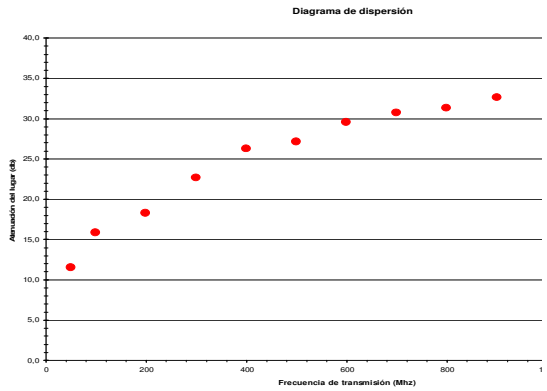
Conceptos preliminares

Modelo determinista

$$Y = \beta_0 + \beta_1 X$$



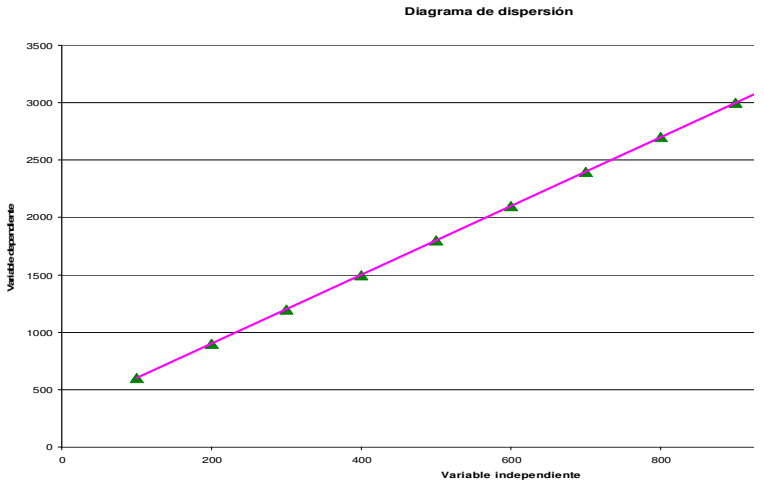
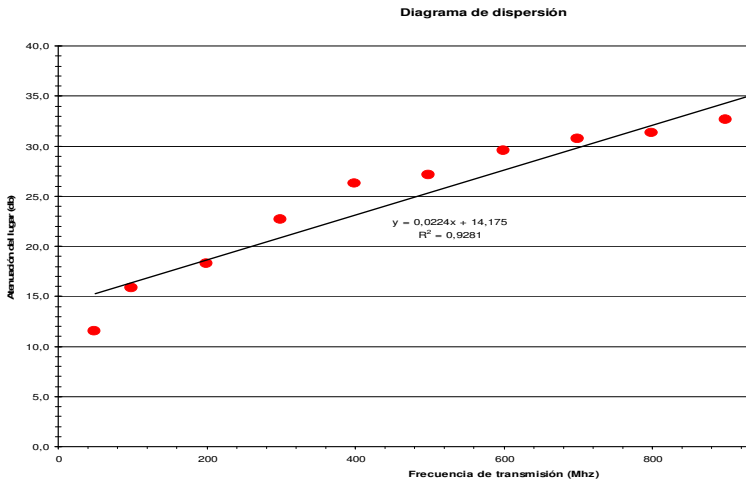


Diagrama de dispersión

Modelo estocástico

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$



Hipótesis de trabajo

(a) $E(\epsilon) = 0$

(b) $\sigma_\epsilon = \text{cte} \; \forall \; x_i$

(c) ϵ tiene una distribución normal $N(0, \sigma^2)$

(d) los errores asociados a cualesquiera dos observaciones son independientes

(e) las y_i son variables aleatorias independientes con distribución $N(\beta_0 + \beta_1 x_i, \sigma^2)$

Modelo de regresión lineal

¡¡¡ Buscar estimaciones de los coeficientes de regresión !!!

$$\beta_0 \rightarrow \hat{\beta}_0$$
$$\beta_1 \rightarrow \hat{\beta}_1$$

Error real o residual = $\varepsilon_i = y_i - (\beta_0 + \beta_1 x_i)$

Error observado = $e_i = y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i) = y_i - \hat{y}_i$

Análisis de regresión lineal simple

(1º) Proponer el modelo probabilístico hipotético

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

(2º) Estimar los parámetros desconocidos de la componente determinista del modelo de hipótesis

$$E(Y|X) = \hat{\beta}_0 + \hat{\beta}_1 X$$

por mínimos cuadrados

(3º) Distribución de probabilidad de la componente de error aleatorio ϵ

(4º) Pruebas de la utilidad del modelo de hipótesis ($\Leftrightarrow$ ¿¿ X contribuye con información a la predicción de Y ??)

(a) Inferencias $\Rightarrow$ intervalo de confianza para β_0 y β_1

(b) Medidas descriptivas numéricas sobre la idoneidad del modelo

(5º) Uso del modelo

(a) Predicción de valores individuales (dando un valor estimado entre un límite inferior y un límite superior para un coeficiente de confianza (prefijado))

(b) Estimar el valor medio de Y, E(Y), para un valor específico de X = x_i

Método de mínimos cuadrados

$$Y = \beta_0 + \beta_1 X + \varepsilon \,/\, E(Y) = E(Y \mid X) = Y = \beta_0 + \beta_1 X$$

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X \longleftarrow \text{Recta de mínimos cuadrados}$$

$\mathbf{SS_E}$ = suma de cuadrados del error = $\sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} \left[y_i - \left(\hat{\beta}_0 + \hat{\beta}_1 x_i \right) \right]^2 =$

$$= SS_{YY} - \hat{\beta}_1 SS_{XY} = \sum_{i=1}^{n} (y_i - \bar{Y})^2 - \frac{\hat{\beta}_1}{n} \left[n \sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i \right] = SS_{YY}(1 - r^2)$$

Minimizar $SS_E \Rightarrow$

$$\hat{\beta}_1 = \frac{SS_{XY}}{SS_{XX}} = \frac{\sum_{i=1}^{n} (x_i - \bar{X})(y_i - \bar{Y})}{\sum_{i=1}^{n} (x_i - \bar{X})^2} = \frac{\sum_{i=1}^{n} (x_i - \bar{X}) y_i}{\sum_{i=1}^{n} (x_i - \bar{X}) x_i} = \frac{\sum_{i=1}^{n} x_i y_i - \frac{1}{n} \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} x_i^2 - \frac{1}{n} \left(\sum_{i=1}^{n} x_i \right)^2} = \frac{\sum_{i=1}^{n} x_i y_i - n \bar{X} \bar{Y}}{\sum_{i=1}^{n} x_i^2 - n \bar{X}^2} = r \sqrt{\frac{SS_{YY}}{SS_{XX}}}$$

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X} = \frac{\sum_{i=1}^{n} y_i \sum_{i=1}^{n} x_i^2 - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} x_i y_i}{n \sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i \right)^2} = \sum_{i=1}^{n} \left[\frac{1}{n} - \frac{\bar{X}(x_i - \bar{X})}{SS_{XX}} \right] y_i$$

Interpretación de SS_E

$$SS_E = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n}\left[y_i - \left(\hat{\beta}_0 + \hat{\beta}_1 x_i\right)\right]^2 = \sum_{i=1}^{n}\left[(y_i - \bar{Y}) - \hat{\beta}_1\left(x_i - \bar{X}\right)\right]^2 =$$

$$= \sum_{i=1}^{n}\left(y_i - \bar{Y}\right)^2 - 2\hat{\beta}_1\sum_{i=1}^{n}\left(x_i - \bar{X}\right)\left(y_i - \bar{Y}\right) + \hat{\beta}_1^2\sum_{i=1}^{n}\left(x_i - \bar{X}\right)^2 =$$

$$= \sum_{i=1}^{n}\left(y_i - \bar{Y}\right)^2 - \hat{\beta}_1^2\sum_{i=1}^{n}\left(x_i - \bar{X}\right)^2 = SS_{TC} - SS_R$$

suma de cuadrados debida a la regresión

suma total corregida de cuadrados

$$SS_T = \sum_{i=1}^{n} y_i^2 = n\bar{Y}^2 + SS_E + SS_R = SS_M + SS_E + SS_R$$

suma total de cuadrados

suma de cuadrados debida a la media

suma de cuadrados debida a la regresión

suma de cuadrados debida al error (≡ desviaciones alrededor de la línea de regresión)

$$\Leftrightarrow y_i = \bar{Y} + (\hat{y}_i - \bar{Y}) + (y_i - \hat{y}_i)$$

Estimación de σ^2

σ^2 mide la variación de los valores de Y respecto de la línea

$$E(Y) = \beta_0 + \beta_1 X$$

Usualmente σ^2 no se conoce, pero debe ser estimada. Un estimador insesgado es:

$$\hat{s}^2 = \frac{SS_E}{n-2}$$

siendo $SS_E = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = SS_{YY} - \hat{\beta}_1 SS_{XY}$ con $SS_{XY} = \sum_{i=1}^{n} x_i y_i - n\overline{XY}$ y

$$SS_{YY} = \sum_{i=1}^{n}(y_i - \bar{y}_i)^2 = \sum_{i=1}^{n} y_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} y_i\right)^2$$

Si se desea hacer estimaciones intervalares y/o contraste de hipótesis sobre la varianza del error del modelo se usa la variable aleatoria

$$\chi^2 = \frac{SS_E}{\sigma^2} = \frac{(n-2)\hat{s}^2}{\sigma^2}$$

que tiene una distribución χ^2 con ν = (n -2) grados de libertad

Propiedades de los estimadores de mínimos cuadrados

Los valores de Y tienen una distribución $N(\beta_0 + \beta_1 X, \sigma^2)$

$$\hat{\beta}_1 = \frac{1}{SS_{XX}} \sum_{i=1}^{n} (x_i - \overline{X}) y_i = \frac{(x_1 - \overline{X})}{SS_{XX}} y_1 + \frac{(x_2 - \overline{X})}{SS_{XX}} y_2 + \cdots + \frac{(x_n - \overline{X})}{SS_{XX}} y_n$$

$\hat{\beta}_0$ y $\hat{\beta}_1$ son estimadores insesgados

$$E(\hat{\beta}_0) = \beta_0$$

$$\sigma^2(\hat{\beta}_0) = \frac{\sigma^2}{n}\left[\frac{1}{SS_{XX}} \sum_{i=1}^{n} x_i^2\right]$$

$$E(\hat{\beta}_1) = \beta_1$$

$$\sigma^2(\hat{\beta}_1) = \frac{\sigma^2}{SS_{XX}}$$

Utilidad del modelo de trabajo

- Inferencias respecto a $\hat{\beta}_0$ y $\hat{\beta}_1$
- Diversos problemas de contraste de hipótesis
- El coeficiente de determinación
- Estimación del valor medio de Y
- Predicción de valores de Y
- Predicción inversa
- Pérdida del ajuste
- Regresión a través de un punto

Inferencias respecto a $\hat{\beta}_1$

¿Qué sucede con los valores de β_0 y β_1 del modelo probabilístico generado si X no contribuye con información a la predicción de Y?

$$\mathbf{H_0}: \ \beta_1 = \beta_1^*$$
$$\mathbf{H_a}: \ \beta_1 \neq \beta_1^*$$

Si no se rechaza H_0 ello no implica que X e Y no estén relacionados ya que se puede cometer un error de tipo I o ambas estar relacionadas de forma no lineal

<u>Estadística de prueba</u>: con las hipótesis de trabajo (planteadas) la distribución en el muestreo de $\hat{\beta}_1$ tiende a una distribución $N\left(\beta_1, \frac{\sigma}{\sqrt{SS_{XX}}}\right)$

Estadístico de prueba

$$\begin{cases} \sigma \text{ conocida} & z = \dfrac{\hat{\beta}_1 - \text{valor de hipótesis de } \beta_1}{\sigma / \sqrt{SS_{XX}}} \\ \sigma \text{ desconocida} & t = \dfrac{\hat{\beta}_1 - \text{valor de hipótesis de } \beta_1}{\hat{s} / \sqrt{SS_{XX}}} \ / \nu = n-2 \quad g.d.l. \end{cases}$$

β_1^* es el valor de hipótesis de β_1

Inferencias respecto a $\hat{\beta}_0$

$$\mathbf{H_0}: \ \beta_0 = \beta_0^*$$
$$\mathbf{H_a}: \ \beta_0 \neq \beta_0^*$$

<u>Estadística de prueba</u>: con las hipótesis de trabajo (planteadas) la distribución en el muestreo de $\hat{\beta}_0$ tiende a una distribución

$$N\left(\beta_0, \sigma\sqrt{\frac{1}{n}+\frac{\overline{X}^2}{SS_{XX}}}\right)$$

Estadístico de prueba

$$\begin{cases} \sigma \text{ conocida} & z = \dfrac{\hat{\beta}_0 - \text{valor de hipótesis de } \beta_0}{\sigma\sqrt{\dfrac{1}{n}+\dfrac{\overline{X}^2}{SS_{XX}}}} \\ \sigma \text{ desconocida} & t = \dfrac{\hat{\beta}_0 - \text{valor de hipótesis de } \beta_0}{\hat{s}\sqrt{\dfrac{1}{n}+\dfrac{\overline{X}^2}{SS_{XX}}}} \ / \nu = n-2 \quad g.d.l. \end{cases}$$

β_0^* es el valor de hipótesis de β_0

Contraste de hipótesis en regresión lineal simple

H_0	Estadístico	Región de rechazo
$\beta_0 = \beta_0^{'}$	$t = \dfrac{\beta_0 - \beta_0^{'}}{s_{\hat{\beta}_0}}$	$t \geq t_{1-\frac{\alpha}{2},n-2}$ o $t \leq t_{\frac{\alpha}{2},n-2}$
$\beta_1 = \beta_1^{'}$	$t = \dfrac{\beta_1 - \beta_1^{'}}{s_{\hat{\beta}_1}}$	
$\mu_{Y\|X=X_0} = \mu_0$	$t = \dfrac{\hat{\beta}_0 - \hat{\beta}_1 X_0 - \mu_0}{s_Y}$	
$\beta_0 = \beta_0^{'}$ y $\beta_1 = \beta_1^{'}$	$F = \dfrac{n\left(\hat{\beta}_0 - \beta_0^{'}\right)^2 + 2n\overline{X}\left(\hat{\beta}_0 - \beta_0^{'}\right)\left(\hat{\beta}_1 - \beta_1^{'}\right) + \sum_{i=1}^{n} x_i^2 \left(\hat{\beta}_1 - \beta_1^{'}\right)^2}{2\hat{\sigma}^2}$	$F \geq F_{1-\alpha;2,n-2}$
$\beta_1 = 0$	$F = \dfrac{MS_R}{MS_E}$	$F \geq F_{1-\alpha;1,n-2}$

Coeficiente de correlación de Pearson

Mide fortaleza de la asociación (≡ tendencia) lineal entre dos variables

$$\mathbf{0} \le |r| \le \mathbf{1} \qquad \begin{cases} r > \mathbf{0} & \textit{correlación positiva} \\ r < \mathbf{0} & \textit{correlación negativa} \\ r = \mathbf{0} & \textit{no existe correlación} \end{cases}$$

$$r = \frac{SS_{XY}}{\sqrt{SS_{XX}}\sqrt{SS_{YY}}} = \frac{s_{XY}}{s_X s_Y} = \hat{\beta}_1 \sqrt{\frac{SS_{XX}}{SS_{YY}}}$$

Una correlación elevada no implica causalidad, ¡¡ tan sólo marca una tendencia lineal entre las variables X e Y !!

Inferencias sobre ρ

Si es ρ el coeficiente de correlación poblacional

Realizar inferencias sobre ρ equivale a efectuar inferencias sobre β_1
La única diferencia está en la escala de medición que se aplica

$\mathbf{H_0}$: $\rho = 0$
$\mathbf{H_a}$: (a) $\rho > 0$
(b) $\rho < 0$
(c) $\rho \neq 0$

$$t = \frac{r-\rho}{s_R} = \frac{r}{s_R} = \frac{\hat{\beta}_1}{s_{\hat{\beta}_1}} = \frac{\hat{\beta}_1}{\hat{s}/\sqrt{SS_{XX}}} = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

con ν = n - 2 grados de libertad

$$s_R = \sqrt{\frac{1-r^2}{n-2}}$$

Si la muestra (X, Y) se escoge al azar de una población normal bivariable
(si ambas distribuciones marginales de X e Y son normales)

Coeficiente de determinación

¿Hasta dónde se pueden reducir los errores de la población Y aprovechando la información proporcionada por X?

$$r^2 = \frac{SS_{YY} - SS_E}{SS_{YY}} = 1 - \frac{SS_E}{SS_{YY}} = \frac{SS_R}{SS_{YY}}$$

Es decir, r^2 representa la proporción (reducción de la suma de los cuadrados de las desviaciones de los valores de Y respecto de sus valores estimados $\hat{Y}$ que se pueden atribuir a una relación lineal entre Y y X

Por ejemplo, si $r^2 = 0.2 \Rightarrow$ la suma de los cuadrados de las desviaciones de los valores de Y respecto de los valores estimados se redujo en un 20 %, al utilizar $\hat{Y}$ en lugar de $\overline{Y}$ para predecir Y

Estimación y predicción

(A) Usar el modelo para *estimar* el valor medio de Y, E(Y), para un valor dado de X

(B) *Predecir* un valor de Y (en particular) para un valor dado de X

Se usa el mismo valor para (A) y (B) pero la diferencia radica en la exactitud relativa de ambos conceptos (estimar y predecir)

La desviación estándar de la distribución de muestreo del estimador $\hat{Y}$ del valor medio de Y para un valor $X = x_p$ es

$$\sigma_{\hat{\mu}_{Y|X}} = \sigma\sqrt{\frac{1}{n} + \frac{(x_p - \bar{X})^2}{SS_{XX}}}$$

$$\Downarrow$$

$$s_{\hat{\mu}_{Y|X}} = \hat{\sigma}\sqrt{\frac{1}{n} + \frac{(x_p - \bar{X})^2}{SS_{XX}}}$$

La desviación estándar del error de predicción para el predictor $\hat{Y}$ de un valor individual de Y para un valor $X = x_p$ es

$$\sigma_{\hat{Y}} = \sigma\sqrt{1 + \frac{1}{n} + \frac{(x_p - \bar{X})^2}{SS_{XX}}}$$

$$\Downarrow$$

$$s_{\hat{Y}} = \hat{\sigma}\sqrt{1 + \frac{1}{n} + \frac{(x_p - \bar{X})^2}{SS_{XX}}}$$

Intervalos de confianza en estimación y predicción

Intervalo de confianza de (1-α) para el valor medio de Y cuando X = x_p:

$$\hat{Y} \mp t_{\frac{\alpha}{2},n-2}\hat{\sigma}\sqrt{\frac{1}{n}+\frac{(x_p-\overline{X})^2}{SS_{XX}}}$$

Intervalo de predicción de (1-α) para una Y individual cuando X = x_p:

$$\hat{Y} \mp t_{\frac{\alpha}{2},n-2}\hat{\sigma}\sqrt{1+\frac{1}{n}+\frac{(x_p-\overline{X})^2}{SS_{XX}}}$$

CUIDADO AL HACER ESTIMACIONES/PREDICCIONES FUERA DEL INTERVALO DONDE SE HA EFECTUADO LA REGRESIÓN

El intervalo de confianza para $\mu_{Y|X}$ es menor que el que corresponde a $\hat{Y}$ ya que este último tiene en cuenta la variabilidad de las y_i individuales porque $s^2_{\hat{Y}} = s^2_{\hat{\mu}_{Y|X}} + \hat{\sigma}^2$. Los intervalos de $\mu_{Y|X}$ e $\hat{Y}$ son más cortos cuando $X = \overline{X}$