Michael_Ghattas_WP5

Michael Ghattas

2024-10-11

Start:

Step 1: Create training and testing sets with 75% data for training

```
# Load titanic data
library(titanic)
df = titanic_train

# Set a seed for reproducibility
set.seed(1)

# Split data into training and testing sets
train.index = sample(1:nrow(df), 0.75 * nrow(df))
df.train = df[train.index, ]     # Training set
df.test  = df[-train.index, ]    # Testing set
```

Step 2: Determine the odds of survival for men vs women using a table

```
# Create a table showing the survival counts based on gender
table(df.train$Sex, df.train$Survived)
```

```
##
##            0   1
##   female  68 181
##   male   342  77
```

```
# Calculate odds ratio for men vs women
odds_men <- sum(df.train$Survived[df.train$Sex == 'male'] == 1) / sum(df.train$Survived[df.train$Sex ==
odds_women <- sum(df.train$Survived[df.train$Sex == 'female'] == 1) / sum(df.train$Survived[df.train$Se
odds_ratio <- odds_men / odds_women

odds_ratio
```

```
## [1] 0.08458531
```

Step 3: Fit a logistic regression model with sex as a predictor

```
# Fit logistic regression model
modelsex = glm(Survived ~ Sex, data = df.train, family = binomial)
summary(modelsex)
```

```
## 
## Call:
## glm(formula = Survived ~ Sex, family = binomial, data = df.train)
## 
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)    
## (Intercept)   0.9790     0.1422   6.883 5.86e-12 ***
## Sexmale      -2.4700     0.1901 -12.992  < 2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
## 
## (Dispersion parameter for binomial family taken to be 1)
## 
##     Null deviance: 891.15  on 667  degrees of freedom
## Residual deviance: 691.76  on 666  degrees of freedom
## AIC: 695.76
## 
## Number of Fisher Scoring iterations: 4
```

Step 4: Fit main effects and interaction models

```
# Main effects model
mainmodel = glm(Survived ~ Sex + as.factor(Pclass) + Age, data = df.train, family = binomial)
summary(mainmodel)
```

```
## 
## Call:
## glm(formula = Survived ~ Sex + as.factor(Pclass) + Age, family = binomial, 
##     data = df.train)
## 
## Coefficients:
##                     Estimate Std. Error z value Pr(>|z|)    
## (Intercept)         3.630530   0.455114   7.977 1.50e-15 ***
## Sexmale            -2.515121   0.235301 -10.689  < 2e-16 ***
## as.factor(Pclass)2 -1.237601   0.319284  -3.876 0.000106 ***
## as.factor(Pclass)3 -2.488687   0.324794  -7.662 1.83e-14 ***
## Age                -0.034653   0.008759  -3.956 7.62e-05 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
## 
## (Dispersion parameter for binomial family taken to be 1)
## 
##     Null deviance: 731.09  on 538  degrees of freedom
## Residual deviance: 489.83  on 534  degrees of freedom
##   (129 observations deleted due to missingness)
## AIC: 499.83
##
## Number of Fisher Scoring iterations: 5
```

```
# Two-way interactions model
twowaymodel = glm(Survived ~ (Sex + as.factor(Pclass) + Age)^2, data = df.train, family = binomial)
summary(twowaymodel)
```

```
##
## Call:
## glm(formula = Survived ~ (Sex + as.factor(Pclass) + Age)^2, family = binomial,
##     data = df.train)
##
## Coefficients:
##                           Estimate Std. Error z value Pr(>|z|)
## (Intercept)                2.87701    1.05675   2.723  0.00648 **
## Sexmale                   -2.82570    1.03607  -2.727  0.00638 **
## as.factor(Pclass)2         1.29929    1.31356   0.989  0.32260
## as.factor(Pclass)3        -2.47743    1.02134  -2.426  0.01528 *
## Age                        0.01938    0.02608   0.743  0.45733
## Sexmale:as.factor(Pclass)2 -0.90216    1.03958  -0.868  0.38549
## Sexmale:as.factor(Pclass)3  2.21895    0.86624   2.562  0.01042 *
## Sexmale:Age               -0.03357    0.02203  -1.524  0.12750
## as.factor(Pclass)2:Age    -0.07335    0.02969  -2.471  0.01349 *
## as.factor(Pclass)3:Age    -0.04660    0.02368  -1.968  0.04908 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 731.09  on 538  degrees of freedom
## Residual deviance: 451.61  on 529  degrees of freedom
##   (129 observations deleted due to missingness)
## AIC: 471.61
##
## Number of Fisher Scoring iterations: 6
```

Step 5: Compare the two models using drop in deviance and ANOVA

```
# Use ANOVA to compare models
anova(mainmodel, twowaymodel, test = "Chisq")
```

```
## Analysis of Deviance Table
##
## Model 1: Survived ~ Sex + as.factor(Pclass) + Age
## Model 2: Survived ~ (Sex + as.factor(Pclass) + Age)^2
##   Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
## 1       534     489.83
## 2       529     451.61  5    38.22 3.407e-07 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

Step 6: Interpret the odds ratio for age on survival

```
# Extracting and interpreting the odds ratio for age
exp(coef(mainmodel)["Age"])
```

```
##       Age
## 0.9659403
```

End.