Andrew ElHabr
ECON 7022
12/7/17

Problem Set 3

I unselect Question T4 - the fourth theoretical exercise. Also, I choose not to do Question P4 in the coding exercises. The programming is done in R.
I worked with Cyrus Rich for a few hours on this assignment.
Some of this work was rushed, so please note this. All of this work was conducted on Saturday, Sunday, and Thursday. I wish I could have dedicated more time to the assignment because I felt that I learned a lot.

Part I: Reading Assignment [50 points]

Question R1: Goldfarb Tucker Guide [50 points]

- There are several steps in the etiquettes for DiD and IV that are similar, including explaining and defending the experiment, conducting multiple robustness, checks, discussing the assumptions behind homogeneous treatment effects, explaining why the treated population is inherently interesting, and clearly recognizing the main assumptions and limitations for each technique. The main differences between the two etiquettes lie in their beginning steps. In addition, defending the experiment can take on very different forms for each etiquette. For IV, we must test for power and overidentification and do a reduced-form regression of the dependent variable on the instruments. This is quite different than DiD because of the completely different nature of the IV method. In IV, we really care about ensuring that we have useful instruments, i.e. not weak, that the exclusion restriction holds for our instruments, and that we do not saturate our model with instruments, which can lead to unexpected bias. There is simply no analogy for these concerns in DiD because only in IV do we have assumptions about the exogeneity of the instrumental variable and the relevance of this variable that simply are not part of the DiD model. As for DiD, explaining and defending the experiment is usually easier to do and more transparent. Also, for the DiD etiquette, we check that raw data indicates that the treatment and control groups are similar in their covariates prior to treatment, that pre-treatment patterns are about the same for both groups, and that the variation of the outcomes is sensible. Finally, the DiD etiquette also calls for presenting baseline estimates of the treatment effect without controls to assess the impact of potential omitted variables. All of these steps centered around the treatment are inherently different than what is seen with IV because the treatment in IV is taken to be the introduction of the instrument itself into the quasi-experiment, which is very different than some outside "shock" that is usually the source of variation for DiD. This primarily has to do with the fact that DiD often has some time-related aspect to its "shock", while IV does not.
- Many of the differences in etiquette that I just discussed for DiD and IV are similar to the differences in etiquette for RD and IV. Once again, the main differences between the two etiquettes lie in their beginning steps, and defending the experiment can take on very different forms for each etiquette. For example, for RD, we have to defend why the source of the threshold is essentially arbitrary and cannot be linked to some underlying discontinuities in behavior. In addition, we have to think carefully about the cases very close to the threshold represent cases on their respective sides of the threshold accurately. There is simply no analogy to this kind of concern in IV. Similarly to DiD, we must chose that the treatment and control groups are similar in their observables so that the counterfactual is valid. Also, the RD etiquette calls for providing baseline estimates and clustering standard errors, which also has no true analogy in IV. The core difference is that treatment is arbitrary for RD, while for IV, we must decide on a good instrument variable, which is what our treatment is for that type of identification

Part II: Theoretical Exercises [100 points]

Question T1: Definitions and Estimators [33 points]

1.1) Linear Probability Model [10 points]

(a)
- It is called a binary (outcome) variable.
- This is called a linear probability model.
-

$$\hat{\boldsymbol{\beta}}_k = \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^K} \frac{1}{n}\sum_{i=1}^{n}(y_i - \boldsymbol{x}_i'\boldsymbol{\beta}_0)^2$$

Thus,

$$\begin{aligned}
\mathbf{0} =& \frac{\partial}{\partial \boldsymbol{\beta}_0}\Big(\frac{1}{n}\sum_{i=1}^{n}(y_i - \boldsymbol{x}_i'\boldsymbol{\beta}_0)^2\Big) \\
=& \frac{\partial}{\partial \boldsymbol{\beta}_0}\Big(\frac{1}{n}\sum_{i=1}^{n}y_i^2 - \boldsymbol{\beta}_0'\frac{2}{n}\sum_{i=1}^{n}\boldsymbol{x}_i y_i + \boldsymbol{\beta}_0'\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{x}_i\boldsymbol{x}_i'\boldsymbol{\beta}_0\Big) \\
=& -\frac{2}{n}\sum_{i=1}^{n}\boldsymbol{x}_i y_i + \frac{2}{n}\sum_{i=1}^{n}\boldsymbol{x}_i\boldsymbol{x}_i'\hat{\boldsymbol{\beta}}_0 \\
=& -\sum_{i=1}^{n}\boldsymbol{x}_i y_i + \sum_{i=1}^{n}\boldsymbol{x}_i\boldsymbol{x}_i'\hat{\boldsymbol{\beta}}_0 \\
\implies \hat{\boldsymbol{\beta}}_0 =& \Big(\sum_{i=1}^{n}\boldsymbol{x}_i\boldsymbol{x}_i'\Big)^{-1}\Big(\sum_{i=1}^{n}\boldsymbol{x}_i y_i\Big)
\end{aligned}$$

- The estimated coefficient $\hat{\boldsymbol{\beta}}_{\mathbf{0},\boldsymbol{k}}$ can be interpreted as the change in probability that a person was retired/unemployed in 2008 as a result in the unit change of $\boldsymbol{x}_{\boldsymbol{k}}$, i.e. the k-th covariate.
- The linear regression model can predict values that are outside the range of [0,1] (and in fact, can predict values anywhere in the range of $(-\infty, \infty)$), which clearly doesn't make sense for the type of dependent variable we have. This is why we usually use a logit or probit model when we are trying to predict probabilities.

1.2) Advertising Beer Again [13 points]

- – The ATE that we want to measure is the expected effect of the advertisement on beer consumption for a given person. Given that

$$Y_i = \begin{cases} Y_{1i} & \text{if } A_i = 1, \\ Y_{0i} & \text{if } A_i = 0, \end{cases} \tag{1}$$

where A_i is the indicator if the advertisement was seen at the football game and Y_i is beer consumption, the ATE can be expressed as

$$\mathbb{E}[Y_{1i} - Y_{0i}].$$

The ATE is unobservable in the sense that we can never truly measure the beer consumption of a person who has seen the advertisement had they never seen the advertisement, and vice versa, i.e. only one of Y_{1i} and Y_{0i} can be observed for person i. This means that we cannot directly measure the ATE on any person.

- We can observe the observed treatment effect of the advertisement, which can be expressed as

$$\mathbb{E}[Y_i|A_i = 1] - \mathbb{E}[Y_i|A_i = 0].$$

- The observed treatment effect of the advertisement can be decomposed in the following way:

$$\mathbb{E}[Y_i|A_i = 1] - \mathbb{E}[Y_i|A_i = 0] = \mathbb{E}[Y_{1i} - Y_{0i}|A_i = 1] + \mathbb{E}[Y_{0i}|A_i = 1] - \mathbb{E}[Y_{0i}|A_i = 0],$$

where $\mathbb{E}[Y_{1i} - Y_{0i}|A_i = 1]$ is the unobserved ATET and $\mathbb{E}[Y_{0i}|A_i = 1] - \mathbb{E}[Y_{0i}|A_i = 0]$ is the unobserved selection bias. Thus, the ATE is not generally the same as the observed treatment effect since $\mathbb{E}[Y_{1i} - Y_{0i}]$ does not necessarily equal $\mathbb{E}[Y_{1i} - Y_{0i}|A_i = 1] + \mathbb{E}[Y_{0i}|A_i = 1] - \mathbb{E}[Y_{0i}|A_i = 0]$. This can only be the case if selection bias is equal to 0.

- The selection bias in a formal sense is

$$\mathbb{E}[Y_{0i}|A_i = 1] - \mathbb{E}[Y_{0i}|A_i = 0].$$

This example of advertising beer in a stadium likely suffers from selection bias since it is likely that the beer consumption of person i, Y_i, would be relatively high for someone who went to the game since these types of people tend to have a higher affinity for beer than people who do not care to go to games, i.e. $\mathbb{E}[Y_{0i}|A_i = 1] \neq \mathbb{E}[Y_{0i}|A_i = 0]$ and $\mathbb{E}[Y_{0i}|A_i = 1] - \mathbb{E}[Y_{0i}|A_i = 0] > 0$.

- LATE is essentially an ATE on a specific subpopulation and does not necessarily generalize to the full population of interest. More formally,

$$LATE = \frac{\mathbb{E}[Y_i|Z_i = 1] - \mathbb{E}[Y_i|Z_i = 0]}{Pr(A_i = 1|Z_i = 1) - Pr(A_i = 1|Z_i = 0)}.$$

- It could estimate a LATE. We need the assumptions of random assignment, exclusion restriction, relevance, and monotonicity.

- BLUE:
  Always takers - people who always buy beer in Publix regardless of if they got a coupon
  Never takes - people who never buy beer in Publix regardless of if they got a coupon
  Defiers - people who only buy beer in Publix if they got a coupon
  Compliers - people who only do not buy beer in Publix if they got a coupon

  GREEN:
  Always takers - people who always buy beer in Publix regardless of if they got a coupon
  Never takes - people who never buy beer in Publix regardless of if they got a coupon
  Defiers - people who only buy beer in Publix if they got a coupon
  Compliers - people who only do not buy beer in Publix if they got a coupon

- One could obtain confidence intervals on the LATE estimates in order to determine if they are statistically significant. This is not covered in the lecture notes. Economic importance can be established if all assumptions are met.

- I would expect to see the larger effect for the blue vouchers the people in the Publix loyalty program are probably always going to Publix while the ones at the stadium may not always go to Publix to buy beer. I would infer that the setting of our advertisement matters when assessing its effect.

- The LATE will equal the ATE occurs under the condition that there are no always-takers. I would suspect that the blue voucher would come closer to satisfying this because it is most likely that people in the loyalty club will buy their beer at Publix no matter if there is an advertisement or not.

- I think this condition would be that the LATE are significantly different for the vouchers.

- We would need to ensure that there are no always-takers. One design that could potentially take on is showing the advertisement to people who do not tend to have a high affinity for beer and not examine the effect for people who go to Publix frequently. Advertising like this could be done near a competing grocery store.

1.3) Panel/DiD Problem [10 points]

1. For the FD model, we consider the previous time period of the standard unobserved effects model in addition to the standard unobserved model:

$$
\begin{aligned}
y_{i(t-1)} =& x_{i(t-1)}\beta + c_i + u_{i(t-1)} \\
\implies y_{it} - y_{i(t-1)} =& (x_{it} - x_{i(t-1)})\beta + (c_i - c_i) + (u_{it} - u_{i(t-1)}) \\
=& (x_{it} - x_{i(t-1)})\beta + (u_{it} - u_{i(t-1)}) \\
\implies \Delta y_i =& \Delta x_i \beta + \Delta u_i.
\end{aligned}
$$

Thus, we find the FD estimate in the following way:

$$
\begin{aligned}
\hat{\beta}_{FD} =& \arg\min_{\beta} \frac{1}{n}\sum_{i=1}^{n}(\Delta y_i - \Delta x_i \beta)^2 \\
\implies 0 =& \frac{\partial}{\partial \beta}\Big(\frac{1}{n}\sum_{i=1}^{n}(\Delta y_i - \Delta x_i \beta)^2\Big) \\
=& \frac{\partial}{\partial \beta}\Big(\frac{1}{n}\sum_{i=1}^{n}(\Delta y_i)^2 - \beta' \frac{2}{n}\sum_{i=1}^{n}\Delta x_i \Delta y_i + \beta' \frac{1}{n}\sum_{i=1}^{n}\Delta x_i \Delta x_i' \beta\Big) \\
=& -\frac{2}{n}\sum_{i=1}^{n}\Delta x_i \Delta y_i + \frac{2}{n}\sum_{i=1}^{n}\Delta x_i \Delta x_i' \hat{\beta}_{FD} \\
=& -\sum_{i=1}^{n}\Delta x_i \Delta y_i + \sum_{i=1}^{n}\Delta x_i \Delta x_i' \hat{\beta}_{FD} \\
\implies \hat{\beta}_{FD} =& \Big(\sum_{i=1}^{n}\Delta x_i \Delta x_i'\Big)^{-1}\Big(\sum_{i=1}^{n}\Delta x_i \Delta y_i\Big).
\end{aligned}
$$

For the FE model, we consider the average over time of the standard unobserved effects model in addition to the standard unobserved model:

$$
\begin{aligned}
\bar{y}_i =& \bar{x}_i \beta + c_i + \bar{u}_i \\
\implies y_{it} - \bar{y}_i =& (x_{it} - \bar{x}_i)\beta + (c_i - c_i) + (u_{it} - \bar{u}_i) \\
=& (x_{it} - \bar{x}_i)\beta + (u_{it} - \bar{u}_i).
\end{aligned}
$$

Thus, we find the FE estimate in the following way:

$$
\begin{aligned}
\hat{\beta}_{FE} =& \arg\min_{\beta} \frac{1}{2n}\sum_{i=1}^{n}\sum_{t=1}^{2}((y_{it}-\bar{y}_i)-(x_{it}-\bar{x}_i)\beta)^2 \\
\implies 0 =& \frac{\partial}{\partial\beta}\Big(\frac{1}{2n}\sum_{i=1}^{n}\sum_{t=1}^{2}((y_{it}-\bar{y}_i)-(x_{it}-\bar{x}_i)\beta)^2\Big) \\
=& \frac{\partial}{\partial\beta}\Big(\frac{1}{2n}\sum_{i=1}^{n}\sum_{t=1}^{2}(y_{it}-\bar{y}_i)^2-\beta'\frac{1}{n}\sum_{i=1}^{n}\sum_{t=1}^{2}(x_{it}-\bar{x}_i)(y_{it}-\bar{y}_i)+\beta'\frac{1}{2n}\sum_{i=1}^{n}\sum_{t=1}^{2}(x_{it}-\bar{x}_i)(x_{it}-\bar{x}_i)'\beta\Big) \\
=& -\frac{1}{n}\sum_{i=1}^{n}\sum_{t=1}^{2}(x_{it}-\bar{x}_i)(y_{it}-\bar{y}_i)+\frac{1}{n}\sum_{t=1}^{2}\sum_{i=1}^{n}(x_{it}-\bar{x}_i)(x_{it}-\bar{x}_i)'\hat{\beta}_{FE} \\
=& -\sum_{i=1}^{n}\sum_{t=1}^{2}(x_{it}-\bar{x}_i)(y_{it}-\bar{y}_i)+\sum_{i=1}^{n}\sum_{t=1}^{2}(x_{it}-\bar{x}_i)(x_{it}-\bar{x}_i)'\hat{\beta}_{FE} \\
\implies \hat{\beta}_{FE} =& \Big(\sum_{i=1}^{n}\sum_{t=1}^{2}(x_{it}-\bar{x}_i)(x_{it}-\bar{x}_i)'\Big)^{-1}\Big(\sum_{i=1}^{n}\sum_{t=1}^{2}(x_{it}-\bar{x}_i)(y_{it}-\bar{y}_i)\Big).
\end{aligned}
$$

Thus,

$$
\begin{aligned}
\hat{\beta}_{FE} =& \Big(\sum_{i=1}^{n}\sum_{t=1}^{2}\Big(x_{it}-\frac{x_{i1}+x_{i2}}{2}\Big)\Big(x_{it}-\frac{x_{i1}+x_{i2}}{2}\Big)'\Big)^{-1}\Big(\sum_{i=1}^{n}\sum_{t=1}^{2}\Big(x_{it}-\frac{x_{i1}+x_{i2}}{2}\Big)\Big(y_{it}-\frac{y_{i1}+y_{i2}}{2}\Big)\Big) \\
=& \Big(\sum_{i=1}^{n}\Big(\frac{x_{i1}-x_{i2}}{2}\Big)\Big(\frac{x_{i1}-x_{i2}}{2}\Big)'+\Big(\frac{x_{i2}-x_{i1}}{2}\Big)\Big(\frac{x_{i2}-x_{i1}}{2}\Big)'\Big)^{-1} \\
& \times\Big(\sum_{i=1}^{n}\Big(\frac{x_{i1}-x_{i2}}{2}\Big)\Big(\frac{y_{i1}-y_{i2}}{2}\Big)+\Big(\frac{x_{i2}-x_{i1}}{2}\Big)\Big(\frac{y_{i2}-y_{i1}}{2}\Big)\Big) \\
=& \Big(\sum_{i=1}^{n}\frac{\Delta x_i\Delta x_i'}{4}+\frac{(-\Delta x_i)(-\Delta x_i)'}{4}\Big)^{-1}\Big(\sum_{i=1}^{n}\frac{\Delta x_i\Delta y_i}{4}+\frac{(-\Delta x_i)(-\Delta y_i)'}{4}\Big) \\
=& \Big(\sum_{i=1}^{n}\frac{\Delta x_i\Delta x_i'}{2}\Big)^{-1}\Big(\sum_{i=1}^{n}\frac{\Delta x_i\Delta y_i}{2}\Big) \\
=& \Big(\sum_{i=1}^{n}\Delta x_i\Delta x_i'\Big)^{-1}\Big(\sum_{i=1}^{n}\Delta x_i\Delta y_i\Big) \\
=& \hat{\beta}_{FD}.
\end{aligned}
$$

2. The bias adjusted estimate of the mean error variance from the FD method is

$$
\hat{s}^2_{FD} = \frac{1}{n-K}\sum_{i=1}^{n}(\Delta\hat{u}_i)^2 = \frac{1}{n-K}\sum_{i=1}^{n}(\Delta y_i-\Delta x_i\hat{\beta}_{FD})^2.
$$

Thus, the error variance estimate from the FD method is

$$
\hat{V}_{\hat{\beta}_{FD}} = \Big(\sum_{i=1}^{n}\Delta x_i\Delta x_i'\Big)^{-1}\hat{s}^2_{FD} = \frac{1}{n-K}\Big(\sum_{i=1}^{n}\Delta x_i\Delta x_i'\Big)^{-1}\sum_{i=1}^{n}(\Delta y_i-\Delta x_i\hat{\beta}_{FD})^2
$$

Similarly, the bias adjusted estimate of the mean error variance error variance estimator from the FE method is

$$
\begin{aligned}
\hat{s}^2_{FE} =& \frac{1}{n-K}\sum_{i=1}^{n}\sum_{t=1}^{2}(\hat{u}_{it})^2 \\
=& \frac{1}{n-K}\sum_{i=1}^{n}\sum_{t=1}^{2}((y_{it}-\bar{y}_i)-(x_{it}-\bar{x}_i)\hat{\beta}_{FE})^2 \\
=& \frac{1}{n-K}\sum_{i=1}^{n}\sum_{t=1}^{2}\Big(\Big(y_{it}-\frac{y_{i1}+y_{i2}}{2}\Big)-\Big(x_{it}-\frac{x_{i1}+x_{i2}}{2}\Big)\hat{\beta}_{FE}\Big)^2 \\
=& \frac{1}{n-K}\sum_{i=1}^{n}\Big[\Big(\Big(\frac{y_{i1}-y_{i2}}{2}\Big)-\Big(\frac{x_{i1}-x_{i2}}{2}\Big)\hat{\beta}_{FE}\Big)^2+\Big(\Big(\frac{y_{i2}-y_{i1}}{2}\Big)-\Big(\frac{x_{i2}-x_{i1}}{2}\Big)\hat{\beta}_{FE}\Big)^2\Big] \\
=& \frac{1}{n-K}\sum_{i=1}^{n}\Big[\Big(\frac{-\Delta y_i+\Delta x_i\hat{\beta}_{FE}}{2}\Big)^2+\Big(\frac{\Delta y_i-\Delta x_i\hat{\beta}_{FE}}{2}\Big)^2\Big] \\
=& \frac{1}{2(n-K)}\sum_{i=1}^{n}\Big(\Delta y_i-\Delta x_i\hat{\beta}_{FE}\Big)^2 \\
=& \frac{1}{2(n-K)}\sum_{i=1}^{n}\Big(\Delta y_i-\Delta x_i\hat{\beta}_{FD}\Big)^2,
\end{aligned}
$$

where the last equality follows from the fact that $\hat{\beta}_{FE}=\hat{\beta}_{FD}$. Thus, $\hat{s}^2_{FE}=\hat{s}^2_{FD}/2$. Thus, the error variance estimate from the FE method is

$$
\begin{aligned}
\hat{V}_{\hat{\beta}_{FE}} =& \Big(\sum_{i=1}^{n}\sum_{t=1}^{2}(x_{it}-\bar{x}_i)(x_{it}-\bar{x}_i)'\Big)^{-1}\hat{s}^2_{FE} \\
=& \frac{1}{2}\Big(\sum_{i=1}^{n}\sum_{t=1}^{2}(x_{it}-\bar{x}_i)(x_{it}-\bar{x}_i)'\Big)^{-1}\hat{s}^2_{FD} \\
=& \frac{1}{2}\Big(\sum_{i=1}^{n}\frac{\Delta x_i\Delta x_i'}{4}+\frac{(-\Delta x_i)(-\Delta x_i)'}{4}\Big)^{-1}\hat{s}^2_{FD} \\
=& \frac{1}{2}\Big(\sum_{i=1}^{n}\frac{\Delta x_i\Delta x_i'}{2}\Big)^{-1}\hat{s}^2_{FD} \\
=& \Big(\sum_{i=1}^{n}\Delta x_i\Delta x_i'\Big)^{-1}\hat{s}^2_{FD} \\
=& \frac{1}{n-K}\Big(\sum_{i=1}^{n}\Delta x_i\Delta x_i'\Big)^{-1}\sum_{i=1}^{n}(\Delta y_i-\Delta x_i\hat{\beta}_{FD})^2 \\
=& \hat{V}_{\hat{\beta}_{FD}}.
\end{aligned}
$$

3. In order for $\hat{\beta}_{FD}$ to be consistent, we must assume that $\mathbb{E}[\Delta x_i\Delta u_i]=0$ (in addition to the assumptions that $(\Delta x_i,\Delta y_i), i=1,...,n$ are iid, $\mathbb{E}[(\Delta y_i)^2]<\infty$ $\mathbb{E}[||\Delta x_i||^2]<\infty$, and $\mathbb{E}[\Delta x_i\Delta x_i']$ is positive definite). Note that by the WLLN,

$$
\begin{aligned}
&\Big(\frac{1}{n}\sum_{i=1}^{n}\Delta x_i\Delta x_i'\Big)^{-1}\xrightarrow{p}\mathbb{E}[\Delta x_i\Delta x_i'] \text{ and} \\
&\Big(\frac{1}{n}\sum_{i=1}^{n}\Delta x_i\Delta u_i\Big)^{-1}\xrightarrow{p}\mathbb{E}[\Delta x_i\Delta u_i]=0.
\end{aligned}
$$

Thus, using the fact that the CMT allows us to combine these convergence results, we have that

$$\begin{aligned}
\hat{\beta}_{FD} - \beta_{FD} =& \Big(\frac{1}{n}\sum_{i=1}^{n} \Delta x_i \Delta u_i\Big)^{-1} \Big(\frac{1}{n}\sum_{i=1}^{n} \Delta x_i \Delta u_i\Big) \\
&\xrightarrow{p} \mathbb{E}[\Delta x_i \Delta x_i'] \mathbb{E}[\Delta x_i \Delta u_i] \\
&= \mathbb{E}[\Delta x_i \Delta x_i'] \times 0 \\
&= 0 \\
\implies \hat{\beta}_{FD} &\xrightarrow{p} \beta_{FD}.
\end{aligned}$$

Since $\hat{\beta}_{FD} = \hat{\beta}_{FE}$, the same result holds for $\hat{\beta}_{FE}$.

Question T2: IV Problems and Method of Moments [17 points]

1.

$$\begin{aligned}
\mathbb{E}[\boldsymbol{x}_i u_i] =& 0 \\
\implies \mathbb{E}[\boldsymbol{x}_i(y_i - \boldsymbol{x}_i'\boldsymbol{\beta})] =& 0
\end{aligned}$$

The method of moments calls for replacing population moments with sample moments. Thus,

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_i(y_i - \boldsymbol{x}_i'\hat{\boldsymbol{\beta}}) =& 0 \\
\implies \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_i y_i - \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i'\hat{\boldsymbol{\beta}} =& 0 \\
\implies \hat{\boldsymbol{\beta}} =& \Big(\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i'\Big)^{-1} \Big(\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_i y_i\Big)
\end{aligned}$$

In addition,

$$\begin{aligned}
\boldsymbol{\Omega} =& \mathbb{E}[\boldsymbol{x}_i \boldsymbol{x}_i' u_i^2] \\
=& \mathbb{E}[\boldsymbol{x}_i \boldsymbol{x}_i'(y_i - \boldsymbol{x}_i'\boldsymbol{\beta})^2].
\end{aligned}$$

Similarly, since the method of moments calls for replacing population moments with sample moments,

$$\begin{aligned}
\hat{\boldsymbol{\Omega}} =& \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i'(y_i - \boldsymbol{x}_i'\hat{\boldsymbol{\beta}})^2 \\
=& \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i' y_i^2 - \frac{2}{n}\sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i'\hat{\boldsymbol{\beta}}'\boldsymbol{x}_i y_i + \frac{1}{n}\sum_{i=1}^{n} \hat{\boldsymbol{\beta}}'\boldsymbol{x}_i \boldsymbol{x}_i'\hat{\boldsymbol{\beta}}\boldsymbol{x}_i \boldsymbol{x}_i'
\end{aligned}$$

2. According to A&P Proposition 5.13.2, smooth functions of sample moments are efficient estimators of their population counterparts. Therefore, yes, $(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\Omega}})$ are efficient estimators of $(\boldsymbol{\beta}, \boldsymbol{\Omega})$ since they are smooth functions of sample moments.

3.

$$\begin{aligned}
\mathbb{E}[\boldsymbol{x}_i u_i] =& \kappa \\
\implies \mathbb{E}[\boldsymbol{x}_i(y_i - \boldsymbol{x}_i'\boldsymbol{\beta})] =& \kappa
\end{aligned}$$

Since the method of moments calls for replacing population moments with sample moments, we have that

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{x}_i(y_i-\boldsymbol{x}_i'\hat{\boldsymbol{\beta}}) =&\kappa\\
\implies \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{x}_i y_i-\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{x}_i\boldsymbol{x}_i'\hat{\boldsymbol{\beta}} =&\kappa\\
\implies \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{x}_i\boldsymbol{x}_i'\hat{\boldsymbol{\beta}} =&\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{x}_i y_i-\kappa\\
\implies \hat{\boldsymbol{\beta}} =&\Big(\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{x}_i\boldsymbol{x}_i\Big)^{-1}\Big(\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{x}_i y_i-\kappa\Big)\\
=&\Big(\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{x}_i\boldsymbol{x}_i\Big)^{-1}\Big(\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{x}_i y_i\Big)-\Big(\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{x}_i\boldsymbol{x}_i\Big)^{-1}\kappa.
\end{aligned}$$

Since the estimator when $\mathbb{E}[\boldsymbol{x}_i u_i]=0$ is efficient, its bias is 0, i.e.

$$\mathbb{E}[\hat{\boldsymbol{\beta}}]=\mathbb{E}\Big[\Big(\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{x}_i\boldsymbol{x}_i\Big)^{-1}\Big(\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{x}_i y_i\Big)\Big]=\boldsymbol{\beta}.$$

Thus, we can find the bias in $\hat{\boldsymbol{\beta}}$ when $\mathbb{E}[\boldsymbol{x}_i u_i]=\kappa$ by comparing it to the original estimator. Thus,

$$\begin{aligned}
\mathbb{E}[\hat{\boldsymbol{\beta}}] =&\mathbb{E}\Big[\Big(\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{x}_i\boldsymbol{x}_i\Big)^{-1}\Big(\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{x}_i y_i\Big)-\Big(\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{x}_i\boldsymbol{x}_i\Big)^{-1}\kappa\Big]\\
=&\mathbb{E}\Big[\Big(\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{x}_i\boldsymbol{x}_i\Big)^{-1}\Big(\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{x}_i y_i\Big)\Big]-\mathbb{E}\Big[\Big(\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{x}_i\boldsymbol{x}_i\Big)^{-1}\kappa\Big]\\
=&\boldsymbol{\beta}-\mathbb{E}\Big[\Big(\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{x}_i\boldsymbol{x}_i\Big)^{-1}\kappa\Big]\\
\implies \text{Bias}(\hat{\boldsymbol{\beta}}) =&\mathbb{E}[\hat{\boldsymbol{\beta}}]-\boldsymbol{\beta}=-\mathbb{E}\Big[\Big(\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{x}_i\boldsymbol{x}_i\Big)^{-1}\kappa\Big].
\end{aligned}$$

4. First, we assume that there is only one endogenous variable since we are only given on instrumental variable. Let $\boldsymbol{x}_{i1}$ be the vector of exogenous variables for which $\mathbb{E}[x_{ik}u_i]=0$ for $k=1,...,K-1$, and let x be the endogenous variables for which $\mathbb{E}[xu_i]=0$ ($x=x_{iK}$). Then

$$\boldsymbol{x_i}=\begin{bmatrix}\boldsymbol{x}_{i1}\\ x\end{bmatrix}.$$

Now, let

$$\boldsymbol{Z_i}=\begin{bmatrix}\boldsymbol{x}_{i1}\\ Z\end{bmatrix},$$

and we know that $\mathbb{E}[\boldsymbol{Z_i}u_i]=0$. Then,

$$\begin{aligned}
\mathbb{E}[\boldsymbol{Z}_i u_i] =&0\\
\implies \mathbb{E}[\boldsymbol{Z}_i(y_i-\boldsymbol{x}_i'\boldsymbol{\beta})] =&0.
\end{aligned}$$

The method of moments calls for replacing population moments with sample moments. Thus,

$$
\begin{aligned}
\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{Z}_i(y_i-\boldsymbol{x}_i'\hat{\boldsymbol{\beta}}_{IV}) =&0\\
\implies \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{Z}_i y_i-\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{Z}_i\boldsymbol{x}_i'\hat{\boldsymbol{\beta}}_{IV} =&0\\
\implies \hat{\boldsymbol{\beta}}_{IV} =&\Big(\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{Z}_i\boldsymbol{x}_i'\Big)^{-1}\Big(\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{Z}_i y_i\Big).
\end{aligned}
$$

5. The first stage is essentially carried out by substituting the x endogenous variable with Z and using the $\boldsymbol{Z}_i$ vector instead of the partially endogenous $\boldsymbol{x}_i$ vector. The second stage is carried out in the part where the method of moments is performed.

6. The expected bias of the 2SLS estimator is

$$
\begin{aligned}
\text{Bias}(\hat{\boldsymbol{\beta}}_{IV}) =&\mathbb{E}[\hat{\boldsymbol{\beta}}_{IV}]-\boldsymbol{\beta}\\
=&\mathbb{E}\Big[\Big(\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{Z}_i\boldsymbol{x}_i'\Big)^{-1}\Big(\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{Z}_i y_i\Big)\Big]-\boldsymbol{\beta}
\end{aligned}
$$

The first $K-1$ entries of this $\text{Bias}(\hat{\boldsymbol{\beta}}_{IV})$ vector are clearly equal to 0 from the previous analysis. It is only the K-th entry of this vector for which we are unsure is equal to 0.

7. We know that

$$
\hat{\boldsymbol{\beta}}_{IV}\xrightarrow{p}\boldsymbol{\beta}+\frac{\sigma_u}{\sigma_x}\frac{\text{Cov}(Z,u)}{\text{Cov}(Z,x)},
$$

so the bias is $\frac{\sigma_u}{\sigma_x}\frac{\text{Cov}(Z,u)}{\text{Cov}(Z,x)}$. In general, we want $\text{Cov}(Z,x)$ to be relatively large and $\text{Cov}(Z,u)$ to be as close to 0 as possible so that the instrument bias can be close to 0. However, if $\text{Cov}(Z,x)$ is almost 0 and potentially on the same order of magnitude as $\text{Cov}(Z,u)$, which we are told is very small and positive, then the instrument bias can be significant.

8. We can only justify the use of this "weak and dusty" IV if

$$
\begin{aligned}
\frac{\sigma_u}{\sigma_x}\frac{\text{Cov}(Z,u)}{\text{Cov}(Z,x)} <&\mathbb{E}\Big[\Big(\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{x}_i\boldsymbol{x}_i\Big)^{-1}\kappa\Big]\\
\implies \text{Cov}(Z,x) >&\frac{\sigma_u}{\sigma_x}\text{Cov}(Z,u)\Big[\mathbb{E}\Big[\Big(\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{x}_i\boldsymbol{x}_i\Big)^{-1}\kappa\Big]\Big]^{-1}.
\end{aligned}
$$

Question T3: The Unknown UK [17 points]

1. It can be a problem that does introduce bias, i.e. omitted variable bias, so running the regression without this unobserved variable is not a useful approach. When we omit UK and add it in with the error term, we now have an endogeneity problem since the expected value of the error term given x_1 cannot be equal to 0. We can even calculate the exact bias since we know $Cov(UK, x_1)$.

2. (a) We usually assume that the residual distribution has 0 mean and finite variance and that $\mathbb{E}[u_i|(x)] =$

0, but we know that is no longer the case here. I do expect a bias, and I will calculate it now.

$$\begin{aligned}
y =& x_1\beta_1 + UK\tau + \epsilon \\
\implies \text{Cov}(y, x_1) =& \beta_1 Var(x_1) + \tau\text{Cov}(UK, x_1) \\
\implies \beta_{1,biased} =& \frac{\text{Cov}(y, x_1) - \tau\text{Cov}(UK, x_1)}{Var(x_1)} \\
=& \beta_{1,unbiased} - \frac{\tau\text{Cov}(UK, x_1)}{Var(x_1)} \\
=& \beta_{1,unbiased} - \frac{\sigma_{UK}}{\sigma_{x_1}}\text{Cov}(UK, x_1) \qquad (???) \\
=& \beta_{1,unbiased} - (-0.4).
\end{aligned}$$

So the expected bias is 0.4.

(b) We need to assume that we are also not omitting any variables from this regression of x_1 on UK. If we do assume that this is the case, the estimate of this regression would be

$$\frac{\text{Cov}(UK, x_1)}{\sigma_{UK}} = \frac{-0.4}{1.5} = -\frac{4}{15} \approx -0.267.$$

3. Using the β that minimizes the sum of squared residuals and following the analysis from earlier theoretical questions,

$$\begin{aligned}
\hat{\beta}_1 =& (X'X)^{-1}X'y \\
=& (X'X)^{-1}X'(X\beta + UK\tau + \epsilon) \\
=& (X'X)^{-1}X'X\beta + (X'X)^{-1}X'UK\tau + (X'X)^{-1}X'\epsilon \\
=& \beta + (X'X)^{-1}X'UK\tau + (X'X)^{-1}X'\epsilon
\end{aligned}$$

4. Therefore, the expected bias can be found with the following:

$$\mathbb{E}[\hat{\beta}_1|X] = \beta + (X'X)^{-1}\mathbb{E}[X'UK]\tau.$$

so the expected bias is $(X'X)^{-1}\mathbb{E}[X'UK]\tau$.

5. From this, it is clear that the expected bias is not equal to 0 since $\mathbb{E}[X'UK] \neq 0$ because $\text{Cov}(UK, x_1) \neq 0$.

Question T4: Linear and Non-Linear GMM [33 points]

4.1) Consider the linear regression model as a special case of GMM. [16 points]

(a)

(b)

(c)

(d)

(e)

4.2) Instrumental Variables [13 points]

(a)

(b)

(c)

(d)

4.3) Non-linear GMM Model [5 points]

(a)

(b)

(c)

(d)

Part III: Applied Questions and Real Problems [100 points]

Question A1: Squirrelnomics - IV [14 points]

(i) A valid instrument Z is a variable that affects an endogenous x_i variable and only affects the dependent variable y_i through its impact on x_i, which is known as the exclusion restriction. We also hope that there is a strong relationship between Z and x_i so that Z is more likely to be an effective instrument.

(ii)

$$\begin{aligned}
y_i =& \beta_0 + \beta_1 x_i + u_i \\
\text{Cov}(Z, y_i) =& \beta_1 \text{Cov}(Z, x_i) + \text{Cov}(Z, u_i) \\
\text{Cov}(Z, y_i) =& \beta_1 \text{Cov}(Z, x_i) && \text{(exogeneity)} \\
\implies \beta_1 =& \frac{\text{Cov}(Z, y_i)}{\text{Cov}(Z, x_i)} \\
\implies \hat{\beta}_{1,IV} =& \frac{\sum_{i=1}^{n}(Z_i - \bar{Z})(y_i - \bar{y})}{\sum_{i=1}^{n}(Z_i - \bar{Z})(x_i - \bar{x})}
\end{aligned}$$

(iii) The two conditions are $\mathbb{E}[Zu_i] = 0$ (exogeneity) and $\mathbb{E}[Zx_i] \neq 0$ (relevance).

(iv) Only the relevance condition can be tested, and this is by checking to make sure that $\text{Cov}[Z, x_i] \neq 0$. The exogeneity condition must be argued.

(v)
- A good instrument for the endogenous number of squirrels per square mile variable is the hawk population per square mile. Good instruments satisfy both exogeneity and relevance.
- A bad instrument for the endogenous number of squirrels per square mile variable is the unemployment rate in the United States. Bad elements usually fail the exogeneity condition. There are even some elements that are exogenous, but are weak because there barely satisfy the relevance condition.
- The good instrument should not have any direct impact on nut tree population because I don't think there is a relation between nut trees and hawks. In addition, it is clearly relevant because hawks are a predator of squirrels. The bad instrument would not satisfy the relevance condition since it is clear that there is no significant relationship between the squirrel population per square mile and the unemployment rate in the United States.

(vi) We have previously shown that as the number of observations becomes large,

$$\hat{\beta}_{OLS} \xrightarrow{p} \beta_1 + \frac{\sigma_u}{\sigma_x} \text{Cov}(x_i, u_i)$$

and

$$\hat{\beta}_{IV} \xrightarrow{p} \beta_1 + \frac{\sigma_u}{\sigma_x} \frac{\text{Cov}(Z,u)}{\text{Cov}(Z,x)}.$$

If we use this bad IV, it may be the case that

$$\text{Cov}(x_i,u_i) < \frac{\text{Cov}(Z_i,u_i)}{\text{Cov}(Z_i,x_i)} \implies \hat{\beta}_{OLS} \neq \hat{\beta}_{IV}.$$

Therefore, we end up with a biased estimator.

Question A2: (Identification, DiD, and RD) Online "Markets" [36 points]

(a)

- There is a high chance of various forms of omitted variable bias with this specification. It is more likely that a younger, more tech-savvy population would review the HPs online much more often than an older population. Also, there are possible regional factors and access to technology factors that are also not being accounted for in this specification. In addition, there is a high amount of heterogeneity in the types of HPs that there are.
- Suitable panel data would be to get data about the per-month rate of reviews for an HP and a per-month rate of one of its performance measures, such as average survival rate. This specification would look like a DiD equation with individual fixed effects since we should control for the different types of HPs:

  $$y_{it} = \mu_i + \beta T_i S_i + \gamma x_{it} + \epsilon_{it}$$

  where μ_i is the individual HP effect, T_i and S_i are dummy variables for treatment group and post-treatment, and x_{it} are the controls. I think that this is better than first idea since we are at least incorporating individual effects, but we still need to specify an actual research design and what these controls are.
- A way that we can implement some kind of "shock" to the system is by having our treatment group of HPs recommend to their patients that they fill out online reviews at a computer in a separate room (so that they do not feel peer-pressured) when they are done with their visits/appointments. The control group of HPs would not have this policy. In addition, we would randomly assign which HPs would go into each group. This might not be feasible in practice since some HPs might not be willing to implement this kind of policy.
- In the ideal experiment, the HPs are able to force every patient they have to fill out an online review. And also, the assignment would be random. This is not much different than what I previously said.
- A threat to identification is the patient types that each HP sees. In the ideal experiment, we would be able to put an even amount of HPs that see high-risk patients (for example, patients with cancer) and low-risk patients, but this may be hard to control for in practice and would certainly lead to some endogeneity issues. Another threat to identification is some outside factor that could influence a certain population of people to feel obliged to fill out more online reviews. This could be some type of news story that gets a lot of publicity. In this case, we are not controlling for this kind of threat, but it is probably unlikely.
- The ideal experiment is probably not feasible here since forcing patients to fill out reviews isn't really ethical, but pushing for patients do these reviews in our treatment group seems pretty good to me. We would also need to be controlling for factors like average patient risk-type, insurance coverage, types of services offered, and regional factors. Also, having equal distributions of the types of HPs in each group would be ideal if we can control this type of thing.
- Since this design follows a DiD specification, we need to make sure that there is no spillover effect into the control group. But the critical identifying assumption to this design is that the treated group is being treated properly, i.e. that pushing patients to fill out online reviews is being performed the same for each HP in the treatment group. This might be too much of a stretch unless all the treated HPs were trained properly.

- I think there needs to be more segmentation about what an HP is. The operations of health care providers can be so different from provider to provider that it isn't quite clear to me if a good research design can be carried out by just looking at all doctors and hospitals. I would recommend starting with a specific types of doctor with this research design and see how effective it is first before moving to hospitals

(b)
- The estimation equation for DiD in regression form is

$$y_{it} = \beta_0 + \beta_1 T_i + \beta_2 S + \beta_3 T_i S_i + \beta_4 x_{it} + \epsilon_{it}$$

  where T and S are indicator variables of being in the treatment group and whether the observation is gathered after treatment has occurred and β_3 is the DiD estimate in which we are interested. In addition, x_{it} are the control variables.

- The estimation equation of RD in regression form is

$$y_i = \alpha + \beta x_i + \delta D_i + u_i$$

  where D_i is an indicator if $x_i \geq x_c$ in which x_c is the threshold.

- I would suggest a DiD that follows what I explained above. To be honest, I am not sure how feasible it is given my current idea. I could probably be convinced that other research designs are more feasible, but I only have access to my own ideas as of now. Overall, I do think DiD could be feasible. It is not quite clear to me how RD would be carried out in this setting, but I am sure that I could be convinced of its feasibility.

-
  - This would be a problem because it would introduce endogeneity since now we would expect health performance metrics to improve once the HPs read what patients are saying about them online. We would not have to worry about this if we knew that an HP did not change how it practices in response to what it reads online about itself. If we also assume that the reviews are not constructive in nature or only leave comments about aspects of the HP that have nothing to do with health performance measures, then we do not have to worry about this. If we recall that for a large number of observations,

$$\hat{\beta}_{OLS} \xrightarrow{p} \beta_1 + \frac{\sigma_u}{\sigma_x} \text{Cov}(x_i, u_i),$$

    so the bias would be $\frac{\sigma_u}{\sigma_x}\text{Cov}(x_i, u_i)$ since we know that $\text{Cov}(x_i, u_i) \neq 0$.
  - I think under my DiD, there would bias to what is seen above because I do not think my ideal experimental design could be carried out well in practice.
  - My suggestion to improve my weak design is to focus on a subset of the HPs that I think I can control more. I outlined this in a previous bullet point.

Question A3: Job Market Papers [50 points]

1. My two letters are A (position 01) and E (position 05). By using the outlined rule, my papers are 1, 2, 5, and 10. (Following the rule, these were picked in the order of 10, 1, 2, 5.) For my own convenience, I have prioritized looking at papers 1, 2, and 5, and leaving out 10.

2. *Paper 1*

   (a) WHAT: He investigates how the digitization of consumer goods has affected the frequency and content of product updates. He examines if product updates are more or less frequent under digitization, and how the content of updates changes under digitization.

   (b) WHY: Economists should care about this problem because it shows that the digitization of a consumer product can effect a firm's behavior by altering the firm's product innovation incentives. In general, this is important as our society develops more digitized consumer goods, and results can indicate how a firm may choose to innovate its products.

(c) HOW: He carries out an empirical analysis of consumer and firm behavior in the context of smartphone apps via Apple. He forms a database of Apple apps and uses natural language processing and machine learning techniques to classify how big a product updates is. In addition, he develops models of app demand and app updating as part of his specification and uses NLLS to solve for model parameters. Simulation is used to construct counterfactuals.

(d) SO WHAT: He finds that changes from digitization result in an increase in the frequency of product updates of 63% to 142%. He also finds that these changes lead to an increase in the relative frequency of major updates compared to minor updates.

Paper 2

(a) WHAT: The authors look at what positions in society women may impact/limit corruption. They also address the question of whether the relationship between female participation and corruption is actually driven by women's access to corruption.

(b) WHY: Corruption can negatively impact economic outcomes such as investment and GDP per capita, especially in poor countries.

(c) HOW: They focus on the interaction of female labor force participation and government corruption. They use an IV approach for female participation and even make a methodological contribution by drawing inferences based on a specific conditional likelihood approach (CLR).

(d) SO WHAT: They provide robust evidence that women's presence in parliament leads to less corruption, while other measures of female participation in economic activities have no effect. They also show that this result extends to 17 other European countries.

Paper 5

(a) WHAT: Consumers of natural gas often use their natural gas without knowing any information about how much they are using.

(b) WHY: A lesson as to how informing people of their own behavior can lead to changes in their behavior, and in this case, it is for their own good and the good of the environment. Also, this can show the importance that informing people has on market efficiency.

(c) HOW: They randomly split customers in California into two groups in which the treated group receives weekly emails about their past, current, and projected natural gas usage over 20 months.

(d) SO WHAT: They find that informed customers reduce energy use by up to 1% compared to the control group, that the treatment effects are largest during the winter when demand for natural gas tends to be high, and that the treatment effects are observed to continue over time.

- Ranking according to the availability of information:
  (1) Paper 5
  (2) Paper 2
  (3) Paper 1

  I think Paper 5 does an excellent job with clearly outlining results in the abstract and conclusions. Paper 2 also did a pretty good job of this, and I found myself spending the most time looking through Paper 1 to learn about specific results. None of them did a bad job though.

- Ranking according to how compelling their question and research strategy was:
  (1) Paper 2
  (2) Paper 1
  (3) Paper 5

  I think Paper 2 is the most compelling to me because I personally think gender equality is a very important issue, and I found their identification method to be both easy to understand and fascinating. I do not care much for Paper 1's topic, but I thought all his modeling and construction of data was really interesting and unique. Finally, I find Paper 5's topic somewhat compelling, but there was nothing too exciting about its research strategy because it was so straight-forward. This is not necessarily a bad thing in terms of the validity of its results, but it made it less compelling.

3. *Paper 1*

   (a) He recognizes that he does that randomized treatment and control groups, so he uses simulation in which he "turns off" his main two aspects of digitization in order to construct counterfactuals. Then he compares the results of this simulation to what was observed to estimate the effect of digitization on a firm's app updating behavior.

   (b) He makes the assumption that the decision of whether to update an app on a weekly basis is made before the realization of the error term of the main regression equation, which is then estimated with NLLS. He also assumes a Markov Perfect Nash Equilibrium in his model of app updating by necessity. He assumes that multi-app developers are treat development of each app separately, and concludes that this assumption is not too detrimental. There are a few other modeling asssumptions described in the paper.

   (c) This first assumption is that the expected value of the error term given the decision to update is equal to 0. The rest of the assumptions are pretty advanced and do not have much to do with the primary structural modeling

   (d) I think his main assumption about the timing of the decision of whether to update an app is pretty fair.

   Paper 2

   - They use a straight forward linear equation for baseline specification. But then use an IV approach to establish causality since they recognize that female participation is an endogenous variable. Their instrument for this variable is an indicator if the country has a dominant language with 2 genders. They also have another IV for female participation specifically in parliament, which happens to be exposure to democratic rights. They recognize that this IV is partially weak, so they use CLR to show that it is not significantly weak.
   - They assume that this IV satisfies exclusion restriction, i.e. the instrument only affects corruption through its effect on female participation, and that it is a relevant IV.
   - These assumptions for the first IV are indicated with the following: the covariance between the indicator dealing with the dominant language and the base model error term is 0, and the covariance between the indicator dealing with the dominant language and femaile participation is significantly different than 0. In this case, it is a relatively big negative number.
   - I think that their IV assumptions are very credible since it is highly unlikely that the gender of a dominant language has any relation to corruption, and they show that there is a strong correlation between the instrument and women participation. I'm sure there are a few outliers, but this holds up in my opinion. Also, they seem to fully understand the limitations of their IVs, which is good.

   Paper 5

   - The identification strategy is observing an average treatment effect via randomized assignments of the treatment and control groups, which removes any type of selection bias. This is often quoted as the gold standard, and they were able to achieve it here. Their specification also accounts for period-specific effects.
   - The key assumption is that they actually have randomly placed subjects in treatment and control groups.
   - This assumption translates to the fact that the estimator for τ that they find truly represents the ATE. An outside concern I have is if these results on the residents of California are generalization to the residents of other states. Perhaps the political affiliation of people in different states may change how the react to information about their natural gas usage
   - I think their main assumption is pretty credible because they show that it is valid by showing that the groups are not statistically different.

4. *Paper 1*

- He recognizes that price changes can be made independently from app updates, so he uses 2SLS approach to exogenize this endogeneity. But besides this, I don't think he used an alternative identification strategy. However, he attempted to answer multiple questions and utilized many techniques outside econometrics to reach his results.
- My main concern is about compounding modeling error. Because he used so many different models and pre-processing techniques, I worry if there is bias in the results. However, I do not have much of a recommendation of how this could be avoided, but I hope he recognizes the limitations of his results.

Paper 2

- No, there is no alternative identification strategy. However, they do implement an IV that checks for the conditional validity of their IVs, which is interesting and robust.
- I think they were pretty thorough with recognizing their limitations, but I suppose I am still concerned about the introduction of bias because of their use of multiple IVs. They mention this, but I would like more discussion.

Paper 5

- They also use a more traditional DiD approach and find almost identical results.
- I didn't actually have any huge concerns, but the DiD identification makes their results more robust in my opinion. I would still like more of a discussion about how their results generalize to other states.

5. Ranking by the most convincing main identification strategy:

   (1) Paper 5

   (2) Paper 2

   (3) Paper 1

   I think Paper 5 clearly had the best strategy since it was able to achieve the gold strategy. The others had fair strategies too, but Paper 5 stands out because it had this gold standard and had an alternative identification strategy.

6. Yes, I would probably invite Paper 5 for a job talk because of how thorough and convincing its conclusions are. This is not meant to offend the authors of the other two papers though.

Part IV: Learning and Coding Exercises [150 points]

Part IV: Learning and Coding Exercises

Andrew

Thu Dec 07 20:54:58 2017

Contents


```r
# Packages. ----
library("dplyr")
library("stringr")
library("tidyr")
# library("readr")
library("ggplot2")
# library("lubridate")
theme_set(theme_minimal())
# theme_set(hrbrthemes::theme_ipsum_rc())
```

```r
# For printing pretty tables.
# printr package simply converts any data frame to a nice looking format.
# library("printr")
library("stargazer")
```

Question P1a: Job Training in R

1) The population of interest is married people, and preferably not married people who are very old. Data needs to be collected in a location and environment in which people are unafraid to admit to having an affair. I would suggest packaging a survey that gaurantees a small financial incentive with a timeshare vacation advertisement. This is because I believe that timeshare vacations are usually taken by married couples who may have kids, and there is not very much selection bias in who would take these types of surveys. Of course, we might see that richer people would respond to the survey more than poor people since richer people tend to be able to afford a timeshare vacation more easily, but there is no obvious reason to me to believe that rich people have more or less affairs than poor people. Of course, this would need to be confirmed through some outside research.

```r
load("affairs.RData")
```

2a) A Summary Table

```r
stargazer(data, type = "text")
```

```
##
## ===============================================
## Statistic  N    Mean    St. Dev.  Min    Max
## -----------------------------------------------
## id        601 1,059.722 914.905    4    9,029
## male      601   0.476    0.500     0      1
## age       601  32.488    9.289   17.500 57.000
## yrsmarr   601   8.178    5.571   0.125  15.000
## kids      601   0.715    0.452     0      1
## relig     601   3.116    1.168     1      5
## educ      601  16.166    2.403     9      20
## occup     601   4.195    1.819     1      7
## ratemarr  601   3.932    1.103     1      5
## naffairs  601   1.456    3.299     0      12
## affair    601   0.250    0.433     0      1
## vryhap    601   0.386    0.487     0      1
## hapavg    601   0.323    0.468     0      1
## avgmarr   601   0.155    0.362     0      1
## unhap     601   0.110    0.313     0      1
## vryrel    601   0.116    0.321     0      1
## smerel    601   0.316    0.465     0      1
## slghtrel  601   0.215    0.411     0      1
## notrel    601   0.273    0.446     0      1
## -----------------------------------------------
```

b) How many men and women are there?

```
cnt_male <- sum(data$male)
cnt_male
```

```
## [1] 286
```

```
cnt_female <- nrow(data) - cnt_male
cnt_female
```

```
## [1] 315
```

There are 286 men and 315 women.

c) How many people have kids?

```
cnt_kids <- sum(data$kids)
cnt_kids
```

```
## [1] 430
```

There are 430 people with kids.

d) What is the average age of respondents? What is the average education of women?

```
avg_age <- mean(data$age)
avg_age
```

```
## [1] 32.48752
```

```
avg_educ_female <-
  data %>% filter(male == 0) %>% summarise(avg_educ = mean(educ)) %>% pull(avg_educ)
avg_educ_female
```

```
## [1] 15.25714
```

32.4875208 is the average age of respondents. 15.2571429 is the average education of women.

e) Create a barplot of the number of people by attitudes towards religion.

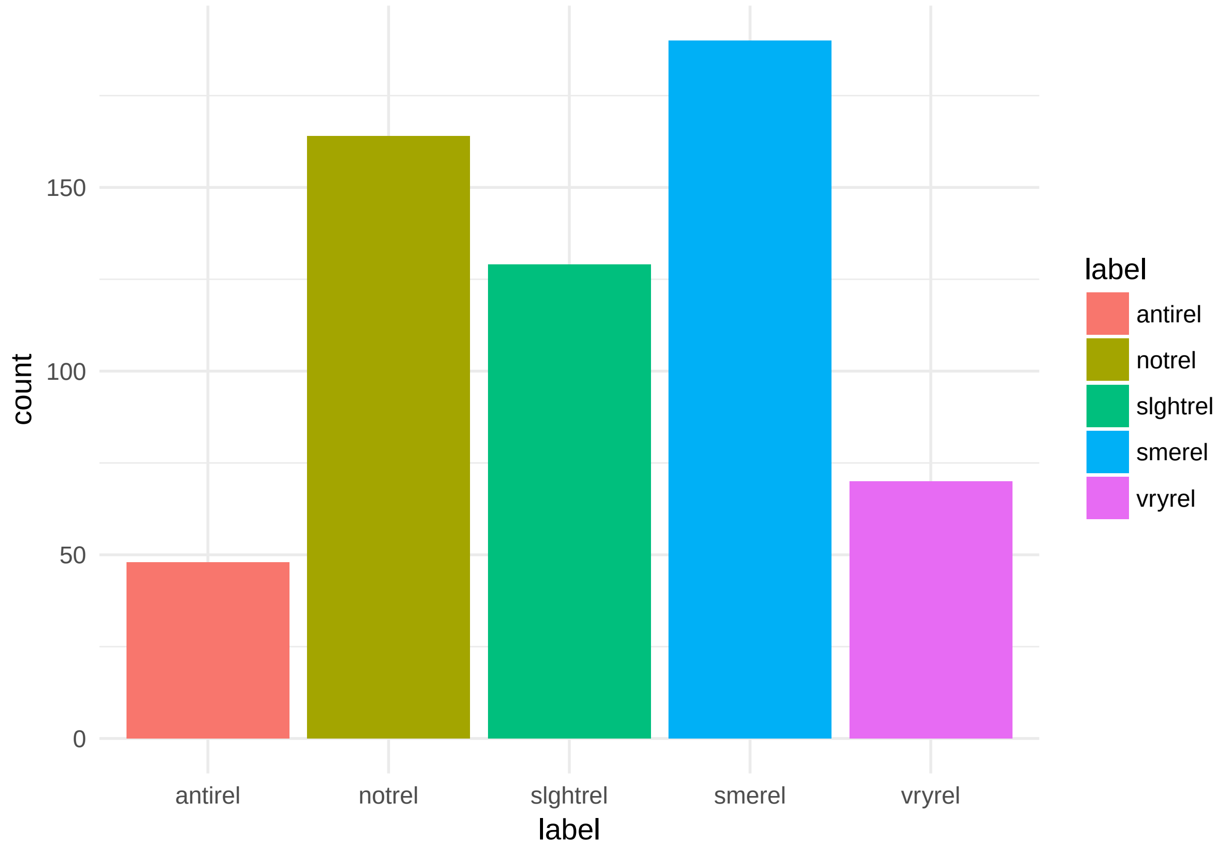
```
# Note that relig variable (with values from 2 to 5) already contains the same info.
# notrel = 2, slghtrel = 3, smerel = 4, veryrel = 5
# Note that clearning the data like this is not totally necessary.
#data$antirel = as.list(data[,"relig"==1])
relig_tidy <-
  data %>%
  as_tibble() %>%
  select(relig, ends_with("rel")) %>%
  mutate(antirel = ifelse(relig == 1, 1, 0)) %>%
  rename(num = relig) %>%
  gather(label, bool, -num) %>%
  filter(bool != 0) %>%
```

```r
  select(-bool) %>%
  mutate_at(vars(label), funs(as.factor))

# Just inspecting the data...
relig_tidy %>%
  group_by(label) %>%
  summarise_at(vars(num),
               funs(
                 cnt = n(),
                 mean,
                 median,
                 q1 = quantile(., 0.25),
                 q3 = quantile(., 0.75)
               ))
```

```
## # A tibble: 5 x 6
##      label   cnt  mean median    q1    q3
##     <fctr> <int> <dbl>  <dbl> <dbl> <dbl>
## 1  antirel    48     1      1     1     1
## 2   notrel   164     2      2     2     2
## 3 slghtrel   129     3      3     3     3
## 4   smerel   190     4      4     4     4
## 5   vryrel    70     5      5     5     5
```

```r
viz_relig <-
  relig_tidy %>%
  ggplot() +
  geom_bar(aes(x = label, fill = label))
viz_relig
```



f) Create a bar plot of the number of people by happiness in marriage.

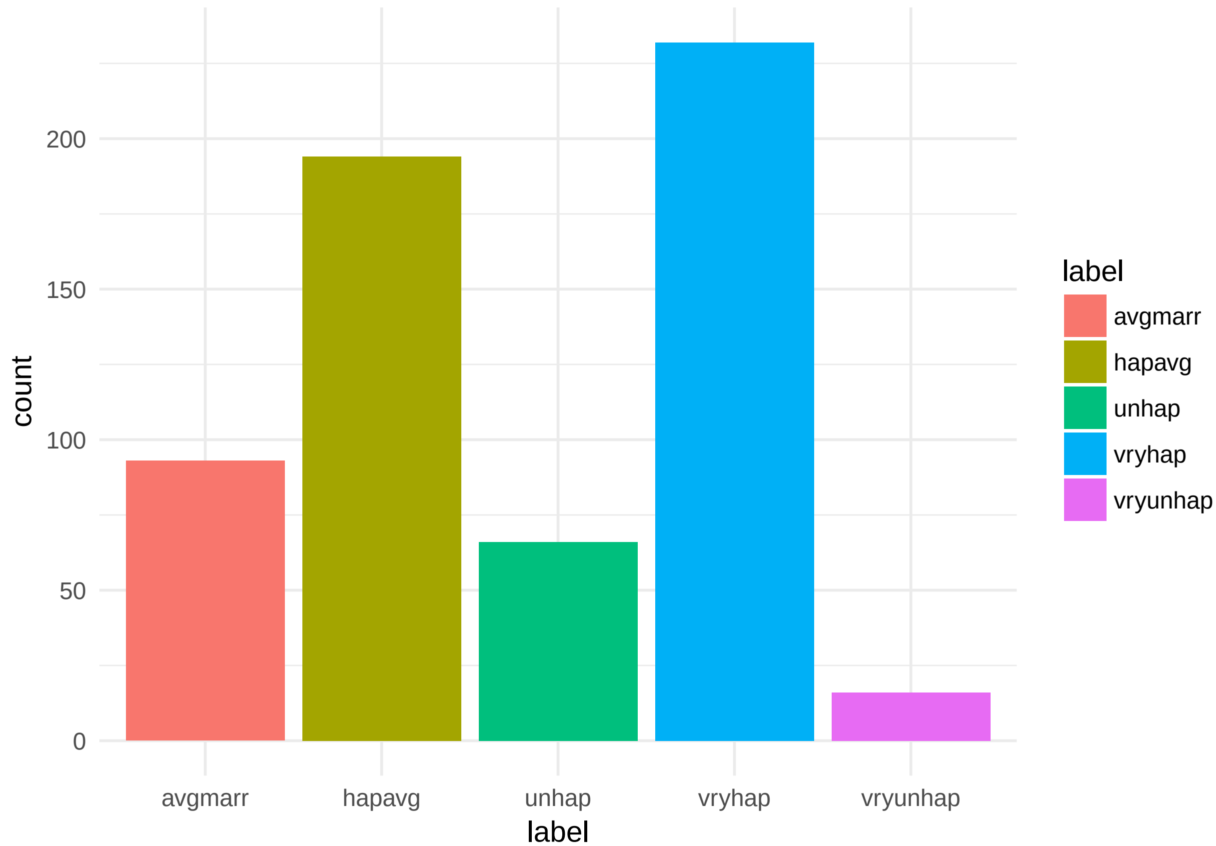
```
# Note that ratemarr variable (with values from 2 to 5) already contains the same info.
# unhap = 2, avgmarr = 3, hapavg = 4, vryhap = 5
# This is identical code to that used for relig_tidy.
hap_tidy <-
  data %>%
  as_tibble() %>%
  filter(yrsmarr > 0) %>%
  select(ratemarr, contains("hap"), avgmarr) %>%
  mutate(vryunhap = ifelse(ratemarr == 1, 1, 0)) %>%
  rename(num = ratemarr) %>%
  gather(label, bool, -num) %>%
  filter(bool != 0) %>%
  select(-bool) %>%
  mutate_at(vars(label), funs(as.factor))

hap_tidy %>%
  group_by(label) %>%
  summarise_at(vars(num),
               funs(
                 cnt = n(),
                 mean,
                 median,
```

```
                 q1 = quantile(., 0.25),
                 q3 = quantile(., 0.75)
               ))
```

```
## # A tibble: 5 x 6
##       label   cnt  mean median    q1    q3
##      <fctr> <int> <dbl>  <dbl> <dbl> <dbl>
## 1  avgmarr    93     3      3     3     3
## 2   hapavg   194     4      4     4     4
## 3    unhap    66     2      2     2     2
## 4   vryhap   232     5      5     5     5
## 5 vryunhap    16     1      1     1     1
```

```
viz_hap <-
  hap_tidy %>%
  ggplot() +
  geom_bar(aes(x = label, fill = label))
viz_hap
```



Q P1a: Job Training in R

1) Clean the environment

```
rm(list = ls())
```

2) Set your working directory and load the ggplot2 package.

```
setwd(getwd())
# This library is already loaded, but doing it again doesn't hurt.
library("ggplot2")
```

3) Load the “jtrain2.RData” dataset.

```
load("jtrain2.RData")
```

4) Describe the data.

How can you check that the data was loaded properly?

Multiple commands can be used, including `str()`, `head()`, `tail()`, and `summary()`.

```
str(data)
```

```
## 'data.frame':    445 obs. of  19 variables:
##  $ train   : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ age     : int  37 22 30 27 33 22 23 32 22 33 ...
##  $ educ    : int  11 9 12 11 8 9 12 11 16 12 ...
##  $ black   : int  1 0 1 1 1 1 1 1 1 0 ...
##  $ hisp    : int  0 1 0 0 0 0 0 0 0 0 ...
##  $ married : int  1 0 0 0 0 0 0 0 0 1 ...
##  $ nodegree: int  1 1 0 1 1 1 0 1 0 0 ...
##  $ mosinex : int  13 13 13 13 13 13 6 6 14 13 ...
##  $ re74    : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ re75    : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ re78    : num  9.93 3.6 24.91 7.51 0.29 ...
##  $ unem74  : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ unem75  : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ unem78  : int  0 0 0 0 0 0 1 0 0 0 ...
##  $ lre74   : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ lre75   : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ lre78   : num  2.3 1.28 3.22 2.02 -1.24 ...
##  $ agesq   : int  1369 484 900 729 1089 484 529 1024 484 1089 ...
##  $ mostrn  : int  13 13 13 13 13 13 6 6 14 13 ...
##  - attr(*, "datalabel")= chr ""
##  - attr(*, "time.stamp")= chr "25 Jun 2011 23:03"
##  - attr(*, "formats")= chr  "%9.0g" "%9.0g" "%9.0g" "%9.0g" ...
##  - attr(*, "types")= int  251 251 251 251 251 251 251 251 254 254 ...
##  - attr(*, "val.labels")= chr  "" "" "" "" ...
```

```
## - attr(*, "var.labels")= chr  "=1 if assigned to job training" "age in 1977" "years of education" "=
## - attr(*, "version")= int 10
```

```r
head(data)
```

```
##   train age educ black hisp married nodegree mosinex re74 re75     re78
## 1     1  37   11     1    0       1        1      13    0    0  9.93005
## 2     1  22    9     0    1       0        1      13    0    0  3.59589
## 3     1  30   12     1    0       0        0      13    0    0 24.90950
## 4     1  27   11     1    0       0        1      13    0    0  7.50615
## 5     1  33    8     1    0       0        1      13    0    0  0.28979
## 6     1  22    9     1    0       0        1      13    0    0  4.05649
##   unem74 unem75 unem78 lre74 lre75      lre78 agesq mostrn
## 1      1      1      0     0     0  2.295566  1369     13
## 2      1      1      0     0     0  1.279792   484     13
## 3      1      1      0     0     0  3.215249   900     13
## 4      1      1      0     0     0  2.015723   729     13
## 5      1      1      0     0     0 -1.238599  1089     13
## 6      1      1      0     0     0  1.400318   484     13
```

```r
tail(data)
```

```
##     train age educ black hisp married nodegree mosinex      re74    re75
## 440     0  44    9     1    0       1        1      21 12.260800 10.8572
## 441     0  21    9     1    0       0        1      23 31.886402 12.3572
## 442     0  28   11     1    0       0        1      24 17.491499 13.3713
## 443     0  29    9     0    1       0        1      23  9.594309 16.3412
## 444     0  25    9     1    0       1        1      22 24.731600 16.9466
## 445     0  22   10     0    0       1        1      22 25.720901 23.0320
##         re78 unem74 unem75 unem78     lre74    lre75    lre78 agesq mostrn
## 440 12.35930      0      0      0 2.506407 2.384828 2.514409  1936      0
## 441  0.00000      0      0      1 3.462180 2.514239 0.000000   441      0
## 442  0.00000      0      0      1 2.861715 2.593111 0.000000   784      0
## 443 16.90030      0      0      0 2.261170 2.793689 2.827332   841      0
## 444  7.34396      0      0      0 3.208082 2.830067 1.993878   625      0
## 445  5.44880      0      0      0 3.247304 3.136885 1.695395   484      0
```

```r
summary(data)
```

```
##      train             age             educ          black       
##  Min.   :0.0000   Min.   :17.00   Min.   : 3.0   Min.   :0.0000  
##  1st Qu.:0.0000   1st Qu.:20.00   1st Qu.: 9.0   1st Qu.:1.0000  
##  Median :0.0000   Median :24.00   Median :10.0   Median :1.0000  
##  Mean   :0.4157   Mean   :25.37   Mean   :10.2   Mean   :0.8337  
##  3rd Qu.:1.0000   3rd Qu.:28.00   3rd Qu.:11.0   3rd Qu.:1.0000  
##  Max.   :1.0000   Max.   :55.00   Max.   :16.0   Max.   :1.0000  
##       hisp            married          nodegree        mosinex     
##  Min.   :0.00000   Min.   :0.0000   Min.   :0.000   Min.   : 5.00  
##  1st Qu.:0.00000   1st Qu.:0.0000   1st Qu.:1.000   1st Qu.:14.00  
##  Median :0.00000   Median :0.0000   Median :1.000   Median :21.00  
##  Mean   :0.08764   Mean   :0.1685   Mean   :0.782   Mean   :18.12  
##  3rd Qu.:0.00000   3rd Qu.:0.0000   3rd Qu.:1.000   3rd Qu.:23.00  
##  Max.   :1.00000   Max.   :1.0000   Max.   :1.000   Max.   :24.00  
##       re74              re75             re78            unem74      
##  Min.   : 0.0000   Min.   : 0.000   Min.   : 0.000   Min.   :0.0000  
##  1st Qu.: 0.0000   1st Qu.: 0.000   1st Qu.: 0.000   1st Qu.:0.0000  
```

```
##  Median : 0.0000   Median : 0.000   Median : 3.702   Median :1.0000
##  Mean   : 2.1023   Mean   : 1.377   Mean   : 5.301   Mean   :0.7326
##  3rd Qu.: 0.8244   3rd Qu.: 1.221   3rd Qu.: 8.125   3rd Qu.:1.0000
##  Max.   :39.5707   Max.   :25.142   Max.   :60.308   Max.   :1.0000
##      unem75           unem78            lre74              lre75
##  Min.   :0.0000   Min.   :0.0000   Min.   :-0.8093   Min.   :-2.5991
##  1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.: 0.0000   1st Qu.: 0.0000
##  Median :1.0000   Median :0.0000   Median : 0.0000   Median : 0.0000
##  Mean   :0.6494   Mean   :0.3079   Mean   : 0.4198   Mean   : 0.2771
##  3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.: 0.0000   3rd Qu.: 0.1995
##  Max.   :1.0000   Max.   :1.0000   Max.   : 3.6781   Max.   : 3.2245
##      lre78            agesq          mostrn
##  Min.   :-3.107   Min.   : 289   Min.   : 0.000
##  1st Qu.: 0.000   1st Qu.: 400   1st Qu.: 0.000
##  Median : 1.309   Median : 576   Median : 0.000
##  Mean   : 1.136   Mean   : 694   Mean   : 7.688
##  3rd Qu.: 2.095   3rd Qu.: 784   3rd Qu.:15.000
##  Max.   : 4.099   Max.   :3025   Max.   :24.000
```

Generate a table with summary statistics.

```r
stargazer(data, type = "text")
```

```
##
## ============================================
## Statistic  N   Mean   St. Dev.  Min    Max
## --------------------------------------------
## train     445  0.416   0.493     0      1
## age       445 25.371   7.100     17     55
## educ      445 10.196   1.792     3      16
## black     445  0.834   0.373     0      1
## hisp      445  0.088   0.283     0      1
## married   445  0.169   0.375     0      1
## nodegree  445  0.782   0.413     0      1
## mosinex   445 18.124   5.312     5      24
## re74      445  2.102   5.364   0.000  39.571
## re75      445  1.377   3.151   0.000  25.142
## re78      445  5.301   6.631   0.000  60.308
## unem74    445  0.733   0.443     0      1
## unem75    445  0.649   0.478     0      1
## unem78    445  0.308   0.462     0      1
## lre74     445  0.420   0.886   -0.809 3.678
## lre75     445  0.277   0.797   -2.599 3.225
## lre78     445  1.136   1.136   -3.107 4.099
## agesq     445 693.978 429.782   289   3,025
## mostrn    445  7.688   9.656     0     24
## --------------------------------------------
```

Which command can you use to figure out how many people in the sample particpated in the job training program?

The `sum()` function can be used.

```r
cnt_train <- sum(data$train)
cnt_train
```

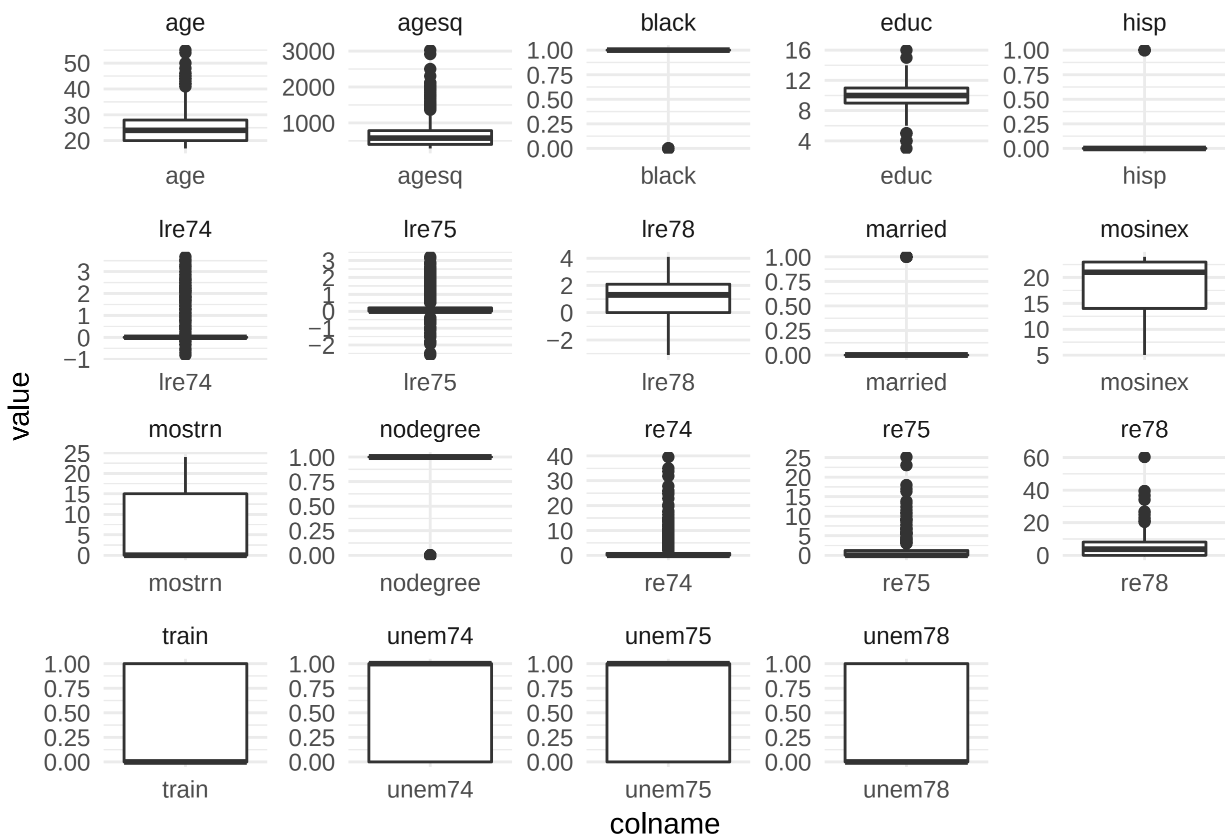
```
## [1] 185
```

185 people participated in the job training program.

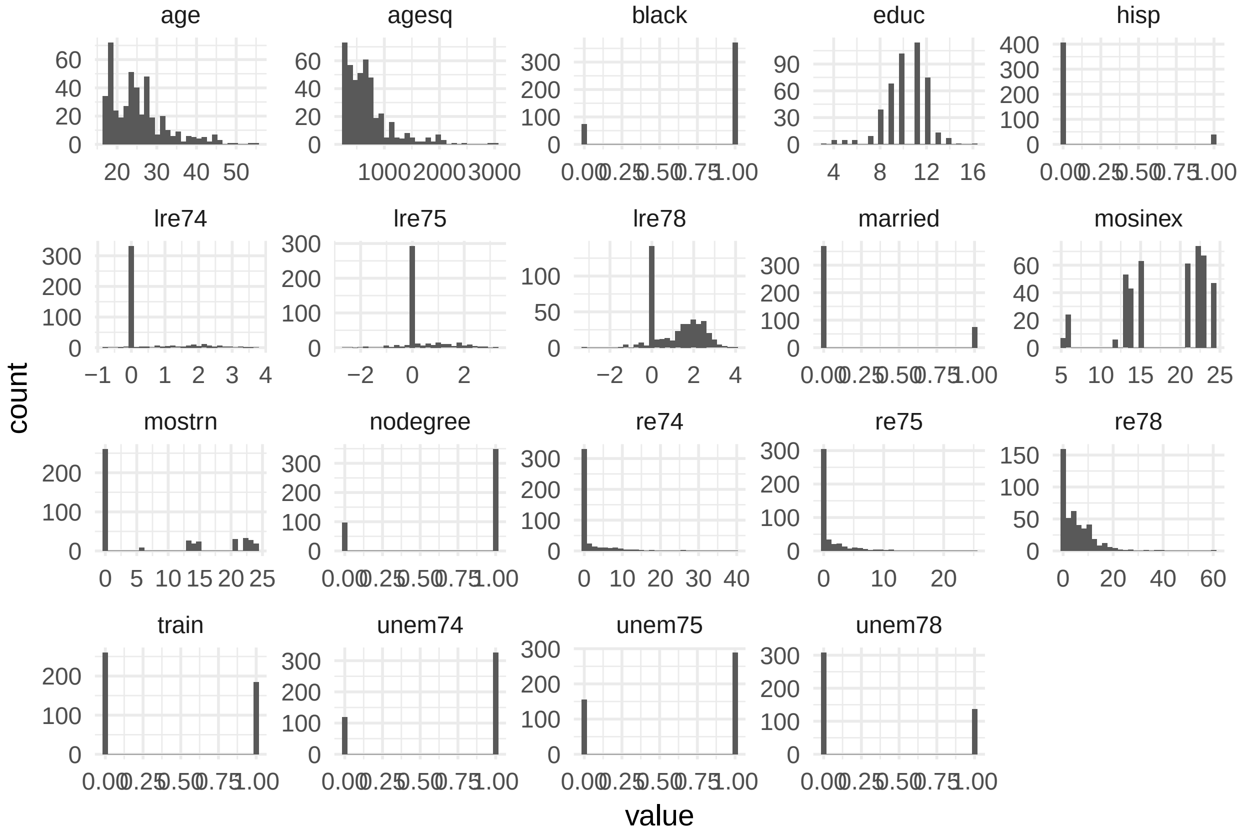
How to detect outliers?

Use ggplot2 to visualize the data (with `geom_boxplot()`, `geom_histogram()`, or another appropriate function).

```r
data %>%
  # mutate_all(scale) %>%
  gather(colname, value) %>%
  ggplot() +
  geom_boxplot(aes(x = colname, y = value)) +
  facet_wrap( ~ colname, scales = "free")
```

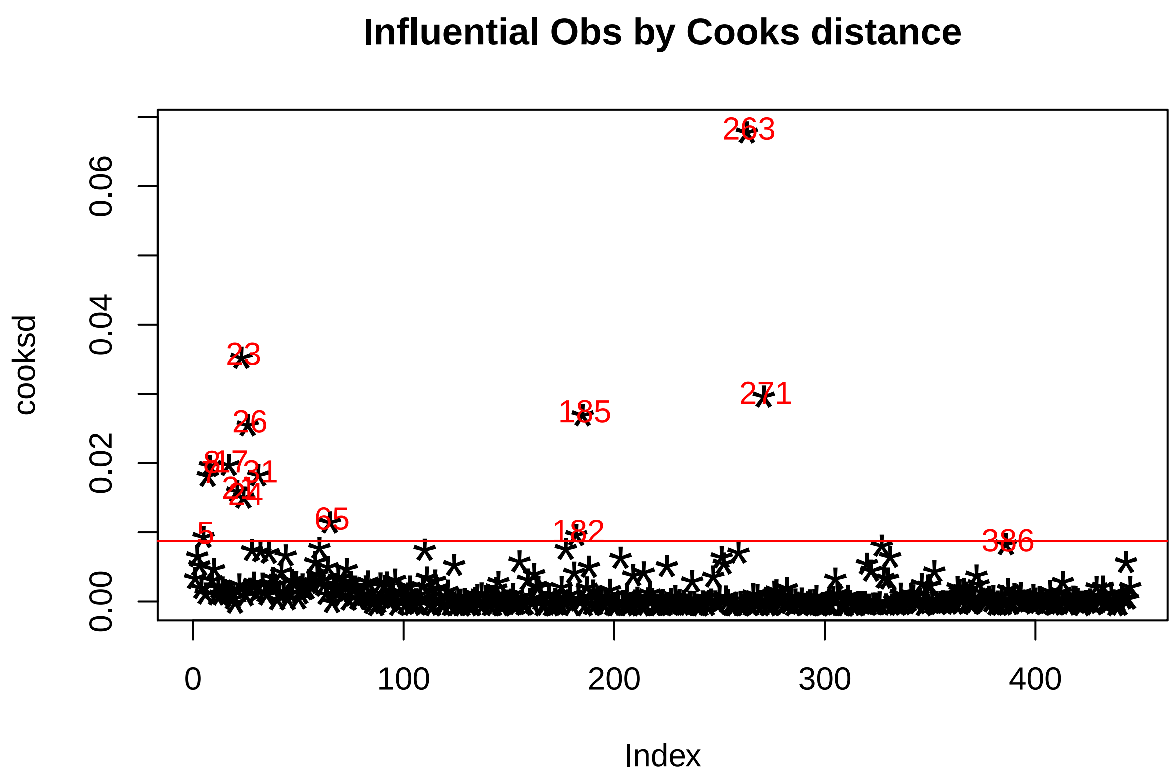



```r
data %>%
  # mutate_all(scale) %>%
  gather(colname, value) %>%
  ggplot() +
  geom_histogram(aes(x = value)) +
  facet_wrap( ~ colname, scales = "free")
```




```r
# Alternatively, use statistical measures (combined with some plotting).
# (Reference: http://r-statistics.co/Outlier-Treatment-With-R.html)

# Using "Multivariate Model Approach" with Cook's distance.
lm_0 <- lm(train ~ ., data = data)
cooksd <- cooks.distance(lm_0)
plot(cooksd,
     pch = "*",
     cex = 2,
     main = "Influential Obs by Cooks distance")
abline(h = 4 * mean(cooksd, na.rm = T), col = "red")
text(
  x = 1:length(cooksd) + 1,
  y = cooksd,
  labels = ifelse(cooksd > 4 * mean(cooksd, na.rm = T), names(cooksd), ""),
  col = "red"
)
```




```
influential <-
  as.numeric(names(cooksd)[(cooksd > 4 * mean(cooksd, na.rm = T))])
head(data[influential, ])
```

```
##    train age educ black hisp married nodegree mosinex re74 re75     re78
## 5      1  33    8     1    0       0        1      13    0    0  0.28979
## 7      1  23   12     1    0       0        0       6    0    0  0.00000
## 8      1  32   11     1    0       0        1       6    0    0  8.47216
## 17     1  27   13     1    0       0        0       6    0    0 14.58190
## 21     1  23   11     1    0       0        1       6    0    0  0.00000
## 23     1  38    9     0    0       0        1       6    0    0  6.40895
##    unem74 unem75 unem78 lre74 lre75      lre78 agesq mostrn
## 5       1      1      0     0     0 -1.238599  1089     13
## 7       1      1      1     0     0  0.000000   529      6
## 8       1      1      0     0     0  2.136786  1024      6
## 17      1      1      0     0     0  2.679781   729      6
## 21      1      1      1     0     0  0.000000   529      6
## 23      1      1      0     0     0  1.857695  1444      6
```

It is hard to point out any outliers from the boxplots and historgrams, but the Cook's distance shows us that there about a dozen outliers.

```
# Using "Outlier's Test" approach.
# car::outlierTest(lm_0)

# Using "outliers package" approach.
# outliers::scores(data, type = "chisq", prob = 0.95)
```

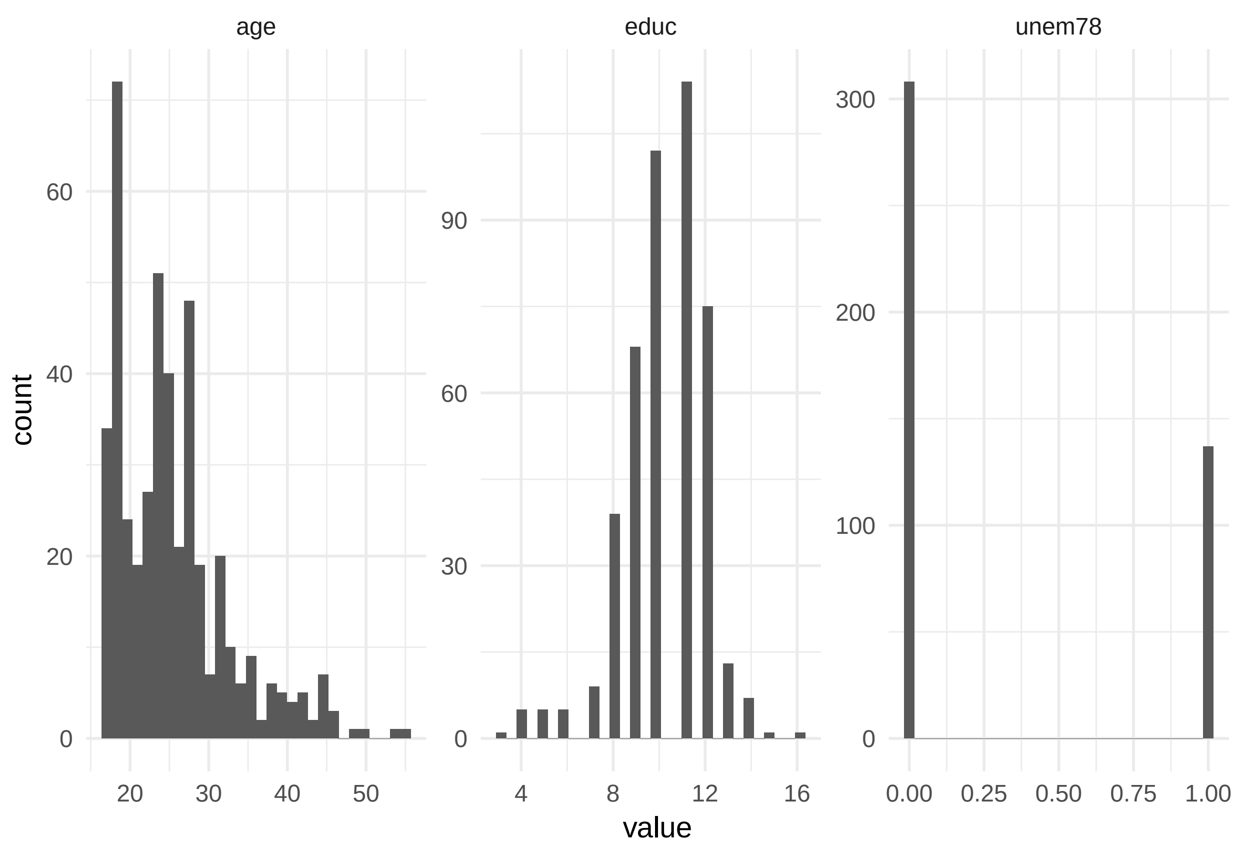
```r
# outliers::scores(data, type = "z", prob = 0.95)
# outliers::scores(data, type = "t", prob = 0.95)

# The row-column values with FALSE are the outliers that should be ommitted.
```

How can you analyze the distribution of `unem78`, `age`, and `educ`?

Use ggplot2 again (with `geom_histogram()` or another appropriate function).

```r
data %>%
  select(unem78, age, educ) %>%
  gather(colname, value) %>%
  ggplot() +
  geom_histogram(aes(x = value)) +
  facet_wrap( ~ colname, scales = "free")
```



5) Run a linear regression model.

```r
fmla_p1 <-
  as.formula("unem78 ~ train + unem74 + unem75 + age + educ + black + hisp + married")
lm_p1 <- lm(fmla_p1, data = data)
summary(lm_p1)
```

```
##
```

```
## Call:
## lm(formula = fmla_p1, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.4106 -0.3546 -0.2428  0.5908  0.9709
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.632e-01  1.761e-01   0.927   0.3546
## train       -1.117e-01  4.431e-02  -2.521   0.0121 *
## unem74       3.869e-02  7.160e-02   0.540   0.5892
## unem75       1.596e-02  6.673e-02   0.239   0.8111
## age          4.332e-05  3.155e-03   0.014   0.9891
## educ         1.442e-04  1.237e-02   0.012   0.9907
## black        1.888e-01  8.134e-02   2.322   0.0207 *
## hisp        -3.770e-02  1.087e-01  -0.347   0.7289
## married     -2.544e-02  5.967e-02  -0.426   0.6701
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4554 on 436 degrees of freedom
## Multiple R-squared:  0.0462, Adjusted R-squared:  0.0287
## F-statistic:  2.64 on 8 and 436 DF,  p-value: 0.007796
```

```r
stargazer(lm_p1, type = "text", omit.stat = c("ser"))
```

```
##
## ========================================
##                  Dependent variable:
##              ---------------------------
##                        unem78
## ----------------------------------------
## train                 -0.112**
##                        (0.044)
##
## unem74                  0.039
##                        (0.072)
##
## unem75                  0.016
##                        (0.067)
##
## age                    0.00004
##                        (0.003)
##
## educ                   0.0001
##                        (0.012)
##
## black                  0.189**
##                        (0.081)
##
## hisp                   -0.038
##                        (0.109)
##
## married                -0.025
```

```
##                         (0.060)
##
## Constant                 0.163
##                         (0.176)
##
## ----------------------------------------
## Observations              445
## R2                       0.046
## Adjusted R2              0.029
## F Statistic    2.640*** (df = 8; 436)
## ========================================
## Note:        *p<0.1; **p<0.05; ***p<0.01
```

6) Interpret the `train` coefficient. Did you get the same result?

The interpretation for the output in table 1 is that participation in the training program has a significant negative impact on unemployment probabilities in 1978, i.e. a positive effect on employment probabilities in 1978. No, I didn't get the exactly the same result because it seems to me that table 1 has filtered the data to 300 observsatvions. (as evident from the degrees of freedom), which is less than the original 445 rows. Nevertheless, the coefficient for "train" is nearly the same as that for the single variable model. In addition, the R^2 and adjusted R^2 values for the full model are better than those for the single variable model.

```r
fmla_p1b <- as.formula("unem78 ~ train")
lm_p1b <- lm(fmla_p1b, data = data)
summary(lm_p1b)
```

```
##
## Call:
## lm(formula = fmla_p1b, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.3538 -0.3538 -0.2432  0.6462  0.7568
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.35385    0.02849  12.419   <2e-16 ***
## train       -0.11060    0.04419  -2.503   0.0127 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4594 on 443 degrees of freedom
## Multiple R-squared:  0.01394,    Adjusted R-squared:  0.01172
## F-statistic: 6.265 on 1 and 443 DF,  p-value: 0.01267
```

```r
stargazer(lm_p1b, type = "text", omit.stat = c("ser"))
```

```
##
## ========================================
##                  Dependent variable:
##               ---------------------------
##                         unem78
## ----------------------------------------
## train                  -0.111**
##                         (0.044)
```

```
##
## Constant              0.354***
##                        (0.028)
##
## ----------------------------------------
## Observations             445
## R2                      0.014
## Adjusted R2             0.012
## F Statistic     6.265** (df = 1; 443)
## ========================================
## Note:        *p<0.1; **p<0.05; ***p<0.01
```

```
# Should convert the response variable to a factor
# (in compliance with best practices), even if it binary.
# Nevertheless, the result doesn't change.
glm_p1b <-
  glm(fmla_p1b,
      data = data,
      # data = mutate_at(data, vars(unem78), funs(as.factor)),
      family = "binomial")
summary(glm_p1b)
```

```
##
## Call:
## glm(formula = fmla_p1b, family = "binomial", data = data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.9346  -0.9346  -0.7466   1.4414   1.6815
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -0.6022     0.1297  -4.643 3.44e-06 ***
## train        -0.5328     0.2149  -2.479   0.0132 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 549.47  on 444  degrees of freedom
## Residual deviance: 543.17  on 443  degrees of freedom
## AIC: 547.17
##
## Number of Fisher Scoring iterations: 4
```

```
stargazer(glm_p1b, type = "text", omit.stat = c("ser"))
```

```
##
## =============================================
##                       Dependent variable:
##                   ---------------------------
##                             unem78
## ---------------------------------------------
## train                      -0.533**
##                             (0.215)
```

```
##
## Constant                     -0.602***
##                               (0.130)
##
## ---------------------------------------------
## Observations                    445
## Log Likelihood               -271.583
## Akaike Inf. Crit.             547.166
## =============================================
## Note:             *p<0.1; **p<0.05; ***p<0.01
```

7) Run a probit regression.

I can just use an internet search to find out how to do a probit regression. (Reference: http://r-statistics.co/Probit-Regression-With-R.html)

```
probit_1 <-
  glm(fmla_p1,
      data = mutate_at(data, vars(unem78), funs(as.factor)),
      family = binomial(link = "probit"))
summary(probit_1)
```

```
##
## Call:
## glm(formula = fmla_p1, family = binomial(link = "probit"), data = mutate_at(data,
##     vars(unem78), funs(as.factor)))
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.0609  -0.9303  -0.7353   1.3236   2.2690
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.0103286  0.5337665  -1.893   0.0584 .
## train       -0.3365876  0.1315169  -2.559   0.0105 *
## unem74       0.1060920  0.2116999   0.501   0.6163
## unem75       0.0636098  0.1959552   0.325   0.7455
## age          0.0006753  0.0091777   0.074   0.9413
## educ        -0.0018905  0.0363625  -0.052   0.9585
## black        0.6336700  0.2744517   2.309   0.0210 *
## hisp        -0.1649211  0.3772195  -0.437   0.6620
## married     -0.0777715  0.1775290  -0.438   0.6613
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 549.47  on 444  degrees of freedom
## Residual deviance: 526.63  on 436  degrees of freedom
## AIC: 544.63
##
## Number of Fisher Scoring iterations: 4
```

Question P2a: Mock Monte Carlo

1) Generate a 5D MVN. 5,000 observations, with the given bilaterial correlations.

See the custom function below.

a) Compute the sample covariances for all pairs and the sample variances.

```
generate_stuff <- function(n = 1, seed = 42) {
  # This is the mostly same setup as in Assignment 2.
  # n <- 1000
  mus <- c(0, 0, 0, 0, 0)
  sigmas <-
    matrix(c(
      1.00, 0.20, 0.10, 0.35, 0.00,
      0.20, 1.00, 0.00, 0.40, 0.00,
      0.10, 0.00, 1.00, 0.00, 0.40,
      0.35, 0.40, 0.00, 1.00, 0.60,
      0.00, 0.00, 0.40, 0.60, 1.00
    ), ncol = 5)

  set.seed(42)
  m_0 <- MASS::mvrnorm(n, mu = mus, Sigma = sigmas, empirical = FALSE)
  m <- cbind(m_0, rep(1, n)) %>% as_tibble()
  varnames <- c("x1", "x2", "x3", "z1", "z2")
  colnames(m) <- c(varnames, "one")
  head(m)
  summary(m)

  # Now it's different from Assignment 2.
  eps1 <- rnorm(n, mean = 0, sd = 1)
  eps2 <- rnbinom(n, size = 2, prob = 0.6)
  eps3 <- 0.15 * m$x1 + eps1

  rhs_constant <- with(m, 0.2 * x1 + 2 * x2 + 0.7 * x3)
  y1 <- rhs_constant + eps1
  y2 <- rhs_constant + eps2
  y3 <- rhs_constant + eps3

  data <- cbind(m, eps1, eps2, eps3, y1, y2, y3) %>% as_tibble()

  varnamesy = c(varnames, "y1", "y2", "y3")
  vars <- sapply(sapply(data[, varnamesy], var), round, 4)
  covs <- round(cov(data[, varnamesy]), 4)

  list(vars = vars, covs = covs, data = data)
}

n_5000_results <- generate_stuff(5000)
n_5000_results$vars
```

```
##     x1     x2     x3     z1     z2     y1     y2     y3
```

```
## 1.0188 1.0245 1.0142 1.0021 1.0071 5.9845 6.8949 6.2214
```

```r
n_5000_results$covs
```

```
##         x1      x2     x3     z1      z2     y1     y2     y3
## x1  1.0188  0.2220 0.0982 0.3552 -0.0092 0.7133 0.6994 0.8661
## x2  0.2220  1.0245 0.0266 0.4017 -0.0089 2.1176 2.0751 2.1510
## x3  0.0982  0.0266 1.0142 0.0126  0.4052 0.7984 0.7629 0.8132
## z1  0.3552  0.4017 0.0126 1.0021  0.5964 0.8695 0.8732 0.9228
## z2 -0.0092 -0.0089 0.4052 0.5964  1.0071 0.2587 0.2768 0.2574
## y1  0.7133  2.1176 0.7984 0.8695  0.2587 5.9845 4.8467 6.0915
## y2  0.6994  2.0751 0.7629 0.8732  0.2768 4.8467 6.8949 4.9516
## y3  0.8661  2.1510 0.8132 0.9228  0.2574 6.0915 4.9516 6.2214
```

b) Now generate 3 more samples with 50, 500, and 100,000 observations.

Report the three new sample variance-covariance matrices.

```r
n_50_results <- generate_stuff(50)
n_50_results$vars
```

```
##     x1     x2     x3     z1     z2     y1     y2     y3
## 1.1017 0.9949 0.6940 1.3717 0.9856 5.0870 6.6154 5.4093
```

```r
n_50_results$covs
```

```
##        x1      x2      x3      z1     z2     y1     y2     y3
## x1 1.1017  0.3916  0.0530  0.5865 0.2152 0.9916 0.9027 1.1568
## x2 0.3916  0.9949 -0.0358  0.6422 0.2555 1.8750 2.1906 1.9337
## x3 0.0530 -0.0358  0.6940 -0.0921 0.2845 0.4781 0.4475 0.4860
## z1 0.5865  0.6422 -0.0921  1.3717 0.8040 1.1350 1.3983 1.2230
## z2 0.2152  0.2555  0.2845  0.8040 0.9856 0.7561 0.9610 0.7884
## y1 0.9916  1.8750  0.4781  1.1350 0.7561 5.0870 4.6462 5.2358
## y2 0.9027  2.1906  0.4475  1.3983 0.9610 4.6462 6.6154 4.7816
## y3 1.1568  1.9337  0.4860  1.2230 0.7884 5.2358 4.7816 5.4093
```

```r
n_500_results <- generate_stuff(500)
n_500_results$vars
```

```
##     x1     x2     x3     z1     z2     y1     y2     y3
## 0.9706 1.0100 0.9825 0.9463 1.0175 5.9230 6.7142 6.1206
```

```r
n_500_results$covs
```

```
##         x1      x2      x3      z1      z2     y1     y2     y3
## x1  0.9706  0.1773  0.0220  0.3524 -0.0433 0.5859 0.5130 0.7315
## x2  0.1773  1.0100 -0.0528  0.4030 -0.0119 2.0561 2.0590 2.0827
## x3  0.0220 -0.0528  0.9825 -0.0307  0.3984 0.6396 0.5368 0.6429
## z1  0.3524  0.4030 -0.0307  0.9463  0.5614 0.8481 0.7954 0.9010
## z2 -0.0433 -0.0119  0.3984  0.5614  1.0175 0.2545 0.1793 0.2480
## y1  0.5859  2.0561  0.6396  0.8481  0.2545 5.9230 4.7021 6.0108
## y2  0.5130  2.0590  0.5368  0.7954  0.1793 4.7021 6.7142 4.7790
## y3  0.7315  2.0827  0.6429  0.9010  0.2480 6.0108 4.7790 6.1206
```

```r
n_100000_results <- generate_stuff(100000)
n_100000_results$vars
```

```
##     x1     x2     x3     z1     z2     y1     y2     y3
```

```
## 0.9985 1.0055 1.0042 1.0039 1.0004 5.7428 6.9747 5.9689
```

```r
n_100000_results$covs
```

```
##        x1      x2      x3      z1     z2     y1     y2     y3
## x1 0.9985  0.2047  0.1009  0.3516 0.0017 0.6787 0.6740 0.8285
## x2 0.2047  1.0055 -0.0019  0.4089 0.0032 2.0494 2.0526 2.0801
## x3 0.1009 -0.0019  1.0042 -0.0001 0.4025 0.7197 0.7197 0.7348
## z1 0.3516  0.4089 -0.0001  1.0039 0.6026 0.8840 0.8899 0.9368
## z2 0.0017  0.0032  0.4025  0.6026 1.0004 0.2828 0.2892 0.2830
## y1 0.6787  2.0494  0.7197  0.8840 0.2828 5.7428 4.7412 5.8446
## y2 0.6740  2.0526  0.7197  0.8899 0.2892 4.7412 6.9747 4.8423
## y3 0.8285  2.0801  0.7348  0.9368 0.2830 5.8446 4.8423 5.9689
```

2) Next, use the smallest sample, and run a regression for all three variables.

Which coefficient beta3 do you expect regression 1, which one do you find, and why?

For regression 1, I expect that the coefficient for `x3` will be 0.7 because 0.7 is the value assigned to it when generating the data and because the error term `eps1` used to generate `y1` has a mean value of 0 and finite variance. (Similarly, I would expect the coefficient value for `x1` to be 0.2, and the coefficient value for `x2` to be 2.) The regressed value turns out to be approximately equal to the expected value.

(Note that I'm not considering `z1` and `z2` to be variables here, and that I do not elimitate the intercept term.)

Regression 2?

For regression 2, I expect that the coefficient value of `x3` (and also for `x1` and `x2`) will be different. They will be biased by the negative binomial form of `eps2`. Predicting `x3`'s value prior to running a regression is difficult due to the interaction with the other covariates.

Regression 3?

For regression 3, I expect that the coefficient value of `x3` will be the same as it is for regression 1 (i.e. 0.7). (Similarly, I would not expect the coefficient value of `x2` to be different from that observed for regression. On the other hand, I expect that the coefficient value of `x1` will be shifted by an addititve factor of `0.15` because the `eps3` term used to generate `y3` is derived from `x1`.) Indeed, the regressed coefficient value is approximately equal to the expected value for `x3`.

```r
# Need to exclude the epsilon terms.
generate_p2_fmla <- function(i, varnames = c("x1", "x2", "x3") ) {
  formula(paste0("y", i, " ~ ", paste(varnames, collapse = " + ")))
}
fmlas_p2 <- lapply(1:3, generate_p2_fmla)

lm_50_1 <- lm(fmlas_p2[[1]], data = n_50_results$data)
summary(lm_50_1)
```

```
##
## Call:
## lm(formula = fmlas_p2[[1]], data = n_50_results$data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
```

```
## -2.36459 -0.65916  0.05563  0.60765  2.57993
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.007007   0.154439  -0.045 0.964010
## x1           0.213357   0.158001   1.350 0.183509
## x2           1.828328   0.166117  11.006 1.77e-14 ***
## x3           0.766997   0.184800   4.150 0.000142 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.073 on 46 degrees of freedom
## Multiple R-squared:  0.7876, Adjusted R-squared:  0.7737
## F-statistic: 56.84 on 3 and 46 DF,  p-value: 1.649e-15
```

```
# Alternatively...
# broom::tidy(lm_50_1)
# broom::glance(lm_50_1)

lm_50_2 <- lm(fmlas_p2[[2]], data = n_50_results$data)
summary(lm_50_2)
```

```
##
## Call:
## lm(formula = fmlas_p2[[2]], data = n_50_results$data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.6295 -1.0148 -0.1097  0.7621  2.5873
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.16699    0.17526   6.659 2.97e-08 ***
## x1          -0.01114    0.17930  -0.062 0.950714
## x2           2.23369    0.18851  11.849 1.41e-15 ***
## x3           0.76105    0.20971   3.629 0.000712 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.218 on 46 degrees of freedom
## Multiple R-squared:  0.7896, Adjusted R-squared:  0.7759
## F-statistic: 57.55 on 3 and 46 DF,  p-value: 1.32e-15
```

```
coef(lm_50_2)[[4]]
```

```
## [1] 0.7610459
```

```
lm_50_3 <- lm(fmlas_p2[[3]], data = n_50_results$data)
summary(lm_50_3)
```

```
##
## Call:
## lm(formula = fmlas_p2[[3]], data = n_50_results$data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
```

```
## -2.36459 -0.65916  0.05563  0.60765  2.57993
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.007007   0.154439  -0.045 0.964010
## x1           0.363357   0.158001   2.300 0.026053 *
## x2           1.828328   0.166117  11.006 1.77e-14 ***
## x3           0.766997   0.184800   4.150 0.000142 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.073 on 46 degrees of freedom
## Multiple R-squared:  0.8002, Adjusted R-squared:  0.7872
## F-statistic: 61.42 on 3 and 46 DF,  p-value: 4.046e-16
```

```
stargazer(lm_50_1, lm_50_2, lm_50_3, type = "text")
```

```
##
## ===========================================================
##                                   Dependent variable:
##                              -----------------------------
##                                        fmlas_p2
##                                 (1)       (2)       (3)
## -----------------------------------------------------------
## x1                             0.213    -0.011    0.363**
##                               (0.158)   (0.179)   (0.158)
##
## x2                            1.828***  2.234***  1.828***
##                               (0.166)   (0.189)   (0.166)
##
## x3                            0.767***  0.761***  0.767***
##                               (0.185)   (0.210)   (0.185)
##
## Constant                       -0.007   1.167***   -0.007
##                               (0.154)   (0.175)   (0.154)
##
## -----------------------------------------------------------
## Observations                     50        50        50
## R2                             0.788     0.790     0.800
## Adjusted R2                    0.774     0.776     0.787
## Residual Std. Error (df = 46)  1.073     1.218     1.073
## F Statistic (df = 3; 46)      56.844*** 57.551*** 61.416***
## ===========================================================
## Note:                           *p<0.1; **p<0.05; ***p<0.01
```

3) Now increase the sample size gradually…

The regression coefficient values become closer to the expected values as the sample size increases. The Law of Large Numbers explains why the estimated values approach the theoretical values as the sample size increases– simulated outcomes tend to converge to the theoretical outcomes as the number of trials increases. beta_3 definitely moves closer to the result that I expected to find for regression 1 and 3 because of this Law of Large Numbers. However, for regression 2, I did not really know what to suspect. I would have guessed that the strange distribution of the errors may not have had a big effect, and that turns out to be the case. beta3 converges to 0.2 for regressions 1 and 2, and it converges to 0.35 for regression 3.

```r
ns <- c(100, 1000, 10000)
num_fmlas <- 3
i <- 1
j <- 1
while(i <= num_fmlas) {
  fmla_i <- generate_p2_fmla(i)
  while(j <= length(ns)) {
    n <- ns[j]
    n_results <- generate_stuff(n)
    lm_n_ij <- lm(fmla_i, data = n_results$data)
    coefs_row <- c(i, j, n, coef(lm_n_ij))
    names(coefs_row) <- c("regression", "iteration", "n", names(coef(lm_n_ij)))
    # coefs_row
    if(i == 1 & j == 1) {
      coefs_df <- coefs_row %>% t() %>% as_tibble()
    } else {
      coefs_df <- rbind(coefs_df, coefs_row) # bind_rows(coefs_df, coefs_row)
    }
    coefs_df
    # cat("coefs_row:", coefs_row, "\n")
    j <- j + 1
    # cat("j:", j, "\n")
  }
  j <- 1
  i <- i + 1
  # cat("i:", i, "\n")
}
coefs_df <- coefs_df %>% select(-iteration)
coefs_df
```

```
## # A tibble: 9 x 6
##   regression     n `(Intercept)`         x1       x2        x3
##        <dbl> <dbl>         <dbl>      <dbl>    <dbl>     <dbl>
## 1          1   100  -0.021923571 0.14090241 1.954422 0.9566007
## 2          1  1000   0.022120964 0.25643001 1.996594 0.7002272
## 3          1 10000  -0.003778981 0.20234141 2.001969 0.7000587
## 4          2   100   1.434525371 0.05814892 2.369247 0.7660117
## 5          2  1000   1.265511515 0.17613815 2.009189 0.6834309
## 6          2 10000   1.321341807 0.19321977 1.992660 0.7067510
## 7          3   100  -0.021923571 0.29090241 1.954422 0.9566007
## 8          3  1000   0.022120964 0.40643001 1.996594 0.7002272
## 9          3 10000  -0.003778981 0.35234141 2.001969 0.7000587
```

4) Is there any fundamentally biased coefficient in one of the three regressions?...

Yes. The `x1` variable is fundamentally biased in regression 3. This is because the error term `eps3` used in generating `y3` is derived from `x1`. Thus, the coefficient value for `x1` is transformed from it's "unbiased" value of 0.2. (In this case, the coefficient value of x1 is shifted additively by 0.15, giving a value of 0.35.) One might say that the endogenous variable `x1` is a consequence of measurement error in the dependent variable and/or omitted variable bias (depending on one's interpretation of the how the data was simulated.)

To estimate the coefficient correctly, an IV might be used to "uncorrelate" the endogenous variable `x1` from the unobserved error term (in this case, `eps3`) in regression 3. 2SLS can be used to refit the model. To do this, `x1` must be regressed on all other variables (which must be exogenous), along with any IVs. In this case, `z1` seems to be an appropriate choice for an IV since it is correlated with `x1`. The fitted values from the regression on the endogenous variable `x1` are subsequently used in the original regression model in place of the observed `x1` values.

However, upon implementing this scheme, it is apparent that the resultant model is not much different. Presumably, this is because the `z1` IV is correlated with the `x2` independent variable. Thus, the 2SLS model is "weak". A better IV(s) would need to be used to improve the model. In addition, one might say that the IV z1 does not satisfy the IV.1 Relevance criteria for the `x1` endogenous variable. Also, the IV.1 Exogeneity criteria is only approximately true (because `cov(z1, eps3)` is not close "enough" to 0).

A weak instrument's test, a Hausman's test, and a Sargan test are implemented (with the `summary()` function called on the variable created by the `ivreg()` function). It is observed that the Hausman test only barely rejects the null hypothesis (that the endogenous variable `x1` is uncorrelated with the unobserved error term), which indicates that the `x1` term might not be considered truly endogenous.

The difficulty with this problem is that it could be a case of "over-identification"– there are possibly more instrument variables (i.e. `z1`, `z2`, etc.) than there are endogenous variables (`x1`). GMM might be a more appropriate method.

```r
# This is the IV estimate.
# It doesn't seem to be correct (probably because there are more than one covariate).
with(n_50_results$data, cov(z1, x1) / cov(z1, y3))
```

```
## [1] 0.4795631
```

```r
# "Manual" implementation of IV 2SLS
lm_50_3_x1 <- lm(x1 ~ x2 + x3 + z1, data = n_50_results$data)
x1_hat <- fitted.values(lm_50_3_x1)

fmla_3_v2 <- generate_p2_fmla(3, varnames = c("x1_hat", "x2", "x3"))
fmla_3_v2
```

```
## y3 ~ x1_hat + x2 + x3
## <environment: 0x0000000028398860>
```

```r
n_50_results_v2 <- n_50_results
n_50_results_v2$data["x1_hat"] <- x1_hat
lm_50_3_v2 <- lm(fmla_3_v2, data = n_50_results_v2$data)
coef(lm_50_3_v2)
```

```
## (Intercept)      x1_hat          x2          x3
## -0.01827754  0.08855318  1.93744502  0.79360136
```

```r
# Alternative implementation, using the AER package.
# References:
# 1) https://rpubs.com/wsundstrom/t_ivreg
# 2) https://bookdown.org/ccolonescu/RPoE4/random-regressors.html#the-instrumental-variables-iv-method
# 3) https://www.r-bloggers.com/instrumental-variables-in-r-exercises-part-1/
# Additionally, see the AER package documentation for how to construct the ivreg() formula, etc..
# Basically, ivreg(Y ~ X + W | W + Z, ... ), where X is endogenous variable(s),
# Z is instrument(s), and W is exogenous controls (not instruments).
lm_50_3_ivreg <- AER::ivreg(y3 ~ x1 + x2 + x3 | x2 + x3 + z1, data = n_50_results$data)
lm_50_3_ivreg
```

```
##
## Call:
## AER::ivreg(formula = y3 ~ x1 + x2 + x3 | x2 + x3 + z1, data = n_50_results$data)
```

```
##
## Coefficients:
## (Intercept)          x1          x2          x3
##    -0.01828     0.08855     1.93745     0.79360
```

```r
summary(lm_50_3_ivreg, diagnostics = TRUE)
```

```
##
## Call:
## AER::ivreg(formula = y3 ~ x1 + x2 + x3 | x2 + x3 + z1, data = n_50_results$data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.36962 -0.76687  0.05446  0.69433  2.78412
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.01828    0.16038  -0.114 0.909760
## x1           0.08855    0.45311   0.195 0.845915
## x2           1.93745    0.23997   8.074  2.3e-10 ***
## x3           0.79360    0.19512   4.067 0.000185 ***
##
## Diagnostic tests:
##                  df1 df2 statistic p-value
## Weak instruments   1  46     6.848   0.012 *
## Wu-Hausman         1  45     0.445   0.508
## Sargan             0  NA        NA      NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.108 on 46 degrees of freedom
## Multiple R-Squared: 0.7871,  Adjusted R-squared: 0.7732
## Wald test: 55.99 on 3 and 46 DF,  p-value: 2.169e-15
```

```r
# Note that the covariance of the IV z1 and the error eps3 is near 0, which is desirable.
# Note taht the covariance of the IV z1 and the response y3 is non-0, which is desirable.
# Note the relaitvely high correlation of the chosen IV z1 with the x2 independent variable.
# (On the other hand, the correlation between z1 and x3 is low, which is desireable.)
# Presumably, it would be nice if z1 were not so highly correlated with the
# other independent variables, so as to isolate the effect of x1.
with(n_50_results$data, cov(z1, eps3))
```

```
## [1] -0.1141569
```

```r
with(n_50_results$data, cov(z1, y3))
```

```
## [1] 1.222972
```

```r
with(n_50_results$data, cov(z1, x1))
```

```
## [1] 0.5864923
```

```r
with(n_50_results$data, cov(z1, x2))
```

```
## [1] 0.6421617
```

```r
with(n_50_results$data, cov(z1, x3))
```

```
## [1] -0.09213261
```

Q P2b: IV Exercise

1) Suggest an IV method.... Explain what is needed for identification.

It seems like linear 2SLS is an appropriate IV method. For this method we should regress q1 on all the z variables, and then use the fitted values from this regression as variables for our second stage regression, in which we also include all of the original exogenous variables. Once again, for identification we need the assumptions that the instruments are exogeneous and relevant. We can actually test if our IVs are relevant, but we need to make arguments to explain why they are exogenous.

2) Describe the economic assumptions for the previous part if the model is extended...

These assumptions are exogeneity and relevance, which have been outlined many times already in this problem set. By relevance, we mean that there is some type of significant relationship between the IVs, family background variables, and IQ. I think it is fair to say that this is true here. Now, we must also have exogeneity, i.e. that the family background variables only affect wages through IQ. In general, we would expect there to be some type of correlation between family background and wages, but if we control for IQ, we believe that most of this effect goes away. I think that this is also a reasonable assumption because something like the education of your parents should not affect your wages without considering their education on your IQ.

3) Implementation and discussion.

Here is the implementation:

```r
wage_educ <- readr::read_csv("wage_educ.csv")
glimpse(wage_educ)
```

```
## Observations: 935
## Variables: 17
## $ wage    <int> 769, 808, 825, 650, 562, 1400, 600, 1081, 1154, 1000, ...
## $ hours   <int> 40, 50, 40, 40, 40, 40, 40, 40, 45, 40, 43, 38, 45, 38...
## $ iq      <int> 93, 119, 108, 96, 74, 116, 91, 114, 111, 95, 132, 102,...
## $ kww     <int> 35, 41, 46, 32, 27, 43, 24, 50, 37, 44, 44, 45, 40, 24...
## $ educ    <int> 12, 18, 14, 12, 11, 16, 10, 18, 15, 12, 18, 14, 15, 16...
## $ exper   <int> 11, 11, 11, 13, 14, 14, 13, 8, 13, 16, 8, 9, 4, 7, 9, ...
## $ tenure  <int> 2, 16, 9, 7, 5, 2, 0, 14, 1, 16, 13, 11, 3, 2, 9, 2, 9...
## $ age     <int> 31, 37, 33, 32, 34, 35, 30, 38, 36, 36, 38, 33, 30, 28...
## $ married <int> 1, 1, 1, 1, 1, 1, 0, 1, 1, 1, 1, 1, 0, 1, 1, 1, 1, 1, ...
## $ black   <int> 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ...
## $ south   <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ...
## $ urban   <int> 1, 1, 1, 1, 1, 1, 1, 1, 0, 1, 0, 0, 0, 1, 1, 1, 1, 1, ...
## $ sibs    <int> 1, 1, 1, 4, 10, 1, 1, 2, 2, 1, 1, 1, 2, 3, 1, 1, 3, 2,...
## $ brthord <int> 2, NA, 2, 3, 6, 2, 2, 3, 3, 1, 1, 2, NA, 1, 1, 2, 3, 3...
## $ meduc   <int> 8, 14, 14, 12, 6, 8, 8, 8, 14, 12, 13, 16, 12, 10, 12,...
## $ feduc   <int> 8, 14, 14, 12, 11, NA, 8, NA, 5, 11, 14, NA, 12, 10, 1...
## $ lwage   <dbl> 6.645091, 6.694562, 6.715384, 6.476973, 6.331502, 7.24...
```

```r
# Building up the formulas here.
fmla_2b_shared_rhs <- "exper + tenure + educ + married + south + urban + black"
fmla_2b_shared_rhs_iq <- str_c(fmla_2b_shared_rhs, " + iq")
fmla_2b_shared_rhs_kww <- str_c(fmla_2b_shared_rhs, " + kww")
```

```r
fmla_2b_shared_iq <- str_c("log(wage) ~ ", fmla_2b_shared_rhs_iq)
fmla_2b_shared_kww <- str_c("log(wage) ~ ", fmla_2b_shared_rhs_kww)
fmla_2b_shared_rhs_2 <- str_c(fmla_2b_shared_rhs, " + sibs + meduc + feduc")

#fmla_2b_iq_ols <- as.formula(fmla_2b_shared_iq)
fmla_2b_iq_vireg <- as.formula(str_c(fmla_2b_shared_iq, " | ", fmla_2b_shared_rhs_2))

#fmla_2b_kww_ols <- as.formula(fmla_2b_shared_kww)
fmla_2b_kww_vireg <- as.formula(str_c(fmla_2b_shared_kww, " | ", fmla_2b_shared_rhs_2))

#lm_2b_iq_ols <- lm(fmla_2b_iq_ols, data = wage_educ)
#summary(lm_2b_iq_ols)
lm_2b_iq_ivreg <- AER::ivreg(fmla_2b_iq_vireg, data = wage_educ)
summary(lm_2b_iq_ivreg, diagnosics = TRUE)
```

```
##
## Call:
## AER::ivreg(formula = fmla_2b_iq_vireg, data = wage_educ)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -2.134057 -0.217364  0.005651  0.231091  1.402072
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.471615   0.468913   9.536  < 2e-16 ***
## exper        0.016219   0.004008   4.047 5.76e-05 ***
## tenure       0.007675   0.003096   2.479   0.0134 *
## educ         0.016181   0.026198   0.618   0.5370
## married      0.190101   0.046759   4.066 5.33e-05 ***
## south       -0.047992   0.036742  -1.306   0.1919
## urban        0.186938   0.032799   5.700 1.76e-08 ***
## black        0.040027   0.113868   0.352   0.7253
## iq           0.015437   0.007708   2.003   0.0456 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3878 on 713 degrees of freedom
## Multiple R-Squared: 0.1546,  Adjusted R-squared: 0.1451
## Wald test: 25.81 on 8 and 713 DF,  p-value: < 2.2e-16
```

```r
#lm_2b_kww_ols <- lm(fmla_2b_kww_ols, data = wage_educ)
#summary(lm_2b_kww_ols)
lm_2b_kww_ivreg <- AER::ivreg(fmla_2b_kww_vireg, data = wage_educ)
summary(lm_2b_kww_ivreg, diagnosics = TRUE)
```

```
##
## Call:
## AER::ivreg(formula = fmla_2b_kww_vireg, data = wage_educ)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -2.319371 -0.238608  0.003009  0.252612  1.496516
##
```

```
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  5.217818   0.162759  32.059  < 2e-16 ***
## exper        0.006868   0.006747   1.018 0.309044
## tenure       0.005115   0.003774   1.355 0.175766
## educ         0.026081   0.025505   1.023 0.306857
## married      0.160527   0.052976   3.030 0.002532 **
## south       -0.091887   0.032215  -2.852 0.004466 **
## urban        0.148400   0.041160   3.605 0.000333 ***
## black       -0.042445   0.089370  -0.475 0.634975
## kww          0.024944   0.015058   1.657 0.098045 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3874 on 713 degrees of freedom
## Multiple R-Squared: 0.1563,  Adjusted R-squared: 0.1468
## Wald test:  25.7 on 8 and 713 DF,  p-value: < 2.2e-16
```

In both cases, the married and urban factors are statistically significant when looking at log wages. Also, our IVs for IQ significantly affect log wages through IQ, but not through KWW. This would be more interesting if I knew what KWW was, but it does make sense that these family background factors can effect log wages through IQ.

Q P3: Panel Exercise

1)

$$newst_i = \theta_t + \beta_1 vfst_{it} + \alpha_i + \mu_{it}$$

$$tc_i = \theta_t + \beta_1 vfst_{it} + \alpha_i + \mu_{it}$$

where α_i represents fixed factors that affect the economic climate, θ_t represents a different intercept for each time period, t represents the time period (i.e.before/after implementation) i is an index for household.

2)

The male household indicator variable `maleh` can be added to the formuia as a fixed effect.

$$newst_i = \theta_t + \beta_1 vfst_{it} + \alpha_i + maleh_i + \mu_{it}$$

3)

The assumption that village funds are random is probably naive. It is likely that a number of factors are related to the initial amount of village funds, including: average household size in a village and the the number of households in a village (i.e. population), the economic standing of each household, (i.e. poverty levels) etc. If the assumption of random distribution of funds is indeed false, then it will make the estimate of the coefficient for `vfst` unreliable/inconsistent.

To provide more detail, the bias in the described case comes at an "entity"-specific (and not necessarily a time-variant) level of abstraction. Thus, if this bias is not accounted for, then model estimates are doomed to be shifted. Therefore, the fitted values of new short term credit and total consumption will be wrong, i.e. misidentified.

4)

A linear fixed effects model incorporating household attributes seems like an appropriate method for modeling this panel data. Ideally, valid IVs could also be used to help remove bias from the dependent variable, but it seems more probable than not that most IVs will be weak and will not lead to improvement with the model. Rather, controlling for circumstances with fixed effect seems appropriate. A random effect model seems unecessary because it seems probable that the data is highly time-variant or impossible to model with proper controls.

Differencing might also be tested since it–like a fixed effects framework– can be useful for accounting for entity-level attributes (by modeling them as entity-specific constants via dummy variables).

The proposed procedure avoids the issues of the previous models because the previously unobserved entity-specfic effects (i.e. caused by ommitted variables and sampling bias) are accounted for by controls incorporated into the model.

5) Implementation of OLS, FD, and FE

It is noteworthy that the FD and FE models estimate similar values for predicting `newst`, but not for predicting `tc`. This could indicate that there is still some bias that is not accounted for in the `tc` panel data models. OLS gives pretty different estimates than the other two because the way the models are carried out is different, so this was expected.

```
rm(list = ls())
library("plm")
load("microcredit.Rdata")
glimpse(dt.microcredit)
```

```
## Observations: 4,718
## Variables: 21
## $ caseid  <dbl> 7030707001, 7030707001, 7030707001, 7030707001, 703070...
## $ village <dbl> 7, 7, 7, 7, 7, 7, 7, 7, 7, 7, 7, 7, 7, 7, 7, 7, 7, 7, ...
## $ year    <int> 1, 2, 3, 4, 5, 6, 7, 1, 2, 3, 4, 5, 6, 7, 1, 2, 3, 4, ...
## $ newst   <dbl> 70000, 0, 240000, 90000, 320000, 120000, 20000, 0, 0, ...
## $ tc      <dbl> 97613.04, 86019.23, 125328.72, 161232.47, 142776.30, 7...
## $ vfst    <dbl> 0, 0, 0, 0, 0, 2, 2, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ...
## $ madult  <dbl> 1, 1, 1, 2, 2, 2, 2, 1, 1, 1, 1, 1, 1, 1, 1, 1, 2, 2, ...
## $ fadult  <dbl> 1, 1, 1, 1, 1, 1, 1, 2, 3, 3, 3, 3, 2, 2, 1, 1, 3, 4, ...
## $ kids    <dbl> 2, 2, 2, 1, 1, 1, 1, 1, 0, 1, 2, 2, 2, 2, 0, 4, 6, 6, ...
## $ maleh   <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, ...
## $ farm    <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, ...
## $ ageh    <dbl> 37, 39, 40, 41, 40, 41, 42, 47, 48, 48, 49, 50, 51, 52...
## $ age2h   <dbl> 1369, 1521, 1600, 1681, 1600, 1681, 1764, 2209, 2304, ...
## $ educh   <dbl> 4, 7, 7, 7, 7, 7, 7, 4, 7, 7, 7, 7, 7, 7, 4, 7, 7, 0, ...
## $ d1      <dbl> 1, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, ...
## $ d2      <dbl> 0, 1, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 1, 0, 0, ...
## $ d3      <dbl> 0, 0, 1, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 1, 0, ...
## $ d4      <dbl> 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 1, ...
## $ d5      <dbl> 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, ...
## $ d6      <dbl> 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, ...
## $ d7      <dbl> 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, ...
```

```
stargazer(dt.microcredit, type = "text")
```

```
##
## ==================================================================================
```

```
## Statistic   N         Mean             St. Dev.          Min            Max
## ---------------------------------------------------------------------------------
## caseid    4,718 33,593,909,871.000 18,077,185,474.000 7,030,707,001 53,060,510,031
## village   4,718        6.444              3.506             1              14
## year      4,718        4.000              2.000             1              7
## newst     4,718      22,220.430         53,335.420        0.000      1,023,000.000
## tc        4,686      71,396.010         95,845.120       550.363     2,766,750.000
## vfst      4,718        0.273              0.691            0.000          8.000
## madult    4,718        1.473              0.884             0              7
## fadult    4,718        1.582              0.758             0              6
## kids      4,718        1.546              1.198             0              9
## maleh     4,718        0.726              0.446            0.000          1.000
## farm      4,718        0.658              0.475             0              1
## ageh      4,718       54.289             13.478           23.000         93.000
## age2h     4,718      3,128.954          1,520.405        529.000       8,649.000
## educh     4,718        6.034              3.160            0.000         16.000
## d1        4,718        0.143              0.350             0              1
## d2        4,718        0.143              0.350             0              1
## d3        4,718        0.143              0.350             0              1
## d4        4,718        0.143              0.350             0              1
## d5        4,718        0.143              0.350             0              1
## d6        4,718        0.143              0.350             0              1
## d7        4,718        0.143              0.350             0              1
## ---------------------------------------------------------------------------------
```

```r
dt.microcredit %>%
  as_tibble() %>%
  group_by(village, year) %>%
  summarise_at(vars(vfst, newst, tc), funs(n()))
```

```
## # A tibble: 98 x 5
## # Groups:   village [?]
##    village  year  vfst newst    tc
##      <dbl> <int> <int> <int> <int>
##  1       1     1    35    35    35
##  2       1     2    35    35    35
##  3       1     3    35    35    35
##  4       1     4    35    35    35
##  5       1     5    35    35    35
##  6       1     6    35    35    35
##  7       1     7    35    35    35
##  8       2     1   124   124   124
##  9       2     2   124   124   124
## 10       2     3   124   124   124
## # ... with 88 more rows
```

```r
# microcredit_treated <- dt.microcredit
dt.microcredit %>% as_tibble() %>% group_by(caseid, year)
```

```
## # A tibble: 4,718 x 21
## # Groups:   caseid, year [4,718]
##         caseid village  year  newst        tc  vfst madult fadult  kids
##          <dbl>   <dbl> <int>  <dbl>     <dbl> <dbl>  <dbl>  <dbl> <dbl>
##  1 7030707001       7     1  70000  97613.04     0      1      1     2
##  2 7030707001       7     2      0  86019.23     0      1      1     2
```

```
##  3 7030707001       7     3 240000 125328.72     0       1       1     2
##  4 7030707001       7     4  90000 161232.47     0       2       1     1
##  5 7030707001       7     5 320000 142776.30     0       2       1     1
##  6 7030707001       7     6 120000  71481.65     2       2       1     1
##  7 7030707001       7     7  20000  70854.34     2       2       1     1
##  8 7030707003       7     1      0 361963.34     0       1       2     1
##  9 7030707003       7     2      0  44275.89     0       1       3     0
## 10 7030707003       7     3      0  96192.15     0       1       3     1
## # ... with 4,708 more rows, and 12 more variables: maleh <dbl>,
## #   farm <dbl>, ageh <dbl>, age2h <dbl>, educh <dbl>, d1 <dbl>, d2 <dbl>,
## #   d3 <dbl>, d4 <dbl>, d5 <dbl>, d6 <dbl>, d7 <dbl>
```

```
# microcredit_pdf <- pdata.frame(micocredit_treated, index = c("village", "year"))
microcredit_pdf <- data.frame(dt.microcredit, index = c("caseid", "year"))
# microcredit_pdf
# table(index(microcredit_pdf), useNA = "ifany")

plm_rhs_constant <- str_c("vfst + d1 + d2 + d3 + d4 + d5 + d6 + d7")
plm_newst_fmla <- str_c("newst ~ ", plm_rhs_constant) %>% as.formula()
# plm_newst_fmla_no1 <- str_c("newst ~ ", plm_rhs_constant, " + 0") %>% as.formula()

# Must specify index for fixed effects model, although not for others(?)
plm_newst_ols <- plm(plm_newst_fmla, data = microcredit_pdf, index = c("caseid", "year"), model = "pooli
plm_newst_fe <- plm(plm_newst_fmla, data = microcredit_pdf, index = c("caseid", "year"), model = "withi
plm_newst_fd <- plm(plm_newst_fmla, data = microcredit_pdf, index = c("caseid", "year"), model = "fd")
# plm_newst_fd_no1 <- plm(plm_newst_fmla_no1, data = microcredit_pdf, index = c("caseid", "year"), mode

stargazer(
  plm_newst_ols,
  plm_newst_fe,
  plm_newst_fd,
  # plm_newst_fd_no1,
  type = "text",
  omit = "d.*",
  no.space = FALSE,
  omit.labels = c("Time Dummies"),
  column.labels = c("Pooling", "FE", "FD") #, "FD w/o Intercept")
)
```

```
##
## ==========================================================================================
##                                         Dependent variable:
##                ---------------------------------------------------------------------------
##                                                newst
##                      Pooling                     FE                       FD
##                        (1)                      (2)                      (3)
## ------------------------------------------------------------------------------------------
## vfst             17,738.680***            13,544.110***            13,883.430***
##                   (1,385.738)              (1,229.837)              (1,442.177)
##
## Constant         10,446.690***
##                   (2,444.415)
##
## ------------------------------------------------------------------------------------------
## Time Dummies         Yes                      Yes                      Yes
```

```
## -----------------------------------------------------------------------------------------
## Observations          4,718                    4,718                    4,044
## R2                    0.040                    0.041                    0.026
## Adjusted R2           0.039                    -0.121                   0.024
## F Statistic  27.992*** (df = 7; 4710) 24.637*** (df = 7; 4037) 17.800*** (df = 6; 4037)
## =========================================================================================
## Note:                                                     *p<0.1; **p<0.05; ***p<0.01
```

```r
plm_tc_fmla <- str_c("tc ~ ", plm_rhs_constant) %>% as.formula()

plm_tc_ols <- plm(plm_tc_fmla, data = microcredit_pdf, index = c("caseid", "year"), model = "pooling")
plm_tc_fe <- plm(plm_tc_fmla, data = microcredit_pdf, index = c("caseid", "year"), model = "within")
plm_tc_fd <- plm(plm_tc_fmla, data = microcredit_pdf, index = c("caseid", "year"), model = "fd")
# plm_tc_fd_no1 <- plm(plm_tc_fmla_no1, data = microcredit_pdf, index = c("caseid", "year"), model = "f

stargazer(
  plm_tc_ols,
  plm_tc_fe,
  plm_tc_fd,
  # plm_tc_fd_no1,
  type = "text",
  omit = "d.*",
  no.space = FALSE,
  omit.labels = c("Time Dummies"),
  column.labels = c("Pooling", "FE", "FD")
)
```

```
##
## ======================================================================================
##                                        Dependent variable:
##              -------------------------------------------------------------------------
##                                                tc
##                      Pooling                   FE                      FD
##                        (1)                    (2)                     (3)
## --------------------------------------------------------------------------------------
## vfst              8,905.364***            4,024.817*              -1,427.061
##                   (2,537.446)            (2,316.107)             (2,377.469)
##
## Constant         66,054.210***
##                   (4,475.447)
##
## --------------------------------------------------------------------------------------
## Time Dummies          Yes                    Yes                     Yes
## --------------------------------------------------------------------------------------
## Observations         4,686                  4,686                   4,012
## R2                   0.010                  0.013                   0.008
## Adjusted R2          0.008                  -0.155                  0.006
## F Statistic  6.433*** (df = 7; 4678) 7.499*** (df = 7; 4005) 5.366*** (df = 6; 4005)
## ======================================================================================
## Note:                                                  *p<0.1; **p<0.05; ***p<0.01
```

6) Importance of time trend?

It is important to incporate a time trend in order to account for possible seasonality in the data (i.e. correlation among lagged intervals). Not accounting for time-variant effects would lead to misleading deductions.

7) Test for Serial Correlation

It seems like there is some serial correlation (although this deduction could be to improper coding implementation). This indicates that the FD model might be better to use.

```
# Method suggested in lecture slides.
u <- residuals(plm_newst_fd)
microcredit_pdf_lag1 <- microcredit_pdf
microcredit_pdf_lag1[as.double(names(u)), "u"] <- u
plm_newst_fd_lag1 <- plm(u ~ lag(u, 1), microcredit_pdf_lag1, index = c("caseid", "year"), model = "pool
lm(u ~ lag(u, 1), microcredit_pdf_lag1) %>% summary()
```

```
##
## Call:
## lm(formula = u ~ lag(u, 1), data = microcredit_pdf_lag1)
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -1.198e-09 -2.200e-11 -1.500e-11 -7.000e-12  6.353e-08
##
## Coefficients:
##              Estimate Std. Error   t value Pr(>|t|)
## (Intercept) 7.232e-24  1.575e-11 0.000e+00        1
## lag(u, 1)   1.000e+00  2.803e-16 3.567e+15   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.002e-09 on 4042 degrees of freedom
##   (674 observations deleted due to missingness)
## Multiple R-squared:      1,  Adjusted R-squared:      1
## F-statistic: 1.272e+31 on 1 and 4042 DF,  p-value: < 2.2e-16
```

```
summary(plm_newst_fd_lag1)
```

```
## Pooling Model
##
## Call:
## plm(formula = u ~ lag(u, 1), data = microcredit_pdf_lag1, model = "pooling",
##     index = c("caseid", "year"))
##
## Balanced Panel: n = 674, T = 5, N = 3370
##
## Residuals:
##       Min.    1st Qu.     Median    3rd Qu.       Max.
## -665919.94   -8385.05     598.76    4994.82 1016424.20
##
## Coefficients:
##                Estimate  Std. Error t-value Pr(>|t|)
## (Intercept) -5.1780e-10  8.8568e+02   0.000        1
## lag(u, 1)   -4.4919e-01  1.5210e-02 -29.533   <2e-16 ***
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Total Sum of Squares:    1.1209e+13
## Residual Sum of Squares: 8.9034e+12
## R-Squared:      0.2057
## Adj. R-Squared: 0.20546
## F-statistic: 872.185 on 1 and 3368 DF, p-value: < 2.22e-16
```

```r
# Two alternative tests.
lmtest::dwtest(plm_newst_fd)
```

```
##
##  Durbin-Watson test
##
## data:  plm_newst_fd
## DW = NA, p-value = NA
## alternative hypothesis: true autocorrelation is greater than 0
```

```r
Box.test(residuals(plm_newst_fd), type = "Ljung-Box")
```

```
##
##  Box-Ljung test
##
## data:  residuals(plm_newst_fd)
## X-squared = 649.66, df = 1, p-value < 2.2e-16
```

```r
# Repeating the lecture method for tc model.
u <- residuals(plm_tc_fd)
microcredit_pdf_lag1 <- microcredit_pdf
microcredit_pdf_lag1[as.double(names(u)), "u"] <- u
plm_tc_fd_lag1 <- plm(u ~ lag(u, 1), microcredit_pdf_lag1, index = c("caseid", "year"), model = "poolin
lm(u ~ lag(u, 1), microcredit_pdf_lag1)
```

```
##
## Call:
## lm(formula = u ~ lag(u, 1), data = microcredit_pdf_lag1)
##
## Coefficients:
## (Intercept)    lag(u, 1)
##           0            1
```

```r
summary(plm_tc_fd_lag1)
```

```
## Pooling Model
##
## Call:
## plm(formula = u ~ lag(u, 1), data = microcredit_pdf_lag1, model = "pooling",
##     index = c("caseid", "year"))
##
## Unbalanced Panel: n = 672, T = 1-5, N = 3330
##
## Residuals:
##      Min.   1st Qu.    Median   3rd Qu.      Max.
## -862137.6  -17418.5   -3516.8   12433.6  920665.1
##
## Coefficients:
```

```
##                Estimate  Std. Error t-value Pr(>|t|)
## (Intercept) -190.286011 1196.766076  -0.159   0.8737
## lag(u, 1)     -0.243675    0.012788 -19.055   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Total Sum of Squares:    1.7604e+13
## Residual Sum of Squares: 1.5873e+13
## R-Squared:      0.098368
## Adj. R-Squared: 0.098097
## F-statistic: 363.085 on 1 and 3328 DF, p-value: < 2.22e-16
```