Regresión lineal simple

Tema 4. Regresión lineal simple

Contenidos

- El objeto del análisis de regresión
- La especificación de un modelo de regresión lineal simple
- Estimadores de mínimos cuadrados: construcción y propiedades
- Inferencias sobre el modelo de regresión:
  - Inferencia sobre la pendiente
  - Inferencia sobre la ordenada al origen o
  - Estimación de una respuesta promedio
  - Predicción de una nueva respuesta

Regresión lineal simple

Objetivos de aprendizaje

- Saber construir un modelo de regresión lineal simple que describa cómo influye una variable X sobre otra variable Y
- Saber obtener estimaciones puntuales de los parámetros de dicho modelo
- Saber contruir intervalos de confianza y resolver contrastes sobre dichos parámetros
- Saber estimar el valor promedio de Y para un valor de X
- Saber predecir futuros de la variable respuesta, Y

Introducción

Un modelo de regresión es un modelo que permite describir cómo influye una variable X sobre otra variable Y.

- X: Variable independiente o explicativa
- Y: Variable dependiente o respuesta

El objetivo es obtener estimaciones razonables de Y para distintos valores de X a partir de una muestra de n pares de valores $(x_1, y_1), \dots, (x_n, y_n)$.

Introducción

Ejemplos

- Estudiar cómo influye la estatura del padre sobre la estatura del hijo.
- Estimar el precio de una vivienda en función de su superficie.
- Predecir la tasa de paro para cada edad.
- Aproximar la calificación obtenida en una materia según el número de horas de estudio semanal.
- Prever el tiempo de computación de un programa en función de la velocidad del procesador.

Introducción

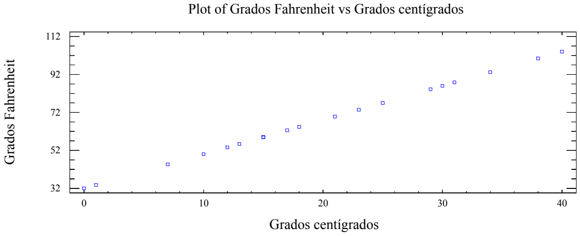
Tipos de relación

- Determinista: Conocido el valor de X, el valor de Y queda perfectamente establecido. Son del tipo:

$$y = f(x)$$

Ejemplo: La relación existente entre la temperatura en grados centígrados (X) y grados Fahrenheit (Y) es:

$$y = 1{,}8x + 32$$



Introducción

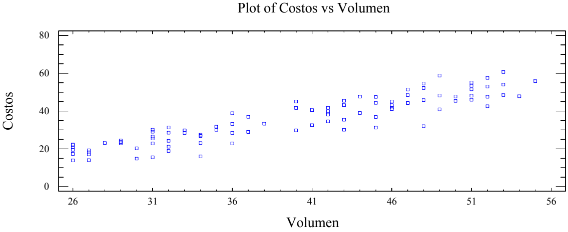
Tipos de relación

- No determinista: Conocido el valor de X, el valor de Y no queda perfectamente establecido. Son del tipo:

$$y = f(x) + u$$

donde u es una perturbación desconocida (variable aleatoria).

Ejemplo: Se tiene una muestra del volumen de producción (X) y el costo total (Y) asociado a un producto en un grupo de empresas.



Existe relación pero no es exacta.

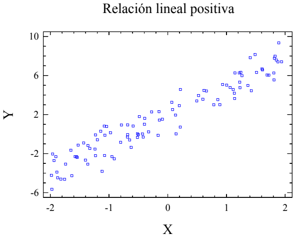
Introducción

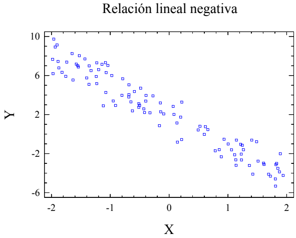
Tipos de relación

- Lineal: Cuando la función $f(x)$ es lineal,

$$f(x) = \beta_0 + \beta_1 x$$

  - Si $\beta_1 > 0$ hay relación lineal positiva.
  - Si $\beta_1 < 0$ hay relación lineal negativa.



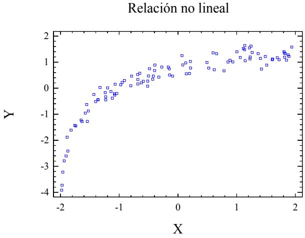


Los datos tienen un aspecto recto.

Introducción

Tipos de relación

- No lineal: Cuando la función $f(x)$ no es lineal. Por ejemplo, $f(x) = log(x)$, $f(x) = x^2 + 3, \ldots$

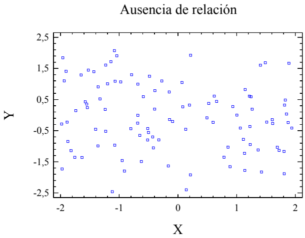


Los datos no tienen un aspecto recto.

Introducción

Tipos de relación

- Ausencia de relación.



El modelo de regresión lineal simple

El modelo de regresión lineal simple supone que,

$$\boxed{y_i = \beta_0 + \beta_1 x_i + u_i}$$

donde:

- y_i representa el valor de la variable respuesta para la observación i-ésima.
- x_i representa el valor de la variable explicativa para la observación i-ésima.
- u_i representa el error para la observación i-ésima que se asume normal,
  $$u_i \sim N(0, \sigma)$$
- β_0 y β_1 son los coeficientes de regresión:
  - β_0 : intercepto
  - β_1 : pendiente

Los parámetros que hay que estimar son: β_0, β_1 y σ.

El modelo de regresión lineal simple

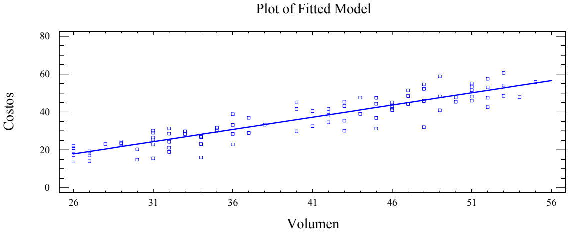
El objetivo es obtener estimaciones $\hat{\beta}_0$ y $\hat{\beta}_1$ de β_0 y β_1 para calcular la recta de regresión:

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

que se ajuste lo mejor posible a los datos.

Ejemplo: Supongamos que la recta de regresión del ejemplo anterior es:

$$\text{Costo} = -15{,}65 + 1{,}29 \text{ Volumen}$$



Se estima que una empresa que produce 25 mil unidades tendrá un costo:

$$\text{costo} = -15{,}65 + 1{,}29 \times 25 = 16{,}6 \text{ mil euros}$$

El modelo de regresión lineal simple

La diferencia entre cada valor y_i de la variable respuesta y su estimación $\hat{y}_i$ se llama residuo:

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

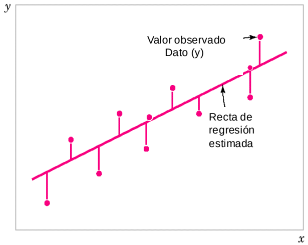


Figure 6-6 Deviations of the data from the estimated regression model.

Ejemplo (cont.): Indudablemente, una empresa determinada que haya producido exactamente 25 mil unidades no va a tener un gasto de exactamente 16,6 mil euros. La diferencia entre el costo estimado y el real es el residuo. Si por ejemplo el costo real de la empresa es de 18 mil euros, el residuo es:

$$e_i = 18 - 16{,}6 = 1{,}4 \text{mil euros}$$

Hipótesis del modelo de regresión lineal simple

- Linealidad: La relación existente entre X e Y es lineal,

$$f(x) = \beta_0 + \beta_1 x$$

- Homogeneidad: El valor promedio del error es cero,

$$E[u_i] = 0$$

- Homocedasticidad: La varianza de los errores es constante,

$$Var(u_i) = \sigma^2$$

- Independencia: Los errores son independientes,

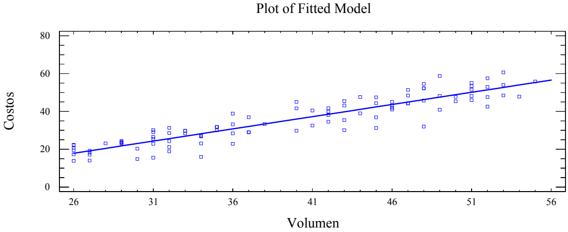
- Normalidad: Los errores siguen una distribución normal,

$$u_i \sim N(0, \sigma)$$

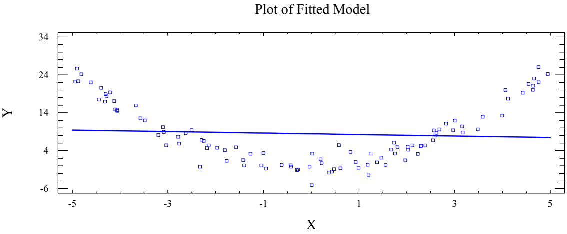
Hipótesis del modelo de regresión lineal simple

Linealidad

Los datos deben ser razonablemante rectos.



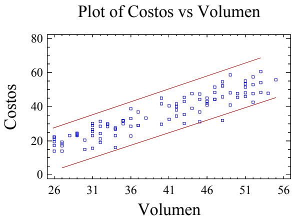
Si no, la recta de regresión no representa la estructura de los datos.



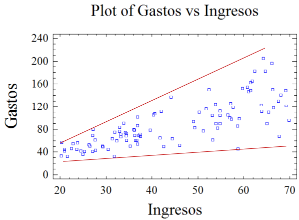
Hipótesis del modelo de regresión lineal simple

Homocedasticidad

La dispersión de los datos debe ser constante para que los datos sean homocedásticos.



Si no se cumple, los datos son heterocedásticos.



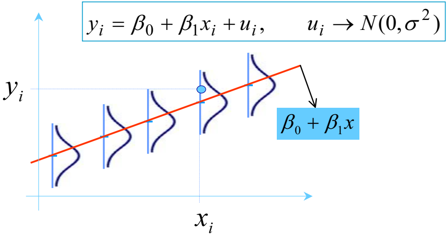
Hipótesis del modelo de regresión lineal simple

Independencia

- Los datos deben ser independientes.
- Una observación no debe dar información sobre las demás.
- Habitualmente, se sabe por el tipo de datos si son adecuados o no para el análisis.
- En general, las series temporales no cumplen la hipótesis de independencia.

Normalidad

- Se asume que los datos son normales a priori.



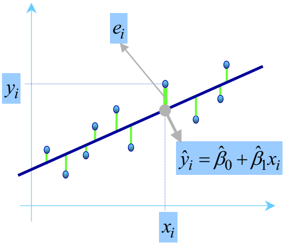
Estimadores de mínimos cuadrados

Gauss propuso en 1809 el método de mínimos cuadrados para obtener los valores $\hat{\beta}_0$ y $\hat{\beta}_1$ que mejor se ajustan a los datos:

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

El método consiste en minimizar la suma de los cuadrados de las distancias verticales entre los datos y las estimaciones, es decir, minimizar la suma de los residuos al cuadrado,

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} \left(y_i - \left(\hat{\beta}_0 + \hat{\beta}_1 x_i\right)\right)^2$$

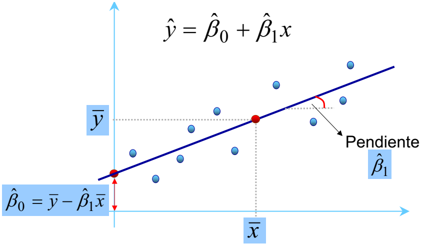


Estimadores de mínimos cuadrados

El resultado que se obtiene es:

$$\boxed{\hat{\beta}_1 = \frac{s(x,y)}{s_{xX}} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}}$$

$$\boxed{\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}}$$



Estimación de la varianza

Para estimar la varianza de los errores, σ^2, podemos utilizar,

Un estimador insesgado de σ^2 es la varianza residual S^2 o sigma cuadrado estimado,

$$\boxed{\hat{\sigma}^2 = \frac{\sum_{i=1}^{n} \mathrm{r}_i^2}{n-2}}$$

Estimación de la varianza

Ejercicio 4.2

Calcula la varianza residual en el ejercicio 4.1.

Resultados

Calculamos primero los residuos, e_i, usando la recta de regresión,

$$\hat{y}_i = 74{,}116 - 1{,}3537 x_i$$

x_i	30	28	32	25	25	25	22	24	35	40
y_i	25	30	27	40	42	40	50	45	30	25
$\hat{y}_i$	33.5	36.21	30.79	40.27	40.27	40.27	44.33	41.62	26.73	19.96
e_i	-8.50	-6.21	-3.79	-0.27	1.72	-0.27	5.66	3.37	3.26	5.03

La varianza residual es:

$$s_R^2 = \frac{\sum_{i=1}^{n} e_i^2}{n-2} = \frac{207{,}92}{8} = 25{,}99$$

Estimación de la varianza

Ejercicio

Calcula la varianza residual

Resultados

Calculamos primero los residuos, e_i, usando la recta de regresión,

$$\hat{y}_i = 74{,}116 - 1{,}3537 x_i$$

x_i	30	28	32	25	25	25	22	24	35	40
y_i	25	30	27	40	42	40	50	45	30	25
$\hat{y}_i$	33.5	36.21	30.79	40.27	40.27	40.27	44.33	41.62	26.73	19.96
e_i	-8.50	-6.21	-3.79	-0.27	1.72	-0.27	5.66	3.37	3.26	5.03

La varianza residual es:

$$s^2 = \frac{\sum_{i=1}^{n} \mathrm{r}_i^2}{n-2} = \frac{207{,}92}{8} = 25{,}99$$

Inferencias sobre el modelo de regresión

- Hasta ahora sólo hemos obtenido estimaciones puntuales de los coeficientes de regresión.
- Usando intervalos de confianza podemos obtener una medida de la precisión de dichas estimaciones.
- Usando contrastes de hipótesis podemos comprobar si un determinado valor puede ser el auténtico valor del parámetro.

Inferencia para la pendiente

El estimador $\hat{\beta}_1$ sigue una distribución normal porque es una combinación lineal de normales,

$$\hat{\beta}_1 = \sum_{i=1}^{n} \frac{(x_i - \bar{x})}{s_{XX}} y_i = \sum_{i=1}^{n} w_i y_i$$

donde $y_i = \beta_0 + \beta_1 x_i + u_i$, que cumple que $y_i \sim N\left(\beta_0 + \beta_1 x_i, \sigma^2\right)$.
Además, $\hat{\beta}_1$ es un estimador insesgado de β_1,

$$E\left[\hat{\beta}_1\right] = \sum_{i=1}^{n} \frac{(x_i - \bar{x})}{s_{XX}} E\left[y_i\right] = \beta_1,$$

y su varianza es,

$$Var\left[\hat{\beta}_1\right] = \sum_{i=1}^{n} \left(\frac{(x_i - \bar{x})}{s_{XX}}\right)^2 Var\left[y_i\right] = \frac{\sigma^2}{s_{XX}}$$

Por tanto,

$$\hat{\beta}_1 \sim N\left(\beta_1, \frac{\sigma^2}{s_{XX}}\right)$$

Intervalo de confianza para la pendiente

Queremos ahora obtener el intervalo de confianza para β_1 de nivel $1-\alpha$. Como σ^2 es desconocida, la estimamos con s^2. El resultado b ´asico cuando la varianza es desconocida es:

$$\frac{\hat{\beta}_1 - \beta_1}{\sqrt{\dfrac{s^2}{s_{XX}}}} \sim t_{n-2}$$

que nos permite obtener el intervalo de confianza p ara β_1:

$$\hat{\beta}_1 \pm t_{n-2,\alpha/2}\sqrt{\frac{s^2}{s_{XX}}}$$

La longitud del intervalo disminuirá si:

- Aumenta el tamaño de la muestra.
- Aumenta la varianza de las x_i.
- Disminuye la varianza residual.

Contrastes sobre la pendiente

Usando el resultado anterior podemos resolver contrastes sobre β_1. En particular, si el verdadero valor de β_1 es cero entonces Y no depende linealmente de X. Por tanto, es de especial interés el contraste:

$$H_0 : \beta_1 = 0$$
$$H_1 : \beta_1 \neq 0$$

La región de rechazo de la hipótesis nula es:

$$\left| \frac{\hat{\beta}_1}{\sqrt{s^2 / s_{\mathrm{xx}}}} \right| > t_{n-2,\alpha/2}$$

Equivalentemente, si el cero está fuera del intervalo de confianza para β_1 de nivel $1-\alpha$, rechazamos la hipótesis nula a ese nivel. El p-valor del contraste es:

$$p\text{-valor} = 2\,\mathrm{P}\left(t_{n-2} > \left| \frac{\hat{\beta}_1}{\sqrt{s^2 / s_{\mathrm{xx}}}} \right| \right)$$

Inferencia para el intercepto

El estimador $\hat{\beta}_0$ sigue una distribución normal porque es una combinación lineal de normales,

$$\hat{\beta}_0 = \sum_{i=1}^{n} \left(\frac{1}{n} - w_i \right) y_i$$

donde $w_i = (x_i - \bar{x}) \,/\, s_{XX}$ y donde $y_i = \beta_0 + \beta_1 x_i + u_i$, que cumple que $y_i \sim N\left(\beta_0 + \beta_1 x_i, \sigma^2\right)$. Además, $\hat{\beta}_0$ es un estimador insesgado de β_0,

$$E\left[\hat{\beta}_0\right] = \sum_{i=1}^{n} \left(\frac{1}{n} - \bar{x} w_i \right) E\left[y_i\right] = \beta_0$$

y su varianza es,

$$Var\left[\hat{\beta}_0\right] = \sum_{i=1}^{n} \left(\frac{1}{n} - \bar{x} w_i \right)^2 Var\left[y_i\right] = \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{s_{xx}} \right)$$

y por tanto,

$$\hat{\beta}_0 \sim N\left(\beta_0, \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{s_{XX}} \right)\right)$$

Intervalo de confianza para el intercepto

Queremos ahora obtener el intervalo de confianza para β_0 de nivel $1-\alpha$. Como σ^2 es desconocida, la estimamos con s^2. El resultado básico cuando la varianza es desconocida es:

$$\frac{\hat{\beta}_0 - \beta_0}{\sqrt{s^2\left(\frac{1}{n} + \frac{\bar{x}^2}{s_{XX}}\right)}} \sim t_{n-2}$$

que nos permite obtener el intervalo de confianza para β_0:

$$\hat{\beta}_0 \pm t_{n-2,\alpha/2}\sqrt{s^2\left(\frac{1}{n} + \frac{\bar{x}^2}{s_{XX}}\right)}$$

La longitud del intervalo disminuirá si:

- Aumenta el tamaño de la muestra.
- Aumenta la varianza de las x_i.
- Disminuye la varianza residual.
- Disminuye la media de las x_i.

Contrastes sobre el intercepto

Usando el resultado anterior podemos resolver contrastes sobre β_0. En particular, si el verdadero valor de β_0 es cero entonces la recta de regresión pasa por el origen. Por tanto, es de especial interés el contraste:

$$H_0 : \beta_0 = 0$$
$$H_1 : \beta_0 \neq 0$$

La región de rechazo de la hipótesis nula es:

$$\left| \frac{\hat{\beta}_0}{\sqrt{s^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{s_{XX}} \right)}} \right| > t_{n-2,\alpha/2}$$

Equivalentemente, si el cero está fuera del intervalo de confianza para β_0 de nivel $1-\alpha$, rechazamos la hipótesis nula a ese nivel. El p-valor es:

$$p\text{-valor} = 2\,\Pr\left(t_{n-2} > \left| \frac{\hat{\beta}_0}{\sqrt{s^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{s_{XX}} \right)}} \right| \right)$$

Estimación de una respuesta promedio y predicción de una nueva respuesta

Se distiguen dos tipos de problemas:

1. Estimar el valor medio de la variable Y para cierto valor $X = x_0$.
2. Predecir el valor que tomará la variable Y para cierto valor $X = x_0$.

Por ejemplo, en el ejercicio 4.1:

1. ¿Cuál será el precio medio del kg. de harina para los años en que se producen 30 ton. de trigo?
2. Si un determinado año se producen 30 ton. de trigo, ¿cuál será el precio del kg. de harina?

En ambos casos el valor estimado es:

$$\begin{aligned} \hat{y}_0 &= \hat{\beta}_0 + \hat{\beta}_1 x_0 \\ &= \bar{y} + \hat{\beta}_1 (x_0 - \bar{x}) \end{aligned}$$

Pero la precisión de las estimaciones es diferente.

Estimación de una respuesta promedio

Teniendo en cuenta que:

$$\begin{aligned} Var\left(\hat{y}_0\right) &= Var\left(\bar{y}\right) + \left(x_0 - \bar{x}\right)^2 Var\left(\hat{\beta}_1\right) \\ &= \sigma^2 \left(\frac{1}{n} + \frac{\left(x_0 - \bar{x}\right)^2}{s_{\mathrm{xx}}}\right) \end{aligned}$$

El intervalo de confianza para la respuesta promedio es:

$$\hat{y}_0 \pm t_{n-2,\alpha/2} \sqrt{s_R^2 \left(\frac{1}{n} + \frac{\left(x_0 - \bar{x}\right)^2}{s_{\mathrm{xx}}}\right)}$$

Predicción de una nueva respuesta

La varianza de la predicción de una nueva respuesta es el error cuadrático medio de la predicción:

$$E\left[(y_0 - \hat{y}_0)^2\right] = Var\left(y_0\right) + Var\left(\hat{y}_0\right)$$
$$= \sigma^2 \left(1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{s_{\mathrm{xx}}}\right)$$

El intervalo de confianza para la predicción de una nueva respuesta es:

$$\hat{y}_0 \pm t_{n-2,\alpha/2} \sqrt{s_R^2 \left(1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{s_{\mathrm{xx}}}\right)}$$

La longitud de este intervalo es mayor que la del anterior (menos precisión) porque no corresponde a un valor medio sino a uno específico.

Estimación de una respuesta promedio y predicción de una nueva respuesta

En rojo se muestran los intervalos para las medias estimadas y en rosa los intervalos de predicción. Se observa que la amplitud de estos últimos es considerablemente mayor.