

# Assignment 2: Titanic Data

*Shane Weisz*

*29/08/2019*

## Contents

<b>Question 1</b>	<b>1</b>
(a) Mean and Variance of a Bernoulli Random Variable . . . . .	1
(b) Derivation of IWLS Components . . . . .	2
(c) R-Function Executing the IWLS Updating Equation . . . . .	3
(d) Verification of glm_fit() Producing Valid Estimates . . . . .	3
(e) Odds of Survival of 1st and 2nd vs 3rd Class Male Passengers Aged 25 . . . . .	4
(f) Model Building of Logistic Regression Model of the Data . . . . .	5
(g) Tree-based Methods . . . . .	7
(i) Fitting a Default Classification Tree . . . . .	7
(ii) Interpretation of Fitted Tree . . . . .	8
(h) Predictions on Test Dataset . . . . .	9
<b>Appendix: R-Code</b>	<b>9</b>

## Question 1

This assignment involves an investigation into a dataset regarding survival logs for passengers aboard the Titanic.

### (a) Mean and Variance of a Bernoulli Random Variable

We first calculate the mean and variance of a random variable  $Y$  that has the Bernoulli distribution.

Let  $Y \sim Be(\pi), \pi \in [0, 1]$ .

Then the density function for  $Y$  is

$$f(y, \pi) = \pi^y(1 - \pi)^{1-y},$$

which can be written as

$$\begin{aligned} f(y, \pi) &= \exp(y \log(\frac{\pi}{1 - \pi}) + \log(1 - \pi)) \\ &= \exp(a(y)b(\pi) + c(\pi) + d(y)), \end{aligned}$$

where

$$\begin{aligned} a(y) &= y \\ b(\pi) &= \log\left(\frac{\pi}{1-\pi}\right) \\ c(\pi) &= \log(1-\pi) \\ d(y) &= 0. \end{aligned}$$

Thus  $Y$  is from the exponential family of distributions, and so we can calculate the expectation and variance of  $Y$  using the appropriate results.

It follows that:

$$E[Y] = -\frac{c'(\pi)}{b'(\pi)} = -\frac{-\frac{1}{1-\pi}}{\frac{1}{1-\pi}} = \pi$$

and, with appropriate simplification,

$$\text{Var}[Y] = -\frac{b''(\pi)c'(\pi) - c''(\pi)b'(\pi)}{[b'(\pi)]^3} = \frac{\frac{2\pi-1}{\pi^2(1-\pi)^2} \cdot -\frac{1}{1-\pi} + \frac{1}{(1-\pi)^2} \cdot \frac{1}{\pi(1-\pi)}}{\left[\frac{1}{\pi(1-\pi)}\right]^3} = \pi(1-\pi).$$

## (b) Derivation of IWLS Components

We first derive an expression for the weights  $w_{ii}$  in terms of  $\mu_i$ .

We know that  $w_{ii} = \frac{1}{\text{Var}[Y_i]} \left(\frac{\partial \mu_i}{\partial \eta_i}\right)^2$ .

For a Bernoulli random variable  $Y_i$  with mean  $\mu_i$ , from (a) we know that

$$\text{Var}[Y_i] = \mu_i(1 - \mu_i).$$

Now, we are given that the link function is  $g(\mu_i) = \eta_i = \text{logit}(\mu_i) = \log\left(\frac{\mu_i}{1-\mu_i}\right)$ . Rearranging, we get that  $\mu_i = \frac{e^{\eta_i}}{1+e^{\eta_i}}$ , from which we obtain

$$\frac{\partial \mu_i}{\partial \eta_i} = \frac{e^{\eta_i}}{(1 + e^{\eta_i})^2} = \mu_i(1 - \mu_i),$$

in terms of  $\mu_i$ .

It thus follows that

$$\begin{aligned} w_{ii} &= \frac{1}{\text{Var}[Y_i]} \left(\frac{\partial \mu_i}{\partial \eta_i}\right)^2 \\ &= \frac{1}{\mu_i(1 - \mu_i)} (\mu_i(1 - \mu_i))^2 \\ &= \mu_i(1 - \mu_i) \end{aligned}$$

Now, we also have that  $\mathbf{U}^{(m-1)}$  is the vector with elements

$$\begin{aligned}\mathbf{U}_j &= \sum_{i=1}^N \frac{(Y_i - \mu_i)}{\text{Var}[Y_i]} x_{ij} \left( \frac{\partial \mu_i}{\partial \eta_i} \right) \\ &= \sum_{i=1}^N \frac{(Y_i - \mu_i)}{\mu_i(1 - \mu_i)} x_{ij} (\mu_i(1 - \mu_i)) \\ &= \sum_{i=1}^N (Y_i - \mu_i) x_{ij},\end{aligned}$$

and so it follows that  $\mathbf{U}^{(m-1)}$  can be written as  $\mathbf{X}^T(\mathbf{y} - \boldsymbol{\mu})$  where  $\mathbf{y} = [y_1, \dots, y_n]^T$  is the vector of responses and  $\boldsymbol{\mu} = [\mu_1, \dots, \mu_n]^T$

### (c) R-Function Executing the IWLS Updating Equation

The function below uses the results calculated in (a) and (b) above to execute the Iterative Weighted Least Squares updating equation for a Bernoulli response variable using the logit link function.

```
glm_fit = function(Y, X, beta_start, k = 10)
{
  # Prepare elements for the updating procedure here:
  beta = matrix(beta_start, nrow = length(X[1,]), ncol = 1)
  W     = matrix(0, nrow = length(Y), ncol = length(Y))

  # Evaluate the updating equation:
  for(i in 2:k)
  {
    mu = exp(X %*% beta) / (1 + exp(X %*% beta))
    diag(W) = mu * (1 - mu)
    J = t(X) %*% W %*% X
    U = t(X) %*% (Y - mu)
    RHS = J %*% beta + U
    beta = solve(J, RHS)
  }
  # Process some of the results here:
  std_errors = diag(sqrt(solve(J)))
  # Return relevant content:
  return(list(estimates = beta, standard_errors = std_errors))
}
```

### (d) Verification of glm\_fit() Producing Valid Estimates

By comparing the estimates produced by R's `glm()` function to those produced by my `glm_fit()` function in the table below, we can verify that the estimates produced by `glm_fit()` are correct.

*Note that we have relevelled the factor variables such that 'PClass3' and 'male' are treated as the reference categories for PClass and Sex respectively.*

Table 1: Verification of Calculated Estimates

	R's glm()		My glm_fit()	
	Estimates	Std. Errors	Estimates	Std. Errors
(Intercept)	-0.6004694	0.3826258	-0.6004694	0.3826307
Sexfemale	1.5235214	0.3679527	1.5235214	0.3679529
Age	-0.0408113	0.0122555	-0.0408113	0.0122557
Pclass1	1.3664381	0.5588384	1.3664381	0.5588483
Pclass2	0.1312443	0.4673816	0.1312443	0.4673822
Fare	-0.0048846	0.0041761	-0.0048846	0.0041764
Sexfemale:Pclass1	2.8503559	0.9846157	2.8503561	0.9847822
Sexfemale:Pclass2	2.3713523	0.7645393	2.3713523	0.7645402

### (e) Odds of Survival of 1st and 2nd vs 3rd Class Male Passengers Aged 25

For the given model, we have that the odds are calculated as

$$\exp(B_0 + B_1x_i^{SEX} + B_2x_i^{AGE} + B_3x_i^{PClass1} + B_4x_i^{PClass2} + B_5x_i^{SEX \cdot PClass1} + B_6x_i^{SEX \cdot PClass2}),$$

with ‘PClass3’ and ‘male’ being treated as the reference categories for PClass and Sex respectively.

We first calculate the ratio of the odds of survival for a *First Class* male passenger aged 25 to the odds of *Third Class* male passengers of the same age as follows:

$$OR_{1st,3rd} = \exp(\beta_3) = 2.7432664.$$

This can be interpreted as First Class male passengers of the same age being 2.743 times as likely to survive than Third Class passengers.

An approximate 95% confidence interval for this odds ratio,  $OR_{1st,3rd}$ , is calculated as

$$\exp(\beta_3 \pm 1.96 \cdot se(\beta_3)) = \exp(1.0091493 \pm 1.96 \cdot 0.4739372),$$

yielding a confidence interval of (1.0835388, 6.9453078). Since this confidence interval does not contain 1, we can interpret this interval as suggesting that at the 95% confidence level, there is evidence in the data of a statistically significant higher probability of a male passenger surviving for a given age if they are a First Class passenger compared to if they were a Third Class passenger.

Similarly, we can calculate the ratio of the odds of survival for a *Second Class* male passenger aged 25 to the odds of *Third Class* male passengers of the same age

$$OR_{2nd,3rd} = \exp(B_4) = 1.0847294,$$

and obtain an approximate 95% confidence interval using

$$\exp(\beta_4 \pm 1.96 \cdot se(\beta_4)) = \exp(0.0813306 \pm 1.96 \cdot 0.4647138),$$

producing a confidence interval of (0.4362636, 2.6970798).

We can interpret these results as suggesting Second Class male passengers of the same age are 1.0847 times as likely to survive than Third Class passengers. However, since the appropriate confidence

interval contains 1, there is insufficient evidence at the 95% confidence level that a Second Class male for a given age has a higher likelihood of surviving compared to a Third Class male passenger of the same age.

In summary, the data suggests that at the 95% confidence level, First Class male passengers are more likely to survive than Third Class male passengers for a given age, however there is insufficient evidence at this confidence level that Second Class male passengers have a greater chance of surviving than Third Class male passengers with the same age.

## (f) Model Building of Logistic Regression Model of the Data

We now proceed to build an appropriate logistic regression model of the Titanic data, and identify significant variables for predicting survival of passengers on the Titanic.

We start by fitting a model using the logit link function with a single predictor, Sex. Sex has been chosen as the first predictor to include since the data shows that 73.88% of females survived as opposed to 17.59% of males - thus indicating that Sex is an excellent candidate for a predictor of survival for passengers aboard the Titanic. After fitting this model with just a sex term, we can calculate the difference in deviance between this model and a model with just a deviance, yielding a statistic  $\Delta D = 113.71$ , which is highly significant when compared to the chi-squared distribution with one degree of freedom. As a result, the data suggests that we should certainly include a sex term in our model. We shall call this model *Model 1*, that is

$$\text{Model 1 : } \textit{Survived} \sim \textit{Sex}.$$

To further guide the model building process, we will use some results based on Exploratory Data Analysis (EDA) conducted to gain insights into the patterns and relationships in the data. Firstly, we observed that the proportion of survivors are highest for 1st Class passengers, then 2nd Class passenger, and lowest for 3rd Class passengers, which suggests that passenger class may be a useful predictor of survival. Similarly, we note that 40.55% of passengers below the age of 50 survived, compared to just 25.8% of passengers over 50 - which also suggests that age may be an important predictor in predicting survival. The proportions of survivors from the different embarking ports are relatively similar though, which suggests this may carry the least information regarding survival probabilities, which seems plausible.

We now conduct a statistical approach of deciding which predictors are significant and should be included in the model, with the context provided by the above data exploration as guidance. We currently have a model that only contains a sex term (*Model 1*), so we now consider which further predictors should be added to the model. We fit five different models where we individually add one of sex, age, passenger class, fare and embarked respectively, and consider which of these models yields the greatest reduction in deviance from *Model 1*, and whether this reduction is statistically significant. The result is that including a passenger class term results in the largest reduction in deviance from *Model 1*, with a difference of  $\Delta D = 19.859$  which yields an extremely small p-value ( $< 0.00005$ ) when compared to the chi-squared distribution with 2 degrees of freedom (due to the model having 2 extra parameters). This finding coincides with what we expected from our EDA, that is, that passenger class is a significant term in predicting survival of passengers. As such, we add passenger class as a term in our model to form *Model 2*,

$$\text{Model 2 : } \textit{Survived} \sim \textit{Sex} + \textit{PassengerClass}.$$

We then employ the same approach as above to determine if any further individual predictors should be added to our model. The results of testing the appropriate deviance statistics from adding individual terms (one of age, fare and embarked port respectively) to the model, yields that age has a statistically significant effect in predicting survival, with a small p-value of 0.001984 arising from the appropriate chi-squared test on the difference in deviances compared to *Model 2*. However, adding further individual terms (fare or embarked port) can be seen to not have a significant reduction in deviance statistics, which leaves us with *Model 3* below,

$$\text{Model 3 : } \textit{Survived} \sim \textit{Sex} + \textit{PassengerClass} + \textit{Age}.$$

Before continuing further, we notice that the currently fitted model, *Model 3*, captures the terms that our EDA had hinted at being useful in predicting survival.

We now consider adding interaction terms to the model. Using the same technique as above in determining which terms would be significant additions to the model (fitting models with one of a sex & passenger class, sex & age, and age and passenger class interaction term), we find that first adding an interaction term for sex and passenger class, and then for sex and age, yields significant decreases in deviances compared to the nested models. However, adding a final interaction term for age and passenger class does not result in a significant improvement to the model, with a change in deviance of  $\Delta D = 3.6845$  not being deemed significant in comparison to a chi-squared distribution with 2 degrees of freedom - with an associated p-value of 0.1448. As such, we now have fitted a model *Model 4* which appears as follows,

$$\text{Model 4 : } \textit{Survived} \sim \textit{Sex} + \textit{PassengerClass} + \textit{Age} + \textit{Sex} \times \textit{PassengerClass} + \textit{Sex} \times \textit{Age}.$$

Lastly, we now consider different choices for link functions. For logistic regressions, the predominant link functions chosen from are typically the logit function, the probit function and the complementary log-log function. We fit *Model 4* with each of these respective link functions, and compare AICs. We see that using the complementary log-log function in fact yields the lowest AIC, with an AIC value of 321.295 compared to 322.235 and 321.616 respectively for the probit and logit functions. As such, we will opt to use the complementary log-log function as the link function for our model.

The final model chosen thus has the following structure:

$$\text{Final Model : } \textit{Survived} \sim \textit{Sex} + \textit{PassengerClass} + \textit{Age} + \textit{Sex} \times \textit{PassengerClass} + \textit{Sex} \times \textit{Age}.$$

The corresponding model (fitted using the complementary log-log link function) can be described more mathematically as follows:

$$\begin{aligned} g(\mu_i) &= \eta_i \\ &= x_i^T \beta \\ &= \beta_0 + \beta_1 x_i^{\textit{Sex}} + \beta_2 x_i^{\textit{PClass1}} + \beta_3 x_i^{\textit{PClass2}} + \beta_4 x_i^{\textit{Age}} + \beta_5 x_i^{\textit{Sex} \times \textit{PClass1}} + \beta_6 x_i^{\textit{Sex} \times \textit{PClass2}} + \beta_7 x_i^{\textit{Sex} \times \textit{Age}}, \end{aligned}$$

where

$$\beta_0 = -0.37270$$

$$\beta_1 = 0.56459$$

$$\beta_2 = 1.23395$$

$$\beta_3 = 0.12041$$

$$\beta_4 = -0.05444$$

$$\beta_5 = 1.87860$$

$$\beta_6 = 2.09530$$

$$\beta_7 = 0.04400$$

and  $g(\mu_i)$  is the complementary log-log link function.

We can interpret the coefficients of the model to gather insights into the nature of the relationship between the predictors and survival of the passengers. The negative coefficient for age indicates that the older the passenger, the smaller their chance of survival. The coefficients for sex and passenger classes suggest that females have a higher chance of survival than males, and being in 1st or 2nd Class results in a higher predicted survival probability than being in 3rd Class.

One of our key questions of interest was if socio-economic status had any bearing on the likelihood of survival for male passengers. Given that it was deemed necessary to add an interaction term for sex and passenger class due to this addition resulting in a statistically significant reduction in model deviance, it appears that there *is* a relationship between socio-economic status and the probability of male passengers surviving. We can interpret the coefficients of the interaction terms for sex and passenger class to gain insight into the nature of this relationship, and in so doing, answer our question of interest. From the coefficient of 1.8786 for the sex and 1st Class passengers interaction term, and 2.0953 for the sex and 2nd Class passengers interaction term (compared to just 0.56459 for female passengers compared to male passengers in isolation), we can clearly see that socio-economic status had a large effect on the likelihood of survival for male passengers. 1st and 2nd Class female passengers clearly had a much greater survival probability than males from these socio-economic classes (over-and-above the effect accounted for just by sex in isolation), which hence provides us with an answer to our question of interest.

## **(g) Tree-based Methods**

### **(i) Fitting a Default Classification Tree**

A default classification tree for the data has been fitted and plotted below.

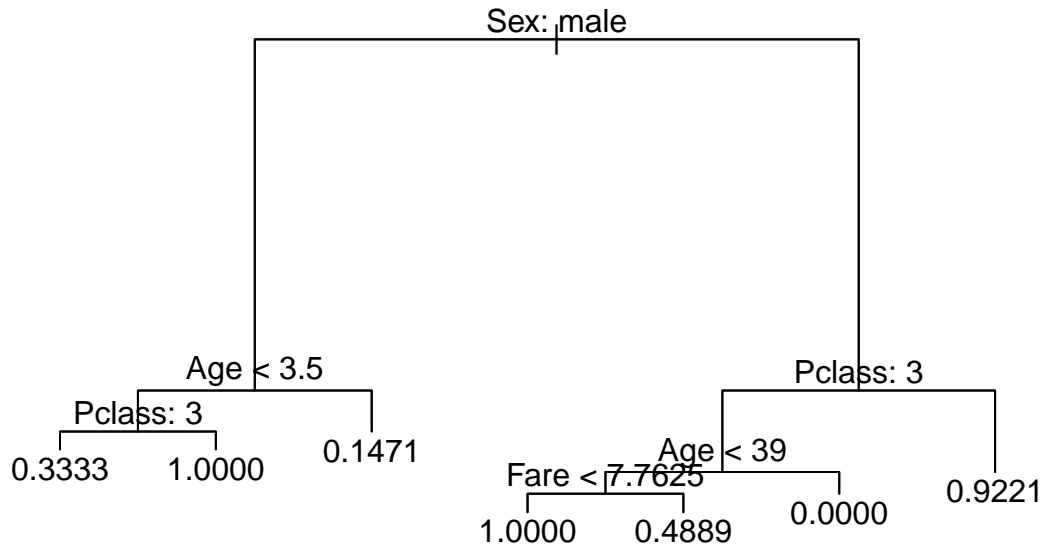


Figure 1: Default Classification Tree for Survival of Passengers

## (ii) Interpretation of Fitted Tree

By interpreting the appropriate nodes and branches of the tree, we can show why the tree plotted above makes sense. The first split is made on sex, with males along the left side of the tree and females on the right. This matches our previous findings that accounting for sex will lead to the greatest reduction in RSS (due to females being more likely to survive than males). Now, analyzing the predictions at some of the terminal nodes, we can see whether the predictions make sense in relation to our earlier observations in the report. We observe that the model predicts that males over the age of 3.5 years old will not survive (a predicted survival response of 0.1471) - this matches our intuition and earlier findings about the likelihood of survival for males. We can also see that females from the 1st class or 2nd class are predicted to survive with high likelihood (0.9221) - this certainly corresponds to what we found earlier in that both females and being in 1st or 2nd class increases the probability of survival.

However, by considering some of the predictors involved in lower tree cuts, it would appear that some of these do not seem relevant based on what our intuition and reasoning suggests, but rather could be seen as an example of *overfitting*. For example, the node for 'Age < 3.5' appears to be highly specific to a small subset of the training data, and thus it seems plausible that this node is evidence of the model being fitted to the training data too well, that is, is an example of overfitting. The region corresponding to 'Age < 3.5' being further split into subregions by passenger class seems to be further evidence supporting our inclination that the model has been overfitted to the training data. Similarly, having a specific terminal node corresponding to females in 3rd class with an age less than 39 and a fare less than 7.7625, appears too particular to the training data and thus also potentially an indication of the presence of overfitting.



## (h) Predictions on Test Dataset

The final GLM model we fitted in part (f) was a logistic regression model with a complimentary log-log link function with the following predictors:

$$\text{Survived} \sim \text{Sex} + \text{Pclass} + \text{Age} + \text{Sex} \cdot \text{Pclass} + \text{Sex} \cdot \text{Age}.$$

If we use this model to make predictions on the test data set, we see that the model correctly predicts whether or not a passenger survived for **279** out of the **350** observations. This equates to a prediction accuracy of the model with respect to survival of **0.7971429**.

Now, in comparison, if we use the default tree model fitted in part (g) to make predictions on the test data set, we see that this model correctly predicts whether or not a passenger survived for **278** out of the **350** observations. This equates to a prediction accuracy of the model with respect to survival of **0.7942857**.

Thus the default tree model performs slightly worse than the chosen GLM model when evaluated on the test dataset.

## Appendix: R-Code

```
# a)
rm(list = ls(all = TRUE))
dat = read.table('Assignment_2_Titanic_SetA.txt', h = TRUE)

# c)
glm_fit = function(Y, X, beta_start, k = 10)
{
  # Prepare elements for the updating procedure here:
  beta = matrix(beta_start, nrow = length(X[1,]), ncol = 1)
  W     = matrix(0, nrow = length(Y), ncol = length(Y))

  # Evaluate the updating equation:
  for(i in 2:k)
  {
    mu = exp(X %*% beta) / (1 + exp(X %*% beta))
    diag(W) = mu * (1 - mu)
    J = t(X) %*% W %*% X
    U = t(X) %*% (Y - mu)
    RHS = J %*% beta + U
    beta = solve(J, RHS)
  }
  # Process some of the results here:
  std_errors = diag(sqrt(solve(J)))
  # Return relevant content:
  return(list(estimates = beta, standard_errors = std_errors))
}
```

```

# d)
# Run R's glm and obtain results
dat$Sex = factor(dat$Sex)
dat$Pclass = factor(dat$Pclass)

# relevel - better for the odds ratios
dat$Pclass = relevel(dat$Pclass, "3")
dat$Sex = relevel(dat$Sex, "male")

fit = glm(Survived ~ Sex + Age + Pclass + Fare + Sex*Pclass, family = binomial(link = "logit"))
check_estimates = fit$coefficients
check_std_errors = summary(fit)$coefficients[,2] # standard errors stored in 2nd column of summary

# Run glm_fit and obtain results
beta_start = rep(0, 6)
Y = dat$Survived
X = model.matrix(fit)
my_output = glm_fit(Y, X, beta_start)
my_estimates = unlist(my_output["estimates"])
my_std_errors = unlist(my_output["standard_errors"])

# Create table of results
library(knitr)
library(kableExtra)
col_headings = rep(c("Estimates", "Std. Errors"),2)
results_mtx = matrix(c(check_estimates, check_std_errors, my_estimates, my_std_errors), ncol = 4)
colnames(results_mtx) = col_headings
rownames(results_mtx) = names(fit$coefficients)
kable(results_mtx, format = 'latex', booktabs = T, align="c", escape = F, caption = "Verification of results")

# e)
fitOR = glm(Survived ~ Sex + Age + Pclass + Sex*Pclass, family = binomial(link = "logit"), data = dat)

beta3 = fitOR$coef['Pclass1']
OR1 = exp(beta3)
OR1

beta4 = fitOR$coef['Pclass2']
OR2 = exp(beta4)
OR2

se3 = sqrt(vcov(fitOR)['Pclass1','Pclass1'])
conf1 = exp(beta3 + c(-1,1)*1.96*se3)
conf1

```

```

se4 = sqrt(vcov(fitOR)['Pclass2','Pclass2'])
beta4
conf2 = exp(beta4 + c(-1,1)*1.96*se4)
conf2

# f)

# EDA
library(ggplot2)
c(mean(dat$Survived[dat$Sex == 'female']),
  (mean(dat$Survived[dat$Sex == 'male'])))

c(mean(dat$Survived[dat$Parch < 2]),
  (mean(dat$Survived[dat$Parch >= 2])))

pclass_means = c(mean(dat$Survived[dat$Pclass == 1]),
                  mean(dat$Survived[dat$Pclass == 2]),
                  mean(dat$Survived[dat$Pclass == 3]))
pclass_means

unique(dat$Embarked)
em_means = c(mean(dat$Survived[dat$Embarked== 'S']),
              mean(dat$Survived[dat$Embarked == 'C']),
              mean(dat$Survived[dat$Embarked == 'Q']))
em_means

num50      = length((dat$Age >= 50)[(dat$Age >= 50) == TRUE])
num50
age_means = c(mean(dat$Survived[dat$Age <= 50]),
              mean(dat$Survived[dat$Age >= 50]))
age_means

# MODEL BUILDING
model_intercept = glm(Survived ~ 1, family = binomial(link = "logit"), data = dat)
model1 = glm(Survived ~ Sex, family = binomial(link = "logit"), data = dat)
anova(model_intercept, model1, test = "Chisq")

model2a = glm(Survived ~ Sex + Pclass, family = binomial(link = "logit"), data = dat)
model2b = glm(Survived ~ Sex + Age, family = binomial(link = "logit"), data = dat)
model2c = glm(Survived ~ Sex + Fare, family = binomial(link = "logit"), data = dat)
model2d = glm(Survived ~ Sex + Parch, family = binomial(link = "logit"), data = dat)
model2e = glm(Survived ~ Sex + Embarked, family = binomial(link = "logit"), data = dat)
AIC(model1, model2a, model2b, model2c, model2d, model2e)
anova(model1, model2a, test = "Chisq")
anova(model1, model2b, test = "Chisq")
anova(model1, model2c, test = "Chisq")
anova(model1, model2d, test = "Chisq")

```

```

anova(model1, model2e, test = "Chisq")

model2 = model2a
model2i = glm(Survived ~ Sex + Pclass, family = binomial(link = "probit"), data = dat)
model2ii = glm(Survived ~ Sex + Pclass, family = binomial(link = "cloglog"), data = dat)
AIC(model2, model2i, model2ii)
summary(model2)

model3a = glm(Survived ~ Sex + Pclass + Age, family = binomial(link = "logit"), data = dat)
model3b = glm(Survived ~ Sex + Pclass + Fare, family = binomial(link = "logit"), data = dat)
model3c = glm(Survived ~ Sex + Pclass + Parch, family = binomial(link = "logit"), data = dat)
model3d = glm(Survived ~ Sex + Pclass + Embarked, family = binomial(link = "logit"), data = dat)
AIC(model2, model3a, model3b, model3c, model3d)
anova(model2, model3a, test = "Chisq")
anova(model2, model3b, test = "Chisq")
anova(model2, model3c, test = "Chisq")
anova(model2, model3d, test = "Chisq")
model3 = model3a

model4a = glm(Survived ~ Sex + Pclass + Age + Fare, family = binomial(link = "logit"), data = dat)
model4b = glm(Survived ~ Sex + Pclass + Age + Parch, family = binomial(link = "logit"), data = dat)
model4c = glm(Survived ~ Sex + Pclass + Age + Embarked, family = binomial(link = "logit"), data = dat)
AIC(model3, model4a, model4b, model4c)
anova(model3, model4a, test = "Chisq")
anova(model3, model4b, test = "Chisq")
anova(model3, model4c, test = "Chisq")
model4 = model3

model5a = glm(Survived ~ Sex + Pclass + Age + Sex*Pclass, family = binomial(link = "logit"), data = dat)
model5b = glm(Survived ~ Sex + Pclass + Age + Sex*Age, family = binomial(link = "logit"), data = dat)
model5c = glm(Survived ~ Sex + Pclass + Age + Pclass*Age, family = binomial(link = "logit"), data = dat)
AIC(model4, model5a, model5b, model5c)
anova(model4, model5a, test = "Chisq")
anova(model4, model5b, test = "Chisq")
anova(model4, model5c, test = "Chisq")
model5 = model5a

model6a = glm(Survived ~ Sex + Pclass + Age + Sex*Pclass + Sex*Age, family = binomial(link = "logit"), data = dat)
model6b = glm(Survived ~ Sex + Pclass + Age + Sex*Pclass + Pclass*Age, family = binomial(link = "logit"), data = dat)
AIC(model5, model6a, model6b)
anova(model5, model6a, test = "Chisq")
anova(model5, model6b, test = "Chisq")
model6 = model6a

model7a = glm(Survived ~ Sex + Pclass + Age + Sex*Pclass + Sex*Age + Pclass*Age, family = binomial(link = "logit"), data = dat)
AIC(model6, model7a)
anova(model6, model7a, test = "Chisq")

```

```

model7 = model6

model8a = glm(Survived ~ Sex + Pclass + Age + Sex*Pclass + Sex*Age, family = binomial(link = "logit"))
model8b = glm(Survived ~ Sex + Pclass + Age + Sex*Pclass + Sex*Age, family = binomial(link = "logit"))
AIC(model7, model8a, model8b)
model8 = model8b

final_model = model8
summary(final_model)

# g)
# TREE BASED METHODS

library(tree)
# Fit the default model
tree_model = tree(Survived ~ ., data = dat)
# tree_model

# Plot the tree with text
plot(tree_model)
text(tree_model, pretty = 0)

# h)
# PREDICTIONS

glm_model = glm(Survived ~ Sex + Pclass + Age + Sex*Pclass + Sex*Age, family = binomial(link = "logit"))
tree_model = tree(Survived ~ ., data = dat)

test_data = read.table('Assignment_2_Titanic_SetB.txt', h = TRUE)
N = length(test_data[,1])

test_data$Sex = factor(test_data$Sex)
test_data$Pclass = factor(test_data$Pclass)
test_data$Pclass = relevel(test_data$Pclass, "3")
test_data$Sex = relevel(test_data$Sex, "male")

# Make GLM predictions
prediction_probs_glm = predict(glm_model, newdata = test_data, type="response")
predictions_glm = round(prediction_probs_glm) # change from probabilities to 0s and 1s

# Make Tree predictions
prediction_probs_tree = predict(tree_model, newdata = test_data)
predictions_tree = round(prediction_probs_tree) # change from probabilities to 0s and 1s

# View first 20 predictions vs test set
# cbind(head(test_data$Survived, 20), head(predictions_glm, 20), head(predictions_tree, 20))

```

```

# GLM Prediction Accuracy
tf_vector_glm      = test_data$Survived == predictions_glm
# tf_vector_glm    = dat$Survived == round(fitted.values(glm_model)) # fitted values on o
num_correct_glm    = length(tf_vector_glm[tf_vector_glm == TRUE])
prediction_accuracy_glm = num_correct_glm/N

# Tree Prediction Accuracy
tf_vector_tree     = test_data$Survived == predictions_tree
# tf_vector        = dat$Survived == round(fitted.values(tree_model)) # fitted values on o
num_correct_tree   = length(tf_vector_tree[tf_vector_tree == TRUE])
prediction_accuracy_tree = num_correct_tree/N

c("GLM Accuracy" = prediction_accuracy_glm,
  "Tree Accuracy" = prediction_accuracy_tree)

pred = data.frame(cbind(predictions_glm, predictions_tree))
write.table(pred, 'Titanic_Pred_WSZSHA001.csv', quote = F, row.names = F, sep = ',')

```