ISYE 6501
Homework 5

September 20, 2019

Question 8.1

It is common in education research to estimate what scholars in the field term Education Production Functions (EPFs). The EPF is typically a linear regression model expressing a quantitative measure of achievement (e.g. SAT scores) as a function of individual, school, and parental characteristics. The most important predictors are parental education, student innate ability, peers effects, teacher quality, and coursework. These need be operationalized, a task which is accomplished by researchers within the data constraints and with a sense of what metrics are the best.

- Parental education often times takes the form of a categorical variable with levels for the highest level of parental education (i.e. doctoral, college, some college, high school, less than high school).
- Student innate ability is hard to pin down, but theoretical models of student achievement all convene that taking it out from regression equations would cause misspecification. In some cases, for instance in studies of PhD students' outcomes, GRE test scores are used. A measure of IQ is potentially the best, but – unfortunately for some researchers and fortunately for some other – there are ethical barriers preventing the collection of this information.
- Peers' effects are also a riddle. What researchers do a lot is taking class average parental education. Other much more complicated workarounds involve estimating class fixed effects (only viable when a panel of data is available).
- Teacher quality can be operationalized by years of experience, highest degree, professional qualification, SAT scores, etc.
- Coursework is little perused, but, again, it is hard to assume that prior course-taking should not be factored in when appraising student outcomes. An IQ of 150 (maybe 180 would do, who knows) would hardly make up for a lack of math exposure, nor it would guarantee that an international student is fluent in English if he never received English instruction prior to attending college in the US.

A last and quick aside. We must notice how peers and teacher effects create a problem for linear regression because their inclusion results in clustered error terms. This is above and beyond the scope of this answer and maybe of this class, but it is the reason why educational researchers have recently turned to Hierarchical Linear Modeling (HLM). Other researchers, mostly trained in economics, stick with linear regression and use cluster-robust standard errors and other variance-inflation techniques.

Question 8.2

In this answer, we will follow the approach that a social scientist (not necessarily a data scientist working in the industry) would take.[1] She starts off by writing down a theoretical model – an economist will typically write an objective function and derive first order conditions – and clearly stating her hypotheses. Then, she will use empirical data to put these hypotheses to a test by operationalizing the model's variables with specific observables. In linear regression, the tests come in the form of the t-statistics associated with the regression parameters, embedding the null hypothesis that the relationship between the regressor and the dependent variable is zero in the population. The distinction between sample and population is crucial: the social scientists want to make inferences about a given population but all she has is a sample.[2]

The primary goal of the social scientist is different than maximizing the predictive power of her model, and is that of specifying a regression equation that properly captures the data generating process for the phenomenon of interest. In the case of the *uscrime* dataset, our researcher would look up the criminology and crime economics literature to gain guidance on the model specification. None of us is extremely familiar with either of those; however, we came up with some hypotheses hinging on common sense and the little bit of the scientific insight we have on crime.

These were our hypotheses:

- We suspected that crime is a decreasing function of household wealth because of the differential opportunity cost associated with committing an offense for poor and rich individuals.
- We also suspected that the effect of wealth almost entirely channels the effect of education on crime (i.e. higher levels of education dovetail into greater wealth *hence* dictating criminal behavior) and therefore the two are unlikely to be jointly significant in a regression model.
- We suspected states with warmer climate to exhibit comparatively higher levels of crime and therefore expected Southern states to exhibit greater criminal activity.
- We suspected that the effect of police expenditures on crime rates is a positive one, upon conditioning on the probability of imprisonment. We should break this statement down a little bit more. Greater police expenditure would have, by themselves, a duplicitous impact, that of increasing the probability of an offense being caught (thus knocking down criminal records) and that of increasing the number of undetected crimes that are identified (thus bumping them up). In the light of this, we included the probability of imprisonment in the regression equation, and expected it to enter negatively. The conditional effect of police expenditure that we expected was then a positive.
- Finally, we suspected that states with higher percentages of males in their teens and early twenties would exhibit higher crime rates because this age group is more prone to delinquency.

The model can be formally written as:

$$Crime_s = f(Wealth_s, South_s, Pol\_Exp_s, Impr\_Rate_s, Male_s) \tag{1}$$

where the subscript s indexes the state. We expect that $\frac{\partial Crime_s}{\partial Wealth_s} < 0$, $\frac{\partial Crime_s}{\partial South_s} > 0$, $\frac{\partial Crime_s}{\partial Pol\_Exp_s} > 0$, $\frac{\partial Crime_s}{\partial Impr\_Rate_s} < 0$, and $\frac{\partial Crime_s}{\partial Male_s} > 0$. At this point, we operationalized our variables:

- The dependent variable *Crime* reports number of offenses per 100,000 population in 1960, which is

[1] Full code in Appendix.

[2] Please note that the t-statistic in a bivariate regression model with one dummy-coded regressor is the same statistic generated from a t-test or difference-of-means test

why do not need to separately account for population in the model.

- The variable *Wealth* reports median value of transferable assets in tens of $.
- *So* is an indicator for Southern state, which we recoded to be a factor variable.[3]
- *Po1* are records of police per capita expenditures in $ (in 1960).
- *Prob* is the ratio of imprisonments over total offenses with domain $[0, 1]$ and was recoded into the variable *Prob2* with domain $[0, 100]$.
- Finally, M captures the percentage of total population aged 14-24.

The empirical model reads:

$$\widehat{Crime} = b_0 + b_1 Wealth + b_2 So + b_3 Po1 + b_4 Prob2 + b_5 M \tag{2}$$

We run the baseline model first, and then tweaked it one variable at a time to test for (some) of the potential pathways highlighted before. In the social sciences, we would typically need to declare the expected structure of the error term, which is zero-mean and constant variance in this case because we have no apparent reasons to suspect that residuals cluster into groups. Each coefficient b_k represents the partial derivative of the estimation equation with respect to the associated regressor k and informs us on the changes in $\widehat{Crime}$ engendered by unitary changes in k while the other regressors are held constant.

```
##Call:
##lm(formula = data$Crime ~ Wealth + M + factor(So) + Po1 + Prob2,
##    data = data)
##
##Residuals:
##    Min      1Q  Median      3Q     Max
##-484.6 -133.8   28.3  106.1  552.8
##
##Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
##(Intercept) -506.15599  798.77649   -0.63   0.5300
##Wealth        -0.00751    0.07989   -0.09   0.9256
##M             67.92082   41.10827    1.65   0.1065
##factor(So)1  152.43537  111.52050    1.37   0.1795
##Po1           77.95461   22.83285    3.41   0.0015 **
##Prob2        -48.65614   20.37981   -2.39   0.0219 *
##---
##Signif. codes:
##0    ***    0.001    **    0.01    *    0.05    .    0.1          1
##
##Residual standard error: 243 on 39 degrees of freedom
##Multiple R-squared:  0.486,    Adjusted R-squared:  0.42
##F-statistic: 7.36 on 5 and 39 DF,  p-value: 0.0000611
```

The sample coefficient on *So* tells us that holding wealth, police expenditures, probability of imprisonment, and proportion of male aged 14-24 constant, Southern states in the sample have about 152 more offenses per 100,000 population,[4] thus supporting our hypothesis. Although the t-statistic fails statistical significance by landing 1.37 units right of zero, the result might still be of substantive significance because of the small sample comprising 45 observations.

[3]Estimation results do not usually change significantly with dummy-coded regressors treated as integers, but they do change meaningfully with factor variables treated as such. Therefore, it is good practice to always convert non-numeric variables to factors.

[4]The phrasings "per 100,000 population" "and holding..." are omitted going forward.

The effect of wealth is that of decreasing the expected number of offenses. Specifically, we would say that a 10K increase in household wealth (remember that the variable is coded in tens of $) decreases expected offenses by a meager 1/20 of an offense. Practically, what the output really informs us about is that wealth has a statistically indistinguishable impact from zero (p-value = 0.93).
The effect of police expenditure is that of increasing the expected number of offenses, an effect which is significant at the 1% level (p-value < 0.01). This means that for each additional per capita dollar of police spending, the number of expected offenses decreases by about 78 (we are talking per capita $ here, so the intervention is far from inexpensive).
Lastly, the effect of imprisonment probability matches expectations and is statistically significant. One percentage-point greater chance of imprisonment converts into about 49 less offenses (please recall the re-coding of the variable *Prob* into *Prob2*).
The models R^2, or ratio between model sum of squares and total sum of squares, is 48.6%, meaning that the model explains about 48.6% of the variation in the dependent variable (the remaining 51.4% is left to the residuals and is stored in the residual sum of squares). The residuals of the models align decently to the 45 degree reference line in the Q-Q plot (see Figure 1). Had residuals happened to be more sparse, this would have questioned the model specification and raised concerns over omitted variables. That is, we would have omitted some crucial explanatory variables, whose relation with the dependent variable resurfaced in the error term.

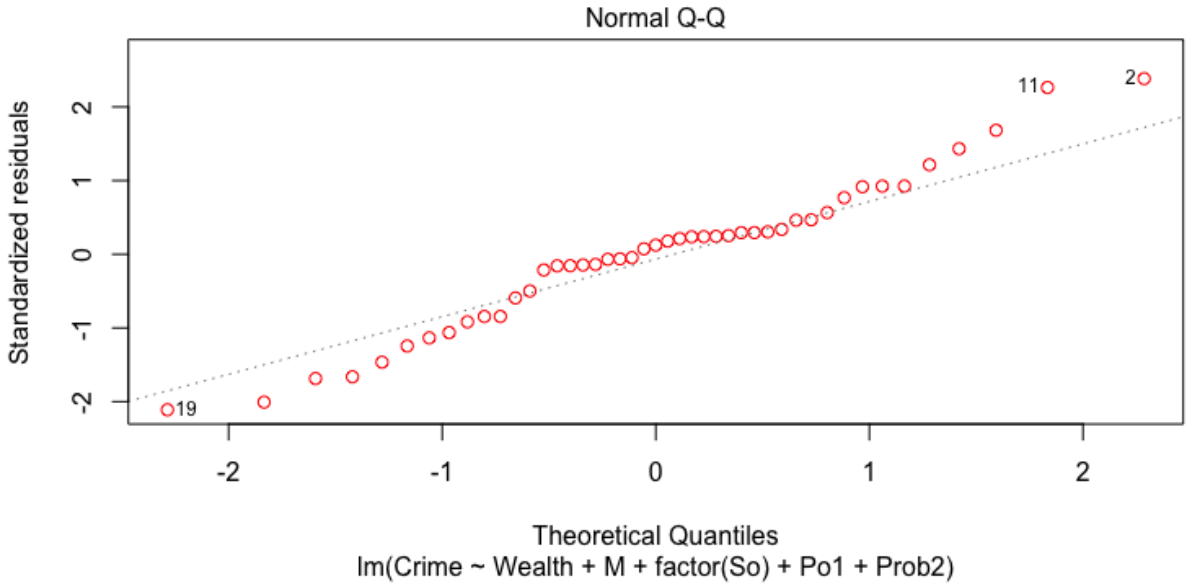

Figure 1: Q-Q Plot

Moving forward, our social scientist would now gain further insight into her hypotheses by leaving out-/adding one variable at a time to the baseline model. The t-statistics directly reported in the regression outputs are robust enough of a test in most occasions; however, a likelihood ratio test could be necessary

when evidence is inconclusive. For instance, we hypothesized that wealth mainly channels the effect of education, which we can test by running a model with average years of education (*Ed*) as an explanatory variable. We expected both variables to come out as insignificant given that they are highly correlated (r = 0.709). However, wealth moved closer to statistical significant relative to the baseline model while average years of education was extremely close to statistical significance (p-value = 0.07). More so, the sample parameter estimate on *Ed* turned out to be positive, meaning that the direct effect of education net of its indirect effect through wealth is actually that of increasing number of offenses. Puzzled by these results, we run a further model which leaves out *Wealth* and specifies *Ed*. This one time, the estimate is more clearly insignificant, and the sample coefficient reduces in size albeit retaining its rather puzzling positive sign. This means a couple things: first of all, that we might hold on to our hypothesis of a negative pathway running from education to number of offenses through wealth (had the coefficient on *Ed* increased further, we could have not legitimately done so); secondly and more importantly, that better data is needed to solve the puzzle. So far is a cursory token of how social scientists approach regression modeling, that is: stating hypotheses and tying their hands to them before jumping into the data. Some of them will not necessarily abide by the golden rule and would tweak and twist the data until they speak (their) truth. This is the furthest from good science, and does not shed positive light on the profession.

Moving on, we added lagged police expenditures in year 1959 (*Po2*), and *Po1* reduced in significance, almost certainly because of multicollinearity. Consistently with path-dependency of police expenditures, running a model with *Po2* and without *Po1* gave by far and large the same results of the baseline model, suggesting that the two predictors are almost interchangeable. Lastly, we run further models with unemployment rates in 1960 (*U1*) and 1959 (*U2*) to parse out the potential effect of what economists term the "intensive margin" of labor participation. In a gist, whether we are employed or not tells about the intensive margin of our labor participation while our income represents its extensive margin. We can imagine models of crime microeconomics that support differential effects of the two on the chance of committing a crime. Anyways, neither the inclusion of *U1* nor of *U1* to the baseline model increased its predictive performance or altered the other estimates meaningfully.

Wrapping up, estimates for our baseline model support three of our hypothesis: the positive relation of warmer climates and police expenditures on offenses, and the negative relation of imprisonment probability and offenses. Surprisingly enough, it does not support our hypothesis that household wealth dictates lower levels of offenses. The social scientist would put the lid on this by calling for further research and for using better and larger datasets. That is also how we conclude.

Appendix

```
# reading data
data <- read.delim("uscrime.txt", header=TRUE)

-------------------- DATA PREPARATION --------------------
# removing outliers
# we concluded in HW3 that obs 4 (Crime = 1969) and 26 (Crime = 1993) are outliers
# we want to be conservative and therefore we remove them from the data
data <- data[-c(4,26),]
# avoiding scientific notation for results and requiring 3 digits
options(scipen=4, digits = 3)

# recoding So to factor
data$So = as.factor(data$So)

# recoding Prob to percentage-points into Prob2
Prob2 <- data$Prob*100
data <- cbind(data, Prob2)
data <- within(data, rm(Prob))

-------------------- LINEAR MODEL --------------------
# baseline model & QQ-plot

model1 <- lm(Crime ~ Wealth + M + factor(So) + Po1 + Prob2, data = data)
summary(model1)
plot(model1, which=2, col=c("red"))


# Including Ed
model2 <- lm(Crime ~ Wealth + Ed + M + factor(So) + Po1 + Prob2, data = data)
summary(model2)
model2 <- lm(Crime ~ Ed + M + factor(So) + Po1 + Prob2, data = data)
summary(model2)

# Including Po2
model3 <- lm(Crime ~ Wealth + M + factor(So) + Po1 + Po2 + Prob2, data = data)
summary(model3)

Including U1 & U2
model4 <- lm(Crime ~ Wealth + U1 + M + factor(So) + Po1 + Prob2, data = data)
summary(model4)
model4 <- lm(Crime ~ Wealth + U2 + M + factor(So) + Po1 + Prob2, data = data)
summary(model4)
model4 <- lm(Crime ~ Wealth + U1 + U2 + M + factor(So) + Po1 + Prob2, data = data)
summary(model4)
```