Week 14

Ali

August 13, 2015

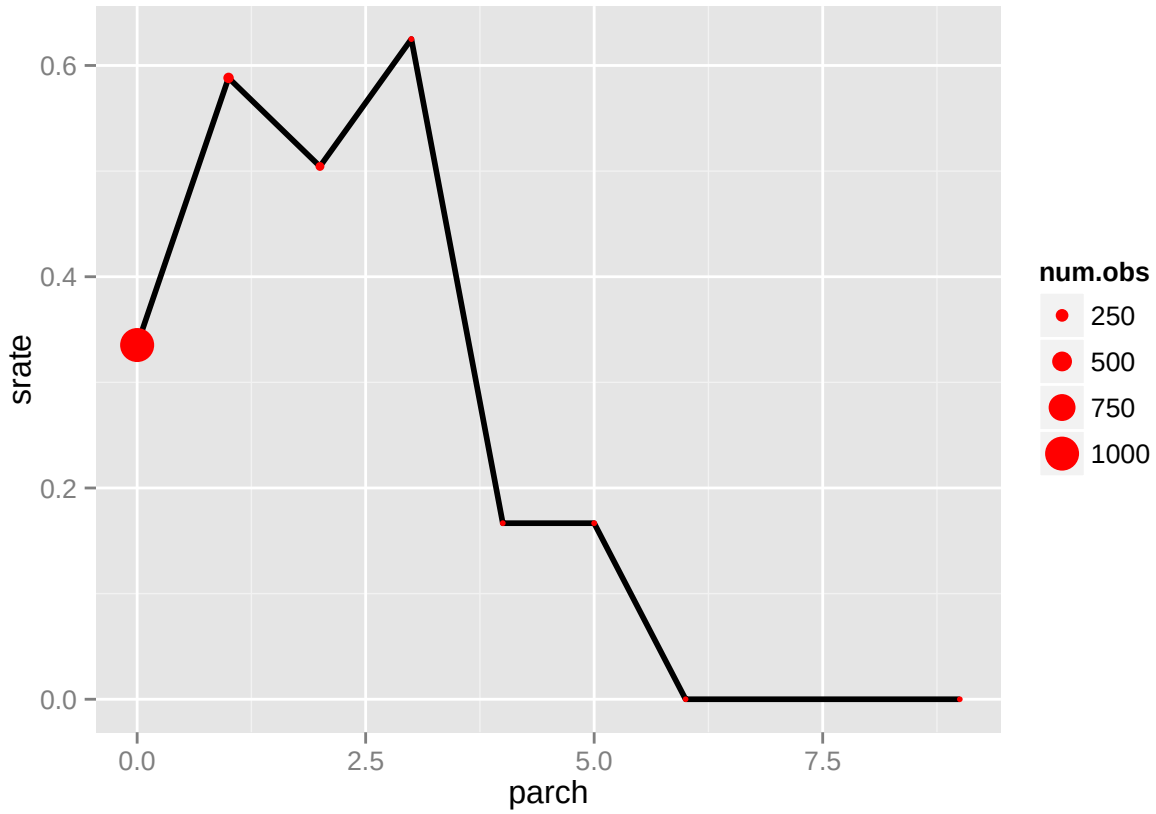
```
#install.packages("PASWR")
library(PASWR)
```

```
## Loading required package: e1071
## Loading required package: MASS
## Loading required package: lattice
```

```
#? titanic3
#summary(titanic3)

# I am lazy:
T= titanic3

table(T$parch)
```

```
##
##    0    1    2    3    4    5    6    9
## 1002  170  113    8    6    6    2    2
```

```
Table = table(T$parch,T$survived)
D = data.frame( matrix(Table,ncol=2))
D$parch = as.numeric( rownames(Table))
D$num.obs = D[,1]+D[,2]
D$srate = D[,2]/D$num.obs

library(ggplot2)
ggplot(D , aes(parch,srate, size = num.obs) ) +
      geom_line(size=1)+
      geom_point(col='red')
```




```
model1 = glm(survived~parch, data=T , family = binomial)
summary(model1)
```

```
##
## Call:
## glm(formula = survived ~ parch, family = binomial, data = T)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.7022  -0.9516  -0.9516   1.4215   1.4215
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -0.55753    0.06289  -8.865  < 2e-16 ***
## parch        0.19318    0.06625   2.916  0.00355 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1741.0  on 1308  degrees of freedom
## Residual deviance: 1732.3  on 1307  degrees of freedom
## AIC: 1736.3
##
## Number of Fisher Scoring iterations: 4
```

```
exp(coef(model1))
```

```
## (Intercept)       parch
##   0.5726225   1.2130983
```

But we remember that there was a quadratic form in the relationship

```
model2 = glm(survived~parch+I(parch^2), data=T , family = binomial)
summary(model2)
```

```
##
## Call:
## glm(formula = survived ~ parch + I(parch^2), family = binomial,
##     data = T)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.3041  -0.9102  -0.9102   1.4705   2.3967
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -0.66698    0.06625 -10.068  < 2e-16 ***
## parch        1.08623    0.18222   5.961 2.50e-09 ***
## I(parch^2)  -0.30312    0.07169  -4.228 2.36e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1741.0  on 1308  degrees of freedom
## Residual deviance: 1693.3  on 1306  degrees of freedom
## AIC: 1699.3
##
## Number of Fisher Scoring iterations: 6
```

```
exp(coef(model2))
```

```
## (Intercept)       parch  I(parch^2)
##   0.5132575   2.9630757   0.7385125
```

Which model is better? From the AIC, we can say that model2 is better:

```
AIC(model2)-AIC(model1)
```

```
## [1] -37.00068
```

Also, let's do the anova:

```r
anova(model1,model2, test = 'Chisq')
```

```
## Analysis of Deviance Table
##
## Model 1: survived ~ parch
## Model 2: survived ~ parch + I(parch^2)
##   Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
## 1      1307     1732.3
## 2      1306     1693.3  1   39.001 4.237e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Finally, we want to see how the predicted probabilities look like.

```r
#predicting probabilities in the two models
newD = data.frame( parch = 0:9)
newD$pr1 = predict(model1, newdata=newD , type='response')
newD$pr2 = predict(model2, newdata=newD , type='response')

#for a pretty plot
# I am going to make a long-form data
NEWD= rbind(newD,newD)
NEWD$predicted.probability = c(newD$pr1,newD$pr2)
NEWD$model = c(rep('model1', 10) , rep('model2',10))
ggplot(NEWD, aes(parch,predicted.probability, col=model))+
    geom_line(size=2)+
    scale_x_continuous(breaks=0:9)
```