ENWONGO EKAH

ENWONGO EKAH

SCHOOL OF PUBLIC POLICY

PLCY 610 – QUANTITATIVE ASPECT OF PUBLIC POLICY

FINAL EXAMS

DR. ANTHONY SANFORD

MAY 19TH, 2021

PROBLEM 1

Part A

Using the output from the regression in R, create a table to show the results.

Solution:

Table 1: REGRESSION TABLE EXPLAINING THE USE OF ELECTRONICS IN CLASSROOMS ON TEST SCORES

		DEPENDENT VARIABLE
		Test Scores of Students
	Hours of Electronics	-0.02857
		(0.07780)
	Constant	20.89035
		(1.544)
	Observations	150
	Adjusted R^2	-0.005844
	Note	*p<0.005, **p<0.001, *** p<0.001

I got the regression table above from running the codes that were provided in the prompt with my date of birth on R and I will use the data above to answer the questions below.

ENWONGO EKAH

Part B

Explain the meaning (in your own words) of statistical and substantive significance. Comment on both the statistical and substantive significance of your results (from Part A).

Solution:

Statistical significance means with a large enough sample size even truly microscopic differences can be statistically significant. If we have a statistically significant, we are determining if there is a relationship between our two variables. Substantive significant; we are referring to not just if there is a relationship but how large is the size and how applicable it is to the context. Whether or not it is big enough to be valid in a real-world context. On average, a 20.8 point in test scores is associated with a -0.016 hours of electronic use in a classroom. The slope estimate suggests that on average 1.5 Sd in test scores is associated with 0.07 Sd hours in classroom. There is a relationship between test scores but it is not statistically significant and substantive enough.

Part C.

Is the model a good fit? How can you tell?

Solution:

The model is not a good fit. The X variable is not doing a good job of predicting the Y variable. It is not explaining the variation and prediction in Y. There is no evidence that the model explains a very large percentage of the total percentage of the total variability in Y.

Part D.

Interpret any other information that you think is relevant in the table and have not discussed so far.

Solution:

The R^2 indicates that the hours of electronic use do not explain very much the variation in test scores.

Part E.

Based on the results you obtained, what is your conclusion? Should you or should you not ban electronics from the classroom? In answering this question, make sure that you use proper

statistical terminology and as much information from the table that you created as necessary to completely answer the question.

Solution:

The F statistics 0.1343 has a p value of 0.7145 larger than our alpha at a 0.05 level. There is insufficient evidence to conclude that a non-zero correlation exists. This means that we fail to reject the null that the regression model does not explain that electronics are distracting and should be banned from the classroom. It is not statistically significant, and it is not substantive enough so electronics should not be banned from the classroom. However, it does not mean anything about our interpretation of individual coefficient.

Part F

Why might the model specification not be accurate? Is there a variable(s) that we could add to the regression model that might help us obtain a better or more reliable conclusion?

Solution:

At least one of them is equal to zero and it means it shouldn't be in the model. Just considering the information provided, we don't have enough evidence to conclude that this is unrelated to other omitted variable. There are other variables that could be missing in the regression model that could help predict the Y more. It could be the number of hours a student studies after the class, it could be the fact that the student does not even study or there are other things that could be distracting in the class apart from the electronics. I don't think electronics should be banned from the classroom.

Part G

Is a linear specification appropriate here? To answer this question, I do not need you to run any tests. All I want you to do is tell me what the assumptions for a linear model are and how (answer this using your words and your intuition – no need to run any code) you would go about testing these assumptions.

Solution:

We always assume linear relationships. The baseline assumptions for a regression are Random sampling/ independence. If we don't have a representative of a random sampling from a population, we cannot make inferences, we will have a bias. Zero conditional mean; there is no bias, these differences shouldn't be correlated in the model X and Y. We assume that all our relationships are not linear and if there are not, if not we have a problem as all of our calculations assume a linear relationship. We can check this out visually by plotting a data and checking the residuals to see if we have a pattern. We can incorporate non- linear relationships by first transforming the variables. Normality: for small samples we have to rely on the assumptions that the errors are normally distributed. In large samples, the CLT will kick in. In order to diagnose

non-normality, we can plot a histogram of the residuals and can drop/ change observations in order to make a more normal fit or use a transformation. The assumption of homoskedasticity tells us that for each value of X, the error terms have the same variance. In order to test for this, we will look at the residuals, if there is a cone like shape or triangle, then there is heteroskedasticity.

ENWONGO EKAH

PROBLEM 2

Part A

Create a two-way table recording the "successes" (political), "failures" (non-political), and totals for each comedian. Place the comedians in the columns and political/not political in the rows.

Solution:

Table 2: **Two Way Table Recording the Political and Non-Political Nature of Comedians**

	Leno	Letterman	Stewart	Total
Political	315	136	83	534
Non- Political	998	512	169	1679
Total	1313	648	252	2,213

Part B

State the null and alternative hypothesis and create a table with the expected counts assuming the null hypothesis is true.

Solution

H_0: μ = Comedians are more political than others

H_a: $\mu \neq$ Comedians are equally political.

Table 3: Expected Counts for Null Hypothesis

	Leno	Letterman	Stewart	Total
Political	316.83	156.36	60.81	534
Non- Political	996.17	491.64	191.19	1679
Total	1313	648	252	2,213

Part C

Conduct and interpret a χ2test to assess the null hypothesis. Report theχ2test statistic and p-value and interpret your findings.

Solution:

Chi Square

x^2 Obs $\sum \frac{(observed\ count - expected\ count)^2}{expected\ count}$

ENWONGO EKAH

$$\frac{(315-316.83)^2}{316.83}+\frac{(136-156.36)^2}{156.36}+\frac{(83-60.81)^2}{60.81}+\frac{(998-996.17)^2}{996.17}+\frac{(512-491.64)^2}{491.64}+\frac{(169-191.19)^2}{191.19}$$

$x^2 = 14.1809$

I conducted a chi test for proportion, and I also ran a chi test in R, and these are my findings.

This test statistics 14.1809 is larger than the critical value (5.99) for a chi test distribution of 2 degrees of freedom. As a result, we have a small p value at 0.00813. With this, we have enough evidence to reject the null hypothesis that some of the comedians are more political than others.

Part D

In a table or a bar plot, record the conditional probabilities for each comedian. What does this information tell you about the political nature of each comedian? How does it supplement the findings from Part C?

Solution:

Table 4: Conditional Probabilities for each Comedian

	Leno	Letterman	Stewart
Political	24%	21%	33%
Non- Political	76%	79%	67%
Total	100%	100%	100%

The information tells us the jokes of comedians make given that they are more political or not. The three comedians are more likely to tell non-political jokes. The chi test helps us test proportions, but conditional probabilities help us determine if the two variables are dependent or independent. If there is a conditional relationship between comedians and their political nature.

ENWONGO EKAH

PROBLEM 3

Part A

Regress Democracy scores (FHREVERS) on unlogged GDP/capita (GDP90) and report the results in a table recording the coefficients, standard errors, and p-values. Interpret the association between GDP/capita and average Freedom House score by discussing the statistical and practical/substantive significance.

Solution:

Table 5: Regression of Democracy Scores on Unlogged GDP/Capita

	Dependent Variable		
	Democracy Score		
	Coefficient	std error	P values
GDP90	0.048360	0.005293	4.5e-16***
INTERCEPT	2.679011	0.206059	<2e-16***
N	150		
R^2	0.3563		
Note	p<0.5* p< 0.01** p<0.001***		

I ran a regression of democracy scores on unlogged GDP on R and the data on the table above were part of the findings.

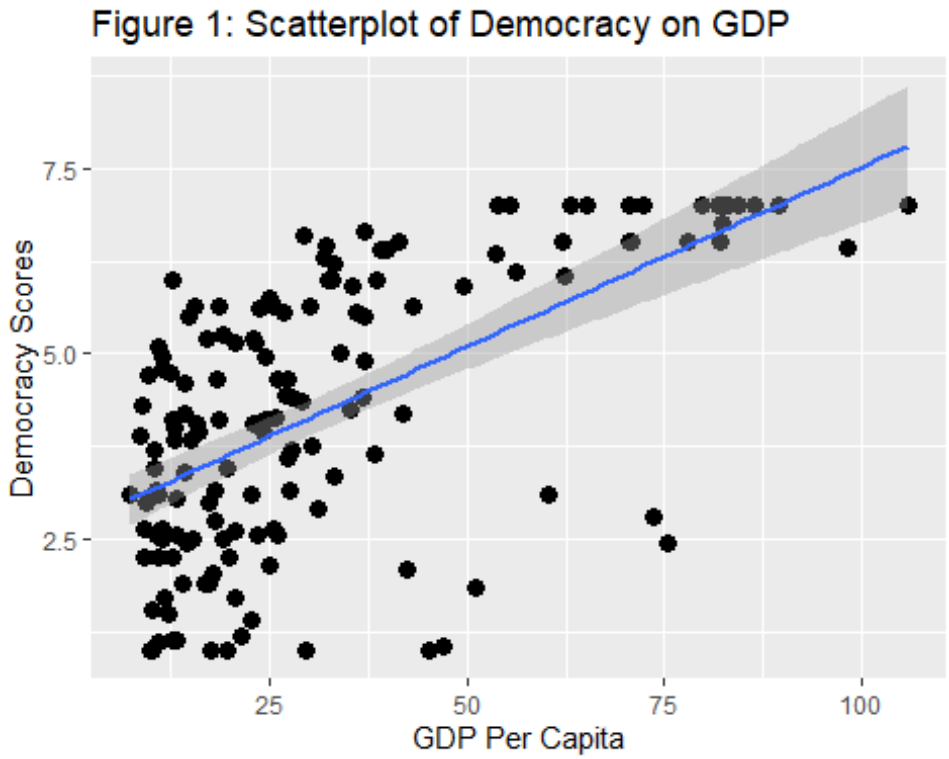
The coefficient suggests that on average 1 unit in increase in GDP/ capita is associated with a 0.048360 change in democracy score. The P value is statistically significant at the 0.5 level and seems potentially substantively significant.

ENWONGO EKAH

Part B

Generate a scatter plot of the two variables with the regression line superimposed. Does the regression model look like a good fit? For example, does it appear as though the association is linear? Is the spread of the observed values of FHREVERS generally equal around the regression line across the values of GDP/capita?

Solution:



I generated a scatterplot on R shown in Figure 1 above, to show the relationships between two variable and hence if the model is a good fit. Looking at the scatterplot above it doesn't seem to be a good fit. The GDP per capita variable is skewed to the left. The relationship between the independent and dependent variable does not seem linear. Most of the dataset is clustered at the lower levels of GDP. The data points to a nonlinear type of relationship. The variation around the regression line is different. The spread of the observed values of democracy scores does not conform to the required assumptions for linear regression.

ENWONGO EKAH

Part C

Do your findings from part B indicate why, in Fish's analysis, GDP/capita is logged? More generally, why are measures of income often transformed? What is often true about their distributions?

Solution:

Yes, it does. Measures of income are often transformed to make it less skewed. It makes the distribution and data more normal.

Part D

Attempt to replicate the coefficients from Model 1, Table 3 (pg. 13) from the Fish paper. Remember that Fish's measure of Economic Development I slogged GDP/capita (GDP90LGN). Record your standard errors and the number of observations included in the regression (N). Note that your standard errors will not match those presented by Fish as he was using robust standard errors. If you were unable to replicate some coefficients, attempt to explain why and discuss whether the differences indicate different conclusions than those reached in the article.

Solution:

Table 6: Regression of Democracy Scores on Hypothesized Determinants

Coefficients:		
	Estimates	Std. Error
Intercept	1.77987	1.36773
GDP90LGN	0.69984	0.55076
ETHLING	-0.39135	0.39142
GRW7598P	0.07250	0.04431
BRITCOL	0.23542	0.26454
POSTCOM	0.37856	0.33288
OPEC	-1.25026	0.45785
MUSLIM	-1.2461	0.25158
GDP90	0.2051	0.1547
Adj R^2	0.55	
N	150	
Note	p<0.5* p< 0.01** p<0.001***	

I attempted to replicate the Model 1, from Fish article by running a regression on R. I was unable to replicate some coefficients because Fish used a robust standard error. This does not mean

there is a different conclusion from what was presented in the article because changing the unit of measurement does not change the interpretation of the regression results.

Part E

Interpret the R2 value of the regression. Does it provide evidence about the strength of the association between Muslim heritage and democratic outcomes?

Solution:

Our adjusted R^2 is 0.55, it means our model is explaining the prediction in our Y. it does provide evidence about the strength of the association based in the regression model that Muslim heritage is associated to a democratic outcome. It indicates a negative relationship between Muslim heritage and democracy score.

Part F

How do the robust standard errors used by Fish differ from the standard errors you recorded? Which set (Fish's or yours) indicate greater uncertainty about theβ's? Does the statistical significance of any of the results change?

Solution:

There is always some uncertainty associated with our βs. The robust standard errors recorded by Fish differs slightly from the ones I recorded from trying to replicate the model. Fish's model indicated greater uncertainty about the β. My Standard errors were smaller indicating more information. The statistical significance of the results did not change as the p value was significant at a $p<0.001$ level.

ENWONGO EKAH

PROBLEM 4

Part A

A widely believed statistic is that the average American family has 2.5 children. Test to see if Congress is representative of the average American family in this respect by conducting a two tailed test of this hypothesis at the α=.05 level. In addition to interpreting the results of the hypothesis test, make sure to include in your answer a formal statement of the null hypothesis and the alternative hypothesis, and the corresponding confidence interval for your estimate of the population mean. How exactly do you interpret the confidence interval?

Solution:

One Sample t- test

$H_0 : \mu =$ Congress is representative of the average American family

$H_a : \mu \neq$ Congress is not representative of the average American family.

α = .05

TABLE 7:

Sample Mean	Critical T-Value	Degrees of Freedom	T Test Statistics	95% Confidence Interval	P Value
2.49	1.963	432	-0.13117	2.33, 2.65	0.8957

The table above shows the data of the results of the one sample t-test I conducted in R. The results states that the T value is 1.963 and the t test statistics is -0.131 is in the non-rejection area. The p value of 0.8957 is statistically insignificant because it is greater than the alpha, .05 indicating that we would see a sample mean value as extreme or more extreme than 2.49 if we conduct 100 studies with a sample size of 432. Therefore, we fail to reject the null hypothesis that congress is representative of the average American family. If we were to repeat the procedure in repeated samples, 95% of the constructed confidence intervals would cover the true parameter value.

ENWONGO EKAH

Part B

Mr. Republican states that Republicans tend to have more family values than Democrats and thus have more children on average. Test this hypothesis using a two-tailed test at theα=.05 level (e.g.H0 is: "do Republicans and Democrats have the same number of children?"). Include in your answer a formal statement of the null hypothesis and the alternative hypothesis.

Solution:

Two Sample t- test

$H_0 : \mu =$ Republicans and Democrats have the same number of children

$H_a : \mu \neq$ Republicans and democrats do not have the same number of children.

$\alpha = .05$

TABLE 8

Sample Mean	Critical T-Value	Degrees of Freedom	T Test Statistics	95% Confidence Interval	P Value
2.49	1.963	510.8	11.66	0.803, 1.123	<2.2e-16

I conducted a two-sample t test to test in R and the table above is the output. The test statistic is 11.66 and the critical value is 1.963. We reject the null that Republicans and democrats do have the same number of children. The purpose of the T test was to see if Republicans and Democrats have the same number of children. The alternative hypothesis mean difference is not equals to zero. The P value goes ahead to certify this as it is closer to zero and less than the alpha .05. Therefore, if we were to create a 95% confidence interval, we will not see zero in the confidence interval.

Part C

Regress a representative's AAUW score (aauw) on his or her number of female children (ngirls) and report your results (including standard errors and p-values) in a nicely formatted table. What is the relationship between the number of female children and a representative's AAUW score? Interpret the coefficient by discussing the statistical and practical/substantive significance.

ENWONGO EKAH

Solution:

Table 9: Regression of Representative Score on Number of Female Children

	Dependent Variable		
	Representative Score		
	Coefficient	std error	P values
NGIRLS	-2.856	1.790	0.111
INTERCEPT	50.930	3.041	<2e-16***
N	430		
R^2	0.00581		
Note	p<0.5* p< 0.01** p<0.001***		

Table 9 above shows results of the regression analysis I conducted in R. The sign indicates an inverse relationship between legislative scores and number of female children. The size of slope coefficient indicates that on average, having one female child is associated with a -2.856 change in representative scores. It is statistically significant at a 0.5 level of significance, but it is not substantively significant as the relationship is not large enough to be applicable.

Part D

Regress a representative's aauw score on his or her number of female children and his or her total number of children (totchi) and report your results (including standard errors and p-values) in a nicely formatted table. Compare your results from this model to those in Part C and provide a substantive interpretation for the coefficient on the number of female children.

ENWONGO EKAH

Solution:

TABLE 10: REGRESSION OF REPRESENTATIVE SCORE ON HIS FEMALE CHILDREN AND TOTAL NUMBER OF CHILDREN

	Representative Score		
	Coefficient	std error	P values
NGIRLS	5.782	2.564	0.0246*
TOTCHI	-8.073	1.751	5.32e-06***
INTERCEPT	60.036	3.568	<2e-16***
N	430		
R^2	0.04828		
Note	p<0.5* p< 0.01** p<0.001***		

Regressing a representative's aauw score on his or her number of female children and his or total number of children, the table above is the result of her analysis.

Holding total number of children constant, number of female children is associated with a 5.782 increase in legislative scores, and it is statistically significant at a 0.5 level. Compared to part c, having female children does not make legislators to be more liberal on women issues. In this model, theR^2 has increased to 0.04828 indicating that the model has explained more of the variation in legislative scores and so we see the substantive effect.

Part E

Assuming the random sampling, linearity, normality, and homoscedasticity assumptions are met, discuss the plausibility of using this model to make causal claims about the effect of the number of daughters. In other words, do you think it is likely or unlikely that there is omitted variable bias (e.g. does the model satisfy the zero conditional mean assumption)? Please be specific. If

you believe a causal claim is plausible, discuss what assumptions you have to make. If you believe a causal claim is not plausible, explain why.

Solution:

No, this model does not satisfy the zero conditional mean assumption that there is no bias. Satisfying this assumption can be difficult because there should be nothing in the error term that is correlated to our x and y. Just considering the information provided, causal claim is not possible as there is likely that there are omitted variable bias. Research will have to demonstrate that the way a legislative vote is not related to other omitted variables that could influence his/her voting more liberal on women issues and not just because he/she has female children. It could be that there is more awareness on issues regarding women, it could be that there are lots of advocates pushing for those policies. Associated between two variables itself is insufficient evidence to claim causation.

ENWONGO EKAH

PROBLEM 5

Part A

To get a sense of whether the program helped individuals to hold jobs, regress real earnings in 1978 (re78) on the training program treatment (nsw). Interpret the coefficient on the dummy variable by discussing the statistical and practical/substantive significance. In other words, what is the "treatment effect"?

Solution

Table 11: REGRESSION OF EARNINGS IN 1978 ON TRAINING PROGRAM TREATMENT

	Coefficient		
	Estimate	std error	P values
NSW	1794.3	632.9	0.00479**
INTERCEPT	4554.8	408.0	<2e-16***
N	445		
R^2	0.01782		
Note	p<0.5* p< 0.01** p<0.001***		

On average the coefficient for nsw is 1794.3 and a standard error of 632.9. The coefficient suggests that on average, the participants of the treatment group earned $1794.3 more than those in the control group. It is statistically significant because the F statistic is 8.039 has a P value of 0.00478 larger than our alpha at 0.05and therefore we reject the null at the 0.5 significant level. It is substantive significant because earning $1794.3 is quite a bit of money.

Part B

Now repeat Part A but include the remaining covariates in the regression model. Record your findings in a table that includes the coefficient estimates, standard errors, and 95% confidence intervals.

Solution

Table 12: REGRESSION OF EARNINGS IN 1978 ON TRAINING PROGRAM TREATMENT WITH OTHER COVARIATES

Coefficient			
	Estimate	Std.Error	Confidence Intervals
Intercept	2.214e+02	2.633e+03	[-4.953513e+03, 5396.3706406]
Nsw	1.672e+03	6.343e+02	[4.253810e+02, 2918.7032654]
Age	5.367e+01	4.530e+01	[3.537343e+01, 142.7087528]
Educ	4.029e+02	1.774e+02	[5.423361e+01, 751.6605976]
Black	-2.039e+02	1.164e+03	[-4.326903e+03, 247.9704339]
Hisp	4.246e+02	1.561e+03	[-2.643077e+03, 3492.3745726]
Married	-1.467e+02	8.810e+02	[-1.878150e+03, 1584.8267121]
re74	1.236e-01	8.698e-01	[-4.738242e-02, 0.29455278]
re75	1.946e-02	1.489e-01	[-2.732313e-01, 0.3121483]
u74	1.381e+03	1.186e+03	[-9.49244e+02, 3711.2423281]
u75	-1.072e+03	1.023e+03	[-3.082314e+03, 938.6796888]
N	445		
R^2	0.03652		
Note	p<0.05* p<0.01** p<0.001***		

Part C

ENWONGO EKAH

Interpret the coefficient for the treatment indicator (nsw). Do the results change with these control variables? What does this indicate about the experimental set-up employed by the program evaluators?

Solution:

Holding all other variables constant, the coefficient for nsw suggests that on average the participants earned 1.672 more than the control group. The results seem to change with the control variables. This is not surprising as all the covariates seems likely to be related to earnings. There could be other omitted variables that are not controlled for like location, that could still influence the earnings so we cannot say that the experimental set-up employed does not mean that participating in the treatment led to having higher earnings than the control group.

Part D

To further investigate the efficacy of the experimental set-up, create a table recording the mean of each covariate for the treated and control groups.

Solution:

TABLE 13: Mean Values for Each Covariate

	Treated (nsw=1)	Control (nsw=0)
Age	25.81622	25.05385
Educ	10.34595	10.08846
Black	0.8432432	0.8269231
Hispanic	0.05945946	0.1076923
Married	0.1891892	0.1538462
re74	2095.574	2107.027
re75	1532.056	1266.909
u74	0.7081081	0.75
u75	0.6	0.6846154

PART E

The data set nsw_psid_withtreated.csv retains the treated units from the experiment (i.e. those eligible applicants who got the training), but the control units from the experiment have been

replaced by a non-experimental sample from the Population Survey of Income Dynamics. In other words, rather than comparing eligible individuals who applied to the NSW centers for support, this analysis is now comparing NSW-eligible individuals who got the NSW training with the entire U.S. labor force. In effect, instead of randomly assigning treatment at NSW centers and then comparing eligible participants with and without the treatment, we will measure the effect of the NSW treatment by comparing the outcomes of eligible participants who received the treatment to all individuals in the labor force. Using this new data, re-estimate Parts A and B. How do the results differ?

Solution:

TABLE 14: REGRESSION ANALYSIS OF EARNINGS ON THE ENTIRE US LABOR FORCE

	Coefficient		
	Estimate	std error	P values
NSW	-15204.8	1154.6	<2e-**
INTERCEPT	21553.9	1154.6	<2e-16***
N	2675		
R^2	0.06057		
Note	p<0.5* p< 0.01** p<0.001***		

REGRESSION ANALYSIS WITH TOTAL US LABOR WITH ALL COVARIATES

	Estimate	Std.error	Confidence Intervals
Intercept	9.5536e+02	1.371e+03	[-1733.9, 3641.11]
Nsw	1.154e+02	1.007e+03	[-1858.91, 2089.74]
Age	-8.977e+01	2.194e+01	[-132.7858, -46.74]
Educ	5.141e+02	7.644e+02	[364.2285, 664.0194]

Black	-4.542e+02	4.969e+02	[1428.5292,520.0973]
Hisp	2.197e+03	1.092e+03	[56.8375, 4337.9084]
married	1.205e+03	5.855e+03	[56.7444, 232.8248]
re74	3.126e-01	3.163e-02	[0.25059, 0.37464]
re75	5.436e-01	3.090e-02	[0.48306, 0.60425]
u74	2.390e+03	1.024e+03	[380.7544, 4398.3066]
u75	-1.462e+03	9.472e+03	[-3319.2776, 395.3475]
N	2675		
R^2	0.5855		
Note	p<0.05* p<0.01** p<0.001***		

The results when comparing the outcome of the eligible participants seems to differ. In part A, there seem to be on average a negative decrease in earnings, makes it seem that there is an inverse relationship in participating in the treatment and increase in earnings, statistically significant and substantive. In part B, controlling for other variables, there was a decrease to an average a coefficient of 1.672 with a standard error of 6.643, statistically significant but not substantive. Comparing those Part B to our new data set, the nsw on average shows earnings of 1.154 with the standard error of 1.007 there seem to be a decrease in the earnings but not large enough when compared with the entire population. The R^2 value has also increased in the data sets indicating that this regression model has explained more of the variation. The F statistics is large, 378.8 with a P value close to zero, it is statistically significant, and we reject the null. Controlling for variables, the information in Part B and the entire labor force shows that the treatment was not substantially significant as the dollar amount earned was too little to make an impact.

Part F

Using the data in nsw_psid_withtreated.csv, repeat part D, by creating a table with means of each variable in the treatment and the control.

Solution:

Table 15: Mean Values for Each Covariate of US Labor Population

	Treated (nsw=1)	Control (nsw=0)
Age	25.81622	34.8506
Educ	10.34595	12.11687
Black	0.8432432	0.2506024
Hispanic	0.05945946	0.03253012
Married	0.1891892	0.8662651
re74	2095.574	19428.75
re75	1532.056	19063.34
u74	0.7081081	0.08634538

u75	0.6	0.1

Part G

What does this table indicate about the differences between the entire U.S. labor force and NSW-eligible individuals within the study (e.g. those that were going to NSW centers)? In other words, how were those who were participating in the experiment different than the broader population of workers?

Solution:

The table above seems to indicate that the mean value for each covariates of the US labor force seems larger than those participating in the experiment. The ultimate goal of experiment is to make inferences and causation. Participating in the treatment group didn't seem to have make any difference when compared to the larger population, there could be other omitted variables missing that were not controlled for. It looks like there may be an association between individuals who applied to get the treatment and earnings. This data cannot be replicated to the larger population and we can't conclude a causal claim that participating in the experiment led to an increase in earnings.