LABORATORIO 2

Regresiones: conceptos básicos

-

Incerteza y error

- La **exactitud** (*accuracy*) se refiere a la proximidad entre el valor medido y el valor verdadero o aceptado de una magnitud. El **error** representa la inexactitud de una medición.
- La **precisión** (*precision*) es una medida de qué tan bien se puede determinar una medición (independientemente que permita o no acercarse al valor verdadero de la magnitud). Es el grado de consistencia y concordancia entre mediciones independientes de la misma cantidad; también la fiabilidad o reproducibilidad del resultado.
- La **incerteza** (*uncertainty*) asociada a una medición tiene en cuenta tanto la exactitud como la precisión de la medición.

Incerteza en una sola medición

Toda medición tiene una incerteza asociada, cualquiera sea la precisión del instrumento de medición.

La incerteza de una sola medición está limitada por la precisión y exactitud del instrumento de medición y del método utilizado.

No existe una regla general para determinar la incerteza en todas las mediciones posibles.

El experimentador es quien mejor puede evaluar y cuantificar la incerteza de una medición en función de todos los posibles factores que afectan el resultado. Por lo tanto, la persona que realiza la medición tiene la obligación de emitir el mejor juicio posible e informar la incerteza de una manera que explique claramente qué es lo que representa.

En general, la incerteza debe indicar aproximadamente un intervalo de confianza del 68%.

Incerteza en mediciones repetidas

El promedio es generalmente la mejor estimación del valor "verdadero" de una magnitud.
Generalmente, cuantas más mediciones se hagan, mejor será esta estimación, pero no tiene sentido hacer más mediciones que las necesarias para la precisión requerida y la exactitud del método e instrumental.

Valor medio

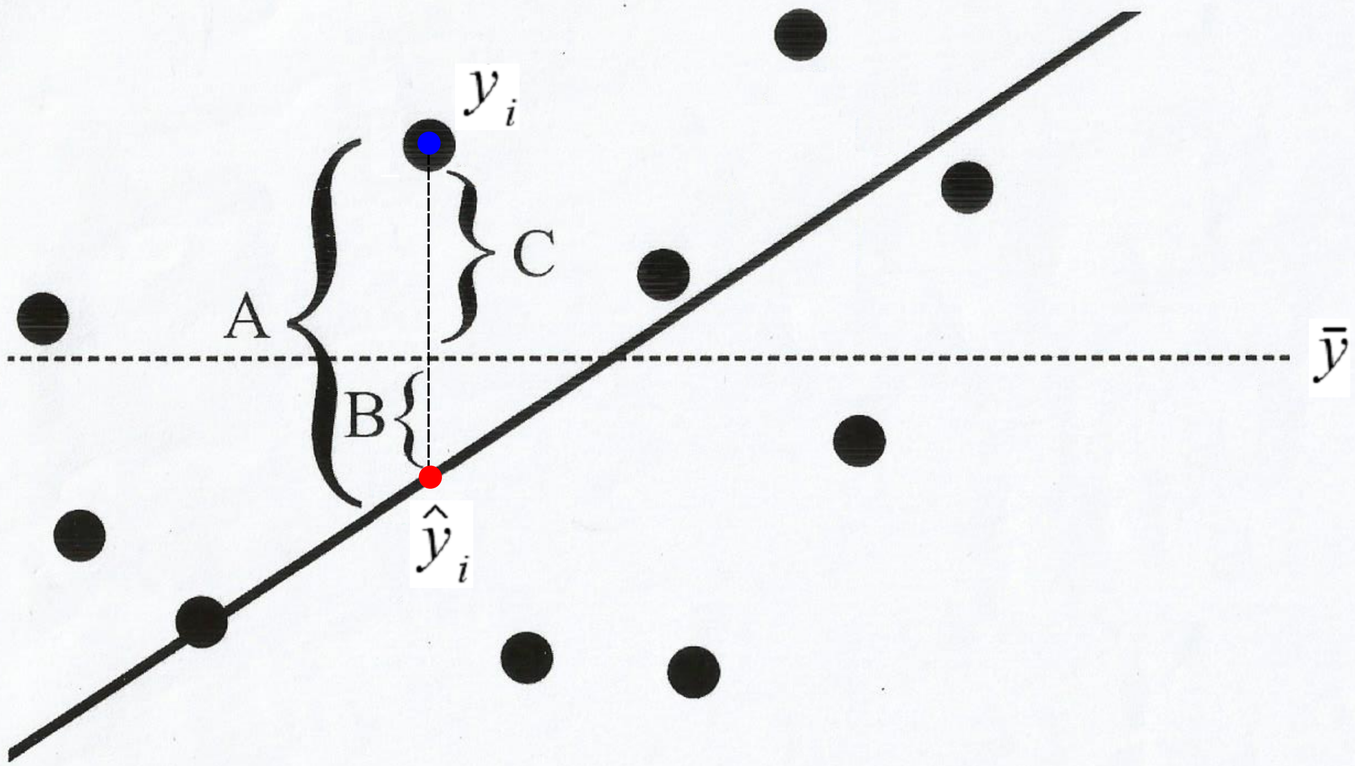
$$\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$$

Desviación estandar

$$\sigma = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})^2}$$

Error del promedio

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{N}}$$



A = variación total en y
B = variación explicada en y
C = residuo no explicado en y

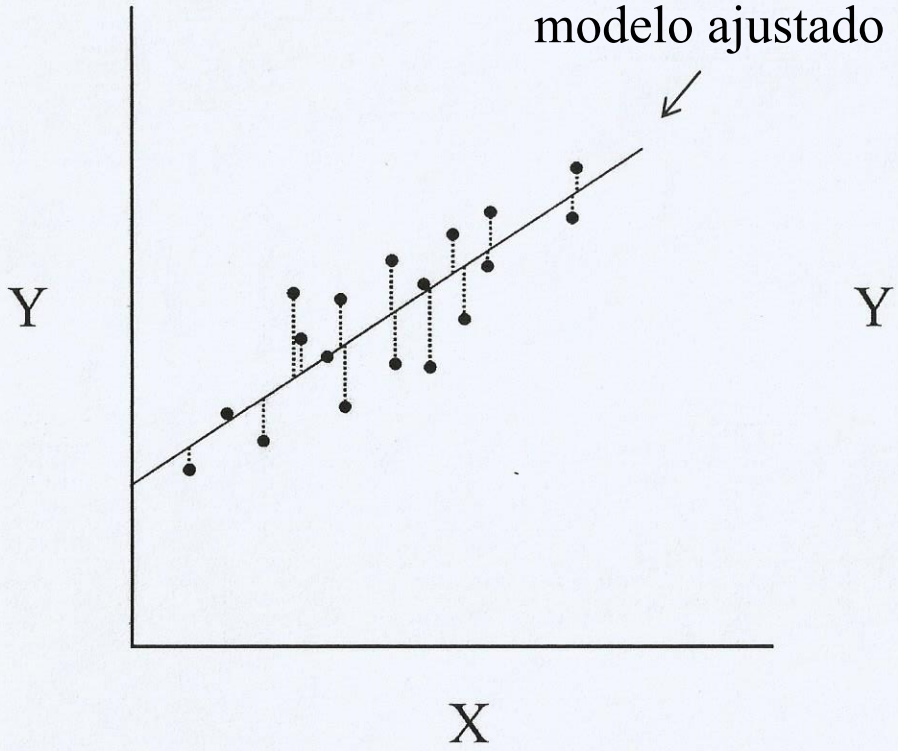
(A) Residual sum of squares: $RSS = \Sigma\left(y_i - \hat{y}_i\right)^2 = SSE$

(B) Explained sum of squares: $ESS = \Sigma\left(\hat{y}_i - \bar{y}\right)^2 = SSR$

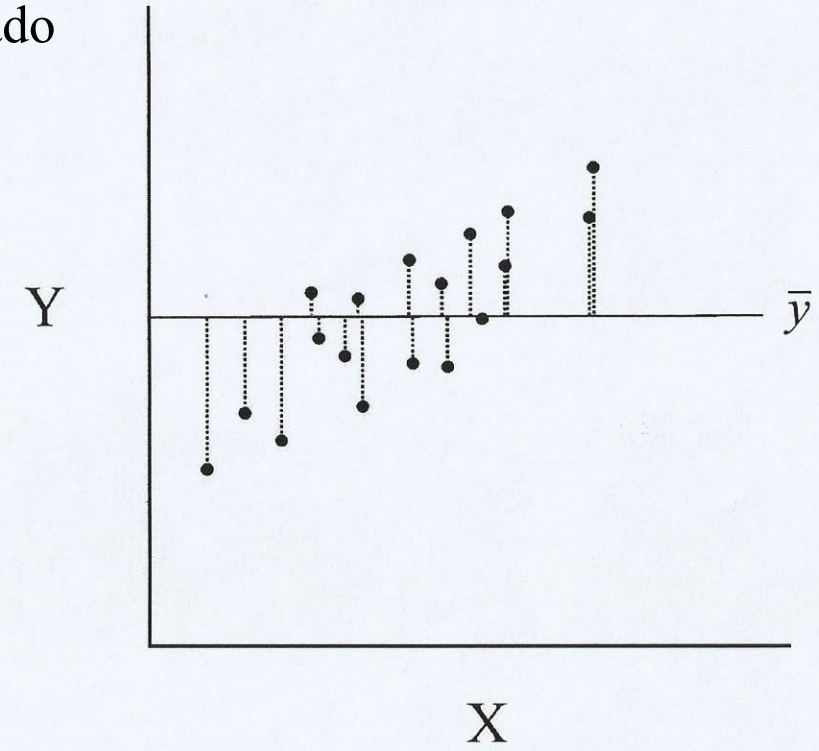
(C) Total sum of squares: $TSS = \Sigma\left(y_i - \bar{y}\right)^2 = SST$

Partición de la suma de cuadrados: $TSS = ESS + RSS$

RSS: Suma de cuadrados de residuos



TSS: Suma de cuadrados total



$$\sum_{i=1}^{n}(y_i-\bar{y})^2=\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2+\sum_{i=1}^{n}(y_i-\hat{y}_i)^2$$

TSS: suma de cuadrados total
(o *SST*)

ESS: suma de cuadrados explicados
(o *SSR*, suma de cuadrados de la regresión)

RSS: suma de cuadrados de residuos
(o *SSE*, suma de cuadrados de los errores,
no explicados por el modelo)

Vale para regresiones lineales, o en general, para variables con errores gaussianos y con parámetros independientes siempre que los residuos sean también independientes.

Cuadrado mínimos

Sea y_i los valores medidos de una variable que se distribuyen normalmente alrededor de los valores del modelo $\hat{y}_i$. La probabilidad de haber obtenido cada uno de esos valores es

$$dP_i = \frac{1}{\sqrt{2\pi}\,\sigma}\exp\left(-\frac{(y_i-\hat{y}_i)^2}{2\sigma^2}\right)dy$$

siendo σ la dispersión de la distribución. Si las mediciones son independientes, la probabilidad de haber obtenido ese conjunto de valores es

$$dP = dP_1 dP_2 \ldots dP_n$$

$$dP = \left(\frac{dy}{\sqrt{2\pi}\,\sigma}\right)^n \prod_{i=1}^{n}\left[\exp\left(-\frac{(y_i-\hat{y}_i)^2}{2\sigma^2}\right)\right] = \left(\frac{dy}{\sqrt{2\pi}\,\sigma}\right)^n \exp\left(-\sum_{i=1}^{n}\frac{(y_i-\hat{y}_i)^2}{2\sigma^2}\right)$$

Utilizando el principio de máxima probabilidad (básicamente que lo que se observó fue lo más probable) hay que pedir dP máximo, lo que implica que el exponente debe ser mínimo

$$\sum_{i=1}^{n}\frac{(y_i-\hat{y}_i)^2}{2\sigma^2} = \text{mínimo}$$

Si la dispersión es constante se tiene

$$RSS = \sum_{i=1}^{n}(y_i-\hat{y}_i)^2 = \text{mínimo}$$

sino se utiliza los cuadrados pesados con la dispersión.

Teorema del Límite Central o Teorema Central del Límite

El error aleatorio en la medición de una magnitud en general proviene de numerosas variaciones debidas a muchos factores. Es decir, el error no es debido a una sola causa.
Entonces, la magnitud a medir puede considerarse como una variable aleatoria que resulta de la suma de varias variables aleatorias.
El Teorema del Límite Central dice que en el límite cuando la cantidad de variables aleatorias tiende a infinito entonces la suma de ellas seguirá una distribución normal. Vale también para variables aleatorias discretas. Por ejemplo, el caso de una moneda que sigue una distribución binomial con $p = 0.5$ (cuando está balanceada) al arrojarla repetidamente la probabilidad de obtener k caras en n tiros tiende a una gaussiana cuando n tiende a infinito. Lo mismo la suma de los valores de un dado arrojado n veces.
En la práctica, muchas veces se obtiene una razonable distribución gaussiana con menos de la suma de 10 variables aleatorias arbitrarias.

La relación entre la variación explicada, *ESS*, y la total, *TSS*, en la variable *y*, se llama coeficiente de correlación, R^2 (para regresiones lineales se escribe en minúsculas, r^2)

$$R^2 = \frac{\overset{ESS}{\overset{\downarrow}{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}}}{\underset{TSS}{\underset{\uparrow}{\sum_{i=1}^{n}(y_i - \bar{y})^2}}}$$

$$R^2 = \frac{ESS}{TSS} = 1 - \frac{RSS}{TSS}$$

Minimizando *RSS* se obtiene R^2 máximo

R^2 comparado con $\overline{R}^2$ ajustado

El coeficiente de regresión R^2 indica la proporción de variación de la variable y que es debida a una variación de las variables x.

Hay quienes, en cambio, prefieren usar el $\overline{R}^2$ ajustado, que es una "penalización" para modelos que tienen muchos parámetros.

Partiendo de

$$R^2 = 1 - \frac{RSS}{TSS}$$

$$1 - R^2 = \frac{RSS}{TSS}$$

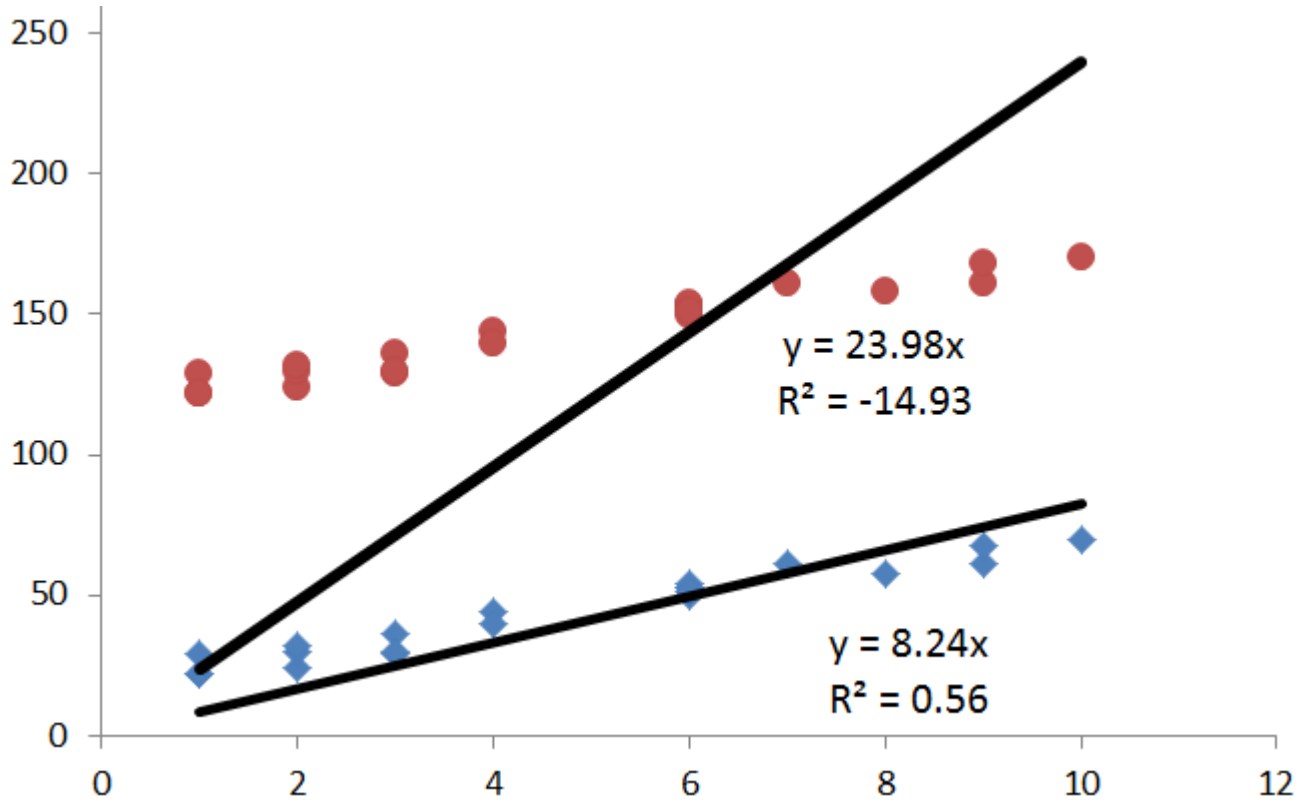
Se define $\overline{R}^2$ ajustado como

$$1 - \overline{R}^2 = \frac{RSS/(n-v)}{TSS/(n-1)}$$

$$\overline{R}^2 = 1 - (1 - R^2)\frac{n-1}{n-v}$$

donde n es la cantidad de datos y v la cantidad de parámetros del modelo.

Un valor negativo de R^2 significa que el ajuste es peor que usar el valor medio



La manera más común de terminar con un valor negativo de R^2 es forzar a la línea de regresión que pase por un punto específico.
El resultado es que la suma del error al cuadrado de la regresión es mayor que si utilizara el valor medio y, por lo tanto, el resultado es un valor negativo de R^2.

Pruebas de significancia de la regresión

Hay que probar varias hipótesis en las regresiones

1- Que la variación explicada por el modelo no sea al azar (*F*-test)

2- Que los parámetros de la regresión sean significativamente distintos de 0 (*t*-test en cada parámetro)

3- Que el valor de R^2 sea cercano a 1 (test Chi-cuadrado)

4- Que el residuo tenga una distribución aleatoria

Algunas consideraciones:

- La estadística *F* dice si le dependencia es al azar o no.

- La estadística *t* dice si cada parámetro es significativo o no para el modelo.

- El coeficiente de correlación, R^2, dice la proporción del del residuo que explica el modelo.

- El análisis del residuo, entre otras cosas, puede determinar si hay o no otras variables no consideras en el modelo.

Hipótesis más comunes en las regresiones:

1. Los errores (residuos) no varían con la variable independiente.

2. Los residuos son independientes, lo que significa que el valor de un residuo no influye sobre el valor de otro.

3. Los residuos siguen una distribución normal.

Test de normalidad (distribución normal)

- Si la variable tiene una distribución normal, se pueden usar estadísticas paramétricas que se basan en esta suposición.
- Si una variable falla la prueba de normalidad, es fundamental observar el histograma y la función de probabilidad normal para ver si sólo un valor atípico o un pequeño subconjunto de valores atípicos ha causado la no normalidad.
- Si no hay valores atípicos, se puede intentar una transformación para que los datos sean normales.
- Finalmente, si una transformación no es una alternativa viable, se pueden utilizar métodos no paramétricos que no requieran normalidad.

Hay numerosos procedimientos para pruebas de normalidad de datos. No siempre un histograma es la mejor herramienta para evaluar la normalidad, ya que hay muchas elecciones subjetivas para construirlo. Además, los histogramas generalmente necesitan tamaños de muestra grandes para detectar datos no normales. Por eso se debe recurrir a otras herramientas (box plot, density trace, normal probability plot).

Se puede verificar la significancia del modelo usando la estadísitica F de la siguiente manera

$$F = \frac{ESS\,/\,(v-1)}{RSS\,/\,(n-v)} \qquad df = v-1, n-v$$

donde v es la cantidad de parámetros del modelo y n el tamaño de la muestra. Por ejemplo, en una regresión lineal $v = 2$.

¿Cómo se interpretan los valores *F* en el análisis de una regresión?

En general, en regresiones el test *F* en regresión compara el ajuste de diferentes modelos. A diferencia del test *t* que evalua solo un coeficiente de regresión a la vez, el test *F* evalua múltiples coeficientes simultáneamente.

En particular el test *F* de significacia general es una forma específica del test *F*. Compara un modelo sin predictores con el modelo que se usa. Un modelo de regresión que no contiene predictores también se conoce como modelo de solo intercepción.

Las hipótesis para lel test *F* de la significancia general son las siguientes:

Hipótesis nula: el ajuste del modelo de solo intercepción y el modelo son iguales.
Hipótesis alternativa: el ajuste del modelo de solo intercepción es significativamente peor en comparación con el modelo.

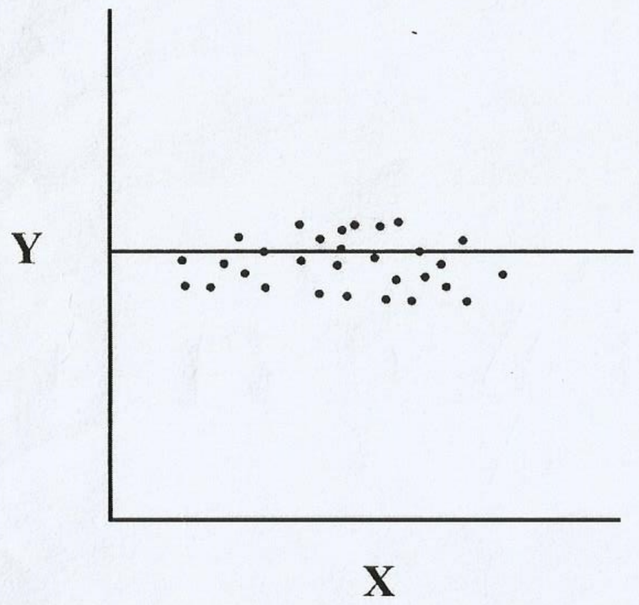
Se puede verificar si la hipótesis nula para un parámetro b usando la disitribución t de la siguiente manera

$$t = \frac{b}{s_b}$$ donde s_b es la desviación estandar del parámetro

$$s_b = \sqrt{\frac{RSS/_{n-v}}{\sum x^2}}$$

Ejemplo: cálculo de s_b para una modelo lineal

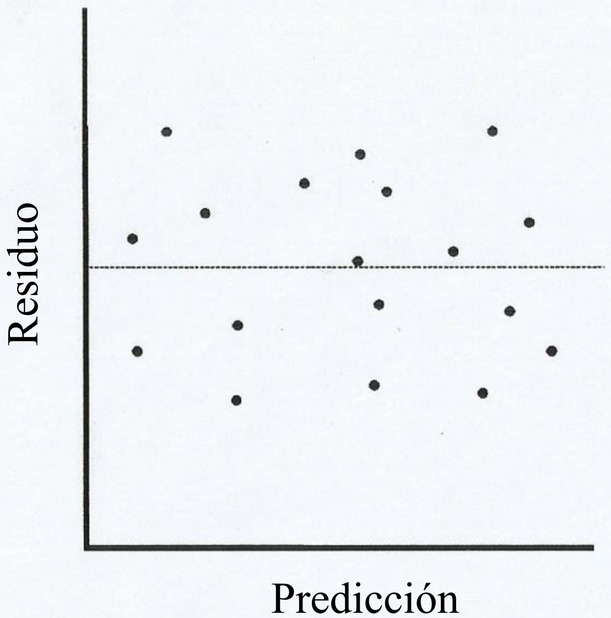
$$df = n - v$$



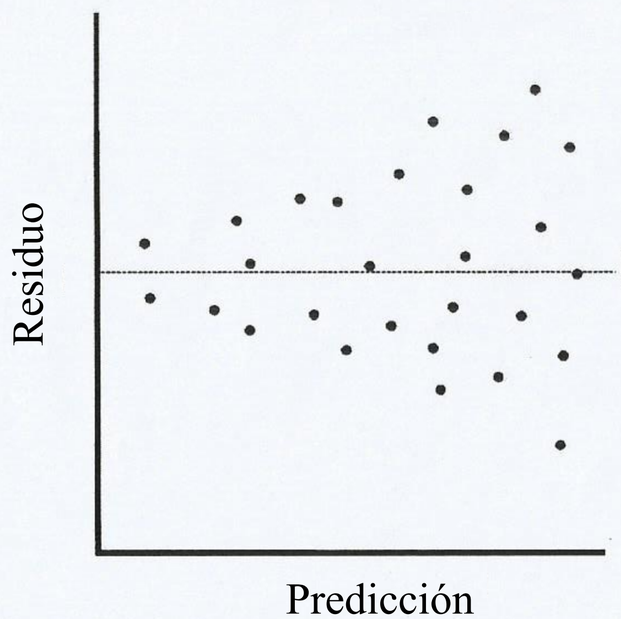
Cuando $b = 0$ no hay cambios en y al variar x

Gráficos de residuos

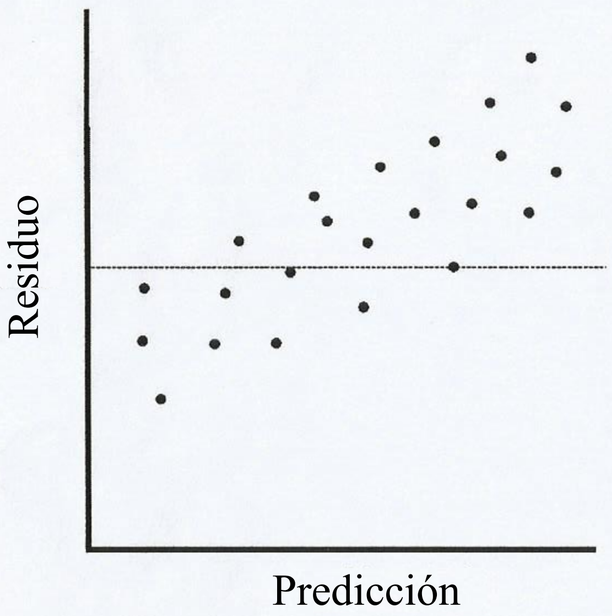
Residuos distribuidos normalmente



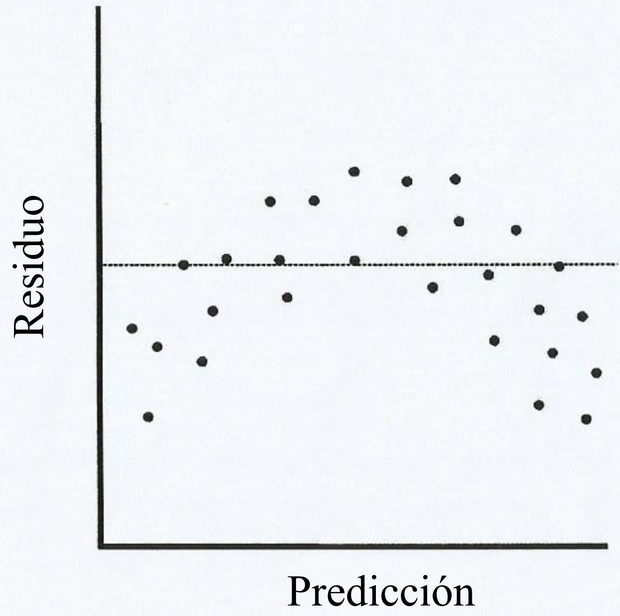
La dispersión del error no es constante



Falta alguna variable en el modelo



El modelo no es el más adecuado



Prueba de bondad del ajuste: χ^2 (Chi-cuadrado)

La prueba de bondad de ajuste de Chi-cuadrado es una prueba no paramétrica que se utiliza para averiguar si el valor observado es significativamente diferente del valor esperado. En la prueba de bondad de ajuste de chi-cuadrado, los datos de la muestra se dividen en intervalos. Luego, se compara la cantidad de puntos medida en cada intervalo con la cantidad esperada.
Hipótesis nula: supone que no hay diferencia significativa entre el valor observado y el esperado.
Hipótesis alternativa: asume que existe una diferencia significativa entre el valor observado y el esperado.
Si el valor calculado de Chi-Cuadrado es mayor que un valor mínimo prefijado, se rechaza la hipótesis nula y se asume que existe una diferencia significativa entre la frecuencia observada y la esperada.

$$\chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}$$

O_i valores observados
E_i valores esperados

los grados de libertad son $l = n - 1$, con n la cantidad de intervalos.

Grados de libertad

Los grados de libertad indican la cantidad de valores independientes que pueden variar en un análisis manteniendo las restricciones existentes.

Los grados de libertad son una combinación de cuantos datos hay y cuantos parámetros se necesitan estimar. O sea, indica cuanta información independiente entra en la estimación de los parámetros. Por lo tanto, para tener estimaciones más precisas se requieren muchos grados de libertad.

Como hemos visto, los grados de libertad también definen las distribuciones de probabilidad para las estadísticas de prueba de varias pruebas de hipótesis. Por ejemplo, las pruebas de hipótesis utilizan la distribución *t*, la distribución *F* y la distribución chi-cuadrado para determinar la significancia estadística. Los grados de libertad para *TSS* son *n*-1 (*n*: cantidad de datos), para *RSS* son *n*-*v* (*v*: cantidad de parámetros del modelo), para *ESS* son *v*-1, y para chi-cuadrado son *k*-1 (*k*: cantidad de intervalos).

¿Cómo se interpretan los valores *p* en el análisis de una regresión?

El valor *p* para cada variable del modelo prueba la hipótesis nula de que el coeficiente es igual a cero (sin efecto). Un valor bajo ($p < 0.05$) indica que se puede rechazar la hipótesis nula. En otras palabras, un predictor que tiene un valor *p* bajo no es debido al azar, y, que probablemente sea un parámetro significativo al modelo porque los cambios en el valor del predictor están relacionados con cambios en la variable de respuesta.
Por el contrario, un valor *p* mayor sugiere que los cambios en el predictor no están asociados con cambios en la respuesta (probabilidad alta de que la dependencia observada sea al azar).
Por eso, normalmente, se utilizan los valores *p* para determinar qué términos mantener en el modelo de regresión.

Intervalos de confianza, intervalos de predicción e intervalos de tolerancia.

Intervalos de confianza:

Intervalo de confianza es un rango de valores derivado de la estadística de la muestra que dice que, con cierta probabilidad, se encuentre el parámetro desconocido de una población, como por ejemplo, el valor medio, o la desviación estandar, como así también los coeficientes de una regresión, etc. Un intervalo de confianza del 95% predice que el valor del parámetro desconocido tiene un probabilidad del 95% de estar dentro de esos límites, pero no predice que el 95% de las observaciones futuras estarán dentro del rango.
Los intervalos de confianza solo evalúan el error de muestreo en relación con el parámetro de interés. Por lo tanto, a medida que aumenta el tamaño de la muestra, el error de muestreo disminuye y los intervalos se vuelven más estrechos. Si se pudiera aumentar el tamaño de la muestra hasta igualar la población, no habría error de muestreo y el intervalo de confianza tendría un ancho cero.

Intervalos de confianza, intervalos de predicción e intervalos de tolerancia.

Intervalos de predicción:

Un intervalo de predicción es un rango que probablemente contenga el valor de respuesta de una única nueva observación dada por la configuración especificada de los predictores en el modelo.
Por ejemplo, un intervalo de predicción con un nivel del 95% indica que existe un 95% de probabilidad de que la próxima medición se encuentre en ese intervalo.

Intervalos de tolerancia:

Un intervalo de tolerancia es un rango que probablemente contenga una proporción específica de la población. Para generar intervalos de tolerancia, se debe especificar tanto la proporción de interés de la población como el nivel de confianza. El nivel de confianza es la probabilidad de que el intervalo cubra realmente la proporción especificada.