Tema 8: Regresión Lineal Simple y Múltiple

1. Introducción

- El análisis de regresión es una técnica estadística que sirve para estudiar la relación existente entre dos o más variables, siendo un caso particularmente sencillo cuando se estudia la relación entre sólo dos variables, que denotaremos por X e Y.
- Si además la relación funcional entre las variables en estudio es de tipo lineal, hablaremos de **regresión lineal simple** (en el caso de dos variables X e Y) y **regresión lineal múltiple** (en el caso de 3 o más variables: Y, X_1, X_2, ..., X_n). Aunque el requisito de "relación lineal" pueda parecer muy restrictivo, veremos más adelante que muchas relaciones de otro tipo se pueden convertir en lineales mediante transformaciones sencillas.

Este tema está dedicado al desarrollo y estudio de un modelo de regresión lineal simple, resultados que serán generalizados con el modelo de regresión múltiple.

2. Planteamiento del modelo

En general, el análisis de regresión requiere el planteamiento de un modelo matemático formal. A continuación mostramos la expresión del modelo de regresión lineal simple mediante un ejemplo.

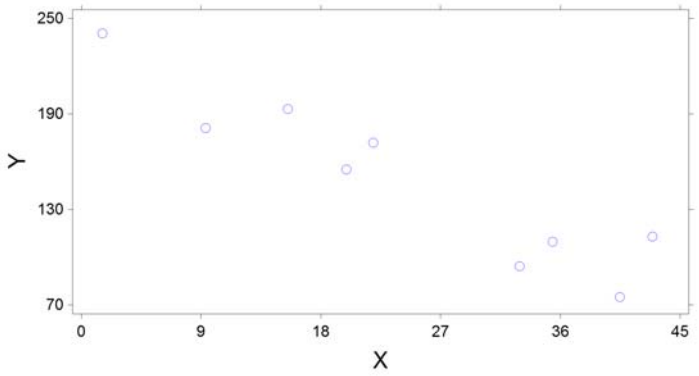
- `Ejemplo` : Consideremos un experimento en el que tomamos simultáneamente medidas sobre dos variables X e Y. Si tomamos una muestra de tamaño n, tendremos n pares de la forma:

  $$(x_1, y_1),\ (x_2, y_2),\ldots,\ (x_n, y_n).$$

  Por ejemplo, considérense los datos de la siguiente tabla (artículo publicado en *Wear*, vol 152, 1992, 171-181). En ella, "y" es el volumen de desgaste del acero dulce y "x" es la viscosidad del aceite:±

x	y $(\times 10^{-4} mm^3)$
1.6	240
9.4	181
15.5	193
20.0	155
22.0	172
35.5	110
43.0	113
40.5	75
33.0	94

  El diagrama de puntos (o diagrama de dispersión) de los n pares, revela información sobre el tipo de relación existente entre ambas variables:



- Supongamos que los valores observados de la variable X son fijos, (x_1, x_2, ..., x_n), denominados niveles del regresor X. Entonces, para cada valor x_i se tienen distintos valores posibles para la variable Y. Por ejemplo, para $x_1 = 1{,}6$ tenemos $y_1 = 240$, pero si repitiéramos el experimento para una viscosidad del aceite de 1,6 nos podría salir un valor distinto de volumen de desgaste del acero, aunque próximo al anterior, debido a variaciones incontrolables en las condiciones de medición.

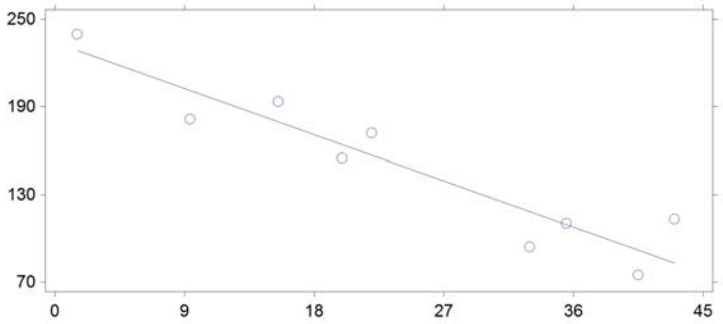
- Por tanto, para cada valor x_i se tiene una variable aleatoria que denotaremos por

$$Y_i = (Y \,|\, x_i)$$

de manera que las observaciones (y_1, y_2, ..., y_n) son una realización de las variables (Y_1, Y_2, ..., Y_n).

- Siguiendo con el ejemplo anterior, el diagrama de puntos refleja que los datos se concentran de forma aleatoria alrededor de una recta. Parece razonable suponer que las medias de las variables Y_i se concentran en torno a una recta (denominada la **recta de regresión**), de ecuación

$$y = \beta_0 + \beta_1 x$$



es decir, las medias de las variables Y_i están relacionadas con x_i por la siguiente relación lineal:

$$E\,(Y_i) = \mu_i = \beta_0 + \beta_1 x_i$$

o en general:

$$E\,(Y \,|\, X = x) = \mu_{Y|x} = \beta_0 + \beta_1 x.$$

- Modelo: En el modelo de regresión lineal simple, todos los factores o causas que influyen en la variable respuesta (Y) pueden descomponerse en dos grupos: el primero contiene la información relativa a la relación lineal existente entre las variables X e Y, y el segundo contiene a todos los factores desconocidos que influyen, de manera pequeña, en la variable respuesta "perturbando" el modelo. En otras palabras, el **modelo de regresión lineal simple** viene dado por:

$$\boxed{Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \quad i = 1, 2, ..., n} \tag{1}$$

o en general:

$$(Y|\, X = x) = \beta_0 + \beta_1 x + \varepsilon$$

donde Y es la *variable respuesta* (variable dependiente), x es el valor prefijado del *regresor* X (variable independiente), ε es una perturbación aleatoria o error que recoge la parte no determinista y β_0, β_1 son los *coeficientes de regresión*.

2.1. Hipótesis del modelo

El modelo de regresión lineal simple (1) requiere las siguientes hipótesis sobre los errores aleatorios (ε_i):

1. *Normalidad*: Los errores aleatorios, ε_i, siguen una distribución Normal.
2. *Homocedasticidad*: Los errores aleatorios, ε_i, tienen varianza constante dada por σ^2.
3. *Independencia*: Los errores aleatorios, ε_i, son todos independientes.

Estas hipótesis pueden resumirse en la siguiente:

$$\varepsilon_i \sim N(0, \sigma) \text{ e independientes, para todo } i = 1, ..., n$$

o equivalentemente:

$$Y_i \sim N(\beta_0 + \beta_1 x_i, \sigma) \text{ e independientes, } \forall i.$$

En general, los parámetros del modelo (β_0, β_1 y σ^2) son desconocidos, de manera que deben estimarse a partir de los datos muestrales.

3. Estimación de los parámetros del modelo

El método utilizado para estimar los coeficientes de regresión, β_0 y β_1, es el denominado *"criterio de mínimos cuadrados"*, es decir, los estimadores serán aquellos que minimizan la suma de las distancias verticales al cuadrado de los puntos de la nube a la recta de ecuación $y = \beta_0 + \beta_1 x$.

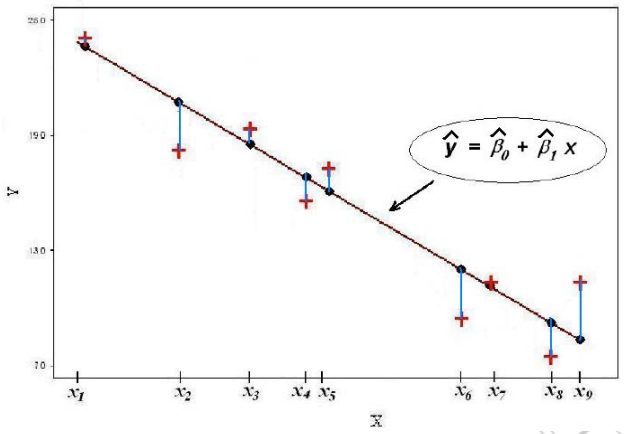
Según este criterio, debemos **minimizar** la siguiente función:

$$Minimizar_{(\beta_0, \beta_1)} \quad D(\beta_0, \beta_1) \tag{2}$$

con

$$D(\beta_0, \beta_1) = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$

que mide la suma de las distancias verticales al cuadrado.



Si llamamos $\widehat{\beta}_0$ y $\widehat{\beta}_1$ a los estimadores de β_0 y β_1, respectivamente, éstos representan la solución al problema (2), de manera que deben satisfacer las siguientes relaciones:

$$\begin{cases} \left.\dfrac{\partial D}{\partial \beta_0}\right|_{\widehat{\beta}_0,\,\widehat{\beta}_1} = -2\sum_{i=1}^{n}(y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_i) = 0 \\ \left.\dfrac{\partial D}{\partial \beta_1}\right|_{\widehat{\beta}_0,\,\widehat{\beta}_1} = -2\sum_{i=1}^{n}(y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_i)x_i = 0 \end{cases}$$

o equivalentemente, las denominadas **Ecuaciones Normales:**

$$\begin{cases} n\widehat{\beta}_0 + \widehat{\beta}_1 \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i \\ \widehat{\beta}_0 \sum_{i=1}^{n} x_i + \widehat{\beta}_1 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} y_i x_i \end{cases}$$

La única solución a las ecuaciones normales anteriores es

$$\boxed{\widehat{\beta}_1 = \frac{\overline{xY} - \overline{x}\,\overline{Y}}{s_x^2} = \frac{s_{xy}}{s_x^2}} \quad , \qquad (3)$$

$$\boxed{\widehat{\beta}_0 = \overline{Y} - \widehat{\beta}_1 \overline{x}}$$

donde $s_x^2 = \dfrac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n}$ representa la varianza de los valores fijados para el regresor X y s_{xy} es la llamada covarianza entre X e Y.

Definición 1 *Llamaremos* ***recta de regresión estimada o ajustada*** *de Y sobre X a la recta de ecuación:*

$$\widehat{y} = \widehat{\beta}_0 + \widehat{\beta}_1 x$$

donde $\widehat{\beta}_0$ y $\widehat{\beta}_1$ representan las estimaciones de los coeficientes de regresión para las observaciones $(x_1, y_1), ..., (x_n, y_n)$, es decir,

$$\boxed{\widehat{\beta}_1 = \frac{s_{xy}}{s_x^2}, \quad \widehat{\beta}_0 = \overline{y} - \widehat{\beta}_1 \overline{x}} \qquad (4)$$

Sustituyendo las expresiones de (4) en la ecuación de la recta ajustada, podemos obtener la siguiente expresión alternativa para dicha recta:

$$\boxed{(\widehat{y} - \overline{y}) = \frac{s_{xy}}{s_x^2}(x - \overline{x})}$$

lo que prueba que la recta de regresión estimada siempre pasa por el punto $(\overline{x}, \overline{y})$.

Definición 2 *A partir de la recta de regresión estimada, para cada par de observaciones* (x_i, y_i) *podemos definir los* ***valores ajustados*** *mediante:*

$$\widehat{y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 x_i$$

Es evidente que los valores observados y_i *se corresponden con los valores ajustados* $\widehat{y}_i$ *más un error:*

$$y_i = \widehat{y}_i + e_i = \widehat{\beta}_0 + \widehat{\beta}_1 x_i + e_i \quad i = 1, ..., n$$

donde $e_i = y_i - \widehat{y}_i$ *recibe el nombre de* ***residuo*** *y describe el error del ajuste del modelo en la* i*-ésima observación.*

4. Distribución de los estimadores

A partir de la expresión de los estimadores de los parámetros de regresión dada en (3):

$$\widehat{\beta}_1 = \frac{s_{xy}}{s_x^2}, \quad \widehat{\beta}_0 = \overline{Y} - \widehat{\beta}_1 \overline{x}$$

y teniendo en cuenta las hipótesis del modelo de regresión lineal simple:

$$Y_i \sim N(\beta_0 + \beta_1 x_i, \sigma) \text{ e independientes, } \forall i$$

se deduce que los estimadores $\widehat{\beta}_1$ y $\widehat{\beta}_0$ siguen distribuciones Normales con los siguientes parámetros:

$$\widehat{\beta}_1 \sim N\left(\beta_1, \sqrt{\frac{\sigma^2}{n s_x^2}}\right)$$

$$\widehat{\beta}_0 \sim N\left(\beta_0, \sqrt{\frac{\sigma^2}{n}(1 + \frac{\overline{x}^2}{s_x^2})}\right)$$

Por tanto, dos formas de disminuir la varianza de estos estimadores consisten en, por una parte, aumentar n (número de pares de observaciones), o bien, aumentar s_x^2 (tomar los valores x_i más dispersos).

Si consideraremos el estimador centrado para la varianza σ^2:

$$\widetilde{\sigma^2} = \frac{n}{n-2}\,\widehat{\sigma^2} = \frac{1}{n-2}\sum_{i=1}^{n}\left(Yi - (\widehat{\beta}_0 + \widehat{\beta}_1 x_i)\right)^2,$$

que verifica:

$$\frac{(n-2)\widetilde{\sigma^2}}{\sigma^2} \sim \chi^2_{n-2}$$

5. Pruebas de hipótesis en la regresión lineal simple

En el contexto de regresión lineal, suelen realizarse contrastes sobre los parámetros de regresión, siendo de particular importancia la denominada prueba de significación de la regresión que mostraremos más adelante.

5.1. Contrastes sobre la ordenada en el origen, β_0

Supongamos que se desea probar la hipótesis de que la ordenada en el origen, β_0, sea igual a un cierto valor prefijado, $\beta_{0,0}$, es decir, queremos llevar a cabo el contraste:

$$\begin{cases} H_0 : \beta_0 = \beta_{0,0} \\ H_1 : \beta_0 \neq \beta_{0,0} \end{cases}.$$

el estadístico del contraste anterior sigue una distribución t de Student con $n-2$ grados de libertad:

$$t_0 = \frac{\widehat{\beta}_0 - \beta_{0,0}}{\sqrt{\frac{\widetilde{\sigma^2}}{n}(1 + \frac{\overline{x}^2}{s_x^2})}} \sim t_{n-2}$$

Por tanto, el criterio de decisión para el contraste planteado es el siguiente:

- Si $|t_0| \leq t_{n-2,1-\alpha/2}$ se acepta H_0
- Si $|t_0| > t_{n-2,1-\alpha/2}$ se rechaza H_0

.

5.1.1. Caso especial, $\beta_0 = 0$

Como caso especial podemos contrastar si la ordenada en el origen se puede considerar nula, es decir, si la recta de regresión pasa por el origen de coordenadas. El contraste en este caso es:

$$\begin{cases} H_0 : \beta_0 = 0 \\ H_1 : \beta_0 \neq 0 \end{cases}$$

y el estadístico del contraste viene dada por:

$$t_0 = \frac{\widehat{\beta}_0}{\sqrt{\frac{\widetilde{\sigma^2}}{n}(1 + \frac{\overline{x}^2}{s_x^2})}} \sim t_{n-2}$$

Comentario 3 *Aunque hemos desarrollado el contraste correspondiente a $\beta_0 = 0$, conviene que la decisión de incluir o no este parámetro se tome a priori en función de las características del problema.*

En el caso de decidir que la constante de la recta de regresión es nula, $\beta_0 = 0$, el modelo de regresión simple quedaría de la siguiente manera:

$$Y_i = \beta_1 x_i + \varepsilon_i \quad i = 1, 2, ..., n.$$

Para el modelo simplificado, el estimador de β_1 viene dado por:

$$\widehat{\beta}_1 = \frac{\overline{xY}}{\overline{x^2}} \sim N\left(\beta_1, \sqrt{\frac{\sigma^2}{n\overline{x^2}}}\right)$$

y el estimador de la varianza común σ^2 es:

$$\tilde{\sigma}^2 = \frac{1}{n-1}\sum (Y_i - \hat{\beta}_1 x_i)^2, \quad \text{con} \quad \frac{(n-1)\,\tilde{\sigma}^2}{\sigma^2} \sim \chi^2_{n-1}$$

5.2. Contrastes sobre la pendiente, β_1

Supongamos que se desea probar la hipótesis de que la pendiente de la recta de regresión, β_1, sea igual a un cierto valor prefijado, $\beta_{1,0}$, es decir:

$$\begin{cases} H_0 : \beta_1 = \beta_{1,0} \\ H_1 : \beta_1 \neq \beta_{1,0} \end{cases}.$$

El estadístico del contraste anterior sigue una distribución t de Student con $n-2$ grados de libertad:

$$t_0 = \frac{\widehat{\beta}_1 - \beta_{1,0}}{\sqrt{\frac{\widetilde{\sigma^2}}{ns_x^2}}} \sim t_{n-2}$$

Por tanto, el criterio para decidir sobre el contraste planteado es:

- Si $|t_0| \leq t_{n-2,1-\alpha/2}$ se acepta H_0
- Si $|t_0| > t_{n-2,1-\alpha/2}$ se rechaza H_0

.

5.2.1. Caso especia, $\beta_1 = 0$

Como caso especial de contraste sobre la pendiente, destaca la *prueba de significación de la regresión*, que indica si realmente existe relación lineal entre las variables X e Y del problema.

El contraste en este caso viene dado por:

$$\begin{cases} H_0 : \beta_1 = 0 \\ H_1 : \beta_1 \neq 0 \end{cases}.$$

Si se tiene que $\beta_1 = 0$, puede significar que la relación existente entre X e Y viene dada por una recta paralela al eje de abcisas, o bien, que la relación entre X e Y no es lineal.

En este caso, el estadístico del contraste viene dado por:

$$t_0 = \frac{\widehat{\beta}_1}{\sqrt{\frac{\widetilde{\sigma^2}}{ns_x^2}}} \sim t_{n-2}.$$

6. Estimación por intervalos en la regresión lineal simple

Una de las principales aplicaciones del análisis de regresión consiste en realizar predicciones para nuevos valores del regresor o regresores. En esta sección, veremos cómo construir intervalos de confianza para los parámetros de regresión β_0 y β_1, para la respuesta promedio de observaciones futuras e intervalos de predicción.

6.1. Intervalos de confianza para β_0 y β_1

Teniendo en cuenta las distribuciones que siguen los estimadores $\widehat{\beta}_0$, $\widehat{\beta}_1$ y $\widetilde{\sigma^2}$, se tiene que:

$$\frac{\widehat{\beta}_0 - \beta_0}{\sqrt{\frac{\widetilde{\sigma^2}}{n}(1 + \frac{\overline{x}^2}{s_x^2})}} \sim t_{n-2}$$

y por otra parte

$$\frac{\widehat{\beta}_1 - \beta_1}{\sqrt{\frac{\widetilde{\sigma^2}}{ns_x^2}}} \sim t_{n-2}$$

A partir de las relaciones anteriores podemos construir intervalos de confianza para los parámetros de regresión β_0 y β_1, del siguiente modo. Fijado un nivel de significación α, un intervalo de confianza para β_0 al $100(1-\alpha)\,\%$ viene dado por:

$$\left(\widehat{\beta}_0 \pm t_{n-2;1-\alpha/2}\sqrt{\frac{\widetilde{\sigma^2}}{n}(1 + \frac{\overline{x}^2}{s_x^2})}\right)$$

y un intervalo de confianza para β_1 al $100(1-\alpha)\,\%$ viene dado por:

$$\left(\widehat{\beta}_1 \pm t_{n-2;1-\alpha/2}\sqrt{\frac{\widetilde{\sigma^2}}{ns_x^2}}\right).$$

6.2. Intervalos de confianza para la respuesta media

Fijado un valor x_0 de X, queremos obtener un intervalo de confianza para la respuesta media en el valor x_0, es decir, un intervalo de confianza para

$$E(Y|\, x_0) = \mu_{Y|x_0} = \beta_0 + \beta_1 x_0.$$

Se trata pues de un intervalo de confianza alrededor de la recta de regresión.

Un estimador puntual de la respuesta media $\mu_{Y|x_0}$ viene dado por:

$$\widehat{\mu}_{Y|x_0} = \widehat{\beta}_0 + \widehat{\beta}_1 x_0$$

este estimador sigue una distribución Normal por ser combinación lineal de distribuciones Normales:

$$\widehat{\mu}_{Y|x_0} \sim N\left(\mu_{Y|x_0}, \sqrt{\sigma^2\left(\frac{1}{n} + \frac{(x_0 - \overline{x})^2}{ns_x^2}\right)}\right).$$

Por otra parte, sabemos que:

$$\frac{\widehat{\mu}_{Y|x_0} - \mu_{Y|x_0}}{\sqrt{\widetilde{\sigma^2}\left(\frac{1}{n} + \frac{(x_0 - \overline{x})^2}{ns_x^2}\right)}} \sim t_{n-2}.$$

Entonces, fijado un nivel de significación α, un intervalo de confianza al $100(1-\alpha)\,\%$ para la respuesta media $\mu_{Y|x_0}$ en el valor x_0 viene dado por:

$$\left(\widehat{\mu}_{Y|x_0} \pm t_{n-2;1-\alpha/2}\sqrt{\widetilde{\sigma^2}\left(\frac{1}{n}+\frac{(x_0-\overline{x})^2}{ns_x^2}\right)}\right).$$

Obsérvese que este intervalo crece a medida que lo hace la distancia de x_0 a $\overline{x}$.

6.3. Intervalos de predicción

Una aplicación importante del modelo de regresión es la predicción de nuevas o futuras observaciones de Y para un valor x_0 de X. Esto es, dado un valor x_0 de X, tenemos asociada una variable aleatoria

$$Y_0 = (Y|\,x_0) = \beta_0 + \beta_1 x_0 + \varepsilon_0 \quad \text{con } \varepsilon_0 \sim N(0,\sigma),$$

que representa la variable respuesta en el punto x_0. Queremos obtener una estimación por intervalo para las futuras observaciones de Y_0.

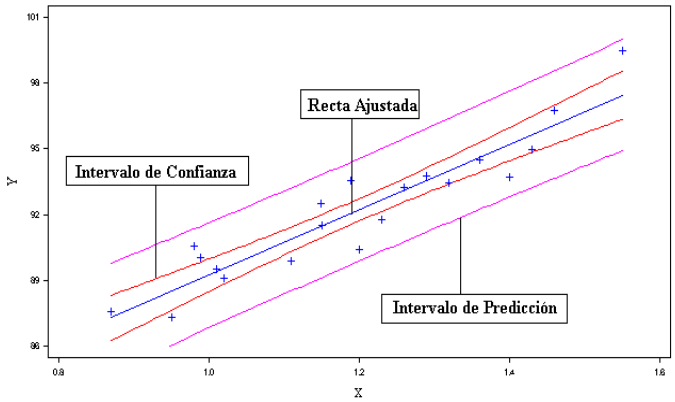
Obsérvese que el intervalo de confianza que se plantea en este apartado es distinto del que se planteó en el apartado anterior, pues el intervalo de confianza para $\mu_{Y|x_0}$ se refiere a la respuesta promedio en x_0 y no a observaciones futuras en dicho punto.

Un intervalo de predicción para una observación futura y_0 en x_0 al $100(1-\alpha)\,\%$ de confianza, viene dado por:

$$\left[\widehat{y}_0 \pm t_{n-2;1-\alpha/2}\sqrt{\widetilde{\sigma^2}\left(1+\frac{1}{n}+\frac{(x_0-\overline{x})^2}{ns_x^2}\right)}\right]$$

donde $\widehat{y}_0$ es el valor que toma el estimador $\widehat{Y}_0$ para la muestra dada.

Comentario 4 *El intervalo de predicción, al igual que el intervalo de confianza para la respuesta media, es mínimo cuando $x_0 = \overline{x}$ y aumenta a medida que crece la distancia de x_0 a $\overline{x}$. Además, el intervalo de predicción en x_0 siempre es más grande que el intervalo de confianza para la respuesta media en x_0, ya que el intervalo de predicción depende tanto del error del modelo ajustado como del error asociado a las observaciones futuras, mientras que el intervalo de confianza sólo depende del error del modelo ajustado.*



Tanto en los intervalos de confianza como en los de predicción, hay que tener especial cuidado a la hora de realizar estimaciones para valores de x_0 fuera de la región que contiene los datos originales. Puede ocurrir que un modelo que ajusta bien dentro de una región no ajuste bien fuera de ella.

7. Validación del modelo: análisis de los residuos

La validez de los resultados vistos en las secciones anteriores depende del cumplimiento de las hipótesis del modelo de regresión lineal simple, vistas en una sección anterior. Aunque las hipótesis se establecen sobre los errores aleatorios, ε_i, del modelo de regresión, usaremos los residuos del modelo ajustado, $e_i = y_i - \widehat{y}_i$, para estudiar la validez de cada hipótesis ya que se corresponden con una realización de las variables ε_i, $i = 1, ..., n$.

Por tanto, el análisis de los residuos tiene como finalidad comprobar que se verifican tales hipótesis, que recordamos por orden creciente de importancia: *Normalidad, Homocedasticidad* e *Independencia*. A estas hipótesis podríamos añadir la de *Linealidad,* entendiendo como tal que la relación existente entre el regresor X y la variable respuesta Y es de tipo lineal.

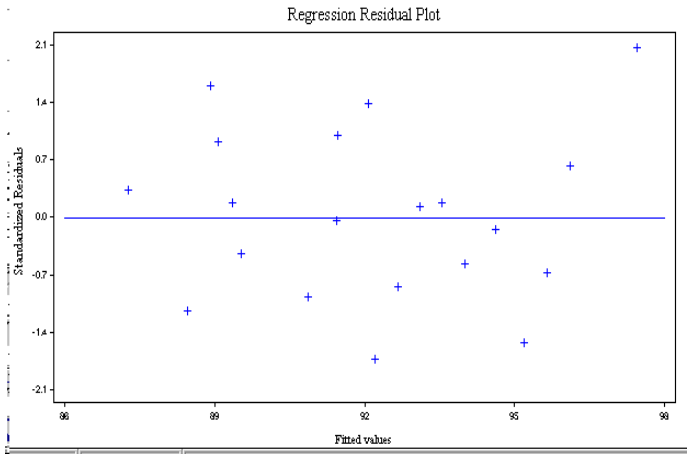
Veamos a continuación algunos métodos gráficos y analíticos que nos permiten comprobar la validez de cada una de las hipótesis anteriores.

7.1. Hipótesis de Normalidad

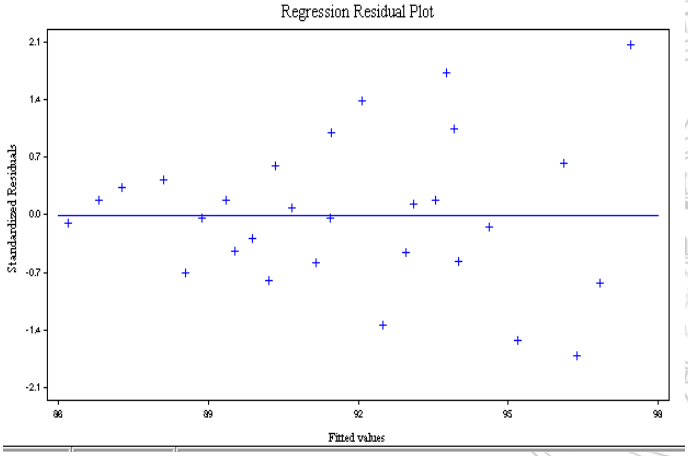
- Como primer método gráfico sencillo para estudiar la Normalidad podemos indicar la representación del histograma: si los datos provienen de una distribución aproximadamente Normal, la forma del histograma se asemeja a la campana de Gauss (función de densidad de una Normal).
- Otros métodos gráficos implementados en la mayoría de software estadístico son los gráficos de cuantiles (o gráficos Q-Q). En el caso de una distribución Normal, estos diagramas deben mostra los puntos de la nube aproximadamente alineados en torno a una recta.
- Por último, mencionaremos la posiblidad de realizar un contraste no paramétrico de Normalidad a través de los test de Kolmogorov-Smirnov o de Shapiro-Wilks, que también suelen estar implementados en cualquier software estadístico.
- En general conviene tener en cuenta varios de estos métodos para decidir sobre la validez de la hipótesis, aunque la robustez de los estimadores del modelo de regresión permite que la hipótesis de Normalidad pueda relajarse.

7.2. Hipótesis de Homocedasticidad

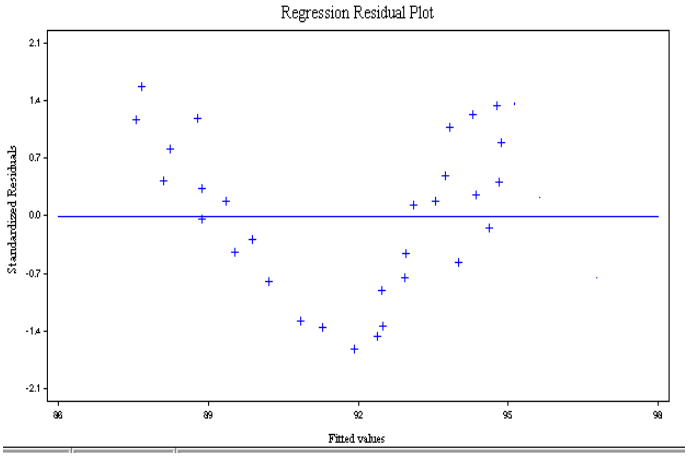
- Con el fin de detectar una desigual varianza entre los residuos, suelen emplearse gráficas de los residuos frente a los valores ajustados.
- Una gráfica como la siguiente representa la situación ideal (varianzas iguales):



- Gráficas como la siguiente detectan una desigual varianza entre las observaciones:

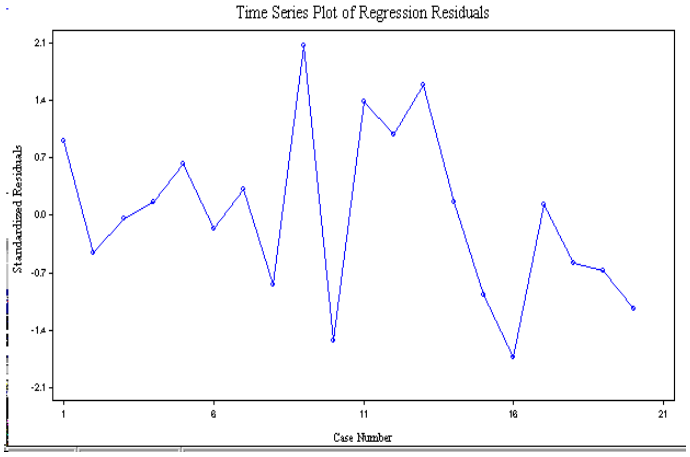


- O como en la siguiente, un modelo inadecuado, siendo necesario en este caso añadir al modelo teórico términos de orden superior.



7.3. Hipótesis de Independencia

Para contrastar la hipótesis de independencia suele usarse un gráfico temporal de los residuos en función del orden de recopilación de datos. En el caso de errores independientes, el gráfico temporal no debe presentar tendencias, rachas o ciclos. El siguiente gráfico representa una situación ideal en el que cabe destacar una gran aleatoriedad entre los residuos.



7.3.1. Test de Durbin Watson

- Para detectar la presencia de autocorrelación en una serie de datos (dependencia entre los datos) la prueba más utilizada es la de Durbin Watson.
- Esta prueba plantea el siguiente contraste:

$$\begin{cases} H_0 : \rho = 0 \\ H_1 : \rho \neq 0 \end{cases}.$$

siendo ρ el coeficiente de correlación entre términos consecutivos de la serie.

- El estadístico del contraste viene dado por:

$$d = \frac{\sum_{i=2}^{n}(e_i - e_{i-1})^2}{\sum_{i=1}^{n} e_i^2}$$

- Los valores críticos de este contraste tanto para el caso de regresión simple como múltiple se encuentran tabulados.
- **Regla práctica:** Debe cumplirse $d \in (1{,}5, 2{,}5)$ para asumir independencia.

8. Bondad del ajuste: el coeficiente de determinación

Una vez estimados los parámetros del modelo, que nos proporcionana la recta de regresión ajustada, y validadas las hipótesis del modelo, surge una pregunta de manera natural: *¿cuán bueno es el modelo ajustado a nuestros datos?*

Una medida numérica de la bondad del ajuste obtenido en un modelo de regresión lineal es el denominado coeficiente de determinación o R^2.

Definición 5 *Se define el coeficiente de determinación, R^2, asociado a un ajuste lineal como:*

$$R^2 = 1 - \frac{SC_{Residual}}{SC_{Total}} = 1 - \frac{\sum_{i=1}^{n}(y_i - \widehat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

expresión que mide la proporción de variabilidad de los datos que no viene explicada por los residuos.

Propiedades del coeficiente de determinación:

1. Está siempre entre cero y uno, $0 \leq R^2 \leq 1$, pues $0 \leq \frac{SC_{Residual}}{SC_{Total}} \leq 1$.
2. Si $R^2 = 1 \Rightarrow SC_{Residual} = 0$, por tanto el modelo se ajusta a los datos observados de manera exacta.

3. Si $R^2 = 0 \Rightarrow SC_{Total} = SC_{Residual}$, por tanto los resultados obtenidos para la variable Y dependen únicamente de los valores residuales y no de los valores obtenidos para la variable X.

Para el modelo de regresión lineal simple **con ordenada en el origen no nula**, $Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$, el coeficiente de determinación adopta las siguientes expresiones:

$$R^2 = \frac{s_{xy}^2}{s_x^2 s_y^2}$$

Obsérvese que en particular se tiene que:

$$R^2 = \frac{SC_{Regresión}}{SC_{Total}},$$

así que mide la proporción de variabilidad de los datos explicada por el modelo de regresión.

El coeficiente de determinación R^2 debe usarse con cuidado, pues siempre es posible conseguir $R^2 = 1$ agregando el suficiente número de regresores al modelo. Por ejemplo, es posible obtener un ajuste perfecto de n puntos ajustando un polinomio de grado n. Sin embargo, esto no significaría que el modelo sea mejor, pues debemos atender al Principio de Parsimonia.

Principio de Parsimonia: *Si dos modelos explican igual de bien un conjunto de datos, debemos optar por aquel que contenga menos parámetros.*

9. Modelos Linealizables

El estudio realizado en las secciones anteriores parte de la suposición de la relación lineal entre las variables. Sin embargo, esta suposición no resulta muy general, aunque existen modelos no lineales que pueden transformarse en lineales sin más que realizar transformaciones oportunas en las variables estudiadas.

Algunos ejemplos son los siguientes:

9.1. Modelo Exponencial

Supongamos que la relación existente entre X e Y viene dada por:

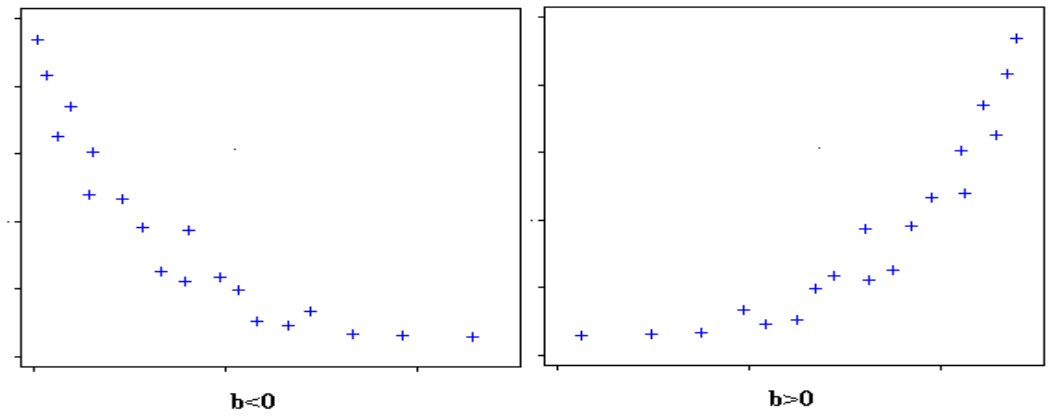
$$Y = ae^{bX}$$

Para transformar en lineal el modelo exponencial basta tomar logaritmos en la relación anterior, obteniéndose:

$$\ln(Y) = \ln(a) + bX$$

lo que se traduce en una relación lineal entre la nueva variable $Y' = \ln(Y)$ y la variable X.

Estos modelos se caracterizan por tener el siguiente gráfico de dispersión:



9.2. Modelo Potencial

Según este modelo, la relación entre X e Y sería de la forma:

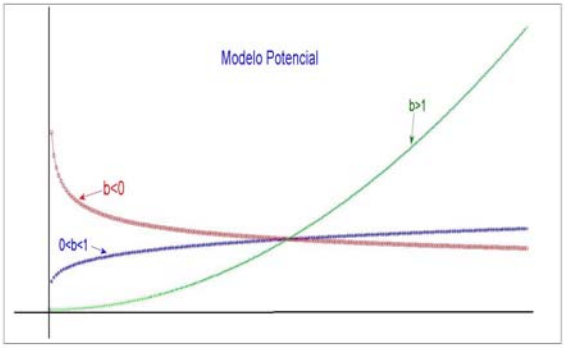
$$Y = aX^b$$

Al igual que el modelo exponencial, para linealizar el modelo potencial es necesario tomar logaritmos en la relación original, obteniéndose:

$$\ln(Y) = \ln(a) + b\ln(X)$$

de manera que tendremos una relación lineal entre las nuevas variables $Y' = \ln(Y)$ y $X' = \ln(X)$.

La gráfica de dispersión dependerá de los valores del parámetro b.



9.3. Modelo Logístico

Igual que en los casos anteriores, intentaremos transformar el modelo:

$$Y = \frac{e^{a+bX}}{1+e^{a+bX}}$$

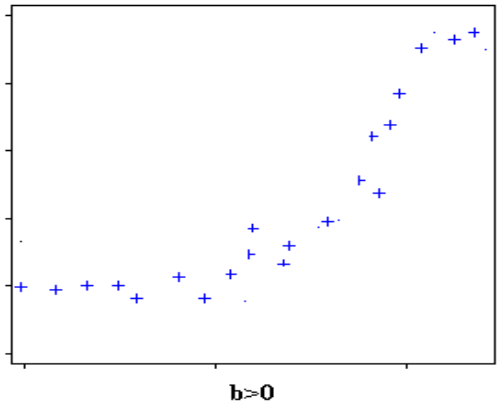
en uno lineal. Para ello hemos de considerar previamente la siguiente relación:

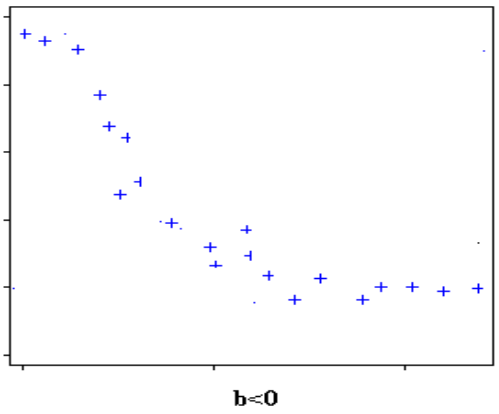
$$\frac{Y}{1-Y} = \frac{\frac{e^{a+bX}}{1+e^{a+bX}}}{1-\frac{e^{a+bX}}{1+e^{a+bX}}} = e^{a+bX}$$

por tanto, tomando logaritmos en la expresión anterior obtenemos la siguiente relación:

$$\ln\left(\frac{Y}{1-Y}\right) = a + bX$$

Este modelo se caracteriza por el siguiente gráfico de dispersión:





10. Regresión Lineal Múltiple

10.1. Introducción

Este capítulo supone una generalización del tema anterior de regresión lineal simple. El estudio del modelo de regresión lineal múltiple permite expresar, mediante una función lineal, el comportamiento de una variable Y respecto a otro conjunto de variables que denotaremos por X_1, X_2, ..., X_n.

Desde el punto de vista de la experimentación y la investigación, existen multitud de situaciones en las que la variable aleatoria objeto de estudio depende de manera lineal de varias variables.

10.2. Planteamiento del modelo

Pretendemos estudiar el comportamiento de una variable Y (**variable respuesta**) para valores dados de otras variables, que denotaremos por $X_1, X_2, ..., X_k$ (**regresores**), los cuales pueden ser variables aleatorias cuyos valores van a ser observados de manera conjunta con los valores de la variable Y o por el contrario pueden ser variables de control cuyos valores van a ser seleccionados por el experimentador.

Supongamos que se fijan n niveles distintos para los regresores:

$$\begin{array}{lcccc} \text{Nivel } 1 \rightsquigarrow & X_1 = x_{1,1} & X_2 = x_{1,2} & \cdots & X_k = x_{1,k} \\ \text{Nivel } 2 \rightsquigarrow & X_1 = x_{2,1} & X_2 = x_{2,2} & \cdots & X_k = x_{2,k} \\ \vdots & \vdots & \vdots & & \vdots \\ \text{Nivel } n \rightsquigarrow & X_1 = x_{n,1} & X_2 = x_{n,2} & \cdots & X_k = x_{n,k} \end{array}$$

Por analogía con el modelo de regresión lineal simple, para cada nivel $(x_{i,1}, x_{i,2}, ..., x_{i,k})$ se tiene una variable aleatoria que denotaremos por:

$$Y_i = (Y \,|\, X_1 = x_{i,1}, X_2 = x_{i,2}, ..., X_k = x_{i,k})$$

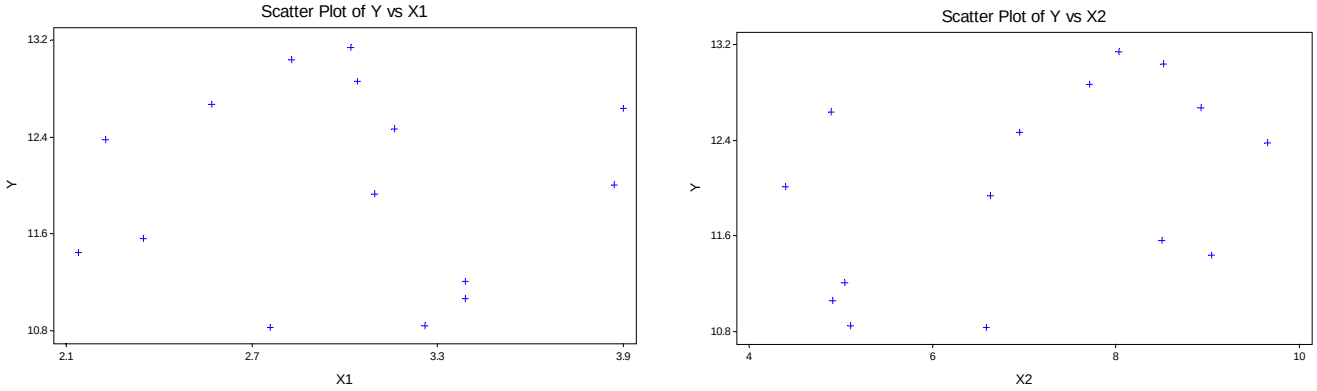
y dispondremos de n observaciones de la siguiente forma:

$$\begin{array}{cccccc} X_1 & X_2 & \cdots & X_k & & Y \\ x_{1,1} & x_{1,2} & \cdots & x_{1,k} & \rightsquigarrow & y_1 \\ x_{2,1} & x_{2,2} & \cdots & x_{2,k} & \rightsquigarrow & y_2 \\ \vdots & \vdots & & \vdots & & \\ x_{n,1} & x_{n,2} & \cdots & x_{n,k} & \rightsquigarrow & y_n \end{array}$$

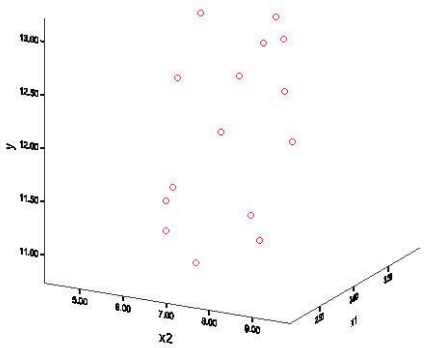
siendo las observaciones $(y_1, y_2, ..., y_n)$ una realización de las variables $(Y_1, Y_2, ..., Y_n)$.

Si disponemos de un único regresor, la nube de puntos o diagrama de dispersión resulta de gran utilidad para visualizar la posible relación entre dicho regresor y la variable respuesta. Sin embargo, cuando se trabaja con más de dos regresores, resulta imposible una representación gráfica del conjunto de datos. A modo de ejemplo, mostraremos la nube de puntos correspondiente a un problema en el que intervienen dos regresores.

Ejemplo 6 ***(Datos de Hamilton):*** *Este ejemplo muestra cómo una variable Y puede depender de dos variables regresoras conjuntamente, pero no de forma individual. En los correspondientes diagramas de dispersión se observa que no existe relación lineal entre Y y X_1, y tampoco entre Y y X_2:*

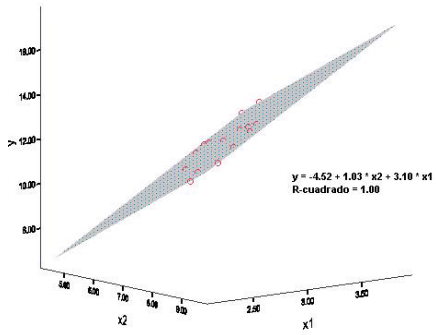


Sin embargo, el diagrama de dispersión tridimensional de Y en función de X_1 y X_2 viene dado por:



que al rotarlo convenientemente, muestra que los puntos de la nube se concentran de forma casi

perfecta alrededor de un plano:



Por tanto, parece razonable suponer que las medias de las variables Y_i están sobre un mismo plano:

$$E(Y_i) = \mu_i = \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} \quad i = 1, 2, ..., n$$

o en general:

$$E(Y | X_1 = x_1, X_2 = x_2) = \mu_{Y|x_1,x_2} = \beta_0 + \beta_1 x_1 + \beta_2 x_2.$$

Por analogía con el modelo de regresión lineal simple, se tiene el concepto de **hiperplano de regresión** como extensión al caso multivariante de la recta de regresión.

Definición 7 *Se llama* ***hiperplano de regresión*** *a la esperanza condicionada de la variable aleatoria Y dados $X_1 = x_1, X_2 = x_2, ..., X_k = x_k$, que viene dada por un hiperplano de ecuación:*

$$E(Y/X_1 = x_1, X_2 = x_2, ..., X_n = x_n) = \beta_0 + \beta_1 x_1 + ... + \beta_k x_k$$

donde los parámetros β_i reciben el nombre de ***coeficientes de regresión parcial.***

Obsérvese que cada coeficiente β_i mide el cambio esperado en la variable respuesta (Y) por unidad de cambio de X_i, cuando los restantes regresores permanecen constantes ($X_1 = x_1$, $X_2 = x_2$,..., $X_{i-1} = x_{i-1}$, $X_{i+1} = x_{i+1}$,...,$X_k = x_k$).

En la práctica, todas las observaciones no se encuentran sobre el hiperplano de regresión, aunque sí presentan una cierta tendencia. Por tanto, el modelo debe incluir un término que englobe el error debido a factores desconocidos por el experimentador y que influyen, de manera pequeña, en la variable respuesta:

$$Y = \beta_0 + \beta_1 X_1 + ... + \beta_n X_k + \varepsilon$$

10.3. Modelo e Hipótesis

Desde un punto de vista teórico, un problema de **Regresión Lineal Múltiple** con k regresores y n observaciones, puede modelizarse de la siguiente manera:

$$Y_i = \beta_0 + \beta_1 x_{i,1} + ... + \beta_k x_{i,k} + \varepsilon_i \quad i = 1, 2, ..., n$$

donde los errores o perturbaciones aleatorias ε_i verifican las siguientes **hipótesis**:

1. *Normalidad*: Todos los errores ε_i siguen una distribución Normal de media cero, $i = 1, ..., n$.

2. *Homocedasticidad*: Todos los errores ε_i tienen la misma varianza, es decir, $Var(\varepsilon_i) = \sigma^2$ para todo $i = 1, ..., n$.

3. *Independencia*: Todos los errores ε_i son independientes.

Además de estas hipótesis, será necesaria una hipótesis adicional sobre los valores fijados para los distintos regresores, la cual introduciremos en una sección posterior cuando surja de manera natural.

10.4. Estimación de los parámetros

En general, los parámetros del modelo (β_0, β_1, ..., β_k y σ^2) son desconocidos, de manera que deben estimarse a partir de datos muestrales. Una vez que se tienen estimaciones de los parámetros, el modelo de regresión lineal múltiple suele utilizarse para predecir observaciones futuras de la variable respuesta Y o bien para estimar la respuesta promedio para un nivel particular de los regresores.

Al igual que en el modelo de regresión lineal simple, comenzaremos calculando los estimadores de los coeficientes de regresión β_0, β_1, ..., β_k, por medio del criterio de mínimos cuadrados, cuya función viene dada por:

$$D(\beta_0, \beta_1, ..., \beta_k) = \sum_{i=1}^{n} (y_i - \beta_0 - \sum_{j=1}^{k} \beta_j x_{i,j})^2.$$

En efecto, teniendo en cuenta las hipótesis del modelo de regresión lineal múltiple, se tiene una distribución Normal para las variables Y_i :

$$Y_i \sim N(\beta_0 + \sum_{j=1}^{k} \beta_j x_{i,j}, \sigma^2) \text{ e independientes, } \forall i = 1, ..., n,$$

Con el fin de estimar dichos parámetros, debemos resolver las $k + 1$ ecuaciones lineales siguientes:

$$\begin{aligned} \frac{\partial D}{\partial \beta_0}\Big|_{\hat{\beta}_0,...,\hat{\beta}_k} &= -2\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \sum_{j=1}^{k}\hat{\beta}_j x_{i,j}) = 0 \\ \frac{\partial D}{\partial \beta_j}\Big|_{\hat{\beta}_0,...,\hat{\beta}_k} &= -2\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \sum_{j=1}^{k}\hat{\beta}_j x_{i,j})x_{i,j} = 0 \quad \text{para } j = 1, ..k, \end{aligned}$$

obteniéndose lo que se conoce como **Ecuaciones Normales** asociadas al modelo de mínimos cuadrados:

$$n\hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^{n} x_{i,1} + \hat{\beta}_2 \sum_{i=1}^{n} x_{i,2} + ... + \hat{\beta}_k \sum_{i=1}^{n} x_{i,k} = \sum_{i=1}^{n} y_i$$

$$\hat{\beta}_0 \sum_{i=1}^{n} x_{i,1} + \hat{\beta}_1 \sum_{i=1}^{n} x_{i,1}^2 + \hat{\beta}_2 \sum_{i=1}^{n} x_{i,1}x_{i,2} + ... + \hat{\beta}_k \sum_{i=1}^{n} x_{i,1}x_{i,k} = \sum_{i=1}^{n} y_i x_{i,1}$$

$$\vdots$$

$$\hat{\beta}_0 \sum_{i=1}^{n} x_{i,k} + \hat{\beta}_1 \sum_{i=1}^{n} x_{i,k}x_{i,1} + \hat{\beta}_2 \sum_{i=1}^{n} x_{i,k}x_{i,2} + ... + \hat{\beta}_k \sum_{i=1}^{n} x_{i,k}^2 = \sum_{i=1}^{n} y_i x_{ik}$$

sistema lineal formado por $k+1$ ecuaciones y $k+1$ incógnitas.

Una vez que obtengamos los estimadores de los parámetros de regresión, denotaremos por:

$$\widehat{y}_i = \widehat{\beta}_0 + \sum_{j=1}^{k} \widehat{\beta}_j x_{i,j}$$

a los **valores ajustados** y por:

$$e_i = y_i - \widehat{y}_i$$

a los **residuos** asociados al modelo.

10.5. Enfoque matricial de la regresión lineal múltiple

Con el fin de obtener la expresión de los estimadores de los coeficientes de regresión así como su distribución, esperanza y varianza, resulta aconsejable trabajar con notación matricial.

Si denotamos por:

$$\mathbf{Y} = \begin{bmatrix} Y_1 \\ Y_2 \\ . \\ . \\ Y_n \end{bmatrix} \quad \mathbf{X} = \begin{bmatrix} 1 & x_{1,1} & x_{1,2} & ... & x_{1,k} \\ 1 & x_{2,1} & x_{2,2} & ... & x_{2,k} \\ . & . & . & & . \\ . & . & . & & . \\ 1 & x_{n,1} & x_{n,2} & ... & x_{n,k} \end{bmatrix} = (\mathbf{1}, \mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_k)$$

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ . \\ . \\ \beta_k \end{bmatrix} \quad \boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ . \\ . \\ \varepsilon_n \end{bmatrix}$$

el **modelo de regresión lineal múltiple en forma matricial** queda de la siguiente forma:

$$\boxed{\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \quad \text{con } \boldsymbol{\varepsilon} \sim \mathbf{N}(\mathbf{0}, \sigma^2\mathbf{I}_n) \Leftrightarrow \mathbf{Y} \sim \mathbf{N}(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}_n).}$$

Además de simplificar los cálculos, la notación matricial nos será de gran utilidad para proporcionar una interpretación geométrica del procedimiento, facilitando a su vez la determinación de las distribuciones muestrales.

El estimador del vector de parámetros $\boldsymbol{\beta}$ viene dado por:

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$$

de manera que la estimación puntual de los parámetros de regresión para los datos muestrales es:

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$

Para el caso de σ^2, su estimador vendrá dado por:

$$\widetilde{s^2} = \frac{1}{n-(k+1)} \sum_{i=1}^{n} (y_i - \widehat{y}_i)^2 ,$$

siendo:

$$\widehat{y}_i = \widehat{\beta}_0 + \sum_{j=1}^{k} \widehat{\beta}_j x_{i,j}$$

Nota.- De manera análoga al caso de regresión lineal simple se pueden realizar contrastes e intervalos de confianza sobre cada uno de los parámetros del modelo así como para el resultado promedio del la respuesta e incluso observaciones individuales. Este tipo de pruebas las realizaremos únicamente utilizando el software de las prácticas.

10.6. Construcción y selección del modelo

10.6.1. Métodos para la selección del modelo

Un problema importante en la regresión lineal múltiple es la elección del conjunto de regresores o predictores. Por una parte, es deseable que el modelo final contenga un número suficiente de regresores de forma que sea válido su uso, por ejemplo para la predicción de nuevas observaciones. Sin embargo, también sería conveniente usar el menor número de variables para que el modelo sea más manejable (principio de parsimonia).

Veremos varios métodos para la selección de variables, los cuales proporcionan "buenas" soluciones en los dos sentidos anteriores.

10.6.2. Regresión por el mejor conjunto de regresores

Este método resulta bastante intuitivo y adecuado cuando el número de regresores candidatos es pequeño. El método consiste en calcular todas las regresiones posibles tomando un regresor, dos regresores,..., todos los regresores. A continuación, se selecciona el "mejor" modelo de regresión atendiendo a algún criterio que mida la adecuación del modelo.

El principal inconveniente de este método es que si el número de regresores candidatos es k, el número de regresiones posibles puede ser demasiado elevado (2^k).

Usando el coeficiente de determinación múltiple R^2 Si denotamos por R_p^2 al coeficiente de determinación de un modelo de regresión con p coeficientes de regresión, sabemos que R_p^2 aumenta a medida que p crece. La selección del modelo de regresión consiste en ir añadiendo regresores al modelo, calculando en cada caso el valor del coeficiente de determinación, hasta que el incremento de R^2 debido a la incorporación de un nuevo regresor sea pequeño. La forma de estudiar el crecimiento del coeficiente de determinación se suele hacer por medio de una gráfica R_p^2 contra p.

Otra opción consiste en utilizar el coeficiente de determinación ajustado $R^2_{ajustado}$, que se relaciona con el coeficiente de determinación mediante:

$$R^2_{corregido} = R_p^2 - \frac{p(1-R_p^2)}{n-p-1}$$

donde p representa el número de variables regresoras y R_p^2 al coeficiente de determinación de un modelo de regresión con p coeficientes de regresión.

Si usamos el coeficiente de determinación ajustado o corregido, seleccionaremos como mejor modelo aquel que proporcione un mayor valor de dicho coeficiente.

Usando el error cuadrático medio CM_{Res} Si denotamos por:

$$CM_{Res}(p) = \frac{SC_{Res}(p)}{n-p}$$

al error cuadrático medio para un modelo con p coeficientes de regresión, generalmente $CM_{Res}(p)$ disminuye si p crece, aunque no siempre es cierto, pues al añadir un nuevo regresor, el coeficiente $CM_{Res}(p)$ pierde un grado de libertad de manera que puede aumentar su valor en lugar de disminuir.

El criterio de selección del modelo consiste en determinar aquél que tenga menor error cuadrático medio, criterio que coincide con la selección del modelo de mayor coeficiente de determinación ajustado.

10.6.3. Regresión por pasos

Esta técnica se basa en introducir de forma progresiva variables en el modelo, de forma que al introducir una nueva variable en el modelo la influencia de las ya presentes es reevaluada mediante un contraste, pudiéndose rechazar alguna de las variables ya incluidas. Para ello, hay que definir un criterio de entrada y de salida para los regresores.

Nosotros utilizaremos siempre como criterio de entrada/salida de variables el criterio AIC (Akaike's information criterion).

- Criterio AIC: Recordemos que este criterio evalúa para cada modelo la siguiente expresión:

$$AIC(p) = nlog[s_e^2] + 2p$$

  siendo n el número de observaciones, s_e^2 la estimación de la varianza asociada a los residuos ($e_i = y_i - \widehat{y}_i$) y p el número de parámetros estimados (p-1 regresores + término independiente). Observar que este método penaliza los modelos con más regresores. Según este criterio, un modelo será mejor que otro si su AIC es menor.

Dirección:

- Atrás/adelante: Se parte de un modelo con todos los regresores y se elimina aquél que produce una menor disminución del AIC. Seguidamente se prueba si dicho criterio mejora o no al introducir alguno de los regresores eliminados en etapas anteriores.
- Adelante/atrás: Se parte de un modelo sin regresores y se introduce aquél que produce una mayor disminución del AIC. Seguidamente se prueba si dicho criterio mejora o no al sacar alguno de los regresores introducidos en etapas anteriores.
- Atrás: Se parte de un modelo con todos los regresores y se elimina aquél que produce una menor disminución del AIC.
- Adelante: Se parte de un modelo sin regresores y se introduce aquél que produce una mayor disminución del AIC.