DM DEPARTAMENTO DE MATEMÁTICA TÉCNICO LISBOA

Linear Models

Second exam

February 6, 2025
8:00 – 10.00

1. Consider a multiple linear regression model with 4 covariates $x_1, \dots, x_4$ and a data set of size $n = 33$. It is also known that 47.59% of the the response variable variability is explained by that regression model. (3.0)

Show that the test statistic F for the hypothesis H_0 that the 4 covariates do not jointly influence the response variable can be written as

$$F = c\frac{R^2}{1-R^2}$$

where R^2 is the coefficient of determination and c is a real constant whose value you should find.

Use this result to see if there is statistical evidence against the hypothesis H_0.

Considering $H_0 : \forall i > 0 : \beta_i = 0$ versus $H_1 : \exists i > 0 : \beta_i \neq 0$, a test statistic is

$$F = \frac{n-p}{p-1}\frac{SSR}{SSE} \overset{H_0}{\sim} F_{(p-1,n-p)}$$

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

For $n = 33$ and $p = 5$ we have $c = 28/4 = 7$, $F_0 = 7\frac{0.4759}{1-0.4759} = 6.3562$ and the p-value=$P(F > 6.3562 \mid H_0) = 1 - F_{F_{(4,28)}}(6.3562) \approx 9 \times 10^{-4}$ that shows that there is clear statistical evidence to reject H_0.

2. In a study of patient satisfaction in a certain hospital, 120 patients were randomly selected to evaluate the relationship between a patient satisfaction index (y), patient's age ($x_1 \in \{22, \dots, 55\}$, in years), severity of illness ($x_2 \in [40, 65]$, an index), anxiety level ($x_3 \in [1.5, 3.0]$, an index) and sex (x_4, 0/1=male/female). Note that larger values of y, x_2 and x_3 are associated with more satisfaction, increased severity of illness and more anxiety, respectively.

To analyse the collected data, a few multiple regression models were fitted in R.

(a) Consider the following output for model $\mathcal{M}_1$ and comment on the significance of this model, the influence of each covariate on the response variable and the overall quality of fit. (2.0)

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

Residuals:
     Min       1Q   Median       3Q      Max 
-26.8916 -13.7378  -0.0774  13.0060  26.2355 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  58.3017    19.4308   3.000  0.00330 ** 
x1           -1.0685     0.1454  -7.349 3.03e-11 ***
x2            0.9058     0.3276   2.765  0.00662 ** 
x3           -7.8722     4.2835  -1.838  0.06865 .  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 14.87 on 116 degrees of freedom
Multiple R-squared:  0.3689,	Adjusted R-squared:  0.3526 
F-statistic:  22.6 on 3 and 116 DF,  p-value: 1.358e-11
```

The hypothesis $H_0 : \beta_1 = \beta_2 = \beta_3 = 0$ is clearly rejected by a p-value = 1.36×10^{-11} which means that together the covariates are able to explain some variation of the response variable.
From the t tests results we can conclude that, individually, x_1 and x_2 seem to have a significative contribution. The situation is less clear regarding x_3 (anxiety level).
However, $R^2 = 0.3689$ suggests that the inclusion of other terms or covariates in the model may lead to some substantial improvement.

(b) Use the following ANOVA table to compute a measure of the relative reduction in the variability of y provided by the inclusion of x_3 in the model, given that x_1 and x_2 are already included. Comment the obtained result. (2.5)

```
Analysis of Variance Table

Response: y
           Df  Sum Sq Mean Sq F value    Pr(>F)    
x1          1 12824.0 12824.0 58.0094 7.684e-12 ***
x2          1  1417.9  1417.9  6.4137   0.01266 *  
x3          1   746.7   746.7  3.3775   0.06865 .  
Residuals 116 25643.8   221.1                      
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

$R^2_{3|1,2} = \frac{SSR(x_3|x_1,x_2)}{SSE(x_1,x_2)} = \frac{SSR(x_1,x_2,x_3)-SSR(x_1,x_2)}{SSE(x_1,x_2)} = \frac{SSE(x_1,x_2)-SSE(x_1,x_2,x_3)}{SSE(x_1,x_2)} =$
$= 1 - \frac{25643.8}{25643.8+746.7.419} \approx 0.0283$
The inclusion of x_3 only allows to explain 2.83% of the variability of y that was left unexplained by the inclusion of x_1 and x_2. This relates to the conclusions in (a) and shows that the importance of x_3 in this regression analysis remains doubtful.

(c) A first-order regression model with all the covariates, $\mathcal{M}_2$, was fitted with the following results. Use an appropriate hypothesis test to compare the models $\mathcal{M}_1$ and $\mathcal{M}_2$ and comment on the inclusion of the covariate x_4. (2.5)

```
Call:
lm(formula = y ~ ., data = patsat)

Residuals:
     Min       1Q   Median       3Q      Max 
-10.1576  -2.9589  -0.4664   3.4654  10.8031 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)  84.84406    6.08998  13.932  < 2e-16 ***
x1           -1.07029    0.04517 -23.696  < 2e-16 ***
x2            0.58319    0.10224   5.704 9.23e-08 ***
x3           -6.00463    1.33195  -4.508 1.58e-05 ***
x41         -27.94340    0.84759 -32.968  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.619 on 115 degrees of freedom
Multiple R-squared:  0.9396,	Adjusted R-squared:  0.9375 
F-statistic: 447.3 on 4 and 115 DF,  p-value: < 2.2e-16
```

```
Analysis of Variance Table

Response: y
           Df  Sum Sq Mean Sq  F value    Pr(>F)    
x1          1 12824.0 12824.0  601.043 < 2.2e-16 ***
x2          1  1417.9  1417.9   66.453 4.941e-13 ***
x3          1   746.7   746.7   34.995 3.476e-08 ***
x4          1 23190.2 23190.2 1086.891 < 2.2e-16 ***
Residuals 115  2453.7    21.3                       
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We are considering the test of
$H_0 : E[y \mid \mathbf{x}] = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$ (model R)
against
$H_1 : E[y \mid \mathbf{x}] = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4$ (model F)
The test statistic is $F^* = \frac{df_F}{df_R - df_F} \frac{SSE(R) - SSE(F)}{SSE(F)} \overset{H_0}{\sim} F_{(df_R - df_F,\, df_F)}$ with $df_F = 115$ and $df_R = 116$.
The observed value of the test statistic is $F_o^* = 115 \frac{25643.8 - 2453.7}{2453.7} \approx 1086.87.581$ with a p-value$=1 - F_{F_{(1,\,115)}}(1086.87) \approx 0$.
The null hupothesis is clearly rejected which shows that x_4 may be highly influential on the response variable. Of course, that is also clear by the large increase in R^2 from 0.3689 to 0.9396.
Also, with the inclusion of x_4, the contribution of the anxiety level x_3 became significant.

(d) A third model, $\mathcal{M}_3$, was also fitted with the following results: (2.0)

```
Call:
lm(formula = y ~ . + x3 * x4, data = patsat)

Residuals:
     Min       1Q   Median       3Q      Max 
-11.9442  -3.4315   0.1599   2.5049   8.5186 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)  70.49242    6.62123  10.646  < 2e-16 ***
x1           -1.06262    0.04221 -25.172  < 2e-16 ***
x2            0.63184    0.09616   6.571 1.56e-09 ***
x3           -1.17639    1.68788  -0.697    0.487    
x41          -2.91373    5.96835  -0.488    0.626    
x3:x41      -10.46116    2.47245  -4.231 4.72e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.313 on 114 degrees of freedom
Multiple R-squared:  0.9478,    Adjusted R-squared:  0.9455 
F-statistic: 414.1 on 5 and 114 DF,  p-value: < 2.2e-16
```

Write down the expressions that define models $\mathcal{M}_2$ and $\mathcal{M}_3$ and explain the differences between them from a modelling perspective considering each level of the binary covariate.

Consider the results from the fitting of models $\mathcal{M}_1$, $\mathcal{M}_2$ and $\mathcal{M}_3$. What do they suggest about the possible role of covariates x_3 and x_4 in this regression analysis?

The models define the following response surfaces:

$$\mathcal{M}_2 : E[y \mid \mathbf{x}] = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 =$$

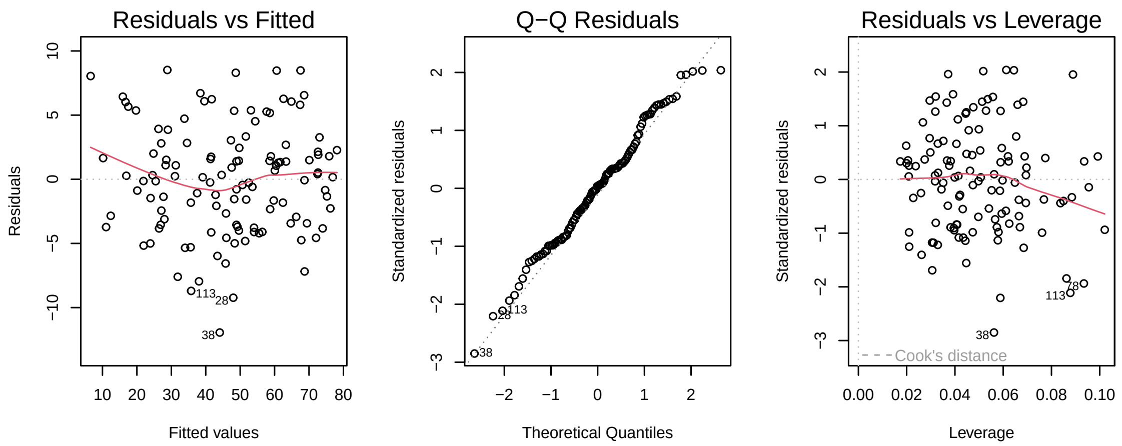
$$= \begin{cases} \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3, & x_4 = 0 (\text{men}) \\ (\beta_0 + \beta_4) + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3, & x_4 = 1 (\text{women}) \end{cases}$$

$$\mathcal{M}_3 : E[y \mid \mathbf{x}] = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_3 x_4 =$$

$$= \begin{cases} \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3, & x_4 = 0 (\text{men}) \\ (\beta_0 + \beta_4) + \beta_1 x_1 + \beta_2 x_2 + (\beta_3 + \beta_5) x_3, & x_4 = 1 (\text{women}) \end{cases}$$

With model $\mathcal{M}_2$, the difference in the mean response between men and women is constant (β_4) for the same values of the other covariates, not depending on those values. On the other hand, model $\mathcal{M}_3$ allows the effect of the anxiety level x_3 to be different between men and women.
The results indicate that x_3 (whose relevance, at first, seemed unclear) and x_4 where found significant in model $\mathcal{M}_2$. Model $\mathcal{M}_3$ shows that, in fact, their interaction may be a even more relevant term in a linear regression model.

(e) Using the previous results and the following diagnostic plots for model $\mathcal{M}_3$ can you find clear reasons to question the validity of the main assumptions or the quality of fit of that model? (2.0)



The first plot does not exhibit any clear sign of heterokedasticity in the observed data and the shape of the cloud of points, evenly spread around zero, also seems to support to use of a linear predictor.
The normality assumption is not questioned by the QQ-plot in the second figure with almost all of the points closely concentrated around the identity line.
The Residuals vs Leverage plot is also reassuring, without any highly influential observations (the Cook's distance level curves don't even show up).
The numerical results show that the regression model is significant and capable of explaining much of the observed variability in y ($R^2 \approx 0.95$).

(f) Consider model $\mathcal{M}_3$ and compute a 99% confidence interval for the difference of mean satisfaction with the hospital service between women and men with any age and severity of illness index and an anxiety index of $x_3 = 2.5$. What can you conclude from the result? (3.0)

Note: the values in the last 2 rows and last 2 columns of the matrix $(\mathbf{X}'\mathbf{X})^{-1}$ are these:

```
                x41     x3:x41
x41       1.914897 -0.7862590
x3:x41 -0.786259  0.3286171
```

We want to estimate $E[y \mid (x_1, x_2, 2.5, 1)] - E[y \mid (x_1, x_2, 2.5, 0)] = \beta_4 + 2.5\beta_5 = \mathbf{c}'\boldsymbol{\beta}$, with $\mathbf{c}' = (0,0,0,0,1,2.5)$ (from (d)). The confidence interval is given by

$$\mathbf{c}'\hat{\boldsymbol{\beta}} \pm F^{-1}_{t_{(114)}}(0.995) \times \hat{\sigma}\sqrt{\mathbf{c}'\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{c}}$$

We have $\mathbf{c}'\hat{\boldsymbol{\beta}} = -29.07$, $F^{-1}_{t_{(114)}}(0.995) = 2.6196$ and $\hat{\sigma} = 4.313$.

$\mathbf{c}'\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{c} = (1, 2.5)\begin{pmatrix} 1.9149 & -0.7863 \\ -0.7863 & 0.3286 \end{pmatrix}\begin{pmatrix} 1 \\ 2.5 \end{pmatrix} \approx 0.0375$

The requested interval is $CI_{0.99}(\beta_4 + 2.5\beta_5) = [-31.25, -26.88]$ that shows that the mean satisfaction of men and women, under the given conditions, is clearly different, with the women showing in average much less satisfaction with the hospital service than men.

3. In a study of the productivity of companies that produce electronic equipment, a measure of productivity was obtained from 27 randomly selected companies that were classified according to the level of their average expenditure for research and development in the past three years (low, moderate, high). The study results are summarized below. (3.0)

Level of expenditure	n_i	$\bar{y}_{i\bullet}$
Low	9	6.877778
Medium	12	8.133333
High	6	9.200000

A single-factor ANOVA model was fitted to the data producing the following results:

```
Call:
lm(formula = productivity ~ expenditure, data = elec)

Residuals:
     Min       1Q   Median       3Q      Max
-1.43333 -0.50556  0.02222  0.53333  1.32222

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)         9.2000     0.3266  28.167  < 2e-16 ***
expenditureLow     -2.3222     0.4217  -5.507 1.16e-05 ***
expenditureMedium  -1.0667     0.4000  -2.666   0.0135 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.8001 on 24 degrees of freedom
Multiple R-squared:  0.5671,    Adjusted R-squared:  0.531
F-statistic: 15.72 on 2 and 24 DF,  p-value: 4.331e-05
```

Describe the ANOVA model that was fitted, identifying and naming the particular encoding of the factor that was used. Explain the 3 values in the column `Estimate` of the output and comment on the influence of the expenditure level on a company's productivity.

Let yij be the observable productivity in company j with a level i of expenditure. The single-factor ANOVA model that was fitted can be expressed as $y_{ij} = \mu_i + E_{ij} = \mu + \alpha_i + E_{ij}$ with $E_{ij} \sim N(0, \sigma^2)$ uncorrelated, for $i = 1, 2, 3$ and $j = 1, \ldots, n_i$.
Since the estimate of the intercept parameter is equal to $\bar{y}_{3\bullet} = \hat{\mu}_3$, it means that this level (High) was used as a reference level and, so, a reference level encoding was applied correponding to the identifiability restriction $\alpha_3 = 0$. Aditionally, we have $\alpha_i = \mu_i - \mu$ for $i = 1, 2$.
Under this model, it is estimated a mean productivity of 9.2 units for an high level expenditure company. For a medium level company, there's an estimated decrease of 1.07 units from that value and of 2.32 units for a low level company. This results are naturally aligned with the clear rejection of the hypothesis $H_0 : \alpha_1 = \alpha_2 = 0$ that shows that the level of expenditure of a company is influential in its productivity.

Formulae

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

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

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

2. $R^2 = \dfrac{SSR}{SST}$

3. $H_0 : \forall i > 0 : \beta_i = 0$ versus $H_1 : \exists i > 0 : \beta_i \neq 0$

   $F = \dfrac{n-p}{p-1}\dfrac{SSR}{SSE} \overset{H_0}{\sim} F_{(p-1,n-p)}$

4. $\dfrac{\mathbf{c}'\hat{\boldsymbol{\beta}} - \mathbf{c}'\boldsymbol{\beta}}{\hat{\sigma}\sqrt{\mathbf{c}'\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{c}}} \sim t_{(n-p)}$ with $\mathbf{c} \in \mathbb{R}^p$

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

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

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

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

8. $y_{ij} = \mu_i + E_{ij} = \mu + \alpha_i + E_{ij}$ with $E_{ij} \sim N(0, \sigma^2)$ uncorrelated

   $\hat{\mu}_i = \bar{y}_{i\bullet}$

   $\hat{\mu} = \bar{y}_{\bullet\bullet}$