ECON C142 FINAL PROJECT

```
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

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

Part 1

1 - Models for having 3+ kids

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)
```

Calculating r-square, adj-r-square, and rmse vals

```
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.sq
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) - difference in probabilities for age 35 vs 30, and associated standard errors

Model 1:

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.

```r
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
```

```r
i_se_m3
```

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

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

(ii) - difference in probabilities for 16 vs 12 years of education, and associated standard errors

The methods are the same as above, just with different random variables/coefficients. The only difference is that the Standard Error of the difference for model 2 must be recomputed with the same method as model 3 in part (i). So, rederivations are not done.

Model 1:

```
ii_m1 <- m1_lm$coefficients["educm"]*(16-12)
ii_se_m1 <- 4*summary(m1_lm)$coefficients["educm", "Std. Error"]

ii_m1
```

```
##       educm
## -0.05464697
```

```
ii_se_m1
```

```
## [1] 0.00180017
```

For M1, the difference is -0.05464697 and the SE is 0.00180017.

Model 2:

```
ii_m2 <- m2_lm$coefficients["factor(educm)16"] - m2_lm$coefficients["factor(educm)12"]
vcov_mat <- vcov(m2_lm)
var_x16 <- vcov_mat["factor(educm)16", "factor(educm)16"]
var_x12 <- vcov_mat["factor(educm)12", "factor(educm)12"]
cov_x16_x12 <- vcov_mat["factor(educm)16", "factor(educm)12"]

ii_se_m2 <- sqrt(var_x16 + var_x12 - 2*cov_x16_x12)

ii_m2
```

```
## factor(educm)16
##      0.02868078
```

```
ii_se_m2
```

```
## [1] 0.003687303
```

For M2, the difference is 0.02868078 and the SE is 0.003687303.

Model 3:

```
ii_m3 <- m3_lm$coefficients["factor(educm)16"] - m3_lm$coefficients["factor(educm)12"]
vcov_mat <- vcov(m3_lm)
var_x16 <- vcov_mat["factor(educm)16", "factor(educm)16"]
var_x12 <- vcov_mat["factor(educm)12", "factor(educm)12"]
cov_x16_x12 <- vcov_mat["factor(educm)16", "factor(educm)12"]

ii_se_m3 <- sqrt(var_x16 + var_x12 - 2*cov_x16_x12)

ii_m3
```

```
## factor(educm)16
##      0.03305081
```

```
ii_se_m3
```

```
## [1] 0.003730534
```

For M3, the difference is 0.03305081 and the SE is 0.003730534.

(iii) Differences for first child at age 20 vs 25

Same stuff as section (ii), just with different variables.

Model 1:

```
iii_m1 <- m1_lm$coefficients["agefstm"]*(25-20)
iii_se_m1 <- 5*summary(m1_lm)$coefficients["agefstm", "Std. Error"]

iii_m1
```

```
##    agefstm
## -0.1924384
```

```
iii_se_m1
```

```
## [1] 0.001976325
```

For M1, the difference is -0.1924384 and the SE is 0.001976325

Model 2:

```
iii_m2 <- m2_lm$coefficients["agefstm"]*(25-20)
iii_se_m2 <- 5*summary(m2_lm)$coefficients["agefstm", "Std. Error"]

iii_m2
```

```
##    agefstm
## -0.2032896
```

```
iii_se_m2
```

```
## [1] 0.002052919
```

For M2, the difference is -0.2032896 and the SE is 0.002052919

Model 3:

```
iii_m3 <- m3_lm$coefficients["factor(agefstm)25"] - m3_lm$coefficients["factor(agefstm)20"]
vcov_mat <- vcov(m3_lm)
var_x25 <- vcov_mat["factor(agefstm)25", "factor(agefstm)25"]
var_x20 <- vcov_mat["factor(agefstm)20", "factor(agefstm)20"]
cov_x25_x20 <- vcov_mat["factor(agefstm)25", "factor(agefstm)20"]

iii_se_m3 <- sqrt(var_x25 + var_x20 - 2*cov_x25_x20)

iii_m3
```

```
## factor(agefstm)25
##        -0.2219466
```

```
iii_se_m3
```

```
## [1] 0.005244743
```

For M3, the difference is -0.2219466 and the SE is 0.005244743

Graphs

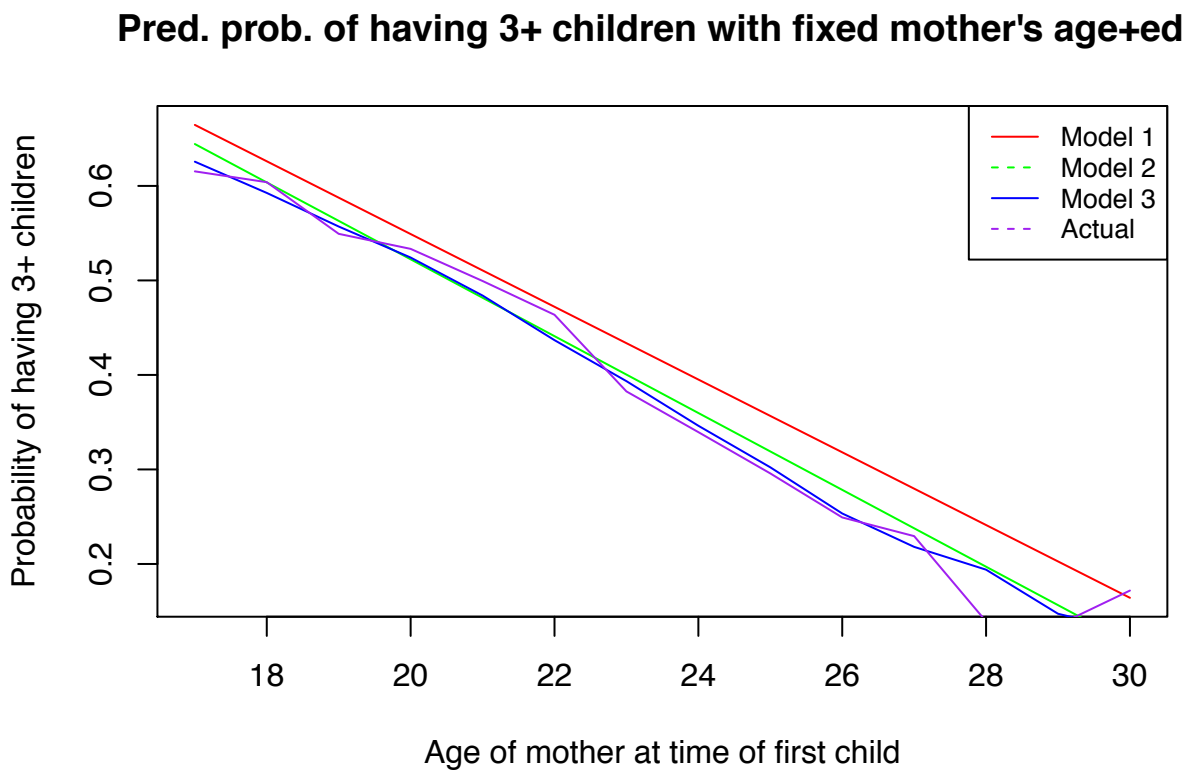
Graphing the actual and predicted probabilities for M1, M2, M3 of having >3 kids for 35 y/o mothers with 12 years of education, over agefstm values 17-30.

Computing the predicted probabilities:

```
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)
}
```

Graphing the actual predicted 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)
```



A comparison of the models:

In comparing M1, M2, and M3, we first theoretically consider what "binning" generally does and how it relates to our estimation of the CEF. 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 saw in lecture 3 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 differ, so it is important to consider the linearity of the population CEF versus the linearity of the sample CEF.

Now, apply this general theory to our situation.

1. Is the effect of 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"). Thus, M3 estimates a nonlinear model. 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, there is no way to argue this analytically - we can just test that thought empirically below.

Last is the 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 morekids, using samesex as instrumental variable

(a) - extend model 3, conduct 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 between model 1 and model 2 (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) using samesex as an "exogenous" determinant of family size

Use same model as (a), but add samesex and re-estimate the model.

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

(i) the average effect of having first two children of the same sex on the probability that morekids=1 is seen by looking at the coefficient on samesex = 0.06880994

(ii) test the claim that families only care about having at least 1 son:

I find there is a significant difference in the impact of two daughters vs two sons. To test this claim, we specify a regression 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.

(iii) sex composition of children being random

If the gender of children is truly random, then I would expect that $samesex = 1$ couldn't be easily explained. In a linear probability model, 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". The question asks to include variables about age, education, race, and ages of first having children, so I include those in the model.

```
sexrand1 <- lm(samesex ~ agem + aged + educm + educd + agefstm + agefstd + blackm + blackd + whitem + wh
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 children on decision to work

(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
```

Comparing OLS models W1 and W2 (not including and including controls)

```
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 (W1) of having morekids is -0.1134461. With controls (W2), 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

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

```
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***	−0.0102412***	
	(0.001984425)	(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

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

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***	−0.009463403***	
	(0.001906481)	(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

Proof on next page.

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_{oi}, implying $E[z_i \vec{x}_{oi}] = \vec{0}$!

(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),

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

CONTROLS

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

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

By FOC:

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

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

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

We only care about this.

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

$$E[z_i y_i] = \hat{\delta}_z E[z_i z_i] + E[z_i \vec{\delta}_{cont}^{\,T} \vec{x}_{oi}]$$

$= 0$, since $E[z_i \vec{x}_{oi}] = \vec{0}$. Can ignore this!

Rearrange

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

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

$$x_i = \pi_0 + \pi_z z_i + \vec{\pi}_{controls}^{\,T} \vec{x}_{0i} + \xi_i$$

By FOC:

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

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

$$E[z_i x_i] = \hat{\pi}_z E[z_i z_i] + E[z_i \hat{\vec{\pi}}_{cont}^{\,T} \vec{x}_{0i}]$$

$\hat{\pi}_z = E[z_i z_i]^{-1} E[z_i x_i]$ $\quad = 0$, since $E[z_i \vec{x}_{0i}] = \vec{0}$.

Same as the no controls case!

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

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

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

(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 theory says that they would be equal only if instrumental variable *samesex* were uncorrelated with the controls; thus, we can infer that *samesex* is correlated somehow with the controls.

I can think of two ways to test whether the two estimates are statistically different - the first is less formal and relies on an F-statistic, and the second is a formal test called the Wald test.

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 + a
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!

```
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) relationship between causal effect of having extra kids on the decision to work, and observational comparisons

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

(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. (below is a basic explanation of how this manifests, which has been gone over in lecture; I'm not sure how thorough we need to be on the project, so skip ahead 1 paragraph or so if this is not necessary).

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

$$\hat{\beta}_1 = \beta_1 + \frac{\sum(x_i - \bar{x})(\beta_2 w_i + u_i)}{\sum(x_i - \bar{x})x_i}$$

$$= \beta_1 + \beta_2 \frac{\sum(x_i - \bar{x})w_i}{\sum(x_i - \bar{x})x_i} + \frac{\sum(x_i - \bar{x})u_i}{\sum(x_i - \bar{x})x_i}$$

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.

(b) - proof that adding together dependent variables can be estimated using the separate dependent variables

Proof is on page below.

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

Note that for

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

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

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 result from (b) indicates 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.

```
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
```

```
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_1 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.

```
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
```

```
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) Stuff about always-takers, compliers, never-takers, etc.

We need to look 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) I just use the manual subset methods for convenience's sake.

```
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)
```

```r
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
```

```r
e12
```

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

```r
e1315
```

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

```r
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 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.

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

As seen in lecture, 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.

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

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

```
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

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

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

```
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.

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

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

```
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.

```
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
```

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

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

Part 2

Train vs Holdout sampling, construct logwage variable

```
df$logwaged <- log(df$waged)
train_df <- subset(df, (rv < 0.75) & (educd==16))
holdout_df <- subset(df, (rv >= 0.75) & (educd==16))
```

1 - OLS model

I predict the wages for 26 and 35 year olds across states. The predicted logwaged for the 26 year olds and the 35 year olds are presented below in alphabetical order, on separate pages.

```
big_ols <- lm(logwaged ~ aged + factor(st) + factor(st)*aged, train_df)

wage_preds_26 <- c()
wage_preds_35 <- c()
dummystuff_states <- c(rep(0, 50))


# for all 51 states except for maine (which was dropped due to colinearity)
for (stval in c(1:50)) {
  dummystuff_states[stval] <- 1
  p_26 <- sum(big_ols$coefficients*c(1, 26, dummystuff_states, dummystuff_states*26))
  p_35 <- sum(big_ols$coefficients*c(1, 35, dummystuff_states, dummystuff_states*35))
  dummystuff_states[stval] <- 0

  wage_preds_26 <- c(wage_preds_26, p_26)
  wage_preds_35 <- c(wage_preds_35, p_35)
}

# now, compute the estimate for maine
p_maine_26 <- sum(big_ols$coefficients*c(1, 26, dummystuff_states, dummystuff_states*26))
p_maine_35 <- sum(big_ols$coefficients*c(1, 35, dummystuff_states, dummystuff_states*35))

wage_preds_26 <- c(p_maine_26, wage_preds_26)
wage_preds_35 <- c(p_maine_35, wage_preds_35)
statenames_ordered <- c("Maine", "New Hampshire", "Vermont", "Massachusetts", "Rhode Island", "Connecti

names(wage_preds_26) <- statenames_ordered
names(wage_preds_35) <- statenames_ordered
```

For 26 year olds:

```
print(wage_preds_26[order(names(wage_preds_26))])
```

```
##              Alabama               Alaska              Arizona
##             3.255210             3.379091             3.215963
##             Arkansas           California             Colorado
##             3.045248             3.295639             3.296744
##          Connecticut             Delaware District of Columbia
##             3.270291             3.297570             3.203715
##              Florida              Georgia               Hawaii
##             3.233833             3.312156             3.076689
##                Idaho             Illinois              Indiana
##             3.153528             3.380433             3.353176
##                 Iowa               Kansas             Kentucky
##             3.250982             3.102928             3.165418
##            Louisiana                Maine             Maryland
##             3.470353             2.954342             3.327719
##        Massachusetts             Michigan            Minnesota
##             3.233423             3.401370             3.292395
##          Mississippi             Missouri              Montana
##             3.068487             3.148381             3.151380
##             Nebraska               Nevada        New Hampshire
##             3.103156             3.333509             3.271760
##           New Jersey           New Mexico             New York
##             3.469317             3.079843             3.371386
##       North Carolina         North Dakota                 Ohio
##             3.153423             3.319536             3.314245
##             Oklahoma               Oregon         Pennsylvania
##             3.164010             3.217800             3.234738
##         Rhode Island       South Carolina         South Dakota
##             3.267129             3.206108             2.932713
##            Tennessee                Texas                 Utah
##             3.160647             3.301835             3.182767
##              Vermont             Virginia           Washington
##             2.986728             3.208262             3.369900
##        West Virginia            Wisconsin              Wyoming
##             3.196579             3.241246             3.475992
```

For 35 year olds:

```
print(wage_preds_35[order(names(wage_preds_35))])
```

```
##              Alabama               Alaska              Arizona
##             3.450434             3.809361             3.464678
##             Arkansas           California             Colorado
##             3.375278             3.546530             3.515803
##          Connecticut             Delaware District of Columbia
##             3.635999             3.562872             3.491736
##              Florida              Georgia               Hawaii
##             3.465410             3.490871             3.382588
##                Idaho             Illinois              Indiana
##             3.416003             3.629209             3.540743
##                 Iowa               Kansas             Kentucky
##             3.436029             3.447877             3.492016
##            Louisiana                Maine             Maryland
##             3.637904             3.309506             3.613921
##        Massachusetts             Michigan            Minnesota
##             3.562858             3.624990             3.533993
##          Mississippi             Missouri              Montana
##             3.378596             3.494107             3.273587
##             Nebraska               Nevada        New Hampshire
##             3.346868             3.613800             3.465162
##           New Jersey           New Mexico             New York
##             3.692557             3.340800             3.592482
##       North Carolina         North Dakota                 Ohio
##             3.445414             3.383634             3.562874
##             Oklahoma               Oregon         Pennsylvania
##             3.435719             3.466948             3.565245
##         Rhode Island       South Carolina         South Dakota
##             3.517245             3.434039             3.374267
##            Tennessee                Texas                 Utah
##             3.444289             3.556598             3.429743
##              Vermont             Virginia           Washington
##             3.313484             3.566447             3.552018
##        West Virginia            Wisconsin              Wyoming
##             3.522362             3.448916             3.398177
```

2 - k-fold CV LASSO model

Some CV setup:

```
#LASSO, CV
#specify search space for lambdas (for LASSO)
lambdas <- 10^seq(4, -4, by = -.1)

#10-fold cross validation to find optimal lambda (alpha=1 means lasso only)
lasso_cv_fit <- cv.glmnet(formula = logwaged ~ aged + factor(st) + aged:factor(st), alpha = 1, lambda =

#obtain lasso coefficients for the model with optimal lambda (in terms of CV error)
coefs <- coef(lasso_cv_fit$glmnet.fit, s=lasso_cv_fit$lambda.min)
```

We use the obtained coefficients similarly to how we did in (1). First, we calculate the predicted values, and then present the predicted logwaged for 26 and 35 year olds below on separate pages.

```
wage_preds_26_lasso <- c()
wage_preds_35_lasso <- c()
dummystuff_states <- c(rep(0, 51))


# for all 51 states
for (stval in c(1:51)) {
  dummystuff_states[stval] <- 1
  p_26 <- sum(coefs*c(1, 26, dummystuff_states, dummystuff_states*26))
  p_35 <- sum(coefs*c(1, 35, dummystuff_states, dummystuff_states*35))
  dummystuff_states[stval] <- 0

  wage_preds_26_lasso <- c(wage_preds_26_lasso, p_26)
  wage_preds_35_lasso <- c(wage_preds_35_lasso, p_35)
}

statenames_ordered <- c("Maine", "New Hampshire", "Vermont", "Massachusetts", "Rhode Island", "Connecti

names(wage_preds_26_lasso) <- statenames_ordered
names(wage_preds_35_lasso) <- statenames_ordered
```

LASSO preds for 26 year olds:

```
print(wage_preds_26_lasso[order(names(wage_preds_26_lasso))])
```

```
##              Alabama               Alaska              Arizona
##             3.224690             3.457508             3.218805
##             Arkansas           California             Colorado
##             3.114227             3.291922             3.269549
##          Connecticut             Delaware District of Columbia
##             3.341595             3.293763             3.262172
##              Florida              Georgia               Hawaii
##             3.225929             3.255779             3.130973
##                Idaho             Illinois              Indiana
##             3.168674             3.371715             3.299550
##                 Iowa               Kansas             Kentucky
##             3.217433             3.178295             3.227012
##            Louisiana                Maine             Maryland
##             3.400244             3.045065             3.335025
##        Massachusetts             Michigan            Minnesota
##             3.291355             3.376327             3.280027
##          Mississippi             Missouri              Montana
##             3.122133             3.222545             3.099126
##             Nebraska               Nevada        New Hampshire
##             3.104288             3.326819             3.239996
##           New Jersey           New Mexico             New York
##             3.441155             3.094438             3.342190
##       North Carolina         North Dakota                 Ohio
##             3.190730             3.201781             3.308531
##             Oklahoma               Oregon         Pennsylvania
##             3.184164             3.220552             3.290753
##         Rhode Island       South Carolina         South Dakota
##             3.269549             3.199231             3.075853
##            Tennessee                Texas                 Utah
##             3.190501             3.295230             3.184547
##              Vermont             Virginia           Washington
##             3.062872             3.291627             3.306926
##        West Virginia            Wisconsin              Wyoming
##             3.261042             3.222173             3.245425
```

LASSO preds 35 year olds:

```
print(wage_preds_26_lasso[order(names(wage_preds_26_lasso))])
```

```
##              Alabama               Alaska              Arizona
##             3.224690             3.457508             3.218805
##             Arkansas           California             Colorado
##             3.114227             3.291922             3.269549
##          Connecticut             Delaware District of Columbia
##             3.341595             3.293763             3.262172
##              Florida              Georgia               Hawaii
##             3.225929             3.255779             3.130973
##                Idaho             Illinois              Indiana
##             3.168674             3.371715             3.299550
##                 Iowa               Kansas             Kentucky
##             3.217433             3.178295             3.227012
##            Louisiana                Maine             Maryland
##             3.400244             3.045065             3.335025
##        Massachusetts             Michigan            Minnesota
##             3.291355             3.376327             3.280027
##          Mississippi             Missouri              Montana
##             3.122133             3.222545             3.099126
##             Nebraska               Nevada        New Hampshire
##             3.104288             3.326819             3.239996
##           New Jersey           New Mexico             New York
##             3.441155             3.094438             3.342190
##       North Carolina         North Dakota                 Ohio
##             3.190730             3.201781             3.308531
##             Oklahoma               Oregon         Pennsylvania
##             3.184164             3.220552             3.290753
##         Rhode Island       South Carolina         South Dakota
##             3.269549             3.199231             3.075853
##            Tennessee                Texas                 Utah
##             3.190501             3.295230             3.184547
##              Vermont             Virginia           Washington
##             3.062872             3.291627             3.306926
##        West Virginia            Wisconsin              Wyoming
##             3.261042             3.222173             3.245425
```

3 - Best state for 26/35 year old men?

For men who are 26, the evidence is split between OLS/LASSO on whether Alaska or Wyoming is better (with LASSO predicting Alaska, and OLS showing Wyoming). For 35 year olds, both models agree that Alaska is the best state.

4 - Model evaluation on the holdout sample

First, I just calculate OLS RMSES:

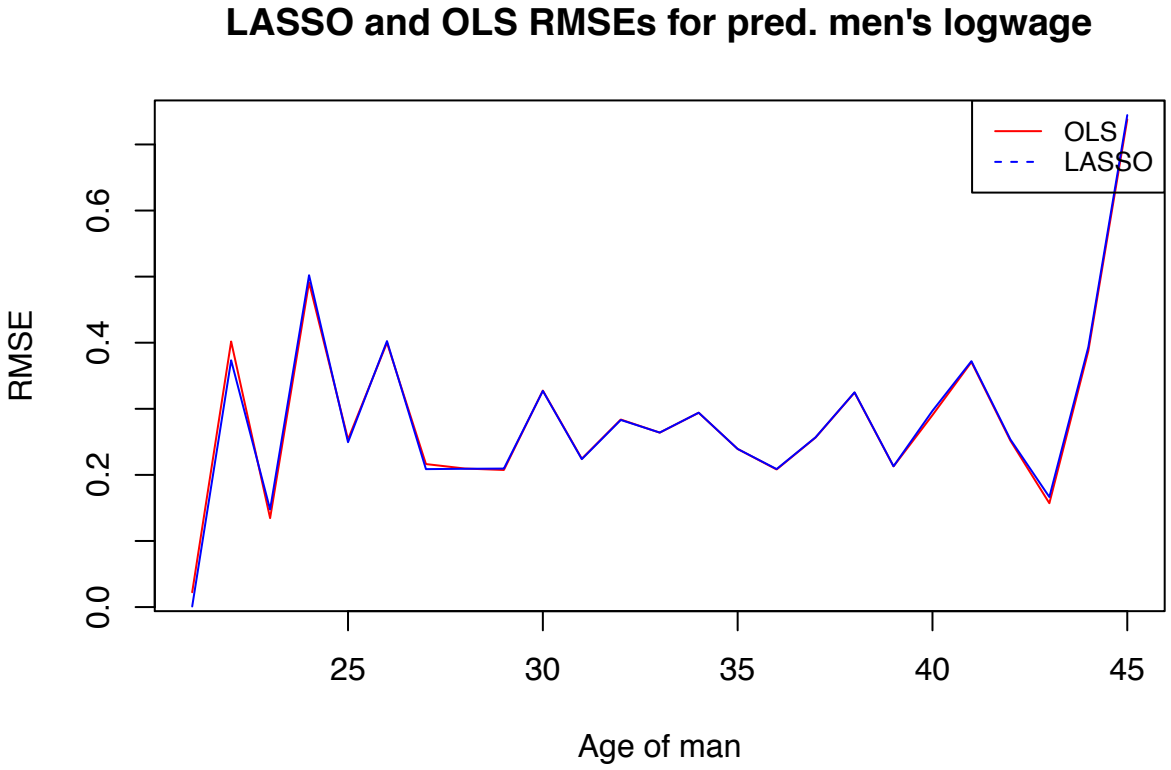
```
ols_rmses <- c()
lasso_rmses <- c()
lmin <- lasso_cv_fit$lambda.min
for (a in c(21:45)) {
  data_for_age <- subset(holdout_df, aged == a)
  y_actual <- data_for_age$logwaged

  y_pred_ols <- predict(big_ols, newdata=data_for_age)
  ols_rmses <- c(ols_rmses, mean((y_pred_ols - y_actual)^2))

}
```

LASSO's predict function is being rude to me so I'm going to do the prediction manually. We have the coefs from previous portions.

```
lasso_rmses <- c()
dummystuff_states <- c(rep(0, 51))
ordered_stcodes_util <- sort(unique(df$st))
for (a in c(21:45)) {
  data_for_age <- subset(holdout_df, aged == a)
  y_actual <- data_for_age$logwaged
  y_preds_lasso <- c()

  for (row in 1:nrow(data_for_age)) {
    state_of_interest <- data_for_age[row, "st"]
    corresponding_coeff_index <- match(state_of_interest, ordered_stcodes_util)

    dummystuff_states[corresponding_coeff_index] <- 1
    individual_pred <- sum(coefs*c(1, a, dummystuff_states, dummystuff_states*a))
    dummystuff_states[corresponding_coeff_index] <- 0

    y_preds_lasso <- c(y_preds_lasso, individual_pred)
  }

  lasso_rmses <- c(lasso_rmses, mean((y_preds_lasso - y_actual)^2))
}
```

We now graph the OLS and LASSO RMSEs.

```
x_axis = c(21:45)
plot(x_axis, ols_rmses, type="l", col="red", main="LASSO and OLS RMSEs for pred. men's logwage", xlab="A
lines(x_axis, lasso_rmses, col="blue")

legend("topright", legend=c("OLS", "LASSO"),
       col=c("red", "blue"), lty=1:2, cex=0.8)
```



We notice that OLS and LASSO are both relatively... the same on the holdout sample, with OLS performing marginally worse at some points and LASSO performing better at others. Here's an explanation of why that might be. Check out the following - a frequency count of how often each age appears in the training dataset:

```
table(train_df$aged)
```

```
##
##   21   22   23   24   25   26   27   28   29   30   31   32   33   34   35
##    2    5   15   36   98  223  414  718 1124 1530 2051 2673 2805 1983 1978
##   36   37   38   39   40   41   42   43   44   45
## 1685 1347  797  545  337  216  139   90   58   54
```

Recall that a big use of LASSO is generalizability; LASSO helps to select for only the important variables and 'zero-out' small idiosyncracies in the original dataset. So, it follows that when there are fewer data points available, OLS may overfit to the idiosyncracies in the data, whereas LASSO would not overfit as much. This is why we would expect LASSO to generalize better to the holdout data when there are fewer data points to go off of (at ages 21, 22, 23, etc) and same or worse when there is more data available (e.g. the age range in the 30s, where there is lots of training data).