MAST90104: A First Course in Statistical Learning

Week 11 Lab and Workshop

Practical questions

1. The `cornnit` dataset in the `faraway` package contains data on the effect of nitrogen on the yield of corn. Fit a gamma regression to this data, using the `glm` command and store the model fit as `gmod`, using the canonical link function. Hint: **consider transforming the predictor variable first**.

   (a) Extract the Pearson residuals from the fitted model using the `residuals` function, then use them to estimate the dispersion parameter. Check that your answer agrees with the summary output from your model.
   Solution:

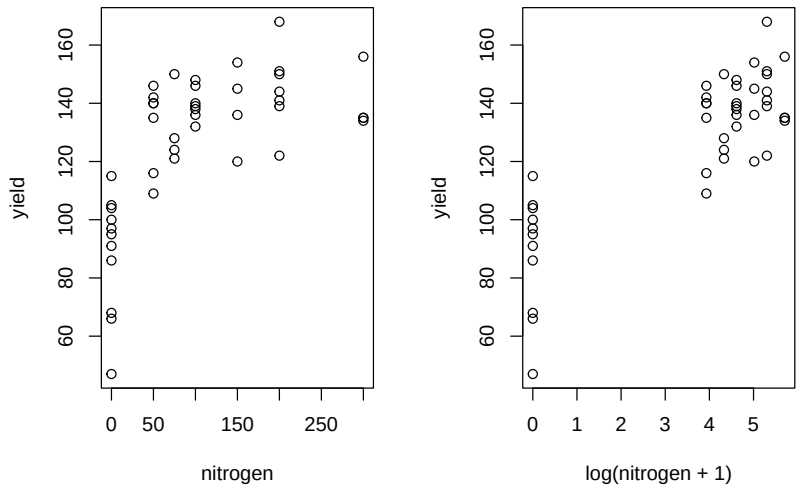

Figure 1: Scatter plot of yield against nitrogen on left panel and yield again log(nitrogen + 1) on right panel.

```
> library(faraway)
> data("cornnit")
> fivenum(cornnit$nitrogen)
[1]   0.0  25.0  87.5 175.0 300.0
> par(mfrow=c(1,2))
> plot(yield~nitrogen,data=cornnit)
> plot(yield~log(nitrogen+1),data=cornnit)
#Judging from scatter plots, log(nitrogen+1) would be a suitable transformation to
#ensure linear relationship between response and predictor
> gmod <- glm(yield~log(nitrogen+1), family=Gamma, data = cornnit)
> summary(gmod)

Call:
glm(formula = yield ~ log(nitrogen + 1), family = Gamma, data = cornnit)
```

```
Coefficients:
                  Estimate  Std. Error  t value  Pr(>|t|)
(Intercept)       1.106e-02  4.349e-04  25.439  < 2e-16 ***
log(nitrogen + 1) -7.902e-04  9.538e-05  -8.285 2.25e-10 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for Gamma family taken to be 0.01874595)

Null deviance: 2.40614  on 43  degrees of freedom
Residual deviance: 0.91727  on 42  degrees of freedom
AIC: 383.74

Number of Fisher Scoring iterations: 4
> sum(residuals(gmod, type = "pearson")^2)/(42)
[1] 0.01874588
```

The estimate for ϕ is indeed 0.01875.

(b) The command `anova(gmod, test="F")` will compare your model against the intercept-only model, using an F test. Using the deviances and dispersion estimates reported by `summary(gmod)`, check that the F statistic reported by the `anova` function is correct.

Solution:

```
> (2.40614-0.91727)/0.01874595/1
[1] 79.42356
> anova(gmod,test="F")
Analysis of Deviance Table

Model: Gamma, link: inverse

Response: yield

Terms added sequentially (first to last)


                 Df Deviance Resid. Df Resid. Dev      F   Pr(>F)
NULL                                43    2.40614
log(nitrogen + 1)  1   1.4889        42    0.91727 79.423 3.11e-11 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

The F-statistic is indeed 79.423.

(c) Now do some diagnostic plots. Can you identify a potential outlier?

Solution: Based on the jack-knife residual plot and the Cook's distance plot, it is clear that both observation 21 is an influential outliers.

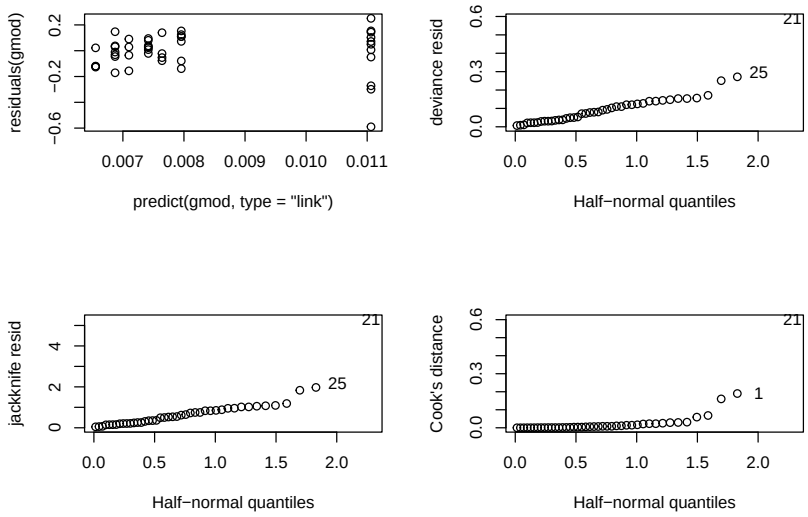

Figure 2: Residual plots for Gamma regression model

(d) Fit a linear model to the `cornnit` data. Which do you prefer, the linear model or the gamma model, and why?

Solution:

```
> gmod.norm <- lm(yield ~ log(nitrogen+1), data=cornnit)
> summary(gmod.norm)

Call:
lm(formula = yield ~ log(nitrogen + 1), data = cornnit)

Residuals:
Min      1Q  Median      3Q     Max
-42.335 -10.261   2.126  10.558  25.665

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)         89.335      4.227   21.13  < 2e-16 ***
log(nitrogen + 1)   10.201      1.017   10.03 1.03e-12 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 14.34 on 42 degrees of freedom
Multiple R-squared:  0.7055,Adjusted R-squared:  0.6985
F-statistic: 100.6 on 1 and 42 DF,  p-value: 1.025e-12
> par(mfrow=c(2,2))
> plot(predict(gmod.norm, type="response"), residuals(gmod.norm))
> halfnorm(residuals(gmod.norm), ylab="deviance resid")
> halfnorm(rstudent(gmod.norm), ylab="jackknife resid")
> halfnorm(cooks.distance(gmod.norm), ylab="Cook's distance")
> plot(predict(gmod.norm, type="response")~predict(gmod, type = "response"))
> abline(0,1)
```

Observation 21 appears to be an influential observation under both the gamma and Gaussian regression models, so in this respect linear regression doesn't improve the model fit. However, the deviance residuals are much smaller for the Gamma regression than in the linear regression. Hence, we prefer the gamme regression model.

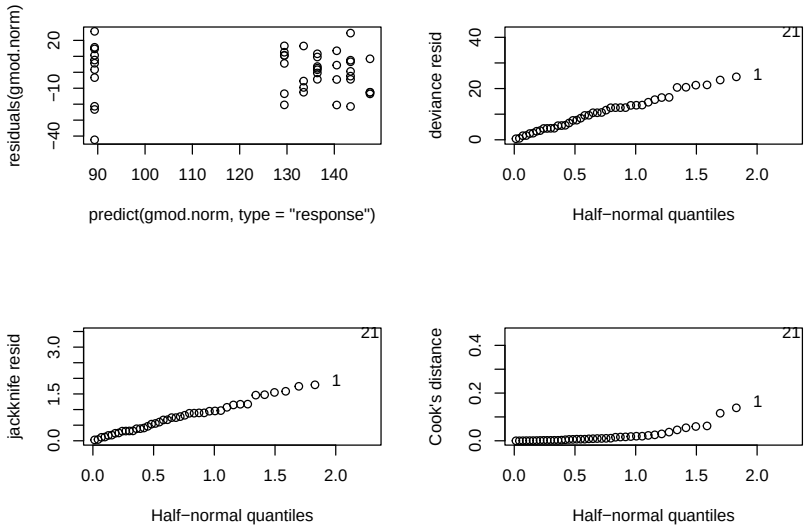

Figure 3: Residual plots for linear (Gaussian) regression model

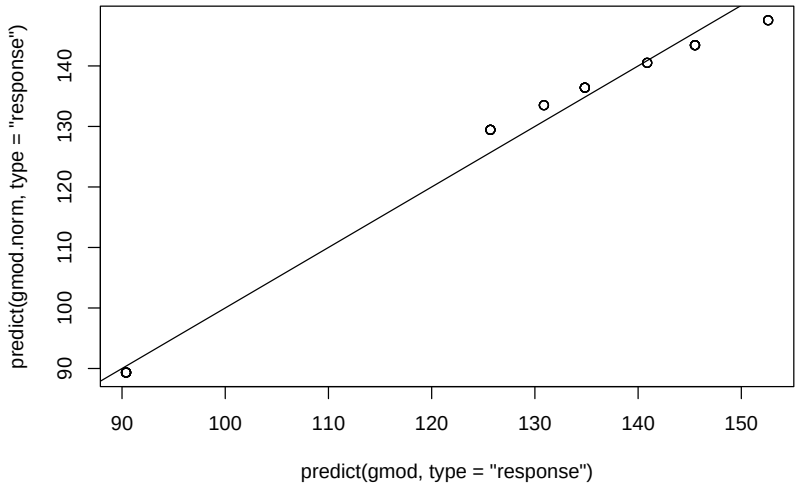

Figure 4: Correspondence of predicted values between gamma regression and linear regression models

2. The `Articles` dataset in the `Rchoice` package contains data on the publication counts (`art`) of research scientists and their respective gender (`fem`), marital status (`mar`), number of children (`kid5`), prestige of graduate program (`phd`), and the number of articles published by their mentors (`ment`).

   (a) Fit a Poisson regression model with `art` as the response variable using the canonical link function.

   ```
   #install.packages("Rchoice")
   > library(Rchoice)
   > data("Articles")
   > mod.pois <- glm(art~., family=poisson, data = Articles)
   ```

   (b) Perform stepwise selection using AIC criterion starting from the full Poisson regression model with all predictors. Write down the equation of your your final regression model.

```
> step(mod.pois)
Start:  AIC=3314.11
art ~ fem + mar + kid5 + phd + ment

Df Deviance    AIC
- phd   1   1634.6 3312.3
<none>      1634.4 3314.1
- mar   1   1640.8 3318.5
- fem   1   1651.5 3329.2
- kid5  1   1656.5 3334.2
- ment  1   1766.2 3444.0

Step:  AIC=3312.35
art ~ fem + mar + kid5 + ment

Df Deviance    AIC
<none>      1634.6 3312.3
- mar   1   1640.8 3316.6
- fem   1   1651.8 3327.5
- kid5  1   1656.7 3332.4
- ment  1   1776.7 3452.5

Call:  glm(formula = art ~ fem + mar + kid5 + ment, family = poisson,
data = Articles)

Coefficients:
(Intercept)          fem          mar         kid5
0.34517     -0.22530      0.15218     -0.18499
ment
0.02576

Degrees of Freedom: 914 Total (i.e. Null);  910 Residual
Null Deviance:     1817
Residual Deviance: 1635  AIC: 3312
> mod.pois.final <- glm(art ~ fem + mar + kid5 + ment, family = poisson,data = Articles)
> summary(mod.pois.final)

Call:
glm(formula = art ~ fem + mar + kid5 + ment, family = poisson,
data = Articles)

Coefficients:
Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.34517    0.06012   5.741 9.41e-09 ***
fem         -0.22530    0.05461  -4.125 3.70e-05 ***
mar          0.15218    0.06107   2.492   0.0127 *
kid5        -0.18499    0.04014  -4.609 4.05e-06 ***
ment         0.02576    0.00195  13.212  < 2e-16 ***
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

Null deviance: 1817.4  on 914  degrees of freedom
Residual deviance: 1634.6  on 910  degrees of freedom
AIC: 3312.3
```

```
Number of Fisher Scoring iterations: 5
```

The fitted final regression model is $y_i \sim \text{Poisson}(\mu_i)$, $i = 1, \ldots, n$, where

$$\log(\mu_i) = \beta_0 + \beta_1 \texttt{fem}_i + \beta_2 \texttt{mar}_i + \beta_3 \texttt{kid5}_i + \beta_4 \texttt{ment}_i.$$

(c) The `glm.nb` command in the `MASS` library fits the following model: $y_i \sim \text{NegBin}(\mu_i, k)$, where $\mu_i > 0$, $k > 0$, and

$$p(y_i = y) = \frac{\Gamma(y+k)}{y!\Gamma(k)} \left(\frac{\mu_i}{\mu_i + k}\right)^y \left(\frac{k}{\mu_i + k}\right)^k, \quad y = 0, 1, 2, \ldots$$

Here, $E(y_i) = \mu_i$ and $\text{Var}(y_i) = \mu_i + \mu_i^2/k$. The default log link function is $g(\mu) = \log(\mu)$. Using `glm.nb`, fit a Negative Binomial regression model with `art` as the response variable using the log link function.

```
> library(MASS)
> mod.nb <- glm.nb(art~., data = Articles)
```

(d) Perform stepwise selection using AIC criterion starting from the full Negative Binomial regression model with all predictors. Write down the equation of your your final regression model.

```
> step(mod.nb,scope=~.)
Start:  AIC=3133.92
art ~ fem + mar + kid5 + phd + ment

Df Deviance    AIC
- phd   1   1004.5 3132.1
<none>      1004.3 3133.9
- mar   1   1007.6 3135.3
- fem   1   1013.2 3140.8
- kid5  1   1015.5 3143.1
- ment  1   1079.1 3206.7

Step:  AIC=3132.1
art ~ fem + mar + kid5 + ment

Df Deviance    AIC
<none>      1004.4 3132.1
- mar   1   1007.7 3133.3
+ phd   1   1004.2 3133.9
- fem   1   1013.3 3139.0
- kid5  1   1015.6 3141.3
- ment  1   1085.3 3211.0

Call:  glm.nb(formula = art ~ fem + mar + kid5 + ment, data = Articles,
init.theta = 2.264120074, link = log)

Coefficients:
(Intercept)          fem          mar         kid5
0.30333     -0.21667      0.14694     -0.17680
ment
0.02943

Degrees of Freedom: 914 Total (i.e. Null);  910 Residual
Null Deviance:     1109
Residual Deviance: 1004  AIC: 3134
> mod.nb.final <- glm.nb(formula = art ~ fem + mar + kid5 + ment, data = Articles)
> summary(mod.nb.final)
```

```
Call:
glm.nb(formula = art ~ fem + mar + kid5 + ment, data = Articles,
init.theta = 2.26411693, link = log)

Coefficients:
Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.303328   0.081389   3.727 0.000194 ***
fem         -0.216673   0.072624  -2.983 0.002850 **
mar          0.146944   0.081765   1.797 0.072312 .
kid5        -0.176797   0.052826  -3.347 0.000818 ***
ment         0.029430   0.003108   9.470  < 2e-16 ***
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for Negative Binomial(2.2641) family taken to be 1)

Null deviance: 1108.9  on 914  degrees of freedom
Residual deviance: 1004.4  on 910  degrees of freedom
AIC: 3134.1

Number of Fisher Scoring iterations: 1


Theta:  2.264
Std. Err.:  0.271

2 x log-likelihood:  -3122.096
```

The final fitted regression model is: $y_i \sim \text{NegBin}(\mu_i, k)$, where

$$\log(\mu_i) = \beta_0 + \beta_1 \texttt{fem}_i + \beta_2 \texttt{mar}_i + \beta_3 \texttt{kid5}_i + \beta_4 \texttt{ment}_i.$$

(e) Which model would you prefer – the Poisson or Negative Binomial? Justify your answer with a suitable residual plot. By assessing the deviance residual-half normal quantile plots, the Poisson regression yields more observations with extreme residuals. Furthermore, the AIC of the final NegBin regression model is lower than that for the final Poisson regression model. Hence, we prefer the NegBin model.

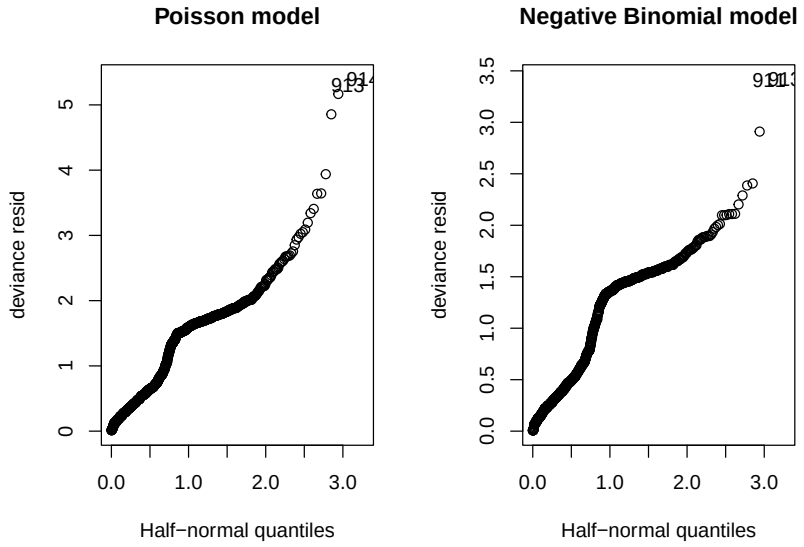

Figure 5: Deviance residual plots for final Poisson (left) and negative binomial regression (right) models

Workshop questions

1. Refer to Q2(c) from practical. By consider k as fixed, show that the negative binomial distribution belongs to the exponential family.
   Solution: The pmf can be alrebraically manipulated as

$$\begin{aligned} p(y_i = y) &= \frac{\Gamma(y+k)}{y!\Gamma(k)} \left(\frac{\mu_i}{\mu_i + k}\right)^y \left(\frac{k}{\mu_i + k}\right)^k \\ &= \exp\left[y \log\{\mu/(\mu+k)\} + k \log\{1 - \mu/(\mu+k)\} + \log\Gamma(y+k) - \log(y!) - \log\Gamma(k)\right] \end{aligned}$$

   Hence, $\theta = \log\{\mu/(\mu+k)\}$ and $\phi = 1$. Moreover, $b(\theta) = -k\log\{1 - e^\theta\}$, $a(\phi) = \phi$, and

$$c(y;\phi) = \log\Gamma(y+k) - \log(y!) - \log\Gamma(k).$$

2. Prove the Lemma in page 17 of Lecture 8.
   Solution: Note that the density of Y is of the form

$$f(y;\theta,\phi) = \exp\left[\frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi)\right]$$

   By noting that f is a proper density, we have

$$\int \exp\left[\frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi)\right] \, \mathrm{d}y = 1.$$

   where we swop $\int$ with $\sum$ for discrete y. Taking derivative wrt to θ on both side and interchanging integral and derivative on LHS, we have

$$\int \tfrac{\partial}{\partial\theta} \exp\left[\frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi)\right] \, \mathrm{d}y = 0.$$

   Note the integrand on the LHS can be expressed as

$$\tfrac{\partial}{\partial\theta} \exp\left[\frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi)\right] = \frac{y - b'(\theta)}{a(\phi)} f(y;\theta,\phi)$$

   Hence we have

$$\int \frac{y - b'(\theta)}{a(\phi)} f(y;\theta,\phi) \, \mathrm{d}y = 0.$$

And hence

$$b'(\theta) \underbrace{\int f(y;\theta,\phi)\,\mathrm{d}y}_{=1} = \int y f(y;\theta,\phi)\,\mathrm{d}y.$$

Then,

$$b'(\theta) = \int y f(y;\theta,\phi)\,\mathrm{d}y = E(Y).$$

To show the formula for variance, note that we previously worked out:

$$\int (y - b'(\theta)) f(y;\theta,\phi)\,\mathrm{d}y = 0.$$

Taking derivative wrt to θ on both side and interchanging integral and derivative on LHS, we have

$$\int \left\{ \frac{(y - b'(\theta))^2}{a(\phi)} - b''(\theta) \right\} f(y;\theta,\phi)\,\mathrm{d}y = 0.$$

and hence

$$\int (y - b'(\theta))^2 f(y;\theta,\phi)\,\mathrm{d}y = b''(\theta)a(\phi).$$

Since $E(Y) = b'(\theta)$, we have

$$\int (y - E(Y))^2 f(y;\theta,\phi)\,\mathrm{d}y = b''(\theta)a(\phi).$$

By noting that the definition of variance is $E\{(Y - E(Y))^2\}$, we have $Var(Y) = b''(\theta)a(\phi)$.

3. Refer to Q2 from practical. The following R output details the fit of two Poisson regression models (with some details redacted).

```
> mod.pois.workshop <- glm(art ~ fem + ment, family = poisson,data = Articles)
> summary(mod.pois.workshop)

Call:
glm(formula = art ~ fem + ment, family = poisson, data = Articles)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.34909    0.04191   8.329  < 2e-16 ***
fem         -0.18445    0.05235  -3.523 0.000426 ***
ment         0.02510    0.00193  13.005  < 2e-16 ***
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

Null deviance: 1817.4  on 914  degrees of freedom
Residual deviance: 1657.0  on 912  degrees of freedom
AIC: 3330.7

Number of Fisher Scoring iterations: 5

> mod.pois.workshop2 <- glm(art ~ fem, family = poisson(link = "inverse"),data = Articles)
> summary(mod.pois.workshop2)

Call:
glm(formula = art ~ fem, family = poisson(link = "inverse"),
data = Articles)
```

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.53118    0.01742  30.499  < 2e-16 ***
fem          0.14895    0.03241   4.595 4.32e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

Null deviance: <Redacted>  on 914  degrees of freedom
Residual deviance: 1794.4  on 913  degrees of freedom
AIC: 3466.1

Number of Fisher Scoring iterations: 7
> qchisq(0.95,1)
[1] 3.841459
```

(a) Write down the equation of the two fitted regression models, including the MLEs of their respective coefficients.
Solution: For `mod.pois.workshop`, the equation is $y_i \sim \text{Poisson}(\mu_i)$, where

$$\log(\mu_i) = \beta_0 + \beta_1 \texttt{fem}_i + \beta_2 \texttt{ment}_i,$$

where MLEs are $\widehat{\beta}_0 = 0.34909$, $\widehat{\beta}_1 = -0.18445$, and $\widehat{\beta}_2 = 0.02510$. For `mod.pois.workshop2`, the equation is $y_i \sim \text{Poisson}(\mu_i)$, where

$$1/\mu_i = \beta_0 + \beta_1 \texttt{fem}_i,$$

where MLEs are $\widehat{\beta}_0 = 0.53118$ and $\widehat{\beta}_1 = 0.14895$.

(b) Can we use a likelihood ratio test to compare `mod.pois.workshop2` against `mod.pois.workshop`? If yes, compute the test statistic, write down its null distribution, and state your conclustion. If no, suggest an alternative approach to compare the two models based on the above output.
Solution: No, we can't compare the likelihood ratios of the two models because they are based on different link functions (log-link versus inverse-link). We can use their respective AICs to compare the two models. In fact, since `mod.pois.workshop` has lower AIC (3330.7), we prefer `mod.pois.workshop`.

(c) Compare `mod.pois.workshop2` against the intercept-only model.
Solution: Note that the redacted null deviance for `mod.pois.workshop2` equals to the null deviance of `mod.pois.workshop` because deviance of the intercept-only model is invariant to the choice of link function. Our test is H_0 : intercept-only model is correct is adequate against H_1 : intercept-only model is incorrect and `mod.pois.workshop2` is correct. Our test statistic is $T = 1817.4 - 1794.4 = 23$. Under H_0, the test statistic follows χ^2_1. We reject H_0 if $T > \chi^2_{1;0.05} = 3.841459$. Therefore, we reject H_0 and conclude that `mod.pois.workshop2` provides a significantly better fit to the data.