Primer taller

jueves, 10 de abril de 2025 1:56 p. m.

De respuesta a las preguntas formuladas a continuación en base a la teoría tratada en clase. **Provea una interpretación de ser necesario.**

1. Considere el siguiente modelo de regresión lineal múltiple con k variables regresoras, $p=(k+1)$ parámetros asociados $Y_i=\beta_0+\beta_1X_{i1}+\beta_2X_{i2}+\ldots+\beta_kX_{ik}+\varepsilon_i$, $\varepsilon_i \overset{iid}{\sim} N(0,\sigma^2)$.

   - (a) Escriba el modelo de forma matricial junto con sus supuestos. **Especifique las dimensiones de cada componente.**

• $\boxed{Y_i=\beta_0+\beta_1X_{i1}+\ldots+\beta_kX_{ik}+\varepsilon_i;\ \varepsilon_i \overset{iid}{\sim} N(0,\sigma^2)}$; $\boxed{Y_i=\beta_0+\beta_1X_{i1}+\varepsilon_i;\ \varepsilon_i \overset{iid}{\sim} N(0,\sigma^2)}$ (2)

↳ $\mathbb{R}^p$

• X_{ij} → Covariables; $j=1,\ldots,k$ → V. regresora.

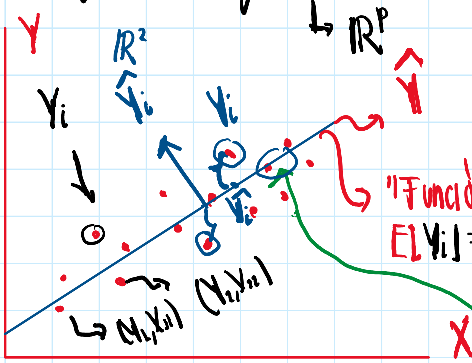
V. respuesta

Observación $i=1,\ldots,n$

Cantidad dada

• k: # Covariables

$p=(k+1)$: # parámetros.



"Función" $E[Y_i]=\beta_0+\beta_1X_{i1}$

$\theta : \hat{\theta}$: $\boxed{\hat{E}[Y_i]=\hat{\beta}_0+\hat{\beta}_1X_{i1}}$

Recta ajustada

$\boxed{Y_i=\beta_0+\beta_1X_{i1}+\ldots+\beta_kX_{ik}+\varepsilon_i;\ \varepsilon_i \overset{iid}{\sim} N(0,\sigma^2)}$;

$$\begin{cases} Y_1=\beta_0+\beta_1X_{11}+\ldots+\beta_kX_{1k}+\varepsilon_1 \\ \vdots \\ Y_n=\beta_0+\beta_1X_{n1}+\ldots+\beta_kX_{nk}+\varepsilon_n \end{cases}; \quad \underline{Y}_n=\begin{bmatrix}Y_1\\ \vdots \\ Y_n\end{bmatrix}_{n\times 1}; \quad \underline{\varepsilon}_n=\begin{bmatrix}\varepsilon_1\\ \vdots \\ \varepsilon_n\end{bmatrix}_{n\times 1}; \quad X=\begin{bmatrix}1 & X_{11} & \ldots & X_{1k}\\ 1 & & \ddots & \\ \vdots & & & \\ 1 & & & X_{nk}\end{bmatrix}_{n\times p}$$

$n\times(k+1)$

• $\boxed{\underline{Y}_n = X\underline{\beta}+\underline{\varepsilon}_n}$; $\underline{\varepsilon}_n \sim N_n(\underline{0}_n, \sigma^2 I)$ $\underline{\beta}=\begin{bmatrix}\beta_0\\ \vdots \\ \beta_k\end{bmatrix}_{(k+1)\times 1}$; $\boxed{\hat{\theta}\rightarrow\theta}$

• $\boxed{\hat{\underline{\beta}}\rightarrow\underline{\beta}}$ (1). MLE: Estadístico (distribución) = Vero. = log... } Minimizar el error

(2). MCO: Matemático (no distribución)

• Con los dos métodos: la estimación $\hat{\underline{\beta}}$. ; { (1). MLE: $\hat{\sigma}^2=\frac{SSE}{n}$ (sesgada). n → # observación

Recta regresión lineal simple: (2). MCO: $\hat{\sigma}^2=\frac{SSE}{n-p}$ (insesgado).

• $SSE=\sum_{i=1}^{n}(y_i-\hat{y}_i)^2$ (minimizarlo).

• En el caso múltiple: El modelo ajustado: $\boxed{\hat{\underline{Y}}_n = X\hat{\underline{\beta}}}$ ≈ $\boxed{\hat{E}[Y_i]=\hat{\beta}_0+\hat{\beta}_1X_{i1}}$

Recta ajustada

$\boxed{SSE=(\underline{Y}_n-\hat{\underline{Y}}_n)^T(\underline{Y}_n-\hat{\underline{Y}}_n)}=(\underline{Y}_n-X\hat{\underline{\beta}})^T(\underline{Y}_n-X\hat{\underline{\beta}})$

• Estimación Varianza (MSE): $\hat{\sigma}^2=\frac{SSE}{n-p}$ (insesgado); $\boxed{MSE=\frac{(\underline{Y}_n-X\hat{\underline{\beta}})^T(\underline{Y}_n-X\hat{\underline{\beta}})}{n-p}}$

Minimizar

• Errores: → No son cuantificables

- Errores: $\rightarrow$ No son cuantificables

- $\boxed{Y_n = X\beta + \varepsilon_n}$; $\varepsilon_n \sim N_n(0_n, \sigma^2 I)$; Residuales: $\hat{\varepsilon}_n \rightarrow \varepsilon_n$ Cuantificar

Minimizar

- $\boxed{\hat{\varepsilon}_n = Y_n - \hat{Y}_n \; ; \; \hat{\varepsilon}_n = Y_n - X\hat{\beta}}$

$\boxed{\theta ; \hat{\theta} \longleftrightarrow E[\hat{\theta}] = \theta}$ Insesgado.

- (b) Demuestre que el estimador $\hat{\boldsymbol{\beta}}$ que se obtiene a través del método de mínimos cuadrados es un estimador insesgado para $\boldsymbol{\beta}$. Analice $\hat{\boldsymbol{\beta}}$.

- $E[\hat{\beta}]$; A partir del MCO: $\boxed{\hat{\beta} = (X^T X)^{-1} X^T Y_n}$; $E[\underbrace{(X^T X)^{-1} X^T}_{\text{Constante}} Y_n] = (X^T X)^{-1} X^T E[Y_n]$

$\boxed{Y_n \sim N_n(X\beta, \ldots)}$: $E[\hat{\beta}] = \underbrace{(X^T X)^{-1}}_{A^{-1}} \underbrace{X^T X}_{A} \beta$: $A^{-1} \cdot A = I$: $\boxed{E[\hat{\beta}] = I \cdot \beta = \beta}$

2. Determine el valor de verdad de las siguientes afirmaciones.

   - (a) Los modelos de regresión múltiple (1).$\mathrm{Y}_i = \beta_0 + \sum_{j=1}^{k} \beta_j (X_{ij})^{1/2} + \varepsilon_i$; (2).$\ln(\mathrm{Y}_i) = \beta_0 + \sum_{j=1}^{k} \beta_j \ln(X_{ij}) + \varepsilon_i; \varepsilon_i \overset{iid}{\sim} N(0, \sigma^2)$ no son modelos lineales, puesto que se altera la forma de la superficie de respuesta. Falsa

   - (b) Bajo los supuestos del modelo de regresión lineal múltiple, $\varepsilon_i \overset{iid}{\sim} N(0, \sigma^2)$, el estimador para los parámetros $\hat{\beta}$ obtenido a través del método de mínimos cuadrados ordinarios es el mismo que el de máxima verosimilitud, así como para la varianza $\hat{\sigma}^2$. **Explique.** Falsa

(1). MLE
(2). MCO

(1). MLE: $\hat{\sigma}^2 = \dfrac{SSE}{n}$ (sesgada). $n \rightarrow$ # observación

(2). MCO: $\hat{\sigma}^2 = \dfrac{SSE}{n-p}$ (insesgado).

Matriz Singular: Matriz invertible

- Para obtener $\hat{\beta}$ con MCO.

$\boxed{H : (\text{Hat}) : X(X^T X)^{-1} X^T}$

$\boxed{\hat{Y}_n = X\hat{\beta}} = X(X^T X)^{-1} X^T Y_n$ M. Hat.

- $\boxed{H^n = H \quad \forall n \in \mathbb{N}}$

   - (c) Se requiere que la matriz $(\mathbf{X}^t\mathbf{X})$ sea singular. Además, la matriz $\mathbf{H}$ es simétrica e idempotente, al igual que $(\mathbf{I_n} - \mathbf{H})$. Verdadera

   - (d) La matriz de varianzas- covarianzas siempre es simétrica respecto a su diagonal principal, además, siempre tiene unos en su diagonal principal. Falsa

   - (e) El parámetro β_0, como la respuesta media de Y es únicamente interpretable si el conjunto de observaciones $(X_1, X_2, ..., X_k) = (0, 0, ..., 0)$ está incluido en el rango experimental. Verdadera

   - (f) Los parámetros $\beta_j, j = 1, ..., k$ indican el cambio en la respuesta media de Y por unidad de incremento en la respectiva variable X_j.

- Matriz Var- Cov:

$\sigma^2 I_{n \times n}$; $\begin{bmatrix} \sigma^2 & 0 & 0 & \\ 0 & \sigma^2 & 0 & \\ 0 & 0 & \sigma^2 & \\ \vdots & & & \ddots \\ & & & & \sigma^2 \end{bmatrix}_{n \times n}$

$\boxed{\text{Afirmación incompleta:}}$

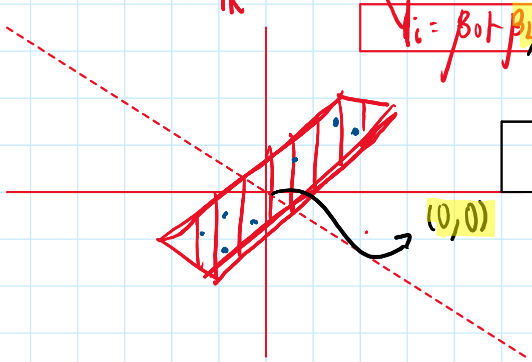
- Mantenido, los demás variables regresoras fijas (βj son efectos parciales)

R^3

$\boxed{Y_i = \beta_0 + \cancel{\beta_1 X_{i1}} + \cancel{\beta_2 X_{i2}} + \varepsilon_i}$

$\Downarrow$

$\boxed{\hat{Y}_i = \hat{\beta}_0}$



(0,0)

1. Determine cuál es la matriz de diseño $(\mathbf{X})$ para este problema en específico. Calcule el vector de parámetros estimados $\hat{\boldsymbol{\beta}}$. Brinde una interpretación adecuada.

Y: Performance
X_1: Strength

1. Determine cuál es la matriz de diseño ($\mathbf{X}$) para este problema en específico. Calcule el vector de parámetros estimados $\hat{\boldsymbol{\beta}}$. Brinde una interpretación adecuada.

$$\boxed{Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \varepsilon_i;\ \varepsilon_i \overset{iid}{\sim} N(0,\sigma^2)}$$

Y: Performance
X_1: Strength
X_2: Skills
X_3: Speed

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{i1} + \hat{\beta}_2 X_{i2} + \hat{\beta}_3 X_{i3} \ : \ \widehat{Perf}_i = \hat{\beta}_0 + \hat{\beta}_1 St_i + \hat{\beta}_2 Sk_i + \hat{\beta}_3 Sp_i$$

- $\hat{Y}_n = X\hat{\beta}$;

$n = 500$; $k = 3$
$p = 4$

Y	X_1	X_2	X_3
Performance	Strength	Skills	Speed
337.4335	60.22040	37.55007	54.67685
418.0430	45.23407	72.26675	53.60737
413.5214	32.09499	86.78852	44.96376
378.6311	63.44574	44.25407	53.85836
428.8907	49.40831	93.26630	43.28241
327.9498	23.50896	42.72160	49.48621
439.1306	59.73309	90.85926	45.66127
387.3193	56.77144	56.27796	61.04297

$$Y_1 = \beta_0 + \beta_1 X_{11} + \beta_2 X_{12} + \beta_3 X_{13} + \varepsilon_i$$
$$\vdots$$
$$Y_{500} = \dots$$

- Con la función $lm(Y \sim X_1 + X_2 + X_3) = \hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{i1} + \hat{\beta}_2 X_{i2} + \hat{\beta}_3 X_{i3}$

```
> X
   (Intercept)       X1       X2       X3
1            1 60.22040 37.55007 54.67685
2            1 45.23407 72.26675 53.60737
3            1 32.09499 86.78852 44.96376
4            1 63.44574 44.25407 53.85836
5            1 49.40831 93.26630 43.28241
6            1 23.50896 42.72160 49.48621
7            1 59.73309 90.85926 45.66127
8            1 56.77144 56.27796 61.04297
9            1 23.36111 88.29830 51.86923
10           1 26.98301 31.57627 62.54522
11           1 42.62999 57.37396 64.38511
12           1 35.95182 74.66441 56.18818
```

500 x4

$$X = \begin{bmatrix} 1 & X_{11} & \dots & X_{1k} \\ 1 & & \ddots & \\ \vdots & & & \\ 1 & X_{n1} & & X_{nk} \end{bmatrix}_{n \times p} \quad n \times (k+1)$$

$$\boxed{\hat{\beta} = (X^T X)^{-1} X^T Y_n}$$

B = solve(t(X) %*% X) %*% t(X) %*% Y

```
                  [,1]
(Intercept) 70.8022541
X1           0.8846324
X2           1.8736830
X3           3.0450232
```

$\hat{\beta}_0$, $\hat{\beta}_1$, $\hat{\beta}_2$, $\hat{\beta}_3$

$$\hat{\beta} = \begin{bmatrix} \beta_0 \\ \vdots \\ \beta_k \end{bmatrix}_{(k+1) \times 1}$$

- $\hat{\beta}_1$: Por un incremento unitario en la fuerza de un estudiante, el cambio promedio en su rendimiento es de 0,884 unidades en presencia de las demás covariables (manteniendo las demás covariables fijas)

2. Calcule la estimación de la respuesta media $\hat{\boldsymbol{y}}$.

- $\hat{Y}_n = X\hat{\beta}$ ~

```
> Y_gorro <- X %*% B
> Y_gorro
       [,1]
1  360.9244
2  409.4584
3  398.7244
4  373.8465
5  421.0579
6  322.3324
7  432.9252
8  412.3484
9  414.8543
10 344.2879
11 412.0689
```

$\hat{Y}_1$
$\hat{Y}_2$
$\vdots$
$\hat{Y}_{500}$

3. Calcule el vector de los errores estimados $\hat{\boldsymbol{\varepsilon}}$.

- $\boxed{\hat{\varepsilon}_n = Y_n - \hat{Y}_n \ ; \ \hat{\varepsilon}_n = Y_n - X\hat{\beta}}$

```
> errores_estimados = Y - Y_gorro
> errores_estimados
          [,1]
1  -23.49083076
2    8.58461340
3   14.79701156
4    4.78466812
5    7.83280779
6    5.61737708
7    6.20544775
8  -25.02909777
9    9.96151621
10   5.76379508
11  15.88812484
12  -3.46925923
```

$\hat{\varepsilon}_1$
$\hat{\varepsilon}_2$
$\vdots$
$\hat{\varepsilon}_{500}$

4. Construya el valor de la estimación para la varianza $\hat{\sigma}^2$.

4. Construya el valor de la estimación para la varianza $\hat{\sigma}^2$.

$\hat{\sigma}^2 = \frac{SSE}{n-p}$ (insesgado) : $\boxed{MSE = \frac{(Y_n - X\hat{\beta})^T (Y_n - X\hat{\beta})}{n-p}}$ $\hat{\varepsilon}_n = Y_n - X\hat{\beta}$

$MSE = \frac{\hat{\varepsilon}_n^T \cdot \hat{\varepsilon}_n}{n-p}$

```
> MSE = (t(errores_estimados) %*% errores_estimados)/(n- p)
> MSE
         [,1]
[1,] 103.6827
```

```
Coefficients:
            Estimate
(Intercept) 70.80225
X1           0.88463
X2           1.87368
X3           3.04502
```

$\approx$ $\boxed{\text{Función Summary}}$