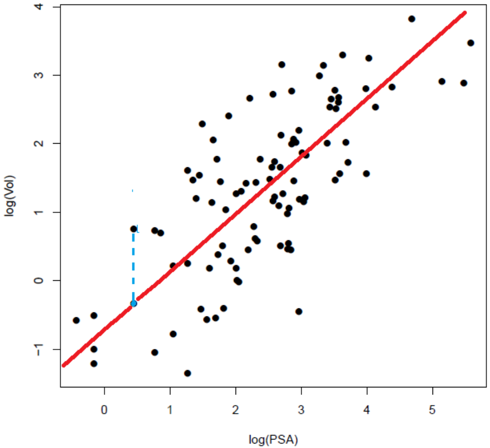
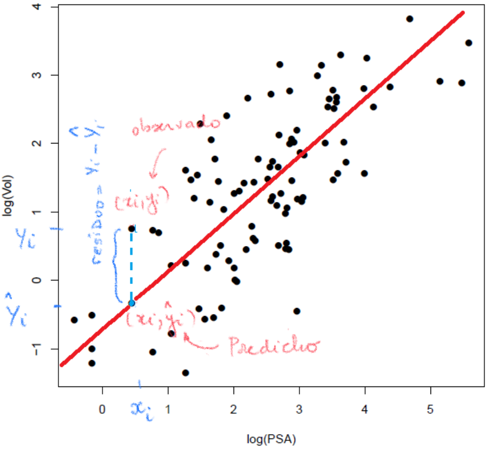
Predichos y Residuos



Predichos y Residuos



Predichos y Residuos

$$\mathbb{E}(y_i) = \mathbf{x}_i'\boldsymbol{\beta} = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}$$

$$\widehat{\mathbb{E}(y_i)} = \mathbf{x}_i'\widehat{\boldsymbol{\beta}} = \widehat{\beta}_0 + \widehat{\beta}_1 x_{i1} + \cdots + \widehat{\beta}_p x_{ip}$$

- i-ésimo valor predicho $\widehat{y}_i = \mathbf{x}_i'\widehat{\boldsymbol{\beta}} = \widehat{\beta}_0 + \widehat{\beta}_1 x_{i1} + \cdots + \widehat{\beta}_p x_{ip}$

- i-ésimo residuo $r_i = y_i - \widehat{y}_i = y_i - (\widehat{\beta}_0 + \widehat{\beta}_1 x_{i1} + \cdots + \widehat{\beta}_p x_{ip})$

Explicando la Variabilidad

$$\begin{aligned} \sum_{i=1}^{n}(y_i - \overline{y})^2 &= \sum_{i=1}^{n}(y_i - \widehat{y}_i)^2 + \sum_{i=1}^{n}(\widehat{y}_i - \overline{y})^2 \\ SCTotal &= SCResiduos + SCRegresion \end{aligned}$$

Coeficiente R^2 Múltiple

$$\begin{aligned}\sum_{i=1}^{n}(y_i-\overline{y})^2 &= \sum_{i=1}^{n}(y_i-\widehat{y}_i)^2+\sum_{i=1}^{n}(\widehat{y}_i-\overline{y})^2\\ SCTotal &= SCResiduos+SCRegresion\end{aligned}$$

$$R^2=\frac{\sum_{i=1}^{n}(\widehat{y}_i-\overline{y})^2}{\sum_{i=1}^{n}(y_i-\overline{y})^2}$$

Coeficiente R^2 Múltiple

$$\begin{aligned}\sum_{i=1}^{n}(y_i-\overline{y})^2 &= \sum_{i=1}^{n}(y_i-\widehat{y}_i)^2+\sum_{i=1}^{n}(\widehat{y}_i-\overline{y})^2\\ SCTotal &= SCResiduos+SCRegresion\end{aligned}$$

$$R^2=\frac{\sum_{i=1}^{n}(\widehat{y}_i-\overline{y})^2}{\sum_{i=1}^{n}(y_i-\overline{y})^2}=1-\frac{\sum_{i=1}^{n}(y_i-\widehat{y}_i)^2}{\sum_{i=1}^{n}(y_i-\overline{y})^2}$$

- $SCTotal$ variabilidad total de la variable Y.
- $SCRes$ variabilidad que no logra explicar el modelo.
- $SCReg$ variabilidad que explica el modelo lineal.

Coeficiente R^2 Múltiple

$$R^2 = \frac{\sum_{i=1}^{n}(\widehat{y}_i - \overline{y})^2}{\sum_{i=1}^{n}(y_i - \overline{y})^2} = 1 - \frac{\sum_{i=1}^{n}(y_i - \widehat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \overline{y})^2}$$

- Este coeficiente mide cuánto de la variabilidad es explicado por la regresión.
- $0 \leq R^2 \leq 1$
- $R^2 \approx 1 \Rightarrow$ el modelo explica muy bien
- $R^2 \approx 0 \Rightarrow$ el modelo provee un ajuste pobre

R^2 representa la proporción de variabilidad de la variable Y que es explicada por el modelo propuesto relativa a cuánta variabilidad tiene Y respecto de su media (es decir, a ser explicada sin aplicar el modelo).

Estimación en un nuevo $\mathbf{x}_o$

En un futuro valor $\mathbf{x}_o = (x_{o1}, \ldots, x_{op})^{\mathsf{t}}$ independiente de los demás datos, ¿qué esperamos observar? $\mathbb{E}(y_o) =?$

Estimación en un nuevo $\mathbf{x}_o$

En un futuro valor $\mathbf{x}_o = (x_{o1}, \ldots, x_{op})^{\mathsf{t}}$ independiente de los demás datos, ¿qué esperamos observar? $\mathbb{E}(y_o) = ?$

$$\psi = \mathbb{E}(y_o) = \mathbf{x}_o^{\mathsf{t}}\boldsymbol{\theta} \Rightarrow \widehat{\psi} = \widehat{\mathbb{E}(y_o)} = \mathbf{x}_o^{\mathsf{t}}\widehat{\boldsymbol{\theta}}$$

Notemos que $\widehat{\boldsymbol{\theta}}$ es insesgado, entonces el estimador propuesto será insesgado de $\mathbb{E}(y_o)$. (Ejercicio)

Estimación en un nuevo $\mathbf{x}_o$

También podríamos interesarnos realizar un IC para la esperanza de una nueva observación $\mathbf{x}_o$ independiente de las demás que cumpla el modelo

$$y_i = \theta_1 x_{i1} + \theta_1 x_{i2} + \cdots + \theta_p x_{ip} + \epsilon_i$$

en $\mathbf{x}_o = (x_{o1}, \ldots, x_{op})^{\mathsf{t}}$ donde $\epsilon_i \sim N(0, \sigma^2)$ independientes.

Como nuestro estimador es $\mathbf{x}_o^{\mathsf{t}} \widehat{\boldsymbol{\theta}}$, tenemos que

$$\mathbf{x}_o^{\mathsf{t}} \widehat{\boldsymbol{\theta}} \sim N(\mathbf{x}_o^{\mathsf{t}} \boldsymbol{\theta}, \sigma^2 \mathbf{x}_o^{\mathsf{t}} (\mathbf{X}^{\mathsf{t}} \mathbf{X})^{-1} \mathbf{x}_o)$$

y es independiente de

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

por lo tanto,

$$T = \frac{\mathbf{x}_o^{\mathsf{t}} \widehat{\boldsymbol{\theta}} - \mathbf{x}_o^{\mathsf{t}} \boldsymbol{\theta}}{S \sqrt{\mathbf{x}_o^{\mathsf{t}} (\mathbf{X}^{\mathsf{t}} \mathbf{X})^{-1} \mathbf{x}_o}} \sim t_{n-p}$$

Estimación en un nuevo $\mathbf{x}_o$

En consecuencia, tenemos el siguiente intervalo de confianza para el valor esperado de la respuesta en $\mathbf{x}_o$

$$\mathbf{x}_o^{\mathsf{t}}\widehat{\boldsymbol{\theta}} \pm t_{n-p,\frac{\alpha}{2}}\ S\sqrt{\mathbf{x}_o^{\mathsf{t}}(\mathbf{X}^{\mathsf{t}}\mathbf{X})^{-1}\mathbf{x}_o}$$

de nivel exacto $1-\alpha$.

Predicción en un nuevo $\mathbf{x}_o$

Podríamos estar interesados en la predicción de y_o, una nueva observación que cumpla el modelo, y en un intervalo para ella, que llamaremos intervalo de predicción o sea $\mathbb{P}(A \leq y_o \leq B) = 1 - \alpha$.

Observemos que el predictor de y_o es $\widehat{y}_o = \mathbf{x}_o^{\mathsf{t}} \widehat{\boldsymbol{\theta}}$. Aquí nos interesa el eeror de predicción y notemos que

$$E(\widehat{y}_o - y_o) = 0$$

¿Qué distribución tiene $\widehat{y}_o - y_o$?

Predicción en un nuevo $\mathbf{x}_o$

Tenemos que

$$\begin{aligned} \widehat{y}_o &\sim N(\mathbf{x}_o^{\mathsf{t}}\boldsymbol{\theta}, \sigma^2 \mathbf{x}_o^{\mathsf{t}}(\mathbf{X}^{\mathsf{t}}\mathbf{X})^{-1}\mathbf{x}_o) \\ y_o &\sim N(\mathbf{x}_o^{\mathsf{t}}\boldsymbol{\theta}, \sigma^2) \end{aligned}$$

y dado que y_o es independiente de las restantes y_i's con las que estimamos, entonces por la independencia entre estas dos normales queda que

$$\widehat{y}_o - y_o \sim N(0, \sigma^2(1 + \mathbf{x}_o^{\mathsf{t}}(\mathbf{X}^{\mathsf{t}}\mathbf{X})^{-1}\mathbf{x}_o))$$

Por lo tanto, el intervalo de predicción de nivel $1-\alpha$ estará dado por

$$\widehat{y}_o \pm t_{n-p,\frac{\alpha}{2}}\ S\sqrt{1 + \mathbf{x}_o^{\mathsf{t}}(\mathbf{X}^{\mathsf{t}}\mathbf{X})^{-1}\mathbf{x}_o}$$

¿Cómo es la longitud del intervalo de predicción respecto de la del intervalo para la media?

Covariables Aleatorias

Asumimos que tenemos:

$$(y_1, \mathbf{X}_1), \ldots, (y_n, \mathbf{X}_n)$$

- **Independencia:**
  $(y_1, \mathbf{X}_1), \ldots, (y_n, \mathbf{X}_n)$ son vectores independientes.

- **Esperanza Condicional:** $(y, \mathbf{X}) \sim (y_i, \mathbf{X}_i)$
  $\mathbb{E}(y \mid \mathbf{X} = \mathbf{x}) = \mathbf{x}^{\mathsf{t}} \boldsymbol{\theta}$

- **Homoscedasticidad:**
  $\mathbb{V}ar(y \mid \mathbf{X} = \mathbf{x}) = \sigma^2$

Covariables Aleatorias

Cuando asumimos normalidad, podemos escribir

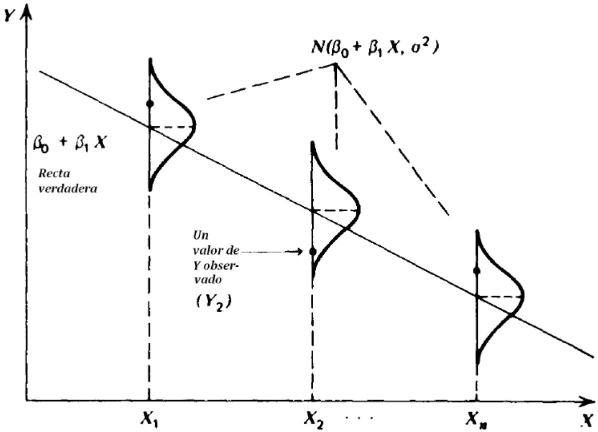
- $y_i \mid \mathbf{X}_i = \mathbf{x} \sim N(\mathbf{x}^{\mathsf{t}}\boldsymbol{\theta}, \sigma^2), i = 1, \ldots, n$

  o si lo pensamos en términos de la notación matricial

- $\mathbf{Y} \mid \mathbf{X} \sim N(\mathbf{X}\boldsymbol{\theta}, \sigma^2\mathbf{I})$

- Dado que al tomar la condicional a las covariables es como tomar las covariables como fijas, podemos aprovechar lo que hemos desarrollado hasta ahora.

Covariables Aleatorias



Trade-off sesgo-varianza

Para ponerlo en palabras:

- Cuando hablamos de sesgo tenemos en cuenta la cercanía al fenomeno real complejo que obtenemos al ajustar con un modelo lineal relativamente simple.

- Cuando hablamos de varianza de $\widehat{m}$ nos referimos a la variabilidad que tendrán nuestras estimaciones del modelo m cuando se basan en distintos conjuntos de datos.

Trade-off sesgo-varianza

Vayamos al ejemplo de R

Trade-off sesgo-varianza

- El ejemplo nos muestra que las predicciones basadas en el modelo lineal simple y en el cuadrático son estables, pero muy sesgadas.
- A la vez, las basadas en polinomios con grado más alto tienen mucho menor sesgo, pero una gran variabilidad.
- Esto pone de manifiesto que los polinomios de menor grado, al ser más rígidos, no pueden capturar la curvatura, pero tampoco el ruido de los datos. Los de mayor grado, al ser más flexibles captan la forma funcional, pero también el ruido.

Trade-off sesgo-varianza

Trade-off: los modelos más complejos tendrán menor sesgo, pero mayor variabilidad.

Trade-off sesgo-varianza

- El ejemplo nos muestra que las predicciones basadas en el modelo lineal simple y en el cuadrático son estables, pero muy sesgadas.
- A la vez, las basadas en polinomios con grado más alto tienen mucho menor sesgo, pero una gran variabilidad.
- Esto pone de manifiesto que los polinomios de menor grado, al ser más rígidos, no pueden capturar la curvatura, pero tampoco el ruido de los datos. Los de mayor grado, al ser más flexibles captan la forma funcional, pero también el ruido.

Trade-off sesgo-varianza

Trade-off: los modelos más complejos tendrán menor sesgo, pero mayor variabilidad.

¿Qué modelo elegimos?