Coursework 3-Did Jack really have to die?

Shizhi Chen-10307389

For this coursework you will use the Titanic data set available to download from Blackboard. The data is over 887 of the real Titanic passengers including the following variables: whether they survived (0/1 variable), their age, passenger-class, their sex and the fare they paid for their ticket.

1.Fit a logistic regression model to predict the survival odds of a passenger using 'sex' as a single predictor.

```
#R code:
#read the data set, show the observations and variables
#check the beginning and the end of the data set
>titan<-read.csv("titanic.csv", header=TRUE)
>attach(titan)
>dim(titan)
>head(titan)
>tail(titan)
#Fit the logistic regression model between Survived and Sex
>model1=glm (Survived~Sex, data=titan, family=binomial)
>summary(model1)
```

Output:

887 8

Survived	Pclass	Name	Sex	Age	Siblings.Spouses.Aboard	Parents.Children.Aboard	Fare
0	3	Mr. Owen Harris Braund	male	22	1	0	7.2500
1	1	Mrs. John Bradley (Florence Briggs Thayer) Cumings	female	38	1	0	71.2833
1	3	Miss. Laina Heikkinen	female	26	0	0	7.9250
1	1	Mrs. Jacques Heath (Lily May Peel) Futrelle	female	35	1	0	53.1000
0	3	Mr. William Henry Allen	male	35	0	0	8.0500
0	3	Mr. James Moran	male	27	0	0	8.4583

	Survived	Pclass	Name	Sex	Age	Siblings.Spouses.Aboard	Parents.Children.Aboard	Fare
882	0	3	Mrs. William (Margaret Norton) Rice	female	39	0	5	29.125
883	0	2	Rev. Juozas Montvila	male	27	0	0	13.000
884	1	1	Miss. Margaret Edith Graham	female	19	0	0	30.000
885	0	3	Miss. Catherine Helen Johnston	female	7	1	2	23.450
886	1	1	Mr. Karl Howell Behr	male	26	0	0	30.000
887	0	3	Mr. Patrick Dooley	male	32	0	0	7.750

```
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
```

The output shows that there are 887 observations and 8 variables in the data set. The estimated model is thus:

$$\log\left(\frac{p_i}{1-p_i}\right) = 1.0566 - 2.5051 \times \text{Sexmale}$$

where, if the person is a male then Sexmale takes value 1 if this person is a female then Sexmale=0.

In this model, the p-value of the estimator of independent variable Sexmale is less than 2×10^{-16}. Therefore, the Sexmale is considered as a highly significant variable.

2.With the help of the fitted model in (1) answer the following questions:

• 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?

$$\text{odds} = \frac{p_i}{1-p_i} = \exp(1.0566 - 2.5051 \times \text{Sexmale})$$

$$\text{probability} = p_i = \frac{\exp(1.0566 - 2.5051 \times \text{Sexmale})}{1 + \exp(1.0566 - 2.5051 \times \text{Sexmale})}$$

```
oddsmale=exp(1.0566-2.5051*1)
oddsmale
pimale=exp(1.0566-2.5051*1)/(1+exp(1.0566-2.5051*1))
pimale
oddsfemale=exp(1.0566-2.5051*0)
oddsfemale
pifemale=exp(1.0566-2.5051*0)/(1+exp(1.0566-2.5051*0))
pifemale
```

```
0.234922407549508
0.190232524823702
2.87657399223717
0.742040264934322
```

Male: Sexmale=1, odds=0.235, the probability of survive=0.190,

Female: Sexmale=0, odds=2.877, the probability of survive=0.742.

3. Now you will fit a more complex model.

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 realistically or not based on the fitted model?

• Show how you calculate the odds and probabilities required to answer this question.

> model2=glm (Survived~Sex + Pclass + Age, family=binomial, data=titan)

> summary (model2)

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

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.6858  -0.6588  -0.4102   0.6386   2.4493

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  4.878511   0.463474  10.526  < 2e-16 ***
Sexmale     -2.589163   0.186933 -13.851  < 2e-16 ***
Pclass      -1.230538   0.124957  -9.848  < 2e-16 ***
Age         -0.034361   0.007134  -4.816 1.46e-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:  801.61  on 883  degrees of freedom
AIC: 809.61

Number of Fisher Scoring iterations: 5
```

Model:

$\log\left(\frac{p_i}{1-p_i}\right) = 4.878511 - 2.589163 \times \text{Sexmale} - 1.230538 \times \text{Pclass} - 0.034361 \times \text{Age}$

Where, Sexmale=1 or 0, Pcalss=1 or 2 or 3, Age∈[0,100].

Interpret each of the coefficient estimates:

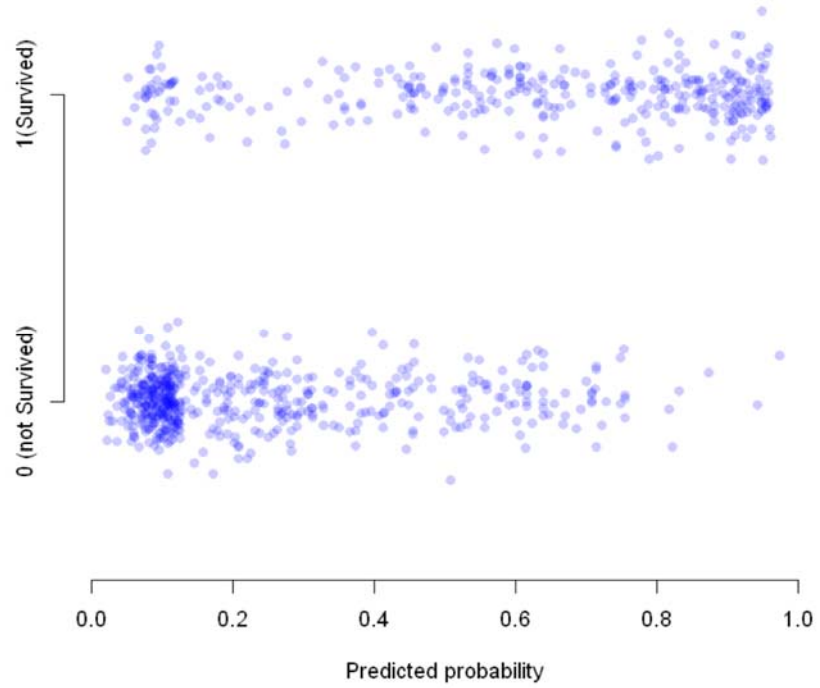
The P value of three predictors (sex, age and passenger class) all very small, so these predictors are considered have the significant correlation with the person survival probability. From the perspective of the coefficients, the three predictions are negatively correlated with survival probability. To be specific, the passenger's sex and class have the strong negative correlation with survival probability. Male passengers have a much lower survival probability than women. Passengers in first class have a higher probability of survival than second class and second class passengers have a higher probability of survival than third class. Age also has a slight impact on survival probability, and older passengers have a slightly higher survival probability than younger passengers.

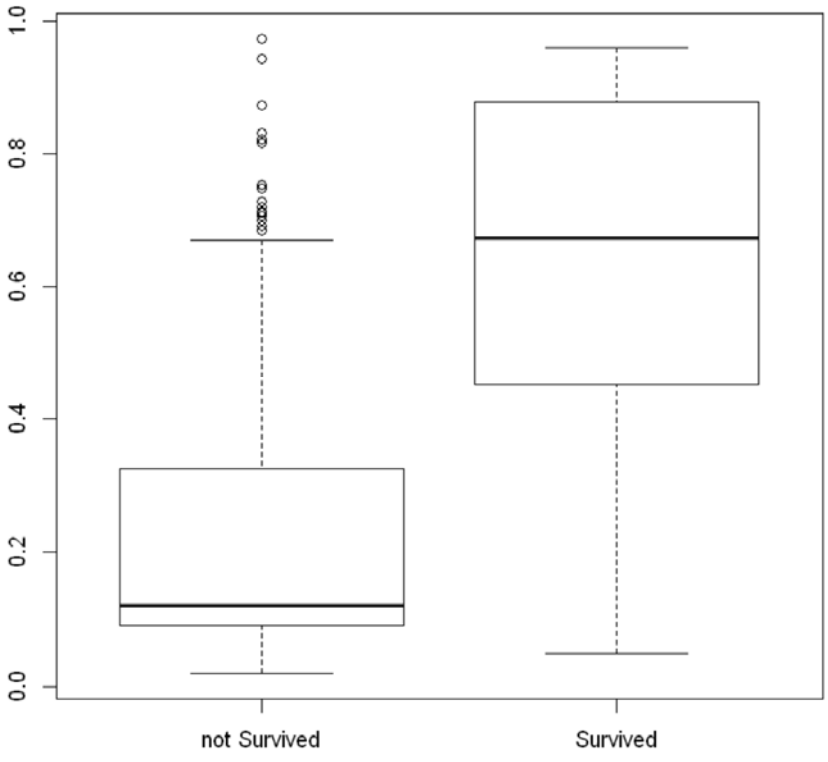
#Diagnostics code:

```
>jitter = rnorm (nrow(titan), sd = 0.08)
>plot (model2$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 (not Survived)", "1(Survived)"))

> plot (factor (Survived, labels = c("not Survived", "Survived")), model2$fitted.values)





Interpret about the plots:

From the plot and boxplot, we can see that most of the people predicted by the model with a survival probability below 0.2 are actually not surviving and people who have a survival probability greater than 0.6 in the predictive model are actually survived in the end. This is means that excluding some special outliers, the model is working well.

#Calculate Jack survived odds and probability:

#Jack, Male, travelled in 3 class, 20 years old, input these value

 # According to the Model 2, the Jack's survival odds and probabilities can be calculated by following:

```
#Model2:log(pi/(1-pi))=4.878511-2.589163× Sexmale-1.230538×Pclass-0.034361× Age
odds_Jack_survived=exp(4.878511-2.589163*1-1.230538*3-0.034361*20)
piJack_survived=exp(4.878511-2.589163*1-1.230538*3-0.034361*20)/
(1+exp(4.878511-2.589163*1-1.230538*3-0.034361*20))
odds_Jack_survived
piJack_survived
```

```
0.123750727346888
```

```
0.110122934148445
```

Based on the model, Jack's survival probability is only around 11.01%. Therefore, Jack does not survive in the realistic is considered to be kind of certain.