Homework 4

Ericka Smith

Problem 1

```r
butters <- read_excel("C:/Users/erick/Downloads/butterflies.xlsx")
names(butters)[2] <- "numyears"
```

Part A

```r
summary(glm(LargeSkipper~factor(Region)+numyears, family=poisson, data=butters))
```

```
##
## Call:
## glm(formula = LargeSkipper ~ factor(Region) + numyears, family = poisson,
##     data = butters)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.8183  -0.9886  -0.5465  -0.2068   2.8921
##
## Coefficients:
##                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)     -0.71607    0.33387  -2.145  0.03197 *
## factor(Region)U -1.48039    0.49782  -2.974  0.00294 **
## numyears         0.04875    0.02218   2.198  0.02796 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 81.450  on 55  degrees of freedom
## Residual deviance: 65.026  on 53  degrees of freedom
## AIC: 105.59
##
## Number of Fisher Scoring iterations: 6
```

```r
summary(glm(LargeSkipper~factor(Region)*numyears, family=poisson, data=butters))
```

```
##
## Call:
## glm(formula = LargeSkipper ~ factor(Region) * numyears, family = poisson,
```

```
##     data = butters)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.8090  -0.9928  -0.5422  -0.2006   2.8949
##
## Coefficients:
##                           Estimate Std. Error z value Pr(>|z|)
## (Intercept)              -0.707518   0.353682  -2.000   0.0455 *
## factor(Region)U          -1.525202   0.804930  -1.895   0.0581 .
## numyears                  0.047996   0.024559   1.954   0.0507 .
## factor(Region)U:numyears  0.004086   0.057194   0.071   0.9430
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 81.450  on 55  degrees of freedom
## Residual deviance: 65.021  on 52  degrees of freedom
## AIC: 107.59
##
## Number of Fisher Scoring iterations: 6
```

I fit the first model without an interaction term and the second model with an interaction term. For the one without an interaction term all of the coefficients were significant at level $\alpha = 0.05$ ($p = 0.00294$ for region and $p = 0.02796$ for number of years). For the one with an interaction term both the number of years and the region were significant at level $\alpha = 0.10$ ($p = 0.0581$ for region and $p = 0.0507$ for number of years). The AIC is slightly higher for the model with an interaction term, and since the interaction term itself is not significant, I believe the model without an interaction term is the better fit.

Part B

```r
summary(glm(PearlyHeath~factor(Region)+ numyears, family=poisson, data=butters))
```

Pearly Heath

```
##
## Call:
## glm(formula = PearlyHeath ~ factor(Region) + numyears, family = poisson,
##     data = butters)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.8357  -0.7034  -0.0001  -0.0001   4.6509
##
## Coefficients:
##                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)     -19.58292 1799.75531  -0.011   0.9913
## factor(Region)U  20.21266 1799.75531   0.011   0.9910
## numyears          0.03312    0.01384   2.393   0.0167 *
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 165.240  on 55  degrees of freedom
## Residual deviance:  64.977  on 53  degrees of freedom
## AIC: 135.39
##
## Number of Fisher Scoring iterations: 17
```

```r
summary(glm(PearlyHeath~factor(Region)*numyears, family=poisson, data=butters))
```

```
##
## Call:
## glm(formula = PearlyHeath ~ factor(Region) * numyears, family = poisson,
##     data = butters)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.8357  -0.7034  -0.0001  -0.0001   4.6509
##
## Coefficients:
##                           Estimate Std. Error z value Pr(>|z|)
## (Intercept)              -1.930e+01  2.507e+03  -0.008    0.994
## factor(Region)U           1.993e+01  2.507e+03   0.008    0.994
## numyears                 -1.281e-08  2.224e+02   0.000    1.000
## factor(Region)U:numyears  3.312e-02  2.224e+02   0.000    1.000
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 165.240  on 55  degrees of freedom
## Residual deviance:  64.977  on 52  degrees of freedom
## AIC: 137.39
##
## Number of Fisher Scoring iterations: 17
```

I fit the first model without an interaction term and the second model with an interaction term. For the one without an interaction term only the coefficient for the number of years was significant at level $\alpha = 0.05$ ($p = 0.0167$). For the one with an interaction term none of the coefficients were significant. The AIC is slightly higher for the model with an interaction term (137.39 without vs. 135.39 with). I believe the model without an interaction term is clearly the better fit.

```r
summary(glm(Ringlet~factor(Region)+ numyears, family=poisson, data=butters))
```

Ringlet

```
##
## Call:
## glm(formula = Ringlet ~ factor(Region) + numyears, family = poisson,
##     data = butters)
```

```
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -4.3884  -2.1747  -0.5139   0.8776   8.2825
##
## Coefficients:
##                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)      1.60294    0.09086  17.641  < 2e-16 ***
## factor(Region)U  0.39832    0.09241   4.310 1.63e-05 ***
## numyears         0.03912    0.00519   7.537 4.81e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 404.86  on 55  degrees of freedom
## Residual deviance: 333.14  on 53  degrees of freedom
## AIC: 531.71
##
## Number of Fisher Scoring iterations: 5
```

```r
summary(glm(Ringlet~factor(Region)*numyears, family=poisson, data=butters))
```

```
##
## Call:
## glm(formula = Ringlet ~ factor(Region) * numyears, family = poisson,
##     data = butters)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -4.1575  -2.2862  -0.4898   0.9268   8.1686
##
## Coefficients:
##                          Estimate Std. Error z value Pr(>|z|)
## (Intercept)              1.663278   0.110603  15.038  < 2e-16 ***
## factor(Region)U          0.298516   0.142583   2.094   0.0363 *
## numyears                 0.033326   0.008227   4.051 5.11e-05 ***
## factor(Region)U:numyears 0.009673   0.010606   0.912   0.3618
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 404.86  on 55  degrees of freedom
## Residual deviance: 332.31  on 52  degrees of freedom
## AIC: 532.88
##
## Number of Fisher Scoring iterations: 6
```

I fit the first model without an interaction term and the second model with an interaction term. For the one without an interaction term all of the coefficients were significant at level $\alpha = 0.05$ ($p = 1.63 * 10^{-05}$ for region and $p = 4.81 * 10^{-14}$ for number of years). For the one with an interaction term both the number of years and the region were significant at level $\alpha = 0.05$ ($p = 0.0363$ for region and $p = 5.11 * 10^{-05}$ for number of years), but the interactino term was not significant ($p = 0.3618$). The AIC is slightly higher for the model with an interaction term (532.88 compared to 531.71), and since the interaction term itself is not significant, I believe the model without an interaction term is the better fit.

Problem 2 (starts next page)

Problem 2

ⓐ Normal Distribution

$$f_x(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2\sigma^2}(x-\mu)^2\right] = \exp\left[-\frac{1}{2\sigma^2}(x-\mu)^2 - \frac{1}{2}\log(2\pi\sigma^2)\right]$$

$$f_x(x) = \exp\left[\frac{-\mu^2+2x\mu-x^2}{2\sigma^2} - \frac{1}{2}\log(2\pi\sigma^2)\right] = \exp\left[\frac{x\mu-\frac{\mu^2}{2}}{\sigma^2} - \frac{x^2}{2\sigma^2} - \frac{1}{2}\log(2\pi\sigma^2)\right]$$

① Canonical parameter $\theta = \mu$

② dispersion parameter $\phi = \sigma^2$

③ $a(\phi) = \phi = \sigma^2$

$b(\theta) = \frac{\theta^2}{2}$

$c(x, \phi) = -\frac{x^2}{2\phi} - \frac{1}{2}\log(2\pi\phi)$

④ $\mu(\theta) = b'(\theta) = \theta$

⑤ variance function

$Var(x) = b''(\theta)a(\phi) = 1 \cdot \phi$, $V(\mu) = b''(\theta) = 1$

⑥ canonical link function $g(\mu) = \theta$

(b) Binomial(n,p)

$$f(y) = \binom{n}{y} p^y (1-p)^{n-y} = \exp\left[\log\binom{n}{y} + y\log p + (n-y)\log(1-p)\right]$$

$$= \exp\left[y\log\frac{p}{1-p} + n\log(1-p) + \log\binom{n}{y}\right] = \exp\left[\left(\frac{\bar{y}\log\frac{p}{1-p} - (-\log(1-\pi))}{1/n}\right) + \log\binom{n}{n\bar{y}}\right]$$

① $\theta = \log\frac{p}{1-p}$

② $\phi = 1$

③ $a(\phi) = \frac{1}{n}$

$b(\theta) = -\log(1-p) = \log(1+e^\theta)$

$c(y,\phi) = \log$

④ $\mu(\theta) = p$

⑤ $Var(\mu) = b''(\theta) = \frac{e^\theta}{1+e^\theta}\frac{1}{1+e^\theta} = p(1-p)$

⑥ $g(\mu) = \log\left(\frac{p}{1-p}\right)$

(c) Poisson

$$f(y) = e^{-\lambda} \frac{\lambda^y}{y!} = \exp\{y\log(\lambda) - \lambda - \log(y!)\}$$

(1) $\theta = \log \lambda$

(2) $\phi = 1$

(3) $a(\phi) = 1$

$b(\theta) = e^{\theta} = \lambda$

$c(y, \phi) = -\log(y!)$

(4) $\mu(\theta) = \lambda = e^{\theta}$

(5) $Var(\mu) = b''(\theta) = \lambda$

(6) $g(\mu) = \log \lambda$

(d) Exponential

$$f(y) = \lambda e^{-\lambda y} = \exp(-\lambda y + \log\lambda) = \exp\left(\frac{y}{\mu} - \log\mu\right)$$

(1) $\theta = \frac{1}{\mu}$

(2) $\phi = 1$

(3) $a(\phi) = 1$

$b(\theta) = \frac{\lambda^2}{2}$

$c(y, \phi) = -\log\mu$

(4) $\mu(\theta) = b'(\theta) = \lambda$

(5) $Var(\mu) = b''(\theta) = \lambda^2$

(6) $g(\mu) = -\frac{1}{\mu}$

(e) Inverse Gaussian

$$f(y) = \sqrt{\frac{\lambda}{2\pi y^3}} \exp\left[-\frac{\lambda(y-\mu)^2}{2\mu^2 y} \right] \qquad y, \mu, \lambda > 0$$

$$= \exp\left[-y \frac{\lambda}{2\mu^2} + \frac{\lambda}{\mu} - \frac{\lambda}{2y} + \frac{1}{2}\log(\lambda) - \frac{1}{2}\log(2\pi y^3) \right]$$

$$= \exp\left[\frac{y\left(-\frac{1}{\mu^2}\right) + \frac{2}{\mu}}{2/\lambda} - \frac{\lambda}{2y} + \frac{\log(\lambda) - \log 2\pi y^3}{2} \right]$$

(1) $\theta = \frac{1}{\mu^2}$

(2) $\phi = \frac{2}{\lambda}$

(3) $a(\phi) = \phi = \frac{2}{\lambda}$

$b(\theta) = -\frac{2}{\mu} = -2\sqrt{-\theta}$

$c(y, \phi) = -\frac{\log(\phi/2) + \log 2\pi y^3}{2} - \frac{1}{\phi y}$

(4) $\mu(\theta) = b'(\theta) = (-\theta)^{-1/2}$

(5) $Var(\mu) = b''(\theta) = \frac{1}{2}(-\theta)^{-3/2}$

(6) $g(\mu) = -\frac{1}{\mu^2}$

(f) Negative Binomial

$$P(Y=y) = \binom{y+r-1}{y}(1-p)^y p^r = \exp\left[y\log(1-p) + r\log p + \log\binom{y+r-1}{y}\right]$$

$$= \exp\left\{\theta y - \left(-r\log(1-e^\theta)\right) + \log\binom{y+r-1}{y}\right\}$$

(1) $\theta = \log(1-p)$

(2) $\phi = 1$

(3) $a(\phi) = \phi = 1$

$b(\theta) = -r\log(1-e^\theta)$

$c(y,\phi) = \log\binom{y+r-1}{y}$

(4) $\mu(\theta) = b'(\theta) = \dfrac{re^\theta}{1-e^\theta} = \dfrac{r(1-p)}{p}$

(5) $Var(\mu) = b''(\theta) = \dfrac{re^\theta}{(1-e^\theta)^2} = \dfrac{r(1-p)}{p^2}$

(6) $g(\mu) = \log\left(\dfrac{\mu}{r+\mu}\right)$