Université Côte d'Azur MSc 1 Data Science & IA
2022-2023

Statistical inference practice, part 2: mid-term exam
Duration 1h30
A4 pages (can be written on both sides)
Any kind of calculator allowed
Table of repartition function of normal distribution allowed

Exercise 1 (13 points)

Let consider the following regression model:

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \ldots + \beta_p x_i^p + \varepsilon_i, \;\; i = 1, \ldots, n$$

with $y_i \in \mathbb{R}$, $x_i \in \mathbb{R}$, $p \in \mathbb{N}^*$ (x_i^p meaning x_i at the power p) and ε_i independent and identically distributed random variables, with $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$

1. Write the model in matrix form $\mathbf{y} = X\beta + \varepsilon$
2. Give the distribution of $\mathbf{y}$
3. Is it still possible to fit a multiple regression model? (explain your answer)

Let assume that we have some real data represented in Figure 1.

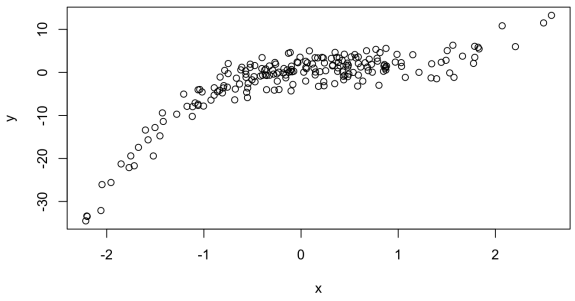

Figure 1: Scatter plot of the data

We now fit the model on $n = 200$ data with $p = 6$.
Then R give us the following summary

```
Call:
lm(formula = y ~ x + I(x^2) + I(x^3) + I(x^4) + I(x^5), data = d)
```

```
Residuals:
    Min      1Q  Median      3Q     Max
-6.1389 -1.4593  0.1208  1.4576  4.6349

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.59138    0.20743   2.851  0.00483 **
x            3.29630    0.38984   8.455 6.61e-15 ***
I(x^2)      -2.81103    0.26406 -10.645  < 2e-16 ***
I(x^3)       0.63126    0.27186   2.322  0.02127 *
I(x^4)      -0.05581    0.04416  -1.264  0.20780
I(x^5)       0.10027    0.03294   3.044  0.00266 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.05 on 194 degrees of freedom
Multiple R-squared:  0.9313,Adjusted R-squared:  0.9295
F-statistic: 525.7 on 5 and 194 DF,  p-value: < 2.2e-16
```

4. Explain how are obtained the different values related to the coefficient of variable x (for each value give a detailed explanation)

   ```
   Coefficients:
               Estimate Std. Error t value Pr(>|t|)
   x            3.29630    0.38984   8.455 6.61e-15 ***
   ```

5. Give a 95% confidence interval on β_1. (justify why it is not mandatory to use Student distribution here)

6. Based on the summary :

   (a) Give the value of s

   (b) Give the value of R^2

   (c) What test is performed in the last line and what is your conclusion?

   ```
   F-statistic: 525.7 on 5 and 194 DF,  p-value: < 2.2e-16
   ```

7. What is the advantage of using the adjusted R^2 compared with using the R^2?

The value of $(X'X)^{-1}$ is the following:

```
             (Intercept)             x        I(x^2)        I(x^3)        I(x^4)        I(x^5)
(Intercept)  1.024112e-02  1.276427e-03 -0.0088421548 -1.053081e-03  1.103871e-03  9.738012e-05
x            1.276427e-03  3.617186e-02 -0.0015977162 -2.220144e-02  9.748356e-06  2.388357e-03
I(x^2)      -8.842155e-03 -1.597716e-03  0.0165956557  1.733767e-03 -2.474542e-03 -1.303479e-04
I(x^3)      -1.053081e-03 -2.220144e-02  0.0017337667  1.759120e-02  4.018354e-06 -2.056922e-03
I(x^4)       1.103871e-03  9.748356e-06 -0.0024745419  4.018354e-06  4.641252e-04 -2.398747e-05
I(x^5)       9.738012e-05  2.388357e-03 -0.0001303479 -2.056922e-03 -2.398747e-05  2.582698e-04
```

8. Based on this matrix and (on previous information ...) give the estimated covariance matrix of the random vector composed with $\hat{\beta}_1$ and $\hat{\beta}_2$

9. Thus deduce the estimated variance of $\beta_1 - \beta_2$

10. Thus propose a way to test $H_0 : \beta_1 = \beta_2$ against $H_1 : \beta_1 \neq \beta_2$ and give your conclusion

11. If you want to test if the order of the consider polynomial is $p = 3$ versus $p = 5$ what test could you perform ? (explain just the method, don't give the result of the test.

Exercise 2 (7 points)

Let consider the following dataset where we aim at predicting the class (die or live) based on the age, the sex, and the fact that the people take steroid or not. Here is a summary of the data

```
  class          age            sex      steriod
 die : 32   Min.   : 7.00   female: 16   no :76
 live:122   1st Qu.:32.00   male  :138   yes:78
            Median :39.00
            Mean   :41.27
            3rd Qu.:50.00
            Max.   :78.00
```

We fit a logistic regression model, and we obtain the following summary

```
Call:
glm(formula = class ~ age + sex + steriod, family = "binomial",
    data = d)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.1324   0.0001   0.5150   0.7322   1.2147

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  20.40283 1543.00813   0.013  0.98945
age          -0.04692    0.01773  -2.646  0.00815 **
sexmale      -17.48818 1543.00786  -0.011  0.99096
steriodyes     0.65788    0.42650   1.543  0.12295
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 157.39  on 153  degrees of freedom
Residual deviance: 138.86  on 150  degrees of freedom
```

```
AIC: 146.86

Number of Fisher Scoring iterations: 17
```

1. Based on the fitted model give the probability of death for this person, and also give its predicted class

   ```
   age    sex steriod
    30 female      no
   ```

2. All being equal, are men more at risk that woman?
3. Give the estimated odds-ratio for the risk of death between male and women.
4. Based on the output, give the value of the likelihood
5. Would you keep this fitted model? (justify your answer)

We now fit the model only why the age variable and get the following results

```
Call:
glm(formula = class ~ age, family = "binomial", data = d)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.0192   0.4589   0.5717   0.7069   1.2410

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  3.17880    0.75622   4.204 2.63e-05 ***
age         -0.04266    0.01623  -2.629  0.00857 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 157.39  on 153  degrees of freedom
Residual deviance: 150.18  on 152  degrees of freedom
AIC: 154.18

Number of Fisher Scoring iterations: 4
```

6. Based on the available information what test could you perform to choose between the first and the second model? (give the value of the test statistic considered)
7. From statistical information criterion point of view, which model would you keep?