Lista 03 - FECD B - 2026

Renato Assunção

1. Recall that $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top \boldsymbol{Y}$. Therefore, $\hat{\boldsymbol{Y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{X}(\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top \boldsymbol{Y} = \boldsymbol{H}\boldsymbol{Y}$, where $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top$. $\boldsymbol{H}$ is called the *hat matrix* because it puts a hat on $\boldsymbol{Y}$. Show the following basic properties of the hat matrix:

   - $\boldsymbol{H}$ is an $n \times n$ matrix.
   - $\boldsymbol{H}^2 = \boldsymbol{H}$ (idempotent).
   - $\boldsymbol{H}^\top = \boldsymbol{H}$ (symmetric).
   - $\boldsymbol{H}$ is the orthogonal projection matrix onto $\mathcal{C}(\boldsymbol{X})$ (Hint: show that: (a) $\boldsymbol{H}\boldsymbol{\beta}$ is always a linear combination of columns of $\boldsymbol{X}$ and therefore belongs to $\mathcal{C}(\boldsymbol{X})$; (b) show that $\boldsymbol{H}$ satisfies the defining property of the unique orthogonal projection matrix: $\boldsymbol{H}\boldsymbol{Y} \perp \boldsymbol{Y} - \boldsymbol{H}\boldsymbol{Y}$)

2. Show that $\boldsymbol{r} = (\boldsymbol{I} - \boldsymbol{H})\boldsymbol{Y} = (\boldsymbol{I} - \boldsymbol{H})\boldsymbol{\xi}$.

3. Show that $\|\boldsymbol{r}\|^2 = \boldsymbol{r}'\boldsymbol{r} = \boldsymbol{\xi}'(\boldsymbol{I} - \boldsymbol{H})\boldsymbol{\xi}$

4. One of the main advantages of the linear regression model is the possibility of gaining a mechanistic understanding of the data generating process. The fitted coefficient estimates the impact on Y of a unit change in the feature. Unfortunately, this interpretation is complicated because the fitted coefficient of a feature depends on which other features are present in the model. Therefore, the effect of a feature is not univocaly determined, but context dependent.

   In this exercise, you will use the *Concrete Compressive Strength Dataset*, found here: `https://archive.ics.uci.edu/ml/machine-learning-databases/concrete/compressive/Concrete_Data.xls`.

   The goal is to compare the fitted coefficient of "Cement"in three different models:

   - Univariate model, with a single feature: Compressive Strength Cement
   - Model with a few covariates: e.g., Compressive Strength Cement + Water + Age
   - Full model: Compressive Strength all variables

   The exercise illustrates multicollinearity and omitted variable bias. When features are themselves correlated (which usually they are), the fitted coefficient of a given feature is influenced by the presence of the other correlated features in the model. You observe that the coefficient for "Cement"changes depending on what else is in the model. This is so because it's capturing not just the effect of Cement but also the influence of the other correlated variables included.

5. OK, the coeffcients change but is the change relevant or insignificant in practice? SHow a plot of how the coefficient changes visually (bar plot or line plot) including the confidence intervals of each of them.

6. How much is the difference between the coefficient of "cement"from the full model and the model with two other features, compared to the range of the Y response? Write the code to obtain the delta between the coefficients and compare it with the range of Y and the standard deviation of Y. For this dataset, the impact will NOT be large.

7. A different example now, where we stress the data and make the difference in estimated coefficients to be large. Gere duas features bastante correlacionadas entre si. Por exemplo, gere 100 valores i.i.d. seguindo uma $U(0, 1)$ como x_1 e os correspondets valores de x_2 como sendo $x_{i2} = 3 + 2x_{i1} + \eta_i$ onde as variáveis $\eta_1, \ldots \eta_{100}$ são i.i.d. com distribuição $N(0, \sigma^2 = 0.05^2)$. A seguir gere o vetor resposta com as duas variáveis sendo importantes para determinar y. Por exemplo, $Y = 10 + 3x_1 + 3x_2 + \varepsilon$ onde $\varepsilon \sim N(0, 1)$.

   (a) Faça um matriz de scatterplots mostrando (y, x_1), (y, x_2) e (x_1, x_2).

   (b) Obtenha a correlação entre x_1 e y e entre x_2 e y. São muito correlacionadas, certo? Veja que a correlação é positiva: aumento de x_1 ou de x_2 levam a aumento de y. Espera-se que os coeficientes da regressão de y usando ambas, x_1 e x_2, sejam positivos.

   (c) Faça o ajuste de regressão com x_1 e x_2.

(d) Faça um loop para repetir a análise 300 vezes: deixe x_1 e x_2 fixos e gere vários (300) vetores $\mathbf{y}$. Para cada vetor $\mathbf{y}$, obtenha os coeficientes β_1 e β_2. Repita isto várias (300) vezes salvando os coeficientes em cada passo. No final, faça um histograma dos 300 valores de β_1 e outro histograma com os 300 valores de β_2. Faça também um scatterplot dos 300 valores (β_1, β_2) obtidos em cada ajuste de regressão. Compare os resultados com o que está escrito no item 6.

DICA: Resultado com 300 simulações do modelo de regressão com x_1 e x_2 correlacionados. Veja a grande variabilidade das estimativas de β_1 e β_2. Embora os valores verdadeiros sejam $\beta_1 = \beta_2 = 3$, vemos que os coeficientes estimados $\hat{\beta}_j$ podem ser muito diferentes, muito maiores ou até mesmo bem negativos. Além disso, as estimativas são muito negativamente correlacionadas: quando temos uma amostra em que $\hat{\beta}_1 >> 3 = \mathbb{E}(\hat{\beta}_1)$ teremos também $\hat{\beta}_2 << 3 = \mathbb{E}(\hat{\beta}_2)$.

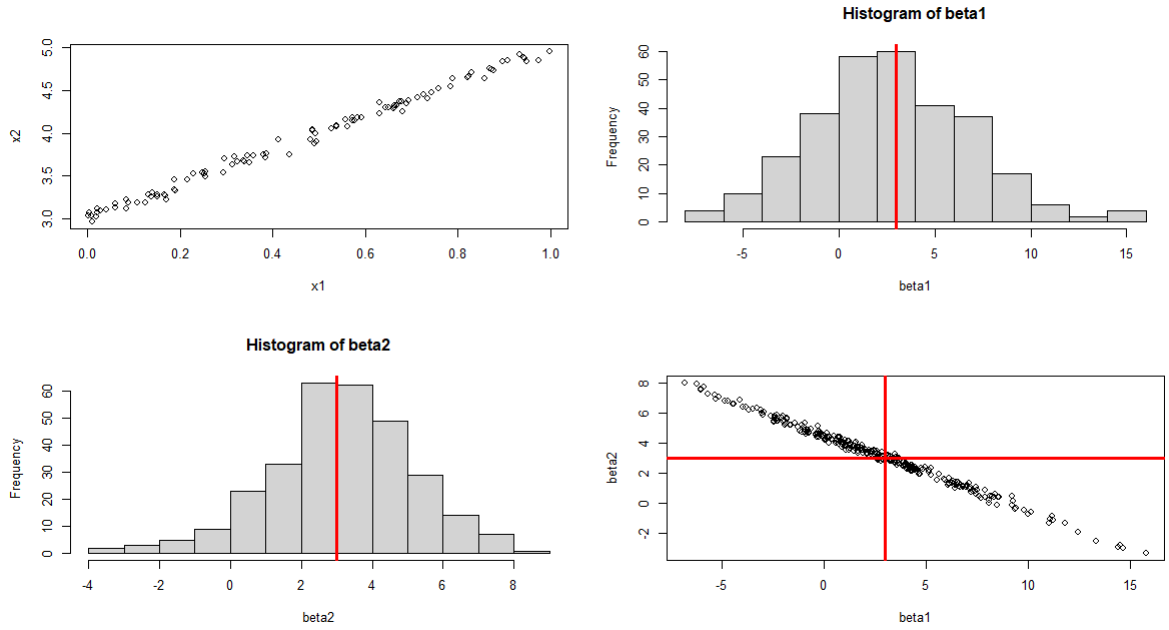

Figura 1: Exercício 4.

8. The aim of this exercise is to make sure that you understand the mechanics of ICs and hypothesis tests. Consider that you have only the first two columns of the table in Figure 2. Verify the values presented in each of the other columns for the feature "cement". That is, use the correct formulas to verify that you obtain the same values as you see in the table.

	coef	std err	t	P>|t|	[0.025	0.975]
const	-23.1638	26.588	-0.871	0.384	-75.338	29.010
Cement	0.1198	0.008	14.110	0.000	0.103	0.136
Blast Furnace Slag	0.1038	0.010	10.245	0.000	0.084	0.124
Fly Ash	0.0879	0.013	6.988	0.000	0.063	0.113
Water	-0.1503	0.040	-3.741	0.000	-0.229	-0.071
Superplasticizer	0.2907	0.093	3.110	0.002	0.107	0.474
Coarse Aggregate	0.0180	0.009	1.919	0.055	-0.000	0.036
Fine Aggregate	0.0202	0.011	1.883	0.060	-0.001	0.041
Age	0.1142	0.005	21.046	0.000	0.104	0.125

Figura 2: Cement fitted coefficients