Assignment for Regression

-Hreinn Gauti Bjarnason -Emil Ørum Thomsen -Emma Sofie Severin Pagaard -Philip von Brockdorff

15/03/2020

```
titanic <-read.csv("http://www.math.ku.dk/~susanne/titanic.txt",header =TRUE,
                   colClasses=c("factor","factor","factor","numeric","integer","integer"))
titanic2 <- titanic %>%
  mutate(sib1 = as.factor(ifelse(sibsp >0,1,0)),
         parch1 = as.factor(ifelse(parch>0,1,0)))
```

Introduction

We will in this report analyse data regarding passengers of Titanic. Each observation corresponds to a passenger. The dataset only regards passengers and not any crewmembers. The data contains information on seven variables:

- **Pclass:** Which class the passenger was on (1,2,3).
- **Survived:** (1=yes, 0=no).
- **Sex:** Female or Male.
- **Age:** Age of the passenger.
- **SibSp:** Amount of siblings or spouses the passenger had onboard.
- **Parch:** Amount of parents or children the passenger had onboard.

Where we consider *Pclass*, *Sex*, *Survived* as categorical factors and the rest as numeric.
We will draw inferens regarding Survival.

Parameters to consider

To consider what parameters to include we will look at contigency tables of *survived* $\times$ other factors, while conditioning on the other factors.

```
kable(table(titanic$survived,titanic$sex),caption ="Sample count of survival\\label{tab:example} ")%>%
  kable_styling(latex_options = "HOLD_position")
```

Table 1: Sample count of survival

	female	male
0	127	682
1	339	161

```
kable((addmargins(prop.table(table(titanic$survived,titanic$sex),margin=2),margin=1)),caption ="Probabi
  kable_styling(latex_options = "HOLD_position")
```

Table 2: Probability of survival given sex

	female	male
0	0.2725322	0.8090154
1	0.7274678	0.1909846
Sum	1.0000000	1.0000000

On table 2 we see that the conditional probability of dying if you are a man (0.809) is much higher than it is for women $P(Survived = 0|Sex = F) = 0.273$, even higher than the conditional probability of surviving if you are a woman $P(Survived = 1|Sex = F) = 0.727$. From table 1 we can see that the sample odds ratio of men dying compared to women dying is $\frac{682\times339}{127\times161} = 11.3$ meaning men were much more likely to die than women.

Table 3: Probability of survival given pclass

	1	2	3
0	0.380805	0.5703971	0.7447109
1	0.619195	0.4296029	0.2552891
Sum	1.000000	1.0000000	1.0000000

Table 4: Probability of survival given sibsp

	0	1	2	3	4	5	8
0	0.6531987	0.4890282	0.547619	0.7	0.8636364	1	1
1	0.3468013	0.5109718	0.452381	0.3	0.1363636	0	0
Sum	1.0000000	1.0000000	1.000000	1.0	1.0000000	1	1

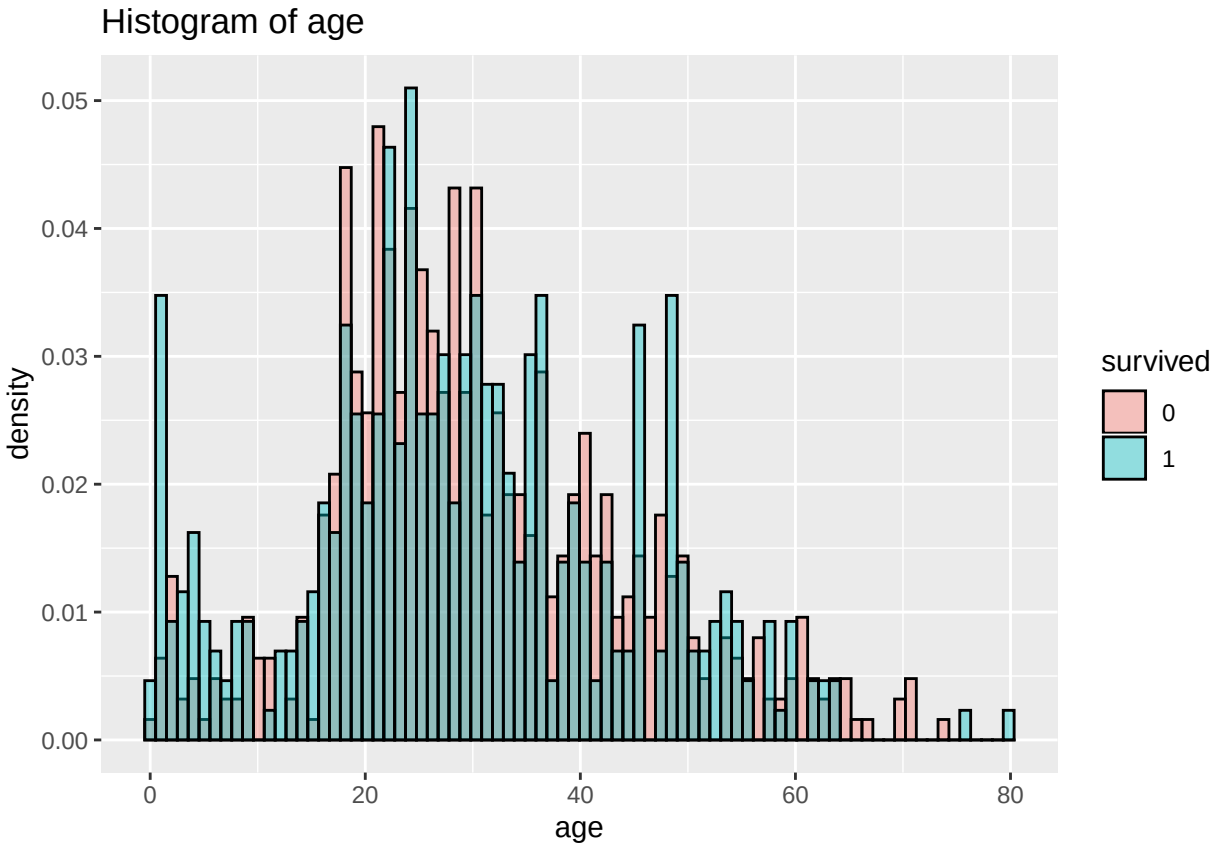
Table 5: Probability of survival given parch

	0	1	2	3	4	5	6	9
0	0.6646707	0.4117647	0.4955752	0.375	0.8333333	0.8333333	1	1
1	0.3353293	0.5882353	0.5044248	0.625	0.1666667	0.1666667	0	0
Sum	1.0000000	1.0000000	1.0000000	1.000	1.0000000	1.0000000	1	1

Now in table 3 we see that first class passengers had a higher conditional survival probability than dying. While the other classes had the reverse effect. In regards to sibsp and parch we note that both $P(Surv = 1|Sibsp = 0)$ and $P(Survived = 1|Parch = 0)$ are approximately $\frac{1}{3}$, while $P(Survived = 1|Sibsp = 1), P(Surv = 1|Parch = 1)$ is roughly $\frac{1}{2}$, likewise for passengers with $Sibsp = 2, Parch = 2$. Perhaps, showing that couples without or with a single child were more likely to survive, as were the single child of a couple.
Higher values of $Parch$ and $Sibsp$ show larger conditional probablities for death (except $Parch = 3$). Furthermore, we note that the amount for these are rather low, so we do not infer much more regarding these.

For the variable *age* we look a histogram, seen below, and conclude that there is a difference between the once who survived and those who didn't. Therefore we will consider the variable in our models.

```
ggplot(titanic, aes(age, fill = survived, y = ..density..)) + geom_histogram(color = "black", alpha = 0
  ggtitle("Histogram of age")
```



In conclussion, all factors have an influence on survival, hence we will choose to include all.

As we noted previously, the larger families are a miniority of the dataset. We have in this study chosen to change both *Parch* and *Sibsp* to binary factors where 0 is as given (no siblings/spouses/children/parents), and 1 indicates that the passenger had one or more. These changes of course give rise to new contigency tables

Table 6: Probability of survival given sib1

	0	1
0	0.6531987	0.5430622
1	0.3468013	0.4569378
Sum	1.0000000	1.0000000

Table 7: Probability of survival given parch1

	0	1
0	0.6646707	0.465798
1	0.3353293	0.534202
Sum	1.0000000	1.000000

Missing data

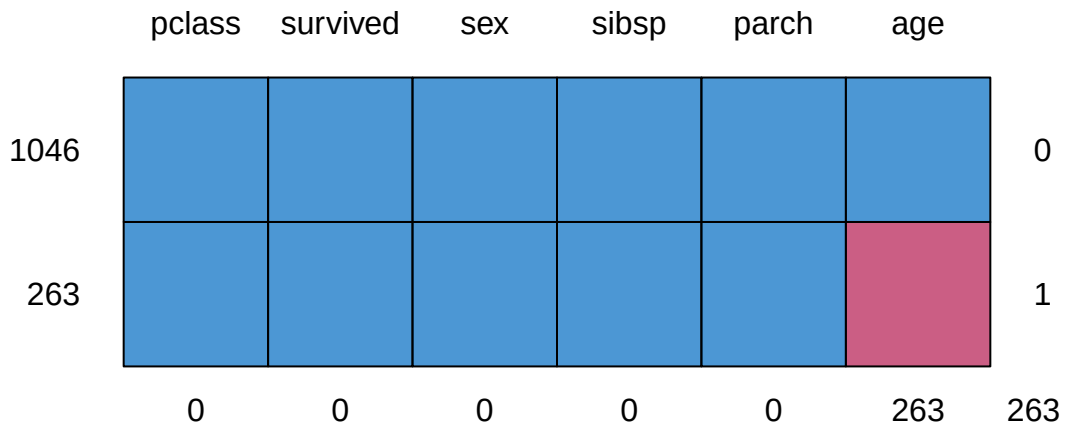
The data was collected by asking the passengers aboard about their age and asking the relatives of the passengers who passed away about their loved ones' age. Therefore, if all the members of a family died, the ages of them all will most likely be missing.

By looking at the summary of our dataset we can see that there is only one variable which has missing data, *age*. The reason why the missing values are not "missing not at random" is because the missing value does not imply a particular value, ie. people with an NA could be spread among all agegroups. It is also not "completely missing at random" because people who died are much more likely to be missing than people who didn't die so it isn't distributed as a "coinflip". We therefore conclude that the missing data is "missing at random".

The plot below shows where the missing data is located, where 1 denotes missing data and 0 denotes a value which is different from NA.

MICE

```
kable(md.pattern(titanic))
```



	pclass	survived	sex	sibsp	parch	age	
1046	1	1	1	1	1	1	0
263	1	1	1	1	1	0	1
	0	0	0	0	0	263	263

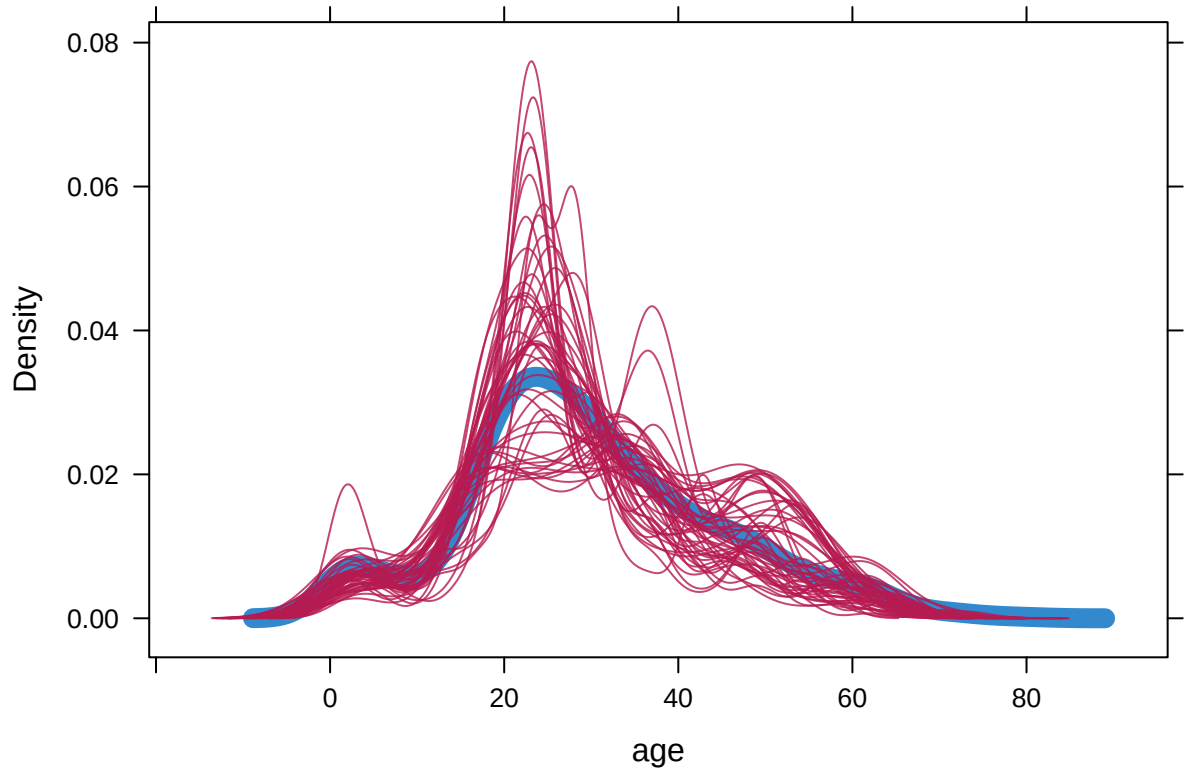
Table 8:

sex	mean	Var	count
female	28.68709	212.4878	466
male	30.58523	203.9350	843

```
kable(titanic %>%
  group_by(sex)%>%
  summarise(mean = mean(age, na.rm =T),
            Var = sd(age, na.rm = T)^2,
            count = n()),caption ="\\label{tab:example}")
```

One way to handle missing values is by imputing the mean of the variable of the non-missing data. However, that is typically not a good idea if the variance is too big. In the table above (Table 8) the variance for the age variable is relatively high. Therefore we decided to use a method called Multiple Imputation by Chained Equations (MICE). As seen in the Figure below, we use the mice function from the package **MICE** to generate 50 new complete datasets from the original one.

```
densityplot(miceMod, thicker = 10)
```



We fit a linear regression model on the variable age on each of the 50 new data sets and then use a function called pool to find the estimate of all of them combined. The estimates for that model can be seen as the following:

```
# We fit the regression model to age
fit <-  with(miceMod, exp = lm(age ~pclass + survived + sex + sib1 + parch1))
```

```
#find an estimate combined of all the models
pooled <- pool(fit)

#define our combined model
pooled_lm <-  fit$analyses[[1]]
pooled_lm$coefficients <-  summary(pooled)$estimate
```

Using that model we predict the missing values of our dataset. We note that the difference between the new and old data is minimal, as seen below:

```
#Make a data set of only the missing values in our dataset
testdata <- titanic2%>%
  filter(is.na(age))
#predict for the missing value using our mice model
newpred<- predict(pooled_lm, newdata = testdata)
NewTitanic <- titanic2
NewTitanic$age[is.na(NewTitanic$age) == TRUE] <- newpred
summary(titanic2$age)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##    0.17   21.00   28.00   29.88   39.00   80.00     263
```

```
summary(NewTitanic$age)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.17   22.00   28.64   29.69   36.50   80.00
```

```
NewTdata2 <- NewTitanic %>%
  dplyr::select(-c(sibsp, parch))
```

We should perhaps consider whether or not this result is good. In a sense we could question whether we can assume that the "missing" data (observation with age as NA) comes from the same (or very similiar) distribution as the rest of the data. For instance, much of the missing data is passengers of third class (so the missing data has proportionally more third class passengers than the rest of the data). We later show that *age* and *pclass* are correlated, hence we might expect the age to be different for the missing data. Furthermore, by inspecting the sibsp and parch variables, the data appears to contain two large families (one with 2 parents and 9 children, and the other with 2 parents and 6 children).

```
subset(titanic,titanic$parch==9)
```

```
##      pclass survived    sex age sibsp parch
## 1180      3        0   male  NA     1     9
## 1181      3        0 female  NA     1     9
```

```
subset(titanic,titanic$sibsp==8)
```

```
##      pclass survived    sex  age sibsp parch
## 1171      3        0   male   NA     8     2
## 1172      3        0   male 14.5     8     2
## 1173      3        0 female   NA     8     2
## 1174      3        0 female   NA     8     2
## 1175      3        0 female   NA     8     2
## 1176      3        0 female   NA     8     2
## 1177      3        0   male   NA     8     2
## 1178      3        0   male   NA     8     2
## 1179      3        0   male   NA     8     2
```

```
subset(titanic,titanic$parch==6)
```

```
##     pclass survived    sex age sibsp parch
## 832      3        0   male  40     1     6
## 833      3        0 female  43     1     6
```

```
subset(titanic,titanic$sibsp==5)
```

```
##     pclass survived    sex age sibsp parch
## 826      3        0   male   9     5     2
## 827      3        0   male   1     5     2
## 828      3        0   male  11     5     2
## 829      3        0 female  10     5     2
## 830      3        0 female  16     5     2
## 831      3        0   male  14     5     2
```

The ages of the largest family is mostly missing, except one 14.5 year old boy. Hence we might assume that the two observations with 9 children/parents are rather old, while the 9 observations with 8 siblings/spouses are young. However the (assumed) parents age is imputed as approximately 20, so is the age of their children.

```
subset(NewTitanic,NewTitanic$parch==9)
```

```
##      pclass survived    sex      age sibsp parch sib1 parch1
## 1180      3        0   male 20.07578     1     9    1      1
## 1181      3        0 female 20.70665     1     9    1      1
```

```
subset(NewTitanic,NewTitanic$sibsp==8)
```

```
##      pclass survived    sex      age sibsp parch sib1 parch1
## 1171      3        0   male 20.07578     8     2    1      1
## 1172      3        0   male 14.50000     8     2    1      1
## 1173      3        0 female 20.70665     8     2    1      1
## 1174      3        0 female 20.70665     8     2    1      1
## 1175      3        0 female 20.70665     8     2    1      1
## 1176      3        0 female 20.70665     8     2    1      1
## 1177      3        0   male 20.07578     8     2    1      1
## 1178      3        0   male 20.07578     8     2    1      1
## 1179      3        0   male 20.07578     8     2    1      1
```

We mention this to highlight how we should not trust that the imputed data is necesarily completely correct. But we do instead get values which look like the known values. Furthermore, as the data is missing we cannot prove nor disprove that our imputations are correct.

We proceed with the imputed data as is.

Modeling

Since our goal is to predict *survival*, which is binary, a logistic regression model is fitting.

We denote $\eta = X^T\beta$ with the log-odds link function, which is given by the expression $log(\frac{p}{1-p}) = \eta$. By simple rewriting, the expression for the mean p is given by the logistic function $p = \frac{e^\eta}{1+e^\eta}$. When we get the coefficients of our model, we will therefore need to make the transformation to get the correct value for p.

We previously argued that we should include all factors. Let us now consider any multiplicative effects on these factors.

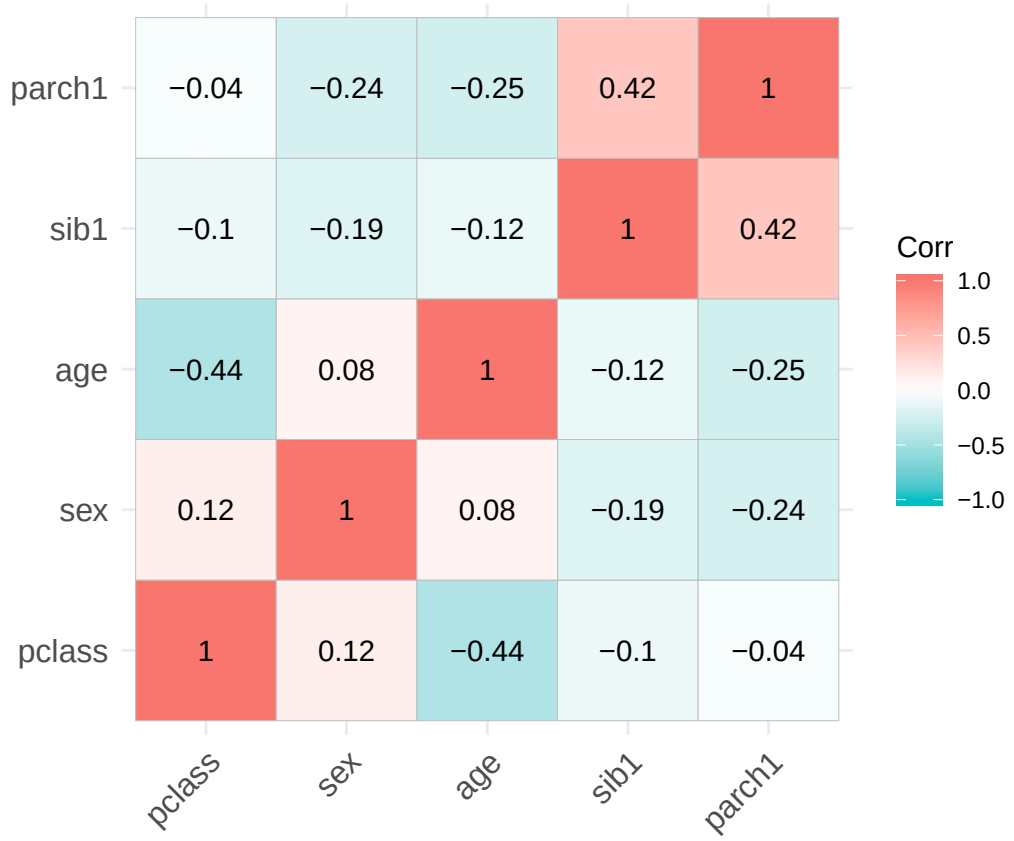
Multiplicative effect

We now consider the correlation plot.

Correlation Plot

We chose to use the Spearman correlation, rather than the Pearson correlation, as we have both continuous and categorical variables.

```
cp <- cor(data.matrix(na.omit(NewTdata2 %>%
                                dplyr::select(-c(survived)), method = "spearman")))
ggcorrplot(cp,method = c("square"), colors = c("#00BFC4", "white","#F8766D"), lab = TRUE)
```



Looking at the correlation plot, we may assume that there is a multiplicative effect for factors with large absolute value. Immediately we see that $parch1$ and $sib1$ are highly correlated likewise is $pclass$ and age. Furthermore, by the low values we see that

- $pclass$ appears to be uncorrelated to the factors: $sex, sib1, parch1$.
- sex appears to be uncorrelated to the factors: $pclass, age$.
- age appears to be uncorrelated to the factors: $sex, sib1$.
- $sib1$ appears to be uncorrelated to the factors: $pclass, age$.
- $parch1$ appears to be uncorrelated to the factor: $pclass$.

This gives us some idea as to which pairwise multiplicative effects to consider. But how about multiplicative effects beyond the pairwise?

One way to come up with an appropriate model is to start with the full model (the model with multiplicative effect among all factors) and do backwards selection. In our case that model would seem rather overwhelming, since we have so many factors and it would result in a model with a quintuple interaction term. Furthermore, by the observations above, there are already some multiplicative effects which we do not need to consider.

So far we only have the two interactions $age \times pclass$ and $sib1 \times parch1$. For the remaining interactions, we consider adding interactions which we did not previously exclude. These are: $sib1 \times parch1 \times sex$ and $age \times parch1$.

This gives us the following model which we do model reduction upon:

$$M_1 = sib1 \times parch1 \times sex + age \times parch1 + pclass \times age$$

Model reduction

We now proceed with the model reduction. We use the drop1 function and evalute the term to be deleted by the corresponding AIC value.

```
M1a=glm(survived~sex*sib1*parch1+age*pclass+age*parch1,family=binomial(link=logit),data=NewTdata2)
drop1(M1a)
```

```
## Single term deletions
##
## Model:
## survived ~ sex * sib1 * parch1 + age * pclass + age * parch1
##                 Df Deviance    AIC
## <none>               1193.3 1221.3
## age:pclass       2   1196.1 1220.1
## parch1:age       1   1193.3 1219.3
## sex:sib1:parch1  1   1194.5 1220.5
```

```
M1b=glm(survived~sex*sib1*parch1+age*pclass+age+parch1,family=binomial(link=logit),data=NewTdata2)
drop1(M1b)
```

```
## Single term deletions
##
## Model:
```

```
## survived ~ sex * sib1 * parch1 + age * pclass + age + parch1
##                 Df Deviance    AIC
## <none>               1193.3 1219.3
## age:pclass       2   1196.1 1218.1
## sex:sib1:parch1  1   1194.5 1218.5
```

```
M1c=glm(survived~sex*sib1*parch1+age+pclass+age+parch1,family=binomial(link=logit),data=NewTdata2)
drop1(M1c)
```

```
## Single term deletions
##
## Model:
## survived ~ sex * sib1 * parch1 + age + pclass + age + parch1
##                 Df Deviance    AIC
## <none>               1196.1 1218.1
## age              1   1236.3 1256.3
## pclass           2   1340.2 1358.2
## sex:sib1:parch1  1   1197.5 1217.5
```

```
M1c=glm(survived~sex*sib1*parch1+age+pclass+age+parch1-sex:sib1:parch1,family=binomial(link=logit),data=
drop1(M1c)
```

```
## Single term deletions
##
## Model:
## survived ~ sex * sib1 * parch1 + age + pclass + age + parch1 -
##     sex:sib1:parch1
##             Df Deviance    AIC
## <none>           1197.5 1217.5
## age          1   1236.8 1254.8
## pclass       2   1340.3 1356.3
## sex:sib1     1   1197.5 1215.5
## sex:parch1   1   1204.3 1222.3
## sib1:parch1  1   1200.0 1218.0
```

```
M1d=glm(survived~sex+sib1+sex*parch1+sib1*parch1+age+pclass+age+parch1-sex:sib1:parch1,family=binomial(
drop1(M1d,test="Chisq")
```

```
## Single term deletions
##
## Model:
## survived ~ sex + sib1 + sex * parch1 + sib1 * parch1 + age +
##     pclass + age + parch1 - sex:sib1:parch1
##             Df Deviance    AIC     LRT  Pr(>Chi)
## <none>           1197.5 1215.5
## age          1   1236.9 1252.9  39.334 3.572e-10 ***
## pclass       2   1340.3 1354.3 142.741 < 2.2e-16 ***
## sex:parch1   1   1206.2 1222.2   8.673  0.003229 **
## sib1:parch1  1   1200.2 1216.2   2.675  0.101944
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
M1e=glm(survived~sex*parch1+sib1+age+pclass,family=binomial(link=logit),data=NewTdata2)
drop1(M1e,test="Chisq")
```

```
## Single term deletions
##
## Model:
## survived ~ sex * parch1 + sib1 + age + pclass
##            Df Deviance    AIC     LRT  Pr(>Chi)
## <none>          1200.2 1216.2
## sib1        1   1204.7 1218.7   4.433  0.035252 *
## age         1   1239.0 1253.0  38.763 4.786e-10 ***
## pclass      2   1350.0 1362.0 149.764 < 2.2e-16 ***
## sex:parch1  1   1209.1 1223.1   8.871  0.002897 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The final model has slightly higher AIC than the second last (1218,5 vs 1217.6), but we prefer a simpler model and remove the multiplicative effect on $sex \times parch1$ by considering the p-value. Of course, if we had used a stricter parameter penalization we would make the same decision.

Thus we have

$$Model_{final} : S = sex \times parch1 + pclass + age + sib1$$

The expectation of our response variable, *survived*, given the covariates $\bar{X}$ is then given as

$$\begin{aligned} g(E[Y|X]) = \beta X^T \implies E[Y|X] &= g^{-1}(\beta X^T) \\ &= \frac{e^{\beta X^T}}{1 + e^{\beta X^T}} \end{aligned}$$

where $\bar{\beta}$ and $\bar{X}$ are given as

$$\bar{\beta} = \begin{pmatrix} \beta_0 \\ \beta_{sex} \\ \beta_{parch} \\ \beta_{age} \\ \beta_{sibsp} \\ \beta_{pclass} \\ \beta_{sex \times parch} \end{pmatrix} \qquad \bar{X} = \begin{pmatrix} X_0 \\ X_{sex} \\ X_{parch} \\ X_{age} \\ X_{sibsp} \\ X_{pclass} \\ X_{sex \times parch} \end{pmatrix}$$

And g is the logit link-function.

```r
M1e <- glm(survived~sex*parch1+sib1+age+pclass,family=binomial(link=logit),data=NewTdata2)
summary(M1e)$coef
```

```
##                    Estimate  Std. Error    z value     Pr(>|z|)
## (Intercept)      3.96845784 0.359028204  11.053332 2.112272e-28
## sexmale         -2.78482234 0.184798307 -15.069523 2.569615e-51
## parch11         -0.29187356 0.235494897  -1.239405 2.151955e-01
## sib11           -0.35482850 0.169816349  -2.089484 3.666420e-02
## age             -0.03942299 0.006572358  -5.998302 1.993917e-09
## pclass2         -1.33577917 0.220348479  -6.062121 1.343380e-09
## pclass3         -2.40774476 0.214090298 -11.246398 2.412433e-29
## sexmale:parch11  0.98711877 0.328257419   3.007148 2.637110e-03
```

Let us now consider the estimates for our model. The intercept corresponds to a 0 year old lone female passenger on first class. The other estimates (except $male \times parch1$) are negative and imply a decreased probability of survival. The intercept corresponds to the best case scenario in terms of survival.

Interestingly, we see that both $sib1$ and $parch1$ are negative. This may seem contradicting to the contingency tables we considered earlier, namely table 6 and table 7. To explain why this occours, we will again investigate some contingency tables.

```
ftable(NewTitanic$survived,NewTitanic$sex,NewTitanic$sib1)
```

```
##             0   1
##
## 0 female   65  62
##   male    517 165
## 1 female  197 142
##   male    112  49
```

```
ftable(NewTitanic$survived,NewTitanic$sex,NewTitanic$parch1)
```

```
##             0   1
##
## 0 female   75  52
##   male    591  91
## 1 female  218 121
##   male    118  43
```

For instance,

$$P(survived = 1|sex = F, sib1 = 0) = 197/(65 + 197) = 0.752$$

while

$$P(survived = 1|sex = F, sib1 = 1) = 142/(62 + 142) = 0.696.$$

Likewise for $parch1$ we see that

$$P(survived = 1|sex = F, parch1 = 0) = 218/(218 + 75) = 0.744$$

while

$$P(survived = 1|sex = F, parch1 = 1) = 121/(121 + 52) = 0.699.$$

So we may conclude that women infact had better surivival probability when being alone.

On the other hand

$$P(survived = 1|sex = M, parch1 = 0) = 118/(118 + 591) = 0.166$$

while

$$P(survived = 1|sex = M, parch1 = 1) = 43/(43 + 91) = 0.321.$$

Giving us the contrary conclussion, hence the positive estimate for $male : parch1$.

However for men we also see that

$$P(survived = 1|sex = M, sib1 = 0) = 112/(112 + 517) = 0.17$$

while

$$P(survived = 1|sex = M, sib1 = 1) = 49/(49 + 165) = 0.22.$$

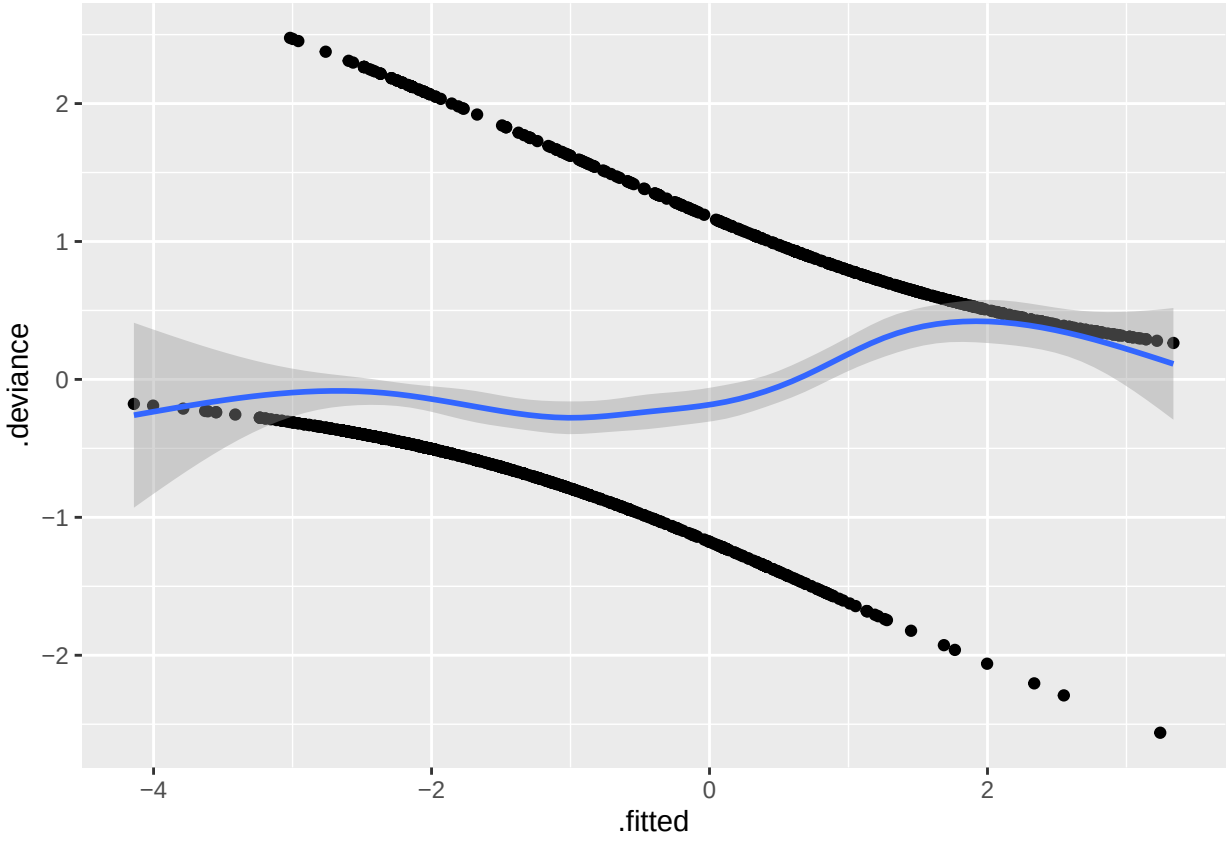
This is also positive, but we do not see quite as large difference as for $parch1$. Thus the interaction term for $sex \times sib1$ is not as significant and vanishes as we do model reduction.

Residuals

For the model diagnostics we look at the residuals. Below we have plotted the residuals against the fitted values.

```
N4b=transform(
  NewTdata2,
  .fitted=predict(M1e),
  .deviance=residuals(M1e),
  .pearson=residuals(M1e,type="pearson")
)
p1 <-  qplot(.fitted,.deviance, data=N4b)+
  geom_smooth(size=1)
p1
```

```
## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
```



The marked line almost shows a logistic regression line, which is what we want. However because we have chosen a somewhat simple model, the regression line has some unwanted waves.

Confidence intervals

We now consider confidence intervals for the estimates giving by our final model.

```
confint(M1e)
```

```
## Waiting for profiling to be done...
```

```
##                       2.5 %      97.5 %
## (Intercept)      3.27907963  4.68746891
## sexmale         -3.15363197 -2.42858293
## parch11         -0.75268188  0.17173638
## sib11           -0.69060188 -0.02436630
## age             -0.05250453 -0.02671563
## pclass2         -1.77266849 -0.90816376
## pclass3         -2.83491789 -1.99489956
## sexmale:parch11  0.34000058  1.62809369
```

```
confint.default(M1e)
```

```
##                       2.5 %      97.5 %
## (Intercept)      3.26477549  4.67214019
## sexmale         -3.14702036 -2.42262431
## parch11         -0.75343508  0.16968795
## sib11           -0.68766243 -0.02199458
## age             -0.05230457 -0.02654140
## pclass2         -1.76765425 -0.90390409
## pclass3         -2.82735403 -1.98813548
## sexmale:parch11  0.34374606  1.63049149
```

Here we see that the interval for parch1 includes 0, suggesting that the factor is insignificant. This is also the case for male:

```
confint(M1e)[3,]+confint(M1e)[8,]
```

```
## Waiting for profiling to be done...
## Waiting for profiling to be done...
```

```
##      2.5 %     97.5 %
## -0.4126813  1.7998301
```

```
confint.default(M1e)[3,]+confint.default(M1e)[8,]
```

```
##     2.5 %    97.5 %
## -0.409689  1.800179
```

From these observations we could be inclined to remove the factor parch1. But by our previous model reduction we chose to keep it (since the interaction between *parch*1 and *sex* was significant), and we will do so for our further analysis.

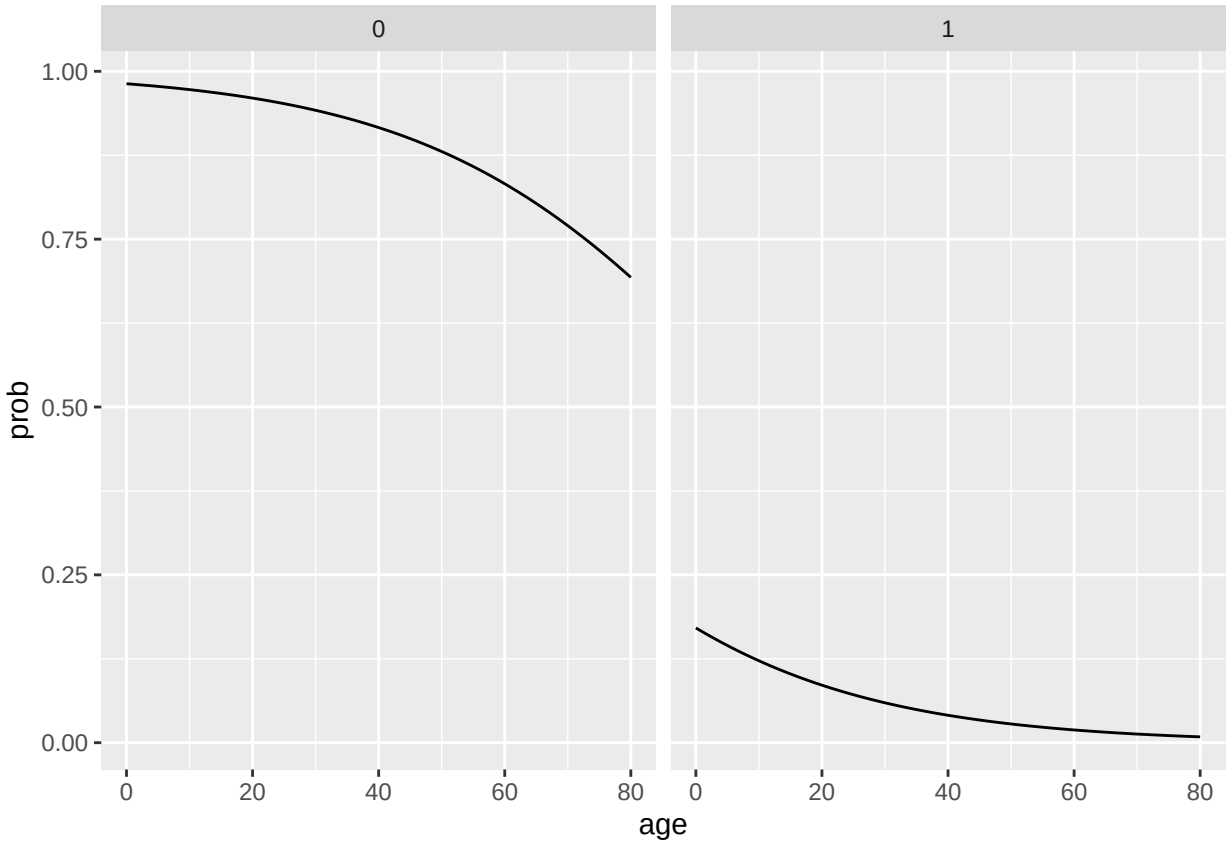
Predicting

We now show how our model does on predicting the survivability over different ages. We plot the probability of survival for a passenger with fixed factors and age on the x-axis. We illustrate this by considering the following plot:

```
G0 <- expit(predict(M1e, newdata = data.frame(pclass=as.factor(1),sex = "female",
                    sib1=as.factor(0),parch1=as.factor(0), age = seq(0,80,1))))
G1 <- expit(predict(M1e, newdata = data.frame(pclass=as.factor(3),sex = "male",
                    sib1=as.factor(1),parch1=as.factor(0), age = seq(0,80,1))))

G00 <- t(rbind(G0, seq(0,80,1), rep(0,81)))
G01 <- t(rbind(G1, seq(0,80,1), rep(1,81)))

G <- as.data.frame(rbind(G00,G01))
colnames(G) <- c("prob", "age", "Case")
ggplot(data = G, aes(age, prob)) + geom_line() + facet_wrap(~Case)
```



Where we picked the limit of the x-axis by the minimum (0.17) and maximum (80) age of the passengers.

On the left we have the best case scenario (lone female on first class) and on the right the worst case scenario (male with no parent(s)/child(ren) and with some sibling(s)/spouse(s) on third class). In both cases we see how survivability is reduced as age is increased. Most notably we see the large difference for passenger sex and class.

Cross validation

To meassure how well our model predicts the data we will use cross validation. We use the package boot and do 5 fold cross validation. The call cv.glm does the following: First we create a 5-fold partition of the data set, for each i'th partition we fit a model from the remaining 4 partitions, which we then use to predict the values of the i'th partition, ultimately giving us the i'th emperical error. We do this for each of the five partition and take the mean of the emperical errors.

```
cv.err.5 <- cv.glm(NewTdata2, M1e,K=5)$delta
cv.err.5
```

```
## [1] 0.1484866 0.1481548
```

Giving us an emperical error of 0.148. Now let us compare this to the full model.

$$M_{full} = sex \times age \times pclass \times parch \times sibsp,$$

with parch and sibsp not transformed into the binary factor.

```
Mfull=glm(survived~sex*age*pclass*sibsp*parch,family=binomial(link=logit), data=NewTitanic)
#NewTitanic is just NewTdata2 but with parch and sibsp included.
cv.err.5s <- cv.glm(NewTitanic, Mfull,K=5)$delta
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```
cv.err.5s
```

```
## [1] 0.1425899 0.1409174
```

Giving us an emperical error of 0.144, as well as a warning of perfect seperation.

Notice that this is only slightly lower than what we got with our much simpler model. Now considering the residual plot as well as the emperical error we conclude that our model performs rather well.

Conclusion

We have thus investigated the survival rate on Titanic. We have seen how the saying "Women and children first" holds true, seeing as the majority of deaths were men and age had a negative influence on survival. We have also seen how the first class passengers had a better survival rate than the other two classes. These observations were mostly as expected before analysing the data. However, with a transformation of the factors, we saw how there was a difference for single male vs female passengers on the boat. I.e., being without spouse/sibling/parent/child proved to yield better survival probability unless you were male with parent(s)/child(ren).