# Day 3: GLMs for Binary Data Exercise

Load the `Titanic-Dataset.csv` file. This CSV file contains information about a subset of the passengers on the Titanic. This dataset has 12 variables:

- `PassengerId`: Passenger ID Number
- `Survived`: 1 = survived, 0 = did not survive
- `Pclass`: Passenger class
- `Name`
- `Sex`
- `Age`
- `SibSp`: Number of siblings / spouses aboard the Titanic
- `Parch`: Number of parents / children aboard the Titanic
- `Ticket`: Ticket number
- `Fare`: Fare paid for ticket
- `Cabin`: Cabin number
- `Embarked`: Embarkation point - "S" = Southampton, England; "C" = Cherbourg, France; "Q" = Queenstown, Ireland

We are interested in better understanding which factors may have influence whether someone survived the disaster. Fit a logistic regression model to predict each passenger's survival status, provide an interpretation of the coefficients included in the model, and assess the performance of the model. Next, fit a probit regression model, and compare its performance to the results of the logistic regression model. Which model would you choose to use? Which variables were the most important for predicting survival? Does this make sense to you?

```r
knitr::opts_chunk$set(message = FALSE)
knitr::opts_chunk$set(warning = FALSE)
```

```r
library(tidyverse)
library(pROC)
```

```r
titanic <- read_csv("Titanic-Dataset.csv") %>%
  mutate(
    Survived = as.factor(Survived),
    Pclass = as.factor(Pclass),
    Sex = as.factor(Sex),
    Embarked = as.factor(Embarked)
  )
```

## Data Exploration

First, we note that the `PassengerId` and `Name` variables should have no direct impact on the survival status of their respective passengers. Ticket number will likely be difficult to analyze as well. We will exclude them from the analysis. A closer look at the `Cabin` variable shows that nearly all cabin numbers have the form of a letter followed by a series of numbers. While each individual cabin number will be difficult to incorporate into a model, it is possible that the preceding letter is associated with a certain location on the ship (i.e. a certain deck or level), which could be closely related to survival. While this may be an interesting source of information, there are likely too many missing values, so we do not include this variable in the analysis either.

We begin by splitting the dataset into a training and test set, using a 75/25 train/test split.

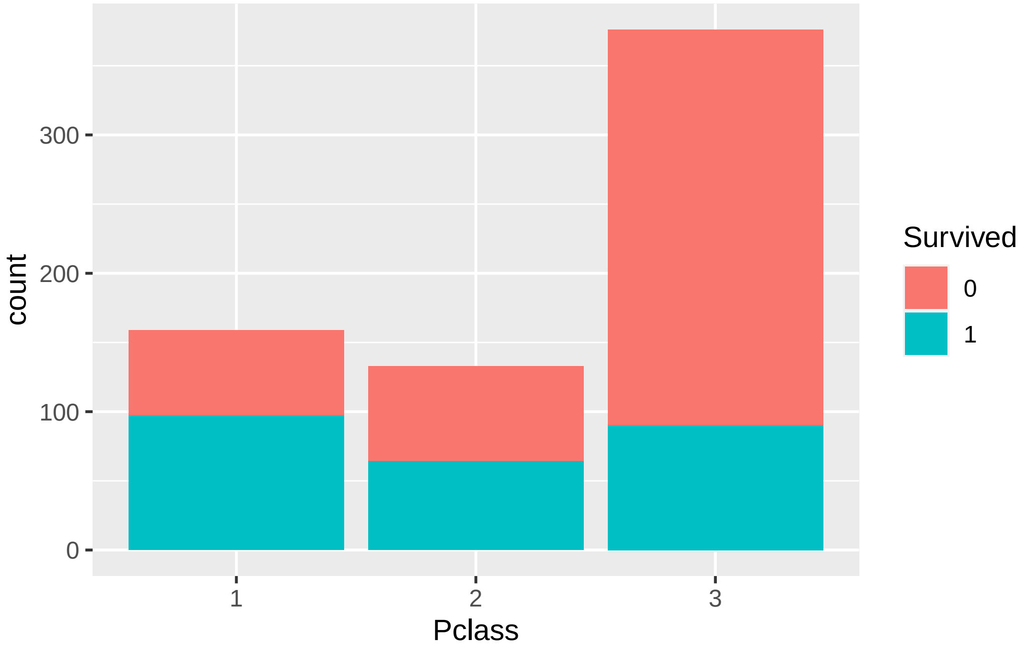
```r
set.seed(100)
train_ids <- sample(
  titanic$PassengerId,
  size = nrow(titanic) * 0.75
)

titanic_train <- titanic %>%
  filter(PassengerId %in% train_ids)

titanic_test <- titanic %>%
  filter(!(PassengerId %in% train_ids))
```

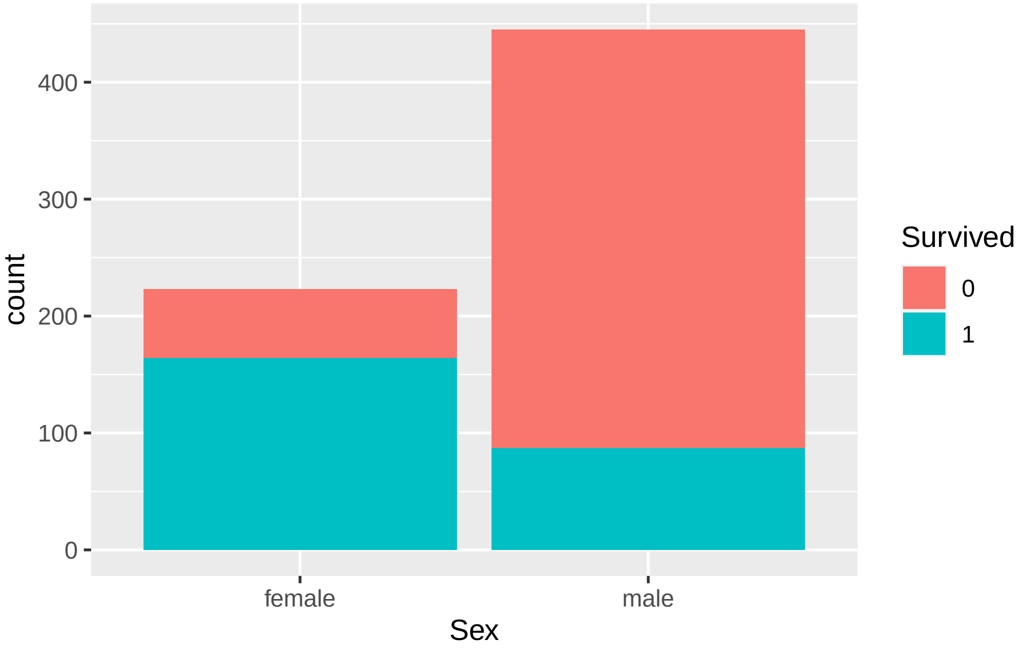
Next, we explore some potential relationships between different variables and survival. Historical accounts of the disaster traditionally emphasize the importance of passenger class on whether people survived or not, and also note that the crew prioritized saving women and children before the men on the ship. Because of this, we begin by examining these variables. The bar plot below indicates that there were many more 3rd class passengers than 1st class or 2nd class, and that, while the proportions of those who survived in 1st and 2nd class are relatively comparable, with 1st class showing higher rates of survival, the proportion of 3rd class who survived is far lower. Because of this, passenger class will likely be a useful variable in our model.

```
ggplot(titanic_train) +
  geom_bar(aes(x = Pclass, fill = Survived))
```



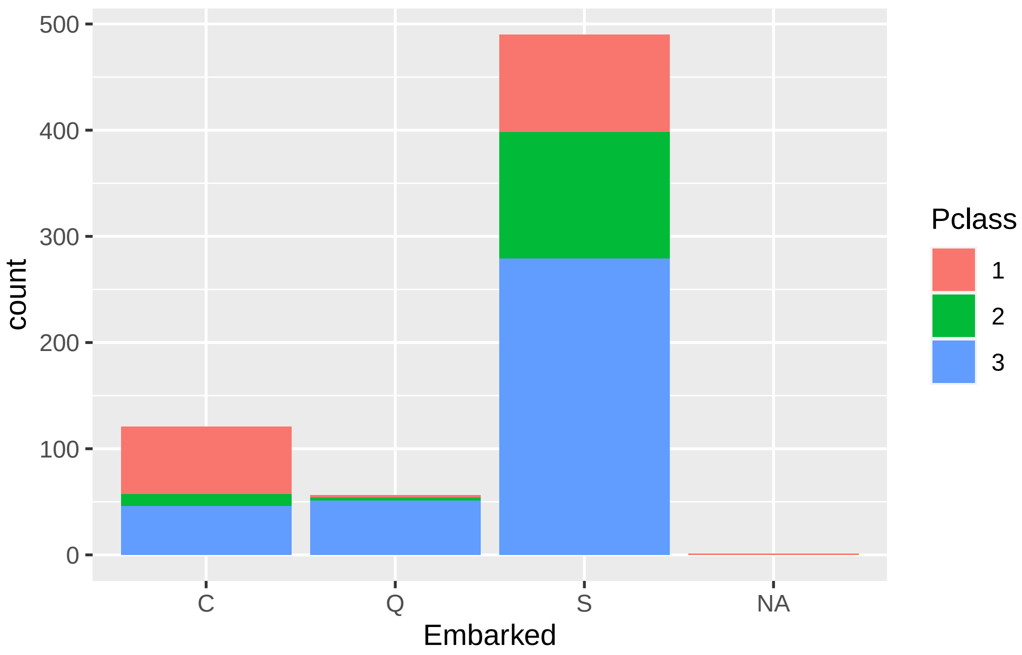
In addition to the large difference in survival rates for the different passenger classes, the plot below shows that there are dramatic differences in survival by sex, with women surviving at much higher rates than men.

```
ggplot(titanic_train) +
  geom_bar(aes(x = Sex, fill = Survived))
```



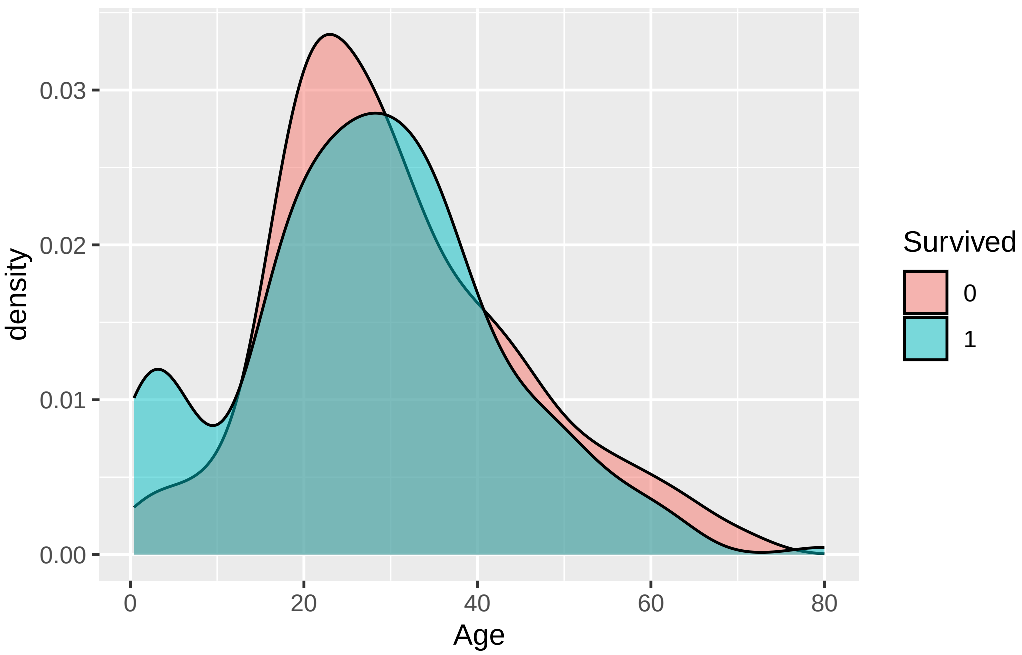
Additionally, the bar plot below shows how passenger class differs by embarkation point. Here, we see that the most common boarding point was Southampton, with Cherbourg being the next most common. Likely due to regional economic differences, there are large disparities in the class composition of the passengers boarding at each location. Almost all passengers boarding at Queenstown are in 3rd class, while the passengers boarding at Cherbourg are nearly evenly split between 1st and 3rd class passengers. Slightly over half of the passengers from Southampton are in 3rd class, while a majority of the remaining passengers are in 2nd class. These distinctions may be seen in the model results, though it is hard to imagine embarkation point influencing survival rates in any way other than passenger class or possibly cabin location.

```r
ggplot(titanic_train) +
  geom_bar(aes(x = Embarked, fill = Pclass))
```



Finally, we examine the impact of age on survival. The density plots below indicate that there is an increase in the number of passengers under 10 years old, in the group of passengers who survived. This does not carry over to the group who did not survive.

```
ggplot(titanic_train) +
  geom_density(aes(x = Age, fill = Survived), alpha = 0.5)
```



### Logistic Regression

With some basic exploration complete, we move to fitting the logistic regression model. In this model, we use all available predictors except for those excluded as mentioned above.

```
logistic_mod1 <- glm(
  Survived ~ Pclass +
    Sex +
    Age +
    SibSp +
    Parch +
    Fare +
    Embarked,
  data = titanic_train,
  family = binomial(link = "logit"),
  na.action = na.omit
)
```

```
summary(logistic_mod1)
```

```
Call:
glm(formula = Survived ~ Pclass + Sex + Age + SibSp + Parch +
    Fare + Embarked, family = binomial(link = "logit"), data = titanic_train,
    na.action = na.omit)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.6256  -0.6740  -0.4112   0.6666   2.4107

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)  4.3903568  0.5974191   7.349 2.00e-13 ***
Pclass2     -1.2019506  0.3718577  -3.232  0.00123 **
Pclass3     -2.3500773  0.3794341  -6.194 5.88e-10 ***
Sexmale     -2.5469946  0.2532901 -10.056  < 2e-16 ***
Age         -0.0427651  0.0092833  -4.607 4.09e-06 ***
SibSp       -0.3444173  0.1498881  -2.298  0.02157 *
Parch       -0.0307946  0.1414381  -0.218  0.82764
Fare        -0.0004588  0.0029127  -0.158  0.87484
EmbarkedQ   -1.0504986  0.6372802  -1.648  0.09927 .
EmbarkedS   -0.4146896  0.3048495  -1.360  0.17373
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 730.08  on 543  degrees of freedom
Residual deviance: 500.87  on 534  degrees of freedom
  (124 observations deleted due to missingness)
AIC: 520.87

Number of Fisher Scoring iterations: 5
```

From the model summary, we see that passenger class, sex, and age all have highly significant regression coefficients, and that the number of siblings or spouses on board also has a significant coefficient. We also see that, while the coefficients for embarkation point are quite different from 0, they are not statistically significant after accounting for passenger class in the model.

```
exp(logistic_mod1$coefficients)
```

```
(Intercept)     Pclass2     Pclass3     Sexmale         Age       SibSp
80.66919842  0.30060728  0.09536179  0.07831668  0.95813644  0.70863316
```

```
      Parch        Fare   EmbarkedQ   EmbarkedS 
 0.96967468  0.99954133  0.34976331  0.66054531 
```

By exponentiating the regression coefficients, we see that the most important predictor variables are Passenger Class and Sex. Compared to the group used in the intercept, a female passenger 1st class, and with all other variables held constant, a passenger in 3rd class has roughly one tenth the odds of survival as a passenger in first class, and a male passenger has 0.078 times the odds of survival as a female passenger. We also see that for each year increase in age, the odds of survival decrease by 5%, reflecting the influence of a higher survival rate among children.

```
logistic_mod1_pred <- predict(
  logistic_mod1,
  newdata = titanic_test,
  type = "response"
)

# accuracy calculation
mean(
  round(logistic_mod1_pred) == titanic_test$Survived,
  na.rm = TRUE
)
```

```
[1] 0.8571429
```

```
mean(
  round(logistic_mod1_pred)[titanic_test$Survived == 1] == titanic_test$Survived[titanic_t
  na.rm = TRUE
)
```

```
[1] 0.7808219
```

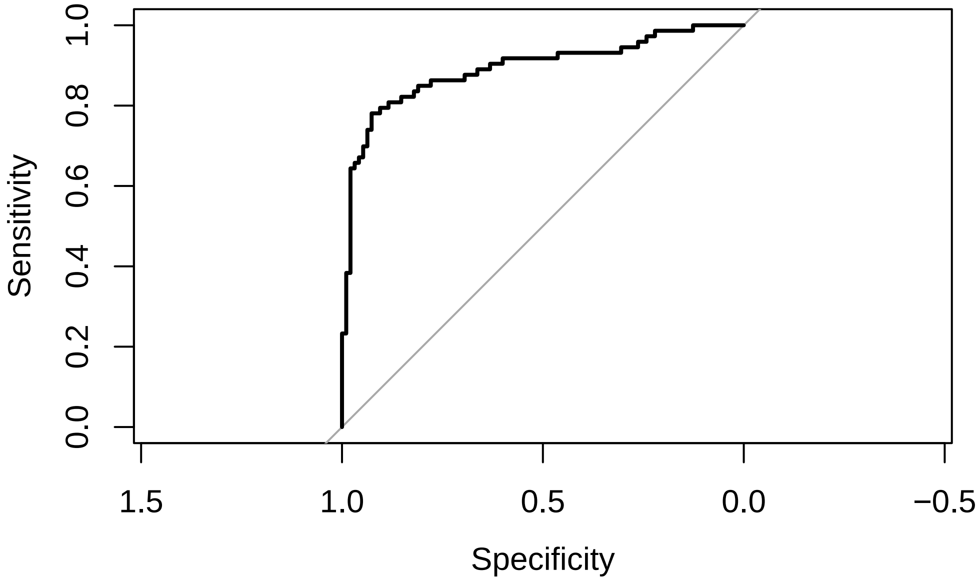
```
mean(
  round(logistic_mod1_pred)[titanic_test$Survived == 0] == titanic_test$Survived[titanic_t
  na.rm = TRUE
)
```

```
[1] 0.9157895
```

Above, we calculate the model's fitted values when applied to the test dataset. After calculating the accuracy, sensitivity, and specificity, we see that the model has an accuracy of 85.7%, sensitivity of 78.1%, and specificity of 91.6%. Note that the calculations for sensitivity and specificity are very similar to the accuracy computation, with the difference being that we only include the indices where `[titanic_test$Survived == 1]` is true or not.

```r
logistic_mod1_roc <- roc(
  response = titanic_test$Survived,
  predictor = logistic_mod1_pred
)

plot(logistic_mod1_roc)
```



In addition to simply calculating the sensitivity and specificity values, we can also view the ROC curve, as shown above. We see that the AUC for this model is 0.8926, indicating an effective model.

```r
logistic_mod1_roc
```

```
Call:
```

```
roc.default(response = titanic_test$Survived, predictor = logistic_mod1_pred)

Data: logistic_mod1_pred in 95 controls (titanic_test$Survived 0) < 73 cases (titanic_test$Su
Area under the curve: 0.8926
```

### Probit Regression

Next, we turn to probit regression.

```r
probit_mod1 <- glm(
  Survived ~ Pclass +
    Sex +
    Age +
    SibSp +
    Parch +
    Fare +
    Embarked,
  data = titanic_train,
  family = binomial(link = "probit"),
  na.action = na.omit
)
```

```r
summary(probit_mod1)
```

```

Call:
glm(formula = Survived ~ Pclass + Sex + Age + SibSp + Parch +
    Fare + Embarked, family = binomial(link = "probit"), data = titanic_train,
    na.action = na.omit)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.7135  -0.6896  -0.4081   0.6732   2.4351

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)  2.5639872  0.3360698   7.629 2.36e-14 ***
Pclass2     -0.6962295  0.2168725  -3.210  0.00133 **
Pclass3     -1.3343296  0.2165517  -6.162 7.20e-10 ***
Sexmale     -1.5167499  0.1427829 -10.623  < 2e-16 ***
Age         -0.0241704  0.0052874  -4.571 4.85e-06 ***
```

```
SibSp       -0.1977931  0.0849449  -2.328  0.01989 *
Parch       -0.0331959  0.0839178  -0.396  0.69242
Fare        -0.0002754  0.0016963  -0.162  0.87103
EmbarkedQ   -0.6683986  0.3738172  -1.788  0.07377 .
EmbarkedS   -0.2529084  0.1756036  -1.440  0.14980
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 730.08  on 543  degrees of freedom
Residual deviance: 501.66  on 534  degrees of freedom
  (124 observations deleted due to missingness)
AIC: 521.66

Number of Fisher Scoring iterations: 5
```

Examining the model summary, we see coefficient values that are roughly similar in direction and significance level, if not in magnitude, to the logistic regression. However, due to the use of the probit link function, these coefficients are difficult to interpret. Instead, we look to the accuracy, sensitivity, and specificity measurements. These are calculated using the same code as for logistic regression. We see that this model has accuracy of 85.1%, sensitivity of 78.1%, and specificity of 90.5%. Compared to the logisitc regression model, we see that we have very similar sensitivity, but slightly lower specificity.

```r
probit_mod1_pred <- predict(
  probit_mod1,
  newdata = titanic_test,
  type = "response"
)

mean(round(probit_mod1_pred) == titanic_test$Survived, na.rm = TRUE)
```

```
[1] 0.8511905
```

```r
mean(
  round(probit_mod1_pred)[titanic_test$Survived == 1] == titanic_test$Survived[titanic_tes
  na.rm = TRUE
)
```

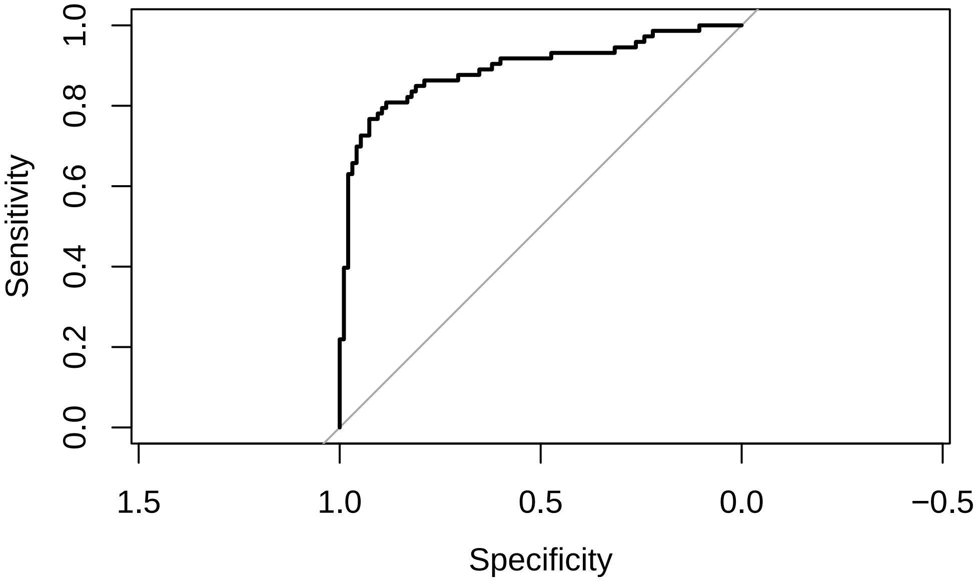
```
[1] 0.7808219
```

```r
mean(
  round(probit_mod1_pred)[titanic_test$Survived == 0] == titanic_test$Survived[titanic_tes
  na.rm = TRUE
)
```

```
[1] 0.9052632
```

```r
probit_mod1_roc <- roc(
  response = titanic_test$Survived,
  predictor = probit_mod1_pred
)

plot(probit_mod1_roc)
```



The ROC curve provides us with an AUC of 0.8921, also slightly lower than the logisitic regression model. Taken together, this tells us that the logistic regression has slightly better performance than the probit regression model in this setting. Also accounting for the improved interpretability of the logistic model, it seems reasonable to prefer the logistic regression model in this case.

```
probit_mod1_roc
```

```

Call:
roc.default(response = titanic_test$Survived, predictor = probit_mod1_pred)

Data: probit_mod1_pred in 95 controls (titanic_test$Survived 0) < 73 cases (titanic_test$Surv
Area under the curve: 0.8921
```