

## Homework 7: Regression Revisited

Answer the following questions in a .pdf or .docx, explaining all of your answers and putting any tables and figures in the document as necessary. When data is called for to answer applied questions, I will provide it in bblearn. Turn in your R code that created all of the tables and figures separately, and be sure that it runs from source in such a way that it loads the data and performs all the tests without me fiddling with it. Make sure to document your R source code using `#` comments if you want partial credit.

For regression, I will be asking to you perform and interpret regressions using traditional theory-based techniques to calculate uncertainty. Please include full tables with information I request in a proper regression table in your .docx or .pdf. For today, I will have you perform regressions on a data set on US presidential elections and county level demographic characteristics, `USfederalelectionscontext-2018.csv`. I have also included a .txt file that describes the variables in that dataset.

1. First, I'd like to know if there is a strong correlation between the number of votes received by the Democratic governor candidate in a county in 2014 and the number of votes received by Hillary Clinton in 2016 and also the Republican governor candidate in a county in 2014 and the number of the votes received by Donald Trump in 2016. Please calculate and report the bivariate correlation between these quantities.
2. Now, I want to predict the percentage of total votes received by Donald Trump in various counties. To do so, we will need to convert our raw vote totals from 2016 to percentages of the total vote using the `mutate()` command (remember to multiply by 100 too).

Run a regression predicting percentage of the vote received by Trump in a county with percent black, percent hispanic, percent 65 and older, median household income, less than college education percent, and percent rural. Present that table including coefficients, standard errors, null hypothesis tests using the \*, \*\*, \*\*\* nomenclature responding to .10, .05, .01  $\alpha$ , and 99% confidence intervals. Please present your estimates with 2 digits of decimals. (if using `tab_model`, to get the stars you want, you can alter the asterisks with the code `p.threshold = c(.10, .05, .01)` and simply put the CI levels you want in `show.ci =` )

3. In the previous table, for which variables can we reject the null hypothesis that their coefficients equal 0 at the .01 level of significance?

4. For each percentage increase in black percentage population in a county, what do we predict about Trump's vote percent? What about for each percentage increase in hispanic population percentage?
5. Please include a plot with the predictions and 95% confidence intervals from this model across the range of the variable percentage hispanic for counties with a median household income of \$35,336 and a median household income of \$60,300.
6. Now run an identical model as before, but interacting hispanic percent with rural percent. Please present this model in a table with the same stylistic choices previously discussed in question 2.
7. Now include a plot with the predictions from this model and 95% confidence intervals across the range of the variable percentage hispanic for counties with a median household income of \$35,336 and a median household income of \$60,300. In words, what is it showing relative to the non-interactive model?