TEMA 15

REGRESIÓN LINEAL.

ESTIMACIÓN Y CONTRASTES

CONTRATE PARA LA INDEPENDENCIA LINEAL

ESTIMACIÓN DE LOS PARÁMETROS DEL MODELO

CONTRASTE DEL COEFICIENTE DE REGRESIÓN (LA PENDIENTE DE LA RECTA)

INTERVALO DE CONFIANZA PARA EL COEFICIENTE DE REGRESIÓN.

INTERVALO DE CONFIANZA DE LAS MEDIAS CONDICIONADAS.

INTERVALO DE CONFIANZA DE LA PREDICCIÓN

CORRELACIÓN Y REGRESIÓN LINEAL

Introducción.

Como ya se estudió en el tema 4 de la asignatura, la correlación y la regresión lineal son técnicas estadísticas para estudiar la (posible) relación existente entre dos o más variables continuas. Más precisamente:

- La correlación pretende evaluar si existe asociación lineal entre dos variables. Tal asociación se da si los valores de una variable cambian proporcionalmente a los valores de la otra. La intensidad de esta asociación se mide mediante el *coeficiente de correlación lineal.*
- La regresión simple, supuesto que existe asociación lineal entre dos variables, una de las cuales es la *variable tratamiento X* y la otra es la *variable respuesta Y*, tiene como objetivo es encontrar una ecuación (*ecuación de regresión lineal*) que, de modo al menos aproximado, permita predecir el valor de *Y* una vez que se conoce el valor de *X*.

Para llevar a cabo un estudio de correlación o regresión lineal en la práctica, es preciso disponer de una muestra de n sujetos experimentales. Al sujeto i-ésimo de esta muestra se le aplica un valor x_i de la variable tratamiento X (medida en escala continua), y se le mide el valor y_i que se obtiene en la variable respuesta Y (también medida en escala continua), supuestamente asociada con X. De esta forma, el punto de partida para nuestro estudio es una tabla de datos como la siguiente:

	X	**Y**
Sujeto 1	x_1	y_1
...		
Sujeto i	x_i	y_i
....	...	...
Sujeto n	x_n	y_n

El hecho de que la relación entre las variables *X e Y* sea lineal se expresa matemáticamente de la forma: $y_i = \alpha + \beta \cdot x_i + \varepsilon_i$ con $i=1,2,...,n$, siendo ε_i variables aleatorias independientes con distribución de probabilidad común $N(0,\sigma)$. Este es el llamado *modelo de regresión lineal simple*. La variable X recibe el nombre de *variable independiente, variable explicativa o regresor*, mientras que la variable Y recibe el nombre de *variable dependiente o variable respuesta.*

Geométricamente, el modelo lineal significa que si se dibujan los n puntos (x_i,y_i) en un gráfico, estos puntos deben estar dispersos a lo largo de una recta no horizontal de ecuación $y_i=\alpha+\beta x_i$. Las variables ε_i miden precisamente la distancia de cada punto a la recta, y representan los *errores* (también llamados *residuos*) en la predicción de Y a partir de la X utilizando este modelo. Estos errores se deben a que, en la práctica, la recta es una simple aproximación de la *verdadera* relación existente entre X e Y. En fenómenos complejos, como los que se estudian en Medicina, Física, Informática,…, lo habitual es que en la relación entre la X y la Y intervengan otras muchas variables no observadas, aunque se espera que cada una de ellas tenga poca influencia relativa sobre esta relación.

De acuerdo con el teorema central del límite, en estas condiciones es esperable que el efecto global de todas estas variables (que es lo que da lugar a los residuos) se manifieste como una variable aleatoria con distribución normal. De hecho, cuando se detecta falta de normalidad en los residuos ello puede ser señal de la existencia de alguna variable que no ha sido tenida en cuenta y que tiene un efecto importante sobre la relación entre X e Y.

Correlación.

Como ya estudiamos en el tema 4, el *coeficiente de correlación lineal de Pearson* mide la intensidad de la asociación lineal entre dos variables. De un modo más gráfico, este coeficiente mide lo bien (o mal) que se ajustan los puntos observados a una recta. Dada una muestra de n observaciones (x_i, y_i) se define el coeficiente de correlación lineal de esa muestra como:

$$r = \frac{S_{XY}}{S_X S_Y}$$

donde: $S_X^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$ y $S_Y^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2$ son las varianzas respectivas de las variables X e Y, y $S_{XY} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \frac{1}{n}\sum_{i=1}^{n}x_i y_i - \overline{xy}$ es la covarianza entre X e Y.

De acuerdo con esta definición el coeficiente de correlación es *adimensional* (no depende de las unidades de medida). Además, puede probarse que este coeficiente cumple la siguiente importante propiedad: $-1 \le r \le 1$

- Cuando $r>0$, entre las dos variables existe una asociación lineal positiva, esto es, cuando el valor de una aumenta, el valor de la otra aumenta proporcionalmente.
- Cuando $r<0$, entre las dos variables existe una asociación lineal negativa: cuando el valor de una aumenta, el valor de la otra disminuye proporcionalmente.
- El valor $r=0$ corresponde a la ausencia de asociación lineal entre las variables. La nube de puntos se distribuye a lo largo de una recta horizontal, o bien se distribuye a lo largo de otra figura geométrica distinta de la recta.
- En general, cuanto más próximo a *0* es el coeficiente de correlación, menor asociación lineal hay en los datos; cuanto más se aproxima a *1* ó a *-1* mayor es la asociación lineal (en sentido positivo o negativo, respectivamente).

En el tema 4 indicamos además criterios a tener en cuenta ***cuándo no usar el coeficiente de correlación*** (pág.70).

Contrate para la independencia lineal

El coeficiente de correlación muestral r sirve no solo para medir la intensidad de la asociación lineal entre las variables observadas, sino también para contrastar la hipótesis de si el coeficiente de correlación lineal entre esas variables *en la población* es nulo. Si llamamos ρ a este coeficiente en la población, este contraste se plantea como:

$$H_0: \rho = 0$$

frente a la hipótesis alternativa: $H_1{:}\rho \neq 0$

Teniendo en cuenta que bajo la hipótesis nula de independencia lineal, el estadístico:

$$t = \sqrt{\frac{(n-2)\hat{\rho}^2}{1-\hat{\rho}^2}}$$

tiene una distribución $t(n\text{-}2)$ (T-student con n-2 grados de libertad).

Se construye el siguiente test de hipótesis:

Aceptar H_0 si $|t| \leq t_{n-2,\ \alpha/2}$ y rechazar en caso contrario.

- Aceptar la hipótesis nula H_0 en este contexto significa que el coeficiente de correlación muestral r no es lo suficientemente grande en valor absoluto como para concluir que en la población de que se ha extraído la muestra exista una asociación lineal significativa entre las variables X e Y.

- En términos de p-valor para el contraste $H_0{:}\rho=0$ frente a $H_1{:}\rho\neq 0$. Si α es el nivel de significación con que se desea realizar el contraste, la decisión se toma de acuerdo con el siguiente criterio:

Si p-valor $\geq \alpha/2 \Rightarrow$ *Aceptar* H_0
Si p-valor $< \alpha/2 \Rightarrow$ *Rechazar* H_0

Regresión simple.

Como ya hemos dicho, el modelo de regresión lineal simple se enuncia en la forma:

$$y_i = \alpha + \beta \cdot x_i + \varepsilon_i \qquad i=1,2,...,n$$

siendo ε_i variables aleatorias independientes con distribución de probabilidad común $N(0,\sigma)$. El coeficiente α recibe el nombre de *ordenada en el origen* y el coeficiente β el nombre de *coeficiente de regresión.* La ecuación $y=\alpha+\beta x$ es la ecuación de una recta en la que α representa el valor de y cuando $x=0$ y β representa el cambio que se produce en el valor de y cuando x se incrementa en una unidad.

Estimación de los parámetros del modelo

De acuerdo con la descripción que se acaba de dar, los parámetros poblacionales del modelo que deben estimarse a partir de los datos muestrales son α, β y σ. En el tema 4 de la asignatura utilizamos el método de mínimos cuadrados, para estimar dichos parámetros obteniéndose los estimadores puntuales de α y β:

$$\hat{\beta} = \frac{S_{XY}}{S_X^2} = b \qquad\qquad \hat{\alpha} = \overline{y} - \hat{\beta} \cdot \overline{x} = a$$

(Estimadores centrados para los correspondientes parámetros, es decir, $E[\hat{\beta}] = \beta$ y $E[\hat{\alpha}] = \alpha$).

Para cada valor de la variable independiente *X*, x_i, se define la predicción de *Y* por:

$$\hat{y}_i = \hat{\alpha} + \hat{\beta} \cdot x_i = \mathrm{E}\,[y \mid x_i]$$

Se tiene entonces la siguiente estimación centrada de la varianza de los errores:

$$\hat{\sigma}^2 = S^2_{\text{Residual}} = \frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \frac{1}{n-2}\sum_{i=1}^{n}(e_i)^2 = S_e^2$$

siendo S_e^2 es un estimador centrado para σ^2 y recibe el nombre de varianza residual. La varianza residual mide la variabilidad de los valores de *y* con respecto a la recta de regresión. Es, por tanto, una medida de la variabilidad de *Y* que no puede explicarse por su relación con *X*.

<u>Contraste de la regresión</u>

Si el parámetro β es cero, significa que los diferentes valores de la variable independiente *X* no cambian el valor de la variable dependiente *Y*. En este caso se dice que no hay regresión lineal. En este sentido, formulamos el siguiente contraste de hipótesis llamado *contraste de la regresión:*

$$H_0 : \beta = 0$$

frente a la hipótesis alternativa:

$$H_1 : \beta \neq 0$$

- Para este caso y bajo la hipótesis nula, se construye el estadístico

$$T_{\text{Observado}} = \frac{\hat{\beta} - \beta_0}{\hat{\sigma}_{\hat{\beta}}} \sim t(n-2)\,; \qquad \hat{\sigma}_{\hat{\beta}} = \frac{\hat{\sigma}}{\sqrt{n}\,S_x} = \frac{S_e^2}{\sqrt{n}\,S_x}$$

De modo que si $T_{observado} \leq t_{n-2,\alpha} \Rightarrow$ se acepta H_0 y se rechazará en caso contrario

- Otra forma de realizar el contraste (véase págs. 277-278) sería calculando:

$$F_{observado} = \frac{R^2(n-2)}{1-R^2} = \frac{\sum_{i=1}^{n}(\hat{y}_i - \overline{y})^2}{\frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} = (n\text{-}2)\cdot SC_{\text{Regresión}} / SC_{\text{Residual}},$$

entonces si H_0 es cierta $F_{\text{observado}}$ sigue una distribución F de Fisher-Snedecor con 1 y n-2 grados de libertad. (siendo α = nivel de significación para el contraste). En este caso la regla de decisión sería:

- Si $F_{observado} \leq F_{1,n-2,\alpha} \Rightarrow$ se acepta $H_0 : \beta=0$
- Si $F_{observado} > F_{1,n-2,\alpha} \Rightarrow$ se rechaza $H_0 : \beta=0$

Obsérvese que el error de tipo I de este test de hipótesis es:

$$P\left(F > F_{1,n-2,\alpha} \mid H_0 \ \ es\ cierta\right) = \alpha$$

- En términos de p-valor al realizar el contraste $H_0 : \beta=0$ frente a $H_1 : \beta\neq0$ se considera el criterio:

Si p-valor $\geq \alpha \Rightarrow$ Aceptar H_0
Si p-valor $< \alpha \Rightarrow$ Rechazar H_0

Intervalo de confianza para el coeficiente de regresión.

En las hipótesis del modelo, el valor de $\hat{\beta}$ es una variable aleatoria con distribución normal de media β y varianza $\sigma_\beta^2 = \dfrac{\sigma^2}{nS_x^2}$ (es decir, $\hat{\beta} \sim N(\beta, \sigma_\beta^2)$). Ello permite construir un intervalo de confianza a nivel $1-\alpha$ para β de la siguiente forma:

$$\left[\hat{\beta} \pm t_{n-2,\alpha/2} \cdot \frac{\hat{\sigma}}{\sqrt{n}\, S_x}\right]$$

donde $\hat{\sigma}^2 = S^2_{\text{Residual}} = S_e^2$ y siendo $t_{n-2,\alpha/2}$ el cuantil 1-α/2 de la distribución t de Student con n-2 grados de libertad. Si este intervalo contuviese al valor *0*, entonces β podría ser *0*, ó lo que es lo mismo, no hay evidencia de que $\beta\neq0$. Ello significa que el contraste:

$H_0 : \beta=0$ frente a $H_1 : \beta\neq0$

podría resolverse también del siguiente modo, definiendo el estadístico:

$$t = \hat{\beta}\frac{\sqrt{n}S_x}{\hat{\sigma}}$$

Aceptar H_0 si $t \leq t_{n-2,\alpha/2}$, y rechazar H_0 en caso contrario.

Este test de hipótesis es equivalente al anterior.

Coeficiente de determinación (R²)

A partir de la ecuación de regresión puede comprobarse que:

$$\sum_{i=1}^{n}\left(y_i-\overline{y}\right)^2 = \sum_{i=1}^{n}\left(y_i-\hat{y}_i\right)^2 + \sum_{i=1}^{n}\left(\hat{y}_i-\overline{y}\right)^2$$
$$SC_{TOTAL} = SC_{RESIDUAL} + SC_{REGRESION}$$

En esta igualdad:

- **SC$_{\mathbf{TOTAL}}$ = *SUMA DE CUADRADOS TOTAL:*** *representa la variabilidad total de la variable Y. Esta cantidad es, de hecho, el numerador de la varianza de Y, y por tanto es efectivamente una medida de la variabilidad presente en dicha variable.*
- **SC$_{\mathbf{RESIDUOS}}$ = *SUMA DE CUADRADOS RESIDUAL:*** *es la suma de los cuadrados de las distancias entre los valores observados y predichos de la variable Y. Representa por tanto una medida del grado de desajuste entre los puntos observados y la recta que más se aproxima a ellos.*
- **SC$_{\mathbf{REGRESIÓN}}$ = *SUMA DE CUADRADOS EXPLICADA POR LA REGRESIÓN:*** *es la suma de los cuadrados de las distancias entre las predicciones y la media de Y. Representa la variabilidad en Y debida a (o explicada por) la presencia de la relación lineal entre X e Y (si no existiera esta relación lineal entonces* $\hat{y}_i = \overline{y} \ \ \forall i$ *y sería* $SC_E = 0$*).*

Una forma de medir la bondad de ajuste de los datos experimentales al modelo lineal teórico es a través del coeficiente de determinación, que se define de la siguiente manera:

$$R^2 = \sum_{i=1}^{n}(\hat{y}_i-\overline{y})^2 / \sum_{i=1}^{n}(y_i-\overline{y})^2 = SC_{\text{Regresión}} / SC_{\text{Total}}$$

Este coeficiente, de acuerdo con el análisis de la varianza en la regresión, evalúa la proporción de variabilidad explicada por el modelo. Obviamente toma valores en el intervalo [0,1]. Un coeficiente de determinación igual a uno indica que los datos experimentales están en la recta de regresión (ajuste total), mientras que un coeficiente nulo indica que no hay regresión lineal. Esta última afirmación se contrasta mediante el contraste de la regresión (pág. 415), para lo cual se puede utilizar la siguiente tabla ANOVA para la regresión, que aportan como salida la mayoría de los software de estadística y que se corresponde con el método de contraste indicado en la página 275

- Si $F_{observado} \leq F_{1,n-2,\alpha} \Rightarrow$ se acepta $H_0 : \beta=0$, en caso contrario se acepta, siendo:

$$F_{observado} = \frac{R^2(n-2)}{1-R^2} = MC_{\text{Regresión}} / MC_{\text{Residual}},$$

Tabla ANOVA para el Análisis de Regresión:

Fuente	S. Cuadrados	Grados libertad	Media de Cuadrados	Estadístico
Regresión	$SC_{Regr} = ns_X^2 \widehat{\beta}_1^2$	1	$MC_{Regr} = \frac{SC_{Regr}}{1}$	$\frac{MC_{Regr}}{MC_{Resid}}$
Residuos	$SC_{Resid} = ns_Y^2 \left(1 - \frac{s_{XY}^2}{s_X^2 s_Y^2}\right)$	$n-2$	$MC_{Resid} = \frac{SC_{Resed}}{n-2}$	
Total corregida	$SC_{Total} = SC_{Regr} + SC_{Resid}$	$n-1$		

siendo $\widehat{\beta}_1 = \frac{s_{XY}}{s_X^2}$, $\widehat{\beta}_0 = \overline{Y} - \widehat{\beta}_1 \overline{X}$, $\widehat{Y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 X_i$, el coeficiente de determinación

$$R^2 = \frac{SC_{Regr}}{SC_{Total}} = \frac{\sum \left(\widehat{Y}_i - \overline{Y}\right)^2}{\sum \left(Y_i - \overline{Y}\right)^2} = \frac{s_{XY}^2}{s_X^2 s_Y^2}$$

y la varianza residual

$$s_r^2 = \frac{n}{n-2} s_Y^2 \left(1 - \frac{s_{XY}^2}{s_X^2 s_Y^2}\right)$$

Estimación de las medias condicionadas.

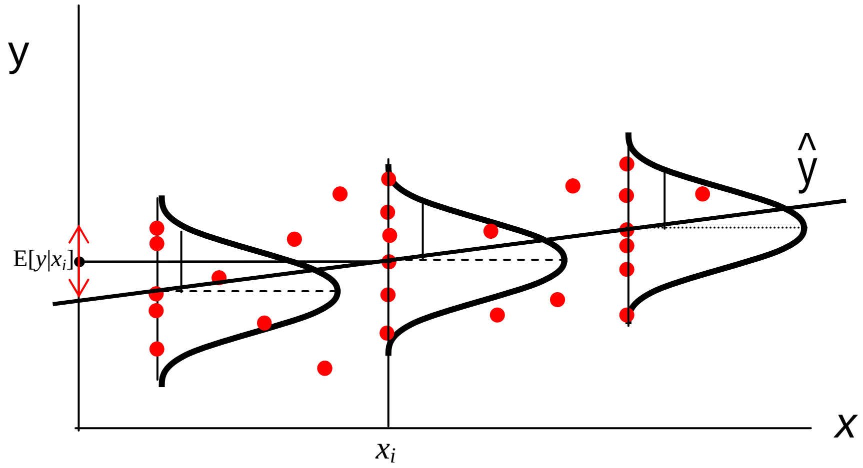
Una vez que se ha estimado la recta de regresión, puede utilizarse para estimar el valor medio de Y que corresponde a un valor $X=x_i$ concreto. Si, por ejemplo, X fuese la edad e Y fuese la tensión arterial, podría usarse la recta de regresión para estimar *el valor medio de la tensión arterial en personas de una edad x_i determinada*. Este valor recibe el nombre de *media de Y condicionada por $X=x_i$*, $E[y \mid x_i]$. Esta estimación se obtiene simplemente sustituyendo el valor de x_i en la ecuación de regresión:

$$E[y \mid X = x_i] = \hat{y} = \hat{\alpha} + \hat{\beta} x_i = \overline{y} + \frac{S_{XY}}{S_X^2}(x_i - \overline{x})$$

El intervalo de confianza a nivel 1-α para esta estimación viene dado por:

$$\left[\hat{y} \pm t_{n-2,\alpha/2} \cdot S_e \sqrt{\frac{1}{n} + \frac{(x_i - \overline{x})^2}{nS_X^2}}\right]$$

siendo $t_{n-2,\alpha/2}$ el cuantil 1-α/2 de la distribución t de Student con n-2 grados de libertad.



Intervalo de confianza de la Predicción

La recta de regresión puede utilizarse también para la predicción de una observación individual de la variable respuesta Y en función de X. Si, como en el apartado anterior, X fuese la edad e Y la tensión arterial, la recta permitiría predecir cuál es el valor de la tensión arterial de *una persona* de edad x_i. La predicción individual del modelo para $X=x_i$ es:

$$\hat{y}_i = \hat{\alpha} + \hat{\beta}\, x_i = \overline{y} + \frac{S_{XY}}{S_X^2}(x_i - \overline{x})$$

que como se aprecia es análoga a la estimación de la media condicionada para ese valor x_i. El intervalo de confianza al nivel 1-α viene dado en este caso por:

$$\left[\hat{y}_i \pm t_{n-2,\alpha/2} \cdot S_e \sqrt{1 + \frac{1}{n} + \frac{(x_i - \overline{x})^2}{n S_X^2}} \right]$$

Obsérvese que el intervalo de confianza para la predicción individual es más amplio que para la estimación de la media condicionada. Ello es así pues se espera menor variabilidad en valores medios que en valores individuales.

Para la validez del contraste de regresión y de los intervalos de confianza que se obtienen para los parámetros, predicciones, etc., es preciso que se cumplan las condiciones que se especifican a continuación:

(a) La relación es aproximadamente lineal. Esto se comprueba en la mayoría de los casos simplemente representando gráficamente X frente a Y. También resulta útil representar los residuos de la regresión: $e_i = y_i - \hat{y}_i = y_i - \hat{\alpha} - \hat{\beta}\, x_i$ frente a los valores x_i. Si se aprecia alguna relación entre los e_i y los x_i ello indica la posible existencia de relación no lineal entre X e Y.

(b) La relación es homocedástica, esto es, la varianza de los residuos se mantiene constante. Ello significa que los errores de predicción o residuos e_i no están relacionados con los valores predichos y_i. Esto se comprueba trazando un gráfico de e_i frente a y_i. Si, por ejemplo, a medida que aumentan los valores y_i aumentan también los e_i ello indica que esta condición no se verifica. Esta situación puede habitualmente remediarse realizando una transformación de la variable Y y recalculando la recta de regresión y los residuos. Una transformación que habitualmente suele dar buen resultado es la transformación $z=ln(y)$.

(c) Los residuos de la regresión siguen una distribución normal. Esta hipótesis puede verificarse mediante un contraste adecuado (Kolmogorov-Smirnov, Bonferroni) o, gráficamente, representando los residuos en papel probabilístico normal. En este caso los residuos deben aparecer dispuestos a lo largo de una recta. Fuertes desviaciones de esta recta indican falta de normalidad en los residuos. La transformación logarítmica citada en el apartado anterior puede también remediar en algunos casos la falta de normalidad en los residuos.

(d) Los residuos son mutuamente independientes. En el caso de que los datos utilizados para ajustar la regresión correspondan a medidas únicas realizadas sobre individuos separados, normalmente no hay razón para suponer que se viole esta hipótesis de independencia, pues la medida tomada sobre un individuo no tiene por qué afectar ni verse afectada por medidas tomadas sobre otros individuos. Sí que puede haber problemas con la independencia de los residuos en los casos siguientes:

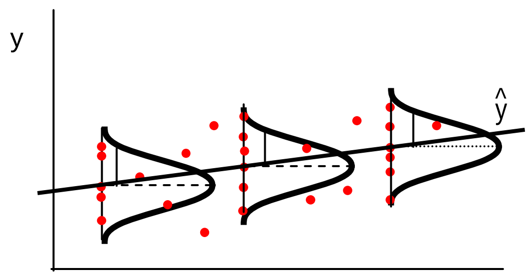
- Las observaciones corresponden a datos ordenados en el tiempo, en cuyo caso puede haber tendencia a que residuos negativos sean seguidos de residuos negativos y residuos positivos sean seguidos por residuos positivos.
- Se han tomado varias observaciones sobre un mismo individuo, en cuyo caso los residuos correspondientes a estos datos pueden estar relacionados. Una posible solución en este caso es promediar los valores obtenidos para un mismo individuo y tratar este promedio como una observación única.

Consecuencias del incumplimiento de las hipótesis.

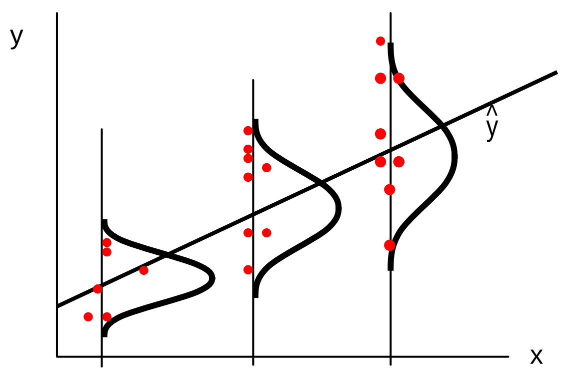
La falta de normalidad o de homoscedasticidad, si no son muy graves, no afectan seriamente a las estimaciones de los parámetros de la regresión, aunque sí que afectan a los intervalos de confianza y al contraste de regresión. Particularmente este último es poco fiable sobre todo en aquellos casos en que el estadístico de contraste se encuentra próximo al valor crítico del contraste.

La falta de linealidad indica que el modelo de regresión no es adecuado para el análisis de los datos y que deben considerarse otros métodos (regresión no lineal, regresión múltiple).

Por último, la falta de independencia de los residuos constituye un problema grave pues serían inválidas las estimaciones así como los contrastes e intervalos de confianza. En caso de dependencia temporal, deben analizarse los datos mediante modelos de series temporales. En caso de varias medidas sobre el mismo individuo, deben utilizarse métodos diseñados para el análisis de datos con medidas repetidas.



Varianza constante (homocedasticidad)



Varianza no constante (heterocedasticidad)