

**Advanced Regression  
Problem Set 1  
Due: September 21, 2009**

*I.* The U.S. is currently experiencing a wave of immigration rivaling those at the turn of the 20<sup>th</sup> century. This has engendered several movements across the U.S. challenging current immigration policy and seeking to limit immigration flows.<sup>1</sup>

A central question in this debate is whether, or to what extent, immigrants exact economic and social burdens on U.S. native born citizens. What are the earnings of immigrants, and does the average immigrant experience lower wages (and slower wage growth) than observationally equivalent natives? Consider the models below (Table 1.1), which have been estimated by Ordinary Least Squares (OLS) regression. Specifically, the models regress a measure of hourly wages on a dummy variable indicating immigrant status, controlling for a number of demographic and human capital characteristics. These data are drawn from the 1990 Census Public Use Micro Sample (PUMS).

*a)* Interpret the slope coefficient on “immigrant” for each model. Is the coefficient statistically significant? If so, at what level? State clearly the hypothesis being tested. Please provide an explanation for why the coefficient changes the way it does as you move from model (1) to model (5).

*b)* Fill in the table.

*c)* Does the addition of education, experience and experience-squared add significant explanatory power to model (2) over model (1)?

*d)* What is the value of the R-squared in model (2)? Describe what the R-squared means in general terms (and provide equations), and then interpret the value for this model.

*e)* Interpret the coefficient on “education” in model (5). Is it statistically significant? If so, at what level? Do you agree with the way “education” is measured? Why or why not?

*f)* Which model is best? Why?

---

<sup>1</sup> For example, some states recently attempted to sue the federal government to recover the costs they incur because of large numbers of immigrants who require social services yet pay little in taxes.

**Table 1.1**  
**The Impact of Immigrant Status on Wages**

	(1)	(2)	(3)	(4)	(5)
Immigrant (=1)	-0.129 (0.006)	-0.014 (0.006)	-0.008 (0.006)	-0.009 (0.006)	-0.140 (0.009)
Experience		0.033 (0.001)	0.033 (0.001)	0.031 (0.001)	0.031 (0.001)
Experience-squared		Yes	Yes	Yes	Yes
Education		0.113 (0.001)	0.114 (0.001)	0.113 (0.001)	0.113 (0.001)
Female (=1)			-0.611 (0.003)	-0.607 (0.003)	-0.607 (0.003)
Married (=1)				0.056 (0.004)	0.056 (0.004)
Years in U.S.					0.008 (0.0003)
Residual/Unexplained Sum of Squares	211,732	191,064	?	167,370	?
R-Squared	0.0016	?	0.2100	?	0.2119

Note: The dependent variable is measured as the natural logarithm of hourly wages [ $\ln(\text{hourly wage})$ ]. Immigrant is a dummy variable that equals 1 if the individual is an immigrant to the U.S. Experience is the number of years each individual has been at the current job. Education is measured as the number of years of completed education. Female is a dummy variable that equals 1 if the individual is female. Married is a dummy variable that equals 1 if the individual is currently married. Years in the U.S. is a continuous variable indicating the total number of years each individual has resided in the U.S. Data are drawn from the 1990 PUMS Census Data. All analyses are based on 253,885 observations.

2. From the early-1970s throughout the 1980s, pressure had been building to reform existing welfare programs by changing the incentives in order to encourage work. However, relatively little empirical evidence existed on the effectiveness of employment incentive programs in moving families from welfare to work. Much of what was known came from three policy experiments in Arkansas, Maryland (Baltimore), and Virginia conducted during the 1980s. In each case, recipients of Aid to Families with Dependent Children (AFDC) were randomly assigned to a control or treatment (experimental) group. Those assigned to the control group had no obligation to seek or hold employment. However, those in the treatment group were expected to look for a job, often with assistance from program administrators in the local welfare office.

Consider the dataset **welfare1.dta**. This dataset contains information on program participants (treatment and control group members) from the three states in which the experiments were conducted. For this problem, let's concentrate on the experimental results from Maryland. In Maryland, Baltimore City's experimental group had a choice of activities to engage in: help with a job search, workfare (working in an entry level job for free), or education and job training activities. Total program costs were about \$1,000 per participant, and failure to participate would result in a partial loss of welfare benefits.

In this problem, you will be walked through several commands in order to increase your familiarity with Stata. Here are some useful tips to consider before you go through this problem:

### Opening a Dataset in Stata

The data are in Stata format. You know this because the data file ends with the “.dta” extension. You should first download the data from the web onto your hard drive, and then open the dataset from Stata using the “open” command under the “file” menu. Do not double-click the Stata file icon to try to open it; this will not work.

### Checking Your Data

After you have opened the dataset, check to see if the data came in without any problems. Type `describe` in your command window. Also, type the `edit` command to make sure the data have loaded properly. Finally, type `summarize` to produce a table of summary statistics. Conduct an ocular logic test to make sure the variable means, minimum and maximum values, and standard deviations make sense.

*a)* As a first step, describe your data. This will allow you to see the variable names and labels, the total number of variables included in the dataset, and the number of observations. To describe your dataset, type the command `describe` in the command window of Stata. When you do this, Stata will display all of the variables and some information about the dataset. You may notice that Stata scrolls to fill the results window, and includes a “–more–” at the bottom-left-hand-corner of the results window. In order to see additional results (your output may be larger than what can be displayed in the results window), hit the space bar or click the “–more–” with your mouse. These additional results will be displayed, and another “–more–” may appear at the bottom of the results window. If so, press the space bar again.

Now look for the following pieces of information:

1. How many observations (individuals, in this case) are there in the dataset?
2. How many variables are in the dataset?

*b)* Now you need to reduce the size of the dataset to include observations from Maryland (since we want to limit the analysis to Maryland program participants). First, you need to figure out how to identify the Maryland cases. To do this, enter the command:

```
tab state
```

Clearly, Maryland cases are designated with the code “2”. In order to drop all the non-Maryland cases, enter the command:

```
keep if state == 2
```

At this point, Stata will tell you how many observations were dropped. Also, notice the double equals sign. This is required anytime you have a condition after the “if” statement.

As an alternative, you could have achieved the same result by typing (you will need to reopen your dataset in order to obtain these results if you have already entered the command `keep if state == 2`):

```
drop if state == 1 <press enter, then type>
drop if state == 3
```

Note that two separate commands are necessary to complete this step. We could have accomplished the same thing by typing this single command:

```
drop if state == 1 | state == 3
```

where the “|” represents the logical operator “or.”

Type `describe` again to check the number of remaining observations. How many are there?

*c)* Since this is a randomized experiment, we need to check the quality of the random assignment to treatment and control groups. This is accomplished by looking at each background characteristic and verifying that the assignment to a given group is not associated with background characteristics. To do this, you will need to conduct several chi-squared tests. Conduct chi-squared tests to see if assignment to the treatment or control group is statistically independent of:

1. Whether or not the participant had received AFDC payments for two years or less;
2. Whether or not the participant had been an AFDC recipient for more than two years;
3. Race;
4. Age;
5. Whether or not the participant has a high school diploma;
6. Whether or not the participant has children less than 6 years old; and
7. Whether or not the participant has children less than 12 years old.

To conduct a chi-squared test, you should, for example, enter the following command:

```
tab exp nonwht, chi2
```

This will produce a cross-tabulation between the assignment and race variables, along with a chi-squared statistic and p-value. Recall the null hypothesis being tested: that assignment to treatment/control groups is statistically independent of race. As expressed, the Stata command above will only generate the number of observations associated with each assignment-race combination. If you want to see the row and column percentages, you should type the following command:

```
tab exp nonwht, row col chi2
```

Does this appear to be a good randomized experiment? Why or why not?

*d)* Now test to see if the program had an impact on welfare participants in Maryland. In this case, use a t-test to compare the treatment and control groups in terms of average labor market earnings in quarters 7 to 10 after random assignment. To do this, you first need to create two new variables for the earnings of each group.

```
gen earn7_10t = earn7_10 if exp == 1
```

```
gen earn710c = earn7_10 if exp == 0
```

Again, notice the double equals sign (=) after the “if” statement. Furthermore, did you notice that Stata told you how many missing observations were generated? Does this make sense? And for whom were missing values generated? To convince yourself, enter the following command:

```
tab exp
```

You’ll notice that the missing observations generated correspond to the sample size of the treatment or the control group.

Before conducting the t-test, we need to make an assumption. Do you think the distribution of earnings for the treatment and control groups has the same amount of dispersion? In other words, can you assume that they have equal variances? Probably not (why?). If we’re going to assume unequal variances in earnings, the relevant t-test command is:

```
ttest earn710t = earn710c, unpaired unequal
```

The `unpaired` option is needed in all t-tests. The `unequal` option reflects the assumption we made above.

Alternatively, we could perform the following t-test:

```
ttest earn710t = earn710c, unpaired
```

Of course in this formulation, we have assumed equal variances across treatment and control group earnings.

There is an alternative (and perhaps easier) way of conducting this t-test. It allows you to perform the test without having to generate new variables. The Stata command is the following:

```
ttest earn7_10, by(exp) unequal
```

This command simply tells Stata to run a t-test comparing average earnings across the treatment and control group members, while continuing to assume unequal variances in the earnings distributions.

After conducting the test, interpret your results. What was the hypothesis being tested? Your initial level of statistical significance? What do you find? Are you able to reject the null hypothesis? Does the provision of work incentives improve earnings?

*e)* Did the experimental treatment have an effect on the average AFDC payment size in periods 7 to 10? Conduct an appropriate test.

*f)* Did the treatment affect the proportion of people on AFDC 10 quarters after initial assignment? Conduct an appropriate test.

*g)* Did the treatment affect the proportion of people with a job 10 quarters after initial assignment? Conduct an appropriate test.

*h)* What do you think? Is the program treatment in Baltimore effective? Why or why not?

3. Since the late 1970s, the gap in wages and employment rates between black and white young men has grown. One of many explanations for these large racial differences is the relative absence of job contacts for young men in minority groups. The purpose of this problem is to use a dataset to investigate the impact of employment contacts and social networking on wages for white and black youths.

The dataset you will use is a Stata file called **boston.dta**. The data come from the 1989 National Bureau of Economic Research Study of Disadvantaged Youth in Boston. The survey was designed to evaluate how disadvantaged youths fared in Boston's tight labor market during the late-1980s. The original data come from a sample of 1,200 individuals ages 17 to 24 from three high-poverty neighborhoods. The sample (N=264) contained in this data file consists of 150 white and 114 black employed young men. The file contains information on hourly wages, race, whether the young men received their job through a personal contact (a friend or relative), whether they are holding a union job, how long they have been at their current job, the number of years of employment experience they have had, whether they are currently enrolled in school, educational attainment (represented by four dummy variables), and occupation (represented by eight dummy variables, including a missing category).

a) Prior to carrying out a multiple regression analysis, verify whether and to what extent:<sup>2</sup>

1. Black youths earn less than white youths;
2. Wages of jobs found through contacts are higher than those found without contacts;
3. Repeat 2. for white and black youths separately; and
4. Determine whether contacts are associated with any of the occupational groupings.

b) Then, use regression analysis to build models that examine:

1. The extent to which the black-white difference in wages (log of wages) is reduced when one controls for the demographic characteristics of sample members, i.e. only education, whether still in school, and work experience.
2. Is the wage differential influenced when "job contact" is added to the model? Interpret the coefficient on "job contact." Is the variable statistically significant? If so, at what level? Do job contacts have an economically significant impact on wages?
3. Now estimate a regression in which you add the characteristics of the job (tenure, union, and occupation) to the model you estimated earlier. What happens to the coefficient on "black"? Is it still statistically significant?
4. Finally, it could be the case that the impact of job contacts on wages operates differently for black and white youths. That is, job contacts may have differential effects on wages depending on the race of the sample member. How would test this question? Create all necessary variables and conduct the relevant regression analysis. What do you find?

Your objective in this problem is to perform an analysis of the data by systematically building a regression model and justifying the inclusion of specific variables. This will require you to use F-tests. See the notes at the end of this problem set for a discussion of issues in Stata.

---

<sup>24</sup> Some of these questions will require you to perform t-tests or chi-square tests. You can do these either within a regression framework, or as separate hypothesis tests. See the notes at the end of the problem set for a discussion issues related to STATA.

Here are some useful tips to consider before you go through this problem:

### Running a Regression in Stata

Performing a basic regression in Stata is relatively simple. If the dependent variable is “lnwage” and the key independent variable is “black,” you would type the following command:

```
regress lnwage black
```

Stata produces output that is critical to fully and correctly interpret your results. Be sure to go through this output carefully to make sure you understand it. You may also want to calculate some additional statistics. First, to test a hypothesis about one coefficient in your regression, type the following command after Stata has displayed the regression results:

```
test black
```

This command tests the null hypothesis that the coefficient on “black” is equal to zero. Stata produces an F-statistic, along with a p-value. Try this. Then, take the square root of the F-statistic and compare it to the t-statistic reported in the regression output. What do you find?

You can perform mathematical calculations in Stata by using the `display` command. Type:

```
display sqrt(11.03)
```

### Performing a Joint F-test After a Regression

If you are interested in testing the null hypothesis that a set of coefficients on two or more variables are jointly equal to zero, you can run an F-test after Stata presents the regression results. For example, say you are interested in estimating the following regression model and asking whether the addition of “contact” and “educational attainment” together add significant explanatory power to the model. Typing the following would get you an answer:

```
regress lnwage black contact hsgrad scol colgr  
  
test contact hsgrad scol colgr
```

This regresses `lnwage` on `black` `contact` `hsgrad` `scol` `colgr` and then performs an F-test of the hypothesis that `contact` `hsgrad` `scol` `colgr` are jointly equal to zero.

### Performing a Chi-squared Test

A chi-squared test is designed to test for statistical independence in the distribution of data across a set of qualitative categories. For example, in this problem you are asked to test if “contact” is evenly distributed across various occupations. This requires a chi-squared test on the cross-tabulation of contact and occupation (you will have to generate a new categorical variable for occupation since the dataset contains only dummy variables). To perform this test, enter the following command:

```
tab contact occupation, chi2
```