Team_Essay_3

Sahil, Ariel, Ilya, Lekha, Anthony, Trang

3/18/2021

Introduction

Trang.

Formula and Basics

Trang.

Loading required R packages

```
library("readxl")
library("tidyverse")
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.0 --

## v ggplot2 3.3.3     v purrr   0.3.4
## v tibble  3.0.6     v dplyr   1.0.4
## v tidyr   1.1.2     v stringr 1.4.0
## v readr   1.4.0     v forcats 0.5.1

## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library("flexplot")
```

```
##
## Attaching package: 'flexplot'

## The following object is masked from 'package:ggplot2':
##
##     flip_data
```

```
library("caret")
```

```
## Loading required package: lattice
```

```
##
## Attaching package: 'caret'

## The following object is masked from 'package:purrr':
##
##     lift
```

Data description

Examples of data and problem:

- Name: represents name of passenger
- PClass: represents the passenger class (1st, 2nd, 3rd)
- Age: represents age of the passenger in years (N/A if unknown)
- Sex: represents the sex of the passenger in terms of Male/Female
- Survived: represents survival status of passenger in terms of survived(1)/didn't survive(0)

```
titanic <- read_excel("Titanic.xlsx")
titanic <- titanic[1:1313,]
index <- titanic$Age != "NA"
titanic <- titanic[index, ]
titanic$PClass = as.factor(titanic$PClass)
titanic$Sex = as.factor(titanic$Sex)
titanic$Survived = as.factor(titanic$Survived)
titanic$Age = as.factor(titanic$Age)
str(titanic)
```

```
## tibble [756 x 5] (S3: tbl_df/tbl/data.frame)
##  $ Name    : chr [1:756] "Allen, Miss Elisabeth Walton" "Allison, Miss Helen Loraine" "Allison, Mr Hu
##  $ PClass  : Factor w/ 3 levels "1st","2nd","3rd": 1 1 1 1 1 1 1 1 1 1 ...
##  $ Age     : Factor w/ 75 levels "0.17","0.33",..: 28 18 30 24 5 48 66 39 60 73 ...
##  $ Sex     : Factor w/ 2 levels "female","male": 1 1 2 1 2 2 1 2 1 2 ...
##  $ Survived: Factor w/ 2 levels "0","1": 2 1 1 1 2 2 2 1 2 1 ...
```

Visualization

Anthony needs to do the visualization.

Analysis

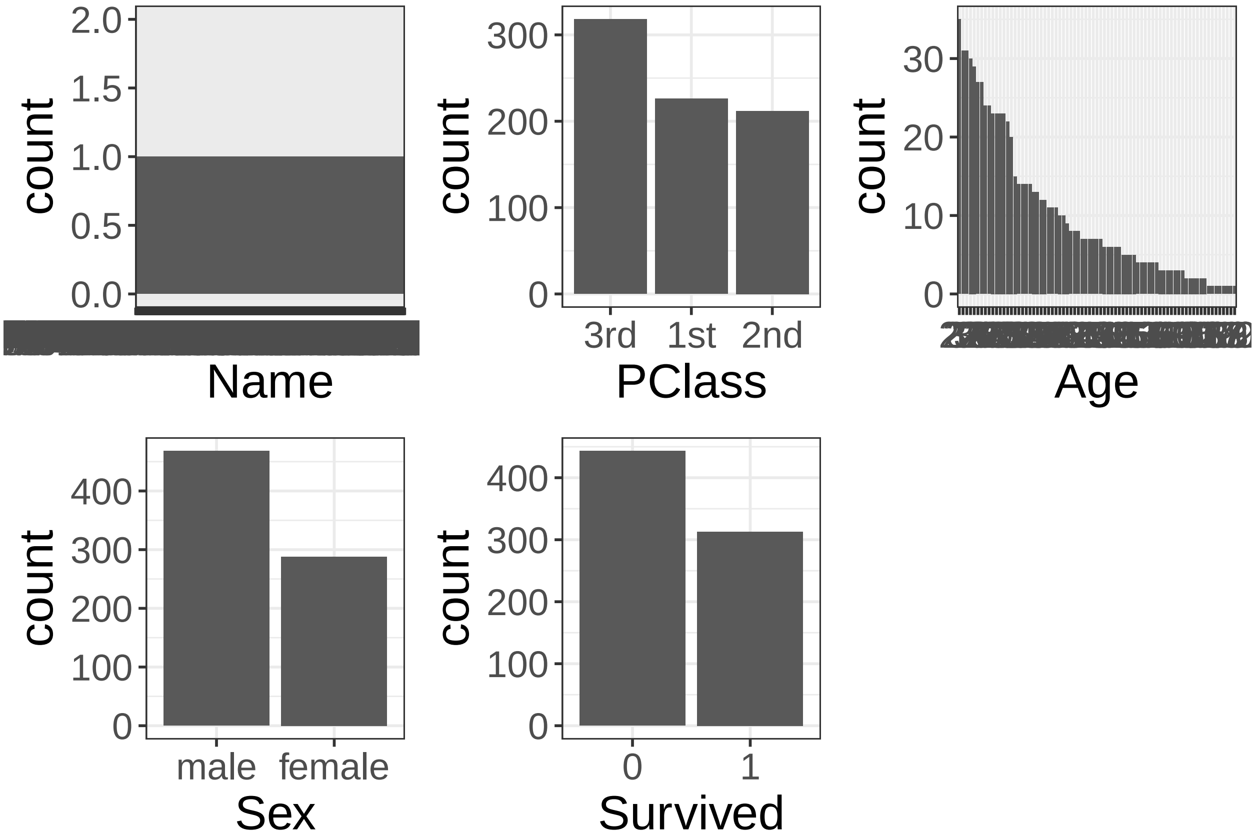
This is the bi-variate dependency result.

```
a = flexplot(Name~1, data = titanic)
b = flexplot(PClass~1, data = titanic)
c = flexplot(Age~1, data = titanic)
d = flexplot(Sex~1, data = titanic)
e = flexplot(Survived~1, data = titanic)
require(cowplot)
```

```
## Loading required package: cowplot
```

```
plot_grid(a, b, c, d, e, nrow=2)
```

```
## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, : font
## width unknown for character 0x9

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, : font
## width unknown for character 0x9

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, : font
## width unknown for character 0x9

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, : font
## width unknown for character 0x9

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, : font
## width unknown for character 0x9

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, : font
## width unknown for character 0x9

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, : font
## width unknown for character 0x9

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, : font
## width unknown for character 0x9

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, : font
## width unknown for character 0x9

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, : font
## width unknown for character 0x9

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, : font
## width unknown for character 0x9

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, : font
## width unknown for character 0x9

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, : font
## width unknown for character 0x9

## Warning in grid.Call.graphics(C_text, as.graphicsAnnot(x$label), x$x, x$y, :
## font width unknown for character 0x9
```



This neat visualization shows us the number of people that are male/female, the number of people that survived, the number of people aged between ranges and the number of people belonging to 1st, 2nd and 3rd class respectively.

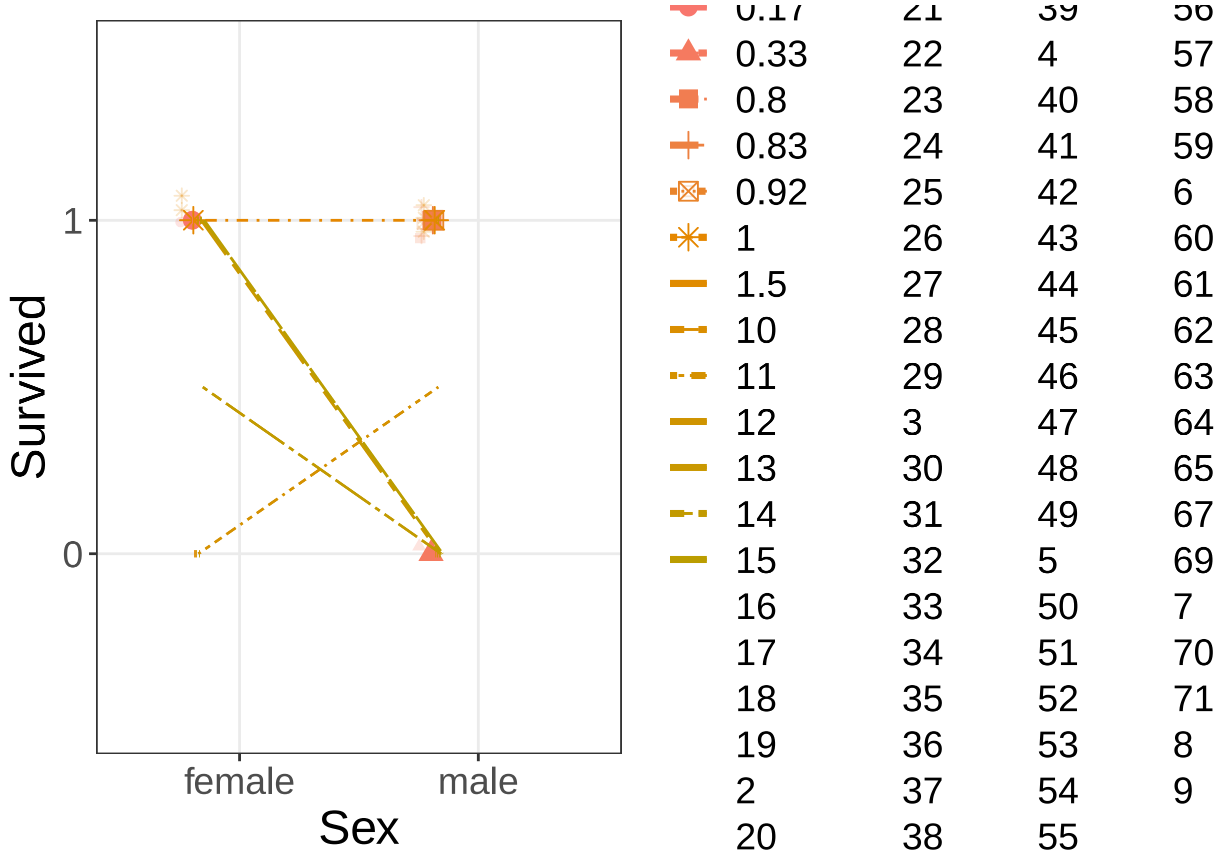
This is the multivariate dependency result.

```
f = flexplot(Survived~Sex + Age + PClass, data = titanic, method="Binomial", se=F, jitter=c(0, 0.2))
plot_grid(f)
```

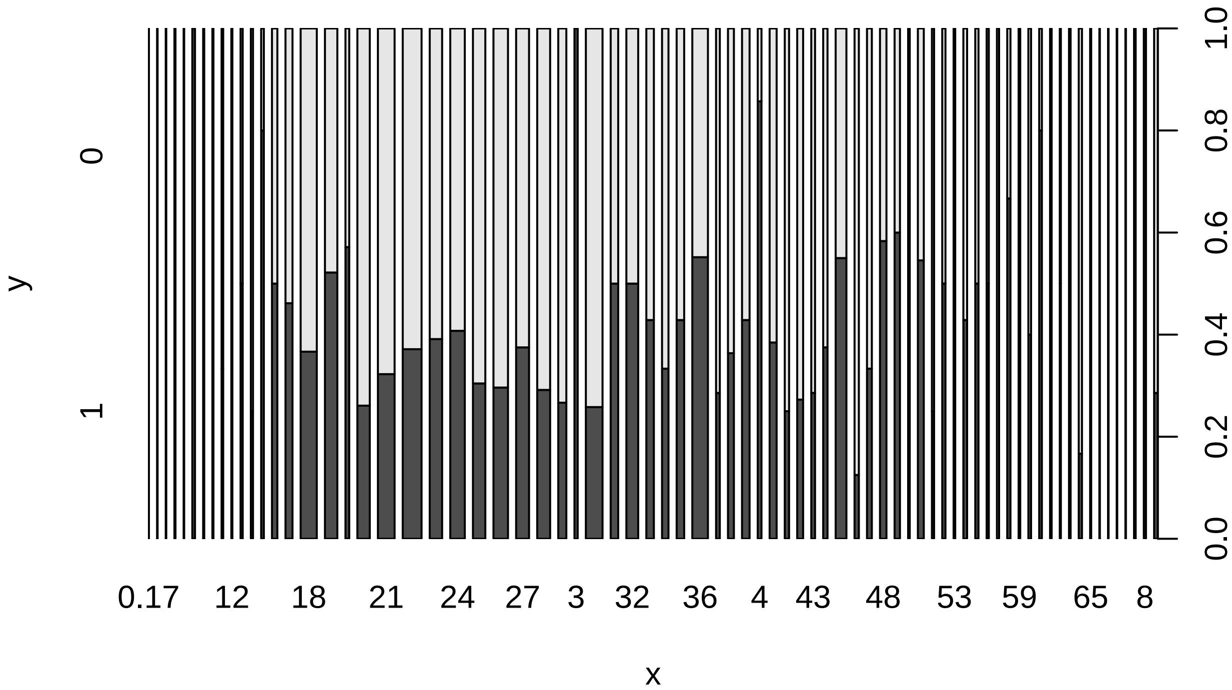
```
## Warning: The shape palette can deal with a maximum of 6 discrete values because
## more than 6 becomes difficult to discriminate; you have 75. Consider
## specifying shapes manually if you must have them.
```

```
## Warning: Removed 745 rows containing missing values (geom_point).
```

```
## Warning: Removed 192 rows containing missing values (geom_point).
```

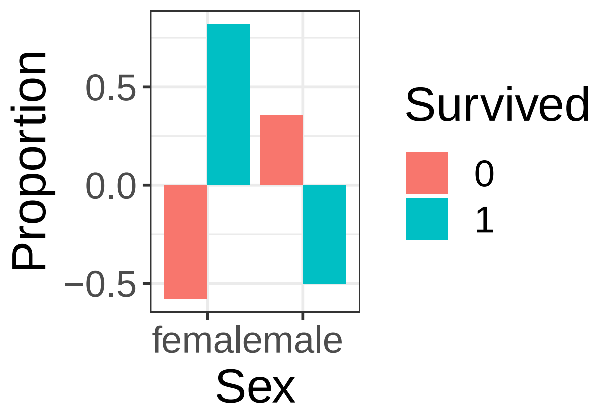


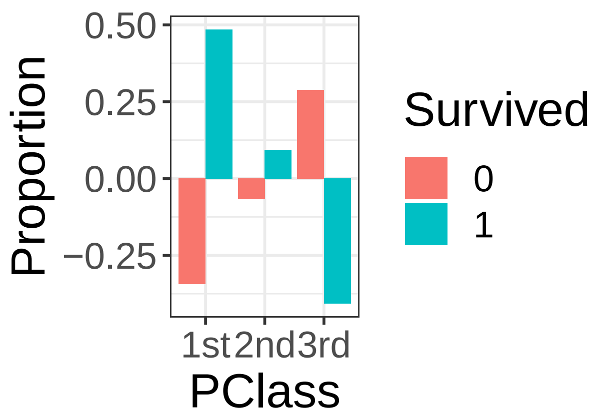

```
f = plot(x = titanic$Age, y = titanic$Survived)
```




```r
g = flexplot(Survived~Sex, data = titanic, method="Binomial", se=F, jitter=c(0, 0.2))
h = flexplot(Survived~PClass, data = titanic, method="Binomial", se=F, jitter=c(0, 0.2))
plot_grid(f, g, h)
```

```
## Warning in as_grob.default(plot): Cannot convert object of class table into a
## grob.
```





Computation and Model assessment

```
model_2 <- glm(Survived~ PClass + Sex, data=titanic, family=binomial)
summary(model_2)
```

```
##
## Call:
## glm(formula = Survived ~ PClass + Sex, family = binomial, data = titanic)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.1315  -0.7091  -0.4332   0.6732   2.1967
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   2.1628     0.2190   9.873  < 2e-16 ***
## PClass2nd    -0.7937     0.2334  -3.401 0.000672 ***
## PClass3rd    -1.8604     0.2319  -8.022 1.04e-15 ***
## Sexmale      -2.6213     0.1966 -13.335  < 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: 1025.57  on 755  degrees of freedom
## Residual deviance:  723.59  on 752  degrees of freedom
## AIC: 731.59
##
## Number of Fisher Scoring iterations: 4
```

```
anova(model_2)
```

```
## Analysis of Deviance Table
##
## Model: binomial, link: logit
##
## Response: Survived
##
## Terms added sequentially (first to last)
##
##
##        Df Deviance Resid. Df Resid. Dev
## NULL                     755    1025.57
## PClass  2   78.026       753     947.55
## Sex     1  223.951       752     723.59
```

```
model_2_probs = predict(model_2, type="response")
model_2_predict <- rep("0", nrow(titanic))
model_2_predict[model_2_probs > .6] <- "1"
table(Predicted = model_2_predict, Reference = titanic$Survived)
```

```
##          Reference
## Predicted   0   1
##         0 428 142
##         1  15 171
```

First things first we construct our binomial model which relates whether the passenger survived to the passenger's boarding class and sex. When we look at the summary and plots of this model, we see that not only are the predictors significant, but our residuals are quite close to our actual values. However, the AIC value is pretty high which can indicate that our model is not actually a good fit. Nonetheless, what matters is actually testing our model's predictive capabilities. We do this by calculating the probability of that someone on the titanic survived. Then, for any probability that is greater than 60% that the passenger survived, we predict that the passenger actually did survive. Finally, we create a confusion matrix to see how many survived and deceased passengers we correctly predicted along with the passengers we incorrectly predicted. Using this, we can calculate the overall accuracy of our predictive model that, when computed, gives us an accuracy of around 79%.

Interpretation and Model evaluation

We interpret the results given by the model as:

- Sex and PClass are the 2 significant predictors, hence we make the model which includes only them.
- Name and Age are the 2 insignificant predictors, hence we thought it was safe to remove them during model creation.

- People who survived mostly belonged to the 1st and 3rd Class and were females.
- Our model is not perfect since AIC can be much smaller than the current value. (AIC: 731.59)
- The residual deviance also could have been lower, but this is due to more number of outliers. (Residual deviance: 723.59 on 752 degrees of freedom)

Coefficients significance

```r
summary(model_2)
```

```
##
## Call:
## glm(formula = Survived ~ PClass + Sex, family = binomial, data = titanic)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.1315  -0.7091  -0.4332   0.6732   2.1967
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   2.1628     0.2190   9.873  < 2e-16 ***
## PClass2nd    -0.7937     0.2334  -3.401 0.000672 ***
## PClass3rd    -1.8604     0.2319  -8.022 1.04e-15 ***
## Sexmale      -2.6213     0.1966 -13.335  < 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: 1025.57  on 755  degrees of freedom
## Residual deviance:  723.59  on 752  degrees of freedom
## AIC: 731.59
##
## Number of Fisher Scoring iterations: 4
```

The coefficients and their respective P-values for our model:

- PClass2nd: -0.7936711 with a P-value of 0.00067
- PClass3rd: -1.8603793 with a P-value of 1.04e-15
- Sexmale: -2.6212723 with a P-value of 2e-16

We can see that the P-values are very close if not equal to 0 which shows that all of the coefficients are significant.

Prediction and Model accuracy

```r
table(Predicted = model_2_predict, Reference = titanic$Survived)
```

```
##          Reference
## Predicted   0   1
##         0 428 142
##         1  15 171
```

From the computed confusion matrix, we can see that our model predicted 599 out of the 756 total data points correctly. Essentially, we can say that our model was able to predict whether a passenger aboard the titanic was able to survive with an accuracy of 79%.

Conclusion and Summary

First things first, we modified our data to remove any data points not containing a specified age for the passenger. From this, we then constructed our binomial model to relate each passenger's survival verdict to their boarding class and sex. We graphically depicted the distribution of values for each of the variables, from which we deduced the observations that there were more males than females, there were less of survivors overall, and most of the passengers were in their youth (below 30). We then observed visually the chances of survival for each age, followed by the proportions of those who survived based on their class and sex separately. We deduced that most of those who survived were female passengers of first class or third class. We then constructed a generalized linear model using the logit function given that the response variable is binomial, and from finding the sex and passenger class to be the two significant predictor variables, we built our model such that it contains only these variables. When assessing the model and the computation, including observing the details of the model summary, we found the following values as the coefficients of our model: B0: 2.1628 B1: -0.7937 B2: -1.8604 B3: -2.6213 We plotted the residuals of the model only containing the two predictor variables sex and passenger class, and found that the residuals are quite close to our actual values. The residuals follow constant variance; this suggests that our model is a good fit. This is contradicted by looking at our AIC value, which was actually higher than expected. Nonetheless, what matters is actually testing our model's predictive capabilities. We do this by passing to the model all of the passengers with the corresponding data as input, calculating the probability of survival for each passenger, and building the model's confusion matrix. This matrix tells us how many passengers' survival verdict the model predicted correctly and incorrectly, for which we set the minimum cutoff probability value for the passenger to be considered a survivor as 60%. According to this matrix, we found that our model was able to determine the survival status of most of the passengers accurately, as it correctly predicted 428 and 171 to not have survived and to have survived respectively, while only 15 and 142 were incorrectly predicted in the same way. Therefore, we found the overall accuracy of our predictive model to be 79%, and hence a satisfactory model to be used for predicting a passenger's survival chances on the Titanic.

Reference

https://www.guru99.com/r-generalized-linear-model.html http://www.john-ros.com/Rcourse/lm.html