiente de Determinación varía entre 0 y 1, aunque es bastante común expresarlo en porcentaje. Un R^2 mayor del 70 % indica una buena asociación lineal entre las variables, luego la variable X puede usarse para predecir Y. Hay que tener presente que el R^2 es afectado por la presencia de valores atípicos.

También R^2 indica qué porcentaje de la variabilidad de la variable de respuesta Y es explicada por su relación lineal con X, mientras más alto sea este valor mejor es la predicción de Y usando X.

Existen otras medidas para medir la precisión de la predición de un modelo de regresión, pero son discutidas en este texto.

9.2.4 Intervalos de Confianza para el valor medio de Y e Intervalo de Predicción

A nivel poblacional para cada valor de la variable X existe una población de valores de Y, la cual se asume que se distribuye normalmente con cierta media y varianza constante σ^2. Lo que se busca es establecer un intervalo de confianza para dicha media asumiendo que la relación entre X e Y es lineal. Dado un valor X_o de la variable X es natural pensar, que un estimado del valor medio de las Y's es $\hat{Y}_o = \hat{\alpha} + \hat{\beta}X_o$. Usando las propiedades distribucionales de este estimado se puede establecer que un intervalo de confianza del 100 (1-α) % para el valor medio de todos los valores Y dado que $X = X_0$ es como sigue:

$$\hat{Y}_0 \pm t_{(1-\alpha/2,n-2)} s \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}}$$

Por otro lado muchas veces estamos interesados en estimar solamente un valor de Y correspondiente a un valor dado X_0. El estimado puntual será el mismo $\hat{Y}_o$, y usando

propiedades distribucionales de $\hat{Y}_o - Y_o$ se obtiene que un Intervalo de confianza del 100 (1-α) % para el valor predicho de Y dado que $X = X_0$ es de la forma:

$$\hat{Y}_0 \pm t_{(1-\alpha/2,n-2)} s \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}}$$

Este intervalo de confianza es llamado intervalo de predicción.

Es más riesgoso hacer predicciones para un sólo valor que para un valor medio, por esta razón el intervalo de predicción de Y es más ancho que el intervalo de confianza para el valor medio.

El botón **Options** de la ventana **regression** permite hallar estos intervalos de confianza. La Figura 9.4 muestra la ventana de diálogo que aparece cuando se oprime el botón **Options**. En este ejemplo se trata de determinar el intervalo de confianza e intervalo de predicción para el precio de la casa cuando ésta tiene un área de 3,500 pies cuadrados usando un nivel de confianza del 95 %. Para ello hay que seleccionar las opciones **Confidence limits** y **Prediction limits.**

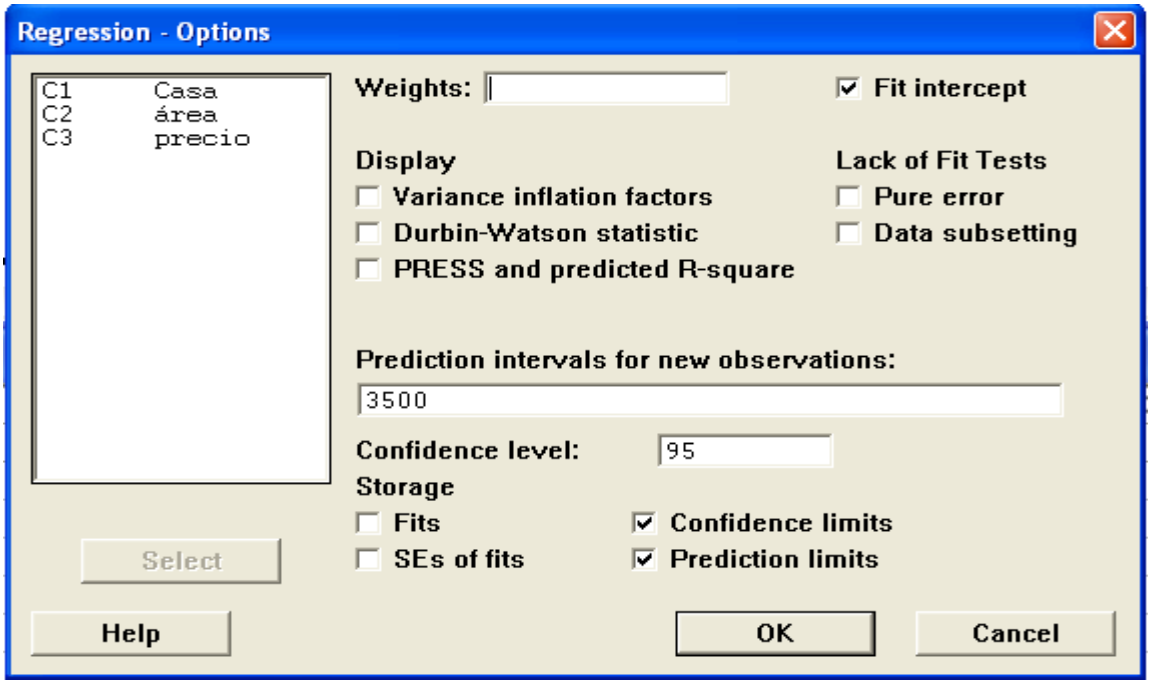

Figura 9.4. Ventana de diálogo que se obtiene al oprimir **options** en **regression.**

En la ventana **session** aparecerá el siguiente resultado **:**

```
Predicted Values for New Observations

 New Obs      Fit      SE Fit         95.0% CI             95.0% PI
 1         207998     10861    ( 184536,  231461)   ( 169518,  246479) X
 X  denotes a row with X values away from the center

Values of Predictors for New Observations

New Obs       Area
1             3500
```

Interpretación*: Hay un 95 % de confianza de que el valor medio de todas las casas de 3,500 píes cuadrado de área caiga entre 184,536 y 231,461.*
Hay un 95 % de confianza de que el valor de una casa de 3,500 píes cuadrados caiga entre 169,518 y 2246,479.

Asímismo, la opción **Fitted line Plot** del menú de **Regression** permite hallar bandas de confianza tanto para el valor predicho como para el valor medio de las Y. Para esto se deben elegir las opciones **Display Confidence Interval** y **Display Prediction Interval** al oprimir el boton **Options**. Con las bandas de confianza se pueden tener intervalos de confianzas para cualquier valor dado de X. Para el presente ejemplo se obtiene:

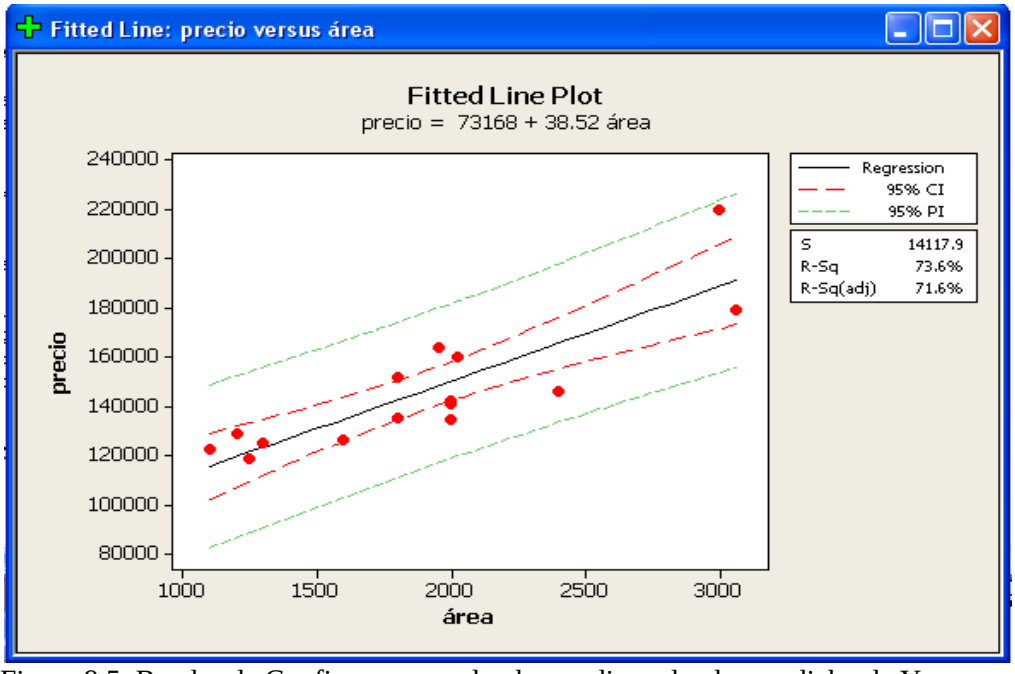

Figura 9.5 Bandas de Confianza para el valor medio y el valor predicho de Y

Notar que las bandas de confianza son anchas en los extremos del eje X y angostas en el centro del mismo. En realidad las bandas se van angostando cuando los valores de X que se toman están cerca del promedio $\bar{x}$.

9.3 Análisis de Residuales

Un residual r_i es la diferencia entre el valor observado Y_i y el valor estimado por la linea de regresión $\hat{Y}_i$, es decir, $r_i = Y_i - \hat{Y}_i$. El residual puede ser considerado como el error aleatorio e_i observado. También se acostumbra usar el **Residual estandarizado**, el cual se obtiene al dividir el residual entre la desviación estándar del residual, y el **Residual estudentizado "deleted",** que es similar al anterior pero eliminando de los cálculos la observación cuyo residual se desea hallar.

El análisis de residuales permite cotejar si las suposiciones del modelo de regresión se cumplen.

Se puede detectar:

a) Si efectivamente la relación entre las variables X e Y es lineal.
b) Si hay normalidad de los errores.
c) Si hay valores anormales en la distribución de errores.
d) Si hay varianza constante (propiedad de Homocedasticidad) y
e) Si hay independencia de los errores.

El análisis de residuales se puede llevar a cabo gráficamente o en forma analítica. En este texto sólo consideraremos un análisis gráfico, las cuales pueden obtenerse de dos maneras. La primera manera es escogiendo el botón **Graphs** de la ventana de diálogo **Regression**.

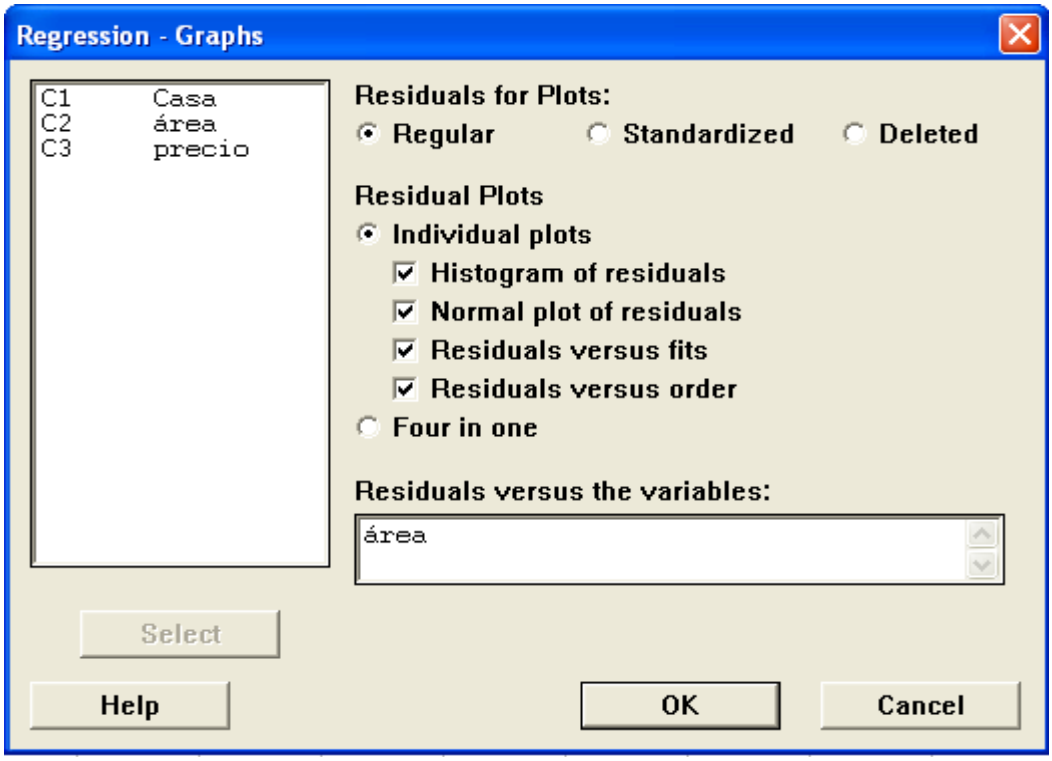

Figura 9.6. Ventana de diálogo que aparece al oprimir **Graphs** en **Regression.**

Hay tres posibles elecciones de residuales y hasta 5 plots de residuales que se pueden hacer. Las ventanas de gráficas aparecerán en cascada como se muestra en la Figura 9.7. En esta manera cada plot de residual sale en una ventana gráfica separada. Hay cinco plots que se usan:

1) **Plot de Normalidad**: Permite cotejar normalidad. Si los puntos están bien cerca de una línea recta se concluye, que hay normalidad.
2) **Histograma de Residuales**: También permite cotejar normalidad. Cuando el histograma es simétrico, con un único pico en el centro, se concluye que hay normalidad.
3) **Plot de Residuales versus los valores predichos (FITS)**: Se usa para detectar si hay datos anormales, cuando hay datos que caen bastantes alejados, tanto en el sentido vertical como horizontal. También permite detectar si la varianza de los errores es constante con respecto a la variable de respuesta.
4) **Plot de Residuales versus el indice de la observación**: Es más específico para detectar que observación es un dato anormal. Si se usan residuales estandarizados, entonces un dato con residual más allá de 2 ó -2 es considerado un "outlier" en el sentido vertical**.**
5) **Plot de Residuales versus la variable predictora**: Es usado para detectar datos anormales así como si la varianza de los errores es constante con respecto a la variable predictora.

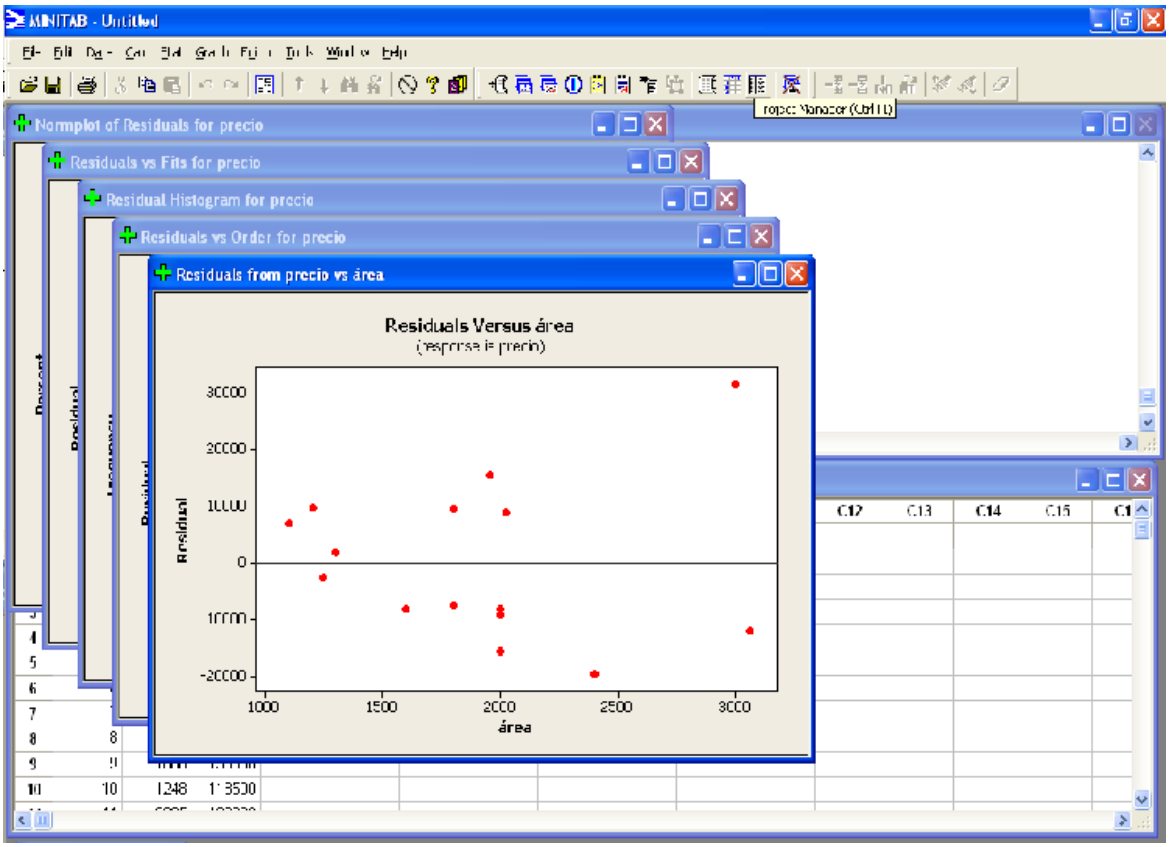

Figura 9.7. Plots de residuales mostrados en cascada.

La segunda manera de obtener los plots de residuales es, escogiendo la opcion **Four in One** en el botón **Graph** de la ventana de diálogo de **Regression** como aparece en la Figura 9.8.

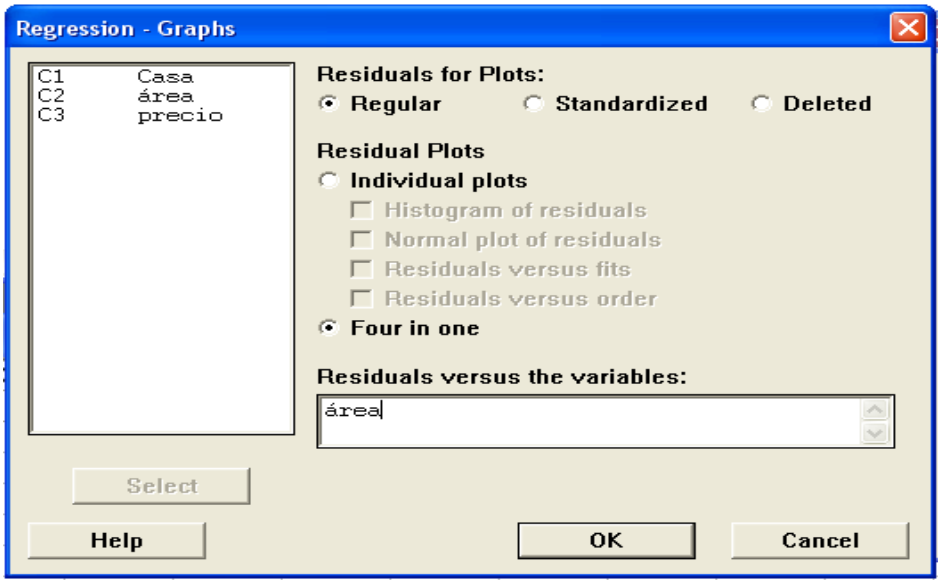

Figura 9.8. Ventana de diálogo de la opcion **graph** de **regression**.

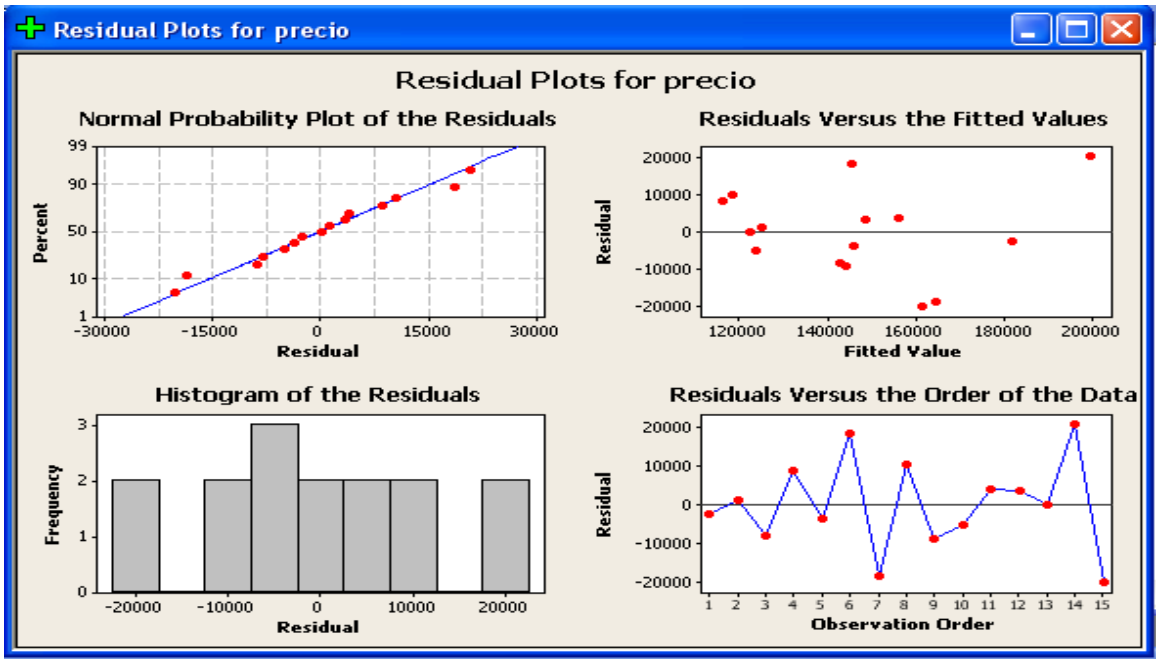

Figura 9.9. Plots de residuales en una misma ventana.

Aparecerán en una misma página los cuatro primeros plots de la lista mencionada anteriormente, como se muestra en la Figura 9.9.

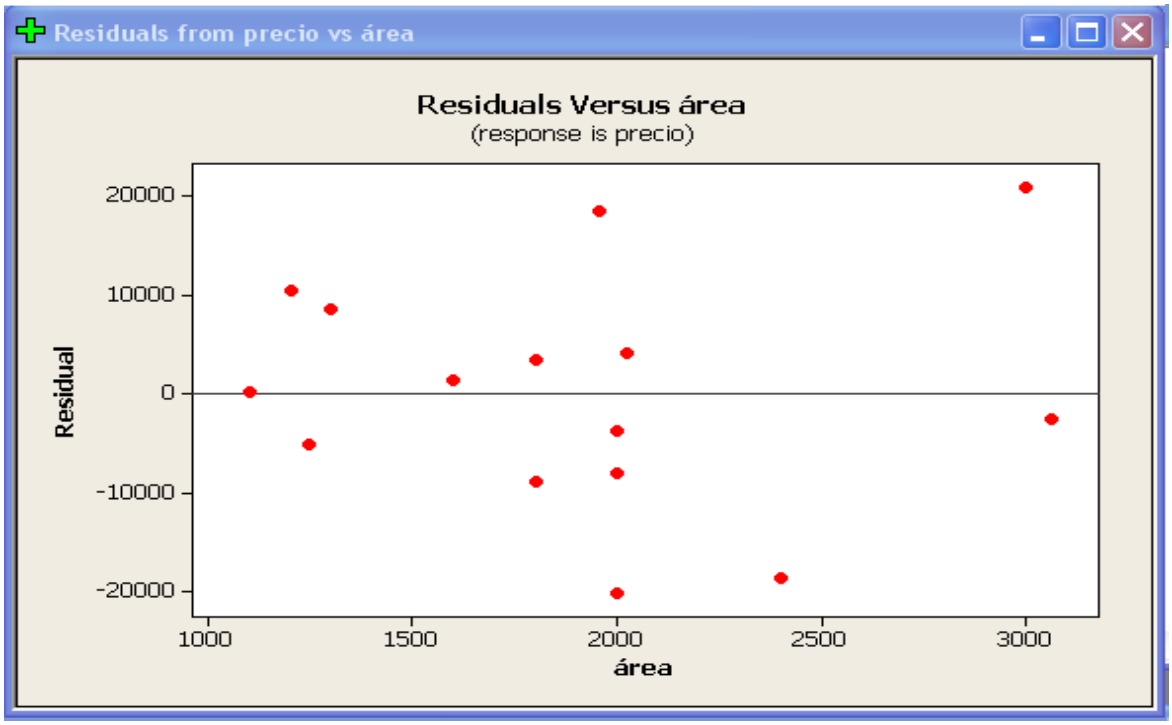

Figura 9.10. Plots de residuales versus la variable area.

***Interpretación**: Los puntos del plot de normalidad no caen cerca de una línea recta y en el extreno superior se detecta un "outlier". Similarmente, el histograma no es simétrico con un pico central y también muestra un "outlier" en el extremo superior. En conclusión, no hay normalidad de los errores. El plot de residuales versus el índice de la observación muestra que la observacion 14 es un "outlier", pues el residual estandarizado cae más allá de dos. El plot de los residuales versus los valores predichos muestra que la varianza de los errores no es constante con respecto a la variable de respuesta, pues tiende ha aumentar cuando el valor de la variable de respuesta aumenta.*

Hay maneras de corregir algunas de las anomalías encontradas en el análisis de residuales, las cuales pueden ser leídas en un texto especializado de regresión.

9.4 Modelos No Lineales y Transformaciones

Cuando se construyen modelos de regresión el objetivo es conseguir un modelo con R^2 alto que se aproxime a 100 %, asumiendo que no hay datos atípicos presentes. Si no se desea incluir variables predictoras adicionales en el modelo, hay dos alternativas:

i) Tratar de usar modelos polinómicos de grado mayor o igual a dos, y
ii) Transformando las variables tanto la predictora como la de respuesta.

9.4.1 Regresión Cuadrática

Un modelo cuadrático es de la forma:

$$Y = a + bX + cX^2 + \varepsilon$$

donde a, b y c son constantes a estimar. Usando la técnica de mínimos cuadrados se pueden obtener fórmulas explícitas para calcular a, b y c.

En **MINITAB,** para obtener la ecuación del modelo cuadrático, hay que elegir la opción **Quadratic** en la ventana de diálogo de **Fitted Line Plot** que es una opción del menú **Regression.** La ventana de diálogo se muestra en la Figura 9.11.

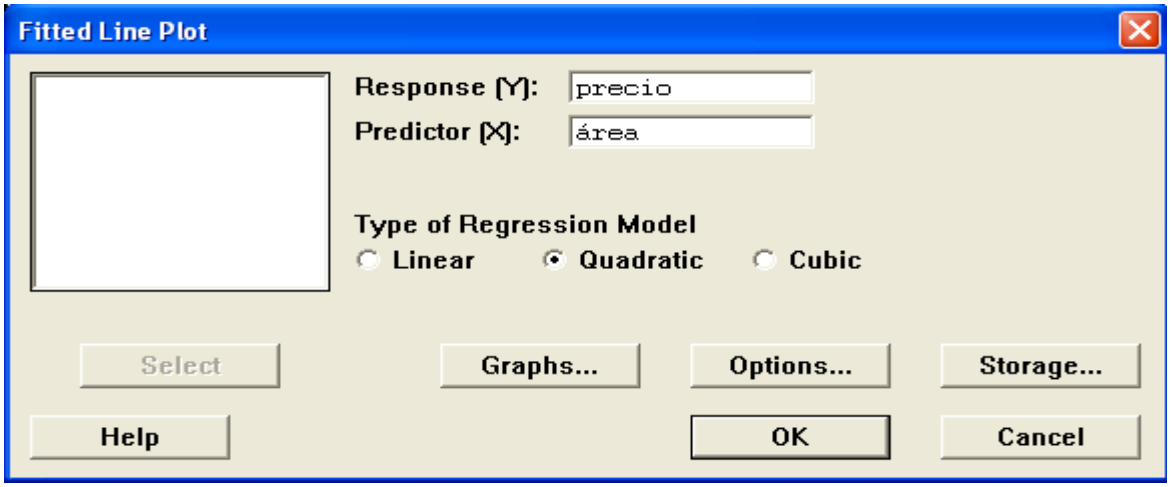

Figura 9.11. Ventana de diálogo para hacer una regresión cuadrática.

Ejemplo 9.2. Ajustar un modelo cuadrático para los datos del Ejemplo 9.1.

La ventana de diálogo se muestra en la Figura 9.11, y los resultados en la ventana **session** serán:

Polynomial Regression

```
precio = 117591 - 8.29281 area + 1.13E-02 area**2
R-Sq = 76.5 %

Analysis of Variance

SOURCE            DF        SS         MS          F          P
Regression         2  7.52E+09  3.76E+09    19.4906  1.70E-04
Error             12  2.31E+09  1.93E+08
Total             14  9.83E+09

SOURCE       DF    Seq SS          F          P
Linear        1  7.24E+09    36.3308  4.25E-05
Quadratic     1  2.77E+08    1.43495  0.254083
```

Además se obtiene el siguiente plot:

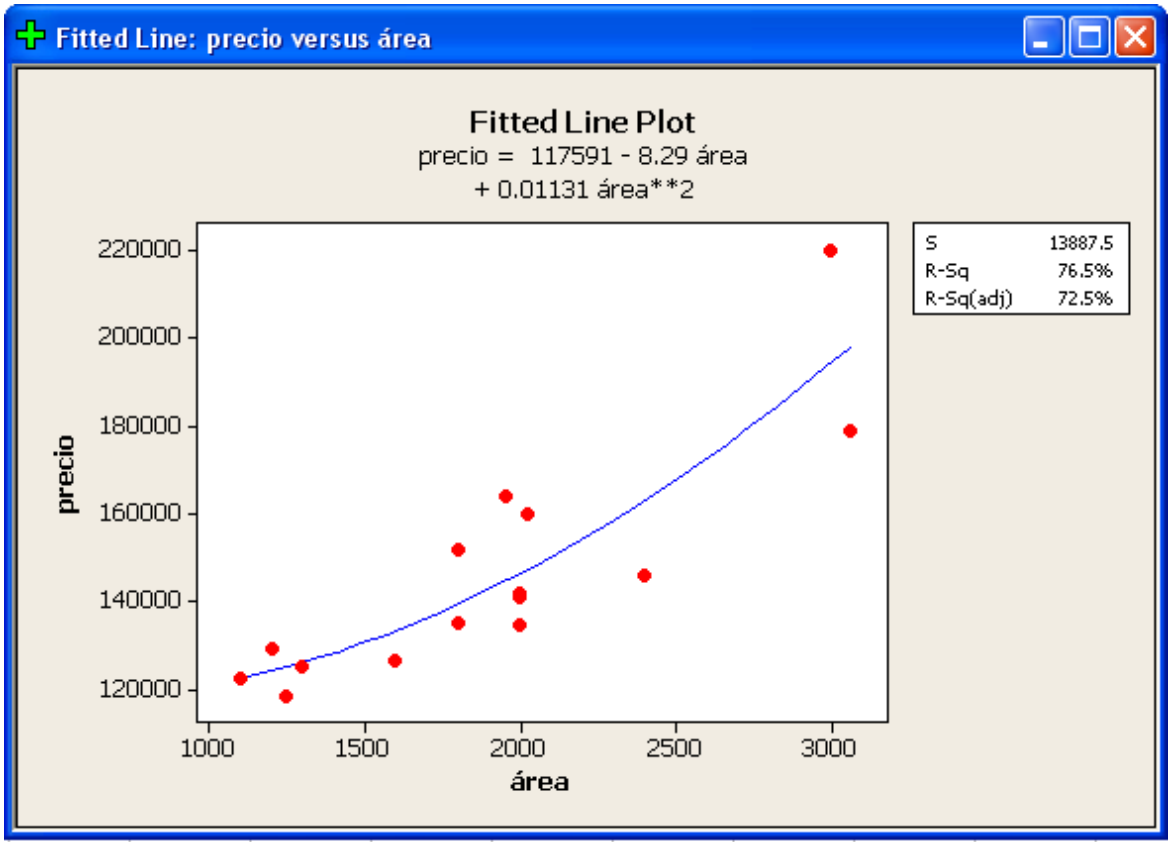

Figura 9.12. Regresión Cuadrática para el Ejemplo 9.1.

Interpretación: *El R^2 del modelo cuadrático es 76.5% comparado con 73.6% del modelo lineal (ver ejemplo 9.1), se ha ganado un 3% en confiabilidad, lo cual no es un aumento sustancial y se puede seguir usando un modelo lineal ya que hacer inferencias con él es mucho más simple que con un modelo cuadrático.*

También se pueden tratar modelos polinómicos más generales (el modelo cúbico sigue después del cuadrático), pero debido a que éstos presentan muchos cambios en la tendencia no son muy adecuados. Otro problema es que se puede llegar a un modelo "sobreajustado", es decir a un modelo que tiene un R^2 perfecto porque pasa por todos los puntos, pero que al momento de predecir fracasa terriblemente. Por ejemplo, si tenemos 8 observaciones, un modelo polinómico de grado 9 tendría un R^2 de 100%.

9.4.2 Modelos Nolineales que pueden ser transformados en lineales

La segunda alternativa para aumentar el R^2 consiste en usar modelos no lineales que pueden ser convertidos en lineales, a través de transformaciones tanto de la variable independiente como dependiente.
Después de hacer un plot para visualizar la relación entre X e Y se puede elegir entre los siguientes modelos linealizables:

Nombre del modelo	Ecuacion del Modelo	Transformación	Modelo Linealizado
Exponencial	$Y=\alpha e^{\beta X}$	Z=Ln Y X=X	Z=Ln α +βX
Logarítmico	Y= α +βLog X	Y=Y W=Log X	Y= α +βW
Doblemente Logarítmico	$Y=\alpha X^{\beta}$	Z=Log Y W=Log X	Z= Log α +βW
Hiperbólico	Y= α +β/X	Y=Y W=1/X	Y= α +βW
Inverso	Y=1/(α +βX)	Z=1/Y X=X	Z=α +βX

Para predecir el valor de Y usando el modelo linealizado hay que aplicar la inversa de la transformación correspondiente al mismo.

Ejemplo 9.3. Los siguientes datos representan como ha cambiado la poblacion en Puerto Rico desde 1930 hasta 1990.

Año	Población
1930	1543913
1940	1869255
1950	2210703
1960	2349544
1970	2712033
1980	3196520
1990	3522037

Se desea establecer un modelo para predecir la población de Puerto Rico en el año 2000.

Solución:
Observando el diagrama de puntos de población versus años que aparece en la figura de abajo.

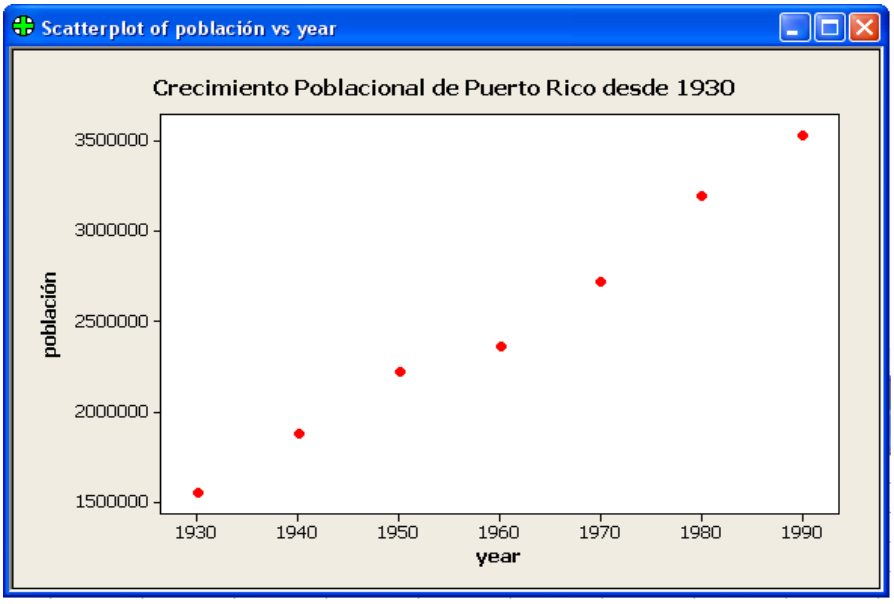

Figura 9.13. Crecimiento poblacional de Puerto Rico

El plot sugiere que podemos ajustar los datos al modelo exponencial:

$$Poblac=\alpha e^{\beta year}$$

Y el modelo linealizado da como ecuación:

$$Ln(Poblac) = - 11.4 + 0.0133\ year$$

con un R^2 del 98.9%, mejorando el R^2 del modelo lineal que era de 98.7%. Para predecir la población para el año 2000 se obtiene que:

$Ln(Poblac) = -11.4 + 0.0133 * 2000 = -11.4 + 26.6 = 15.2$

luego $Poblac = e^{15.2} = 3{,}992{,}787$. Así, 3,992,787 será la población estimada de PR para el año 2000.

9.5 Regresión lineal múltiple

Frecuentemente una sola variable predictora no es suficiente para explicar el comportamiento de la variable de respuesta. Por ejemplo, para explicar la nota que un estudiante saca en un examen lo primero que uno piensa es en el número de horas que estudio para tomarlo (X_1), pero también puede influir el número de créditos que lleva (X_2), el número de horas semanales que mira televisión (X_3), el número de horas que se divierte (X_4), el número de personas que viven con el o ella (X_5), etc. La idea en regresión lineal múltiple es usar más de una variable predictora para explicar el comportamiento de la variable de respuesta.

El modelo de regresión lineal múltiple con p variables predictoras $X_1,\ldots X_p$, es de la siguiente forma:

$$Y = b_0 + b_1 X_1 + b_2 X_2 + b_3 X_3 + ... + b_p X_p + \varepsilon$$

Las constantes $b_0, b_1, ..., b_p$, llamadas coeficientes de regresión, se estiman usando el método de mínimos cuadrados, y usando n observaciones de la forma $y_i, x_{i1}, x_{i2}, ..., x_{ip}$, donde $i = 1, ..., n$. La cantidad ε es una variable aleatoria con media 0 y varianza σ^2. Usando notación vectorial y matricial se puede escribir una fórmula explícita para los coeficientes de regresión, pero esto cae más allá del alcance de este texto. Se hará uso de **MINITAB** para hallar dichos coeficientes.

Interpretación del coeficiente de regresión estimado β_j

El estimado del coeficiente de regresión poblacional b_j, con $j = 1, ..., p$, se representará por β_j. Este estimado indica el cambio promedio en la variable de respuesta Y cuando la variable predictora X_j cambia en una unidad adicional asumiendo que las otras variables predictoras permanecen constantes.

Ejemplo 9.4 Se desea explicar el comportamiento de la variable de respuesta IGS (Indice General del Estudiante admitido a la Universidad de Puerto Rico) de acuerdo a X_1 (puntaje en la parte de aptitud matemática del College Borrad), X_2 (puntaje en la parte de aprovechamiento matemático) y X_3 (Tipo de Escuela; 1: Pública, 2: Privada). La muestra de 50 observaciones está disponible en el archivo **igs** de la página del texto.

Solución:
La ventana de diálogo de **Regression** se completa como se muestra en la siguiente figura:

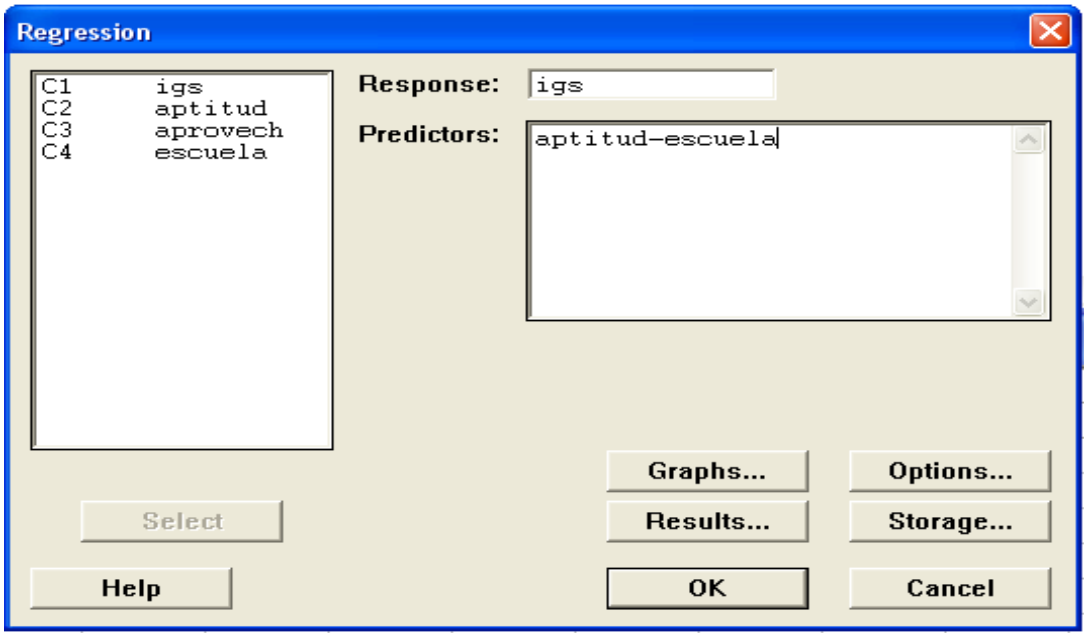

Figura 9.14. Ventana de diálogo para la regresión multiple del ejemplo 9.4

En la ventanita de **Response** se escribe la columna que contiene los datos de la variable dependiente *igs*, y en **Predictors**, se escriben las columnas que contienen las variables dependientes.

La ventana **session** cuando se ejecuta una regresión tendrá un contenido como el que sigue:

Regression Analysis: igs versus escuela, aprovech, aptitud

```
The regression equation is
igs = 136 + 1.93 escuela + 0.197 aprovech + 0.0569 aptitud


Predictor      Coef  SE Coef      T      P
Constant     135.93    24.50   5.55  0.000
escuela       1.933    3.091   0.63  0.535
aprovech    0.19698  0.03152   6.25  0.000
aptitud     0.05688  0.03140   1.81  0.077


S = 10.8896   R-Sq = 56.0%   R-Sq(adj) = 53.2%
```

```
Analysis of Variance

Source           DF       SS      MS      F      P
Regression        3   6952.0  2317.3  19.54  0.000
Residual Error   46   5454.8   118.6
Total            49  12406.9


Source    DF  Seq SS
escuela    1    52.9
aprovech   1  6510.1
aptitud    1   389.0


Unusual Observations

Obs  escuela     igs     Fit  SE Fit  Residual  St Resid
 18     1.00  263.00  286.58    6.47    -23.58     -2.69RX
 27     1.00  347.00  315.10    2.95     31.90      3.04R
 48     2.00  285.00  307.09    2.76    -22.09     -2.10R

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

Interpretación: *El coeficiente de una variable predictora indica el cambio promedio en la variable de respuesta igs cuando, se incrementa en una unidad la variable predictora asumiendo que las otras variables permanecen constantes. En este ejemplo, el aumento promedio en el igs es de 0.0569 por cada punto adicional en la parte de aptitud matemática, asumiendo que las otras dos variables permanecen constantes, asímismo el aumento promedio en el igs es de 0.197 por cada punto adicional en la parte de aprovechamiento matemático asumiendo que las otras variables permanezcan constantes y hay un aumento promedio de 1.93 en el igs cuando nos movemos de escuela pública a privada asumiendo que las otras variables permanecen constantes.*

Aún cuando el R^2 es bajo del 56%, eligiendo el botón **Options** se puede predecir el *igs* de un estudiante para hacer predicciones de la variable de respuesta Y para valores dados de las variables predictoras.

Por ejemplo el *igs* estimado de un estudiante que obtuvo 600 puntos en la prueba de aptitud y 750 en la prueba de aprovechamiento y que proviene de escuela privada será 321.66, como lo muestra el contenido de la ventana **session:**

```
Predicted Values for New Observations

New
Obs     Fit  SE Fit       95% CI            95% PI
  1  321.66    4.05  (313.51, 329.81)  (298.28, 345.05)


Values of Predictors for New Observations

New
Obs  escuela  aprovech  aptitud
  1     2.00       750      600
```

Estimación de la varianza σ^2

La estimación de la varianza de los errores σ^2 es crucial para hacer inferencias acerca de los coeficientes de regresión. Si en nuestro modelo hay *p* variables predictoras entonces, σ^2 es estimada por:

$$s^2 = \frac{\sum_{i=1}^{n}(\hat{Y}_i - Y_i)^2}{n-p-1} = \frac{SSE}{n-p-1} = MSE$$

Aquí, SSE representa la suma de cuadrados del error y MSE representa el cuadrado medio del error.

9.6 Inferencia en regresión lineal múltiple

9.6.1 Prueba de hipótesis de que cada coeficiente de regresión es cero

En este caso la hipótesis nula es $H_0 : \beta_j = 0$ ($j = 1,..., p$), o sea, la variable X_j no es importante en el modelo, versus la hipótesis alterna $H_a : \beta_j \neq 0$, que significa que la variable X_j si es importante. La prueba estadística es la prueba de *t* dada por:

$$t = \frac{\hat{\beta}_j}{s.e(\hat{\beta}_j)}$$

MINITAB da el valor de la prueba estadística y de los "p-values" correspondientes. En el Ejemplo 9.4 los "P-values" de la prueba de t que son mayores que .05 sugieren que las variables *Escuela* y aptitud no contribuyen al modelo, pues se acepta la hipótesis nula de que dicho coeficiente es cero. La variable *aprovechamiento* si es importante en el modelo ya que su "P-value" es menor que .05.

9.6.2 Prueba de hipótesis de que todos los coeficientes de regresión son ceros.

En este caso la hipótesis nula es $H_0 : \beta_1 = \beta_2 = ... = \beta_p = 0$, o sea, que el modelo no sirve, versus la hipótesis alterna H_a: Al menos uno de los coeficientes es distinto de cero, o sea, al menos una de las variables del modelo sirve.

La prueba estadística es la prueba de F que se obtiene al hacer la tabla del análisis de varianza para la regresión múltiple. La suma de cuadrados de Regresión tiene *p* grados de libertad que es igual al número de variables predictoras en el modelo. La Suma de Cuadrados del Total tiene $n-1$ grados de libertad y la suma de cuadrados del error tiene $n-p-1$ grados de libertad. Si la hipótesis nula es cierta, entonces:

$$F = \frac{\frac{SSR}{p}}{\frac{SSE}{n-p-1}} = \frac{MSR}{MSE}$$

Se distribuye como una F con p grados de libertad en el numerador y $n-p-1$ grados de libertad en el denominador.

En el Ejemplo 9.4, el "P-value" de la Prueba de F es 0.0000. Esto lleva a la conclusión de que el al menos una de las variables predictoras presentes en el modelo es importante para predecir el *igs*.

Por otro lado, el R^2 del 56% indica que el modelo no es muy confiable para hacer predicciones, porque sólo el 56% de la variación en el *igs* es explicada por su relación con las variables predoctoras.

9.6.3 Prueba de hipótesis para un subconjunto de coeficientes de regresión

Algunas veces estamos interesados en probar si algunos coeficientes del modelo de regresión son iguales a 0 simultánemente. Por ejemplo, si el modelo tiene *p* variables predictoras y quisiéramos probar si los k primeros coeficientes son ceros. O sea, $H_0 : \beta_1 = \beta_2 = ... = \beta_k = 0$. En este caso al modelo que tiene las p variables se le llama el **modelo completo** y al modelo que queda, asumiendo que la hipótesis nula es cierta, se le llama **modelo reducido.** Para probar si la hipótesis nula es cierta se usa una prueba de F que es llamada F-parcial. **La prueba de F parcial** se calcula por:

$$F_p = \frac{\frac{SSR(C) - SSR(R)}{k}}{\frac{SSE(C)}{n-p-1}} = \frac{\frac{SSR(C) - SSR(R)}{k}}{MSE(C)}$$

Donde, SSR(C) y MSE(C), representan la suma de cuadrados de regresión y el cuadrado medio del error del modelo completo, respectivamente, y SSR(R) es la suma de cuadrados de regresión del modelo reducido. Si F_p es mayor que $F_{1-\alpha}$, usando k grados de libertad para el numerador y $n-p-1$ para el denominador, entonces se rechaza H_o en caso contrario se acepta.

MINITAB no tiene una opción que haga directamente la prueba de F parcial. Hay que calcular los dos modelos de regresión y usar las sumas de cuadrados de regresión de ambos modelos para calcular la prueba de F parcial usando **Calculator.**

Ejemplo 9.5. Usando los datos del Ejemplo 9.4, probar la hipótesis $H_0 : \beta_1 = \beta_2 = 0$, versus Ha: al menos uno de los dos: β_1 o β_2 no es cero. Interpretar sus resultados.

Solución:

$H_0 : \beta_1 = \beta_2 = 0$ (significa que las variables: aptitud y aprovechamiento no influyen simultáneamente en la predicción del *igs*).

H_a: al menos uno de los dos: β_1 o β_2 no es cero (significa que al menos una de las dos variables influye en el comportamiento de Y)

En este caso p=3, k=2, p-k = 1, y de la tabla del análisis de varianza del Ejemplo 9.4, SSR(C) = 6952 y MSE(C) = 118.6. Para obtener SSR(R), se hace la regresión simple entre Y = *igs* y X = *aptitud* y de la tabla del análisis de varianza se obtiene SSR(R) = 203. Luego la prueba de F parcial será igual a $F_p = (6952 - 203/2)/118.6 = 29.128$. Por otro lado, para obtener la F con 2 g.l en el numerador y 46 en el denominador se usa la secuencia **calc ▸ probability distributions ▸ F** y se obtiene una $F = 3.1996$. Luego, se rechaza la hipótesis nula y se concluye, que al 5% de significación hay suficiente evidencia estadística para afirmar que al menos una de las dos variables (aptitud o aprovechamiento) influye en el comportamiento de la variable de respuesta Y.

En forma similar a la regresión lineal simple se pueden hacer predicciones de la variable de respuesta asignando valores adecuados a las variables predictoras. Asímismo, las gráficas que se usan para analizar los residuales pueden ser obtenidas usando la secuencia **stat ▸ regression ▸ regression**. Luego escoger opción **Graph** en la ventana de diálogo de **Regresión.** Escoger la opción **"*Four in one*"**. Para el Ejemplo 9.4 las gráficas resultantes son las siguientes:

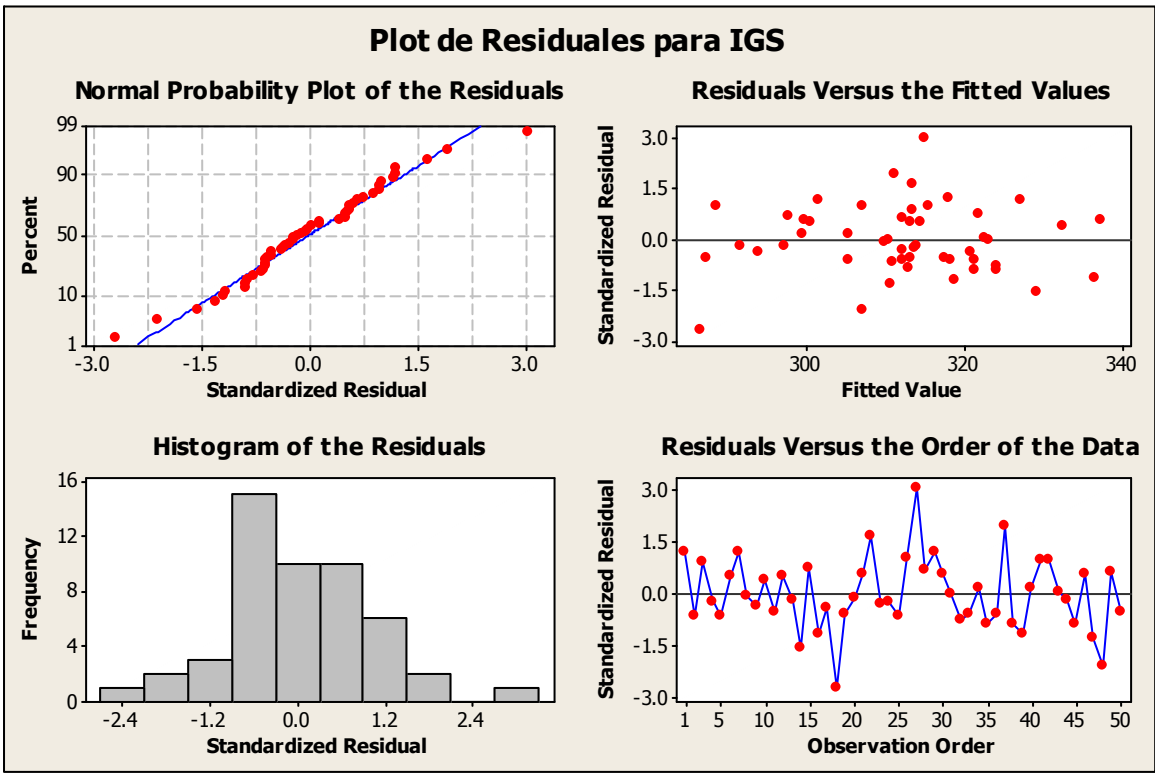

Figura 9.15. Análisis de Residuales para el Ejemplo 9.4

Interpretación: *El plot de normalidad y el histograma de los residuales indican que hay algo de normalidad en la distribución de los errores, debido a que los puntos no se alejan mucho de una línea recta en el primer plot y algo de simetría que se puede ver en el segundo. Sin embargo es clara la presencia de los "outliers" en ambos extremos, lo cual afecta la condición de normalidad.*

El plot de residuales versus el orden de la observación sugiere que las observaciones 18, 27 y 48 son "outliers" en el sentido vertical, estos "outliers" también se pueden notar en el plot de residuales versus valores predichos ("fits").

El plot de residuales versus valores predichos sugiere que la varianza de los errores es constante, porque no hay un patrón definido que siguen los puntos.

9.7 Selección de variables en Regresión Múltiple

Una buena propiedad de un modelo de regresión lineal es que permita explicar el comportamiento de la variable de respuesta Y lo mejor posible, haciendo uso del menor número de variables predictoras posibles, esta propiedad es llamada "parsimonía".

Existen dos métodos generales de lograr este objetivo: los métodos "stepwise" y el método de los mejores subconjuntos.

9.7.1 Los metodos "stepwise"

Comprenden los siguientes métodos:

Método de eliminación hacia atrás ("Backward Elimination"): Aqui en el paso inicial se incluyen en el modelo a todas las variables predictoras y en cada paso se elimina la variable cuyo "P-value" es más grande para la prueba de t o cuyo valor de la prueba t menor que 2 en valor absoluto. Una variable que es eliminada del modelo ya no puede volver a entrar en un paso subsiguiente. El proceso termina cuando todos los "P-values" son menores que .05, o cuando todos los valores de la prueba t son mayores que 2 en valor absoluto. Lo anterior también se puede hacer con una prueba F-parcial, puesto que $F = t^2$ (cuando el numerador tiene grados de libertad igual a 1). Luego, el método terminará cuando todas las F son mayores que 4.

Método de Selección hacia adelante ("Forward Selection"): Aqui en el paso inicial se considera una regresión lineal simple que incluye a la variable predictora que da la correlación más alta con la variable de respuesta. Luego se incluye una segunda variable en el modelo, que es aquella variable dentro de las no incluidas aún, que da el "P-value" más bajo para la prueba *t* o el valor de la prueba de *t* más grande en valor absoluto. Y así se siguen incluyendo variables, notando que una vez que ésta es incluida ya no puede ser sacada del modelo. El proceso termina cuando los "P-values" para la prueba t de todas las variables que aún no han sido incluidas son mayores que .05 ó la prueba de t es menor que 2 para dichas variables. Si se usa la prueba de F, entonces el proceso termina cuando todas las F son menores que 4.

Método Paso a Paso ("**Stepwise")**: Es una modificación del método "Forward", donde una variable que ha sido incluida en el modelo en un paso previo puede ser eliminada posteriormente. En cada paso se cotejan si todas las variables que están en el modelo deben permanecer alli. La mayoría de las veces, pero no siempre, los tres métodos dan el mismo resultado para el mejor modelo de regresión.

En **MINITAB**, la opción ***Stepwise*** del submenú **Regression** selecciona el mejor modelo de regresión usando los métodos "**Stepwise".** En el siguiente ejemplo se muestra el método **"stepwise"** paso por paso y luego directamente usando la opción **stepwise.**

Ejemplo 9.6. El conjunto de datos **grasa** contiene 13 variables que sirven para predecir el porcentaje de grasa en el cuerpo humano.

Columna	Nombre	
C1	grasa	VARIABLE DE RESPUESTA
C2	edad	en años
C3	peso	en libras
C4	altura	en pulgadas
C5	cuello	en cms
C6	pecho	en cms
C7	abdomen	en cms
C8	cadera	en cms
C9	muslo	en cms

C10	rodilla	en cms
C11	tobillo	en cms
C12	biceps	en cms
C13	antebrazo	en cms
C14	muñeca	en cms

Se tomaron las mediciones en 250 sujetos.

Se trata de hallar el mejor modelo de regresión usando los métodos "Stepwise".

Solución:

A) Usando el método de eliminación hacia atrás.

Primero, haremos paso a paso el método "Backward" y luego directamente con las opciones que están disponibles en **MINITAB**.

Paso 1. Se hace la regresión con todas las variables

```
The regression equation is
grasa = - 18.2 + 0.0621 edad - 0.0884 peso - 0.0696 altura - 0.471 cuello
           - 0.0239 pecho + 0.955 abdomen - 0.208 cadera + 0.236 muslo
           + 0.015 rodilla + 0.174 tobillo - 1.62 muñeca + 0.182 biceps
           + 0.452 antebrazo

Predictor        Coef       StDev          T        P
Constant       -18.19       17.35       -1.05    0.296
edad           0.06208      0.03235      1.92    0.056
peso          -0.08844      0.05353     -1.65    0.100
altura        -0.06959      0.09601     -0.72    0.469
cuello        -0.4706       0.2325      -2.02    0.044
pecho         -0.02386      0.09915     -0.24    0.810
abdomen        0.95477      0.08645     11.04    0.000
cadera        -0.2075       0.1459      -1.42    0.156
muslo          0.2361       0.1444       1.64    0.103
rodilla        0.0153       0.2420       0.06    0.950
tobillo        0.1740       0.2215       0.79    0.433
muñeca        -1.6206       0.5349      -3.03    0.003
biceps         0.1816       0.1711       1.06    0.290
antebraz       0.4520       0.1991       2.27    0.024

S = 4.305       R-Sq = 74.9%     R-Sq(adj) = 73.5%


Se elimina del modelo la variable rodilla, pués su "P-value"=0.950 es el mayor.
```

Paso 2. Regresion sin la variable rodilla

```
The regression equation is
grasa = - 17.9 + 0.0626 edad - 0.0876 peso - 0.0691 altura - 0.473 cuello
           - 0.0244 pecho + 0.954 abdomen - 0.207 cadera + 0.239 muslo
           + 0.176 tobillo - 1.62 muñeca + 0.181 biceps + 0.453 antebrazo

Predictor        Coef       StDev          T        P
Constant       -17.93       16.84       -1.06    0.288
edad           0.06259      0.03125      2.00    0.046
peso          -0.08758      0.05165     -1.70    0.091
altura        -0.06907      0.09545     -0.72    0.470
cuello        -0.4728       0.2293      -2.06    0.040
```

```
pecho           -0.02442     0.09855      -0.25    0.805
abdomen          0.95440     0.08606      11.09    0.000
cadera          -0.2071      0.1455       -1.42    0.156
muslo            0.2386      0.1384        1.72    0.086
tobillo          0.1763      0.2179        0.81    0.419
muñeca          -1.6181      0.5323       -3.04    0.003
biceps           0.1808      0.1703        1.06    0.289
antebraz         0.4532      0.1979        2.29    0.023

S = 4.296       R-Sq = 74.9%     R-Sq(adj) = 73.6%

Se elimina del modelo la variable pecho, pués su "p-value"=0.805 es el mayor.
```

Paso 3. Regresión sin las variables rodilla y pecho

```
The regression equation is
grasa = - 19.7 + 0.0625 edad - 0.0927 peso - 0.0638 altura - 0.475 cuello
           + 0.944 abdomen - 0.200 cadera + 0.245 muslo + 0.179 tobillo
           - 1.61 muñeca + 0.177 biceps + 0.448 antebrazo

Predictor        Coef       StDev          T        P
Constant       -19.69       15.24         -1.29    0.198
edad             0.06249     0.03118       2.00    0.046
peso            -0.09271     0.04723      -1.96    0.051
altura          -0.06378     0.09285      -0.69    0.493
cuello          -0.4754      0.2287       -2.08    0.039
abdomen          0.94421     0.07545      12.51    0.000
cadera          -0.2004      0.1427       -1.41    0.161
muslo            0.2451      0.1356        1.81    0.072
tobillo          0.1785      0.2173        0.82    0.412
muñeca          -1.6149      0.5311       -3.04    0.003
biceps           0.1771      0.1693        1.05    0.297
antebraz         0.4477      0.1963        2.28    0.023

S = 4.288       R-Sq = 74.9%     R-Sq(adj) = 73.7%

Se elimina del modelo la variable altura, pués su "p-value"=0.493 es el mayor.
```

Paso 4. Regresion sin las variables rodilla, pecho y altura

```
The regression equation is
grasa = - 26.0 + 0.0651 edad - 0.107 peso - 0.467 cuello + 0.958 abdomen
           - 0.179 cadera + 0.259 muslo + 0.185 tobillo - 1.66 muñeca
           + 0.186 biceps + 0.453 antebrazo


Predictor        Coef       StDev          T        P

Constant        -26.00       12.15         -2.14    0.033
edad              0.06509     0.03092       2.11    0.036
peso             -0.10740     0.04207      -2.55    0.011
cuello           -0.4675      0.2281       -2.05    0.042
abdomen           0.95772     0.07276      13.16    0.000
cadera           -0.1791      0.1391       -1.29    0.199
muslo             0.2593      0.1339        1.94    0.054
tobillo           0.1845      0.2169        0.85    0.396
muñeca           -1.6567      0.5271       -3.14    0.002
biceps            0.1862      0.1686        1.10    0.271
antebraz          0.4530      0.1959        2.31    0.022

S = 4.283       R-Sq = 74.8%     R-Sq(adj) = 73.8%

Se elimina del modelo la variable tobillo, pues su "p-value"=.396 es el mayor.
```

Paso 5. Regresión sin incluir las variables: rodilla, pecho, altura y tobillo

```
The regression equation is
grasa = - 23.3 + 0.0635 edad - 0.0984 peso - 0.493 cuello + 0.949 abdomen
           - 0.183 cadera + 0.265 muslo - 1.54 muñeca + 0.179 biceps
           + 0.451 antebrazo

Predictor        Coef       StDev          T        P
Constant       -23.30       11.73         -1.99    0.048
edad             0.06348     0.03084       2.06    0.041
peso            -0.09843     0.04070      -2.42    0.016
cuello          -0.4933      0.2260       -2.18    0.030
abdomen          0.94926     0.07204      13.18    0.000
cadera          -0.1829      0.1389       -1.32    0.189
muslo            0.2654      0.1336        1.99    0.048
muneca          -1.5421      0.5093       -3.03    0.003
biceps           0.1789      0.1683        1.06    0.289
antebraz         0.4515      0.1958        2.31    0.022

S = 4.281       R-Sq = 74.8%     R-Sq(adj) = 73.8%

Se elimina del modelo la variable biceps, pués su "p-value"=.289 es el mayor.
```

Paso 6. Regresión sin incluir las variables: rodilla, pecho, altura, tobillo y biceps

```
The regression equation is
grasa = - 22.7 + 0.0658 edad - 0.0899 peso - 0.467 cuello + 0.945 abdomen
           - 0.195 cadera + 0.302 muslo - 1.54 muneca + 0.516 antebrazo

Predictor        Coef       StDev          T        P
Constant       -22.66       11.71         -1.93    0.054
edad             0.06578     0.03078       2.14    0.034
peso            -0.08985     0.03991      -2.25    0.025
cuello          -0.4666      0.2246       -2.08    0.039
abdomen          0.94482     0.07193      13.13    0.000
cadera          -0.1954      0.1385       -1.41    0.159
muslo            0.3024      0.1290        2.34    0.020
muñeca          -1.5367      0.5094       -3.02    0.003
antebraz         0.5157      0.1863        2.77    0.006

S = 4.282       R-Sq = 74.7%     R-Sq(adj) = 73.8%

Se elimina del modelo la variable cadera, pués su "p-value"=.159 es el mayor.
```

Paso 7. Regresión sin incluir las variables: rodilla, pecho, altura, tobillo, biceps y cadera.

```
The regression equation is
grasa = - 33.3 + 0.0682 edad - 0.119 peso - 0.404 cuello + 0.918 abdomen
           + 0.222 muslo - 1.53 muneca + 0.553 antebrazo

Predictor        Coef       StDev          T        P
Constant      -33.258       9.007        -3.69    0.000
edad            0.06817     0.03079       2.21    0.028
peso           -0.11944     0.03403      -3.51    0.001
cuello         -0.4038      0.2206       -1.83    0.068
abdomen         0.91788     0.06950      13.21    0.000
muslo           0.2220      0.1160        1.91    0.057
muneca         -1.5324      0.5104       -3.00    0.003
```

```
antebraz        0.5531      0.1848        2.99    0.003

S = 4.291       R-Sq = 74.4%     R-Sq(adj) = 73.7%
```

Se elimina del modelo la variable cuello, pués su "p-value"=.068 es el mayor.

Paso 8. Regresión sin incluir las variables: rodilla, pecho, altura, tobillo, biceps, cadera y cuello.

```
The regression equation is
grasa = - 38.3 + 0.0629 edad - 0.136 peso + 0.912 abdomen + 0.220 muslo
           - 1.78 muñeca + 0.489 antebrazo

Predictor        Coef       StDev          T        P
Constant      -38.322       8.612      -4.45    0.000
edad            0.06290     0.03080     2.04    0.042
peso           -0.13648     0.03288    -4.15    0.000
abdomen         0.91179     0.06975    13.07    0.000
muslo           0.2202      0.1166      1.89    0.060
muneca         -1.7788      0.4947     -3.60    0.000
antebraz        0.4891      0.1823      2.68    0.008

S = 4.311       R-Sq = 74.1%     R-Sq(adj) = 73.5%
```

Se elimina del modelo la variable muslo, pués su "p-value".060 es el mayor.

Paso 9. Regresión sin incluir las variables: rodilla, pecho, altura, tobillo, biceps, cadera, cuello y muslo.

```
The regression equation is
grasa = - 31.0 + 0.0410 edad - 0.111 peso + 0.939 abdomen - 1.83 muñeca
           + 0.508 antebrazo

Predictor        Coef       StDev          T        P
Constant      -30.970       7.724      -4.01    0.000
edad            0.04100     0.02869     1.43    0.154
peso           -0.11095     0.03014    -3.68    0.000
abdomen         0.93901     0.06860    13.69    0.000
muñeca         -1.8296      0.4965     -3.68    0.000
antebraz        0.5085      0.1830      2.78    0.006

S = 4.334       R-Sq = 73.7%     R-Sq(adj) = 73.2%
```

Se elimina del modelo la variable edad, pués su "p-value"=.154 es el mayor.

Paso 10. Regresión sin incluir las variables: rodilla, pecho, altura, tobillo, biceps, cadera, cuello, muslo y edad.

```
The regression equation is
grasa = - 34.9 - 0.136 peso + 0.996 abdomen - 1.51 muñeca + 0.473 antebrazo

Predictor        Coef       StDev          T        P
Constant      -34.854       7.245      -4.81    0.000
peso           -0.13563     0.02475    -5.48    0.000
abdomen         0.99575     0.05607    17.76    0.000
muñeca         -1.5056      0.4427     -3.40    0.001
antebraz        0.4729      0.1817      2.60    0.010
```

```
S = 4.343       R-Sq = 73.5%     R-Sq(adj) = 73.1%
```

El proceso termina, porque todos los "p-values" son menores que 0.05 o las pruebas t en valor absoluto son mayores que 2. El mejor modelo para predecir el porcentaje de grasa en el cuerpo será el que incluyea las variables:peso, circunferencia de abdomen, nuñeca y antebrazo.

Ahora, haremos todo lo anterior en forma directa. La ventana de diálogo para hacer selección de variables en **MINITAB** se obtiene al elegir la opción **Stepwise** del menú **regresión.** La ventana de diálogo se completara como se muestra en la Figura 9.16

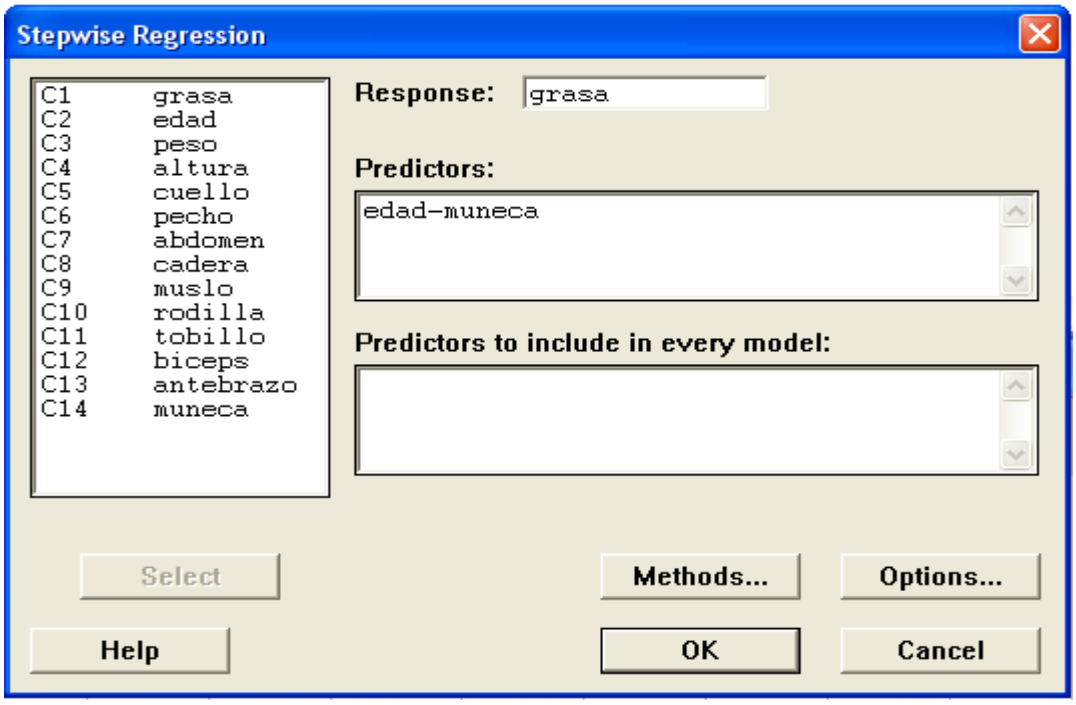

Figura 9.16. Ventana de diálogo para el método "Stepwise"

Al seleccionar **Methods** aparece la ventana de diálogo de la Figura 9.17:

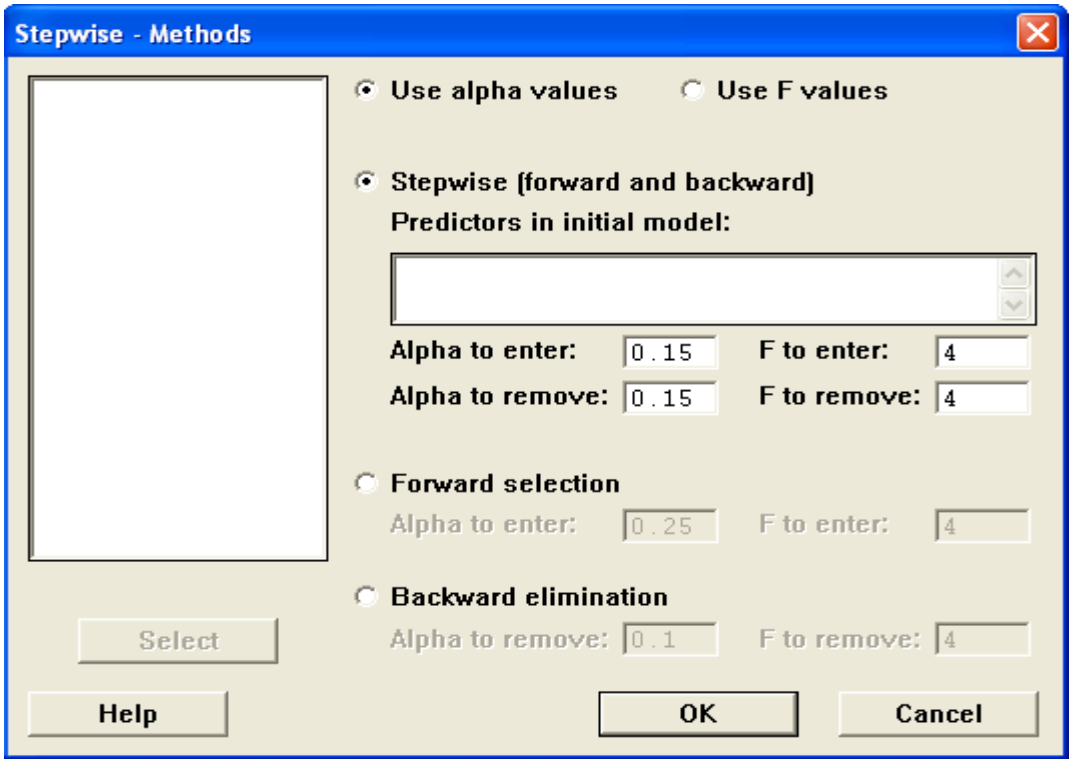

Figura 9.17. Ventana de diálogo que aparece al oprimir **methods** en "**stepwise**".

En el método de eliminación hacia atrás se selecciona **Backward Elimination**. Aparece seleccionado 0.15 en **Alpha to remove**. Este 0.15 es el nivel de significación que se usa en la prueba de F al momento de decidir si se elimina o no una variable del modelo. Este valor puede ser cambiado por el usuario. Si se elige un valor más pequeño de "alpha" entonces, es probable que el modelo incluya más variables predictoras, lo contrario ocurre si se elige un "alpha" grande.

En las versiones anteriores de **MINITAB** se usaba un valor de 4.0 en **F to Remove**. Este valor corresponde a un "alpha" de 0.05 cuando se tiene una F con 1 grado de libertad en el numerador y grados de libertad del denominador relativamente grande, mayor que 30. Con este cambio **MINITAB** ha adoptado la técnica de hacer "stepwise" que aparece en la mayoría de los programas estadísticos.

Para los datos de la hoja de trabajo **grasa.mtw** en donde se trata de ver qué medidas del cuerpo sirven para determinar el porcentaje de grasa en el cuerpo humano, el método de eliminación hacia atrás da los siguientes resultados:

Stepwise Regression: grasa versus edad, peso, ...

```
Backward elimination.  Alpha-to-Remove: 0.05

 Response is  grasa   on 13 predictors, with N =  252

    Step          1        2        3        4        5        6        7
Constant     -18.19   -17.93   -19.69   -26.00   -23.30   -22.66   -33.26
```

edad	0.062	0.063	0.062	0.065	0.063	0.066	0.068
T-Value	1.92	2.00	2.00	2.11	2.06	2.14	2.21
P-Value	0.056	0.046	0.046	0.036	0.041	0.034	0.028
peso	-0.088	-0.088	-0.093	-0.107	-0.098	-0.090	-0.119
T-Value	-1.65	-1.70	-1.96	-2.55	-2.42	-2.25	-3.51
P-Value	0.100	0.091	0.051	0.011	0.016	0.025	0.001
altura	-0.070	-0.069	-0.064				
T-Value	-0.72	-0.72	-0.69				
P-Value	0.469	0.470	0.493				
cuello	-0.47	-0.47	-0.48	-0.47	-0.49	-0.47	-0.40
T-Value	-2.02	-2.06	-2.08	-2.05	-2.18	-2.08	-1.83
P-Value	0.044	0.040	0.039	0.042	0.030	0.039	0.068
pecho	-0.024	-0.024					
T-Value	-0.24	-0.25					
P-Value	0.810	0.805					
abdomen	0.955	0.954	0.944	0.958	0.949	0.945	0.918
T-Value	11.04	11.09	12.51	13.16	13.18	13.13	13.21
P-Value	0.000	0.000	0.000	0.000	0.000	0.000	0.000
cadera	-0.21	-0.21	-0.20	-0.18	-0.18	-0.20	
T-Value	-1.42	-1.42	-1.41	-1.29	-1.32	-1.41	
P-Value	0.156	0.156	0.161	0.199	0.189	0.159	
muslo	0.24	0.24	0.25	0.26	0.27	0.30	0.22
T-Value	1.64	1.72	1.81	1.94	1.99	2.34	1.91
P-Value	0.103	0.086	0.072	0.054	0.048	0.020	0.057
rodilla	0.02						
T-Value	0.06						
P-Value	0.950						
tobillo	0.17	0.18	0.18	0.18			
T-Value	0.79	0.81	0.82	0.85			
P-Value	0.433	0.419	0.412	0.396			
biceps	0.18	0.18	0.18	0.19	0.18		
T-Value	1.06	1.06	1.05	1.10	1.06		
P-Value	0.290	0.289	0.297	0.271	0.289		
antebraz	0.45	0.45	0.45	0.45	0.45	0.52	0.55
T-Value	2.27	2.29	2.28	2.31	2.31	2.77	2.99
P-Value	0.024	0.023	0.023	0.022	0.022	0.006	0.003
muneca	-1.62	-1.62	-1.61	-1.66	-1.54	-1.54	-1.53
T-Value	-3.03	-3.04	-3.04	-3.14	-3.03	-3.02	-3.00
P-Value	0.003	0.003	0.003	0.002	0.003	0.003	0.003
S	4.31	4.30	4.29	4.28	4.28	4.28	4.29
R-Sq	74.90	74.90	74.90	74.85	74.77	74.66	74.45
R-Sq(adj)	73.53	73.64	73.75	73.81	73.84	73.82	73.71
C-p	14.0	12.0	10.1	8.5	7.2	6.4	6.3

Step	8	9	10
Constant	-38.32	-30.97	-34.85
edad	0.063	0.041	
T-Value	2.04	1.43	
P-Value	0.042	0.154	
peso	-0.136	-0.111	-0.136
T-Value	-4.15	-3.68	-5.48

P-Value	0.000	0.000	0.000
altura			
T-Value			
P-Value			
cuello			
T-Value			
P-Value			
pecho			
T-Value			
P-Value			
abdomen	0.912	0.939	0.996
T-Value	13.07	13.69	17.76
P-Value	0.000	0.000	0.000
cadera			
T-Value			
P-Value			
muslo	0.22		
T-Value	1.89		
P-Value	0.060		
rodilla			
T-Value			
P-Value			
tobillo			
T-Value			
P-Value			
biceps			
T-Value			
P-Value			
antebraz	0.49	0.51	0.47
T-Value	2.68	2.78	2.60
P-Value	0.008	0.006	0.010
muneca	-1.78	-1.83	-1.51
T-Value	-3.60	-3.68	-3.40
P-Value	0.000	0.000	0.001
S	4.31	4.33	4.34
R-Sq	74.10	73.72	73.50
R-Sq(adj)	73.46	73.19	73.07
C-p	7.7	9.2	9.3

Interpretación: *El método termina en 10 pasos. La primera variable eliminada del modelo es rodilla, cuyo valor de la prueba t, 0.06, es el más pequeño de todos, luego se eliminan, pecho, altura, tobillo, biceps, cadera, cuello, muslo y edad en ese orden. El mejor modelo será:*

Grasa= 34.85 -.136 peso+ .996 abdomen +0.47 antebrazo - 1.51muñeca

El cual tiene un R^2 de 73.50, mientras que el modelo completo con 13 variable predictoras tiene un R^2 de 74.90%, se ha perdido un 1.40% de confiablidad en las predicciones pero se ha economizado 9 variables, lo cual es más conveniente.

B) Usando el método "Forward"

Haciendo paso a paso el método "Forward":

Paso 1. Se halla la regresión simple con la variable predictora más altamente correlacionada con la variable de respuesta En este caso, es *abdomen* que tiene correlación 0.803 con *grasa.*

```
The regression equation is
grasa = - 39.3 + 0.631 abdomen

Predictor        Coef       StDev          T        P
Constant      -39.280       2.660       -14.77    0.000
abdomen         0.63130     0.02855      22.11    0.000

S = 4.877       R-Sq = 66.2%     R-Sq(adj) = 66.0%
```

Paso 2. Se halla todas las regresiones con dos variables predictoras, una de las cuales es abdomen. Aqui se muestran sólo dos de las 12 regresiones posibles.

Con la variables aabdomen y pecho

```
The regression equation is
grasa = - 30.3 + 0.818 abdomen - 0.261 pecho

Predictor        Coef       StDev          T        P
Constant      -30.274        4.057       -7.46     0.000
abdomen         0.81794      0.07006      11.67    0.000
pecho          -0.26066      0.08961     -2.91     0.004
S = 4.806       R-Sq = 67.3%     R-Sq(adj) = 67.0%
```

Con las variables adomen y peso

```
The regression equation is
grasa = - 46.0 + 0.990 abdomen - 0.148 peso

Predictor        Coef       StDev          T        P
Constant      -45.952       2.605       -17.64    0.000
abdomen         0.98950     0.05672      17.45    0.000
peso           -0.14800     0.02081      -7.11    0.000

S = 4.456       R-Sq = 71.9%     R-Sq(adj) = 71.7%
```

Notar que el valor absoluto de la prueba *t* para la variable pecho es 2.91 (p-value = .004), y para la variable peso es 7.11 (p-value = 0.000). La variable peso entra al modelo porque es aquella con valor de t más grande en valor absoluto entre todas las variables que aún no estaban incluídas.

Paso 3. Se hallan todas las regresiones con tres variables predictoras, las dos incluidas en los dos pasos anteriores y cada una de las variables no incluidas aún. Aqui se muestran sólo dos de las 11 regresiones posibles.

```
The regression equation is
grasa = - 45.8 + 0.990 abdomen - 0.148 peso - 0.002 cadera

Predictor        Coef       StDev          T        P
Constant      -45.846       7.059       -6.49     0.000
abdomen         0.98974     0.05866      16.87    0.000
peso           -0.14763     0.03087      -4.78    0.000
cadera         -0.0020      0.1199       -0.02    0.987

S = 4.465       R-Sq = 71.9%     R-Sq(adj) = 71.5%
```

Regression Analysis

```
The regression equation is
grasa = - 27.9 + 0.975 abdomen - 0.114 peso - 1.24 muneca

Predictor        Coef       StDev          T        P
Constant      -27.930       6.817      -4.10    0.000
abdomen        0.97513      0.05615    17.37    0.000
peso          -0.11446      0.02364    -4.84    0.000
muñeca        -1.2449       0.4362     -2.85    0.005

S = 4.393       R-Sq = 72.8%     R-Sq(adj) = 72.4%
```

La variable muñeca entra al modelo porque es aquella con el valor de t más grande en valor absoluto entre todas las variables que aún no estaban incluídas.

Paso 4. Se hallan todas las regresiones con cuatro variables predictoras, las tres incluídas en los tres pasos anteriores y cada una de las variables no incluidas aún. Aqui se muestran sólo dos de las 10 regresiones posibles.

Regression Analysis

```
The regression equation is
grasa = - 35.1 + 0.979 abdomen - 0.144 peso - 1.10 muñeca + 0.158 muslo

Predictor        Coef       StDev          T        P
Constant      -35.117       8.414      -4.17    0.000
abdomen         0.97856     0.05607    17.45    0.000
peso           -0.14355     0.03096    -4.64    0.000
muñeca         -1.0990      0.4467     -2.46    0.015
muslo           0.1585      0.1092      1.45    0.148

S = 4.383       R-Sq = 73.0%     R-Sq(adj) = 72.6%
```

Regression Analysis

```
The regression equation is
grasa = - 34.9 + 0.996 abdomen - 0.136 peso - 1.51 muñeca + 0.473 antebrazo

Predictor        Coef       StDev          T        P
Constant      -34.854       7.245      -4.81    0.000
abdomen         0.99575     0.05607    17.76    0.000
peso           -0.13563     0.02475    -5.48    0.000
muñeca         -1.5056      0.4427     -3.40    0.001
antebraz        0.4729      0.1817      2.60    0.010

S = 4.343       R-Sq = 73.5%     R-Sq(adj) = 73.1%
```

La variable antebrazo entra al modelo porque es aquella con el valor de *t* más grande en valor absoluto entre todas las variables que aún no estaban incluídas.

Aquí termina el proceso porque al hacer las regresiones de grasa con las cuatro variables consideradas hasta ahora y cada una de las 9 variables no incluidas hasta ahora se obtienen "p-values" para la prueba *t* mayores de 0.05.

Para hacer selección hacia adelante en **MINITAB** se sigue la secuencia **STAT ▸ Regression ▸ Stepwise ▸ Methods** y luego se elige **Forward Selection.** En la ventanita **Alpha-to-Enter** aparece 0.25, que es el nivel de significación que usa la prueba de F para decidir si una variable debe o no entrar en el modelo. Este valor puede ser cambiado por el usuario, tomando en cuenta que si elige un valor de "alpha" más pequeño es más probable que el modelo incluya un menor número de variables que cuando se escoge una "alpha" más grande.

En las versiones anteriores de **MINITAB** se usaba un valor de 4.0 en **F to Enter**. Este valor corresponde a un "alpha" de 0.05 cuando se tiene una F con 1 grado de libertad en el numerador y grados de libertad del denominador relativamente grande, mayor que 30.

Para los datos de la hoja de trabajo **grasa.mtw**, el método de selección hacia adelante da los siguientes resultados, usando "alpha" = 0.05.

Stepwise Regression: grasa versus edad, peso, ...

```
Forward selection.  Alpha-to-Enter: 0.05

 Response is  grasa   on 13 predictors, with N =  252
    Step          1        2        3        4
Constant     -39.28   -45.95   -27.93   -34.85

abdomen        0.631    0.990    0.975    0.996
T-Value        22.11    17.45    17.37    17.76
P-Value        0.000    0.000    0.000    0.000

peso                   -0.148   -0.114   -0.136
T-Value                 -7.11    -4.84    -5.48
P-Value                 0.000    0.000    0.000

muneca                           -1.24    -1.51
T-Value                          -2.85    -3.40
P-Value                          0.005    0.001

antebraz                                   0.47
T-Value                                    2.60
P-Value                                   0.010

S              4.88     4.46     4.39     4.34
R-Sq          66.17    71.88    72.77    73.50
R-Sq(adj)     66.03    71.65    72.44    73.07
C-p            72.9     20.7     14.2      9.3
```

C) Usando el método "Stepwise".

Para llevar a cabo en **MINITAB** selección de variables usando el método "stepwise" se sigue la secuencia **STAT ▸ Regression ▸ Stepwise ▸ Methods** y luego se elige **Stepwise.** Aparece la ventana de diálogo de la Figura 9.18. En las ventanitas **Alpha-to-Enter** y **Alpha to-Remove,** aparece el mismo valor 0.15, el cual puede ser cambiado por el usuario. El valor de **Alpha-to-Enter** debe ser menor que **Alpha to-Remove.** En las versiones anteriores de MINITAB aparecían las ventanitas **F-to-Enter** y **F-to-Remove** donde se asignaba el valor de 4.0

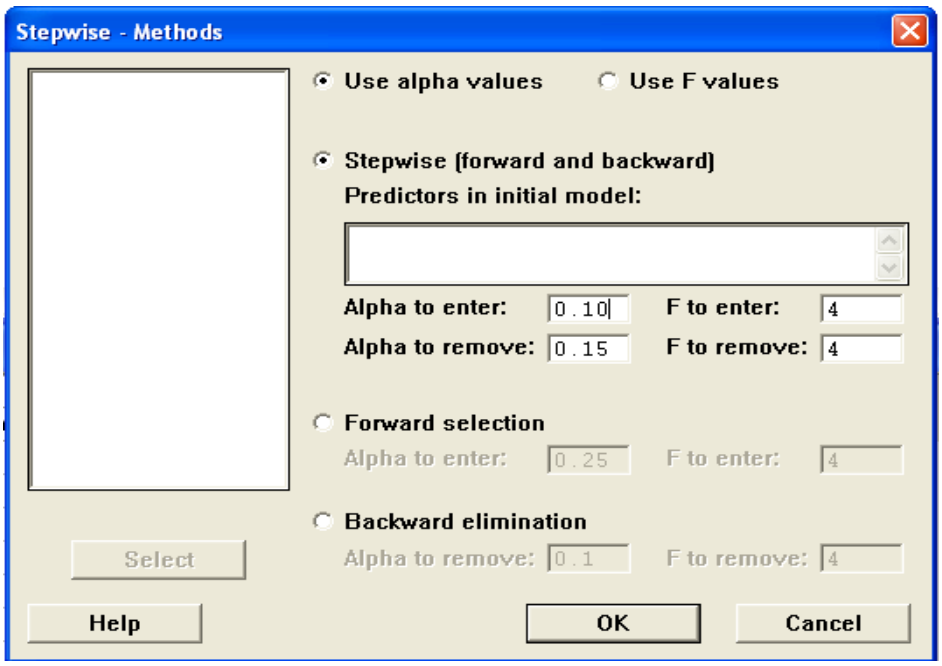

Figura 9.18. Ventana de diálogo para hacer selección "stepwise".

Para el conjunto de datos **grasa** el método "stepwise" usando **Alpha-to-Enter** = 0.10 y **Alpha to-Remove** = 0.05**,** produce los siguientes resultados:

Stepwise Regression: grasa versus edad, peso, ...

```
  Alpha-to-Enter: 0.1  Alpha-to-Remove: 0.15

 Response is  grasa   on 13 predictors, with N =  252


    Step          1        2        3        4        5
Constant     -39.28   -45.95   -27.93   -34.85   -30.65

abdomen        0.631    0.990    0.975    0.996    1.008
T-Value        22.11    17.45    17.37    17.76    17.89
P-Value        0.000    0.000    0.000    0.000    0.000

peso                   -0.148   -0.114   -0.136   -0.123
T-Value                 -7.11    -4.84    -5.48    -4.75
P-Value                 0.000    0.000    0.000    0.000

muneca                           -1.24    -1.51    -1.25
T-Value                          -2.85    -3.40    -2.66
P-Value                          0.005    0.001    0.008

antebraz                                   0.47     0.53
```

```
T-Value                                          2.60     2.86
P-Value                                         0.010    0.005

cuello                                                   -0.37
T-Value                                                  -1.65
P-Value                                                  0.100

S              4.88     4.46     4.39     4.34     4.33
R-Sq          66.17    71.88    72.77    73.50    73.79
R-Sq(adj)     66.03    71.65    72.44    73.07    73.26
C-p            72.9     20.7     14.2      9.3      8.6
```

9.7.2 Método de los mejores subconjuntos.

La opción **Best Subsets** del submenú **Regression** del menú **Stat** se usa para seleccionar los mejores modelos para un número dado de variables de acuerdo a 3 criterios:

El coeficiente de Determinación. El mejor modelo es aquél con $R^2 = \frac{SSR}{SST}$ más alto pero con el menor número de variables posibles. Por decir, si con 3 variables predictoras se obtiene un R^2 de .84 y con 4 variables se obtiene un R^2 de .87 se debería preferir el primer modelo porque la cuarta variable ha incrementado el R^2 pero por muy poco.

El coeficiente de Determinación Ajustado. Es una variante del R^2 y que a diferencia de éste no aumenta necesariamente al incluir una variable adicional en el modelo. Se calcula por:

$$R^2_{Ajust} = \frac{MSR}{MST} = 1-(1-R^2)\frac{n-1}{n-p-1}$$

La manera de usar este criterio es similar al anterior.

El **Coeficiente C_p de Mallows**. Se calcula por:

$$C_p = \frac{SSE_p}{s^2} + 2(p+1) - n$$

Donde SSE_p es la suma de cuadrados del error del modelo que incluye p variables predictoras y s^2 es la varianza estimada del error en el modelo que incluye todas las variables.

El mejor modelo es aquel para el cual se cumple aproximadamente $C_p = p+1$, pero con el menor número de variables posibles. Notar que la igualdad anterior también se cumple cuando se usa el modelo completo.

Para el ejemplo anterior, la ventana de diálogo aparece a continuación:

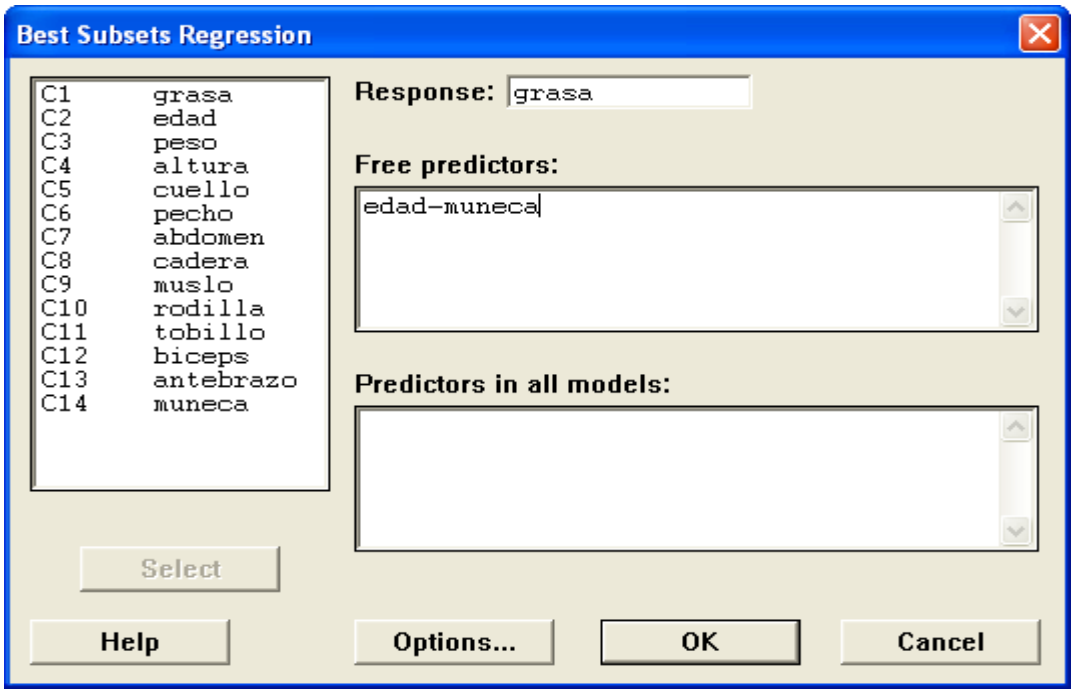

Figura 9.19. Ventana de diálogo para **Bests Subsets**, usando los datos del Ejemplo 9.6

y la ventana **session** contendrá los siguientes resultados:

Best Subsets Regression

Response is grasa

Vars	R-Sq	Adj. R-Sq	C-p	s	edad	peso	altura	cuello	pecho	abdomen	cadera	muslo	rodilla	tobillo	biceps	antebrazo	muneca
1	66.2	66.0	72.9	4.8775						X							
1	49.4	49.2	232.2	5.9668					X								
2	71.9	71.7	20.7	4.4556		X				X							
2	70.2	70.0	36.6	4.5866						X							X
3	72.8	72.4	14.2	4.3930		X				X							X
3	72.4	72.0	18.0	4.4251		X		X		X							
4	73.5	73.1	9.3	4.3427		X				X						X	X
4	73.3	72.8	11.4	4.3609		X				X					X		X
5	73.8	73.3	8.6	4.3276		X		X		X						X	X
5	73.7	73.2	9.2	4.3336	X	X				X						X	X
6	74.1	73.5	7.7	4.3111	X	X				X		X				X	X
6	74.1	73.4	8.0	4.3138	X	X		X		X						X	X
7	74.4	73.7	6.3	4.2906	X	X		X		X		X				X	X
7	74.3	73.6	7.4	4.2998	X	X		X		X					X	X	X
8	74.7	73.8	6.4	4.2819	X	X		X		X	X	X				X	X
8	74.6	73.8	7.0	4.2872	X	X		X		X		X			X	X	X
9	74.8	73.8	7.2	4.2808	X	X		X		X	X	X			X	X	X
9	74.7	73.8	7.7	4.2851	X	X	X	X		X	X	X				X	X
10	74.8	73.8	8.5	4.2832	X	X		X		X	X	X		X	X	X	X

10	74.8	73.8	8.7	4.2850	X	X	X	X		X	X	X			X	X	X
11	74.9	73.7	10.1	4.2879	X	X	X	X		X	X	X		X	X	X	X
11	74.8	73.7	10.5	4.2920	X	X		X	X	X	X	X		X	X	X	X
12	74.9	73.6	12.0	4.2963	X	X	X	X	X	X	X	X		X	X	X	X
12	74.9	73.6	12.1	4.2968	X	X	X	X		X	X	X	X	X	X	X	X
13	74.9	73.5	14.0	4.3053	X	X	X	X	X	X	X	X	X	X	X	X	X

Interpretación: *De acuerdo al R^2 el mejor modelo podría ser aquél con las dos variables predoctoras peso y abdomen que aún cundo su R^2 es de 71.9 está cerca del mayor posible que es de 74.9 y además es donde el R^2 ha tenido un mayor incremento. Un resultado similar cuando se usa el R^2 ajustado. De acuerdo al C_p de Mallows, el mejor modelo es aquél que tiene las siguientes 6 variables predictoras: edad, peso, muslo, abdomen, antebrazo y cadera con un valor de $C_p = 7.7$ muy próximo a $p+1=6+1=7$.*

EJERCICIOS

Para conseguir los archivos de datos accesar a la siguiente dirección en la internet www.math.uprm.edu/~edgar.datos.html o mandar un mensaje al autor.

1. Los siguientes datos corresponden el tiempo de experiencia en dias de 10 técnicos recientemente contratados por una compañia de electricidad, y el tiempo (en minutos) que demoran en hacer una instalación.

experiencia (X) :	5	2	3	10	7	6	5	7	1	8
tiempo de demora (Y) :	30	42	35	20	28	31	32	19	39	25

   a) Construir un diagrama de dispersión ("scatterplot") de los datos.
   b) Hallar la línea de cuadrados mínimos que representa la relación entre la experiencia y el tiempo de demora.
   c) Calcular el coeficiente de Determinación e interpretar el resultado.
   d) Probar usando un 5% de significación si la pendiente de ésta relación es cero.
   e) Si se sabe que un técnico tiene 5 dias de experiencia, ¿En cuánto tiempo se espera que realice una instalación?
   f) Hallar el intervalo de confianza del 95% del tiempo medio de duración para todos los ténicos que tienen 5 días de experiencia. Calcular también el intervalo de prediccion. Interpretar sus resultados.
   g) Hacer un análisis de varianza y sacar sus conclusiones.

2. La tienda "Sweet Dreams", especializada en vender dulces y regalos, registra durante 12 días el número de personas que entran a la tienda y la cantidad de venta (en dólares) de dulces en cada uno de esos días.

# de personas (X) :	174	112	166	138	172	90	148	116	196	116	124	95
ventas (Y) :	145.2	83.2	120.5	113.6	119	67	109.3	96.8	140.8	77.8	105	98.6

   a) Construir un diagrama de dispersión ("scatterplot") de los datos.
   b) Hallar la linea de cuadrados mínimos para aproximar la relación entre el número de personas que entran a la tienda y la venta de dulces por día.
   c) Probar a un 5% de nivel de significancia si la pendiente es cero.
   d) Probar a un 5% de nivel de significancia si el intercepto es cero.
   e) Calcular el coeficiente de correlación entre el número de personas y las ventas.
   f) Calcular el coeficiente de Determinación e interpretar éste resultado.
   g) Si el número de personas que entran a la tienda es de 130, predecir las ventas de ese día a un 95% de confianza.
   h) Obtener las bandas de confianza para el valor medio y de predicción
   i) Realizar un análisis de varianza y sacar sus conclusiones.

3. En un país se eligen 10 pueblos al azar y se anota el ingreso personal promedio de los habitantes (en miles) y la tasa de divorcio (por cada 1000 personas). Los datos están en el archivo **divorcio.**

   a) Hacer un plot de los datos
   b) Hallar el coeficiente de correlación r e interpretarlo
   c) Hallar la línea de regresión estimada e interpretar las constantes a y b
   d) Probar si la pendiente de la línea de regresión es cero.
   e) Trazar la línea de regresión sobre el plot de los puntos
   f) ¿Cuánto es el coeficiente de Determinación y qué significa?
   g) ¿Cuál será la tasa de divorcio estimada de un pueblo en donde el ingreso promedio anual es 12,500.
   h) Hallar además el intervalo de confianza del valor medio y el intervalo de predicción. Interpretar cada uno de ellos.
   i) Obtener la gráfica de las bandas de confianza.
   j) ¿Qué conclusión se obtendrá de la siguiente tabla de análisis de varianza?
   k) Hallar una regresión que pase por el origen e interpretar el resultado.

4. En un pueblo se eligen 15 personas al azar y se anota su salario mensual (X) y la cantidad que ahorran mensualmente (Y):

Salario	Ahorro
800	150
850	100
900	280
1200	400
1500	350
1700	500
1900	635
2000	600
2300	750
2500	680
2700	900
3000	800
3200	300
3500	1200
5000	1000

   a) Hallar la línea de regresión. e interpretar sus coeficientes.
   b) Trazar la linea de regresión por encima del diagrama de puntos.
   c) Probar la hipótesis de que la pendiente es cero. Comentar su resultado
   d) Hacer una regresión que pase por el orígen e interpretar la pendiente
   e) Asigne un valor adecuado a la variable predictora y halle un intervalo de confianza del 90 por ciento para el valor medio de la variable de respuesta e intrepretar el resultado.

f) Asigne un valor adecuado a la variable predictora y halle un intervalo de predicción del 95% para un valor individual de la variable, de respuesta e interpretar su resultado.
g) Obtenga las bandas de confianza para el valor medio y de predicción y explicar para qué se usan..
h) Interpretar el coeficiente de determinación
i) Hacer un análisis de residuales y comentar sus resultados
j) Si existen "outliers" eliminar uno de ellos y explicar su efecto en los cálculos del coeficiente de determinación y de la linea de regresión.
k) Hacer una regresión cuadrática y compararla con la regresión lineal

5. El conjunto de datos **brain** contiene las variables:
MRI (X), conteo en pixels del 18 scans de resonancia magnética del cerebro de una persona
Score_IQ, (Y) score en un test de inteligencia.
Mientras más alto sea el conteo de pixels más grande es el cerebro de las personas.

a) Hallar la línea de regresión. e interpretar los coeficientes de la linea de regresión
b) Trazar la línea de regresión encima del diagrama de puntos.
c) Probar la hipótesis de que la pendiente es cero. Comentar su resultado
d) Hacer una regresión que pase por el origen e interpretar la pendiente
e) Asigne un valor adecuado a la variable predictora y halle un intervalo de confianza del 90 por ciento para el valor medio de la variable, de respuesta e intrepretar el resultado.
f) Asigne un valor adecuado a la variable predictora y halle un intervalo de prediccion del 95% para un valor individual de la variable, de respuesta e interpretar su resultado.
g) Obtenga las bandas de confianza para el valor medio y de prediccion y explicar para qué se usan.
h) Interpretar el coeficiente de determinación
i) Hacer un análisis de residuales y comentar sus resultados
j) Si existen "outliers" eliminar uno de ellos y explicar su efecto en los cálculos del coeficiente de determinación y de la linea de regresión.
k) Hacer una regresión cuadrática y compararla con la regresión lineal

6. El conjunto de datos **pesobajo** contiene las variables:
peso, (Y): peso del recién nacido en gramos
duración (X): duración del período de gestación

a) Hallar la linea de regresión. e interpretar los coeficientes de la linea de regresión
b) Trazar la linea de regresión encima del diagrama de puntos.
c) Probar la hipótesis de que la pendiente es cero. Comentar su resultado
d) Hacer una regresión que pase por el origen e interpretar la pendiente

e) Asigne un valor adecuado a la variable predictora y halle un intervalo de confianza del 90 por ciento para el valor medio de la variable, de respuesta e intrepretar el resultado.
f) Asigne un valor adecuado a la variable predictora y halle un intervalo de predicción del 95% para un valor individual de la variable, de respuesta e interpretar su resultado.
g) Obtenga las bandas de confianza para el valor medio y de predicción y explicar para qué se usan.
h) Interpretar el coeficiente de determinación
i) Hacer un análisis de residuales y comentar sus resultados
j) Si existen "outliers" eliminar uno de ellos y explicar su efecto en los cálculos del coeficiente de determinación y de la linea de regresión.
k) Hacer una regresión cuadrática y compararla con la regresión lineal

7. En la siguiente tabla se presentan las presiones arteriales Sistólica y Diastólica de 20 personas

persona	pres. Sisto	pres. Dias	Persona	pres. Sisto	pres. Dias
1	130	80	11	120	75
2	100	70	12	130	95
3	130	80	13	130	80
4	140	80	14	140	90
5	130	70	15	110	80
6	115	75	16	160	95
7	120	85	17	150	110
8	125	75	18	130	95
9	110	65	19	125	75
10	125	70	20	130	80

a) Construya un diagrama de dispersión (“scatteplot”) para los datos.
b) Hallar la Regresión lineal, considerando como variable dependiente la Presión Arterial Diastólica.
c) Interpretar los coeficientes de la regresión obtenida en la parte a).
d) Trazar la línea de regresión estimada encima del diagrama de Dispersión.
e) Estime la presión Arterial Diastólica de una persona que tiene una presión Arterial Sistólica de 128.
f) Determine un intervalo al 95% para el valor medio de la variable, de respuesta si la presión arterial Sistólica es de 128, interpretarlo.
g) Graficar las bandas de confianza para el valor medio y de predicción.
h) Realizar un análisis de Residuales.

8. La siguiente tabla muestra el número (en cientos) de bacterias que sobreviven después de ser expuestas a rayos X de 200 kilovoltios por períodos de tiempo T de 6 minutos de duración cada uno:

Tiempo	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Bacterias	355	211	197	166	142	106	104	60	56	38	36	32	21	19	13

a) Hacer un plot de los datos que relacione el número de bacterias sobrevivientes versus el tiempo.
b) Ajustar varios modelos que pueden ser linealizados y decidir acercar el mejor modelo para representar la relación entre las variables.
c) Predecir el número de bacterias sobrevivientes después de 18 periodos de exposición

9. Usar los archivos de datos **homedat.mtw, salary.mtw y pulse.mtw** que están dentro de **MINITAB.** Para **homedat** escoger Y=c1, para **salary** escoger Y=C7 y no usar las columnas c1 y c2, para **Pulse** escoger Y=c2.

a) Hallar el modelo de regresión múltiple e interpretar tres de los coeficientes de regresión.
b) Interpretar el coeficiente de Determinación.
c) Probar que todos los coeficientes del modelo de regresión son ceros. Comentar el resultado.
d) Probar que cada uno de los coeficientes del modelo de regresión es cero. Comentar el resultado.
e) Probar la hipótesis Ho: $B_2=B_4=0$. Comentar su resultado.
f) Hallar un Intervalo de Confianza para el valor medio de Y y el valor Predicho del 99% para Y, escogiendo valores adecuados de las variables predictoras. Comentar sus resultados
g) Usar los métodos "Backward" y "Forward" para elegir el modelo de Regresión. Interpretar la salida de **MINITAB**. Osea explicar cada paso del método y porqué es que se detiene.

10. Dada la siguiente información:

Y: medida de severidad de la enfermedad respiratoria
X1: años de educación
X2: número de personas en el edificio donde vive la persona
X3: medida de la calidad del aire (un número grande indica pobre calidad)
X4: nivel de nutrición
X5: 0 es no fuma y 1 si fuma.

Y	X1	X2	X3	X4	X5
40	7	25	22	94	0
67	7	33	61	18	1
30	6	19	30	103	0
71	15	29	50	17	1
47	11	21	43	109	0
53	10	24	54	0	1
39	8	21	28	33	0
55	14	22	35	21	1
47	10	26	22	76	0
56	9	32	43	97	1
43	8	22	48	104	0
41	8	19	27	37	0

51	9	28	32	87	1
48	8	22	62	131	0
36	8	19	37	53	0

a) Hallar la regresión lineal múltiple. Comentar los coeficientes.
b) Hacer un análisis de residuales y comentar sus resultados.
c) Aplicar el método "stepwise" para elegir el mejor modelo. Comentar los resultados.
d) Aplicar el método de "Los mejores subconjuntos" para elegir el mejor modelo. Comentar sus resultados.

11. El archivo de datos **rendimiento** contiene la siguiente información:

Y=rendimiento de la enfermera
X1=firmeza de carácter
X2=entusiasmo
X3=ambición
X4=habilidad para comunicarse
X5=habilidad para resolver problemas
X6=iniciativa

a) Hallar la regresión lineal múltiple. Comentar los coeficientes.
b) Probar las hipótesis de que las variables entusiasmo e iniciativa no son importantes para predecir el rendimiento de la enfremera.
c) Hacer una análisis de residuales y comentar sus resultados.
d) Aplicar el método "stepwise" para elegir el mejor modelo. Comentar los resultados.
e) Aplicar el método de "Los mejores subconjuntos" para elegir el mejor modelo. Comentar sus resultados.

12. El archivo de datos **detroit** que aparece en la página de internet del texto contiene la siguiente información acerca de la tasa de homicidio en Detroit entre 1966 y 1973

FTP - Full-time police per 100,000 population
UEMP - % unemployed in the population
LIC - Number of handgun licences per 100,000 population
CLEAR - % homicides cleared by arrests
WM - Number of white males in the population
NMAN - Number of non-manufacturing workers in thousands
GOV - Number of government workers in thousands
HE - Average hourly earnings
HOM - Number of homicides per 100,000 of population

a) Hallar la regresión lineal múltiple considerando Y=HOM. Comentar los coeficientes.
b) Hacer un análisis de residuales y comentar sus resultados.

c) Aplicar el método "stepwise" para elegir el mejor modelo. Comentar los resultados.
d) Aplicar el método de "Los mejores subconjuntos" para elegir el mejor modelo. Comentar sus resultados.

13. Los siguientes datos corresponden al precio de venta (en dólares) de 25 propiedades. Para cada una de ellas se tomó datos acerca del número de cuartos, años de antiguedad, área total de la propiedad (en metros cuadrados) y área patio exterior (en metros cuadrados)

Precio	Cuartos	Antiguedad	Área	Patio
108360	5	41	463	243
460800	20	7	1779	340
189000	5	33	594	379
611440	20	32	1775	395
198000	5	28	520	175
360000	10	32	1250	150
130500	4	41	730	426
331846	11	12	515	160
504000	20	9	1175	750
714000	32	36	1750	1400
672000	26	37	1121	821
321600	13	28	1200	400
348000	9	38	1600	469
207840	6	11	550	100
387600	11	12	1180	280
195000	5	9	530	150
424200	20	31	1500	160
161280	4	35	600	100
224400	8	10	908	158
186840	4	29	650	100
111000	4	41	658	248
132000	4	25	460	80
887000	14	5	11200	8820
96600	4	41	762	372
336600	4	42	910	510

a) Construir diagramas de dispersión entre el precio y el área total, el precio y la antiguedad de la propiedad.
b) Hallar el modelo de Regresión Lineal Múltiple e interpretar los coeficientes de Regresión.
c) Presentan los datos evidencia suficiente para concluir que los coeficientes de regresión son distintos de cero? , use un $\alpha = 0.05$.
d) Hacer un análisis de varianza, e interpretar los resultados.

14. Los siguientes datos corresponden a las mediciones de peso (en libras), estatura (en pulgadas) y edad de 26 personas

Peso (y)	Talla (x1)	Edad (x2)
123	4.7	17
111	4.9	19
130	4.9	19
150	5.1	19
164	5.3	23
151	5	23
147	5.2	26
138	5.1	27
159	5.2	28
160	5.1	28
150	4.8	28
175	5	28
152	4.9	29
156	5.2	30
145	4.8	30
143	5.3	30
171	5.4	30
172	5.2	30
177	5.5	31
202	5.3	36
199	5.5	38
174	5.1	40
186	5.3	44
170	5.2	44
210	5.3	50
199	5.4	55

a) Hallar un modelo de regresión lineal múltiple de la variable peso en función de las variables predictoras; estatura y edad. Interpretar los coeficientes.
b) Hacer un Análisis de Residuales y comentar sus resultados.