STA2201H Winter 2022 Assignment 1

Due: 11:59pm, 6 February 2022

What to hand in: .Rmd file and the compiled pdf

How to hand in: Submit files via Quercus

1 Overdispersion

Suppose that the conditional distribution of outcome Y given an unobserved variable θ is Poisson, with a mean and variance θ, so

$$Y|\theta \sim \text{Poisson}(\mu\theta)$$

a) Assume $E(\theta) = 1$ and $Var(\theta) = \sigma^2$. Using the laws of total expectation and total variance, show $E(Y) = \mu$ and $Var(Y) = \mu(1 + \mu\sigma^2)$.
b) Assume θ is Gamma distributed with α and β as shape and scale parameters, respectively. Show the unconditional distribution of Y is Negative Binomial.
c) In order for $E(Y) = \mu$ and $Var(Y) = \mu(1 + \mu\sigma^2)$, what must α and β equal?

2 Child support payments

Consider the situation where we are interested in the effect of divorced fathers' income on their child support payments. In this question we will be simulating the 'true' relationship between these two factors in a population, simulating survey data, then calculating the estimated relationship based on that surveyed data.

a) Simulate data on income and child support payments for a population of 1000 fathers. Simulate both variables on the log scale, and assume

   - Log of Income is normally distributed with a mean of log(10000) and a standard deviation of log(100)
   - Log of Payments are normally distributed with a mean of log(3500) and a standard deviation of log(30)

   Plot a histogram of both variables.

b) Create a scatter plot of log payments versus log income and add a line of best fit. Briefly describe what you observe.

c) Simulate a set of fathers who are surveyed from total population in the following way:

   - Transform log income and log payments to z-scores
   - Create a new variable called `survey` that is a logical variable equal to `TRUE` if the sum of the two z-scores, plus some random noise (i.e. plus a draw from a standard normal distribution) is greater than 0.

   Explain what the calculation for `survey` is doing, and what real-life sampling situation it is emulating. Summarize the mean payments and income for surveyed and non-surveyed fathers, and briefly comment.

d) Illustrate with the regressions and a plot the estimated effect of income on payments for the surveyed fathers. How does this differ to the same relationship estimated from the total population?

e) Discuss briefly what you observe and broader implications for drawing inferences from survey data.

Note: this example was inspired by a classic paper in sociology:

Lin, I. F., & Seltzer, J. A. (1999). Causes and effects of nonparticipation in a child support survey. Journal of Official Statistics, 15(2), 143

3 Hurricanes

In 2014 the following paper was published in PNAS:

> Jung, K., Shavitt, S., Viswanathan, M., & Hilbe, J. M. (2014). Female hurricanes are deadlier than male hurricanes. Proceedings of the National Academy of Sciences, 111(24), 8782-8787.

As the title suggests, the paper claimed that hurricanes with female names have caused a greater loss of life. In this question you will be investigating the data set used for the regression part of their analysis.

You can download the from the paper's supporting information here: https://www.pnas.org/content/111/24/8782/tab-figures-data

You should skim the whole paper but you will probably find it useful to read the sections on the Archival Study in the most depth (both in the main text and 'Materials and Methods' section).

a) Create three graphs in ggplot that help to visualize patterns in deaths by femininity, minimum pressure, and damage. Discuss what you observe based on your visualizations.
b) Run a Poisson regression with `deaths` as the outcome and `femininity` as the explanatory variable. Interpret the resulting coefficient estimate. Check for overdispersion. If it is an issue, run a quasi-Poisson regression with the same variables. Interpret your results.
c) Reproduce Model 4 (as described in the text and shown in Table S2).[1] Report the estimated effect of femininity on deaths assuming a hurricane with median pressure and damage ratings.
d) Using Model 4, predict the number of deaths caused by Hurricane Sandy. Interpret your results.
e) Describe at least two strengths and two weaknesses of this paper, focusing on the archival analysis. What was done well? What needed improvement?
f) Are you convinced by the results? If you are, explain why. If you're not, describe what additional data and/or analyses you would like to see to further test the author's hypothesis.

[1] I was able to reproduce the coefficient estimates using the data available but the standard errors were slightly different, so don't worry if that is what you find.

4 Vaccinations

This question relates to COVID-19 vaccination rates in the United States. We are interested in exploring factors that are associated with differences in vaccine coverage by US county.

- You can download the latest data on vaccination coverage here: https://data.cdc.gov/Vaccinations/COVID-19-Vaccinations-in-the-United-States-County/8xkx-amqh/data. Note that this is updated most days so depending on when you download it, it might be slightly different from others (that's okay). For the purposes of the assignment, please consider data from the 15th of January. Also note that on the same webpage you should be able to find a data dictionary. We will be interested in people who have had at least two vaccinations, which refers to columns that have the `Series_Complete` prefix.
- The class repo has a dataset `acs` that contain a range of different demographic, socioeconomic, and health variables by county. These were obtained from the American Community Survey (ACS) via the R package `tidycensus`. For reference, the extraction code can be found in the repo (`acs.R`)

a) Perform some exploratory data analysis (EDA) using a dataset combining the vaccination and ACS data, and summarize your observations with the aid of 3-4 key tables or graphs.
b) Build a regression model at the county level to help investigate patterns in the full vaccination rate for the population aged 18+ (that is, people aged 18+ who have received at least two vaccines). There is no one right answer here, but you should justify the outcome measure you are using (e.g. counts, proportions, rates, etc) and your distributional assumptions about the outcome measure (e.g. binary, poisson, normal, etc). You should also discuss briefly your model building strategy; what covariates you considered and why (motivated by your EDA)[2], and how the candidate model was chosen. Interpret your findings, including visualizations where appropriate.
c) Use your model from b) to predict the proportion of the population aged 18+ in Ada County, Idaho. Briefly discuss how good you think this prediction is, and why.
d) Give a brief summary of your analysis. What other variables may be of interest to investigate in future?
e) Now consider the situation of analysing vaccination rates at the **state** level. Consider the three following options:

    1) Regression at the state level, outcome used is the total population 18+ fully vaccinated
    2) Regression at the state level, outcome used is the average of the county level full vaccination rates of 18+ population
    3) Regression at the county level, outcome used is the total population 18+ fully vaccinated, and include as a covariate a categorical variable (fixed effect) which indicates which state a county is in.

    Without performing these regressions, briefly discuss how you think these three approaches would differ in terms of the granularity of information used and the type of outcome measure. In your opinion which is the most appropriate analysis, or does it depend on the question being asked?

[2]Note that the vaccines dataset also has a `Metro` variable which you are welcome to use in your analyses.