Chapter 4 Part 1

04/28/2021

Estimation and Inference Poisson Regression

- We first consider a model for which $\log(\lambda)$ is linear in age.
- We then will determine whether a model with a quadratic term in age provides a significant improvement based on trends we observed in the exploratory data analysis.
- Lets fit the model in R!

The Fitted Model

- R reports an estimated regression equation for the linear Poisson model as:

$$log(\hat{\lambda}) = 1.55 - 0.0047\text{age}$$

Interpreting Coefficients

- How can the coefficient estimates be interpreted in terms of this example?
- As done when interpreting slopes in the LLSR models, we consider how the estimated mean number in the house, λ, changes as the age of the household head increases by an additional year.
- But in place of looking at change in the mean number in the house, with a Poisson regression we consider the log of the mean number in the house and then convert back to original units.

Interpreting Coefficients

- Consider a comparison of two models—one for a given age (x) and one after increasing age by 1 ($x+1$):

$$\begin{aligned} log(\lambda_X) &= \beta_0 + \beta_1 X \\ log(\lambda_{X+1}) &= \beta_0 + \beta_1 (X+1) \\ log(\lambda_{X+1}) - log(\lambda_X) &= \beta_1 \\ log\left(\frac{\lambda_{X+1}}{\lambda_X}\right) &= \beta_1 \\ \frac{\lambda_{X+1}}{\lambda_X} &= e^{\beta_1} \end{aligned} \tag{1}$$

- We will not calculate this in R

Interpreting Coefficients

- These results suggest that by exponentiating the coefficient on age we obtain the multiplicative factor by which the mean count changes.
- In this case, the mean number in the house changes by a factor of $e^{-0.0047} = 0.995$ or decreases by 0.5% (since $1 - .995 = .005$) with each additional year older the household head is
  - OR we predict a 0.47% *increase* in mean household size for a 1-year *decrease* in age of the household head (since $1/.995 = 1.0047$).

Interpreting Coefficients

- The quantity on the left-hand side of this euqation is referred to as a **rate ratio** or **relative risk**, and it represents a percent change in the response for a unit change in X.
- In fact, for regression models in general, whenever a variable (response or explanatory) is logged, we make interpretations about multiplicative effects on that variable, while with unlogged variables we can reach our usual interpretations about additive effects.

Understanding Model Output

- The standard errors for the estimated coefficients are included in Poisson regression output.
- Lets scroll back up to that in R
- Here the standard error for the estimated coefficient for age is 0.00094. We can use the standard error to construct a confidence interval for β_1.
  - A 95% CI provides a range of plausible values for the age coefficient and can be constructed:

$$(\hat{\beta}_1 - Z^* \cdot SE(\hat{\beta}_1), \quad \hat{\beta}_1 + Z^* \cdot SE(\hat{\beta}_1))$$

$$(-0.0047 - 1.96 * 0.00094, \quad -0.0047 + 1.96 * 0.00094)$$

$$(-0.0065, -0.0029).$$

Interval in Context

- Exponentiating the endpoints yields a confidence interval for the relative risk; i.e., the percent change in household size for each additional year older.
- Thus $(e^{-0.0065}, e^{-0.0029}) = (0.993, 0.997)$ suggests that we are 95% confident that the mean number in the house decreases between 0.7% and 0.3% for each additional year older the head of household is.
- It is best to construct a confidence interval for the coefficient and then exponentiate the endpoints because the estimated coefficients more closely follow a normal distribution than the exponentiated coefficients.

Interval Methods

- There are other approaches to constructing intervals in these circumstances, including profile likelihood, the delta method, and bootstrapping, and we will discuss some of those approaches later.
- In this case, for instance, the profile likelihood interval is nearly identical to the Wald-type (normal theory) confidence interval above.
- Lets get this profile likelihood, it is in fact, the default in R confint
- Note: confint is a function in base R and mosaic. The mosaic version has a couple of method options one of which are the wald version we saw manually above

More on Testig the Model

- If there is no association between age and household size, there is no change in household size for each additional year, so λ_X is equal to λ_{X+1} and the ratio λ_{X+1}/λ_X is 1.
- In other words, if there is no association between age and household size, then $\beta_1 = 0$ and $e^{\beta_1} = 1$.
- Note that our interval for e^{β_1}, (0.993,0.997), does not include 1, so the model with age is preferred to a model without age; i.e., age is significantly associated with household size.
- Note that we could have similarly confirmed that our interval for β_1 does not include 0 to show the significance of age as a predictor of household size.

Testing using Tests Statistics

- Another way to test the significance of the age term is to calculate a **Wald-type statistic**.
- A Wald-type test statistic is the estimated coefficient divided by its standard error.
  - This should seem familiar
- When the true coefficient is 0, this test statistic follows a standard normal distribution for sufficiently large n.

Testing using Tests Statistics

- The estimated coefficient associated with the linear term in age is $\hat{\beta}_1 = -0.0047$ with standard error $SE(\hat{\beta}_1) = 0.00094$.
- The value for the Wald test statistic is then $Z = \hat{\beta}_1 / SE(\hat{\beta}_1) = -5.026$, where Z follows a standard normal distribution if $\beta_1 = 0$.
- In this case, the two-sided p-value based on the standard normal distribution for testing $H_0 : \beta_1 = 0$ is almost 0 ($p = 0.000000501$).
- Therefore, we have statistically significant evidence (Z = -5.026, p $<$ .001) that average household size decreases as age of the head of household increases.