Tema 9
Regresión múltiple

Carlos Montes – uc3m

1. Introducción

Modelo de regresión simple

$$y_i = a + bx_i + e_i$$

El valor que resulta de aplicar la recta a+bx al valor $x=x_i$ es la predicción:

$$\hat{y}(x_i) \quad o \quad \hat{y}$$

1. Introducción

La recta que predice el valor de y cuando $x=x_i$ puede expresarse como:

$$\hat{y}_i = a + bx$$

Luego el residuo puede expresarse como:

$$e_i = y_i - \hat{y}_i = y_i - (a + bx_i)$$

1. Introducción

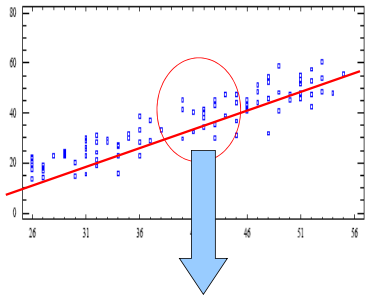
Los valores de la variable Y pueden dividirse en dos partes:

- Parte lineal o determinista (explicada por la variable X)
- Parte aleatoria (parte de Y no explicada linealmente por X)

$$Y = a + bx + e$$

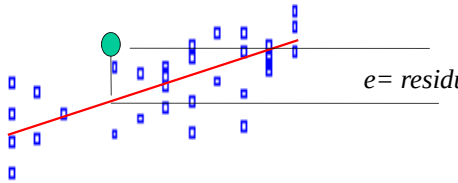
Cuando en el modelo de regresión simple asumimos que e sigue una normal, le denominamos *modelo lineal general* (de regresión simple)

1. Introducción



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

real *recta*



1. Introducción

El coeficiente R^2 (coeficiente de determinación) indica la proporción de Y que es explicada por X.

Es el coeficiente de correlación al cuadrado.

Analysis of Variance

Source	Sum of Squares	Df	Mean Square	F-Ratio	P-Value
Model	528,476	1	528,476	20,33	0,0020
Residual	207,925	8	25,9906		
Total (Corr.)	736,4	9			

Correlation Coefficient = -0,84724
R-squared = 71,7647 percent
Standard Error of Est. = 5,0991

R^2 = 71,76%

1. Introducción

$$R^2 = 1 - \frac{\sum_{i=1}^{n} e_i^2}{\sum_{i=1}^{n} (y_i - y)^2}$$

Analysis of Variance

Source	Sum of Squares	Df	Mean Square	F-Ratio	P-Value
Model	528,476	1	528,476	20,33	0,0020
Residual	207,925	8	25,9906		
Total (Corr.)	736,4	9			

Correlation Coefficient = -0,84724
R-squared = 71,7647 percent
Standard Error of Est. = 5,0991

R^2 = 71,76%

2. Modelo lineal general

En un modelo de regresión múltiple, queremos conocer el valor de una variable respuesta
a partir de más de una variable explicativa.

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + ... + \beta_k x_{ki} + e_i$$

Fijo *Variable*

$x_1, x_2 ... x_n$ son las variables independientes o explicativas, que pueden ser cualitativas o cuantitativas.

$$E(e_i) = 0$$

También se le denomina "recta de regresión" aunque no sea una recta, sino un hiperplano.

2. Modelo lineal general

Observado x_i, el valor esperado de y_i es:

$$E(y_i \setminus x_i) = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + ... + \beta_k x_{ki} + E(e_i \setminus x_i)$$

$$\downarrow$$

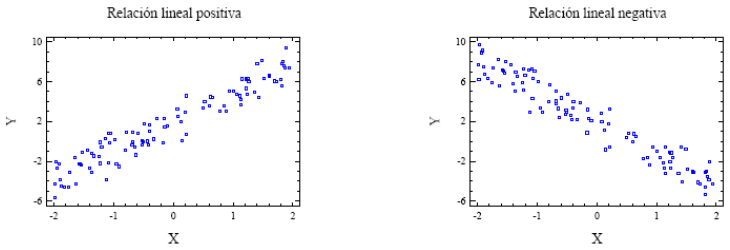
0

2. Modelo lineal general

El modelo se basa en una serie de hipótesis:

1) Linealidad

Las variables explicativas X influyen en Y según una combinación lineal.



2. Modelo lineal general

2) E(e)=0

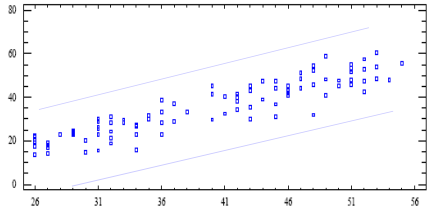
3) cov(X,e) =0

El término de error representa el resto de variables no contenidas en $X=(X_1 ..., X_n)$, luego contiene información independiente de X.

4) Homocedasticidad

La varianza de los errores es constante,
y no depende del nivel de las variables.

La nube de puntos de los datos tiene una anchura más o menos constante a lo largo de la recta de regresión.



$$\text{var}(e) = E\left[\left(e - E(e)^2\right)\right] = E(e^2{}_i) = \sigma^2 \ \ \forall i$$

$$\text{var}(y_i \setminus X = x_i) = \sigma^2$$

5) Normalidad

Los errores se distribuyen
según una distribución normal.

$$e_i \rightarrow N(0, \sigma^2)$$

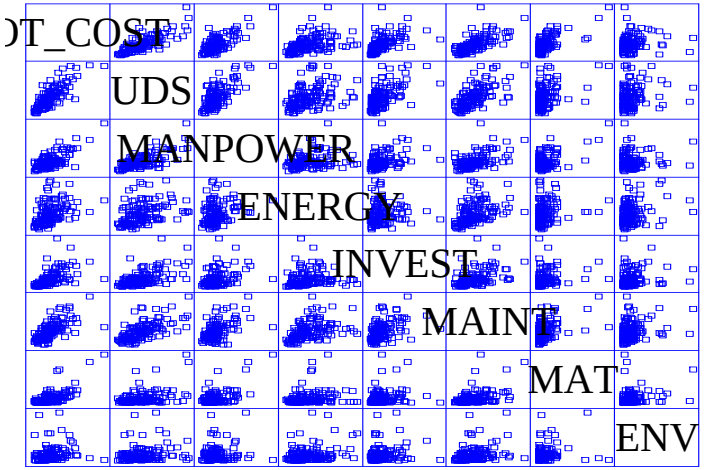
Es una hipótesis razonable por el Teorema Central del Límite: la suma de muchas causas (variables) pequeñas tiende a distribuirse normalmente.

6) Independencia

La secuencia de valores de e_i es independiente.

2. Modelo lineal general

Cada celda del grafico matricial representa la relación bilateral entre dos variables.



3.1. Estimación del modelo. Definición.

1) Estimación de los coeficientes β

Para estimar los coeficientes β empleamos el método de los mínimos cuadrados ordinarios (MCO).

Dada una muestra de n datos:

$$(y_1, x_{11}..., x_{k1})$$
$$(y_2, x_{12}..., x_{k2})$$
$$...$$
$$(y_n, x_{1n}..., x_{kn)}$$

3.1. Estimación del modelo. Definición.

Los valores que asignamos a β_1…, β_k son aquellos que minimizan los errores al cuadrado.

Buscamos el mínimo de la función:

$$S(\beta) = \sum_{i=1}^{n} \left[y_i - \left(\hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \hat{\beta}_2 x_{2i} + ... + \beta_k x_{ki} \right) \right]^2 = \sum_{i=1}^{n} e^2{}_i$$

Puede demostrarse que el vector de estimadores de β que minimiza $S(\beta)$ es:

$$\hat{\beta} = \left(X^T X \right)^{-1} X^T Y$$

Modelo de regresión de altura (Y) sobre peso (X_1)

$$\hat{Y} = 138{,}4 + 0{,}53 X_1 + e$$

La estatura de un individuo que pese 80 kg es una variable aleatoria normal de media estimada:

138.4 + 0.53·80=180.4 cm

Los individuos que pesan 1 kg más tienden a medir

0.53 cm más

Modelo de regresión de altura (Y) sobre peso (X_1) y talla de zapato (X_2)

$$\hat{Y} = 77{,}7 + 0{,}13X_1 + 2{,}16X_2 + e$$

La estatura media de un individuo de 80 kg que calza un 37 es:

77.7+0.13·80+2.16·37=168.02 cm

Si calza un 43:

77.7+0.13·80+2.16·43=181.98 cm

2) Estimación de σ²

$$e_i = N(0, \sigma^2)$$

$$e_i = y_i - \left(\hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \ldots + \hat{\beta}_k x_{ki}\right)$$

Como la media de e es 0, la varianza muestral de los residuos será:

$$\frac{\sum_{i=1}^{n} e^2{}_i}{n}$$

Pero para que sea insesgado, debemos usar la varianza residual:

$$\hat{S}_R{}^2 = \frac{\sum_{1}^{n} e_i{}^2}{n - p}$$

Donde p es el número de parámetros beta.

*Analizamos datos de una muestra de n=82 automóviles. Entre las variables se encuentra **velmax**, que es la velocidad máxima (km/h) que puede alcanzar el vehículo. Queremos construir un modelo lineal que prediga esa velocidad máxima a partir de la variable **Potencia** (caballos de vapor -cv-) y la variable **Peso** (kg) del vehículo.*

El modelo de regresión tal y como lo muestra Statgraphics es:

Multiple Regression Analysis

Dependent variable: velmax

Parameter	Estimate	Standard Error	T Statistic	P-Value
CONSTANT	155,465	1,3399	116,027	0,0000
Potencia	0,519647	0,00966429	53,7698	0,0000
Peso	-0,0252839	0,00148786	-16,9935	0,0000

Analysis of Variance

Source	Sum of Squares	Df	Mean Square	F-Ratio	P-Value
Model	40555,0	2	20277,5	2698,00	0,0000
Residual	593,746	79	7,51577		
Total (Corr.)	41148,8	81			

R-squared = 98,5571 percent
R-squared (adjusted for d.f.) = 98,5205 percent
Standard Error of Est. = 2,74149
Mean absolute error = 1,99442
Durbin-Watson statistic = 1,18907

velmax=155.5 + 0,52·Potencia − 0.025·Peso + e

velmax=155.5 + 0,52·Potencia − 0.025·Peso + e

Para un mismo peso del vehículo, cada caballo de potencia adicional permite aumentar la velocidad máxima

0.52 km/h por término medio.

Para una misma potencia, cada kg adicional disminuye la velocidad máxima en 0.025 km/h.

velmax=155.5 + 0,52·Potencia − 0.025·Peso + e

La velocidad máxima de un coche que pesa 1500 kg y tenga 100 CV de potencia es una variable aleatoria de media

155.5 + 0,52·100 − 0.025·1500= 170 km/h

Su varianza es la del término de error

Parameter	Estimate	Standard Error	T Statistic	P-Value
CONSTANT	155,465	1,3399	116,027	0,0000
Potencia	0,519647	0,00966429	53,7698	0,0000
Peso	-0,0252839	0,00148786	-16,9935	0,0000

Analysis of Variance

Source	Sum of Squares	Df	Mean Square	F-Ratio	P-Value
Model	40555,0	2	20277,5	2698,00	0,0000
Residual	593,746	79	7,51577		
Total (Corr.)	41148,8	81			

R-squared = 98,5571 percent
R-squared (adjusted for d.f.) = 98,5205 percent
Standard Error of Est. = 2,74149
Mean absolute error = 1,99442
Durbin-Watson statistic = 1,18907

En la tabla: 7.51577

$$S_R^2 = \frac{593.746}{79} = 7.51577$$

$$S_R^2 = 2.74149^2 = 7.51577$$

3.2. Coeficiente de determinación corregido

$$R^2 = 1 - \frac{\sum_1^n e_i^{\,2}}{\sum_1^n (y_i - \bar{y})^2} = corr\,(\hat{y}, y)^2$$

A medida que aumenta el número de variables, el coeficiente puede aumentar aunque las variables no sean significativas.

Por ello se define el coeficiente de determinación corregido o ajustado:

3.2. Coeficiente de determinación corregido

$$\overline{R}^2 = 1 - \frac{\hat{S}^2_R}{\hat{S}^2_y} \longrightarrow \textit{cuasivarianza de Y}$$

Aunque no es exactamente el % de la variabilidad de Y explicada por las variables independientes, se interpreta de manera informal de la misma manera.

El objetivo es conseguir el máximo número de variables explicativas que consigan aumentar el coeficiente.


```
Multiple Regression Analysis
-------------------------------------------------------------------------------
Dependent variable: velmax
-------------------------------------------------------------------------------
                                       Standard          T
Parameter              Estimate          Error       Statistic        P-Value
-------------------------------------------------------------------------------
CONSTANT                155,465         1,3399         116,027         0,0000
Potencia                0,519647     0,00966429         53,7698         0,0000
Peso                  -0,0252839     0,00148786        -16,9935         0,0000
-------------------------------------------------------------------------------

                         Analysis of Variance
-------------------------------------------------------------------------------
Source             Sum of Squares     Df  Mean Square     F-Ratio       P-Value
-------------------------------------------------------------------------------
Model                     40555,0      2       20277,5     2698,00         0,0000
Residual                  593,746     79       7,51577
-------------------------------------------------------------------------------
Total (Corr.)             41148,8     81

R-squared = 98,5571 percent
R-squared (adjusted for d.f.) = 98,5205 percent
Standard Error of Est. = 2,74149
Mean absolute error = 1,99442
Durbin-Watson statistic = 1,18907
```

4. Contraste sobre los parámetros

Los coeficientes: $\beta = (\beta_0, \beta_1 ..., \beta_k)^T$

Son parámetros poblacionales, y por tanto desconocidos.

Para su estimación usamos: $\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1 ..., \hat{\beta}_k)^T$

4. Contraste sobre los parámetros

Basándonos en la distribución de β_i podemos hacer el contraste de si una variable es o no "significativa".

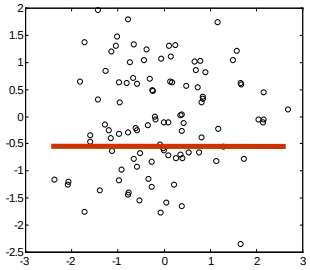
Variable significativa es aquella que aporta información sobre Y no incluida en el resto de las variables.

Por tanto, será relevante incluirla en la regresión.

4. Contraste sobre los parámetros

Una variable será no significativa si:

$\beta_i=0$



Carlos Montes – uc3m

4. Contraste sobre los parámetros

Hacemos el contraste:

H_0: $\beta_i=0$

H_1: $\beta_i\neq0$

Ya que el estadístico: $t=\dfrac{\hat{\beta}_i}{s_R\sqrt{1/(n-1)s^2_{x_i}}}$

Sigue una distribución Z en muestras grandes y una t_{n-p} en poblaciones normales.

4. Contraste sobre los parámetros

Multiple Regression Analysis

Dependent variable: velmax

Parameter	Estimate	Standard Error	T Statistic	P-Value
CONSTANT	155,465	1,3399	116,027	0,0000
Potencia	0,519647	0,00966429	53,7698	0,0000
Peso	-0,0252839	0,00148786	-16,9935	0,0000

Analysis of Variance

Source	Sum of Squares	Df	Mean Square	F-Ratio	P-Value
Model	40555,0	2	20277,5	2698,00	0,0000
Residual	593,746	79	7,51577		
Total (Corr.)	41148,8	81			

R-squared = 98,5571 percent
R-squared (adjusted for d.f.) = 98,5205 percent
Standard Error of Est. = 2,74149
Mean absolute error = 1,99442
Durbin-Watson statistic = 1,18907

p < 0.05

Rechazamos Ho. Las dos variables son significativas.

4. Contraste sobre los parámetros

Multiple Regression Analysis

Dependent variable: velmax

Parameter	Estimate	Standard Error	T Statistic	P-Value
CONSTANT	155,465	1,3399	116,027	0,0000
Potencia	0,519647	0,00966429	53,7698	0,0000
Peso	-0,0252839	0,00148786	-16,9935	0,0000

Analysis of Variance

Source	Sum of Squares	Df	Mean Square	F-Ratio	P-Value
Model	40555,0	2	20277,5	2698,00	0,0000
Residual	593,746	79	7,51577		
Total (Corr.)	41148,8	81			

R-squared = 98,5571 percent
R-squared (adjusted for d.f.) = 98,5205 percent
Standard Error of Est. = 2,74149
Mean absolute error = 1,99442
Durbin-Watson statistic = 1,18907

/t/>2

Rechazamos Ho. Las dos variables son significativas.

5. Diagnosis del modelo

Es la comprobación de las hipótesis del modelo, lo que garantiza que éste va a describir la verdadera relación entre variables:

- Linealidad.
  - Gráfico de residuos frente a valores predichos.

  Si el modelo está bien ajustado, no debe presentar ninguna estructura.
- Homocedasticidad.
  - Gráfico de residuos frente a valores ajustados.
  - Gráfico de residuos frente a X.
- Independencia. Existen contrastes específicos como el de Durbin-Watson.
- Normalidad.
  - Gráfico probabilístico normal de los residuos.

5. Diagnosis del modelo

A la hora de analizar la normalidad:

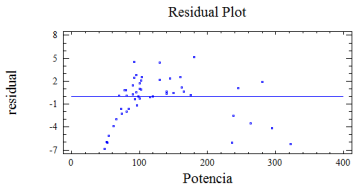
• Puede ser suficiente con comprobar que los residuos sean unimodales y simétricos.

• Si queremos calcular probabilidades con el modelo normal, necesitamos asegurar la normalidad de los residuos mediante el test de la chi cuadrado.

6.1. Transformaciones. Generalidades.

Cuando las hipótesis del modelo no se cumplen es necesario transformar los datos, de manera que los datos transformados cumplan las hipótesis.

Hay que detectar las variables que no tienen un comportamiento lineal.



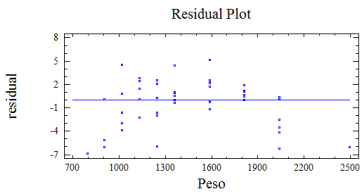


Gráfico de residuos frente a X

6.1. Transformaciones. Generalidades.

Como en la regresión simple, las más utilizadas son:

- Logaritmo

  $y' = \ln x \quad x' = \ln y$

  Frecuente para evitar la falta de linealidad o heterocedasticidad
- Cuadrado

  $y' = y^2 \quad x' = x^2$
- Inversa

  $y' = 1/y \quad x' = 1/x$
- Raíz cuadrada

  $y' = \sqrt{y} \quad x' = \sqrt{x}$

  Muy útil cuando los datos proceden de una Poisson.

6.2. Transformaciones. Gráfico de componentes

El gráfico de componentes *(component effect plot)* representa:

X_i frente a: $e_i + \hat{\beta}_i(X_i - \overline{X}_i)$

Puede interpretarse como la variable Y a la que se la ha eliminado la influencia de todas las variables, salvo la $X_{i,}$

6.2. Transformaciones. Gráfico de componentes

$$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + e$$

Influencia de X_2

$$y - \beta_1 X_1 = \beta_2 X_2 + e$$

Eliminamos la influencia de X_1

y sin influencia de ninguna variable

Los programas informáticos permiten, además, la realización de otro tipo de regresiones.

7. Regresión con variables binarias

* Variable binaria o dicotómica es aquella que toma solo 2 valores (0 y 1).
* Se utilizan para describir la ausencia o presencia de una propiedad.

7. Regresión con variables binarias

Al estudiar la altura, podemos plantearnos en qué medida es explicada por el sexo.

Podemos separar el modelo en dos, uno para cada valor de la variable binaria.

$$E(altura \mid chica) = 165.313 + 14.0367 \times 0 = 165.313$$
$$E(altura \mid chico) = 165.313 + 14.0367 \times 1 = 179.3497$$

Multiple Regression Analysis

Dependent variable: altura

Parameter	Estimate	Standard Error	T Statistic	P-Value
CONSTANT	165,313	0,856112	193,097	0,0000
sexo	14,0367	1,05129	13,3519	0,0000

Podemos completar el modelo añadiendo otras variables explicativas, por ejemplo, peso.

$$Altura = \beta_0 + \beta_1 \cdot sexo + \beta_2 \cdot peso + e$$

Multiple Regression Analysis

Dependent variable: altura

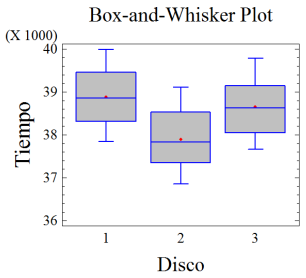
Parameter	Estimate	Standard Error	T Statistic	P-Value
CONSTANT	150,306	3,12145	48,1528	0,0000
peso	0,267968	0,0540419	4,95853	0,0000
sexo	9,28133	1,34214	6,91531	0,0000

Entre un chico y una chica del mismo peso, el chico tiene una altura 9,28 cm mayor.

* Para comparar G grupos podemos construir G variables binarias que denotaremos por D_g , donde cada una toma el valor 1 en aquellos elementos que pertenezcan al grupo *g-ésimo* y 0 al resto.

* Hay que introducir un máximo de G-1 variables binarias, porque si no la primera columna (de unos, correspondiente al término constante de la regresión) será combinación lineal de las otras.

Se quiere comparar el comportamiento de tres discos duros con el fin de ver cuál es el más rápido. Para ello se graba un fichero de 200 MB en cada uno de ellos y se cronometra el tiempo de descarga. Se repite ese experimento un número de veces con cada disco.
Los resultados se encuentran en el fichero Discosduros.sf3. ¿Cuál es el disco duro más rápido?



Queremos ver si el tiempo del disco 2 es significativamente mejor.

$$Y = \beta_0 + \beta_1 D_1 + \beta_3 D_3 + e$$

$$E(Y \mid disco\ 2) = \beta_0$$

$$E(Y \mid disco\ 1) = \beta_0 + \beta_1$$

$$E(Y \mid disco\ 3) = \beta_0 + \beta_3$$

$$Y = \beta_0 + \beta_1 D_1 + \beta_3 D_3 + e$$

Multiple Regression Analysis

Dependent variable: Tiempo

Parameter	Estimate	Standard Error	T Statistic	P-Value
CONSTANT	37896,3	75,572	501,46	0,0000
Disco=1	978,018	107,235	9,12029	0,0000
Disco=3	747,922	118,785	6,29644	0,0000

El 2 es significativamente mejor

$$Y = \beta_0 + \beta_2 D_2 + \beta_3 D_3 + e$$

Multiple Regression Analysis

Dependent variable: Tiempo

Parameter	Estimate	Standard Error	T Statistic	P-Value
CONSTANT	38874,4	76,0809	510,96	0,0000
Disco=2	-978,018	107,235	-9,12029	0,0000
Disco=3	-230,096	119,109	-1,93181	0,0548

* El disco 2 tiene una duración media significativamente inferior en 978,018 unidades de tiempo.
* La diferencia del disco 1 con el disco 3 no parece ser significativa al tener el p-valor mayor que 0,05.
* No podemos asegurar que el 1 y el 3 sean diferentes entre si.