S632 HW4

Erik Parker

Feburary 24th, 2018

1. ELM 5.1: The dataset *discoveries* lists the numbers of "great" inventions and scientific discoveries in each year from 1860 to 1959.
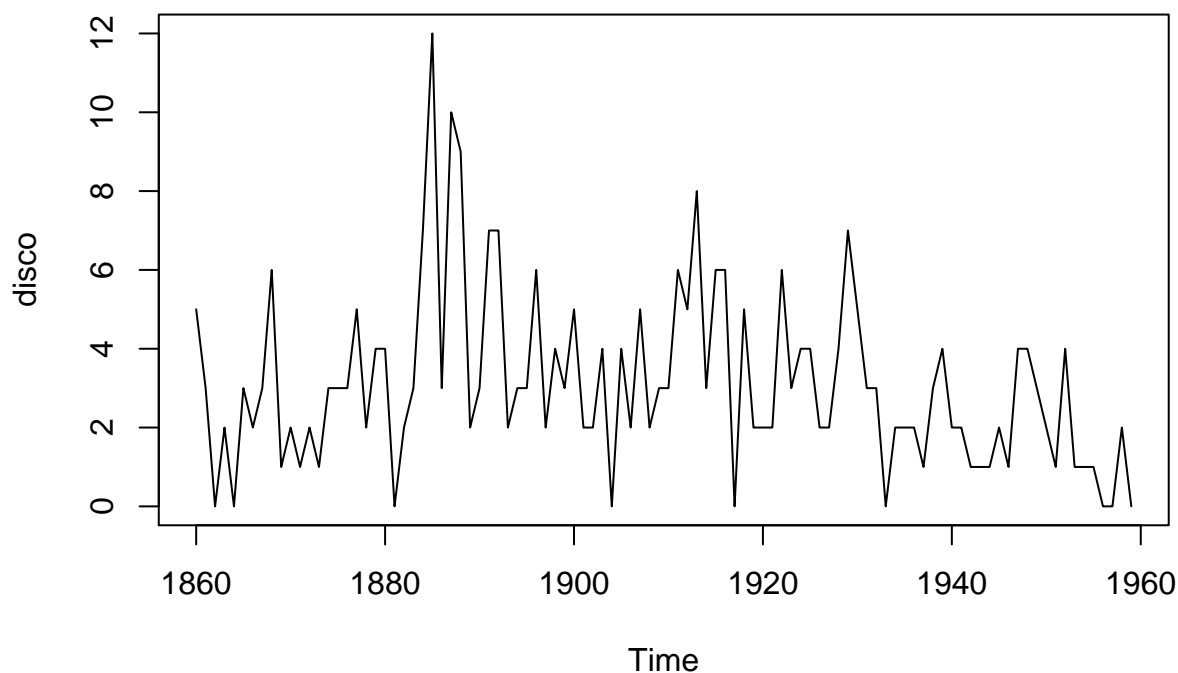
```
rm(list = ls())

library(ggplot2)
library(faraway)
library(dplyr)
library(pscl)
library(alr4)

disco <- discoveries
```

a) Plot the discoveries over time and comment on the trend.

```
plot.ts(disco)
```


There isn't a super well defined trend to the number of discoveries made over time, though there does seem to be a slight spike in the number around the late 1880's and early 1890's, followed by a decrease back to the baseline rate of around 2-6 discoveries per year. Overall though, there is no general trend of increase or decrease over the entire range.

b) Fit a poisson response model with a constant term. Then compute the mean number of discoveries per year. What is the relationship between the mean and coefficient seen in the

model?

```r
mean(disco)
```

```
## [1] 3.1
```

```r
m1 <- glm(disco ~ 1, family = poisson)

exp(m1$coefficients)
```

```
## (Intercept)
##         3.1
```

> The general interpretation of the coefficient seen in this model, μ change by a factor of e^{β_n} when X_n increases by 1 unit, can just be summarized as e^{β_0} because all we have is an intercept in this model. And then we see that the mean number of discoveries each year is the same as the coefficient obtained.

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

```r
1 - pchisq(deviance(m1), df.residual(m1))
```

```
## [1] 3.79455e-05
```

> This test of fit results in a very small p-value well below 0.05, allowing us to conclude that this "null" model with only a constant term does not fit sufficiently well. This tells us that the rate of discoveries can't be considered constant over time.

d) Make a table of how many years had 1,2,3,etc discoveries. Collapse eight or more into a single category. Under poisson, calculate the expected number of years with each number of discoveries. Plot the observed against the expected, how well do they agree?

```r
t1 <- table(disco)

t1 <- as.data.frame(t1)

sum(t1[9:12, ]$Freq)
```

```
## [1] 4
```

```r
t2 <- c(8, 4)

t1 <- rbind(t1, t2)

t1 <- t1[-(9:12), ]

expected <- c(dpois(0:7, 3.1) * 100, ppois(7, 3.1, lower.tail = FALSE) * 100)

te <- as.data.frame(expected)

colnames(te) <- "Freq"

te$disco <- c(0, 1, 2, 3, 4, 5, 6, 7, 8)

combined <- rbind(t1, te)
```
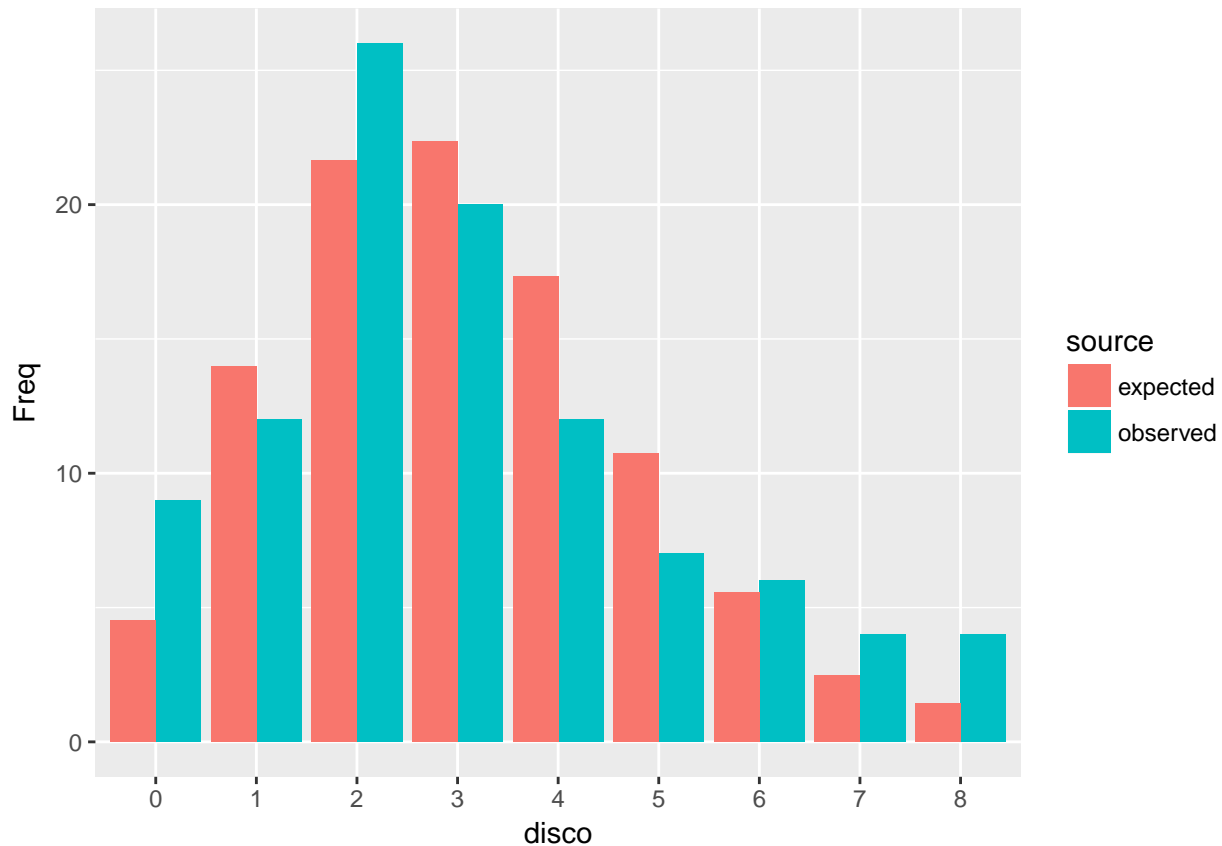
```
combined$source <- rep(c("observed", "expected"), each = 9)

ggplot(combined, aes(x = disco, y = Freq, fill = source)) + geom_col(position = position_dodge())
```


The observed and expected distributions agree relatively well. There were more observations observed at both the extreme high, and low ends, but in general these two distribuions are quite similar.

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

```
chisq.test(x = t1$Freq, p = c(dpois(0:7, 3.1), ppois(7, 3.1, lower.tail = FALSE)))
```

```
##
##  Chi-squared test for given probabilities
##
## data:  t1$Freq
## X-squared = 14.517, df = 8, p-value = 0.06924
```

This chi-squared test returns a p-value of 0.069, which if we consider an alpha of 0.05 allows us to barely conclude that the null hypothesis can't be rejected and so we say that the observed and expected distributions aren't significantly different from one another.

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

3

```
discover <- as.data.frame(disco)

colnames(discover) <- "discoveries"

discover$year <- seq.int(1860, 1959, 1)

m2 <- glm(discoveries ~ year + I(year^2), family = poisson, discover)

sumary(m2)
```

```
##               Estimate  Std. Error z value  Pr(>|z|)
## (Intercept) -1.4822e+03  3.1634e+02 -4.6855 2.793e-06
## year         1.5610e+00  3.3179e-01  4.7048 2.541e-06
## I(year^2)   -4.1061e-04  8.6989e-05 -4.7203 2.355e-06
##
## n = 100 p = 3
## Deviance = 132.83843 Null Deviance = 164.68460 (Difference = 31.84618)
```

```
Anova(m2)
```

```
## Analysis of Deviance Table (Type II tests)
##
## Response: discoveries
##           LR Chisq Df Pr(>Chisq)
## year        24.302  1  8.236e-07 ***
## I(year^2)   24.477  1  7.519e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Above we can see both that the estimated coefficient for the quadratic term is highly significant when added to a model already containing the intercept and non-quadratic year, and that the quadratic year term is significant from the Anova command - telling us that the reduced model lacking the quadratic term is not sufficient to explain the response. The significance of the quadratic year term tells us that there is a relationship between the year and the discovery rate, so a quadratic trend in discovery is present.
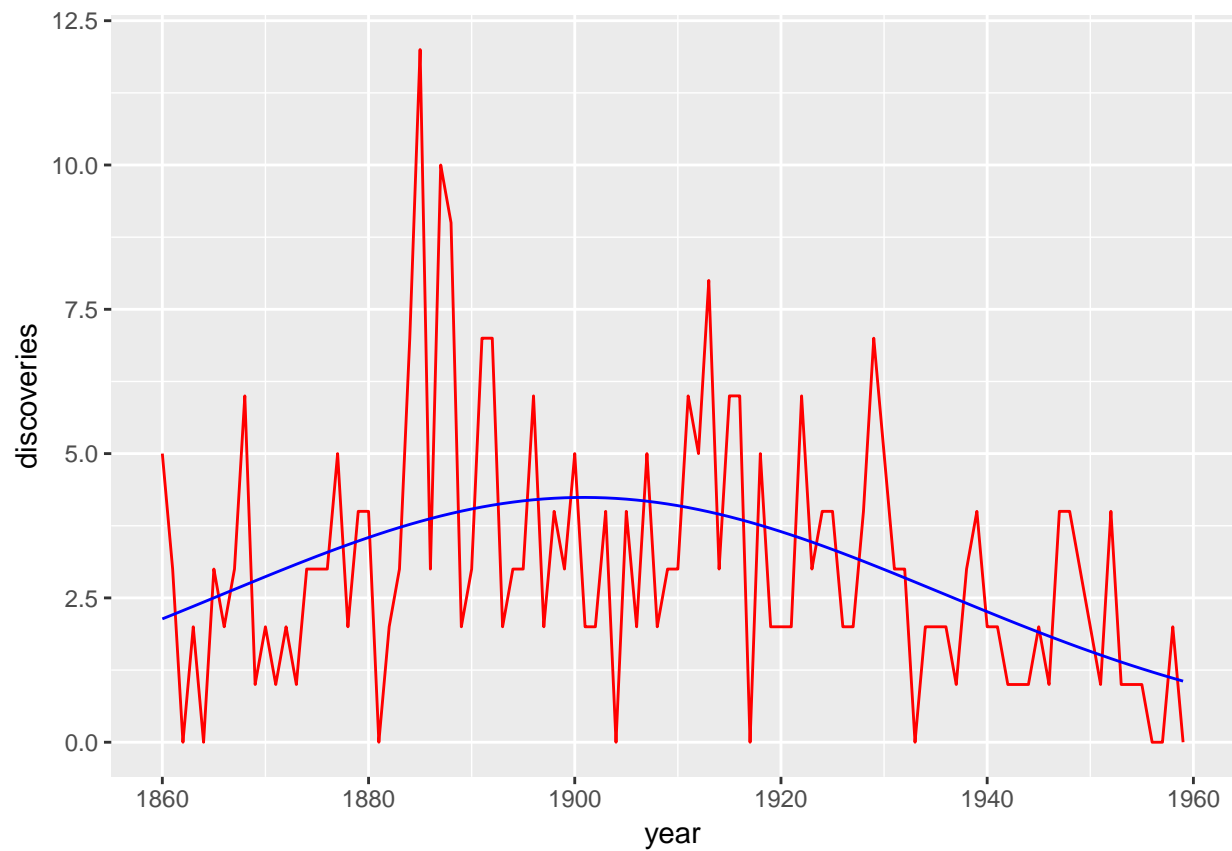
g) Compute the predicted number of discoveries each year and show these predictions as a line drawn over the data. Comment on what you see.

```
predicted <- predict(m2, type = "response")

predicted <- as.data.frame(predicted)

predicted$year <- seq.int(1860, 1959, 1)

ggplot() + geom_line(data = discover, aes(x = year, y = discoveries), color = "red") +
    geom_line(data = predicted, aes(x = year, y = predicted), color = "blue") +
    xlab("year") + ylab("discoveries")
```

The predicted number of discoveries each year, as determined by the quadratic Poisson model from part f, fits the observed data quite well overall. The quadratic hump of the predicted line is a little bit off, but it still closely mirrors the increase in discovery rates seen in the late 1880's, as noticed back in part a.