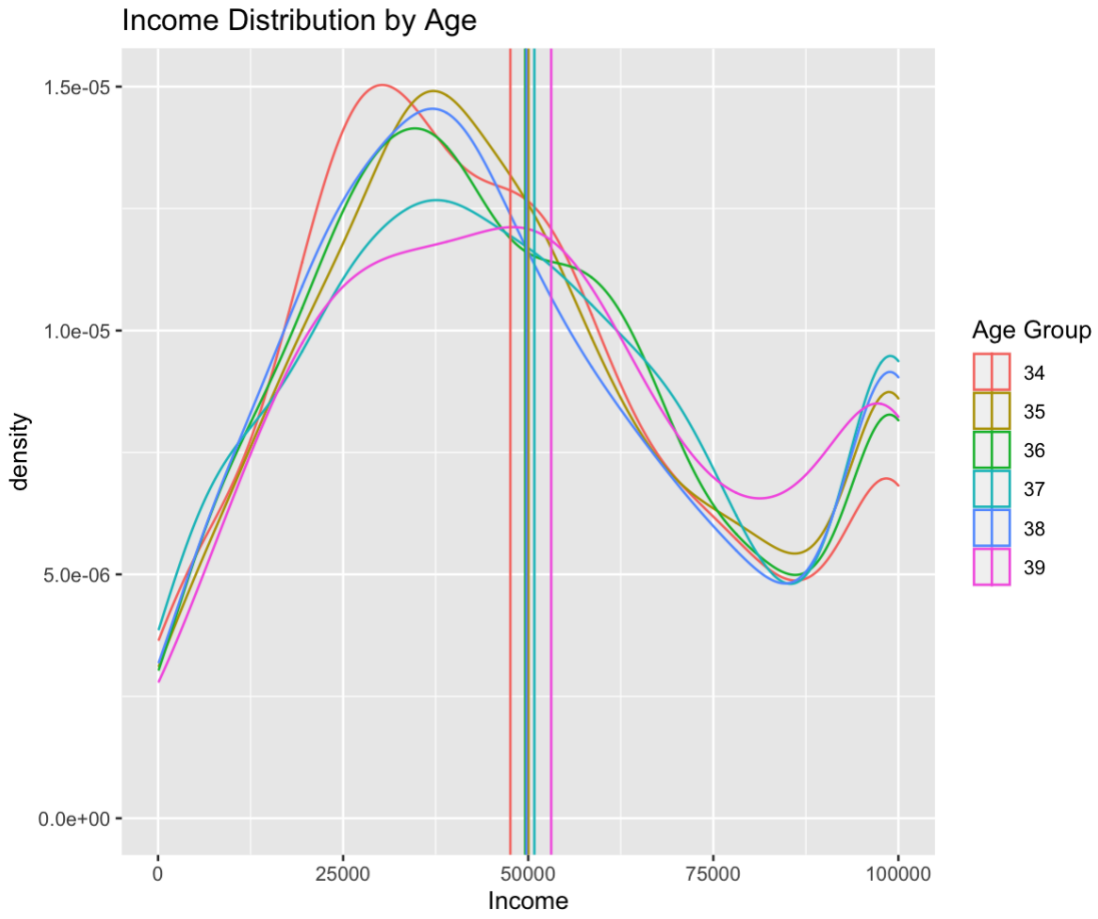
Assignment 4
ECON 613
Julian Sauvage

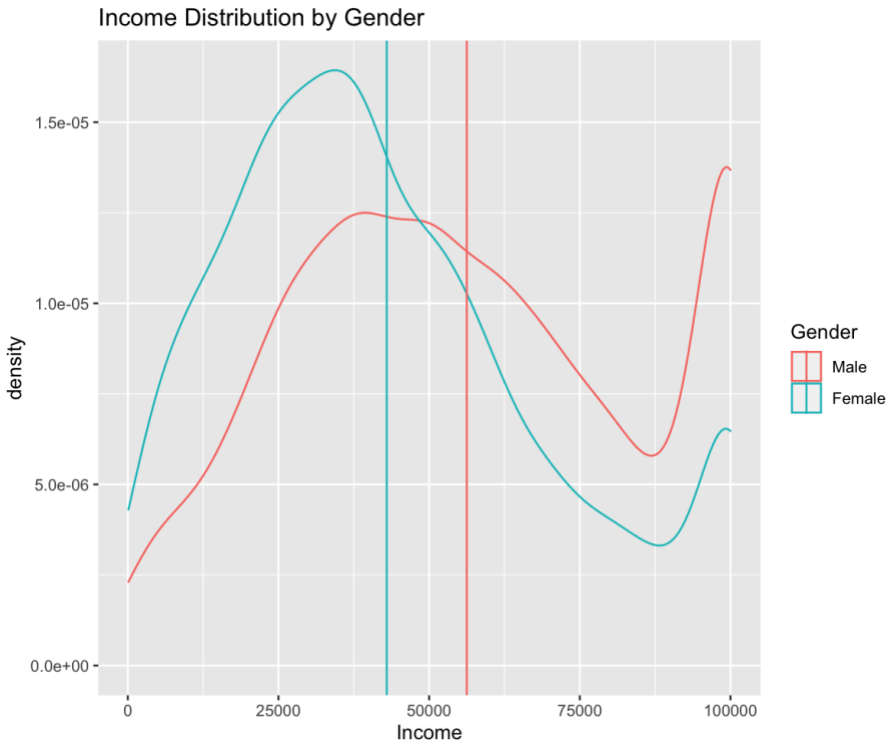
Exercise 1

In the following income distributions, each group's mean is represented by the appropriately colored vertical line.

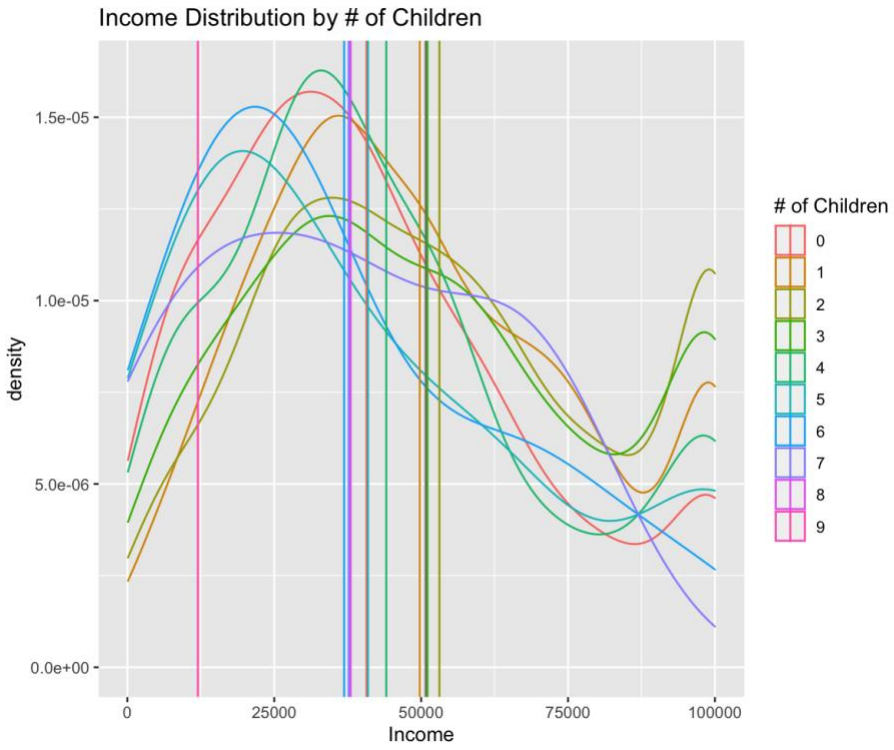
Income by Age Group:



Income by Gender:



Income by # of Children:



Share of “0” in income data by age group:

age	Count0	Total	Share0
34	6	793	0.007566204
35	8	1112	0.007194245
36	8	1114	0.007181329
37	8	1123	0.007123776
38	4	1018	0.003929273
39	2	252	0.007936508

Share of “0” in income data by gender group:

sex	Count0	Total	Share0
Male	21	2800	0.007500000
Female	15	2612	0.005742726

Share of “0” in income data by number of children:

kids	Count0	Total	Share0
0	8	537	0.014897579
1	9	1147	0.007846556
2	8	1393	0.005743001
3	5	623	0.008025682

- No “0” income observations for respondents with more than 3 children in household

Share of “0” in income data by marital status:

marstat	Count0	Total	Share0
Never-married	11	1947	0.005649718
Married	20	2683	0.007454342
Separated	4	93	0.043010753
Divorced	1	650	0.001538462

Interpretation of visualizations:

Men not only make more than women on average, but their income is also more likely to have their income top-coded, relative to women. This suggests that if we do not account for top-coded nature of the data, we will undervalue gender disparities in the income distribution.

In terms of age, we see that the median and mean wage increases alongside age, though these increases appear small. There are no major differences in the proportion of top-coded incomes based on age, though individuals who are younger than 35 are marginally less likely to have reached the upper-most income bracket.

Individuals with 1-3 children have a higher average income relative to those without children, but average income decreases as more children are in the household. These dynamics are likely due to a selection mechanism whereby individuals with higher income decide to have additional children, up to a point. There is also potentially a correlation with age, whereas an individual ages their income increases alongside their likelihood of having additional children. The large dropoff in mean income associated with having 9 children is likely due to small sample size.

For the most part there are no large discrepancies in terms of the share of “0”s appearing in the income data. We do see that individuals who are married or separated are more likely to report no income, though the differences are not dramatic. It is worth pointing out that no respondents with more than 3 children reported an income of “0”.

Exercise 2

1) Linear Model specification:

Income = educ_years + work_exp + age + sex + marstat + kids

Results:

```
lm(formula = income ~ educ_years + work_exp + age + sex + marstat +
    kids, data = nlsy)

Residuals:
   Min     1Q Median     3Q    Max
-80461 -16678  -2392  16832  99194

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)        9941.38    9822.38   1.012   0.3115
educ_years         2177.26      96.47  22.569  < 2e-16 ***
work_exp           1090.76      74.16  14.709  < 2e-16 ***
age                 140.71     268.50   0.524   0.6003
sexFemale        -18180.36     797.54 -22.796  < 2e-16 ***
marstatMarried     9279.29     970.95   9.557  < 2e-16 ***
marstatSeparated    581.65    2908.91   0.200   0.8415
marstatDivorced    6237.97    1300.60   4.796 1.68e-06 ***
marstatWidowed      973.06    6542.14   0.149   0.8818
kids                707.08     356.52   1.983   0.0474 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 24270 on 3922 degrees of freedom
Multiple R-squared:  0.2956,    Adjusted R-squared:  0.294
F-statistic: 182.9 on 9 and 3922 DF,  p-value: < 2.2e-16
```

This regression suggests that every additional year of education raises wages by $2,400, while an additional year of work experience raises wages by about $1100. It also suggests that being female reduces wages by about $18,000 dollars, although having children raises wages by $707 per child. It also suggests that individuals who are married have significantly higher incomes than those who are single, this is also true of individuals who are divorced.

A selection problem may exist wherein women are more likely to leave the labor force upon having children. This would produce a strong negative relationship between income and being female. Another selection problem that may arise is that individuals who have higher incomes are more likely to decide to get married and start a family, thus we see income is higher for respondents who are/were married and/or have children. The causal path may be that higher income leads to marriage/children, rather than marriage/children leads to higher income.

2) Heckman model can deal with selection problem by estimating first the likelihood that an individual is employed (income > 0), then incorporates these predicted probabilities in a standard regression to correct for selection bias.

3) Heckman model specification:

1st step: Pr(Income > 0) = work_exp + age + sex + marstat + kids
2nd step: Income = educ_years + work_exp + age + sex + marstat + kids + inverseMillsRatio

Results:

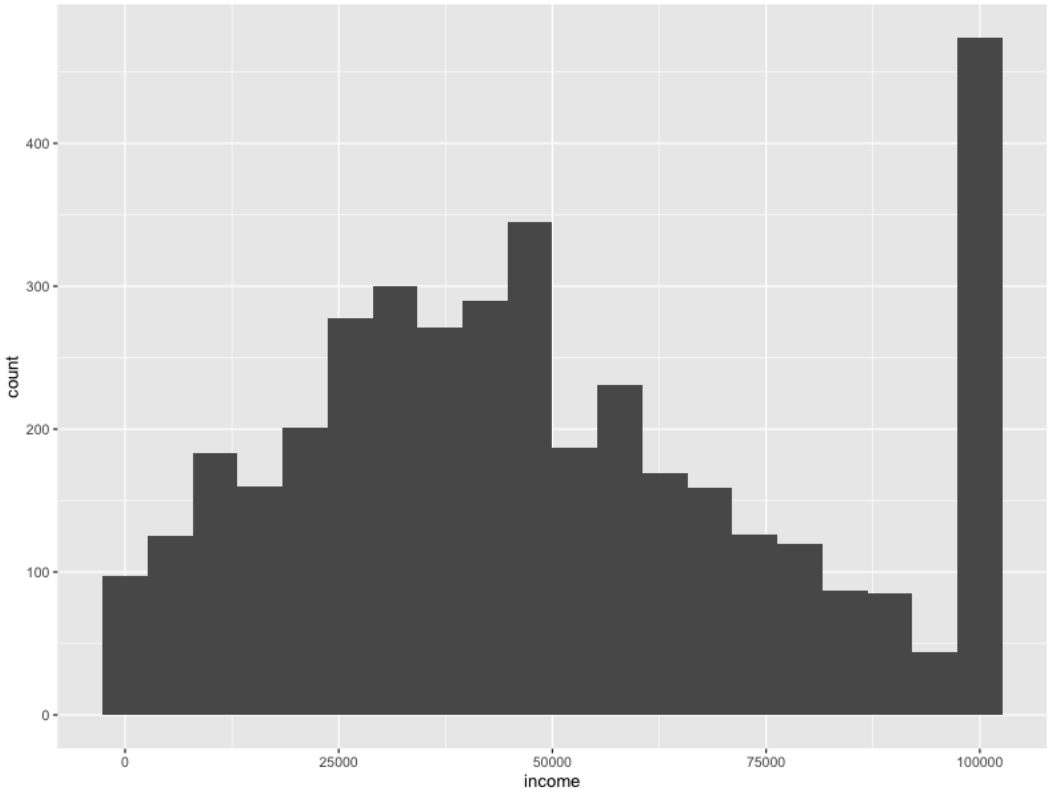
```
lm(formula = income ~ educ_years + work_exp + age + sex + marstat +
    kids + imr, data = nlsy)

Residuals:
   Min     1Q Median     3Q    Max
-79282 -16684  -2327  16799  98848

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)       29648.47   12171.17   2.436  0.01490 *
educ_years         2165.31      96.49  22.441  < 2e-16 ***
work_exp            921.98      96.39   9.565  < 2e-16 ***
age                -245.99     303.19  -0.811  0.41721
sexFemale        -17687.94     816.93 -21.652  < 2e-16 ***
marstatMarried    11287.23    1216.19   9.281  < 2e-16 ***
marstatSeparated  17245.32    6745.11   2.557  0.01060 *
marstatDivorced    4256.72    1487.44   2.862  0.00424 **
marstatWidowed    -1628.70    6605.46  -0.247  0.80526
kids               -531.72     575.89  -0.923  0.35591
imr             -177505.81   64837.89  -2.738  0.00622 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 24250 on 3921 degrees of freedom
Multiple R-squared:  0.297,     Adjusted R-squared:  0.2952
F-statistic: 165.6 on 10 and 3921 DF,  p-value: < 2.2e-16
```

Interpretation: We see that after the selection-bias correction process, there is no longer a significant relationship between having children and a respondent's income. The negative effect of being female is also marginally smaller.

The difference exists because we incorporate the probability of selection into the workforce as a function of a respondent's characteristics. We then incorporate those estimates into our regression. The selection model results may differ because individuals with children are much more likely to work than their childless counterparts, and after accounting for this trend, we see that having children is no longer associated with an increase in one's income.

Exercise 3

1) Histogram of income:



All incomes greater than $100,000 are censored and set equal to $100,000. Another note is that there are far more individuals with income between $0-$50,000 compared to individuals between $50,000-$99,999; this may suggest income has a bimodal distribution based on education or gender.

2) To deal with the censoring problem, I will again use the Heckman Two-step Estimator.

1st step: Pr(Income < $100,000) = educ_years + work_exp + age + marstat + kids
2nd step: Income = educ_years + work_exp + age + sex + marstat + invMillsRatio

Results:

```
lm(formula = income ~ educ_years + work_exp + age + sex + marstat +
    imr, data = nlsy)

Residuals:
    Min      1Q  Median      3Q     Max
-106856  -16124   -1982   15739   85597

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)       80173.11   11197.89   7.160 9.61e-13 ***
educ_years          548.91     167.71   3.273  0.00107 **
work_exp            383.85      94.59   4.058 5.04e-05 ***
age               -1350.54     290.69  -4.646 3.50e-06 ***
sexFemale        -18575.40     772.58 -24.043  < 2e-16 ***
marstatMarried    -1905.42    1355.38  -1.406  0.15986
marstatSeparated    877.17    2859.70   0.307  0.75906
marstatDivorced    3164.84    1304.40   2.426  0.01530 *
marstatWidowed     9180.16    6470.01   1.419  0.15601
imr               86022.65    7297.85  11.787  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 23870 on 3922 degrees of freedom
Multiple R-squared:  0.319,     Adjusted R-squared:  0.3175
F-statistic: 204.2 on 9 and 3922 DF,  p-value: < 2.2e-16
```

Interpretation:

In this model, we see that the intercept value is much greater, which explains why age now has a negative relationship with income. The negative relationship associated with age is counter-balanced by the strong positive effect of additional education years and years of work experience. However, the coefficients associated with education/work experience are much smaller than in prior specifications, and more in line with existing literature. We still see that being female has a strong negative effect on wages, though there are a number of factors that we do not include in the model that may explain some of this difference (i.e. job type, job performance, geographic location, etc.).

We also can see that most of the significant effects associated with marital status have vanished after the adjusting for censoring, and the sign of the coefficient associated with marriage has reversed. This suggests that the unmarried individuals make up a larger proportion of censored observations, and after incorporating a measure of their true distribution the relationship between marriage and income falls away.

Exercise 4

1) Ability bias

Individuals with higher innate ability are also more likely to pursue more education because there are positive returns to education. Nonetheless, even if high-ability individuals had average education, they would still have higher than average earnings (assuming employers are able to differentiate between low/high ability workers). When trying to understand the determinants of wages we are often unable to incorporate implicit measures of ability, and the effect of innate ability is subsumed into the effect of education, leading to a biased estimate of the effect of schooling.

2) Model Specification:

Income = work_exp + educ + age + marstat + sex + ethnicity

Results:

	Within Est.	Between Est.	First-Difference Est.
Work Experience	621.91***	1477.35***	547.58***
	(56.09)	(83.50)	(56.28)
Education	3485.48***	9080.57***	908.76
	(474.81)	(225.18)	(526.78)
Age	2638.93***	1144.61***	1657.56***
	(51.08)	(143.13)	(145.55)
Marital Status	668.91*	2098.00***	536.05
	(324.52)	(390.22)	(332.13)
Female		−15255.53***	
		(626.23)	
Ethnicity (Black as reference level)			
Hispanic		5493.97***	
		(907.42)	
Mixed Race (non-Hispanic)		6405.71*	
		(3264.03)	
Non-Black/Non-Hispanic		6692.08***	
		(771.40)	
(Intercept)		−11476.16*	2205.17***
		(4667.77)	(340.10)
R^2	0.19	0.29	0.01
Adj. R^2	−0.07	0.29	0.01
Num. obs.	31035	7490	23545

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

Interpretation:

The within and first-difference estimator eliminate ability bias by allowing income to vary based on unobserved omitted variables. The within estimator and first-difference estimator are similar for the most part, however the within estimator has a much larger coefficient on age and education. The first-difference estimator is optimal here because it allows for consistent estimates even when our omitted variables are correlated with our other independent variables, as we might assume in this case.

We can see that only the Between estimator gives us an estimate of the effect of sex and ethnicity on income. This is because the within and first-difference estimators use a fixed-effect style estimation which control for time-invariant individual characteristics such as sex and ethnicity. On the other hand, the between estimator simply regresses the average independent variables on the average dependent variable, thus while the sex/gender are unchanged we can see variation that arises across the distribution.