Pooled Panel Regression

R Tutorials for Applied Statistics

1 Introduction

Cross sectional data in when a sample has multiple sampling units (observations) measured once in a single time period.

Time series data is when a sample contains measures a single sampling unit measured over multiple time periods. For example, the closing price of the Dow Jones Industrial Average for each of the last 100 days is a time series.

Panel data is when a sample has multiple sampling units measures over multiple time periods. For example, the a group of people could be selected for a sample and their monthly earnings, experience, education and other factors could be measured for multiple time periods.

Pooled panel combines observations from multiple time periods into a single sample, and treats it as a cross section. That is, if a single person is measured in 2012 and 2013 for some variable, each of these measures appears as its own independent observation in the sample, thereby creating a larger sample size than the number of individuals measured.

There are more advanced panel methods that we will discuss in future tutorials that treat more carefully the possible dependence of multiple observations of a single individual across time.

2 Example: Wages, education, and gender over time

The code below downloads and loads an R data set that includes variables from the U.S. Current Population Survey for usual hourly earnings, educational attainment, age, sex, and race for individuals in 1980, 1990, 2000, and 2010.

```
load(url("http://murraylax.org/datasets/cpswages.RData"))
```

We will use a pooled panel regression to predict usual hourly earnings based on education, age, and sex. Usual hourly earnings will certainly be different on average depending on the time period, so we must construct a regression model that accounts for this. Reasons that usual hourly earnings will differ include:

1. Usual hourly earnings are in nominal terms for one. Rising price levels over each decade lead to higher average nominal earnings.
2. Even if usual hourly earnings are measured or converted to real terms, real per-capita compensation rose along with standard of living over each of the decades in the sample.
3. Finally, any differences in economic conditions in each of the time periods may affect the average of the outcome variable.

3 Dummy-up categorical variables

To account for these differences in the average outcome variable based on time, we create multiple dummy variables so that we can account for every time period. Generally, one should create *one fewer* dummy variables than there are categories. For example, if sex has two categories, “male” and “female”, one should create a single dummy variable for one of the categories, where the value is equal to 1 if the observation fits the category, and 0 if it does not.

There are *four* time periods, so we will create *three* dummy variables for time periods. We will let the first time period, 1980, serve as the base for comparison, and create dummy variables for *year1990*, *year2000*, and *year2010*. Each of these variables is equal to 1 if the observation is for the associated year, and 0 otherwise. When all of these dummy variables are equal to 0, we know the observation is for the omitted (i.e. base) time period, 1980.

When we call the `lm()` function, we can use the function `factor()` to tell `lm()` to treat the multiple categories for year and multiple dummy variables. The `lm()` function by default makes the first year the omitted category.

4 Estimating the model with time dummies

The following call to `lm()` estimates a regression model predicted the natural log of usual hourly earnings (`usualwage`) based on education (`educ`), a dummy variable for female (`female`), and dummy variables for the time variable.

```
lmwages <- lm(log(usualwage) ~ educ + I(educ^2) + age + I(age^2) + female
              + factor(year), data=cpswages)
```

Both `educ` and `age` appear linearly and with a squared term, as in `I(educ^2)`. The `I()` function tells R to evaluate the expression within the `I()` function, whose result is then used *as is* as an explanatory variable in `lm()`. Squared terms are included to allow for the possibility of increasing or decreasing returns to education and age.

Let us examine a summary of the regression results:

```
summary(lmwages)
```

```
##
## Call:
## lm(formula = log(usualwage) ~ educ + I(educ^2) + age + I(age^2) +
##     female + factor(year), data = cpswages)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.2719 -0.3202  0.0381  0.3909  5.1548
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)       1.236e-02  1.067e-01   0.116    0.908
## educ              1.235e-03  1.933e-03   0.639    0.523
## I(educ^2)         5.599e-05  1.210e-05   4.628 3.78e-06 ***
## age               5.524e-02  4.176e-03  13.228  < 2e-16 ***
## I(age^2)         -5.202e-04  4.986e-05 -10.433  < 2e-16 ***
## female           -2.502e-01  2.123e-02 -11.783  < 2e-16 ***
## factor(year)1990  4.384e-01  3.015e-02  14.541  < 2e-16 ***
## factor(year)2000  8.155e-01  3.135e-02  26.011  < 2e-16 ***
## factor(year)2010  1.084e+00  2.873e-02  37.726  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.745 on 4974 degrees of freedom
## Multiple R-squared:  0.403,  Adjusted R-squared:  0.402
## F-statistic: 419.6 on 8 and 4974 DF,  p-value: < 2.2e-16
```

Each of the coefficients on the time dummy variables are positive and statistically significant. These coefficients measure the difference in (log) usual wages of each time period *relative to the base period, 1980*. Therefore, relative to 1980, usual wages were 43.8% higher in 1990; usual wages were 81.6% higher in 2000 relative to 1980; and finally, usual wages were 108.4% higher in 2010 than 1980. Remember that these differences include both the effects of inflation and improvements in real wages / standard of living.

5 Comparing marginal effects across time periods

We can see from the above regression results that the coefficient on female is negative and statistically significant. With a coefficient on `female` equal to -0.25, regression suggests that women earn on average 25% less than men. This is after accounting for education and wage.

This is an average over the entire sample period, that includes data from 1980 through 2010. Is there evidence that the gender gap has changed over time? To answer this, we interact our time dummies with the variable for female:

```
lmwages_gap <- lm(log(usualwage) ~ educ + I(educ^2) + age + I(age^2) + female
             + factor(year) + female:factor(year), data=cpswages)
```

Notice the inclusion of `female:factor(year)` in the regression equation. I intentionally saved the output of the `lm()` function under a different name, `lmwages_gap`, so that we can compare the results later. Let us look at a summary of the results:

```
summary(lmwages_gap)
```

```
##
## Call:
## lm(formula = log(usualwage) ~ educ + I(educ^2) + age + I(age^2) +
##     female + factor(year) + female:factor(year), data = cpswages)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.2492 -0.3186  0.0380  0.3835  5.1745
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)               3.756e-02  1.074e-01   0.350  0.72651
## educ                      1.556e-03  1.935e-03   0.804  0.42135
## I(educ^2)                 5.372e-05  1.212e-05   4.434 9.46e-06 ***
## age                       5.522e-02  4.174e-03  13.230  < 2e-16 ***
## I(age^2)                 -5.202e-04  4.983e-05 -10.440  < 2e-16 ***
## female                   -3.293e-01  4.150e-02  -7.936 2.56e-15 ***
## factor(year)1990          4.236e-01  4.113e-02  10.299  < 2e-16 ***
## factor(year)2000          7.609e-01  4.211e-02  18.068  < 2e-16 ***
## factor(year)2010          1.018e+00  3.833e-02  26.553  < 2e-16 ***
## female:factor(year)1990  4.015e-02  6.027e-02   0.666  0.50531
## female:factor(year)2000  1.233e-01  6.250e-02   1.972  0.04863 *
## female:factor(year)2010  1.481e-01  5.681e-02   2.606  0.00918 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7446 on 4971 degrees of freedom
## Multiple R-squared:  0.404,  Adjusted R-squared:  0.4027
## F-statistic: 306.3 on 11 and 4971 DF,  p-value: < 2.2e-16
```

Each of the interaction terms for the year dummy variable and female measure the *additional* marginal effect that being female has on (log) usual wages *relative to the base year, 1980*.

- The coefficient for `female` equal to -0.329 suggests that women earned on average 32.9% less per hour than men in 1980.
- The coefficient on the interaction term `female:factor(year)1990` equal to 0.040 suggests that the gender gap was 4 percentage points smaller in 1990 versus 1980. This is not statistically significantly different from zero, meaning we failed to find evidence that the gender gap decreased from 1980 to 1990.
- Add to coefficient on `female` the coefficient on the interaction term `female:factor(year)1990` = -0.329 + 0.040 = -0.289. This suggests that women earned on average 28.9% less than men.
- The coefficient on `female:factor(year)2000` equal to 0.123 suggests that the gender gap was 12.3 percentage points smaller in 2000 versus 1980. *Again, note that the point of comparison is 1980, the base year*. This is statistically significant, so we can say we have statistical evidence that the gender gap is smaller in 2000 versus 1980.
- Add to coefficient on `female` the coefficient on the interaction term `female:factor(year)2000` = -0.329 + 0.123 = -0.206. This suggests that women earned on average 20.6% less than men.
- The coefficient on `female:factor(year)2010` equal to 0.148 suggests that the gender gap was 14.8 percentage points smaller in 2010 versus 1980. *Again, note that the point of comparison is 1980, the base year*. This is statistically significant, so we can say we have statistical evidence that the gender gap is smaller in 2010 versus 1980.
- Add to coefficient on `female` the coefficient on the interaction term `female:factor(year)2010` = -0.329 + 0.148 = -0.181. This suggests that women earned on average 18.1% less than men.

6 Structural change across time

Is the marginal effect for female different across time? We can use a joint F-test to conduct this test. Our hypothesis (written in plain English) is given by,

H_0 : The population coefficients for all the interaction terms are equal to zero
H0:The population coefficients for all the interaction terms are equal to zero

H_1 : At least one of the population coefficients for the above interaction terms is not equal to zero
H1:At least one of the population coefficients for the above interaction terms is not equal to zero

The joint F-test compares the sum of squared explained between the nested models: the unrestricted model that includes all the interaction terms with female and the time dummies, and the restricted model with excluded these interaction terms.

We estimate the f-test with the following call to `anova()`:

```
anova(lmwages, lmwages_gap)
```

```
## Analysis of Variance Table
##
## Model 1: log(usualwage) ~ educ + I(educ^2) + age + I(age^2) + female +
##     factor(year)
## Model 2: log(usualwage) ~ educ + I(educ^2) + age + I(age^2) + female +
##     factor(year) + female:factor(year)
##   Res.Df    RSS Df Sum of Sq      F  Pr(>F)
## 1   4974 2760.7
## 2   4971 2756.0  3    4.6969 2.8239 0.03731 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The p-value is equal to 0.0373. This is less than 0.05, so we reject the null hypothesis. There is statistical evidence that the effect being female has on wages is different over time.