The Regression Tool

Yona Rubinstein

July 2016

Regressions

- *Regression analysis* is one of the most commonly used statistical techniques in social and behavioral sciences.
- Its main objective is to explore the relationship between a **dependent** variable and *one* or *more* **independent** variables.
- Regression analysis establishes a *correlation* between phenomena.
- In most cases regression is used with *observational data*.
- Therefore, regression estimates may or may not have a *causal interpretation*. And as the saying goes, *correlation* is not *causation*.
- Setting aside the **causality problem** for the moment, we start with the mechanical properties of regression analysis.
- Focus on the concepts and NOT on the formulations!!

Road Map

- There are universal features of the *Population Regression* (and its sample analog) and not with a researcher's interpretation of his output,
- Specifically:
  - The **connection** between the *Population Regression Function* (PRF) and the *Conditional Expectation Function* (CEF);
  - Why estimates / regression coefficients change as covariates are added or removed from the regression model;
  - The link between regression and other "control" strategies;
  - The sampling distribution of regression estimates.
- In this note we'll focus on the first two questions.

The Conditional Expectation Function

- The connection between schooling, type of a job, cognitive abilities (IQ), family backgrounds and average earnings has considerable predictive power, in spite of the enormous variation in individual circumstances.
- Of course, the fact that more educated people earn more than less educated people *does not mean* that schooling causes earnings to increase.
- The question of whether the earnings-schooling or earninigs-job or earninigs-IQ relationships are *causal* is of major importance, of course!
- Yet, even without resolving that it's clear that education, family backgrounds etc., IQ, occupation and industry predict earnings in a narrow statistical sense.
- This predictive power is summarized by the *Conditional Expectation Function* (hereafter, CEF).

CEF Definition

- The **CEF** for a dependent variable Y_i given a vector of covariates X_i is the <u>*population average*</u> of Y_i with X_i **held fixed**.
- The <u>*population average*</u> can be thought of as the mean in an infinitely large sample, or the average in a completely enumerated finite population.
- The CEF is written as:

$$E\left[Y_i \mid X_i\right].$$

- The CEF is a function of X_i.
- Sometimes we might work with a particular value of X_i, for example the mean wages of MBA students:

$$E\left[WAGES_i \mid EDUCATION_i = MBA\right]$$

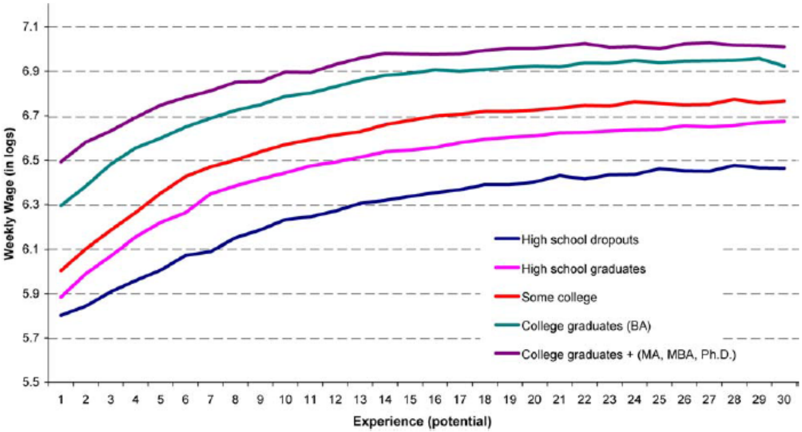
- In our previous examples we examine the gap in mean log hourly wages between sectors (finance; others):

$$E\left[Y_i \mid F_i = 1\right] - E\left[Y_i \mid F_i = 0\right]$$

where F_i is a binary (zero-one) variable that equals 1 if worker i is

Populations and Samples

- **Expectation** is a *population* concept.
- In practice, data usually come in the form of *samples* and rarely consist of an entire population. We therefore use *samples* to make *inferences* about the *population*.
- For example, the *sample* CEF is used to learn about the *population* CEF.
- We postpone a discussion of the formal inference step taking us from sample to population.
- Let's move to the data.
- Figure 1, taken from Rubinstein and Weiss (2008) shows white males, mean weekly wages (in logs) by education and experience for full-time full-year workers (52 weeks), using the CPS, March supplement for the years 1964–2002.



Weekly Wage (in logs)
7.1
6.9
6.7
6.5
6.3
6.1
5.9
5.7
5.5
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30
Experience (potential)
High school dropouts
High school graduates
Some college
College graduates (BA)
College graduates + (MA, MBA, Ph.D.)

Conditional Means in Reality

- These wage profiles have the general shape found in previous studies.
- The CEF in the figure above captures the fact that - despite the large variation in person specific circumstances - average wages are well ranked by educational attainments.
- People with more schooling generally earn more, on average for any level of potential experience.
- Mean wages increase rapidly (by approximately 80 percent) over the first 10 to 15 years of a career. As careers progress, we find little change in mean wages.
- The average earnings gain associated with a year of schooling is typically about 8 percent.

The Decomposition Property

- Perhaps the most important **property** of the *CEF* is the *decomposition property*. That is, any random variable Y_i can be expressed as:
$$Y_i = E\left[Y_i \mid X_i\right] + e_i.$$
- The term ε_i is subject's i deviation from the population mean, conditional on her X.
- The *conditional expectation* theorem means that:
  - ε_i is mean-independent of X_i, i.e., $E\left[\varepsilon_i \mid X_i\right] = 0$ and that,
  - ε_i is uncorrelated with any function of X_i.
- This theorem says that **any** *random variable* Y_i can be decomposed into (i) a piece that's "explained by X_i", i.e., the CEF, and (ii) a "left over" which is *orthogonal* to (i.e., uncorrelated with) any function of X_i.
- The CEF is a good summary of the relationship between Y_i and X_i for a number of reasons. First, and foremost, we are used to thinking of averages as providing a representative value for a random variable.

Fitting the Data

- More formally, the **CEF** is the "*best*" predictor of Y_i given X_i in the sense that it provides the "**best**" **fit** to the data. The CEF does that by solving a **M**inimum **M**ean **S**quared **E**rror (MMSE) *prediction problem*.
- Let $m(X_i)$ be any function of X_i. For example $b_0 + b_X X_i$.
- The CEF minimizes the Mean Square Error of the prediction of Y given X. In that case the CEF solves:

$$E[Y_i \mid X_i] = \arg\min \left[E\left(Y_i - m(X_i)\right)^2 \right] \tag{1}$$

- An example: consider the following linear function:

$$m(X_i) = b_0 + b_X X_i$$

- In this particular case the CEF solves:

$$E[Y_i \mid X_i] = \arg\min \left[\left(Y_i - b_0 - b_X X_i\right)^2 \right] \tag{2}$$

- We minimize the **MSE** by the choosing b_0 and b_X.

The ANOVA Property: The Between and within Groups Variation

- Last but not least, a property of the CEF, closely related to both the CEF decomposition and prediction properties, is the *Analysis-of-Variance* (ANOVA) theorem.
- The variance of Y_i can be decomposed to (i) the variances of the CEF, that is $E[Y_i \mid X_i]$ and (ii) the variance of the "left over" ε_i.
- The variance of Y_i, that is $V(Y_i)$, equals to:

$$V(Y_i) = V(E[Y_i \mid X_i]) + V(\varepsilon_i). \tag{3}$$

- The first term in the right hand side is variance of the predicted value of Y_i conditional on X_i. This reflects, for example, differences in mean earnings between workers with different levels of education.
- The second term reflects, for example, the variation in earnings among workers with the same level of schooling.

The Population Regression Function (PRF)

- A popular prediction of the population conditional means is the *Population Regression Function*.
- The *Population Regression Function* portraits the *linear relationship* between the Y and X in the population.
- Illustrated in the simple bivariate case where the regression vector includes only the single regressor, X, and a constant:

$$Y_i = b_0 + b_X X + \varepsilon_i. \tag{4}$$

- The parameters b_0 and b_X are the *constant* and the *slope* of the linear function respectively.

The PRF (cont.)

- The parameters b_0 and b_X are the *constant* and the *slope* of the linear function respectively.
- These parameters (b_0 and b_X) provide the *best fitting line* generated by minimizing expected squared errors, that is:

$$(b_0, b_X) = \arg\min \left[E \left(Y_i - b_0 - b_X X_i \right)^2 \right]. \qquad (5)$$

- The $b's$ that minimize (5) are known as the **O***rdinary* **L***east* **S***quares*, **OLS**, estimators.
- We dnote those by b_0^{OLS} and b_X^{OLS}.

The PRF (cont.)

- The parameters of the *Population Regression Function* decompose Y_i into two components:
- First, the average Y conditional on subject's X_i, that is $b_0 + b_X X_i$.
- Second, a "left over term" ε_i. This term is also known as the "error term". The error term is zero mean. The error term is also orthogonal / uncorrelated to X.
- Employed with the above the *Population Regression Function* exhibits the following linear form:

$$Y_i = b_0 + b_X X_i + \varepsilon_i. \tag{6}$$

- *Note* that b_X *does not* provide the *causal impact* of X on Y.

The PRF and the CEF

- The *PRF* is tightly linked to the *CEF*.
- In the case that the *conditional expectation function* of Y is a linear function of X then the *PRF* is also the *CEF*.
- Specifically, assuming that the CEF exhibits the following linear form:

$$E\left[Y_i \mid X_i\right] = b_0 + b_X X_i, \tag{7}$$

then the *PRF* equals to the *CEF*.

- We fit the data by solving:

$$\begin{aligned}\left(b_0^{OLS}, b_X^{OLS}\right) &= \arg\min E\left(Y_i - E\left[Y_i \mid X_i\right]\right)^2 \\ &= \arg\min\left[E\left(Y_i - b_0 - b_X X_i\right)^2\right].\end{aligned} \tag{8}$$

The PRF and the CEF (cont.)

- Note that the term in $(\cdot)$ is the residual ε_i:

$$\varepsilon_i = Y_i - b_0 - b_X X_i = Y_i - E\left[Y_i \mid X_i\right]$$

- Therefore we implicitly look for (b_0, b_X) that minimize the variance of the residual:

$$\begin{aligned} \left(b_0^{OLS}, b_X^{OLS}\right) &= \arg\min E\left(Y_i - E\left[Y_i \mid X_i\right]\right)^2 \quad (9) \\ &= \arg\min \left[E\left(\varepsilon_i\right)^2\right] = \arg\min \left[var\left(\varepsilon_i\right)\right]. \end{aligned}$$

- Where:

$$var\left(\varepsilon_i\right) = \sigma_\varepsilon^2 = E\left(\varepsilon_i\right)^2. \quad (10)$$

The Key Properties of the Population Regression Function

- The *PRF* provides the **BEST** prediction of Y as a linear function of X.
- The First Order Conditions imply that:

$$
\begin{aligned}
E(\varepsilon_i) &= 0, \\
Cov(\varepsilon_i, X_i) &= 0.
\end{aligned}
$$

- Therefore, by construction, the left-over term, the residual (ε_i), is **zero mean** and **uncorrelated** with X.
- Let's use these properties to recover the *bs*.

Recovering the Population Regression Parameters

- The covariance between Y_i and X_i, in the simple bivariate case where the regression, equals to:

$$Cov\left(Y_i, X_i\right) = Cov\left(\left(b_0 + b_X X_i + \varepsilon_i\right), X_i\right)$$

- Since X_i and ε_i are **uncorrelated by construction** then the covariance between Y_i and X_i equals to:

$$Cov\left(Y_i, X_i\right) = b_X Var\left(X_i\right)$$

- Therefore β_X equals to:

$$b_X^{OLS} = \frac{Cov\left(Y_i, X_i\right)}{Var\left(X_i\right)} \tag{11}$$

- Using the zero mean property we receive that β_0 equals to:

$$b_0^{OLS} = \bar{Y} - b_X \bar{X}. \tag{12}$$

Regression Coefficients as Conditional Means:

- Consider the lot hourly wages of workers in the financial sector and others sectors. Let's assume that the log hourly wage exhibits the following form:

$$Y_i = b_0 + b_F F_i + \varepsilon_i, \tag{13}$$

  - Y_i is the log hourly wage of worker i and
  - F_i is an indicator that equals 1 if person i works in the financial sector and 0 otherwise.

- What are the OLS regression coefficients b_0 and b_F?
- We already know that

$$b_F^{OLS} = \frac{Cov\left(Y_i, F_i\right)}{Var\left(F_i\right)} \tag{14}$$

- What is the variance of F_i? The variance of a binary variable is:

$$Var\left(F_i\right) = \bar{F} \cdot \left(1 - \bar{F}\right), \tag{15}$$

  where $\bar{F} = E\left(F_i\right)$ is the probability of a random workers to be working in the financial sector.

Regression Coefficients as Conditional Means (cont.):

- We already know that

$$b_F^{OLS} = \frac{Cov\left(Y_i, F_i\right)}{Var\left(F_i\right)} \tag{17}$$

- What is the variance of F_i? The variance of a binary variable is:

$$Var\left(F_i\right) = \bar{F} \cdot \left(1 - \bar{F}\right), \tag{18}$$

  where $\bar{F} = E\left(F_i\right)$ is the probability of a random workers to be working in the financial sector.

- What is the covariance of a binary and continuos variable? The covariance is:

$$Cov\left(Y_i, F_i\right) = E\left(Y_i \cdot F_i\right) - E\left(Y_i\right) \cdot E\left(F_i\right). \tag{19}$$

Regression Coefficients...(cont.):

- The covariance of a binary variable (F_i) and a continuous variable (Y_i) equals to:

$$Cov\,(Y_i, F_i) = \bar{F} \cdot Y_1 - [\bar{F} \cdot Y_1 + (1 - \bar{F}) \cdot Y_0] \cdot \bar{F}. \qquad (20)$$

  - $Y_0 = E\,(Y_i | F_i = 0)$
  - $Y_1 = E\,(Y_i | F_i = 1)$

- By substituting (15) and (20) into (14) we obtain that the **OLS** estimator for b_F^{OLS}, equals to the gap in mean outcome between groups:

$$b_F^{OLS} = \frac{\bar{F} \cdot (1 - \bar{F}) \cdot (Y_1 - Y_0)}{\bar{F} \cdot (1 - \bar{F})} = Y_1 - Y_0 \qquad (21)$$

Regression Coefficients...(cont.):

- What about the "constant" b_0^{OLS}? The constant equals to:

$$\begin{aligned} b_0^{OLS} &= \bar{F} Y_1 + (1 - \bar{F}) Y_0 - (Y_1 - Y_0) \bar{F} \qquad (22) \\ &= \bar{F} Y_1 + Y_0 - \bar{F} Y_0 - \bar{F} Y_1 + \bar{F} Y_0 = Y_0. \end{aligned}$$

- Hence, the **OLS** estimator for b_0, b_0^{OLS}, equals to mean outcome of the benchmark group.

The Multivariate Regression

- The *Regression Function* in the multivariate case, i.e., with more than one non-constant regressor, is:

$$Y_i = b_0 + b_{X_1} X_{1i} + b_{X_2} X_{2i} + ... + b_{X_K} X_{Ki} + \varepsilon_i, \tag{23}$$

  where X_1 to X_K is a vector of K explanatory variables.

- The coefficient b_{X_1} is the bivariate slope coefficient for the corresponding regressor, after "*partialling out*" all the other variables in the model.
- What does it mean?
- Naturally X_1 might be correlated with other explanatory variables $(X_2 ... X_K)$.

The Multivariate Regression (cont.)

- Therefore, let's decompose X_{1i} into two main components: (i) the part that is correlated with all other explanatory variables ;(ii) a *residual* that is **not correlated by construction** with **all other variables** $(\tilde{X}_{1i})$:

$$X_{1i} = \underbrace{a_0 + a_{X_2} X_{2i} + ... + a_{X_K} X_{Ki}}_{\Downarrow \atop \text{correlated with } X_2 - X_K} + \underbrace{\tilde{X}_{1i}.}_{\Downarrow \atop \text{residual}}$$

- In general the coefficient b_{X1} or any other slope equals to the covariance between Y_i and the "residual" $\tilde{X}_{1i}$ over the variance of $\tilde{X}_{1i}$:

$$b_{X1} = \frac{Cov\left(Y_i, \tilde{X}_{1i}\right)}{Var\left(\tilde{X}_{1i}\right)}. \tag{24}$$

- A particular case is when X_1 is the only explanatory variable. In that case $\tilde{X}_{1i} = X_{1i}$.

Estimating the PRF for Finance in a Wage Equation

- Use the NLSY79 data found in Moodle "MG4A4_NLSY79_2555.dta".
- Consider the following PRF:

$$Y_{it} = b_0 + b_F F_{it} + b_S S_{it} + b_A A_i + b_{E1} EXP_{it} + b_{E2} EXP_{it}^2 + \varepsilon_{it}. \quad (25)$$

- Where:
  - Y_{it} = the log hourly wage in year t (Wt)
  - F_{it} = a binary indicator that equals to 1 if person i works in finance in year t (FINANCE).
  - S_{it} = years of schooling completed by person i (MAXsclyrs);
  - A_i = AFQT score (percentile) of person i (afqt06);
  - EXP_{it} = potential labor market experience (exp1) ;
  - EXP_{it}^2 = potential labor market experience square (exp2).

An Example using the NLSY79.

- Let's estimate two models:
- The first model:

$$Y_{it} = a_0 + a_F F_{it} + \varepsilon_{it}. \tag{26}$$

- The second model:

$$Y_{it} = b_0 + b_F F_{it} + b_S S_{it} + b_A AFQT_i + b_{E1} EXP_{it} + b_{E2} EXP_{it}^2 + \varepsilon_{it}. \tag{27}$$

- Note that ε_{it} in the first and the second equations are not necessarily identical.
- The corresponding STATA codes are:

```
reg _Wit    _FINANCEit
reg _Wit    _FINANCEit    _Sit _AFQTi _EXP1it _EXP2it
```

- What is the difference between a_F and b_F? Estimates are found in the next page.

BEFORE WE RUN A REGRESSION DES AND SUM THE DATA

- Use the 'des' command

```
> id year _SEXi _RACEi _SEXi _AFQTi _Sit _EXP1it _EXP2it
> _FINANCEit _TYPEWRKRit _FTFYit _HRSWRKit _WAGERit _HWRit _Wit ;

              storage   display    value
variable name   type    format     label      variable label
-------------------------------------------------------------------------------------
id              float   %9.0g                 ID# (1-12686) 79
year            float   %9.0g
_SEXi           byte    %9.0g      vlR0214800
                                              SEX OF R 79
_RACEi          byte    %23.0g     vlR0214700
                                              RACL/ETHNIC COHORT /SCRNR 79
_SEXi           byte    %9.0g      vlR0214800
                                              SEX OF R 79
_AFQTi          float   %9.0g                 afqt79
_Sit            byte    %9.0g                 years of schooling completed
_EXP1it         float   %9.0g                 experience in the labor market
_EXP2it         float   %9.0g                 experience square in the labor market
_FINANCEit      byte    %9.0g                 Finance
_TYPEWRKRit     byte    %26.0g     vTYPEWRKRt
                                              TYPE OF WORKER IN CURRENT YEAR, 25+
_FTFYit         byte    %9.0g                 1 IF WKS>=50 & HRS P WK>=35
_HRSWRKit       int     %9.0g      vlR0215710
                                              # OF HRS WRKD IN P-C YR 79
_WAGERit        float   %9.0g                 TOT INC WAGES AND SALRY CPI ADJUSTED 2010
_HWRit          float   %9.0g                 HOURLY WAGE CPI ADJUSTED 2010
_Wit            float   %9.0g                 log hourly wage cpi adjusted (2010)
```

BEFORE WE RUN A REGRESSION DES AND SUM THE DATA

- Use the 'sum' command

```
> sum
> id year _SEXi _RACEi _SEXi _AFQTi _Sit _EXP1it _EXP2it
> _FINANCEit _TYPEWRKRit _FTFYit _HRSWRKit _WAGERit _HWRit _Wit ;

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
          id |     21,112    4069.948    3220.139          5      12686
        year |     21,112    1995.619    8.129035       1982       2012
       _SEXi |     21,112           1           0          1          1
      _RACEi |     21,112           3           0          3          3
       _SEXi |     21,112           1           0          1          1
-------------+---------------------------------------------------------
      _AFQTi |     21,112     .573526    .2794696        .01        .99
        _Sit |     21,112    13.86714    2.596317          6         20
     _EXP1it |     21,112    15.42947    8.188321          0         41
     _EXP2it |     21,112     305.114    297.3525          0       1681
  _FINANCEit |     21,112    .0425351    .2018112          0          1
-------------+---------------------------------------------------------
 _TYPEWRKRit |     21,112           1           0          1          1
     _FTFYit |     21,112           1           0          1          1
   _HRSWRKit |     21,112    2467.709    512.9577       2000       8736
    _WAGERit |     21,112    60703.17    47044.19   70.34064     333309
      _HWRit |     21,112    24.81519    18.17846   .0262191   160.2447
-------------+---------------------------------------------------------
        _Wit |     21,112    3.029333    .5991676  -3.641266   5.076702
```

Estimate the Regression Model using OLS

```
> reg _Wit _FINANCEit, robust  ;

Linear regression                               Number of obs     =     21,112
                                                F(1, 21110)       =     334.25
                                                Prob > F          =     0.0000
                                                R-squared         =     0.0177
                                                Root MSE          =     .59384

------------------------------------------------------------------------------
             |               Robust
        _Wit |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
  _FINANCEit |   .3955093   .0216333    18.28   0.000     .3531063    .4379123
       _cons |    3.01251   .0041631   723.62   0.000      3.00435    3.02067
------------------------------------------------------------------------------

. estimates store OLS1, title((CRUDE));

. reg _Wit _FINANCEit _AFQTi _Sit _EXP1it _EXP2it, robust ;

Linear regression                               Number of obs     =     21,112
                                                F(5, 21106)       =    1309.51
                                                Prob > F          =     0.0000
                                                R-squared         =     0.2807
                                                Root MSE          =     .50822

------------------------------------------------------------------------------
             |               Robust
        _Wit |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
  _FINANCEit |   .1951821   .0178065    10.96   0.000     .1602801    .2300842
      _AFQTi |   .4582582    .016138    28.40   0.000     .4266266    .4898898
```

Report the Regression Coefficients

	OLS1	OLS2
_FINANCEit	**0.396*** **	**0.195*** **
	(0.022)	(0.018)
_AFQTi		0.458***
		(0.016)
_Sit		0.079***
		(0.002)
_EXP1it		0.058***
		(0.002)
_EXP2it		-0.001***
		(0.000)
Constant	3.013***	1.082***
	(0.004)	(0.031)
Observations	21112	21112
R-square	0.018	0.281

The STATA code for the table above.

```
reg _Wit _FINANCEit, robust ;
estimates store OLS1, title((CRUDE));
reg _Wit _FINANCEit _AFQTi _Sit _EXP1it _EXP2it, robust ;
estimates store OLS2, title((RESID1));
estout OLS1 OLS2 using "$RESULTS/MG4A4 REGRESSION TOOL
NLSY79.txt",
replace keep
(_cons _FINANCEit _AFQTi _Sit _EXP1it _EXP2it)
cells(b(star fmt(3)) se(par fmt (3)))
varlabels(_cons Constant)
stats(N r2 , fmt(0 3 0) labels(Observations R-square))
starlevels(* 0.10 ** 0.05 *** 0.01)
prehead("Table 434" "REGRESSION TOOL")
posthead("")
prefoot("")
postfoot(""
"Notes."
"TBA"
```

Full vs. Partial Relationship: an Example

- Now let's estimate the following two equations:

1. First let's regress FINANCE on all other explanatory variables / Right Hand Side (RHS) variables:

$$F_{it} = d_0 + d_S S_{it} + d_A A_i + d_{E1} EXP_{it} + d_{E2} EXP_{it}^2 + \tilde{F}_{it}. \qquad (28)$$

2. Second, using the **estimated residual** $\tilde{F}_{it}$ estimate the following equation:

$$Y_{it} = c_0 + c_F \tilde{F}_{it} + \varepsilon_{it}. \qquad (29)$$

- The corresponding STATA codes and results are found in the next page.
- Note that according the regression coefficient c_F measures the partial relationship between log hourly wages and workers' industry (finance) conditional on years of school completed, AFQT scores and potential experience in the population sample the PRF Explain (to yourself) the exercise.

Does it Work? If so, what do we learn from that?

```
> reg _FINANCEit _AFQTi _Sit _EXP1it _EXP2it, robust ;

Linear regression                               Number of obs     =     21,112
                                                F(4, 21107)       =     132.56
                                                Prob > F          =     0.0000
                                                R-squared         =     0.0253
                                                Root MSE          =     .19926

------------------------------------------------------------------------------
             |               Robust
  _FINANCEit |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      _AFQTi |   .0571408   .0054779    10.43   0.000     .0464036    .0678779
        _Sit |   .0076226    .000669    11.39   0.000     .0063113    .0089339
     _EXP1it |  -3.59e-06    .000805    -0.00   0.996    -.0015815    .0015743
     _EXP2it |   2.90e-06   .0000206     0.14   0.888    -.0000374    .0000432
       _cons |  -.0967689   .0114747    -8.43   0.000    -.1192602   -.0742776
------------------------------------------------------------------------------

. predict _FINANCERit, resid ;

. reg _Wit _FINANCERit, robust  ;

Linear regression                               Number of obs     =     21,112
                                                F(1, 21110)       =      72.38
                                                Prob > F          =     0.0000
                                                R-squared         =     0.0042
                                                Root MSE          =     .59792

------------------------------------------------------------------------------
             |               Robust
        _Wit |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
 _FINANCERit |   .1951821   .0229414     8.51   0.000     .1502152     .240149
       _cons |   3.029333   .0041151   736.16   0.000     3.021267    3.037398
------------------------------------------------------------------------------
```

Regression Coefficients and Treatment Effects

- The note above shows that the **OLS** regression provides the **best** *linear approximation* to the **CEF**.
- This tool, however, **does not** *necessarily* help us reaching our deeper goal – estimating the causal impact of X on Y.
- In our leading example – the finance wage premium – the regression coefficient on finance provides the crude (or the residual) estimated gap in mean log hourly wages between those who work in the financial industry $(F_i = 1)$ and others $(F_i = 0)$.
- The **univariate OLS** estimates the gap in mean crude log hourly wages.
- When we estimate the following equation:

$$Y_{it} = a_0 + a_F F_{it} + \varepsilon_{it}, \tag{30}$$

we estimate the following parameter:

$$a_F^{OLS} = Y_1 - Y_0 = E[Y_{it} \mid F_{it} = 1] - E[Y_{it} \mid F_{it} = 0] \tag{31}$$

Regression and Treatment Effects (cont.)

- The **multivariate OLS** estimates the gap in mean residual log hourly wages.
- When we estimate the following equation:

$$Y_{it} = b_0 + b_F F_{it} + b_S S_{it} + b_A AFQT_i + b_{E1} EXP_{it} + b_{E2} EXP_{it}^2 + \varepsilon_{it}. \tag{32}$$

we estimate the following parameter:

$$b_F^{OLS} = Y_{1,X} - Y_{0,X} = E\left[Y_{it} \mid X_{it}, F_{it} = 1\right] - E\left[Y_{it} \mid X_{it}, F_{it} = 0\right] \tag{33}$$

where Y_R stands for the residualized Y.

- Question: when can we think of a regression coefficient as approximating the *causal effect* that can be revealed in an ideal experiment?
- See next lectures to learn more about this key question.