MAST30027: Modern Applied Statistics

Week 3 Lab

1. The dataset `discoveries` lists the number of great scientific discoveries for the years 1860 to 1959, as chosen by "The World Almanac and Book of Facts", 1975 Edition. Has the discovery rate remained constant over time?

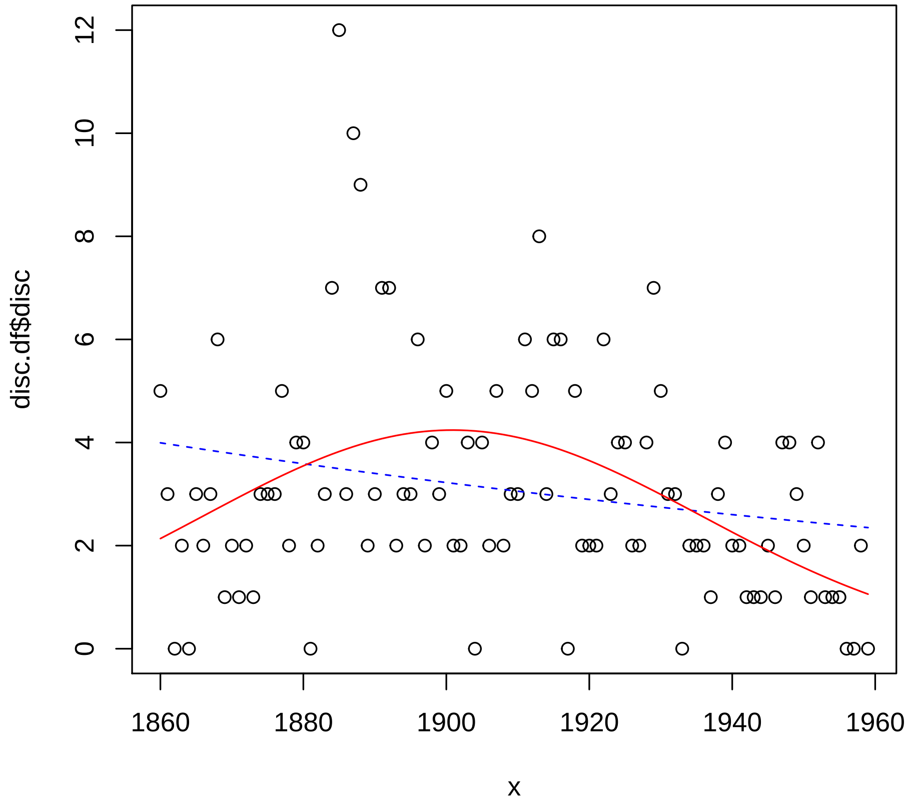
   To answer this question, fit a poisson regression model with a log link, and use the deviance to compare a null model with models including the year and year squared as predictors.

   Solution First we fit two models, the first including the year and the second the year and the year squared. The plot gives the fitted rates in each case.

```
> data(discoveries)
> disc.df <- data.frame(year=1860:1959, disc=discoveries)
> model1 <- glm(disc ~ year, family=poisson, disc.df)
> summary(model1)

Call:
glm(formula = disc ~ year, family = poisson, data = disc.df)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.8112  -0.9482  -0.3533   0.6637   3.5504

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) 11.354807   3.775677   3.007  0.00264 **
year        -0.005360   0.001982  -2.705  0.00683 **
---
Signif. codes:  0 âĂŸ***âĂŹ 0.001 âĂŸ**âĂŹ 0.01 âĂŸ*âĂŹ 0.05 âĂŸ.âĂŹ 0.1 âĂŸ âĂŹ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 164.68  on 99  degrees of freedom
Residual deviance: 157.32  on 98  degrees of freedom
AIC: 430.32

Number of Fisher Scoring iterations: 5

> model2 <- glm(disc ~ year + I(year^2), family=poisson, disc.df)
> summary(model2)

Call:
glm(formula = disc ~ year + I(year^2), family = poisson, data = disc.df)

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) -1.482e+03  3.163e+02  -4.685 2.79e-06 ***
year         1.561e+00  3.318e-01   4.705 2.54e-06 ***
I(year^2)   -4.106e-04  8.699e-05  -4.720 2.35e-06 ***
---
Signif. codes:  0 âĂŸ***âĂŹ 0.001 âĂŸ**âĂŹ 0.01 âĂŸ*âĂŹ 0.05 âĂŸ.âĂŹ 0.1 âĂŸ âĂŹ 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
```

```
> x <- disc.df$year
> plot(x, disc.df$disc)
> beta1 <- model1$coefficients
> lines(x, exp(beta1[1] + beta1[2]*x), col="blue", lty=2)
> beta2 <- model2$coefficients
> lines(x, exp(beta2[1] + beta2[2]*x + beta2[3]*x^2), col="red")
```



From the plot both year and year squared look significant, but we need to quantify this observation. For a poisson model the deviance only looks χ^2 if the responses are large enough to look vaguely normal, which they are not in this case. None-the-less, we can use deviance differences to perform likelihood ratio tests. From the above, the null model has deviance 164.68, the model with just year has deviance 157.32, and the model with year and year squared has deviance 132.84. We test the significance of adding year and then year squared:

```
> pchisq(164.68-157.32, 1, lower.tail=FALSE)

[1] 0.006669079

> pchisq(157.32-132.84, 1, lower.tail=FALSE)
```

```
[1] 7.508521e-07
```

There is strong evidence that year improves the model, and very strong evidence that year squared has something to add. We conclude that there is strong evidence that the discovery rate has changed over time.

2. The `ships` dataset from the `MASS` package gives the number of damage incidents and aggregate months of service for different types of ships broken down by year of construction and period of operation. Load the dataset using the commands `library(MASS)` then `data(ships)`.

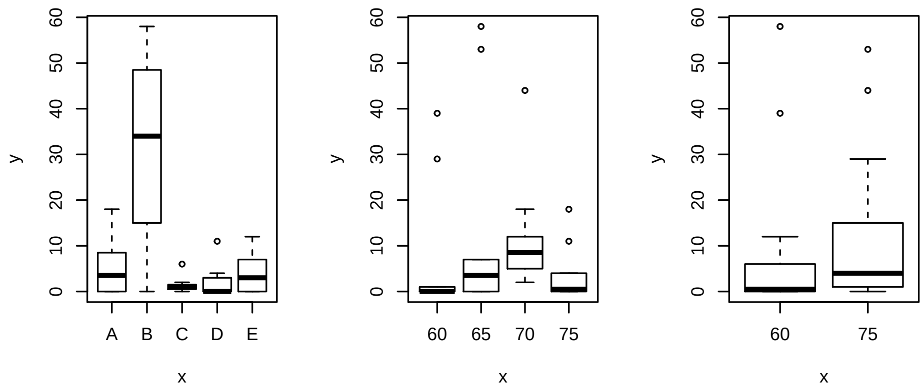
   Develop a model for the rate of incidents (i.e. a poisson regression model with log link), describing the effect of the important predictors.

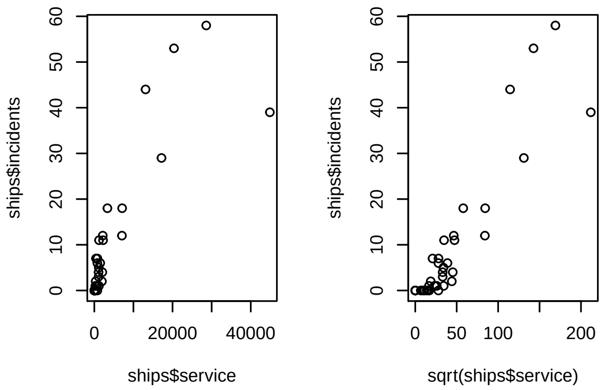
   Solution After loading and inspecting the data, it seems that `year` and `period` are really ordered factors rather than numerical predictors, so we alter these variables appropriately.

```
> library(MASS)
> data(ships)
> ships$year <- factor(ships$year, levels=seq(60, 75, 5), ordered=TRUE)
> ships$period <- factor(ships$period)
```

   Next we explore the relations between the variables. All the variables look important, and we note that applying a square root transform to `service` improves the relation between `service` and `incidents`

```
> par(mfrow=c(2,3))
> plot(ships$type, ships$incidents)
> plot(ships$year, ships$incidents)
> plot(ships$period, ships$incidents)
> plot(ships$service, ships$incidents)
> plot(sqrt(ships$service), ships$incidents)
> par(mfrow=c(1,1))
```





We can fit now a log-poisson model. From the Wald tests each variable looks significant. We could confirm this using likelihood ratio tests based on the deviance.

```
> ships$rootserv <- sqrt(ships$service)
> model <- glm(incidents ~ type + year + period + rootserv, family=poisson, ships)
> summary(model)

Call:
glm(formula = incidents ~ type + year + period + rootserv, family = poisson,
    data = ships)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.1509  -1.2833  -0.7905   0.2751   2.6875

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.207853   0.234122   0.888 0.374649
typeB       -0.121206   0.250163  -0.485 0.628024
typeC       -1.005644   0.329657  -3.051 0.002284 **
typeD       -0.574643   0.289933  -1.982 0.047481 *
typeE       -0.025521   0.236667  -0.108 0.914127
year.L       0.654626   0.194109   3.372 0.000745 ***
year.Q      -0.822592   0.122829  -6.697 2.13e-11 ***
year.C      -0.128340   0.097295  -1.319 0.187142
period75     0.726592   0.125831   5.774 7.73e-09 ***
rootserv     0.021648   0.002202   9.830  < 2e-16 ***
```

```
---
Signif. codes:  0 âĂŸ***âĂŹ 0.001 âĂŸ**âĂŹ 0.01 âĂŸ*âĂŹ 0.05 âĂŸ.âĂŹ 0.1 âĂŸ âĂŹ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 730.253  on 39  degrees of freedom
Residual deviance:  67.035  on 30  degrees of freedom
AIC: 184.9

Number of Fisher Scoring iterations: 5
```

Note that because `year` is ordered R has used linear, quadratic and cubic contrasts. You can see them exactly using `contrasts`

```
> contrasts(ships$year)

             .L   .Q         .C
[1,] -0.6708204  0.5 -0.2236068
[2,] -0.2236068 -0.5  0.6708204
[3,]  0.2236068 -0.5 -0.6708204
[4,]  0.6708204  0.5  0.2236068
```

Next we look for interactions.

```
> model1 <- glm(incidents ~ type + year + period + rootserv + type:year, family=poisson, ships)
> pchisq(deviance(model) - deviance(model1), df.residual(model) - df.residual(model1), lower.tail=FALSE)

[1] 0.0001099668

> model2 <- glm(incidents ~ type + year + period + rootserv + type:period, family=poisson, ships)
> pchisq(deviance(model) - deviance(model2), df.residual(model) - df.residual(model2), lower.tail=FALSE)

[1] 0.08820292

> model3 <- glm(incidents ~ type + year + period + rootserv + type:rootserv, family=poisson, ships)
> pchisq(deviance(model) - deviance(model3), df.residual(model) - df.residual(model3), lower.tail=FALSE)

[1] 0.003187932

> model4 <- glm(incidents ~ type + year + period + rootserv + year:period, family=poisson, ships)
> pchisq(deviance(model) - deviance(model4), df.residual(model) - df.residual(model4), lower.tail=FALSE)

[1] 0.0001018208

> model5 <- glm(incidents ~ type + year + period + rootserv + year:rootserv, family=poisson, ships)
> pchisq(deviance(model) - deviance(model5), df.residual(model) - df.residual(model5), lower.tail=FALSE)

[1] 0.0153112

> model6 <- glm(incidents ~ type + year + period + rootserv + period:rootserv, family=poisson, ships)
> pchisq(deviance(model) - deviance(model6), df.residual(model) - df.residual(model6), lower.tail=FALSE)

[1] 0.4123239

> model7 <- glm(incidents ~ type + year + period + rootserv + type:year + period:year, family=poisson, shi
> pchisq(deviance(model1) - deviance(model7), df.residual(model1) - df.residual(model7), lower.tail=FALSE)

[1] 0.0005265296

> model8 <- glm(incidents ~ type + year + period + rootserv + type:year + period:year + type:rootserv, fam
> pchisq(deviance(model7) - deviance(model8), df.residual(model7) - df.residual(model8), lower.tail=FALSE)

[1] 0.07904069
```

```
> model9 <- glm(incidents ~ type + year + period + rootserv + type:year + period:year + year:rootserv, fam
> pchisq(deviance(model7) - deviance(model9), df.residual(model7) - df.residual(model9), lower.tail=FALSE)

[1] 0.8730395

> summary(model7)

Call:
glm(formula = incidents ~ type + year + period + rootserv + type:year +
    period:year, family = poisson, data = ships)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-1.80944  -0.00785  -0.00005   0.00847   2.06533

Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)     -9.556e+00  3.042e+03  -0.003    0.997
typeB            5.649e+00  2.668e+03   0.002    0.998
typeC            3.815e+00  2.668e+03   0.001    0.999
typeD           -5.655e+00  4.666e+03  -0.001    0.999
typeE           -3.376e-01  3.779e+03   0.000    1.000
year.L           1.206e+00  8.164e+03   0.000    1.000
year.Q          -2.107e+01  6.085e+03  -0.003    0.997
year.C          -1.162e-01  2.721e+03   0.000    1.000
period75         5.714e+00  1.463e+03   0.004    0.997
rootserv         1.426e-02  1.341e-02   1.063    0.288
typeB:year.L    -1.467e+01  7.158e+03  -0.002    0.998
typeC:year.L    -1.451e+01  7.158e+03  -0.002    0.998
typeD:year.L     4.069e+00  1.043e+04   0.000    1.000
typeE:year.L    -1.669e+00  1.014e+04   0.000    1.000
typeB:year.Q     1.019e+01  5.335e+03   0.002    0.998
typeC:year.Q     1.040e+01  5.335e+03   0.002    0.998
typeD:year.Q     1.028e+01  9.333e+03   0.001    0.999
typeE:year.Q    -1.288e+00  7.559e+03   0.000    1.000
typeB:year.C    -4.192e+00  2.386e+03  -0.002    0.999
typeC:year.C    -5.540e+00  2.386e+03  -0.002    0.998
typeD:year.C    -1.430e+01  8.091e+03  -0.002    0.999
typeE:year.C     2.112e-01  3.380e+03   0.000    1.000
year.L:period75  1.364e+01  3.926e+03   0.003    0.997
year.Q:period75  1.044e+01  2.926e+03   0.004    0.997
year.C:period75  4.229e+00  1.309e+03   0.003    0.997

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 730.25  on 39  degrees of freedom
Residual deviance:  10.53  on 15  degrees of freedom
AIC: 158.4

Number of Fisher Scoring iterations: 18
```

Curiously, although the `type:year` and `period:year` interactions are significant, none of the Wald tests are significant in the model with interactions. This suggests dependency between our predictors. We look for a more parsimoneous model using `step`.

```
> model10 <- step(model7)

Start:  AIC=158.4
incidents ~ type + year + period + rootserv + type:year + period:year

             Df Deviance    AIC
- rootserv    1   11.694 157.56
<none>            10.530 158.40
```

```
- type:year   12   45.965 169.83
- year:period  3   28.151 170.02

Step:  AIC=157.56
incidents ~ type + year + period + type:year + year:period

              Df Deviance    AIC
<none>             11.694 157.56
- year:period  3   72.163 212.03
- type:year   12  123.483 245.35

> summary(model10)

Call:
glm(formula = incidents ~ type + year + period + type:year +
    year:period, family = poisson, data = ships)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-1.86294  -0.03467  -0.00005   0.03221   2.18897

Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)        -9.0814  3030.6260  -0.003    0.998
typeB               6.9806  2733.6363   0.003    0.998
typeC               3.7125  2733.6364   0.001    0.999
typeD              -5.7949  4742.5185  -0.001    0.999
typeE              -0.5124  3865.9456   0.000    1.000
year.L              0.7141  8132.0229   0.000    1.000
year.Q            -21.2793  6061.2520  -0.004    0.997
year.C             -0.1840  2710.6743   0.000    1.000
period75            5.5950  1308.4060   0.004    0.997
typeB:year.L      -16.0763  7335.1160  -0.002    0.998
typeC:year.L      -15.0317  7335.1161  -0.002    0.998
typeD:year.L        4.0042 10660.4227   0.000    1.000
typeE:year.L       -1.8407 10373.4206   0.000    1.000
typeB:year.Q       10.3502  5467.2727   0.002    0.998
typeC:year.Q       10.4696  5467.2728   0.002    0.998
typeD:year.Q       10.5783  9485.0369   0.001    0.999
typeE:year.Q       -1.3731  7731.8912   0.000    1.000
typeB:year.C       -3.9304  2445.0387  -0.002    0.999
typeC:year.C       -5.6421  2445.0388  -0.002    0.998
typeD:year.C      -14.2745  8141.6974  -0.002    0.999
typeE:year.C        0.1601  3457.8069   0.000    1.000
year.L:period75    14.9779  3510.8218   0.004    0.997
year.Q:period75    10.2033  2616.8120   0.004    0.997
year.C:period75     4.1899  1170.2739   0.004    0.997

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 730.253  on 39  degrees of freedom
Residual deviance:  11.694  on 16  degrees of freedom
AIC: 157.56

Number of Fisher Scoring iterations: 18
```

We see that, given the `type:year` and `period:year` interactions, `rootserv` is no longer significant. Formally, the reason for this is that `rootserv` itself can be predicted using `type`, `year`, `period`, `type:year` and `period:year`, so it is no longer needed when it comes to predicting `incidents`. Having said that, there is a clear scientific reason for wanting `rootserv` in the model, so given that the AIC for `model10` is not much smaller than that for `model7`, I would be inclined to keep it.

The fact that the individual parameters in `model10` are all close to zero is not necessarily a problem, but does suggest that some of these levels could be grouped. Testing that two levels of a factor are the same is not as easy for a glm as for a linear model, but can still be done indirectly using likelihood ratio tests. What we have to do is fit a model where the levels are combined, and then see if it performs significantly worse.

3. The `infert` dataset from the `survival` package presents data from a study of infertility after spontaneous and induced abortion. Using a logistic regression model, analyse and report on the factors related to infertility based on this data. (Don't use the factor stratum, as it is confounded with the other predictors.)

   Solution The response is `case`, with 1 indicating infertility and 0 fertility. The data comes from a case-control study, the aim of which was to estimate the effect of the number of prior induced and spontaneous abortions on the probability of becoming infertile. In the original study it was believed that education, age and parity (something numeric, whatever it is) were confounding variables, so the cases were separated into 83 strata based on these variables, and two controls were recruited from each stratum. (One control from one of the strata was subsequently omitted from the dataset, for reasons unexplained.)

   Because of how the data were collected, the observations are *not* independent, so a logistic regression model is not actually appropriate. None-the-less we will carry on as if it is.

```
> library(survival)
> data(infert)
> model1 <- glm(case ~ age+parity+education+spontaneous+induced,
+               data = infert, family = binomial())
> summary(model1)

Call:
glm(formula = case ~ age + parity + education + spontaneous +
    induced, family = binomial(), data = infert)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.7603  -0.8162  -0.4956   0.8349   2.6536

Coefficients:
                 Estimate Std. Error z value Pr(>|z|)
(Intercept)      -1.14924    1.41220  -0.814   0.4158
age               0.03958    0.03120   1.269   0.2046
parity           -0.82828    0.19649  -4.215 2.49e-05 ***
education6-11yrs -1.04424    0.79255  -1.318   0.1876
education12+ yrs -1.40321    0.83416  -1.682   0.0925 .
spontaneous       2.04591    0.31016   6.596 4.21e-11 ***
induced           1.28876    0.30146   4.275 1.91e-05 ***
---
Signif. codes:  0 âĂŸ***âĂŹ 0.001 âĂŸ**âĂŹ 0.01 âĂŸ*âĂŹ 0.05 âĂŸ.âĂŹ 0.1 âĂŸ âĂŹ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 316.17  on 247  degrees of freedom
Residual deviance: 257.80  on 241  degrees of freedom
AIC: 271.8

Number of Fisher Scoring iterations: 4

> model2 <- glm(case ~ parity+education+spontaneous+induced,
+               data = infert, family = binomial())
> summary(model2)

Call:
glm(formula = case ~ parity + education + spontaneous + induced,
```

```
    family = binomial(), data = infert)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.8372  -0.8194  -0.4737   0.8909   2.5822

Coefficients:
                 Estimate Std. Error z value Pr(>|z|)
(Intercept)        0.2646     0.8669   0.305   0.7602
parity            -0.8043     0.1964  -4.095 4.22e-05 ***
education6-11yrs  -1.1494     0.7868  -1.461   0.1441
education12+ yrs  -1.6123     0.8185  -1.970   0.0489 *
spontaneous        1.9882     0.3048   6.523 6.90e-11 ***
induced            1.2329     0.2986   4.128 3.66e-05 ***
---
Signif. codes:  0 âĂŸ***âĂŹ 0.001 âĂŸ**âĂŹ 0.01 âĂŸ*âĂŹ 0.05 âĂŸ.âĂŹ 0.1 âĂŸ âĂŹ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 316.17  on 247  degrees of freedom
Residual deviance: 259.43  on 242  degrees of freedom
AIC: 271.43

Number of Fisher Scoring iterations: 4
```

```
> pchisq(deviance(model2) - deviance(model1), 1, lower.tail=FALSE)
```

```
[1] 0.2019603
```

Continuing in this manner we find that all the remaining variables are significant at the 5% level (using the χ^2 test).