Stat 8053, Fall 2013: Poisson Regression (Faraway, 2.1 & 3)

The random component $y|x \sim \text{Po}(\lambda(x))$, so $\text{E}(y|x) = \lambda(x) = \text{Var}(y|x)$.

Linear predictor $\eta(x) = \mathbf{x}'\boldsymbol{\beta}$

Link function $\log(\text{E}(y|x)) = \log(\lambda(x)) = \eta(x)$, for the log-link usual with Poisson data.

Log-likelihood function Sum/product over the n observations

$$\begin{aligned}
\log(L(\boldsymbol{\beta})) &= \log\left(\prod \frac{\exp(-\lambda(x))\lambda(x)^y}{y!}\right) \\
&= \sum\left[-\lambda(x) + y\log(\lambda(x)) - \log(y!)\right] \\
&= \sum\left[-\exp(\eta(x)) + y\eta(x) - \log(y!)\right] \\
&= \sum\left[-\exp(\mathbf{x}'\boldsymbol{\beta}) + y(\mathbf{x}'\boldsymbol{\beta}) - \log(y!)\right]
\end{aligned}$$

Differentiate with respect to $\boldsymbol{\beta}$, and set the result to zero:

$$\begin{aligned}
\sum\left[-\exp(\mathbf{x}'\boldsymbol{\beta}) + y(\mathbf{x}'\boldsymbol{\beta}) - \log(y!)\right] &= 0 \\
\mathbf{X}'\mathbf{y} &= \mathbf{X}'\hat{\mu}
\end{aligned}$$

where $\hat{\mu} = \exp(\mathbf{X}\hat{\boldsymbol{\beta}})$ is the vector of estimated means.

Deviance

$$G^2 = 2\sum\left[y\log(y/\hat{\mu}) - (y - \hat{mu})\right]$$

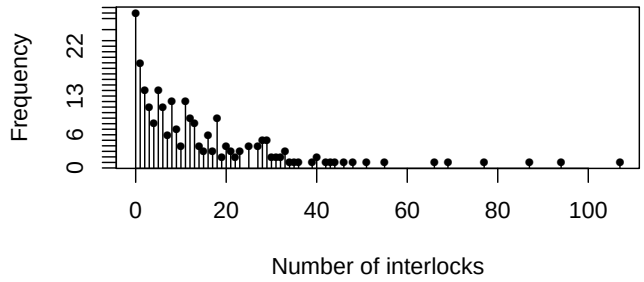
Interlocking Directors

```
library(alr4)
str(Ornstein)
```

```
'data.frame':        248 obs. of  4 variables:
 $ assets    : int  147670 133000 113230 85418 75477 40742 40140 26866 24500 23700 ...
 $ sector    : Factor w/ 10 levels "AGR","BNK","CON",..: 2 2 2 2 2 4 9 2 9 8 ...
 $ nation    : Factor w/ 4 levels "CAN","OTH","UK",..: 1 1 1 1 1 1 1 1 1 4 ...
 $ interlocks: int  87 107 94 48 66 69 46 16 77 6 ...
```

For 248 Canadian companies, `interlocks` is the number of other companies that share the same member of the board of directors.

```
tab <- xtabs( ~ interlocks, Ornstein)
plot((x<-as.numeric(names(tab))), tab, type="h", xlab="Number of interlocks", ylab="Frequency")
points(x, tab, pch=16, cex=.75)
```

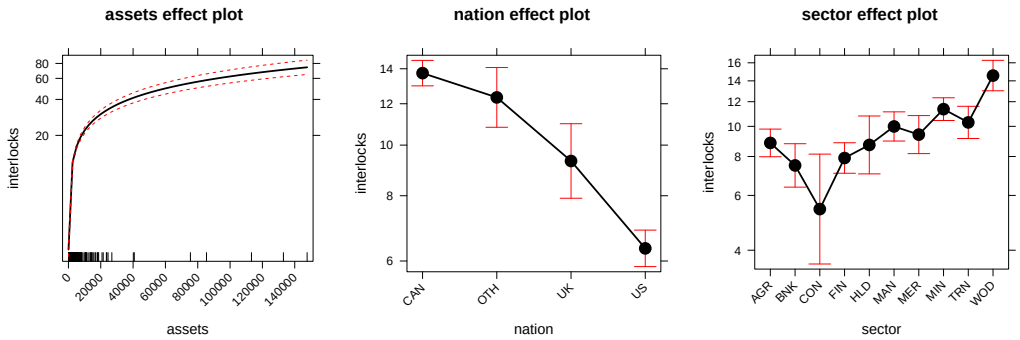



```
p1 <- glm(interlocks ~ log(assets) + nation + sector, family=poisson, data=Ornstein)
Anova(p1)
```

```
Analysis of Deviance Table (Type II tests)

Response: interlocks
            LR Chisq Df Pr(>Chisq)
log(assets)      731  1     <2e-16
nation           276  3     <2e-16
sector           103  9     <2e-16
```

```
plot(allEffects(p1, xlevels=list(assets=60)), rows=1, cols=3, rotx=45)
```



Nearly all companies had assets less than $C40 billion. Try refitting without the mammoth companies.

```
p2 <- update(p1, subset=assets < 30000)
Anova(p2)
plot(allEffects(p2, xlevels=list(assets=60)), rows=1, cols=3, rotx=45)
```

Contingency Tables

A *contingency table* is a cross-classification of counts. For now, ignore `count11`, and look only at PhD's in 2008-09:

```
tab2 <- xtabs(count ~ paste(type,citizen) + sex, AMSsurvey)
cbind(tab2, frate=round(tab2[,1]/(tab2[,1] + tab2[,2]), 2))
```

```
             Female Male frate
I(Pr) Non-US     25   79  0.24
I(Pr) US         20   87  0.19
I(Pu) Non-US     29  130  0.18
I(Pu) US         35  132  0.21
II Non-US        50   89  0.36
II US            47   96  0.33
III Non-US       39   53  0.42
III US           32   47  0.41
IV Non-US       105  122  0.46
IV US            54   71  0.43
Va Non-US        12   28  0.30
Va US            14   34  0.29
```

```
print(summary(tab2), digits=4)
```

```
Call: xtabs(formula = count ~ paste(type, citizen) + sex, data = AMSsurvey)
Number of cases in table: 1430
Number of factors: 2
Test for independence of all factors:
        Chisq = 71.42, df = 11, p-value = 6.569e-11
```

The same calculation can be done fitting a Poisson regression model:

```
str(data2 <- as.data.frame(tab2))
```

```
'data.frame':        24 obs. of  3 variables:
 $ paste.type..citizen.: Factor w/ 12 levels "I(Pr) Non-US",..: 1 2 3 4 5 6 7 8 9 10 ...
 $ sex                 : Factor w/ 2 levels "Female","Male": 1 1 1 1 1 1 1 1 1 1 ...
 $ Freq                : num  25 20 29 35 50 47 39 32 105 54 ...
```

```
names(data2)[1] <- "typebycitizen"
summary(p2 <- glm(Freq ~ typebycitizen + sex, poisson, data2))$coef
```

```
                          Estimate Std. Error z value   Pr(>|z|)
(Intercept)                3.51453    0.10526 33.3876 2.073e-244
typebycitizenI(Pr) US      0.02844    0.13770  0.2065  8.364e-01
typebycitizenI(Pu) Non-US  0.42451    0.12611  3.3661  7.624e-04
typebycitizenI(Pu) US      0.47360    0.12491  3.7914  1.498e-04
typebycitizenII Non-US     0.29008    0.12965  2.2374  2.526e-02
typebycitizenII US         0.31845    0.12887  2.4711  1.347e-02
typebycitizenIII Non-US   -0.12260    0.14313 -0.8566  3.917e-01
typebycitizenIII US       -0.27494    0.14924 -1.8422  6.544e-02
typebycitizenIV Non-US     0.78056    0.11841  6.5921  4.337e-11
typebycitizenIV US         0.18392    0.13272  1.3858  1.658e-01
typebycitizenVa Non-US    -0.95551    0.18605 -5.1357  2.811e-07
typebycitizenVa US        -0.77319    0.17450 -4.4310  9.380e-06
sexMale                    0.73967    0.05655 13.0806  4.251e-39
```

```
obs <- data2$Freq   # observed counts
fit <- predict(p2, type="response")  # expected counts given independence
sum( (obs -fit)^2/ fit )  # Formula for Pearson's X2
```

```
[1] 71.42
```

The model `p2` is of `sex` $\perp\!\!\!\perp$ (`type:citizenship`), and it does not fit.

Purum marriage: BFH, 191

The Purum have five sibs; marriage within a sib is taboo, as are marriages between certain other sibs.

```
sibs <- c("Marrim", "Makan", "Parpa", "Thao", "Kheyang")
wifeSib <- gl(5,5, labels=paste("W", sibs))
husbandSib <- factor(rep(1:5, 5), labels=paste("H", sibs))
count <- c(NA, 5, 17, NA, 6,  5, NA, 0, 16, 2,  NA, 2, NA, 10, 11,  10, NA, NA, NA, 9,  6, 20, 8, 0, 1)
purum <- data.frame(wifeSib, husbandSib, count)
with(purum, tapply(count, list(wifeSib, husbandSib), sum))
```

```
          H Marrim H Makan H Parpa H Thao H Kheyang
W Marrim        NA       5      17     NA         6
W Makan          5      NA       0     16         2
W Parpa         NA       2      NA     10        11
W Thao          10      NA      NA     NA         9
W Kheyang        6      20       8      0         1
```

One (Kheyang, Kheyang) marriage took place, but was in fact taboo. Are `wifeSib` and `husbandSib` independent?

```
m1 <- glm(count ~ wifeSib + husbandSib, poisson, purum)
data.frame(deviance=deviance(m1), df=m1$df.residual, pvalue=pchisq(deviance(m1), m1$df.residual, lower.tail=FALSE))
```

```
  deviance df   pvalue
1    76.25  8 2.77e-13
```

```
purum[names(m1$residuals), "PearsonRes"] <- residuals(m1, type="pearson")
print(with(purum, tapply(PearsonRes, list(wifeSib, husbandSib), sum)), digits=1)
```

```
          H Marrim H Makan H Parpa H Thao H Kheyang
W Marrim        NA      -2    1.94     NA      -0.2
W Makan       0.06      NA   -2.56    3.1      -1.0
W Parpa         NA      -2      NA    0.2       2.6
W Thao       -0.12      NA      NA     NA       0.1
W Kheyang     0.11       4    0.09   -3.0      -1.7
```

Physical disability at admission for stroke

```
score <- data.frame(
     initial = gl(5, 5, labels=paste("Initial", 1:5)),
     final = factor(rep(1:5, 5), labels=paste("Final", 1:5)),
     count = c(11, 23, 12, 15, 8, 9, 10,  4, 1, NA, 6, 4, 4, NA, NA, 4, 5, NA, NA, NA, 5, NA, NA, NA, NA))
with(score, tapply(count, list(initial, final), sum))
```

```
          Final 1 Final 2 Final 3 Final 4 Final 5
Initial 1      11      23      12      15       8
Initial 2       9      10       4       1      NA
Initial 3       6       4       4      NA      NA
Initial 4       4       5      NA      NA      NA
Initial 5       5      NA      NA      NA      NA
```

```
 summary(g2 <- glm(count ~ ., poisson, score))

Call:
glm(formula = count ~ ., family = poisson, data = score)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.0048  -0.2195   0.0197   0.3937   1.0692

Coefficients:
                 Estimate Std. Error z value Pr(>|z|)
(Intercept)         2.751      0.206   13.36  < 2e-16
initialInitial 2   -0.933      0.241   -3.87  0.00011
initialInitial 3   -1.263      0.301   -4.20  2.7e-05
initialInitial 4   -1.429      0.366   -3.90  9.5e-05
initialInitial 5   -1.142      0.492   -2.32  0.02040
finalFinal 2        0.336      0.239    1.41  0.15926
finalFinal 3       -0.272      0.292   -0.93  0.35186
finalFinal 4       -0.310      0.317   -0.98  0.32703
finalFinal 5       -0.672      0.409   -1.64  0.10063

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 49.5566  on 14  degrees of freedom
Residual deviance:  9.5958  on  6  degrees of freedom
  (10 observations deleted due to missingness)
AIC: 83.55

Number of Fisher Scoring iterations: 5

 data.frame(deviance=deviance(g2), df=g2$df.residual, pvalue=pchisq(deviance(g2), g2$df.residual, lower.tail=FALSE))

  deviance df pvalue
1    9.596  6 0.1427
```