DM
DEPARTAMENTO
DE MATEMÁTICA
TÉCNICO LISBOA

Linear Models

First exam

January 23, 2025
8:00 – 10.00

1. Consider the full rank linear model $\mathbf{y}_{n\times 1} = \mathbf{X}_{n\times p}\,\boldsymbol{\beta}_{p\times 1} + \mathbf{E}_{n\times 1}$ with $\mathbf{E}_{n\times 1} \sim N_n\left(\mathbf{0}, \sigma^2\mathbf{V}\right)$, where $\mathbf{V}$ is a fixed $n \times n$ positive definite matrix (note that for any positive definite matrix $\mathbf{V}$ there exists a non-singular matrix $\mathbf{B}$ such that $\mathbf{V} = \mathbf{B}\mathbf{B}'$).

(a) Apply the linear transformation $\mathbf{z} = \mathbf{B}^{-1}\mathbf{y}$ to the initial model and show that the minimum squares estimator of $\boldsymbol{\beta}$ is $\hat{\boldsymbol{\beta}} = \left(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{y}$. (2.0)

Applying the transformation, we get the linear model $\mathbf{z} = \mathbf{B}^{-1}\mathbf{X}\boldsymbol{\beta} + \mathbf{B}^{-1}\mathbf{E}$.
$E[\mathbf{z}] = \mathbf{B}^{-1}E[\mathbf{y}] = \mathbf{B}^{-1}\mathbf{X}\boldsymbol{\beta}$
$Var[\mathbf{z}] = \mathbf{B}^{-1}Var[\mathbf{y}]\left(\mathbf{B}^{-1}\right)' = \sigma^2\mathbf{B}^{-1}\mathbf{V}\left(\mathbf{B}^{-1}\right)' = \sigma^2\mathbf{B}^{-1}\mathbf{B}\mathbf{B}'\left(\mathbf{B}^{-1}\right)' = \sigma^2\mathbf{I}$
The transformed model is a linear model with the Gauss-Markov structure and design matrix $\mathbf{B}^{-1}\mathbf{X}$ and so $\hat{\boldsymbol{\beta}} = \left[(\mathbf{B}^{-1}\mathbf{X})'\mathbf{B}^{-1}\mathbf{X}\right]^{-1}(\mathbf{B}^{-1}\mathbf{X})'\mathbf{B}^{-1}\mathbf{y} = \left(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{y}$.

(b) Describe a situation in which this model with a diagonal matrix $\mathbf{V}$ might be useful. What else can be achieved if we drop the restriction to a diagonal matrix? Discuss the expected difficulties of adopting either of these choices. (2.0)

A diagonal matrix $\mathbf{V}$ can be used to model heterocedasticity. To do that we must be able to fix all the elements of the diagonal. Usually, this may be possible when a non-constant variance can be related to some covariate, such as a temporal one, for example, or to some particular ordering of the observations.
The use of a more general matrix can also help modelling correlated data. That would require to fix $\frac{n(n-1)}{2}$ additional constants (the out of diagonal correlations) which is, in general, a difficult task. Sometimes, it is possible to define a distance between observations and fix those correlations as some function of that distance. A typical case is when there is spatial information in the covariates such as geographical coordinates, for example.

2. The yield of a chemical process (y in g) is supposed to be related to one of the reagents concentration (x_1 in g/dm^3), the operating temperature (x_2 in $^\circ F$) and the presence of a certain catalyst (x_3). To analyse the relationship between those variables, it was collected data from $n = 120$ replications of the process with $1 \le x_1 \le 2$ and $150 \le x_2 \le 180$, with the following summary statistics:

```
       x1                x2           x3          y        
 Min.   :0.02463   Min.   :150.4   0:66   Min.   : 7.61  
 1st Qu.:1.21424   1st Qu.:156.5   1:54   1st Qu.:10.24  
 Median :1.44909   Median :163.8          Median :11.45  
 Mean   :1.48362   Mean   :164.7          Mean   :11.67  
 3rd Qu.:1.78813   3rd Qu.:173.4          3rd Qu.:13.05  
 Max.   :1.98890   Max.   :179.4          Max.   :16.00  
```

A researcher started its analysis by fitting a first-order regression model with the following output:

```

Call:
lm(formula = y ~ x1 + x2 + x3, data = chemproc)

Residuals:
     Min       1Q   Median       3Q      Max 
-2.36654 -0.48415 -0.04879  0.48291  2.26716 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) -2.029897   1.388455  -1.462    0.146    
x1           1.603116   0.232375   6.899 2.94e-10 ***
x2           0.061241   0.008343   7.341 3.16e-11 ***
x31          2.749099   0.151229  18.178  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.8241 on 116 degrees of freedom
Multiple R-squared:  0.7918,    Adjusted R-squared:  0.7865 
F-statistic: 147.1 on 3 and 116 DF,  p-value: < 2.2e-16
```

(a) Explain the fitted model taking into account each level of the binary covariate x_3 and suggest a possible improvement of that model. (2.0)

This model defines the response surface

$$E[y \mid \mathbf{x}] = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 =$$

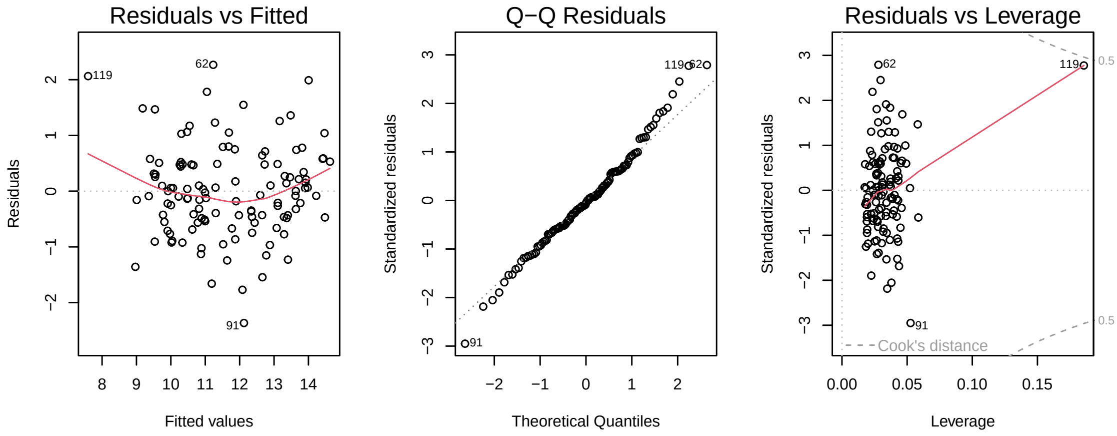
$$= \begin{cases} \beta_0 + \beta_1 x_1 + \beta_2 x_2, & x_3 = 0 \\ (\beta_0 + \beta_3) + \beta_1 x_1 + \beta_2 x_2, & x_3 = 1 \end{cases}$$

So, to model the mean yield of the chemical process, we are fitting two parallel planes as functions of both continuous covariates, one for each level of the binary covariate. This implies that

$$E[y \mid x_1, x_2, 1] - E[y \mid x_1, x_2, 0] = \beta_3$$

which means that the effect of the catalyst in the mean response is the same for any values of (x_1, x_2). This feature is completely introduced by design and can fail to capture adequately the effect of the catalyst on the chemical reaction. A way to remove that parallelism could be the inclusion of interaction terms in the model.

(b) Comment the previous results and the following diagnostic plots. Propose some remedial measure for any problem that you may find. (2.0)



In the `lm` function output we see that the regression is significant and that each covariate seems to provide valuable explanation for the observed variation of the response variable. Also, the hypothesis $\beta_0 = 0$ is not rejected at the usual significance levels as it could be expected. The effect of the covariates allows to explain almost 80% of the total variability in y.

The curved smooth line in the first plot (Residuals vs Fitted) raises some doubts on the adequacy of a linear predictor. Also, in the Q-Q plot there seems to be some mild deviance from normality in the right tail of the residuals distribution. The Residuals vs Leverage plot shows that we have an influential observation (observation #119) that is also highlighted in the two first plots and that can be causing the previously referred potential problems. So, this observation should be considered an outlier. This is probably related to the reagent concentration covariate (x_1) because in the data summary statistics we have a minimum sample value of 0.02463 when we were told that the values of x_1 should be in the interval $[1,2]$. This can be due to a recording error and then the reasonable measure would be to remove that observation and repeat the analysis.

(c) Use the following ANOVA table to compute and interpret the coefficient of partial determination between y and x_3, given that x_1 and x_2 are already included in the model. (2.0)

```
Analysis of Variance Table

Response: y
           Df  Sum Sq Mean Sq F value    Pr(>F)
x1          1  37.324  37.324  54.959 2.181e-11 ***
x2          1  37.918  37.918  55.834 1.613e-11 ***
x3          1 224.419 224.419 330.454 < 2.2e-16 ***
Residuals 116  78.778   0.679
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

$$R^2_{3|1,2} = \frac{SSR(x_3 \mid x_1, x_2)}{SSE(x_1,x_2)} = \frac{SSR(x_1,x_2,x_3) - SSR(x_1,x_2)}{SSE(x_1,x_2)} = \frac{SSE(x_1,x_2) - SSE(x_1,x_2,x_3)}{SSE(x_1,x_2)} = 1 - \frac{78.778}{78.778 + 224.419} \approx 0.74$$

The inclusion of the covariate x_3 provides explanation for 74% of the variation of the response variable that was not explained by x_1 and x_2.

The researcher also used the `predict` function in `R` to produce the following results:

x_1	x_2	x_3	$\hat{y}$	se$(\hat{y})$
1	150	0	8.7593	0.1883
1	150	1	11.5084	0.1939

(d) Use the Bonferroni method for simultaneous estimation and a 90% global confidence level to obtain joint interval estimates for the mean response of reactions with $x_1 = 1$ and $x_2 = 150$, with and without the catalyst. (2.0)

From Bonferroni's inequality, to achieve a 90% global confidence level we need to compute individual confidence intervals with a common confidence level γ such that $1 - 2(1 - \gamma) = 0.9$, that is, $\gamma = 0.95$. Those intervals are given by

$$\hat{y} \pm F^{-1}_{t_{(116)}}(0.975) \times se(\hat{y})$$

with $\hat{y} = \mathbf{x}_0'\hat{\boldsymbol{\beta}}$ and $F^{-1}_{t_{(116)}}(0.975) = 1.981$.
For $\mathbf{x}_0' = (1, 150, 0)$ we have $A_0 = [8.386, 9.132]$, for $\mathbf{x}_0' = (1, 150, 1)$ we have $A_1 = [11.124, 11.892]$ and the rectangle $A_0 \times A_1$ is the required 90% joint confidence region.

(e) Compute a 95% prediction interval for the yield of a future reaction with $\mathbf{x} = (1, 150, 1)$. Compare it with the related interval computed in (d) and, in particular, explain their differences. (2.0)

The prediction interval is given by

$$\hat{y}_0 \pm F^{-1}_{t_{(116)}}(0.975) \times se(\hat{y}_0 - y_0)$$

with $se(\hat{y}_0 - y_0) = \sqrt{MSE\left(1 + \mathbf{x}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0\right)} = \sqrt{MSE + se^2(\hat{y}_0)}$.
This leads to the interval $[9.832, 13.185]$ that is related to the interval A_1 computed in (d) for the same values of the covariates. Both intervals are centered at the estimated mean response value, 11.5084, but the later one is wider because it takes into account not only the uncertainty in the estimation of the mean response, $E[y \mid \mathbf{x}_0]$, but also the added uncertainty of estimating the variation of the response variable distribution around that expected value.

(f) It is also admitted that the effect of the catalyst may be related to the other covariates. To analyse this conjecture, some interaction terms were added to the initial model, which led to the following results: (2.0)

```
Analysis of Variance Table

Response: y
           Df  Sum Sq Mean Sq  F value    Pr(>F)    
x1          1  37.324  37.324  58.3521 7.388e-12 ***
x2          1  37.918  37.918  59.2815 5.410e-12 ***
x3          1 224.419 224.419 350.8560 < 2.2e-16 ***
x1:x3       1   5.235   5.235   8.1851  0.005025 ** 
x2:x3       1   0.625   0.625   0.9766  0.325125    
Residuals 114  72.918   0.640                       
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Use an appropriate test of hypotheses to compare both models and draw your conclusions at a significance level $\alpha = 0.05$.

We are now considering a new model with a response surface

$$E[y \mid \mathbf{x}] = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_{13} x_1 x_3 + \beta_{23} x_2 x_3$$

that would remove the parallelism referred in (a). To compare both models we can test the linear hypotheses $H_0 : \beta_{13} = \beta_{23} = 0$ (reduced model, R) vs. $H_1 : \beta_{13} \neq 0$ or $\beta_{23} \neq 0$ (full model, F). The test statistic is $F^* = \frac{\frac{SSE(R)-SSE(F)}{df_R - df_F}}{\frac{SSE(F)}{df_F}} \overset{H_0}{\sim} F_{(df_R - df_F,\, df_F)}$ with $df_F = 114$ and $df_R = 116$.

The observed value of the test statistic is $F_o^* = \frac{\frac{78.778-72.918}{2}}{\frac{72.918}{114}} \approx 4.581$ with a p-value$=1 - F_{F_{(2,\ 114)}}(4.581) = 0.0117822$. So, at a significance level of 0.05, there is evidence to reject H_0 which means that the effect of the catalyst may not be constant for all values of the other two covariates.

3. Four methods ($i = 1, \ldots, 4$) to measure the content of magnesium in a chemical compound (y in mg) are being compared. Each method was used a few times (n_i) and the following is a summary of the 18 recorded measures:

Method (i)	n_i	$\bar{y}_{i\bullet}$
1	4	78.410
2	4	80.755
3	5	76.580
4	5	84.600

(a) Describe the model you think is most appropriate for analysing this data. (2.0)

For a continuous response variable and a single categorical covariate (or factor) we can use the single-factor analysis of variance model.

Let Y_{ij} be the observable value of the response variable of the j^{th} observation for the i^{th} factor level, with $i = 1, \dots, 4$ and $j = 1, \dots, n_i$ with $n_1 = n_2 = 4$ and $n_3 = n_4 = 5$.

The model can be defined as

$$Y_{ij} = \mu + \alpha_i + E_{ij}$$

for $i = 1, \dots, 4$ and $j = 1, \dots, n_i$ with $E_{ij} \sim N(0, \sigma^2)$, uncorrelated.

This is an non-identifiable model that requires the addition of identifiability restrictions. Setting $\sum_{i=1}^{4} n_i \alpha_i = 0$ we have $\mu = \frac{\sum_{i=1}^{4} n_i \mu_i}{n}$ with $\mu_i = E[Y_{ij}]$, an overall mean response value, and $\alpha_i = \mu_i - \mu$ that represents the effect of level i of the factor.

(b) Use the following results to assess whether all methods lead to the same average magnesium measurement, and if not, compare them two by two. (2.0)

```
Analysis of Variance Table

Response: mag
          Df  Sum Sq Mean Sq F value    Pr(>F)
method     3 176.310  58.770   11.07 0.0005452 ***
Residuals 14  74.326   5.309
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = mag ~ method, data = mag_meas)

$method
      diff        lwr        upr     p adj
2-1  2.345 -2.3905797  7.0805797 0.4971465
3-1 -1.830 -6.3225654  2.6625654 0.6462325
4-1  6.190  1.6974346 10.6825654 0.0063169
3-2 -4.175 -8.6675654  0.3175654 0.0725483
4-2  3.845 -0.6475654  8.3375654 0.1056328
4-3  8.020  3.7843687 12.2556313 0.0004010
```

The hypothesis $H_0 : \alpha_1 = \dots = \alpha_4 = 0$ is rejected with a p-value of 5.5×10^{-4} showing evidence of the existence of a factor effect on the response variable, that is, the four methods don't lead to the same mean measurements.

From Tukey's pairwise comparisons we can conclude with a global 95% confidence level that Method 4 provides mean results that are significantly different from the ones obtained with Methods 1 and 3.

Formulae

1. $(\mathbf{A}')^{-1} = \left(\mathbf{A}^{-1}\right)'$

2. $(\mathbf{AB})^{-1} = \mathbf{B}^{-1}\mathbf{A}^{-1}$

3. $E[\mathbf{Ay} + \mathbf{b}] = \mathbf{A}E[\mathbf{y}] + \mathbf{b}$

4. $Var[\mathbf{Ay} + \mathbf{b}] = \mathbf{A}Var[\mathbf{y}]\mathbf{A}'$

5. $\underset{n\times 1}{\mathbf{y}} = \underset{n\times p}{\mathbf{X}} \underset{p\times 1}{\boldsymbol{\beta}} + \underset{n\times 1}{\mathbf{e}}$ with $\mathbf{e} \sim N_n\left(\mathbf{0}, \sigma^2\mathbf{I}\right)$ and $r(\mathbf{X}) = p$

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

   $$\hat{\sigma}^2 = \frac{SSE}{n-p} = \frac{\left(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}\right)'\left(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}\right)}{n-p} = MSE$$

6. $P\left(\bigcap_{i=1}^{n} A_i\right) \geq 1 - \sum_{i=1}^{n} P(\bar{A}_i)$

7. $\dfrac{\mathbf{x}_0'\hat{\boldsymbol{\beta}} - \mathbf{x}_0'\boldsymbol{\beta}}{se(\mathbf{x}_0'\hat{\boldsymbol{\beta}})} \sim t_{(n-p)}$ with $se(\mathbf{x}_0'\hat{\boldsymbol{\beta}}) = \sqrt{MSE\,\mathbf{x}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0}$

8. $\dfrac{\hat{y}_0 - y_0}{se(\hat{y}_0 - y_0)} \sim t_{(n-p)}$ with $se(\hat{y}_0 - y_0) = \sqrt{MSE\left(1 + \mathbf{x}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0\right)}$

9. $H_0 : \mathbf{C}\boldsymbol{\beta} = \mathbf{m}$ (R) versus $H_1 : \mathbf{C}\boldsymbol{\beta} \neq \mathbf{m}$ (F)

   $$F^* = \frac{\frac{SSE(R) - SSE(F)}{df_R - df_F}}{\frac{SSE(F)}{df_F}} \overset{H_0}{\sim} F_{(df_R - df_F,\, df_F)}$$

10. $R^2_{p|1,\ldots,p-1} = \dfrac{SSR(x_p \mid x_1, \ldots, x_{p-1})}{SSE(x_1, \ldots, x_{p-1})}$

11. $SSR(x_1, \ldots, x_{p-1}) = SSR(x_1, \ldots, x_{p-3}) + SSR(x_{p-2}, x_{p-1} \mid x_1, \ldots, x_{p-3})$