Linear Models

Alex Fitz

Contents


```
if (!require(fitdistrplus))
  {install.packages("fitdistrplus")}
  library(fitdistrplus)

if (!require(lmtest))
  {install.packages("lmtest")}
  library(lmtest)

if (!require(titanic))
  {install.packages("titanic")}
  library(titanic)
```

Section 1: Linear Regression

Question 1

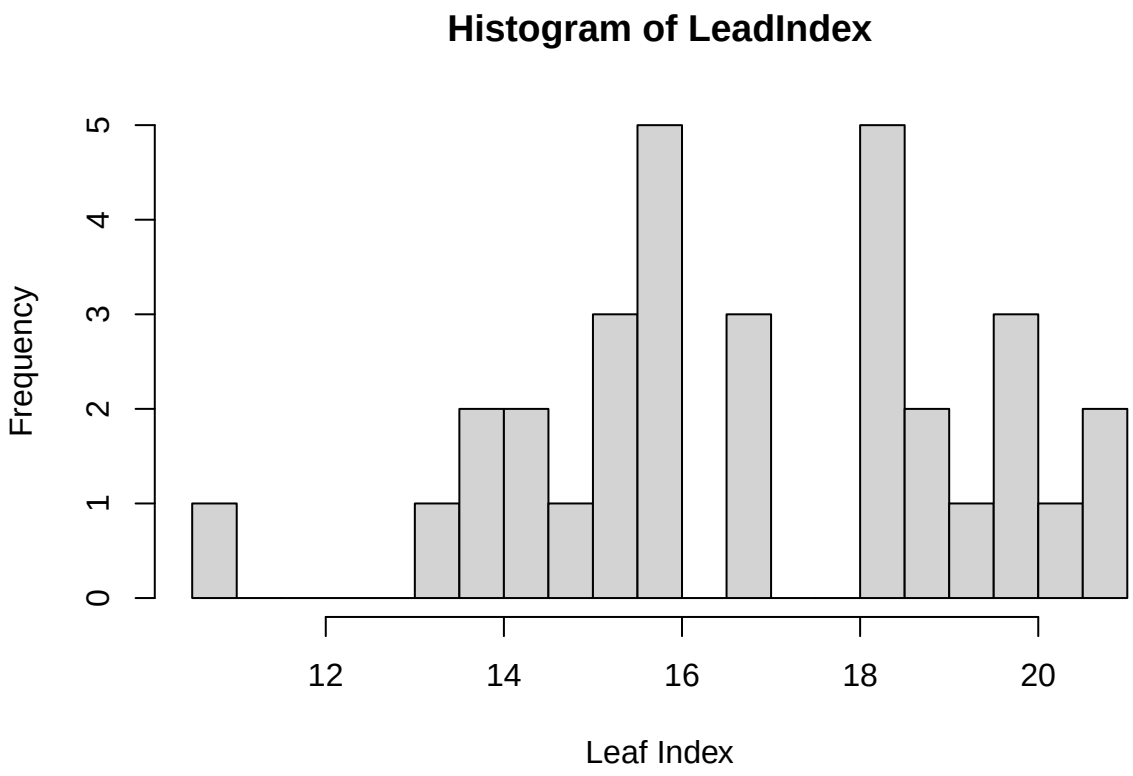
In a study of genetic variation in sugar maple, seeds were collected from native trees in the eastern United States and Canada and planted in a nursery in Wooster, Ohio. The time of leafing out of these seedlings can be related to the latitude and mean July temperature of the place of origin of the seed. The variables are X1 = latitude, X2 = July mean temperature, and Y = weighted mean index of leafing out time. (Y is a measure of the degree to which the leafing out process has occurred. A high value is indicative that the leafing out process is well advanced.) The data is in the file maple.txt.

Part A Question: Read the data into R and save it to a variable called "maple". Make a histogram of the variable LeafIndex. Use qqnorm, qqline, and shapiro.test to assess the normality of this variable. Is the data approximately normally distributed?

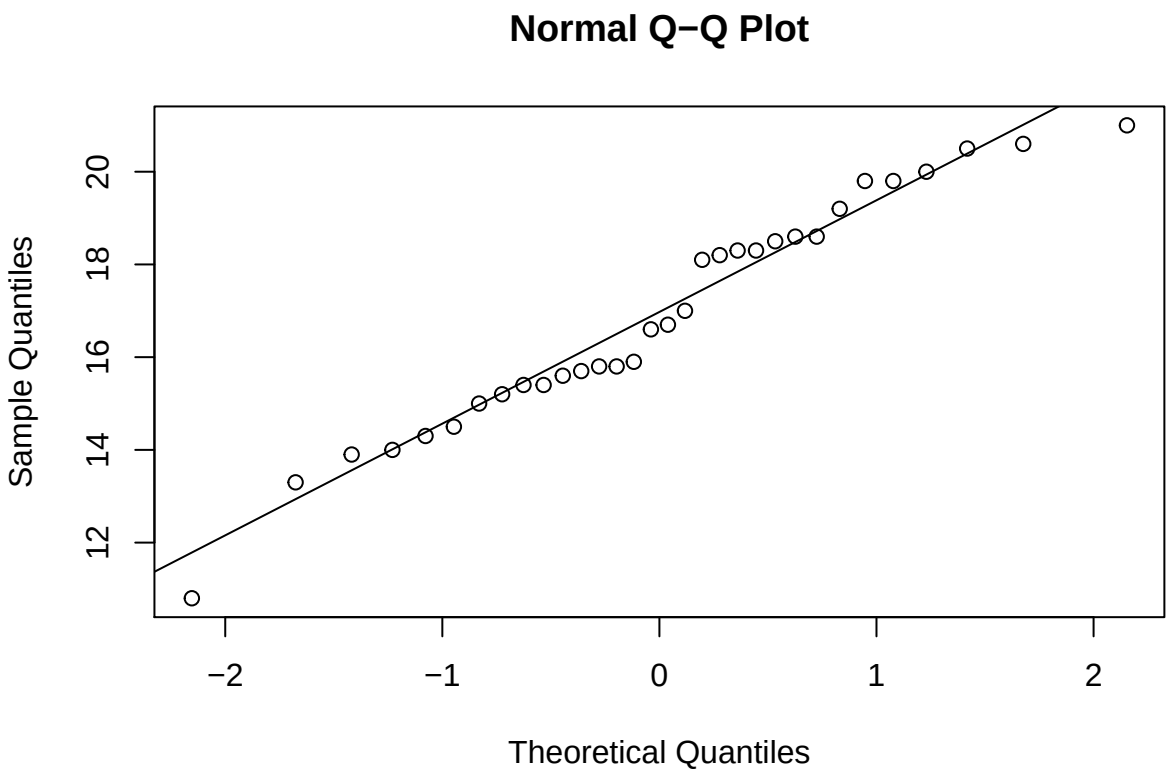
```
dir <- "~/Documents/GitHub/code_examples/R/RMarkdown Examples"
maple <- read.table(paste(dir,"/maple.txt",sep=""), header = T)
head(maple)
```

```
##     Location Latitude JulyTemp LeafIndex
## 1     Quebec    46.95     66.7      20.5
## 2 NovaScotia    45.54     64.3      21.0
## 3 NovaScotia    45.00     61.8      18.3
## 4      Maine    45.42     65.0      18.6
## 5    Vermont    44.50     63.5      18.6
## 6   Michigan    46.37     65.5      19.8
```

```
hist(maple$LeafIndex,freq=T, xlab="Leaf Index", breaks = 30, main = "Histogram of LeadIndex")
```




```
{qqnorm(maple$LeafIndex)
qqline(maple$LeafIndex)}
```



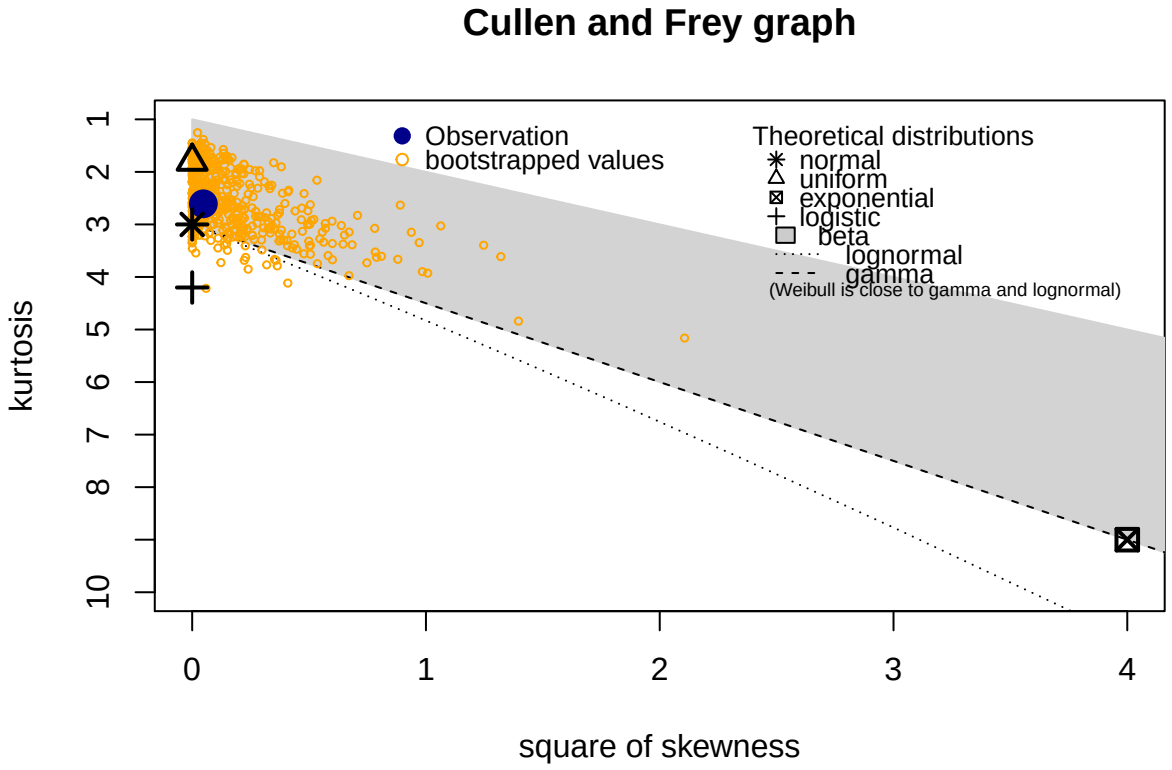

```
shapiro.test(maple$LeafIndex)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  maple$LeafIndex
## W = 0.96576, p-value = 0.3913
```

Here we have a null hypothesis that the distribution is normal, and we get a pvalue of 0.39, so we fail to reject our null hypothesis and the data is not normally distributed.

Part B **Question:** Load the library "fitdistrplus" and use the command descdist to evaluate the distribution of LeafIndex and save the results to a variable called "fit.norm". What value will you fill in for the function input argument "discrete"? Why? Please use boot = 500 (to calculate 500 bootstrap values of the skewness and kurtosis of LeafIndex observations. A bootstrap samples values from the observations with replacement, meaning that an observation can be chosen more than once. This is a technique to assess variance in our estimate of skewness and kurtosis).

```
#Library/Installation is at beginning of script
fit.norm<-descdist(maple$LeafIndex, discrete = FALSE, boot=500)
```



Discrete should be false for this because the response variable (LeafIndex) is continuous.

Part C **Question:** The descdist results provide a guideline about the distribution of our data. According to the plot that is output after using this command, two distributions (out of normal, uniform, exponential, logistic, beta, lognormal, and gamma) are good candidates for LeafIndex: normal and beta. Please look up the beta distribution on Wikipedia (which has excellent entries for probability distributions):

And tell me – what is the interval or range of values that a beta distributed variable can take? Please write a short argument here about why a beta distribution is unlikely for LeafIndex.

Answer: Beta distribution is defined on the interval [0,1] parametrized by positively shaped distributions. The LeafIndex is not bounded from [0,1] while the beta distribution is, so it is very unlikely for it to take a beta distribution.

Part D **Question:** Now that we are more confident about the normality of our response variable, we proceed with regression. Regress LeafIndex on Latitude (using lm) and save the results to a variable called "mod1".

```
mod1<-lm(LeafIndex~Latitude,data=maple)
summary(mod1)
```

```
##
## Call:
## lm(formula = LeafIndex ~ Latitude, data = maple)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.2348 -0.8488  0.0773  1.0074  3.3305
```

```
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.66716    3.05202  -0.546    0.589
## Latitude     0.45369    0.07427   6.108 1.03e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.673 on 30 degrees of freedom
## Multiple R-squared:  0.5543, Adjusted R-squared:  0.5394
## F-statistic: 37.31 on 1 and 30 DF,  p-value: 1.031e-06
```

Part E **Question:** Is Latitude a useful predictor of leaf index? Give the model estimate of the regression beta coefficient and the standard error around that estimate – do not just write that out. To get full credit, please index this information either from mod1 or by saving summary(mod1) to a new variable and indexing that variable. Does LeafIndex increase or decrease with Latitude?

```
s_mod1<-summary(mod1)
s_mod1$coefficients[,1:2]
```

```
##               Estimate Std. Error
## (Intercept) -1.6671556 3.05202062
## Latitude     0.4536868 0.07427497
```

Here we have a positive beta coefficient and the standard error around that beta coefficient. This means that latitude is a useful predictor of leaf index. Since the beta coefficient is positive, leafindex increases with latitude.

Part F **Question:** Regress LeafIndex on JulyTemp (using lm) and save the results to a variable called "mod2". Is JulyTemp a useful predictor of leaf index? Give the model estimate of the regression beta coefficient and the standard error around that estimate. Does LeafIndex increase or decrease with JulyTemp?

```
mod2<-lm(LeafIndex ~ JulyTemp, data= maple)
s_mod2<-summary(mod2)
s_mod2$coefficients[,1:2]
```

```
##              Estimate Std. Error
## (Intercept) 40.742971  4.4549781
## JulyTemp    -0.333177  0.0620633
```

Looking at the JulyTemp regressed against LeafIndex, it is also a good indicator since it has a negative beta coeffient, and also a small standard error around that coefficient. Since the beta coefficient is negative that means that LeafIndex will decrease with JulyTemp.

Part G **Question:** Regress LeafIndex on JulyTemp and Latitude (using lm) and save the results to a variable called "mod3". Give the model estimates of the regression beta coefficients and the standard error around those estimates.

```
mod3<-lm(LeafIndex ~ JulyTemp + Latitude, data=maple)
s_mod3<-summary(mod3)
s_mod3$coefficients[,1:2]
```

```
##                Estimate  Std. Error
## (Intercept) 13.7318390 11.42026248
## JulyTemp    -0.1352401  0.09676374
## Latitude     0.3139276  0.12388008
```

Part H **Question:** Use AIC (command AIC) and likelihood ratio tests to determine a minimal adequate model. Why did you include or exclude a given predictor? Report the AIC values then report the results of the likelihood ratio test.

```
cat("AIC of mod1:", AIC(mod1),"\n")
```

```
## AIC of mod1: 127.6754
```

```
cat("AIC of mod2:", AIC(mod2),"\n")
```

```
## AIC of mod2: 131.9904
```

```
cat("AIC of mod3:", AIC(mod3))
```

```
## AIC of mod3: 127.5894
```

When examining the AIC's we can see that we get the lowest AIC for mod3, but that was very similar to mod1. Using AIC to measure the likelihood is good, because it penalizes the user for each predictor variable.

```
#Here we use mod3 as maximum likelihood model
cat("Likelihood Ratio (mod3-mod1):",pchisq(2 * (logLik(mod3) - logLik(mod1)), df = 1, lower.tail=FALSE)
```

```
## Likelihood Ratio (mod3-mod1): 0.1486588
```

```
cat("Likelihood Ratio (mod3-mod2):",pchisq(2 * (logLik(mod3) - logLik(mod2)), df = 1, lower.tail=FALSE)
```

```
## Likelihood Ratio (mod3-mod2): 0.0114056
```

After using the AIC's, we determined that mod1 and mod3 were both very good predictors of leadindex, only differing in their AIC by a very small number. Using the likelihod ratio test we could whether or not mod3 and mod1 differed from each other in approximately a chi-squared distribution. Here we see that there was a p-value of 0.148, thus meaning that mod1 does not differ from mod3 in a statistically significant way. For my minimum adequate model I would include only Latitude, because it is able to predict leafindex almost similarly to when Julytemp is included, but it is better since its only one predictor. So, mod1 is my minimum adequate model.

Part I **Question:** Construct a null model (and save that to a variable called "mod_null"), then use AIC and likelihood ratio tests to determine whether the minimal adequate model differs in log-likelihood from the null model. Is the minimal adequate model more or less likely than the null model, given our observed data?

```
mod_null<-lm(LeafIndex~1,data=maple)

cat("Mod_Null:", AIC(mod_null), "\n")
```

```
## Mod_Null: 151.535
```

```
cat("Mod3:", AIC(mod1))
```

```
## Mod3: 127.6754
```

This indicates that the AIC of the minimum adequate model (mod3) is lower than the null model, which indicates that the minimum adequate model is more likely to occur than the null model, given our observed data.

```
cat("Likelihood Ratio (Minimum Adequate Model-Null Model):",pchisq(2 * (logLik(mod1) - logLik(mod_null)),
```

```
## Likelihood Ratio (Minimum Adequate Model-Null Model): 3.671648e-07
```

Using the likelihood ratio test, we get a pvalue of 3.67e-07 when comparing the minimum adequate model to the null model in an approximately chi-square distribution. This means that the minimum adequate model differs from the null model in a statistically significant way.

Section 2: Generalized Linear Models

Question 2

Titanic was a British passenger liner that sank in 1912 after colliding with an iceberg. Only 31% of passengers survived in this disaster. Let's try to predict who was more likely to survive and why.

- Survived - Survival (0 = No; 1 = Yes).
- Pclass - Passenger Class (1 = 1st; 2 = 2nd; 3 = 3rd)
- Name - Name
- Sex - Sex
- Age - Age
- Sibsp - Number of Siblings/Spouses Aboard
- Parch - Number of Parents/Children Aboard
- Ticket - Ticket Number
- Fare - Passenger Fare
- Cabin - Cabin
- Embarked - Port of Embarkation (C = Cherbourg; Q = Queenstown; S = Southampton)

Part A Question: The response variable (Survived) is binomially distributed. We will use a logit-link function. Please fit a GLM (using glm) of Survived using Pclass, Fare, Age, Sex, SibSp, Parch, and Embarked as predictors. Save the results to a variable called "mod1".

```
#Package installed at beginning of script
data(titanic_train)

mod1<-glm(Survived ~ Pclass+Fare+Age+Sex+SibSp+Parch+Embarked, data=titanic_train,family=binomial)
summary(mod1)
```

```
##
## Call:
## glm(formula = Survived ~ Pclass + Fare + Age + Sex + SibSp +
##     Parch + Embarked, family = binomial, data = titanic_train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.7233  -0.6439  -0.3772   0.6288   2.4457
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)  17.894850 607.855474   0.029  0.97651
## Pclass       -1.199251   0.164619  -7.285 3.22e-13 ***
## Fare          0.001432   0.002531   0.566  0.57165
## Age          -0.043350   0.008232  -5.266 1.39e-07 ***
## Sexmale      -2.638476   0.222256 -11.871  < 2e-16 ***
## SibSp        -0.363208   0.129017  -2.815  0.00487 **
```

```
## Parch        -0.060270   0.123900  -0.486  0.62666
## EmbarkedC   -12.257443 607.855250  -0.020  0.98391
## EmbarkedQ   -13.080988 607.855453  -0.022  0.98283
## EmbarkedS   -12.658656 607.855228  -0.021  0.98339
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 964.52  on 713  degrees of freedom
## Residual deviance: 632.34  on 704  degrees of freedom
##   (177 observations deleted due to missingness)
## AIC: 652.34
##
## Number of Fisher Scoring iterations: 13
```

Part B **Question:** Save the results of summary(mod1) to a new variable called "res1". Give the model estimates of the regression beta coefficients and the standard error around those estimates (index res1 to give me this information).

```
res1<-summary(mod1)
res1$coefficients[,1:2]
```

```
##                 Estimate   Std. Error
## (Intercept)  17.894849676 6.078555e+02
## Pclass       -1.199250912 1.646190e-01
## Fare          0.001431586 2.530963e-03
## Age          -0.043349967 8.232089e-03
## Sexmale      -2.638476351 2.222565e-01
## SibSp        -0.363208369 1.290171e-01
## Parch        -0.060269766 1.239004e-01
## EmbarkedC   -12.257443068 6.078552e+02
## EmbarkedQ   -13.080987800 6.078555e+02
## EmbarkedS   -12.658656450 6.078552e+02
```

Part C **Question:** Use the step command to conduct a stepwise regression model reduction and save the results to a new variable called "mod2", then save the results of summary(mod2) to a new variable called "res2".

```
mod2<-step(mod1)
```

```
## Start:  AIC=652.34
## Survived ~ Pclass + Fare + Age + Sex + SibSp + Parch + Embarked
##
##            Df Deviance    AIC
## - Embarked  3   635.81 649.81
## - Parch     1   632.58 650.58
## - Fare      1   632.68 650.68
## <none>          632.34 652.34
## - SibSp     1   640.91 658.91
## - Age       1   662.75 680.75
## - Pclass    1   686.64 704.64
```

```
## - Sex       1   808.42 826.42
##
## Step:  AIC=649.81
## Survived ~ Pclass + Fare + Age + Sex + SibSp + Parch
##
##          Df Deviance    AIC
## - Parch   1   636.07 648.07
## - Fare    1   636.62 648.62
## <none>        635.81 649.81
## - SibSp   1   645.25 657.25
## - Age     1   667.62 679.62
## - Pclass  1   695.26 707.26
## - Sex     1   815.18 827.18
##
## Step:  AIC=648.07
## Survived ~ Pclass + Fare + Age + Sex + SibSp
##
##          Df Deviance    AIC
## - Fare    1   636.72 646.72
## <none>        636.07 648.07
## - SibSp   1   647.23 657.23
## - Age     1   667.86 677.86
## - Pclass  1   699.21 709.21
## - Sex     1   820.07 830.07
##
## Step:  AIC=646.72
## Survived ~ Pclass + Age + Sex + SibSp
##
##          Df Deviance    AIC
## <none>        636.72 646.72
## - SibSp   1   647.29 655.29
## - Age     1   669.44 677.44
## - Pclass  1   742.29 750.29
## - Sex     1   823.84 831.84
```

```
res2<-summary(mod2)
res2
```

```
##
## Call:
## glm(formula = Survived ~ Pclass + Age + Sex + SibSp, family = binomial,
##     data = titanic_train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.7714  -0.6445  -0.3836   0.6276   2.4585
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  5.600846   0.543441  10.306  < 2e-16 ***
## Pclass      -1.317398   0.140900  -9.350  < 2e-16 ***
## Age         -0.044385   0.008155  -5.442 5.26e-08 ***
## Sexmale     -2.623483   0.214524 -12.229  < 2e-16 ***
## SibSp       -0.376119   0.121080  -3.106  0.00189 **
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 964.52  on 713  degrees of freedom
## Residual deviance: 636.72  on 709  degrees of freedom
##   (177 observations deleted due to missingness)
## AIC: 646.72
##
## Number of Fisher Scoring iterations: 5
```

Part D **Question:** Construct a null model with no predictors and save the results of the glm of that null model to a new variable called "mod_null".

```
mod_null<-glm(Survived~1, data=titanic_train, family=binomial)
summary(mod_null)
```

```
##
## Call:
## glm(formula = Survived ~ 1, family = binomial, data = titanic_train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.9841  -0.9841  -0.9841   1.3839   1.3839
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -0.47329    0.06889   -6.87  6.4e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1186.7  on 890  degrees of freedom
## Residual deviance: 1186.7  on 890  degrees of freedom
## AIC: 1188.7
##
## Number of Fisher Scoring iterations: 4
```

Part E **Question:** Give me with a detailed numerical description of the model comparison results using AIC and likelihood tests (for mod1, mod2, and mod_null). (i) Tell me the minimal adequate model, (ii) whether that differs from the null model, (iii) using AIC and likelihood ratio tests. and (iv) tell me the estimate of the effects (beta coefficients) and give the standard error around that estimate.

```
cat("AIC (mod_null):", AIC(mod_null), "\n")
```

```
## AIC (mod_null): 1188.655
```

```
cat("AIC (mod1):", AIC(mod1),"\n")
```

```
## AIC (mod1): 652.343
```

```
cat("AIC (mod2):", AIC(mod2),"\n")
```

```
## AIC (mod2): 646.7193
```

When examining the AIC values, we see that both mod1 and mod2 have similar values, while the null model is significantly larger. Next, we will conduct the likelihood ratio test.

Here would would use mod1 as our maximum likelihood model, because it has the most predictors.

```
cat("Likelihood Ratio (Mod1-Mod2):",pchisq(2 * (logLik(mod1) - logLik(mod2)), df = 5, lower.tail=FALSE)
```

```
## Likelihood Ratio (Mod1-Mod2): 0.4966041
```

When examining the results, we see that there is a pvalue of 0.496 when comparing mod1, to mod2. This means that as we decrease the number of predictors to find our minimum adequate model, there was very little change in the difference of the models. This agrees with our calculated AIC values. So here mod2 (Survived ~ Pclass + Age + Sex + SibSp) is our minimum adequate model, as we would pick our model with the least amount of predictors because there is a penalty for more predictors along with it being harder for researchers to collect unecessary data. To show that our minimum adequate model differs from our null model, we calculate the pvalue given an chi-square approximate distribution.

```
cat("Likelihood Ratio (Mod2-Null Model):",pchisq(2 * (logLik(mod2) - logLik(mod_null)), df = 3, lower.ta
```

```
## Likelihood Ratio (Mod2-Null Model): 7.175315e-119
```

Here we can an extremely small pvalue, showing that the distributions are different.

```
#Beta Coefficients
res2$coefficients[,1:2]
```

```
##                Estimate  Std. Error
## (Intercept)  5.6008462 0.543441202
## Pclass      -1.3173981 0.140900143
## Age         -0.0443847 0.008155269
## Sexmale     -2.6234834 0.214524258
## SibSp       -0.3761192 0.121080490
```

Part F **Question:** Provide a verbal overview of the effects of each predictor in the minimal adequate model. How do you interpret the effect of each predictor? For this, use the sign of the beta coefficient (negative or positive) to describe how that predictor influences your probability of survival.

```
cbind(mod2$coefficients, confint(mod2)[,1], confint(mod2)[,2])
```

```
## Waiting for profiling to be done...
## Waiting for profiling to be done...
```

```
##                   [,1]        [,2]        [,3]
## (Intercept)  5.6008462  4.56877277  6.70182493
## Pclass      -1.3173981 -1.60095422 -1.04776701
## Age         -0.0443847 -0.06073701 -0.02871785
## Sexmale     -2.6234834 -3.05476185 -2.21256757
## SibSp       -0.3761192 -0.62152704 -0.14560851
```

First, above is a table of the beta coefficient, and an estimate of the effect given our confidence interval around the effect size.

When we examine the beta coefficient, we can find the odds ratio. Looking at class, which given our probability of survival means that those which were a lower class (1st) had a 13% higher chance of survival than (2nd) class. We can note that for Age there is very little relationship between survival as it is closely dispersed

around 0, but the slight relationship that exists is negatively associated with age, so every year older someone is means they are 0.4% less likely to survive. When looking at sex, we see that the beta coefficient is negative, so males are 26.2% less likely to survive. We also note a negative beta coefficient for the number of siblings, which means that for each sibling someone has, it means they are 3.7% less likely to survive.