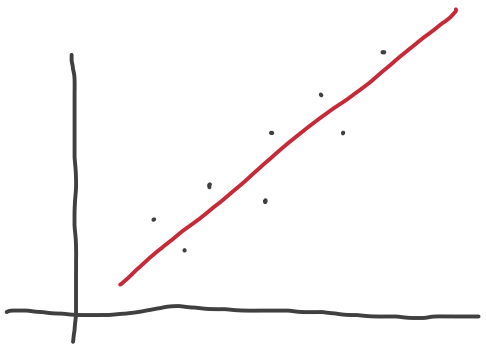
Regresión lineal II

jueves, 5 de mayo de 2022 14:28

Recordemos



Buscamos la recta de regresión:

$$Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_r X_r + \varepsilon \Rightarrow E(Y) = \beta_0 + \beta_1 X_1 + \ldots + \beta_r X_r$$

Tal que ε (los residuos) sean mínimos.

Para ello, escogemos

$$\hat{\beta} = (X' \cdot X)^{-1} X' Y$$

$$\varepsilon = Y - \hat{Y}$$

Coeficiente de determinación

Clase anterior

Descomposición de Bi's

Con R^2 podemos saber cómo de bien ajustan los datos para una futura predicción.

Ahora, nos preguntamos cuáles son las variables que tienen una aportación significativa en el ajuste.

Para ello, utilizamos las siguientes pruebas de hipótesis:

1.) ¿ Alguna variable es significativa ?

$H_0: \beta_1 = \beta_2 = \dots = \beta_r = 0$

$H_1: \exists\, \beta_i \neq 0 \quad i = 1, \dots, r$

Para comprobarlo: (ANOVA)

$$\frac{SC_{EXP}/r}{SC_{res}/(n-1-r)} \sim F_{r,n-1-r}$$

Es decir, si $\dfrac{SC_{EXP}/r}{SC_{res}/(n-1-r)} > F_{r,n-1,-r}$ entonces se rechaza la $H_0 \Rightarrow$ Alguna β es significativa.

2) ¿ La variable i es significativa ?

La variable X_i NO influye en el modelo $\Leftrightarrow \beta_i = 0$

$H_0: \beta_i = 0$

$H_A: \beta_i \neq 0$

Para comprobarlo, se cumple:

$$\frac{\hat{\beta}_i}{S_{\hat{\beta}_i}} \sim t_{n-1-r} \qquad \left(\frac{\bar{x} - m}{S/\sqrt{n}} \sim t_{n-1} \right)$$

Por tanto, si $\left|\dfrac{\hat{\beta}_i}{S_{\hat{\beta}_i}}\right| > t_{n-1-r}(\alpha/2)$ se rechaza H_0 y se deduce que la variable X_i influye sobre el modelo

Métodos para seleccionar variables: stepwise | (con) backward, fordward (sin)

Inferencia del modelo

} Intervalo de confianza: Intervalo entre el que se estima que estará un valor con cierta conf.

- Para β_i's:

Los intervalos de conf. para β_i con conf. $(1-\alpha)\cdot 100\%$ se definen como:

$$\hat{\beta}_i \pm \sqrt{\hat{var}(\beta_i)} \cdot \sqrt{(r+1)\cdot F_{r+1,\,n-r-1}(\alpha)} \qquad i = 1, \dots, r$$

$\sqrt{\hat{var}(\beta_i)}$ ↓ i-ésimo elemento de la diagonal de $s^2(X'X)^{-1}$

- Para $E(Y)$ (Intervalo de conf.)

$$E(Y) = \beta_0 + \beta_1 X_1 + \dots + \beta_r X_r$$

$$E(Y_0|X_0) = \beta_0 + \beta_1 x_{01} + \dots + \beta_r X_{0r} = X_0'\beta$$

El estimador de $X_0'\beta$ es $X_0'\hat{\beta}$.

Los intervalos de conf. para $E(Y_0|X_0)$ están dado por:

$$X_0'\hat{\beta} \pm t_{n-r-1}(\alpha/2)\sqrt{(X_0'(X'X)^{-1}X_0)s^2}$$

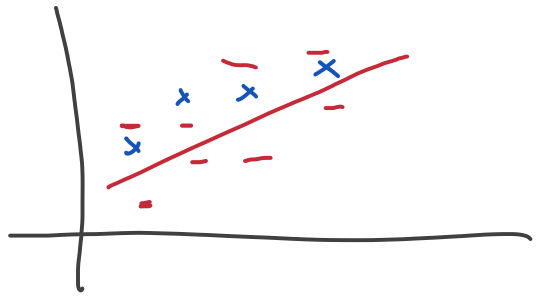
"Asegura que el rango incluirá (a un $(1-\alpha)\cdot 100\%$ de conf.) la respuesta media"

- Para $\hat{Y}$ (Intervalo de predicción)

Los intervalos de conf. para $\hat{Y}$ con conf. $(1-\alpha)\cdot 100\%$ están dados por:

$$X_0'\hat{\beta} \pm t_{n-r-1}(\alpha/2)\cdot\sqrt{s^2(1+X_0'(X'X)^{-1}X_0)}$$

$x_0 \pm t_{n-r-1}(\alpha/2)\, s\sqrt{1 + x_0^T (X^T X)^{-1} x_0}$



"Asegura que al $(1-\alpha)\cdot 100\%$ que este rango incluye el valor de la nueva obs"

Si cogieramos 100 muestras, en 95 de ellas el valor real estaría en el int. de conf.

- Los intervalos de confianza expresan la incertidumbre del muestreo en cantidades estimadas a partir de muchos puntos de datos. Cuantos más datos, menor incertidumbre de muestreo y, por tanto, más estrecho el intervalo.
- Los intervalos de predicción, además de la incertidumbre del muestreo, también expresan la incertidumbre en torno a un único valor, lo que los hace más amplios que los intervalos de confianza.

Tipos de variables

De momento sólo hemos trabajado con variables numéricas. Otros tipos:

* Variables categóricas (cualitativas):

P.e: Para predecir el valor de un apartamento podemos incluir el barrio.

Si tenemos k categorías (Chapinero, Usaquén...) => Creamos k-1 variables dummy.

Ejemplo: Resistencia de una viga

y: Resistencia

X_1: Densidad del hormigón

X_2 : Tipo de hormigón (A, B, C)

El modelo sería : Añadimos v. dummy

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 \Rightarrow Y = \beta_0 + \beta_1 X_1 + \beta_2 Z_1 + \beta_3 Z_2$$

Tipo de hormigón	A	B
A	0	0
B	1	0
C	0	1

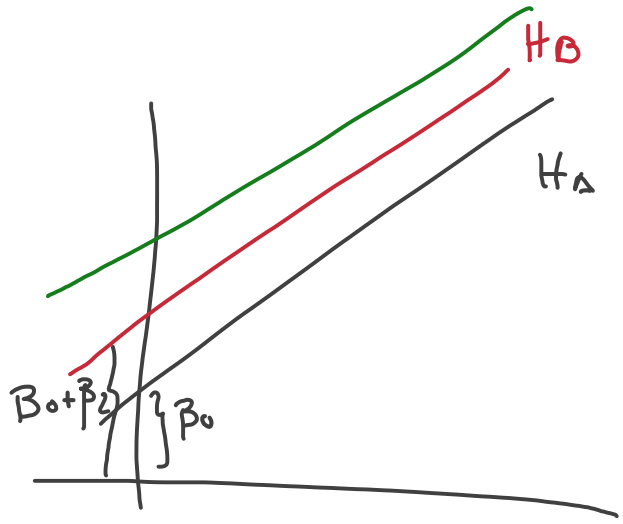
H_C

Por tanto:

$$E(Y \mid X_1, H_A) = \beta_0 + \beta_1 X_1$$
$$E(Y \mid X_1, H_B) = \beta_0 + \beta_1 X_1 + \beta_2$$
$$E(Y \mid X_1, H_C) = \beta_0 + \beta_1 X_1 + \beta_3$$

↓

Misma pendiente



β_2, β_3 : Dif. de resistencia media obtenida entre usar el hormigón B/C en vez de hormigón A

Otro tipo: variables cuadráticas, interacción de variables...

Validación del modelo

Supuestos:

i) Variables X_i's NO pueden estar relacionadas entre sí.

ii) $E(Y) = \beta' X$ (Def. Regresión)

iii) $\varepsilon \sim$ Normal

iv) $Var(\varepsilon) =$ cte (Homocedasticidad)

v) Errores (residuos) indeps.

Vamos a estudiarlos uno a uno.

i) Problema de correlación lineal entre v. explicativas

Problema:

Si hay dos variables corr. $\Rightarrow$ $X'X$ es singular $\Rightarrow$

$\Rightarrow$ No tiene inversa $\Rightarrow$

$\Rightarrow$ no se puede calcular β

Supongamos que están corr. de la siguiente manera:

$X_1 = 0.2\, X_2$

Podemos escribir:

$$E(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 = \beta_0 + \beta_1 \cdot 0.2\, X_2 + \beta_2 X_2$$

Con este modelo <u>NO</u> podríamos estimar el efecto verdadero de X_1 y podría afectar a la predicción final.

Aunque la correlación NO sea exacta, también afecta a los β_i's y el modelo tiene predicciones inexactos y muy sensibles.

Detección de la correlación

* Con R(m. corr), $|r_{ij}| > 0.7$
* Con VIF (variance inflation factor) $\Rightarrow$ Diag. de R^{-1}

  $VIF_i = \text{diag}_i\, R^{-1}$

  $\hookrightarrow$ la dif. es que el VIF mira la corr. de una variable

con todas las demás, NO individualmente.

- Si VIF > 10 => Alta colinealidad => Eliminar del modelo

* Utilizar PCA

iii) Los residuos siguen una normal.

Problema: Si los residuos NO siguen una normal =>

=> Estimador de min. cuadrados ≠ MLE

=> β poco eficientes

Detección: QQ plot, test de normalidad

Nota: Suele pasar si hay datos anómalos.

iv) Heterocedasticidad

Problema: Los test de hipótesis NO se pueden aplicar

Resultados erróneos

Detección: Gráfico de residuos no disperso (no aleat.)

v) Falta de indeps. en los residuos

Detección: No haya patrones en residuals / fitted

Resumiendo:

- ¿Para qué voy a utilizar este análisis? Para predecir una variable numérica a partir de otras
- ¿Cómo voy a interpretar los resultados? Con R^2. Los

- ¿Cómo voy a interpretar los [illegible] coef. de. B, ...
- ¿En qué casos lo voy a poder utilizar? Cuando se cumplan los supuestos i-v.