Vrije Universiteit Amsterdam	Statistical Data Analysis, Exam II
Faculty of Sciences	30 May 2018

Use of a basic calculator is allowed. Graphical calculators, laptops, e-readers, mobile phones, smartphones, smartwatches etc. are not allowed. This exam consists of 4 questions on 6 pages (27 points). You have 2 hours to write the exam.

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

GOOD LUCK!

Question 1 [2+2+2=6 points]
Indicate for each of the following statements whether it is correct or incorrect.
Motivate your answers shortly.

a. In the context of multiple linear regression, the definition of collinearity is that the error term in the regression equation has a variance of almost zero.

b. For particular distributions underlying the data, the two-sided signed rank test for a location parameter θ, i.e. for testing $H_0 : \theta = \theta_0$ vs. $H_a : \theta \neq \theta_0$, can be more powerful than the t test for H_0 vs. H_a. (Here, $\theta_0 \in \mathbb{R}$ is arbitrary.)

c. In a general contingency table: extreme cell entries can be identified by comparing the cells' contributions of the chi-square statistic with quantiles of the standard normal distribution.

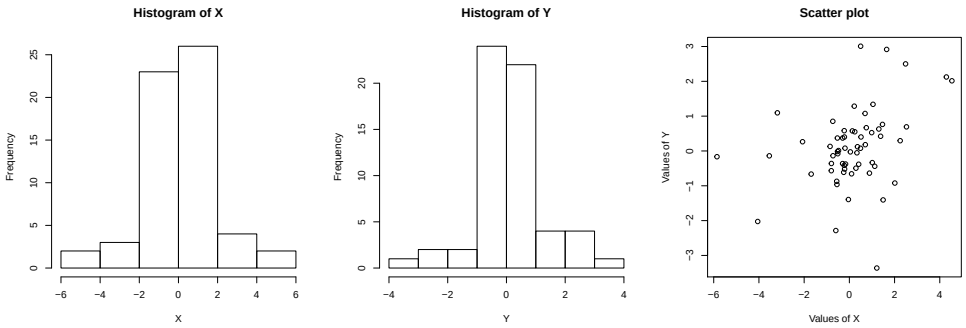

Figure 1: Histograms and a scatter plot of (X_i, Y_i) for $i = 1, \ldots, 60$.

Question 2 [3+3=6 points]
Figure 1 shows the data of the independent and identically distributed pairs $(X_1, Y_1), \ldots, (X_{60}, Y_{60})$ in histograms and a scatter plot.

a. Suppose we want to quantify the difference of the location parameters of the samples $X_1, \ldots, X_{60}$ and $Y_1, \ldots, Y_{60}$. Assume for the moment that the X- and Y-variables are independent. Indicate which of the following confidence intervals for the difference of location parameters you prefer, and explain why. Also explain for the other intervals why they are not preferred.

   i. The confidence interval $C_1 = [-0.489, 0.543]$ based on two-sided, two-sample t-tests.

   ii. The confidence interval $C_2 = [-0.274, 0.453]$ based on two-sided, two-sample Wilcoxon tests.

   iii. The confidence interval $C_3 = [-0.496, 0.372]$ based on two-sided median tests with location shift alternatives.

b. Now suppose that we doubt the independence assumption between the X- and the Y-variables which is why a statistical test shall be applied. Indicate for each of the following tests whether they are applicable for testing the null hypothesis H_0 : "X_1 and Y_1 are independent" vs. H_a : "X_1 and Y_1 are dependent", and explain why (not):

   i. Kendall's rank correlation test.

   ii. The two-sample Kolmogorov-Smirnov test.

   iii. The t-test for $\tilde{H}_0 : \beta = 0$ vs. $\tilde{H}_a : \beta \neq 0$ in the linear regression model $Y_i = \beta X_i + e_i$, where e_i are the i.i.d. error terms.

Question 3 [2+2+2+2=8 points]
In a small elementary school in the U.S. all $n = 39$ of the 4^{th} year students (boys and girls) were asked whether they are left- or right-handed. Table 1 below presents the counts of occurrences in each of the categories boy/girl and left-/right-handedness.

Handedness \ Sex	Boy	Girl	total
Left	5	3	8
Right	15	16	31
total	20	19	39

Table 1: Handedness counts among boys and girls

a. Formulate a suitable model (A, B, or C) of multinomial distribution(s) and state the corresponding null and alternative hypotheses for investigating whether there is a relationship between gender and handedness.
You may formulate your hypotheses either in words or in formulas.
Hint: Before one takes a closer look at a random school class, the numbers of boys and girls as well as the number of left- and right-handed children is unknown.

The test for hypotheses from part a. can be based on the test statistic

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

which has approximately a $\chi^2_{(r-1)(c-1)}$-distribution under H_0. In the remainder of this Question 3, we choose $\alpha = 5\%$ as significance level.

b. In order to use the above-described asymptotic $\chi^2_{(r-1)(c-1)}$-distribution as an approximation of the distribution of the test statistic under the null hypothesis, a certain rule of thumb should be satisfied. Explain what this rule of thumb is, and check whether it is satisfied for the present data.

c. Irrespectively of your conclusion in part b., test the hypothesis in part a. with the help of the chi-square test. You may use that the value of the test statistic is $X^2 \approx 0.099$ for the present data, and use a suitable critical value from Table 2. What is the test conclusion?

k	α						
	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

Table 2: α-quantiles of χ^2_k-distribution for indicated values of α and k.

d. If the rule of thumb is not satisfied, a more reliable test can be performed with the help of simulations, in particular: a certain bootstrap. Explain briefly the main idea behind this simulation approach for contingency tables and use the $B = 100$ ordered realisations of the bootstrapped test statistic below to find a reliable p-value that could be used to test the null hypothesis. Compute the bootstrap p-value.
Hint: There is no need to test the null hypothesis again!
(Remember: the value of the statistic based on the data is $X^2 \approx 0.099$.)

```
 [1] 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00
[11] 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00
[21] 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.10
[31] 0.10 0.10 0.10 0.10 0.10 0.10 0.10 0.10 0.10 0.10
[41] 0.10 0.10 0.10 0.10 0.10 0.10 0.10 0.10 0.10 0.10
[51] 0.10 0.10 0.10 0.10 0.23 0.23 0.23 0.23 0.23 0.23
[61] 0.23 0.23 0.23 0.23 0.23 0.23 0.23 0.23 0.23 0.23
[71] 0.23 0.23 0.23 0.23 0.23 0.23 0.23 0.23 1.23 1.23
[81] 1.23 1.23 1.23 1.23 1.23 1.23 1.23 1.23 1.23 1.62
[91] 1.62 1.62 1.62 1.62 1.62 3.62 3.62 3.62 3.62 4.26
```

(R output of $B = 100$ bootstrap realisations of the statistic X^2 for the present dataset.)

Question 4 [3+2+2=7 points]
We wish to analyse the dataset `fish` that contains data about $n = 35$ caught fish of the species Abramis brama. Figure 2 (see next page) shows the scatter plots of all pairs of measurements: `Weight` (the weight; in grams), `Length` (the length from the nose to the beginning of the tail; in cm), `Height` (the maximal height; in cm), `Width` (the maximal width; in cm). For these data we assume a multiple linear regression model where the response variable is `Weight`, and the other variables are the possible explanatory variables.

a. The R output **in the appendix** resulted from the `summary` commands applied to differently fitted linear models. We want to use the step up procedure to find a suitable linear model, and we choose the significance level $\alpha = 5\%$. Explain which variable(s) you would add to the empty model in the first step, and explain your choice. Write down the fitted linear model after this first step.

For arbitrary reasons, we focus on the particular linear model (ignoring the units gram and cm)

$$\texttt{Weight} = -1012.1 + 34.0 \cdot \texttt{Length} + 111.8 \cdot \texttt{Width} + \mathrm{e} \tag{1}$$

for the remaining part of this Question 4.

b. Several diagnostic plots for Model (1) are presented in Figure 3. Which assumptions of the multiple linear regression model can be checked with the first two plots? Do you think these assumptions are appropriate for these data? Motivate your answers.

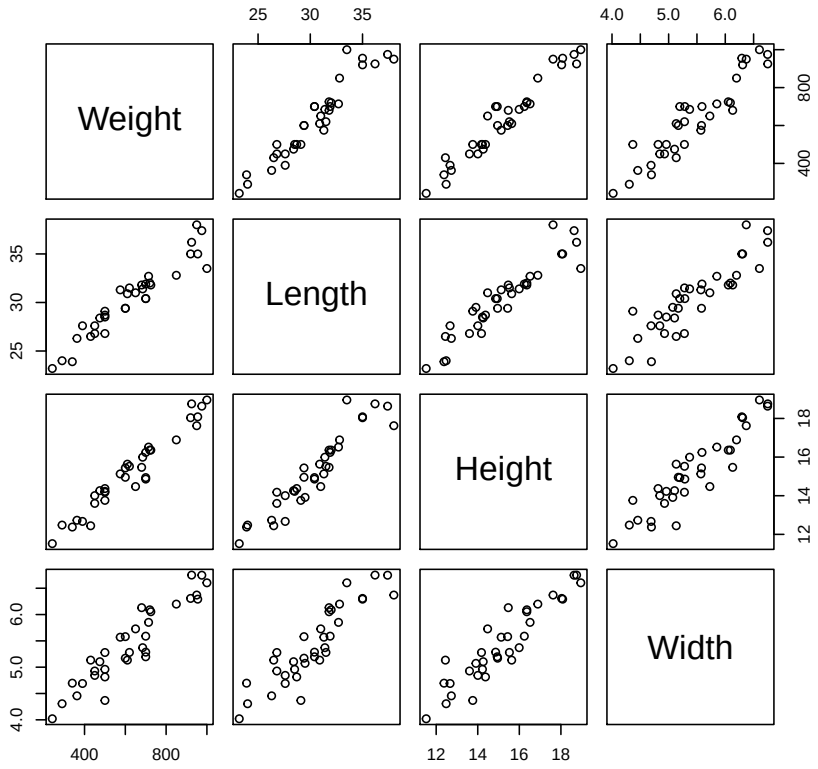

Figure 2: Scatter plots of fish data.

c. What do Figure 2 and the last plot in Figure 3 tell you about the possible dangers or advantages of including the variable `Height` into Model (1)? Motivate your answers.

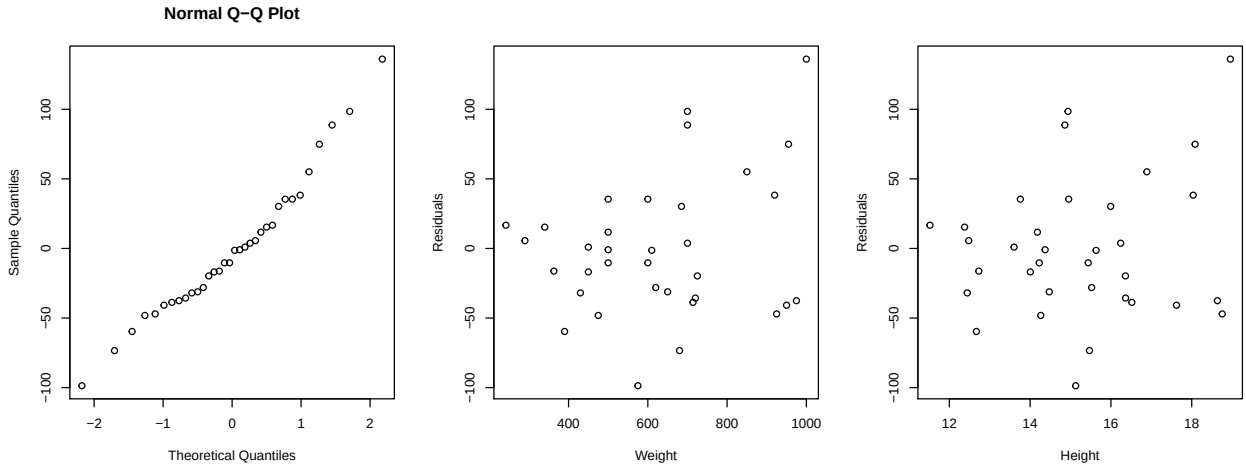

Figure 3: Plots related to the model of part b.: a normal QQ-plot of the residuals, the residuals against the response variable `Weight`, and the residuals against the variable `Height`.

Appendix

R-code for Question 4a.

```
> lm(Weight ~ Length, data = fish)
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -1015.050     91.184  -11.13 1.54e-12
Length         54.108      2.986   18.12  < 2e-16
---
Multiple R-squared:  0.9112,Adjusted R-squared:  0.9084
F-statistic: 328.4 on 1 and 32 DF,  p-value: < 2.2e-16

> lm(Weight ~ Height, data = fish)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -914.590     67.847  -13.48 9.64e-15
Height       101.218      4.421   22.89  < 2e-16
---
Multiple R-squared:  0.9425,Adjusted R-squared:  0.9407
F-statistic: 524.1 on 1 and 32 DF,  p-value: < 2.2e-16

> lm(Weight ~ Width, data = fish)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -812.37      97.99   -8.29 1.80e-09
Width         264.50      17.86   14.81 7.16e-16
---
Multiple R-squared:  0.8726,Adjusted R-squared:  0.8686
F-statistic: 219.2 on 1 and 32 DF,  p-value: 7.159e-16

> lm(Weight ~ Length + Height, data = fish)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -994.758     64.675 -15.381 4.66e-16
Length        20.218      6.284   3.218  0.00302
Height        66.197     11.558   5.727 2.67e-06
---
Multiple R-squared:  0.9569,Adjusted R-squared:  0.9541
F-statistic: 343.8 on 2 and 31 DF,  p-value: < 2.2e-16

> lm(Weight ~ Length + Width, data = fish)
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -1012.143     75.609 -13.387 2.00e-14
Length         33.961      5.678   5.981 1.29e-06
Width         111.828     28.362   3.943 0.000428
---
Multiple R-squared:  0.9409,Adjusted R-squared:  0.9371
F-statistic: 246.6 on 2 and 31 DF,  p-value: < 2.2e-16

> lm(Weight ~ Height + Width, data = fish)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -931.73      64.02 -14.554 2.11e-15
Height         77.49      11.00   7.044 6.55e-08
Width          69.57      29.87   2.329   0.0266
---
Multiple R-squared:  0.951,Adjusted R-squared:  0.9479
F-statistic:   301 on 2 and 31 DF,  p-value: < 2.2e-16

> lm(Weight ~ Length + Height + Width, data = fish)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -997.515     62.134 -16.054 2.84e-16
Length        17.658      6.184   2.855 0.007726
Height        52.761     13.165   4.008 0.000374
Width         52.391     27.592   1.899 0.067249
---
Multiple R-squared:  0.9615,Adjusted R-squared:  0.9576
F-statistic: 249.7 on 3 and 30 DF,  p-value: < 2.2e-16
```

Table of t-quantiles

α / df	0.900	0.925	0.950	0.975	0.990	α / df	0.900	0.925	0.950	0.975	0.990
1	3.078	4.165	6.314	12.706	31.821	46	1.300	1.464	1.679	2.013	2.410
2	1.886	2.282	2.920	4.303	6.965	47	1.300	1.463	1.678	2.012	2.408
3	1.638	1.924	2.353	3.182	4.541	48	1.299	1.463	1.677	2.011	2.407
4	1.533	1.778	2.132	2.776	3.747	49	1.299	1.462	1.677	2.010	2.405
5	1.476	1.699	2.015	2.571	3.365	50	1.299	1.462	1.676	2.009	2.403
6	1.440	1.650	1.943	2.447	3.143	51	1.298	1.462	1.675	2.008	2.402
7	1.415	1.617	1.895	2.365	2.998	52	1.298	1.461	1.675	2.007	2.400
8	1.397	1.592	1.860	2.306	2.896	53	1.298	1.461	1.674	2.006	2.399
9	1.383	1.574	1.833	2.262	2.821	54	1.297	1.460	1.674	2.005	2.397
10	1.372	1.559	1.812	2.228	2.764	55	1.297	1.460	1.673	2.004	2.396
11	1.363	1.548	1.796	2.201	2.718	56	1.297	1.460	1.673	2.003	2.395
12	1.356	1.538	1.782	2.179	2.681	57	1.297	1.459	1.672	2.002	2.394
13	1.350	1.530	1.771	2.160	2.650	58	1.296	1.459	1.672	2.002	2.392
14	1.345	1.523	1.761	2.145	2.624	59	1.296	1.459	1.671	2.001	2.391
15	1.341	1.517	1.753	2.131	2.602	60	1.296	1.458	1.671	2.000	2.390
16	1.337	1.512	1.746	2.120	2.583	61	1.296	1.458	1.670	2.000	2.389
17	1.333	1.508	1.740	2.110	2.567	62	1.295	1.458	1.670	1.999	2.388
18	1.330	1.504	1.734	2.101	2.552	63	1.295	1.457	1.669	1.998	2.387
19	1.328	1.500	1.729	2.093	2.539	64	1.295	1.457	1.669	1.998	2.386
20	1.325	1.497	1.725	2.086	2.528	65	1.295	1.457	1.669	1.997	2.385
21	1.323	1.494	1.721	2.080	2.518	66	1.295	1.456	1.668	1.997	2.384
22	1.321	1.492	1.717	2.074	2.508	67	1.294	1.456	1.668	1.996	2.383
23	1.319	1.489	1.714	2.069	2.500	68	1.294	1.456	1.668	1.995	2.382
24	1.318	1.487	1.711	2.064	2.492	69	1.294	1.456	1.667	1.995	2.382
25	1.316	1.485	1.708	2.060	2.485	70	1.294	1.456	1.667	1.994	2.381
26	1.315	1.483	1.706	2.056	2.479	71	1.294	1.455	1.667	1.994	2.380
27	1.314	1.482	1.703	2.052	2.473	72	1.293	1.455	1.666	1.993	2.379
28	1.313	1.480	1.701	2.048	2.467	73	1.293	1.455	1.666	1.993	2.379
29	1.311	1.479	1.699	2.045	2.462	74	1.293	1.455	1.666	1.993	2.378
30	1.310	1.477	1.697	2.042	2.457	75	1.293	1.454	1.665	1.992	2.377
31	1.309	1.476	1.696	2.040	2.453	76	1.293	1.454	1.665	1.992	2.376
32	1.309	1.475	1.694	2.037	2.449	77	1.293	1.454	1.665	1.991	2.376
33	1.308	1.474	1.692	2.035	2.445	78	1.292	1.454	1.665	1.991	2.375
34	1.307	1.473	1.691	2.032	2.441	79	1.292	1.454	1.664	1.990	2.374
35	1.306	1.472	1.690	2.030	2.438	80	1.292	1.453	1.664	1.990	2.374
36	1.306	1.471	1.688	2.028	2.434	81	1.292	1.453	1.664	1.990	2.373
37	1.305	1.470	1.687	2.026	2.431	82	1.292	1.453	1.664	1.989	2.373
38	1.304	1.469	1.686	2.024	2.429	83	1.292	1.453	1.663	1.989	2.372
39	1.304	1.468	1.685	2.023	2.426	84	1.292	1.453	1.663	1.989	2.372
40	1.303	1.468	1.684	2.021	2.423	85	1.292	1.453	1.663	1.988	2.371
41	1.303	1.467	1.683	2.020	2.421	86	1.291	1.453	1.663	1.988	2.370
42	1.302	1.466	1.682	2.018	2.418	87	1.291	1.452	1.663	1.988	2.370
43	1.302	1.466	1.681	2.017	2.416	88	1.291	1.452	1.662	1.987	2.369
44	1.301	1.465	1.680	2.015	2.414	89	1.291	1.452	1.662	1.987	2.369
45	1.301	1.465	1.679	2.014	2.412	90	1.291	1.452	1.662	1.987	2.368

Table 3: Quantiles of t-distributions with 1 to 90 degrees of freedom (df).