Econ 142 Final Project

Sofia Guo[*]

May 16, 2019

Part I:

1. First, we specify M1, M2 and M3:

$$
\begin{aligned}
M1 &: morekids_i = \beta_0 + \beta_1 educm_i + \beta_2 agem_i + \beta_3 agefstm_i + u_i \\
M2 &: morekids_i = \beta_0 + \delta_1[educm = 0] + \cdots + \delta_{21}[educm = 20] + \beta_1 agem_i + \beta_2 agefstm_i + u_i \\
M3 &: morekids_i = \beta_0 + \delta_1[educm = 0] + \cdots + \delta_{21}[educm = 20] + \beta_1[agem = 21] + \cdots + \beta_{15}[agem = 35] + \\
&\qquad \gamma_1[agefstm = 15] + \cdots + \gamma_{19}[agefstm = 33] + u_i
\end{aligned}
$$

Table 1: Linear Probability Model Regression Results

	Dependent variable:		
	morekids	morekids	morekids
	(1)	(2)	(3)
RMSE	0.4494	0.448	0.4479
AIC	156401.4133	155653.7828	155621.2145
Observations	126,302	126,302	126,302
R^2	0.077	0.083	0.083
Adjusted R^2	0.077	0.083	0.083
Note:			*p<0.1; **p<0.05; ***p<0.01 *Table omits coefficients

Just from looking at the RMSE, we see that M2 and M3 perform better as they have a 0.01 lower statistic than M1. This is reflected in the AIC statistic as well, and when we compare M2 and M3, M3 wins in terms of having the lowest RMSE and AIC. Knowing this, we can look at the differences in estimated and actual probabilities and see what the models give us:

(i-iii). Calculating differences in probablities

Taking the absolute value of differences between groups, it seems like M1 produces higher differences in predicted probabilites between different subgroups than M2 or M3 except when looking at fstm differences; however, the spread between the three model results is not that large. The starkest difference (in group probability differences) we see is that adding education as a dummy instead of continuous variable reduces the estimated differences between groups (0.0471 from M1 vs ~0.028 for M2 and M3). We see that this decrease is due to an increase in the estimated probability for 35 year old mothers with 16 years of education (a 0.02 jump), indicating that M1 somehow underestimates this probability because it does not use dummies for education.

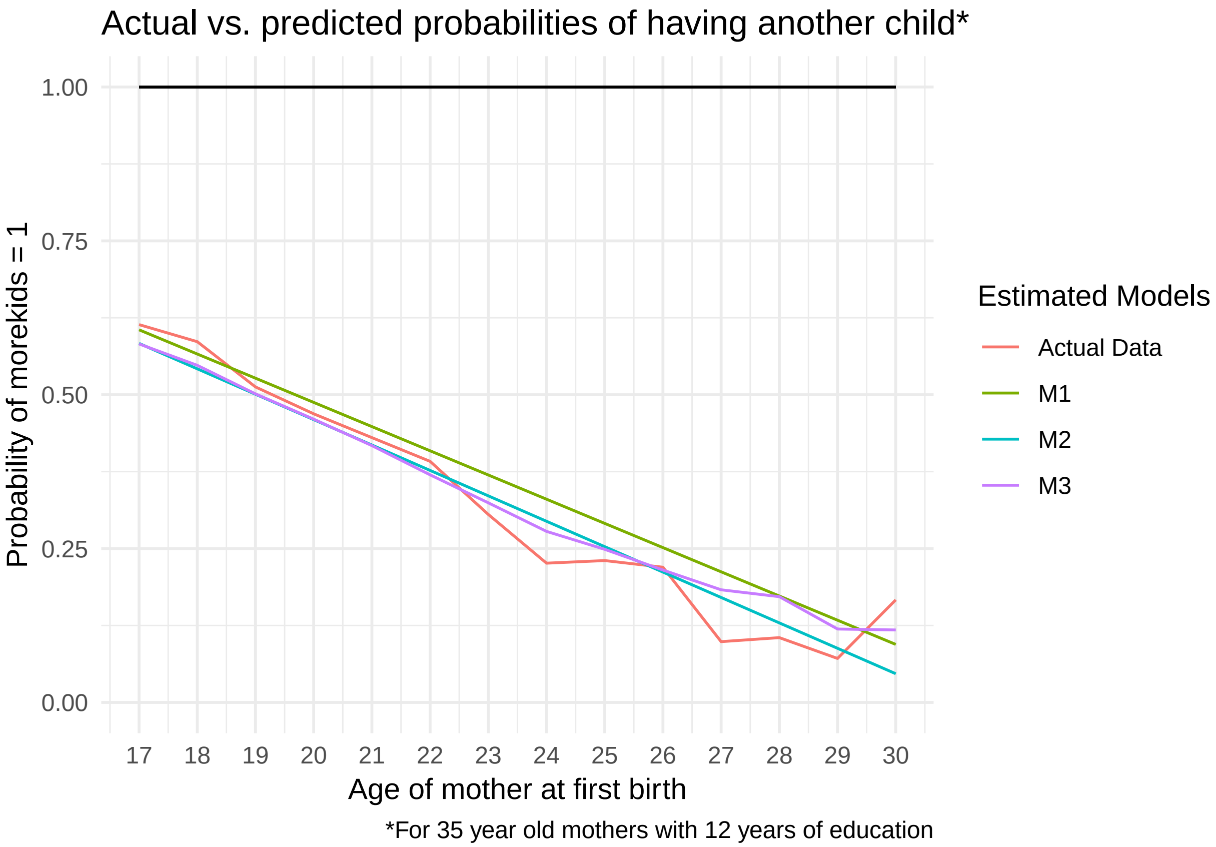
[*] I collaborated with Samuel Zhu, Alina, and Josh on much of this project and thank them for their enormous help.

Table 2: Estimated differences* in predicted probabilities for each model

Model	Agem30	Agem35	Agemdiff	Educ12	Educ16	Educdiff	Fstm20	Fstm25	Fstmdiff
M1	0.305	0.456	0.1508	0.330	0.283	0.0471	0.348	0.151	0.1966
SE1	0.001	0.002		0.001	0.002		0.001	0.003	
M2	0.519	0.670	0.1509	0.301	0.328	0.0271	0.563	0.356	0.2064
SE2	0.029	0.029		0.002	0.004		0.029	0.029	
M3	0.206	0.354	0.1479	-0.030	0.000	0.0297	0.548	0.337	0.2115
SE3	0.043	0.043		0.033	0.033		0.029	0.029	

[a] *Absolute value of differences were taken to ensure consistency

Pred. vs. actual prob:


Actual vs. predicted probabilities of having another child*

*For 35 year old mothers with 12 years of education

Knowing that our previous results point to M2 or M3 as the stronger models, we can see that this graph supports that conclusion; M1 consistently overestimates the probability of $morekids = 1$, while M2 and M3 are much closer (on average) to the actual data line (which are the proportions of mothers in that subgroup by age that have $morekids = 1$). An interesting observation is that they diverge once the age of mother at first birth passes 26; M3 proceeds to overestimate probabilites compared to M2, on average. This is possibly due to the variation introduced by using dummies for all independent variables (which we can confirm by seeing M3 being slightly more responsive to fluctuations in the data than M2 for agefstm > 26).

Based on this evidence, we conclude that M2 has the best absolute predictive capabilities of all the models, while M3 is more sensitive to trend fluctuations but overestimates probabilites in the higher agefstm categories.

Extending the model

2. (a) We define an extended model that also includes dad's age, education, and 3 variables for mother's race/ethnicity: blackm, hispm, othracem, then exclude the dad variables and the race variables:

$$
\begin{aligned}
M3\ new: morekids_i = {} & \beta_0 + \delta_1[educm = 0] + \cdots + \delta_{21}[educm = 20] + \beta_1[agem = 21] + \cdots + \beta_{15}[agem = 35] + \\
& \gamma_1[agefstm = 15] + \cdots + \gamma_{19}[agefstm = 33] + \\
& \lambda_1 aged_i + \lambda_2 educd_i + \lambda_3 blackm_i + \lambda_4 hispm_i + \lambda_5 othracem_i + u_i
\end{aligned}
$$

Table 3: Extended Linear Probability Model Regression Results*

	Dependent variable:		
	morekids	morekids	morekids
	(1)	(2)	(3)
aged	0.0001		0.0003
	(0.0005)		(0.0005)
educd	-0.002^{***}		-0.003^{***}
	(0.001)		(0.001)
blackm	0.086^{***}	0.087^{***}	
	(0.005)	(0.005)	
hispm	0.102^{***}	0.104^{***}	
	(0.008)	(0.008)	
othracem	0.055^{***}	0.055^{***}	
	(0.007)	(0.007)	
Constant	0.405^{***}	0.398^{***}	0.444^{***}
	(0.038)	(0.036)	(0.038)
RMSE	0.44699	0.44701	0.44781
Observations	126,302	126,302	126,302
R^2	0.087	0.087	0.084
Adjusted R^2	0.087	0.086	0.083
Residual Std. Error	0.447 (df = 126245)	0.447 (df = 126247)	0.448 (df = 126248)
F Statistic	214.597^{***} (df = 56; 126245)	222.328^{***} (df = 54; 126247)	217.225^{***} (df = 53; 126248)
Note:			*p<0.1; **p<0.05; ***p<0.01 *Table omits dummy coefficients

We can see just by the coefficients that it is likely the dad variables are not very important, but to make sure we perform partial F-tests on both reduced models:

Table 4: Partial F-test Results for Extended M3 (adding dad variables)

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
2	126245	25235.3	2	2.172235	5.433535	0.0043687

Table 5: Partial F-test Results for Extended M3 (adding race variables)

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
2	126245	25235.3	3	92.46535	154.1926	0

Adding the dad variables does not yield a very significant F-stat (5.43 with a high p-value of 0.0043) while adding the race variables yield a very significant F-stat (154.19 with a p-value of 0). This tells us that indeed, the two dad variables have low predictive power and can be excluded while the race ethnicity variables are more important.

(b) Evaluate the potential use of *samesex* as an exogenous determinant of family size:

(i). We add *samesex* to the extended model 3 and reestimate:

Table 6: Linear Probability Model Regression Results* w/samesex

	Dependent variable:
	morekids
samesex	0.064^{***}
	(0.003)
aged	0.0001
	(0.0005)
educd	-0.002^{***}
	(0.001)
blackm	0.087^{***}
	(0.005)
hispm	0.103^{***}
	(0.008)
othracem	0.056^{***}
	(0.007)
Constant	0.367^{***}
	(0.038)
RMSE	0.44585
Observations	126,302
R^2	0.092
Adjusted R^2	0.091
Residual Std. Error	0.446 (df = 126244)
F Statistic	223.230^{***} (df = 57; 126244)
Note:	*p<0.1; **p<0.05; ***p<0.01 *Table omits dummy coefficients

The estimated average effect of having the first two children of the same sex on the probability that $morekids = 1$ is 6.4%, ceteris paribus.

(ii). To test the claim that families care more about having at least 1 son and thus are more likey to have another kid, we can conduct a t-test on the following null hypothesis:

$$H_0 : \hat{\beta}_{boys2} = \hat{\beta}_{girls2}$$

If the t-stat is significant (>1.96 for a 95% Confidence interval), then we reject the null and find that there is a significant difference in having 2 daughters vs 2 sons on the probability of having an additional child. We estimate the extended model 3 and drop *samesex* to avoid perfect multicollinearity:

Table 7: Linear Probability Model Regression Results* w/boys2 and girls2

	Dependent variable:
	morekids
boys2	0.052^{***}
	(0.003)
girls2	0.077^{***}
	(0.003)
aged	0.0001
	(0.0005)
educd	-0.002^{***}
	(0.001)
blackm	0.087^{***}
	(0.005)
hispm	0.103^{***}
	(0.008)
othracem	0.056^{***}
	(0.007)
Constant	0.368^{***}
	(0.038)
RMSE	0.44577
Observations	126,302
R^2	0.092
Adjusted R^2	0.092
Residual Std. Error	0.446 (df = 126243)
F Statistic	220.328^{***} (df = 58; 126243)
Note:	*p<0.1; **p<0.05; ***p<0.01 *Table omits dummy coefficients

We can now conduct the t-test by calculating the t-stat:

$$\begin{aligned} t &= \frac{\hat{\beta}_{boys2} - \hat{\beta}_{girls2}}{SE_{\hat{\beta}_{boys2}}} \\ &= \frac{0.052 - 0.077}{0.003} \\ &= -8.33 \end{aligned}$$

Since the reported standard errors are the same (0.003) for each beta, we will get the same absolute value t-stat. Since $8.33 > 1.96$, we can reject the null $H_0 : \hat{\beta}_{boys2} = \hat{\beta}_{girls2}$ and conclude that there is a statistically significant difference between the estimated effect of boys2 versus girls2 on families having another child, ceteris paribus. The model estimates that families with 2 first girls are 0.077-0.052 = 0.025 or 2.5% more likely to have another child compared to families with 2 first boys.

(iii). We define three models that have samesex as the dependent variable and information on parents' age, education, race and age at first birth; we interact all the race terms to simplify the regression:

$$\begin{aligned} M4: samesex_i = {} & \beta_0 + \beta_1 agem_i + \beta_2 aged_i + \beta_3 educm_i + \beta_4 educd_i + \beta_5 blackm_i * blackd_i + \beta_6 whitem_i * whited_i \\ & + \beta_7 othracem_i * othraced_i + \beta_8 hispm_i * hispd_i + \beta_9 agefstm_i + \beta_{10} agefstd_i + u_i \\ M5: samesex_i = {} & \beta_0 + \beta_1 agem_i + \beta_2 aged_i + \sum_{i=0}^{20} \gamma_i [educm = i] + \sum_{i=0}^{20} \lambda_i [educd = i] + \beta_3 blackm_i * blackd_i \\ & + \beta_4 whitem_i * whited_i + \beta_5 othracem_i * othraced_i + \beta_6 hispm_i * hispd_i + \beta_7 agefstm_i + \beta_8 agefstd_i + u_i \\ M6: samesex_i = {} & \beta_0 + \beta_1 agem_i + \beta_2 aged_i + \sum_{i=0}^{20} \gamma_i [educm = i] + \sum_{i=0}^{20} \lambda_i [educd = i] + \beta_3 blackm_i * blackd_i \\ & + \beta_4 whitem_i * whited_i + \beta_5 othracem_i * othraced_i + \beta_6 hispm_i * hispd_i \\ & + \sum_{i=0}^{14} \delta_i \sum_{j=15}^{32} [agefstm = j] + \sum_{i=0}^{27} \alpha_i \sum_{j=15}^{43} [agefstd = j] + u_i \end{aligned}$$

Next, we run the models and obtain the RMSE's:

Table 8: Linear Probability Model Regression Results* predicting samesex

	Dependent variable:		
	samesex	samesex	samesex
	(1)	(2)	(3)
agem	−0.003	−0.003	−0.003
	(0.003)	(0.003)	(0.003)
aged	0.003	0.003	0.003
	(0.003)	(0.003)	(0.003)
blackm	0.105	0.103	0.101
	(0.064)	(0.064)	(0.064)
blackd	0.011	0.014	0.011
	(0.040)	(0.040)	(0.040)
whitem	0.032	0.033	0.037
	(0.043)	(0.043)	(0.043)
whited	0.036	0.039	0.044
	(0.041)	(0.041)	(0.041)
othracem	0.031	0.031	0.030
	(0.032)	(0.032)	(0.032)
othraced	0.023	0.025	0.026
	(0.031)	(0.031)	(0.031)
hispm			
hispd			
blackm:blackd	−0.075	−0.074	−0.065
	(0.067)	(0.067)	(0.067)
whitem:whited	−0.021	−0.022	−0.027
	(0.040)	(0.040)	(0.040)
othracem:othraced	−0.018	−0.020	−0.017
	(0.041)	(0.041)	(0.042)
hispm:hispd	0.040	0.039	0.042
	(0.046)	(0.046)	(0.046)
Constant	0.502^{***}	0.527^{***}	0.653^{***}
	(0.078)	(0.060)	(0.088)
RMSE	0.49997	0.49987	0.49916
Observations	126,302	126,302	126,302
R^2	0.0001	0.001	0.003
Adjusted R^2	−0.00001	0.0001	−0.0003
Residual Std. Error	0.500 (df = 126284)	0.500 (df = 126247)	0.500 (df = 125843)
F Statistic	0.905 (df = 17; 126284)	1.203 (df = 54; 126247)	0.923 (df = 458; 125843)
Note:			*p<0.1; **p<0.05; ***p<0.01 *Table omits dummy coefficients

Immediately, we see that none of these models produce significant estimates for the regressors. All the models have high RMSE's and poor fits. While this result could support the idea that $samesex = 1$ is a random event (none of the demographic information on parents), we check to see if the predicted probabilities are around 50%:

Table 9: Estimated predicted probabilities* for samesex = 1

Model	Predicted.Probability	Standard.error
M4	0.453	0.045
M5	0.512	0.058
M6	0.550	0.064

[a] *Using mean values of continous variables

Even given the poor fit of the models, our predicted probabilities yield around 50%, plus or minus 5%. This makes an adequate case for the assertion that the probability of having a boy/girl is roughly random, at least in this sample.

3. Compare OLS vs. IV

(a). We define 2 linear probability models:

$$
\begin{aligned}
W1 &: workedm_i = \beta_0 + \beta_1 morekids_i + u_i \\
W2 &: workedm_i = \beta_0 + \beta_1 morekids_i + \beta_2 educm_i + \sum_{i=0}^{11} \gamma_i [educm = i] + \beta_3 educd_i + \sum_{i=0}^{11} \delta_i [educd = i] \\
&\quad + \beta_4 agem_i + \beta_5 agem_i^2 + \beta_6 agefstm_i + \beta_7 agefstm_i^2 + \beta_8 aged_i + \beta_9 aged_i^2 + \\
&\quad + \beta_{10} agefstd_i + \beta_{11} agefstd_i^2 + \beta_{12} blackm_i + \beta_{13} hispm_i + \beta_{14} othracem_i + u_i
\end{aligned}
$$

Table 10: Estimated effect of morekids=1 on workedm

Model	Morekids.Coef
W1	4.03e-15
W2	6.75e-15

Comparing the estimated coefficients for morekids (the effect of having more kids on workedm), we see that the model without controls underestimates the effect, or that $\beta_{morekids,W1} < \beta_{morekids,W2}$. Thinking about OVB and short vs. long regressions, this phenomenon is explained by a negatively biased relationship between morekids and the residual in W1; the unexplained variables in u_{W1} are negatively correlated with morekids, which would explain why the estimated beta for morekids increases as more regressors are added.

(b). We define and test a simple causal model:

$$
\begin{aligned}
Full\ Model &: workedm_i = \beta_0 + \beta_1 morekids_i + u_i \\
First\ Stage &: morekids_i = \pi_0 + \pi_1 samesex_i + \eta_i \\
Reduced\ Form &: workedm_i = \delta_0 + \delta_1 samesex_i + v_i
\end{aligned}
$$

Table 11: FS, RF, and IV results

	Dependent variable:		
	morekids	workedm	workedm
	OLS	*OLS*	*instrumental variable*
	(1)	(2)	(3)
samesex	0.062^{***}	0.000	
	(0.003)	(0.000)	
morekids			0.000
			(0.000)
Constant	0.292^{***}	1.000^{***}	1.000^{***}
	(0.002)	(0.000)	(0.000)
Observations	126,302	126,302	126,302
R^2	0.004	0.500	−Inf.000
Adjusted R^2	0.004	0.500	−Inf.000
Residual Std. Error (df = 126300)	0.467	0.000	0.000
F Statistic (df = 1; 126300)	555.460^{***}	$126{,}300.100^{***}$	
Note:			*p<0.1; **p<0.05; ***p<0.01

Table 12: Estimated coefficient on morekids using IV

Regression	Morekids.Coef
Ratio	8.81e-14
IV	1.82e-13

The calculated ratio $\frac{\hat{\delta}_1}{\hat{\pi}_1}$ is slightly different from $\hat{\beta}_{IV}$ as seen in the table above; given that all missing values were removed at the beginning, it is unclear why there is this discrepancy. However, a possible explanation could be the case that for very small estimates, there is a larger margin of error. Overall, if one looks at the beta estimates, they are very strongly 0 (to the 13th or 14th) decimal place, so one could say that they "match" on a general scale.

(c). We estimate the model with the added controls from (a):

Table 13: FS, RF, and IV results (with 15 controls)

	Dependent variable:		
	morekids	workedm	workedm
	OLS	*OLS*	*instrumental variable*
	(1)	(2)	(3)
samesex	0.064^{***}		
	(0.003)		
morekids		0.000	0.000
		(0.000)	(0.000)
Constant	0.149^{***}	1.000^{***}	1.000^{***}
	(0.016)	(0.000)	(0.000)
Observations	126,302	126,302	126,302
R^2	0.090	0.500	−Inf.000
Adjusted R^2	0.090	0.500	−Inf.000
Residual Std. Error (df = 126285)	0.446	0.000	0.000
F Statistic (df = 16; 126285)	780.428^{***}	$7,892.814^{***}$	
Note:			*p<0.1; **p<0.05; ***p<0.01

Table 14: Estimated coefficient on morekids using IV w/15 controls

Regression	Morekids.Coef
Ratio	1.06e-13
IV	1.69e-13

This time around, the calculated ratio is much closer to the IV estimate. However, one thing to note is that the first stage regression for both (b) and (c) show that *samesex* is not a good instrument due to its 0 effect on morekids.

(d). We want to prove that given a truly random instrument, if a vector of controls x_{Oi} is included in an IV regression, it will not be substantially different from the estimated $\hat{\beta}_{IV}$ given by a non-controlled IV regression. Since we assume our instrument z_i is orthogonal to the control vector x_{Oi}, we begin the proof by stating the following expectation and the FOC of the reduced form equation:

$$\begin{aligned}
&RF_{controls} : y_i = \delta_0 + \delta_{1cont} z_i + x_{Oi} + v_i \\
&E[z_i x_{Oi}] = 0 \\
&FOC : E[z_i v_i] = 0 \\
&\Rightarrow E[z_i(y_i - \delta_0 - \delta_{1cont} z_i - x_{Oi})] = 0 \\
&\Rightarrow E[z_i y_i - \delta_0 z_i - z_i x_{Oi}] = \delta_{1cont} E[z_i^2] \\
&\Rightarrow E[z_i y_i] - \delta_0 E[z_i] - E[z_i x_{Oi}] = \delta_{1cont} E[z_i^2] \\
&\Rightarrow E[z_i y_i] - \delta_0 E[z_i] - 0 = \delta_{1cont} E[z_i^2] \\
&\Rightarrow \delta_{1cont} = \frac{E[z_i y_i] - \delta_0 E[z_i]}{E[z_i^2]}
\end{aligned}$$

$$\begin{aligned}
&RF_{no\ controls} : y_i = \delta_0 + \delta_1 z_i + v_i \\
&FOC : E[z_i v_i] = 0 \\
&\Rightarrow E[z_i(y_i - \delta_0 - \delta_1 z_i)] = 0 \\
&\Rightarrow E[z_i y_i - \delta_0 z_i] = \delta_1 E[z_i^2] \\
&\Rightarrow E[z_i y_i] - \delta_0 E[z_i] = \delta_1 E[z_i^2] \\
&\Rightarrow E[z_i y_i] - \delta_0 E[z_i] = \delta_1 E[z_i^2] \\
&\Rightarrow \delta_{1cont} = \delta_1 = \frac{E[z_i y_i] - \delta_0 E[z_i]}{E[z_i^2]} \blacksquare .
\end{aligned}$$

We conduct the same proof with the first stage regression:

$$\begin{aligned}
&FS_{controls} : x_i = \pi_0 + \pi_{1cont} z_i + x_{Oi} + \eta_i \\
&E[z_i x_{Oi}] = 0 \\
&FOC : E[z_i \eta_i] = 0 \\
&\Rightarrow E[z_i(x_i - \pi_0 - \pi_{1cont} z_i - x_{Oi})] = 0 \\
&\Rightarrow E[z_i x_i - \pi_0 z_i - z_i x_{Oi}] = \pi_{1cont} E[z_i^2] \\
&\Rightarrow E[z_i x_i] - \pi_0 E[z_i] - E[z_i x_{Oi}] = \pi_{1cont} E[z_i^2] \\
&\Rightarrow E[z_i x_i] - \pi_0 E[z_i] - 0 = \pi_{1cont} E[z_i^2] \\
&\Rightarrow \pi_{1cont} = \frac{E[z_i x_i] - \pi_0 E[z_i]}{E[z_i^2]}
\end{aligned}$$

$$\begin{aligned}
&FS_{no\ controls} : x_i = \pi_0 + \pi_1 z_i + \eta_i \\
&FOC : E[z_i \eta_i] = 0 \\
&\Rightarrow E[z_i(x_i - \pi_0 - \pi_1 z_i)] = 0 \\
&\Rightarrow E[z_i x_i - \pi_0 z_i] = \pi_1 E[z_i^2] \\
&\Rightarrow E[z_i x_i] - \pi_0 E[z_i] = \pi_1 E[z_i^2] \\
&\Rightarrow E[z_i x_i] - \pi_0 E[z_i] = \pi_1 E[z_i^2] \\
&\Rightarrow \pi_{1cont} = \pi_1 = \frac{E[z_i x_i] - \pi_0 E[z_i]}{E[z_i^2]} \blacksquare .
\end{aligned}$$

Now that we've shown the coefficients for the RF and FS are equal regardless of controls, we can therefore conclude that $\hat{\beta}_{IV\ no\ controls} = \hat{\beta}_{IV\ with\ controls} = \frac{\hat{\delta}_1}{\hat{\pi}_1} = \frac{\hat{\delta}_{1cont}}{\hat{\pi}_{1cont}} \blacksquare$.

(e). From (b), the IV coefficient is estimated at 1.82e-13, while from (c) the IV coefficient is estimated at 1.69e-13. There is a very minute difference of approximately 0.12e-13, which in the larger picture, given that these coefficients are all basically 0, satisfies the result we proved in (d) that $\hat{\beta}_{IV\ no\ controls} = \hat{\beta}_{IV\ with\ controls}$.

Formally, we can conduct a t-test on the statistical significance of their difference:

$$H_0 : \hat{\beta}_{IV\ no\ controls} = \hat{\beta}_{IV\ with\ controls}$$

We calculate the t-stat with the $SE_{\beta_{IV\ no\ controls}}$:

$$\begin{aligned} t &= \frac{\hat{\beta}_{IV\ no\ controls} - \hat{\beta}_{IV\ with\ controls}}{SE_{\hat{\beta}_{IV\ no\ controls}}} \\ &= \frac{1.82 * 10^{-13} - 1.69 * 10^{-13}}{1.98 * 10^{-12}} \\ &= 0.006565657 \end{aligned}$$

We calculate the t-stat with the $SE_{\beta_{IV\ with\ controls}}$:

$$\begin{aligned} t &= \frac{\hat{\beta}_{IV\ with\ controls} - \hat{\beta}_{IV\ no\ controls}}{SE_{\hat{\beta}_{IV\ with\ controls}}} \\ &= \frac{1.69 * 10^{-13} - 1.82 * 10^{-13}}{1.92 * 10^{-12}} \\ &= 0.006770833 \end{aligned}$$

Neither t-stat is significant (greater than 1.96), allowing us to fail to reject $H_0 : \hat{\beta}_{IV\ no\ controls} = \hat{\beta}_{IV\ with\ controls}$. As shown in our approximated analysis and proofs, the results of the hypothesis test show that we cannot conclude the estimates with and without controls are statistically significantly different from each other.

(f). Both the OLS and IV yield coefficients that are essentially 0. While the IV regression is not particularly strong given that π_1 is 0 with or without controls (the instrument is not correlated with the independent variable), even the OLS regression yields a 0 effect of *morekids* on *workedm*. We can interpret this as our regressions showing little to no effect of having extra kids on a mother's decision to work.

4. Compare OLS vs. IV (Family)

(a). We define 2 OLS models where $y_i = [earningsm_i, earninsd_i, famearns_i]$, $x_i = morekids_i$ and x_{Oi} is a vector of 15 controls defined in (3). For the IV models, we use *samesex* as an instrument for *morekids* and the same 15 controls as the OLS:

$$OLS1 : y_i = \beta_0 + \beta_1 x_i + u_i$$
$$OLS2 : y_i = \beta_0 + \beta_1 x_i + x_{Oi} + u_i$$

We estimate them for each dependent variable:

Table 15: IV and OLS results* for morekids on 3 dependent vars

Dependent.Var	OLS	IV	OLS.cont	IV.cont
earningsm	-3018.855	-1552.796	-3972.369	-1240.394
earningsd	-2394.837	-2646.003	-1589.691	-2915.451
famearn	-5413.692	-4198.800	-5562.061	-4155.845

[a] *The control variables are the same 15 in (3).

We think of a possible explanation for why the IV estimate for earnings is less negative than the OLS counterpart for mothers but more negative than the OLS for fathers. Because OLS measures the average effect of morekids on earnings while IV measures the effect on the subset of families where $samesex = 1$ (because the instrument only turns on in the first stage when $samesex = 1$). Thus, the predicted values of *morekids* used in the reduced form to estimate the coefficient result from a smaller group of families than the OLS draws from.

For mothers in the subset of families with 2 children of the same sex, their estimated average earnings dropped by less than the whole group possibly because they had already made adjustments of their income levels prior to having the additional child. For example, the mother may already have taken a part time job given that she knew they were going to have another kid, so her earnings did not drop as much as the whole group which didn't necessarily make those adjustment (weren't expecting to have more kids).

For fathers in the subset of families with 2 children of the same sex, their estimated average earnings dropped by more than the whole group possibly due to a larger share of his earnings (versus the mother's) going to support the additional child i.e. he pays for more stuff. Combined with the reduction in the mother's earnings (if she reduced her work hours to take care of more children), the impact on the father's earnings is amplified (more negative than the OLS) for this subset.

(b). Prove that $\hat{\beta}_3 = \hat{\beta}_1 + \hat{\beta}_2$ given $y_{3i} = y_{1i} + y_{2i}$ and each y_i is defined by a vector of controls x_i:

$$
\begin{aligned}
&y_{1i} = \beta_1 x_i + u_{1i} \\
&y_{2i} = \beta_2 x_i + u_{2i} \\
&y_{3i} = \beta_3 x_i + u_{3i} \\
&\hat{\beta}_3 = E[x_i^2]^{-1}E[x_i y_{3i}] \ by \ definition \\
&\Rightarrow y_{3i} = (\beta_1 x_i + u_{1i}) + (\beta_2 x_i + u_{2i}) \\
&\Rightarrow y_{3i} = x_i(\beta_1 + \beta_2) + (u_{1i} + u_{2i}) \\
&\Rightarrow \underbrace{(u_{1i} + u_{2i})}_{u_i} = y_{3i} - x_i(\beta_1 + \beta_2) \\
&FOC : E[x_i u_i] = 0 \\
&\Rightarrow E[x_i(y_{3i} - x_i(\beta_1 + \beta_2))] = 0 \\
&\Rightarrow E[x_i y_{3i}] = E[x_i^2(\beta_1 + \beta_2)] \\
&\Rightarrow (\hat{\beta}_1 + \hat{\beta}_2)E[x_i^2] = E[x_i y_{3i}] \\
&\Rightarrow (\hat{\beta}_1 + \hat{\beta}_2) = E[x_i^2]^{-1}E[x_i y_{3i}] = \hat{\beta}_3 \blacksquare.
\end{aligned}
$$

(c). For the OLS with controls, we want to show that $\hat{\beta}_3$ is estimate for $famearn$ because it is equal to $\hat{\beta}_1 + \hat{\beta}_2$ where $\hat{\beta}_1$ is the estimate for mothers' earnings and $\hat{\beta}_2$ is the estimate for fathers' earnings. From table 15, we know that the $\hat{\beta}_1 = -3972.369$ and $\hat{\beta}_2 = -1589.691$. Thus, $\hat{\beta}_1 + \hat{\beta}_2 = -5562.06$ which is the estimate for morekids using $famearn$. For OLS without controls, the same holds as $\hat{\beta}_1 = -3018.855$ and $\hat{\beta}_2 = -2394.837$. Thus, $\hat{\beta}_1 + \hat{\beta}_2 = -5413.692$ which is the estimate for morekids using $famearn$. To calculate the share of the total effect which is driven by mothers, we take $\frac{\hat{\beta}_1}{\hat{\beta}_3}$ for both the control and non control group:

$$
\begin{aligned}
OLS_m &: \frac{\hat{\beta}_1}{\hat{\beta}_3} = \frac{3018.855}{5413.692} = 0.5576 \\
OLS_{mcontrol} &: \frac{\hat{\beta}_1}{\hat{\beta}_3} = \frac{3972.369}{5562.061} = 0.7141
\end{aligned}
$$

To calculate the share of the total effect which is driven by fathers, we take $\frac{\hat{\beta}_1}{\hat{\beta}_3}$ for both the control and non control group:

$$
\begin{aligned}
OLS_d &: \frac{\hat{\beta}_1}{\hat{\beta}_3} = \frac{2394.837}{5413.692} = 0.4423 \\
OLS_{dcontrol} &: \frac{\hat{\beta}_1}{\hat{\beta}_3} = \frac{1589.691}{5562.061} = 0.2858
\end{aligned}
$$

As expected, the share varies with and without controls, but we can state this as: the mothers drive approximately 55.76% to 71.41% of the decline in family earnings, depending on if there are controls added to the regression. Meanwhile, the fathers drive approximately 44.23% to 28.58% of the decline in family earnings, without and with controls, respectively.

(d). We conduct the same caluclations as in (c). First, we verify that the sum of the father and mother IV estimates with and without controls add to the estimate on $famearn$. Indeed, for the IV without controls, we see that $\hat{\beta}_1 + \hat{\beta}_2 = 1552.796 + 2646.003 = 4198.800 = \hat{\beta}_3$. We also see that for the IV with controls, $\hat{\beta}_1 + \hat{\beta}_2 = 1240.394 + 2915.451 = 4155.845 = \hat{\beta}_3$. Next, we calculate the respective shares of total lost income that mothers and fathers make up in the IV regressions:

$$IV_m : \frac{\hat{\beta}_1}{\hat{\beta}_3} = \frac{1552.796}{4198.800} = 0.3698$$
$$IV_{mcontrol} : \frac{\hat{\beta}_1}{\hat{\beta}_3} = \frac{1240.394}{4155.845} = 0.2984$$

To calculate the share of the total effect which is driven by fathers, we take $\frac{\hat{\beta}_1}{\hat{\beta}_3}$ for both the control and non control group:

$$IV_d : \frac{\hat{\beta}_1}{\hat{\beta}_3} = \frac{2646.003}{4198.800} = 0.6301$$
$$IV_{dcontrol} : \frac{\hat{\beta}_1}{\hat{\beta}_3} = \frac{2915.451}{4155.845} = 0.7015$$

The share varies with and without controls, but we can state this as: the mothers drive approximately 36.98% to 29.84% of the decline in family earnings, depending on if there are controls added to the regression. Meanwhile, the fathers drive approximately 63.01% to 70.15% of the decline in family earnings, without and with controls, respectively.

This is interestingly contradictory to the OLS ratios where the mothers make up a majority share under OLS, but for IV, they make up only ~30%. Again, one could attribute this difference to the way IV and OLS are estimated. In this case, IV seems to give fathers' earnings more weight (more negative), indicating that the subset of families with 2 children of the samesex experience greater decreases in the father's earnings proportional to the mother's than the whole sample.

(e). LATE analysis: (i-iv). We can calculate the fraction of AT, NT, and C by first dividing the sample into a treatment and control group, where everyone in the treatment group has $samesex = 1$ while everyone in the control group has $samesex = 1$. Then, we can take from lecture what we know about the fractions to calculate:

$$\begin{aligned} C :& E[morekids_{treatment} - morekids_{control}] \\ AT :& E[morekids_{control}] \\ NT :& 1 - E[morekids_{treatment}] \end{aligned}$$

Table 16: Estimated always takers (AT), compliers (C), and never takers (NT)

LessThan12	EqTo12	Between13And15	MoreThan16	WholeSample	Category
0.065	0.066	0.072	0.050	0.062	C
0.432	0.283	0.255	0.190	0.292	AT
0.503	0.651	0.673	0.759	0.646	NT

(f). In order to calculate the means for compliers, we use what we learned in class where the δ_1 in the following “goofy” IV regression recovers the fraction of compliers, x_i is a vector of dummies with which you want to find the compliers within, and *morekids* is estimated with the instrument *samesex*:

$$IV : x_i * morekids = \delta_0 + \delta_1 morekids_i + u_i$$

After running the goofy regression for each complier group, we can calculate the all-families means by taking the mean of the dummy columns and obtain the following results:

Table 17: Means of compliers vs. all families

Sample	educm.12	mean.educm.	mean.agefstm.	frac.agefstm.21.	frac.hisp.	frac.black.	frac.white.	frac.Utah.
Compliers	0.164	12.333	20.258	0.591	0.024	0.065	0.883	0.011
All Families	0.153	12.584	20.621	0.529	0.025	0.068	0.877	0.009

Looking at our results, we see that the compliers vs. overall family means are pretty close and kind of trade off in terms of which is higher. There are slightly more lower educated mothers in the compliers group, which explains their slightly lower mean education. The compliers-only also have a higher share of under-21 first time mothers, which could be paired with the lower education levels as a potential explanation (lower educated mothers have children earlier). Most of both the compliers only and all families samples are white (at a whopping 88%), allowing us to make the observation that the majority these lower educated young first time mothers were white.

Part II:

1. Estimate OLS

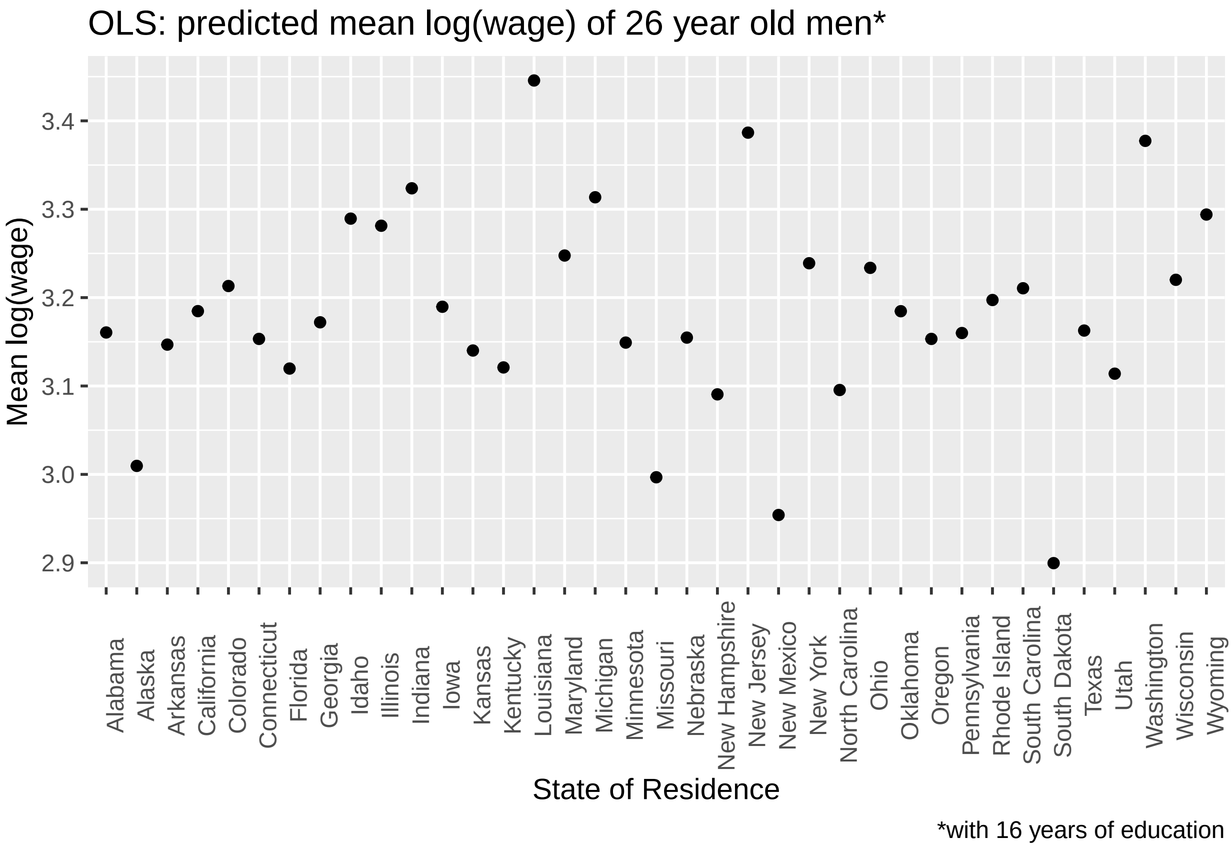
We estimate the OLS model below:

$$logwage_i = \beta_0 + \beta_1 aged_i + \sum_{i=1}^{51} \delta_i [st = i] + \sum_{i=1}^{51} \gamma_i aged * [st = i] + u_i$$

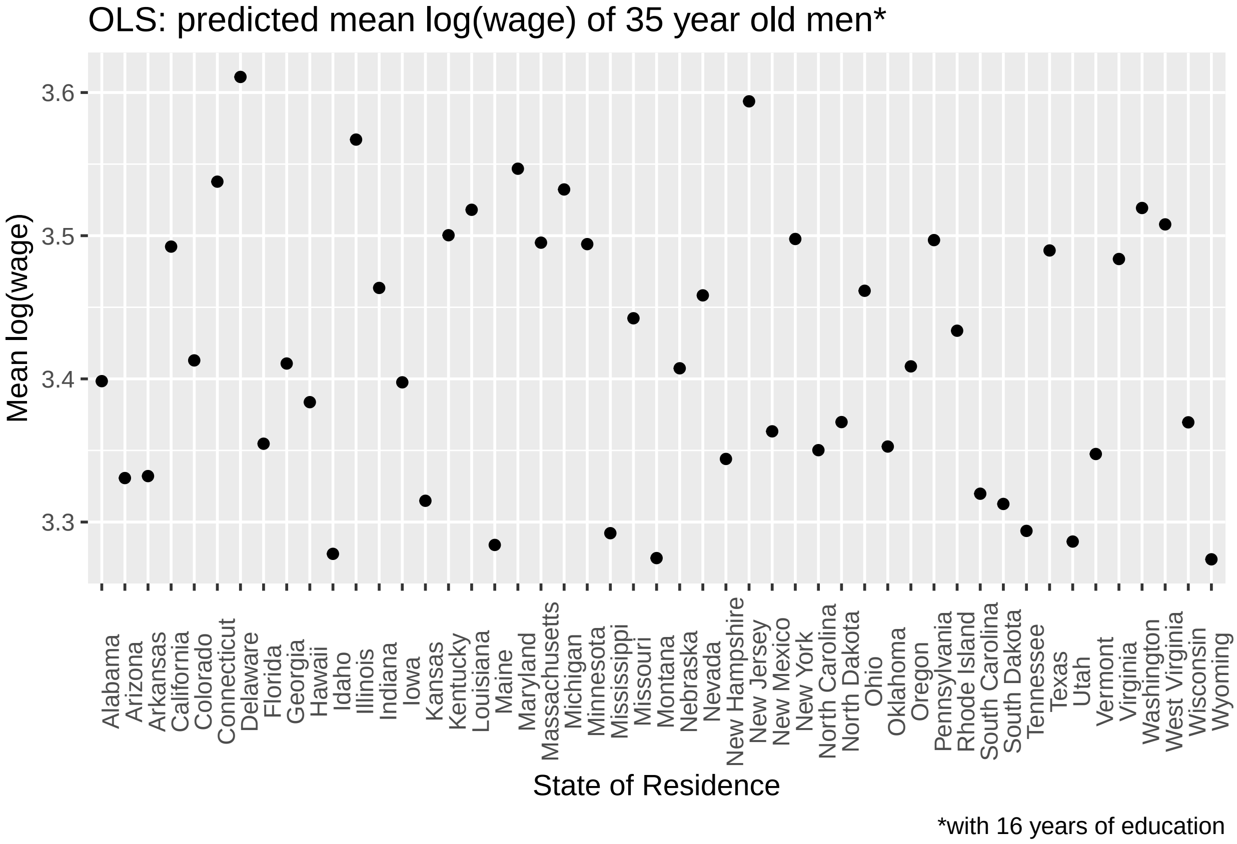
Table 18: Simple OLS Regression Results*

	Dependent variable:
	logwaged
aged	0.036^{**}
	(0.018)
Observations	10,710
R^2	0.063
Adjusted R^2	0.054
Residual Std. Error	0.513 (df = 10608)
F Statistic	7.059^{***} (df = 101; 10608)
Note:	*p<0.1; **p<0.05; ***p<0.01
	*Table omits dummy coefficients

We then use the model to predict the earnings of men aged 26 in each state, and men of age 35 in each state:


OLS: predicted mean log(wage) of 26 year old men*

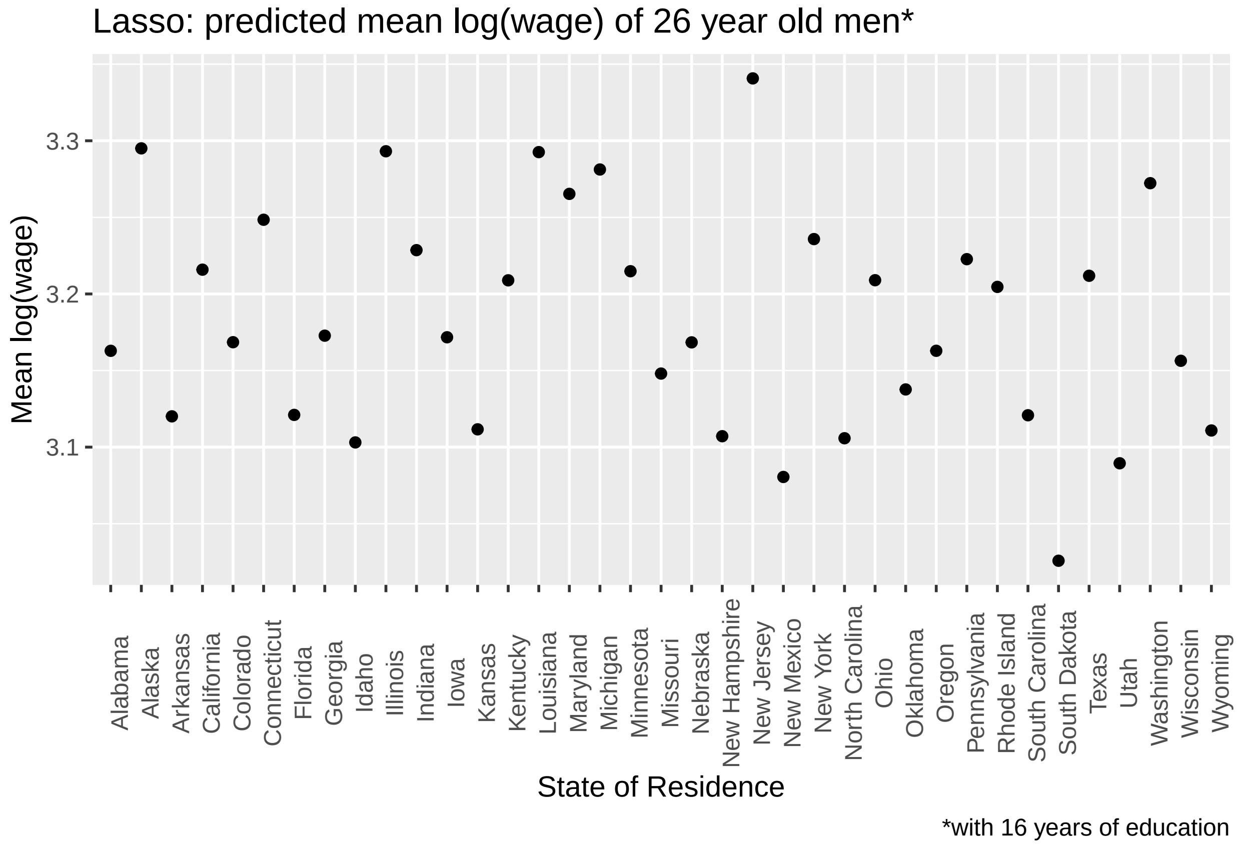
*with 16 years of education

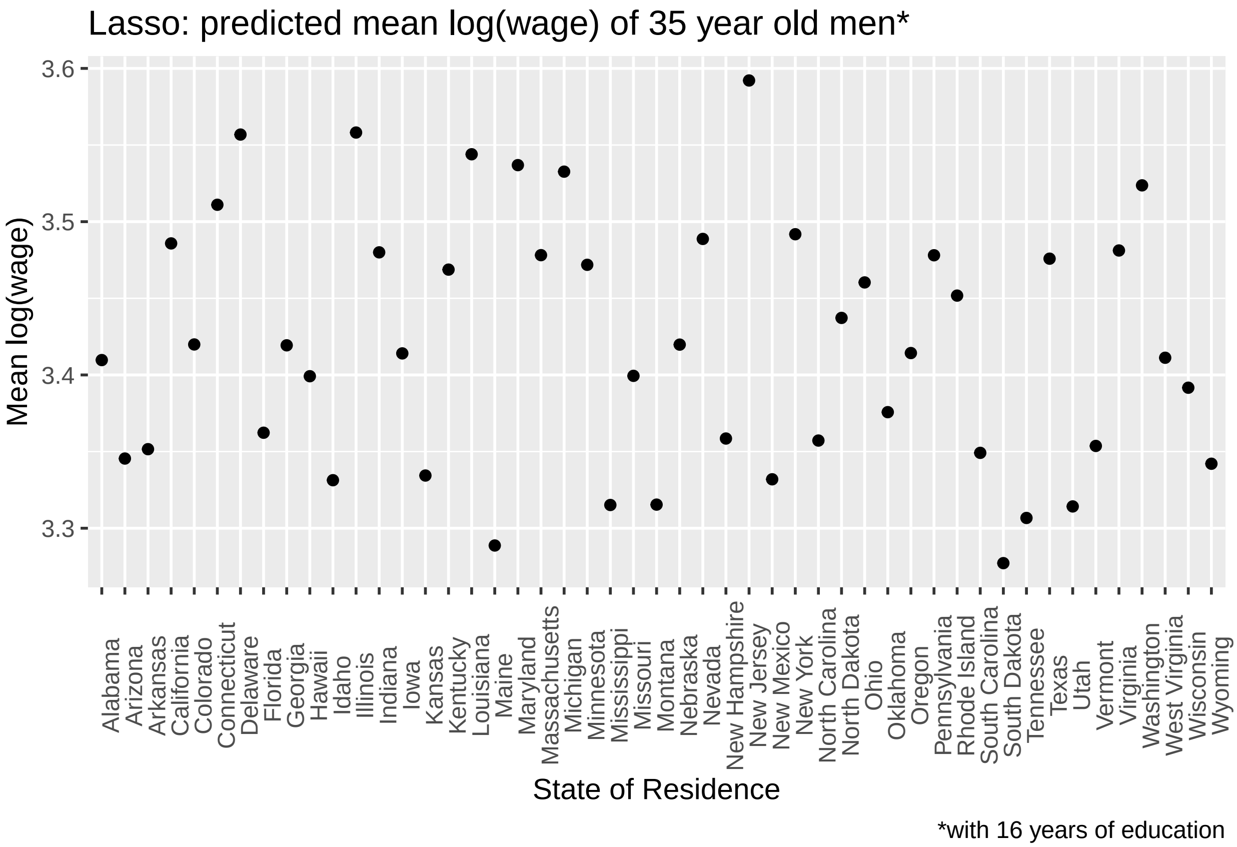

OLS: predicted mean log(wage) of 35 year old men*

*with 16 years of education

2. K-fold CV and Lasso:

Using the `cv.glmnet()` function, we find a lasso model that we use to predict the logwages of 26 and 35 year old men in our estimation sample. First, we create a dummy for each state, then add all the dummies that interact $aged * statedummy$ to our estimation data frame. Then we predict the log wages in a similar fashion as I.1 and graph them:





We see that the lasso and OLS results are pretty similar, with lasso giving slightly lower predictions for each state.

3. Best state

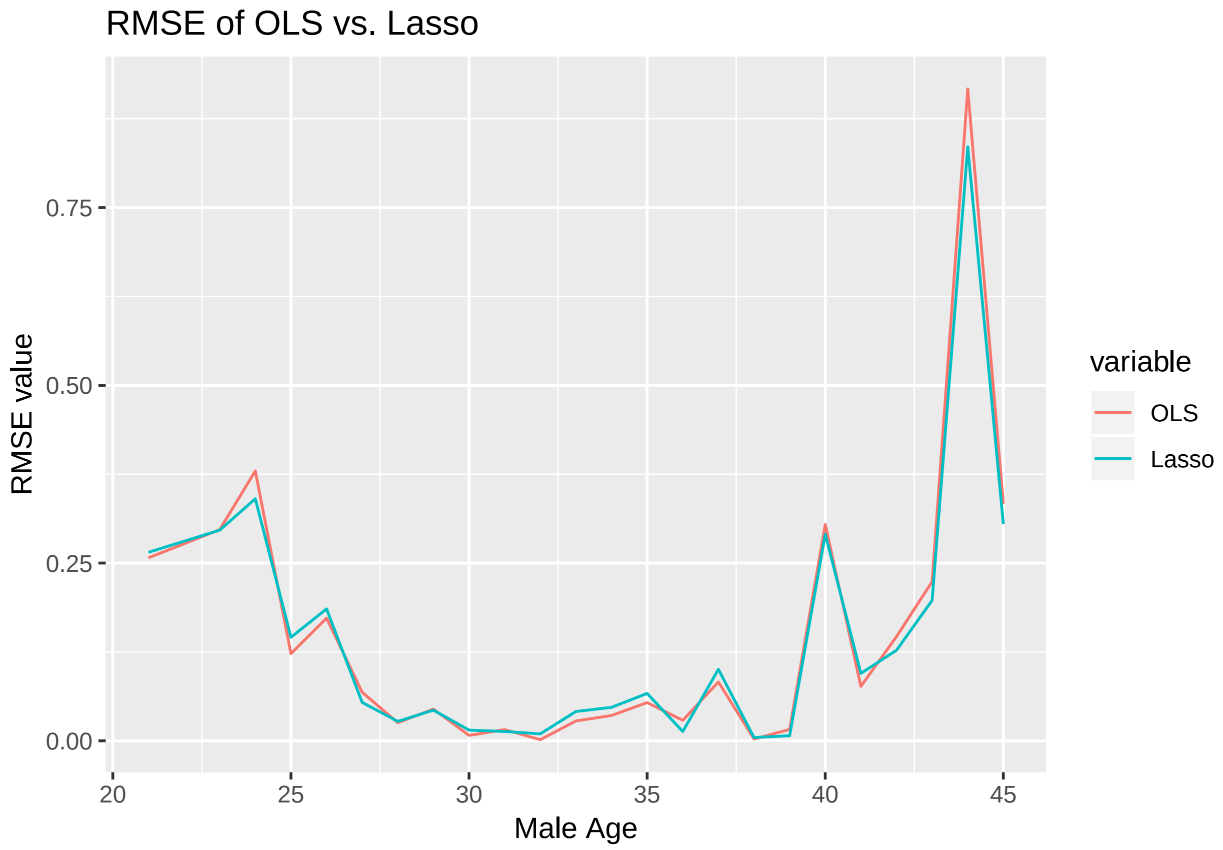
Table 19: States with the highest male mean log wage by age group

Model	Age	State Name	meanlogwage
OLS	26	Louisiana	3.446
LASSO	35	New Jersey	3.341
OLS	26	Delaware	3.611
LASSO	35	New Jersey	3.592

It seems like Alaska is a surprisingly high wage state for men of these age groups, with the exception of the OLS for 26 year old men. The best state for 26 year old men with a college degree is Alaska (3.48), while the best state for 35 year old men with a college degree is also Alaska. This could have to do with the concentration of oil/energy business in the state which would employ men of these education levels (not more advanced degrees, which would probably go to other industries like tech or finance) for drilling/mining.

4. Predict logwage on holdout

We use the same prediction method as before the get a vector of predictions and subsequent RMSE's for each men's age, then graph the RMSE results:



We can see that for ages 20 to 22, lasso has poor predictive power. However, starting from 22 until the mid-30's, the two models are nearly matched with OLS being slightly higher. They trade off again with lasso doing slightly worse until the late 30's, then it prevails as a little more accurate over OLS for the rest of the age range. Overall, both are pretty well matched, and one could say that for more consistent predictive power, OLS is superior on the lower tail end of the distribution.

Appendix:

Include all code used to generate figures, tables and calculations.

```
#load libraries
library(dplyr)
library(ggplot2)
library(reshape2)
library(psych)
library(stargazer)
library(qpcR) #for RMSE function
library(kableExtra)
library(AER)
library(glmnet)
library(fastDummies)
library(zoo)

#load master_wset
master <- read.csv("/Users/sofia/Box/Cal (sofiaguo@berkeley.edu)/2018-19/Spring 2019/Econ 142/Final Pro
  na.omit()

#check means
#chk_means <- describe(master)

#household counts by kid number and outcome
#chk_counts <- master %>%
 # group_by(KIDCOUNT, morekids) %>%
 # summarize(n())
```

```
$$
\begin{aligned}
\begin{split}
M1:morekids_{i} &= \beta_{0} + \beta_{1}educm_{i} + \beta_{2}agem_{i} + \beta_{3}agefstm_{i} + u_{i} \\
M2:morekids_{i} &= \beta_{0} + \delta_{1}[educm = 0] + \dots + \delta_{21}[educm = 20] + \beta_{1}agem_
M3:morekids_{i} &= \beta_{0} + \delta_{1}[educm = 0] + \dots + \delta_{21}[educm = 20] + \beta_{1}[agem
&\gamma_{1}[agefstm = 15] + \dots + \gamma_{19}[agefstm = 33] + u_{i}
\end{split}
\end{aligned}
$$
```

```
#estimate the model
M1 <- lm(morekids ~ educm + agem + agefstm, data = master)
M2 <- lm(morekids ~ factor(educm) + agem + agefstm, data = master)
M3 <- lm(morekids ~ factor(educm) + factor(agem) + factor(agefstm), data = master)

#calculate RMSE
RMSE1 <- RMSE(M1)
RMSE2 <- RMSE(M2)
RMSE3 <- RMSE(M3)

#calculate AIC
AIC1 <- AIC(M1)
AIC2 <- AIC(M2)
AIC3 <- AIC(M3)
```

```
stargazer(M1,
          M2,
          M3,
          type = "latex", title = "Linear Probability Model Regression
          Results",
          header = F, omit = c('educm','agem', 'agefstm', 'Constant'), notes = "*Table omits
         coefficients",
         omit.stat = c('f', 'ser'),
          multicolumn = F,
          font.size = "small",
          column.sep.width = '1pt',
          add.lines = list(c('RMSE',round(RMSE1, 4), round(RMSE2, 4), round(RMSE3, 4)),
                           c('AIC', round(AIC1, 4), round(AIC2, 4), round(AIC3, 4))))
```

```
## (i). Restricting mother's age to 30 and 35:
#new datasets with 30 and 35
newdata1 <- with(master, data.frame(agem = c(30, 35), educm=mean(educm), agefstm=mean(agefstm)))
newdata2 <- with(master, data.frame(agem = c(30, 35), educm=c(0,0), agefstm=mean(agefstm)))
newdata3 <- with(master, data.frame(agem = c(30, 35), educm=c(0,0), agefstm = c(30,30))) #set agefstm a


#predict the models with the age restrictions
pred_m1 <- predict(M1, newdata1, type="response", se.fit = T)
pred_m2 <- predict(M2, newdata2, type="response", se.fit=T)
pred_m3 <- predict(M3, newdata3, type="response", se.fit=T)

#report the difference in probabilities and the standard errors
agem_pred1_1 <- pred_m1$fit[1]
agem_pred1_2 <- pred_m1$fit[2]
agem_diff1 <- round(abs(agem_pred1_2 - agem_pred1_1), 4)
agem_se1 <- pred_m1$se.fit

agem_pred2_1 <- pred_m2$fit[1]
agem_pred2_2 <- pred_m2$fit[2]
agem_diff2 <- round(abs(agem_pred2_2 - agem_pred2_1), 4)
agem_se2 <- pred_m2$se.fit


agem_pred3_1 <- pred_m3$fit[1]
agem_pred3_2 <- pred_m3$fit[2]
agem_diff3 <-  round(abs(agem_pred3_2 - agem_pred3_1), 4)
agem_se3 <- pred_m3$se.fit
```

```
## (ii).Restricting mother's education to 12 and 16:

#new datasets with 12 and 16
newdata1 <- with(master, data.frame(agem = mean(agem), educm=c(12, 16), agefstm=mean(agefstm)))
newdata2 <- with(master, data.frame(agem = mean(agem), educm=c(12, 16), agefstm=mean(agefstm)))
newdata3 <- with(master, data.frame(agem = c(30, 30), educm=c(12,16), agefstm = c(30,30))) #set agem as


#predict the models with the age restrictions
pred_m1 <- predict(M1, newdata1, type="response", se.fit = T)
pred_m2 <- predict(M2, newdata2, type="response", se.fit=T)
```

```
pred_m3 <- predict(M3, newdata3, type="response", se.fit=T)

#report the difference in probabilities and the standard errors
educ_pred1_1 <- pred_m1$fit[1]
educ_pred1_2 <- pred_m1$fit[2]
educ_diff1 <- round(abs(educ_pred1_2 - educ_pred1_1), 4)
educ_se1 <- pred_m1$se.fit


educ_pred2_1 <- pred_m2$fit[1]
educ_pred2_2 <- pred_m2$fit[2]
educ_diff2 <- round(abs(educ_pred2_2 - educ_pred2_1), 4)
educ_se2 <- pred_m2$se.fit

educ_pred3_1 <-pred_m3$fit[1]
educ_pred3_2 <- pred_m3$fit[2]
educ_diff3 <- round(abs(educ_pred3_2 - educ_pred3_1), 4)
educ_se3 <- pred_m3$se.fit
```

```
## (iii).Restricting mother's age at first birth to 20 and 25:
#new datasets with 20 and 25
newdata1 <- with(master, data.frame(agem = mean(agem), educm=mean(educm), agefstm=c(20, 25)))
newdata2 <- with(master, data.frame(agem = mean(agem), educm=c(0,0), agefstm=c(20, 25)))
newdata3 <- with(master, data.frame(agem = c(30, 30), educm=c(0,0), agefstm=c(20, 25))) #set agem as ar


#predict the models with the age restrictions
pred_m1 <- predict(M1, newdata1, type="response", se.fit = T)
pred_m2 <- predict(M2, newdata2, type="response", se.fit=T)
pred_m3 <- predict(M3, newdata3, type="response", se.fit=T)

#report the difference in probabilities and the standard errors
fstm_pred1_1 <- pred_m1$fit[1]
fstm_pred1_2 <- pred_m1$fit[2]
fstm_diff1 <- round(abs(fstm_pred1_2 - fstm_pred1_1), 4)
fstm_se1 <- pred_m1$se.fit

fstm_pred2_1 <- pred_m2$fit[1]
fstm_pred2_2 <- pred_m2$fit[2]
fstm_diff2 <- round(abs(fstm_pred2_2 - fstm_pred2_1), 4)
fstm_se2 <- pred_m2$se.fit

fstm_pred3_1 <- pred_m3$fit[1]
fstm_pred3_2 <- pred_m3$fit[2]
fstm_diff3 <- round(abs(fstm_pred3_2 - fstm_pred3_1), 4)
fstm_se3 <- pred_m3$se.fit
```

```
#make data frame of results
mod_results_df <- data.frame("Model" = c("M1", "SE1", "M2", "SE2", "M3", "SE3"),
                             "Agem30" = c(agem_pred1_1, agem_se1[[1]], agem_pred2_1, agem_se2[[1]], ager
                             "Agem35" = c(agem_pred1_2, agem_se1[[2]], agem_pred2_2, agem_se2[[2]], ager
                             "Agemdiff" = c(agem_diff1, "", agem_diff2, "", agem_diff3, ""),
                             "Educ12" = c(educ_pred1_1, educ_se1[[1]], educ_pred2_1, educ_se2[[1]], edu
                             "Educ16" = c(educ_pred1_2, educ_se1[[2]], educ_pred2_2, educ_se2[[2]], edu
                             "Educdiff" = c(educ_diff1, "", educ_diff2, "", educ_diff3, ""),
                             "Fstm20" = c(fstm_pred1_1, fstm_se1[[1]], fstm_pred2_1, fstm_se2[[1]], fst
                             "Fstm25" = c(fstm_pred1_2, fstm_se1[[2]], fstm_pred2_2, fstm_se2[[2]], fst
                             "Fstmdiff" = c(fstm_diff1, "", fstm_diff2, "", fstm_diff3, ""))
```

```
#display table of results
kable(mod_results_df, "latex", booktabs = T, align = "c", caption = "Estimated differences* in predicte
  kable_styling(latex_options = 'hold_position')%>%
   add_footnote("*Absolute value of differences were taken to ensure consistency")
```

```
#calculate the predicted probabilities
pred_m1_4 <- rep(0, 14)
pred_m2_4 <- rep(0, 14)
pred_m3_4 <- rep(0, 14)

for(i in 17:30){
  newdata <- with(master, data.frame(agem=35, educm=12, agefstm=i))
  pred <- predict(M1, newdata, type = "response")
  pred_m1_4[i-16] <- as.numeric(pred[1])
  pred <- predict(M2, newdata, type = "response")
  pred_m2_4[i-16] <- as.numeric(pred[1])
  pred <- predict(M3, newdata, type = "response")
  pred_m3_4[i-16] <- as.numeric(pred[1])
}

pred_m1_4 <- unlist(pred_m1_4)
pred_m2_4 <- unlist(pred_m2_4)
pred_m3_4 <- unlist(pred_m3_4)

#calculate the actual probabilities
master_agem35_educ12_mk1 <- master %>%
  filter(agem == 35 & educm == 12 & morekids ==1)

master_agem35_educ12 <- master %>%
  filter(agem == 35 & educm == 12)

#calculate the number of actual mothers with morekids ==1 vs both 1 and 0 for agefstm = 17:30
numerator <- rep(0,14)
denominator <- rep(0, 14)

for(i in 17:30) {
  numerator[i-16] <- nrow(subset(master_agem35_educ12_mk1, agefstm==i))
  denominator[i-16] <- nrow(subset(master_agem35_educ12, agefstm==i))
}

actual_prob <- unlist(numerator)/unlist(denominator)
```

```
#make data frame of predicted vs. actual probabilites for each model
probs_df <- data.frame("Agefstm" = 17:30,
                       "Actualprob" = actual_prob,
                       "Predprob1" = pred_m1_4,
                       "Predprob2" = pred_m2_4,
                       "Predprob3" = pred_m3_4)
probs_df_melt <- melt(probs_df, id.vars = "Agefstm")
```

```
ggplot(probs_df_melt, aes(Agefstm, value, group = variable)) +
  geom_line(aes(color = variable)) +
  geom_line(aes(x=Agefstm, y =1)) +
  labs(x="Age of mother at first birth", y="Probability of morekids = 1", title = 'Actual vs. predicted
       caption = "*For 35 year old mothers with 12 years of education") +
  scale_y_continuous(limits = 0:1) +
  scale_x_continuous(limits=c(17, 30), breaks = 17:30) +
  scale_color_discrete(name="Estimated Models",
                         breaks=c("Actualprob", "Predprob1", "Predprob2", "Predprob3"),
                         labels=c("Actual Data", "M1", "M2", "M3")) +
  theme_minimal()
```

```
$$
\begin{aligned}
\begin{split}
M3 \ new :morekids_{i} &= \beta_{0} + \delta_{1}[educm = 0] + \dots + \delta_{21}[educm = 20] + \beta_{
&\gamma_{1}[agefstm = 15] + \dots + \gamma_{19}[agefstm = 33] + \\
&\lambda_{1}aged_{i} + \lambda_{2}educd_{i} + \lambda_{3}blackm_{i} + \lambda_{4}hispm_{i} + \lambda_{5
\end{split}
\end{aligned}
$$
```

```
#estimate extended model
M3_ext <- lm(morekids ~ factor(educm) + factor(agem) + factor(agefstm) + aged + educd + blackm + hispm +

M3_ext_dad <- lm(morekids ~ factor(educm) + factor(agem) + factor(agefstm)+ blackm + hispm + othracem, d

M3_ext_race <- lm(morekids ~ factor(educm) + factor(agem) + factor(agefstm) + aged + educd, data = maste

#get RMSE
RMSE_ext1 <- RMSE(M3_ext)
RMSE_ext2 <- RMSE(M3_ext_dad)
RMSE_ext3 <- RMSE(M3_ext_race)

#get F-stats
anova_M3_ext_dad <- as.data.frame(anova(M3_ext_dad,M3_ext))[2,]
anova_M3_ext_race <- as.data.frame(anova(M3_ext_race, M3_ext))[2,]
```

```
stargazer(M3_ext,
          M3_ext_dad,
          M3_ext_race,
          type = "latex", title = "Extended Linear Probability Model Regression
          Results*",
          header = F, omit = c('educm','agem', 'agefstm'),
notes = "*Table omits dummy
         coefficients",
          multicolumn = F,
          font.size = "small",
          column.sep.width = '1pt',
          add.lines = list(c('RMSE', round(RMSE_ext1, 5), round(RMSE_ext2, 5), round(RMSE_ext3, 5))))
```

```
#display table of results for second ANOVA
kable(anova_M3_ext_dad, "latex", booktabs = T, align = "c", caption = "Partial F-test Results for Exten
  kable_styling(latex_options = 'hold_position')
```

```
#display table of results for second ANOVA
kable(anova_M3_ext_race, "latex", booktabs = T, align = "c", caption = "Partial F-test Results for Exter
  kable_styling(latex_options = 'hold_position')
```

```
#estimate extended model with samesex
M3_ext_sex <- lm(morekids ~ factor(educm) + factor(agem) + factor(agefstm) + samesex + aged + educd + bl

#get RMSE
RMSE_ext4 <- RMSE(M3_ext_sex)
```

```
stargazer(M3_ext_sex,
          type = "latex", title = "Linear Probability Model Regression
          Results* w/samesex",
          header = F, omit = c('educm','agem', 'agefstm'),
notes = "*Table omits dummy
         coefficients",
          multicolumn = F,
          font.size = "small",
          column.sep.width = '1pt',
          add.lines = list(c('RMSE', round(RMSE_ext4, 5))))
```

```
$$
H_{0}: \hat\beta_{boys2} = \hat\beta_{girls2}
$$

#estimate extended model with boys2, dropping samesex
M3_ext_bg <- lm(morekids ~ factor(educm) + factor(agem) + factor(agefstm) + boys2 + girls2 + aged + educ

#get RMSE
RMSE_ext5 <- RMSE(M3_ext_bg)
```

```
stargazer(M3_ext_bg,
          type = "latex", title = "Linear Probability Model Regression
          Results* w/boys2 and girls2",
          header = F, omit = c('educm','agem', 'agefstm'),
notes = "*Table omits dummy
         coefficients",
          multicolumn = F,
          font.size = "small",
          column.sep.width = '1pt',
          add.lines = list(c('RMSE', round(RMSE_ext5, 5))))
```

```
$$
\begin{aligned}
\begin{split}
t &= \frac{\hat\beta_{boys2} - \hat\beta_{girls2}}{SE_{\hat\beta_{boys2}}} \\
&= \frac{0.052 - 0.077}{0.003} \\
&= -8.33
\end{split}
\end{aligned}
$$

$$
\begin{aligned}
\begin{split}
```

```
M4: samesex_{i} &= \beta_{0} + \beta_{1}agem_{i} + \beta_{2}aged_{i} + \beta_{3}educm_{i} + \beta_{4}edu
&+ \beta_{7}othracem_{i}*othraced_{i} + \beta_{8}hispm_{i}*hispd_{i} + \beta_{9}agefstm_{i} + \beta_{10
M5: samesex_{i} &= \beta_{0} + \beta_{1}agem_{i} + \beta_{2}aged_{i} + \sum_{i=0}^{20}\gamma_{i}[educm 
+ \beta_{5}othracem_{i}*othraced_{i} + \beta_{6}hispm_{i}*hispd_{i} + \beta_{7}agefstm_{i} +\beta_{8}ag
M6: samesex_{i} &= \beta_{0} + \beta_{1}agem_{i} + \beta_{2}aged_{i} + \sum_{i=0}^{20}\gamma_{i}[educm 
&+ \beta_{4}whitem_{i}*whited_{i} + \beta_{5}othracem_{i}*othraced_{i} + \beta_{6}hispm_{i}*hispd_{i} \
&+ \sum_{i=0}^{14}\delta_{i}\sum_{j=15}^{32}[agefstm = j] +\sum_{i=0}^{27}\alpha_{i}\sum_{j=15}^{43}[ag
\end{split}
\end{aligned}
$$
```

```
#run the models
M4 <- lm(samesex ~ agem + aged + educm + educd + blackm*blackd + whitem*whited +othracem*othraced + hisp

M5 <- lm(samesex ~ agem + aged + factor(educm) + factor(educd) + blackm*blackd + whitem*whited + othrac

M6 <- lm(samesex ~ agem + aged + factor(educm)*factor(educd) + blackm*blackd + whitem*whited + othracem

#get rmse's
RMSE_4 <- RMSE(M4)
RMSE_5 <- RMSE(M5)
RMSE_6 <- RMSE(M6)
```

```
stargazer(M4,
          M5,
          M6,
          type = "latex", title = "Linear Probability Model Regression
          Results* predicting samesex",
          header = F, omit = c('educm', 'educd', 'agefstd', 'agefstm'),
notes = "*Table omits dummy
         coefficients",
          multicolumn = F,
          font.size = "small",
          column.sep.width = '1pt',
          add.lines = list(c('RMSE', round(RMSE_4, 5), round(RMSE_5, 5), round(RMSE_6, 5))), table.plac
```

```
#create new data to use for prediction
newdata4 <- with(master, data.frame(agem = mean(agem), aged = mean(aged), educm=mean(educm), educd = mea

newdata5 <- with(master, data.frame(agem = mean(agem), aged = mean(aged), educm= 0, educd = 0, blackm = 

newdata6 <- with(master, data.frame(agem = mean(agem), aged = mean(aged), educm=0, educd = 0, blackm = 

#predict the probabilities of samesex =1
pred_M4 <- predict(M4, newdata4, type="response", se.fit = T)
pred_M5 <- predict(M5, newdata5, type="response", se.fit = T)
pred_M6 <- predict(M6, newdata6, type="response", se.fit = T)

#extract the predicted probabilities given the means of continuous variables
predprob_M4 <- pred_M4[[1]]
predprob_M5 <- pred_M5[[1]]
predprob_M6 <- pred_M6[[1]]

#extract the standard errors
```

```
SE_M4 <- pred_M4[[2]]
SE_M5 <- pred_M5[[2]]
SE_M6 <- pred_M6[[2]]

#report in a dataframe
samesex_prob_df <- data.frame(
  "Model" = c("M4", "M5", "M6"),
  "Predicted Probability" = c(predprob_M4, predprob_M5, predprob_M6),
  "Standard error" = c(SE_M4, SE_M5, SE_M6)
)
```

```
kable(samesex_prob_df, "latex", booktabs = T, align = "c", caption = "Estimated predicted probabilities
  kable_styling(latex_options = 'hold_position')%>%
   add_footnote("*Using mean values of continous variables")
```

```
$$
\begin{aligned}
\begin{split}
W1: workedm_{i} &= \beta_{0} + \beta_{1}morekids_{i} + u_{i} \\
W2: workedm_{i} &= \beta_{0} + \beta_{1}morekids_{i} + \beta_{2}educm_{i} + \sum_{i=0}^{11}\gamma_{i}[e
&+ \beta_{4}agem_{i} + \beta_{5}agem_{i}^2 + \beta_{6}agefstm_{i} + \beta_{7}agefstm_{i}^2 + \beta_{8}ag
&+ \beta_{10}agefstd_{i} + \beta_{11}agefstd_{i}^2 + \beta_{12}blackm_{i} + \beta_{13}hispm_{i} + \beta_
\end{split}
\end{aligned}
$$
```

```
#create new column with restricted education
master_w <- master %>%
  mutate(educm12 = as.numeric(educm < 12),
         educd12 = as.numeric(educd < 12))

#run models W1 and W2
W1 <- lm(workedm ~ morekids, data = master_w)
W2 <- lm(workedm ~ educm + educm12 + educd + educd12 + agem + poly(agem, 2) + agefstm + poly(agefstm, 2

#get rmse
RMSE_W1 <- RMSE(W1)
RMSE_W2 <- RMSE(W2)

#format coefficients for morekids so they display in stargazer
W1$coefficients[[2]] <- format(W1$coefficients[[2]], width = 4, digits = 3)
W2$coefficients[[21]] <- format(W2$coefficients[[21]], width = 4, digits = 3)

#create dataframe of results for table
W_df <- data.frame("Model" = c("W1", "W2"),
                   "Morekids Coef" = c(W1$coefficients[[2]], W2$coefficients[[21]]))
```

```
#display table of coefficients
kable(W_df, "latex", booktabs = T, align = "c", caption = "Estimated effect of morekids=1 on workedm", 
  kable_styling(latex_options = 'hold_position')
```

```
$$
\begin{aligned}
\begin{split}
Full \ Model: workedm_{i} &= \beta_{0} + \beta_{1}morekids_{i} + u_{i} \\
First \ Stage: morekids_{i} &= \pi_{0} + \pi_{1}samesex_{i} + \eta_{i} \\
Reduced \ Form: workedm_{i} &= \delta_{0} + \delta_{1}samesex_{i} + v_{i}
\end{split}
\end{aligned}
$$
```

```
#run each IV and 2SLS model
FS <- lm(morekids ~ samesex, data = master)
RF <- lm(workedm ~ samesex, data = master)
IV <- ivreg(workedm ~ morekids|samesex, data = master)
```

```
stargazer(FS,
          RF,
          IV,
          type = "latex", title = "FS, RF, and IV results",
          header = F,
          multicolumn = F,
          font.size = "small",
          column.sep.width = '1pt',
          table.placement = "H")
```

```
#calculate the ratio
IV_ratio <- format(RF$coefficients[[2]]/FS$coefficients[[2]], width = 4, digits = 3)

#format coefficients for morekids to display in table
RF$coefficients[[2]] <- format(RF$coefficients[[2]], width = 4, digits = 3)
FS$coefficients[[2]] <- format(FS$coefficients[[2]], width = 4, digits = 3)
IV$coefficients[[2]] <- format(IV$coefficients[[2]], width = 4, digits = 3)

#create dataframe of results for table
IV_df <- data.frame("Regression" = c("Ratio", "IV"),
                   "Morekids Coef" = c(IV_ratio, IV$coefficients[[2]]))
```

```
#display table of coefficients
kable(IV_df, "latex", booktabs = T, align = "c", caption = "Estimated coefficient on morekids using IV"
  kable_styling(latex_options = 'hold_position')
```

```
#estimate the IV model with controls
RF_cont <- lm(workedm ~ morekids + educm + educm12 + educd + educd12 + agem + poly(agem, 2) + agefstm +
FS_cont <- lm(morekids ~ samesex + educm + educm12 + educd + educd12 + agem + poly(agem, 2) + agefstm +
IV_cont <- ivreg(workedm ~ morekids + educm + educm12 + educd + educd12 + agem + poly(agem, 2) + agefst
```

```
stargazer(FS_cont,
          RF_cont,
          IV_cont,
          type = "latex", title = "FS, RF, and IV results (with 15 controls)",
          omit = c("educm", "educm12", "educd", "educd12", "agem", "poly(agem, 2)1", "poly(agem, 2)2",
          header = F,
          multicolumn = F,
          font.size = "small",
          column.sep.width = '1pt',
          table.placement = "H")
```

```
#calculate the ratio
IV_cont_ratio <- format(RF_cont$coefficients[[2]]/FS_cont$coefficients[[2]], width = 4, digits = 3)
```

```
#format coefficients for morekids to display in table
RF_cont$coefficients[[2]] <- format(RF_cont$coefficients[[2]], width = 4, digits = 3)
FS_cont$coefficients[[2]] <- format(FS_cont$coefficients[[2]], width = 4, digits = 3)
IV_cont$coefficients[[2]] <- format(IV_cont$coefficients[[2]], width = 4, digits = 3)

#create dataframe of results for table
IV_cont_df <- data.frame("Regression" = c("Ratio", "IV"),
                   "Morekids Coef" = c(IV_cont_ratio, IV_cont$coefficients[[2]]))
```

```
#display table of coefficients
kable(IV_cont_df, "latex", booktabs = T, align = "c", caption = "Estimated coefficient on morekids using
  kable_styling(latex_options = 'hold_position')
```

```
$$
\begin{aligned}
\begin{split}
&RF_{controls}: y_{i} = \delta_{0} + \delta_{1cont}z_{i} + x_{0i} + v_{i} \\
&E[z_{i}x_{0i}] = 0 \\
&FOC: E[z_{i}v_{i}] = 0 \\
&\Rightarrow E[z_{i}(y_{i} - \delta_{0} - \delta_{1cont}z_{i} - x_{0i})] = 0\\
&\Rightarrow E[z_{i}y_{i} - \delta_{0}z_{i} - z_{i}x_{0i}] = \delta_{1cont}E[z_{i}^2] \\
&\Rightarrow E[z_{i}y_{i}] - \delta_{0}E[z_{i}] - E[z_{i}x_{0i}] = \delta_{1cont}E[z_{i}^2] \\
&\Rightarrow E[z_{i}y_{i}] - \delta_{0}E[z_{i}] - 0 = \delta_{1cont}E[z_{i}^2] \\
&\Rightarrow \delta_{1cont} = \frac{E[z_{i}y_{i}] - \delta_{0}E[z_{i}]}{E[z_{i}^2]} \\
\\
&RF_{no \ controls}: y_{i} = \delta_{0} + \delta_{1}z_{i} + v_{i} \\
&FOC: E[z_{i}v_{i}] = 0 \\
&\Rightarrow E[z_{i}(y_{i} - \delta_{0} - \delta_{1}z_{i})] = 0\\
&\Rightarrow E[z_{i}y_{i} - \delta_{0}z_{i}] = \delta_{1}E[z_{i}^2] \\
&\Rightarrow E[z_{i}y_{i}] - \delta_{0}E[z_{i}] = \delta_{1}E[z_{i}^2] \\
&\Rightarrow E[z_{i}y_{i}] - \delta_{0}E[z_{i}] = \delta_{1}E[z_{i}^2] \\
&\Rightarrow \delta_{1cont} = \delta_{1} = \frac{E[z_{i}y_{i}] - \delta_{0}E[z_{i}]}{E[z_{i}^2]} \blacks
\end{split}
\end{aligned}
$$

$$
\begin{aligned}
\begin{split}
&FS_{controls}: x_{i} = \pi_{0} + \pi_{1cont}z_{i} + x_{0i} + \eta_{i} \\
&E[z_{i}x_{0i}] = 0 \\
&FOC: E[z_{i}\eta_{i}] = 0 \\
&\Rightarrow E[z_{i}(x_{i} - \pi_{0} - \pi_{1cont}z_{i} - x_{0i})] = 0\\
&\Rightarrow E[z_{i}x_{i} - \pi_{0}z_{i} - z_{i}x_{0i}] = \pi_{1cont}E[z_{i}^2] \\
&\Rightarrow E[z_{i}x_{i}] - \pi_{0}E[z_{i}] - E[z_{i}x_{0i}] = \pi_{1cont}E[z_{i}^2] \\
&\Rightarrow E[z_{i}x_{i}] - \pi_{0}E[z_{i}] - 0 = \pi_{1cont}E[z_{i}^2] \\
&\Rightarrow \pi_{1cont} = \frac{E[z_{i}x_{i}] - \pi_{0}E[z_{i}]}{E[z_{i}^2]} \\
\\
&FS_{no \ controls}: x_{i} = \pi_{0} + \pi_{1}z_{i} + \eta_{i} \\
&FOC: E[z_{i}\eta_{i}] = 0 \\
&\Rightarrow E[z_{i}(x_{i} - \pi_{0} - \pi_{1}z_{i})] = 0\\
&\Rightarrow E[z_{i}x_{i} - \pi_{0}z_{i}] = \pi_{1}E[z_{i}^2] \\
&\Rightarrow E[z_{i}x_{i}] - \pi_{0}E[z_{i}] = \pi_{1}E[z_{i}^2] \\
&\Rightarrow E[z_{i}x_{i}] - \pi_{0}E[z_{i}] = \pi_{1}E[z_{i}^2] \\
&\Rightarrow \pi_{1cont} = \pi_{1} = \frac{E[z_{i}x_{i}] - \pi_{0}E[z_{i}]}{E[z_{i}^2]} \blacksquare. \
\end{split}
\end{aligned}
$$

$$
H_{0}: \hat\beta_{IV \ no \ controls} = \hat\beta_{IV \ with \ controls}
$$
We calculate the t-stat with the $SE_{\beta_{IV \ no \ controls}}$:

$$
\begin{aligned}
\begin{split}
t &= \frac{\hat\beta_{IV \ no \ controls} - \hat\beta_{IV \ with \ controls}}{SE_{\hat\beta_{IV \ no \ c
&= \frac{1.82*10^{-13} - 1.69*10^{-13}}{1.98*10^{-12}} \\
&= 0.006565657
\end{split}
\end{aligned}
$$
We calculate the t-stat with the $SE_{\beta_{IV \ with \ controls}}$:

$$
\begin{aligned}
\begin{split}
t &= \frac{\hat\beta_{IV \ with \ controls} - \hat\beta_{IV \ no \ controls}}{SE_{\hat\beta_{IV \ with \
&= \frac{1.69*10^{-13} - 1.82*10^{-13}}{1.92*10^{-12}} \\
&= 0.006770833
\end{split}
\end{aligned}
$$

$$
\begin{aligned}
\begin{split}
&OLS1: y_{i} = \beta_{0} + \beta_{1}x_{i} + u_{i} \\
&OLS2: y_{i} = \beta_{0} + \beta_{1}x_{i} + x_{0i} + u_{i}\\
\end{split}
\end{aligned}
$$
```

```
#this section was done with Alina and Josh
#estimate earningsm on morekids
OLS1 <- lm(earningsm ~ morekids, data = master_w)
OLS2 <- lm(earningsm ~ morekids + educm + educm12 + educd + educd12 + agem + poly(agem, 2) + agefstm + p
IV1 <- ivreg(earningsm ~ morekids | samesex, data = master_w)
IV2 <- ivreg(earningsm ~ morekids + educm + educm12 + educd + educd12 + agem + poly(agem, 2) + agefstm +

#estimate earningsd on morekids
OLS3 <- lm(earningsd ~ morekids, data = master_w)
OLS4 <- lm(earningsd ~ morekids + educm + educm12 + educd + educd12 + agem + poly(agem, 2) + agefstm + p
IV3 <- ivreg(earningsd ~ morekids | samesex, data = master_w)
IV4 <- ivreg(earningsd ~ morekids + educm + educm12 + educd + educd12 + agem + poly(agem, 2) + agefstm +

#estimate earningsd on morekids
```

```
OLS5 <- lm(famearn ~ morekids, data = master_w)
OLS6 <- lm(famearn ~ morekids + educm + educm12 + educd + educd12 + agem + poly(agem, 2) + agefstm + pol
IV5 <- ivreg(famearn ~ morekids | samesex, data = master_w)
IV6 <- ivreg(famearn ~ morekids + educm + educm12 + educd + educd12 + agem + poly(agem, 2) + agefstm + p

#extract coefficients to place in table

#create dataframe for display
OLS_IV_df <- data.frame("Dependent Var" = c("earningsm", "earningsd", "famearn"),
                        "OLS" = c(OLS1$coefficients[[2]], OLS3$coefficients[[2]], OLS5$coefficients[[2]]
                        "IV" = c(IV1$coefficients[[2]], IV3$coefficients[[2]], IV5$coefficients[[2]]) ,
                        "OLS cont" = c(OLS2$coefficients[[2]], OLS4$coefficients[[2]], OLS6$coefficients
                        "IV cont" = c(IV2$coefficients[[2]], IV4$coefficients[[2]], IV6$coefficients[[2]
```

```
#display table of results
kable(OLS_IV_df, "latex", booktabs = T, align = "c", caption = "IV and OLS results* for morekids on 3 de
  kable_styling(latex_options = 'hold_position')%>%
   add_footnote("*The control variables are the same 15 in (3).")
```

```
$$
\begin{aligned}
\begin{split}
&y_{1i} = \beta_{1}x_{i} + u_{1i} \\
&y_{2i} = \beta_{2}x_{i} + u_{2i} \\
&y_{3i} = \beta_{3}x_{i} + u_{3i} \\
&\hat\beta_{3} = E[x_{i}^2]^{-1}E[x_{i}y_{3i}] \ by \ definition \\
&\Rightarrow y_{3i} = (\beta_{1}x_{i} + u_{1i}) + (\beta_{2}x_{i} + u_{2i}) \\
&\Rightarrow y_{3i} = x_{i}(\beta_{1} + \beta_{2}) + (u_{1i} + u_{2i}) \\
&\Rightarrow \underbrace{(u_{1i} + u_{2i})}_{u_{i}} = y_{3i} - x_{i}(\beta_{1} + \beta_{2})\\
&FOC: E[x_{i}u_{i}] = 0 \\
&\Rightarrow E[x_{i}(y_{3i} - x_{i}(\beta_{1} + \beta_{2}))] = 0 \\
&\Rightarrow E[x_{i}y_{3i}] = E[x_{i}^2(\beta_{1} + \beta_{2})] \\
&\Rightarrow (\hat\beta_{1} + \hat\beta_{2})E[x_{i}^2]  = E[x_{i}y_{3i}] \\
&\Rightarrow (\hat\beta_{1} + \hat\beta_{2}) = E[x_{i}^2]^{-1} E[x_{i}y_{3i}] = \hat\beta_{3} \blacksqua
\end{split}
\end{aligned}
$$

$$
\begin{aligned}
\begin{split}
&OLS_{m}: \frac{\hat\beta_{1}}{\hat\beta_{3}} = \frac{3018.855}{5413.692} = 0.5576\\
&OLS_{mcontrol}: \frac{\hat\beta_{1}}{\hat\beta_{3}} = \frac{3972.369}{5562.061} = 0.7141\\
\end{split}
\end{aligned}
$$
$$
\begin{aligned}
\begin{split}
&OLS_{d}: \frac{\hat\beta_{1}}{\hat\beta_{3}} = \frac{2394.837}{5413.692} = 0.4423\\
&OLS_{dcontrol}: \frac{\hat\beta_{1}}{\hat\beta_{3}} = \frac{1589.691}{5562.061} = 0.2858\\
\end{split}
\end{aligned}
$$
```

```
$$
\begin{aligned}
\begin{split}
&IV_{m}: \frac{\hat\beta_{1}}{\hat\beta_{3}} = \frac{1552.796}{4198.800} = 0.3698\\
&IV_{mcontrol}: \frac{\hat\beta_{1}}{\hat\beta_{3}} = \frac{1240.394}{4155.845} = 0.2984\\
\end{split}
\end{aligned}
$$

To calculate the share of the total effect which is driven by fathers, we take $\frac{\hat\beta_{1}}{\ha

$$
\begin{aligned}
\begin{split}
&IV_{d}: \frac{\hat\beta_{1}}{\hat\beta_{3}} = \frac{2646.003}{4198.800} = 0.6301\\
&IV_{dcontrol}: \frac{\hat\beta_{1}}{\hat\beta_{3}} = \frac{2915.451}{4155.845} = 0.7015\\
\end{split}
\end{aligned}
$$

$$
\begin{aligned}
\begin{split}
C: &E[morekids_{treatment} - morekids_{control}] \\
AT: &E[morekids_{control}] \\
NT: &1 - E[morekids_{treatment}] \\
\end{split}
\end{aligned}
$$
```

```
#define the treatment and control
treatment <- subset(master_w, samesex == 1)
control <- subset(master_w, samesex == 0)

#(i-iii) calculate the compliers, AT, NT
comp <- (sum(treatment$morekids)/length(treatment$morekids)) - (sum(control$morekids)/length(control$mo
at <- (sum(control$morekids)/length(control$morekids))
nt <- 1 - (sum(treatment$morekids)/length(treatment$morekids))

#create vector for displaying in table
tot_frac <- c(comp, at, nt)

#(iv) fractions for subgroups
#treatment w/educm restrictions
tl12 = subset(master_w[which(master_w$educm<12), ], samesex == 1)
t12 = subset(master_w[which(master_w$educm==12), ], samesex == 1)
t13 = subset(master_w[which(master_w$educm>12&master_w$educm<15), ], samesex == 1)
t16 = subset(master_w[which(master_w$educm>16), ], samesex == 1)

#control w/educm restrictions
cl12 = subset(master_w[which(master_w$educm<12), ], samesex == 0)
c12 = subset(master_w[which(master_w$educm==12), ], samesex == 0)
c13 = subset(master_w[which(master_w$educm>12&master_w$educm<15), ], samesex == 0)
```

```
c16 = subset(master_w[which(master_w$educm>16), ], samesex == 0)

#calculate the AT, C and NT ratios
#compliers
compl12 = (sum(tl12$morekids)/length(tl12$morekids)) - (sum(cl12$morekids)/length(cl12$morekids))
comp12 = (sum(t12$morekids)/length(t12$morekids)) - (sum(c12$morekids)/length(c12$morekids))
comp13 = (sum(t13$morekids)/length(t13$morekids)) - (sum(c13$morekids)/length(c13$morekids))
comp16 = (sum(t16$morekids)/length(t16$morekids)) - (sum(c16$morekids)/length(c16$morekids))

#AT
atl12 = (sum(cl12$morekids)/length(cl12$morekids))
at12 = (sum(c12$morekids)/length(c12$morekids))
at13 = (sum(c13$morekids)/length(c13$morekids))
at16 = (sum(c16$morekids)/length(c16$morekids))

#NT
ntl12 = 1 - (sum(tl12$morekids)/length(tl12$morekids))
nt12 = 1 - (sum(t12$morekids)/length(t12$morekids))
nt13 = 1 - (sum(t13$morekids)/length(t13$morekids))
nt16 = 1 - (sum(t16$morekids)/length(t16$morekids))

#combine into a table
compll <- c(LessThan12 = compl12, EqTo12 = comp12, Between13And15 = comp13,
MoreThan16 = comp16)
att <- c(LessThan12 = atl12, EqTo12 = at12, Between13And15 = at13, MoreThan16 = at16)
ntt <- c(LessThan12 = ntl12, EqTo12 = nt12, Between13And15 = nt13, MoreThan16 = nt16)

#create table for display
frac_df <- as.data.frame(rbind(C = compll, AT = att, NT = ntt)) %>%
  mutate(WholeSample = tot_frac)
```

```
#display data table
kable(frac_df, "latex", booktabs = T, align = "c", caption = "Estimated always takers (AT), compliers (
  kable_styling(latex_options = 'hold_position')
```

```
$$
IV: x_{i}*morekids = \delta_{0} + \delta_{1}morekids_{i} + u_{i}
$$
```

```
#create the interactions for the dependent variables
master_w_iv <- master_w %>%
  mutate(educ_12_more = educm12*morekids,
         educ_12_mean = educm*morekids,
         agefstm_mean = agefstm*morekids,
         agefstm_21_more = (ifelse(agefstm < 21, 1, 0))*morekids,
         hispm_mean = hispm*morekids,
         blackm_mean = blackm*morekids,
         whitem_mean = whitem*morekids,
         Utahm_mean = ifelse(st == 87, 1, 0)*morekids)

#mean frac of C moms with <12 schooling
reg_4 <- ivreg(educ_12_more ~ morekids + educm + educm12 + educd + educd12 + agem + poly(agem, 2) + age

#mean education of C moms
```

```
reg_4_1 <- ivreg(educ_12_mean ~ morekids + educm + educm12 + educd + educd12 + agem + poly(agem, 2) + ag

#mean age of C moms at first birth
reg_4_2 <- ivreg(agefstm_mean ~ morekids + educm + educm12 + educd + educd12 + agem + poly(agem, 2) + ag

#fraction of C moms with agefstm < 21
reg_4_3 <- ivreg(agefstm_21_more ~ morekids + educm + educm12 + educd + educd12 + agem + poly(agem, 2) +

#fraction of C moms hispanic
reg_4_4 <- ivreg(hispm_mean ~ morekids + educm + educm12 + educd + educd12 + agem + poly(agem, 2) + age

#fraction of C moms black
reg_4_5 <- ivreg(blackm_mean ~ morekids + educm + educm12 + educd + educd12 + agem + poly(agem, 2) + ag

#fraction of C moms white non hispanic
reg_4_6 <- ivreg(whitem_mean ~ morekids + educm + educm12 + educd + educd12 + agem + poly(agem, 2) + ag

#fraction of C mom from Utah
reg_4_7 <- ivreg(Utahm_mean ~ morekids + educm + educm12 + educd + educd12 + agem + poly(agem, 2) + age

#calculate mean for all families
frac_c_12 <- mean(master_w$educm12)
frac_c_educ <- mean(master_w$educm)
frac_c_agefstm <- mean(master_w$agefstm)
frac_c_agefstm21 <- mean(as.numeric(master_w$agefstm < 21))
frac_c_hisp <- mean(master_w$hispm)
frac_c_black <- mean(master_w$blackm)
frac_c_white <- mean(master_w$whitem)
frac_c_utah <- mean(as.numeric(master_w$st == 87))

#create dataframe for comparisons
compliers_df <- data.frame("Sample" = c("Compliers", "All Families"),
                           "educm<12" = c(reg_4$coefficients[[2]], frac_c_12),
                           "mean(educm)" = c(reg_4_1$coefficients[[2]], frac_c_educ),
                           "mean(agefstm)" = c(reg_4_2$coefficients[[2]], frac_c_agefstm),
                           "frac(agefstm<21)" = c(reg_4_3$coefficients[[2]], frac_c_agefstm21),
                           "frac(hisp)" = c(reg_4_4$coefficients[[2]], frac_c_hisp),
                           "frac(black)" = c(reg_4_5$coefficients[[2]], frac_c_black),
                           "frac(white)" = c(reg_4_6$coefficients[[2]], frac_c_white),
                           "frac(Utah)" = c(reg_4_7$coefficients[[2]], frac_c_utah))
```

```
#display table of results
kable(compliers_df, "latex", booktabs = T, align = "c", caption = "Means of compliers vs. all families"
  kable_styling(latex_options = 'hold_position')
```

```
$$
\begin{aligned}
\begin{split}
logwage_{i} = \beta_{0} + \beta_{1}aged_{i} + \sum_{i=1}^{51}\delta_{i}[st = i] + \sum_{i=1}^{51}\gamma_
\end{split}
\end{aligned}
$$
```

```
#filter and sample from the
master_d16 <- master %>%
  filter(educd == 16)

master_d16_1 <- master_d16 %>% mutate(logwaged = log(waged),
         rv = sample(1:nrow(master_d16), size = nrow(master_d16), replace = F)/nrow(master_d16))

#create the holdout and estimation samples
estimation <- subset(master_d16_1, rv <0.75)
holdout <- subset(master_d16_1, rv >=0.75)

#estimate the model
OLS_simple <- lm(logwaged ~ aged + factor(st) + factor(st)*aged, data = estimation)
```

```
stargazer(OLS_simple,
          type = "latex", title = "Simple OLS Regression Results*",
          header = F, omit = c('st', "aged:factor(st)"),
notes = "*Table omits dummy
         coefficients",
          multicolumn = F,
          font.size = "small",
          column.sep.width = '1pt')
```

```
#new datasets with 30 and 35
OLS_simple_newdata_26 <- subset(estimation, aged == 26)
OLS_simple_newdata_35 <- subset(estimation, aged == 35)

#predict the models with the age restrictions
pred_OLS_aged_26 <- predict(OLS_simple, OLS_simple_newdata_26)
pred_OLS_aged_35 <- predict(OLS_simple, OLS_simple_newdata_35)

#append the predict columns
OLS_simple_newdata_26$pred_logwage26 <- pred_OLS_aged_26
OLS_simple_newdata_35$pred_logwage35 <- pred_OLS_aged_35

#find the means by state
OLS_26 <- OLS_simple_newdata_26 %>%
  group_by(st) %>%
  summarize(pred_logwaged = mean(pred_logwage26))
OLS_35 <- OLS_simple_newdata_35 %>%
  group_by(st) %>%
  summarize(pred_logwaged = mean(pred_logwage35))

#read in state classifications
state_names <- read.csv("/Users/sofia/Box/Cal (sofiaguo@berkeley.edu)/2018-19/Spring 2019/Econ 142/Final

#change column names
colnames(state_names) <- c("st", "State Name")

#create dataframes
logwage_df26 <- merge(OLS_26, state_names, by = "st")
logwage_df35 <- merge(OLS_35, state_names, by = "st")
```

```
#visualize the predictions
ggplot(logwage_df26, aes(`State Name`, pred_logwaged))+
  geom_point() +
  labs(x="State of Residence", y="Mean log(wage)", title = 'OLS: predicted mean log(wage) of 26 year ol
       caption = "*with 16 years of education") +
  theme(axis.text.x = element_text(angle=90))
```

```
#visualize the predictions
ggplot(logwage_df35, aes(`State Name`, pred_logwaged))+
  geom_point() +
  labs(x="State of Residence", y="Mean log(wage)", title = 'OLS: predicted mean log(wage) of 35 year ol
       caption = "*with 16 years of education") +
  theme(axis.text.x = element_text(angle=90))
```

```
#create state dummies
estimation_dum <- dummy_cols(estimation, select_column="st") %>%
  dplyr::select(c(st, aged, logwaged, 43:93))

#create interaction columns
estimation_dum$int11 = estimation_dum$aged*estimation_dum$st_11
estimation_dum$int12 = estimation_dum$aged*estimation_dum$st_12
estimation_dum$int13 = estimation_dum$aged*estimation_dum$st_13
estimation_dum$int14 = estimation_dum$aged*estimation_dum$st_14
estimation_dum$int15 = estimation_dum$aged*estimation_dum$st_15
estimation_dum$int16 = estimation_dum$aged*estimation_dum$st_16
estimation_dum$int21 = estimation_dum$aged*estimation_dum$st_21
estimation_dum$int22 = estimation_dum$aged*estimation_dum$st_22
estimation_dum$int23 = estimation_dum$aged*estimation_dum$st_23
estimation_dum$int31 = estimation_dum$aged*estimation_dum$st_31
estimation_dum$int32 = estimation_dum$aged*estimation_dum$st_32
estimation_dum$int33 = estimation_dum$aged*estimation_dum$st_33
estimation_dum$int34 = estimation_dum$aged*estimation_dum$st_34
estimation_dum$int35 = estimation_dum$aged*estimation_dum$st_35
estimation_dum$int41 = estimation_dum$aged*estimation_dum$st_41
estimation_dum$int42 = estimation_dum$aged*estimation_dum$st_42
estimation_dum$int43 = estimation_dum$aged*estimation_dum$st_43
estimation_dum$int44 = estimation_dum$aged*estimation_dum$st_44
estimation_dum$int45 = estimation_dum$aged*estimation_dum$st_45
estimation_dum$int46 = estimation_dum$aged*estimation_dum$st_46
estimation_dum$int47 = estimation_dum$aged*estimation_dum$st_47
estimation_dum$int51 = estimation_dum$aged*estimation_dum$st_51
estimation_dum$int52 = estimation_dum$aged*estimation_dum$st_52
estimation_dum$int53 = estimation_dum$aged*estimation_dum$st_53
estimation_dum$int54 = estimation_dum$aged*estimation_dum$st_54
estimation_dum$int55 = estimation_dum$aged*estimation_dum$st_55
estimation_dum$int56 = estimation_dum$aged*estimation_dum$st_56
estimation_dum$int57 = estimation_dum$aged*estimation_dum$st_57
estimation_dum$int58 = estimation_dum$aged*estimation_dum$st_58
estimation_dum$int59 = estimation_dum$aged*estimation_dum$st_59
estimation_dum$int61 = estimation_dum$aged*estimation_dum$st_61
estimation_dum$int62 = estimation_dum$aged*estimation_dum$st_62
estimation_dum$int63 = estimation_dum$aged*estimation_dum$st_63
estimation_dum$int64 = estimation_dum$aged*estimation_dum$st_64
estimation_dum$int71 = estimation_dum$aged*estimation_dum$st_71
```

```
estimation_dum$int72 = estimation_dum$aged*estimation_dum$st_72
estimation_dum$int73 = estimation_dum$aged*estimation_dum$st_73
estimation_dum$int74 = estimation_dum$aged*estimation_dum$st_74
estimation_dum$int81 = estimation_dum$aged*estimation_dum$st_81
estimation_dum$int82 = estimation_dum$aged*estimation_dum$st_82
estimation_dum$int83 = estimation_dum$aged*estimation_dum$st_83
estimation_dum$int84 = estimation_dum$aged*estimation_dum$st_84
estimation_dum$int85 = estimation_dum$aged*estimation_dum$st_85
estimation_dum$int86 = estimation_dum$aged*estimation_dum$st_86
estimation_dum$int87 = estimation_dum$aged*estimation_dum$st_87
estimation_dum$int88 = estimation_dum$aged*estimation_dum$st_88
estimation_dum$int91 = estimation_dum$aged*estimation_dum$st_91
estimation_dum$int92 = estimation_dum$aged*estimation_dum$st_92
estimation_dum$int93 = estimation_dum$aged*estimation_dum$st_93
estimation_dum$int94 = estimation_dum$aged*estimation_dum$st_94
estimation_dum$int95 = estimation_dum$aged*estimation_dum$st_95
```

```
#from https://www.datacamp.com/community/tutorials/tutorial-ridge-lasso-elastic-net

# Load libraries, get data & set seed for reproducibility ---------------------
set.seed(123)    # seed for reproducibility

#transform x and y into matrices
x <- as.matrix(estimation_dum[,-3])
y <- as.matrix(estimation_dum[,3])

# Setting alpha = 1 implements lasso regression
lasso_cv <- cv.glmnet(x, y, alpha = 1)

lambda_opt <- lasso_cv$lambda.min

#fit the model with optimal lambda
mod_2 <- glmnet(x, y, alpha=1, lambda = lambda_opt, standardize = T)

#calculate the RMSE
y_hat <- predict(mod_2, x)
RMSE <- sqrt(mean(y_hat - y)^2)
```

```
#predict earnings for age 26 and age 35
#subset the data for each age
est_26 <- estimation_dum %>%
  filter(aged == 26)

est_35 <- estimation_dum %>%
  filter(aged == 35)

#estimate the logwage given the lasso model
est_26$pred_26_lasso <- predict(mod_2, as.matrix(est_26[ ,-3]))
est_35$pred_35_lasso <- predict(mod_2, as.matrix(est_35[ ,-3]))

#create dataframes for graphing
est_26_pred_df <- est_26 %>%
  group_by(st) %>%
  summarize(mean_pred = mean(pred_26_lasso)) %>%
  merge(state_names, by="st")

#create dataframes for graphing
est_35_pred_df <- est_35 %>%
  group_by(st) %>%
  summarize(mean_pred = mean(pred_35_lasso)) %>%
  merge(state_names, by="st")
```

```
#visualize the predictions
ggplot(est_26_pred_df, aes(`State Name`, mean_pred))+
  geom_point() +
  labs(x="State of Residence", y="Mean log(wage)", title = 'Lasso: predicted mean log(wage) of 26 year 
       caption = "*with 16 years of education") +
  theme(axis.text.x = element_text(angle=90))
```

```
#visualize the predictions
ggplot(est_35_pred_df, aes(`State Name`, mean_pred))+
  geom_point() +
  labs(x="State of Residence", y="Mean log(wage)", title = 'Lasso: predicted mean log(wage) of 35 year 
       caption = "*with 16 years of education") +
  theme(axis.text.x = element_text(angle=90))
```

```
#find the max lwage from OLS for 26
OLS_h26 <- logwage_df26 %>%
  arrange(desc(pred_logwaged)) %>%
  slice(1)

colnames(OLS_h26) <- c("st", "meanlogwage", "State Name")

#find the max lwage from lasso for 26
LA_h26 <- est_26_pred_df %>%
  arrange(desc(mean_pred)) %>%
  slice(1)
colnames(LA_h26) <- c("st", "meanlogwage", "State Name")

#find the max lwage from OLS for 35
OLS_h35 <- logwage_df35 %>%
  arrange(desc(pred_logwaged)) %>%
  slice(1)
colnames(OLS_h35) <- c("st", "meanlogwage", "State Name")

#find the max lwage from lasso for 35
LA_h35 <- est_35_pred_df %>%
  arrange(desc(mean_pred)) %>%
  slice(1)
colnames(LA_h35) <- c("st", "meanlogwage", "State Name")

#create dataframe
best_df <- as.data.frame(rbind(OLS_h26, LA_h26, OLS_h35, LA_h35)) %>%
  mutate(Model = c("OLS", "LASSO", "OLS", "LASSO"),
         Age = c(26, 35, 26, 35)) %>%
  dplyr::select(c("Model", "Age", "State Name", "meanlogwage"))
```

```
#display table of results
kable(best_df, "latex", booktabs = T, align = "c", caption = "States with the highest male mean log wag
  kable_styling(latex_options = 'hold_position')
```

```
#create state dummies
hold_dum <- dummy_cols(holdout, select_column="st") %>%
  dplyr::select(c(st, aged, logwaged, 43:93))

#create interaction columns
hold_dum$int11 = hold_dum$aged*hold_dum$st_11
hold_dum$int12 = hold_dum$aged*hold_dum$st_12
hold_dum$int13 = hold_dum$aged*hold_dum$st_13
hold_dum$int14 = hold_dum$aged*hold_dum$st_14
hold_dum$int15 = hold_dum$aged*hold_dum$st_15
hold_dum$int16 = hold_dum$aged*hold_dum$st_16
hold_dum$int21 = hold_dum$aged*hold_dum$st_21
hold_dum$int22 = hold_dum$aged*hold_dum$st_22
hold_dum$int23 = hold_dum$aged*hold_dum$st_23
hold_dum$int31 = hold_dum$aged*hold_dum$st_31
hold_dum$int32 = hold_dum$aged*hold_dum$st_32
hold_dum$int33 = hold_dum$aged*hold_dum$st_33
hold_dum$int34 = hold_dum$aged*hold_dum$st_34
hold_dum$int35 = hold_dum$aged*hold_dum$st_35
hold_dum$int41 = hold_dum$aged*hold_dum$st_41
hold_dum$int42 = hold_dum$aged*hold_dum$st_42
hold_dum$int43 = hold_dum$aged*hold_dum$st_43
hold_dum$int44 = hold_dum$aged*hold_dum$st_44
hold_dum$int45 = hold_dum$aged*hold_dum$st_45
hold_dum$int46 = hold_dum$aged*hold_dum$st_46
hold_dum$int47 = hold_dum$aged*hold_dum$st_47
hold_dum$int51 = hold_dum$aged*hold_dum$st_51
hold_dum$int52 = hold_dum$aged*hold_dum$st_52
hold_dum$int53 = hold_dum$aged*hold_dum$st_53
hold_dum$int54 = hold_dum$aged*hold_dum$st_54
hold_dum$int55 = hold_dum$aged*hold_dum$st_55
hold_dum$int56 = hold_dum$aged*hold_dum$st_56
hold_dum$int57 = hold_dum$aged*hold_dum$st_57
hold_dum$int58 = hold_dum$aged*hold_dum$st_58
hold_dum$int59 = hold_dum$aged*hold_dum$st_59
hold_dum$int61 = hold_dum$aged*hold_dum$st_61
hold_dum$int62 = hold_dum$aged*hold_dum$st_62
hold_dum$int63 = hold_dum$aged*hold_dum$st_63
hold_dum$int64 = hold_dum$aged*hold_dum$st_64
hold_dum$int71 = hold_dum$aged*hold_dum$st_71
hold_dum$int72 = hold_dum$aged*hold_dum$st_72
hold_dum$int73 = hold_dum$aged*hold_dum$st_73
hold_dum$int74 = hold_dum$aged*hold_dum$st_74
hold_dum$int81 = hold_dum$aged*hold_dum$st_81
hold_dum$int82 = hold_dum$aged*hold_dum$st_82
hold_dum$int83 = hold_dum$aged*hold_dum$st_83
hold_dum$int84 = hold_dum$aged*hold_dum$st_84
hold_dum$int85 = hold_dum$aged*hold_dum$st_85
hold_dum$int86 = hold_dum$aged*hold_dum$st_86
hold_dum$int87 = hold_dum$aged*hold_dum$st_87
hold_dum$int88 = hold_dum$aged*hold_dum$st_88
hold_dum$int91 = hold_dum$aged*hold_dum$st_91
```

```
hold_dum$int92 = hold_dum$aged*hold_dum$st_92
hold_dum$int93 = hold_dum$aged*hold_dum$st_93
hold_dum$int94 = hold_dum$aged*hold_dum$st_94
hold_dum$int95 = hold_dum$aged*hold_dum$st_95
```

```
#Alina, Josh and I worked on this part together
RMSE_ols_all = c(0)
RMSE_lasso_all = c(0)
plot1 <- data.frame(Age = sort(unique(hold_dum$aged)))
k=1
for(i in sort(unique(hold_dum$aged))){
  hold_dum_temp <- subset(hold_dum, aged == i)
  hold_dum_temp$pred <- predict(OLS_simple, hold_dum_temp)

  y_hat = hold_dum_temp$pred
  y = hold_dum_temp$logwaged
  RMSE_ols_all[k] <- sqrt(mean(y_hat-y)^2)

  temp_train = hold_dum_temp[ ,1:105]

  x_train = as.matrix(temp_train[, -3])
  y_train = as.matrix(temp_train[ , 3])

  y_hat_l <- predict(mod_2, x_train)
  RMSE_lasso_all[k] = sqrt(mean(y_hat_l - y_train)^2)

  k=k+1
}

plot_2 <- cbind(plot1, OLS = RMSE_ols_all, Lasso = RMSE_lasso_all)

test_data_long_1 <- melt(plot_2, id="Age")  # convert to long format
```

```
#also collaborated code with Alina and Josh
ggplot(data=test_data_long_1,
       aes(x=Age, y=value, colour=variable)) +
  geom_line() +
  xlab("Male Age") +
  ylab("RMSE value") +
  ggtitle("RMSE of OLS vs. Lasso")
```