Analysis of 2010 Census Bureau Data - What is the effect of having more children on earnings and employment?

This project is half exploration, half tutorial. The guiding question is, “what’s the effect of having 3 or more children (variable”morekids=1" from here on out) on income/the decision for a mother to work?" I walk through every component required to conduct this analysis, including a lot of derivations, proofs, some fun unsolved stats problems, and the boring old R implementations. I recommend you read the below for a brief summary before diving in. I think the most interesting parts by far are the last two sections regarding model comparison and CACE; the earlier stuff has some fun proofs, and basic data explorations and “checks” for instrument validity, but is not as exciting.

The first question any econometrician/statistician must have is, “how do I represent my existing data?” For every variable, there are numerous ways to represent it. For a variable like education; maybe the difference between 1 or 2 years of high school doesn’t really matter, but the difference between 3 (incomplete) and 4 (graduating) years of high school. I play around with some simple linear probability models to explore different ways to represent our existing data, and briefly discuss how I might be able to use some variables differently.

The second question is, “what controls do we want to include in our model?” To control for omitted variables bias, some might say, “include every control you can. Duh!” But it is nevertheless important to check for the significance of certain control variables and think about what that might mean. If you have some independent variable (morekids), you don’t want it to be highly correlated with a certain variable that you then omit from your analysis.

Third, we ask, “What variable can we use as a valid instrument for morekids=1?” I provide some derivations, empirical evidence, and logical explanations for the relevance and exogeneity of samesex=1 (the sex of the first two children being same). I then, well, do the actual regressions.

After establishing that samesex is a valid instrument to use in our models, we then ask, “ok, but what do we include in the model? What controls do we include, and how does that relate to the explorations we’ve done earlier?” In this section, I provide some quick and dirty proofs to justify what inclusions I make to the model, explain the variation I see in results with and without inclusion of certain controls, and talk about some fun stats stuff.

Finally, I get heavy into the Complier Average Causal Effects framework (CACE). This framework helps us look at how instrument relevance might differ across different populations by looking at “compliers”, for which the instrument does lead to a certain effect on the instrumented variable. If instrument relevance (aka proportion of “compliers” - we’ll explain later) differs among different demographics/subjects, then we have good reason to consider that our estimate effect sizes will be different across different demographics too.

```r
library(AER)
library(stargazer)
library(dplyr)
library(glmnetUtils)
library(knitr)
opts_chunk$set(tidy.opts=list(width.cutoff=60),tidy=TRUE)
df = read.csv("sample.csv")
```

RMSE helper function

```r
rmse <- function(res) {
  return(sqrt(crossprod(res)/length(res)))
}
```

Part 1

In this section, I'll explore a basic lienar probability model and describe how different parameterizations for this model can make certain calculations different, easier, harder, etc.

1 - Construct models for having 3+ kids

First, I construct some basic models to look at the effects of education, age, and age of first marriage on the probability of having more than 3 kids. This is a simple linear probability model. This is not meant to be a causal regression, but meant to explore the effects of creating indicator variables for numerical variables such as "years of education". This will inform our model creation later. In model 1, we use mother's education, age, and age of first child as independent variables. In Model 2, we factor years of education into dummy AKA indicator variables, and in Model 3, we factor all the variables into separate indicator variables.

Making the models:

```
m1_lm <- lm(morekids ~ educm + agem + agefstm, df)
m2_lm <- lm(morekids ~ factor(educm) + agem + agefstm, df)
m3_lm <- lm(morekids ~ factor(educm) + factor(agem) + factor(agefstm), df)
```

I then calculate the normal and adjusted R-Squared, RMSE, and Akaike Information Criterion values. From these values alone, it seems like model 3 (where every variable, such as age and education, are changed into indicator "bins") has a lot of explanatory power within the dataset. We will later analyze whether this generalizes out of the dataset.

```
r_squared_vals <- c(summary(m1_lm)$r.squared, summary(m2_lm)$r.squared, summary(m3_lm)$r.squared)
names(r_squared_vals) <- c("Model 1", "Model 2", "Model 3")
adj_rsqua_vals <- c(summary(m1_lm)$adj.r.squared, summary(m2_lm)$adj.r.squared, summary(m3_lm)$adj.r.squ
names(adj_rsqua_vals) <- c("Model 1", "Model 2", "Model 3")
rmse_vals <- c(rmse(m1_lm$residuals), rmse(m2_lm$residuals), rmse(m3_lm$residuals))
names(rmse_vals) <- c("Model 1", "Model 2", "Model 3")
aic_vals <- c(AIC(m1_lm), AIC(m2_lm), AIC(m3_lm))
names(aic_vals) <- c("Model 1", "Model 2", "Model 3")

print(r_squared_vals)
```

```
##    Model 1    Model 2    Model 3
## 0.07045997 0.07534509 0.07609582
```

```
print(adj_rsqua_vals)
```

```
##    Model 1    Model 2    Model 3
## 0.07044817 0.07525905 0.07589260
```

```
print(rmse_vals)
```

```
##   Model 1   Model 2   Model 3
## 0.4662791 0.4650522 0.4648634
```

```
print(aic_vals)
```

```
##  Model 1  Model 2  Model 3
## 310228.9 309020.9 308888.9
```

(i) - Probabilities of having 3+ kids for different ages of women

Now, I demonstrate how to use these different models to answer basic questions, e.g. what's the difference in the probability of having 3+ kids for a 30 year old woman, compared to a 35 year old?

For model 1, the difference in probabilities is considered holding all other coefficients constant, which will cancel out when the probability expressions are subtracted from each other.

$$Diff = 35\beta_{agem} - 30\beta_{agem}$$

. For the standard error,

$$SE(Diff) = \sqrt{Var(Diff)} = \sqrt{Var((35-30)\beta_{agem})} = \sqrt{25 * Var(\beta_{agem})} = 5 * \sqrt{Var(\beta_{agem})} = 5 * SE(\beta_{agem})$$

```
i_m1 <- m1_lm$coefficients["agem"]*(35-30)
i_se_m1 <- 5*summary(m1_lm)$coefficients["agem", "Std. Error"]

unname(i_m1)
```

```
## [1] 0.1486871
```

```
i_se_m1
```

```
## [1] 0.001555809
```

For M1, the difference is 0.1486871 and the SE is 0.001555809.

Model 2:

Difference in probabilities, and standard error, are the same formula as in Model 1. Despite Model 2 containing dummies, we once again hold all of them constant and they cancel out, leading to the same expression for the Diff.

```
i_m2 <- m2_lm$coefficients["agem"]*(35-30)
i_se_m2 <- 5*summary(m2_lm)$coefficients["agem", "Std. Error"]

unname(i_m2)
```

```
## [1] 0.1486034
```

```
i_se_m2
```

```
## [1] 0.00155253
```

For M2, the difference is 0.1486034 and the SE is 0.00155253

Model 3:

When taking the difference in probabilities, we hold other coefficients constant, changing only the dummy variables for agem==35 and agem==30. Thus, the expression for the difference becomes $Diff = X_{agem=35} - X_{agem=30}$, and thus,

$$SE(Diff) = \sqrt{Var(Diff)} = \sqrt{Var(X_{agem=35} - X_{agem=30})}$$

$$= \sqrt{Var(X_{agem=35}) + Var(X_{agem=30}) - 2Cov(X_{agem=35}, X_{agem=30})}$$

. Since the two RVs $X_{agem=35}$ and $X_{agem=30}$ are not independent, the covariance is nonzero, and we have to manually work this expression out.

```
i_m3 <- m3_lm$coefficients["factor(agem)35"] - m3_lm$coefficients["factor(agem)30"]
vcov_mat <- vcov(m3_lm)
var_x35 <- vcov_mat["factor(agem)35", "factor(agem)35"]
var_x30 <- vcov_mat["factor(agem)30", "factor(agem)30"]
cov_x35_x30 <- vcov_mat["factor(agem)35", "factor(agem)30"]

i_se_m3 <- sqrt(var_x35 + var_x30 - 2*cov_x35_x30)

i_m3
```

```
## factor(agem)35
##      0.1391881
```

```
i_se_m3
```

```
## [1] 0.00441031
```

For M3, the difference is 0.1391881 and the SE is 0.00441031.

We see here that the inclusion of dummy variables makes answering certain questions a bit more involved, but not by much. And it's obvious that the inclusion of these dummy variables would make performance on the training dataset much better; it provides the model with more degrees of freedom to overfit to. The question is, does it perform better outside of this dataset?

Graphs

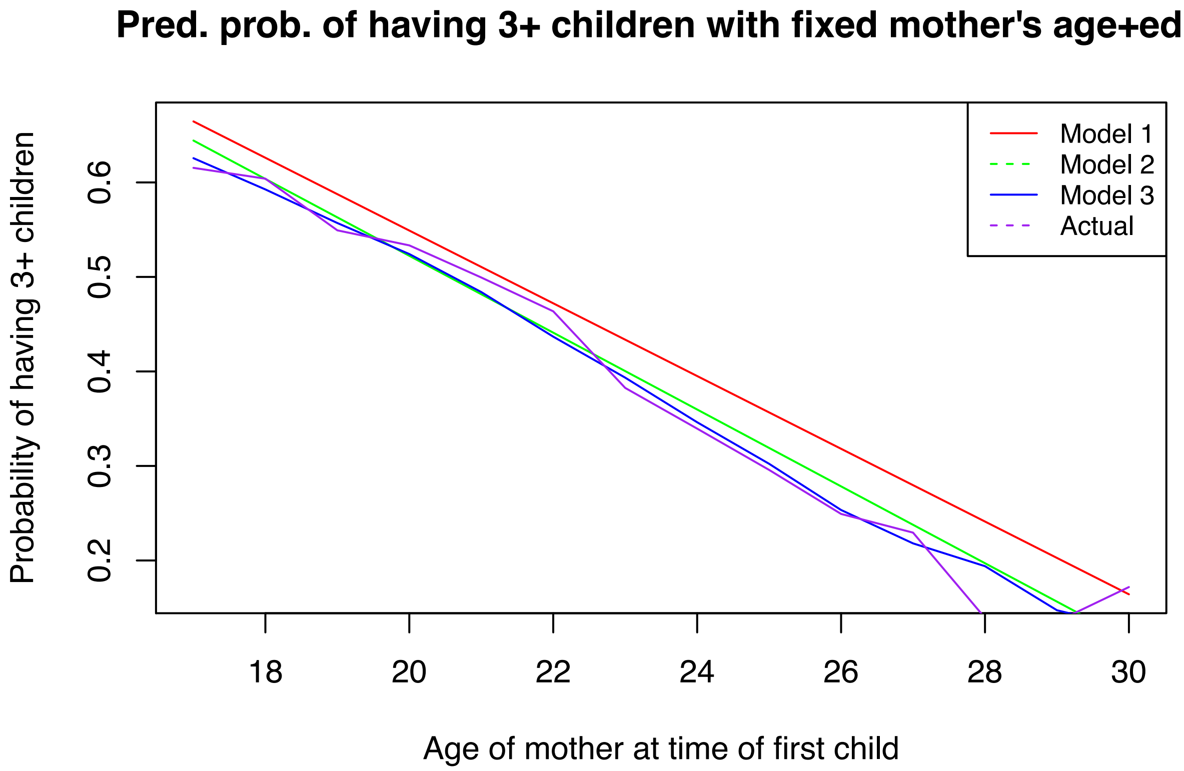
Here I graph the actual and predicted probabilities for M1, M2, M3 of having 3+ kids for 35 y/o mothers with 12 years of education, over age of first child values 17-30.

First, compute the predicted probabilities using the linear probability models from earlier. I then compute the actual empirical probability of having 3+ kids for these mothers.

```r
m1_preds <- c()
m2_preds <- c()
m3_preds <- c()

dummystuffm2 <- c(rep(0, 20))
dummystuffm2[12] <- 1

dummystuffm3edu <- c(rep(0, 20))
dummystuffm3edu[12] <- 1

dummystuffm3age <- c(rep(0, 14))
dummystuffm3age[14] <- 1

dummystuffm3afval <- c(rep(0, 18))

actual_probs <- c()

# new <- data.frame(agem=c(35)*14)
for (afval in c(17:30)) {
  m1_p <- sum(m1_lm$coefficients*c(1, 12, 35, afval))
  m2_p <- sum(m2_lm$coefficients*c(1, dummystuffm2, 35, afval))

  dummystuffm3afval[afval-15] <- 1
  m3_p <- sum(m3_lm$coefficients*c(1, dummystuffm3edu, dummystuffm3age, dummystuffm3afval))
  dummystuffm3afval[afval-15] <- 0

  m1_preds <- c(m1_preds, m1_p)
  m2_preds <- c(m2_preds, m2_p)
  m3_preds <- c(m3_preds, m3_p)

  av <- mean(subset(df, (educm == 12) & (agefstm == afval) & (agem == 35))$morekids)

  actual_probs <- c(actual_probs, av)
}
```

Graph the values.

```
x_axis = c(17:30)
plot(x_axis, m1_preds, type="l", col="red", main="Pred. prob. of having 3+ children with fixed mother's
lines(x_axis, m2_preds, col="green")
lines(x_axis, m3_preds, col="blue")
lines(x_axis, actual_probs, col="purple")

legend("topright", legend=c("Model 1", "Model 2", "Model 3", "Actual"),
       col=c("red", "green", "blue", "purple"), lty=1:2, cex=0.8)
```



Now we get to the fun stuff: a comparison of the models:

In comparing M1, M2, and M3, we first theoretically consider what "binning" and creating extra indicator variables generally does, and how it relates to our estimation of the conditional expectation function. We next consider how this general theory relates to the models/situation at hand to evaluate pros and cons of the three models. We finally take an empirical stance by re-fitting and re-validating these models on a train and test set of data to observe how they overfit and generalize.

First, the general theory.

1. Binning the age (as done in M3) causes us to lose the assumption that the effect of age is constant. It posits that for movements between different ages, you could have different sized effects.

2. Each of those age effect estimators will have a high standard error (we will compare these next), in part because there is more variation around a singular age bin with fewer datapoints, in part because there are just more degrees of freedom.

3. Finally, and most importantly, binning the age will create a different model from using age as a continuous variable *IF THE CONDITIONAL EXPECTATION FUNCTION E[Y/X] IS NON-LINEAR*. We know that if you have dummies for every possible data point, you fit the mean of each datapoint, thus obtaining the CEF. This is super important! If the CEF is non-linear, the models will fundamentally differ. In theory, you at least want the form of your sample CEF to match the form of the population

CEF, so it is important to consider the distribution/linearity of the sample data as compared to the population data.

Now, apply this general theory to our situation.

1. Is the effect of age of first child (agefstm) on having more children constant? I would expect it mostly is. The change in probability from, say, 23-25 has no reason to be different than the change in probability from 33-35. I don't see any specific ages where there may be a discontinuity or a particularly prominent jump/change in slope. This means M1 and M2 are more likely to be "True" in my opinion, since they hold agefstm's effect to be constant.

2. Here are the standard errors for model 1 and model 2: 3.9526506×10^{-4}, and 4.1058372×10^{-4}. In contrast, here are some of the standard errors for model 3: for agefstm = 25, 0.0104234, and for agefstm = 30, 0.0236415. Way bigger SEs! Another reason to prefer M1 and M2, for lower variability in estimates.

3. The sample CEF is clearly nonlinear (look at the plotted "actual probabilities"). M3 estimates a nonlinear model as well, which matches the sample CEF. However, I hypothesize that this might be bad, because though the sample CEF is nonlinear, the population CEF might be linear, and these nonlinearities are the result of randomness in the data. Thus, M3 would be capturing little trends/idiosyncracies that do not generalize well compared to M1 and M2. However, instead of arguing this analytically, we can just test that thought empirically below.

Empirical validation:

```
train1_df <- subset(df, (rv < 0.75))
holdout1_df <- subset(df, (rv >= 0.75))

m1_vlm <- lm(morekids ~ educm + agem + agefstm, train1_df)
m2_vlm <- lm(morekids ~ factor(educm) + agem + agefstm, train1_df)
m3_vlm <- lm(morekids ~ factor(educm) + factor(agem) + factor(agefstm), train1_df)

m1_rmse_val <- mean((predict(m1_vlm, newdata=holdout1_df) - holdout1_df$morekids)^2)
m2_rmse_val <- mean((predict(m2_vlm, newdata=holdout1_df) - holdout1_df$morekids)^2)
m3_rmse_val <- mean((predict(m3_vlm, newdata=holdout1_df) - holdout1_df$morekids)^2)

m1_rmse_train <- mean((predict(m1_vlm) - train1_df$morekids)^2)
m2_rmse_train <- mean((predict(m2_vlm) - train1_df$morekids)^2)
m3_rmse_train <- mean((predict(m3_vlm) - train1_df$morekids)^2)

c(m1_rmse_train, m2_rmse_train, m3_rmse_train)
```

```
## [1] 0.2173874 0.2162571 0.2160741
```

```
c(m1_rmse_val, m2_rmse_val, m3_rmse_val)
```

```
## [1] 0.2175042 0.2163478 0.2162144
```

Well, I have to eat my words. Model 3 does better on both the training and validation sets. Guess it is capturing nonlinearities that exist in the data!

2 - Richer set of models for 3+ kids, using samesex as instrumental variable

So we saw in part 1 that turning education/age/age of first marriage into indicator variables provides us with more explanatory power in both training and validation data. Thus, for our more powerful model, we'll choose to do that. Before we get to the instrumental variables analysis, though, we need to first develop a good first-stage regression - and to do that, we need to consider adding in more control variables beyond education/age/age of first marriage. So, lets do that! We further consider the education and age of the father, as well as the race of the mother, to extend our Model 3.

(a) - extend model 3, conduct F-tests

We want to extend our model, but we also want to see if our extensions are significant. We accomplish this with some F-tests.

```
m3_ext_lm <- lm(morekids ~ factor(educm) + factor(agem) + factor(agefstm) + educd + aged + blackm + hisp

m3_ext_nodad_lm <- lm(morekids ~ factor(educm) + factor(agem) + factor(agefstm) + blackm + hispm + othra

m3_ext_norace_lm <- lm(morekids ~ factor(educm) + factor(agem) + factor(agefstm) + educd + aged, df)
```

Performing some F-tests using the built in anova() function reveals that taking out the 'dad' variables doesn't change the SSR in a statistically significant way, while taking out the 'race' variables does.

```
ftest1<-anova(m3_ext_lm, m3_ext_nodad_lm)
ftest2<-anova(m3_ext_lm, m3_ext_norace_lm)

ftest1
```

```
## Analysis of Variance Table
##
## Model 1: morekids ~ factor(educm) + factor(agem) + factor(agefstm) + educd +
##     aged + blackm + hispm + othracem
## Model 2: morekids ~ factor(educm) + factor(agem) + factor(agefstm) + blackm +
##     hispm + othracem
##   Res.Df   RSS Df Sum of Sq      F Pr(>F)
## 1 236401 50996
## 2 236403 50996 -2  -0.46012 1.0665 0.3442
```

There is not a significant difference for getting rid of dad variables.

```
ftest2
```

```
## Analysis of Variance Table
##
## Model 1: morekids ~ factor(educm) + factor(agem) + factor(agefstm) + educd +
##     aged + blackm + hispm + othracem
## Model 2: morekids ~ factor(educm) + factor(agem) + factor(agefstm) + educd +
##     aged
##   Res.Df   RSS Df Sum of Sq      F    Pr(>F)
## 1 236401 50996
## 2 236404 51097 -3   -101.17 156.32 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Now, there is a significant difference between the two models when the race/ethnicity variables are removed!

(b) Evaluating samesex as an instrument - examining the instrument relevance condition

Use same model as (a), but add samesex (the first two children being of the same sex) and re-estimate the model. This will give us our first-stage regression.

```
m3_ext_samesex_lm <- lm(morekids ~ factor(educm) + factor(agem) + factor(agefstm) + educd + aged + black
```

Some would say that families only care about having at least one son. In this scenario, if they have 2 girls (samesex=1, girls2=1) they are more likely to have more children (morekids=1) than if they had 2 boys (samesex=1, boys2=1). Being able to show something like this would indicate that samesex is a relevant instrument, meaning it does induce an effect on the variable we are instrumenting for (having 3+ kids).

Here, we can use a regression to test this claim. I find there is a significant difference in the impact of two daughters vs two sons. The test is conducted as follows:

$$y[morekids = 1] = \beta_0 + \beta' X + \beta_g girls2 + \beta_b boys2$$

If it is true that families only care about having 1 son, then we would expect families who have 2 boys to be less likely to have more children, and we would expect families with 2 girls to want to more children. In stats terms, our null hypothesis is $H_0 : \beta_g - \beta_b \leq 0$ (families are either ambivalent, or only care about having 1 girl), and our alternative hypothesis is $H_A : \beta_g - \beta_b > 0$. This is a ONE-TAILED test.

There is a weaker set of hypotheses

$$H_0 : \beta_g - \beta_b = 0$$

$$H_A : \beta_g - \beta_b \neq 0$$

. I am personally interested in testing this first, just to rule out the possibility that the coefficients are equal. Then we can retest a 1-tailed test to see if our stronger alternative hypothesis is true.

To evaluate the above hypotheses, we run the regression to obtain $diff = \hat{\beta}_g - \hat{\beta}_b$. To see if this difference is statistically significant, we find the standard error of the difference:

$$SE = \sqrt{Var(diff)} = \sqrt{Var(\hat{\beta}_g) + Var(\hat{\beta}_b) - 2 * Cov(\hat{\beta}_g, \hat{\beta}_b)}$$

where all the Variance and Covariance terms are reported from the regression. Then, we take the ratio of the difference to the standard error, and compare it against our desired t values for a 95% confidence interval in the two-tailed AND one-tailed cases.

```
m3_ext_bgtest_lm <- lm(morekids ~ factor(educm) + factor(agem) + factor(agefstm)
                       + educd + aged + blackm + hispm + othracem + girls2 + boys2, df)

diff <- m3_ext_bgtest_lm$coefficients["girls2"] - m3_ext_bgtest_lm$coefficients["boys2"]
vcov_mat <- vcov(m3_ext_bgtest_lm)
var_g <- vcov_mat["girls2", "girls2"]
var_b <- vcov_mat["boys2", "boys2"]
cov_gb <- vcov_mat["girls2", "boys2"]
se <- sqrt(var_g + var_b - 2*cov_gb)
unname(diff)
```

```
## [1] 0.02127117
```

```
unname(diff/se)
```

```
## [1] 7.925297
```

The difference between β_g and β_b is 0.02127117, and statistically significant at the $p = 0.001$ level for both the one-sided and two-sided t-tests (the t-statistic is 7.925297/2 and 7.925297 respectively). Further, the difference is positive, implying that the strong alternative hypothesis is true - if a family already has two boys, they are less likely to have more children than if a family has two girls. Samesex is clearly a relevant instrumental variable in that it does induce having more children.

(ii) Sex composition of children being random

Another necessary condition of our instrument is that it is exogenous to our model. In other words, we want the sex of children to be randomly assigned here, such that $samesex = 1$ couldn't be "explained" by any variables in the model. Statistically, this means that if I had a linear probability model for samesex, no matter what variables I included, I would expect the explanatory power (R-squared) to be essentially zero and for no coefficients to have a statistically significant effect. I would also expect the intercept to estimate the mean (0.5) if sexes were pretty much random. Finally, I would expect the F-statistic to be non-significant, meaning that we stick with the null hypothesis that "all coefficients in the model are equal to zero".

So, let's test this! We create a model for samesex and see if it is exogenous to the other variables in the model:

```r
sexrand1 <- lm(samesex ~ agem + aged + educm + educd + agefstm + agefstd + blackm + blackd + whitem + w
summary(sexrand1)
```

```
##
## Call:
## lm(formula = samesex ~ agem + aged + educm + educd + agefstm +
##     agefstd + blackm + blackd + whitem + whited + hispm + hispd,
##     data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.5286 -0.5056  0.4891  0.4944  0.5243
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  5.052e-01  1.297e-02  38.955   <2e-16 ***
## agem        -1.262e-03  1.891e-03  -0.668    0.504
## aged         5.748e-04  1.877e-03   0.306    0.759
## educm       -7.208e-05  5.818e-04  -0.124    0.901
## educd        2.247e-04  4.370e-04   0.514    0.607
## agefstm      1.399e-03  1.893e-03   0.739    0.460
## agefstd     -7.875e-04  1.883e-03  -0.418    0.676
## blackm      -1.391e-02  2.336e-02  -0.596    0.552
## blackd       1.567e-02  2.334e-02   0.671    0.502
## whitem      -1.706e-04  9.273e-03  -0.018    0.985
## whited       7.797e-03  9.628e-03   0.810    0.418
## hispm        1.039e-02  1.606e-02   0.647    0.518
## hispd       -4.727e-03  1.607e-02  -0.294    0.769
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5 on 236446 degrees of freedom
## Multiple R-squared:  4.332e-05,  Adjusted R-squared:  -7.428e-06
## F-statistic: 0.8536 on 12 and 236446 DF,  p-value: 0.5946
```

It is clear that there is no explanatory power (the R-squared values are incredibly low), the intercept is about 0.5 (implying that the mean of samesex is being estimated at 0.5), none of the coefficients are even close to significant, and the F-statistic is also insignificant. Thus, samesex appears random, not something that is correlated or explained with race/ethnicity or education or age or age when first conceiving.

3 - OLS vs IV models - effect of having 3+ children on the decision to work

Now, we have good reason to believe that having two children of the same sex (samesex=1) is a valid instrument for modeling the effect of having 3+ kids. Samesex appears randomly assigned (instrument exogeneity) and does seem to have an effect on having 3+ kids (instrument relevance). Let's use this instrument to estimate the effects of having more children on the mother's decision to work, and let's compare this to some basic OLS/linear probability models too!

(a) Linear probability models for the event that mom works

First, manually add some variables for our controls:

```
df$lowedm <- ifelse(df$educm < 12, 1, 0)
df$lowedd <- ifelse(df$educd < 12, 1, 0)
df$agem2 <- df$agem^2
df$aged2 <- df$aged^2
df$agefstm2 <- df$agefstm^2
df$agefstd2 <- df$agefstd^2
```

We construct some basic linear models, one of which is a straightforward regression of having 3+ kids on the decision to work, another of which includes many control regressors.

```
w1_lm <- lm(workedm ~ morekids, df)
w2_lm <- lm(workedm ~ educm + lowedm + educd + lowedd + agem + agem2 + agefstm +
              agefstm2 + aged + aged2 + agefstd + agefstd2 + blackm + hispm + othracem + morekids, df)
```

The effect size without controls (model w1_lm) of having morekids is -0.1134461. With controls (w2_lm), the effect size is -0.1585621. Thus, because these values are different, we conclude that model W1 is like a "short regression", with omitted variables being captured in the residual, and the coefficient estimate attempting to "explain" those omitted variables. W2 is the "long regression" where those omitted variables are present and the coefficient estimate loses some of that bias.

(b) Using samesex as an IV for a simple causal model

Now, let's use samesex as our instrumental variable. We first set up the first stage and reduced form regressions (for 2-stage least squares, which is like "manually" doing an instrumental variables regression); we also straightforwardly use an ivreg package in R just to show that the answers obtained from the 2SLS method will match the results from the package.

First stage, reduced form, and IV models using samesex as an IV.

```r
fs_ssiv_lm <- lm(morekids ~ samesex, df)
rf_ssiv_lm <- lm(workedm ~ samesex, df)
ssiv_lm <- ivreg(workedm ~ morekids | samesex, data=df)
```

The ratio of the reduced form coefficient (workedm regressed on samesex) and the first stage coefficient (morekids regressed on samesex) is

$$\frac{-0.0102412}{0.06735206} = -0.15205474$$

which matches the reported coefficient in the instrumental variables regression exactly.

Table below:

Table 1: First stage, reduced, and IV models

	Dependent variable:		
	morekids	workedm	
	OLS	*OLS*	*instrumental variable*
	(1)	(2)	(3)
samesex	0.06735206^{***} (0.001984425)	-0.0102412^{***} (0.002051684)	
morekids			-0.1520547^{***} (0.03030033)
Note:			*p<0.1; **p<0.05; ***p<0.01

(c) Using samesex as an IV for a model with many controls

Now let's go wild and add in all the control variables we can.

```
fs_ssiv_cont_lm <- lm(morekids ~ samesex + educm + lowedm + educd + lowedd + agem + agem2 + agefstm + ag
rf_ssiv_cont_lm <- lm(workedm ~ samesex + educm + lowedm + educd + lowedd + agem + agem2 + agefstm + age
ssiv_cont_lm <- ivreg(workedm ~ morekids + educm + lowedm + educd + lowedd + agem + agem2 + agefstm + ag
```

The ratio of the reduced form and the first stage is

$$\frac{-0.009463403}{0.06895031} = -0.137249608$$

which still matches the reported coefficient in the instrumental variables regression exactly. The effect sizes are slightly different, which we'll discuss later.

Table below:

Table 2: First stage, reduced, and IV models with controls

	Dependent variable:		
	morekids	workedm	
	OLS	*OLS*	*instrumental variable*
	(1)	(2)	(3)
samesex	0.06895031^{***} (0.001906481)	-0.009463403^{***} (0.00200602)	
morekids			-0.1372496^{***} (0.02876707)
Note:			*p<0.1; **p<0.05; ***p<0.01

(d) Proof that a truly random IV's estimate is approx the same with/without controls

So, we note that the effect sizes are different. Is this a result of random variation, or something we actually need to worry about? I prove a theoretical result here that a random IV's estimate should be the same in expectation, regardless of the inclusion of control variables.

To prove this, we must show $\hat{\beta}_{IV} = \frac{\hat{\delta}_z}{\hat{\pi}_z}$ is the same whether or not controls are included.

To show that, we individually show $\hat{\delta}_z$ (reduced form reg coef) is the same, with or without the controls. We then do the same for $\hat{\pi}_z$. Remember - assume z_i is orthog. to controls x_{0i}, implying $E[z_i \vec{x}_{0i}] = \vec{0}$!

① Show $\hat{\delta}_z$ is same, for no controls and controls:

NO CONTROLS:

$$y_i = \delta_0 + \delta_z z_i + \eta_i$$

By FOC (and derived a billion times in class already),

$$\boxed{\hat{\delta}_z = E[z_i z_i]^{-1} E[z_i y_i]}$$

CONTROLS

$$y_i = \delta_0 + \delta_z z_i + \delta'_{controls} x_{0i} + \eta_i$$

☆ can ignore $\hat{\delta}_0$ term WLOG because data can always be demeaned s.t. $\hat{\delta}_0 = 0$.

By FOC:

$$E[z_i \eta_i] = 0$$

$$E[z_i(y_i - (\hat{\delta}_z z_i + \vec{\hat{\delta}}_{cont.}^T \vec{x}_{0i}))] = 0$$

$$E[x_{0ji}(y_i - (\hat{\delta}_z z_i + \vec{\hat{\delta}}_{cont}^T \vec{x}_{0i}))] = 0 \quad \text{for all controls } x_{01i} \dots x_{0ji} \dots$$

We only care about this.

$$E[z_i y_i] - E[z_i(\hat{\delta}_z z_i + \vec{\delta}_{cont}^T \vec{x}_{0i})] = 0$$

$$\underbrace{E[z_i y_i] = \hat{\delta}_z E[z_i z_i]}_{\text{Re arrange}} + \underbrace{E[z_i \vec{\delta}_{cont}^T \vec{x}_{0i}]}_{= 0 \text{ since } E[z_i \vec{x}_{0i}] = \vec{0}\text{. Can Ignore this!}}$$

$$\boxed{\hat{\delta}_z = E[z_i z_i]^{-1} E[z_i y_i]}$$

Which is same as the no controls case!

2) Show $\hat{\pi}_z$ same for no cont. and cont.

NO CONTROLS

$$x_i = \pi_0 + \pi_z z_i + \xi_i$$

FOC leads to

$$\hat{\pi}_z = E[z_i z_i]^{-1} E[z_i x_i]$$

CONTROLS

$$x_i = \pi_0 + \pi_z z_i + \overrightarrow{\pi_{controls}}^T \overrightarrow{x_{0i}} + \xi_i$$

* can ignore $\hat{\pi}_0$ term WLOG because data can always be demeaned s.t. $\hat{\pi}_0 = 0$

By FOC:

$$E[z_i \xi_i] = 0$$

$$E[z_i(x_i - (\hat{\pi}_z z_i + \overrightarrow{\hat{\pi}_{cont}}^T \overrightarrow{x_{0i}}))] = 0$$

$$E[z_i x_i] = \hat{\pi}_z E[z_i z_i] + \underbrace{E[z_i \overrightarrow{\hat{\pi}_{cont}}^T \overrightarrow{x_{0i}}]}_{=0,\ \text{since } E[z_i \overrightarrow{x_{0i}}] = \vec{0}.}$$

$$\hat{\pi}_z = E[z_i z_i]^{-1} E[z_i x_i]$$

Same as the no controls case!

3) Thus, since $\hat{\delta}_z$ and $\hat{\pi}_z$ are same with or without controls, and

$\hat{\beta}_{IV} = \dfrac{\hat{\delta}_z}{\hat{\pi}_z}$, $\hat{\beta}_{IV}$ is same with or without controls.

Scanned by CamScanner

(e) Evaluate the difference in IV estimates in parts (b) and (c), based on part (d)

The IV estimates in (b) and (c) are not exactly equal. The question is, "does that difference matter? is it significant?" The theory says that these IV estimates are only equivalent in expectation, and only if instrumental variable *samesex* were uncorrelated with the controls (which we established in earlier parts by showing how none of those controls have a significant effect on samesex).

Beyond that theory, however, we should test whether the two estimates are statistically different. The first method I use is less formal and relies on an F-statistic, and the second is a formal test called the Wald test.

NOTE: This is actually a much, much harder and more important problem than it appears at a first glance - I could write a blog post about this. In summary, there isn't a good test to compare coefficients for the same variable across different models. If there were, we could answer really important questions, such as "does the effect size/coefficient of a variable change in a significant way if new variables M are added and variables N are removed?" In this case, the models we are comparing are NESTED (a special case in which variables are exclusively only added or removed) meaning we can use a Wald test. But the general form is really difficult to answer.

First, let us look at the relationship between *samesex* and the controls. If there is a statistically significant relationship between the two, we can assume that this correlation affected the estimates in a statistically significant way. Right off the bat, however, I'm guessing that there won't be much of a connection, because *samesex* is "truly randomly" assigned, as we saw in part 2(b)(iii). We observe the F-statistic for a model regressing *samesex* on the controls.

```
aux_iv_lm <- lm(samesex ~ educm + lowedm + educd + lowedd + agem + agem2 + agefstm + agefstm2 + aged +
summary(aux_iv_lm)
```

```
##
## Call:
## lm(formula = samesex ~ educm + lowedm + educd + lowedd + agem +
##     agem2 + agefstm + agefstm2 + aged + aged2 + agefstd + agefstd2 +
##     blackm + hispm + othracem, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.5328 -0.5050  0.4793  0.4949  0.5148
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  8.028e-01  8.184e-02   9.810   <2e-16 ***
## educm       -8.396e-04  6.949e-04  -1.208   0.2270
## lowedm      -8.240e-03  3.922e-03  -2.101   0.0356 *
## educd        2.252e-04  5.318e-04   0.424   0.6719
## lowedd      -9.648e-04  3.782e-03  -0.255   0.7987
## agem        -1.316e-02  6.146e-03  -2.141   0.0323 *
## agem2        2.051e-04  9.858e-05   2.080   0.0375 *
## agefstm     -1.894e-03  5.123e-03  -0.370   0.7116
## agefstm2     7.796e-05  1.109e-04   0.703   0.4822
## aged        -2.402e-03  4.484e-03  -0.536   0.5923
## aged2        4.370e-05  6.164e-05   0.709   0.4784
## agefstd     -2.811e-03  3.866e-03  -0.727   0.4671
## agefstd2     3.954e-05  6.854e-05   0.577   0.5640
## blackm      -5.898e-03  4.760e-03  -1.239   0.2153
## hispm        3.388e-05  6.745e-03   0.005   0.9960
## othracem    -5.266e-03  6.236e-03  -0.845   0.3984
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 0.5 on 236443 degrees of freedom
## Multiple R-squared:  0.0001031,  Adjusted R-squared:  3.967e-05
## F-statistic: 1.625 on 15 and 236443 DF,  p-value: 0.0589
```

Recall that the reported F-statistic below is the result of a test where the null hypothesis is "all coefficients are zero". Since the F-statistic is not significant, it appears that the relationship between *samesex* and these controls is mostly spurious.

What this means is that our assumption from part (d) - namely, $E[z_i * x_{Oi}] = 0$ is probably true - however, since this is true only in expectation, we probably got some random variation to make it != 0, though not in any significant way.

As for individual terms, it appears that *samesex* is correlated with $educm < 12$, and both terms for the mother's age. However, the super low R-squared values also imply that these controls don't really explain much of the variance in *samesex*.

You can also think of it as, we are testing the hypothesis $\beta_{samesex,controls} - \beta_{samesex,nocontrol} = 0$. However, this can't be formally tested with a good ol' t-test; it requires the standard error of the difference, which is the square root of

$$Var(\beta_{ss,c} - \beta_{ss,nc}) = Var(\beta_{ss,c}) + Var(\beta_{ss,nc}) - 2 * Cov(\beta_{ss,c}, \beta_{ss,nc})$$

. Though the variances can be retrieved from the original regressions, the covariance cannot be.

HOWEVER, this is what is known as a "nested model", and coefficients can be compared with a Wald test. The Wald test asks the basic question, "does reducing these other parameters to zero significantly reduce the model fit?" This is fortunately easy in R!

```r
waldtest(ssiv_cont_lm,ssiv_lm)
```

```
## Wald test
##
## Model 1: workedm ~ morekids + educm + lowedm + educd + lowedd + agem +
##     agem2 + agefstm + agefstm2 + aged + aged2 + agefstd + agefstd2 +
##     blackm + hispm + othracem | samesex + educm + lowedm + educd +
##     lowedd + agem + agem2 + agefstm + agefstm2 + aged + aged2 +
##     agefstd + agefstd2 + blackm + hispm + othracem
## Model 2: workedm ~ morekids | samesex
##   Res.Df  Df  Chisq Pr(>Chisq)
## 1 236442
## 2 236457 -15 8168.9  < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The results above show that only the "morekids" variable is statistically significant. Thus, we conclude that the difference between the "base" model and the larger model it is nested in is a significant difference.

(f) Some observational comparisons as evidence for the effect of having extra kids on the decision to work

Based on the OLS and IV models, I conclude that having more kids does have a causal effect on a mother's decision to work. This explains a lot of the observational comparisons between mothers with 2 versus 3+ kids. Let's compare the mean values for mothers with 2 kids versus mothers with 3+ kids by subsetting the dataframe and subtracting the column means for some variables we are interested in.

```
withmorekids = select(subset(df, morekids==1), educm, workedm, hrsweekm, annhrsm, earningsm, faminc, exp
withlesskids = select(subset(df, morekids==0), educm, workedm, hrsweekm, annhrsm, earningsm, faminc, exp

colMeans(withmorekids) - colMeans(withlesskids)
```

```
##         educm       workedm      hrsweekm       annhrsm     earningsm 
## -6.338885e-01 -1.134461e-01 -3.729825e+00 -1.832065e+02 -3.662022e+03 
##        faminc          expm        lowedm 
## -5.403542e+03  1.324117e+00  9.606827e-02
```

It appears from the output above that if you have more kids, education/propensity to work/hours worked/earnings all tend to be lower (the average is lower for people with morekids than lesskids, so the difference is negative). The mothers with morekids tend to have more years of work after completing school though, and a higher propensity to have low education (the positive values).

4 - compare OLS and IV models for morekids effect on earnings

We've set up all this theoretical groundwork, so now let's have some fun. We go wild in this section and construct a bunch of different models for the effect of having more kids on mother's earnings, dad's earnings, total family earnings, etc.

(a) - estimate 2 OLS and 2 IV models for mother's, father's, and total earnings.

```
earnm_ols <- lm(earningsm ~ morekids, df)
earnm_cont_ols <- lm(earningsm ~ morekids + educm + lowedm + educd + lowedd + agem + agem2 + agefstm + a

earnd_ols <- lm(earningsd ~ morekids, df)
earnd_cont_ols <- lm(earningsd ~ morekids + educm + lowedm + educd + lowedd + agem + agem2 + agefstm + a

famearn_ols <- lm(famearn ~ morekids, df)
famearn_cont_ols <- lm(famearn ~ morekids + educm + lowedm + educd + lowedd + agem + agem2 + agefstm + a


earnm_iv <- ivreg(earningsm ~ morekids|samesex, data=df)
earnm_cont_iv <- ivreg(earningsm ~ morekids + educm + lowedm + educd + lowedd + agem + agem2 + agefstm +

earnd_iv <- ivreg(earningsd ~ morekids|samesex, data=df)
earnd_cont_iv <- ivreg(earningsd ~ morekids + educm + lowedm + educd + lowedd + agem + agem2 + agefstm +

famearn_iv <- ivreg(famearn ~ morekids|samesex, data=df)
famearn_cont_iv <- ivreg(famearn ~ morekids + educm + lowedm + educd + lowedd + agem + agem2 + agefstm +
```

A table of the results is presented here:

Table 3: Effects of morekids on mother's, father's, and family earnings for no control and control models

	No Controls		Controls	
	OLS	*IV*	*OLS*	*IV*
	(1)	(2)	(3)	(4)
earningsm	-3662.02 ***	−3654.2***	−4882.4582***	−3097.6551 ***
	(<2e-16)	(8.03e-05)	(< 2e-16)	(0.000372)
earningsd	-1779.1 ***	−3155.8	−74.193	−3596.780 *
	(<2e-16)	(0.118)	(0.5761)	(0.04470)
famearns	-5441.16 ***	−6810.1**	−4956.651***	−6694.435 ***
	(<2e-16)	(0.00146)	(< 2e-16)	(0.000392)

Note: *p<0.1; **p<0.05; ***p<0.01

The IV estimate is less negative than the OLS estimate for mothers because of the “direction” of the omitted variables bias.

Specifically, recall that IV estimates are used when there is correlation between the causal variable of interest and the error term, causing the OLS estimate to be biased. In math terms, for a regression $y_i = \beta_0 + \beta_1 x_i + \eta_i$, our estimate $\hat{\beta_1}$ will be as follows

$$\hat{\beta_1} = \beta_1 + \frac{\sum(x_i - \bar{x})\eta_i}{\sum(x_i - \bar{x})x_i}$$

However, that last term might be nonzero if there is some omitted variable, leading to bias. Specifically, lets say the real model is $y_i = \beta_0 + \beta_1 x_i + \beta_2 w_i + u_i$. This means $\eta_i = \beta_2 w_i + u_i$, and when we plug back our value for η_i, we get

$$\begin{aligned}\hat{\beta_1} &= \beta_1 + \frac{\sum(x_i - \bar{x})(\beta_2 w_i + u_i)}{\sum(x_i - \bar{x})x_i} \\ &= \beta_1 + \beta_2 \frac{\sum(x_i - \bar{x})w_i}{\sum(x_i - \bar{x})x_i} + \frac{\sum(x_i - \bar{x})u_i}{\sum(x_i - \bar{x})x_i}\end{aligned}$$

In this “true” model, the last term is “truly” zero in expectation, but the second term is non-zero. The question is, **is the second term positive or negative?**

The sign of the omitted variable bias in the formula above determines whether the IV estimate (which gets rid of that bias) is higher or lower than the biased OLS estimate. In our regression, there is some variable w_i which correlates with *morekids* DIFFERENTLY for men vs women, meaning $\sum x_i w_i > 0$ for one, and is < 0 for another. I hypothesize it could be something like, parent’s earnings? Perhaps if you were a well-off male as a child, you may choose to have more kids just because you can, whereas if you were a better off female as a child, you would have fewer kids due to more opportunities for career pursuits/different values.

A separate question came up to me in this process. Family earnings is just (at least in these “typical” families which construct the dataset) dad’s earnings + mom’s earnings. Can I get the effect on family’s earnings just by estimating the effects of the dad’s earnings, the mom’s earnings, and adding the effect sizes together?

(b) - proof that coefficient for the sum of two dependent variables can be estimated using the separate dependent variables

The answer to that question is yes. Proof is below, on two pages.

Let $y_{3i} = y_{2i} + y_{1i}$ (as given).

Note: that for

$$y_{3i} = x_i \beta_3 + u_{3i}$$

$$\hat{\beta}_3 = E[x_i x_i']^{-1} E[x_i y_{3i}] \text{ by the FOC.}$$

Then, plug in $y_{3i} = y_{2i} + y_{1i}$

$$\hat{\beta}_3 = E[x_i x_i']^{-1} E[x_i (y_{2i} + y_{1i})]$$

$$\hat{\beta}_3 = E[x_i x_i']^{-1} \left(E[x_i y_{2i}] + E[x_i y_{1i}] \right)$$

$$\hat{\beta}_3 = \underbrace{E[x_i x_i']^{-1} E[x_i y_{2i}]}_{= \hat{\beta}_2 \text{ by FOC for } \hat{\beta}_2} + \underbrace{E[x_i x_i']^{-1} E[x_i y_{1i}]}_{= \hat{\beta}_1 \text{ by FOC for } \hat{\beta}_1}$$

$$\hat{\beta}_3 = \hat{\beta}_2 + \hat{\beta}_1 \quad !$$

(c) - verify that estimated OLS effect on mother/dad earnings add up to the OLS effect on family earnings.

The results from (b) indicate that since $famearn = earningsm + earningsd$, then $\beta_{morekids,fam} = \beta_{morekids,m} + \beta_{morekids,d}$, and the fraction $\frac{\beta_{morekids,m}}{\beta_{morekids,fam}}$ defines the percentage of the family effect composed of the effect on mothers. We calculate this for both the OLS regressions, with and without controls, and show this to be true.

```r
beta_fam <- famearn_ols$coefficients["morekids"]
beta_m <- earnm_ols$coefficients["morekids"]
beta_d <- earnd_ols$coefficients["morekids"]

beta_fam_cont <- famearn_cont_ols$coefficients["morekids"]
beta_m_cont <- earnm_cont_ols$coefficients["morekids"]
beta_d_cont <- earnd_cont_ols$coefficients["morekids"]

c(beta_fam, beta_m, beta_d)
```

```
##  morekids  morekids  morekids
## -5441.158 -3662.022 -1779.136
```

```r
c(beta_fam_cont, beta_m_cont, beta_d_cont)
```

```
##    morekids    morekids    morekids
## -4956.65107 -4882.45821   -74.19286
```

By looking at the output above, we verify that $-5441.158 = -3662.022 + -1779.136$ and $-4956.65107 = -4882.45821 + -74.19286$.

We see that with and without controls, β_m is way more influential. Without controls, the effect of having morekids on family earnings is $3662.022/5441.158 = 67\%$ driven by the kids' effect on the mother's earnings. With controls, the effect of morekids on family earnings is $4882.45821/4956.65107 = 98.5\%$ driven by the kids' effect on the mother.

(d) - repeating (c), but for the IV estimates this time

The logic earlier still holds for the IV case. Proof below, on a new page:

we now want to show

$$\hat{\beta}_{3IV} = \hat{\beta}_{2IV} + \hat{\beta}_{1IV}$$

$\hat{\beta}_{3IV} = \frac{\hat{\delta}_3}{\hat{\pi}_3}$, and similarly for β_2. and β_1, so: showing this is equivalent.

$$\frac{\hat{\delta}_3}{\hat{\pi}_3} = \frac{\hat{\delta}_2}{\hat{\pi}_2} + \frac{\hat{\delta}_1}{\hat{\pi}_1}$$

Note that $\hat{\pi}_3 = \hat{\pi}_2 = \hat{\pi}_1$, because they are defined by $x_i = \pi_0 + \pi_1 z_i + \eta_i$. This doesn't change across $\hat{\pi}_{1,2,3}$.

Thus, it is equivalent to show

$\hat{\delta}_3 = \hat{\delta}_2 + \hat{\delta}_1$, which is exactly like what we proved in part (b), just with changed variable names. The β was swapped for δ, and the x was swapped for z, but the proof still holds.

Proof for (b) proves $\hat{\delta}_3 = \hat{\delta}_2 + \hat{\delta}_1$, proving that $\hat{\beta}_{3_{IV}} = \hat{\beta}_{2_{IV}} + \hat{\beta}_{1_{IV}}$ ✓.

Now, verify that this is true with the coefficients we have.

```r
beta_fam <- famearn_iv$coefficients["morekids"]
beta_m <- earnm_iv$coefficients["morekids"]
beta_d <- earnd_iv$coefficients["morekids"]

beta_fam_cont <- famearn_cont_iv$coefficients["morekids"]
beta_m_cont <- earnm_cont_iv$coefficients["morekids"]
beta_d_cont <- earnd_cont_iv$coefficients["morekids"]

c(beta_fam, beta_m, beta_d)
```

```
##  morekids  morekids  morekids
## -6810.088 -3654.247 -3155.842
```

```r
c(beta_fam_cont, beta_m_cont, beta_d_cont)
```

```
##  morekids  morekids  morekids
## -6694.435 -3097.655 -3596.780
```

By looking at the output above, we verify that $-6810.088 = -3654.247 + -3155.842$ and $-6694.435 = -3097.655 + -3596.780 $.

We see that with and without controls, β_m is way more influential. Without controls, the effect of having morekids on family earnings is $3654.247/6810.088 = 53.66\%$ driven by the kids' effect on the mother's earnings. With controls, the effect of morekids on family earnings is $3097.655/6694.435 = 46.27\%$ driven by the kids' effect on the mother.

(e) Complier Average Causal Effects framework

Great. We have these effect sizes. But, there are other fun questions to answer! Very relevant to policy is, "Are these effect sizes the same for all people?" It's a very important question to ask; one might find that even though the average effect of some variable X is some number Beta (what we've done above), the effect of X on different types of people can differ drastically.

We can use the Complier Average Causal Effects framework (CACE) to break down how the effect sizes differ across different family/mother demographics. This framework thinks of the instrument (samesex) as an assignment of a subject to some group. You're either in the samesex=1 group, or not. It then sees the dependent variable (morekids) as a treatment that people in these groups have to do.

Within these groups, there is the existence of Always-Takers (ATs), Never-Takers (NTs), and Compliers (Cs). The always-takers are people who, even if they weren't assigned the samesex group, would have gone and gotten the treatment (had more kids) anyways. In any universe, they would have chosen to have more kids. The never-takers are people who would never have more kids, regardless of what group they were assigned to (whether samesex=1 or not). And the compliers are the people who act according to the assignment; if assigned samesex=1, they would have more kids, and if not, then they'd stick to having less kids.

The proportion of AT/NT/Cs is pretty important, and knowing these proportions can help us answer a lot of other interesting questions. The "Compliers" are people we are usually interested in from a policy standpoint, in which we can enforce compliance. The next section is just calculation of these proportions and honestly incredibly boring, so skip past those if you'd like. Part (iv) is where it gets cool.

In the CACE framework, we look initially at the first-stage regression without controls (the regression of morekids on samesex, run in the earlier question 3(b)). This regression is

$$MoreKids_i = \beta_0 + \beta_1 * SameSex_i + u_i$$

.

To get the AT/NT/C, we need to look at values of $P(AT) = E[MoreKids_i|SameSex_i = 0]$, $P(C) = E[MoreKids_i|SameSex_i = 1] - E[MoreKids_i|SameSex_i = 0]$, and $P(NT) = 1 - (P(AT) + P(C))$. These values can be calculated by taking means on subsets of our population, or by using the basic first-stage regressions. I will show that these return the same value.

(i) fraction of AT (always-takers)

In the first stage regression, we know that $E[MoreKids_i|SameSex_i = 0] = \pi_0$, since the regression intercept fits the mean of the datapoints for which $SameSex_i = 0$. So, we compare this regression coefficient (calculated earlier in fs_ssiv_lm, in part 3(b)) with the "manual method" of taking the mean $MoreKids_i$ value on the $SameSex = 0$ subset of our data.

```
manual_at <- mean(subset(df, samesex==0)$morekids)
reg_at <- unname(fs_ssiv_lm$coefficients["(Intercept)"])
manual_at
```

```
## [1] 0.3390693
```

```
reg_at
```

```
## [1] 0.3390693
```

33.9% are always-takers. Both methods are equal!

(ii) fraction of NT (never-takers) Repeat the above. This time, the never-takers are all the people who are neither compliers or always-takers. Using the next part (where we calculate the compliers), $Pr(NT) = 1 - (Pr(AT) + Pr(C)) = 0.59357864$.

(iii) fraction of C (compliers)

In the first stage regression, we know that $E[MoreKids_i|SameSex_i = 1] - E[MoreKids_i|SameSex_i = 0] = \pi_1$. The interpretation compliers are the people who we expect to have more kids iff they had 2 of the samesex. Like in part (i), compare the regression method with the manual subset method.

```
manual_c <- mean(subset(df, samesex==1)$morekids) - mean(subset(df, samesex==0)$morekids)
reg_c <- unname(fs_ssiv_lm$coefficients["samesex"])
manual_c
```

```
## [1] 0.06735206
```

```
reg_c
```

```
## [1] 0.06735206
```

(iv) Compare AT/NT/C for four subgroups, based on mother's education level (<12, =12, b/w 13 and 15, and >16)

This is the cool stuff. We are now asking: given a mother's education level, what's the proportion who will "comply" (have more kids if their first two kids are the same sex)? What's the proportion who will always have more kids, no matter what? Etc. The R below is very simple; I just take subsets of the data based on these education levels and compute some means, which I then interpret:

```
at_le12 <- mean(subset(df, (educm < 12) & (samesex == 0))$morekids)
c_le12 <- mean(subset(df, (educm < 12) & (samesex==1))$morekids) - at_le12
nt_le12 <- 1 - (at_le12 + c_le12)
le12 <- c(at_le12, c_le12, nt_le12)
names(le12) <- c("AT <12", "C <12", "NT <12")

at_e12 <- mean(subset(df, (educm == 12) & (samesex == 0))$morekids)
c_e12 <- mean(subset(df, (educm == 12) & (samesex==1))$morekids) - at_e12
nt_e12 <- 1 - (at_e12 + c_e12)
e12 <- c(at_e12, c_e12, nt_e12)
names(e12) <- c("AT =12", "C =12", "NT =12")

at_1315 <- mean(subset(df, (educm > 13) & (educm < 15) & (samesex == 0))$morekids)
c_1315 <- mean(subset(df, (educm > 13) & (educm < 15) & (samesex==1))$morekids) - at_1315
nt_1315 <- 1 - (at_1315 + c_1315)
e1315 <- c(at_1315, c_1315, nt_1315)
names(e1315) <- c("AT >13<15", "C >13<15", "NT >13<15")

at_g16 <- mean(subset(df, (educm >= 16) & (samesex == 0))$morekids)
c_g16 <- mean(subset(df, (educm >= 16) & (samesex==1))$morekids) - at_g16
nt_g16 <- 1 - (at_g16 + c_g16)
g16 <- c(at_g16, c_g16, nt_g16)
names(g16) <- c("AT >16", "C >16", "NT >16")

le12
```

```
##     AT <12      C <12     NT <12
## 0.46863377 0.07180795 0.45955828
```

```
e12
```

```
##     AT =12      C =12     NT =12
## 0.32562136 0.07144493 0.60293371
```

```
e1315
```

```
## AT >13<15  C >13<15 NT >13<15
## 0.3000108 0.0784443 0.6215449
```

```
g16
```

```
##     AT >16      C >16     NT >16
## 0.2652341 0.0449317 0.6898342
```

From the results above, an interesting trend appears to be here - if the mother's education is low, they are more likely to always have more kids regardless of if their first children are the same sex (be an always-taker), and as they have more education they are less likely to be always-takers. They are also less likely to be compliers and more likely to be never-takers if they have higher education. What I take from this is that the more education the mother has, the less likely that having two kids of the samesex will cause them to have more.

HOWEVER, I doubt these results - deniers (a 4th mysterious group! These people do the opposite of what they are assigned to do; if they are assigned samesex=1, they will instead choose to stop having kids, and vice versa) do exist within this dataset. It is incredibly plausible that if the first two kids are the same sex, the parents decide to not have more children - either because they wanted two children of the same sex, or they decided that having 2 boys was absolutely intolerable and they couldn't handle any more. The CACE framework has an inconvenient assumption that deniers do not exist, so these results aren't really airtight.

(f) - Calculate means for overall set of compliers using wonky ivreg method

Finally, we get to some SUPER powerful stuff you can do with CACE. You can find the demographic characteristics of compliant mothers!

For any arbitrary variable x_i, we want to do an instrumental variables regression in which we are interested in $x_i * morekids_i$ as an outcome variable, and $samesex_i$ as the instrumental variable. We calculate these means for compliers and compare them to the means for all families.

Fraction of complier mothers with $<$ 12 years of schooling is 0.1598666, as calculated below. Compared to the average fraction of all mothers with $<$ 12 years of schooling (0.1708711), compliant mothers are more likely to have less schooling.

```
df$lowedmmorekidsint = df$lowedm*df$morekids
modlowedm <- ivreg(lowedmmorekidsint ~ morekids | samesex, data=df)
unname(modlowedm$coefficients[2])
```

```
## [1] 0.1598666
```

```
mean(df$lowedm)
```

```
## [1] 0.1708711
```

Mean education of complier mothers is 12.2245 years, as seen below. This is slightly less than the overall mean education of mothers, which is 12.42263.

```
df$educmmorekidsint = df$educm*df$morekids
modeducm <- ivreg(educmmorekidsint ~ morekids | samesex, data=df)
unname(modeducm$coefficients[2])
```

```
## [1] 12.2245
```

```
mean(df$educm)
```

```
## [1] 12.42263
```

Mean age of first childbirth of complier mothers is 20.64157. This is less than the mean first childbirth age for all mothers, which is 20.84279.

```
df$agefstmmorekidsint = df$agefstm*df$morekids
modagefstm <- ivreg(agefstmmorekidsint ~ morekids | samesex, data=df)
unname(modagefstm$coefficients[2])
```

```
## [1] 20.64157
```

```
mean(df$agefstm)
```

```
## [1] 20.84279
```

Fraction of complier mothers who had their first child before 21 is 0.5269768. This is greater than the fraction of all mothers with a first child before 21, which is 0.4971686. This backs up the previous calculation - compliers tend to have given birth earlier.

```
df$bef21 <- ifelse(df$agefstm < 21, 1, 0)
df$bef21morekidsint = df$bef21*df$morekids
modbef21 <- ivreg(bef21morekidsint ~ morekids | samesex, data=df)
unname(modbef21$coefficients[2])
```

```
## [1] 0.5269768
```

```
mean(df$bef21)
```

```
## [1] 0.4971686
```

Fraction of complier mothers who are hispanic is 0.01996363. Compared to the fraction of all mothers who are hispanic, 0.02509526, it seems that the complier population is less likely to be hispanic.

```r
df$hispmorekidsint = df$hispm*df$morekids
modhisp <- ivreg(hispmorekidsint ~ morekids | samesex, data=df)
unname(modhisp$coefficients[2])
```

```
## [1] 0.01996363
```

```r
mean(df$hispm)
```

```
## [1] 0.02509526
```

Fraction of complier mothers who are black is 0.0275295. Compared to the fraction of all mothers who are black, 0.04971264, it seems that the complier population is less likely to be black

```r
df$blackmmorekidsint = df$blackm*df$morekids
modblack <- ivreg(blackmmorekidsint ~ morekids | samesex, data=df)
unname(modblack$coefficients[2])
```

```
## [1] 0.0275295
```

```r
mean(df$blackm)
```

```
## [1] 0.04971264
```

Fraction of complier mothers who are white non-hispanic is 0.9269914. Compared to the fraction of all mothers who are white non-hispanic, 0.8967263, it seems that the complier population is more likely to be white than the regular population.

```r
df$wnhmorekidsint = df$wnhm*df$morekids
modwnh <- ivreg(wnhmorekidsint ~ morekids | samesex, data=df)
unname(modwnh$coefficients[2])
```

```
## [1] 0.9269914
```

```r
mean(df$wnhm)
```

```
## [1] 0.8967263
```

Fraction of complier mothers from Utah is 0.007350648, compared to fraction of all mothers from utah, 0.007350648.

```r
df$utah <- ifelse(df$st == 87, 1, 0)
df$utahmorekidsint = df$utah*df$morekids
modutah <- ivreg(utahmorekidsint ~ morekids | samesex, data=df)
unname(modutah$coefficients[2])
```

```
## [1] 0.007350648
```

```r
mean(df$utah)
```

```
## [1] 0.01051768
```

So, the grand question - what does it all mean?

The “complier” population are those who, upon having 2 kids of the samesex, actually will “comply” and have a 3rd child or more. Some people are more or less likely to comply than one would expect. We can come up with any amount of interesting hypothetical reasonings for these different numbers; maybe white people tend to be compliers in higher proportions than their population because of some cultural factors. None of that is really provable. But it is important to see empirically. Beyond this specific narrow research question, the CACE has application for evaluating instruments such as, “X type of workplace health insurance plan is offered with fewer benefits”. How many people will “comply” with this assignment and consequently seek out certain types of private insurance? Maybe people in certain states, or of certain ages, are more likely to? The CACE framework attempts to answer those things.