Análisis de regresión

7 de septiembre
(semana 5)

Plan de trabajo

1. Inferencia sobre los coeficientes de regresión
2. Predicción en la regresión lineal simple
3. Bondad de ajuste

Nota:

- Esta semana, a más tardar el domingo, les entrego sus propuestas de proyecto (siguiente entrega: semana 11)
- Esta semana también subiré algunas soluciones a los problemas del taller 1.
- **Semana 6 (12 – 18 septiembre):**
  - Quiz por MOODLE. Vale el doble de un quiz de clase.
  - Sesión dudas pre-parcial 1 (viernes 16 o sábado 17)
- **Semana universitaria (19 – 25 septiembre):** ¿clase el miércoles?
- **Semana 7 (26 sept. – 2 octubre):** Parcial 1

Estimaciones

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x}_m)\, y_i}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

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

Estimadores

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x}_m)\, Y_i}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

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

Una definición importante y un lema

Mejor estimador linealmente insesgado (BLUE)

Sean $Y_1, Y_2, \ldots, Y_n$ variables aleatorias involucradas en un proceso de estimación. Se dice que un estimador de la forma $\sum_{i=1}^{n} \phi_i Y_i$ con constantes ϕ_i no aleatorias y conocidas es un **BLUE** si

1) es un estimador insesgado,

2) es el estimador con menor varianza dentro de todos los estimadores lineales insesgados.

- Como siempre la definición no es constructiva. Usaremos el lema del máximo (Lema 11.2.7 de Casella & Berger) más adelante para verificar que un estimador es BLUE.

Coeficientes de regresión (I)

Teorema de Gauss-Markov

La solución al problema de minimización propuesto es:

$$\begin{cases} \hat{\beta}_1 = \dfrac{\sum_{i=1}^{n} \left(x_i - \overline{x}_n \right) Y_i}{\sum_{i=1}^{n} \left(x_i - \overline{x}_n \right)^2}, \\ \hat{\beta}_0 = \overline{Y}_n - \hat{\beta}_1 \overline{x}_n \end{cases}$$

bajo las condiciones del método descrito anteriormente, corresponde a los mejores estimadores linealmente insesgados de los parámetros del modelo de regresión lineal simple.

Además, son consistentes y tienen distribuciones asintóticas normales (por el TCL de Hájek - Sidak).

Idea de la prueba (I)

Es claro que los estimadores encontrados son lineales.

$$\begin{cases} \hat{\beta}_1 = \dfrac{\sum_{i=1}^{n}\left(x_i-\bar{x}_n\right)Y_i}{\sum_{i=1}^{n}\left(x_i-\bar{x}_n\right)^2} = \sum_{i=1}^{n} c_i Y_i, \text{ con } c_i = \dfrac{\left(x_i-\bar{x}_n\right)}{\sum_{i=1}^{n}\left(x_i-\bar{x}_n\right)^2} \\ \hat{\beta}_0 = \bar{Y}_n - \hat{\beta}_1\bar{x}_n = \dfrac{1}{n}\sum_{i=1}^{n} Y_i - \bar{x}_n\sum_{i=1}^{n} c_i Y_i = \sum_{i=1}^{n}\left(\dfrac{1}{n}-\bar{x}_n c_i\right)Y_i \end{cases}.$$

Ahora, vamos a probar que son insesgados:

$$E\left[\hat{\beta}_1\right] = \frac{\sum_{i=1}^{n}\left(x_i-\bar{x}_n\right)E\left[Y_i\right]}{\sum_{i=1}^{n}\left(x_i-\bar{x}_n\right)^2} = \frac{\sum_{i=1}^{n}\left(x_i-\bar{x}_n\right)\left(\beta_0+\beta_1 x_i\right)}{\sum_{i=1}^{n}\left(x_i-\bar{x}_n\right)^2} = \frac{\beta_0\sum_{i=1}^{n}\left(x_i-\bar{x}_n\right)+\beta_1\sum_{i=1}^{n}x_i\left(x_i-\bar{x}_n\right)}{\sum_{i=1}^{n}\left(x_i-\bar{x}_n\right)^2} = \beta_1.$$

$$E\left[\hat{\beta}_0\right] = E\left[\bar{Y}_n\right] - E\left[\hat{\beta}_1\right]\bar{x}_n = \frac{1}{n}\sum_{i=1}^{n}\left(\beta_0+\beta_1 x_i\right) - \beta_1\bar{x}_n = \frac{1}{n}n\beta_0 + \beta_1\bar{x}_n - \beta_1\bar{x}_n = \beta_0.$$

Idea de la prueba (II)

Centremos la atención en el proceso de estimación de β_1.

Todo estimador lineal $\tilde{\beta}_1$ debe ser de la forma $\tilde{\beta}_1 = \sum_{i=1}^{n} \phi_i Y_i$.

Para ser insesgado, debe satisfacer que $E\left[\tilde{\beta}_1\right] = \sum_{i=1}^{n} \phi_i E\left[Y_i\right] = \beta_1$.

De este modo, debe satisfacer que $\sum_{i=1}^{n} \phi_i \left(\beta_0 + \beta_1 x_i\right) = \beta_0 \sum_{i=1}^{n} \phi_i + \beta_1 \sum_{i=1}^{n} \phi_i x_i = \beta_1$.

De allí se obtienen dos restricciones para los coeficientes:

$(1) \sum_{i=1}^{n} \phi_i = 0,\ (2) \sum_{i=1}^{n} \phi_i x_i = 1.$

Ahora, $Var\left(\tilde{\beta}_1\right) = Var\left(\sum_{i=1}^{n} \phi_i Y_i\right) = \sum_{i=1}^{n} \phi_i^2 Var\left(Y_i\right) = \sigma^2 \sum_{i=1}^{n} \phi_i^2$.

Por ende, un BLUE se puede encontrar solucionando el problema de minimizar

$\sum_{i=1}^{n} \phi_i^2$ sujeto a las restricciones (1) y (2). El lema del máximo garantiza que dicha solución es la que se obtiene con el estimador de mínimos cuadrados.

Otro lema importante

Distribución normal de un estimador

Sean $Y_1, Y_2, \ldots, Y_n$ variables aleatorias **independientes con distribución normal**. Se dice que un estimador de la forma $\sum_{i=1}^{n} \phi_i Y_i$ con constantes ϕ_i no aleatorias tiene distribución normal.

Prueba :

Como $Y_1, Y_2, \ldots, Y_n$ son normales e independientes, el vector $\mathbf{Y} = \left(Y_1, Y_2, \ldots, Y_n\right)^T$ tendrá entonces distribución normal multivariada. Adicionalmente, cualquier combinación lineal de sus componentes también tendrá distribución normal, es decir, $\mathbf{a}^T \mathbf{Y} + b = \sum_{i=1}^{n} a_i Y_i + b$ tiene distribución normal univariada.

Coeficientes de regresión (II)

Teorema de Gauss-Markov (II)

Si además se tiene que $\{\varepsilon_i\}$ son una m.a. $N\left(0,\sigma^2\right)$, $\hat{\beta}_0$ y $\hat{\beta}_1$ son UMVUEs de sus respectivos parámetros y son los mismos estimadores ML. Además, tienen distribución normal bivariada con

$$\left(\hat{\beta}_0,\hat{\beta}_1\right)^T \sim MVN_2\left(\begin{pmatrix}\beta_0\\ \beta_1\end{pmatrix},\begin{bmatrix}\dfrac{\sigma^2\sum_{i=1}^{n}x_i^2}{n\sum_{i=1}^{n}\left(x_i-\overline{x}_n\right)^2} & \dfrac{-\overline{x}_n\sigma^2}{\sum_{i=1}^{n}\left(x_i-\overline{x}_n\right)^2}\\ \dfrac{-\overline{x}_n\sigma^2}{\sum_{i=1}^{n}\left(x_i-\overline{x}_n\right)^2} & \dfrac{\sigma^2}{\sum_{i=1}^{n}\left(x_i-\overline{x}_n\right)^2}\end{bmatrix}\right)$$

Idea de la prueba (I)

Si $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ son una m.a. $N\left(0, \sigma^2\right)$; $Y_1, Y_2, \ldots, Y_n$ heredan esa normalidad y son independientes entre sí, $Y_i \sim N\left(\beta_0 + \beta_1 x_i, \sigma^2\right)$. Entonces,

$$L\left(\beta_0, \beta_1, \sigma^2 \middle| \mathbf{y}, \mathbf{x}\right) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp\left[\frac{-\left(y_i - \beta_0 - \beta_1 x_i\right)^2}{2\sigma^2}\right] \text{ y}$$

$$l\left(\beta_0, \beta_1, \sigma^2 \middle| \mathbf{y}, \mathbf{x}\right) = -\frac{n}{2}\ln 2\pi - n\ln\sigma - \frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(y_i - \beta_0 - \beta_1 x_i\right)^2.$$

Para estimar a β_0, β_1 :

$$\max l\left(\beta_0, \beta_1, \sigma^2 \middle| \mathbf{y}, \mathbf{x}\right) = \max\left\{-\frac{n}{2}\ln 2\pi - n\ln\sigma - \frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(y_i - \beta_0 - \beta_1 x_i\right)^2\right\}$$

$$= \max\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(y_i - \beta_0 - \beta_1 x_i\right)^2\right\} = \min\left\{\sum_{i=1}^{n}\left(y_i - \beta_0 - \beta_1 x_i\right)^2\right\}.$$

Luego, los estimadores ML coinciden con los estimadores MCO (OLS en inglés).

Idea de la prueba (II)

Retomando la densidad de Y, se puede ver que este modelo pertenece a la fam.

exponencial de densidades triparamétrica con $\boldsymbol{\theta} = \left(\beta_0, \beta_1, \sigma^2\right)^T$ porque

$$f_{Y_i}\left(y_i, \boldsymbol{\theta}\right) = \frac{1}{\sqrt{2\pi\sigma}} \exp\left[\frac{-\left(y_i - \beta_0 - \beta_1 x_i\right)^2}{2\sigma^2}\right]$$

$$= \frac{1}{\sqrt{2\pi\sigma}} \exp\left[\frac{-y_i^2}{2\sigma^2} + \frac{y_i\left(\beta_0 + \beta_1 x_i\right)}{\sigma^2} - \frac{\left(\beta_0 + \beta_1 x_i\right)^2}{2\sigma^2}\right]. \text{ De donde se tiene:}$$

$$a\left(\boldsymbol{\theta}\right) = \frac{1}{\sqrt{2\pi\sigma}} \exp\left[-\frac{\left(\beta_0 + \beta_1 x_i\right)^2}{2\sigma^2}\right], \qquad b\left(y\right) = 1,$$

$$\mathbf{c}\left(\boldsymbol{\theta}\right) = \left(\frac{-1}{2\sigma^2}, \frac{\beta_0}{\sigma^2}, \frac{\beta_1}{\sigma^2}\right)^T, \qquad \mathbf{d}\left(y\right) = \left(y_i^2, y_i, x_i y_i\right)^T.$$

Luego, $\mathbf{d}\left(\mathbf{Y}\right) = \left(\sum_{i=1}^n Y_i^2, \sum_{i=1}^n Y_i, \sum_{i=1}^n x_i Y_i\right)^T$ es un vector de estadísticas suficientes y completas.

Como $\hat{\beta}_0, \hat{\beta}_1$ son insesgados y función de $\mathbf{d}\left(\mathbf{Y}\right)$, estos son UMVUEs.

Gracias a la expresión lineal de cada estimador, ambos tienen distribución normal.

La normalidad multivariada se probará más adelante. Finalmente,

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

- $$Var\left(\hat{\beta}_0\right) = Var\left(\bar{Y}_n-\hat{\beta}_1\bar{x}_n\right) = Var\left(\bar{Y}_n\right)+\bar{x}_n^2 Var\left(\hat{\beta}_1\right)-2Cov\left(\bar{Y}_n,\hat{\beta}_1\bar{x}_n\right)$$

$$= \frac{\sigma^2}{n}+\frac{\bar{x}_n^2\sigma^2}{\sum_{i=1}^{n}\left(x_i-\bar{x}_n\right)^2}-2\bar{x}_n Cov\left(\bar{Y}_n,\hat{\beta}_1\right)$$

$$= \sigma^2\left[\frac{\sum_{i=1}^{n}\left(x_i-\bar{x}_n\right)^2+n\bar{x}_n^2}{n\sum_{i=1}^{n}\left(x_i-\bar{x}_n\right)^2}\right]-2\bar{x}_n Cov\left(\frac{1}{n}\sum_{i=1}^{n}Y_i,\frac{\sum_{i=1}^{n}\left(x_i-\bar{x}_n\right)Y_i}{\sum_{i=1}^{n}\left(x_i-\bar{x}_n\right)^2}\right)$$

$$= \sigma^2\left[\frac{\sum_{i=1}^{n}x_i^2}{n\sum_{i=1}^{n}\left(x_i-\bar{x}_n\right)^2}\right]-\frac{2\bar{x}_n}{n\sum_{i=1}^{n}\left(x_i-\bar{x}_n\right)^2}Cov\left(\sum_{i=1}^{n}Y_i,\sum_{i=1}^{n}\left(x_i-\bar{x}_n\right)Y_i\right)$$

$$= \sigma^2 \left[\frac{\sum_{i=1}^{n} x_i^2}{n\sum_{i=1}^{n}\left(x_i - \bar{x}_n\right)^2} \right] - \frac{2\bar{x}_n}{n\sum_{i=1}^{n}\left(x_i - \bar{x}_n\right)^2} \sum_{i=1}^{n}\sum_{j=1}^{n} Cov\left(Y_i, \left(x_j - \bar{x}_n\right) Y_j\right)$$

$$= \sigma^2 \left[\frac{\sum_{i=1}^{n} x_i^2}{n\sum_{i=1}^{n}\left(x_i - \bar{x}_n\right)^2} \right] - \frac{2\bar{x}_n}{n\sum_{i=1}^{n}\left(x_i - \bar{x}_n\right)^2} \sum_{i=1}^{n} Cov\left(Y_i, \left(x_i - \bar{x}_n\right) Y_i\right)$$

$$= \sigma^2 \left[\frac{\sum_{i=1}^{n} x_i^2}{n\sum_{i=1}^{n}\left(x_i - \bar{x}_n\right)^2} \right] - \frac{2\bar{x}_n}{n\sum_{i=1}^{n}\left(x_i - \bar{x}_n\right)^2} \sum_{i=1}^{n} \left(x_i - \bar{x}_n\right) \underbrace{Cov\left(Y_i, Y_i\right)}_{=Var\left(Y_i\right)=\sigma^2} = \sigma^2 \left[\frac{\sum_{i=1}^{n} x_i^2}{n\sum_{i=1}^{n}\left(x_i - \bar{x}_n\right)^2} \right].$$

- $Cov\left(\hat{\beta}_0, \hat{\beta}_1\right) = Cov\left(\bar{Y}_n - \hat{\beta}_1 \bar{x}_n, \hat{\beta}_1\right) = Cov\left(\bar{Y}_n, \hat{\beta}_1\right) - Cov\left(\hat{\beta}_1 \bar{x}_n, \hat{\beta}_1\right)$

$$= 0 - \bar{x}_n Cov\left(\hat{\beta}_1, \hat{\beta}_1\right) = -\bar{x}_n Var\left(\hat{\beta}_1\right) = \left[\frac{-\bar{x}_n \sigma^2}{\sum_{i=1}^{n}\left(x_i - \bar{x}_n\right)^2} \right].$$

Gauss - Markov (1)

$$\begin{cases} Y_k = \mu_k + e_k \\ \mu_k = \beta_0 + \beta_1 x_k \\ e_k \overset{iid}{\sim} \text{con } E[e_k] = 0 \\ \qquad\qquad \text{Var}[e_k] = \sigma^2 \ \forall k \end{cases}$$

$\Downarrow$

- $\hat{\beta}_0, \hat{\beta}_1$ son insesgados → igual
- $\hat{\beta}_0, \hat{\beta}_1$ son BLUEs → son UMVUEs
- $\hat{\beta}_0, \hat{\beta}_1$ son consistentes → igual
- $\hat{\beta}_0, \hat{\beta}_1$ tienen distribución asintótica normal → La distribución normal se tiene para todo tamaño de muestra

Gauss - Markov (2)

$$\begin{cases} Y_k = \mu_k + e_k \\ \mu_k = \beta_0 + \beta_1 x_k \\ e_k \overset{iid}{\sim} N(0, \sigma^2) \ \forall k \end{cases}$$

+

$\frac{(n-2)\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{(n-2)}$ y es independiente, permite hacer inferencia

Coeficientes de regresión (III)

Bajo normalidad, el estimador ML de σ^2 es

$$\hat{\sigma}^2_{ML} = \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right)^2,$$

sin embargo, este estimador es sesgado. Un estimador insesgado de la varianza es

$$\hat{\sigma}^2 = \frac{1}{n-2}\sum_{i=1}^{n}\left(Y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right)^2.$$ Sobre este estimador, es posible probar que:

$$\frac{(n-2)\hat{\sigma}^2}{\sigma^2} \sim \chi^2(n-2),$$ y es independiente de $\hat{\beta}_0$ y $\hat{\beta}_1$.

- Se probará más adelante su insesgamiento, su distribución y su independencia.

Intervalos de confianza para los coeficientes de regresión bajo normalidad

$$\frac{\hat{\beta}_1 - \beta_1}{\sqrt{Var\left(\hat{\beta}_1\right)}} = \sqrt{\sum_{i=1}^{n}\left(x_i - \bar{x}_n\right)^2}\left(\frac{\hat{\beta}_1 - \beta_1}{\sigma}\right) \sim N(0,1)$$

$$\perp \quad \frac{(n-2)\hat{\sigma}^2}{\sigma^2} \sim \chi^2(n-2)$$

$$\frac{\hat{\beta}_0 - \beta_0}{\sqrt{Var\left(\hat{\beta}_0\right)}} = \sqrt{\frac{n\sum_{i=1}^{n}\left(x_i - \bar{x}_n\right)^2}{\sum_{i=1}^{n} x_i^2}}\left(\frac{\hat{\beta}_0 - \beta_0}{\sigma}\right) \sim N(0,1)$$

- **Cantidad pivote:**

$$\frac{\hat{\beta}_i - \beta_i}{\sqrt{V\hat{a}r\left(\hat{\beta}_i\right)}} \sim t(n-2)$$

$$IC_{100(1-\alpha)\%}\left(\beta_i\right) = \hat{\beta}_i \mp t_{1-\frac{\alpha}{2}}(n-2)\sqrt{V\hat{a}r\left(\hat{\beta}_i\right)}$$

Tomando como ejemplo a $\hat{\beta}_1$:

Si $\hat{\beta}_1 \sim N\left(\beta_1, \frac{\sigma^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}\right)$ ← $Var(\hat{\beta}_1)$

¿Por qué no se toma la versión estandarizada como variable pivote?

$$\frac{1}{\sqrt{Var(\hat{\beta}_1)}} = \frac{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}\left(\hat{\beta}_1-\beta_1\right)}{\sigma} \sim N(0,1)$$

R/ Porque depende de σ además de depender de β_1, así que no sirve como variable pivote.

Solución:

$$\frac{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}\left(\hat{\beta}_1-\beta_1\right)}{\sigma} \sim N(0,1) \quad \perp \quad \frac{(n-2)\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{(n-2)}$$

(⊥: independiente)

Luego,

$$\frac{N(0,1)}{\sqrt{\frac{\chi^2_{(n-2)}}{n-2}}} = \frac{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}\left(\hat{\beta}_1-\beta_1\right)}{\hat{\sigma}} \sim t_{(n-2)}$$

Chi-cuad en ↑ dividida por sus grados de libertad

Nota:

Esto es lo que se hace siempre que se va a armar una t, solo que en el denominador siempre va a quedar $\hat{\sigma}$ en vez de σ; entonces, se suele decir simplemente

$$\frac{\hat{\beta}_1-\beta_1}{\sqrt{\widehat{Var}(\hat{\beta}_1)}} \sim t_{(n-2)} \quad \text{con} \quad \widehat{Var}(\hat{\beta}_1) = \frac{\hat{\sigma}^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}$$

Varianza estimada: se reemplaza σ por $\hat{\sigma}$

Pruebas de hipótesis para los coeficientes de regresión bajo normalidad

(Sist. a dos colas) **(Sist. con cola a derecha)** **(Sist. con cola a izquierda)**

$$\begin{cases} H_0 : \beta_i = \tilde{\beta} \\ versus \\ H_1 : \beta_i \neq \tilde{\beta} \end{cases} \qquad \begin{cases} H_0 : \beta_i \leq \tilde{\beta} \\ versus \\ H_1 : \beta_i > \tilde{\beta} \end{cases} \qquad \begin{cases} H_0 : \beta_i \geq \tilde{\beta} \\ versus \\ H_1 : \beta_i < \tilde{\beta} \end{cases}$$

- **Valor calculado:** $t_C = \dfrac{\hat{\beta}_i - \tilde{\beta}}{\sqrt{\hat{Var}\left(\hat{\beta}_i\right)}}$

- **Regla de decisión:**

(Sist. a dos colas) **(Sist. con cola a derecha)** **(Sist. con cola a izquierda)**

τ:"Rechazar H_0 si

$$|t_C| > t_{1-\frac{\alpha}{2}}(n-2)" \qquad t_C > t_{1-\alpha}(n-2)" \qquad t_C < t_{\alpha}(n-2)"$$

$$p - values =$$

$$2p\left(T_C > |t_C| \,\middle\|\, H_0\right) \qquad p\left(T_C > t_C \middle| H_0\right) \qquad p\left(T_C < t_C \middle| H_0\right)$$

Intervalos de confianza para la media

$\mu_i = \beta_0 + \beta_1 x_i$: Valor esperado para una unidad con valor x_i de la covariable.

→ $\hat{\mu}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$: Estimación puntual. → Var. Aleatoria

↦ $E\left[\hat{\mu}_i\right] = E\left[\hat{\beta}_0 + \hat{\beta}_1 x_i\right] = \beta_0 + \beta_1 x_i = \mu_i$

↦ $Var\left[\hat{\mu}_i\right] = Var\left[\hat{\beta}_0 + \hat{\beta}_1 x_i\right] = Var\left(\hat{\beta}_0\right) + x_i^2 Var\left(\hat{\beta}_1\right) + 2x_i \operatorname{cov}\left(\hat{\beta}_0, \hat{\beta}_1\right)$

$\hat{\mu}_i \sim N\left(\mu_i, Var\left[\hat{\mu}_i\right]\right)$

❖ **Cantidad pivote:**

$$\frac{\hat{\mu}_i - \mu_i}{\sqrt{\hat{Var}\left(\hat{\mu}_i\right)}} \sim t\left(n-2\right)$$

$\hat{Var}(\hat{\mu})$ depende de σ^2, luego $\hat{Var}(\hat{\mu})$ se obtiene reemplazando σ^2 por $\hat{\sigma}^2$

$$IC_{100(1-\alpha)\%}\left(\mu_i\right) = \hat{\mu}_i \mp t_{1-\frac{\alpha}{2}}\left(n-2\right)\sqrt{\hat{Var}\left(\hat{\mu}_i\right)}$$

Intervalos de predicción

Sea x^* el valor de la covariable para una **nueva unidad** de la cual no se conoce su respuesta Y^*. ¿Cómo construir un int. de **predicción** para Y^*?

$\mu^* = \beta_0 + \beta_1 x^*$: Valor esperado para una unidad con valor x^*.

$$\hat{\mu}^* \sim N\left(\mu^*, Var\left[\hat{\mu}^*\right]\right)$$

Sea $e^* = Y^* - \hat{\mu}^*$

$$E\left[e^*\right] = E\left[Y^* - \hat{\mu}^*\right] = 0$$

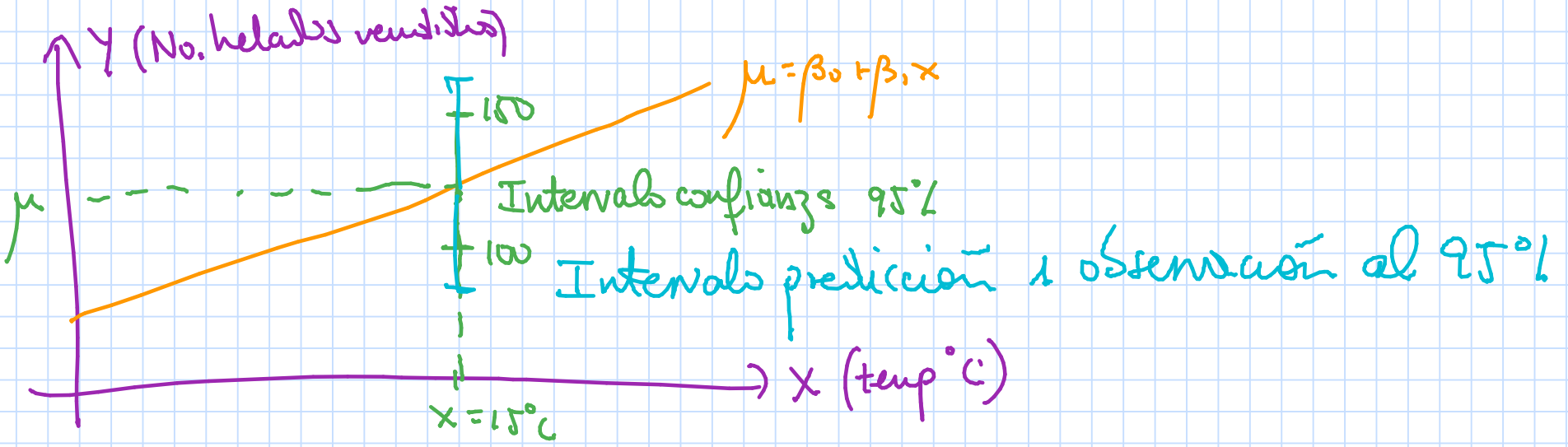
$$Var\left[e^*\right] = Var\left[Y^* - \hat{\mu}^*\right] = Var\left(Y^*\right) + Var\left[\hat{\mu}^*\right] = \sigma^2 + Var\left[\hat{\mu}^*\right]$$

$$e^* \sim N\left(0, \sigma^2 + Var\left[\hat{\mu}^*\right]\right)$$

❖ **Cantidad pivote:**

Misma observación

$$\frac{e^*}{\sqrt{\hat{\sigma}^2 + V\hat{a}r\left(\hat{\mu}^*\right)}} \sim t\left(n-2\right) \quad IP_{100(1-\alpha)\%}\left(Y^*\right) = \hat{\mu}^* \mp t_{1-\frac{\alpha}{2}}\left(n-2\right)\sqrt{\hat{\sigma}^2 + V\hat{a}r\left(\hat{\mu}^*\right)}$$


Y (No. helados vendidos)
$\mu = \beta_0 + \beta_1 x$
150
μ
Intervalo confianza 95%
100
Intervalo predicción 1 observación al 95%
X (temp °C)
x = 15°C

Intervalo de confianza

- Interés en un parámetro (μ) para el valor promedio

- Estimación puntual:

$$\hat{\mu}_k = \hat{\beta}_0 + \hat{\beta}_1 x_k$$

- Variable pivote:

$$Var(\hat{\mu}_k) = Var[\hat{\beta}_0 + \hat{\beta}_1 x_k] =$$

$$Var[\hat{\beta}_0] + x_k^2 Var[\hat{\beta}_1] + 2x_k Cov(\hat{\beta}_0, \hat{\beta}_1)$$

$\hat{\mu}_k$ es una C.L. de $Y_1, \dots, Y_n$

Luego, $\hat{\mu}_k \sim N(\mu_k, Var(\mu_k))$

y $$\frac{\hat{\mu}_k - \mu_k}{\sqrt{\hat{Var}(\hat{\mu}_k)}} \sim t(n-2)$$

Porque toca estimar a σ

$$IC(\mu_k) = \hat{\mu}_k \mp t_{1-\alpha/2}^{(n-2)} \sqrt{\hat{Var}(\hat{\mu}_k)}$$

Intervalo de predicción

- Interés en una variable aleatoria (Y_k) para un individuo fuera de la muestra

- Predicción puntual

$$\hat{\mu}_k = \hat{y}_k = \hat{\beta}_0 + \hat{\beta}_1 x_k$$

- OJO! necesitamos una var. pivote que tenga a Y_k^* y a $\hat{\mu}_k^*$

$$e_k^* = Y_k^* - \hat{\mu}_k^*$$

$$E[e_k^*] = E[Y_k^*] - E[\hat{\mu}_k^*]$$

$$= \beta_0 + \beta_1 x_k - (\beta_0 + \beta_1 x_k) = 0$$

$$Var[e_k^*] = Var[Y_k^*] + Var[\hat{\mu}_k^*]$$

↑ Y_k^* está fuera de la muestra

$$= \sigma^2 + Var[\hat{\mu}_k^*]$$

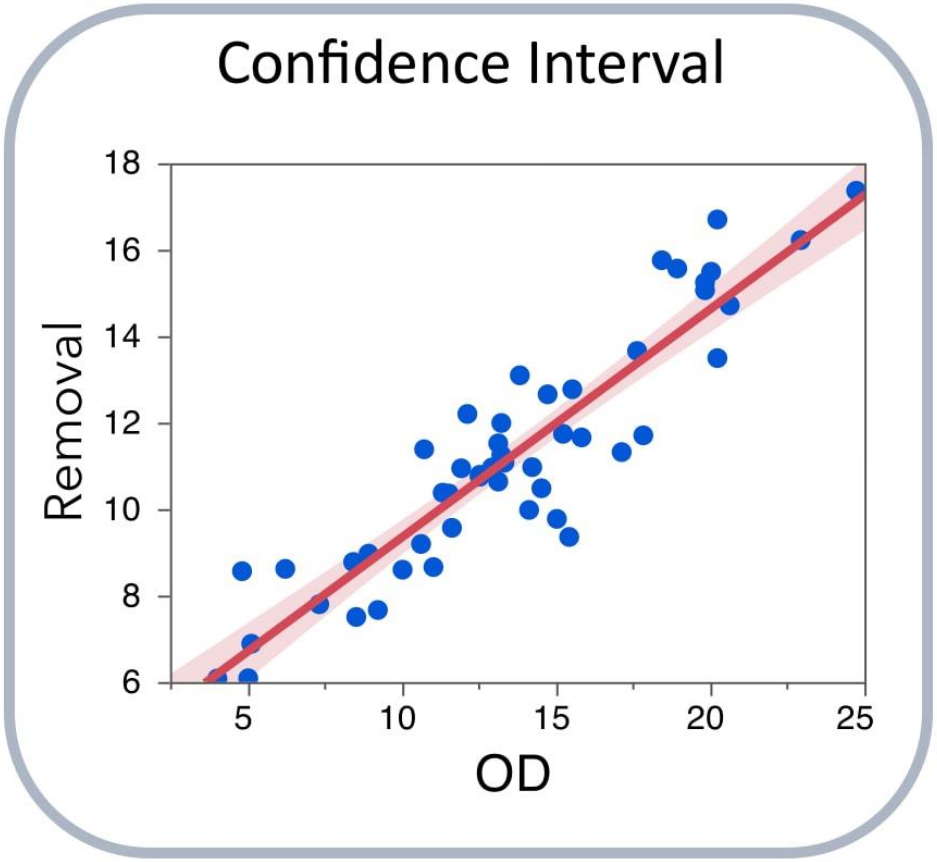
Luego,

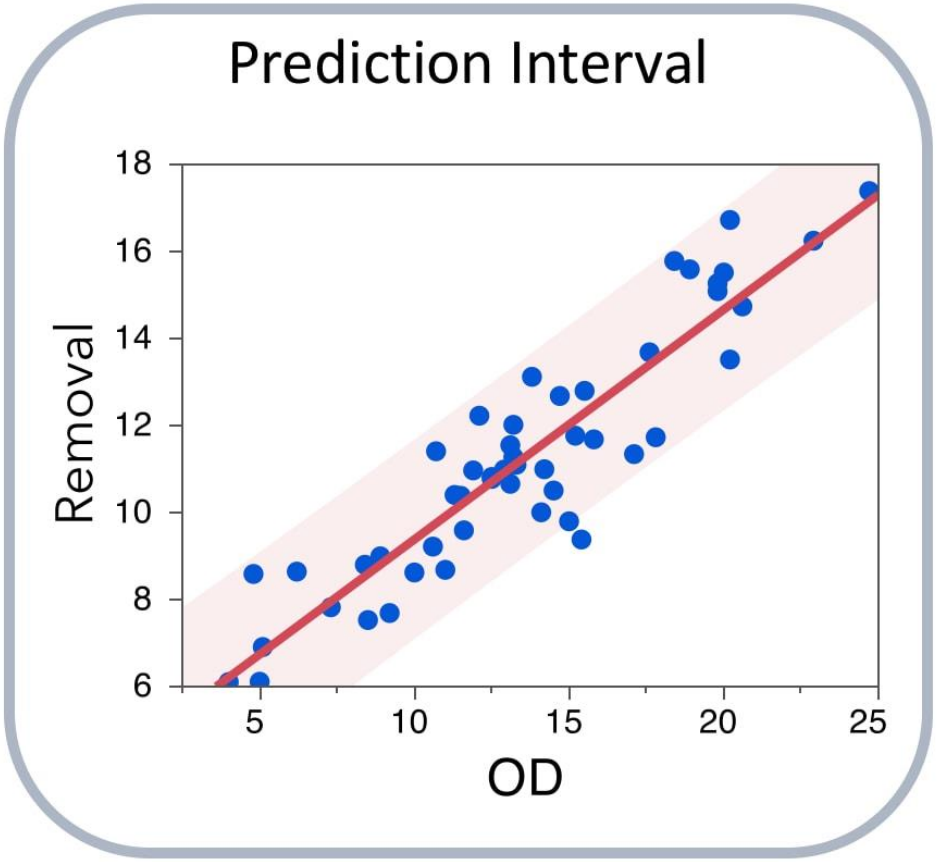
$$e_k^* \sim N(0, \sigma^2 + Var[\hat{\mu}_k^*])$$

$$\frac{Y_k^* - \hat{\mu}_k^*}{\frac{e_k^*}{\sqrt{\hat{\sigma}^2 + \hat{Var}(\hat{\mu}_k^*)}}} \sim t(n-2)$$

$$IP(Y_k^*) = \hat{\mu}_k^* \mp t_{1-\alpha/2}^{(n-2)} \sqrt{\hat{\sigma}^2 + \hat{Var}(\hat{\mu}_k^*)}$$

Bandas de confianza y de predicción





Fuente: https://www.jmp.com/en_us/statistics-knowledge-portal/what-is-regression/interpreting-regression-results.html

Descomposición de varianza (Bondad ajuste)

Teorema de descomposición de varianza

Si se define:

$$\bar{Y} := \frac{1}{n}\sum_{i=1}^{n} Y_i$$

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

Entonces:

Var. sin explicar en el modelo — Vars. explicadas por el modelo

$$\underbrace{\sum_{i=1}^{n}\left(Y_i - \bar{Y}\right)^2}_{SC_{total}} = \underbrace{\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2}_{SC_{error}} + \underbrace{\sum_{i=1}^{n}\left(\hat{Y}_i - \bar{Y}\right)^2}_{SC_{\text{mod}}}.$$

(Fijo. No dep. del mod. de regresión)

- Análogo al cálculo de descomposición de varianza en un modelo ANOVA.

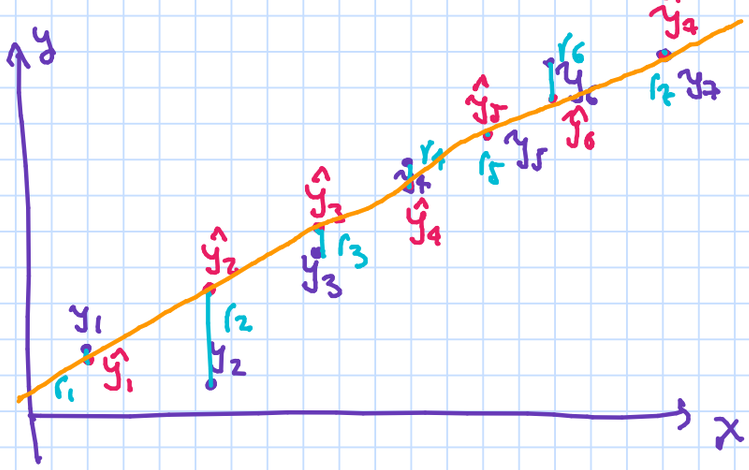
Un "buen" modelo desea tener SC error ↓ SC mod ↑

- Lo vamos a probar en regresión lineal múltiple.

Visualización de la SC total

Asumiendo una muestra aleatoria

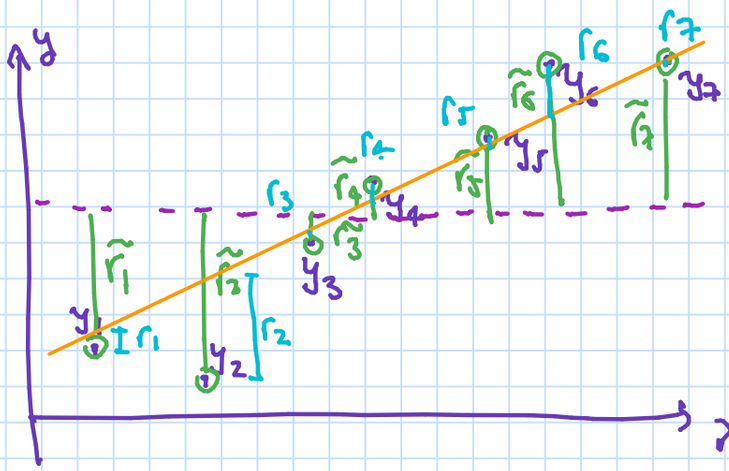
$$SC_{Total} = \sum_{i=1}^{n} (y_i - \bar{y})^2 = \tilde{r}_1^2 + \tilde{r}_2^2 + \dots + \tilde{r}_7^2$$



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

$$SC_{error} = r_1^2 + r_2^2 + r_3^2 + r_4^2 + r_5^2 + r_6^2 + r_7^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

"Error puro"



$$SC_{mod} = (\tilde{r}_1 - r_1)^2 + (\tilde{r}_2 - r_2)^2 + \dots + (\tilde{r}_7 - r_7)^2 = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$$

Hipótesis de significancia de la regresión

Considere el sistema:

$$\begin{cases} H_0: \text{La regresión no tiene valor agregado} \\ \textit{versus} \\ H_1: \text{ La regresión tiene valor agregado} \end{cases} \overset{RLS}{\Leftrightarrow} \begin{cases} H_0: \beta_1 = 0 \\ \textit{versus} \\ H_1: \ \beta_1 \neq 0 \end{cases}$$

Teorema:

Bajo H_0, se tiene que:

$$\frac{SC_{error}}{\sigma^2} = \frac{\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2}{\sigma^2} \sim \chi^2\left(n-2\right), \text{ y}$$

$$\frac{SC_{\text{mod}}}{\sigma^2} = \frac{\sum_{i=1}^{n}\left(\hat{Y}_i - \bar{Y}\right)^2}{\sigma^2} \sim \chi^2\left(1\right); \text{ siendo v. a. independientes.}$$

Prueba F de significancia de la regresión

Prueba F de significancia de la regresión

Considere el sistema: $\begin{cases} H_0 : \text{La regresión no tiene valor agregado} \\ \textit{versus} \\ H_1 : \text{La regresión tiene valor agregado} \end{cases}$.

Si se define $CM_{\text{mod}} = SC_{\text{mod}}/1$ y $CM_{error} = SC_{error}/n-2$

$$F_C := \frac{CM_{\text{mod}}}{CM_{error}} \overset{H_0 \text{ cierta}}{\sim} F(1, n-2).$$

El test τ:"Rechazar H_0 si $f_C > F_{1-\alpha}(1, n-2)$" es un test del $100 \cdot \alpha\%$ de significancia.

El p-valor de este test es $p - value = p\left[F_C > f_C \middle| H_0\right]$.

Prueba F de significancia de la regresión (II)

Teorema: Prueba F de significancia de la regresión

En el modelo de regresión lineal simple, se tiene que:

$$F_C := \frac{CM_{\text{mod}}}{CM_{error}} \overset{H_0 \text{ cierta}}{\sim} F\left(1, n-2\right).$$

Para el sistema: $\begin{cases} H_0 : \beta_1 = 0 \\ versus \\ H_1 : \beta_1 \neq 0 \end{cases}$, se tiene que:

$$T_C = \frac{\hat{\beta}_1}{\sqrt{V\hat{a}r\left(\hat{\beta}_1\right)}} \overset{H_0 \text{ cierta}}{\sim} t\left(n-2\right).$$

En este caso, se tiene que $T_C^2 \overset{H_0 \text{ cierta}}{=} F_C \sim F\left(1, n-2\right).$

Coeficiente de determinación R^2

Coeficiente de determinación

El coeficiente de determinación, R^2, expresa el porcentaje de varianza explicado por el modelo de regresión. Se calcula como:

$$R^2 = \frac{SC_{\text{mod}}}{SC_{total}} = 1 - \frac{SC_{error}}{SC_{total}}$$

- En un modelo de regresión lineal simple tiene sentido su uso, sin embargo, hay que tener cuidado cuando se esté ajustando un modelo de regresión lineal múltiple.

Teorema: Coeficiente de determinación

En el modelo de regresión lineal simple:

$$R^2 = \hat{\rho}^2_{X,Y}$$

Validación del modelo

Residuales

Los residuales del modelo, $\{r_i\}$, se definen como los componentes no explicados por el modelo. Es decir:

$$r_i = y_i - \hat{y}_i = y_i - \left(\hat{\beta}_0 + \hat{\beta}_1 x_i\right)$$

- Aunque no es posible considerar que los residuales tienen exactamente el mismo comportamiento probabilístico que los errores del modelo, se espera que ellos den indicios sobre las propiedades deseables de los errores.

Validación del modelo (II)

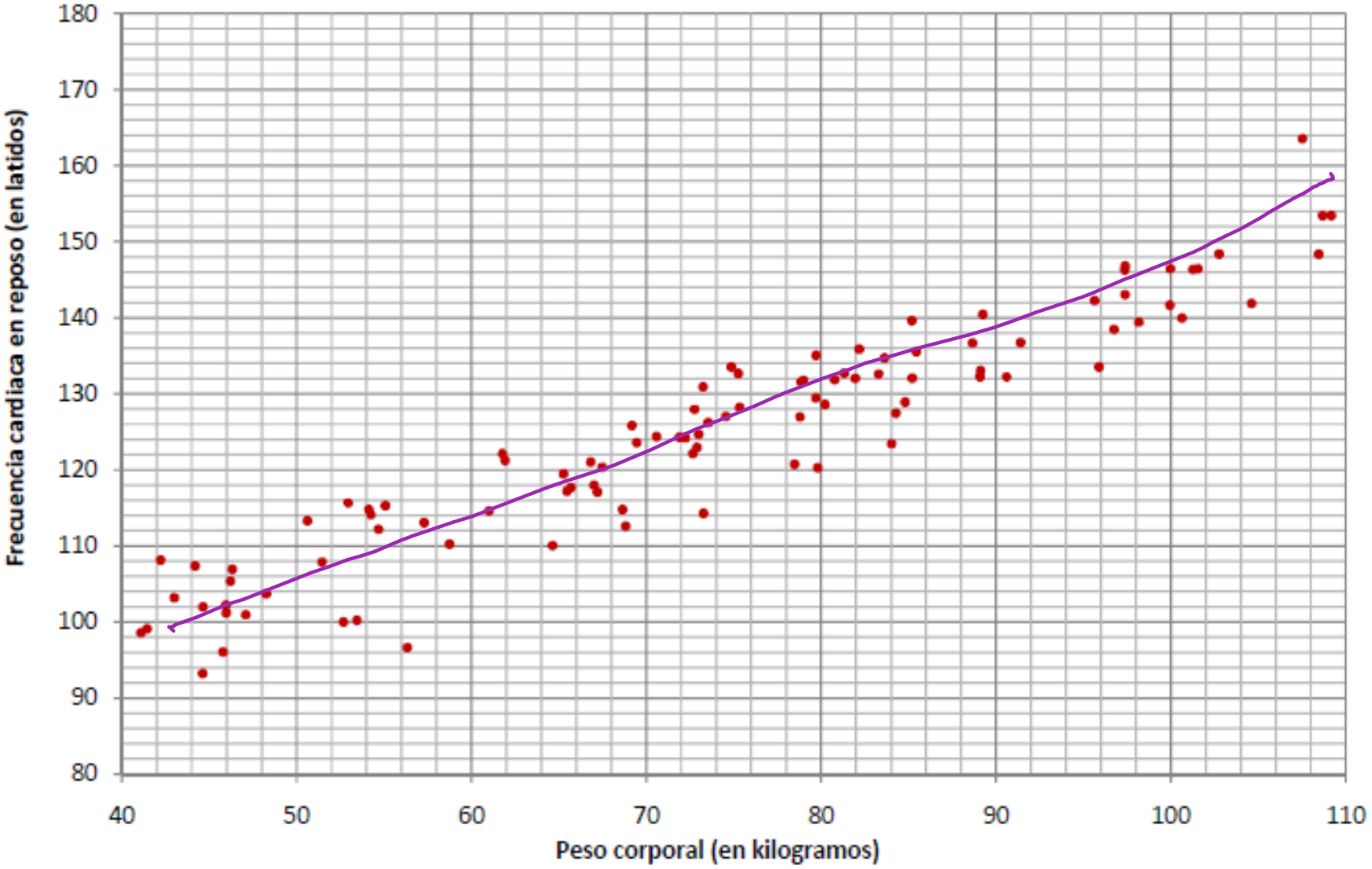
- Validación de la media cero y la ausencia de patrones.
  - Método gráfico (histograma, valores predichos versus residuales)

- Validación de la homoscedasticidad (homogeneidad de varianzas)
  - Método gráfico (valores predichos versus residuales)
  - Prueba de hipótesis de Breusch-Pagan
  - Si la hipótesis de homoscedasticidad no se verifica, es necesario usar mínimos cuadrados ponderados (ver más adelante).

- Validación de la independencia (no correlación serial)
  - Método gráfico (graficar los residuales versus el orden temporal).
  - Prueba de rachas (para aleatoriedad)

- Validación de la distribución normal con media cero
  - QQ plots
  - Prueba de normalidad

Ejercicio en R (I)

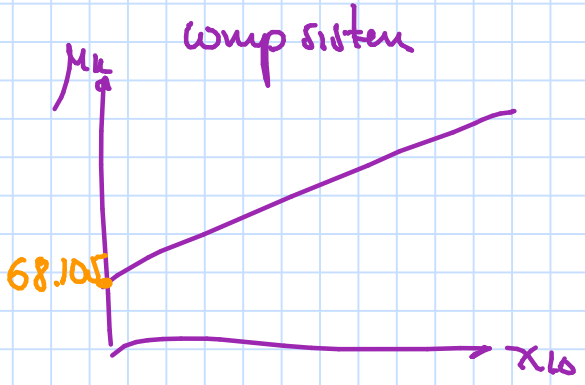


Y_k: "Frec. cardiaca en reposo (en lat/min) del k-ésimo individuo"

x_{1k}: "Peso (en kg) del k-ésimo individuo"

MODELO

$$\begin{cases} Y_k = \mu_k + e_k \\ \mu_k = \beta_0 + \beta_1 x_{1k} \\ e_k \overset{iid}{\sim} N(0, \sigma^2) \quad \forall k \end{cases}$$



$\hat{\beta}_0 = 68.105$ lat/min Si pudiéramos pensar en una persona con 0 kg de peso, el número PROMEDIO de lat/min en reposo sería de 68.105.

$\hat{\beta}_1 = 0.765$ lat/(min · kg) Aumento PROMEDIO de la frec. en reposo por cada kg adicional de peso

$\hat{\sigma} = 4.814$ lat/min Distancia promedio entre el valor observado de la frecuencia y el valor predicho