Sean Cordrey

November 3, 2014

Stat 3654

Roberts

In Class 9 Part 1

Code:

```
setwd("C:\\Users\\Sean\\Desktop\\Stat 3654")

load("fdata.Rdata")


head(final)

attach(final)

logit = glm(disorder ~ som1 + som2 + som3 + som4 + som5, data = final, family = "binomial")

summary(logit)


logit2 = glm(disorder ~ som6 + som7 + som8 + som9, data = final, family = "binomial")

summary(logit2)


logit3 = glm(disorder ~ som10 + som11 + som12 + som13 + som14, data = final, family = "binomial")

summary(logit3)
```

Model 1:

```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.3107  -0.2463   0.0000   0.0957   3.0752

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -4.7196     0.5726  -8.242  < 2e-16 ***
som1          0.5215     0.2351   2.218  0.02656 *
som2          1.2390     0.2970   4.172 3.02e-05 ***
som3          0.5441     0.1171   4.648 3.35e-06 ***
som4          0.5320     0.1468   3.624  0.00029 ***
som5          2.4536     0.4228   5.804 6.48e-09 ***
---
```

```
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 447.70  on 322  degrees of freedom
Residual deviance: 123.75  on 317  degrees of freedom
AIC: 135.75

Number of Fisher Scoring iterations: 8

> logit$coef
(Intercept)        som1        som2        som3        som4        som5
 -4.7195679   0.5214651   1.2390471   0.5441373   0.5320053   2.4535915
> exp(logit$coef)
 (Intercept)         som1         som2         som3         som4         som5
 0.008919032  1.684493784  3.452322274  1.723121264  1.702342539 11.630041636
```

Model 2:

```
Deviance Residuals:
     Min        1Q    Median        3Q       Max
-2.34372  -0.62207   0.00045   0.49654   1.86426

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -1.5442     0.2009  -7.685 1.53e-14 ***
som6          1.9406     0.4662   4.163 3.15e-05 ***
som7          1.0921     0.2536   4.307 1.66e-05 ***
som8          1.1669     0.4176   2.794   0.0052 **
som9          1.1918     0.1925   6.190 6.03e-10 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 447.70  on 322  degrees of freedom
Residual deviance: 270.32  on 318  degrees of freedom
AIC: 280.32

Number of Fisher Scoring iterations: 7
```

Model 3:

```
Deviance Residuals:
     Min        1Q    Median        3Q       Max
-1.84363  -0.26365   0.00067   0.05370   3.09428

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -4.7789     0.6066  -7.878 3.32e-15 ***
som10         1.0075     0.3260   3.090   0.0020 **
som11         0.7396     0.3913   1.890   0.0587 .
som12         0.5288     0.3161   1.673   0.0944 .
som13         1.4370     0.2148   6.689 2.24e-11 ***
som14         1.0204     0.4125   2.474   0.0134 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)
```

```
    Null deviance: 447.70  on 322  degrees of freedom
Residual deviance: 110.32  on 317  degrees of freedom
AIC: 122.32

Number of Fisher Scoring iterations: 8
```

Interpretations:

1) All of the factors are statistically significant. The coefficient for som1 is 1.68 after applying the exponential to it, meaning there is a 68% chance of there being a disorder present if som1 is present.
2) All of the factors are statistically significant.
3) som10, 13 and 14 are statistically significant, while som11, and 12 are not.
4) Comparing the 3 models to each other shows that the second model using som6 through som9 is the best indicator because it has higher significance than the other two. This may be due to only having 4 factors as opposed to 5 though so the first model may turn out to be better. A little more testing would be necessary to conclusively say.

In Class 9 Part 2

Code:

```
final$gp <- runif(dim(final)[1])

testSet <- subset(final, final$gp <= 0.1)

trainSet <- subset(final, final$gp > 0.1)

rm(final)



install.packages("MASS")

library(MASS)



attach(trainSet)

#linear regression

fit<- lm(ssc ~
som1+som2+som3+som4+som5+som6+som7+som8+som9+som10+som11+som12+som13+som14

        +age+gender+location+ethnicity+coder)

summary(fit)
```

```
#step regression

step <- stepAIC(fit, direction = "both")

step

step$anova

#retained variables linear regression

fitRetained<-lm(ssc ~ som1+som2+som3+som4+som5+som9+som10+som12+som13+som14+

                age+location+ethnicity+coder)

summary(fitRetained)


#Prediction

testSet$sscpred <- predict(fitRetained, newdata = testSet)

head(testSet)


library(ggplot2)

ggplot(data = testSet, aes(x = sscpred, y = ssc)) +

  geom_point(color = "red")+

  geom_line(aes(x = ssc, y = ssc), color = "blue")
```

Conclusions:

The model did ok with predicting ssc scores. The predicted values tended to be pretty spread out from the actuals but the overall trend was preserved fairly well. At lower scores the model tended to overestimate and at higher scores tended to underestimate compared to the actuals.