STAA 577: HW2

Matthew Stoebe

Problem 1

$$p(x) = \frac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}}$$

$$\begin{aligned}
\frac{p(x)}{1-p(x)} &= \frac{\frac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}}}{1-\frac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}}} \\
&= \frac{\frac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}}}{\frac{1+e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}}-\frac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}}} \\
&= \frac{\frac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}}}{\frac{1}{1+e^{\beta_0+\beta_1 x}}} \\
&= e^{\beta_0+\beta_1 x}
\end{aligned}$$

Problem 2

(a)

$$\begin{aligned}
odds = .37 &= \frac{p}{1-p} \\
.37 * (1-p) = .37 - .37p &= p \\
.37 = 1p + .37p &= 1.37p \\
p = \frac{.37}{1.37} &= .27
\end{aligned}$$

(b)

$$odds = \frac{p}{1-p} = \frac{.16}{1-.16}$$

$$odds = \frac{.16}{.84} = .19$$

Problem 3

(a)

$$\begin{aligned}\hat{\beta}_0 + \hat{\beta}_1(30) + \hat{\beta}_2(3.25) &= -5 + 0.1 \times 30 + 1 \times 3.25 \\ &= -5 + 3 + 3.25 \\ &= 1.25\end{aligned}$$

$$\hat{p} = \frac{e^{1.25}}{1+e^{1.25}} = \frac{3.4903}{1+3.4903} \approx 0.78$$

(b)

$$\log\left(\frac{0.60}{1-0.60}\right) = \beta_0 + \beta_1(x) + \hat{\beta}_2(3.25)$$

$$0.4054 \approx -5 + 0.1h + 1(3.25)$$

$$h \approx 21.6 \text{ hours}$$

Problem 4

(a)

```
##
## Call:
## glm(formula = survived ~ factor(passenger_class) + gender + ns(age,
##     3) + ns(fare_paid, 3), family = binomial, data = titanic)
##
## Coefficients:
##                          Estimate Std. Error z value Pr(>|z|)
## (Intercept)                1.5998     0.7461   2.144 0.032012 *
## factor(passenger_class)2  -1.4095     0.3270  -4.310 1.63e-05 ***
## factor(passenger_class)3  -2.6304     0.3443  -7.639 2.19e-14 ***
## genderwoman                2.6192     0.1930  13.570  < 2e-16 ***
## ns(age, 3)1               -0.7952     0.4931  -1.613 0.106798    
## ns(age, 3)2               -4.4320     0.9977  -4.442 8.91e-06 ***
## ns(age, 3)3               -3.4341     0.9588  -3.582 0.000341 ***
## ns(fare_paid, 3)1         -1.5955     1.0096  -1.580 0.114041    
## ns(fare_paid, 3)2          0.5982     1.2976   0.461 0.644814    
## ns(fare_paid, 3)3          2.0100     1.9612   1.025 0.305417    
## ---
## 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:  793.67  on 877  degrees of freedom
## AIC: 813.67
##
## Number of Fisher Scoring iterations: 5
```

The fitted beta for female is 2.6192. If we exponentiate this we get an odds value of 13.92. This means that women had signifficantly higher odds of surviving

(b)

```
##         1
## 0.8124541
```

The Probability of surviving under these conditions is .812

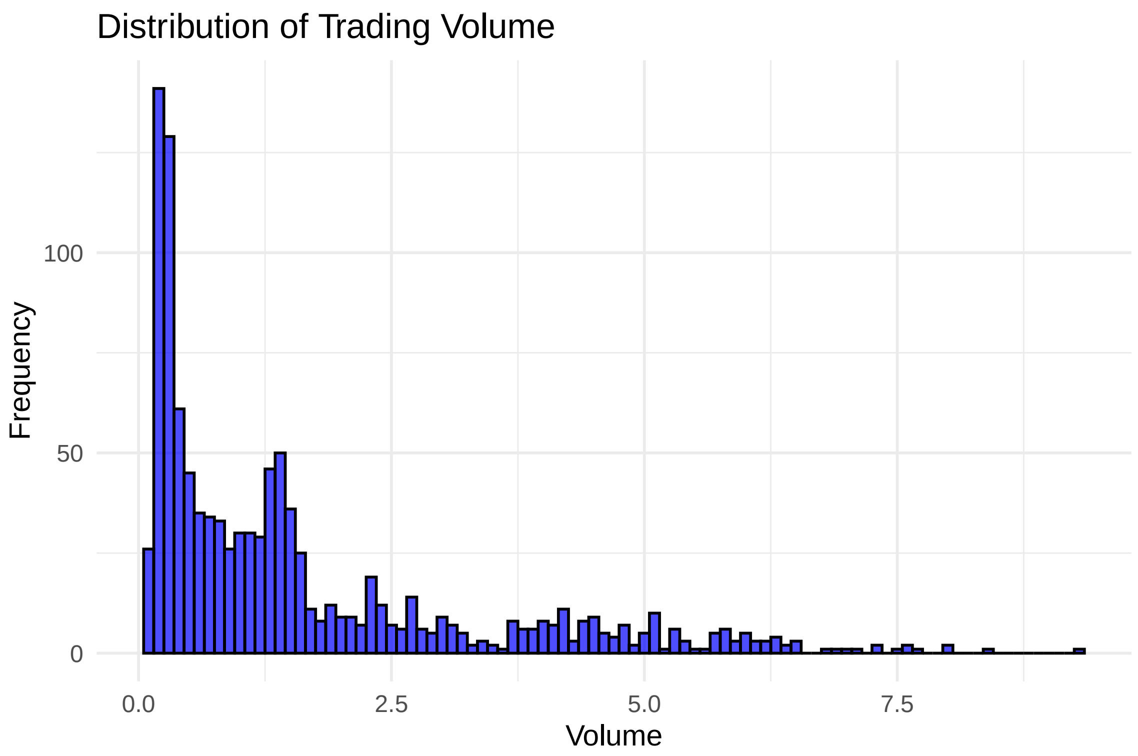
(c)

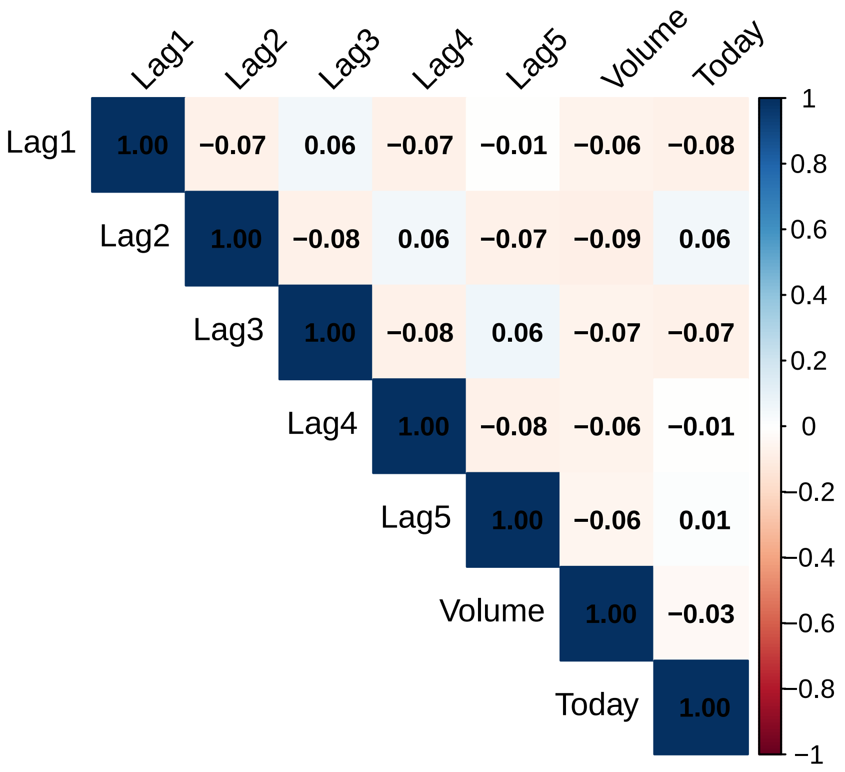
```
##         1
## 0.1196773
```

This Child has approximately a .12 probability of survival

Problem 5

(a)





The Correlation between these features is is so low that I wouldnt even proceed to train a model in this case.

(b)

```
##
## Call:
## glm(formula = Direction ~ Lag1 + Lag2 + Lag3 + Lag4 + Lag5 +
##     Volume, family = binomial, data = Weekly)
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.26686    0.08593   3.106   0.0019 **
## Lag1        -0.04127    0.02641  -1.563   0.1181
## Lag2         0.05844    0.02686   2.175   0.0296 *
## Lag3        -0.01606    0.02666  -0.602   0.5469
## Lag4        -0.02779    0.02646  -1.050   0.2937
## Lag5        -0.01447    0.02638  -0.549   0.5833
## Volume      -0.02274    0.03690  -0.616   0.5377
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1496.2  on 1088  degrees of freedom
## Residual deviance: 1486.4  on 1082  degrees of freedom
## AIC: 1500.4
##
## Number of Fisher Scoring iterations: 4
```

Only the Lag2 feature appears significant. this is not surprising as the correlation was so low above. This feature set is quite weak.

(c)

```
##          Actual
## Predicted Down  Up
##      Down   54  48
##      Up    430 557
```

```
## [1] 0.5610652
```

This Model is terrible and overly optimistic. It predicts the marekt to go up almost 90% of thhe time when it only goes up around 60% of the time.

The correct prediction rate is 56 percent which is terrible. This is almost as bad as always predicting the market to go up.

Additionally, We are evaluating on the train set here which makes this performance even more terrible.

(d)

```
##          Actual
## Predicted Down Up
##      Down    3  0
##      Up     17 32
## [1] 0.6730769
```

This model supposedly performs better but we are using a poor eval metric. Again this is overly optimistic predicting up in all but 3 cases. Recall may be 100% but our precision is terribly low.

Problem 6

```r
heart_train <- read.csv("heart_training.csv")
heart_test <- read.csv("heart_test.csv")

glm.final <- glm(target ~ age + slope + thalach, data = heart_train, family = binomial)
summary(glm.final)
prob.test <- predict(glm.final, heart_test, type = "response")
pred.test <- ifelse(prob.test > 0.5, 1, 0)

cat("Final Accuracy", mean(pred.test == heart_test$target))
```

Test Data Accuracy is .8166. This is technically data leakage as we are doing feature and model selection on the test set.

Appendix

```r
library(knitr)
# install the tidyverse library (do this once) install.packages('tidyverse')
library(tidyverse)
library(splines)

# set chunk and figure default options
knitr::opts_chunk$set(echo = FALSE, message = FALSE, warning = FALSE, fig.width = 4,
    fig.height = 4, tidy = TRUE)
## R code for problem 4a
titanic = read.csv("titanic.csv")

glmOut <- glm(survived ~ factor(passenger_class) + gender + ns(age, 3) + ns(fare_paid,
    3), data = titanic, family = binomial)

summary(glmOut)

# R code for problem 4b
new.df <- data.frame(passenger_class = 2, gender = "woman", age = 25, fare_paid = 20)

prob.survive <- predict(glmOut, newdata = new.df, type = "response")
prob.survive

# R code for problem 4c
new.df <- data.frame(passenger_class = 3, gender = "man", age = 15, fare_paid = 10)
```

```r
prob.survive <- predict(glmOut, newdata = new.df, type = "response")
prob.survive

## R code for problem 5a
library(ISLR)
library(reshape2)  # For reshaping data

ggplot(Weekly, aes(x = Volume)) + geom_histogram(binwidth = 0.1, fill = "blue", color = "black",
    alpha = 0.7) + theme_minimal() + ggtitle("Distribution of Trading Volume") +
    xlab("Volume") + ylab("Frequency")

library(corrplot)
cor_matrix <- cor(Weekly[, c("Lag1", "Lag2", "Lag3", "Lag4", "Lag5", "Volume", "Today")])

corrplot(cor_matrix, method = "color", type = "upper", tl.col = "black", tl.srt = 45,
    addCoef.col = "black", number.cex = 0.8)

# R code for problem 5b
glm.weekly <- glm(Direction ~ Lag1 + Lag2 + Lag3 + Lag4 + Lag5 + Volume, data = Weekly,
    family = binomial)
summary(glm.weekly)

# R code for problem 5c
prob <- predict(glm.weekly, type = "response")
pred <- ifelse(prob > 0.5, "Up", "Down")

table(Predicted = pred, Actual = Weekly$Direction)
mean(pred == Weekly$Direction)  # fraction of correct predictions

# R code for problem 5d
train <- (Weekly$Year <= 2009)
Weekly.train <- Weekly[train, ]
Weekly.test <- Weekly[!train, ]

glm.train <- glm(Direction ~ Lag2, data = Weekly.train, family = binomial)
prob.test <- predict(glm.train, Weekly.test, type = "response")
pred.test <- ifelse(prob.test > 0.5, "Up", "Down")

table(Predicted = pred.test, Actual = Weekly.test$Direction)
mean(pred.test == Weekly.test$Direction)

# R code for problem 7


heart_train <- read.csv("heart_training.csv")
heart_test <- read.csv("heart_test.csv")

glm.final <- glm(target ~ age + slope + thalach, data = heart_train, family = binomial)
summary(glm.final)
prob.test <- predict(glm.final, heart_test, type = "response")
pred.test <- ifelse(prob.test > 0.5, 1, 0)

cat("Final Accuracy", mean(pred.test == heart_test$target))
```