1a. Prova - Estat para CC

Renato Assunção - DCC, UFMG

Março de 2014

1. Considere um modelo de regressão linear em que a matriz de desenho $\mathbf{X}$ é de dimensão $n \times 1$ com apenas uma única coluna. Esta coluna é a coluna de 1's. Isto é,

$$\mathbf{X} = \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix}$$

- Explique geometricamente o que é o espaço $\mathcal{M}(\mathbf{X})$ neste modelo particular.
- Obtenha a matriz $\mathbf{H}$ de projeção ortogonal de $\mathbf{Y}$ no espaço $\mathcal{M}(\mathbf{X})$.
- Obtenha o vetor projetado $\hat{\mathbf{Y}} = \mathbf{HY}$.
- Qual a expressão de de $\hat{\boldsymbol{\beta}}$?
- O que é o vetor de resíduos $\mathbf{r}$ neste caso?

Solução:

- $\mathcal{M}(\mathbf{X}) = \{c\mathbf{1} : c \in \mathbb{R}\}$. Isto é, $\mathcal{M}(\mathbf{X})$ é formado pelos múltiplos do vetor $\mathbf{1}$.
- Por definição,

$$\mathbf{H} = \mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}' = \mathbf{1}(\mathbf{1}'\mathbf{1})^{-1}\mathbf{1}' \quad \text{e} \quad \mathbf{1}'\mathbf{1} = (1,\ldots,1)\begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix} = n$$

  Então

$$\mathbf{H} = \frac{1}{n}\mathbf{1}\mathbf{1}' = \frac{1}{n}\begin{bmatrix} 1 & 1 & \ldots & 1 \\ 1 & 1 & \ldots & 1 \\ \vdots & \vdots & \ldots & \vdots \\ 1 & 1 & \ldots & 1 \end{bmatrix}$$

- Temos $\hat{\mathbf{Y}} = \mathbf{HY} = \bar{Y}\mathbf{1}$ onde $\bar{Y} = (Y_1 + \ldots + Y_n)/n$.
- $\hat{\boldsymbol{\beta}} = \bar{Y}$
- Temos

$$\mathbf{r} = \mathbf{Y} - \bar{Y}\mathbf{1} = \begin{bmatrix} Y_1 - \bar{Y} \\ Y_2 - \bar{Y} \\ \vdots \\ Y_n - \bar{Y} \end{bmatrix}$$

---

2. Considere o modelo de regressão usual com matriz de desenho $\mathbf{X}$ de dimensão $n \times p$ cuja primeira coluna é o vetor $\mathbf{1}$. Mostre que a soma dos resíduos $\sum_i r_i$ é igual a zero.

   Solução: O vetor de resíduos é dado por $\mathbf{r} = (\mathbf{I} - \mathbf{H})\,\mathbf{Y}$ onde $\mathbf{H}$ é a matriz de projeção ortogonal no espaço das combinações lineares das colunas de $\mathbf{X}$.

   A soma $\sum_i r_i$ é igual ao produto interno dos vetores $\mathbf{1}$ e $\mathbf{r}$:

$$\begin{aligned} r_1 + \ldots + r_n &= \mathbf{r}'\mathbf{1} \\ &= \mathbf{Y}'(\mathbf{I} - \mathbf{H})'\,\mathbf{1} \\ &= \mathbf{Y}'(\mathbf{I} - \mathbf{H})\,\mathbf{1} \quad \text{pois } \mathbf{H} \text{ é simétrica} \\ &= \mathbf{Y}'(\mathbf{1} - \mathbf{H1}) \\ &= \mathbf{Y}'(\mathbf{1} - \mathbf{1}) \\ &= 0 \end{aligned}$$

   A penúltima passagem é justificada pois $\mathbf{1}$ pertence ao espaço $\mathcal{M}(\mathbf{X})$ já que $\mathbf{1}$ é umas das colunas de $\mathbf{X}$. Assim, a projeção ortogonal de $\mathbf{1}$ em $\mathcal{M}(\mathbf{X})$ é o próprio $\mathbf{1}$.

---

3. Mostre que o vetor de resíduos $\mathbf{r}$ é ortogonal ao vetor ajustado $\hat{\mathbf{Y}}$ e conclua que eles são vetores aleatórios independentes. Assuma o modelo de regressão usual com matriz de desenho $\mathbf{X}$ $n \times p$ cuja primeira coluna é o vetor $\mathbf{1}$.

   Solução: Temos

$$\begin{aligned} \langle \mathbf{r}, \hat{\mathbf{Y}} \rangle &= \mathbf{r}'\hat{\mathbf{Y}} \\ &= \mathbf{Y}'(\mathbf{I} - \mathbf{H})'\,\mathbf{HY} \\ &= \mathbf{Y}'(\mathbf{I} - \mathbf{H})\,\mathbf{HY} \quad \text{pois } \mathbf{H} \text{ é simétrica} \\ &= \mathbf{Y}'(\mathbf{H} - \mathbf{H}^2)\mathbf{Y} \\ &= \mathbf{Y}'(\mathbf{H} - \mathbf{H})\mathbf{Y} \\ &= 0 \end{aligned}$$

   Como $\mathbf{r}$ e $\hat{\mathbf{Y}}$ são transformações lineares do vetor gaussiano multivariado $\mathbf{Y}$, a distribuição conjunta do vetor $(\mathbf{r}, \hat{\mathbf{Y}})$ de dimensão $2n$ é uma normal gaussiana com matriz de covariância $2n \times 2n$. O bloco $(1, 2)$ de dimensão $n \times n$ desta matriz representa a matriz de covariância entre $\mathbf{r}$ e $\hat{\mathbf{Y}}$ e ele é dado por

$$\begin{aligned} \text{Cov}(\mathbf{r}, \hat{\mathbf{Y}}) &= \text{Cov}\left((\mathbf{I} - \mathbf{H})\mathbf{Y}, \mathbf{H})\,\mathbf{Y}\right) \\ &= (\mathbf{I} - \mathbf{H})\text{Cov}(\mathbf{Y}, \mathbf{Y})\mathbf{H}' \\ &= (\mathbf{I} - \mathbf{H})\sigma^2\mathbf{I}\mathbf{H}' \\ &= \sigma^2(\mathbf{I} - \mathbf{H})\mathbf{H} \\ &= \sigma^2(\mathbf{H} - \mathbf{H}^2) \\ &= \sigma^2(\mathbf{H} - \mathbf{H}) \\ &= \mathbf{0} \end{aligned}$$

   No caso gaussiano, covariância (ou correlação) nula implica independência. Assim, conclímos que $\mathbf{r}$ e $\hat{\mathbf{Y}}$ são independentes.

---

4. Seja $X_1, \ldots X_n$ v.a.'s i.i.d. $N(0, 1)$. Defina $Y_1 = X_1$ e $Y_i = X_i - X_{i-1}$ para $i = 2, \ldots, n$. Encontre a distribuição do vetor $\mathbf{Y} = (Y_1, \ldots, Y_n)'$.

Solução: O vetor $\mathbf{X}$ segue uma distribuição gaussiana multivariada $N_n(\mathbf{0}, \mathbf{I})$ e

$$\mathbf{Y} = \begin{bmatrix} Y_1 \\ Y_2 \\ Y_3 \\ Y_4 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 & \dots & 0 & 0 \\ -1 & 1 & 0 & 0 & \dots & 0 & 0 \\ 0 & -1 & 1 & 0 & \dots & 0 & 0 \\ 0 & 0 & -1 & 1 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \dots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \dots & -1 & 1 \end{bmatrix} \begin{bmatrix} X_1 \\ X_2 \\ X_3 \\ X_4 \\ \vdots \\ X_n \end{bmatrix} = \mathbf{AX}$$

Portanto, $\mathbf{Y}$ também é um vetor gaussiano $N_n(\mathbf{A0}, \mathbf{AIA}') = N_n(\mathbf{0}, \mathbf{AA}')$. A matriz de covariância é dada por

$$\mathbf{AA}' = \begin{bmatrix} 1 & -1 & 0 & 0 & \dots & 0 & 0 \\ -1 & 2 & -1 & 0 & \dots & 0 & 0 \\ 0 & -1 & 2 & -1 & \dots & 0 & 0 \\ 0 & 0 & -1 & 2 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \dots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \dots & -1 & 2 \end{bmatrix}$$

---

5. Se $X_1, \dots, X_n$ são v.a.'s tais que $\mathbb{V}(X_1) = \sigma^2$ e satisfazendo $X_{i+1} = \rho X_i$ onde $\rho \in (0, 1)$ é uma constante e $i = 2, \dots, n$. Encontre a matriz de covariância $\mathbb{V}(\mathbf{X})$ do vetor $\mathbf{X}$.

   Solução: Como $X_2 = \rho X_1$, $X_3 = \rho X_2 = \rho^2 X_1$ e, em geral, $X_i = \rho^{i-1} X_1$ para $i > 1$, temos as variâncias dadas por

   $$\mathbb{V}(X_i) = \mathbb{V}(\rho^{i-1} X_1) = \rho^{2(i-1)} \mathbb{V}(X_1) = \rho^{2(i-1)} \sigma^2 \, .$$

   Quanto às covariâncias, temos

   $$\text{Cov}(X_i, X_j) = \text{Cov}(\rho^{i-1} X_1, \rho^{j-1} X_1) = \rho^{i-1} \text{Cov}(X_1, X_1) \rho^{j-1} = \rho^{i+j-2} \mathbb{V}(X_1) = \rho^{i+j-2} \sigma^2$$

---

6. No modelo abaixo,

   $$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

   onde a matriz de desenho $\mathbf{X}$ tem dimensão $n \times p$, o vetor $\boldsymbol{\varepsilon}$ tem distribuição normal multivariada com vetor esperado $\boldsymbol{\mu} = (0, \dots, 0)'$ e matriz de covariância igual a matriz diagonal com elementos

   $$\text{diag}(\mathbb{V}(\boldsymbol{\varepsilon}) = \sigma^2 (c_1, c_2, \dots, c_n)$$

   onde $c_i > 0$ são constantes CONHECIDAS. Assim, os erros não possuem variância constante.

   Ignorando a matriz de covariância diferente da usual, aplica-se a fórmula matricial para obter o estimador de mínimos quadrados de $\boldsymbol{\beta}$. Mostre que este estimador é não viciado para estimar $\boldsymbol{\beta}$. EXTRA BONUS: Obtenha a matriz de covariância do estimador $\hat{\boldsymbol{\beta}}$.

   Solução: Como os erros $\boldsymbol{\varepsilon}$ possuem esperança zero, temos

   $$\mathbb{E}(\mathbf{Y}) = \mathbb{E}\left(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}\right) = \mathbf{X}\boldsymbol{\beta} + \mathbb{E}\left(\boldsymbol{\varepsilon}\right) = \mathbf{X}\boldsymbol{\beta}$$

   e portanto

   $$\mathbb{E}\left(\hat{\boldsymbol{\beta}}\right) = \mathbb{E}\left(\left(\mathbf{X}'\mathbf{X}\right)^{-1} \mathbf{X}'\mathbf{Y}\right) = \left(\mathbf{X}'\mathbf{X}\right)^{-1} \mathbf{X}' \mathbb{E}\left(\mathbf{Y}\right) = \left(\mathbf{X}'\mathbf{X}\right)^{-1} \mathbf{X}'\mathbf{X}\boldsymbol{\beta} = \boldsymbol{\beta}$$

   BONUS:

   $$\mathbb{V}\left(\hat{\boldsymbol{\beta}}\right) = \left(\mathbf{X}'\mathbf{X}\right)^{-1} \mathbf{X}' \mathbb{V}(\mathbf{Y}) \mathbf{X} \left(\mathbf{X}'\mathbf{X}\right)^{-1} = \left(\mathbf{X}'\mathbf{X}\right)^{-1} \mathbf{X}'\mathbf{V}\mathbf{X} \left(\mathbf{X}'\mathbf{X}\right)^{-1}$$

---

7. Suponha que $\mathbf{Y} = (Y_1, Y_2)$ seja uma vetor aleatório gaussiano com valor esperado $\boldsymbol{\mu} = (10, 15)$ e matriz de covariância

$$\mathbb{V}(\mathbf{Y}) = \begin{bmatrix} 4 & 1.6 \\ 1.6 & 1 \end{bmatrix}$$

Dois pontos que fazem da amostra são $\mathbf{y}_1 = (12, 15)$ e $\mathbf{y}_2 = (10, 17)$, ambos a uma distância euclidiana de 2 unidades de $\boldsymbol{\mu}$. Qual deles está a uma distância estatística maior de $\boldsymbol{\mu}$?

Solução:

$$d^2(\mathbf{y}_1, \boldsymbol{\mu}) = (\mathbf{y}_1 - \boldsymbol{\mu})'\mathbb{V}(\mathbf{Y})^{-1}(\mathbf{y}_1 - \boldsymbol{\mu}) = (2, 0)\mathbb{V}(\mathbf{Y})^{-1} \begin{pmatrix} 2 \\ 0 \end{pmatrix} = (2)^2 \left[\mathbb{V}(\mathbf{Y})^{-1}\right]_{11}$$

Uma fórmula similar para o segundo ponto.

---

8. Se colocarmos mais atributos na matriz $\mathbf{X}$ (sempre mantendo as colunas lineamente independentes), podemos garantir que o R^2 vai sempre aumentar. Explique porque isto acontece.

Solução: Seja $\mathbf{X}^*$ a matriz de desenho aumentada. O espaço vetorial $\mathcal{M}(\mathbf{X}^*)$ inclui $\mathcal{M}(\mathbf{X})$ como sub-espaço vetorial. Assim, minimizar a distância entre $\mathbf{Y}$ e um elemento de $\mathbf{X}^*$ inclui todas as possíveis soluções restritas apenas a $\mathcal{M}(\mathbf{X})$ e o comprimento do vetor de resíduos deve ser, no mínimo, a solução encontrada usando apenas $\mathcal{M}(\mathbf{X})$. Como o R^2 é obtido como 1 menos a razão entre o comprimento ao quadrado do vetor de resíduos sobre $\sum_i (Y_i - \bar{Y})^2$, e este denominador não muda nos dois casos, devemos ter R^2 aumentando a medida que colocamos mais atributos na matriz de desenho.