Multivariate Statistics: Exercise 2

Eduard Szöcs (szoecs@uni-landau.de)

January 20, 2016

1 Part I - Generalized Linear Model

The number of cones stripped by squirrels (`SqCones`) is a surrogate for squirrel abundance and we want to study the effects of tree diameter (`DBH`) on habitat selection of squirrels. For more information please refer to the original publication (Flaherty et al. 2012) or the solution to exercise 1.
In Exercise 1.2 we saw that a Gaussian distribution may not be appropriate to model the relationship between stripped cones and tree diameter - mainly because of left skewed residuals (no negative values). As we have count data a Poisson GLM might better fit to this data.

You can find the data in the *data* folder of this exercise (`RedSquirrels.txt`). It is the same data as in exercise 1, please do not modidy the raw data file!

1) Read the data into R.
2) Answer the following questions!

q1_1 Fit a Poisson GLM to `SqCones` explained by `DBH`. Use the formula interface and the `data` argument!

q1_2 Figure 1 shows the residuals vs. fitted values, is there any conspicuous pattern in this plot?

```
TRUE or FALSE
```

q1_3 What is the maximum value of cook's distance for this model? Please give a numeric answer, round to two digits, e.g.

```
25.35
```

You can use the `?round` function to round to two digits.

q1_4 Are there any issues with overdispersion with this model?

```
TRUE or FALSE
```

q1_5 Which of the following model equations is correct?

1. $SqCones = 2.331 + 1.972 * DBH$
2. $SqCones = e^{2.331+1.972*DBH}$
3. $SqCones = 2.331 + e^{1.972*DBH}$

Give your result as single digit number!

```
1 or 2 or 3
```

q1_6 Plot raw data and the model predictions into one plot!

2 Part II - Generalized Linear Model - More then one predictor

Although a Poisson might not the best choice here, we'll stick with the Poisson model and add some more predictors to the model.

q2_1 Explore the relationships between the four predictor variables (Ntrees, DBH, TreeHeight, CanopyCover) graphically.

Create a matrix of scatterplots using the `pairs()` function. The plots above the diagonal should be scatterplots with an added smoother. The diagonal should contain histogram of the distribution of the variables. The plot below the diagonal should print the correlation between variables. (Hint: All you need can be found in the help of `?pairs`).

q2_2 What is the biggest absolute (pearson) correlation between the four predictor variables? Round to two digits, e.g.

```
0.15
```

q2_3 Fit a possion model predicting SqCones with the four variables Ntrees, DBH, TreeHeight and CanopyCover (the predictors should enter in this order in the model). Use the formula interface and the `data` argument! Use the `rsq_data` data.frame.

q2_4 Figure 2 shows the fitted vs. residuals of this model.

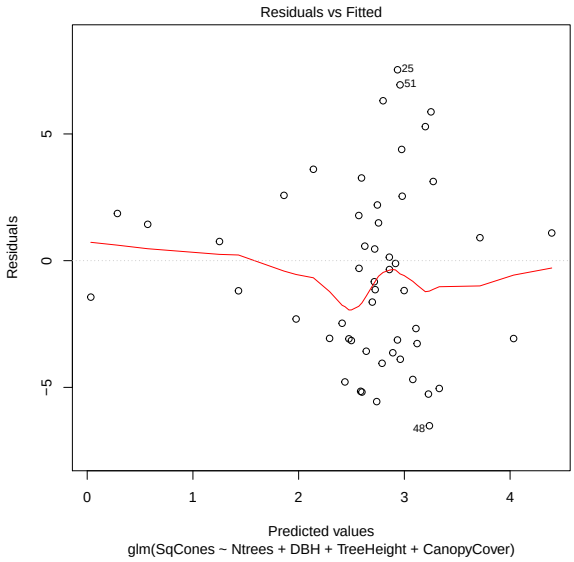

Figure 1: Residual vs. fitted values of the red squirrel poisson glm.

Is there anything conspicuous on this plot?

```
<TRUE or FALSE>
```

q2_5 Are there problems with collinearity in this model? (judge by VIFs)? (TRUE = Yes, FALSE = NO)

```
<TRUE or FALSE>
```

q2_6 Which of the following model equations is correct?

1. $SqCones = -5.501 + 0.003 * DBH + 2.485 * DBH + 0.037 * Treeheight + 0.072 * CanopyCover$
2. $SqCones = -5.501 - e^{0.003*DBH+2.485*DBH+0.037*Treeheight+0.072*CanopyCover}$
3. $SqCones = e^{-5.501+0.003*DBH+2.485*DBH+0.037*Treeheight+0.072*CanopyCover}$
4. none of the above

Give your result as single digit number!

```
1 or 2 or 3 or 4
```

q2_7 Which of the variables is most important for the red squirrels?

Regression coefficients in linear models are usually not directly comparable, because the estimates depend on the variances and these usually differ between input variables (e.g. DHB is measure in meters and ranges from 0.1 to 0.8 - CanopyCover is measured in % and ranges from 50% to 100%) [1] (you'll find the publication in the data folder - **read it!**). A simple approach is to scale input variables to zero mean and unit variance.

Scale the four predictors to zero mean and unit variance and store them in a new data.frame. The new data.frame should have the same structure as the unscaled one.

q2_8 Refit the model from q2_3 using the scaled predictors! Which variable is most important?

1. Ntrees
2. DBH
3. Treeheight
4. CanopyCover

Give your result as single digit number!

```
1 or 2 or 3 or 4
```

3 Part III - Generalized Linear Model - Dealing with overdispersion

The Poisson GLM in part I showed strong overdispersion. Perhaps you checked also the model from part II: overdispersion is also present there.
One possibility to deal with overdispersion mentioned in the lecture is using quasi-likelihood estimation. However, another possibility is to use the negative-binomial distribution. The negative binomial distribution is more flexible due to an additional dispersion parameter and can be fitted using a true likelihood (quasipoisson is based on a quasi-likelihood, therefore an AIC cannot be calculated). Negative binomial GLM fits in many cases good to overdispersed count data.

q3_1 Fit the model from q2_3 using quasipoisson. Use the formula interface and the `data` argument! Use the `rsq_data` data.frame.

q3_2 Compare the summaries of the models q2_3 and q3_1, which statement is correct?

1. Parameter estimates, AIC and Standard Errors are the same
2. Parameter estimates are the same, AIC and Standard Errors differ
3. AIC and Standard Errors are the same, Parameter estimates differ
4. Parameter estimates, AIC and Standard Errors differ

```
1 or 2 or 3 or 4
```

q3_3 Fit the model from q2_3 using a negative-binomial distribution. (`glm.nb()` from the MASS package. Note, you do not need to specify the family argument in this case!). Use the formula interface and the `data` argument! Use the `rsq_data` data.frame.

q3_4 Compare the summaries of the models q2_3 and q3_3, which statement is correct?

1. Parameter estimates, AIC and Standard Errors are the same
2. Parameter estimates are the same, AIC and Standard Errors differ
3. AIC and Standard Errors are the same, Parameter estimates differ
4. Parameter estimates, AIC and Standard Errors differ

```
1 or 2 or 3 or 4
```

q3_5 The negative binomial model from q3_3 seems to fit good to the data (have you checked the model visually?). Refit the model from q3_3 using the scaled predictors! Which variable is most important?

1. Ntrees
2. DBH
3. Treeheight
4. CanopyCover

Give your result as single digit number!

```
1 or 2 or 3 or 4
```

q3_6 Plot data and model! For easiness and representaion (2 dimensional space) use **only CanopyCover (unscaled)** as predictor! Your plot should include: raw data, model predictions, a 95% interval. A bonus would be the model equation.) This has not been covered in the lecture! Please use a internet search-machine of your choice and try to solve the task!

References

[1] Holger Schielzeth. Simple means to improve the interpretability of regression coefficients: Interpretation of regression coefficients. *Methods in Ecology and Evolution*, 1(2):103–113, February 2010.