Statistics – Seminar 3

Group 1-2

Task 1: Predicting diabetes

Use the R Data manager to download the dataset Pima.tr in the package MASS. These are data on the presence of diabetes in 532 women of Pima Indian heritage together with certain background variables as well as some laboratory tests.

A) Try to predict the presence (or absence) of diabetes using the (non-laboratory) variables age, number of pregnancies, skin fold thickness and BMI. Try also to give a presentation of your prediction that is easy to understand for someone with very limited statistics knowledge.

A-1) Choose analyze model

Because this data set is a binary outcome, logistic regression was performed.

A-2) Choose an optimal predict model

Table 1.1 shows that age term is the most significantly associated with being diabetes positive. Because the p-value, AIC and Residual deviance is the smallest among the 4 variables.

Table 1.1 *the results of logistic regression by each variable.*

	(1) Age	(2) npreg	(3) skin	(4) BMI
coefficient	0.07216	0.16879	0.04624	0.10482
probability	1.58e-06	0.000242	0.00136	0.000129
AIC	233.94	246.03	248.7	243.93
Residual deviance	229.94	242.03	244.70	239.97

Next, all variables are included in the model. The results can be seen in Figure 1.1.

```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.7411  -0.8234  -0.4918   0.9773   2.2383

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -6.444e+00  1.225e+00  -5.262 1.43e-07 ***
npreg        6.508e-02  5.726e-02   1.137  0.25569
skin         4.254e-05  1.902e-02   0.002  0.99822
bmi          1.076e-01  3.793e-02   2.836  0.00457 **
age          5.936e-02  1.858e-02   3.196  0.00140 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 256.41  on 199  degrees of freedom
Residual deviance: 214.62  on 195  degrees of freedom
AIC: 224.62

Number of Fisher Scoring iterations: 4
```

Figure 1.1 *Summary of model1*

Since just age and BMI have a significant p value, we chose both factors to find an optimal model (Figure 1.2).

```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.7935  -0.8368  -0.5033   1.0211   2.2531

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -6.49870    1.17459  -5.533 3.15e-08 ***
age          0.07104    0.01538   4.620 3.84e-06 ***
bmi          0.10519    0.02956   3.558 0.000373 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 256.41  on 199  degrees of freedom
Residual deviance: 215.93  on 197  degrees of freedom
AIC: 221.93

Number of Fisher Scoring iterations: 4
```

Figure 1.2 *Summary of model 2*

The results of model2 show that age and bmi is remarkably associated with probability of being positive diabetes. Compared to model1, the p-values of coefficients are smaller.

Moreover, the AIC value of model2 is smaller than of model1. The likelihood ratio test (Figure 1.3) also indicates that there is no significant difference between both models, so we can choose the simpler model. In conclusion, model2 is chosen as optimal model to predict diabetes

```
Likelihood ratio test

Model 1: type ~ age + npreg + skin + bmi
Model 2: type ~ age + bmi
  #Df  LogLik Df  Chisq Pr(>Chisq)
1   5 -107.31
2   3 -107.96 -2 1.3002      0.522
```

Figure 1.3 *Likelihood ratio test between model1 and model2*

A-3) Explain with predicted model

Since both p-values of coefficients are very low, it can be seen that BMI and age factors are significantly associated with the outcome with very low p-value. The coefficients of these factors are positive which means that the increase in the age and BMI will be associated with an increased probability of being diabetes-positive.

The logistic equation can be written.

$$logit(p(diabetes \mid bmi, age)) = -6.4987 + 0.1052BMI + 0.0710age$$

$$p(diabetes \mid bmi, age) = \frac{e^{-6.4987+0.1052BMI+0.0710age}}{1 + e^{-6.4987+0.1052BMI+0.0710age}}$$

Let's imagine there are two groups consisting of 3 people, in each group subjects have the same age. The age of the first group is 20, and another group is 50 and each group has three different BMI values. Predict the probability of being diabetes-positive based on the above equation.

1) age 20
   - BMI 20 : 0.04860726
   - BMI 30 : 0.12761000
   - BMI 40 : 0.29517907

2) age 50
   - BMI 20 : 0.3008975
   - BMI 30 : 0.5520281
   - BMI 40 : 0.7791577

Predicted probabilities of being diabetes-positive indicates that the higher BMI level, the higher possibility of being positive. Besides, in the age 20 group, there is no positive response of diabetes but in the age 50 group there are 2 people over 50% probability of being positive.

B) Assess how much better your predictions will be if you have access also to the laboratory data of plasma glucose concentration in an oral glucose tolerance test.

B-1) Create a model

The optimal model from task 1A is used and the variable for plasma glucose concentration (glu) is added. The summary of model3 can be seen in Figure 1.3.

```
Call:
glm(formula = type ~ glu + bmi + age, family = binomial)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.2356  -0.6974  -0.3967   0.6956   2.3878

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -9.405120   1.477177  -6.367 1.93e-10 ***
glu          0.030850   0.006448   4.784 1.72e-06 ***
bmi          0.091871   0.032242   2.849  0.00438 **
age          0.052569   0.016970   3.098  0.00195 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 256.41  on 199  degrees of freedom
Residual deviance: 188.39  on 196  degrees of freedom
AIC: 196.39

Number of Fisher Scoring iterations: 5
```

Figure 1.4 *Summary of model3*

B-2) Compare with the model not included laboratory data

The models can be compared using the AIC, the residual deviance and a likelihood ratio test (Figure 1.5)

(A): Most optimal model not included laboratory data (model2)
(B): Most optimal model included laboratory data (model3)

- AIC
  (A): 221.93
  (B): 196.39

- Residual deviance
  (A): 215.93
  (B): 188.39

```
Likelihood ratio test

Model 1: type ~ glu + bmi + age
Model 2: type ~ age + bmi
  #Df   LogLik Df  Chisq Pr(>Chisq)
1   4  -94.196
2   3 -107.963 -1 27.532  1.545e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 1.5 *Likelihood ratio test between model2 and model3*

As mentioned before, AIC value shows that model3 has more goodness of fit. In terms of R-square value, the smaller residual deviance value, the higher the R-square value, it also indicates model3 is better. Moreover, we can figure out that model3 is more optimal to predict probability of diabetes because p-value is significantly small in the likelihood ratio test.

Furthermore, the accuracy was calculated from a given dataset (Pima.tr) to analyze which model was better. First, we predicted the possibility of diabetes being positive for each model, and if the probability was over 50%, we classified it as positive or negative, and then compared it to the actual value. The accuracies for the models are:

Model2: *0.715*
Model3: *0.76*

Figure 1.6 shows that the accuracy of each model, the left graph is model2, the right one is model3. However, for more precise test, a new dataset is used to assess accuracy of predicted model.

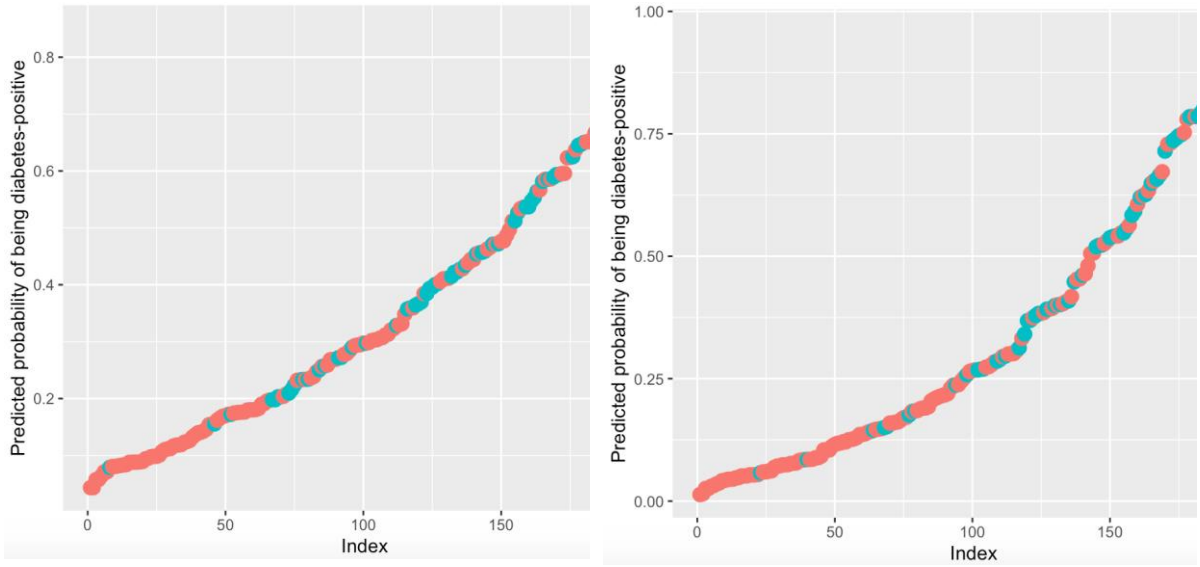

Figure 1.6 *Predicted probability of being diabetes positive with actual value for model2 (left) and model3 (right)*

The accuracies for the Pima.te (test data) are:

Model2: *0.7108434*
Model3: *0.7891566*

In conclusion, the model3 is better for predicting probability of diabetes being positive. Not only is the simple reason for its accuracy, but also the results of accuracy tests using Pima.te are better. It shows that model3 can generalize well for the future data.

However, it is remarkable that the accuracy of the predicted model without laboratory data set is at a similar level with using a laboratory dataset.

Task 2: Time until progression of prostate cancer

With the R Data manager, find the dataset stagec in the package rpart. The dataset contains data on 146 patients with prostate cancer, including the time to disease progression, the age of the patient, the histopathological grade of the tumour according to two different systems (Farrow and Gleason), and the ploidy state (a chromosome classification).

We will study the length of the time to progression of the disease, and how it is related to the ploidy state, the histopathological grade, and age.

A) Make a graphical presentation of the time to progression for the whole group and for each ploidy group.

The given data includes 146 patients with prostate cancer. Table 2-1 shows an overview of the ploidy groups including the number of all patients and the patients with progression (pgstat = 1) as well as the median time of progression (for patients with progression) or last check up (for patients without progression).

Table 2-1: *Overview of the ploidy groups*

Ploidy group	Number of patients	Number of patients with progression	Median time of progression (All patients)	Median time of progression (Patients with progression)
Diploid	67	13	6.1	3.2
Tetraploid	68	34	5.7	3.4
Aneuploid	11	7	3.4	3

It can be seen in figure 2-1, that the distribution of the progression time differs between the three ploidy groups while the median is similar (compare also Figure 2-1).

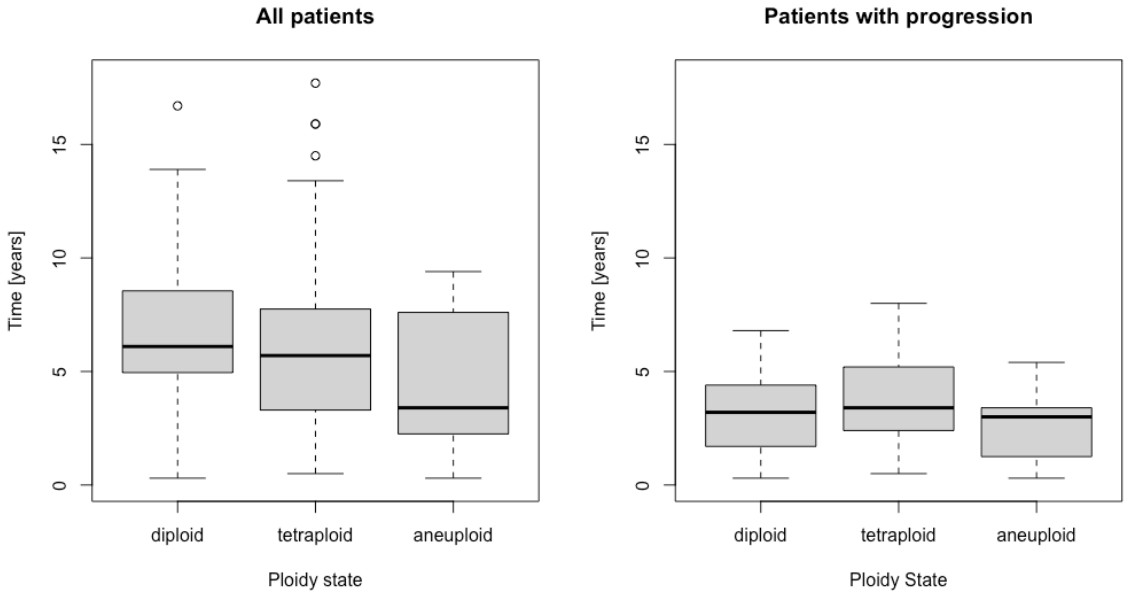

Figure 2-1: *Progression time of the three ploidy groups for all patients (left) and patients with progression (right)*

The median time of progression for all patients is longer than for patients with progression (compare Table 2-1). In Figure 2-2 the difference is visualized.

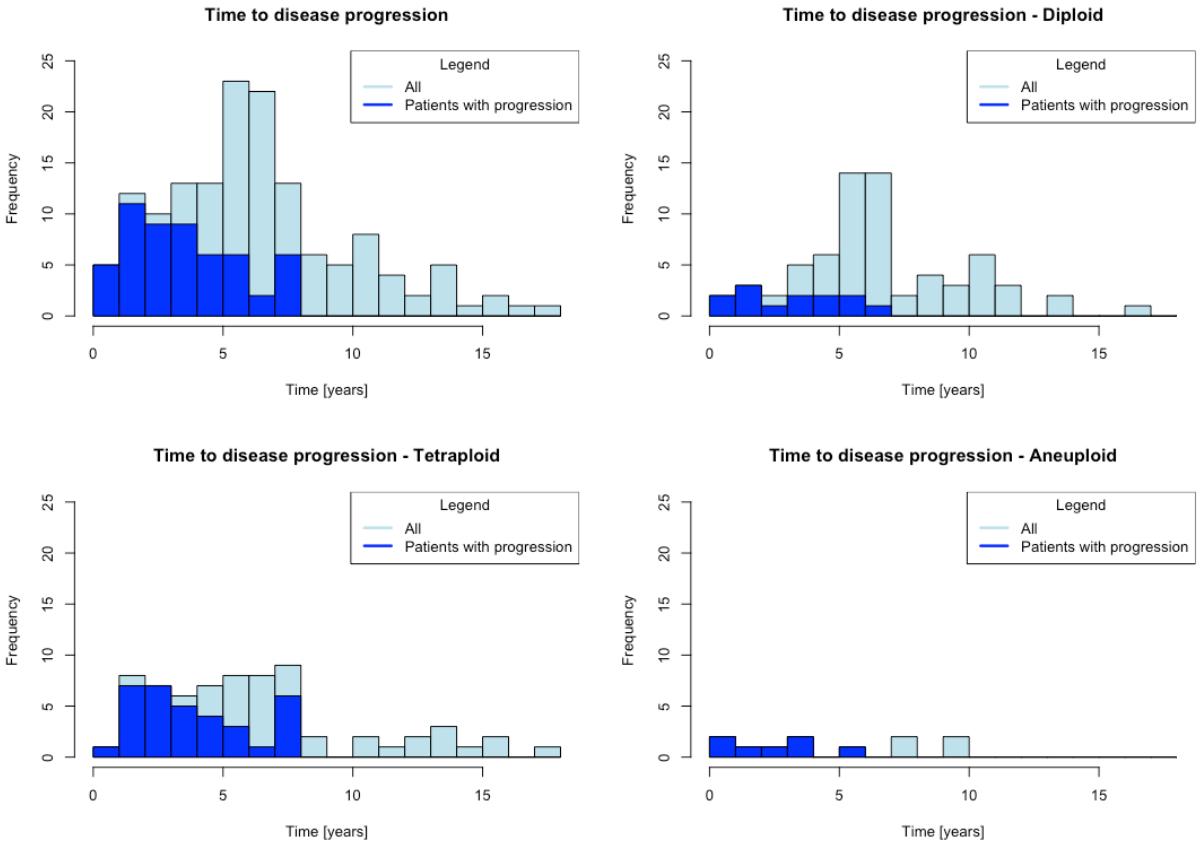

Figure 2-2: *Comparison between the distributions of the progression time between all patients (light blue) and patients with progression (dark blue) for all ploidy groups together and separate*

B) Analyze the time to progression for the different ploidy groups.

The time to progression can be investigated using a survival analysis. Therefore, the package “Survival” is used in R. First, the survival data is generated:

survival_data = Surv(pgtime,pgstat)

As the follow up time “pgtime” is used, whereas “pgstat” indicates whether an event is happening (progression) or not. Next, the data is fitted to analysis the difference between the ploidy groups:

survival_fit = survfit(survival_data ~ ploidy, data=stagec)

The data can be plotted using a survival function (Figure 2-3) or a hazard function (Figure 2-4). Second shows a higher hazard for the aneuploid status and the lowest for the diploid status.

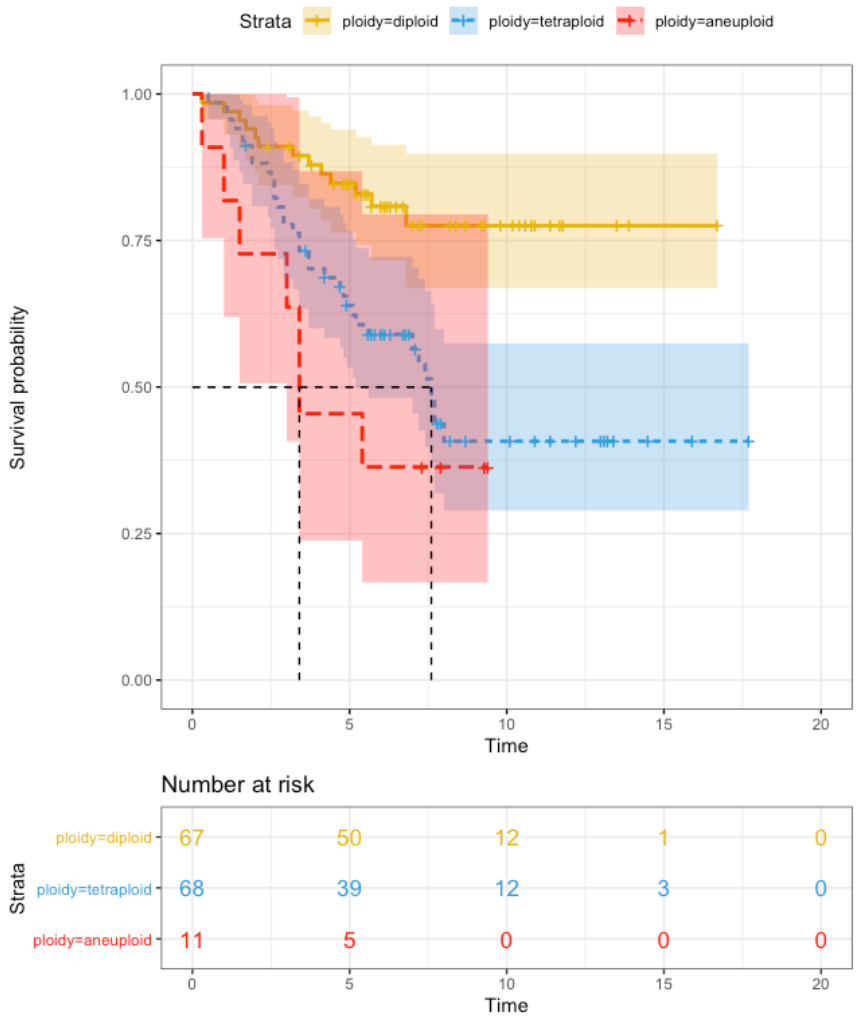

Figure 2-3: *Survival function*

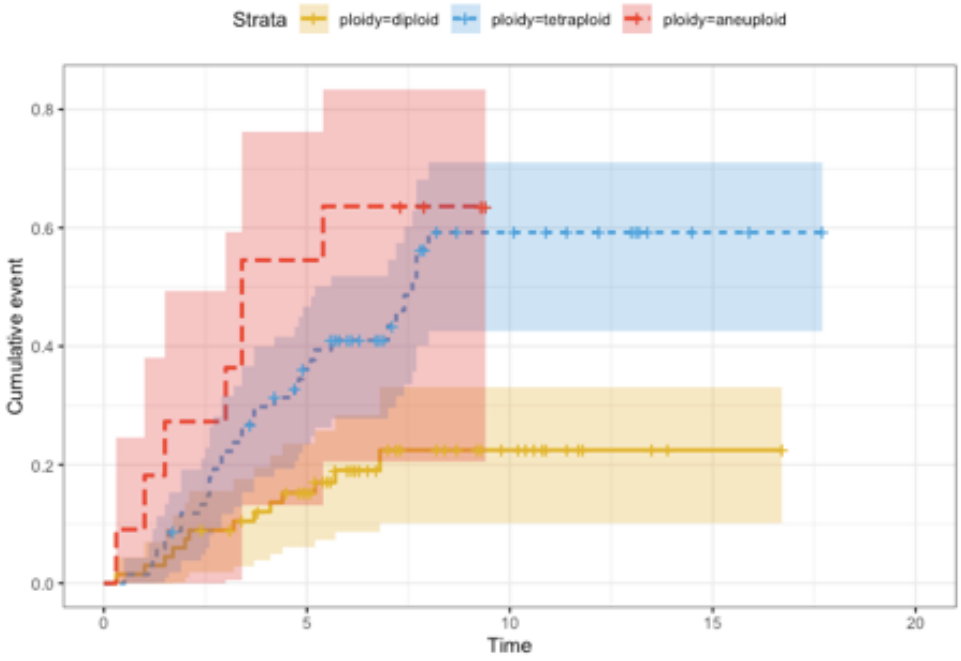

Figure 2-4: *Hazard function*

Even though the plots (Figure 2-3 and 2-4) indicate that there is a difference between the ploidy groups in terms of the progression time, tests are performed to confirm that observation. The null hypothesis is, that there is no difference between the ploidy groups.

First, a simple log-rank test (Figure 2-5) is used. It shows that there is a difference between the groups (p = 6e-04). Another, more detailed, test method is a fit of a cox proportional hazards regression model. The results are shown in Figure 2-6. All three overall-tests show that there is a difference between the groups (the log-rank test is also included). Furthermore, we can see that having a tetraploid status increases the hazard rate (HR) by 2.85 in comparison to having a diploid status. The aneuploid group has an increase of the HR by even 4.34. These results are highly significant.

```
> surv_diff = survdiff(survival_data ~ ploidy, data = stagec)
> surv_diff
Call:
survdiff(formula = survival_data ~ ploidy, data = stagec)

                    N Observed Expected (O-E)^2/E (O-E)^2/V
ploidy=diploid     67       13    26.33      6.75     13.29
ploidy=tetraploid 68       34    24.35      3.82      7.02
ploidy=aneuploid  11        7     3.32      4.08      4.38

 Chisq= 14.8  on 2 degrees of freedom, p= 6e-04
```

Figure 2-5: *Results of the log-rank test*

```
> summary(data_cox)
Call:
coxph(formula = survival_data ~ ploidy, data = stagec)

  n= 146, number of events= 54

                   coef exp(coef) se(coef)     z Pr(>|z|)
ploidytetraploid 1.0454    2.8445   0.3263 3.203  0.00136 **
ploidyaneuploid  1.4681    4.3411   0.4696 3.127  0.00177 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

                 exp(coef) exp(-coef) lower .95 upper .95
ploidytetraploid     2.845     0.3516     1.500     5.393
ploidyaneuploid      4.341     0.2304     1.729    10.897

Concordance= 0.634  (se = 0.036 )
Likelihood ratio test= 14.98  on 2 df,   p=6e-04
Wald test            = 13.33  on 2 df,   p=0.001
Score (logrank) test = 14.87  on 2 df,   p=6e-04
```

Figure 2-6: *Results of cox proportional hazards regression model for simple model*

To sum up, the progression time depends significantly on the ploidy group if no other factors are included. The diploid group has the lowest hazard, while the aneuploid group has the highest hazard.

C) Repeat this analysis taking into account also the age and the histopathological gradings.

Before analyzing the data, the three new factors are assessed. Table 2-2 shows an overview of the median ages in the different ploidy groups.

Table 2-2: *Overview of age in different ploidy groups*

Ploidy group	Median age (All patients)	Median age (Patients with progression)
Diploid	63	62
Tetraploid	64	64
Aneuploid	62	59

The first histopathological grading variant is the farrow grading system (more information). As can be seen in Table 2-3, most patients were diagnosed with grade 2 cancer (59 patients) or grade 3 cancer (79 patients). Most patients with progressions have been diagnosed with stage 3 cancer (39 patients). However, the percentage of patients with progression is highest for grade 4 cancer (100%), while just 49.4% of patients with grade 3 cancer and 15.3% of patients with grade 2 cancer are diagnosed with progression. Just two patients were diagnosed with grade 1 cancer and both haven't been diagnosed with progression.

Table 2-4 shows the median progression time of the different ploidy groups depending on the cancer grading.

Table 2-3: *Number of patients for farrow grade*

Patient group	Number of patients			Number of patients with progression		
Grade	**diploid**	**tetraploid**	**aneuploid**	**diploid**	**tetraploid**	**aneuploid**
1	2	0	0	0	0	0
2	36	22	1	2	7	0
3	27	43	9	9	24	6
4	2	3	1	2	3	1

Table 2-4: *Overview of median progression time for different ploidy groups and grades*

Patient group	**Alle patients**			**Patients with progression**		
Grade	**diploid**	**tetraploid**	**aneuploid**	**diploid**	**tetraploid**	**aneuploid**
1	12.7	-	-	-	-	-
2	6.6	5.95	7.9	3.65	4.7	-
3	5.2	5.6	3.4	3.7	3.55	3.2
4	0.65	2.5	1.5	0.65	2.5	1.5

The second grading system is the gleason grading. It is divided into ten stages. A first look at the data shows a similar trend as for the farrow grading system. Most patients are diagnosed with a middle Gleason score, while the percentage of patients with progression is higher as higher the Gleason score is.

Table 2-5: *Number of patients for gleason grade*

Patient group	**Number of patients**			**Number of patients with progression**		
Grade	**diploid**	**tetraploid**	**aneuploid**	**diploid**	**tetraploid**	**aneuploid**
3	2	0	0	0	0	0
4	4	2	0	1	1	0
5	22	12	1	0	3	0
6	16	15	3	3	8	2
7	14	21	3	5	11	2
8	5	13	3	2	8	2
9	0	4	1	0	3	1
10	2	0	0	2	0	0

A survival test is performed to investigate how the progression time is influenced by age and grading. The fit is analyzed by using the cox proportional hazards regression model. The results are shown in Figure 2-7.

```
> cox2 = coxph(survival_data ~ ploidy + age + grade + gleason, data = stagec)
> summary(cox2)
Call:
coxph(formula = survival_data ~ ploidy + age + grade + gleason,
    data = stagec)

  n= 143, number of events= 54
   (3 observations deleted due to missingness)

                    coef exp(coef) se(coef)      z Pr(>|z|)
ploidytetraploid  0.47165   1.60264  0.34482  1.368  0.17137
ploidyaneuploid   0.71585   2.04593  0.48845  1.466  0.14276
age              -0.03352   0.96703  0.02622 -1.279  0.20104
grade             1.34982   3.85674  0.42346  3.188  0.00143 **
gleason           0.14718   1.15857  0.16569  0.888  0.37437
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

                exp(coef) exp(-coef) lower .95 upper .95
ploidytetraploid     1.603     0.6240    0.8153     3.150
ploidyaneuploid      2.046     0.4888    0.7855     5.329
age                  0.967     1.0341    0.9186     1.018
grade                3.857     0.2593    1.6818     8.844
gleason              1.159     0.8631    0.8373     1.603

Concordance= 0.738  (se = 0.034 )
Likelihood ratio test= 43.25  on 5 df,   p=3e-08
Wald test            = 35.01  on 5 df,   p=1e-06
Score (logrank) test = 38.91  on 5 df,   p=2e-07
```

Figure 2-7: *Results of cox proportional hazards regression model for complex model*

All overall tests show that the model is significant. The only significant influence is the (farrow) grade. A higher grade by one lead to an increase in hazard by 3.86. A higher Gleason grade also leads to an increase in hazard (HR = 1.16), but the result is not significant.

The age leads to a decrease in hazard (by 0.97). However, this result is also not significant. For a visualization of the influence of the three factors (age, Farrow Grade, and Gleason Grade), three survival models were created and plotted (Figure 2-8).

In comparison to the first model, the difference of ploidy states is not significant anymore.

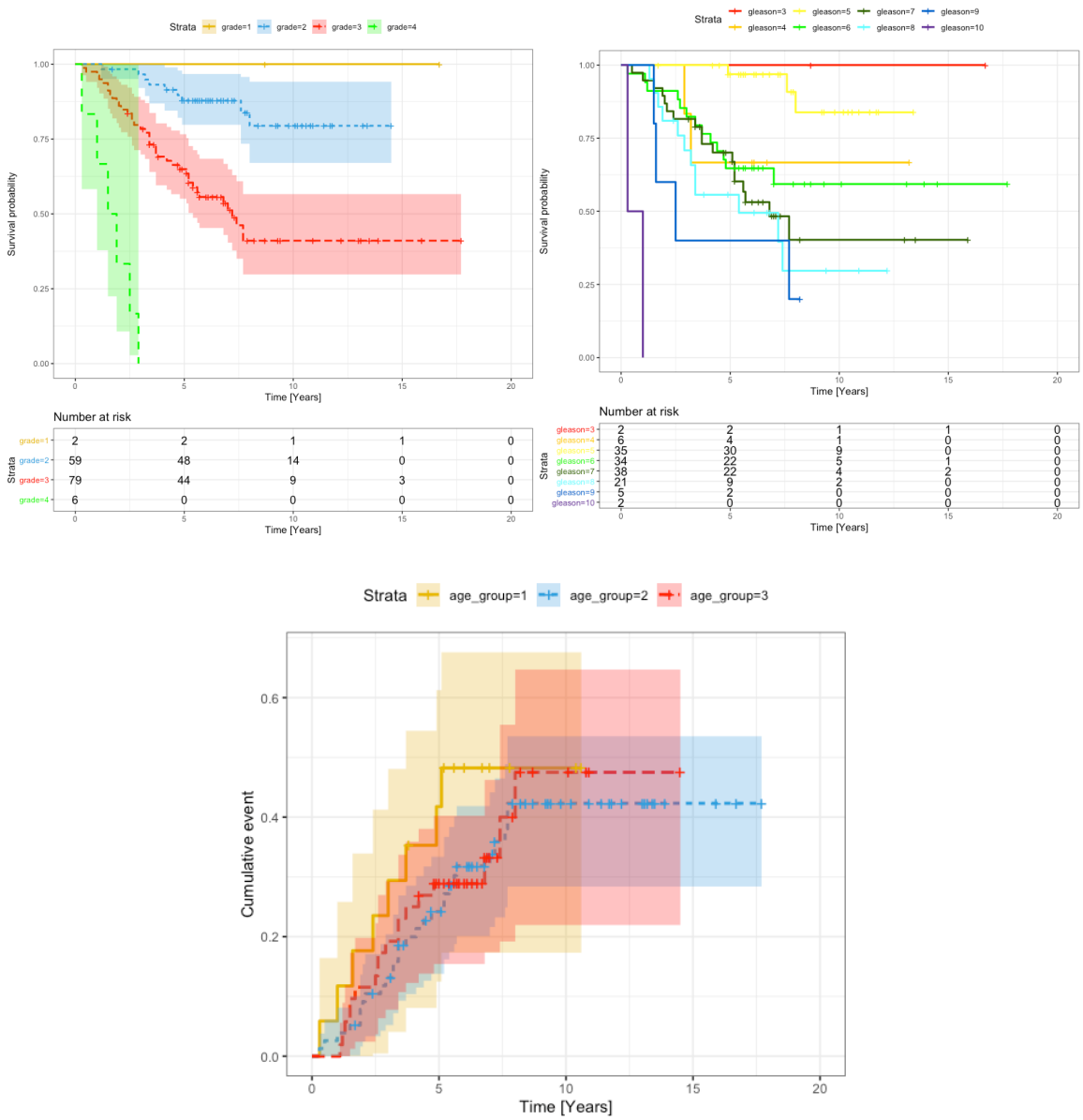

Figure 2-8: *Survival functions for different factors grade, gleason, and age (thress age groups: 46-55 years (group1), 56-65 years (group2) and 66-75 years (groups 3)*

D) Interpret your findings and try to explain any inconsistencies.

The time to disease progression significantly correlates with the Farrow grade. The higher the grade the faster a disease progression occurs. A similar observation was made for the Gleason grading system, but this result was not significant. A possible reason could be the small subject size. Since the Gleason grading is divided into 10 steps, for some grades just a couple of patients were diagnosed.

The influence of the ploidy state was significant in the simple model, but not significant in the complex model. This shows, that after adjusting the model to other factors as age and grading, the ploidy state makes a smaller contribution to the difference in the hazard rate.

Task 3: Analysing data with a large number of variables

You have been tasked with proposing a limited set of proteins that could be used to develop an early diagnostic assay for the severity of a rare inflammatory disease.
To your disposal you have a proteomic dataset from 100 patients consisting of an inflammatory response biomarker, which you assume to represent the severity of the disease, and the expression of 150 endogenous proteins.

Task: Discuss potential regression methods and what would be a suitable methodology for this task. Carry out the analysis in R and present your findings. Discuss what additional analysis you would like to carry out to build confidence in your model.

3.1 First Approach - Correlation

The data (Figure 3.1) is presented as observations of inflammatory markers and 150 protein expressions of 100 subjects. The goal is to correlate high inflammatory responses to certain protein expressions to determine relationship.

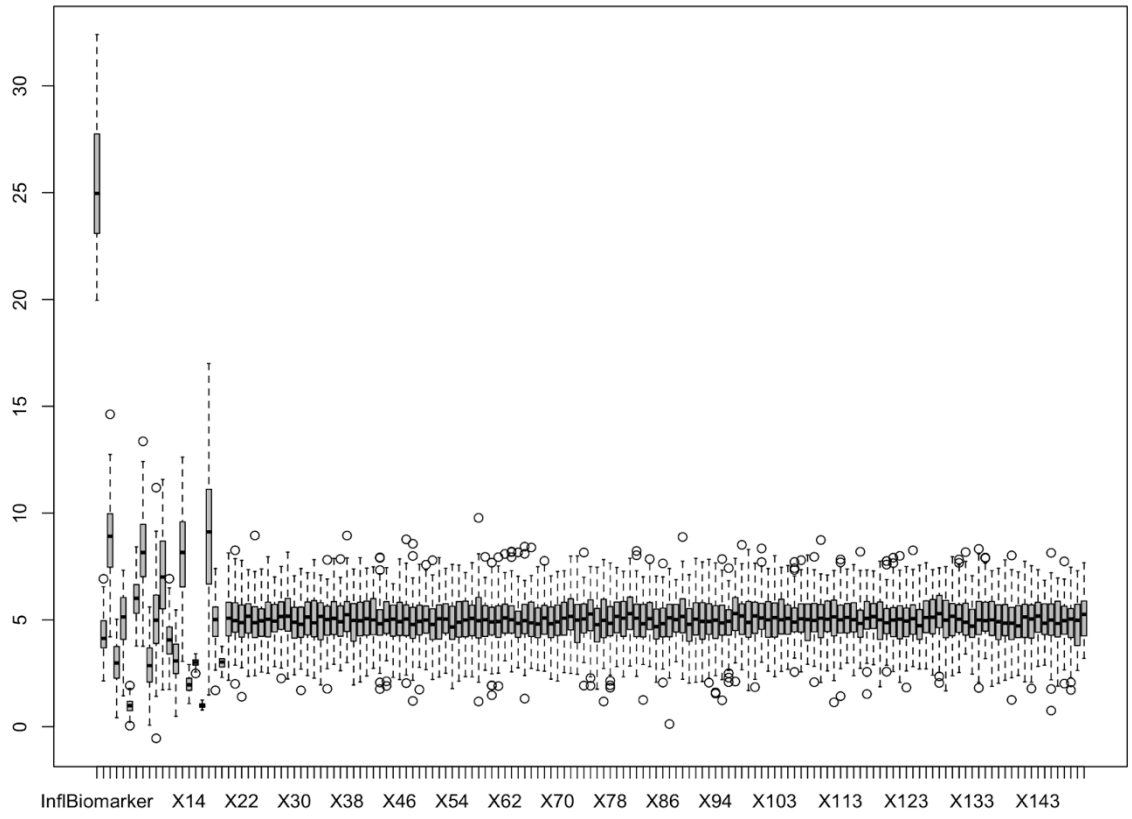

Figure 3.1. Boxplot of raw data.

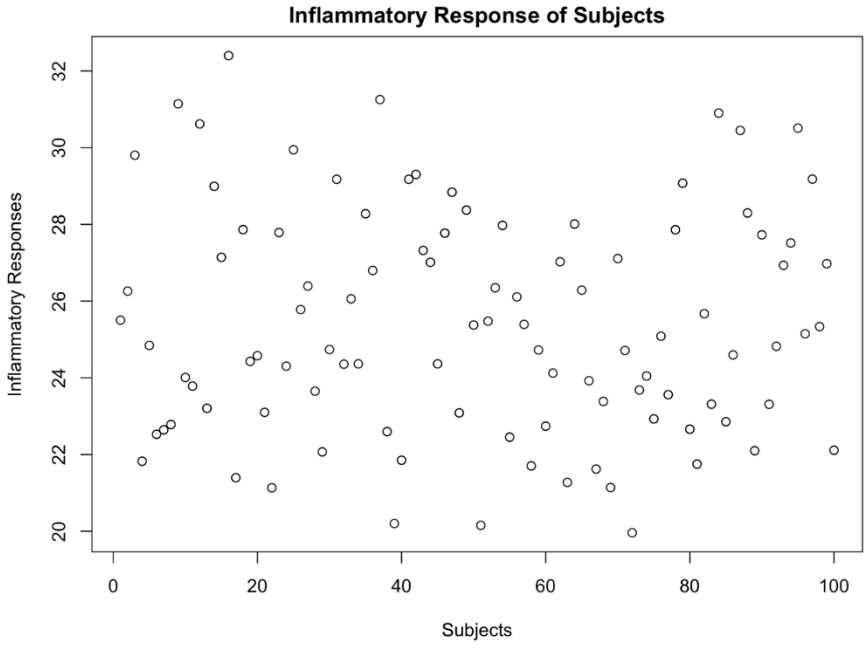

Figure 3.2. Inflammatory response for each of the 100 subjects.

Because we have a large data set with a large number of variables to check for correlation against the inflammatory biomarkers, creating a model with all of the 150 protein variables would make up a large list of data to go through and would make for the possibility of 2^150 different model (also known as combinatoric explosion).

It would be appropriate to quickly calculate the relationships and extract the highest correlations. In order to do that, we first create a correlation matrix. We then extract the relevant element, how inflammatory response relate to proteins. This will provide a 1x151 array.

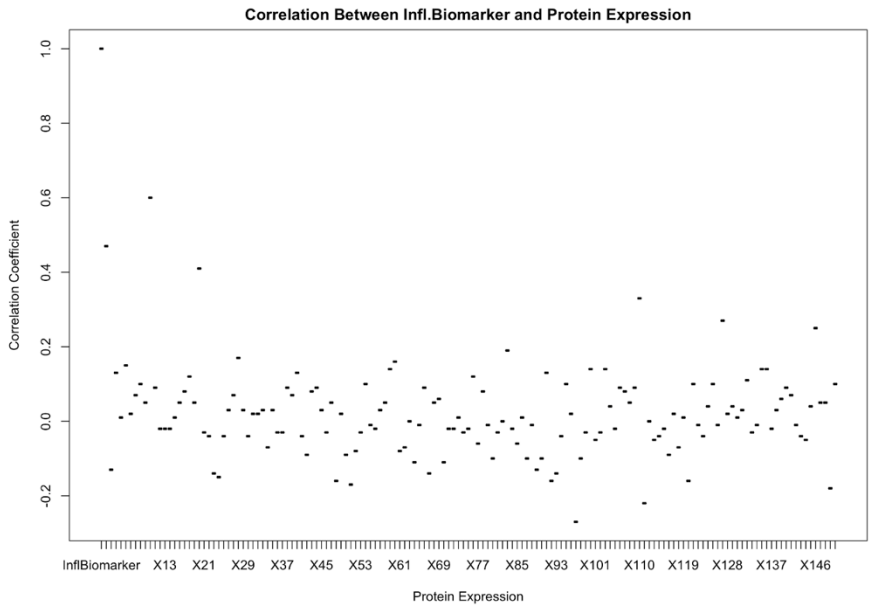

Figure 3.3. Plot of correlation array between inflammatory response and the expression of the 150 different proteins.

As we can see in figure 3.3 the correlation is higher for some, and lower for some proteins. This might indicate that some proteins have a more significant impact on the inflammatory response, either by over- or under expression.

We create a sorted array of the correlation coefficients and extract the ones suitable from figure 3.3. We choose the 6 top overexpression and lowest two under expressions as they seem to have a different correlation than the rest of the proteins. These are:

- Overexpression: X10, X1 X20, X110, X127 and X146
- Underexpression: X58 and X47

A generalized model1 with gaussian identity family was created. The following summary of the model (figure 3.4) was expressed, indicating only 4 significant variables and an AIC score of 330.54.

```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.6787  -0.7726   0.0347   0.8197   3.4698

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  6.18793    1.24479   4.971 3.11e-06 ***
X10          0.89368    0.05948  15.025  < 2e-16 ***
X1           1.63790    0.13560  12.079  < 2e-16 ***
X20          0.84180    0.10237   8.223 1.32e-12 ***
X110         0.23041    0.10886   2.116    0.037 *
X127        -0.01668    0.11308  -0.148    0.883
X146         0.11365    0.11642   0.976    0.332
X58          0.10290    0.09070   1.135    0.260
X47         -0.08666    0.10544  -0.822    0.413
```

Figure 3.4. Summary of model1.

A model2 was created including only the significant variables X10, X1, X20 and X110. This produces the following summary (figure 3.5) and an AIC score of 326.13 indicating a better fit than model1.

```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.5979  -0.6807  -0.0733   0.7573   3.3670

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  6.65597    0.90115   7.386 5.79e-11 ***
X10          0.89416    0.05669  15.772  < 2e-16 ***
X1           1.66163    0.12854  12.927  < 2e-16 ***
X20          0.83943    0.09981   8.410 4.09e-13 ***
X110         0.23037    0.10750   2.143   0.0347 *
```

Figure 3.5. Summary of model2

Next, glm models with different families were tested and assesed according to AIC score.

Family Gamma(link="log") got an AIC score of 323.28 which is an improvement from previous model with gaussian family. None of the other tested families provided a decrease in AIC score. The best fit model with lowest AIC score was the gamma link model, and figure 3.6 shows the regression diagnostics which are okay, but not ideal.

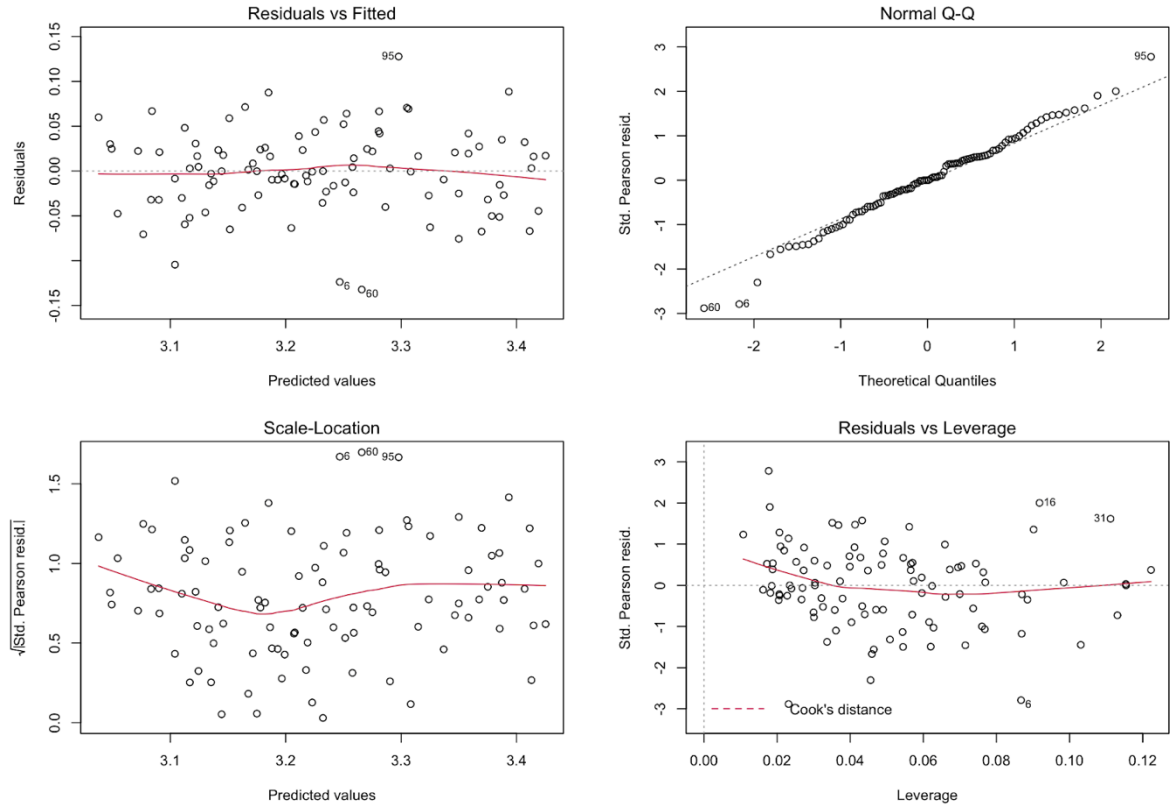

Figure 3.6. Regression diagnostics of gamma link model.

In order to build further confidence in the model, a lot more variables would have to be included. When performing a process of forward selection, backward elimination or a combination of both, the AIC value reaches as low as -5925, but it includes the majority of proteins in the model, if not all, and this is something we do not want. To get a better model, different methods of regression could be tried, such as ridge and lasso.

3.2 Second Approach: Lasso Regression

A second approach is to find the best model using lasso regression. We split the data into train and test sets and calculated the best lambda for the data using the train set, which was 0.3057224. And with the chosen lambda, the following independent variables were significant.

(Intercept)	X1	X5	X10	X20
10.787034716	1.325507538	0.039859668	0.741652321	0.631807562
X48	X54	X63	X110	
-0.010480167	0.023798246	0.004657438	0.088495419	

Figure 3.7. Significant attributes according to Lasso Regression

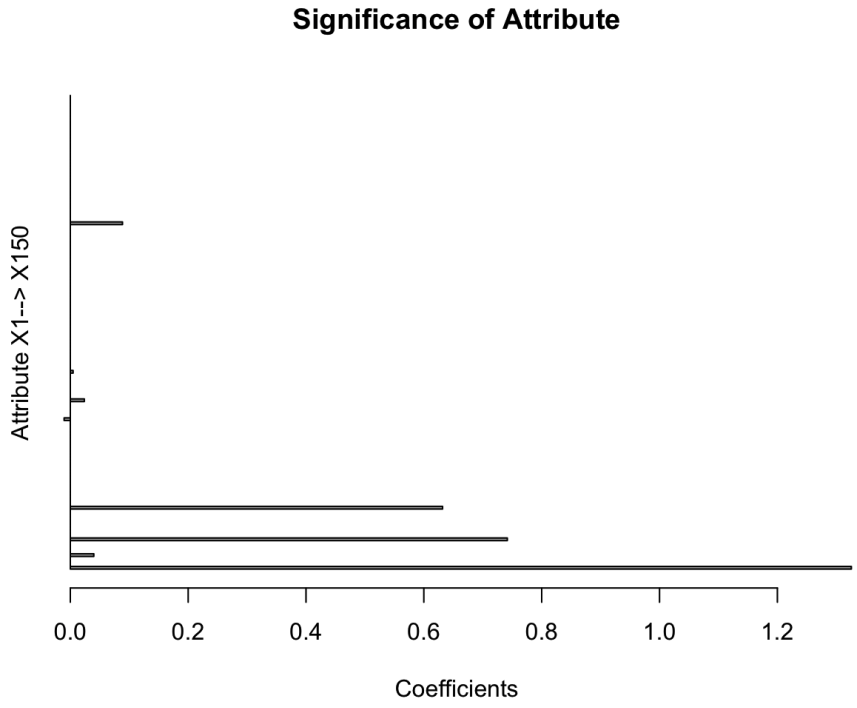

Figure 3.8. Significance of attributes in model with Lasso Regression method.

The lasso model with the significant attributes is:

```
glm(    InflBiomarker~X1+X5+X10+X20+X48+X54+X63+X110,
        family=Gamma(link="log"))
```

which acquired an AIC score of 301.81, which is a definite improvement from the previous method. The regression diagnostics for the lasso model is also a general improvement from the first model, as illustrated in figure 3.8.

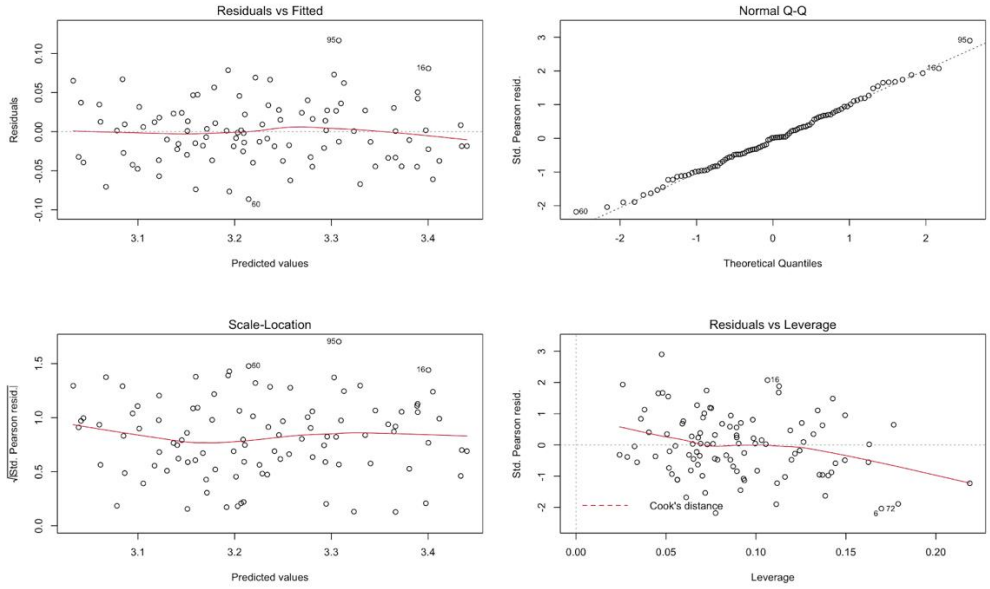

Figure 3.9. Regression diagnostics for Lasso Regression.

3.3 Third Approach: Ridge Regression

The following barchart (figure 3.10) was created with Ridge Regression. As you can see the significance is a lot higher for all of the variables than in the Lasso Regression, and it also includes X16 as the highest contributor. None the less, the model created with Ridge Regression gets a much higher AIC score, so the Lasso Regression is still the most optimal one.

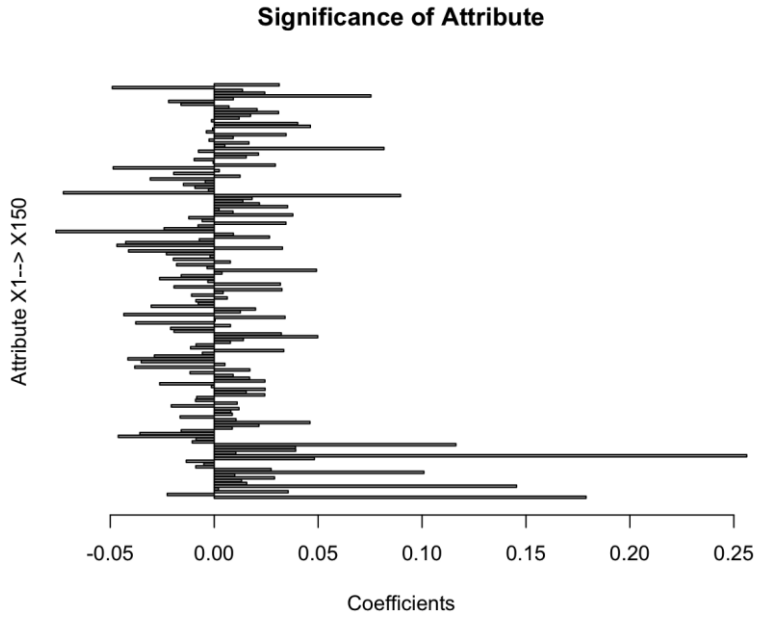

Figure 3.10. Significance of attributes according to Ridge Regression.

Task 4: Modelling complex oncology data

A clinical trial with parallel design was carried out in 100 patients over six months to evaluate a novel treatment for stage 3 and 4 breast cancer as compared to the standard treatment. Progression of disease was measured using tumour size.

The data file, oncdata.csv, contains 300 data points with subject id (named 'Subject'), time of measurement in months ('Months'), measurements of tumour size ('TumourSize'), indicator of stage of disease ('Stage', where stage 3=0, stage 4=1), and indicator of treatment ('Treatment', where control treatment=0 and experimental treatment=1).

A: Using the R library, mle4, develop a linear mixed-effects model to evaluate the treatment effect.

The first step we took was visualizing the data at hand to gain a better understanding of it. In our case the response variable is tumour size and we want to see how it changes over time given a certain treatment (Figure 4-1). The data was from 300 different subjects at the points in time for each subject. We also had additional data of which stage in the disease the subjects were and if they got the treatment or not. The difference between stage 3 (variable 0) and stage 4 (variable 1) is shown in Figure 4-2. The difference between the treatment group (variable 0) and the control group (variable 1) is shown in Figure 4-3).

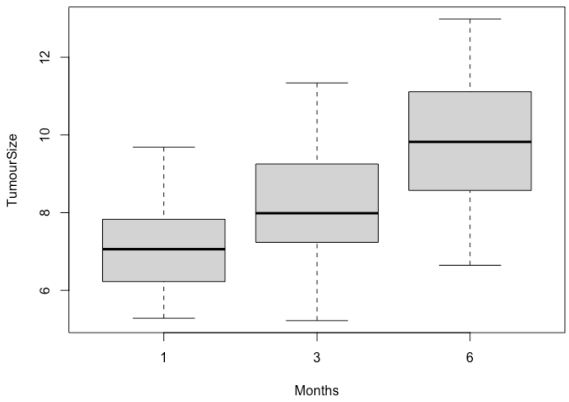

Figure 4-1. *Tumour size in different months*

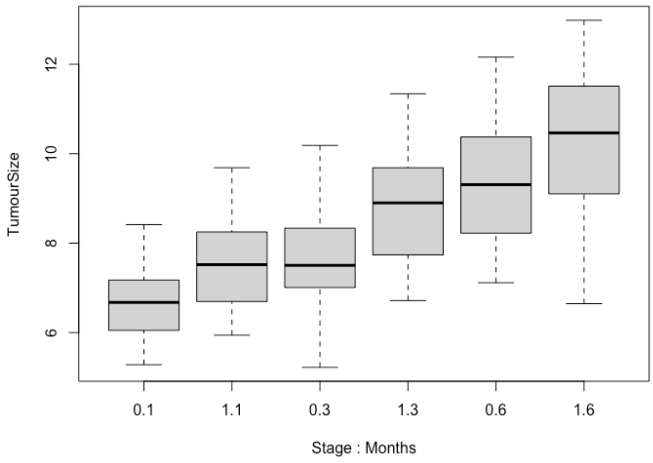

Figure 4-2. *Tumor size in different months for different stages*

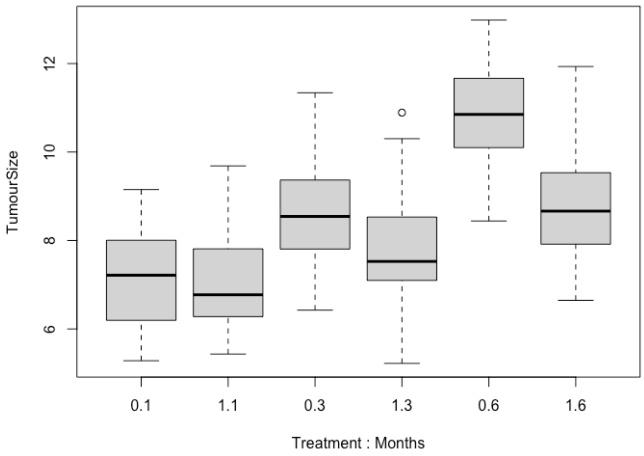

Figure 4-3. *Tumor size in different months for different stages*

In the treatment group we could observe a smaller median tumour size and this was what we wanted to explore further, thus we set treatment as a fixed effect.

When plotting the different stages in disease we could observe that the tumour size mean is higher than the one corresponding to the earlier stage in the disease. When creating our model, we are interested to see if this has an effect on our response variable and thus we choose this to be a fixed effect. We also tried setting subject and stage as crossed random effects because both stages could be found in both treatment groups but we do not necessarily care about the individual differences between stages.

For all models we set the subjects as a random effect, as this is the population of which we are sampling from and want to control for. We do not care about the individual differences within the subjects of the population and we are trying to make inferences from the subjects onto a bigger population.

Lastly, months were set as an fixed effect as we are interested to see how the passing of time affects the tumour size.

After all this evaluation, four different models were chosen. The models can be seen in Figure 4-4.

```
# Mixed-effects models
fit1_model = lmer(TumourSize ~ 1 + Months + Treatment + (1|Stage) + ( 1 | Subject)  , data = oncdata)
fit2_model = lmer(TumourSize ~  1 + Months + Treatment + Stage + ( 1 | Subject)  , data = oncdata)
fit3_model = lmer(TumourSize ~  1 + Months + Treatment + Months:Treatment + (1 | Subject)  , data = oncdata)
fit4_model = lmer(TumourSize ~  1 + Months + Treatment + Stage + Months:Treatment + (1 | Subject)  , data = oncdata)
fit5_model = lmer(TumourSize ~  1 + Months + Stage + Months:Treatment + (1 | Subject)  , data = oncdata)
```

Figure 4-4. *Different mixed-effect models*

The first model was a model where stage and subject were seen as a cross random effect. The second model was a model with stage as a fixed effect. The third model had the interaction of months and treatment added as a fixed effect. The fourth model was the most complex one with stage as a fixed effect and the interaction of months and treatment added. The last model was a reduced variant of the fourth model.

B: Discuss how you can assess goodness-of-fit of the model and how you can compare performance between models.

The model performance was evaluated by using an ANOVA test and comparing the different AIC, BIC and log likelihood scores. This can be seen in the figure below. The best model we could create was the most complex one and had the subjects as a random effect and stage as a fixed effect. This makes sense as the state of your health probably has an effect on how well you respond to treatment.

```
Data: oncdata
Models:
fit1_model: TumourSize ~ 1 + Months:Treatment + (1 | Stage) + (1 | Subject)
fit2_model: TumourSize ~ 1 + Months:Treatment + Stage + (1 | Subject)
fit3_model: TumourSize ~ 1 + Months + Treatment + Months:Treatment + (1 |
fit3_model:     Subject)
fit5_model: TumourSize ~ 1 + Months + Stage + Months:Treatment + (1 | Subject)
fit4_model: TumourSize ~ 1 + Months + Treatment + Stage + Months:Treatment +
fit4_model:     (1 | Subject)
           npar     AIC     BIC  logLik deviance    Chisq Df Pr(>Chisq)
fit1_model    5 1155.22 1173.74 -572.61  1145.22
fit2_model    5 1148.77 1167.29 -569.39  1138.77   6.4487  0    < 2e-16 ***
fit3_model    6  670.00  692.22 -329.00   658.00 480.7727  1    < 2e-16 ***
fit5_model    6  649.08  671.31 -318.54   637.08  20.9157  0    < 2e-16 ***
fit4_model    7  647.77  673.70 -316.89   633.77   3.3140  1    0.06869 .
```

Figure 4-5. *ANOVA test to compare the performance of different models*

When evaluating the fourth model (Figure 4-6) we could see that treatment was not significant but if we would have dropped that variable we would essentially have model 5 which performed worse than the fourth model. Thus, we stayed with the fourth model as our final model. Also, we could see that the interaction between months and treatment (months:treatment) was significant and had a negative value which means that the treatment had an effect on the subjects and that it was contributing to lower tumour size.

```
Linear mixed model fit by REML. t-tests use Satterthwaite's method ['lmerModLmerTest']
Formula: TumourSize ~ 1 + Months + Treatment + Stage + Months:Treatment +      (1 | Subject)
   Data: oncdata

REML criterion at convergence: 652.1

Scaled residuals:
    Min      1Q  Median      3Q     Max
-2.2754 -0.5177 -0.0296  0.5948  2.5314

Random effects:
 Groups   Name        Variance Std.Dev.
 Subject  (Intercept) 0.7968   0.8926
 Residual             0.2141   0.4627
Number of obs: 300, groups:  Subject, 100

Fixed effects:
                 Estimate Std. Error        df t value Pr(>|t|)
(Intercept)       5.89878    0.16881 127.25257  34.943  < 2e-16 ***
Months            0.74659    0.01839 198.00000  40.606  < 2e-16 ***
Treatment         0.37134    0.20555 140.27411   1.807    0.073 .
Stage             0.96293    0.18672  97.00000   5.157 1.33e-06 ***
Months:Treatment -0.38989    0.02600 198.00000 -14.995  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Correlation of Fixed Effects:
            (Intr) Months Trtmnt Stage
Months      -0.363
Treatment   -0.599  0.298
Stage       -0.509  0.000 -0.018
Mnths:Trtmn  0.257 -0.707 -0.422  0.000
```

Figure 4-6 *Summary of model4*

When evaluating the residual plot (Figure 4-7) we could also see a good spread from model 4 as thus assess the goodness-of-fit of the model to be high.

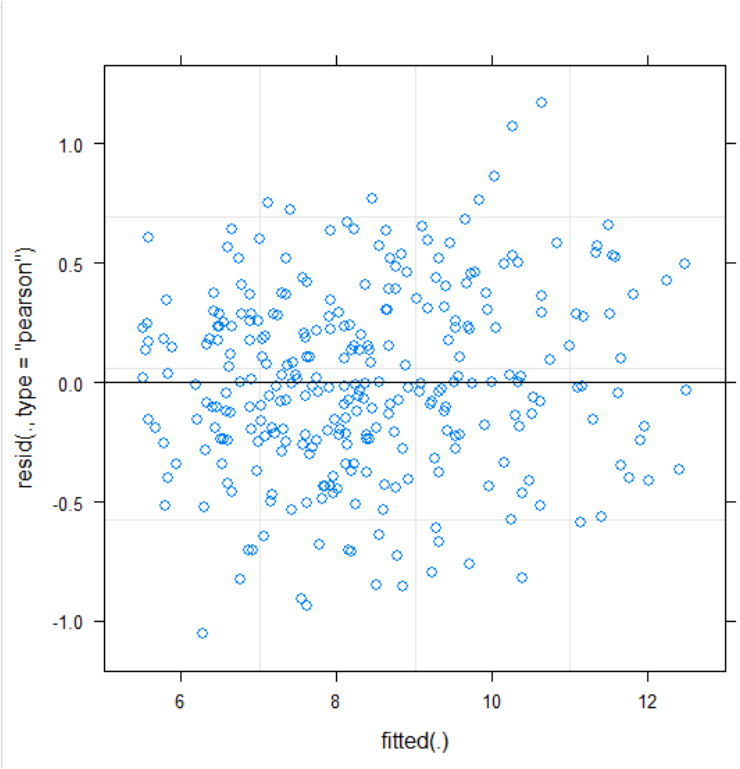

***Figure 4-7**. Residual plot for model4*

C: Discuss alternative methods for analysing the data.

One alternative method for analyzing the data would be to make a generalized linear model and then evaluate its performance similar to the other tasks of this seminar. Another method would be to make a model with treatment as a fixed effect and a model where treatment is not included at all.

Lastly, one could also do a unpaired t-test where you take the difference between the start and end of the time period and compare those between treatment and control groups. All of these are valid options of evaluating treatment effects on tumour size.