Assignment 2

STATS 3860B/9155B

Winter 2020

Question 1

Dataset *seeds*. A Biologist analyzed an experiment to determine the effect of moisture content on seed germination. Eight boxes of 100 seeds each were treated with the same moisture level. 4 boxes were covered and 4 left uncovered. The process was repeated at 6 different moisture levels (nonlinear scale). The data were ordered in blocks of 6 observations per box.

```
library(faraway)
data(seeds)
## creating a new predictor describing the box:
seeds$box <- factor(x=rep(1:8, c(6,6,6,6,6,6,6,6)),
                    levels=c("1","2","3","4","5","6","7","8"))
## removing one observation with missing data
(seeds[is.na(seeds$germ),])
```

```
##    germ moisture covered box
## 47   NA        9     yes   8
```

```
seeds <- seeds[!is.na(seeds$germ),]
str(seeds)
```

```
## 'data.frame':    47 obs. of  4 variables:
##  $ germ    : num  22 41 66 82 79 0 25 46 72 73 ...
##  $ moisture: num  1 3 5 7 9 11 1 3 5 7 ...
##  $ covered : Factor w/ 2 levels "no","yes": 1 1 1 1 1 1 1 1 1 1 ...
##  $ box     : Factor w/ 8 levels "1","2","3","4",..: 1 1 1 1 1 1 2 2 2 2 ...
```

a) The response variable *germ* contains the number of seeds that germinated out of 100. Fit a binomial regression model including *box* and *moisture* as predictors.

```
qa_fit <- glm(cbind(germ,100 - germ) ~ box + moisture, family = binomial, seeds )
summary(qa_fit)
```

```
##
## Call:
## glm(formula = cbind(germ, 100 - germ) ~ box + moisture, family = binomial,
##     data = seeds)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -9.5285  -5.5046   0.7063   5.0465   7.9405
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.593872   0.098319   6.040 1.54e-09 ***
## box2        -0.041493   0.117609  -0.353    0.724
## box3        -0.041493   0.117609  -0.353    0.724
## box4         0.020724   0.117544   0.176    0.860
## box5        -0.090012   0.117697  -0.765    0.444
## box6        -0.062269   0.117643  -0.529    0.597
## box7        -0.069200   0.117655  -0.588    0.556    
## box8         0.067297   0.123410   0.545    0.586    
## moisture    -0.110487   0.008813 -12.537  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1791.0  on 46  degrees of freedom
## Residual deviance: 1624.4  on 38  degrees of freedom
## AIC: 1832.1
##
## Number of Fisher Scoring iterations: 5
```

b) Interpret the estimated coefficients of *moisture* and *box4*.

```
box_4 <- qa_fit$coeff[['box4']]
moisture <- qa_fit$coeff[['moisture']]

odds_box4 <- exp(box_4) - 1 # decrease
odds_moist <- 1 - exp(moisture) # increase
odds_box4
```

```
## [1] 0.02094064
```

```
odds_moist
```

```
## [1] 0.1046023
```

`moisture` - For a one unit change in moisture, the odds of germs will decrease by 10.5%. Given the other predictor variables in the model are held constant. `box4` - When the seeds where in the 4th box, the odds of germs will increase by 2.1%. Given the other predictor variables stay the same.

c) What are the two hypothesis tests we can use to assess the goodness of fit for the model in a)? Perform one of those tests. Is there statistical evidence for lack of fit?

- Pearson's X2 Statistics
- binomial deviance to test

```
pchisq(deviance(qa_fit), df.residual(qa_fit),lower=FALSE)
```

```
## [1] 6.957246e-317
```

Since this p-vlaue seems to be significantly less than than 0.05, we may reject the null hypothesis that there is no evidence for lack of fit.

d) What are the other common causes for a deviance value to be larger than expected besides over/under-dispersion?

- Wrong form of the model: not included right predictors or transformations
- Presence of outliers
- Sparse data

e) Suppose we have eliminated the causes listed in d) as the source of the problem, so that we can now put the blame on over/under-dispersion. Estimate the dispersion parameter and comment if the problem is over or underdispersion.

```
sigma2 <- sum(residuals(qa_fit,type="pearson")^2)/(47-9)
sigma2
```

```
## [1] 35.71223
```

Over dispersion seems to be the problem since the dispersion parameter > 1.

f) Test for the significance of the individual predictors (*moisture* and *box*) accounting for overdispersion.

```
drop1(qa_fit,scale=sigma2,test="F")
```

```
## Warning in drop1.glm(qa_fit, scale = sigma2, test = "F"): F test assumes
## 'quasibinomial' family

## Single term deletions
##
## Model:
## cbind(germ, 100 - germ) ~ box + moisture
##
## scale:  35.71223
##
##          Df Deviance    AIC F value  Pr(>F)
## <none>        1624.4 1832.2
## box       7   1627.0 1818.2  0.0086 1.00000
## moisture  1   1786.5 1834.7  3.7918 0.05892 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Neither predictors are significant.

g) Test for the significance of individual predictors (*moisture* and *box*) ignoring overdispersion. How do the results differ from e)?

```
drop1(qa_fit, test="Chi")
```

```
## Single term deletions
##
## Model:
## cbind(germ, 100 - germ) ~ box + moisture
##          Df Deviance    AIC    LRT Pr(>Chi)
## <none>        1624.4 1832.2
## box       7   1627.0 1820.7   2.57   0.9217
## moisture  1   1786.5 1992.2 162.09   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The Moisture predictor is significant, box is not.

Question 2

Parts a), b), c), f) and g) of Exercise 1 on Page 98 of the textbook. Dataset *discoveries*.

```
data("discoveries")
str(discoveries)
```

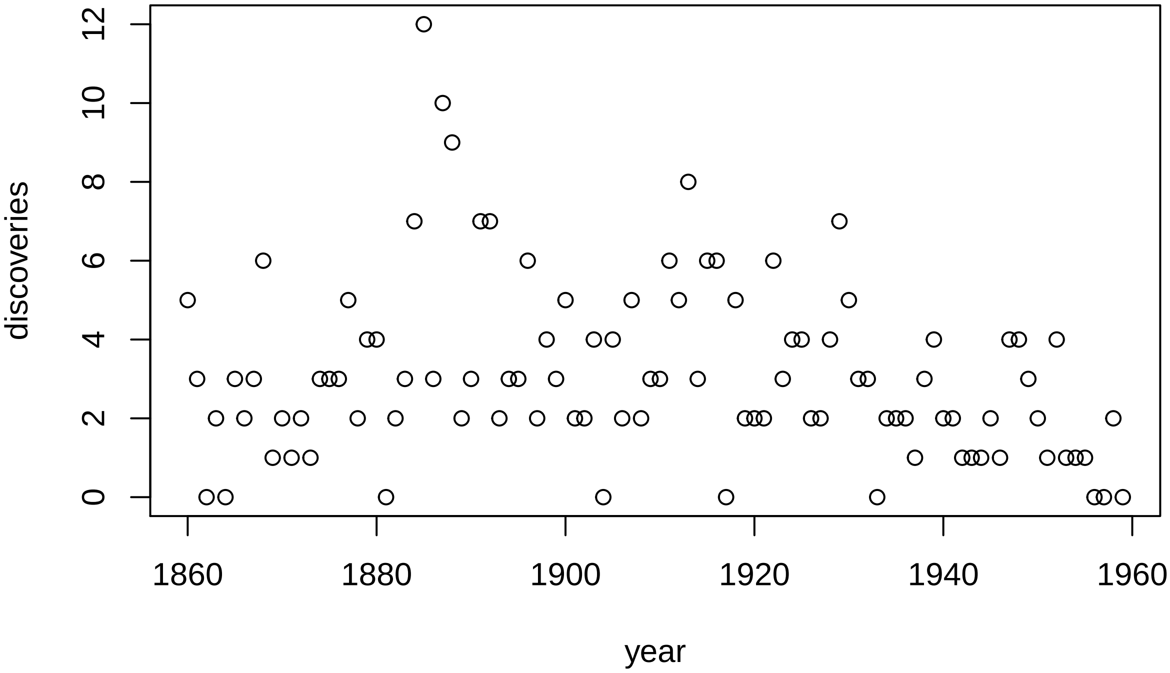
```
##  Time-Series [1:100] from 1860 to 1959: 5 3 0 2 0 3 2 3 6 1 ...
```

```
## creating a dataframe
discoveries <- as.data.frame(cbind(1860:1959,discoveries))
colnames(discoveries) <- c("year","discoveries")
str(discoveries)
```

```
## 'data.frame':    100 obs. of  2 variables:
##  $ year       : num  1860 1861 1862 1863 1864 ...
##  $ discoveries: num  5 3 0 2 0 3 2 3 6 1 ...
```

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

```
plot(discoveries)
```



It seems like the number of discoveries were trending upwards from 1860 to 1900. Then the opposite from 1900 onwards.

(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?

```
qb_fit <- glm(discoveries~1, family=poisson, data=discoveries)
sumary(qb_fit)
```

```
##             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)
```

```
mean(discoveries$discoveries)
```

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

When exponentiated, $e^{1.131402} = 3.1$ *which is the mean*

(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?

```
#plot(qb_fit)
pchisq(deviance(qb_fit), df.residual(qb_fit),lower=FALSE)
```

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

Since the p-value is <0.05 we will reject the null hypothesis. There *is* a lack of fit and the rate of discoveries are probably not constant over time.

(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?

```r
qf_fit <- glm(discoveries~year + I(year^2), family=poisson, discoveries)
#plot(discoveries)
#abline(coefficients(qf_fit))
sumary(qf_fit)
```

```
##                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)
```

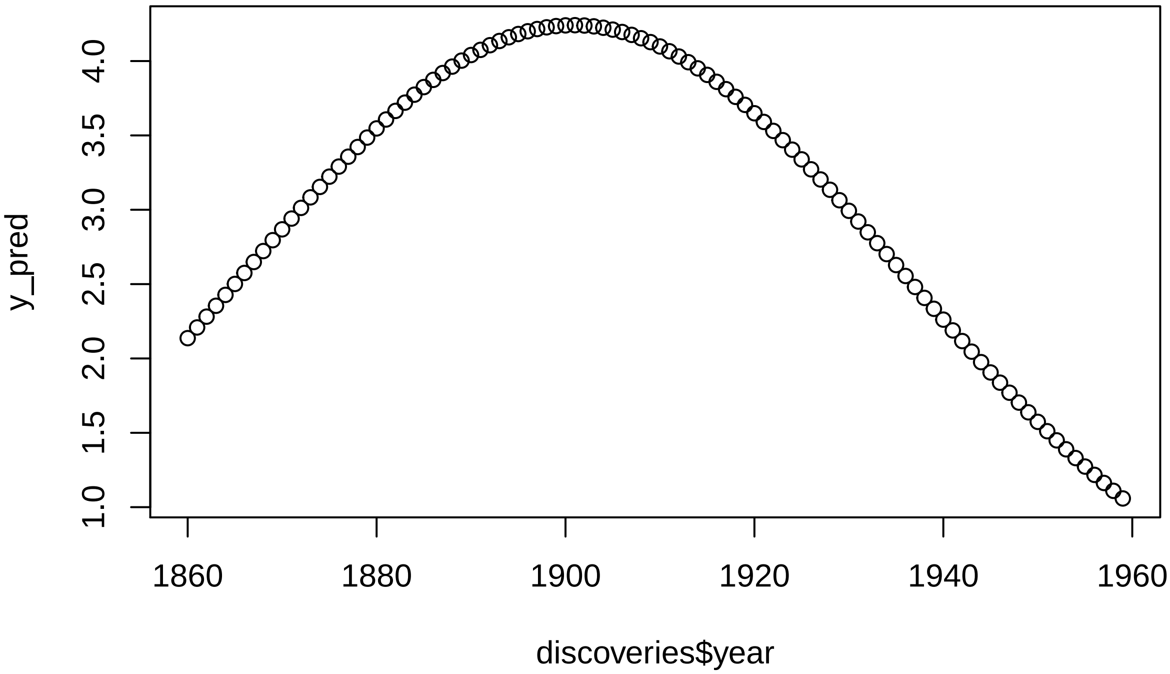
```r
anova(qf_fit, qb_fit, test="Chisq")
```

```
## Analysis of Deviance Table
##
## Model 1: discoveries ~ year + I(year^2)
## Model 2: discoveries ~ 1
##   Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
## 1        97     132.84
## 2        99     164.69 -2  -31.846 1.215e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

It appears the model including the quadtric terms is significantly impactful on the goodness of fit, thus the larger model is better. Due to a small p-vlaue, we will reject the null hypothesis. It appears that the relationship between discoveries(response) and years is more quadratic in nature than linear.

(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.

```r
y_pred = predict(qf_fit,type = "response",newdata = discoveries)
plot(y_pred~discoveries$year)
```



This graph shows that the model argues a negative quadratic relationship between discoveries and time. Which is quite sad if true and this extrapolated to today. What do we have to learn from the end of the 19th century?