| Vrije Universiteit Amsterdam | Statistical Data Analysis, Exam II |
|---|---|
| Faculty of Sciences | 26 May 2021 |

**Use of a basic calculator is allowed. Programable calculators, and communication devices such as mobile phones, smart watches, etc., are not allowed. This exam consists of 4 questions on 5 pages (27 points).**

**Please write all answers in English. Grade =** $\frac{total+3}{3}$**.**

**You have 120 minutes to write the exam.**

**GOOD LUCK!**

**Question 1 [4.5 points]**
Indicate for each of the following statements whether it is correct or incorrect. ***In either case****, briefly motivate your answer.*

a. [1.5 points] The two-sample Kolmogorov-Smirnov test, which tests if two independent random samples have the same distribution, is a non-parametric, distribution-free test.

b. [1.5 points] Fisher's exact test for $2 \times 2$ contingency tables cannot be applied to test hypotheses about dependence of continuous variables.

c. [1.5 points] When you fit a multiple linear regression model, the statistics $\hat{\beta}_0, \ldots, \hat{\beta}_p$, $\hat{\sigma}^2$, and $\mathcal{R}^2$ characterise the underlying dataset in the sense that if you change the data then necessarily (at least one of) these statistics will change.

**Question 2 [5.5 points]**
Consider an i.i.d. sample $X_1, \ldots, X_{20}$ from some distribution $F$. Let $m_q$ represent the quantile of level $q$ of $F$ – i.e., for $q \in [0, 1]$, $F(m_q) = q$ – so that, for instance, $m_{0.5}$ is the median of $F$.

a. [1.5 points] To test $H_0 : m_{0.5} = 0$ we can use the sign test which is based on the test statistic $T = \sum_{i=1}^{20} 1\{X_i > 0\}$. If the median is unique and the probability that an observation $X_i$ is equal to the median is zero, then the distribution of $T$ under the null is $\text{Bin}(n, 1/2)$. If $F$ is strictly increasing and the underlying distribution continuous, do these two assumptions hold? *Justify your answer.*

b. [2 points] Suppose that you are in a situation where the median is unique, but $x_1 = x_2 = x_3 = 0$ so that there are three observations equal to the median that is prescribed under the null. Explain how you would perform the sign test in this case. *Be explicit about the test statistic and its distribution under the null.*

c. [2 points] Suppose now that you want to test $H_0 : m_{0.8} = 0$. (You can assume that $m_{0.8}$ is unique and there is zero probability that any observation is 0.) Suppose that you use the same test statistic $T$ as in the median test (i.e., $T = \sum_{i=1}^{20} 1\{X_i > 0\}$.) What is the distribution of $T$ under this new null hypothesis?

**Question 3 [8 points]**
The contingency table below (Table 1) contains data about hair- and eye colour for $N = 592$ students. (You can think of these as a random sample of students from some population.) We consider 4 categories for each of the two variables, namely: {Black, Brown, Red, Blonde} for hair colour, and {Black, Blue, Hazel, Green} for eye colour. We are interested in investigating the relation between hair- and eye colour within this population.

| Hair \ Eye | Black | Blue | Hazel | Green | Total |
|---|---|---|---|---|---|
| Black | 68 | 20 | 15 | $a$ | 108 |
| Brown | 119 | 84 | 54 | 29 | 286 |
| Red | 26 | 17 | 14 | 14 | $d$ |
| Blonde | $b$ | 94 | $c$ | 16 | 127 |
| Total | 220 | 215 | 93 | 64 | 592 |

Table 1: Distribution of hair- and eye colour across a group of students. *(Note that four entries are missing.)*

a. [1.5 points] You learned of three models for $rc$-contingency tables; which of these three models (A, B, or C) is appropriate here? State the corresponding null and alternative hypotheses for investigating the relationship between hair- and eye colour. *You may formulate your hypotheses in words.*

To perform the test from a., the following test statistic can be used:

$$X^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(N_{ij} - n\hat{p}_{ij})^2}{n\hat{p}_{ij}},$$

whose distribution under $H_0$ is, in certain cases, well approximated by a chi-square distribution. For the remainder of this question, work with the significance level $\alpha = 5\%$ for all tests that you perform.

b. [2 points] In order for the chi-square approximation just mentioned to be good, a certain rule of thumb should be satisfied. Answer the following: i) explain what this rule of thumb is (you don't need to check it); ii) what are the missing values $a$, $b$, $c$, and $d$? iii) which of the missing values $a$, $b$, $c$, and/or $d$ are actually needed to check the rule of thumb?

c. [1 point] To compute the statistic $X^2$ you would have to compute the $n\hat{p}_{ij}$. How are these computed?

d. [2 points] Test the hypotheses from part a. with the help of the chi-square test. For this, use that the value of the test statistic is $X^2 \approx 138.29$ for the present data, and select a suitable critical value from Table 2 (be explicit about which one you use.) What is the conclusion of the test?

| $k$ | $\gamma$ = 0.025 | 0.050 | 0.330 | 0.500 | 0.670 | 0.950 | 0.975 |
|---|---|---|---|---|---|---|---|
| 1 | 0.00 | 0.00 | 0.18 | 0.45 | 0.95 | 3.84 | 5.02 |
| 2 | 0.05 | 0.10 | 0.80 | 1.39 | 2.22 | 5.99 | 7.38 |
| 3 | 0.22 | 0.35 | 1.55 | 2.37 | 3.43 | 7.81 | 9.35 |
| 4 | 0.48 | 0.71 | 2.36 | 3.36 | 4.61 | 9.49 | 11.14 |
| 5 | 0.83 | 1.15 | 3.19 | 4.35 | 5.76 | 11.07 | 12.83 |
| 6 | 1.24 | 1.64 | 4.05 | 5.35 | 6.90 | 12.59 | 14.45 |
| 7 | 1.69 | 2.17 | 4.92 | 6.35 | 8.03 | 14.07 | 16.01 |
| 8 | 2.18 | 2.73 | 5.80 | 7.34 | 9.15 | 15.51 | 17.53 |
| 9 | 2.70 | 3.33 | 6.68 | 8.34 | 10.26 | 16.92 | 19.02 |

Table 2: **Table 2.** $\gamma$-quantiles of $\chi^2_k$-distribution for the listed values of $\gamma$ and $k$.

e. [1.5 points] Consider the bootstrap version of the previous test (i.e., based on the same test statistic.) Answer the following: i) Is there a downside to the bootstrap test compared to the test that you performed in d.? ii) Now suppose that you performed both tests (the one in d. and its bootstrap version) and that the conclusions of the two tests, for the same significance level, are different. In such a case, which test would you prefer and why?

**Question 4 [9 points]**
The dataset `attitude` contains information about how satisfied employees of 30 departments of a large financial organisation report to be about different aspects of their job. Namely, several employees in each department were asked if they agreed with the following statements:

- Employee **complaints** are taken into account;
- There are no employees with special **privileges**;
- There are sufficient **learning** opportunities;
- **Raises** are based on job performance;
- Feedback is not overly **critical**;
- There are opportunities to **advance** in the career;

In addition, employees are asked to report an overall **rating** for their satisfaction. Scores for all 7 measurements are on a $0\% - 100\%$ scale. The data is summarised in Figure 1 where each of the 30 observations corresponds to the average score in that department. (There were roughly 35 replies per department.)

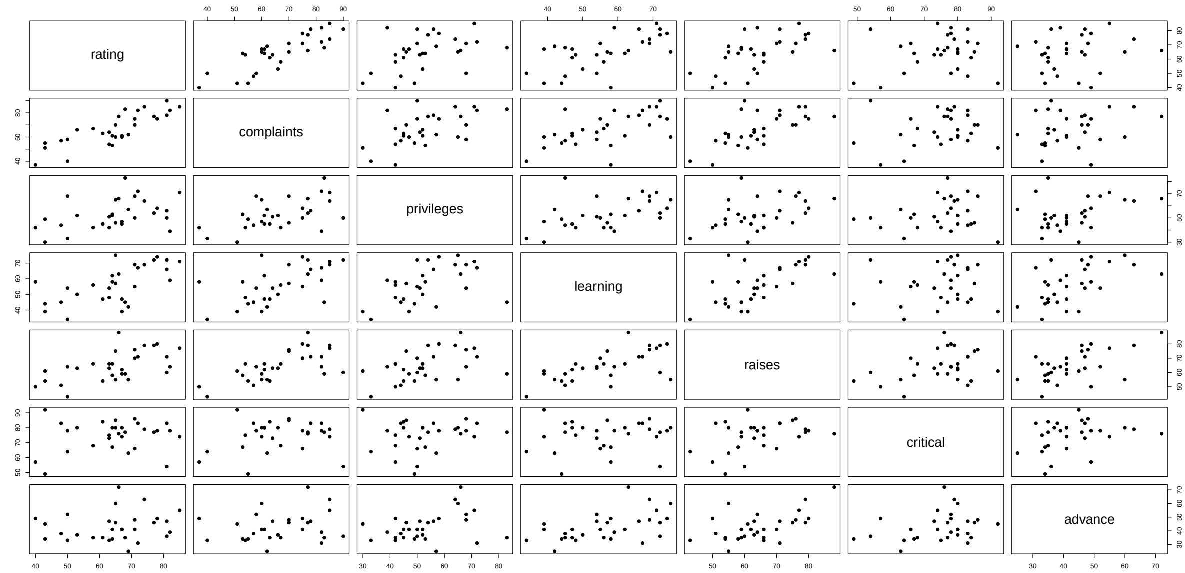


Figure 1: Matrix scatter plot of the dataset `attitude`.

Through a linear regression analysis we would like to find a model that describes the **rating**, with the help of a selection of the remaining variables.

a. [3 points] The following output was obtained in `R` when regressing **rating** linearly on the 6 variables, **complaints**, **privileges**, **leaning**, **raises**, **critical**, and **advance**:

```
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 10.78708   11.58926   0.931 0.361634
complaints   0.61319    0.16098   3.809 0.000903
privileges  -0.07305    0.13572  -0.538 0.595594
learning     0.32033    0.16852   1.901 0.069925
raises       0.08173    0.22148   0.369 0.715480
critical     0.03838    0.14700   0.261 0.796334
advance     -0.21706    0.17821  -1.218 0.235577
```

A step-down procedure to reduce the model is based on a certain test. Describe in words what two models are compared in this test under respectively the null and alternative. Briefly explain the next step of the step-down procedure ***in general***. ***In this particular case***, which (if any) variable(s) would you remove (at a level 5%) from the model in the next step of the step-down procedure?

b. [1 point] Suppose instead that you are performing a step-up procedure. You have the following fit:

```
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 14.37632    6.61999   2.172   0.0385
complaints   0.75461    0.09753   7.737 1.99e-08
Residual standard error: 6.993 on 28 degrees of freedom
Multiple R-squared:  0.6813,Adjusted R-squared:  0.6699
F-statistic: 59.86 on 1 and 28 DF,  p-value: 1.988e-08
```

You now consider including each of the remaining 5 variables one by one and get these 5 fits:

```
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 15.32762    7.16023   2.141   0.0415
complaints   0.78034    0.11939   6.536 5.22e-07
privileges  -0.05016    0.12992  -0.386   0.7025
Residual standard error: 7.102 on 27 degrees of freedom
Multiple R-squared:  0.6831,Adjusted R-squared:  0.6596
F-statistic:  29.1 on 2 and 27 DF,  p-value: 1.833e-07
---
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   9.8709     7.0612   1.398    0.174
complaints    0.6435     0.1185   5.432 9.57e-06
learning      0.2112     0.1344   1.571    0.128
Residual standard error: 6.817 on 27 degrees of freedom
Multiple R-squared:  0.708,Adjusted R-squared:  0.6864
F-statistic: 32.74 on 2 and 27 DF,  p-value: 6.058e-08
---
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 11.98732    8.42257   1.423    0.166
complaints   0.71276    0.13312   5.354 1.18e-05
raises       0.08009    0.17047   0.470    0.642
Residual standard error: 7.093 on 27 degrees of freedom
Multiple R-squared:  0.6839,Adjusted R-squared:  0.6605
F-statistic: 29.21 on 2 and 27 DF,  p-value: 1.769e-07
---
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 14.251378  11.172337   1.276    0.213
complaints   0.754344   0.101120   7.460 5.03e-08
critical     0.001908   0.136069   0.014    0.989
Residual standard error: 7.122 on 27 degrees of freedom
Multiple R-squared:  0.6813,Adjusted R-squared:  0.6577
F-statistic: 28.86 on 2 and 27 DF,  p-value: 1.974e-07
---
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 15.56025    7.89845   1.970   0.0592
complaints   0.76116    0.10177   7.479  4.8e-08
advance     -0.03773    0.13170  -0.287   0.7767
Residual standard error: 7.111 on 27 degrees of freedom
Multiple R-squared:  0.6823,Adjusted R-squared:  0.6587
F-statistic: 28.99 on 2 and 27 DF,  p-value: 1.895e-07
```

Explain how you proceed with the step-up procedure. (Use again significance level 5%.)

c. [2 point] From now on, consider the fit `rating = 14.37632 + 0.75461 * complaints + e`. What is the interpretation of the coefficient 0.75461 in the model? *Comment on both size and sign of that coefficient.*

d. [2 points] Figure 2 shows two diagnostic plots for the fit. Which assumptions of the multiple linear regression model can you check from each of them? Do you think that these assumptions are appropriate for these data? *Justify your answers.*

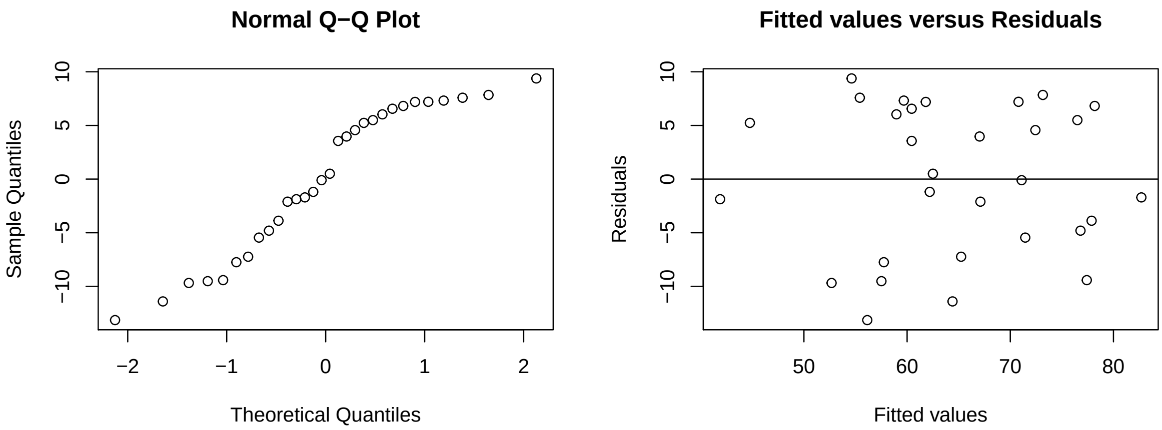


Figure 2: Diagnostic plots for the fit of the dataset `attitude`.

e. [1 point] Below are reported Cook's distance for the 30 observations:

```
            1            2            3            4            5            6
6.580419e-02 7.447877e-05 4.810908e-03 5.677733e-04 2.759077e-02 1.012345e-01
            7            8            9           10           11           12
2.140983e-02 3.734430e-06 3.236385e-02 1.221551e-02 5.819114e-02 1.616134e-02
           13           14           15           16           17           18
6.342030e-02 9.591843e-02 8.451688e-03 5.190153e-03 3.031026e-02 5.237903e-02
           19           20           21           22           23           24
1.189365e-03 4.069269e-02 4.868556e-02 3.594013e-03 4.233754e-02 1.007380e-02
           25           26           27           28           29           30
3.837243e-02 1.270913e-01 2.077871e-02 3.667913e-02 5.062969e-02 2.293365e-02
```

Based on these values, are there any influence points in the dataset? *Justify your answer.*