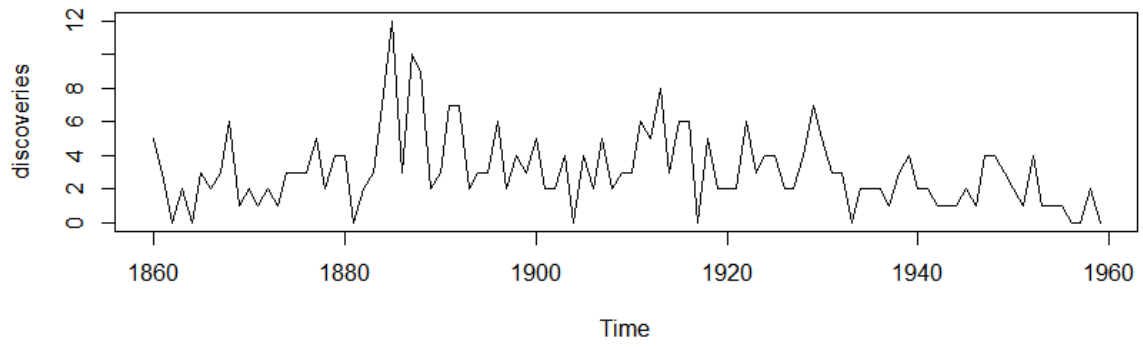
Sample solutions

Stat 8051 Homework 8

Problem 1: Faraway Exercise 3.1

A plot of the time series reveals kind of a fluctuating pattern:



Trying to fit poisson regression models yields a quadratic model if we only consider significant polynomial effects.

```
> n = as.numeric(discoveries)
> discoveries1 = data.frame(n)
> discoveries1$year = 1860:1959
>
> m1.pois = glm(n~poly(year,2), family=poisson, data=discoveries1)
> summary(m1.pois)

Call:
glm(formula = n ~ poly(year, 2), family = poisson, data = discoveries1)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.9066  -0.8397  -0.2544   0.4776   3.3303

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)    1.07207    0.06064  17.681  < 2e-16 ***
```

```
poly(year, 2)1 -2.04766    0.66937  -3.059  0.00222 **
poly(year, 2)2 -3.05975    0.64821  -4.720 2.35e-06 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 164.68  on 99  degrees of freedom
Residual deviance: 132.84  on 97  degrees of freedom
AIC: 407.85

Number of Fisher Scoring iterations: 5
```

There is not much overdispersion as we can see from the value of the dispersion parameter:

```
> (dp <- sum(residuals(m1.pois,type="pearson")^2)/m1.pois$df.res)
[1] 1.305649
```

The significant year effects indicate that the discovery rates vary significantly from constant with respect to year.

Note Doing the half-normal plot and testing for outliers using studentized residuals reveals point 26 as a potential outlier. Removing this point actually improves the quadratic fit (check).

Problem 2: Faraway Exercise 3.2

A poission model with linear effect of dose does very badly in terms of deviance:

```
> m2.pois = glm(colonies~dose, family=poisson, data=salmonella)
> summary(m2.pois)

Call:
glm(formula = colonies ~ dose, family = poisson, data = salmonella)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.6482  -1.8225  -0.2993   1.2917   5.1861

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) 3.3219950  0.0540292  61.485   <2e-16 ***
dose        0.0001901  0.0001172   1.622    0.105
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

```
(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 78.358  on 17  degrees of freedom
Residual deviance: 75.806  on 16  degrees of freedom
AIC: 172.34

Number of Fisher Scoring iterations: 4
```

The value of the dispersion parameter is very high, and we can actually check for its significance using the function `dispersiontest` from package `AER`.

```
> require(AER)
> dispersiontest(m2.pois, alternative="greater")

Overdispersion test

data:  m2.pois
z = 1.913, p-value = 0.02787
alternative hypothesis: true dispersion is greater than 1
sample estimates:
dispersion
  4.522293
```

Fitting a negative binomial model doesn't improve the fit. Instead we can try fitting polynomial link functions. Doing so reveals that a cubic fit gives all polynomial effects significant and decreases the deviance as well:

```
> m21.pois = glm(colonies~poly(dose,3), family=poisson, data=salmonella)
> summary(m21.pois)

Call:
glm(formula = colonies ~ poly(dose, 3), family = poisson, data = salmonella)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-2.43608  -0.85295  -0.07833   0.56028   2.65580

Coefficients:
               Estimate Std. Error z value Pr(>|z|)
(Intercept)      3.3300     0.0455  73.181  < 2e-16 ***
poly(dose, 3)1   0.3826     0.1903   2.011   0.0444 *
poly(dose, 3)2  -0.8648     0.1767  -4.893 9.91e-07 ***
poly(dose, 3)3   0.7745     0.1716   4.514 6.37e-06 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

```
(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 78.358  on 17  degrees of freedom
Residual deviance: 36.055  on 14  degrees of freedom
AIC: 136.59

Number of Fisher Scoring iterations: 4

> dispersiontest(m21.pois, alternative="greater")

Overdispersion test

data:  m21.pois
z = 1.808, p-value = 0.03531
alternative hypothesis: true dispersion is greater than 1
sample estimates:
dispersion
  2.030354
```

The dispersion parameter is still high, so let us now fit a cubic negative-binomial model:

```
> m21.nb = glm.nb(colonies~poly(dose,3), data=salmonella)
> summary(m21.nb)

Call:
glm.nb(formula = colonies ~ poly(dose, 3), data = salmonella,
    init.theta = 28.81281522, link = log)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-1.61674  -0.67164  -0.03553   0.37609   1.67190

Coefficients:
               Estimate Std. Error z value Pr(>|z|)
(Intercept)     3.33054    0.06321  52.687  < 2e-16 ***
poly(dose, 3)1  0.37991    0.26626   1.427 0.153626
poly(dose, 3)2 -0.85836    0.25686  -3.342 0.000832 ***
poly(dose, 3)3  0.75662    0.25465   2.971 0.002966 **
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

(Dispersion parameter for Negative Binomial(28.8128) family taken to be 1)

    Null deviance: 38.447  on 17  degrees of freedom
```

```
Residual deviance: 17.613  on 14  degrees of freedom
AIC: 132.43

Number of Fisher Scoring iterations: 1


              Theta:  28.8
          Std. Err.:  18.9

 2 x log-likelihood:  -122.434
```

Here also the higher powers of `dose` turn out to be significant.

Problem 3: Faraway Exercise 3.7

Here the number of complaints linearly depends on the number of visits, so we fit a rate model (`log(visits)` as offset). The summary is as given below:

```
Call:
glm(formula = complaints ~ residency + gender + revenue + hours +
    offset(log(visits)), family = poisson, data = esdcomp)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.9434  -0.9490  -0.3130   0.7859   1.8036

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -8.1202460  0.8502806  -9.550   <2e-16 ***
residencyY  -0.2090058  0.2011520  -1.039   0.2988
genderM      0.1954338  0.2181525   0.896   0.3703
revenue      0.0015761  0.0028294   0.557   0.5775
hours        0.0007019  0.0003505   2.002   0.0452 *
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 63.435  on 43  degrees of freedom
Residual deviance: 54.518  on 39  degrees of freedom
AIC: 187.3

Number of Fisher Scoring iterations: 5
```

Number of hours turns out to be the only significant variables. Also there is no significant overdispersion, so the poisson model is sufficient, though the fit is not good.

```
> dispersiontest(m32.pois)

Overdispersion test

data:  m32.pois
z = 1.1459, p-value = 0.1259
alternative hypothesis: true dispersion is greater than 1
sample estimates:
dispersion
   1.17846
```