P8110: Applied Regression II
Homework #7 Solution
[10 points]

School administrators study the attendance behavior of high school juniors at two schools. Predictors of the number of days of absence include gender of the student, standardized test scores in math and language arts. The data are saved in "HW7.csv". The columns of variables from left to right are:

ID	= student ID
school	= 1 - school A; 2 - school B
male	= 1 - male; 0 - female
math	= standardized test score in math
langarts	= standardized test score in language arts
daysabs	= days absent

1. **Fit a Poisson regression model with days of school absence as the outcome and school, gender, and standardized test scores in math and language arts as covariates (model 1). Write the model. Is overdispersion a potential problem for this Poisson model? [2 points]**

   Model (1):

   $$\log E(\text{daysabs}_i | \text{male}_i, \ \text{school}_i, \ \text{langarts}_i, \ \text{math}_i)$$
   $$= \ \beta_0 + \beta_1 \text{male}_i + \beta_2 \text{school}_i + \beta_3 \text{langarts}_i + \beta_4 \text{math}_i$$

   where N_i denotes the unit of exposure such as the observation time window. Based on the SAS output, the value of Deviance over DF is 6.4638, which is much larger than 1, that may indicates evidence of over-dispersion.
   SAS output:

Criteria For Assessing Goodness Of Fit			
Criterion	DF	Value	Value/DF
Deviance	311	2010.2269	6.4638
Scaled Deviance	311	2010.2269	6.4638
Pearson Chi-Square	311	2492.5965	8.0148
Scaled Pearson X2	311	2492.5965	8.0148
Log Likelihood		1594.4267	
Full Log Likelihood		-1435.8113	
AIC (smaller is better)		2881.6226	
AICC (smaller is better)		2881.8161	
BIC (smaller is better)		2900.4013	

2. **Refit model 1 with the scale parameter being equal to Deviance divided by DF. Estimate the absence rate ratio between male and female students. Provide the 95% confidence interval and interpret.[3 points]**

   The estimate of the school absence rate ratio between male students and female students adjusted for the standardized test score of language arts, math, and school is

$$RR = \frac{E(\text{daysabs}_i|\text{male}_i = 1, \text{school}_i, \text{langarts}_i, \text{math}_i)}{E(\text{daysabs}_j|\text{male}_j = 0, \text{school}_j, \text{langarts}_j, \text{math}_j)}$$

$$= \frac{\exp(\beta_0 + \beta_1 + \beta_2 + \beta_3 + \beta_4)}{\exp(\beta_0 + \beta_2 + \beta_3 + \beta_4)} = \exp(\beta_1)$$

the foregoing rate ratio $RR = \exp(\beta_1)$.

According to the SAS output, the estimate of β_1 is $\hat{\beta}_1 = -0.4331$ with 95% confidence interval [-0.6733,-0.1928].

Therefore, the absence rate ratio between male students and female students , given the test score in math, language arts, and school is estimated to be $\widehat{RR} = \exp(-0.4331) = 0.6485$ with 95% confidence to lie between $\exp(-0.6733) = 0.5100$ and $\exp(-0.1928) = 0.8246$.

Interpretation: The absence rate among male students is 0.6485 times of that among female students, controlled for the test score in math, language arts, and school, and this rate ratio could be as little as 0.5100 and as much as 0.8246.

SAS output:

Analysis Of Maximum Likelihood Parameter Estimates						
Parameter	**DF**	**Estimate**	**Standard Error**	**Wald 95% Confidence Limits**		**Wald Chi-Square**	**Pr > ChiSq**
Intercept	1	3.3751	0.2193	2.9452	3.8050	236.75	<.0001
school	1	-0.8100	0.1416	-1.0876	-0.5325	32.72	<.0001
male	1	-0.4331	0.1226	-0.6733	-0.1928	12.48	0.0004
math	1	0.0005	0.0046	-0.0086	0.0095	0.01	0.9176
langarts	1	-0.0066	0.0049	-0.0161	0.0030	1.82	0.1771
Scale	0	2.5424	0.0000	2.5424	2.5424		

3. **Refit model 1 using negative binomial regression. Provide a formal test to decide whether a negative binomial model is needed for this data than a Poisson regression model? [2 points]**

   Here we perform a t-test for the over-dispersion parameter α: H_0: $\alpha = 0$ vs. H_a: $\alpha \neq 0$

   From the SAS output, we can find the statistic value of α is 10.10, and the p-value is smaller than 0.0001, so we can draw the conclusion that we reject H_0 at 95% significant level. And a negative binomial model is more appropriate.

   SAS output:

4. **Use the negative binomial model to estimate the absence rate ratio between male and female students and provide 95% confidence interval. Is the conclusion different from the Poisson model in part (2)? [3 points]**

   The estimate of the school absence rate ratio between male students and female students adjusted for the standardized test score of language arts, math, and school is

Parameter Estimates					
Parameter	DF	Estimate	Standard Error	t Value	Approx Pr > |t|
Intercept	1	3.371968	0.256674	13.14	<.0001
school	1	-0.769076	0.143066	-5.38	<.0001
male	1	-0.395827	0.134562	-2.94	0.0033
math	1	0.002034	0.004847	0.42	0.6747
langarts	1	-0.009568	0.005398	-1.77	0.0763
_Alpha	1	1.151089	0.114024	10.10	<.0001

$$
\begin{aligned}
RR &= \frac{E(\text{daysabs}_i|\text{male}_i=1, \text{school}_i, \text{langarts}_i, \text{math}_i)}{E(\text{daysabs}_j|\text{male}_j=0, \text{school}_j, \text{langarts}_j, \text{math}_j)} \\
&= \frac{\exp(\beta_0 + \beta_1 + \beta_2 + \beta_3 + \beta_4)}{\exp(\beta_0 + \beta_2 + \beta_3 + \beta_4)} = \exp(\beta_1)
\end{aligned}
$$

the foregoing rate ratio $RR = \exp(\beta_1)$.

According to the SAS output, the estimate of β_1 is $\hat{\beta}_1 = -0.3958$ with 95% confidence interval [-0.6596,-0.1320].

$(-0.3958 - 1.96*0.1346 = -0.6596$, and $-0.3958 + 1.96*0.1346 = -0.1320)$

Therefore, the absence rate ratio between male students and female students , given the test score in math, langarts, and school is estimated to be $\widehat{RR} = \exp(-0.3958) = 0.6731$ with 95% confidence to lie between $\exp(-0.6596) = 0.5171$ and $\exp(-0.1321) = 0.8763$.

The conclusion is not different from the Poisson model in part (2).

SAS output:

Parameter Estimates					
Parameter	DF	Estimate	Standard Error	t Value	Approx Pr > |t|
Intercept	1	3.371968	0.256674	13.14	<.0001
school	1	-0.769076	0.143066	-5.38	<.0001
male	1	-0.395827	0.134562	-2.94	0.0033
math	1	0.002034	0.004847	0.42	0.6747
langarts	1	-0.009568	0.005398	-1.77	0.0763
_Alpha	1	1.151089	0.114024	10.10	<.0001

SAS code:

```
data hw7;
infile 'HW7.csv' delimiter=',' missover dsd;
input id school male math langarts daysabs;
run;

proc genmod data=hw7;
model daysabs = school male math langarts/ dist=poi link=log type3;
run;

proc genmod data=hw7;
model daysabs = school male math langarts/ dist=poi link=log type3 scale=D;
run;

proc countreg data=hw7;
model daysabs = school male math langarts/ dist=negbin(p=2) ;
run;
```