Chapter 4 Part 4

04/29/2021

Adding a Covariate

- Add more predictors - it could be better
- By controlling for important , we can obtain more precise estimates of the relationship between age and household size.
- We may discover that the relationship between age and household size may differ by levels of a covariate.
- One important covariate to consider is location. As described earlier in the case study, there are 5 different regions that are associated with the `Location` variable: Central Luzon, Metro Manila, Visayas, Davao Region, and Ilocos Region. Assessing the utility of including the covariate `Location` is, in essence, comparing two nested models; here the quadratic model is compared to the quadratic model plus terms for `Location`.
- Lets do this in R

New Model

$$\log(\text{total}) = -0.384 + 0.070 \cdot \text{age} - 0.00070 \cdot \text{age}^2 + 0.061 \cdot \text{IlocosRegion} + 0.054 \cdot \text{MetroManila} + 0.112 \cdot \text{Visayas} - 0.019 \cdot \text{DavaoRegion}$$

- Notice that because there are 5 different locations, we must represent the effects of different locations through 4 indicator variables.
- For example, $\hat{\beta}_6 = -0.0194$ indicates that, after controlling for the age of the head of household, the log mean household size is 0.0194 lower for households in the Davao Region than for households in the reference location of Central Luzon.
- In more interpretable terms, mean household size is $e^{-0.0194} = 0.98$ times "higher" (i.e., 2% lower) in the Davao Region than in Central Luzon, when holding age constant.
- How does R choose the reference level and how can we change it?

Testing if Difference in Location is Significant

- What method could we use to test "is location significant?"
  - What is difficult about testing this with our model?

Testing if Difference in Location is Significant

- To test if the mean household size significantly differs by location, we must use a drop-in-deviance test, rather than a Wald-type test, because four terms (instead of just one) are added when including the `location` variable.
- Lets do this in R.

Testing if Difference in Location is Significant

- Adding the four terms corresponding to location to the quadratic model with age produces a statistically significant improvement ($\chi^2 = 13.144, df = 4, p = 0.0106$), so there is significant evidence that mean household size differs by location, after controlling for age of the head of household.

Roof Type

- Lets add this and testing if it is significant in R
- Does the order which we are adding variables matter?
- Further modeling (not shown) shows that after controlling for location and age of the head of household, mean household size did not differ between the two types of roofing material.

Goodness-of-Fit

- We are skipping the section on poisson residuals as they are not as useful as normal regression residuals
- The residuals used are called deviance residuals (we saw this output in R)
- The model residual deviance can be used to assess the degree to which the predicted values differ from the observed.
- When a model is true, we can expect the residual deviance to be distributed as a χ^2 random variable with degrees of freedom equal to the model's residual degrees of freedom.
- Our model thus far, the quadratic terms for age plus the indicators for location, has a residual deviance of 2187.8 with 1493 df.
- The probability of observing a deviance this large if the model fits is essentially 0, saying that there is significant evidence of lack-of-fit.
- See this i R

Lack of Fit

- There are several reasons why **lack-of-fit** may be observed.
- We may be missing important covariates or interactions; a more comprehensive data set may be needed.
- There may be extreme observations that may cause the deviance to be larger than expected; however, our residual plots did not reveal any unusual points.
- Lastly, there may be a problem with the Poisson model.
  - In particular, the Poisson model has only a single parameter, λ, for each combination of the levels of the predictors which must describe both the mean and the variance. This limitation can become manifest when the variance appears to be larger than the corresponding means. In that case, the response is more variable than the Poisson model would imply, and the response is considered to be **overdispersed**. More on this later.

Linear Least Squares vs. Poisson Regression

Response

LLSR : Normal

PoissonRegression : Counts

Variance

LLSR : Equal for each level of X

PoissonRegression : Equal to the mean for each level of X

Model Fitting

LLSR : $\mu = \beta_0 + \beta_1 x$ using Least Squares

PoissonRegression : $log(\lambda) = \beta_0 + \beta_1 x$ using Maximum Likelihood

Linear Least Squares vs. Poisson Regression

EDA

LLSR : Plot X vs. Y; add line

PoissonRegression : Find $log(\bar{y})$ for several subgroups; plot vs. X

Comparing Models

LLSR : Extra sum of squares F-tests; AIC/BIC

PoissonRegression : Drop in Deviance tests; AIC/BIC

Interpreting Coefficients

LLSR : $\beta_1 =$ change in μ_y for unit change in X

PoissonRegression : $e^{\beta_1} =$ percent change in λ for unit change in X