MAST30027: Modern Applied Statistics

Week 5 Lab

1. Incubation temperature can affect the sex of turtles. An experiment was conducted with three independent replicates for each temperature and the number of male and female turtles born was recorded. The data can be found in the `turtle` dataset in the `faraway` package.

   Check for evidence of overdispersion in a binomial model for the sex of the turtle.

   What problems can arise if you ignore overdispersion?

   Solution: We fit a binomial regression and estimate the dispersion ϕ.

```
> library(faraway)
> data(turtle)
> bmod <- glm(cbind(male, female) ~ temp, data=turtle, family=binomial)
> summary(bmod)

Call:
glm(formula = cbind(male, female) ~ temp, family = binomial,
    data = turtle)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.0721  -1.0292  -0.2714   0.8087   2.5550

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -61.3183    12.0224  -5.100 3.39e-07 ***
temp          2.2110     0.4309   5.132 2.87e-07 ***
---
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: 74.508  on 14  degrees of freedom
Residual deviance: 24.942  on 13  degrees of freedom
AIC: 53.836

Number of Fisher Scoring iterations: 5

> (phihat <- sum( residuals(bmod, type="pearson")^2 )/13)

[1] 2.018641

> pchisq(24.942, 13, lower.tail=F)

[1] 0.02349208
```

   $\hat{\phi}$ is a little bit larger than 1, indicating possible overdispersion. Also, the chi-squared test for model adequacy using the deviance gives a significant result, indicating that there is something left unexplained. This could be overdispersion, or a problem with the model. In this case it looks like a problem with the model, as if we add `temp^2` to the model we can improve the fit, and the test for model adequacey is no longer significant at the 5% level (just).

```
> bmod2 <- glm(cbind(male, female) ~ temp + I(temp^2), data=turtle, family=binomial)
> summary(bmod2)
```

```
Call:
glm(formula = cbind(male, female) ~ temp + I(temp^2), family = binomial,
    data = turtle)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.6703  -0.8875  -0.4194   0.9481   2.2198

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -677.5950   268.7984  -2.521   0.0117 *
temp          45.9173    18.9169   2.427   0.0152 *
I(temp^2)     -0.7745     0.3327  -2.328   0.0199 *
---
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: 74.508  on 14  degrees of freedom
Residual deviance: 20.256  on 12  degrees of freedom
AIC: 51.15

Number of Fisher Scoring iterations: 4
```

```
> (phihat <- sum( residuals(bmod2, type="pearson")^2 )/12)
```

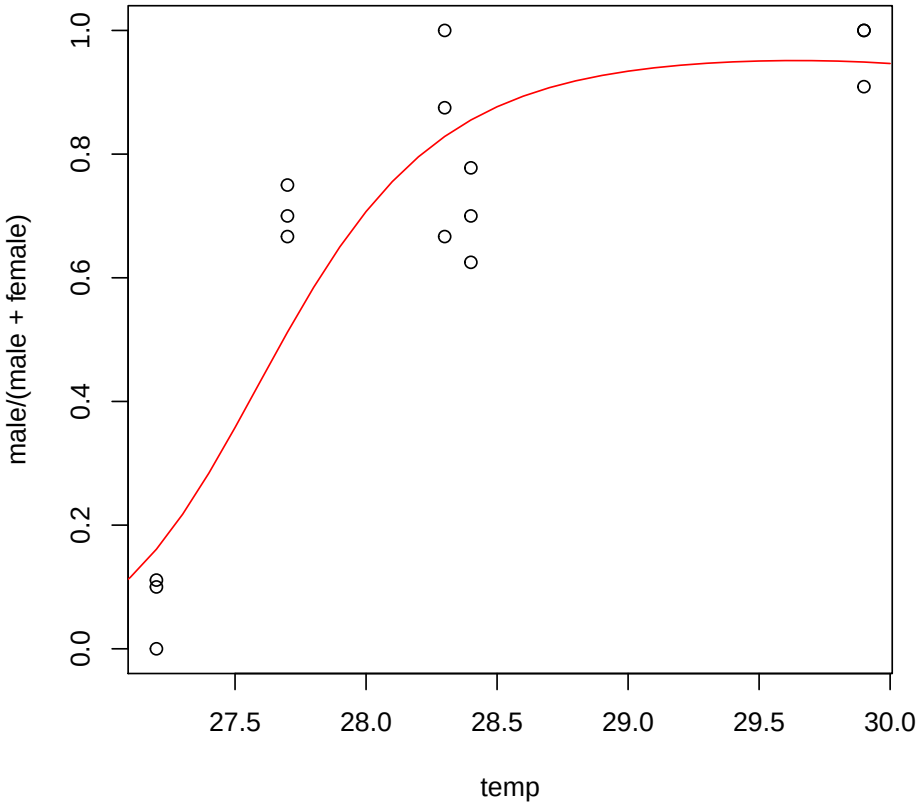
```
[1] 1.438774
```

```
> pchisq(20.256, 12, lower.tail=F)
```

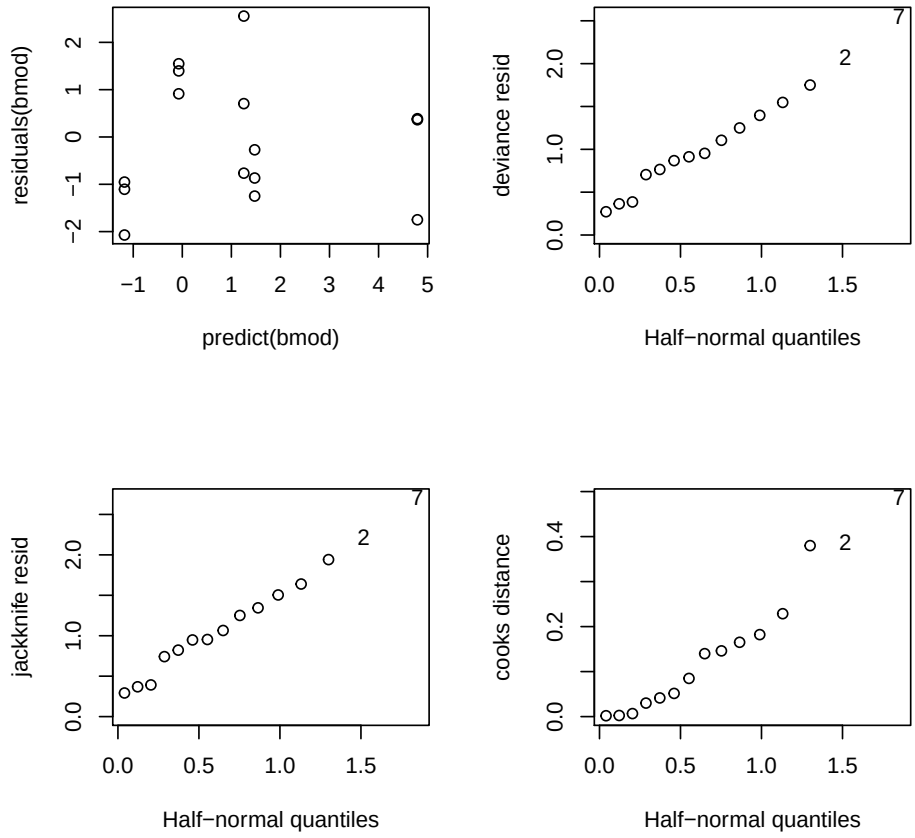
```
[1] 0.06239564
```

The estimate of ϕ has been reduced so overdispersion seems less likely. We can check the fit and residuals for this model

```
> with(turtle, plot(temp, male/(male+female)))
> t <- seq(27, 30, .1)
> lines(t, ilogit(-677.595 + 45.9173*t - 0.7745*t^2), col="red")
```




```
> par(mfrow=c(2,2))
> plot(predict(bmod), residuals(bmod))
> halfnorm(residuals(bmod), ylab="deviance resid")
> halfnorm(rstudent(bmod), ylab="jackknife resid")
> halfnorm(cooks.distance(bmod), ylab="cooks distance")
```



The fit still isn't brilliant for the lower temperatures, but there is no evidence of outliers. Note however that the new model is not without problems of its own. In particular the relationship betweem temperature and the chance of being male is no longer monotonic (though it mostly is in the range of temperatures considered). This makes the model harder to interpret.

In general, if we do not account for overdispersion, then our tests for variable significance will be too sensitive. That is, they may indicate a variable is significant when it really isn't. Similarly, if we do not account for overdispersion, confidence intervals for parameter estimates will be too small See Question 4 for a good example of this.

2. Suppose that $Y_i \sim \text{Poisson}(\lambda_i)$, where $\lambda_i \propto t_i$. For example, if we record the number of burglaries reported in different cities, the observed number will depend on the number of households in these cities. In other cases, the size variable t may be time. For example, if we record the number of customers served by sales people, we must take account of the differing amounts of time worked.

   We can model the rate *per unit time* using a log link via

   $$\log(\lambda_i/t_i) = x_i^T \beta$$

   where x_i are known predictors and β unknown parameters. That is

   $$\log(\lambda_i) = \log t_i + x_i^T \beta.$$

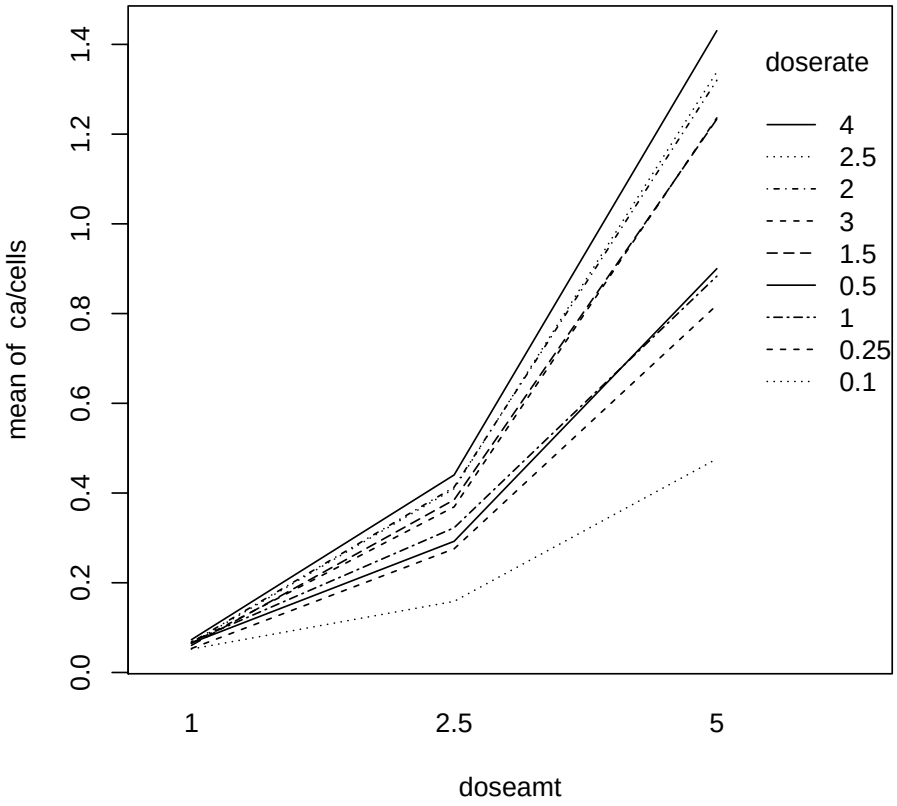
   This is of the form of a Poisson glm with log link, but where the coefficient of $\log t_i$ has been constrained to be 1. This is called a *rate model*.

   In an R model description we can fix the coefficient of a variable to 1 by enclosing it in the `offset` function, viz `y ~ offset(log(t)) + x1 + x2 + ⋯`.

   In Purott and Reeder (1976), some data is presented from an experiment conducted to determine the effect of gamma radiation on the numbers of chromosomal abnormalities (ca) observed. The

number (cells), in hundreds of cells exposed in each run, differs. The dose amount (doseamt) and the rate (doserate) at which the dose is applied are the predictors of interest. We can plot the data as follows

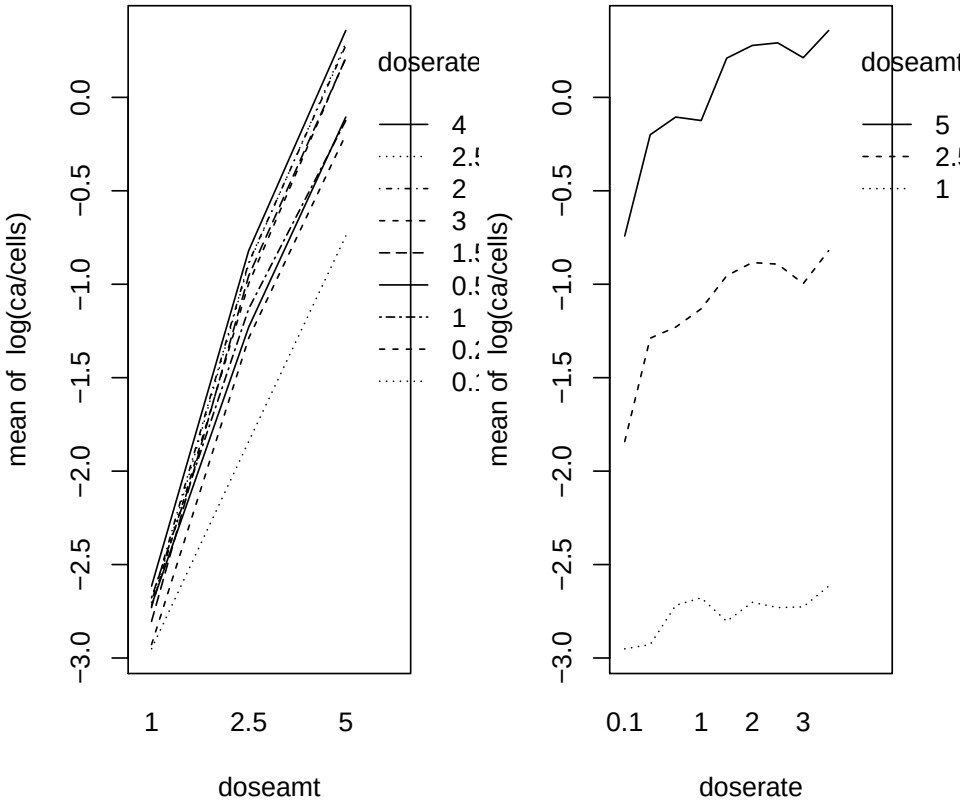
```
> library(faraway)
> data(dicentric)
> with(dicentric, interaction.plot(doseamt, doserate, ca/cells))
```



Fit a rate model to this data.

Solution: Plotting log(ca/cells) against doseamt and doserate helps us judge linearity.

```
> par(mfrow=c(1,2))
> with(dicentric, interaction.plot(doseamt, doserate, log(ca/cells)))
> with(dicentric, interaction.plot(doserate, doseamt, log(ca/cells)))
> par(mfrow=c(1,1))
```



The plots show nice linear relationships between log(ca/cells) and both doseamt and doserate. They also show a possible interaction between doseamt and doserate, since the slope of doserate vs. log(ca/cells) seems to depend on doseamt (and vice versa). We can now fit the model:

```
> model <- glm(ca ~ offset(log(cells)) + doserate*doseamt, family=poisson, data=dicentric)
> summary(model)

Call:
glm(formula = ca ~ offset(log(cells)) + doserate * doseamt, family = poisson,
    data = dicentric)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-5.7308  -2.2842  -0.6264   3.3487   5.8272

Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)     -3.29994    0.06160 -53.567  < 2e-16 ***
doserate         0.06401    0.02922   2.191 0.028476 *
doseamt          0.61224    0.01707  35.862  < 2e-16 ***
doserate:doseamt 0.02715    0.00765   3.549 0.000387 ***
---
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: 4753.00  on 26  degrees of freedom
Residual deviance:  270.26  on 23  degrees of freedom
AIC: 453.67
```

```
Number of Fisher Scoring iterations: 4
```

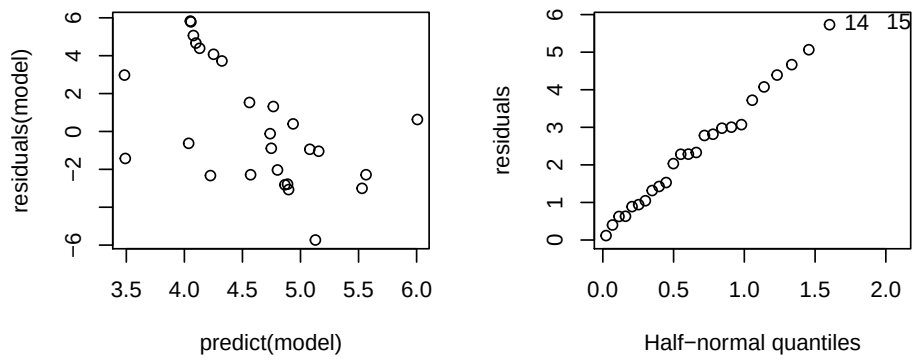
We can test the significance of the interaction using a chi-squared test. Not surprisingly, given its z-value, it appears very significant (but see below).

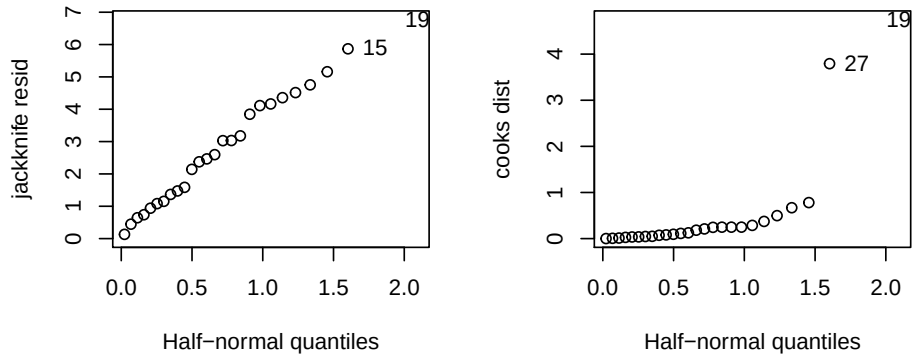
```
> anova(model, test="Chi")

Analysis of Deviance Table

Model: poisson, link: log

Response: ca

Terms added sequentially (first to last)


                 Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
NULL                                26     4753.0
doserate          1    231.3        25     4521.7 < 2.2e-16 ***
doseamt           1   4238.7        24      282.9 < 2.2e-16 ***
doserate:doseamt  1     12.7        23      270.3 0.0003681 ***
---
Signif. codes:  0 âĂŸ***âĂŹ 0.001 âĂŸ**âĂŹ 0.01 âĂŸ*âĂŹ 0.05 âĂŸ.âĂŹ 0.1 âĂŸ âĂŹ 1
```

The deviance of our fitted model is very high. Our counts ca are reasonably large (the smallest is 25), so the deviance should look roughly chi-squared. Thus something is amiss with the model.

The residuals look mostly OK. Points 19 and 27 have a large Cook's distance, but aren't distinguished otherwise. You can check that if you fit a model omitting these points, then the coefficients do not change much and the deviance is still very high.

```
> par(mfrow=c(2,2))
> plot(predict(model), residuals(model))
> halfnorm(residuals(model), ylab="residuals")
> halfnorm(rstudent(model), ylab="jackknife resid")
> halfnorm(cooks.distance(model), ylab="cooks dist")
> par(mfrow=c(1,1))
```





The resaon for the high deviance is overdispersion. We will consider this problem in the Question 4.

3. In Question 3 we fitted a rate model (a type of Poisson regression) to data on the effect of gamma radiation on chromosomal abnormalities.

   Show that these data are overdispersed compared to a Poisson distribution. Next test for an interaction between `doserate` and `doseamt`, firstly without allowing for overdispersion (fixing this dispersion $\phi = 1$), and secondly allowing for overdispersion. Do you get different answers?

   Solution: The fitted model from the Q3:

```
> model <- glm(ca ~ offset(log(cells)) + doserate*doseamt, family=poisson, data=dicentric)
> summary(model)

Call:
glm(formula = ca ~ offset(log(cells)) + doserate * doseamt, family = poisson,
    data = dicentric)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-5.7308  -2.2842  -0.6264   3.3487   5.8272

Coefficients:
                 Estimate Std. Error z value Pr(>|z|)
(Intercept)      -3.29994    0.06160 -53.567  < 2e-16 ***
doserate          0.06401    0.02922   2.191 0.028476 *
doseamt           0.61224    0.01707  35.862  < 2e-16 ***
doserate:doseamt  0.02715    0.00765   3.549 0.000387 ***
---
```

```
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: 4753.00  on 26  degrees of freedom
Residual deviance:  270.26  on 23  degrees of freedom
AIC: 453.67

Number of Fisher Scoring iterations: 4
```

The deviance of our fitted model is very high because of overdispersion. We estimate the dispersion parameter and get something much larger than 1. We then use our estimate to scale the variance of our estimates, and repeat our significance test for the interaction term (using an F test). We see that the interaction no longer appears significant.

Note that the command `drop1` does *not* work for overdispersed models.

```
> (phi <- sum(residuals(model, type="pearson")^2)/23)

[1] 12.97226

> model1 <- glm(ca ~ offset(log(cells)) + doserate*doseamt, family=quasipoisson, data=dicentric)
> summary(model1) # same as summary(model, dispersion=phi)

Call:
glm(formula = ca ~ offset(log(cells)) + doserate * doseamt, family = quasipoisson,
    data = dicentric)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-5.7308  -2.2842  -0.6264   3.3487   5.8272

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     -3.29994    0.22188 -14.873 2.73e-13 ***
doserate         0.06401    0.10524   0.608    0.549
doseamt          0.61224    0.06149   9.957 8.29e-10 ***
doserate:doseamt  0.02715    0.02755   0.985    0.335
---
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 quasipoisson family taken to be 12.97228)

    Null deviance: 4753.00  on 26  degrees of freedom
Residual deviance:  270.26  on 23  degrees of freedom
AIC: NA

Number of Fisher Scoring iterations: 4

> anova(model1, test="F") # same as anova(model, dispersion=phi, test="F")

Analysis of Deviance Table

Model: quasipoisson, link: log

Response: ca

Terms added sequentially (first to last)


                Df Deviance Resid. Df Resid. Dev        F    Pr(>F)
NULL                             26     4753.0
doserate         1    231.3        25     4521.7  17.8319 0.0003232 ***
doseamt          1   4238.7        24      282.9 326.7535  4.33e-15 ***
```

```
doserate:doseamt  1     12.7        23      270.3   0.9781 0.3329689
---
Signif. codes:  0 âĂŸ***âĂŹ 0.001 âĂŸ**âĂŹ 0.01 âĂŸ*âĂŹ 0.05 âĂŸ.âĂŹ 0.1 âĂŸ âĂŹ 1
```

We refit omitting the interaction, to get our final model.

```
> model2 <- glm(ca ~ offset(log(cells)) + doserate + doseamt, family=poisson, data=dicentric)
> (phi2 <- sum(residuals(model2, type="pearson")^2)/24)

[1] 12.72343

> summary(model2, dispersion=phi2)

Call:
glm(formula = ca ~ offset(log(cells)) + doserate + doseamt, family = poisson,
    data = dicentric)

Deviance Residuals:
   Min      1Q  Median      3Q     Max
-6.761  -1.696  -0.401   3.286   5.798

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -3.46115    0.15430 -22.432   <2e-16 ***
doserate     0.15501    0.04881   3.176   0.0015 **
doseamt      0.66230    0.03456  19.163   <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 12.72343)

    Null deviance: 4753.00  on 26  degrees of freedom
Residual deviance:  282.95  on 24  degrees of freedom
AIC: 464.35

Number of Fisher Scoring iterations: 4
```