# NCSU ST 503 HW 12

Probem 5.1,8.5,8.6 Faraway, Julian J. Extending the Linear Model with R CRC Press.

*Bruce Campbell*

*21 November, 2017*

---

## # 1 from Chapter 5.

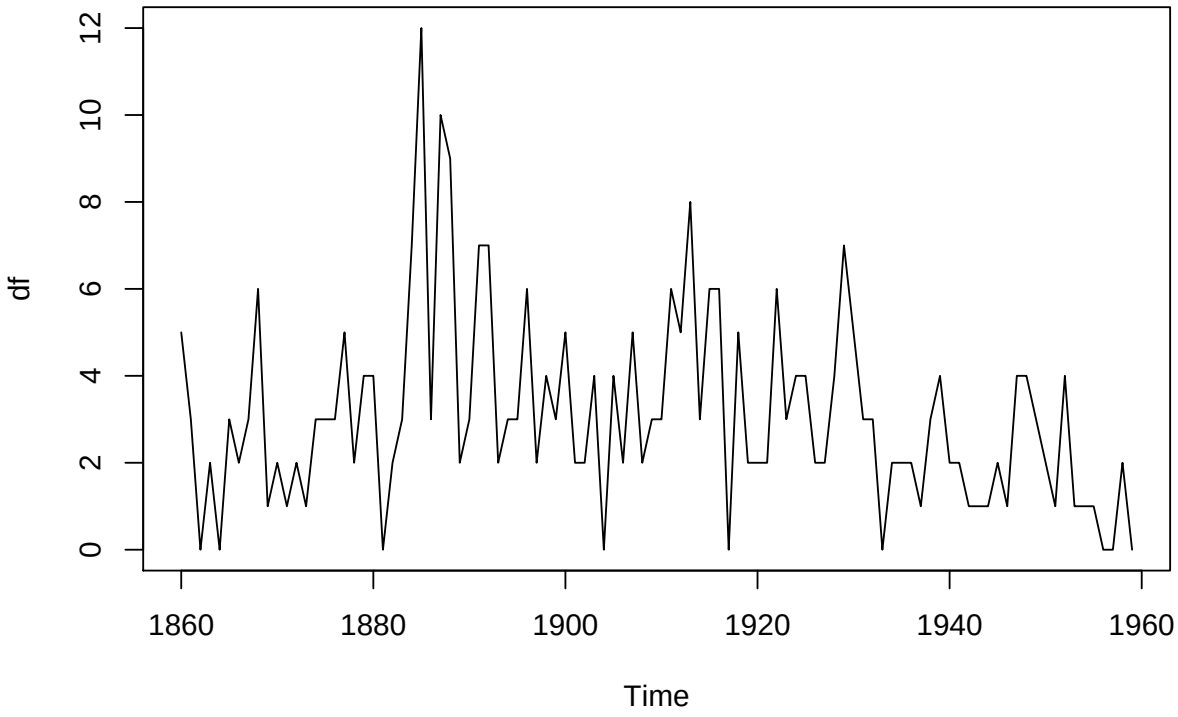
Complete exercises # 5 (a - d) and # 6 (a, b, d - f)) from Chapter 8

## 5.1 discoveries analysis

The dataset discoveries lists the numbers of "great" inventions and scientific discoveries in each year from 1860 to 1959.

### (a) Plot the discoveries over time and comment on the trend, if any.

```
rm(list = ls())
library(faraway)
data("discoveries", package="faraway")
df <- discoveries
plot(df)
```




**(b) Fit a Poisson response model with a constant term. Now compute the mean number of discoveries per year. What is the relationship between this mean and the coefficient seen in the model?**

```
ddf <- data.frame(df)
model.pois <- glm(df ~ 1, family=poisson, ddf)
sumary(model.pois)
```

```
##             Estimate Std. Error z value  Pr(>|z|)
## (Intercept) 1.131402   0.056796   19.92 < 2.2e-16
##
## n = 100 p = 1
## Deviance = 164.68460 Null Deviance = 164.68460 (Difference = 0.00000)
```

```
pander(data.frame(lambda.hat =mean(df)))
```

| lambda.hat |
|---|
| 3.1 |

Our model is $Y_i \sim Pois(\mu_i)$ , $log(\mu_i) = x^\intercal \beta$ and we have that $e^{\beta_0} \sim \hat{\lambda}$

**(c) Use the deviance from the model to check whether the model fits the data. What does this say about whether the rate of discoveries is constant over time?**

The deviance $D \sim \chi^2_{n-1}$ is significant and we conclude the null model is not a good fit for the data. We can conclude - and see in the plot - where the rate changes over time.

**(d) Make a table of how many years had zero, one, two, three, etc. discoveries. Collapse eight or more into a single category. Under an appropriate Poisson distribution, calculate the expected number of years with each number of discoveries. Plot the observed against the expected using a different plotting character to denote the number of discoveries. How well do they agree?**

```
tbl <- table(df)
tt <- tbl[1:9]
sumover8 <- sum(tbl[9:length(tbl)])
tt[9] <- sumover8

pander(tt, caption = "freqss")
```

Table 2: freqss

| 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
|---|---|---|---|---|---|---|---|---|
| 9 | 12 | 26 | 20 | 12 | 7 | 6 | 4 | 4 |

```
propo <- tt /sum(tt)

pander(propo, caption = "proportion")
```

Table 3: proportion

| 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
|---|---|---|---|---|---|---|---|---|
| 0.09 | 0.12 | 0.26 | 0.2 | 0.12 | 0.07 | 0.06 | 0.04 | 0.04 |

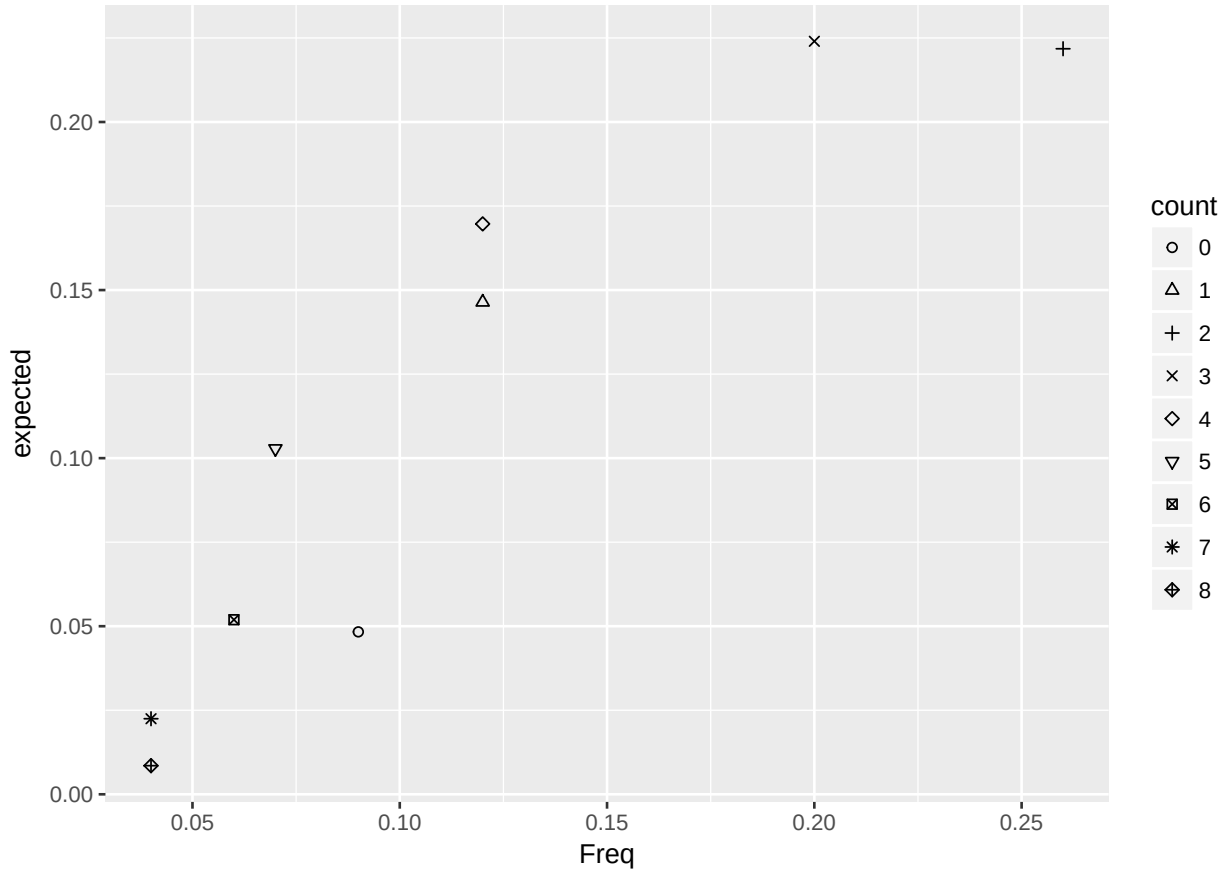
```
lambda <- sum(0:8 * propo)

expected <- dpois(0:8,lambda = lambda)

#n=1000000
```

```
#sum(0:n * dpois(0:n,lambda = lambda))

pander(data.frame(t(expected)), caption = "expected")
```

Table 4: expected

| X1 | X2 | X3 | X4 | X5 | X6 | X7 | X8 | X9 |
|---|---|---|---|---|---|---|---|---|
| 0.04832 | 0.1464 | 0.2218 | 0.224 | 0.1697 | 0.1028 | 0.05193 | 0.02248 | 0.008514 |

```
dfp <-data.frame(count=as.factor(0:8),propo,expected)

(p <- ggplot(dfp, aes(x=Freq, y=expected, shape=count)) +geom_point()+ scale_shape_manu
```




**(e) Use the Pearson's Chi-squared test to check whether the observed numbers are consistent with the expected numbers. Interpret the result.**

We have to deal with the fact that we binned the counts $9 : \infty$

```
observed <- tt

elast <- 1- sum(dpois(0:8,lambda = lambda))

ee <- expected
ee[8]<- elast
expected.counts <- ee*sum(tt)

ctbl <-data.frame(observed,expected.counts)
ctbl$df <- NULL
chisq.test(ctbl)
```

```
##
##  Pearson's Chi-squared test
##
## data:  ctbl
## X-squared = 8.4433, df = 8, p-value = 0.3914
```

We have evidence that the observed numbers are cinsistent with the expected numbers

**(f) Fit a Poisson response model that is quadratic in the year. Test for the signifiance of the quadratic term. What does this say about the presence of a trend in discovery?**

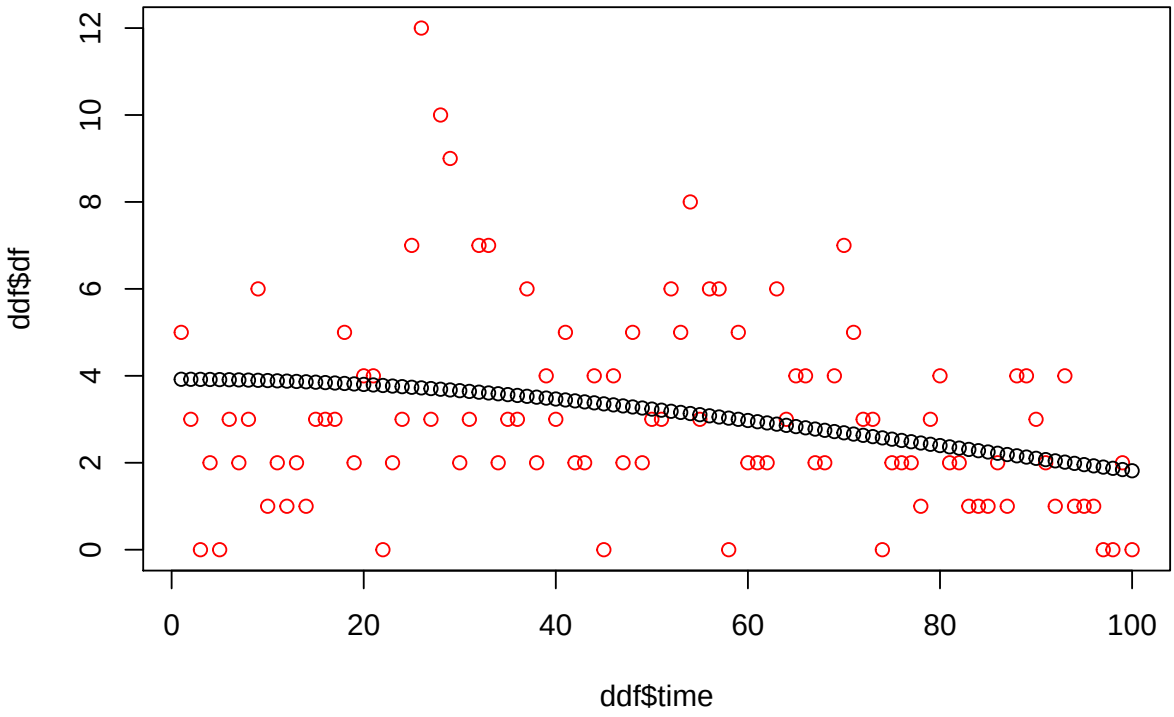
```
ddf <- data.frame(time=1:length(df),df)
model.pois <- glm(df ~ I(time^2), family=poisson, ddf)
sumary(model.pois)
```

```
##                Estimate  Std. Error z value  Pr(>|z|)
## (Intercept)  1.3664e+00  8.0153e-02 17.0476 < 2.2e-16
## I(time^2)   -7.6975e-05  2.0628e-05 -3.7315 0.0001903
##
## n = 100 p = 2
## Deviance = 149.82413 Null Deviance = 164.68460 (Difference = 14.86047)
```

This model confirms our observation that the rate changes

**(g) Compute the predicted number of discoveries each year and show show these predictions as a line drawn over the data. Comment on what you see.**

```
plot(ddf$time,ddf$df,col=2)
points(ddf$time,fitted(model.pois),col=1)
```




We note the rate does change over time in concordance with the data except for a small period of exceptional discovery early in the series.

## 8.5 Galapagos

Again using the Galápagos data, fit a Poisson model to the species response with the five geographic variables as predictors. Do not use the endemics variable. The purpose of this question is to compare six different ways of testing the significance of the elevation predictor, i.e., $H0 : \beta_{Elev} = 0$. In each case, report the p-value.

### (a) Use the z-statistic from the model summary.

```
rm(list = ls())
library(faraway)
data("gala", package="faraway")
df <- gala
model.pois <- glm( Species~ Area+Elevation+Nearest+Scruz+Adjacent, family=poisson, df)
sumary(model.pois)
```

```
##                Estimate  Std. Error  z value  Pr(>|z|)
## (Intercept)  3.1548e+00  5.1750e-02  60.9630 < 2.2e-16
## Area        -5.7994e-04  2.6273e-05 -22.0737 < 2.2e-16
## Elevation    3.5406e-03  8.7407e-05  40.5070 < 2.2e-16
## Nearest      8.8256e-03  1.8213e-03   4.8459 1.261e-06
## Scruz       -5.7094e-03  6.2562e-04  -9.1260 < 2.2e-16
## Adjacent    -6.6303e-04  2.9328e-05 -22.6078 < 2.2e-16
##
## n = 30 p = 6
## Deviance = 716.84577 Null Deviance = 3510.72862 (Difference = 2793.88284)
```

The p-value is < 2.2e-16

**(b) Fit a model without elevation and use the difference in deviances to make the test.**

```
model.pois.reduced <- glm( Species~ Area+Nearest+Scruz+Adjacent, family=poisson, df)
sumary(model.pois.reduced)
```

```
##                Estimate  Std. Error  z value Pr(>|z|)
## (Intercept)  4.3447e+00  3.1137e-02 139.5352   <2e-16
## Area         4.1901e-04  1.2043e-05  34.7930   <2e-16
## Nearest      2.0805e-02  1.6999e-03  12.2390   <2e-16
## Scruz       -6.7806e-03  5.4561e-04 -12.4275   <2e-16
## Adjacent    -2.5642e-06  2.9390e-05  -0.0872   0.9305
##
## n = 30 p = 5
## Deviance = 2389.56888 Null Deviance = 3510.72862 (Difference = 1121.15974)
```

Our test statistic is $\frac{(2389.56888-716.84577)}{\hat{\phi}}$

We need to estimate $\phi$

```
(dp <- sum(residuals(model.pois,type="pearson")^2)/model.pois$df.res)
```

```
## [1] 31.74914
```

**(c) Use the Pearson Chi-squared statistic in place of the deviance in the previous test.**

```
anova(model.pois,test="Chi")
```

```
## Analysis of Deviance Table
##
## Model: poisson, link: log
```

```
##
## Response: Species
##
## Terms added sequentially (first to last)
##
##
##           Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
## NULL                         29     3510.7
## Area       1   895.14        28     2615.6 < 2.2e-16 ***
## Elevation  1   802.05        27     1813.5 < 2.2e-16 ***
## Nearest    1    15.71        26     1797.8 7.378e-05 ***
## Scruz      1   456.37        25     1341.5 < 2.2e-16 ***
## Adjacent   1   624.61        24      716.8 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Again the p-value for *Elevation* is `< 2.2e-16`

**(d) Fit the Poisson model with a free dispersion parameter as described in Section 5.2. Make the test using the model summary.**

```
sumary(model.pois,dispersion=dp)
```

```
##                Estimate  Std. Error z value  Pr(>|z|)
## (Intercept)  3.15480788  0.29158975 10.8193 < 2.2e-16
## Area        -0.00057994  0.00014804 -3.9175 8.947e-05
## Elevation    0.00354059  0.00049251  7.1889 6.530e-13
## Nearest      0.00882557  0.01026214  0.8600    0.3898
## Scruz       -0.00570942  0.00352514 -1.6196    0.1053
## Adjacent    -0.00066303  0.00016525 -4.0123 6.013e-05
##
## Dispersion parameter = 31.74914
## n = 30 p = 6
## Deviance = 716.84577 Null Deviance = 3510.72862 (Difference = 2793.88284)
```

The p-value for *Elevation* is `6.530e-13` so the pvalue is 0.001511532

**(e) Use the sandwich estimation method for the standard errors in the original model. Use these to compute z-statistics.**

```
library(sandwich)

(sebeta <- sqrt(diag(vcovHC(model.pois))))
```

```
##  (Intercept)         Area    Elevation      Nearest        Scruz
## 0.3185550946 0.0012176343 0.0011939774 0.0228021845 0.0065382200
##     Adjacent
## 0.0006212657
```

our z-value is `3.5406e-03/0.0011939774` which yields a pvalue of

**(f) Use the robust GLM estimation method and report the test result from the summary.**

```
library(robust)
rmodpla <- glmRob(Species~ log(Area)+log(Elevation)+Nearest+Scruz+Adjacent, family=poiss
summary(rmodpla)
```

```
##
## Call: glmRob(formula = Species ~ log(Area) + log(Elevation) + Nearest +
##     Scruz + Adjacent, family = poisson, data = gala)
## Deviance Residuals:
##        Min         1Q     Median         3Q        Max
## -5.420e+31  1.914e+01  4.374e+01  9.263e+01  1.911e+02
##
## Coefficients:
##                Estimate Std. Error    z value Pr(>|z|)
## (Intercept)    -7647.58     0.3972 -19254.477   0.0000
## log(Area)       -781.53     0.1334  -5856.749   0.0000
## log(Elevation)   878.42     0.8083   1086.707   0.0000
## Nearest         -371.08     5.3795    -68.980   0.0000
## Scruz             50.85    45.7806      1.111   0.2666
## Adjacent         598.49     0.4496   1331.141   0.0000
##
## (Dispersion Parameter for poisson family taken to be 1 )
##
##     Null Deviance: 21190 on 29 degrees of freedom
##
## Residual Deviance: NaN on 24 degrees of freedom
##
## Number of Iterations: 50
##
## Correlation of Coefficients:
##                (Intercept) log(Area) log(Elevation) Nearest Scruz
## log(Area)      2.5177
## log(Elevation) 0.3972      1.0000
## Nearest        0.3972      1.0000    1.0000
## Scruz          0.3972      1.0000    1.0000         1.0000
```

```
## Adjacent       0.3972      1.0000    1.0000          1.0000  1.0000
```

No clue why this would not converge unless I log transformed elevation and area!

**(g) Compare all six results. Pick the best one and justify your choice.**

unfinished :(