**Week 2 Lab Handout- Simple Regression**

**PA 5032 – Applied Regression**

January 29th, 2021

Wenchen Wang

**PART A: SINGLE-VARIABLE REGRESSION ~20min**

**PART B: HYPOTHESIS TESTING ~40min**

**PART C: REGRESSION OUTPUT TABLES~15min**

Data File: PSID

Contents: 9607 Observations

Variables:

release \_number

family\_interview

reference\_age

house\_value

dollars\_rent

race\_black

reference\_income

family\_income

healthcare\_expenditures

family\_incomeunder100k

**PART A: SINGLE-VARIABLE REGRESSION**

If we wanted to explore the relationship between age and household family income in our PSID sample population, we could run a simple regression:

*regress family\_income age\_reference OR*

*reg family\_income age\_reference*

* What is our model for this regression? Write the equation:
* Interpret the constant or Y-intercept in a sentence.
* Interpret the slope coefficient. Is it statistically significant?
* What is the Adjusted R-squared value? Interpret it in a sentence.(This is also call the TSS- Total Sum of Squares)

**PART B: HYPOTHESIS TESTING**

**(1) T-Test**

We use t-test in statistics when we want to examine whether the observed difference is real (alternative hypothesis) or just due to chance variation (null hypothesis).

A t-statistic (t-value) measures the difference between the data and what is expected on the null hypothesis: **t-value = coefficient/SE**

P-value (the observed significance level) is the chance of getting a test statistic as extreme as or more extreme than the observed one. The chance is computed on the basis that the null hypothesis is right. **Note that P-value does not give the chance of the null hypothesis being right**. Small values of P are evidence against the null hypothesis: they indicate something besides chance was operating to make the difference.

For example, if we wanted to test to see if the actual or true average age of the parameter (in this case the total US population) is 29 years, we could run a t-test using our sample data:

*ttest reference\_age==29*

* What is the “null hypothesis” for this t-test? Our alternative?

T-test example between subgroups of sample, for example, between the high and low family income groups:

*ttest reference\_age, by(family\_incomeunder100k)*

* Is the difference between these two subgroups real?

T-test example with single-variable regression:

*reg family\_income reference\_age*

* When examining the relationship between family income and age, are the coefficients statistically significant? If they are significant, at what percent level?

**(2) F-Test**

The F-statistic is used to test the joint significance of a group of variables. In contrast, the t-test we talked about before tests only the significance of one variable, or the difference between the mean of a variable or between two sub-samples

For example, suppose we regress healthcare expenditures on family income, controlling for the income of the reference person.

* We’re a little suspicious that the effect of family income on healthcare expenditures may be the same as the effect of reference person income. What would be the null hypothesis of our F Test?

*reg healthcare\_expenditure family\_income reference\_income*

H\_0:

**There is a statistically significant difference between the effect of family income on healthcare expenditures and the effect of the reference person income on healthcare expenditures.****The F statistic indicates that we can reject the null hypothesis. Family income and reference income have unique effects.**

* What does the following regression’s F-test tell us? State the null hypothesis and interpret the results.

*reg family\_income reference\_age number\_children*

H\_0 :

Results:

**PART D. REGRESSION OUTPUT TABLES**

To get nice clean regression output, you have two options, **estout and outreg2**. Both of these are packages you can install on Stata using the command: *ssc install [estout OR outreg2]*

You then need to run a set of commands *Immediately* following your regression.

Ex: *ssc install outreg2*

*reg family\_income age\_reference*

*outreg2 using lab2\_reg.doc, replace ctitle(Model 1)*

|  |  |
| --- | --- |
|  | (1) |
| VARIABLES | Model 1 |
| reference\_age | 315.4\*\*\* |
|  | (52.77) |
| Constant | 58,853\*\*\* |
|  | (2,573) |
| Observations | 9,596 |
| R-squared | 0.004 |

Standard errors in parentheses

\*\*\* p<0.01, \*\* p<0.05, \* p<0.1

OR

*ssc install estout*

*reg family\_income age\_reference*

*est store m1, title(Model 1)*

*estout m1, cells(b(star fmt(3)) se(par fmt(2))) replace starlevels(\* 0.1 \*\* 0.05 \*\*\* 0.01) legend label varlabels(\_cons Constant)stats(r2 N)*

