

## Economics 522A, Spring 2007

### Lecture Note 5: Multivariate Data and Least Squares Fit

This material is based on Ruud, Ch.1. The data set and related material can be obtained online at: <http://elsa.berkeley.edu/~ruud/cet/index.html>.

#### 1. CPS Data

CPS: U.S. Current Population Survey (Ruud data from March 1995)

Summary Statistics: see Ruud, Table 1.1.

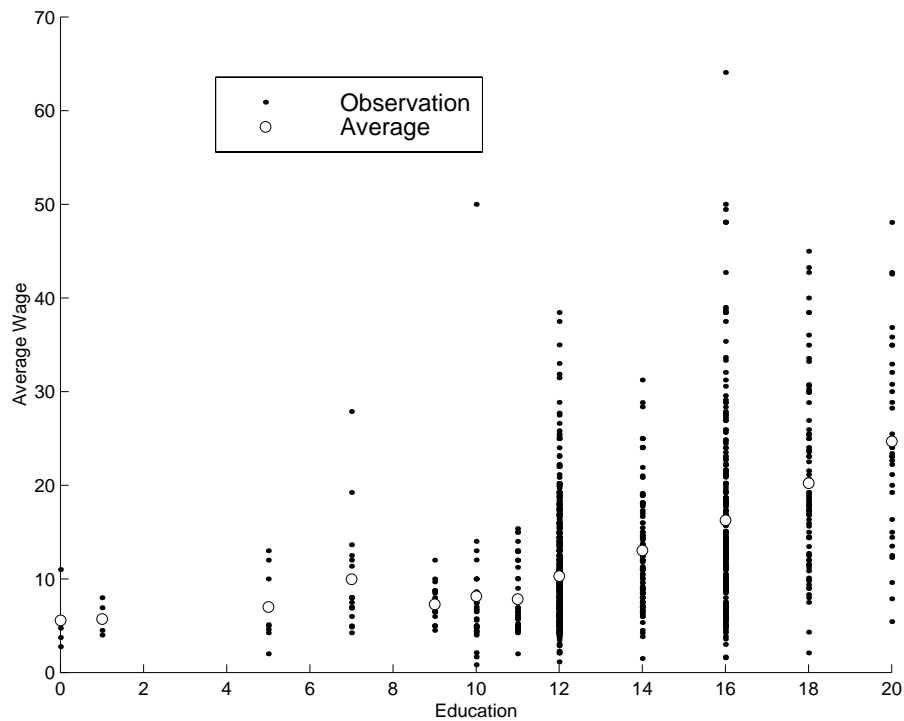
Good practice for empirical work:

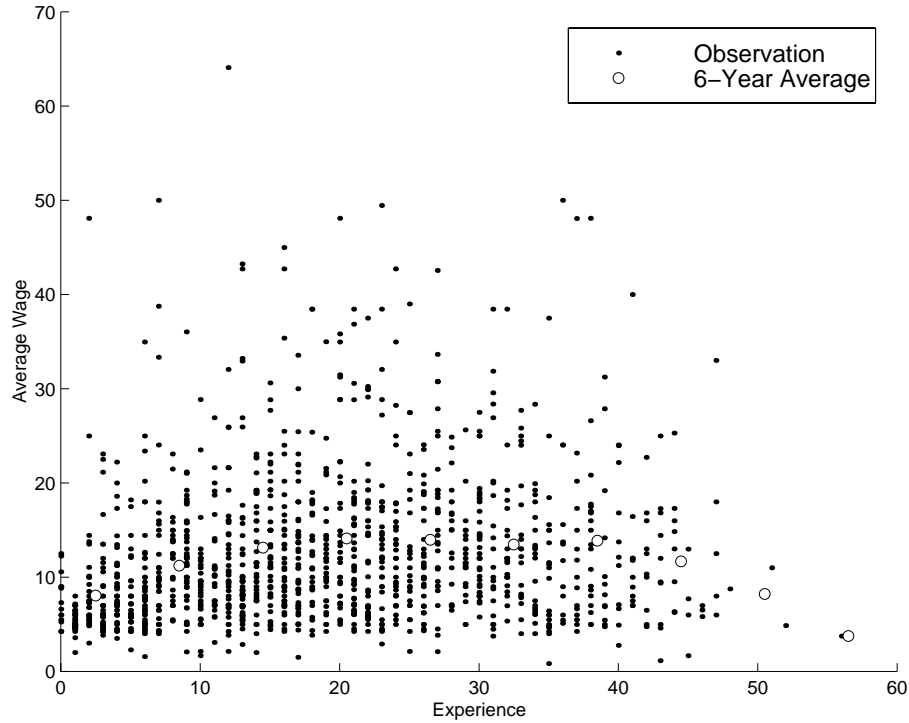
- Explain data set: population from which data are drawn, how the sampling was done, sample selection rules, nonresponse.
- Explain definitions of key variables; be careful about how variables are measured.
- Be careful about units of measurement, top-coding.
- Give summary statistics – variable name, sample mean, sample standard deviation, minimum and maximum values in sample.
- Perform sanity checks: do the numbers make sense??
- Compare summary statistics to aggregate figures.

#### 2. Explore data - cross-tabulations and correlations.

See Tables 1.2-1.5

#### 3. Explore data - plots (reproduced from Ruud, Figures 1.1-1.2)





#### 4. Samples Averages and Least Squares Fits

Sample average:

$$\bar{y}_n = \frac{1}{n} \sum_{i=1}^n y_i.$$

This is the solution to the minimization problem:

$$\min_a \sum_{i=1}^n (y_i - a)^2.$$

Here, we have a variable  $y_i$  which we want to relate to a number of other variables  $x_{i1}, x_{i2}, \dots, x_{ik}$ . We focus on linear functions of the  $x$  variables:

$$\beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_k x_{ik}.$$

Usually, we set  $x_{i1} = 1$ , so that the first term is a linear intercept term. Then we want to find values for the  $\beta_1, \dots, \beta_k$  to get a good approximation to the  $y_i$ :

$$y_i \approx \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_k x_{ik}.$$

Typically, we will not get a perfect fit, so we try to find the values that minimize the sum of squared differences between  $y_i$  and the linear function:

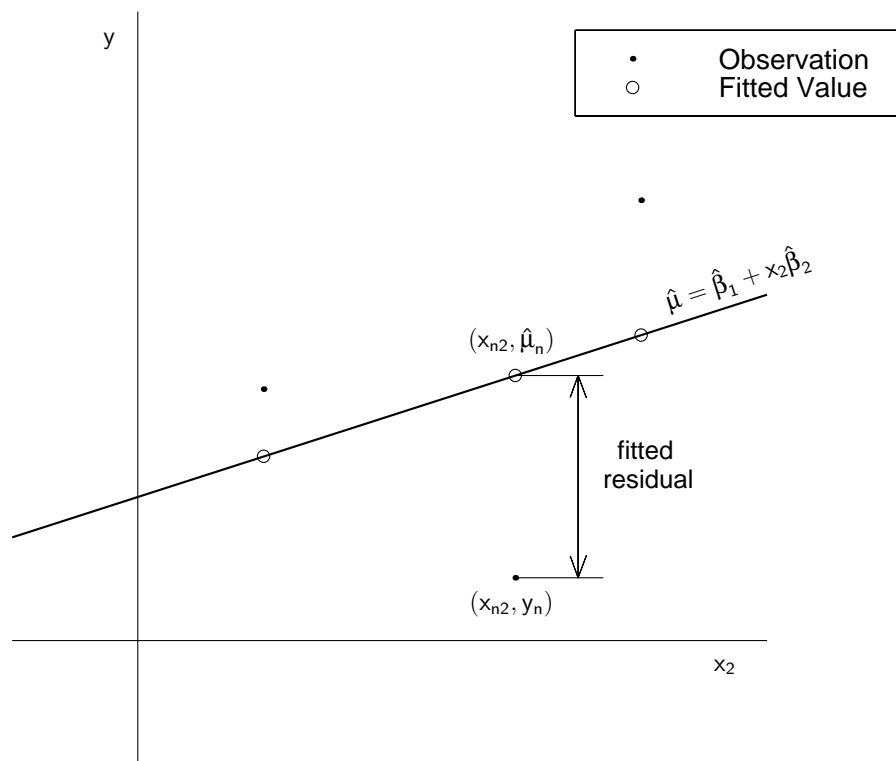
$$\min_{\beta_1, \dots, \beta_k} \sum_{i=1}^n (y_i - \beta_1 x_{i1} - \beta_2 x_{i2} - \dots - \beta_k x_{ik})^2.$$

We denote the solution by

$$\hat{\beta}_1, \hat{\beta}_2, \dots, \hat{\beta}_k.$$

This is called the ordinary least squares (OLS) fit.

Graphical representation (Ruud, Figure 1.3):



## 5. Logarithms and other transforms.

For some variables like wages, we often work with the (natural) logarithm of the variable, instead of the variable itself. Let  $w_i$  be the wage for person  $i$ , and let  $y_i = \log w_i$ . Suppose we approximate:

$$y_i = \log w_i = \beta_1 x_{i1} + \dots + \beta_k x_{ik} + \epsilon_i$$

(here,  $\epsilon_i$  should be viewed as the residual from the linear approximation,  $\epsilon = y_i - \beta_1 x_{i1} - \dots - \beta_k x_{ik}$ .)

This is equivalent to:

$$\begin{aligned} w_i &= \exp(\beta_1 x_{i1} + \dots + \beta_k x_{ik} + \epsilon_i). \\ &= \left(e^{\beta_1}\right)^{x_{i1}} \times \dots \times \left(e^{\beta_k}\right)^{x_{ik}} \times e^{\epsilon_i}. \end{aligned}$$

Also, when  $\beta_j$  is not too far from 0,

$$e^{\beta_j} \approx 1 + \beta_j.$$

Therefore,

$$w_i \approx (1 + \beta_1)^{x_{i1}} \times (1 + \beta_2)^{x_{i2}} \times \cdots \times (1 + \beta_k)^{x_{ik}} \times e^{\epsilon_i}.$$

**So we can interpret the  $\hat{\beta}_j$  as the approximate percentage increase in wage associated with a one unit increase in  $x_{ij}$ , holding fixed the other variables.**

## 6. OLS via Linear Algebra

Recall that OLS solves the minimization problem:

$$\min_{\beta_1, \dots, \beta_k} \sum_{i=1}^n (y_i - \beta_1 x_{i1} - \beta_2 x_{i2} - \cdots - \beta_k x_{ik})^2.$$

We can do this a bit more compactly using matrix notation. Let

$$y = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}, \quad X = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1k} \\ x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & \cdots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nk} \end{bmatrix}, \quad \beta = \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_k \end{pmatrix}.$$

Then

$$X\beta = \begin{bmatrix} \sum_{j=1}^k \beta_j x_{1j} \\ \vdots \\ \sum_{j=1}^k \beta_j x_{nj} \end{bmatrix},$$

So that

$$(y - X\beta)'(y - X\beta) = \sum_{i=1}^n (y_i - \beta_1 x_{i1} - \beta_2 x_{i2} - \cdots - \beta_k x_{ik})^2.$$

The OLS minimization problem can be rewritten as

$$\min_{\beta} (y - X\beta)'(y - X\beta).$$

The solution can be calculated explicitly using matrix calculus:

$$\hat{\beta} = (X'X)^{-1}X'y,$$

provided that the inverse  $(X'X)^{-1}$  exists.