Dr. M. Dettling Summer 2011

Exam Applied Statistical Regression

Approved: Any written material, calculator (without communication facility).
Tables: Attached.
Note: **All tests have to be done at the 5%-level.**
If the question concerns the significance of a factor (or similar) and if nothing else is indicated, you *don't* need to give the null- and alternative hypothesis, the test statistics or the critical values.
Exercise 1 is a multiple-choice exercise. In each sub-exercise, exactly one answer is correct. A correct answer adds 1 *plus*-point and a wrong answer $\frac{1}{2}$ *minus*-point. You get a minimum of 0 points for the whole multiple-choice exercise. Tick the correct answer to the multiple choice exercises in the separately added answer sheet. Do not stay too long at a part where you experience a lot of difficulties.

Good Luck!

1. (8 points)

A multiple regression model of the following form is fitted to a data set.

$$Y_i = \beta_0 + \beta_1 \cdot x_{i,1} + \beta_2 \cdot x_{i,2} + \beta_3 \cdot x_{i,3} + \beta_4 \cdot x_{i,4} + \varepsilon_i, \quad \varepsilon_i \sim \mathcal{N}(0, \sigma^2) \text{ i.i.d.}$$

The model is fitted using the software R and the following summary output is obtained.

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)      ???     0.1960   8.438 3.57e-13
x1            5.3036     2.5316     ??? 0.038834
x2            4.0336     2.4796   1.627 0.107111
x3           -9.3153     2.4657  -3.778 0.000276
x4            0.5884     2.2852   0.257 0.797373

Residual standard error: 1.892 on 95 degrees of freedom
Multiple R-squared: 0.1948,Adjusted R-squared: ???
F-statistic: 5.745 on 4 and 95 DF,  p-value: 0.0003483
```

1) What is the value of the t-statistics of $\hat{\beta}_1$?

a) 0.099　　b) 13.43　　c) 2.095　　d) 0.015

2) How many observations are in the data set?

a) 100　　b) 99　　c) 96　　d) 95

3) Has the null hypothesis $H_0: \ \beta_3 = 0$ to be rejected on a 5% level?

a) Yes　　b) No　　c) No answer possible.

4) What is the estimate of the intercept $\hat{\beta}_0$?

a) 1.654　　b) 0.324　　c) 43.051　　d) 1.591

5) What is the estimate of $Var(\epsilon_i)$.

a) 1.892　　b) 3.579　　c) 1.375　　d) 9.46

6) Which of the following intervals is a two-sided 95% confidence interval for β_3?

a) $-9.315 \pm 1.99 \cdot 0.00028$　　b) $-9.315 \pm 1.99 \cdot \frac{2.466}{\sqrt{95}}$

c) $-9.315 \pm 1.99 \cdot \frac{0.00028}{\sqrt{95}}$　　d) $-9.315 \pm 1.99 \cdot 2.466$

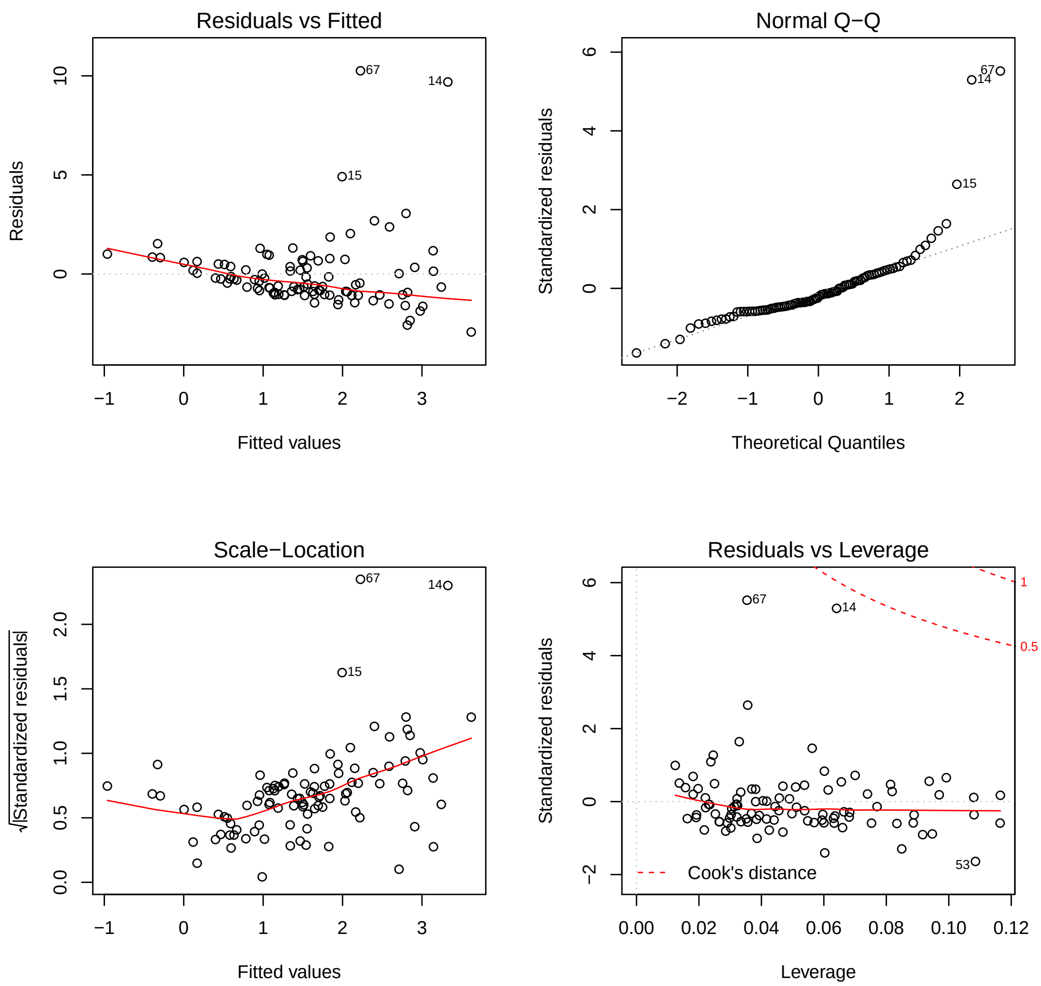
7) Have a look at the residual plots. Are the model assumptions on the ϵ_i fullfilled and if not, what is the main problem?

a) Yes.

b) No, since leverage points exist.

c) No, since the assumption of constant variance of the ε_i is violated.

d) No, since the ε_i are dependent.

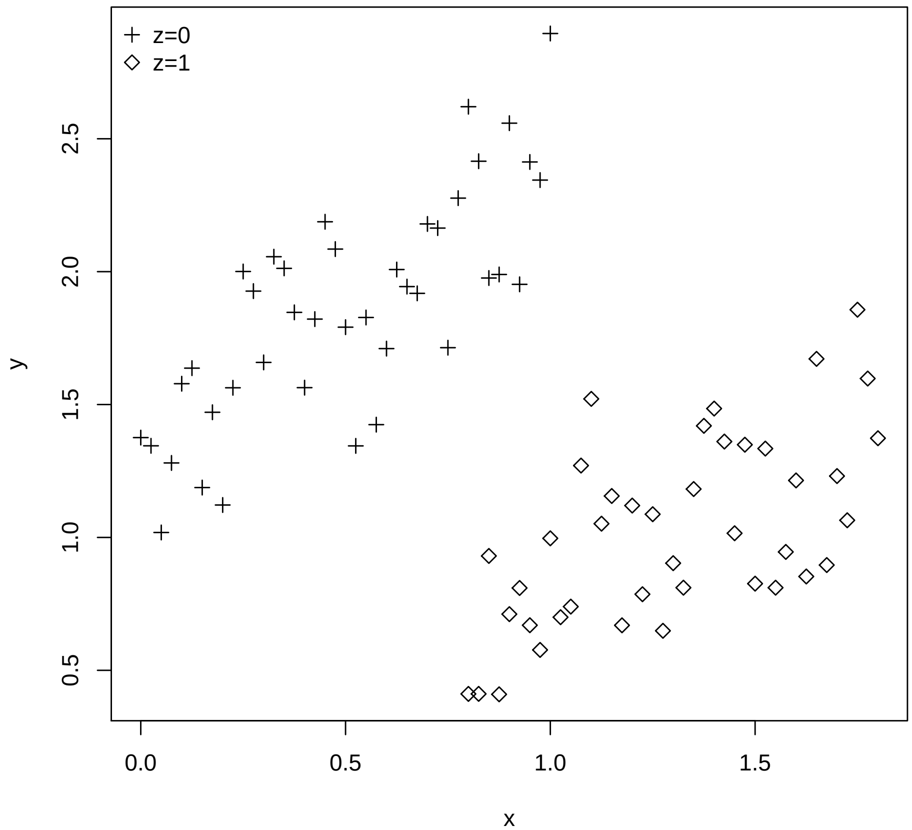


8) You want to repeat the regression, but with a better model and/or adapted data basis. What action do you take?

a) Leave out all non significant variables
b) Investigate the data without leverage points and outliers
c) Add a quadratic term
d) Apply a transformation to the response variable

2. (6 points)

Consider the following scatterplot:



The different symbols in the plot correspond to the values of two different groups. The response variable y and the covariable x are continuous, the indicator variable $z \in \{0, 1\}$ encodes the respective group membership.

a) The covariables x and z are *interacting*. Explain!

b) Are x and z correlated? Explain!

c) What model would you fit to these data? Write down a model equation?

A linear model has been fit to the above data. The R-output is given as follows:

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.31215    0.08485  15.464  < 2e-16
x            1.09296    0.14606   7.483 9.37e-11
z           -1.25344    0.21241  -5.901 8.84e-08
x:z         -0.35241    0.20656  -1.706    0.092

Residual standard error: 0.2766 on 78 degrees of freedom
Multiple R-squared: 0.7755,Adjusted R-squared: 0.7669
F-statistic: 89.82 on 3 and 78 DF,  p-value: < 2.2e-16
```

d) What are the estimated regression lines for the two groups?

e) Is it statistically nesessery to fit two regression lines with different slopes ? Motivate your answer.

We repeat the regression analysis but without interaction of x and z.

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.40026    0.06814  20.550  < 2e-16
x            0.91675    0.10452   8.771 2.72e-13
z           -1.57061    0.10400 -15.103  < 2e-16

Residual standard error: 0.28 on 79 degrees of freedom
Multiple R-squared: 0.7671,Adjusted R-squared: 0.7612
F-statistic: 130.1 on 2 and 79 DF,  p-value: < 2.2e-16
```

f) Which quantities in the R-Output can be used to compare the two models?

3. (8 points)

a) Decide (with *short* explanations) whether the following statements are true or false.

i. We consider the model $y = \beta_o + \beta_1 x + \epsilon$. Let $[-0.01, 1.5]$ be the 95%-confidence interval for β_1. In this case, a t-Test with significance level 1% rejects the null hypothesis $H_0: \ \beta_1 = 0$

ii. The coefficient of determination R^2 is not an appropriate measure to compare the goodness of fit of a single model, fitted onto different data sets.

iii. Complicated models with a lot of parameters are better for prediction then simple models with just a few parameters.

iv. The following formulas specify all the same model: `z ~ x + y + x:y`, `z ~ x*y` and `z ~ (x+y)^2`.

v. It can happen that all individual t-tests in a Regression do not reject the null hypothesis, although the global F-test is significant.

b) A fitted Poisson model has the following form: $\log\left(\widehat{\lambda}\right) = x\widehat{\beta}_1$ with $\widehat{\beta}_1 = 1.8$. How big is the variance of a new observation at $x = 3$?

c) Suppose you have a saturated model, i.e. a model containing the same number of parameters as observations. What would the estimate of σ^2 be? Give an explanation for your answer.

d) The coefficient of determination R^2 for a multiple linear regression with $n = 10$ observations and $p = 5$ predictors is equal to $R^2 = 0.86$ and the F-value is equal to $F = 20.5$ What's the value of the adjusted coefficient of determination $adjR^2$?

4. (11 points)

The swiss military carried out a study in order to analyze which soldiers are fit enough to join the special force team AAD10. In this regard, the dependent binary variable (y) reflects state of fitness of a soldier. $y = 1$ means that the soldier is fit enough for the special force team AAD10, whereas $y = 0$ indicates that the soldier is not fit enough. The following predictor variables were used for the analysis:

- $x1$: The soldiers age (in years older than 18)
- $x2$: The body mass index
- $x3$: The average amount of sport/exercise per week (in hours)

a) Write down the logistic regression model for this case.

b) Look at the following R-Output. Formally, which predictors have a significant influence on the response?

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -15.5543     7.2946  -2.132   0.0330
X1           -0.5859     0.3569     ???      ???
X2            0.5643     0.3317     ???      ???
X3            1.9639     0.8800     ???      ???

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 27.526  on ???  degrees of freedom
Residual deviance: 14.177  on 16   degrees of freedom
AIC: ???

Number of Fisher Scoring iterations: 6
```

c) How many observations were used in this logistic regression?

d) What are the odds for $y = 1$ if x_2 is increased by 1 and the other predictors remain the same?

e) Estimate the probability for $y = 1$ with $x_1 = 3$, $x_2 = 25$ and $x_3 = 2$. What would be your prediction for y in this case?

f) We have $x_1 = 5$ and $x_2 = 25$. Which value do we have to choose for x_3 in order to get a probability of 50% for $y = 1$?

g) Now we calculate the logistic regression without the predictor variable x_1.

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -7.60061    4.43762  -1.713   0.0868
X2           0.08727    0.14484   0.603      ???
X3           1.53255    0.68010   2.253      ???

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 27.526  on ???  degrees of freedom
Residual deviance: 18.158  on 17   degrees of freedom
AIC: ???
```

Which of the two models from above would you prefere concerning AIC? Motivate your answer.

5. (8 points)

We investigate the number of lung cancer deaths by age and smoking status:

	age	deaths	population	smoke
1	40-44	18	656	no
2	45-59	22	359	no
3	50-54	19	249	no
4	55-59	55	632	no
5	60-64	117	1067	no
6	65-69	170	897	no
7	70-74	179	668	no
8	75-79	120	361	no
9	80+	120	274	no
10	40-44	124	3410	yes
11	45-59	140	2239	yes
12	50-54	187	1851	yes
13	55-59	514	3270	yes
14	60-64	778	3791	yes
15	65-69	689	2421	yes
16	70-74	432	1195	yes
17	80+	63	113	yes

Notice that the information for the age group 75–79 with smoke status "yes" is missing.

a) What qualitative relation between the prevalence of skin cancers and the factors `age` and `smoke` do the numbers imply?

We want to model the expected number of cases λ by a Poisson regression model,

$$\log(\lambda) = \log(\text{population}) + \beta + \sum_{j=1}^{8} \gamma_j x_j + \delta z, \tag{1}$$

where the x_j and z are dummy variables to encode the factors `age` and `smoke`.

b) Why do we not include the interaction term `smoke:age`?

c) According to model (1), what is the qualitative effect of the population size on the expected number of lung cancer deaths?

Fitting model (1) to the data, we obtain the following R output:

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -3.70919    0.09144 -40.563  < 2e-16 ***
age45-59     0.57148    0.11496   4.971 6.66e-07 ***
age50-54     1.01766    0.10908   9.330  < 2e-16 ***
age55-59     1.42945    0.09381  15.238  < 2e-16 ***
age60-64     1.68400    0.09035  18.639  < 2e-16 ***
age65-69     2.04282    0.09065  22.534  < 2e-16 ***
age70-74     2.31253    0.09338  24.764  < 2e-16 ***
age75-79     2.59447    0.10067  25.773  < 2e-16 ***
```

```
age80+        2.82207    0.11372  24.816  < 2e-16 ***
smokeyes      0.41044    0.04096  10.021  < 2e-16 ***

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 1789.071  on 17  degrees of freedom
Residual deviance:   12.661  on  8  degrees of freedom
AIC: 153.38

Number of Fisher Scoring iterations: 4
```

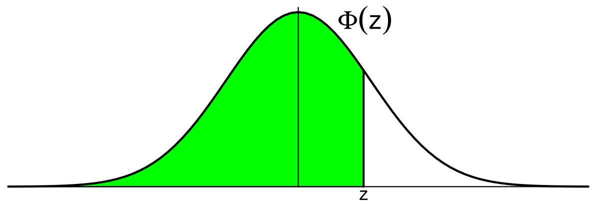
d) Compute the fitted value of the first observation.

e) The effect of `smoke` is significant. According to the fitted model, how much more likely is it that a randomly chosen smoking pearson dies from lung cancer in comparison to a randomly chosen non-smoking pearson, given that both belong to the same age group?

f) If there are 436 person in the age group 75–79 with smoke status "yes", how many of them do you expect to die from lung cancer according to the model?

g) Consider the interval $[191.01, 240.16]$. Is it plausible that this interval is a 95% prediction interval for the number in **f)**? Explain. (Hint: A Poisson distribution with parameter $\lambda > 100$ is well approximated by a normal distribution.)

Table of the cumulative Normal distribution $\Phi(z) = \mathrm{P}\left[Z \leq z\right], \; Z \sim \mathcal{N}(0,1)$



Bsp.: $\mathrm{P}\left[Z \leq 1.96\right] = 0.975$

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990
3.1	0.9990	0.9991	0.9991	0.9991	0.9992	0.9992	0.9992	0.9992	0.9993	0.9993
3.2	0.9993	0.9993	0.9994	0.9994	0.9994	0.9994	0.9994	0.9995	0.9995	0.9995
3.3	0.9995	0.9995	0.9995	0.9996	0.9996	0.9996	0.9996	0.9996	0.9996	0.9997
3.4	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9998