Chapter 4 Part 5 Case study

04/30/2021

Case Study: Campus Crime

- Students want to feel safe and secure when attending a college or university.
- In response to legislation, the US Department of Education seeks to provide data and reassurances to students and parents alike.
- All postsecondary institutions that participate in federal student aid programs are required by the Jeanne Clery Disclosure of Campus Security Policy and Campus Crime Statistics Act and the Higher Education Opportunity Act to collect and report data on crime occurring on campus to the Department of Education.
- In turn, this data is publicly available on the website of the Office of Postsecondary Education.
- We are interested in looking at whether there are regional differences in violent crime on campus, controlling for differences in the type of school.

Data Organization

- Each row of `c_data.csv` contains crime information from a post secondary institution, either a college or university. The variables include:
- `Enrollment` = enrollment at the school
- `type` = college (C) or university (U)
- `nv` = the number of violent crimes for that institution for the given year
- `nvrate` = number of violent crimes per 1000 students
- `enroll1000` = enrollment at the school, in thousands
- `region` = region of the country (C = Central, MW = Midwest, NE = Northeast, SE = Southeast, SW = Southwest, and W = West)
- Lets read it into R

Exploratory Data Analysis

- We want to look at are variables and understand how they relate to one another
- The book presents, usually, most of the plots/tables we will want to look at.
  - In many cases, we may need to look at more
- Lets do this away from the book using some basic R functions, alongside ggplot, but with us motivating the need for each table or plot
  - What plots/tables will be helpful in understanding out data in this example?

Exploratory Data Analysis

- A graph of the number of violent crimes, reveals the pattern often found with distributions of counts of rare events.
  - Many schools reported no violent crimes or very few crimes. A few schools have a large number of crimes making for a distribution that appears to be far from normal.
- Therefore, Poisson regression should be used to model our data; Poisson random variables are often used to represent counts (e.g., number of violent crimes) per unit of time or space (e.g., one year).

Exploratory Data Analysis

- Let's take a look at two covariates of interest for these schools: type of institution and region. In our data, the majority of institutions are universities (65% of the 81 schools) and only 35% are colleges.
- Interest centers on whether the different regions tend to have different crime rates.
- A relative frequency able of school type by region contains the name of each region and each column represents the percentage of schools in that region which are colleges or universities.
    - The proportion of colleges varies from a low of 20% in the Southwest (SW) to a high of 50% in the West (W).

Exploratory Data Analysis

- While a Poisson regression model is a good first choice because the responses are counts per year, it is important to note that the counts are not directly comparable because they come from different size schools.
- This issue sometimes is referred to as the need to account for *sampling effort*; in other words, we expect schools with more students to have more reports of violent crime since there are more students who could be affected.
- We cannot directly compare the 30 violent crimes from the first school in the data set to no violent crimes for the second school when their enrollments are vastly different: 5,590 for school 1 versus 540 for school 2.
- We can take the differences in enrollments into account by including an **offset** in our model, which we will discuss in the next section.
- What is an easy solution to this?

Exploratory Data Analysis

- For the remainder of the EDA, we examine the violent crime counts in terms of the rate per 1,000 enrolled $(\frac{\text{number of violent crimes}}{\text{number enrolled}} \cdot 1000)$.

Exploratory Data Analysis

- Note that there is a noticeable outlier for a Southeastern school (5.4 violent crimes per 1000 students), and there is an observed rate of 0 for the Southwestern colleges which can lead to some computational issues.
- We therefore can combined the SW and SE to form a single category of the South, and we also removed the extreme observation from the data set.

Exploratory Data Analysis

- Tabling and using colored boxplots using school type, region, and crime rate shows that crimes are generally lower at the colleges within a region (with the exception of the Northeast).
- In addition, the regional pattern of rates at universities appears to differ from that of the colleges.

Accounting for Enrollment

- Although working with the observed rates (per 1000 students) is useful during the exploratory data analysis, we do not use these rates explicitly in the model.
- The counts (per year) are the Poisson responses when modeling, so we must take into account the enrollment in a different way.
- An approach is to include a term on the right side of the model called an **offset**, which is the log of the enrollment, in thousands.

Accounting for Enrollment

- There is an intuitive heuristic for the form of the offset.
- If we think of λ as the mean number of violent crimes per year, then $\lambda/\mathrm{enroll1000}$ represents the number per 1000 students, so that the yearly count is adjusted to be comparable across schools of different sizes.
- Adjusting the yearly count by enrollment is equivalent to adding $log(\mathrm{enroll1000})$ to the right-hand side of the Poisson regression equation—essentially adding a predictor with a fixed coefficient of 1:

Accounting for Enrollment Math

$$log(\frac{\lambda}{\text{enroll1000}}) = \beta_0 + \beta_1(\text{type})$$
$$log(\lambda) - log(\text{enroll1000}) = \beta_0 + \beta_1(\text{type})$$
$$log(\lambda) = \beta_0 + \beta_1(\text{type}) + log(\text{enroll1000})$$

Not Modelign $\frac{\lambda}{\text{enroll1000}}$

- While this heuristic is helpful, it is important to note that it is *not* $\frac{\lambda}{\text{enroll1000}}$ that we are modeling.
- We are still modeling $log(\lambda)$, but we're adding an offset to adjust for differing enrollments, where the offset has the unusual feature that the coefficient is fixed at 1.0.
- As a result, no estimated coefficient for `enroll1000` or $log(\text{enroll1000})$ will appear in the output.
- As this heuristic illustrates, modeling $log(\lambda)$ and adding an offset is equivalent to modeling rates, and coefficients can be interpreted that way.

Modeling Assumptions

- In the table by both region and school type, we see that the variances are greatly higher than the mean counts in almost every group.
- Thus, we have reason to question the Poisson regression assumption of variability equal to the mean; we will have to return to this issue after some initial modeling.
- The fact that the variance of the rate of violent crimes per 1000 students tends to be on the same scale as the mean tells us that adjusting for enrollment may provide some help, although that may not completely solve our issues with excessive variance.

Modeling Assumptions

- As far as other model assumptions, linearity with respect to $log(\lambda)$ is difficult to discern without continuous predictors, and it is not possible to assess independence without knowing how the schools were selected.
- What are some continous predictors that may have been useful?

Initial Models

- We are interested primarily in differences in violent crime between institutional types controlling for difference in regions, so we fit a model with region, institutional type, and our offset.
- Note that the central region is the reference level in our model.
- Again, R choosese this since Central is alphabeticaly 1st
- Let us fit this model in R.

Some Interpretations

- From our model, the Northeast and the South differ significantly from the Central region (p= 0.00000037 and p=0.0000924, respectively).
- The estimated coefficient of 0.778 means that the violent crime rate per 1,000 in the Northeast is nearly 2.2 ($e^{0.778}$) times that of the Central region controlling for the type of school.
- A Wald-type confidence interval for this factor can be constructed by first calculating a CI for the coefficient ($0.778 \pm 1.96 \cdot 0.153$) and then exponentiating (1.61 to 2.94).

Tukey's Honestly Significant Differences

- Comparisons to regions other than the Central region can be accomplished by changing the reference region.
- If many comparisons are made, it would be best to adjust for multiple comparisons using a method such as **Tukey's Honestly Significant Differences**, which considers all pairwise comparisons among regions.
- This method helps control the large number of false positives that we would see if we ran multiple t-tests comparing groups. The honestly significant difference compares a standardized mean difference between two groups to a critical value from a studentized range distribution.
- Lets try this in R

Understanding Tukey's Honestly Signifcant Differences Output

- In our case, Tukey's Honestly Significant Differences simultaneously evaluates all 10 mean differences between pairs of regions.
- We find that the Northeast has significantly higher rates of violent crimes than the Central, Midwest, and Western regions, while the South has significantly higher rates of violent crimes than the Central and the Midwest, controlling for the type of institution.
- In the primary model, the University indicator is significant and, after exponentiating the coefficient, can be interpreted as an approximately ($e^{0.280}$) 32% increase in violent crime rate over colleges after controlling for region.
- What if we also had a continuous predictor?
- What else should we check?

Understanding Tukey's Honestly Signifcant Differences Output

- These results certainly suggest significant differences in regions and type of institution. However, the EDA findings suggest the effect of the type of institution may vary depending upon the region, so we consider a model with an interaction between region and type.
- Lets add interaction terms and model in R.

Looking at Our New Model

- These results provide convincing evidence of an interaction between the effect of region and the type of institution.
- A drop-in-deviance test like the one we carried out in the previous case study confirms the significance of the contribution of the interaction to this model.
- We have statistically significant evidence ($\chi^2 = 71.98, df = 4, p < .001$) that the difference between colleges and universities in violent crime rate differs by region.
- For example, our model estimates that violent crime rates are 13.6 ($e^{.196+2.411}$) times higher in universities in the West compared to colleges, while in the Northeast we estimate that violent crime rates are 2.4 ($\frac{1}{e^{.196-1.070}}$) times higher in colleges.
- In the absence of other covariates or extreme observations, we consider overdispersion as a possible explanation of the significant lack-of-fit.