GLM and Logistic Regression

MD MAHFUJUL KARIM SHEIKH

2025-03-10

Contents

Packages

install.packages(“titanic”)

install.packages(“caret”, dependencies = TRUE)

```r
library(tidyverse)
library(titanic)
```

Data

```r
data <- titanic_train

data <- data %>%
  select(Survived, Pclass, Sex, Age, Fare) %>%
  filter(!is.na(Age)) %>%
  mutate(Survived = factor(Survived, levels = c(0, 1)),
         Sex = factor(Sex, levels = c("male", "female")))
```

```r
data %>%
  group_by(Sex, Pclass) %>%
  summarise(mean(Age), n())
```

```
# A tibble: 6 x 4
# Groups:   Sex [2]
  Sex    Pclass `mean(Age)` `n()`
  <fct>   <int>       <dbl> <int>
1 male        1        41.3   101
2 male        2        30.7    99
3 male        3        26.5   253
4 female      1        34.6    85
5 female      2        28.7    74
6 female      3        21.8   102
```

```r
mean(data$Age)
```

```
[1] 29.69912
```

```r
model <- glm(Survived ~ Pclass + Sex + Age, data = data, family = binomial)
summary(model)
```

```
Call:
glm(formula = Survived ~ Pclass + Sex + Age, family = binomial,
    data = data)

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  2.533875   0.456247   5.554 2.80e-08 ***
Pclass      -1.288545   0.139259  -9.253  < 2e-16 ***
Sexfemale    2.522131   0.207283  12.168  < 2e-16 ***
Age         -0.036929   0.007628  -4.841 1.29e-06 ***
---
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: 647.29  on 710  degrees of freedom
AIC: 655.29

Number of Fisher Scoring iterations: 5
```

```r
coef(model)
```

```
(Intercept)      Pclass   Sexfemale         Age 
 2.53387533 -1.28854507  2.52213086 -0.03692902 
```

```r
exp(coef(model))
```

```
(Intercept)      Pclass   Sexfemale         Age 
 12.6022495   0.2756716  12.4551085   0.9637445 
```

```r
exp(coef(model)[-1])  # odds ratio
```

```
    Pclass  Sexfemale        Age 
 0.2756716 12.4551085  0.9637445 
```

Each increase in class decreases survival odds by 72% (since odds = 0.28).
Females were 92% more likely to survive than males.
Males are 92% less likely to survive than females.
Each year increase in age decreases odds of survival by 4%.
Fare has no meaningful effect.

Prediction

Predict survival probabilities for new data:

```r
new_data <- data.frame(
  Pclass = c(1, 2, 3),
  Sex = factor(c("male", "female", "female"), levels = c("male", "female")),
  Age = c(30, 25, 40)
)

new_data$link <- predict(model, newdata = new_data, type = "link") # log(odds)
new_data
```

```
  Pclass    Sex Age       link
1      1   male  30  0.1374597
2      2 female  25  1.5556906
3      3 female  40 -0.2867897
```

```r
new_data$Predicted_Prob <- predict(model, newdata = new_data, type = "response")
new_data
```

```
  Pclass    Sex Age       link Predicted_Prob
1      1   male  30  0.1374597      0.5343109
2      2 female  25  1.5556906      0.8257341
3      3 female  40 -0.2867897      0.4287900
```

```
new_data$terms <- predict(model, newdata = new_data, type = "terms")
new_data
```

```
  Pclass    Sex Age       link Predicted_Prob terms.Pclass   terms.Sex
1      1   male  30  0.1374597      0.5343109   1.59353684 -0.92195540
2      2 female  25  1.5556906      0.8257341   0.30499176  1.60017546
3      3 female  40 -0.2867897      0.4287900  -0.98355331  1.60017546
    terms.Age
1 -0.01111129
2  0.17353380
3 -0.38040146
```

```
test_dat <- titanic_test

test_dat$Predicted <- predict(model, test_dat, type = "response")
test_dat$Survived <- ifelse(test_dat$Predicted >= 0.5, "Yes", "No")
```

Evaluation Metrices

Predict survival using the model (threshold = 0.5):

```
predicted_class <- ifelse(predict(model, type = "response") >= 0.5, 1, 0)  # Convert probability to cla
```

Confusion metrix:

```
library(caret)
caret::confusionMatrix(factor(predicted_class), data$Survived)
```

```
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 356  83
         1  68 207

               Accuracy : 0.7885
                 95% CI : (0.7567, 0.8179)
    No Information Rate : 0.5938
    P-Value [Acc > NIR] : <2e-16

                  Kappa : 0.558

 Mcnemar's Test P-Value : 0.2546

            Sensitivity : 0.8396
            Specificity : 0.7138
         Pos Pred Value : 0.8109
         Neg Pred Value : 0.7527
```

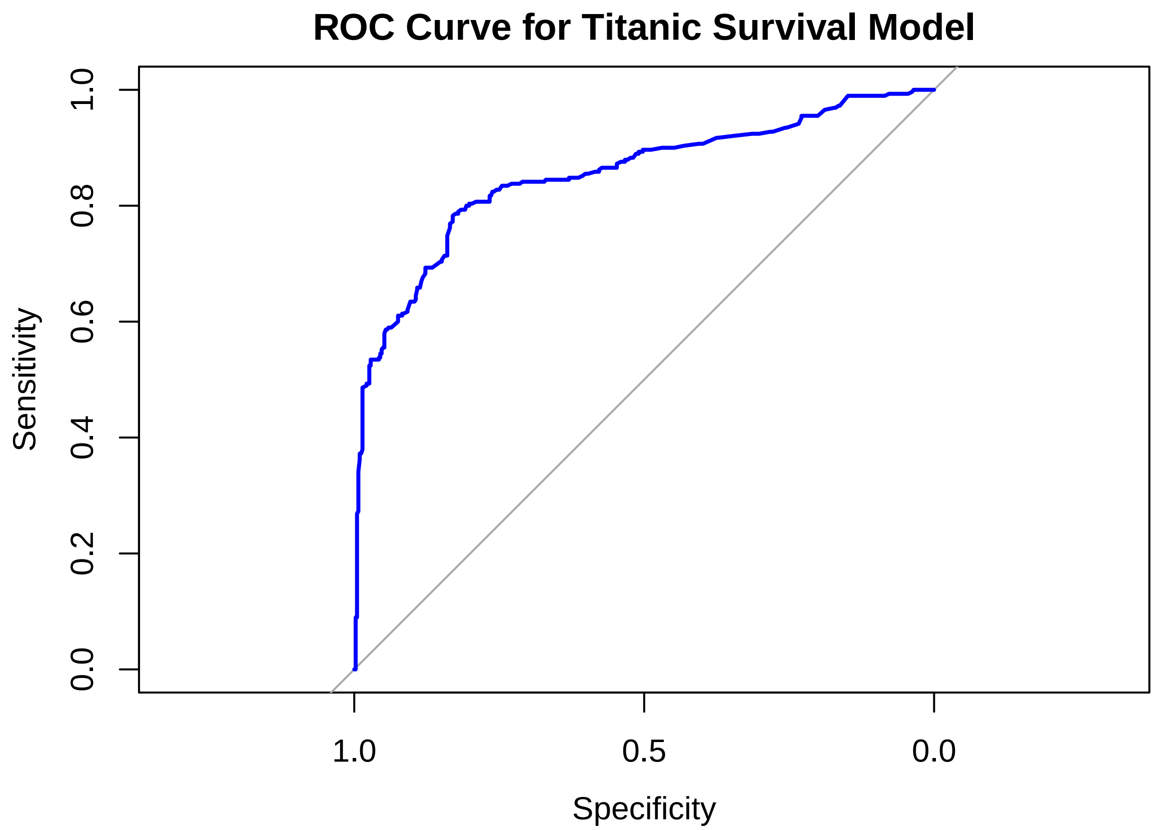
```
            Prevalence : 0.5938
        Detection Rate : 0.4986
  Detection Prevalence : 0.6148
     Balanced Accuracy : 0.7767

      'Positive' Class : 0
```

```r
caret::confusionMatrix(factor(predicted_class), data$Survived, mode = "prec_recall")
```

```
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 356  83
         1  68 207

               Accuracy : 0.7885
                 95% CI : (0.7567, 0.8179)
    No Information Rate : 0.5938
    P-Value [Acc > NIR] : <2e-16

                  Kappa : 0.558

 Mcnemar's Test P-Value : 0.2546

              Precision : 0.8109
                 Recall : 0.8396
                     F1 : 0.8250
             Prevalence : 0.5938
         Detection Rate : 0.4986
   Detection Prevalence : 0.6148
      Balanced Accuracy : 0.7767

       'Positive' Class : 0
```

ROC Curve

```r
library(pROC)

predicted_prob <- predict(model, type = "response")
roc_curve <- roc(data$Survived, predicted_prob)

plot(roc_curve, col = "blue", main = "ROC Curve for Titanic Survival Model")
```




```r
auc_value <- auc(roc_curve)
print(paste("AUC:", auc_value))
```

```
[1] "AUC: 0.852362556929083"
```

McFadden's R-squared

```r
null_model <- glm(Survived ~ 1, data = data, family = binomial)  # Null model (only intercept, no predi
R2 <- 1 - (logLik(model) / logLik(null_model))
print(paste("McFadden's R-squared:", R2 |> round(3)))
```

```
[1] "McFadden's R-squared: 0.329"
```

R squared > 0.2: Good model. R squared > 0.4: Strong model.