BIOS 6612 Homework 4: Poisson regression
Solutions

A study has been carried out to determine the relationship between incidence of non-melanoma skin cancer (outcome) and age and city of residence (predictors) among women. The file `skincancer.csv` contains the data from this study. There are four variables:

- `city`: a factor with two levels for city of residence, either Minneapolis or Dallas.
- `age.group`: a factor with age in years, given in ranges.
- `cases`: number of incident cases of non-melanoma skin cancer (Y_i).
- `py1000`: total person-time at risk, in 1000s of person-years (T_i).

You will need to fit several Poisson regression models to answer the following questions. Provide code and output for your analyses with your answers to this assignment.

1. (**15 points**) For descriptive purposes, the simple regression model

$$\log \mathbb{E}(Y_i) = \log T_i + \beta_0 + \beta_1 \texttt{Dallas}_i$$

is of interest. Let $\texttt{Dallas}_i$ be an indicator variable for city of residence, with $\texttt{Dallas}_i = 1$ if the city of residence is Dallas and 0 if the city of residence is Minneapolis.

   (a) Fit this model; provide estimates and standard errors for each regression coefficients. (**5 points**)

```
> mod0 <- glm(cases ~ offset(log(py1000)) + city,
+             data=nonmelanoma.data,family=poisson)
> summary(mod0)

Call:
glm(formula = cases ~ offset(log(py1000)) + city, family = poisson,
    data = nonmelanoma.data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-23.353   -7.819    4.222   10.239   19.015
```

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.18200    0.04295  -4.237 2.26e-05 ***
cityDallas   0.72350    0.05135  14.090  < 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: 2740.5  on 15  degrees of freedom
Residual deviance: 2525.8  on 14  degrees of freedom
AIC: 2627

Number of Fisher Scoring iterations: 6
```

(b) Interpret this model. Include estimates and confidence intervals for the rate of non-melanoma skin cancer in each city as part of your response. Interpret the intercept or explain why it is not interpretable. (**10 points**)
This model allows for a different rate of non-melanoma skin cancer in each city, but does not distinguish between age groups; that is, the rates are assumed not to vary across age groups. The intercept is the log rate of non-melanoma skin cancer in Minnesota. We estimate a rate of non-melanoma skin cancer of 0.8336 cases per 1000 person-years in Minnesota (95% CI: 0.7663–0.9068). The rate in Dallas is 1.719 cases per 1000 person-years (95% CI: 1.626–1.816).

```
> # rate in Minneapolis
> exp(mod0$coefficients[1])
(Intercept)
     0.8336
> exp(confint.default(mod0)[1,])
 2.5 % 97.5 %
0.7663 0.9068
> # rate in Dallas
> exp(sum(mod0$coefficients))
[1] 1.719
> V0 <- vcov(mod0)
> se.dallasrate <- sqrt(V0[1,1]+V0[2,2]+2*V0[1,2])
> exp(sum(mod0$coefficients) + c(-1,1)*qnorm(1-.05/2)*se.dallasrate)
[1] 1.626 1.816
```

2. (**30 points**) It may be important to adjust for age when modeling the incidence of non-melanoma skin cancer.

(a) Fit a Poisson regression model for the effect of city on rate of non-melanoma skin

cancer adjusting for age. That is, include terms for both age group and city in your model. Report your estimated model coefficients and standard errors. (**5 points**)

```
> mod1 <- glm(cases ~ offset(log(py1000)) + city + age.group,
+             data=nonmelanoma.data,family=poisson)
>
> summary(mod1)

Call:
glm(formula = cases ~ offset(log(py1000)) + city + age.group,
    family = poisson, data = nonmelanoma.data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.5724  -0.5556  -0.0207   0.6778   1.7576

Coefficients:
               Estimate Std. Error z value Pr(>|z|)
(Intercept)     -3.4496     0.2384  -14.47  < 2e-16 ***
cityDallas       0.7868     0.0514   15.30  < 2e-16 ***
age.group25-34   1.3046     0.2735    4.77  1.8e-06 ***
age.group35-44   2.5602     0.2495   10.26  < 2e-16 ***
age.group45-54   3.3305     0.2427   13.72  < 2e-16 ***
age.group55-64   3.8266     0.2413   15.86  < 2e-16 ***
age.group65-74   4.3570     0.2405   18.12  < 2e-16 ***
age.group75-84   4.7971     0.2414   19.87  < 2e-16 ***
age.group85+     4.8988     0.2549   19.22  < 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: 2740.498  on 15  degrees of freedom
Residual deviance:   12.665  on  7  degrees of freedom
AIC: 127.8

Number of Fisher Scoring iterations: 4
```

(b) A previous study found that age-adjusted skin cancer rates in Dallas were double those in Minneapolis. Carry out a hypothesis test of whether results from this study are consistent with those from the previous study. Make sure to explicitly state your null and alternative hypothesis, the test statistic and it's null distribu-

tion, and the conclusion of your hypothesis test. (**10 points**)
From this model, we estimate that the age-adjusted rate ratio for skin cancer between Dallas and Minneapolis is $\exp(0.7868) = 2.196$. This seems close to 2, but we can carry out a Wald test to be statistically precise about that. If β_1 is the coefficient for city, then $H_0 : \beta_1 = \log(2) = 0.6931$. The test statistic is $((0.7868 - 0.6931)/0.0514)^2 = 3.317$, which we can compare to a reference χ^2 with 1 degree of freedom to obtain $p = 0.06858$. We fail to reject the null hypothesis and conclude that the results of the current study are not inconsistent with those of the previous study.

(c) Carry out a test of the hypothesis that rates of skin cancer in the 15–24 and 25–34 age groups are equal. (**10 points**)
We have chosen 15–24 as the reference group, so to evaluate this null hypothesis, we just need to test $H_0 : \beta_2 = 0$, if β_2 is the coefficient associated with the indicator variable for being in the 25–34 age group. The Wald z statistic is 4.77 (from the table of coefficient estimates), with associated $p < 0.0001$. Therefore we reject the null hypothesis and conclude that there is evidence that there is a statistically significant difference in the rate of skin cancer between people ages 15–24 and people ages 25–34.

(d) Based on this model, what is the estimated rate of skin cancer among women in Minneapolis aged 45–54? (**5 points**)
Minneapolis is the reference group for city, so to find the estimated rate for women in this city aged 45–54, we need the the coefficient estimates for the intercept and the relevant age group: that is, the estimated rate is $\exp(-3.4496 + 3.3305) = 0.8877$ cases per 1000 person-years.

3. (**15 points**) The saturated model for this data includes (in addition to main effects) the interaction between age and city of residence.

(a) Fit this model and carry out a likelihood ratio test of the null hypothesis that the rate ratio for skin cancer between Dallas and Minneapolis does not depend on age. (**10 points**)
The previous model (with just main effects for age and city) is nested within the saturated (interaction) model, so we just need to take the difference in deviance statistics between the two models to carry out the LRT. This gives a value of 12.6645 on 7 degrees of freedom, for $p = 0.0807$. We fail to reject the null hypothesis and conclude that there is no statistical evidence that the rate ratio of skin cancer for Dallas versus Minneapolis depends on age.

```
> mod2 <- glm(cases ~ offset(log(py1000)) + city * age.group,
+             data=nonmelanoma.data,family=poisson)
> anova(mod1,mod2,test='LRT')
Analysis of Deviance Table

Model 1: cases ~ offset(log(py1000)) + city + age.group
Model 2: cases ~ offset(log(py1000)) + city * age.group
```

```
Resid. Df Resid. Dev Df  Deviance Pr(>Chi)
1         7  12.664504
2         0   0.000000  7 12.664504 0.080717 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’
1
```

(b) Explain another way to perform this test *without* actually fitting the interaction model. (**5 points**)

Since there are only two covariates here, the interaction model *is* the saturated model. The residual deviance of the model with both main effect terms gives the difference in deviance between the saturated model and the model without the interaction, so you can use the residual deviance from the previous model to perform this same LRT.