Midterm SOCI709

Kate Brandt

March 12, 2019

** Exam started : 12:20 PM, March 12 **

** Exam finished: 3:40 PM, March 12 **

Part A

Q1. The coefficient on gender changes across the different models in Table 1.

(a) In model 1, there are no interaction terms included in the model, so the gender coefficient is the difference in happiness between **gender A and gender B individuals generally in the entire dataset**.

(b) In model 2, the excluded group is low education, childless, gender A individuals. This model includes an interaction term between parenthood and gender B though, so the gender coefficient can **only** refer to the difference in happiness between the **childless, gender A individuals and childless, gender B individuals (regardless of education status)**. If comparing effects of gender on happiness different between genders and parent statuses, the gender and parenthood interaction term must be included.

(c) In model 3, the excluded group is low education, childless, gender A individuals. This model includes an interaction term between parenthood and gender B and education and gender B though, so the gender coefficient can **only** refer to the difference in happiness between **low education, childless, gender A individuals and low education, childless, gender B individuals**. If comparing effects of gender on happiness different between genders, parent statuses, and education levels, the gender and parenthood interaction term and gender and education level interaction term must be included.

Q2. Why does the coefficient on gender increase going from model 3 to model 4? This suggests that the coefficient on gender in model 3 was being under estimated. What was causing it to be underestimated?

Models 3 and 4 are identical with the exception that model 4 includes hours of house work per week in the model. By including this in the model, difference betweeen gender happiness incereases by a great, which points to a possible explanation. First, this means that between genders A and B, gender B exhibits greater happiness overall, when accounting for the other factors in the model. The housework variable says there is a negative relationship between hours of housework per week and happiness; the underestimation of the gender variable in model 3 may be due to an observed tendency of gender B to complete and high amounts of housework. By including housework in the model, we can estimate the separate effects of that variable and gender, and understand that this association between gender and housework may have dampened the size of gender B's coefficient.

Q3. What does model 5 show us about whether the effect of housework differs by gender? Given this result, what is the only really plausible explanation for the comparison of results between models 3 and 4?

Model 5 is identical to model 4, except it includes an interaction between gender and housework. This interaction shows that there is an extremely small (essentially unnoticeable) interaction of gender B and housework hours, where the negative effect of hours of housework on gender B is dampened very slightly. Given this slight difference, we can conclude that the effect of housework operates to the same effect for both genders A and B. This clears up any suspicion of effects that model 4 may have been missing and

confirms that there is a genuinely greater difference in happiness among genders A and B, and there may be an association between high hours of housework and gender B.

Q4. What is the predicted happiness of the following person using model 5: Gender B, Parent (yes), high education, 20 hours of housework per week.

Excluded group: Gender A, childless, low education

Coefficients to include:

constant + Gender B + Parent + High education + GenderBxEdu + GenderBxParent + Housework(20) + GenderBxHousework

15.17 + 5.455+ 5.885 + 4.356 + -2.588 + 3.924 + -0.308(20) + 0.0002(20) = **26.046**

Q5. Using model 5, if a person of gender B increases their education from low to high, how will their predicted happiness change?

Keeping all else constant, the change in education for a person of Gender B can be calcuated as:

[high education] + [GenderBxEducation] =

4.356 + -2.588 = **Happiness predicted to increase by 1.768**

Q6. What is the gender gap among high education parents who do no housework in Model 5?

Comparison Table (to narrow down needed variables)

Gender A	Gender B
~~constant~~	~~constant~~
_____	Gender B
~~Parent~~	~~Parent~~
~~HighEd~~	~~HighEd~~
_____	GenderBxParent
_____	GenderBxHighEd

Formula: GenderB + GenderBxParent + GenderBxHighEd

= 5.455 + 3.924 + -2.588 = **6.791**

Part B

Q7. In Model 1, what is the 90% confidence interval for the effect of ambition0 on income?

```
# Enter in data
ambition0 <- -0.4385718
se <- 0.21142
error <- qt(0.95, df = 2996)*se

# Create upper and lower bounds of interval
lower <- ambition0 - error
upper <- ambition0 + error
```

```
lower
```

```
## [1] -0.7864343
```

```
upper
```

```
## [1] -0.09070928
```

90% CI for ambition0 = (-0.7864343, -0.09070928)

Q8. Looking at the results for Model 1, do you think you have estimated the true causal effect of education on income? Why or why not? If you think there is bias in the effect, what direction do you think the bias would be?

While the OLS model is a good start for exploring the general direction of effects of education on an individual's income, modeling the pooled data cannot properly capture the effect of education on income because it is not able to control for individual effects or changeds over time with changes in education. This model has likely overestimated the effects of education on income because it is unable to capture control for individual level effects that may influence income (for example, personality traits that cause someone to be promoted faster to higher paying jobs).

Additionally, since this is panel data, the pooled OLS is not able to tell us the effect size of an increase in education over time, as a random effects or fixed effects model would. RElated, the random effect u, while not correlated with any of our variables, can cause problems with heteroskedasticity that the random effects model (model 3) will be able to address.

Finally, the inclusion of the "voucher" intervention group in this model is likely creating some sort of bias in these estimates because of the extremely increased prevalence of individuals from low-income families obtaining higher education. This has also likely overestimated the effect of education on income because it changed the relationship between income and educational attainment.

Q9. Does Model 2 estimate the true causal effect of education on income? Why or why not?

Model 2 does not estimate the true causal effect of education on income. Though the instrumental variable of the voucher intervention may have been able to provide valuable insights about the effect of education on income, the design of the intervention was not well suited for this study. Because the voucher was only distributed to part of one sub-group of the study population, there is no way to tell how the intervention may have operated for the other towns with a different socioeconomic status profile. Indeed, we see that individuals in the Big Gap Creek community obtained higher levels of education at nearly twice the rate of those in the other study communities. The effects of the intervetion would have to be studied within the Big Gap Creek community to truly learn effects of this; even then, the intervention would on be able to tell us about how it operates for a low-income community.

Finally, we run into the same problems with pooled data in this model as we do in Model 1.

Q10. Does Model 3 estimate the true effect of education on income? Why or why not? If you think there is bias in the effect, what direction do you think the bias would be?

A random effects model assumes that there is no correlation of any variables with the error term, u. As we know from the set up of this scenario, this is the case. A random effects model is similar to a pooled OLS model, with the advantage that it can properly estimate the time-invariant variables. This is likely why, when comparing model 1 and model 3, we can see that most coefficient values are about the same with the exception of the ambition0 variable; since this is time invariant, the random effects model may have been able to estimate it's effect more properly.

However, because of the similarities of the random effects and pooled OLS models, I suspect that this model is still not the best estimate of education's effect on income. There are still issues with the data regarding the voucher intervention. There still may be an overestimate of education's effect on income because of this. This model is the same as the pooled OLS except that it solves issues of heteroskedasticity in model 1 and improves the estimate of ambition0.

Q11.

(a). Why do ambition0 and parent_inc drop out of the model?

In fixed effects models, time invariant variables drop out because their effects are impossible to estimate since they are constant over time. Since these are the time invariant variables of this study, they are not able to be estimated using fixed effects.

(b) Does Model 4 estimate the true effect of education on income? Why or why not?

Model 4 is the best model to estimate the effect of education on income. This is because it is able to control for individual, unchanging effects and estimate how income changes with education over time. The coefficient estimate is lower than in the other models, proving that the other models were overestimating the effect. This model also solves the problem of the non-random voucher distribution by only estimating effect of educational attainment, without inclusion of parent's income, so there is less bias between the effects of parent's income and individual education.

Of course, there are disadvantages to the fixed effects model because of the innability to estimate the effect of parent's income on educational attainment. However, the study set out to estimate the effect of education on income, and if this is the main relationship being studied, this is a good way to estimate the effect by removing bias of unobserved effects/error.