Tema 2. Problemas de inferencia estadística en el modelo de regresión lineal simple.

1. Inferencia respecto a los parámetros

La construcción de los **intervalos de confianza** para los parámetros β_1, β_0 y σ^2 se realizará a partir de las distribuciones de probabilidad de $\hat{\beta}_1, \hat{\beta}_0$ y s_R^2.

1.1 Intervalo de confianza para β_1

Caso 1: Si σ^2 es conocida, vimos en el apartado 4.1 del Tema 1, que

$$\hat{\beta}_1 \sim N\left(\beta_1, \frac{\sigma^2}{n s_x^2}\right) \Rightarrow \frac{\hat{\beta}_1 - \beta_1}{\frac{\sigma}{\sqrt{n}\, s_x}} \sim N(0,1)$$

Proposición: El intervalo

$$I.C.(\beta_1) = \left[\hat{\beta}_1 - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}\, s_x},\ \hat{\beta}_1 + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}\, s_x}\right],$$

donde $z_{1-\alpha/2}$ es el percentil $(1-\alpha/2)100$ de la ley normal estándar, es un intervalo de confianza para β_1 con un nivel de confianza $1-\alpha$ cuando σ^2 es conocida. (Demostración)

De la fórmula anterior se deduce que **la amplitud del intervalo** depende de:

- El **nivel de confianza**: A mayor nivel de confianza, mayor es la amplitud y, por tanto, menos precisión.
- La **varianza del estimador** $\hat{\beta}_1$: A mayor varianza, mayor amplitud.

Caso 2: Cuando σ^2 es desconocida, ésta se sustituye por su estimador, s_R^2, y entonces

$$\frac{\hat{\beta}_1 - \beta_1}{\frac{s_R}{\sqrt{n}\, s_x}} \sim t_{n-2}$$

Proposición: El intervalo

$$I.C.(\beta_1) = \left[\hat{\beta}_1 - t_{1-\alpha/2}\frac{s_R}{\sqrt{n}\, s_x},\ \hat{\beta}_1 + t_{1-\alpha/2}\frac{s_R}{\sqrt{n}\, s_x}\right],$$

donde $t_{1-\alpha/2}$ es el percentil $(1-\alpha/2)100$ de la ley t de Student con $n-2$ grados de libertad, es un intervalo de confianza para β_1 con un nivel de confianza $1-\alpha$ cuando σ^2 es desconocida. (Demostración análoga)

1.2 Intervalo de confianza para β_0

Análogamente se deducen los intervalos de confianza para el parámetro β_0:

Caso 1: Si σ^2 es conocida, entonces:

$$\hat{\beta}_0 \sim N\left(\beta_0, \frac{\sigma^2}{n}\left(1+\frac{\bar{x}^2}{s_x^2}\right)\right) \Rightarrow \frac{\hat{\beta}_0 - \beta_0}{\frac{\sigma}{\sqrt{n}}\sqrt{1+\frac{\bar{x}^2}{s_x^2}}} \sim N(0,1)$$

Proposición: El intervalo

$$I.C.(\beta_0) = \left[\hat{\beta}_0 - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\sqrt{1+\frac{\bar{x}^2}{s_x^2}},\ \hat{\beta}_0 + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\sqrt{1+\frac{\bar{x}^2}{s_x^2}}\right],$$

donde $z_{1-\alpha/2}$ es el percentil $(1-\alpha/2)100$ de la ley normal estándar, es un intervalo de confianza para β_0 con un nivel de confianza $1-\alpha$ cuando σ^2 es conocida.

Caso 2: Si σ^2 es desconocida, entonces:

$$\frac{\hat{\beta}_0 - \beta_0}{\frac{s_R}{\sqrt{n}}\sqrt{1+\frac{\bar{x}^2}{s_x^2}}} \sim t_{n-2}$$

Proposición: El intervalo

$$I.C.(\beta_0) = \left[\hat{\beta}_0 - t_{1-\alpha/2}\frac{s_R}{\sqrt{n}}\sqrt{1+\frac{\bar{x}^2}{s_x^2}},\ \hat{\beta}_0 + t_{1-\alpha/2}\frac{s_R}{\sqrt{n}}\sqrt{1+\frac{\bar{x}^2}{s_x^2}}\right],$$

donde $t_{1-\alpha/2}$ es el percentil *(1-α/2)100* de la ley *t* de Student con *n-2* grados de libertad, es un intervalo de confianza para β_0 con un nivel de confianza *1-α* cuando σ^2 es desconocida.

Observación: Para $n \geq 30$, aunque σ^2 sea desconocida, la ley *t* de Student suele aproximarse por una ley normal estándar. Es decir, que si $n \geq 30$ entonces $t_{1-\alpha/2} \approx z_{1-\alpha/2}$.

Ejemplo: Hallar los I.C. al 99% para los parámetros β_1, β_0 del **Ejemplo 1** (densidad del tráfico) del Tema 1.

En el **Ejemplo 1** (del Tema 1) se intentaba modelar la velocidad de circulación (*y*) en cierta carretera en función de la densidad del tráfico (*x*).

Para este conjunto de datos se tenía que:

$$n = 24,\ s_x^2 = 635.73,\ s_R^2 = 0.0377,\ \bar{x} = 54.44,\ \ \hat{\beta}_1 = -0.057,\ \hat{\beta}_0 = 8.09.$$

Además, σ^2 es desconocida. Los I.C. deben calcularse al *(1-α)100%=99%,* por tanto, $t_{1-\alpha/2} = t_{0.995} = 2.82$.

$$I.C.(\beta_1) = \left[\hat{\beta}_1 \mp t_{1-\alpha/2}\frac{s_R}{\sqrt{n}\,s_x}\right] = \left[-0.057 \mp 2.82\sqrt{\frac{0.0377}{24\cdot 635.73}}\right] = \left[-0.0614,\ -0.0526\right]$$

$$I.C.(\beta_0) = \left[\hat{\beta}_0 \mp t_{1-\alpha/2}\frac{s_R}{\sqrt{n}}\sqrt{1+\frac{\bar{x}^2}{s_x^2}}\right] = \left[8.09 \mp 2.82\sqrt{\frac{0.0377}{24}\left(1+\frac{54.44^2}{635.73}\right)}\right] = \left[7.8240,\ 8.3559\right]$$

Ejercicio: Hallar los I.C. al 99% para los parámetros β_1, β_0 del **Ejemplo 2** (esperanza de vida) del Tema 1.

Ejercicio 7.

a) Hallar un intervalo de confianza para la pendiente de la recta ***y=5.9+0.59 x***, donde ***y*** es el salario mensual y ***x*** son los años de escolarización, para un nivel de confianza del 90%.

b) Lo mismo para el término independiente.

Ejercicio 8. Hallar los intervalos de confianza al 95% para β_0, β_1 para ***y=79.3210-1.3765 x***, donde ***y*** es la longitud de la línea de la mano y ***x*** es la edad al morir.

1.3 Intervalo de confianza para σ^2

Utilizando el resultado que vimos en el apartado 5 del Tema 1:

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

se puede deducir el intervalo de confianza para la varianza del modelo.

Proposición: El intervalo

$$I.C.(\sigma^2) = \left[\frac{n-2}{x_{1-\alpha/2}}s_R^2,\ \frac{n-2}{x_{\alpha/2}}s_R^2\right],$$

donde $x_{\alpha/2}$, $x_{1-\alpha/2}$ son, respectivamente, los percentiles *(α/2)100* y *(1-α/2)100* de la ley chi-cuadrado con *n-2* grados de libertad, es un intervalo de confianza para σ^2 con un nivel de confianza *1-α*.

Ejemplo: Hallar el I.C. al 90% para el parámetro σ^2 del **Ejemplo 1** (densidad del tráfico) del Tema 1.

Para este conjunto de datos se tenía que $n=24,\ s_R^2=0.0377.$ El I.C. debe calcularse al *(1-α)100%=90%,* por tanto, $x_{\alpha/2}=x_{0.05}=33.9,\ x_{1-\alpha/2}=x_{0.95}=33.9.$.

$$I.C.(\sigma^2)=\left[\frac{n-2}{x_{1-\alpha/2}}s_R^2,\ \frac{n-2}{x_{\alpha/2}}s_R^2\right]=\left[\frac{22}{33.9}0.0377,\ \frac{22}{12.3}0.0377\right]=[0.0245,\ 0.0691]$$

Ejercicio 9. Hallar un intervalo de confianza al 90% para σ^2 con los datos del **Ejercicio 5** (longitud de la línea de la mano).

Ejercicio 10. Hallar un intervalo de confianza al 95% para σ^2 con los datos del **Ejercicio 6** (salario-escolarización).

2. El contraste de regresión

2.1. El contraste de la pendiente (*t*- Student)

El objetivo general de los contrastes de hipótesis es ayudar al investigador a tomar una decisión sobre la población de estudio, a partir de una muestra de ella.

En este caso, **el contraste de hipótesis sobre la pendiente** nos ayudará a decidir si el parámetro β_1 admite un valor concreto β_1^*.

Mediante un contraste de hipótesis podremos responder a preguntas como, por ejemplo:

1) ¿El valor β_1^* es un valor admisible para β_1?
2) ¿Es realmente significativa (es decir, influye) la variable x sobre la variable y?
3) Se sabe de investigaciones anteriores que β_1 vale β_1^*, ¿mi resultado es similar, teniendo en cuenta que a partir de una muestra mi estimación es $\hat{\beta}_1$?

El planteamiento es similar en cualquiera de los tres casos.

2.1.1. Contraste bilateral:

Se plantean dos hipótesis: la hipótesis nula, H_0, y la hipótesis alternativa, H_1:

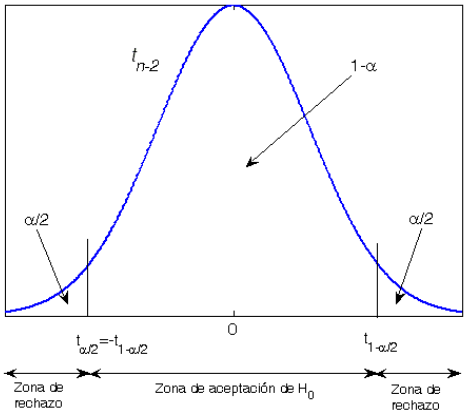
$$\begin{cases} H_0: \beta_1=\beta_1^* \\ H_1: \beta_1\neq\beta_1^* \end{cases}$$

Para resolver el contraste se construye un **estadístico de contraste,** cuya distribución debe ser **conocida** bajo el supuesto de la hipótesis H_0. En este caso, se utiliza:

$$t_{\exp}=\frac{\hat{\beta}_1-\beta_1^*}{\frac{s_R}{\sqrt{n}s_x}}\sim t_{n-2}$$

Cuanto más parecidos sean la estimación $\hat{\beta}_1$ y el valor del parámetro que se quiere contrastar β_1^*, más próximo a cero será el valor de $t_{\exp}$.

Utilizando la ley del estadístico de contraste, se construyen las **zonas de rechazo** y **de aceptación** de H_0, para un nivel de significación α fijado:



Regla de decisión:

Para un nivel de significación α fijado, se rechaza la hipótesis H_0 si

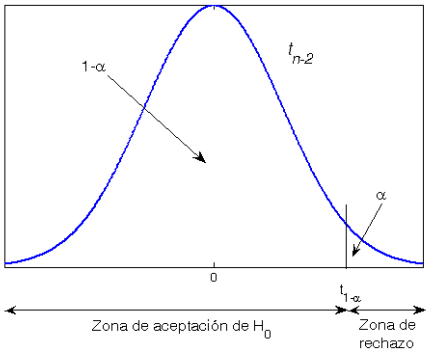
$$\left|t_{\exp}\right|>t_{1-\alpha/2}^{(n-2)}$$

2.1.2. Contraste unilateral izquierdo:

Hipótesis del contraste:

$$\begin{cases} H_0 : \beta_1 = \beta_1^* \\ H_1 : \beta_1 > \beta_1^* \end{cases}$$

$$t_{\exp} = \frac{\hat{\beta}_1 - \beta_1^*}{\frac{s_R}{\sqrt{n}s_x}}$$



Regla de decisión:

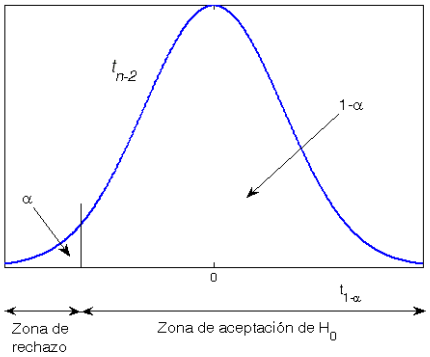
Para un nivel de significación fijado, α, se rechaza la hipótesis H_0 si $t_{\exp} > t_{1-\alpha}^{(n-2)}$

2.1.3. Contraste unilateral derecho:

Hipótesis del contraste:

$$\begin{cases} H_0 : \beta_1 = \beta_1^* \\ H_1 : \beta_1 < \beta_1^* \end{cases}$$

$$t_{\exp} = \frac{\hat{\beta}_1 - \beta_1^*}{\frac{s_R}{\sqrt{n}s_x}}$$



Regla de decisión:

Para un nivel de significación fijado, α, se rechaza la hipótesis H_0 si $t_{\exp} < -t_{1-\alpha}^{(n-2)}$

Observaciones:

- Es de particular interés el **contraste de significación** del parámetro β_1, que consiste en tomar $\beta_1^*=0$ en cualquiera de los contrastes anteriores:

$\begin{cases} H_0 : \beta_1 = 0 \\ H_1 : \beta_1 \neq 0 \end{cases}$	$\begin{cases} H_0 : \beta_1 = 0 \\ H_1 : \beta_1 > 0 \end{cases}$	$\begin{cases} H_0 : \beta_1 = 0 \\ H_1 : \beta_1 < 0 \end{cases}$
bilateral	unilateral izquierdo	unilateral derecho

  En este caso, el no rechazar la hipótesis H_0 y, por tanto, concluir que el valor 0 es un valor admisible para el parámetro β_1, es equivalente a afirmar que **la variable explicativa no influye en la variable respuesta**.

- Los paquetes estadísticos, como Statgraphics, basan el criterio de decisión en el *p-valor*, que para los contrastes unilaterales es la probabilidad de cola asociada al valor del estadístico de test.
  Para un nivel de significación fijado α, se rechaza la hipótesis H_0 si *p-valor* < α.

Ejemplo: Contrastar con los datos del **Ejemplo 1** (densidad del tráfico) la hipótesis nula de que la densidad no influye (linealmente) en la velocidad frente a la alternativa de que un incremento de la densidad produce un descenso en la velocidad, para un nivel de significación del 5%.

x="densidad del tráfico", y="velocidad",

$$n = 24,\ s_x^2 = 635.73,\ s_R^2 = 0.0377,\ \hat{\beta}_1 = -0.057,\ \hat{\beta}_0 = 8.09.$$

Recta de regresión: $y=8.09-0.057x$

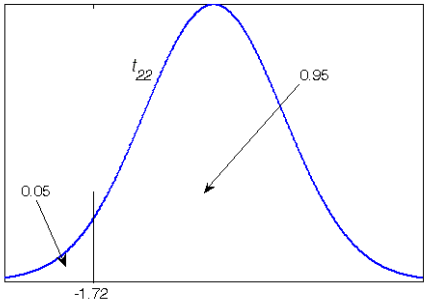
Planteamos el siguiente contraste:

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

A partir de la información muestral, calculamos el estadístico de test:

$$t_{\exp} = \frac{\hat{\beta}_1 - \beta_1^*}{\frac{s_R}{\sqrt{n}s_x}} = \frac{-0.057 - 0}{\sqrt{\frac{0.0377}{24(635.73)}}} = -36.26$$

Comparamos el valor del estadístico de test con el percentil 5% de la ley t de Student con 22 grados de libertad:



$$t_{\exp} = -36.26 < -1.72 = -t_{0.95}^{(22)}$$

¿Cuánto vale el *p-valor*?

$$P(t_{22} < -36.26) = 2.03249E\text{-}21 < \alpha$$

Concluimos que $\beta_1<0$ y, por tanto, la velocidad está influida por la densidad del tráfico, para un nivel de significación del 5%.

Contraste de hipótesis a través del intervalo de confianza

El intervalo de confianza permite realizar el contraste de hipótesis bilateral.

Por ejemplo, para resolver el contraste de significación de β_1, para un nivel de significación fijado α, podemos utilizar el intervalo de confianza de β_1 de la forma siguiente:

Basta verificar si para un nivel de confianza del *(1-α)100%*, el valor 0 está contenido en el I.C. siguiente:

$$I.C.(\beta_1)_{1-\alpha} = \left[\hat{\beta}_1 - t_{1-\alpha/2}^{(n-2)}\frac{s_R}{\sqrt{n}s_x}, \hat{\beta}_1 + t_{1-\alpha/2}^{(n-2)}\frac{s_R}{\sqrt{n}s_x}\right]$$

Si el valor 0 pertenece a este intervalo concluiremos que 0 es un valor admisible para β_1 para el nivel de significación α.

Ejercicio 11. (producción-fertilizantes) Se dispone de los siguientes datos experimentales obtenidos en un campo de cultivo que relacionan la producción con la cantidad de fertilizante aplicado.

Fertilizantes (kg/hect.) ***x***	100	200	300	400	500	600	700
Producción (kg/hect.) ***y***	40	45	50	65	70	70	80

Hallar:

a) La nube de puntos y dibujar una recta que pase "lo más cerca posible" de todos sus puntos y en especial por el centroide.
b) La recta de regresión de ***y*** sobre ***x***. Interpretar los coeficientes.
c) La varianza residual.
d) I.C. al 95% para β_1.
e) I.C. al 95% para σ^2 y para σ.
f) Coeficiente de correlación y coeficiente de determinación. Interpretar los resultados.
g) Si se aplican 350kg/hect. de fertilizante, ¿qué producción se obtendrá? ¿Y con 1000 kg/hect.?
h) Realiza un contraste de hipótesis con un nivel de significación del 5% para comprobar si la producción depende del fertilizante.

2.2. Contraste de regresión (ANOVA)

Definición: Se denomina **contraste de regresión** al contraste de hipótesis sobre la pendiente de la recta de regresión.

$$\begin{cases} H_0 : \beta_1 = 0 \\ H_1 : \beta_1 \neq 0 \end{cases} \Leftrightarrow \begin{cases} H_0 : E(y/x) = \beta_0 \quad \text{constante} \\ H_1 : E(y/x) = \beta_0 + \beta_1 x \end{cases}$$

Aunque en la sección anterior ya hemos deducido este contraste, aquí vamos a resolverlo mostrando su relación con el análisis de la varianza (ANOVA).

Recordemos la descomposición de la variabilidad:

$$VT = \sum_{i=1}^{n}(y_i - \bar{y})^2 = ns_y^2$$

$$VE = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 = n\hat{\beta}_1^2 s_x^2$$

$$VNE = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} e_i^{\,2} = (n-2)s_R^2 = n(s_y^2 - \hat{\beta}_1^2 s_x^2)$$

Por un lado, sabemos que:

$$\frac{VNE}{\sigma^2} \sim \chi^2_{n-2}$$

Y por otro, se puede demostrar (Peña vol. II, apéndice 11B) que cuando $\beta_1=0$ (es decir, bajo el supuesto de H_0):

$$\frac{VT}{\sigma^2} \sim \chi^2_{n-1} \quad \text{y} \quad \frac{VE}{\sigma^2} \sim \chi^2_1$$

Por tanto, si $\beta_1=0$, se construye el siguiente **estadístico de contraste**:

$$F_{\exp} = \frac{\frac{VE}{1}}{\frac{VNE}{n-2}} = \frac{VE}{s_R^2} = \frac{n\hat{\beta}_1^2 s_x^2}{s_R^2} \sim F_{1,n-2} \quad \text{F de Fisher con 1 y n - 2 gr. lib.}$$

La **regla de decisión** ahora es: Para un nivel de significación fijado, α, se rechaza la hipótesis H_0 si $F_{\exp} > F^{\alpha}_{1,n-2}$

Cuando $F_{\exp}$ es **grande** significa la variabilidad explicada por el modelo es mucho mayor que la variabilidad residual.

Este contraste suele realizarse a partir de la tabla ANOVA:

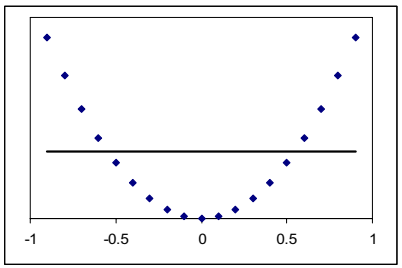
Fuente de variación	Sumas de Cuadrados	Grados de libertad	Cociente o varianza	$F_{\exp}$
VE	$n\hat{\beta}_1^2 s_x^2$	1	$n\hat{\beta}_1^2 s_x^2$	$\frac{n\hat{\beta}_1^2 s_x^2}{s_R^2}$
VNE	$(n-2)s_R^2$	$n-2$	s_R^2	
VT	ns_y^2	$n-1$		

El estadístico $F_{\exp}$ compara el cuadrado del estimador $\hat{\beta}_1$ con su varianza.

Atención!: Este contraste supone **linealidad**, puesto que compara dos modelos:

$$\begin{cases} H_0 : E(y/x) = \beta_0 \quad \text{constante} \quad (\beta_1 = 0) \\ H_1 : E(y/x) = \beta_0 + \beta_1 x \quad (\beta_1 \neq 0) \end{cases}$$

es decir, analiza si **una recta con pendiente no nula representa mejor a los datos que una recta horizontal**. Por tanto, **aceptar H_0** (recta horizontal) **no implica** que x e y sean independientes.



En este caso, se aceptaría H_0 porque la recta horizontal tiene mejor ajuste que cualquier otra recta.

Aunque $\beta_1=0$, existe relación entre x e y.

Para evitar casos patológicos como este, es siempre conveniente **representar los datos** mediante un diagrama de dispersión.

Solamente podremos concluir a partir del contraste de regresión que las variables x e y **son independientes si conjuntamente tienen ley normal bivariante**.

Equivalencia entre el contraste de regresión y el contraste de significación de β_1

Contraste de regresión:

$$F_{\exp} = \frac{n\hat{\beta}_1^2 s_x^2}{s_R^2} \sim F_{1,n-2}$$

Contraste de significación de β_1:

$$t_{\exp} = \frac{\hat{\beta}_1}{s_R/\sqrt{n}s_x} = \frac{\sqrt{n}s_x\hat{\beta}_1}{s_R} \sim t_{n-2}$$

Elevando $t_{\exp}$ al cuadrado se tiene que:

$$t^2_{\exp} = \left(\frac{\sqrt{n}s_x\hat{\beta}_1}{s_R}\right)^2 = F_{\exp}$$

Ejemplo: Con los datos del **Ejercicio 11** (producción-fertilizantes) realizar un contraste de regresión con un nivel de significación del 1%.

y="producción" (en kg/hect), x="fertilizantes" (en kg/hect).
Para este conjunto de datos se tenía que:

$$n=7,\ s_x^2=40000,\ s_R^2=11.0559,\quad \hat{\beta}_1=0.068,\ \hat{\beta}_0=32.8.$$

Recta de regresión estimada: $y=32.8+0.068x$

El contraste de regresión que se plantea es:

$$\begin{cases} H_0: E(y/x)=\beta_0 \\ H_1: E(y/x)=\beta_0+\beta_1 x \end{cases} \Leftrightarrow \begin{cases} H_0: \beta_1=0 \\ H_1: \beta_1\neq 0 \end{cases}$$

Calculamos el estadístico de contraste:

$$F_{\exp}=\frac{n\hat{\beta}_1^2 s_x^2}{s_R^2}=\frac{7\cdot 0.068^2\cdot 40000}{11.0559}=117.1067$$

Puesto que $F_{\exp}=117.1067>16.23=F_{1,5}(0.05)$, se rechaza H_0, con lo que se concluye que la recta con pendiente no nula es mejor para representar estos datos.

Otra forma de resolver este Ejemplo es mediante la tabla ADEVA (ANOVA):

Fuente de variación	Sumas de Cuadrados	Grados de libertad	Cociente o varianza	$F_{\exp}$
VE	1294.72	1	1294.72	117.1067
VNE	55.2795	5	11.0559	
VT	1349.9997	6		

$F_{\exp}=117.1067>16.23=F_{1,5}(0.05)$, se rechaza H_0, sí existe relación lineal entre x e y.

Ejercicio 12. Dada la ecuación de regresión $y=3+4x$ indicar, mediante el contraste de regresión, si hay evidencias de que y depende linealmente de x con un nivel de significación $\alpha=0.01$ y sabiendo que $R^2=0.4$ y $n=25$.

3. Contraste de las hipótesis mediante los residuos: Diagnosis del modelo.

Se realiza mediante el estudio gráfico de los residuos.

Objetivo: validar las hipótesis sobre las que se basa el modelo de regresión.

- Linealidad

$$E(u_i)=0 \quad \text{ó} \quad E(y|x)=\beta_0+\beta_1 x$$

- Normalidad

$$u_i \sim Normal$$

- Homocedasticidad

$$\text{var}(u_i)=\sigma^2$$

- Independencia

$$\text{cov}(u_i,u_j)=0,\quad \forall i\neq j$$

Permite depurar los datos:

- Eliminar atípicos (cuando sea conveniente)
- Mejorar la Normalidad
- Mejorar la homocedasticidad
- Transformar un modelo no lineal en uno que sea lineal.

Gráficos preliminares:

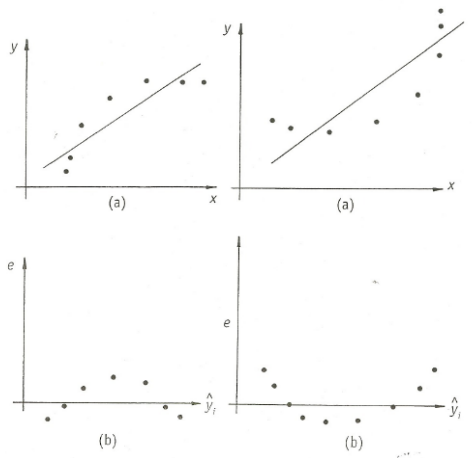
- Conjuntos: diagrama de dispersión (detectar desviaciones de la linealidad)
- Individuales: hitogramas, gráficos de cajas (detectar desviaciones de la normalidad: más de una moda, asimetría, atípicos…)

Gráficos posteriores:

- Residuos vs. Valores predichos ($\hat{y}$)
- Residuos vs. Variable explicativa (x)
- Residuos vs. Sencuencia temporal
- Residuos vs. Otra variable de interés

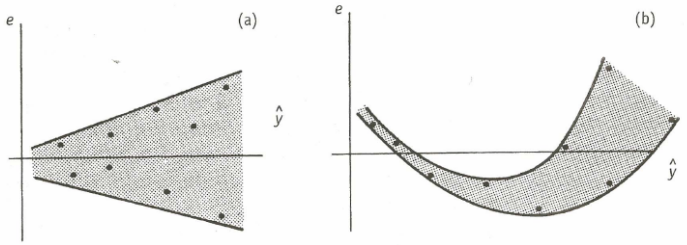
Falta de linealidad

Figura 6.2 Relaciones no lineales y sus residuos



Falta de homocedasticidad

Figura 6.3 Heterocedasticidad, con falta de linealidad en (b)



Falta de independencia entre los residuos

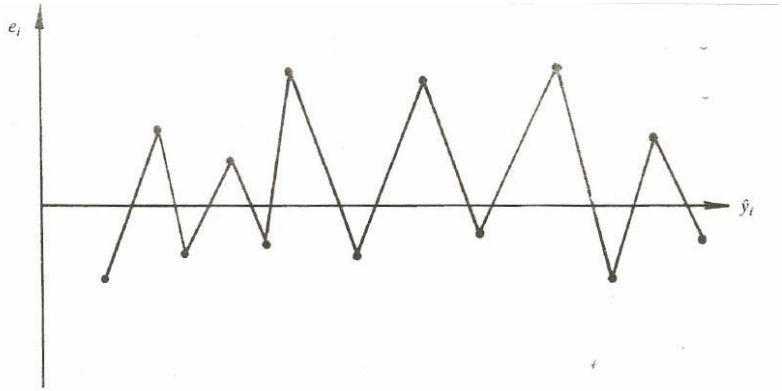


FIGURA 11.22. *Autocorrelación negativa entre los residuos.*

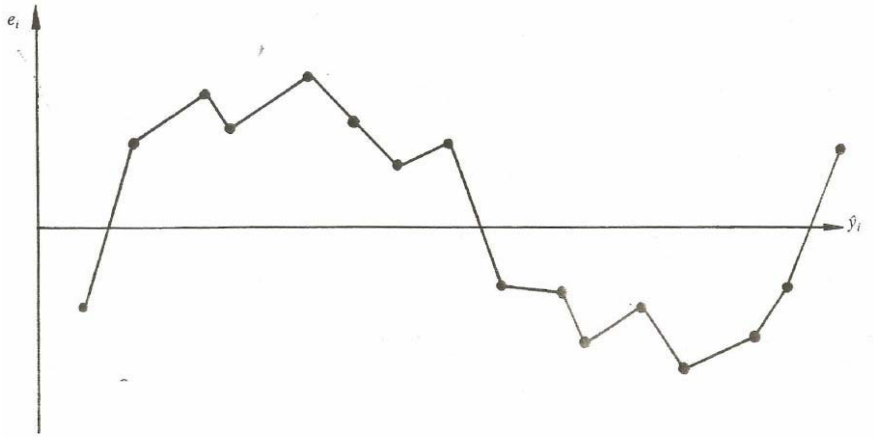


FIGURA 11.23. *Autocorrelación positiva entre los residuos.*

Observación influyente

Es una observación atípica (x_A,y_A) cuya exclusión produce un **cambio drástico** en la recta de regresión.

Posibles causas de una observación influyente

1) error de observación

2) modelo incorrecto:

- **no linealidad**: la relación entre x e y no es lineal cerca de x_A
- **heterocedasticidad**: la varianza aumenta mucho con x.
- variable desconocida que ha tomado un valor distinto en x_A

Figura 6.7 La influencia de un punto atípico en las x

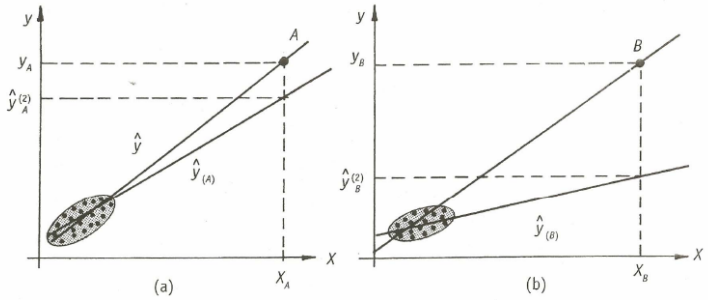
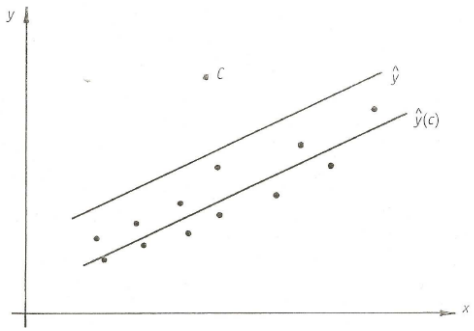


Figura 6.8 La influencia de un punto con respuesta y atípica



Resumen de posibles problemas a tener en cuenta

i. La función de regresión no es lineal.

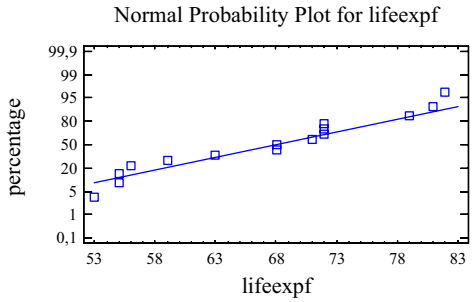
- *Diagrama de dispersión*:
  Observar si los datos se desvían de manera sistemática de la recta estimada, en particular si lo hacen sobre una trayectoria curvilínea.
- *Residuos vs. valores predichos o vs. variable independiente*.

Posible remedio (a nuestro alcance):

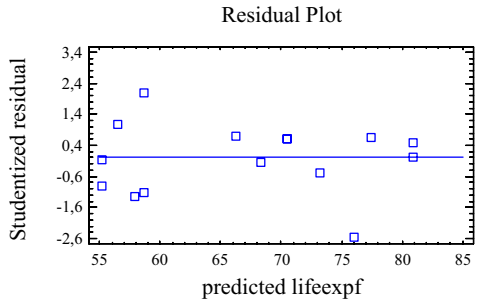
Por lo general, basta **transformar** solamente x.

ii. Los errores no siguen una ley Normal.

- Gráficos descriptivos de los residuos: Histograma, gráfico de caja,…
- Gráficos de Probabilidad Normal (*Normal Probability Plot*). Desviaciones respecto a las colas del gráfico indican posible sesgo o colas muy pesadas.



- Gráfico de residuos estandarizados vs. valores predichos: Si los errores son normales, los residuos estandarizados deberían encontrarse en su mayoría en la franja horizontal (-3, 3).



Posible remedio (a nuestro alcance): Por lo general, basta **transformar** solamente la variable y, puesto que es la distribución de la variable respuesta la que, por hipótesis, debe ser Normal.

iii. Los errores no tienen varianza constante.

- Residuos vs. valores predichos o vs. variable independiente.
  Buscar formas de trompeta en la nube de residuos.
- Residuos al cuadrado o en valor absoluto vs. valores predichos o vs. variable independiente:
  Puede ser más útil realizar este gráfico que el anterior, puesto que el signo de los residuos no tiene importancia y, en cambio, será más fácil de apreciar los cambios en la varianza.

Posible remedio (a nuestro alcance): Por lo general, basta **transformar** solamente la variable y, puesto que es la varianza de la variable respuesta la que, por hipótesis, debe ser constante.

iv. Los errores no son independientes.

- Residuos vs. secuencia temporal

Cuando los errores son independientes, los residuos deberían fluctuar en un patrón aleatorio alrededor del 0. Si se observa algún patrón definido, ello puede ser debido a que los errores están correlacionados.

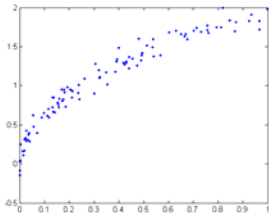
v. Se han omitido variables en el modelo

- Residuos vs. variables omitidas en el modelo.
- Dividir la muestra en grupos según los valores de alguna variable categórica, o variable categorizada, (sexo, edad, …) y examinar los residuos de cada grupo por separado.

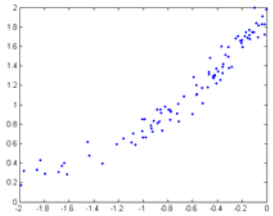
4. Transformaciones

Cuando el diagrama de dispersión entre las dos variables o el de los residuos presenta indicios de incumplimiento de alguna hipótesis básica, entonces hay que (i) **abandonar el modelo inicial por uno menos simple**, o bien (ii) **aplicar alguna transformación a los datos**.

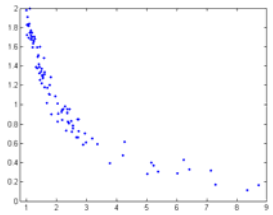
4.1 Corrección de la no linealidad



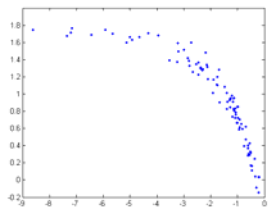
$x \to \ln x$ ó $x \to \sqrt{x}$



$x \to x^2$ ó $x \to e^x$



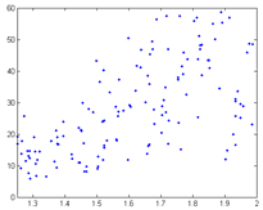
$x \to 1/x$



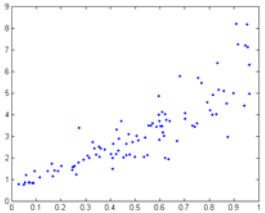
$x \to 1/\sqrt{x}$ ó $x \to e^{-x}$

4.2 Corrección de la no normalidad y la heterocedasticidad

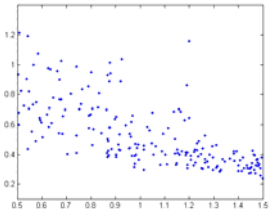
Empezaremos transformando solamente la variable y, y si no es suficiente, transformaremos ambas variables.



$y \to \sqrt{y}$ ó $\begin{cases} y \to \ln y \\ x \to \ln x \end{cases}$

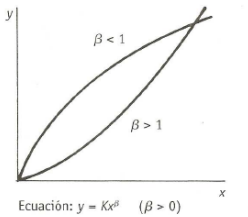


$y \to \ln y$



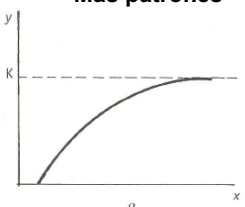
$y \to 1/y$

Más patrones



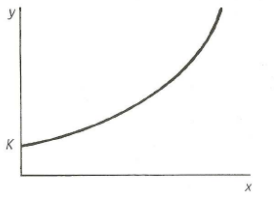
Ecuación: $y = Kx^{\beta}$ $(\beta > 0)$

Lineal con: Ln $y = \beta_0 + \beta \text{Ln}x$



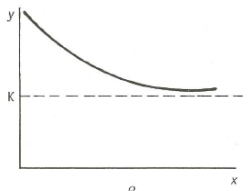
Ecuación: $y = K - \frac{\beta}{x}$ $(\beta > 0)$

Lineal con: $y = K - \beta x^{-1}$



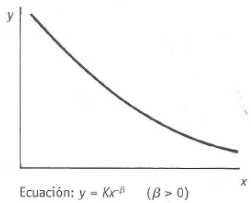
Ecuación: $y = Ke^{\beta x}$ $(\beta > 0)$

Lineal con: Ln $y = \beta_0 + \beta x$



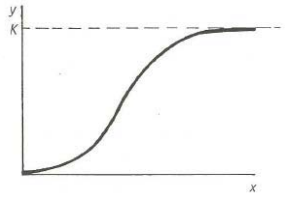
Ecuación: $y = K + \frac{\beta}{x_\beta}$ $(\beta > 0)$

Lineal con: $y = K + \beta x^{-1}$



Ecuación: $y = Kx^{-\beta}$ $(\beta > 0)$

Lineal con: Ln $y = \beta_0 - \beta \text{Ln}x$



Ecuación: $y = Ke^{-\beta/x}$

Lineal con: Ln $y = \beta_0 - \beta x^{-1}$

5. Predición

Un modelo de regresión permite:

1) **Estimar las medias condicionadas**: Estimar las medias de las distribuciones de y para cada valor de x,
2) **Realizar pronósticos**: Prever futuros valores de la variable respuesta.

Tanto la estimación de la media como la predicción de y se obtienen sustituyendo en la ecuación de la recta de regresión. Por tanto, **sus valores numéricos son idénticos**.

Sin embargo, la precisión de estas dos estimaciones es distinta y, por tanto, **los intervalos de confianza serán distintos**.

5.1. Estimación de las medias condicionadas (o variación esperada)

Modelo: $y_i=\beta_0+\beta_1 x_i + u_i$, donde u_i, $i=1,...,n$, son v.a. i.i.d.~$N(0,\sigma^2)$.

Vimos en el Tema 1 que: $E(y/x)= \beta_0+\beta_1 x$,

El nuevo parámetro que se quiere estimar es la media de y condicionada a que x tome un valor concreto x_h, es decir:

$$E(y/x=x_h)= \beta_0+\beta_1 x_h=m_h$$

El **estimador puntual** de m_h es:

$$\hat{y}_h = \hat{\beta}_{0+}\hat{\beta}_1 x_h ,$$

que tiene ley normal, al ser combinación lineal de normales.

Proposición: $\hat{y}_h = \hat{\beta}_{0+}\hat{\beta}_1 x_h$ es un estimador **insesgado** de m_h, cuya varianza es

$$\text{var}(\hat{y}_h)=\frac{\sigma^2}{\hat{n}_h}, \quad \text{donde} \quad \hat{n}_h=\frac{n}{1+\left(\frac{\bar{x}-x_h}{s_x}\right)^2}$$

(Demostración)

Observaciones:

- $\hat{n}_h$ se denomina **número equivalente de observaciones** para la estimación de m_h,
- $\left(\frac{\bar{x}-x_h}{s_x}\right)^2$ es el cuadrado de la **distancia de Mahalanobis** entre el punto x_h y la media de x,
- $\hat{n}_h$ depende de la distancia de Mahalanobis entre el punto x_h y la media de x, y determina la **precisión de la estimación**, es decir:
  Cuánto más cerca está x_h de la media de x (menor distancia de Mahalanobis) → mayor $\hat{n}_h$ → menor $\text{var}(\hat{y}_h)$ y, por tanto, mayor precisión.

Intervalo de confianza para m_h

Utilizaremos la ley de probabilidad de su estimador puntual $\hat{y}_h = \hat{\beta}_{0+}\hat{\beta}_1 x_h$.

Caso 1: σ^2 conocida,

$$\hat{y}_h \sim N\left(m_h, \frac{\sigma^2}{\hat{n}_h}\right) \Rightarrow \frac{\hat{y}_h - m_h}{\sigma/\sqrt{\hat{n}_h}} \sim N(0,1)$$

Caso 2: σ^2 desconocida,

$$\frac{\hat{y}_h - m_h}{s_R/\sqrt{\hat{n}_h}} \sim t_{n-2}$$

En general, se estará en la hipótesis del Caso 2, con lo que el I.C. para m_h al $(1\text{-}\alpha)100\%$ será:

$$I.C.(m_h)=\left[\hat{y}_h \mp t_{1-\frac{\alpha}{2}}\ s_R\sqrt{\frac{1}{\hat{n}_h}}\right]$$

donde $t_{1-\alpha/2}$ es el percentil $(1\text{-}\alpha/2)100\%$ de la ley t de Student con $n\text{-}2$ grados de libertad y

$$\hat{n}_h=\frac{n}{1+\left(\frac{x_h-\bar{x}}{s_x}\right)^2}$$

5.2 Predicción de una nueva observación (pronóstico)

Se quiere prever el **valor futuro** de la variable respuesta cuando $x=x_h$. Por tanto, se tiene una nueva v.a.

$y_h=\beta_0+\beta_1 x_h + u_h$, donde u_h~$N(0,\sigma^2)$ y es independiente de $u_1, u_2, ..., u_n$.

La predicción o pronóstico de la variable respuesta se obtiene sustituyendo en la ecuación de la recta de regresión:

$$\hat{y}_h = \hat{\beta}_{0+}\hat{\beta}_1 x_h$$

pero, ¿y si se quiere un **intervalo de predicción**?

Aclaración: El intervalo que se construye para los valores futuros de la variable respuesta, en realidad, **no es un intervalo de confianza** porque no se está estimando ningún parámetro. Se trata de un **intervalo de valores probables** para la variable aleatoria y_h, que se denomina **intervalo de predicción o de pronóstico**.

Este intervalo al $(1\text{-}\alpha)100\%$ es

$$y_h \in \left[\hat{y}_h \mp t_{1-\frac{\alpha}{2}}\ s_R\sqrt{1+\frac{1}{\hat{n}_h}}\right]$$

donde $t_{1-\alpha/2}$ es el percentil $(1\text{-}\alpha/2)100\%$ de la distribución t de Student con n-2 gr. lib. y

$$\hat{n}_h=\frac{n}{1+\left(\frac{x_h-\bar{x}}{s_x}\right)^2}$$

¿Cómo se deduce este intervalo?

Se considera la variable diferencia, es decir, el residuo $e_h = y_h - \hat{y}_h$, que tiene ley normal al ser combinación lineal de normales.

Se obtienen su esperanza y varianza:

$$E(y_h - \hat{y}_h) = E(\beta_{0+}\beta_1 x_h + u_h - \hat{y}_h) = \underbrace{\beta_{0+}\beta_1 x_h}_{=m_h} + \underbrace{E(u_h)}_{=0} - \underbrace{E(\hat{y}_h)}_{=m_h} = 0$$

$$\text{var}(y_h - \hat{y}_h) = \text{var}(y_h) + \text{var}(\hat{y}_h) = \sigma^2 + \frac{\sigma^2}{\hat{n}_h} = \sigma^2\left(1 + \frac{1}{\hat{n}_h}\right)$$

Observación: La varianza se obtiene como suma de varianzas puesto que $y_h, \hat{y}_h$ son independientes al serlo u_h de $u_1, u_2, \ldots, u_n$.

Por tanto, se tiene que:

$$y_h - \hat{y}_h \sim N\left(0, \sigma^2\left(1 + \frac{1}{\hat{n}_h}\right)\right) \Rightarrow \frac{y_h - \hat{y}_h}{\sigma\sqrt{1 + \frac{1}{\hat{n}_h}}} \sim N(0,1)$$

siempre que σ^2 sea conocida.

Si σ^2 es desconocida, entonces se sustituye por su estimador, y entonces:

$$\frac{y_h - \hat{y}_h}{s_R\sqrt{1 + \frac{1}{\hat{n}_h}}} \sim t_{n-2}$$

Puesto que esta será la situación habitual, el intervalo de predicción para la respuesta concreta y_h es:

$$y_h \in \left[\hat{y}_h \mp t_{1-\frac{\alpha}{2}}\ s_R\sqrt{1 + \frac{1}{\hat{n}_h}}\right]$$

Observaciones:

- El intervalo de predicción es mucho más ancho que el $I.C.(m_h)$, porque se añade la variabilidad de una observación alrededor de la media.
- En ambos intervalos, la amplitud aumenta cuando aumenta la distancia de Mahalanobis entre x_h y la media de x.
- Ambas amplitudes son mínimas cuando $x_h = \bar{x}$.

5.3 Bandas de confianza

Las **bandas de confianza para la predicción de las medias condicionadas** (respuesta media) se obtienen uniendo los extremos de los intervalos construidos para el mismo nivel de confianza 1-α, y para cada valor de x.

Análogamente, también pueden construirse Las **bandas de confianza para los pronósticos** (futuros valores de y)

Figura 6.35 Bandas de confianza para la predicción