insurance2.R

beven

2025-09-06

###################################################  
# Logistic Regression - Titanic Dataset  
###################################################  
  
# Load Libraries  
library(readr)  
library(naniar)

## Warning: package 'naniar' was built under R version 4.4.3

library(mice)

## Warning: package 'mice' was built under R version 4.4.2

##   
## Attaching package: 'mice'

## The following object is masked from 'package:stats':  
##   
## filter

## The following objects are masked from 'package:base':  
##   
## cbind, rbind

library(car)

## Warning: package 'car' was built under R version 4.4.2

## Loading required package: carData

library(tidymodels)

## Warning: package 'tidymodels' was built under R version 4.4.2

## ── Attaching packages ────────────────────────────────────── tidymodels 1.2.0 ──

## ✔ broom 1.0.7 ✔ recipes 1.1.1  
## ✔ dials 1.4.0 ✔ rsample 1.2.1  
## ✔ dplyr 1.1.4 ✔ tibble 3.2.1  
## ✔ ggplot2 3.5.1 ✔ tidyr 1.3.1  
## ✔ infer 1.0.7 ✔ tune 1.2.1  
## ✔ modeldata 1.4.0 ✔ workflows 1.1.4  
## ✔ parsnip 1.3.0 ✔ workflowsets 1.1.0  
## ✔ purrr 1.0.4 ✔ yardstick 1.3.2

## Warning: package 'ggplot2' was built under R version 4.4.3

## Warning: package 'infer' was built under R version 4.4.2

## Warning: package 'modeldata' was built under R version 4.4.2

## Warning: package 'purrr' was built under R version 4.4.2

## Warning: package 'rsample' was built under R version 4.4.2

## Warning: package 'tune' was built under R version 4.4.2

## Warning: package 'workflows' was built under R version 4.4.2

## Warning: package 'workflowsets' was built under R version 4.4.2

## Warning: package 'yardstick' was built under R version 4.4.2

## ── Conflicts ───────────────────────────────────────── tidymodels\_conflicts() ──  
## ✖ purrr::discard() masks scales::discard()  
## ✖ dplyr::filter() masks mice::filter(), stats::filter()  
## ✖ dplyr::lag() masks stats::lag()  
## ✖ dplyr::recode() masks car::recode()  
## ✖ purrr::some() masks car::some()  
## ✖ yardstick::spec() masks readr::spec()  
## ✖ recipes::step() masks stats::step()  
## • Learn how to get started at https://www.tidymodels.org/start/

library(pROC)

## Warning: package 'pROC' was built under R version 4.4.3

## Type 'citation("pROC")' for a citation.

##   
## Attaching package: 'pROC'

## The following objects are masked from 'package:stats':  
##   
## cov, smooth, var

library(ResourceSelection)

## Warning: package 'ResourceSelection' was built under R version 4.4.3

## ResourceSelection 0.3-6 2023-06-27

library(pscl)

## Warning: package 'pscl' was built under R version 4.4.3

## Classes and Methods for R originally developed in the  
## Political Science Computational Laboratory  
## Department of Political Science  
## Stanford University (2002-2015),  
## by and under the direction of Simon Jackman.  
## hurdle and zeroinfl functions by Achim Zeileis.

library(nortest)  
library(corrplot)

## Warning: package 'corrplot' was built under R version 4.4.2

## corrplot 0.95 loaded

###################################################  
# Load & Explore Data  
###################################################  
Titanic <- read\_csv("C:/Users/beven/Downloads/teams/Titanic.csv")

## Rows: 891 Columns: 11

## ── Column specification ────────────────────────────────────────────────────────  
## Delimiter: ","  
## chr (4): Sex, Ticket, Cabin, Embarked  
## dbl (7): PassengerId, Survived, Pclass, Age, SibSp, Parch, Fare  
##   
## ℹ Use `spec()` to retrieve the full column specification for this data.  
## ℹ Specify the column types or set `show\_col\_types = FALSE` to quiet this message.

# Initial Exploration  
head(Titanic)

## # A tibble: 6 × 11  
## PassengerId Survived Pclass Sex Age SibSp Parch Ticket Fare Cabin  
## <dbl> <dbl> <dbl> <chr> <dbl> <dbl> <dbl> <chr> <dbl> <chr>  
## 1 1 0 3 male 22 1 0 A/5 21171 7.25 <NA>   
## 2 2 1 1 female 38 1 0 PC 17599 71.3 C85   
## 3 3 1 3 female 26 0 0 STON/O2. 310… 7.92 <NA>   
## 4 4 1 1 female 35 1 0 113803 53.1 C123   
## 5 5 0 3 male 35 0 0 373450 8.05 <NA>   
## 6 6 0 3 male NA 0 0 330877 8.46 <NA>   
## # ℹ 1 more variable: Embarked <chr>

str(Titanic)

## spc\_tbl\_ [891 × 11] (S3: spec\_tbl\_df/tbl\_df/tbl/data.frame)  
## $ PassengerId: num [1:891] 1 2 3 4 5 6 7 8 9 10 ...  
## $ Survived : num [1:891] 0 1 1 1 0 0 0 0 1 1 ...  
## $ Pclass : num [1:891] 3 1 3 1 3 3 1 3 3 2 ...  
## $ Sex : chr [1:891] "male" "female" "female" "female" ...  
## $ Age : num [1:891] 22 38 26 35 35 NA 54 2 27 14 ...  
## $ SibSp : num [1:891] 1 1 0 1 0 0 0 3 0 1 ...  
## $ Parch : num [1:891] 0 0 0 0 0 0 0 1 2 0 ...  
## $ Ticket : chr [1:891] "A/5 21171" "PC 17599" "STON/O2. 3101282" "113803" ...  
## $ Fare : num [1:891] 7.25 71.28 7.92 53.1 8.05 ...  
## $ Cabin : chr [1:891] NA "C85" NA "C123" ...  
## $ Embarked : chr [1:891] "S" "C" "S" "S" ...  
## - attr(\*, "spec")=  
## .. cols(  
## .. PassengerId = col\_double(),  
## .. Survived = col\_double(),  
## .. Pclass = col\_double(),  
## .. Sex = col\_character(),  
## .. Age = col\_double(),  
## .. SibSp = col\_double(),  
## .. Parch = col\_double(),  
## .. Ticket = col\_character(),  
## .. Fare = col\_double(),  
## .. Cabin = col\_character(),  
## .. Embarked = col\_character()  
## .. )  
## - attr(\*, "problems")=<externalptr>

summary(Titanic)

## PassengerId Survived Pclass Sex   
## Min. : 1.0 Min. :0.0000 Min. :1.000 Length:891   
## 1st Qu.:223.5 1st Qu.:0.0000 1st Qu.:2.000 Class :character   
## Median :446.0 Median :0.0000 Median :3.000 Mode :character   
## Mean :446.0 Mean :0.3838 Mean :2.309   
## 3rd Qu.:668.5 3rd Qu.:1.0000 3rd Qu.:3.000   
## Max. :891.0 Max. :1.0000 Max. :3.000   
##   
## Age SibSp Parch Ticket   
## Min. : 0.42 Min. :0.000 Min. :0.0000 Length:891   
## 1st Qu.:20.12 1st Qu.:0.000 1st Qu.:0.0000 Class :character   
## Median :28.00 Median :0.000 Median :0.0000 Mode :character   
## Mean :29.70 Mean :0.523 Mean :0.3816   
## 3rd Qu.:38.00 3rd Qu.:1.000 3rd Qu.:0.0000   
## Max. :80.00 Max. :8.000 Max. :6.0000   
## NA's :177   
## Fare Cabin Embarked   
## Min. : 0.00 Length:891 Length:891   
## 1st Qu.: 7.91 Class :character Class :character   
## Median : 14.45 Mode :character Mode :character   
## Mean : 32.20   
## 3rd Qu.: 31.00   
## Max. :512.33   
##

Titanic$Sex <- as.factor(Titanic$Sex)  
Titanic$Pclass <- as.factor(Titanic$Pclass)  
Titanic$Embarked <- as.factor(Titanic$Embarked)  
  
  
# Check Age <= 1  
Titanic$Age[Titanic$Age <= 1]

## [1] NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA  
## [16] NA NA NA NA 0.83 NA NA NA NA NA NA NA NA NA NA  
## [31] NA NA NA 1.00 NA NA 1.00 NA NA NA 1.00 NA NA NA NA  
## [46] NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA  
## [61] NA NA NA NA NA NA 0.92 NA NA NA NA NA NA NA NA  
## [76] NA NA NA NA NA NA 1.00 NA 1.00 NA NA NA NA NA NA  
## [91] NA NA NA NA NA NA NA NA NA NA NA NA 0.75 NA NA  
## [106] NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA  
## [121] NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA  
## [136] NA NA NA NA NA NA NA NA 0.75 NA NA NA NA NA NA  
## [151] NA NA NA NA NA NA NA NA NA NA NA NA 0.67 NA NA  
## [166] NA NA NA NA NA 1.00 NA NA NA 0.42 NA NA NA 1.00 NA  
## [181] 0.83 NA NA NA NA NA NA NA NA NA NA

# Check duplicates and missing values  
sum(duplicated(Titanic))

## [1] 0

colSums(is.na(Titanic))

## PassengerId Survived Pclass Sex Age SibSp   
## 0 0 0 0 177 0   
## Parch Ticket Fare Cabin Embarked   
## 0 0 0 687 2

nrow(Titanic)

## [1] 891

ncol(Titanic)

## [1] 11

# Remove Cabin column (10th col)  
n\_titanic <- Titanic[ , -10]  
head(n\_titanic)

## # A tibble: 6 × 10  
## PassengerId Survived Pclass Sex Age SibSp Parch Ticket Fare Embarked  
## <dbl> <dbl> <fct> <fct> <dbl> <dbl> <dbl> <chr> <dbl> <fct>   
## 1 1 0 3 male 22 1 0 A/5 21171 7.25 S   
## 2 2 1 1 female 38 1 0 PC 17599 71.3 C   
## 3 3 1 3 female 26 0 0 STON/O2. … 7.92 S   
## 4 4 1 1 female 35 1 0 113803 53.1 S   
## 5 5 0 3 male 35 0 0 373450 8.05 S   
## 6 6 0 3 male NA 0 0 330877 8.46 Q

str(n\_titanic)

## tibble [891 × 10] (S3: tbl\_df/tbl/data.frame)  
## $ PassengerId: num [1:891] 1 2 3 4 5 6 7 8 9 10 ...  
## $ Survived : num [1:891] 0 1 1 1 0 0 0 0 1 1 ...  
## $ Pclass : Factor w/ 3 levels "1","2","3": 3 1 3 1 3 3 1 3 3 2 ...  
## $ Sex : Factor w/ 2 levels "female","male": 2 1 1 1 2 2 2 2 1 1 ...  
## $ Age : num [1:891] 22 38 26 35 35 NA 54 2 27 14 ...  
## $ SibSp : num [1:891] 1 1 0 1 0 0 0 3 0 1 ...  
## $ Parch : num [1:891] 0 0 0 0 0 0 0 1 2 0 ...  
## $ Ticket : chr [1:891] "A/5 21171" "PC 17599" "STON/O2. 3101282" "113803" ...  
## $ Fare : num [1:891] 7.25 71.28 7.92 53.1 8.05 ...  
## $ Embarked : Factor w/ 3 levels "C","Q","S": 3 1 3 3 3 2 3 3 3 1 ...

colSums(is.na(n\_titanic))

## PassengerId Survived Pclass Sex Age SibSp   
## 0 0 0 0 177 0   
## Parch Ticket Fare Embarked   
## 0 0 0 2

# Percentage missing (Age)  
177/891 \* 100

## [1] 19.86532

###################################################  
# Missing Data Analysis  
###################################################  
# MCAR Test  
mcar\_test(Titanic)

## # A tibble: 1 × 4  
## statistic df p.value missing.patterns  
## <dbl> <dbl> <dbl> <int>  
## 1 572. 39 0 5

# Missing indicators  
data <- n\_titanic  
data$Age\_missing <- ifelse(is.na(data$Age), 1, 0)  
data$Embarked\_missing <- ifelse(is.na(data$Embarked), 1, 0)  
  
# Chi-square association with missingness  
chisq.test(table(data$Age\_missing, data$Pclass))

##   
## Pearson's Chi-squared test  
##   
## data: table(data$Age\_missing, data$Pclass)  
## X-squared = 46.063, df = 2, p-value = 9.945e-11

chisq.test(table(data$Age\_missing, data$Sex))

##   
## Pearson's Chi-squared test with Yates' continuity correction  
##   
## data: table(data$Age\_missing, data$Sex)  
## X-squared = 2.4344, df = 1, p-value = 0.1187

chisq.test(table(data$Age\_missing, data$SibSp))

## Warning in chisq.test(table(data$Age\_missing, data$SibSp)): Chi-squared  
## approximation may be incorrect

##   
## Pearson's Chi-squared test  
##   
## data: table(data$Age\_missing, data$SibSp)  
## X-squared = 45.633, df = 6, p-value = 3.503e-08

chisq.test(table(data$Age\_missing, data$Parch))

## Warning in chisq.test(table(data$Age\_missing, data$Parch)): Chi-squared  
## approximation may be incorrect

##   
## Pearson's Chi-squared test  
##   
## data: table(data$Age\_missing, data$Parch)  
## X-squared = 22.214, df = 6, p-value = 0.001107

chisq.test(table(data$Age\_missing, data$Fare))

## Warning in chisq.test(table(data$Age\_missing, data$Fare)): Chi-squared  
## approximation may be incorrect

##   
## Pearson's Chi-squared test  
##   
## data: table(data$Age\_missing, data$Fare)  
## X-squared = 416.43, df = 247, p-value = 8.527e-11

chisq.test(table(data$Age\_missing, data$Embarked))

##   
## Pearson's Chi-squared test  
##   
## data: table(data$Age\_missing, data$Embarked)  
## X-squared = 107.33, df = 2, p-value < 2.2e-16

###################################################  
# Multiple Imputation (MICE)  
###################################################  
# Convert categorical variables to factors  
n\_titanic$Sex <- as.factor(n\_titanic$Sex)  
n\_titanic$Embarked <- as.factor(n\_titanic$Embarked)  
  
# Define imputation methods  
meth <- make.method(n\_titanic)  
meth["Age"] <- "pmm" # numeric  
meth["Embarked"] <- "polyreg" # categorical  
meth["Sex"] <- "logreg" # binary categorical  
  
# Run MICE  
imp <- mice(n\_titanic, method = meth, m = 5, seed = 123)

##   
## iter imp variable  
## 1 1 Age Embarked  
## 1 2 Age Embarked  
## 1 3 Age Embarked  
## 1 4 Age Embarked  
## 1 5 Age Embarked  
## 2 1 Age Embarked  
## 2 2 Age Embarked  
## 2 3 Age Embarked  
## 2 4 Age Embarked  
## 2 5 Age Embarked  
## 3 1 Age Embarked  
## 3 2 Age Embarked  
## 3 3 Age Embarked  
## 3 4 Age Embarked  
## 3 5 Age Embarked  
## 4 1 Age Embarked  
## 4 2 Age Embarked  
## 4 3 Age Embarked  
## 4 4 Age Embarked  
## 4 5 Age Embarked  
## 5 1 Age Embarked  
## 5 2 Age Embarked  
## 5 3 Age Embarked  
## 5 4 Age Embarked  
## 5 5 Age Embarked

## Warning: Number of logged events: 1

# Get completed data  
completed\_data <- complete(imp)  
colSums(is.na(completed\_data))

## PassengerId Survived Pclass Sex Age SibSp   
## 0 0 0 0 0 0   
## Parch Ticket Fare Embarked   
## 0 0 0 0

completed\_data\_n <- completed\_data[ , -10]  
str(completed\_data\_n)

## 'data.frame': 891 obs. of 9 variables:  
## $ PassengerId: num 1 2 3 4 5 6 7 8 9 10 ...  
## $ Survived : num 0 1 1 1 0 0 0 0 1 1 ...  
## $ Pclass : Factor w/ 3 levels "1","2","3": 3 1 3 1 3 3 1 3 3 2 ...  
## $ Sex : Factor w/ 2 levels "female","male": 2 1 1 1 2 2 2 2 1 1 ...  
## $ Age : num 22 38 26 35 35 28 54 2 27 14 ...  
## $ SibSp : num 1 1 0 1 0 0 0 3 0 1 ...  
## $ Parch : num 0 0 0 0 0 0 0 1 2 0 ...  
## $ Ticket : chr "A/5 21171" "PC 17599" "STON/O2. 3101282" "113803" ...  
## $ Fare : num 7.25 71.28 7.92 53.1 8.05 ...

###################################################  
# Median Imputation (for skewed Age)  
###################################################  
median(n\_titanic$Age, na.rm = TRUE)

## [1] 28

n\_titanic$Embarked=as.character(n\_titanic$Embarked)  
# Imputation: Age (median), Embarked (mode)  
imputed\_data <- n\_titanic %>%  
 replace\_na(list(  
 Age = median(n\_titanic$Age, na.rm = TRUE),  
 Embarked = mode(n\_titanic$Embarked)  
 ))  
  
head(imputed\_data)

## # A tibble: 6 × 10  
## PassengerId Survived Pclass Sex Age SibSp Parch Ticket Fare Embarked  
## <dbl> <dbl> <fct> <fct> <dbl> <dbl> <dbl> <chr> <dbl> <chr>   
## 1 1 0 3 male 22 1 0 A/5 21171 7.25 S   
## 2 2 1 1 female 38 1 0 PC 17599 71.3 C   
## 3 3 1 3 female 26 0 0 STON/O2. … 7.92 S   
## 4 4 1 1 female 35 1 0 113803 53.1 S   
## 5 5 0 3 male 35 0 0 373450 8.05 S   
## 6 6 0 3 male 28 0 0 330877 8.46 Q

colSums(is.na(imputed\_data))

## PassengerId Survived Pclass Sex Age SibSp   
## 0 0 0 0 0 0   
## Parch Ticket Fare Embarked   
## 0 0 0 0

###################################################  
# Logistic Regression Model  
###################################################  
# Fit model  
model <- glm(Survived ~ Pclass + Sex + Age + SibSp + Parch + Fare + Embarked,  
 data = imputed\_data,  
 family = binomial)  
summary(model)

##   
## Call:  
## glm(formula = Survived ~ Pclass + Sex + Age + SibSp + Parch +   
## Fare + Embarked, family = binomial, data = imputed\_data)  
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) 4.062486 0.472734 8.594 < 2e-16 \*\*\*  
## Pclass2 -0.911903 0.297391 -3.066 0.00217 \*\*   
## Pclass3 -2.144097 0.297668 -7.203 5.89e-13 \*\*\*  
## Sexmale -2.710309 0.201224 -13.469 < 2e-16 \*\*\*  
## Age -0.038752 0.007873 -4.922 8.55e-07 \*\*\*  
## SibSp -0.320495 0.109056 -2.939 0.00329 \*\*   
## Parch -0.091313 0.118850 -0.768 0.44231   
## Fare 0.002304 0.002462 0.936 0.34940   
## Embarkedcharacter 12.318160 610.882033 0.020 0.98391   
## EmbarkedQ -0.057728 0.381060 -0.151 0.87959   
## EmbarkedS -0.440140 0.239533 -1.837 0.06614 .   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 1186.66 on 890 degrees of freedom  
## Residual deviance: 784.42 on 880 degrees of freedom  
## AIC: 806.42  
##   
## Number of Fisher Scoring iterations: 13

# Backward Selection  
backward\_model <- stats::step(model, direction = "backward")

## Start: AIC=806.42  
## Survived ~ Pclass + Sex + Age + SibSp + Parch + Fare + Embarked  
##   
## Df Deviance AIC  
## - Parch 1 785.02 805.02  
## - Embarked 3 789.09 805.09  
## - Fare 1 785.37 805.37  
## <none> 784.42 806.42  
## - SibSp 1 794.52 814.52  
## - Age 1 810.58 830.58  
## - Pclass 2 844.15 862.15  
## - Sex 1 1011.87 1031.87  
##   
## Step: AIC=805.02  
## Survived ~ Pclass + Sex + Age + SibSp + Fare + Embarked  
##   
## Df Deviance AIC  
## - Fare 1 785.71 803.71  
## - Embarked 3 789.96 803.96  
## <none> 785.02 805.02  
## - SibSp 1 797.57 815.57  
## - Age 1 811.03 829.03  
## - Pclass 2 848.01 864.01  
## - Sex 1 1017.31 1035.31  
##   
## Step: AIC=803.71  
## Survived ~ Pclass + Sex + Age + SibSp + Embarked  
##   
## Df Deviance AIC  
## - Embarked 3 791.23 803.23  
## <none> 785.71 803.71  
## - SibSp 1 797.58 813.58  
## - Age 1 812.43 828.43  
## - Pclass 2 882.66 896.66  
## - Sex 1 1023.62 1039.62  
##   
## Step: AIC=803.23  
## Survived ~ Pclass + Sex + Age + SibSp  
##   
## Df Deviance AIC  
## <none> 791.23 803.23  
## - SibSp 1 805.53 815.53  
## - Age 1 819.15 829.15  
## - Pclass 2 902.20 910.20  
## - Sex 1 1044.35 1054.35

summary(backward\_model)

##   
## Call:  
## glm(formula = Survived ~ Pclass + Sex + Age + SibSp, family = binomial,   
## data = imputed\_data)  
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) 4.003495 0.397214 10.079 < 2e-16 \*\*\*  
## Pclass2 -1.180514 0.261362 -4.517 6.28e-06 \*\*\*  
## Pclass3 -2.352135 0.243384 -9.664 < 2e-16 \*\*\*  
## Sexmale -2.739590 0.194053 -14.118 < 2e-16 \*\*\*  
## Age -0.039568 0.007795 -5.076 3.86e-07 \*\*\*  
## SibSp -0.354559 0.103566 -3.424 0.000618 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 1186.66 on 890 degrees of freedom  
## Residual deviance: 791.23 on 885 degrees of freedom  
## AIC: 803.23  
##   
## Number of Fisher Scoring iterations: 5

# Odds Ratios with 95% CI  
exp(cbind(OR = coef(backward\_model), confint(backward\_model)))

## Waiting for profiling to be done...

## OR 2.5 % 97.5 %  
## (Intercept) 54.78933059 25.68109969 122.03944134  
## Pclass2 0.30712090 0.18283044 0.50992262  
## Pclass3 0.09516580 0.05848118 0.15201182  
## Sexmale 0.06459685 0.04377309 0.09374566  
## Age 0.96120449 0.94634903 0.97575052  
## SibSp 0.70148257 0.56656771 0.85025176

###################################################  
# Train-Test Split  
###################################################  
set.seed(123)  
data\_split <- initial\_split(imputed\_data, prop = 0.8, strata = Survived)  
  
train\_data <- training(data\_split)  
test\_data <- testing(data\_split)  
  
# Train model  
model <- glm(Survived ~ Pclass + Sex + Age + SibSp + Parch + Fare + Embarked,  
 data = train\_data,  
 family = binomial)  
summary(model)

##   
## Call:  
## glm(formula = Survived ~ Pclass + Sex + Age + SibSp + Parch +   
## Fare + Embarked, family = binomial, data = train\_data)  
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) 3.950153 0.517765 7.629 2.36e-14 \*\*\*  
## Pclass2 -0.952833 0.322676 -2.953 0.00315 \*\*   
## Pclass3 -2.163068 0.319248 -6.776 1.24e-11 \*\*\*  
## Sexmale -2.638641 0.223636 -11.799 < 2e-16 \*\*\*  
## Age -0.037596 0.008819 -4.263 2.01e-05 \*\*\*  
## SibSp -0.239568 0.127330 -1.881 0.05991 .   
## Parch -0.131722 0.131115 -1.005 0.31507   
## Fare 0.001983 0.002526 0.785 0.43247   
## EmbarkedQ -0.285381 0.432042 -0.661 0.50891   
## EmbarkedS -0.336368 0.263703 -1.276 0.20211   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 947.98 on 711 degrees of freedom  
## Residual deviance: 641.34 on 702 degrees of freedom  
## AIC: 661.34  
##   
## Number of Fisher Scoring iterations: 5

# Backward Selection  
backward\_model <- stats::step(model, direction = "backward")

## Start: AIC=661.34  
## Survived ~ Pclass + Sex + Age + SibSp + Parch + Fare + Embarked  
##   
## Df Deviance AIC  
## - Embarked 2 642.96 658.96  
## - Fare 1 642.00 660.00  
## - Parch 1 642.38 660.38  
## <none> 641.34 661.34  
## - SibSp 1 645.23 663.23  
## - Age 1 660.82 678.82  
## - Pclass 2 693.03 709.03  
## - Sex 1 813.44 831.44  
##   
## Step: AIC=658.96  
## Survived ~ Pclass + Sex + Age + SibSp + Parch + Fare  
##   
## Df Deviance AIC  
## - Fare 1 644.04 658.04  
## - Parch 1 644.12 658.12  
## <none> 642.96 658.96  
## - SibSp 1 647.43 661.43  
## - Age 1 663.07 677.07  
## - Pclass 2 697.79 709.79  
## - Sex 1 820.81 834.81  
##   
## Step: AIC=658.04  
## Survived ~ Pclass + Sex + Age + SibSp + Parch  
##   
## Df Deviance AIC  
## - Parch 1 644.88 656.88  
## <none> 644.04 658.04  
## - SibSp 1 648.11 660.11  
## - Age 1 664.99 676.99  
## - Pclass 2 736.05 746.05  
## - Sex 1 825.38 837.38  
##   
## Step: AIC=656.88  
## Survived ~ Pclass + Sex + Age + SibSp  
##   
## Df Deviance AIC  
## <none> 644.88 656.88  
## - SibSp 1 651.01 661.01  
## - Age 1 666.19 676.19  
## - Pclass 2 738.47 746.47  
## - Sex 1 830.76 840.76

summary(backward\_model)

##   
## Call:  
## glm(formula = Survived ~ Pclass + Sex + Age + SibSp, family = binomial,   
## data = train\_data)  
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) 3.889916 0.441686 8.807 < 2e-16 \*\*\*  
## Pclass2 -1.176515 0.286854 -4.101 4.11e-05 \*\*\*  
## Pclass3 -2.376989 0.267722 -8.879 < 2e-16 \*\*\*  
## Sexmale -2.610744 0.213365 -12.236 < 2e-16 \*\*\*  
## Age -0.038921 0.008759 -4.443 8.85e-06 \*\*\*  
## SibSp -0.278010 0.119844 -2.320 0.0204 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 947.98 on 711 degrees of freedom  
## Residual deviance: 644.88 on 706 degrees of freedom  
## AIC: 656.88  
##   
## Number of Fisher Scoring iterations: 5

# Predictions  
pred\_probs <- predict(backward\_model, newdata = test\_data, type = "response")  
pred\_class <- ifelse(pred\_probs > 0.5, 1, 0)  
  
# Confusion Matrix  
table(Predicted = pred\_class, Actual = test\_data$Survived)

## Actual  
## Predicted 0 1  
## 0 95 16  
## 1 15 53

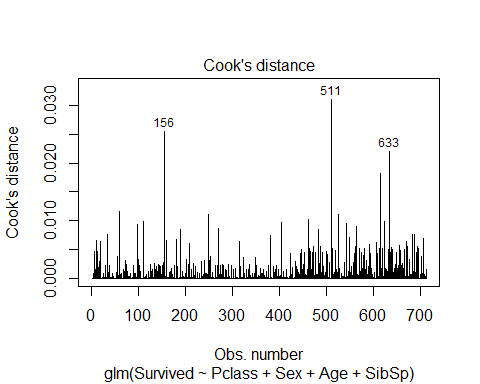
###################################################  
# Diagnostics  
###################################################  
# Frequency table  
table(imputed\_data$Survived)

##   
## 0 1   
## 549 342

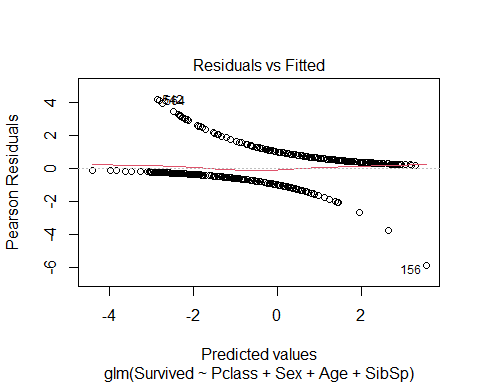
# Multicollinearity (VIF)  
vif(backward\_model)

## GVIF Df GVIF^(1/(2\*Df))  
## Pclass 1.321597 2 1.072197  
## Sex 1.125202 1 1.060755  
## Age 1.313441 1 1.146055  
## SibSp 1.117324 1 1.057036

# Influence diagnostics  
plot(backward\_model, which = 4) # Cook's distance



plot(backward\_model, which = 1) # Residuals vs fitted



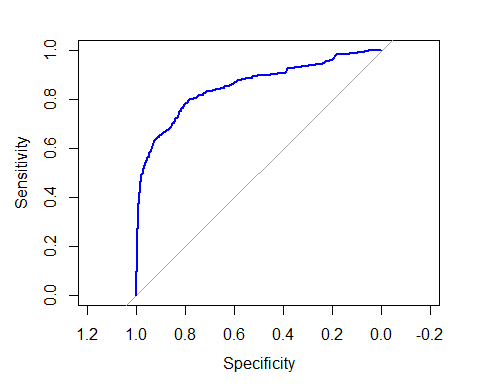
nrow(imputed\_data$Survived)

## NULL

###################################################  
# ROC Curve & AUC  
###################################################  
library(pROC)  
  
# Get predicted probabilities for class = 1  
pred\_probs <- predict(backward\_model, newdata = imputed\_data, type = "response")  
  
# Now both are length 891  
roc\_curve <- roc(imputed\_data$Survived, pred\_probs)

## Setting levels: control = 0, case = 1  
## Setting direction: controls < cases

plot(roc\_curve, col = "blue")



auc(roc\_curve)

## Area under the curve: 0.8539

###################################################  
# Hosmer-Lemeshow Test  
###################################################  
  
hoslem.test(train\_data$Survived, fitted(backward\_model))

##   
## Hosmer and Lemeshow goodness of fit (GOF) test  
##   
## data: train\_data$Survived, fitted(backward\_model)  
## X-squared = 25.503, df = 8, p-value = 0.001277

# Likelihood ratio test  
anova(backward\_model, test = "Chisq")

## Analysis of Deviance Table  
##   
## Model: binomial, link: logit  
##   
## Response: Survived  
##   
## Terms added sequentially (first to last)  
##   
##   
## Df Deviance Resid. Df Resid. Dev Pr(>Chi)   
## NULL 711 947.98   
## Pclass 2 90.021 709 857.96 < 2.2e-16 \*\*\*  
## Sex 1 189.515 708 668.44 < 2.2e-16 \*\*\*  
## Age 1 17.438 707 651.01 2.968e-05 \*\*\*  
## SibSp 1 6.126 706 644.88 0.01332 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

# Pseudo R-squared  
pR2(backward\_model)

## fitting null model for pseudo-r2

## llh llhNull G2 McFadden r2ML r2CU   
## -322.4402059 -473.9904305 303.1004492 0.3197327 0.3466896 0.4711086

###################################################  
# Influence Diagnostics - Cook's Distance  
###################################################  
cooks\_d <- cooks.distance(backward\_model)  
cutoff <- 4 / nrow(imputed\_data)  
influential\_points <- which(cooks\_d > cutoff)  
  
cat("Number of influential points removed:", length(influential\_points), "\n")

## Number of influential points removed: 67

# Remove influential points  
clean\_data <- imputed\_data[-influential\_points, ]  
  
  
###################################################  
# Refit Model on Clean Data  
###################################################  
model\_clean <- glm(Survived ~ Pclass + Sex + Age + SibSp + Parch + Fare + Embarked,  
 data = clean\_data,  
 family = binomial)  
  
# Stepwise Backward Selection  
backward\_model <- stats::step(model\_clean, direction = "backward")

## Start: AIC=747.42  
## Survived ~ Pclass + Sex + Age + SibSp + Parch + Fare + Embarked  
##   
## Df Deviance AIC  
## - Embarked 3 728.83 744.83  
## - Parch 1 725.74 745.74  
## - Fare 1 726.44 746.44  
## <none> 725.42 747.42  
## - SibSp 1 734.11 754.11  
## - Age 1 749.98 769.98  
## - Pclass 2 778.02 796.02  
## - Sex 1 942.63 962.63  
##   
## Step: AIC=744.83  
## Survived ~ Pclass + Sex + Age + SibSp + Parch + Fare  
##   
## Df Deviance AIC  
## - Parch 1 729.34 743.34  
## - Fare 1 730.56 744.56  
## <none> 728.83 744.83  
## - SibSp 1 738.97 752.97  
## - Age 1 753.90 767.90  
## - Pclass 2 782.96 794.96  
## - Sex 1 959.99 973.99  
##   
## Step: AIC=743.34  
## Survived ~ Pclass + Sex + Age + SibSp + Fare  
##   
## Df Deviance AIC  
## - Fare 1 730.75 742.75  
## <none> 729.34 743.34  
## - SibSp 1 741.84 753.84  
## - Age 1 754.28 766.28  
## - Pclass 2 786.40 796.40  
## - Sex 1 964.70 976.70  
##   
## Step: AIC=742.75  
## Survived ~ Pclass + Sex + Age + SibSp  
##   
## Df Deviance AIC  
## <none> 730.75 742.75  
## - SibSp 1 742.09 752.09  
## - Age 1 756.71 766.71  
## - Pclass 2 830.89 838.89  
## - Sex 1 972.19 982.19

summary(backward\_model)

##   
## Call:  
## glm(formula = Survived ~ Pclass + Sex + Age + SibSp, family = binomial,   
## data = clean\_data)  
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) 3.989336 0.410843 9.710 < 2e-16 \*\*\*  
## Pclass2 -1.199349 0.272358 -4.404 1.06e-05 \*\*\*  
## Pclass3 -2.323902 0.252333 -9.210 < 2e-16 \*\*\*  
## Sexmale -2.763884 0.200824 -13.763 < 2e-16 \*\*\*  
## Age -0.039779 0.008134 -4.891 1.00e-06 \*\*\*  
## SibSp -0.320950 0.104749 -3.064 0.00218 \*\*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 1098.10 on 823 degrees of freedom  
## Residual deviance: 730.75 on 818 degrees of freedom  
## AIC: 742.75  
##   
## Number of Fisher Scoring iterations: 5

# Train-test split (clean data)  
set.seed(123)  
data\_split <- initial\_split(clean\_data, prop = 0.8, strata = Survived)  
  
train\_data <- training(data\_split)  
test\_data <- testing(data\_split)  
  
model <- glm(Survived ~ Pclass + Sex + Age + SibSp + Parch + Fare + Embarked,  
 data = train\_data,  
 family = binomial)  
summary(model)

##   
## Call:  
## glm(formula = Survived ~ Pclass + Sex + Age + SibSp + Parch +   
## Fare + Embarked, family = binomial, data = train\_data)  
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) 4.138980 0.559217 7.401 1.35e-13 \*\*\*  
## Pclass2 -0.908701 0.349573 -2.599 0.00934 \*\*   
## Pclass3 -2.132652 0.349259 -6.106 1.02e-09 \*\*\*  
## Sexmale -2.844595 0.238393 -11.932 < 2e-16 \*\*\*  
## Age -0.041349 0.009381 -4.408 1.04e-05 \*\*\*  
## SibSp -0.294121 0.122113 -2.409 0.01601 \*   
## Parch -0.077181 0.128337 -0.601 0.54758   
## Fare 0.002165 0.002955 0.733 0.46370   
## Embarkedcharacter 11.817532 535.411316 0.022 0.98239   
## EmbarkedQ -0.282928 0.450744 -0.628 0.53021   
## EmbarkedS -0.434499 0.283130 -1.535 0.12488   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 876.75 on 657 degrees of freedom  
## Residual deviance: 563.21 on 647 degrees of freedom  
## AIC: 585.21  
##   
## Number of Fisher Scoring iterations: 12

backward\_model <- stats::step(model, direction = "backward")

## Start: AIC=585.21  
## Survived ~ Pclass + Sex + Age + SibSp + Parch + Fare + Embarked  
##   
## Df Deviance AIC  
## - Embarked 3 565.98 581.98  
## - Parch 1 563.58 583.58  
## - Fare 1 563.78 583.78  
## <none> 563.21 585.21  
## - SibSp 1 570.01 590.01  
## - Age 1 584.29 604.29  
## - Pclass 2 605.69 623.69  
## - Sex 1 745.74 765.74  
##   
## Step: AIC=581.98  
## Survived ~ Pclass + Sex + Age + SibSp + Parch + Fare  
##   
## Df Deviance AIC  
## - Parch 1 566.49 580.49  
## - Fare 1 567.04 581.04  
## <none> 565.98 581.98  
## - SibSp 1 573.82 587.82  
## - Age 1 587.40 601.40  
## - Pclass 2 610.86 622.86  
## - Sex 1 758.22 772.22  
##   
## Step: AIC=580.49  
## Survived ~ Pclass + Sex + Age + SibSp + Fare  
##   
## Df Deviance AIC  
## - Fare 1 567.26 579.26  
## <none> 566.49 580.49  
## - SibSp 1 575.97 587.97  
## - Age 1 587.97 599.97  
## - Pclass 2 614.93 624.93  
## - Sex 1 762.80 774.80  
##   
## Step: AIC=579.26  
## Survived ~ Pclass + Sex + Age + SibSp  
##   
## Df Deviance AIC  
## <none> 567.26 579.26  
## - SibSp 1 576.02 586.02  
## - Age 1 589.42 599.42  
## - Pclass 2 650.30 658.30  
## - Sex 1 770.57 780.57

summary(backward\_model)

##   
## Call:  
## glm(formula = Survived ~ Pclass + Sex + Age + SibSp, family = binomial,   
## data = train\_data)  
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) 4.036384 0.463094 8.716 < 2e-16 \*\*\*  
## Pclass2 -1.150403 0.307046 -3.747 0.000179 \*\*\*  
## Pclass3 -2.358556 0.282238 -8.357 < 2e-16 \*\*\*  
## Sexmale -2.863596 0.228891 -12.511 < 2e-16 \*\*\*  
## Age -0.041841 0.009269 -4.514 6.36e-06 \*\*\*  
## SibSp -0.313379 0.116388 -2.693 0.007091 \*\*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 876.75 on 657 degrees of freedom  
## Residual deviance: 567.26 on 652 degrees of freedom  
## AIC: 579.26  
##   
## Number of Fisher Scoring iterations: 5

# Predictions  
pred\_probs <- predict(backward\_model, newdata = test\_data, type = "response")  
pred\_class <- ifelse(pred\_probs > 0.5, 1, 0)  
  
# Confusion Matrix  
y <- table(Predicted = pred\_class, Actual = test\_data$Survived)  
y

## Actual  
## Predicted 0 1  
## 0 89 24  
## 1 13 40

###################################################  
# Odds Ratios (Clean Data)  
###################################################  
exp(cbind(OR = coef(backward\_model), confint(backward\_model)))

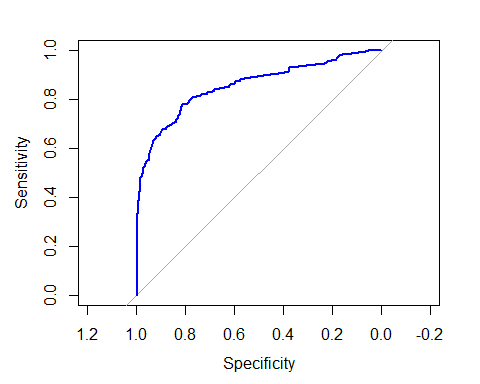
## Waiting for profiling to be done...

## OR 2.5 % 97.5 %  
## (Intercept) 56.62125439 23.50832380 144.76646602  
## Pclass2 0.31650923 0.17191022 0.57392532  
## Pclass3 0.09455670 0.05362824 0.16245341  
## Sexmale 0.05706318 0.03597079 0.08836176  
## Age 0.95902257 0.94136181 0.97625377  
## SibSp 0.73097287 0.57288811 0.90484936

###################################################  
# ROC Curve & AUC (Clean Data)  
###################################################  
  
library(pROC)  
  
# Get predicted probabilities for class = 1  
pred\_probs <- predict(backward\_model, newdata = clean\_data, type = "response")  
  
# Now both are length 891  
roc\_curve <- roc(clean\_data$Survived, pred\_probs)

## Setting levels: control = 0, case = 1  
## Setting direction: controls < cases

plot(roc\_curve, col = "blue")



auc(roc\_curve)

## Area under the curve: 0.8538

###################################################  
# Hosmer-Lemeshow Test (Clean Data)  
###################################################  
hoslem.test(train\_data$Survived, fitted(backward\_model))

##   
## Hosmer and Lemeshow goodness of fit (GOF) test  
##   
## data: train\_data$Survived, fitted(backward\_model)  
## X-squared = 26.561, df = 8, p-value = 0.0008417

# Pseudo R-squared  
pR2(backward\_model)

## fitting null model for pseudo-r2

## llh llhNull G2 McFadden r2ML r2CU   
## -283.6285853 -438.3750420 309.4929133 0.3530002 0.3752189 0.5096912

###################################################  
# Normality Checks - Kolmogorov-Smirnov Test  
###################################################  
# Age  
ks.test(imputed\_data$Age, "pnorm",   
 mean=mean(imputed\_data$Age, na.rm=TRUE),   
 sd=sd(imputed\_data$Age, na.rm=TRUE))

## Warning in ks.test.default(imputed\_data$Age, "pnorm", mean =  
## mean(imputed\_data$Age, : ties should not be present for the one-sample  
## Kolmogorov-Smirnov test

##   
## Asymptotic one-sample Kolmogorov-Smirnov test  
##   
## data: imputed\_data$Age  
## D = 0.14658, p-value < 2.2e-16  
## alternative hypothesis: two-sided

# Fare  
ks.test(imputed\_data$Fare, "pnorm",   
 mean=mean(imputed\_data$Fare, na.rm=TRUE),   
 sd=sd(imputed\_data$Fare, na.rm=TRUE))

## Warning in ks.test.default(imputed\_data$Fare, "pnorm", mean =  
## mean(imputed\_data$Fare, : ties should not be present for the one-sample  
## Kolmogorov-Smirnov test

##   
## Asymptotic one-sample Kolmogorov-Smirnov test  
##   
## data: imputed\_data$Fare  
## D = 0.28185, p-value < 2.2e-16  
## alternative hypothesis: two-sided

# Loop for continuous vars  
cont\_vars <- c("Age", "Fare")  
for (v in cont\_vars) {  
 cat("\nK-S test for", v, ":\n")  
 print(  
 ks.test(imputed\_data[[v]], "pnorm",   
 mean=mean(imputed\_data[[v]], na.rm=TRUE),   
 sd=sd(imputed\_data[[v]], na.rm=TRUE))  
 )  
}

##   
## K-S test for Age :

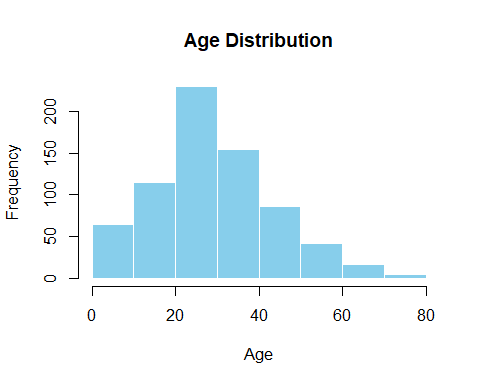
## Warning in ks.test.default(imputed\_data[[v]], "pnorm", mean =  
## mean(imputed\_data[[v]], : ties should not be present for the one-sample  
## Kolmogorov-Smirnov test

##   
## Asymptotic one-sample Kolmogorov-Smirnov test  
##   
## data: imputed\_data[[v]]  
## D = 0.14658, p-value < 2.2e-16  
## alternative hypothesis: two-sided  
##   
##   
## K-S test for Fare :

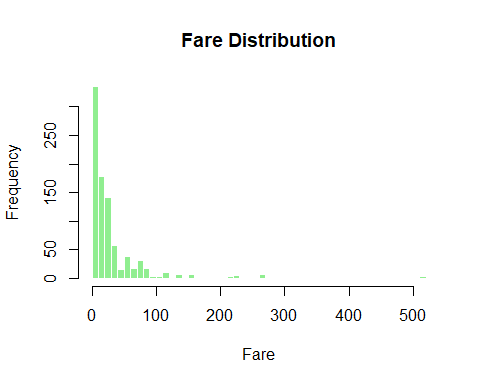
## Warning in ks.test.default(imputed\_data[[v]], "pnorm", mean =  
## mean(imputed\_data[[v]], : ties should not be present for the one-sample  
## Kolmogorov-Smirnov test

##   
## Asymptotic one-sample Kolmogorov-Smirnov test  
##   
## data: imputed\_data[[v]]  
## D = 0.28185, p-value < 2.2e-16  
## alternative hypothesis: two-sided

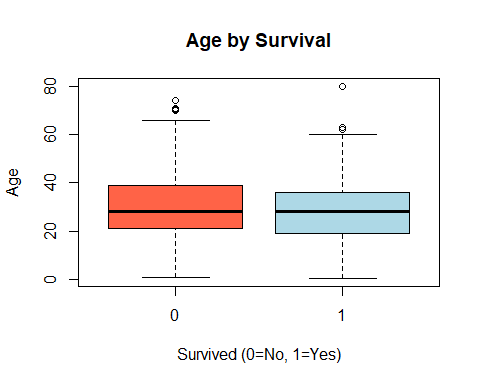
###################################################  
# Visualizations  
###################################################  
# Histogram - Age  
hist(Titanic$Age, main="Age Distribution", xlab="Age",   
 col="skyblue", border="white")



# Histogram - Fare  
hist(Titanic$Fare, main="Fare Distribution", xlab="Fare",   
 col="lightgreen", border="white", breaks=50)

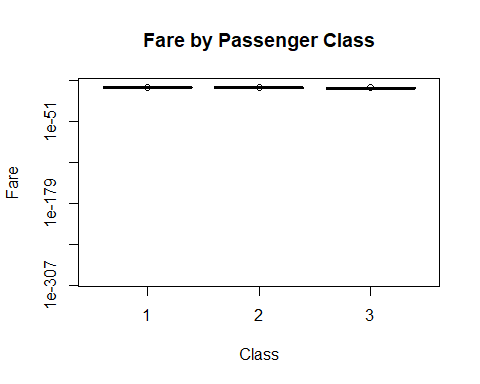


# Boxplot - Age by Survival  
boxplot(Age ~ Survived, data=Titanic,  
 main="Age by Survival", xlab="Survived (0=No, 1=Yes)",   
 ylab="Age", col=c("tomato", "lightblue"))

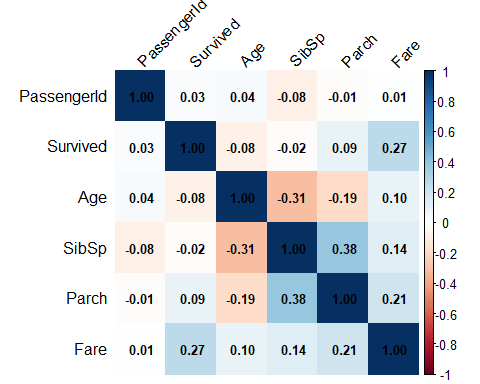


# Boxplot - Fare by Class  
boxplot(Fare ~ Pclass, data=Titanic,  
 main="Fare by Passenger Class", xlab="Class", ylab="Fare",  
 col=c("orange","lightgreen","lightblue"), log="y")

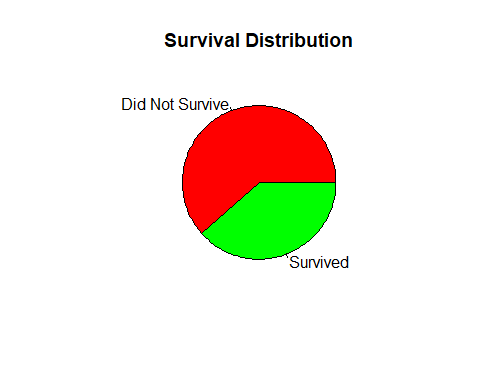
## Warning in plot.window(xlim = xlim, ylim = ylim, log = log, yaxs = pars$yaxs):  
## nonfinite axis=2 limits [GScale(-inf,2.70955,..); log=TRUE] -- corrected now



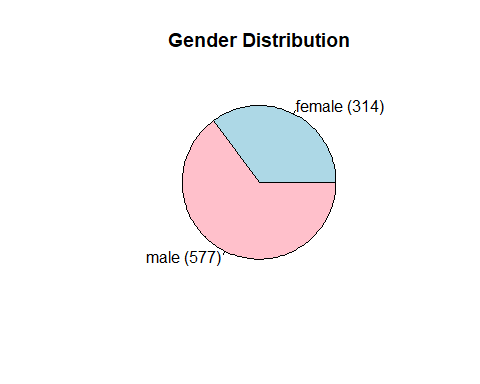
# Correlation Plot  
# Select only numeric columns  
numeric\_vars <- Titanic[, sapply(Titanic, is.numeric)]  
  
# Compute correlation matrix  
corr\_matrix <- cor(numeric\_vars, use = "complete.obs")  
  
# Load library  
library(corrplot)  
  
# Plot correlation heatmap with numbers  
corrplot(corr\_matrix,  
 method = "color", # colored squares  
 type = "full", # full matrix  
 addCoef.col = "black", # add correlation numbers in black  
 tl.col = "black", # text label color  
 tl.srt = 45, # text label rotation  
 number.cex = 0.8) # size of correlation numbers



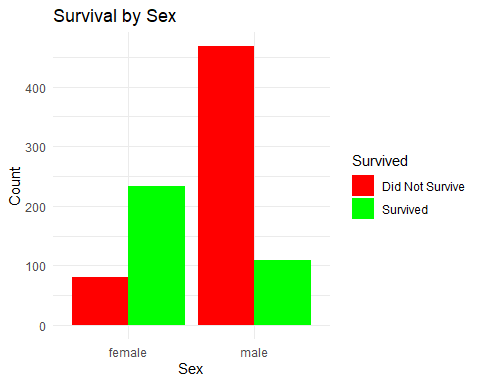
# Pie chart - Survival  
survived\_counts <- table(Titanic$Survived)  
pie(survived\_counts,  
 labels=c("Did Not Survive","Survived"),  
 main="Survival Distribution",  
 col=c("red","green"))



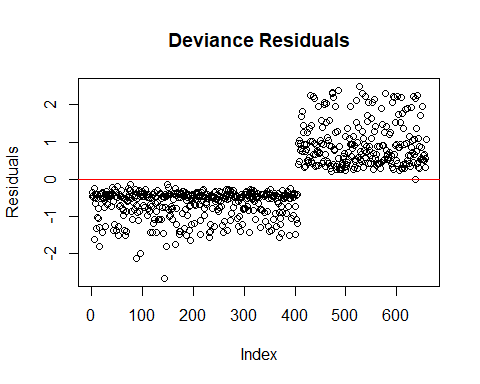
# Pie chart - Sex  
sex\_counts <- table(Titanic$Sex)  
pie(sex\_counts,  
 labels=paste(names(sex\_counts)," (",sex\_counts,")",sep=""),  
 main="Gender Distribution",  
 col=c("lightblue","pink"))



library(ggplot2)  
  
# Clustered bar plot  
ggplot(Titanic, aes(x = Sex, fill = factor(Survived))) +  
 geom\_bar(position = "dodge") +  
 labs(title = "Survival by Sex",  
 x = "Sex",  
 y = "Count",  
 fill = "Survived") +  
 scale\_fill\_manual(values = c("0" = "red", "1" = "green"),  
 labels = c("0" = "Did Not Survive", "1" = "Survived")) +  
 theme\_minimal()



###################################################  
# Residual Analysis  
###################################################  
# Deviance residuals  
dev\_resid <- residuals(model, type="deviance")  
plot(dev\_resid, main="Deviance Residuals", ylab="Residuals")  
abline(h=0, col="red")



# Pearson residuals  
pearson\_resid <- residuals(model, type="pearson")  
plot(pearson\_resid, main="Pearson Residuals", ylab="Residuals")  
abline(h=0, col="red")

