PM592: Regression Analysis for Data Science

Name: Flemming Wu

HW10

Regression for Count Outcomes

Instructions

- Answer questions directly within this document.
- Upload to Blackboard by the due date & time.
- Clearly indicate your answers to all questions.
- If a question requires analysis, attach all relevant output to this document in the appropriate area. Do not attach superfluous output.
- For the purpose of this assignment, statistical evidence refers to a test statistic and associated p-value.
- If a question requires a conclusion, it must be phrased professionally and coherently.
- There are 2 questions and 30 points possible.

Question 1 [20 points]

A study was performed on the number of live births among women living in Fiji. Investigators wanted to determine factors related to the number of live births in this population. The data is located in **fiji.dta**.

$Y_{TOTBORN} = \#$ of live births within the covariate pattern

$$X_{DUR} = \begin{cases} 1, \text{married } 0-4 \text{ years} \\ 2, \text{married } 5-9 \text{ years} \\ 3, \text{married } 10-14 \text{ years} \\ 4, \text{married } 15-19 \text{ years} \\ 5, \text{married } 20-24 \text{ years} \\ 6, \text{married } > 24 \text{ years} \end{cases} \qquad X_{RES} = \begin{cases} 1, \text{lives in Suva} \\ 2, \text{lives in another urban community} \\ 3, \text{lives in a rural area} \end{cases}$$

$$X_{EDUC} = \begin{cases} 1, \text{no formal education} \\ 2, \text{lower primary education} \\ 3, \text{upper primary education} \\ 4, \text{secondary education or higher} \end{cases} \qquad X_N = \#\text{ of women within the covariate pattern}$$

- Use Poisson regression with the "offset" option.

1a. [2 points] Which variable will you use as the offset? Explain why you'd use this particular variable and why it is necessary for the model.

I will use X_N as the offset. Since the research question is to address what covariates relate to the number of live births, the number of women for each covariate pattern should be used as an offset. Since the number of women in each covariate pattern would influence the number of live births, it would make more sense to model the rate of live births instead of raw counts. There would naturally be an overall higher count of live births observed in a covariate pattern that is more common, so this factor should be controlled for in the model.

1b. [4 points] Run a null model (no independent variables) and report the estimate (and 95% CI) for the mean rate of live births per woman. Based on this output and examining the GOF statistics, does it appear that the mean number of live births is equivalent across women in this population? Why or why not?

```
Call:
glm(formula = totborn ~ 1 + offset(log(n)), family = poisson,
    data = fiji)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) 1.376346   0.009712   141.7   <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: 3731.9  on 69  degrees of freedom
Residual deviance: 3731.9  on 69  degrees of freedom
AIC: 4163.3

Number of Fisher Scoring iterations: 5
```

The estimate for the mean live births per woman is $e^{1.37}$= 3.96 (CI = (3.89, 4.04)).

```
> pois_dev_gof(null_model)
$pval
[1] 0

$df
[1] 69

> pois_pearson_gof(null_model)
$pval
[1] 0

$df
[1] 69
```

According to the Pearson Chi-Square goodness of fit test, and the deviance Chi-Square goodness of fit test, the null model in which the mean is predicted for every observation seems to deviate from good fit significantly ($p < 0.01; p < 0.01$). This seems to suggest that the mean number of live births is not equivalent across women in this population.

1c. [7 points] Provide the LR chi-square and p-value for the effect of education on rate of live births, including an interpretation of the effect of education. Use the emmeans commands (see lab assignment) to obtain estimates and 95% CI of rate of live births per woman by education level. Do GOF statistics indicate this model fits well?

First, determine whether to encode education as a dummy variable set, or keep as linear.

```
> educ_lin_model <- glm(totborn ~ educ + offset(log(n)), family = poisson, data = fiji)
> educ_fact_model <- glm(totborn ~ factor(educ) + offset(log(n)), family = poisson, data = fiji)
> summary(educ_lin_model)

Call:
glm(formula = totborn ~ educ + offset(log(n)), family = poisson,
    data = fiji)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  2.03080    0.02215   91.67   <2e-16 ***
educ        -0.33840    0.01110  -30.49   <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: 3731.9  on 69  degrees of freedom
Residual deviance: 2725.5  on 68  degrees of freedom
AIC: 3159

Number of Fisher Scoring iterations: 5

> summary(educ_fact_model)

Call:
glm(formula = totborn ~ factor(educ) + offset(log(n)), family = poisson,
    data = fiji)

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)    1.64728    0.01469  112.11   <2e-16 ***
factor(educ)2 -0.21179    0.02168   -9.77   <2e-16 ***
factor(educ)3 -0.61605    0.02886  -21.35   <2e-16 ***
factor(educ)4 -1.22468    0.05141  -23.82   <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: 3731.9  on 69  degrees of freedom
Residual deviance: 2661.0  on 66  degrees of freedom
AIC: 3098.5

Number of Fisher Scoring iterations: 5

> anova(educ_lin_model, educ_fact_model, test="LRT")
Analysis of Deviance Table

Model 1: totborn ~ educ + offset(log(n))
Model 2: totborn ~ factor(educ) + offset(log(n))
  Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
1        68     2725.5
2        66     2661.0  2   64.541 9.662e-15 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Comparing the AIC of the dummy variable model to the linear model, it appears that using a dummy encoding scheme decreases the AIC by 60.5. Additionally, the likelihood ratio test shows that a dummy

encoding scheme significantly improves model fit ($\chi^2_2 = 64.5, p < 0.001$). I will use the dummy variable set to represent education.

According to the Poisson regression model in which education was encoded as a dummy variable set, women with lower primary education are expected to have $e^{-0.21} = 0.809$ times the mean number of live births per woman that is expected for women with no education (p < 0.001). Women with upper primary education are expected to have $e^{-0.61} = 0.540$ times the mean number of live births per woman that is expected for women with no education (p < 0.001). Finally, women with secondary education or higher are expected to have $e^{-1.22} = 0.294$ times the mean number of live births per woman that is expected for women with no education. Overall, education level is significantly associated with expected rate of live births ($\chi^2_3 = 1070, p < 0.001$) and seems to have a negative relationship with the rate.

The expected rate of live births for each education level are shown below: (The results are transformed from the log-scale because parameter type = "response" was provided. If this is not specified, the results will be given on the log scale.)

```
> emmeans(educ_fact_model, "educ", offset = log(1), type = "response")
 educ rate     SE  df asymp.LCL asymp.UCL
    1 5.19 0.0763 Inf      5.05      5.34
    2 4.20 0.0670 Inf      4.07      4.34
    3 2.80 0.0697 Inf      2.67      2.94
    4 1.53 0.0752 Inf      1.39      1.68

Confidence level used: 0.95
Intervals are back-transformed from the log scale
```

```
> pois_dev_gof(educ_fact_model)
$pval
[1] 0

$df
[1] 66

> pois_pearson_gof(educ_fact_model)
$pval
[1] 0

$df
[1] 66
```

According to the Pearson Chi-Square goodness of fit test, and the deviance Chi-Square goodness of fit test, the Poisson regression model using education level still seems to deviate from good fit significantly ($p < 0.01; p < 0.01$).

1d. [7 points] Provide the LR chi-square and p-value for the effect of education on rate of live births, adjusting for residence and marriage duration. Include an interpretation of the effect of education. Output the margins for education level and provide a professionally formatted table that shows the difference between the adjusted live birth rates from 1d, compared to the unadjusted live birth rates in 1c. Do GOF statistics indicate this model fits well?

Need to determine functional form for 'dur' (length of marriage). Residence is for sure a categorical (factor) variable, as it has no order.

```
> anova(
+   glm(totborn ~ factor(educ) + factor(res) + dur + offset(log(n)), family = poisson, data = fiji),
+   glm(totborn ~ factor(educ) + factor(res) + factor(dur) + offset(log(n)), family = poisson, data = fiji),
+   test="LRT"
+ )
Analysis of Deviance Table

Model 1: totborn ~ factor(educ) + factor(res) + dur + offset(log(n))
Model 2: totborn ~ factor(educ) + factor(res) + factor(dur) + offset(log(n))
  Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
1        63     416.98
2        59      70.67  4   346.32 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

It appears that categorical encoding of 'dur' significantly improves model fit.

```
Call:
glm(formula = totborn ~ educ.f + res.f + dur.f + offset(log(n)),
    family = poisson, data = fiji)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.11710    0.05491  -2.132 0.032969 *
educ.f2      0.02297    0.02266   1.014 0.310597
educ.f3     -0.10127    0.03099  -3.268 0.001082 **
educ.f4     -0.31015    0.05521  -5.618 1.94e-08 ***
res.f2       0.11242    0.03250   3.459 0.000541 ***
res.f3       0.15166    0.02833   5.353 8.63e-08 ***
dur.f2       0.99693    0.05274  18.902  < 2e-16 ***
dur.f3       1.36940    0.05107  26.815  < 2e-16 ***
dur.f4       1.61376    0.05119  31.522  < 2e-16 ***
dur.f5       1.78491    0.05121  34.852  < 2e-16 ***
dur.f6       1.97641    0.05003  39.501  < 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: 3731.852  on 69  degrees of freedom
Residual deviance:   70.665  on 59  degrees of freedom
AIC: 522.14

Number of Fisher Scoring iterations: 4
```

```
> anova(educ_fact_model, educ_adj_model, test = "LRT")
Analysis of Deviance Table

Model 1: totborn ~ educ.f + offset(log(n))
Model 2: totborn ~ educ.f + res.f + dur.f + offset(log(n))
  Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
1        66    2661.00
2        59      70.67  7   2590.3 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

After adjusting for residence and marriage duration the interpretation of the effect of education is as follows: women with lower primary education are expected to have ~~$e^{-0.21} = 1.06$~~ $e^{0.023} = 1.02$ times the mean number of live births per woman that is expected for women with no education (~~p = 0.006~~) (p=0.31), adjusting for residence and marriage duration. Women with upper primary education are expected to have ~~$e^{-0.095} = 0.910$~~ $e^{-0.101} = 0.90$ times the mean number of live births per woman that is expected for women with no education (p = 0.001), adjusting for residence and marriage duration. Finally, women with secondary education or higher are expected to have ~~$e^{-0.406} = 0.667$~~ $e^{-0.31} = 0.73$ times the mean number of live births per woman that is expected for women with no education, adjusting for residence and marriage duration (p < 0.001). Overall, education level is significantly associated with expected rate of live births ($\chi^2_2 = 2241, p < 0.001$).

	totborn			totborn		
Predictors	*Incidence Rate Ratios*	*CI*	*p*	*Incidence Rate Ratios*	*CI*	*p*
(Intercept)	5.19	5.04 – 5.34	**<0.001**	0.89	0.80 – 0.99	**0.033**
educ.f: educ.f2	0.81	0.78 – 0.84	**<0.001**	1.02	0.98 – 1.07	0.311
educ.f: educ.f3	0.54	0.51 – 0.57	**<0.001**	0.90	0.85 – 0.96	**0.001**
educ.f: educ.f4	0.29	0.27 – 0.32	**<0.001**	0.73	0.66 – 0.82	**<0.001**
res.f: res.f2				1.12	1.05 – 1.19	**0.001**
res.f: res.f3				1.16	1.10 – 1.23	**<0.001**
dur.f: dur.f2				2.71	2.45 – 3.01	**<0.001**
dur.f: dur.f3				3.93	3.56 – 4.35	**<0.001**
dur.f: dur.f4				5.02	4.55 – 5.56	**<0.001**
dur.f: dur.f5				5.96	5.39 – 6.59	**<0.001**
dur.f: dur.f6				7.22	6.55 – 7.97	**<0.001**
Observations	70			70		
R^2 Nagelkerke	1.000			1.000		

In comparing the unadjusted model to the adjusted model, it appears that residence and marriage duration do confound the relationship between education level and rate of live births. The parameter estimates for each of the education levels has changed by over 15%. However, the p-values of the parameter estimates still remains significant.

```
> pois_dev_gof(educ_adj_model)
$pval
[1] 0.1421387

$df
[1] 59

> pois_pearson_gof(educ_adj_model)
$pval
[1] 0.1268647

$df
[1] 59
```

According to the Pearson Chi-Square goodness of fit test, and the deviance Chi-Square goodness of fit test, the Poisson regression model using education level and adjusting for residence and education do not deviate from good fit significantly ($p = 0.14, p = 0.13$).

Question 2 [10 points]

Use the same data to run a negative binomial regression with the offset option on the full model (with all covariates). Note: in the `glm.nb()` function, specify the offset within the formula.

You should get a warning message about the iteration limit for theta being reached. This means that the maximum likelihood algorithm is having a difficult time finding an estimate for θ.

2a. [3 points] Provide your model output and explain how this model differs from your model in 1d. What information is θ adding to the model?

```
Call:
MASS::glm.nb(formula = totborn ~ educ.f + res.f + dur.f + offset(log(n)),
    data = fiji, init.theta = 809405.8048, link = log)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.11709    0.05492  -2.132 0.032988 *
educ.f2      0.02297    0.02266   1.014 0.310794
educ.f3     -0.10128    0.03099  -3.268 0.001083 **
educ.f4     -0.31016    0.05521  -5.617 1.94e-08 ***
res.f2       0.11243    0.03250   3.459 0.000541 ***
res.f3       0.15166    0.02833   5.353 8.67e-08 ***
dur.f2       0.99694    0.05275  18.901  < 2e-16 ***
dur.f3       1.36939    0.05107  26.813  < 2e-16 ***
dur.f4       1.61375    0.05120  31.519  < 2e-16 ***
dur.f5       1.78489    0.05122  34.849  < 2e-16 ***
dur.f6       1.97640    0.05004  39.497  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for Negative Binomial(809403.6) family taken to be 1)

    Null deviance: 3730.420  on 69  degrees of freedom
Residual deviance:   70.656  on 59  degrees of freedom
AIC: 524.15

Number of Fisher Scoring iterations: 1


              Theta:  809406
          Std. Err.:  7528400
Warning while fitting theta: alternation limit reached

 2 x log-likelihood:  -500.146
```

The negative binomial model differs from the Poisson model in that the ~~parameter estimates have changed, and also that the p-values for education levels 2 and 3 and residence are no longer significant.~~ The p-values and standard errors for the parameter estimates are slightly higher, but overall they are very similar. The information θ adds to the model is the overdispersion factor, which allows for more variance in the count data than is what is assumed in Poisson. In Poisson regression, the variance is

assumed to be equal to the mean, but in negative binomial regression, the variance is modeled as being equal to the mean plus $\frac{\mu^2}{\theta}$. This means that θ is assumed to be constant over all values of X.

2b. [3 points] What is the value of the dispersion parameter that the model found? What do you think this means in terms of overdispersion?

~~The value of the dispersion parameter that the model found is 17.97 (CI = (13.84, 22.10)). The formula for the variance of Y is given by $Var(Y) = \mu + \frac{\mu^2}{\theta}$, so as $\theta \to \infty$ the negative binomial distribution would converge to a Poisson distribution. Since the estimated value of θ is pretty low, I would argue that there is a high degree of overdispersion in the data.~~

The value of the overdispersion parameter was actually taken to be one. This means that the data was actually not overdispersed and was fit by the Poisson model well.

2c. [4 points] Compare the AIC value for the Poisson and negative binomial models. Based on this value, state which model you prefer, and provide at least 2 reasons why.

The Akaike information criterion (AIC) for the Poisson and negative binomial models is ~~861.24 and 643.29~~ 522.14 and 524.15 respectively. A difference of at least 10 in AIC values for two different models would suggest significant difference in fit, and since the difference is ~~much~~ not greater than 10 for these models, I would argue that there is a no need for the added complexity of the overdispersion parameter in the negative binomial model. Additionally, the lower AIC value for the negative binomial model suggests that a lower amount of information is lost by modeling the data as negative binomial, pointing to the negative binomial model being a better choice.