

## Problem Set #6

PPD 558

*Total points: 36*

1. Using data from from a variety of sources, we will look at the effects of demographics and economic circumstances on voting behavior in the US Presidential Election at the county level.<sup>1</sup> These data are in the file `ps6_pt1_voting.dta`. After you've inspected the data and you're sure you understand what it contains, proceed.

- (a) (4 points) Begin with the regression

$$\begin{aligned}\text{pct\_biden} = & \beta_0 + \beta_1 \text{pct\_poverty} + \beta_2 \text{MedianIncome} \\ & + \beta_3 \text{pct\_white} + \beta_4 \text{pct\_black} \\ & + \beta_5 \text{pct\_asian} + \beta_6 \text{pct\_otherrace} \\ & + \beta_7 \text{pct\_hisplatino} + \varepsilon,\end{aligned}$$

where: `pct_biden` = Percentage of ballots cast for Joe Biden in 2020  
`MedianIncome` = Median income in 2019  
`pct_white` = Percent of the population whose only racial identity is White  
`pct_black` = Percent of the population identifying as Black only  
`pct_asian` = Percent of the population identifying as Asian only  
`pct_otherrace` = Percent identifying as another race, or as having more than one racial identity  
`pct_hisplatino` = Percent of the population identifying as Hispanic or Latino  
`pct_poverty` = Percent of the population with income below the poverty line.

- i. How do you interpret the coefficients you obtained?
- ii. Are you missing any of the  $\beta$ s? Explain.

---

<sup>1</sup>Data are taken from a variety of sources and cleaned by Nic Duquette. County-level estimated race, income, and poverty statistics are from IPUMS-NHGIS (<https://www.nhgis.org/>). County-level election policy data are from Verified Voting (<https://verifiedvoting.org>). Year 2020 presidential vote data is taken from Tony McGovern's GitHub archive (<https://github.com/tonmcg>).

- (b) (2 points) Define a new variable for log of median income. Re-run the regression including this new variable and dropping the `MedianIncome` and `pct_white` variables. Do you prefer this specification?
- (c) (2 points) Implement your regression from part (b) again, but instead of predicting percentage of votes as the outcome, make the dependent variable `total_votes_biden`. What is this variable? Should we prefer this specification?
- (d) (2 points) Are any of our regressors substantially multicollinear? Test for multicollinearity in the regression we implemented in part (b). Should we do anything about what we find?
- (e) (2 points) Are our results due to influential outliers? Implement a rule of thumb for identifying outliers, and run the regression from part (b) with non-outliers only.
- (f) (2 points) Are our standard errors understated due to heteroskedasticity? Test for heteroskedastic residuals, and run a regression correcting for heteroskedasticity. Do your conclusions change?
- (g) (2 points) The variable `early` is equal to 1 if a county has a process for casting a ballot before Election Day, 0 otherwise. Early voting makes casting a ballot easier for people with rigid work schedules or poor polling access. Since 2020, some state legislatures have controversially enacted laws that restrict early voting and other convenient forms of ballot access in the belief that this will reduce votes for Democrats in future elections.

Add `early` as a regressor to the model estimated in part (f). What is the coefficient on `early`? Do you believe this captures a causal effect of early voting on two-party vote share?

- (h) (2 points) Use Stata's factor notation to add a categorical regressor for each state to the early-voting model you estimated in part (g). What do those categorical variables mean? How do you interpret the demographic variables now? What about `early`?
- (i) (2 points) Low-income people are the specific group early voting is most important to. Estimate the regression from part (g) again, but use Stata's factor notation to add an interaction between `pct_poverty` and `early`. What do you observe? How do you interpret this estimate?

2. For our second question, we look at data on professional European soccer players.<sup>2</sup> Open the file `ps6_pt2_futbol.dta`.

Each observation is one professional player as of 2008. The variables include:

`playquality4` = And index from 0 to 3 of the player's quality

`experience` = Years of playing experience

`age` = Age in years

`nation` = Nationality of the player

`residence` = Country where the player resides and plays

`taxrate_domestic` = Marginal tax rate if the player were to live in his country of origin

`taxrate_foreign` = The average marginal tax rate the player would pay in 13 European countries other than their country of origin

- (a) (2 points) Create a new variable, `isdomestic`, that is equal to 1 if the player plays in his country of origin and 0 if he plays in a different country. What percentage of players in the dataset play abroad?
- (b) (2 points) Estimate a linear probability model predicting the chance that a soccer player will play in their home country as a function of age, experience, quality, and tax rates at home and abroad. What do you observe?
- (c) (2 points) Use your model from part (b) to predict the probability that individual players will go stay home. Are your predictions reasonable?
- (d) (2 points) Implement the model from part (b) as a logit. Do the coefficients signs or statistical significance change for any variables?
- (e) (2 points) It's hard to interpret logit coefficients, because they're transformed by an exponent inside a fraction. Calculate the marginal effects of all five regressors at the sample means. Do they look different from the LPM marginal effects?
- (f) (2 points) Players don't just decide whether to stay home; there are many different countries to possibly play in, each of which a player might be more or less interested in.  
Estimate a multinomial logit for the country a player chooses to play in as a function of their age, experience, quality, and tax rates.
- (g) (2 points) As players increase their quality at the margin, which countries are they more or less likely to choose to play in?

---

<sup>2</sup>Source: Kleven, Henrik Jacobsen, Camille Landais, and Emmanuel Saez. 2013. "Taxation and International Migration of Superstars: Evidence from the European Football Market." *American Economic Review*, 103 (5): 1892-1924.

Hint: after you run the correct `margins` command, if you add the following code to your do-file, you will get a nice visualization of your answer.

```
#delimit;
marginsplot,
    scheme(s1color)
    xlabel(1 "Austria"
           2 "Belgium"
           3 "Denmark"
           4 "England"
           5 "France"
           6 "Germany"
           7 "Greece"
           8 "Italy"
           9 "Netherlands"
          10 "Norway"
          11 "Portugal"
          12 "Spain"
          13 "Sweden"
          14 "Switzerland",
    alternate labsize(small))
    xtitle("")
    yline(0, lcolor(blue))
;
```

- (h) (2 points) Tax rates might be especially important for the elite players who are paid the most. Use `margins` one more time to calculate (1) the marginal effect of a higher foreign tax rate on (2) the probability of playing in England (country number 4), (3) at each of the four tiers of player quality. What do you observe?