Coursework: Did Jack really have to die?
Week 3

By Ida Johanne Austad

As emphasized in the lecture the interpretations that follow for the different models made in subtask 1 and 3 assumes "all else held constant". That is, we assume no other variations than the variations in the variables we have chosen to investigate and include in our model.

1. **Fit a logistic regression model to predict the survival odds of a passenger using 'sex' as a single predictor. Write down the model to be estimated and show the summary of the model fit. Write down the estimated model: are the estimates significant?**

See attached code.
The model we are estimating has the following form:

$$\eta = \beta_0 + \beta_1 \times Sex$$

Summary of the model `Titatic_survival_sex` fit:

```
Call:
glm(formula = Survived ~ Sex, family = binomial, data = titan)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.6462  -0.6496  -0.6496   0.7725   1.8218

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   1.0566     0.1290   8.191 2.58e-16 ***
Sexmale      -2.5051     0.1672 -14.980  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1182.77  on 886  degrees of freedom
Residual deviance:  916.12  on 885  degrees of freedom
AIC: 920.12

Number of Fisher Scoring iterations: 4
```

As our response variable is a categorical variable with two options, "survived" or "did not survive" we thus want to use logistic regression and the logit function. The estimated model is thus

$$\log\left(\frac{p_i}{1-p_i}\right) = 1.06 - 2.51 \times Sex$$

We note that the output specifies that the variable will be set to 1 if the sex is male and 0 for female (as we read 'sexmale' in the summary output).

Regarding the estimates significance we note two things in the output. Firstly, the p-values given by the z-test are small (2.58e-16 and < 2e-16). In general, small p-values indicate that it is not likely that we would see the observed relationship between the predictor (Sex) and response (Survival) by chance. Thus the p-value indicates that there is a relationship between Sex and Survival. Secondly, this is reflected in the Significance scores (*** for both) which indicates that the estimates are highly significant (very small p-values).

2. **With the help of the fitted model in (1) answer the following questions (show your calculations, either by hand or with help of R):**
    - **What are the odds and probability that you survive given that you are a male?**
    - **What are the odds and probability of surviving if you are female?**

Calculating the odds and probability of survival given you are a male (Sex = 1):

$$Odds_{male} = \frac{p_i}{1-p_i} = e^{1.06-2.51\times 1} = e^{-1.45} = 0.2346$$

Solving out for p_i gives the following probability:

$$Probability_{male} = p_i = \frac{e^{-1.45}}{1+e^{-1.45}} = \frac{0.2346}{1.2346} = 0.1900$$

Calculating the odds and probability of survival given you are a female (Sex = 0):

$$Odds_{female} = \frac{p_i}{1-p_i} = e^{1.06-2.51\times 0} = e^{1.06} = 2.8864$$

Solving out for p_i gives the following probability:

$$Probability_{female} = p_i = \frac{e^{1.06}}{1+e^{1.06}} = \frac{2.8864}{3.8864} = 0.7427$$

3. **Now you will fit a more complex model. Remember the character 'Jack Dawson' from the movie Titanic? Using our dataset, we will fit a model to answer whether his death is realistic or not. We know Jack was male, 20 years old, and he travelled in the third class. Fit a logistic regression model using three predictors: sex, age and passenger class. Write down the model to be estimated. Summarize the fit, interpret each of the coefficient estimates and run some diagnostics. Then answer the following: Is the outcome that Jack does not survive realistic or not based on the fitted model? Show how you calculate the odds and probabilities required to answer this question.**

See attached code.
Decided to use the `factor()` function for the Pclass variable. This helps R understand that Pclass is not a continuous variable (although it is indicated by integers in our dataset), but rather a categorical variable. The model we are estimating thus has the following form:

$$\eta = \beta_0 + \beta_1 \times Sex + \ \beta_2 \times 2nd\ class + \ \beta_3 \times 3rd\ class + \ \beta_4 \times Age$$

Summary of the model `Titanic_survival_jack` fit:

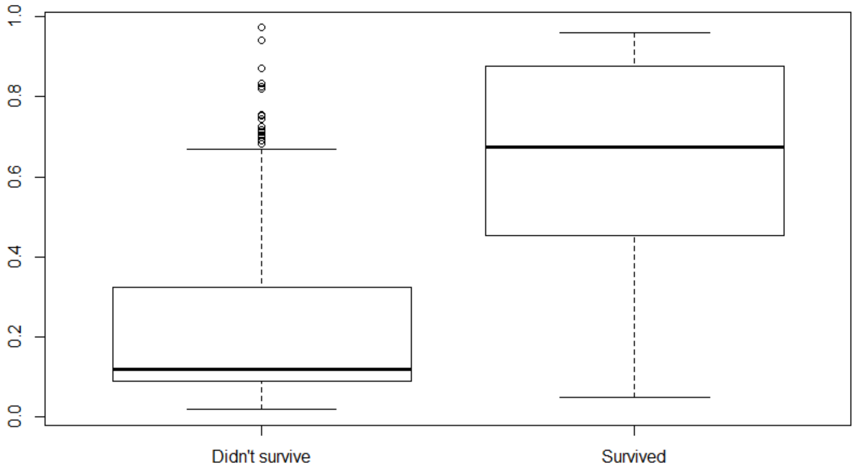
```
Call:
glm(formula = Survived ~ Sex + factor(Pclass) + Age, family = binomial,
    data = titan)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.6811  -0.6653  -0.4137   0.6367   2.4505

Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)      3.63492    0.37045   9.812  < 2e-16 ***
Sexmale         -2.58872    0.18701 -13.843  < 2e-16 ***
factor(Pclass)2 -1.19911    0.26158  -4.584 4.56e-06 ***
factor(Pclass)3 -2.45544    0.25322  -9.697  < 2e-16 ***
Age             -0.03427    0.00716  -4.787 1.69e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1182.77  on 886  degrees of freedom
Residual deviance:  801.59  on 882  degrees of freedom
AIC: 811.59

Number of Fisher Scoring iterations: 5
```

As for the model we made in subtask 1 we see that all our p-values here are very small, and that this is reflected in the significance scores (all ***). This indicates that the coefficient estimates are significant.

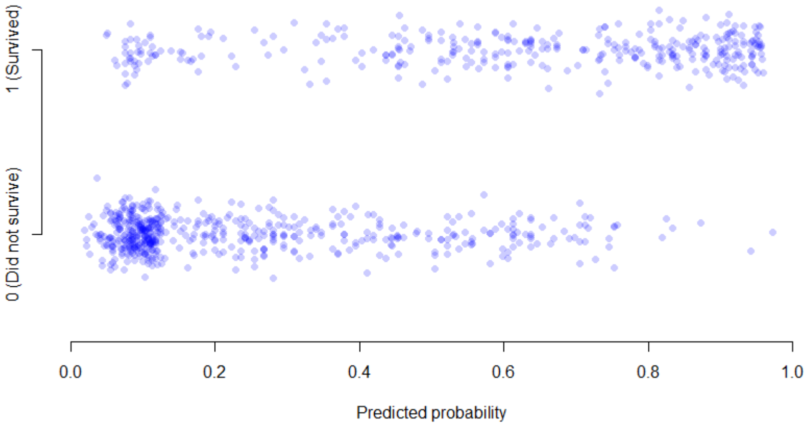
Furthermore, we can see that the estimates for the different coefficients vary in their value, and that all except the intercept are negative. From this we can interpret that sex (if you're a male instead of female), your passenger class (relative to passenger class 1) and age (relative to age 0) has a negative effect on your estimated chance of survival if you were aboard the Titanic. We see that being a male and living on passenger class 3 are the ones that have the “largest” negative estimated coefficients, which were both the case for Jack. Also, the estimated coefficient for age has to be multiplied by the actual age of the passenger, and will thus get a larger and larger negative effect on estimated chance of survival as a passenger’s age increases.

To investigate the model further, we run some diagnostics:

Boxplot:



Jitterplot:



Looking at the boxplots above, where we have separated our data points into those who did survive and those who did not, we can learn several things about our model. We can see that it gives the “middle 50%” (indicated by the area called the interquartile range, i.e. the box) of those who did not survive an estimated probability of survival to well below 0.4. Also the median for these are very close to the lower quartile. We also see that the lower whisker (for those who did not survive) stretches to even lower probabilities. Based on this we can interpret that our model is rather good at predicting the probability for those who did in fact not survive as low. For those who *did* survive our model gave the middle 50% of them a probability of survival of 0.5 or more, although the data for these are more spread out. Here as well, the whisker on the upper side of the interquartile range stretches to even high probabilities. Again, based on this we can interpret that our model is rather good at predicting that those who did in fact survive actually would as the model suggests probability of surviving.

For those who did not survive, we see that we have quite a few outliers (the dots above the upper whisker). That is, our model gave them a high probability of surviving – but they did in fact not. In general we also see that the whiskers for both the survivors and the ones who did not that they extend to almost the entire probability scale. This tells us that although our model indicated a probability of them surviving as high or low, the opposite was for some passengers the case.

The take-aways from the boxplot is reflected in the Jitter-plot as well. The fact that there seems to be a cluster in each of the rows (at the right for the survivors indicating a higher probability, and at the left for the non-survivors – indicating a lower probability) can tell us that our model is rather good predicting the probability of survival. Still, we see that the dots are spread across almost the entire line – which tells us the same as for the whiskers and outliers in the boxplot; some of those who are predicted to not have survived, did actually survive, and vice versa. One explanation for this could be that the event were a ship goes down is complicated, were many variables affect whether a passenger will survive or not – not just the three predictors we have taken into account in our model. Thus, our model will not be able to predict all cases correctly.

Lastly, we want to calculate the odds and probability of Jack surviving given the model above. We use the following relationship:

$$\eta = \log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 \times Sex + \beta_2 \times 2nd\ class + \beta_3 \times 3rd\ class + \beta_4 \times Age$$

We can thus fill in the values of the estimated coefficients from the model above to find η as follows:

$$\eta = 3.635 - 2.589 \times Sex - 1.199 \times 2nd\ class - 2.455 \times 3rd\ class - 0.034 \times Age$$

$$\eta = 3.635 - 2.589 \times 1 - 1.199 \times 0 - 2.455 \times 1 - 0.034 \times 20$$

$$\eta = -2.089$$

$$Odds\ of\ Jack\ surviving = \frac{p_i}{1-p_i} = e^{-2.089} = 0.1238$$

$$Probability\ of\ Jack\ surviving = p_i = \frac{e^{-2.089}}{1+e^{-2.089}} = 0.11$$

All in all, our model suggests that the fact that Jack died in the movie is realistic – as his probability of survival is estimated to be just 11%.

Attached code:

```
#importing data and putting it in the varialble titan
titan <- read.csv("P:/ML_and_stats/Week3/titanic.csv", header = TRUE)
attach(titan)   #to be able to get objects in the dataset can be accessed by name
dim(titan)      #number of rows and coloums
head(titan)     #to look at the first rows in the data

#1 fitting a logistic regression model using sex as predictor
Titanic_survival_sex <- glm(Survived ~  Sex, data = titan, family = binomial)
summary(Titanic_survival_sex)

#3 Fit logistic model using predictors sex, age and passenger class
#Using the factor function as Pclass is a categorical variable and not continuous
TitanicSurvivalJack <- glm(Survived ~  Sex + factor(Pclass) + Age, data = titan, fa
mily = binomial)
summary(TitanicSurvivalJack)

#diagnostics
#jitterplot
set.seed(1)
jitter = rnorm(nrow(titan), sd = 0.08)
plot(TitanicSurvivalJack$fitted.values, Survived + jitter, xlim = 0:1, ylim = c(-0.
5,1.5), axes = FALSE, xlab = "Predicted probability", ylab = "", col =adjustcolor("
blue",0.2), pch = 16)
axis(1)
axis(2, at = c(0,1), labels = c("0 (Did not survive)" , "1 (Survived)"))

#boxplot
plot(factor(Survived,labels = c ("Didn't survive", "Survived")), TitanicSurvivalJac
k$fitted.values)
```