Poisson Regression with R

Now let's look at the output of function glm more closely.

- The output begins with echoing the function call. The information on deviance residuals is displayed next.
- Deviance residuals are approximately normally distributed if the model is specified correctly.In our example, it shows a little bit of skeweness since median is not quite zero.

- Next come the Poisson regression coefficients for each of the variables along with the standard errors, z-scores, p-values and 95% confidence intervals for the coefficients. The coefficient for math is .07.This means that the expected log count for a one-unit increase in math is .07.

- The indicator variable progAcademic compares between prog = "Academic" and prog = "General", the expected log count for prog = "Academic" increases by about 1.1. The indicator variable prog.Vocational is the expected difference in log count ($\approx .37$) between prog = "Vocational" and the reference group (prog = "General").

- The information on deviance is also provided. We can use the residual deviance to perform a goodness of fit test for the overall model. The residual deviance is the difference between the deviance of the current model and the maximum deviance of the ideal model where the predicted values are identical to the observed.

- Therefore, if the residual difference is small enough, the goodness of fit test will not be significant, indicating that the model fits the data. We conclude that the model fits reasonably well because the goodness-of-fit chi-squared test is not statistically significant.
- If the test had been statistically significant, it would indicate that the data do not fit the model well. In that situation, we may try to determine if there are omitted predictor variables, if our linearity assumption holds and/or if there is an issue of over-dispersion.

Poisson Regression with R

```
with(m1, cbind(res.deviance = deviance, df =
  p = pchisq(deviance, df.residual, lower.ta

     res.deviance  df      p
[1,]        189.4 196 0.6182
```

Poisson Regression with R

We can also test the overall effect of prog by comparing the deviance of the full model with the deviance of the model excluding prog. The two degree-of-freedom chi-square test indicates that prog, taken together, is a statistically significant predictor of num_awards.

Poisson Regression with R

```
 update m1 model dropping prog
m2 <- update(m1, . ~ . - prog)
 test model differences with chi square test
anova(m2, m1, test="Chisq")

 Analysis of Deviance Table

 Model 1: num_awards ~ math
 Model 2: num_awards ~ prog + math
   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
 1       198        204
```

Poisson Regression with R

Sometimes, we might want to present the regression results as incident rate ratios and their standard errors, together with the confidence interval. To compute the standard error for the incident rate ratios, we will use the Delta method. To this end, we make use the function deltamethod implemented in R package **msm**.

Poisson Regression with R

```
s <- deltamethod(list(~ exp(x1), ~ exp(x2),


 exponentiate old estimates dropping the p v
rexp.est <- exp(r.est[, -3])
 replace SEs with estimates for exponentiate
rexp.est[, "Robust SE"] <- s
```

Poisson Regression with R

```
rexp.est

                 Estimate Robust SE        LL
 (Intercept)     0.005263   0.00340 0.001484
 progAcademic    2.956065   0.94904 1.575551
 progVocational  1.447458   0.57959 0.660335
 math            1.072672   0.01119 1.050955
```

Poisson Regression with R

The output above indicates that the incident rate for prog = "Academic" is 2.96 times the incident rate for the reference group (prog = "General"). Likewise, the incident rate for prog = "Vocational" is 1.45 times the incident rate for the reference group holding the other variables at constant. The percent change in the incident rate of num_awards is by 7% for every unit increase in math. For additional information on the various metrics in which the results can be presented, and the interpretation of such, please see Regression Models for Categorical Dependent Variables Using Stata, Second Edition by J. Scott Long and Jeremy Freese (2006).

Poisson Regression with R

Sometimes, we might want to look at the expected marginal means. For example, what are the expected counts for each program type holding math score at its overall mean? To answer this question, we can make use of the predict function. First off, we will make a small data set to apply the predict function to it.

Poisson Regression with R

```
(s1 <- data.frame(math = mean(p$math),
  prog = factor(1:3, levels = 1:3, labels =

    math       prog
1 52.65    General
2 52.65   Academic
3 52.65 Vocational
```

Poisson Regression with R

```
predict(m1, s1, type="response", se.fit=TRUE

 $fit
      1       2       3
 0.2114 0.6249 0.3060

 $se.fit
       1        2        3
 0.07050 0.08628 0.08834

 $residual.scale
```

Poisson Regression with R

In the output above, we see that the predicted number of events for level 1 of prog is about .21, holding math at its mean. The predicted number of events for level 2 of prog is higher at .62, and the predicted number of events for level 3 of prog is about .31. The ratios of these predicted counts ($\frac{.625}{.211} = 2.96$, $\frac{.306}{.211} = 1.45$) match what we saw looking at the IRR.

Poisson Regression with R

We can also graph the predicted number of events with the commands below. The graph indicates that the most awards are predicted for those in the academic program (prog $= 2$), especially if the student has a high math score. The lowest number of predicted awards is for those students in the general program (prog $= 1$). The graph overlays the lines of expected values onto the actual points, although a small amount of random noise was added vertically to lessen overplotting.

Poisson Regression with R

calculate and store predicted values

```
p$phat <- predict(m1, type="response")

 order by program and then by math
p <- p[with(p, order(prog, math)), ]
```

Poisson Regression with R

create the plot

```
ggplot(p, aes(x = math, y = phat, colour = p
  geom_point(aes(y = num_awards), alpha=.5,
  geom_line(size = 1) +
  labs(x = "Math Score", y = "Expected numbe
```

Poisson Regression with R

Things to consider

- When there seems to be an issue of dispersion, we should first check if our model is appropriately specified, such as omitted variables and functional forms. For example, if we omitted the predictor variable prog in the example above, our model would seem to have a problem with over-dispersion. In other words, a misspecified model could present a symptom like an over-dispersion problem.

Poisson Regression with R

- Assuming that the model is correctly specified, the assumption that the conditional variance is equal to the conditional mean should be checked. There are several tests including the likelihood ratio test of over-dispersion parameter alpha by running the same model using negative binomial distribution. R package pscl (Political Science Computational Laboratory, Stanford University) provides many functions for binomial and count data including odTest for testing over-dispersion.

Poisson Regression with R

- One common cause of over-dispersion is excess zeros, which in turn are generated by an additional data generating process. In this situation, zero-inflated model should be considered.
- If the data generating process does not allow for any 0s (such as the number of days spent in the hospital), then a zero-truncated model may be more appropriate.

Poisson Regression with R

- Count data often have an exposure variable, which indicates the number of times the event could have happened. This variable should be incorporated into a Poisson model with the use of the offset option.
- The outcome variable in a Poisson regression cannot have negative numbers, and the exposure cannot have 0s.

Poisson Regression with R

- Many different measures of pseudo-R-squared exist. They all attempt to provide information similar to that provided by R-squared in OLS regression, even though none of them can be interpreted exactly as R-squared in OLS regression is interpreted.
- Poisson regression is estimated via maximum likelihood estimation. It usually requires a large sample size.