USCLAP Corrections

Maya Ghanem

12/21/2021

Notes from Gradescope

- In general, this is quite good. One challenge with modeling the % is that linear regression treats them as having equal variance, but the population sizes differ markedly, so the %'s should have different variances. Ideally you'd use a binomial model where n is the population and x is the # with the endpoint. Also note Bonferroni doesn't address lack of independence (to handle that you can control for state in the model). It would be good to show some exploratory data analysis of the correlations among your predictor variables, as you hypothesize about them later in the report.
- Overall this is good. Residuals show some departure from equal variances, and perhaps other approaches could be explored (e.g., logistic regression). Would be good to go into more detail about how exclusion of states with <10 cities affects results – the observations are cities, not states. Maybe you could run both ways and see if results are consistent in a sensitivity analysis?

So we need to rethink the regressions. For things that are a proportion of the population, we could have two responses (1 = has the outcome, 2 = does not have the outcome) and try to model the probability of getting something based on that. But can you make estimates of percentages into predictions for a binary variable?

This is a good answer I found after searching up regressions where the dependent variable is a percentage: https://ezinearticles.com/?Proportions-As-a-Dependent-Variable-in-Regression---Which-Type-of-Model?&id=2101689

If you have the total population, could you then create a binary variable for all the observations within that? I feel like that would be very complicated though...

There is something called the beta distribution that can be used to model continuous percentages, but it seems to be way beyond the scope of the course. I feel like if we did it we would not be able to fully understand it. Maybe the best thing to do would be to

a) focus on ANOVA or
b) focus on one city and create 0,1 variables based on the overall percentage? That could definitely be something that would work...

Maybe we could focus on one state, look at the differences between cities in that state with a chi squared test, and then predict the probability of getting something based on a binary response for all the observations within the cities of that state...

Make sure you understand: the asssumptions, what is the model used for, how to perform it, and how to interpret the results.

Here are things we can do:

1) Start with a focus on ANOVA. See how far that takes you and what significance you can draw from it.
2) For the linear regressions, create a two-limit Tobit model that sets 0 and 100 as the bounds of the model. That way you are accounting for the fact that it is not unlimited.

Here is a link that talks about the tobit model: https://stats.oarc.ucla.edu/r/dae/tobit-models/

3) Select one state, create a binary variable based on the proportions of the cities. Conduct chi squared test and then do logistic regression. Use it to potentially have a commentary about wealth? Insurance? Other socioeconomic factors?

I'm thinking that the original research report was too scattered. Maybe is best to ask: How does state residence affect your access to insurance? By just focusing on one variable, maybe that is a better thing to do...? Is there anything else that we should control for in this scenario? Maybe we should control for other socioeconomic and demographic numbers. Does CDC include this?

Potential research questions:

1) Is there any significant relationship between state residence and healthcare access? You would need to control for income level to determine if this is due to state policies, but you do not have data to do so.
2) Is there any significant relationship between state residence and binge drinking? You could use healthcare access to control for income level/socioeconomic status and then work from there. First do ANOVA to see if there is anything significant, then create a tobit limited linear regression model controlling for health care access to see what the relationship is.
3) Explain why this is significant and present research question.
4) Exploratory data: Maps for Binge Drinking.
5) ANOVA: check assumptions, conduct test, see results
6) TOBIT linear regression: check assumptions, conduct test, see results
7) Conclusions