TMA4315 Generalized Linear Models

Assignment 1:

Linear models for Gaussian data

Deadline: Monday, October 3, 2016 (week 40)

To be delivered to Jacob Skauvold either in the mailbox (Dept. of Mathematical Sciences, 7th floor, SBII) or sent electronically to `skauvold@math.ntnu.no`.

Students may work on the assignment in pairs. Each group needs to hand in a report for Parts 2–8 as well as all source code. If you want to make life easier for yourself, you can write the report using knitr (`http://yihui.name/knitr/`) to combine `R` and LaTeX.

Guidance by the teaching assistant will be given at scheduled times. For more information, please refer to the course website. If necessary, contact Jacob at `skauvold@math.ntnu.no` or Håvard at `hrue@math.ntnu.no`.

In this exercise you will create your own `R`-package to handle linear models for Gaussian data and use this package to analyse a dataset. In a later exercise you will extend this package to handle Binomial and Poisson observations. You can find more information about the package system and classes in `R` at `https://cran.r-project.org/doc/contrib/Leisch-CreatingPackages.pdf`. The text below describes how the package can be created using R-Studio since it makes it easy to rebuild and reload the package. The description is reasonably platform-independent, and should work on Windows, Mac and Linux. R-Studio is a useful integrated development environment (IDE) for `R`.

Part 1: Make basic structure for package

a) Follow these steps to get started:

0. Install R-Studio (already installed on machines in Nullrommet and Banachrommet)
1. Start R-Studio
2. Select File–New Project
3. Select create project from "New Directory"
4. Select project type "R Package"
5. Set "Package name" equal to "myglm" (without the quotation marks)
6. Select the desired directory in the field "Create project as subdirectory of"
7. Select "Create Project"

This will create the required folder structure and minimum files required for an `R`-package. The project folder `myglm` will be a subfolder of the directory chosen. The file "DESCRIPTION" contains a description of the package and you can change it with your information. The folder "man" will contain documentation for your package and can be ignored for now. The folder of interest is "R", which contains the source code for your package.

b) Each time you make a change to the source code of the package, you must build it again and load the library in your `R`-session. In R-Studio this can be done by

1. Select Build–"Build and Reload"

c) Replace the `myglm/R/myglm.R` file with the file found at
`https://www.math.ntnu.no/emner/TMA4315/2016h/Assignment1/myglm.R`.
Remember that you need to rebuild the package in R-Studio before it will be available in your `R`-session. When you write source code that uses the functions in your package, you will first need to load the package with `library(myglm)`.

d) The next time you start R-studio you may have to select File–Load Project and select the file `myglm.Rproj` in the folder `myglm` to use R-studio to build the package.

Part 2: Exploratory analysis of dataset

For this part the `myglm` package is not needed. The following data set consists of observations on five variables for 52 tenure-track professors in a small college[1]. The variables are:

- sx = sex; female (1) or male (2)
- rk = rank; assistant (1), associate (2) or full professor (3)
- yr = number of years in current rank
- yd = number of years since highest degree was earned
- sl = academic salary, in dollars.

a) Read the data into R using the following command:

```
salarydata <- read.table(
"https://www.math.ntnu.no/emner/TMA4315/2016h/Assignment1/salary.dat",
header=T)
```

b) Draw a scatter plot matrix of all the variables and comment briefly on the relationship between some of the variables.
For a given data frame, the function `pairs(mydata)` makes a scatter plot matrix.

c) We want to study how the qualitative variable "academic salary" depends on one or more explanatory variables. Which assumptions do we make about the data when doing such an analysis?

Part 3: Linear regression with `myglm` package

You can use asymptotic expressions for all the tests. This means that instead of the t-test you can use the z-test, and instead of the F-test you can use the χ^2-test. To solve this problem your package must be able to

[1] S. Weisberg, *Applied Linear Regression* New York: John Wiley & Sons, 1985.

1. calculate the estimates of the regression coefficients and the estimated covariance matrix
2. calculate the standard error and z-test for each estimated coefficient
3. calculate the fitted values and the residuals
4. calculate the residual sum of squares (RSS) and degrees of freedom
5. calculate the analysis of variance table
6. calculate the coefficient of determination (R^2) and residual standard error

a) Fit a simple linear regression model to the data with “academic salary” as the response and “number of years in current rank” as predictor. Fill-in the missing parts in the `myglm` function in your `myglm` package required to calculate the estimates of the coefficients and the estimated covariance matrix.

```
> model1 <- myglm(y~x, data=mydata)
> print(model1)
```

should give the same result as

```
> model1 <- lm(y~x, data=mydata)
> print(model1)
```

The function `myglm` returns an object of class `myglm`. So you will have to implement a `print.myglm` function in your package. See the help page `?lm` for more information about this function.

b) What are the estimates and standard errors of the constant and slope for this model? Test the significance of the slope using a z-test. What is the interpretation of the parameters?
Fill-in the missing parts in `summary.myglm` such that

```
> summary(model1)
```

gives a similar table of tests of significances as for `summary` used on an object from `lm`. Note: the result will not be the same since `lm` uses a t-test instead of a z-test.

c) Plot the observed values versus the fitted values for this model.
Implement a `plot` function for the `myglm` class that makes a scatter plot with fitted value on the x-axis and observed value on the y-axis.

d) What is the residual sum of squares (RSS) and the degrees of freedom for this model? What is RSS for the null model? Test the significance of the slope using a χ^2-test. What is the

relationship between the χ^2- and z-statistic? Find the critical values for both tests.
Fill-in the missing parts in the function `anova.myglm` so that it shows the results in a similar way as the `anova` function on an object from `lm`.

e) What is the coefficient of determination for this model and how do you interpret this value? Modify the function `summary.myglm` so that it shows this value.

f) What is the Pearson's linear correlation coefficient for this model and how do you interpret this value?

Part 4: Multiple linear regression

a) Fit a linear regression model to the data with "academic salary" as the response with "number of years in current rank" and "number of years since highest degree was earned" as predictors. You can regress a variable `y` on `x1` and `x2` from a data frame `mydata` with

```
model2 <- myglm(y~x1+x2, data=mydata)
```

Make any changes necessary to make the function `myglm` and the function `print.glm` work also for multiple linear regression.

b) What are the estimates and standard errors of the constant and slopes for this model? Test the significance of the slopes using a z-test. What is the interpretation of the parameters? Make any changes necessary to make the function `summary.myglm` work also for multiple linear regression.

c) What are the gross and net effects of "number of years in current rank" and "number of years since highest degree was earned" on "academic salary"?

d) Find the simple and partial correlations of "academic salary" with "number of years in current rank" and "number of years since highest degree was earned" and interpret the results.

Part 5: One-way analysis of variance

Use your `myglm` package to solve this problem.

a) Fit a linear regression model to the data with “academic salary” as the response and “rank” as predictor.

b) What are the estimates and standard errors of the constant and effects of factors for this model? Test the significance of the factors using a z-test. What is the interpretation of the parameters?

c) Which constraint is used on the parameters in the model above? Try setting $\mu = 0$. What happens to the coefficient of determination and χ^2-statistic, is this model at better fit?
To remove the intercept of a linear model write

```
model3 <- myglm(y~x-1, data=mydata)
```

Part 6: Two-way analysis of variance

Use your `myglm` package to do the calculations.

a) Fit a linear regression model to the data with “academic salary” as the response with “rank” and “sex” as predictors.

b) What are the estimate and standard error of the constant and effects of factors for this model? Test the significance of the factors using a z-test. What is the interpretation of the parameters?

c) Do a hierarchical anova of the model. Is the net effect of “sex” significant?
You might have to make changes to the `anova` function in your package to handle multiple predictors. You can compare with the result from fitting the same model with `lm` and using `anova` on the result. Your values will differ sligthly since you use the χ^2-test instead of the F-test.

d) Change the order of the predictors and redo the previous point. Are the results contradictory?

Part 7: Analysis of covariance

Use your `myglm` package to do the calculations.

a) Fit a linear regression model to the data with “academic salary” as the response with “rank”, “number of years in current rank” and “number of years since highest degree was earned” as predictors.

b) What are the estimate and standard error of the constant, effects of factors and slopes for this model? Test the significance of the factors using a z-test. What is the interpretation of the parameters?

c) Do a hierarchical anova of the model. Is the net effect of “number of years since highest degree was earned” significant?

d) Refit the model with only significant predictors and add an interaction term. What is the interpretation of the parameters?
To add an interaction term to a linear model write

```
model4 <- myglm(y~x1*x2, data=mydata)
```

e) Test the hypothesis of parallelism for this model using an χ^2-test.

Part 8: Regression diagnostics

a) Discuss the different diagnostic plots produced by your final model. Comment on the possible outliers. Will the exclusion of these have any effect on the final model?
For this problem you may fit the model with `lm` and use `plot` on the result. If you want an extra challenge you can implement the same plots for the function `plot.myglm` in your package.