Chapter 4 Part 8 Case study

05/04/2021

Case Study: Weekend Drinking

- Sometimes when analyzing Poisson data, you may see many more zeros in your data set than you would expect for a Poisson random variable.
- For example, an informal survey of students in an introductory statistics course included the question, “How many alcoholic drinks did you consume last weekend?”.
- This survey was conducted on a dry campus where no alcohol is officially allowed, even among students of drinking age, so we expect that some portion of the respondents never drink. The non-drinkers would thus always report zero drinks.

Case Study: Weekend Drinking

- There will also be students who are drinkers reporting zero drinks because they just did not happen to drink during the past weekend.
- Our zeros, then, are a **mixture** of responses from non-drinkers and drinkers who abstained during the past weekend. Ideally, we'd like to sort out the non-drinkers and drinkers when performing our analysis.
- Let us read this data in and start EDA (some)

Research Question

- The purpose of this survey is to explore factors related to drinking behavior on a dry campus.
- What proportion of students on this dry campus never drink? What factors, such as off-campus living and sex, are related to whether students drink?
- Among those who do drink, to what extent is moving off campus associated with the number of drinks in a weekend?
- It is commonly assumed that males' alcohol consumption is greater than females'; is this true on this campus?
- Answering these questions would be a simple matter if we knew who was and was not a drinker in our sample. Unfortunately, the non-drinkers did not identify themselves as such, so we will need to use the data available with a model that allows us to estimate the proportion of drinkers and non-drinkers.

Data Organization

- Each line of `weekendDrinks.csv` contains data provided by a student in an introductory statistics course.
- In this analysis, the response of interest is the respondent's report of the number of alcoholic `drinks` they consumed the previous weekend, whether the student lives `off.campus`, and `sex`.
    - We will also consider whether a student is likely a `firstYear` student based on the `dorm` they live in.
- Let us look more at the data and see a new way to get the *grand mean*, also update our variables

Exploratory Data Analysis

- Now that are variables are set, time to continue the EDA
- There are 77 observations.
- Large sample sizes are preferred for the type of model we will consider, and n=77 is on the small side. We proceed with that in mind.

Drinker vs Non-drinkers

- A premise of this analysis is that we believe that those responding zero drinks are coming from a mixture of non-drinkers and drinkers who abstained the weekend of the survey.
- **Non-drinkers**: respondents who never drink and would always reply with zero.
- **Drinkers**: obviously this includes those responding with one or more drinks, but it also includes people who are drinkers but did not happen to imbibe the past weekend. These people reply zero but are not considered non-drinkers.

EDA

- Beginning the EDA with the response, number of drinks, we find that over 46% of the students reported no drinks during the past weekend. F
- The mean number of drinks reported the past weekend is 2.013.
- Our sample consists of 74% females and 26% males, only 9% of whom live off campus.

EDA

- Why is it natural to expect to us poisson regression?
- You may recall that a Poisson distribution has only one parameter, λ, for its mean and variance. Here we will include an additional parameter, α. We define α to be the true proportion of *non-drinkers* in the population.

Motivating the Zero Inflated Poisson

- In a we have the histogram of our observed response, note the large bar at 0
- In b we have the poisson distribution based on our null hypothesis, that the mean is about 2.017 and that the response is truly poisson
- This circumstance actually arises in many Poisson regression settings.
- We will define λ to be the mean number of drinks *among those who drink*, and α to be the proportion of *non-drinkers* ("true zeros"). - Then, we will attempt to model λ and α (or functions of λ and α) simultaneously using covariates like sex, first-year status, and off-campus residence.
- This type of model is referred to as a **zero-inflated Poisson model** or **ZIP model**.

Modeling

- We first fit a simple Poisson model with the covariates `off.campus` and `sex`.
- Let us do this in R, iftting our regular poisson

Summary of 1st Model Results

- Both covariates are statistically significant, but a goodness-of-fit test reveals that there remains significant lack-of-fit (residual deviance: 230.54 with only 74 df; p$<$.001 based on χ^2 test with 74 df).
- In the absence of important missing covariates or extreme observations, this lack-of-fit may be explained by the presence of a group of non-drinkers.

Zero Inflated Poisson

- A zero-inflated Poisson regression model to take non-drinkers into account consists of two parts:
    - One part models the association, among drinkers, between number of drinks and the predictors of sex and off-campus residence.
    - The other part uses a predictor for first-year status to obtain an estimate of the proportion of non-drinkers based on the reported zeros.

Zero Inflated Poisson

- The form for each part of the model follows. The first part looks like an ordinary Poisson regression model:

$$\mathit{log}(\lambda) = \beta_0 + \beta_1 \text{off.campus} + \beta_2 \text{sex}$$

where λ is the mean number of drinks in a weekend *among those who drink*. The second part has the form

$$\mathit{logit}(\alpha) = \beta_0 + \beta_1 \text{firstYear}$$

where α is the probability of being in the non-drinkers group and $\mathit{logit}(\alpha) = \mathit{log}(\alpha/(1-\alpha))$.
- We'll provide more detail on the logit in the linear regression chapter. - There are many ways in which to structure this model; here we use different predictors in the two pieces, athough it would have been perfectly fine to use the same predictors for both pieces, or even no predictors for one of the pieces.

Does ZIP Work?

- How is it possible to fit such a model?
- We cannot observe whether a respondent is a drinker or not (which probably would've been good to ask).
- The ZIP model is a special case of a more general type of statistical model referred to as a **latent variable model**.
- More specifically, it is a type of a **mixture model** where observations for one or more groups occur together and the group membership is unknown.
- Zero-inflated models are a particularly common example of a mixture model, but the response does not need to follow a Poisson distribution.

Does ZIP Work?

- Likelihood methods are at the core of this methodology, but fitting is an iterative process where it is necessary to start out with some guesses (or starting values).
- In general, it is important to know that models like ZIP exist, although we'll only explore interpretations and fitting options for a single case study here.

General Idea

- Imagine that the graph of the Poisson distribution in Figure (b) is removed from the observed data distribution in Figure (a).
- Some zero responses will remain.
- These would correspond to non-drinkers, and the proportion of all observations these zeros constitute might make a reasonable estimate for α, the proportion of non-drinkers.

General Idea

- The likelihood is used and some iterating in the fitting process is involved because the Poisson distribution in Figure (b) is based on the mean of the observed data, which means it is the average among all students, not only among drinkers.
- The likelihood incorporates the predictors, `sex` and `off.campus`. So there is a little more to it than computing the proportion of zeros, but this heuristic should provide you a general idea of how these kinds of models are fit.
- We will use the R function `zeroinfl` from the package `pscl` to fit a ZIP model.

Model Discussion

- Our model uses firstYear to distinguish drinkers and non-drinkers ("Zero-inflation model coefficients") and off.campus and sex to help explain the differences in the number of drinks among drinkers ("Count model coefficients").
- We could have used the same covariates for the two pieces of a ZIP model, but neither off.campus nor sex proved to be a useful predictor of drinkers vs. non-drinkers after we accounted for first-year status.

Model Discussion

- The “Count model coefficients,” provide information on how the sex and off-campus status of a student who is a drinker are related to the number of drinks reported by that student over a weekend.
- As we have done with previous Poisson regression models, we exponentiate each coefficient for ease of interpretation.

Model Discussion

- For those who drink, the average number of drinks for males is $e^{1.0209}$ or 2.76 times the number for females (Z $= 5.827$, p < 0.001) given that you are comparing people who live in comparable settings, i.e., either both on or both off campus.

Model Discussion

- Among drinkers, the mean number of drinks for students living off campus is $e^{0.4159} = 1.52$ times that of students living on campus for those of the same sex ($Z = 2.021$, $p = 0.0433$).

Model Discussion – Zero Part

- The “Zero-inflation model coefficients” refer to separating drinkers from non-drinkers.
- An important consideration in separating drinkers from non-drinkers may be whether this is their first year, where `firstYear` is a $0/1$ indicator variable.

We have

$$\mathit{log}(\alpha/(1-\alpha)) = -0.6036 + 1.1364 \mathrm{firstYear}$$

Model Discussion – Zero Part

- We are interested in α, the proportion of non-drinkers. Exponentiating the coefficient for the first-year term for this model yields 3.12.
- Here it is interpreted as the odds ($\frac{\alpha}{1-\alpha}$) that a first-year student is a non-drinker is 3.12 times the odds that an upper-class student is a non-drinker.

Model Discussion – Zero Part

- A little algebra (solving the equation with $log(\alpha/(1-\alpha))$ for α), we have

$$\hat{\alpha} = \frac{e^{-0.6036+1.1364(\text{firstYear})}}{1 + e^{-0.6036+1.1364(\text{firstYear})}}.$$

- The estimated chance that a first-year student is a non-drinker is

$$\frac{e^{0.533}}{1 + e^{0.533}} = 0.630$$

or 63.0%, while for non-first-year students, the estimated probability of being a non-drinker is 0.354.
- If you have seen logistic regression, you'll recognize that this transformation is what is used to estimate a probability.

The Vuong Test

- Moving from ordinary Poisson to zero-inflated Poisson has helped us address additional research questions: What proportion of students are non-drinkers, and what factors are associated with whether or not a student is a non-drinker?
- While a ZIP model seems more faithful to the nature and structure of this data, can we quantitatively show that a zero-inflated Poisson is better than an ordinary Poisson model?

The Vuong Test

- We cannot use the drop-in-deviance test we discussed earlier because these two models are not nested within one another.
- Vuong [-@Vuong1989] devised a test to make this comparison for the special case of comparing a zero-inflated model and ordinary regression model.
- Essentially, the Vuong Test is able to compare predicted probabilities of **non-nested** models.
- We use vuong function in R

The Vuong Test Results

- We have structured the Vuong Test to compare Model 1: Ordinary Poisson Model to Model 2: Zero-inflation Model.
- If the two models do not differ, the test statistic for Vuong would be asymptotically standard Normal and the p-value would be relatively large.
- Here the first line of the output table indicates that the zero-inflation model is better ($Z = -2.69, p = .0036$).
- Note that the test depends upon sufficiently large n for the Normal approximation, so since our sample size (n=77) is somewhat small, we need to interpret this result with caution.
- More research is underway to address statistical issues related to these comparisons.

Residual Plot in ZIP

- Fitted values $(\hat{y})$ and residuals $(y - \hat{y})$ can be computed for zero-inflation models and plotted.
- Let us plot in R

Residual Plot in ZIP

- We can see that one observation appears to be extreme (Y=22 drinks during the past weekend).
- Is this a legitimate observation or was there a transcribing error?
- Without the original respondents, we cannot settle this question. It might be worthwhile to get a sense of how influential this extreme observation is by removing Y=22 and refitting the model.

Limitations

- Given that you have progressed this far in your statistical education, the weekend drinking survey question should raise some red flags.
- What are you thoughts?

Limitations

- What time period constitutes the “weekend”?
- Will some students be thinking of only Saturday night, while others include Friday night or possibly Sunday evening?
- What constitutes a drink—a bottle of beer?
- How many drinks will a respondent report for a bottle of wine?
- Precise definitions would vastly improve the quality of this data.
- There is also an issue related to confidentiality.
- If the data is collected in class, will the teacher be able to identify the respondent?
- Will respondents worry that a particular response will affect their grade in the class or lead to repercussions on a dry campus?

Limitations

- In addition to these concerns, there are a number of other limitations that should be noted.
  - Following the concern of whether this data represents a random sample of any population (it doesn't), we also must be concerned with the size of this data set.
  - ZIP models are not appropriate for small samples and this data set is not impressively large.