Dr. M. Dettling Winter 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. (7 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.

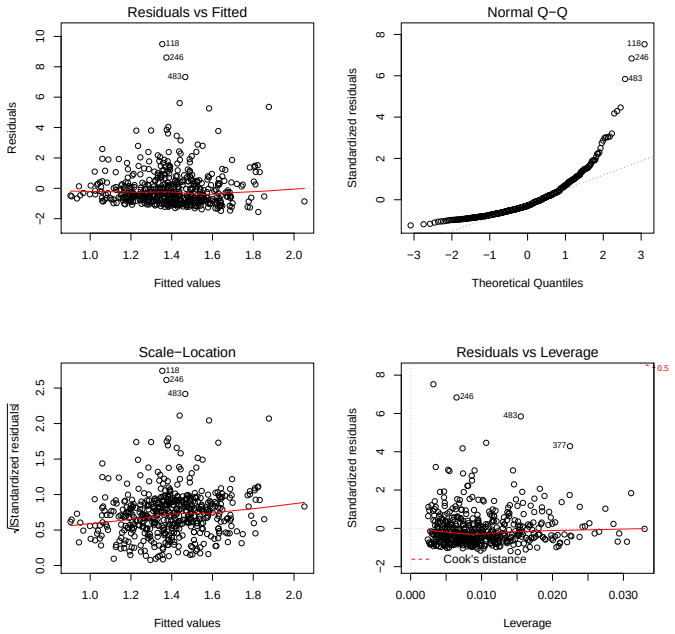
```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)      ???        ???   24.71   <2e-16 ***
x1            5.5407     3.6544     ???    0.130
x2            8.1789        ???     ???      ???
x3           -5.9592     3.5417   -1.68    0.093 .
x4           -0.9570     3.5530   -0.27    0.788
---

Residual standard error: 1.26 on 495 degrees of freedom
Multiple R-squared: 0.0207,Adjusted R-squared: 0.0128
```

1) What is the value of the t-statistics of $\hat{\beta}_1$?
   a) 0.475 b) 20.247 c) 1.516 d) 0.036
2) How many observations are in the data set?
   a) 500 b) 499 c) 496 d) 495
3) Does at least one of the explanatory variables, in the presence of the other predictors, have a significant effect on the response variable Y?
   a) Yes.
   b) No.
   c) One has to do fit four simple linear regression models to answer this question.
   d) It is not possible to make a conclusion.
4) Has the null hypothesis $H_0 : \beta_3 = 0$ to be rejected on a 10% level?
   a) Yes b) No c) No answer possible.
5) Which of the following intervals is a two-sided 95% confidence interval for β_1?
   a) $5.541 \pm 1.96 \cdot 0.13$
   b) $5.541 \pm 1.96 \cdot \frac{3.654}{\sqrt{495}}$
   c) $5.541 \pm 1.96 \cdot \frac{0.13}{\sqrt{495}}$
   d) $5.541 \pm 1.96 \cdot 3.654$

6) 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 outliers exist.
   c) No, since the residuals are not normally distributed.
   d) No, since the ε_i are dependent.

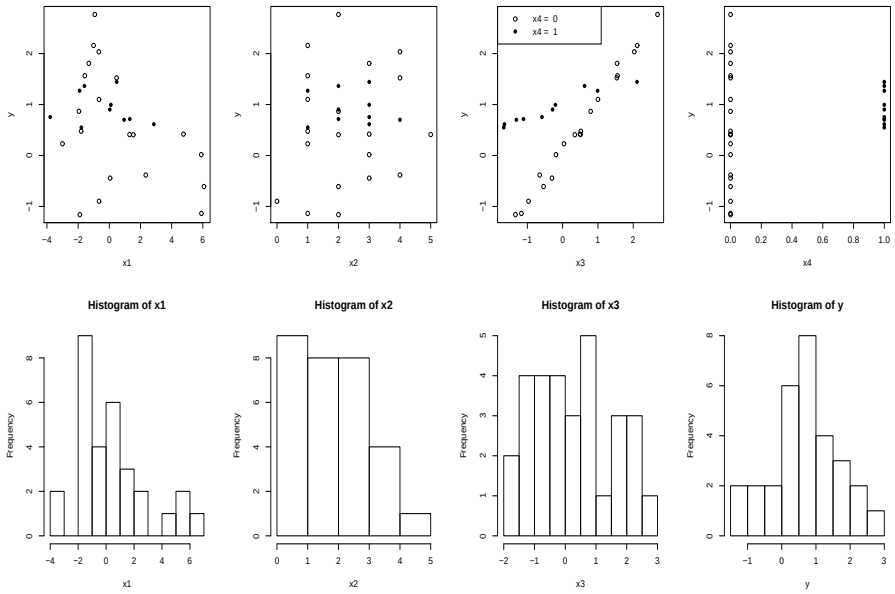


7) You would like to construct a 95% prediction interval for a new observation. Which steps would you take so that the model assumptions hold true?

   a) Investigate the data without leverage points and outliers
   b) Leave out all non significant variables
   c) Add a quadratic term
   d) Apply a transformation to the response variable
   e) If one is interested in prediction intervals rather than confidence intervalls, the model assumptions are not important

2. (9 points)

Have a look at the following plots. Here, `y, x1, x3` are continuous, `x2` is a count variable and `x4` is a factor encoding a group membership.



a) Describe the marginal correlation between `x1` and `y` and `x2` and `y`. What is special about the relation between `x3, x4` and `y`?

b) Which transformations would you propose for `y, x1, x2` and `x3` based on the information above?

A linear model was fitted to the data. Although potential transformations have been applied, the variable names are still `y, x1, x2` and `x3`. A summary of the regression is given here:

```
Call: lm(formula = y ~ x1 + x2 + x3 + x4, data = dat)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.10908    0.29137   0.374    0.711
x1          -0.08118    0.11668  -0.696    0.493
x2           0.14258    0.18882   0.755    0.457
x3           0.71445    0.07399   9.656 6.46e-10 ***
x41          0.90987    0.19307   4.713 7.85e-05 ***
---

Residual standard error: 0.4476 on 25 degrees of freedom
Multiple R-squared: 0.8149,        Adjusted R-squared: ???
F-statistic: 27.52 on 4 and 25 DF,  p-value: 7.775e-09
```

c) Only `x3` and `x4` are significant. Is it wise to exclude the intercept, `x1` and `x2` simultaneously without any further tests? Motivate your answer.

Based on the scatterplots from above a linear model that accounts for all possible 2-way interactions was fitted to the data. A summary of the regression is given here:

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.254985   0.254662   1.001 0.329279
x1          -0.195135   0.199501  -0.978 0.340307
x2          -0.074015   0.203027  -0.365 0.719468
x3           1.074260   0.077373  13.884 2.13e-11 ***
x41          0.945193   0.209688   4.508 0.000241 ***
x1:x2        0.085422   0.143174   0.597 0.557794
x1:x3       -0.081025   0.063053  -1.285 0.214231
x1:x41       0.024652   0.094595   0.261 0.797197
x2:x3        0.001419   0.050911   0.028 0.978047
x2:x41      -0.024258   0.153056  -0.158 0.875741
x3:x41      -0.716400   0.042235 -16.962 6.22e-13 ***
---

Residual standard error: 0.1177 on 19 degrees of freedom
Multiple R-squared: ??? ,       Adjusted R-squared: ???
F-statistic: 193.6 on 10 and 19 DF,  p-value: < 2.2e-16
```

d) Is the multiple R-squared of this model smaller or larger than the one from the previous model without interactions?

e) In a third step the following model was fitted to the data.

$$Y_i = \beta_0 + \beta_3 \cdot x_{i,3} + \beta_4 \cdot x_{i,4} + \beta_{10} x_{i,3} x_{i,4} + \varepsilon_i, \quad \varepsilon_i \sim \mathcal{N}(0, \sigma^2) \text{ i.i.d.}$$

The value of the F-statistic is 715.1 and the residual standard error amounts to 0.1116. This third model is nested in the one from c). Perform a partial F-test to judge if at least one of the predictors from the model in c) that are not contained in this last model have a significant influence on the response. The 95% percentile of the F-distribution with the related degrees of freedom is 2.54.

3. (8 points)

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

i. Leverage points should always be removed from the regression analysis.

ii. One squareroot-transformed the response variable and now wants to calculate confidence intervals on the original scale. He/she therefore just needs to square the confidence intervals on the transformed scale.

iii. In a regression model a factor with 4 levels served as predictor. One of these levels is not significant and should consequently be removed from the analysis.

iv. There are two predictors in a multiple linear regression which are not significant. The global F-test is highly significant. For this reason, it is better not to remove both predictors simultaneously.

b) How many parameters have to be estimated for a multinomial regression model with a response variable with 4 levels, 2 continous predictors, 1 categorial variable with 5 levels and an intercept? How many observations do you at least recommend for fitting this model?

c) Consider the model

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i1} x_{i2} + \beta_4 x_{i4} + \epsilon_i, i = 1, ..., n.$$

One wants to test the null hypothesis $\beta_2 = \beta_3 = \beta_4 = 0$ against the alternative hypothesis $\beta_2 \neq 0$ and/or $\beta_3 \neq 0$ and/or $\beta_4 \neq 0$. Which class of distributions does the corresponding test statistic have?

d) The BIC is defined as: $-2 \max(\log likelihood) + p \log n$. What is p for a model with the following formula: `y` $\sim$ `x1*x2*x3`, where `y` is continuous.

4. (9 points)

We perform a logistic regression with continuous predictors x_1, x_2 and x_3 as main effects and a binary response Y.

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? What about the relevance?

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -0.7129     1.6021  -0.445   0.6563
X1            0.0931     4.1942   0.022   ???
X2            1.8661     1.0135   1.841   ???
X3           -2.8236     1.4140  -1.997   ???

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 27.526  on ???  degrees of freedom
Residual deviance: 13.979  on 16  degrees of freedom
AIC: 21.979

Number of Fisher Scoring iterations: 5
```

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 $x_1 = 3$, $x_2 = 2$ and $x_3 = 1$. What would be your prediction for Y in this case?

f) We have $x_1 = 3$ and $x_2 = 5$. 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)  -0.6875     1.1255  -0.611   0.5413
X2            1.8803     0.7899   2.380   ???
X3           -2.8058     1.1595  -2.420   ???

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 27.526  on ???  degrees of freedom
Residual deviance: 13.979  on 17  degrees of freedom
AIC: 19.979

Number of Fisher Scoring iterations: 5
```

Use the deviance to compare the two nested models. Does the larger model yield an improvement? Motivate your answer.

5. (7 points)

In this task we look at a fictive data example. We have continuous predictors x_1, x_2 and x_3 and a count data response Y. In order to analyze the data we perform a Poisson regression.

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

b) Look at the following R-Output. Does the model fit well? (Hint: The expected value of a χ^2-distributed random variable with ν degrees of freedom is equal to ν.)

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   1.7719     0.8910   1.989   0.0467 *
X1            1.6350    15.3774   0.106   0.9153
X2            1.0897     0.1035  10.528   <2e-16 ***
X3           -4.1656    30.7287  -0.136   0.8922

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 269.730  on 19  degrees of freedom
Residual deviance:  16.068  on 16  degrees of freedom
AIC: 73.272

Number of Fisher Scoring iterations: 5
```

c) According to the fitted model from above, estimate $\mathbf{E}[Y^*]$ for a new observation with $x_1^* = 3$, $x_2^* = 3$ and $x_3^* = 1$.

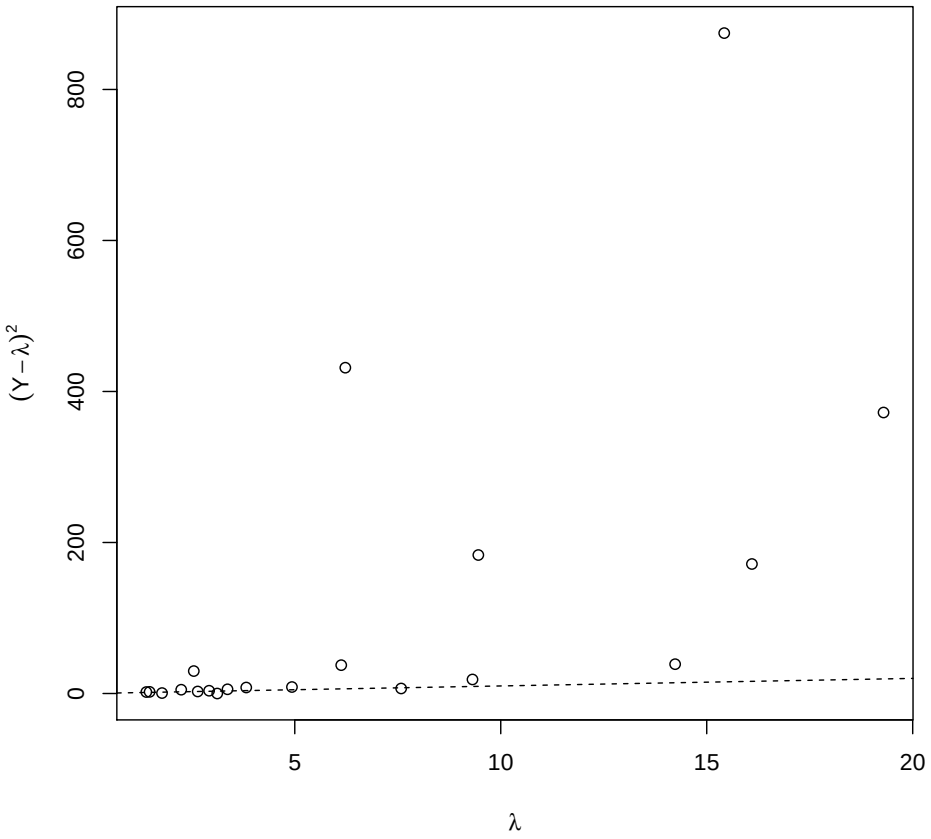
d) Now we look at the model where we drop the predictor x_2.

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)   -0.8354     0.5813  -1.437    0.151
X1            68.8018    14.2170   4.839 1.30e-06 ***
X3          -136.0553    28.5669  -4.763 1.91e-06 ***

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 269.73  on 19  degrees of freedom
Residual deviance: 190.89  on 17  degrees of freedom
AIC: 246.10

Number of Fisher Scoring iterations: 6
```

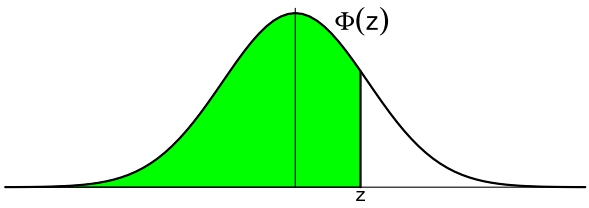


Look at the figure above. Do we have under- or overdispersion in this model?

e) The sum of squared Pearson residuals for the second model equals 211.9841. Estimate the dispersion parameter ϕ.

f) How does the dispersion parameter ϕ in general impact the inference in the case of overdispersion?

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