PSTAT127 Homework 4

Kevin Ayala

2/22/2019

1)

(e) Fit a Poisson response model for the number of incidents with the predictors: log of service, type, year and period. Test whether the parameter associated with the service term can be one. Explain why we are interested in such a test

```r
library(MASS)
?ships
ships
```

```
##    type year period service incidents
## 1     A   60     60     127         0
## 2     A   60     75      63         0
## 3     A   65     60    1095         3
## 4     A   65     75    1095         4
## 5     A   70     60    1512         6
## 6     A   70     75    3353        18
## 7     A   75     60       0         0
## 8     A   75     75    2244        11
## 9     B   60     60   44882        39
## 10    B   60     75   17176        29
## 11    B   65     60   28609        58
## 12    B   65     75   20370        53
## 13    B   70     60    7064        12
## 14    B   70     75   13099        44
## 15    B   75     60       0         0
## 16    B   75     75    7117        18
## 17    C   60     60    1179         1
## 18    C   60     75     552         1
## 19    C   65     60     781         0
## 20    C   65     75     676         1
## 21    C   70     60     783         6
## 22    C   70     75    1948         2
## 23    C   75     60       0         0
## 24    C   75     75     274         1
## 25    D   60     60     251         0
## 26    D   60     75     105         0
## 27    D   65     60     288         0
## 28    D   65     75     192         0
## 29    D   70     60     349         2
## 30    D   70     75    1208        11
## 31    D   75     60       0         0
## 32    D   75     75    2051         4
## 33    E   60     60      45         0
## 34    E   60     75       0         0
## 35    E   65     60     789         7
## 36    E   65     75     437         7
## 37    E   70     60    1157         5
## 38    E   70     75    2161        12
```

```
## 39    E   75     60       0         0
## 40    E   75     75     542         1
```

```r
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.2.1 --
```

```
## v ggplot2 3.0.0     v purrr   0.2.5
## v tibble  1.4.2     v dplyr   0.7.6
## v tidyr   0.8.1     v stringr 1.3.1
## v readr   1.1.1     v forcats 0.3.0
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## x dplyr::select() masks MASS::select()
```

```r
cleanships <- ships %>% filter(service != 0)
modelfit1 <- glm(incidents ~ log(service) + type + year + period, data = cleanships, family = poisson)
summary(modelfit1)
```

```
##
## Call:
## glm(formula = incidents ~ log(service) + type + year + period,
##     family = poisson, data = cleanships)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.2355  -1.0345  -0.4454   0.6005   2.8353
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -8.616856   1.528004  -5.639 1.71e-08 ***
## log(service)  0.886469   0.099297   8.927  < 2e-16 ***
## typeB        -0.330248   0.261301  -1.264   0.2063
## typeC        -0.736295   0.341342  -2.157   0.0310 *
## typeD        -0.284220   0.291989  -0.973   0.3304
## typeE         0.335936   0.242645   1.384   0.1662
## year          0.035468   0.013802   2.570   0.0102 *
## period        0.022079   0.008114   2.721   0.0065 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 614.539  on 33  degrees of freedom
## Residual deviance:  58.114  on 26  degrees of freedom
## AIC: 171.98
##
## Number of Fisher Scoring iterations: 5
```

When we inspect the data closley, we notice that the log of the service is in a close enough range to 1, it is possible to model our rate in which in this case is incidents. This is becuase were hold constant the count response by using the Poisson regression while keeping the coefficeint with offset. Thus incident damage is correlated to service by the data upon further inspection.

(f) Fit the Poisson rate model with all two-way interactions of the three predictors. Does this model fit the data?

```r
modelfit2 <- glm(incidents ~ (type + year + period)^2, data = cleanships, family = poisson(link = "log"
modelfit2
```

```
##
## Call:  glm(formula = incidents ~ (type + year + period)^2, family = poisson(link = "log"),
##     data = cleanships, offset = log(service))
##
## Coefficients:
##  (Intercept)         typeB         typeC         typeD         typeE
##   -34.656444     -0.122240     -0.550223      2.233244     15.123276
##         year        period    typeB:year    typeC:year    typeD:year
##     0.407120      0.367583      0.005232      0.090512     -0.058216
##   typeE:year  typeB:period  typeC:period  typeD:period  typeE:period
##    -0.220308     -0.010416     -0.091131      0.023830      0.006272
##  year:period
##    -0.005096
##
## Degrees of Freedom: 33 Total (i.e. Null);  18 Residual
## Null Deviance:       146.3
## Residual Deviance: 32.12     AIC: 162
```

```r
summary(modelfit2)
```

```
##
## Call:
## glm(formula = incidents ~ (type + year + period)^2, family = poisson(link = "log"),
##     data = cleanships, offset = log(service))
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.8476  -1.0609  -0.1118   0.3878   2.0800
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -34.656444  10.105973  -3.429 0.000605 ***
## typeB         -0.122240   3.451462  -0.035 0.971747
## typeC         -0.550223   6.321104  -0.087 0.930635
## typeD          2.233244   5.577499   0.400 0.688860
## typeE         15.123276   5.234048   2.889 0.003860 **
## year           0.407120   0.147188   2.766 0.005675 **
## period         0.367583   0.133900   2.745 0.006047 **
## typeB:year     0.005232   0.048638   0.108 0.914339
## typeC:year     0.090512   0.093349   0.970 0.332239
## typeD:year    -0.058216   0.076622  -0.760 0.447385
## typeE:year    -0.220308   0.077925  -2.827 0.004696 **
## typeB:period  -0.010416   0.028935  -0.360 0.718873
## typeC:period  -0.091131   0.048570  -1.876 0.060619 .
## typeD:period   0.023830   0.061210   0.389 0.697048
## typeE:period   0.006272   0.036799   0.170 0.864658
## year:period   -0.005096   0.001912  -2.666 0.007679 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
```

```
##
##     Null deviance: 146.328  on 33  degrees of freedom
## Residual deviance:  32.116  on 18  degrees of freedom
## AIC: 161.98
##
## Number of Fisher Scoring iterations: 6
```

Yes it does fit the model, no predictors need to be dropped as non are significant as the p value is very close to 1 or is 1, which means we always reject the null hypothesis

(h) Now fit the rate model with just the main effects and compare it to the interaction model. Which model is preferred?

```r
modelfit3 <- glm(incidents ~ period + year + type, family = poisson(link = "log"),
            data = cleanships, offset = log(service))
summary(modelfit3)
```

```
##
## Call:
## glm(formula = incidents ~ period + year + type, family = poisson(link = "log"),
##     data = cleanships, offset = log(service))
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.5348  -0.9319  -0.3686   0.4654   2.8833
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -10.079076   0.876149 -11.504  < 2e-16 ***
## period        0.023705   0.008091   2.930 0.003392 **
## year          0.042247   0.012826   3.294 0.000988 ***
## typeB        -0.546090   0.178415  -3.061 0.002208 **
## typeC        -0.632631   0.329500  -1.920 0.054862 .
## typeD        -0.232257   0.287979  -0.807 0.419951
## typeE         0.405975   0.234933   1.728 0.083981 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 146.328  on 33  degrees of freedom
## Residual deviance:  59.375  on 27  degrees of freedom
## AIC: 171.24
##
## Number of Fisher Scoring iterations: 5
```

Here we can use the Akaike Information Criteria and compare the 2nd and 3rd model. Under the sencond model we have it so that our AIC is 165 whereas this new third model has a AIC scoore of 146, which is smaller than the second. Thus the third model is superior under the Akaike information criterian.

(i) Fit quasi Poisson versions of the two previous models and repeat the comparison.

```r
#model with no interaction effects
modelfit4 <- glm(incidents ~ period + year + type, family = quasipoisson(link = "log"),
            data = cleanships, offset = log(service))
summary(modelfit4)
```

```
##
## Call:
## glm(formula = incidents ~ period + year + type, family = quasipoisson(link = "log"),
##     data = cleanships, offset = log(service))
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.5348  -0.9319  -0.3686   0.4654   2.8833
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -10.07908    1.36829  -7.366 6.35e-08 ***
## period        0.02370    0.01264   1.876   0.0715 .
## year          0.04225    0.02003   2.109   0.0443 *
## typeB        -0.54609    0.27863  -1.960   0.0604 .
## typeC        -0.63263    0.51458  -1.229   0.2295
## typeD        -0.23226    0.44974  -0.516   0.6098
## typeE         0.40597    0.36690   1.107   0.2783
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for quasipoisson family taken to be 2.438934)
##
##     Null deviance: 146.328  on 33  degrees of freedom
## Residual deviance:  59.375  on 27  degrees of freedom
## AIC: NA
##
## Number of Fisher Scoring iterations: 5
```

```
#model with interacton effects
modelfit5 <-  glm(incidents ~ (type + year + period)^2, data = cleanships, family = quasipoisson(link =
summary(modelfit5)
```

```
##
## Call:
## glm(formula = incidents ~ (type + year + period)^2, family = quasipoisson(link = "log"),
##     data = cleanships)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.5599  -1.7315  -0.2022   0.6934   3.4047
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -8.2059461 20.4138758  -0.402   0.6924
## typeB        15.9400228  6.4500394   2.471   0.0237 *
## typeC         7.8666610 11.7094385   0.672   0.5102
## typeD        -5.0890756 12.3807857  -0.411   0.6859
## typeE         9.9483213  8.5534776   1.163   0.2600
## year          0.1094585  0.3004308   0.364   0.7199
## period        0.0061575  0.2762598   0.022   0.9825
## typeB:year   -0.1781189  0.0880348  -2.023   0.0582 .
## typeC:year   -0.0313661  0.1784830  -0.176   0.8625
## typeD:year    0.0145165  0.1552573   0.093   0.9265
## typeE:year   -0.1486057  0.1319298  -1.126   0.2748
```

```
## typeB:period -0.0289428  0.0678055  -0.427   0.6746
## typeC:period -0.1003367  0.1196822  -0.838   0.4128
## typeD:period  0.0434932  0.1398779   0.311   0.7594
## typeE:period  0.0024020  0.0926317   0.026   0.9796
## year:period   0.0004291  0.0039691   0.108   0.9151
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for quasipoisson family taken to be 5.565217)
##
##     Null deviance: 614.54  on 33  degrees of freedom
## Residual deviance: 109.75  on 18  degrees of freedom
## AIC: NA
##
## Number of Fisher Scoring iterations: 7
```

```r
#model comparison
anova(modelfit4, modelfit5,test = "F")
```

```
## Analysis of Deviance Table
##
## Model 1: incidents ~ period + year + type
## Model 2: incidents ~ (type + year + period)^2
##   Resid. Df Resid. Dev Df Deviance F Pr(>F)
## 1        27     59.375
## 2        18    109.748  9  -50.373
```

With observe a p-value of .2454 between our comparisons of our two quasi poisson models, this indicates that at a .05 alpha level, we fail to reject the null hyphothesis of the main effects models being better. Thus we concolude main effects "modelfit4" quasi poisson is prefered. We note that there exists

(j) Interpret the coefficients of the main effects of the quasi-Poisson model. What factors are associated with higher and lower rates of damage incidents?

```r
#for coefficients tpe b and e
exp(0.32558 - (-0.54334))
```

```
## [1] 2.384334
```

```r
#periods
exp(0.38447)
```

```
## [1] 1.468836
```

Given the information above, we observe that boates that are of type B and have lower indident rates compared to of those that are from type E and D. Based on the data, we know that that type E boats are 2.38 (rounded value) or about twice as likely to get into an incident that ships of type B.

we observe that the rate of incident increases by 1.467, meaning that ships built after 1964 and before 1974 have higher chance of incident, where ships built before have lower insident rates. This is perhaps because older ships were perhaps easier to navigate/maintian since the technology was well known, wherease newer ships with newer tech are harder to maintain since not many people have expertise with recent tech by nature.

2)

```r
 # not the same as the S-PLUS dataset
select <- MASS::select  #needed to define select here since tidyverse and Mass interfere with each othe
longley
```

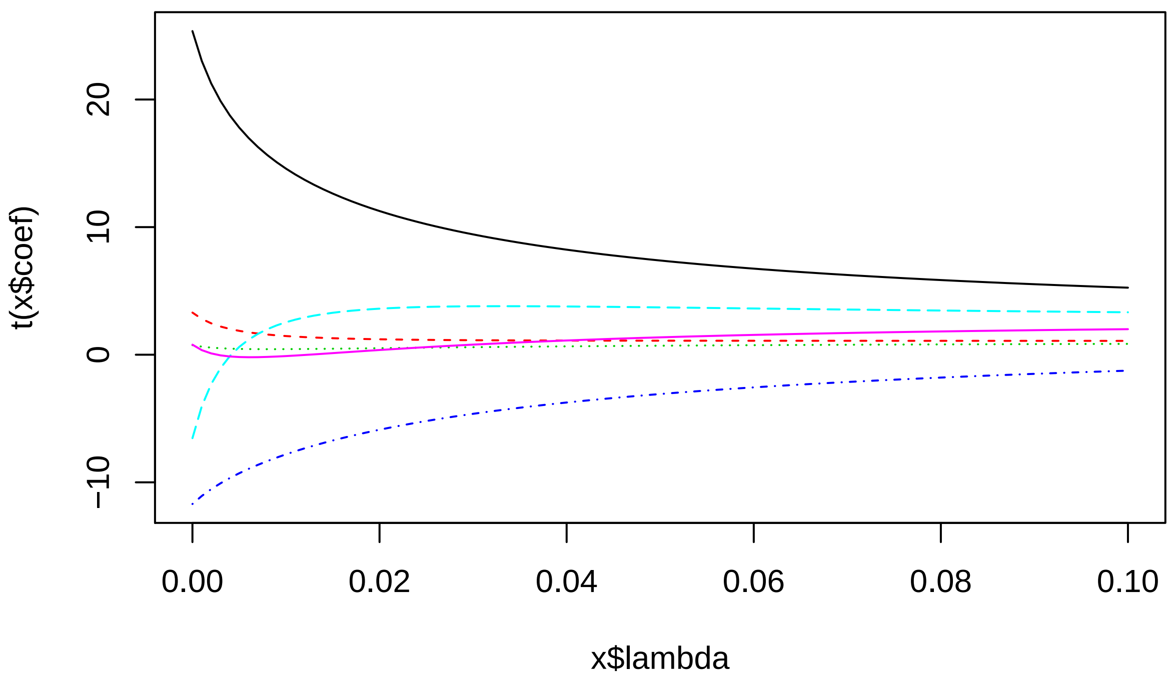
```
##      GNP.deflator     GNP Unemployed Armed.Forces Population Year Employed
```

```
## 1947         83.0 234.289      235.6        159.0      107.608 1947    60.323
## 1948         88.5 259.426      232.5        145.6      108.632 1948    61.122
## 1949         88.2 258.054      368.2        161.6      109.773 1949    60.171
## 1950         89.5 284.599      335.1        165.0      110.929 1950    61.187
## 1951         96.2 328.975      209.9        309.9      112.075 1951    63.221
## 1952         98.1 346.999      193.2        359.4      113.270 1952    63.639
## 1953         99.0 365.385      187.0        354.7      115.094 1953    64.989
## 1954        100.0 363.112      357.8        335.0      116.219 1954    63.761
## 1955        101.2 397.469      290.4        304.8      117.388 1955    66.019
## 1956        104.6 419.180      282.2        285.7      118.734 1956    67.857
## 1957        108.4 442.769      293.6        279.8      120.445 1957    68.169
## 1958        110.8 444.546      468.1        263.7      121.950 1958    66.513
## 1959        112.6 482.704      381.3        255.2      123.366 1959    68.655
## 1960        114.2 502.601      393.1        251.4      125.368 1960    69.564
## 1961        115.7 518.173      480.6        257.2      127.852 1961    69.331
## 1962        116.9 554.894      400.7        282.7      130.081 1962    70.551
```

```r
names(longley)[1] <- "y"
lm.ridge(y ~ ., longley)
```

```
##                          GNP    Unemployed  Armed.Forces     Population
## 2946.85636017   0.26352725    0.03648291    0.01116105    -1.73702984
##          Year       Employed
##   -1.41879853    0.23128785
```

```r
plot(lm.ridge(y ~ ., longley,
              lambda = seq(0,0.1,0.001)))
```




```r
select(lm.ridge(y ~ ., longley,
               lambda = seq(0,0.1,0.0001)))
```

```
## modified HKB estimator is 0.006836982
## modified L-W estimator is 0.05267247
## smallest value of GCV  at 0.0057
```

(a) Write the model that is being fitted (with assumptions).

$$Yi$$

= the ith observation for GNP implicit price deflator (1954=100) where

$$i = 1, ..., 16$$

, for ith obs/row

The Gross National Prouct, (GNP), is denoted by

$$x_{i2}$$

The nunmber of unemployed (unemployed) is denoted by

$$x_{i3}$$

Number of people in armed forces (Armed.Forces) is denoted by

$$x_{i4}$$

noninstitutionalized' population greater or equal to 14 years of age. (population) denoted by

$$x_{i5}$$

the year (time) as (Year) denoted by

$$x_{i6}$$

The numberof people emoployed (Employed) denoted by

$$x_{i7}$$

Our regression model tries to predict/model the number of people employed (Employed), thus the model take the form

$$Y_i = \beta_0 + \sum_{j=1}^{7} B_j x_{ij} + \epsilon_i$$

where $\epsilon \sim N(0, \sigma^2)$ are iid which follows a normal distribution

(b) Write a brief explanation of the patterns you observe in this plot, as

$$\lambda$$

changes, relative to the OLS estimators.

We notice that the estimated coeffeint converges to 0 as

$$\lambda$$

approaches

$$\infty$$

. Why?

When we observe

$$\hat{\beta}$$

(red dashed line and black line) that it gets closer and closer to zero as the value of

$$\lambda$$

increases. This is because the estimates of Beta ridge gets smaller than OLS estimatros. Inversely, we notice that

$$\hat{\beta_{ridge}}$$

in pink and blue get bigger than OLS estimatres. We also notice that for the green line that as the value of

$$\lambda$$

increases, that $\hat{\beta_{ridge}}$ is approximately equal to OLS estimates.