Week 04 Homewrok

Brian Ritz

Tuesday, February 03, 2015

I chose to study the survival rates of people on the RMS Titanic on April 15, 1912. April 15 marks the infamous maritime disaster in which over 1,500 people lost their lives in the cold, vast Atlantic Ocean after the Titanic struck an iceburg and disappeared below the waves in the span of just 2 hours.

I will determine the probability of any one passenger surviving as a function of their sex and fare paid for the trip.

I found my data on Kaggle, a website for data analysis competitions.

```
# import the data from csv
titanic.data <- read.csv("titanic_data.csv", header=TRUE)
# loook at data
dim(titanic.data)
```

```
## [1] 891  12
```

```
colnames(titanic.data)
```

```
##  [1] "PassengerId" "Survived"    "Pclass"      "Name"        "Sex"
##  [6] "Age"         "SibSp"       "Parch"       "Ticket"      "Fare"
## [11] "Cabin"       "Embarked"
```

```r
class(titanic.data$Sex)
```

```
## [1] "factor"
```

```r
class(titanic.data$Fare)
```

```
## [1] "numeric"
```

```r
class(titanic.data$Survived)
```

```
## [1] "integer"
```

Make the model, R will create binary variable for us for Sex because it is a factor.

We will fit the models with three different link funcitons, the logit, the probit, and the complementary log-log. All three functions return similar probabilities when the linear predictor is around zero. The complementary log fucntion returns a slightly higher probabilty when the linear predictor is zero.

Towards the extremes, the probit and log-log fucntions return probabilities closer to 1 or 0 faster than the logit function as the linear predictor differs from zero.

```r
logit.link.model <- glm(Survived ~ Sex+Fare, data=titanic.data, family=binomial(link="logit"))
probit.link.model <- glm(Survived ~ Sex+Fare, data=titanic.data, family=binomial(link="probit"))
clog.link.model <- glm(Survived ~ Sex+Fare, data=titanic.data, family=binomial(link="cloglog"))
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

Look at the coefficients for each model:

```
data.frame(logit=coef(logit.link.model),probit=coef(probit.link.model),clog=coef(clog.link.model))
```

```
##                logit    probit      clog
## (Intercept)  0.64710  0.406092  0.049873
## Sexmale     -2.42276 -1.475478 -1.823584
## Fare         0.01121  0.006668  0.006896
```

All three functions return coefficients on the same order of magnitude for each predictor.

Look at the predicted probabilities:

```
predicted.vals<-data.frame(logit=fitted.values(logit.link.model), probit=fitted.values(probit.link.model),
clog=fitted.values(clog.link.model))

predicted.vals[1:10,]
```

```
##     logit probit   clog
## 1  0.1552 0.1536 0.1634
## 2  0.8095 0.8110 0.8207
## 3  0.6761 0.6769 0.6705
## 4  0.7760 0.7764 0.7804
## 5  0.1564 0.1549 0.1642
## 6  0.1570 0.1555 0.1646
## 7  0.2325 0.2347 0.2155
## 8  0.1766 0.1765 0.1782
```

```
## 9  0.6839 0.6845 0.6786
## 10 0.7280 0.7279 0.7256
```

Most of the predicted values are very close to eachother. The complementary log-log model appears a little further away from the probit and logit models however.

Lets take a closer look at the logit model:

```
summary(logit.link.model)
```

```
##
## Call:
## glm(formula = Survived ~ Sex + Fare, family = binomial(link = "logit"),
##     data = titanic.data)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -2.208  -0.621  -0.582   0.813   1.966
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   0.6471     0.1485    4.36  1.3e-05 ***
## Sexmale      -2.4228     0.1705  -14.21  < 2e-16 ***
## Fare          0.0112     0.0023    4.89  1.0e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
```

```
##
##     Null deviance: 1186.66  on 890  degrees of freedom
## Residual deviance:  884.31  on 888  degrees of freedom
## AIC: 890.3
##
## Number of Fisher Scoring iterations: 5
```

```
exp(coef(logit.link.model)["Sexmale"])
```

```
## Sexmale
## 0.08868
```

Here we see that sex and Fare both play somewhat of a role in predicting whether a passenger will survive the Titanic catastrophe. I conclude that being male made it less likely to survive the disaster. Specifically, being a male reduced your chances of surviving to 0.0887 of what it would be if you were female. Likewise, for every dollar increase in your fare, you were 1.0113 times as likely to survive.