WWS 509 Generalized Linear Models: Precept 6
Introduction to Poisson Models

Kristin E. Bietsch

Office of Population Research, Princeton University

November 2012

1 Introducing the Data

This data is available from Germán's website. It is looking at lung cancer deaths in the population. Variables include these deaths, the population in each group, age groups (40-44 to 80 plus), and smoking status (don't smoke, cigarettes only, pipe/cigar only, pipe/cigar and cigarettes). I use non-smokers age 40-44 as the reference group for the additive model. I have run the necessary models for you, they can be found in Appendix A.

2 A Discussion of Poisson

2.1 Some basics of Poisson data

1. With a Poisson model, we can model **count** data.
2. An important feature of Poisson data is that the mean is equal to the **variance**.
   (a) Because of this, the normal assumption of **homoscedasticity** is not appropriate for Poisson data.
3. Another useful property is that the **sum** of independent Poisson random variables is also Poisson.
   (a) A consequence of this is that we can analyze **individual** and **group** data with equivalent results.
      i. They have the same likelihood function.

2.2 A few more things about Poisson

1. With count data, what kind of numbers will we never have? **Negative**
   (a) Thus, how should we deal withl this? Hint: think back to logits **Take the log of the outcome**
   (b) How would we write this? $log(\mu_i) = x_i'\beta$
   (c) Ok, not if we just wanted the left side in terms of μ? $\mu_i = e^{x_i'\beta}$
2. In the final model, an exponentiated regression coefficient e^{β_j} represents a **multiplicative** effect of the j^{th} predictor on the mean
   (a) Increasing x_j by 1 unit multiplies the mean by a factor of e^{β_j}
   (b) An advantage of the log-link model: with count data, the effects of predictors are often multiplicative rather than **additive**
      i. **Small** effects for small models
      ii. **Large** effects for large models
3. Finally, we can use deviance to test the goodness of fit for the models. We can also use Pearson's chi-squared statistic. And advantage of using deviance is that we can compare **nested** models.

3 Deviance Tables

Table 1: Goodness of Fit of Models

Model	Goodness of Fit	Degrees of Freedom
Null	4055.982	35
Smoking	3910.702	32
Age	191.7219	27
Smoking+Age	21.48617	24

Table 2: Deviance of Models

Model	Deviance	Degrees of Freedom
Smoking (compared to null)	145.82	3
Age (compared to null)	3864.2601	8
S+A (compared to S)	3889.21583	8
S+A (compared to A)	170.23573	3

4 Interpretation

In the null model:

1. What are we modeling? **Deaths divided by the population**

2. What is the offset? **The log of the population in each group**
3. What does the constant represent?**The mean**
4. What am I doing when I write the line of code "quietly sum dr [fw=pop]"? **We are seeing that the constant gives us the mean of deaths divided by population**

Refer to the additive model:

1. What does the constant represent? **The reference group- non-smokers age 40-44**
   (a) What kind of transformation do you need to do to make this interpretable? **Exponentiate the coefficient**
   (b) Do that and interpret. **2.5222356e-7, 2.5 deaths per 10 million people**
2. How much higher is the probability of dying for a 76 year old compared to a 42 year old of the same smoking status? **12.9!!!**
3. How much higher is the probability of dying for a cigar and cigarette smoker compared to a non-smoker of the same age? **24% higher**
4. Compare the probability of dying for a 42 year old non-smoker to a 75 year old pipe only smoker. **1.048968*12.91739=13.5499**