Problem set 7

Your name here

Due 11/15/2022 at 5pm

NOTE1: Start with the file `ps7_2022.Rmd` *(available from the github repository at https://github.com/UChicago-pol-methods/IntroQSS-F22/tree/main/assignments). Modify that file to include your answers. Make sure you can "knit" the file (e.g. in RStudio by clicking on the* `Knit` *button). Submit both the Rmd file and the knitted PDF via Canvas*

To make the results of your knitted problem sets comparable, set the seed to (arbitrarily chosen) 60637:

```
# keep this code as-is
set.seed(60637)
```

Question 1: Best linear predictor vs OLS

Consider the following joint PMF:

$$f_{X,Y}(x,y) = \begin{cases} 1/3 & x=0, y=0 \\ 1/6 & x=0, y=1 \\ 1/6 & x=1, y=1 \\ 1/3 & x=1, y=2 \\ 0 & \text{otherwise} \end{cases}$$

(1a) What are the coefficients (slope and intercept) of the best linear predictor (BLP) of Y given X? (Show your work, which will require computing $\mathrm{E}[X]$, $\mathrm{E}[Y]$, $\mathrm{V}[X]$, and $\mathrm{Cov}[X,Y]$.)

(1b) What is the prediction of the BLP at $X=1$? Confirm that this is the same as $E[Y|X=1]$.

(1c) Make a tibble with the same joint distribution of `x` and `y` as the joint PMF above. Regress `y` on `x` in this dataset, present the results in a regression table using the `huxreg()` command in the `huxtable` package, and confirm check that you recover the coefficients of the BLP.

(1d) Look up the `slice_sample()` command (part of the `dplyr` package, which is part of `tidyverse`). Draw a sample of size 100 (with replacement) from the tibble you created in (1c) and again regress `y` on `x` and store the result. Do the same again but make the sample size 1000. Use `huxreg()` in the `huxtable` package to display the regression results side by side. Comment about the two sets of results and how they relate to the BLP.

Question 2: OLS mechanics

Load the data on presidential elections from the course github. The url is below:

```
url <- "https://raw.githubusercontent.com/UChicago-pol-methods/IntroQSS-F22/main/data/pres_data.csv"
```

2a) Regress the incumbent party's vote share (`incvote`) on the president's approval rating in June (`juneapp`). Store the result and report it using `huxtable::huxreg()`.

2b) Write a function that computes the mean squared residual from a linear prediction of `incvote` based on `juneapp` given a slope and intercept. Use the function to compute the mean squared residual we obtain when we predict `incvote` using `juneapp` with the intercept you estimated in (2a) and a slope of .1.

Hints/suggestions: You may want to start by writing code that takes the dataset, generates predicted `incvote` given a slope and intercept, and computes the mean squared residual. Then wrap this in a function. The arguments to your function should be a `slope` and an `intercept`. Make sure the function returns a numeric value – you might need to use `as.numeric()` to convert the raw result of your code into a number.

2c) Using the function you wrote, compute the mean squared residual for a sequence of slopes between 0 and .3 (by increments of .005) and again using the intercept you computed in (2a). (Hint: you could use `map_dbl`, `map2_dbl`, `sapply`, or a for-loop to do this.) Plot the mean squared residual for each value of the slope, and add a red vertical line at the OLS slope you computed in (2a).

Question 3: Interpretation of regression coefficients

The CSV at https://andy.egge.rs/data/brexit/brexit_data.csv contains results of the 2016 UK Brexit referendum by local authority (collected from the Electoral Commission website) and 2011 census data. It was gathered by Claire Peacock.

3a) Load the data.

3b) Use `group_by()` and `summarize()` to make a table showing, for each `Region`, (i) the mean of `Percent_Leave` and (ii) the number of local authorities. (You may find the `n()` function useful.) Store the table for later use and display it below.

3c) (Law of iterated expectations applied to a sample) Compute the mean of `Percent_Leave` in this dataset in two ways: (i) unconditionally (the analogue of $E[Y]$) and (ii) as the weighted average of the region averages (the analogue of $E[E[Y \mid X]]$).

3d) Regress `Percent_Leave` on `Region`. Output the result using `huxtable::huxreg()`.

3e) Based on your regression, what is the predicted support for Leave in a local authority in London? Compare your answer to the average support for Leave in London authorities in the data.

3f) Make a figure showing `Bachelors_deg_percent` on the horizontal axis and `Percent_Leave` on the vertical axis. Include a dot for each local authority, with the size scaled by `Valid_Votes` and specifying `alpha = .5` in your `geom_point()` command to avoid excessive overplotting. Use `geom_smooth()` to estimate the CEF. Does the relationship look linear?

3g) Do the same, but estimate the CEF separately for Scotland and the rest of the sample. (Hint: create a variable that distinguishes Scotland from other places, and assign it to the `color` aesthetic.) Describe the result in words.

3h) Based on what you found in (3g), run a regression predicting support for Leave in a local authority as a function of the proportion of residents with a bachelors degree and whether the local authority is in Scotland. (Do you include an interaction? Explain why or why not.) Report the result in a regression table as above. According to your model, what is the predicted support for Brexit in a Scottish local authority in which 30% of inhabitants have a bachelor's degree?