DM
DEPARTAMENTO
DE MATEMÁTICA
TÉCNICO LISBOA

Linear Models

First test

October 17, 2025

8:00 – 8:45 (01)

Name		**No.**						

1. . . . . □

2. . . . . □

3. . . . . □

4. . . . . □

5. . . . . □

6. . . . . □

Final mark . . . . □

Formulae

$\mathbf{y} \sim N_n(\mathbf{X}\beta, \sigma^2\mathbf{I})$, with $\beta' = \left(\beta_0, \dots, \beta_{p-1}\right)$

$\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$

$\hat{\sigma}^2 = \frac{(\mathbf{y}-\hat{\mathbf{y}})'(\mathbf{y}-\hat{\mathbf{y}})}{n-p} = \frac{SSE}{n-p}$

$\hat{\mathbf{y}} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = \mathbf{H}\mathbf{y}$

$\sum\left(y_i - \bar{y}\right)^2 = \sum\left(y_i - \hat{y}_i\right)^2 + \sum\left(\hat{y}_i - \bar{y}\right)^2 \iff SST = SSE + SSR$

$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$

$R_a^2 = 1 - \frac{(n-1)SSE}{(n-p)SST}$

$$\frac{\hat{\beta}_j - \beta_j}{se\left(\hat{\beta}_j\right)} \sim t_{(n-p)} \text{ with } se\left(\hat{\beta}_j\right) = \hat{\sigma}\sqrt{(\mathbf{X}'\mathbf{X})_{jj}^{-1}}$$

H_0 : Reduced model (R) vs. H_1 : Larger model (L)

$$F = \frac{df_L}{df_R - df_L}\frac{SSE(R) - SSE(L)}{SSE(L)} \overset{H_0}{\sim} \mathcal{F}_{(df_R - df_L,\ df_L)}$$

DM
DEPARTAMENTO
DE MATEMÁTICA
TÉCNICO LISBOA

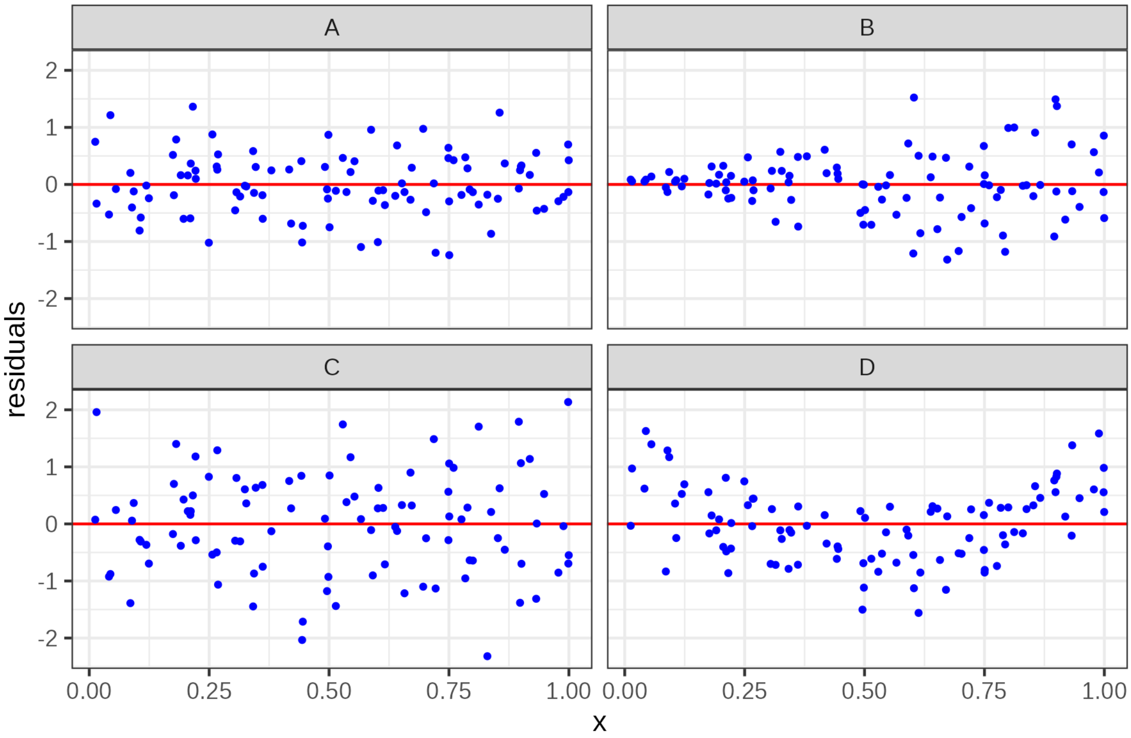
Linear Models

First test

October 17, 2025

8:00 – 8:45 (01)

1. For the same values of a covariate x, four sets of values of a response variable y were generated from different models. The following graphs represent the residuals $(y_i - \hat{y}_i)$ from the fit of simple linear regression models in each case, as a function of the values of the covariate. (4×0.5)



Choose the corresponding graph for each of the following models:

(a) $y = 5x + e,\ e \sim N(0, 0.25)$

(b) $y = 5x + e,\ e \sim N(0, x^2)$

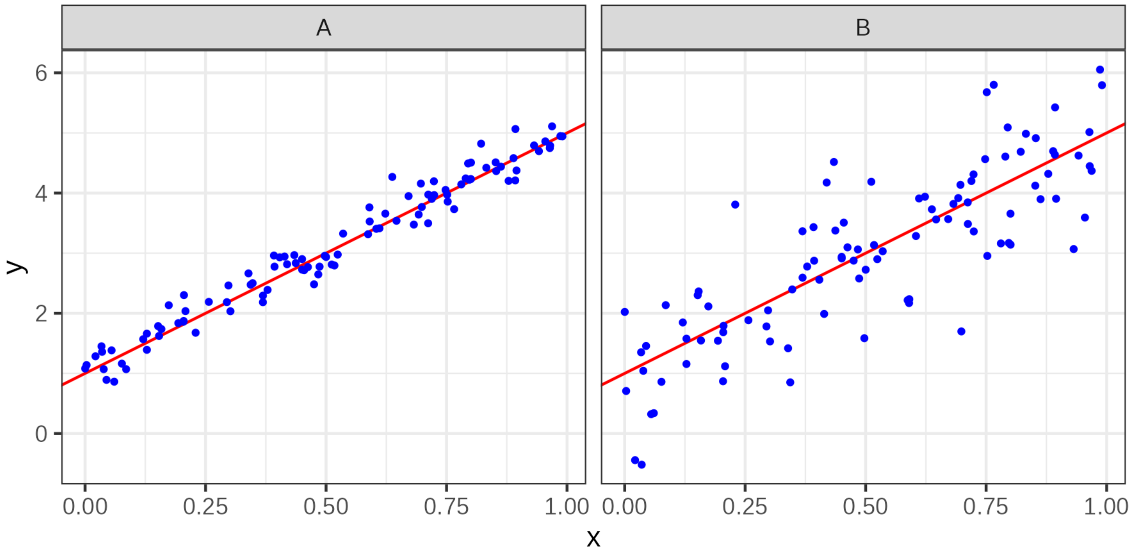
(c) $y = 5x + e,\ e \sim N(0, 1)$

(d) $y = 5x^2 + e,\ e \sim N(0, 0.25)$

Your answer should be a sequence of four letters such as *ABCD* (replace any letter with ∅ to indicate a lack of response).

ABCD

2. In the following two scatter plots, the red lines represent the same fitted regression line, $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$, for two data sets that have the same size and the same values of a covariate x. (4 × 1.0)



For each of the following statistics related to the fitted regression models, state whether its value is larger in model A (A), in model B (B), or if no comparison is possible (C).

(a) The observed value of the t-test statistic for $H_0 : \beta_0 = 0$.

(b) Standard error for $\hat{\beta}_1$.

(c) Residual sum of squares (SSE).

(d) Coefficient of determination (R^2).

Your answer should be a sequence of four letters such as $ABCA$ (replace any letter with $\emptyset$ to indicate a lack of response).

ABBC

3. Consider a multiple linear regression model that includes an intercept term, β_0, i.e., that satisfies $\mathbf{X}\mathbf{u}_1 = \mathbf{1}$ where $\mathbf{u}_1$ is a column vector with zeros in all the components except the first one that is equal to one. (2.0)

Show that the sum of the residuals, $\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}}$, is equal to zero.

The residuals can be written as $\mathbf{e} = (\mathbf{I} - \mathbf{H})\mathbf{y}$ and their sum is equal to $\mathbf{1}'(\mathbf{I} - \mathbf{H})\mathbf{y} = \mathbf{u}_1'\mathbf{X}'(\mathbf{I} - \mathbf{H})\mathbf{y} = \mathbf{u}_1'\mathbf{X}'(\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')\mathbf{y} = \mathbf{u}_1'(\mathbf{X}' - \mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')\mathbf{y} = \mathbf{u}_1'(\mathbf{X}' - \mathbf{X}')\mathbf{y} = 0.$

4. A data set about 181 UN member countries contains information on the following variables: (4.0)

`fr`: fertility rate (average number of children per woman)

`le`: life expectancy at birth (years)

`hdi`: human development index

`pop`: logarithm of the population size (thousands)

To analyse the response variable `fr`, it was fitted a first-order multiple regression model that lead to the following results:

```

Call:
lm(formula = fr ~ ., data = un2024)

Residuals:
     Min       1Q   Median       3Q      Max 
-1.30037 -0.46179 -0.04908  0.39789  1.98225 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  8.88475    0.79600  11.162  < 2e-16 ***
pop          0.01926    0.02190   0.880 0.380293    
le          -0.04741    0.01406  -3.373 0.000914 ***
hdi         -4.43436    0.61608  -7.198 1.68e-11 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.5919 on 177 degrees of freedom
Multiple R-squared:  0.7435,	Adjusted R-squared:  0.7392 
F-statistic: 171.1 on 3 and 177 DF,  p-value: < 2.2e-16
```

```
Analysis of Variance Table

Response: fr
           Df  Sum Sq Mean Sq F value    Pr(>F)    
pop         1   4.402   4.402  12.564 0.0005031 ***
le          1 157.242 157.242 448.815 < 2.2e-16 ***
hdi         1  18.151  18.151  51.807 1.677e-11 ***
Residuals 177  62.012   0.350                      
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Discuss the results of the tests related to the covariate pop in both tables. In particular, you should present the hypotheses tested in each case and explain why there is no contradiction between the results of those tests.

Let $\mathbf{x} = (\texttt{pop}, \texttt{le}, \texttt{hdi})'$.

In the summary table, the hypotheses tested are $H_0 : E[\texttt{fr}|\,\mathbf{x}] = \beta_0 + \beta_2\texttt{le}+\beta_3\texttt{hdi}$ against $H_1 : E[\texttt{fr}|\,\mathbf{x}] = \beta_0 + \beta_1\texttt{pop}+\beta_2\texttt{le}+\beta_3\texttt{hdi}$. That means, it is being tested if the contribution of the pop covariate to the full model may be considered to be non significant. The p-value of the t-test does not lead to the rejection of H_0 at the usual significance levels and, so, it does not seem to be enough evidence to support the maintenance of that covariate in the full model.

In the ANOVA table we have results from sequential F-tests. The hypotheses tested are $H_0 : E[\texttt{fr}|\,\mathbf{x}] = \beta_0$ against $H_1 : E[\texttt{fr}|\,\mathbf{x}] = \beta_0 + \beta_1\texttt{pop}$. This time we are testing if, coming from the null model, it is not worthwhile to include the covariate pop. That null hypothesis is clearly rejected which, at first sight, could seem to contradict the previous result from the t-test. That is not the case since the F-test is applied before the remaining covariates are added to the model and so, if the explanatory effect of those other covariates is not fully accounted, the effect of the first one to be included in the model is tipically found to be relevant. This should not be taken as an assessment of the covariate's contribution to the full model.

5. (. . .) Guided by the previous results, the covariate pop was removed and the fitting of the corresponding reduced model produced the following results: (4.0)

```
Call:
lm(formula = fr ~ le + hdi, data = un2024)

Residuals:
     Min       1Q   Median       3Q      Max 
-1.34111 -0.46133 -0.04221  0.39323  1.99814 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  9.26496    0.66796  13.870  < 2e-16 ***
le          -0.04868    0.01397  -3.484 0.000622 ***
hdi         -4.41066    0.61510  -7.171 1.92e-11 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.5915 on 178 degrees of freedom
Multiple R-squared:  0.7424,	Adjusted R-squared:  0.7395 
F-statistic: 256.5 on 2 and 178 DF,  p-value: < 2.2e-16
```

Explain the observed variations of both versions of R^2 between the first and the second model. Use some numerical results in the previous ANOVA table to support your explanation.

The R^2 decreases from 0.7435 to 0.7424, by the removal of the covariate pop from the full model. This was expected since a reduced model can not explain a larger proportion of variation in the response variable than a larger one. The coefficient of determination only accounts for the SSR (or the SSE) which is always smaller in reduced models nested in some larger model.

On the contrary, the adjusted R^2 has a small increase from 0.7392 to 0.7395. This can be seen as an indication of a better fit of the reduced model since it shows that the increase in the SSE was countered beneficially by the increase of one degree of freedom in the model. This conclusions can be further supported by the small partial SSR for pop, SSR(pop)=4.402 or, better yet, by the partial coefficient of determination $R^2_{pop} = 0.018$ that shows that the covariate has a very low explanatory power.

6. (. . .) Apply a partial F-test to compare the two previous models. Conclude, taking into account that the p-value of that test is equal to 0.38. (4.0)

```
Analysis of Variance Table

Response: fr
           Df  Sum Sq Mean Sq F value    Pr(>F)    
le          1 161.532 161.532 461.647 < 2.2e-16 ***
hdi         1  17.991  17.991  51.418 1.925e-11 ***
Residuals 178  62.283   0.350                      
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We are considering again the test of $H_0 : E[\texttt{fr}|\ \mathbf{x}] = \beta_0 + \beta_2\texttt{le}+\beta_3\texttt{hdi}$ against $H_1 : E[\texttt{fr}|\ \mathbf{x}] = \beta_0 + \beta_1\texttt{pop}+\beta_2\texttt{le}+\beta_3\texttt{hdi}$ now using the test statistic

$$F = \frac{df_L}{df_R - df_L} \frac{SSE(R) - SSE(L)}{SSE(L)} \overset{H_0}{\sim} \mathcal{F}_{(df_R - df_L,\, df_L)}$$

The results are:

```
Analysis of Variance Table

Model 1: fr ~ le + hdi
Model 2: fr ~ pop + le + hdi
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1    178 62.283
2    177 62.012  1   0.27104 0.7736 0.3803
```

that, as before, seem to support the reduced model against the larger one.