Problem Set 9

Daniel Shapiro

11/9/2022

Question 1 Background:

On average, women earn less than men in almost every country – this is referred to as the 'gender gap.' We know less about the mechanism behind this gap. Employers may discriminate against women by not hiring them for high-paying jobs or by paying them less than men for comparable jobs. Women may have different levels of education. Women may also choose occupations that generally have lower levels of monetary compensation – for example, due to social expectations of performing "feminine" jobs, or because these careers offer greater work-life balance.

This exercise employs a sample from a 1994 survey of South African workers. The data are contained in a file called safrica.dta. Variable definitions are as follows:

- wage: average hourly earnings (1994 RZA; 1 US dollar = 3.5 RZA in 1994).
- age: age in years.
- educ: number of years of education (note: 10=completed secondary school, 13=completed university degree)
- exper: years of work experience (age-education-7)
- union: = 1 if working a union job, = 0 otherwise.
- female: = 1 if female, = 0 otherwise.
- married: = 1 if married with spouse present, = 0 otherwise.
- urban: = 1 if lives in an urban area, = 0 otherwise.

1a) Run a regression of wage on female. Interpret the coefficient on the dummy variable. Do you believe this regression meets the assumptions of the OLS model? What kind of inferences would you be able to make given your assessment of the validity of the assumptions?

```
data <- read.dta("safrica.dta")
```

```
model <- lm(wage ~ female, data = data)
summary(model)
```

```
##
## Call:
## lm(formula = wage ~ female, data = data)
##
```

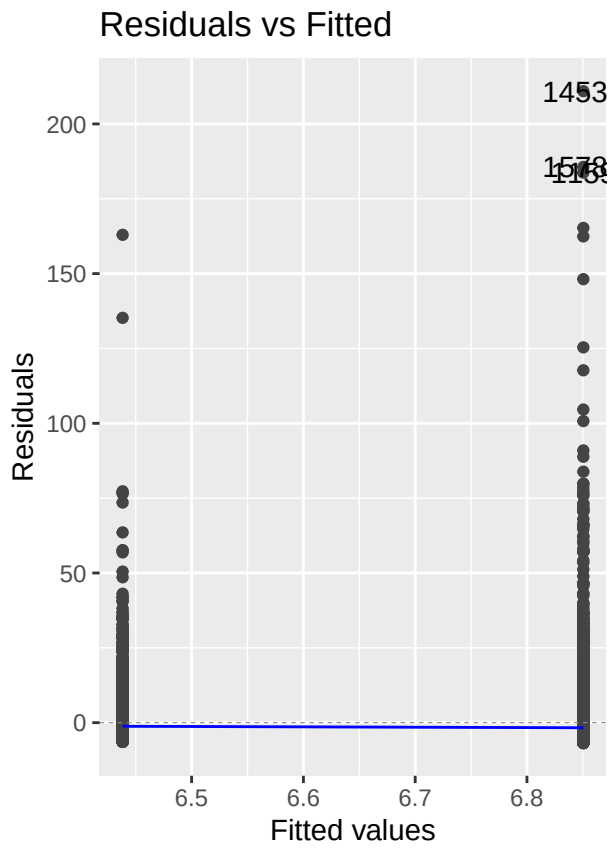
```
## Residuals:
##     Min      1Q  Median      3Q     Max
##  -6.838  -4.394  -1.901   1.812 210.950
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  6.85046    0.06864  99.796  < 2e-16 ***
## female      -0.41216    0.11432  -3.605 0.000313 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.753 on 19946 degrees of freedom
## Multiple R-squared:  0.0006512,  Adjusted R-squared:  0.0006011
## F-statistic:    13 on 1 and 19946 DF,  p-value: 0.0003126
```

The coefficient basically says that for each increase of "1" in the independent variable ("female"), average hourly earnings decrease by about 41 cents (ZAR). This is the technical definition. But the fact that "female" is a binary variable basically means that if a person is female, their average hourly earnings can be expected to be lower.

I'm going to run some tests to see if the assumptions are met:

1) Linearity of parameters

```
autoplot(model, 1)
```



There isn't linearity, because it's a binary variable as the only explanatory one.

2) Random Sampling

We can't really tell for sure, but the data looks relatively random, at least.

3) Variation in X

There is only zero and one, so there really isn't too much variation. That being said, at least there doesn't seem to be a clear pattern of only ones or only zeros.

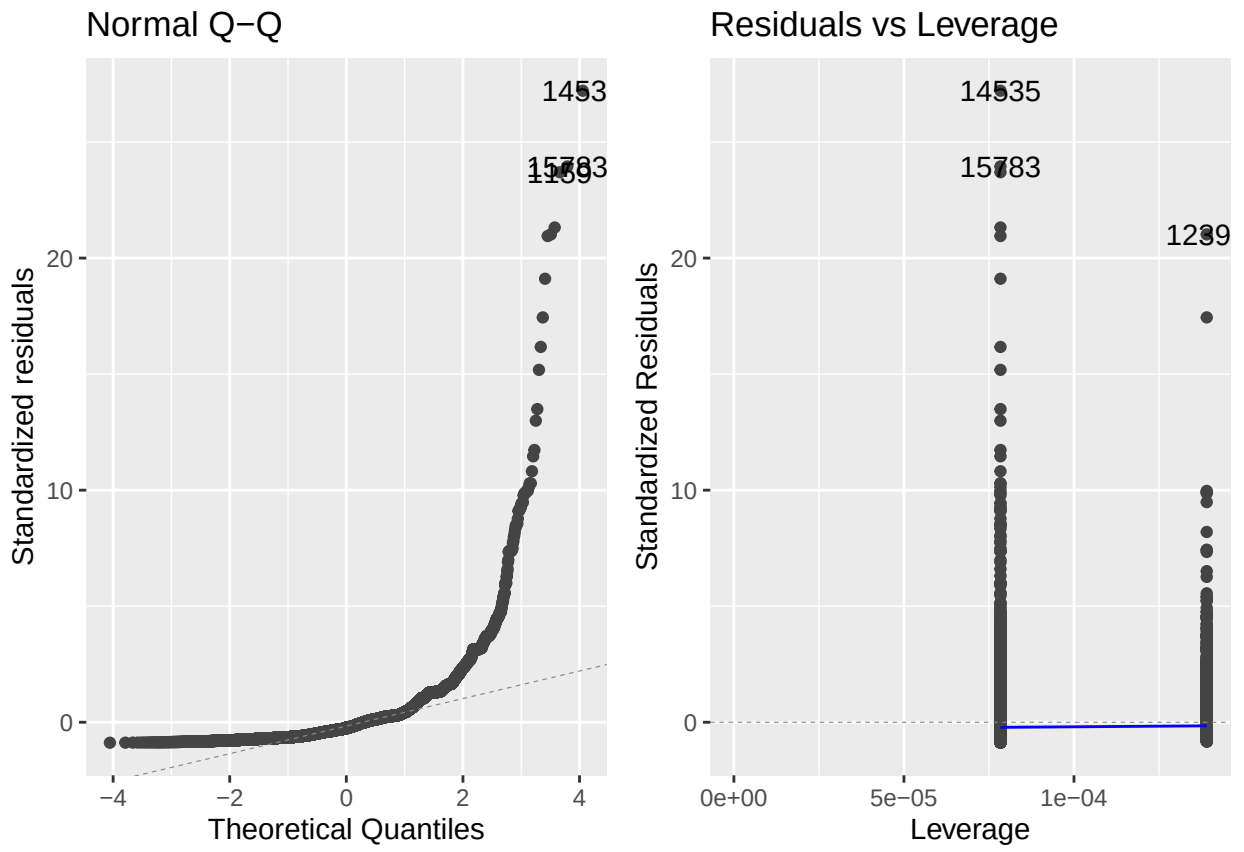
4) Zero conditional mean

I think that this is a key issue; the question itself even admits that the endogeneity problem is quite relevant. I think that we can say that there is likely some relationship between gender and wage, but I think that the direction of the causal arrow is difficult to define. Plus, there are likely a ton of other variables that can explain the outcome.

5) Homoskedasticity, normally distributed errors

No, the error term is really not normally distributed. We can see how far away some of the points are from the dashed line. Also, residuals will not be normally distributed because the wage values are binary. See below:

```
autoplot(model, c(2, 5))
```



The model works in certain ways, but I would find it difficult to make any clear predictions based on this. There are so many other variables that are left out, and the causal arrows are not clear. Much more work is needed.

1b) More education is associated with increased earnings. Run a regression of wage on female and years of education. Interpret the coefficients, explaining how they have changed from a) and why. Does controlling for education explain the gender gap? Why or why not? Does including education change whether the model meets the OLS assumptions? Why or why not?

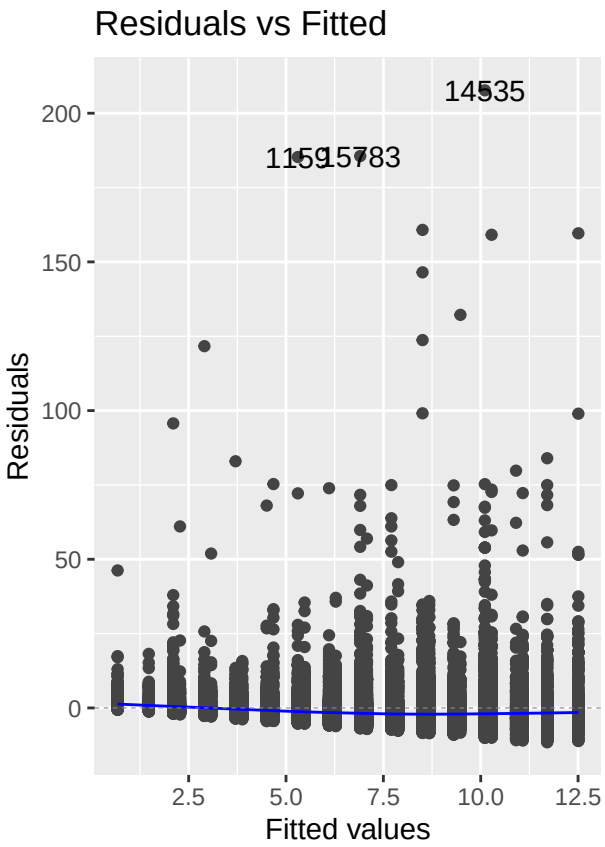
```
model2 <- lm(wage ~ female + educ, data = data)
summary(model2)
```

```
##
## Call:
## lm(formula = wage ~ female + educ, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -11.551  -3.349  -1.056   1.427 207.693
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.10404    0.10622   19.81   <2e-16 ***
## female      -1.42746    0.10784  -13.24   <2e-16 ***
## educ         0.80026    0.01431   55.91   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.209 on 19945 degrees of freedom
## Multiple R-squared:  0.136,  Adjusted R-squared:  0.1359
## F-statistic:  1570 on 2 and 19945 DF,  p-value: < 2.2e-16
```

The coefficient for female is now quite a bit larger. So it means that if you control for education, then the gender pay gap seems to widen.

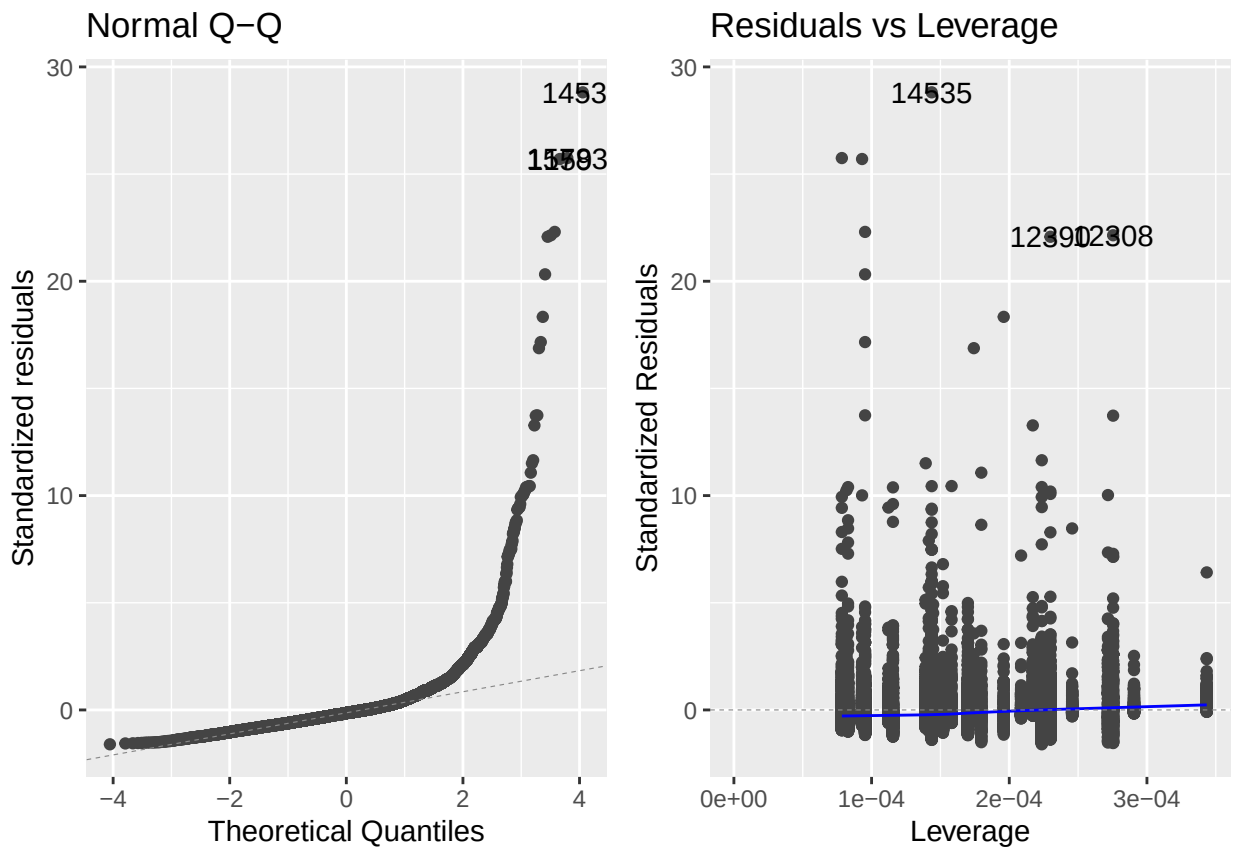
At least to some degree, the OLS assumptions appear to be better in this model. There are just more points, more possible x-values, etc. But the fundamental issues of them (linearity, normally distributed errors, etc.) So let's do some more autoplots.

```
autoplot(model2, 1)
```



The following two autoplots look rather similar to the autoplots for 1a, just with a few more points for residuals vs. leverage. There really isn't too much of a difference between the two questions in this sense.

```
autoplot(model2, c(2, 5))
```



The main area in which I think that we can actually find some real differences is in the Zero Conditional Mean assumptions. We're adding another variable in, so the ommitted variable bias naturally goes down.

1c) Run a regression of wage on female, educ, age, union, marital status, and urban variables in the regression. Does the gender gap persist? Based on this, can you conclude that the mechanism behind the gender gap is discrimination?

```
model3 <- lm(wage ~ female + educ + age + union + married + urban, data = data)

summary(model3)
```

```
##
## Call:
## lm(formula = wage ~ female + educ + age + union + married + urban,
##     data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -12.526  -3.170  -0.979   1.333 206.850
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.167196   0.221640  -5.266 1.41e-07 ***
## female      -1.291026   0.107195 -12.044  < 2e-16 ***
```

```
## educ         0.714917   0.016151  44.264  < 2e-16 ***
## age          0.060526   0.005473  11.058  < 2e-16 ***
## union        0.716726   0.114559   6.256 4.02e-10 ***
## married      1.004422   0.111505   9.008  < 2e-16 ***
## urban        1.542556   0.114078  13.522  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.087 on 19941 degrees of freedom
## Multiple R-squared:  0.1652, Adjusted R-squared:  0.165
## F-statistic: 657.7 on 6 and 19941 DF,  p-value: < 2.2e-16
```

The gender gap definitely does persist even when you add in all of these other factors. But I'm still not sure if we can claim causality here. There are a ton of other factors that go into wages, and it's not clear that obvious discrimination is behind the gap. Also, I don't trust a lot of the variables here. Since a lot of them are binary, I wonder about the degree to which controlling for them really just means choosing one or the other. So here for example, if I'm interpreting it correctly, controlling for union, married, and urban just could mean that we're looking at non-union, unmarried, rural woman (where all of them equal 0)? This is an honest question that I'm just not entirely sure about.

Regardless, I don't think it's enough evidence to firmly say that there is active discrimination. And I think this is a fundamental issue of quantitative analysis. How much can we really figure out if "discrimination" is the answer? There are so many factors that could factor in here – and a lot that we don't even begin to know. Some contextual evidence is necessary as well in order to really make judgments such as these.

1d) Run a regression of wage on age. Interpret the coefficient. Interpret the statistical and practical significance, too.

```
model4 <- lm(wage ~ age, data = data)
summary(model4)
```

```
##
## Call:
## lm(formula = wage ~ age, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
##  -7.584  -4.293  -1.900   1.684 210.907
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 5.194622   0.195380  26.587  < 2e-16 ***
## age         0.042464   0.005283   8.037 9.69e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.743 on 19946 degrees of freedom
## Multiple R-squared:  0.003228,   Adjusted R-squared:  0.003178
## F-statistic:  64.6 on 1 and 19946 DF,  p-value: 9.693e-16
```

Essentially, this coefficient says that for each additional year of age, on average, expected wages go up by about 4 cents per hour. It appears fairly statistically significant (three stars). In terms of practicality, I'd

have to see a graph or something to be able to tell if it matters or not, because it could be that the graph looks similar to the graph we had for wages from a previous problem set, which resembled more of an arc than a straight line.

1e) Run a regression of wage on age and exper. Interpret the coefficient estimate on age, explaining how it has changed from d) and why. Does this change strike you as a problem? If so, do you have any suggestion to cure the problem?

```
model5 <- lm(wage ~ age + exper, data = data)
summary(model5)
```

```
##
## Call:
## lm(formula = wage ~ age + exper, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -12.578  -3.327  -1.059   1.374 207.703
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -7.6889     0.2885  -26.66   <2e-16 ***
## age           0.9162     0.0160   57.27   <2e-16 ***
## exper        -0.8201     0.0143  -57.37   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.174 on 19945 degrees of freedom
## Multiple R-squared:  0.1444, Adjusted R-squared:  0.1443
## F-statistic:  1683 on 2 and 19945 DF,  p-value: < 2.2e-16
```

The age coefficient has expanded drastically. But then the experience coefficient is negative. These results don't really make that much sense. My suspicion is that because age and experience are so intimately linked, there's a lot of colinearity here. So the model gets thrown off here. There are a couple things we could try here. One is putting in an interaction variable (age*exper), which would at least look at them together. The other would be to just take out age altogether and just use the experience variable, because experience presumably follows age pretty closely to some degree but just gives us more information on wages than age does.

Question 2 Background:

An article by Ebonya Washington in the American Economic Review argues that having a daughter (as opposed to a son) might affect how politicians vote on women's issues. In particular, she argues that having a daughter causes Congressional representatives to vote more liberally on women's issues. We will use Washington's data (available on the website as `washington.csv`*). The key variables are:*

- ngirls: Number of daughters
- totchi: Total number of children
- party: Indicator for democrats (1), republicans (2), or other (3)
- aauw: Legislator's voting score as assigned by the American Association of University Women - higher scores indicate being more liberal on women's issues.

2a) Comparing only Republicans and Democrats (dropping the 'Other' category), examine whether legislators of different parties have the same number of children. Do the results surprise you? Why or why not?

```
washington <- read.csv("washington.csv")
```

Here, I'm going to just look at the mean value of children by party; I think that that's an easy initial way to look at it.

```
dems <- washington %>%
  filter(party == 1)

mean(dems$totchi)
```

```
## [1] 2.223301
```

```
reps <- washington %>%
  filter(party == 2)

mean(reps$totchi, na.rm = TRUE)
```

```
## [1] 2.731278
```

Obviously this isn't perfect, but it at least shows a solid pattern. I figured that I could do this instead of a regression because there aren't going to be any *huge* outliers; the data is all between 0 and 10 (apart from one NA, which I dropped) for the "totchi" variable. Anyway, we see that Republicans generally tend to have more kids than Democrats. I think that this is a pretty well-established pattern, namely, that Republicans tend to (on average) have more children than Democrats; not just politicians but ordinary people too. So no, I'm not surprised by this.

2b) Regress a representative's aauw score on the number of female children and report your results. What is the relationship between the number of female children and the AAUW score? Do you think this relationship estimates the causal effect of having a female child?

```
modelb <- lm(aauw ~ ngirls, data = washington)
summary(modelb)
```

```
##
## Call:
## lm(formula = aauw ~ ngirls, data = washington)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -50.964 -45.396  -6.396  49.036  60.171
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   50.964      3.043  16.748   <2e-16 ***
## ngirls        -2.784      1.791  -1.554    0.121
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 41.94 on 432 degrees of freedom
##   (1 observation deleted due to missingness)
## Multiple R-squared:  0.005562,   Adjusted R-squared:  0.00326
## F-statistic: 2.416 on 1 and 432 DF,  p-value: 0.1208
```

Well, the coefficient is generally negative – on average, it indicates that for each extra girl a representative has, their AAUW score is about 2.784 lower. But I have a couple of issues with this. For one, the coefficient is not significant, so that's an immediate red flag. Secondly, however, the AAUW score is on a 1-100 scale with huge gaps between values. It might be easier to look at this relationship on a log scale so that we can diminish the effects of wide swings between 0 and 100. Third, I know that we shouldn't look at this, but the R-squared value is really minimal. There are a ton of other variables that should be taken into account. If this model does measure the effect of having girls on propensity to vote more liberally on womens' issues, it does so in a very weak and inconclusive way.

2c) Now regress a representative's aauw score on the number of female children and total number of children and report your results. What is the relationship between the number of female children and the AAUW score after controlling for the total number of children? Do you think this relationship accurately estimates the causal effect of having a female child? If you believe a causal claim is possible, discuss what assumptions you have to make. If you believe it is not possible, explain why.

```
modelc <- lm(aauw ~ ngirls + totchi, data = washington)
summary(modelc)
```

```
##
## Call:
## lm(formula = aauw ~ ngirls + totchi, data = washington)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -59.982 -37.350  -2.877  41.978  87.970
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   59.982      3.573   16.79  < 2e-16 ***
## ngirls         5.776      2.567    2.25    0.025 *
## totchi        -7.992      1.753   -4.56 6.68e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 41.01 on 431 degrees of freedom
##   (1 observation deleted due to missingness)
## Multiple R-squared:  0.05132,    Adjusted R-squared:  0.04692
## F-statistic: 11.66 on 2 and 431 DF,  p-value: 1.172e-05
```

I'm not sure as how a causal claim can be made here. The coefficient for "ngirls" is technically significant at the 0.05 level, but I think that there's a fair amount of colinearity between totchi and ngirls, since each girl in ngirls also counts as a contribution to totchi. They're directly connected with one another. I think

also the question of endogeneity comes into play here as well – maybe, for example, more liberal, pro-women districts tend to elect representatives with more daughters. There are undoubtedly some other issues here; these are just a few I can think of off the top of my head.

2d) Now add party fixed effects to your regression in part c). How does this change the results? Why? Is it a good idea to add this control?

```
modeld <- lm(aauw ~ ngirls + totchi + as.factor(party), data = washington)
summary(modeld)
```

```
##
## Call:
## lm(formula = aauw ~ ngirls + totchi + as.factor(party), data = washington)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -76.284 -13.168  -0.084  11.625  89.969
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)         89.871      1.862  48.257  < 2e-16 ***
## ngirls               3.138      1.218   2.575   0.0104 *
## totchi              -3.345      0.840  -3.982 8.03e-05 ***
## as.factor(party)2  -72.943      1.896 -38.479  < 2e-16 ***
## as.factor(party)3   17.233     19.508   0.883   0.3775
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 19.43 on 429 degrees of freedom
##   (1 observation deleted due to missingness)
## Multiple R-squared:  0.788,  Adjusted R-squared:  0.786
## F-statistic: 398.5 on 4 and 429 DF,  p-value: < 2.2e-16
```

Adding the party fixed effects REALLY helps the explanatory power of this regression. Even just looking at the R-squared value, we can see that it jumps WAY up from where it was, meaning that adding party really impacts the explanatory power. It does look like the ngirls variable does remain as a positive variable in supporting women's issues, and controlling for parties really helps. There is also less omitted variable bias here.