Week 1. OLS Linear Regression

I. Simple Regression

1. $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$
   $\hat{y}_i = \beta_0 + \beta_1 x_i$

   where
   y_i: dependent variable of individual i
   x_i: independent variable of individual i
   β_0: intercept, the predicted (or expected) value of y when $x = 0$.
   β_1: slope, the predicted (or expected) change in y when x changes by 1 unit, $\frac{\partial y}{\partial x} = b_1$

   ε_i: residual; error term; disturbance $(= \{y_i - \hat{y}_i\} = \{y_i - (\beta_0 + \beta_1 x_i)\})$
   $\hat{y}_i$: the predicted (or expected) value of y for individual i

2. How to find the best fitting line:

   (a) **The Ordinary Least-squares Methods (OLS)** is a technique for fitting the "best" straight line to the sample of x, y observations. It involves minimizing the sum of the squered (vertical) deviations of points from the line:

   $$\text{Minimize} \sum (y_i - \hat{y}_i)^2$$

   (b) $\hat{\beta}_1 = \frac{SS(xy)}{SS(x)} = \frac{cov(x,y)}{\sigma_x^2}$

   where
   $SS(xy) = \sum (x_i - \bar{x})(y_i - \bar{y})$
   $SS(x) = \sum (x_i - \bar{x})^2$
   $cov(x,y) = \frac{1}{n-1} \sum (x_i - \bar{x})(y_i - \bar{y}) =$ covariance between x and y

   (c) $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$

3. Test of Goodness of Fit

   (a) Total variation in y = Explained variation in y + Residual variation in y
   Total sum of squares = Regression sum of squares + Error sum of squares
   $TSS = ESS + RSS$
   $\sum (y_i - \bar{y})^2 = \sum (\hat{y}_i - \bar{y})^2 + \sum (y_i - \hat{y}_i)^2$

   (b) Divided both sides by TSS,
   $$1 = \frac{RSS}{TSS} + \frac{ESS}{TSS}$$

   The coefficient of determination, or R^2 is then defined as the proportion of the total variation in y explained by the regression of y on x:

   $$R^2 = \frac{RSS}{TSS} = 1 - \frac{ESS}{TSS} = 1 - \frac{\sum \varepsilon_i^2}{\sum (y_i - \bar{y})^2} = 1 - \frac{\sum (y_i - \hat{y}_i)^2}{\sum (y_i - \bar{y})^2} = 1 - \frac{\sum (y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i))^2}{\sum (y_i - \bar{y})^2}$$

(c) Like any other proportions, R^2 ranges from 0 to 1.

4. $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$
   $y_i = \hat{\beta}_0 + \hat{\beta}_1 x_i + e_i$

   where
   β_0: parameter (=true value for a population) of intercept
   β_1: parameter (=true value for a population) of slope
   $\hat{\beta}_0$: estimated parameter (=statistic computed based on a sample) of intercept
   $\hat{\beta}_1$: estiamted parameter (=statistic computed based on a sample) of slope

   What we want to know are β_0 and β_1, but what we get from a sample are $\hat{\beta}_0$ and $\hat{\beta}_1$.

5. Test of significance of parameter estimates.

   We test $H_0 : \beta_0 = 0$ and $H_0 : \beta_1 = 0$. To do so, the variances of $\hat{\beta}_0$ and $\hat{\beta}_1$ are required.

$$Var(\hat{\beta}_0) = \sigma_\varepsilon^2 \left(\frac{\sum x_i^2}{n \sum (x_i - \bar{x})^2} \right)$$
$$Var(\hat{\beta}_1) = \sigma_\varepsilon^2 \left(\frac{1}{\sum (x_i - \bar{x})^2} \right)$$

   Since, σ_ε^2 is unknown (sigma is unknown), the residual variance (s_e^2) is used as an estiamte of σ_ε^2.

$$\hat{\sigma}_\varepsilon^2 = s_e^2 = \frac{\sum e^2}{n - k}$$

   where k represents the number of parameter estimates. For simple regression, k is 2 (slope and intercept). Therefore, the estimates of the variance of $\hat{\beta}_0$ and $\hat{\beta}_1$ are given by (Don't ask. Not many sociologists understand how these formulas are driven.):

$$s_{\hat{\beta}_0}^2 = \left(\frac{\sum e_i^2}{n - k} \right) \left(\frac{\sum x_i^2}{n \sum (x_i - \bar{x})^2} \right)$$
$$s_{\hat{\beta}_1}^2 = \left(\frac{\sum e_i^2}{n - k} \right) \left(\frac{1}{\sum (x_i - \bar{x})^2} \right)$$

   Square roots of $s_{\hat{\beta}_0}^2$ and $s_{\hat{\beta}_1}^2$ are the standard errors of the estimate. Using these standard errors, we can do t-test.

$$t_{\hat{\beta}_0} = \frac{\hat{\beta}_0 - \beta_0}{s_{\hat{\beta}_0}} = \frac{\hat{\beta}_0}{s_{\hat{\beta}_0}}$$
$$t_{\hat{\beta}_1} = \frac{\hat{\beta}_1 - \beta_1}{s_{\hat{\beta}_1}} = \frac{\hat{\beta}_1}{s_{\hat{\beta}_1}}$$

   Because we test $H_0 : \beta_0 = 0$ and $H_0 : \beta_1 = 0$, β_0 and β_1 are zero in computing t.

6. Properties of OLS Estimates

   (a) OLS estimates are *best linear unbiased estimates (BLUE).*

   (b) Lack of biase means

   $$E(\hat{\beta}) = \beta$$

   So that Bias $= E(\hat{\beta}) - \beta$

   (c) *Best unbiased* or *efficient* means smallest variance. This is known as the Gauss-Markov theorem and represents the most important justification for using OLS.

   (d) Sometiems, a researcher may want to trade off some bias for a possibly smaller variance and minimize the mean square error, MSE:

   $$MSE(\hat{\beta}) = E(\hat{\beta} - \beta)^2 = var(\hat{\beta}) + (\text{bias } \hat{\beta})^2$$

   (e) $\sum \varepsilon_i = 0$

   (f) The best fit line always goes through $\bar{x}$ and $\bar{y}$.

   (g) OLS is a conditional mean function. $E(y|x) = \beta_0 + \beta_1 x$

   (h) For simple regression, r-squred is equal to the square of correlation coefficient, r (Only for simple regression).

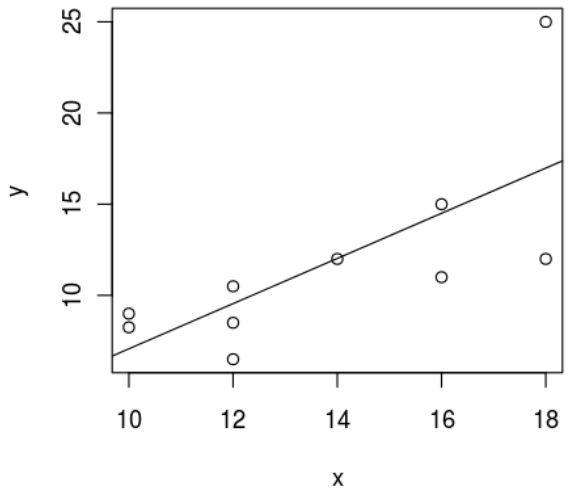
7. An example

Table 1. Employee Hourly Wages and Years of Schooling

Employee ID	Hourly Wage (y)	Years of Schooling (x)
1	8.50	12
2	12.00	14
3	9.00	10
4	10.50	12
5	11.00	16
6	15.00	16
7	25.00	18
8	12.00	18
9	6.50	12
10	8.25	10

Stata Result

```
input id wage sch
 1  8.50 12
 2 12.00 14
 3  9.00 10
 4 10.50 12
 5 11.00 16
 6 15.00 16
 7 25.00 18
 8 12.00 18
 9  6.50 12
10  8.25 10
end
```




```
. twoway scatter wage sch
. twoway (scatter wage sch)(lfit wage sch)

. pwcorr wage sch, sig

             |     wage      sch
-------------+------------------
        wage |   1.0000
             |
             |
         sch |   0.7216   1.0000
             |   0.0185
             |
.
. reg wage sch

      Source |       SS       df       MS              Number of obs =      10
-------------+------------------------------           F(  1,     8) =    8.69
       Model |  128.260795     1  128.260795           Prob > F      =  0.0185
    Residual |  118.045455     8  14.7556818           R-squared     =  0.5207
-------------+------------------------------           Adj R-squared =  0.4608
       Total |   246.30625     9  27.3673611           Root MSE      =  3.8413

------------------------------------------------------------------------------
        wage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         sch |   1.238636    .420123     2.95   0.018     .2698309    2.207442
       _cons |  -5.318182   5.923586    -0.90   0.396      -18.978    8.341632
------------------------------------------------------------------------------
```

The result can be summarized as: $wage = -.3181 + 1.2386sch + e$
or you can write as: $\hat{wage} = -.3181 + 1.2386sch$, omitting the residual term, e.

The effect of schooling is statistically significant at α=.05. It indicates that as years of schooling increases by 1 year, the expected change in hourly wage is \$1.24. When years of schooling is zero, the expected wage is \$-5.32 according to the estimated results. However, this result does not intuitively make sense.

8. Appendix:

Mathematical Proof that:
$b_1 = \frac{SS(xy)}{SS(x)} = \frac{cov(x,y)}{\sigma_x^2}$
$b_0 = \bar{y} - \hat{\beta}_1 \bar{x}$

To get the best fit line, we need to minimize $\sum e^2$. That is to minimize $\sum(y - (b_0 + b_1 x))^2$.

Let's say $S(b_0, b_1) = \sum(y - (b_0 + b_1 x))^2$

(a) How to get b_0:
We need to have a partial derivative of b_0:

$$\frac{\partial S(b_0, b_1)}{\partial b_0} = -2\sum(y - (b_0 + b_1 x))$$

Setting the partial derivative to zero:

$$-2\sum(y - (b_0 + b_1 x)) = 0$$

$$b_0 = \bar{y} - b_1 \bar{x}$$

(b) How to get b_1:
Likewise, we need to have a partial derivative of b_1:

$$\frac{\partial S(b_0, b_1)}{\partial b_1} = -2\sum x(y - (b_0 + b_1 x))$$

Setting the partial derivative to zero:

$$-2\sum x(y - (b_0 + b_1 x)) = 0$$

$$\sum xy - \sum x b_0 - \sum b_1 x^2 = 0$$

$$\sum xy - \sum x(\bar{y} - b_1 \bar{x}) - \sum b_1 x^2 = 0$$

Solve the above *normal equations*:

$$b_1 = \frac{\sum xy - \frac{\sum x \sum y}{n}}{\sum x^2 - \frac{(\sum x)^2}{n}} = \frac{SS(xy)}{SS(x)}$$

II. Multiple Regression

1. Three variable (i.e., 2 independent variables) linear regression model:

   $y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \varepsilon_i$

   where
   x_{1i} and x_{2i} should NOT be exactly linearly associated.

2. How to find the best fitting line:

   (a) OLS parameter estimates can be obtained by minimizing the sum of e^2.

   $$\sum e^2 = \sum (y_i - \hat{y}_i)^2 = \sum (y_i - \{\hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \hat{\beta}_2 x_{2i}\})^2$$

   (b) By using OLS, we can obtain:

   $$\hat{\beta}_1 = \frac{SS(x_1 y)SS(x_2) - SS(x_2 y)SS(x_1 x_2)}{SS(x_1)SS(x_2) - \{SS(x_1 x_2)\}^2} = \left(\frac{s_y}{s_{x_1}}\right)\left(\frac{r_{yx_1} - r_{yx_2} r_{x_1 x_2}}{1 - r^2_{x_1 x_2}}\right)$$

   $$\hat{\beta}_2 = \frac{SS(x_2 y)SS(x_1) - SS(x_1 y)SS(x_1 x_2)}{SS(x_1)SS(x_2) - \{SS(x_1 x_2)\}^2} = \left(\frac{s_y}{s_{x_2}}\right)\left(\frac{r_{yx_2} - r_{yx_1} r_{x_1 x_2}}{1 - r^2_{x_1 x_2}}\right)$$

   $$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}_1 - \hat{\beta}_2 \bar{x}_2$$

   where,
   $SS(x_1 y) = \sum (x_1 - \bar{x}_1)(y - \bar{y})$
   $SS(x_2 y) = \sum (x_2 - \bar{x}_2)(y - \bar{y})$
   $SS(x_1 x_2) = \sum (x_1 - \bar{x}_1)(x_2 - \bar{x}_2)$
   $SS(x_1) = \sum (x_1 - \bar{x}_1)^2$
   $SS(x_2) = \sum (x_2 - \bar{x}_2)^2$

3. How to interpret the estimated coefficients:

   (a) $\hat{\beta}_0$: the expected y when both x_1 and x_2 are zero.

   (b) $\hat{\beta}_1$: the expected change in y when x_1 changes by 1 unit, holding x_2 constant. Or you can say that the expected change in y when x_1 changes by 1 unit, *ceteris paribus*. Or you can say that the expected change in y when x_1 changes by 1 unit, other things being equal.

   $$\frac{\Delta y}{\Delta x_1} = \beta_1$$

   (c) $\hat{\beta}_2$: the expected change in y when x_2 changes by 1 unit, holding x_1 constant. Or you can say that the expected change in y when x_2 changes by 1 unit, *ceteris paribus*. Or you can say that the expected change in y when x_2 changes by 1 unit, other things being equal.

   $$\frac{\Delta y}{\Delta x_2} = \beta_2$$

4. Test of significance of parameter estimates.

   (a) $\hat{\beta}$ differs from sample to sample. $E(\hat{\beta}) = \beta$ means that on average $\hat{\beta}$ (= estimated parameter (i.e., statistic) based on sample) equal to β (true population parameter). This is similar to the idea that $E(\bar{y}) = \mu_{\bar{y}} = \mu_y$ (recall Central Limit Theorem). The distribution of of $\hat{\beta}$ refers to the sampling distribution of $\hat{\beta}$, that is, how $\hat{\beta}$ varies from sample to sample.

   (b) The variance of $\hat{\beta}$ can be computed as follows:

$$Var(\hat{\beta}_1) = \sigma_\varepsilon^2 \left(\frac{\sum x_2^2}{\sum x_1^2 \sum x_2^2 - (\sum x_1 x_2)^2} \right)$$

$$Var(\hat{\beta}_2) = \sigma_\varepsilon^2 \left(\frac{\sum x_1^2}{\sum x_1^2 \sum x_2^2 - (\sum x_1 x_2)^2} \right)$$

   (c) σ_ε^2 is unknown, therefore, the residual variance $\hat{\sigma}_\varepsilon^2 = s_e^2 = \frac{\sum e_i^2}{n-k}$ is used.

$$s_{\hat{\beta}_1}^2 = \left(\frac{\sum e_i^2}{n-k} \right) \left(\frac{\sum x_2^2}{\sum x_1^2 \sum x_2^2 - (\sum x_1 x_2)^2} \right) = \left(\frac{\sum e_i^2}{n-k} \right) \left(\frac{1}{\sum x_1^2 \cdot (1 - r_{x_1 x_2}^2)} \right)$$

$$s_{\hat{\beta}_2}^2 = \left(\frac{\sum e_i^2}{n-k} \right) \left(\frac{\sum x_1^2}{\sum x_1^2 \sum x_2^2 - (\sum x_1 x_2)^2} \right) = \left(\frac{\sum e_i^2}{n-k} \right) \left(\frac{1}{\sum x_2^2 \cdot (1 - r_{x_1 x_2}^2)} \right)$$

   where $s_e^2 = \frac{\sum e_i^2}{n-k}$ is the variance of the observed error terms. It is sometimes called MSE or the mean squared error. k = the number of independent variables plus 1. Thus, for a simple regression, $k = 2$, and for a multiple regression with 2 independent variables, $k = 3$.

   (d) The significance of coefficients estimated is tested with t-test. We test whether $H_0 : \beta_1 = 0$ and $H_0 : \beta_2 = 0$.

$$t_{\hat{\beta}_1} = \frac{\hat{\beta}_1 - \beta_1}{s_{\hat{\beta}_1}} = \frac{\hat{\beta}_1}{s_{\hat{\beta}_1}}$$

$$t_{\hat{\beta}_2} = \frac{\hat{\beta}_2 - \beta_2}{s_{\hat{\beta}_2}} = \frac{\hat{\beta}_2}{s_{\hat{\beta}_2}}$$

   Because we test $H_0 : \beta_1 = 0$ and $H_0 : \beta_2 = 0$, β_1 and β_2 are zero in computing t. That is, we test whether the coefficients estimated (= $\hat{\beta}_1$ and $\hat{\beta}_2$) are statistically different from zero or not.

5. $(1 - \alpha)$ confidence interval for the slope

$$\hat{\beta}_1 \pm t_{(\alpha/2)} s_{\hat{\beta}_1}$$

   where $t_{(\alpha/2)}$ is a critical value based on $d.f. = N - k$

   Note that 5 and 4 are basically the same as the confidence interval for a mean and the hypothesis test of a mean that we discussed previously.

6. The coefficients of multiple determination, R^2

   R^2 is defined as the proportion of the total variation in y "explained" by the multiple regression of y on x_1 and x_2.

$$R^2 = 1 - \frac{\sum \varepsilon_i^2}{\sum (y_i - \bar{y})^2} = 1 - \frac{\sum (y_i - \hat{y}_i)^2}{\sum (y_i - \bar{y})^2} = 1 - \frac{\sum (y_i - \{\hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \hat{\beta}_2 x_{2i}\})^2}{\sum (y_i - \bar{y})^2}$$

$$= \frac{\hat{\beta}_1 \sum (y_i - \bar{y})(x_{1i} - \bar{x}_1) + \hat{\beta}_2 \sum (y_i - \bar{y})(x_{2i} - \bar{x}_2)}{\sum (y_i - \bar{y})^2}$$

   Since the inclusion of additional independent or explanatory variable is likely to increase the $RSS = \sum \hat{y}_i^2$ for the same $TSS = \sum y_i^2$, thus R^2 increases. To factor in the reduction in the degree of freedom as additional independent or explanatory variables are added, the *adjusted* R^2 is computed:

$$\text{adjusted } R^2 = 1 - (1 - R^2)\frac{n-1}{n-k}$$

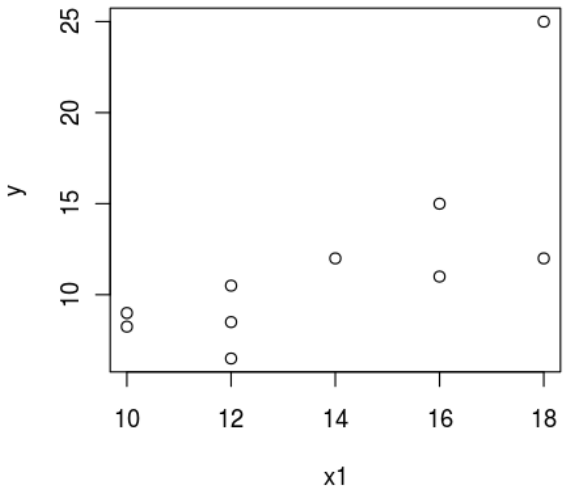
   where n is the number of observations (i.e., total sample size), and k is the number of parameters estimated.

   Therefore, the maximum number of possible independent variables is $n - 1$.

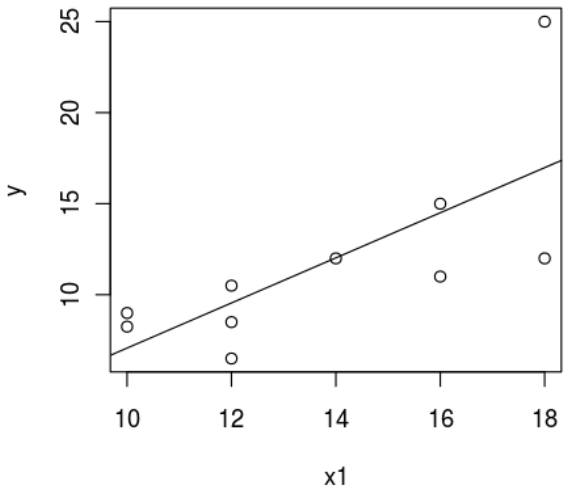
7. $E(y_i|x_1, x_2) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$

   OLS is a conditional mean function. The predicted y_i is a linear addictive function of x_1 and x_2. In other words, $\hat{y}_i$ is an expected value given x_1 and x_2.

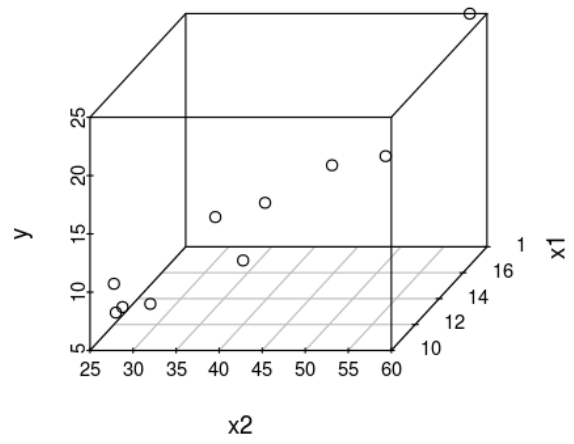
   For example, let's say wage = β_0 + β_1(educ) + β_2(age) + e,
   the $\hat{wage} = \hat{\beta}_0 + \hat{\beta}_1(BA) + \hat{\beta}_2(30)$ is the mean wage for workers with BA at age 30.



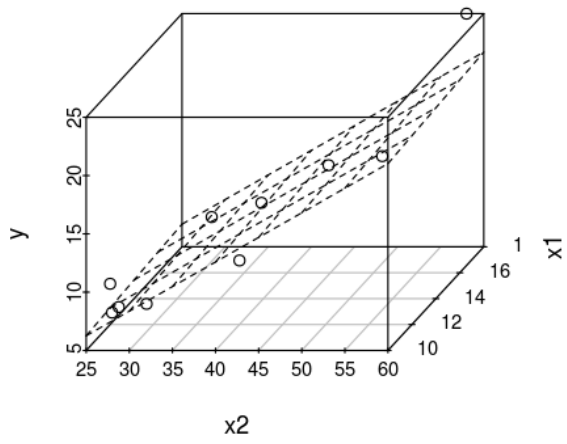
(a) Scatterplot



(b) w/ Best Fit Line



(c) 3D Scatterplot



(d) w/ Best Fit Plane

Let's say we have the following data.

Table 1. Employee Hourly Wages, Years of Schooling, and Age

Employee ID	Hourly Wage (y)	Years of Schooling (x_1)	Age (x_2)
1	8.50	12	25
2	12.00	14	34
3	9.00	10	32
4	10.50	12	40
5	11.00	16	37
6	15.00	16	51
7	25.00	18	58
8	12.00	18	42
9	6.50	12	26
10	8.25	10	28

Stata Codes

```
clear
input id wage sch age
 1 8.5  12  25
 2 12   14  34
 3 9    10  32
 4 10.5 12  40
 5 11   16  37
 6 15   16  51
 7 25   18  58
 8 12   18  42
 9 6.5  12  26
10 8.25 10  28
end

* ------------
* Correlation Coefficients Matrix
* ------------
pwcorr wage sch age, sig

* ------------
* Regression
* ------------
reg wage sch age
predict wagehat
predict residual, resid

* ------------
* Regression with centering of the independent variables
* : which makes b_0 the expected wage when all independent variables
* : are centered at their means (i.e., yhat when all x's = x-means)
* ------------
```

```
egen wagemean= mean(wage)
gen  newwage = wage-wagemean
egen schmean = mean(sch)
gen  newsch  = sch-schmean
egen agemean = mean(age)
gen  newage  = sch-agemean

reg wage sch age
```

Stata Results

```
. reg wage sch age

      Source |       SS       df       MS              Number of obs =      10
-------------+------------------------------           F(  2,     7) =   16.82
       Model |   203.88044     2   101.94022           Prob > F      =  0.0021
    Residual |    42.42581     7     6.06083           R-squared     =  0.8278
-------------+------------------------------           Adj R-squared =  0.7785
       Total |   246.30625     9  27.3673611           Root MSE      =  2.4619

------------------------------------------------------------------------------
        wage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         sch |   .0852087   .4232348     0.20   0.846    -.9155825       1.086
         age |   .4218134   .1194178     3.53   0.010     .1394352    .7041917
       _cons |  -5.134521   3.796749    -1.35   0.218    -14.11241    3.843362
------------------------------------------------------------------------------
```

- `Total SS`: $\sum(y_i - \bar{y})^2 = 246.30625$.
- `Model SS`: $\sum(\hat{y}_i - \bar{y})^2 = 203.88044$.
- `Residual SS`: $\sum(y_i - \hat{y})^2 = \sum e^2 = 42.42581$.
- `MS`: SS/df.
- `MSE` (Mean Square Error): Residual SS / df $= \frac{\sum e^2}{n-k} = 6.06083$.
- `Root MSE`: $\sqrt{MSE} = \sqrt{6.06083} = 2.4619$.
- `R-squared`: The proportion of variation in y explained by x's. That is, 82.8% of the variation in wage ($= \sum(y - \bar{y})^2$) is reduced by factoring in schooling and age.
- `Adj R-squared`: Adjusted R-squared after factoring in the number of covariates.
- `Coef`: Parameter estimated (i.e., $\hat{\beta}$).
- `St.Err.`: Standard error of the parameter estimated.
- `t`: t-value = Coef/Std.Err.
- `P>|t|`: p-value = the probability of type I error for each parameter estimated.
- `F` and `Prob > F` will be discussed next week.

8. Test of the significance of all (or multiple) slopes: F-tests

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \cdots + \beta_{k-1} x_{(k-1)i} + \varepsilon_i$$

(a) t-test tests the significance of individual slope:
$H_0 : \beta_j = 0$

(b) F-test examines the significance of all regression coefficients combined:
$H_0 : \beta_1 = \beta_2 = \cdots = \beta_{k-1} = 0$
H_A : not all coefficients of independent variables = 0

Null Model (NM): $y_i = \beta_0 + \varepsilon_i$
Full Model (FM): $y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \cdots + \beta_k x_{(k-1)i} + \varepsilon_i$

That is, F-test examines whether the amount of residual variance diminishes significantly enough in Full Model compared to the amount of residual variance of Null Model. Note that the parameter estimated of Full Model is $k = 1 +$ the number of independent variables. Therefore, the parameter estimated of Null Model is 1.

(c) How to compute F-test statistic to test $H_0 : \beta_1 = \beta_2 = \cdots = \beta_k = 0$:

$$F = \frac{(ESS_{NM} - ESS_{FM})/(k-1)}{ESS_{FM}/(n-k)} = \frac{(R^2_{FM} - R^2_{NM})/(k-1)}{(1-R^2_{FM})/(n-k)} = \frac{R^2_{FM}/(k-1)}{(1-R^2_{FM})/(n-k)}$$

where ESS refers to the error sum of squares. Note that $R^2_{NM} = 0$ and $\hat{\beta}_0 = \bar{y}$ in Null Model. Recall that $k-1$ is equal to the number of independent variables and $n-k$ is equal to [the total sample size - # of variables - 1.]

If F-test statistic is greater than the critical value, we reject H_0. F-distribution is a ratio of χ^2 distribution. The d.f. of F-test statistic is the same as the number of independent variables.

(d) For example, in the following Stata result,
Null Model: $lnwage_i = \beta_0 + \varepsilon_i$
Full Model: $lnwage_i = \beta_0 + \beta_1(age_i) + \beta_2(yrsch_i) + \varepsilon$

```
. sum lnwage

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
      lnwage |       580    10.18874      1.1232   4.317488   12.92164
```

```
. reg lnwage

      Source |       SS       df       MS              Number of obs =     580
-------------+------------------------------           F(  0,   579) =    0.00
       Model |           0     0           .           Prob > F      =       .
    Residual |  730.453968   579  1.26157853           R-squared     =  0.0000
-------------+------------------------------           Adj R-squared =  0.0000
       Total |  730.453968   579  1.26157853           Root MSE      =  1.1232

------------------------------------------------------------------------------
      lnwage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       _cons |   10.18874   .0466383   218.46   0.000     10.09714    10.28034
------------------------------------------------------------------------------

. reg lnwage age yrsch

      Source |       SS       df       MS              Number of obs =     580
-------------+------------------------------           F(  2,   577) =   86.56
       Model |  168.587183     2  84.2935915           Prob > F      =  0.0000
    Residual |  561.866785   577  .973772591           R-squared     =  0.2308
-------------+------------------------------           Adj R-squared =  0.2281
       Total |  730.453968   579  1.26157853           Root MSE      =   .9868

------------------------------------------------------------------------------
      lnwage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   .0224036   .0031647     7.08   0.000     .0161879    .0286194
       yrsch |   .1614003   .0153952    10.48   0.000     .1311627    .1916378
       _cons |   7.054155   .2422886    29.11   0.000      6.57828     7.53003
------------------------------------------------------------------------------
```

$H_0 : \beta_1 = \beta_2 = 0$ is tested by `F( 2, 577)`. `Prob > F = 0.0000` shows that the probability that both β_1 and β_2 are jointly zero is less than .0001. Note that two separate t-tests (i.e.,$H_0 : \beta_1 = 0$ and $H_0 : \beta_2 = 0$) are not equivalent to testing the joint hypothesis of $H_0 : \beta_1 = \beta_2 = 0$.

In the above Stata result, $F = \frac{(730.453968 - 561.866785)/2}{561.866785/577} = \frac{.2308/2}{(1-.2308)/577} = 86.56$. The critical value at $F(\alpha = .05, 2, 577)$ is 3.01.

(e) F-test also examines the significance of two or more additional independent variables compared to the model which does not have these additional variables.

Restricted Model (RM): $y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \cdots + \beta_k x_{(k-q-1)i} + \varepsilon_i$
Unrestricted Model (UM): $y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \cdots + \beta_k x_{(k-1)i} + \varepsilon_i$

In the above, unrestricted model has q number of more independent variables than the restricted model.

Here we test,

$H_0 : \beta_{k-q} = \beta_{k-q+1} = \beta_{k-q+2} = \cdots = \beta_{k-1} = 0$
H_A : not all q coefficients $= 0$

(f) How to compute F-test statistic to test $H_0 : \beta_{k-q} = \beta_{k-q+1} = \beta_{k-q+2} = \cdots = \beta_{k-1} = 0$:

$$F = \frac{(ESS_{RM} - ESS_{UM})/q}{ESS_{UM}/(n-k)} = \frac{R^2_{UM} - R^2_{RM}/q}{(1 - R^2_{UM})/(n-k)}$$

where q refers to the number of independent variable.

(g) For example, in the following regression, we add `year`, `wtsupp`, and `serial` in addition to `age` and `yrsch` that we controlled for in the previous example.

RM: $lnwage_i = \beta_0 + \beta_1(age_i) + \beta_2(yrsch_i) + \varepsilon$
UM: $lnwage_i = \beta_0 + \beta_1(age_i) + \beta_2(yrsch_i) + \beta_3(year_i) + \beta_4(wtsupp_i) + \beta_5(serial_i) + \varepsilon$

We want to test $H_0 : \beta_3 = \beta_4 = \beta_5 = 0$

```
. reg lnwage age yrsch

      Source |       SS       df       MS              Number of obs =     580
-------------+------------------------------           F(  2,   577) =   86.56
       Model |  168.587183     2  84.2935915            Prob > F      =  0.0000
    Residual |  561.866785   577  .973772591            R-squared     =  0.2308
-------------+------------------------------           Adj R-squared =  0.2281
       Total |  730.453968   579  1.26157853            Root MSE      =   .9868

------------------------------------------------------------------------------
      lnwage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   .0224036   .0031647     7.08   0.000     .0161879    .0286194
       yrsch |   .1614003   .0153952    10.48   0.000     .1311627    .1916378
       _cons |   7.054155   .2422886    29.11   0.000      6.57828     7.53003
------------------------------------------------------------------------------
```

```
. reg lnwage age yrsch year wtsupp serial

      Source |       SS       df       MS              Number of obs =     580
-------------+------------------------------           F(  5,   574) =   34.69
       Model |  169.519941     5  33.9039882           Prob > F      =  0.0000
    Residual |  560.934027   574  .977236981           R-squared     =  0.2321
-------------+------------------------------           Adj R-squared =  0.2254
       Total |  730.453968   579  1.26157853           Root MSE      =  .98855

------------------------------------------------------------------------------
      lnwage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   .0224951   .0031972     7.04   0.000     .0162155    .0287746
       yrsch |   .1630112   .0155505    10.48   0.000     .1324684     .193554
        year |   .0254408   .0830589     0.31   0.759    -.1376957    .1885773
      wtsupp |  -.0000228   .0000437    -0.52   0.603    -.0001086    .0000631
      serial |   1.19e-06   1.47e-06     0.81   0.420    -1.70e-06    4.08e-06
       _cons |  -44.14367   167.0164    -0.26   0.792    -372.1814    283.8941
------------------------------------------------------------------------------
```

(h) Using two regression results, $F = \frac{(561.866785-560.934027)/3}{560.934027/574} = \frac{(.2321-.2308)/3}{(1-.2321)/574} = .318$. This is not statistically significant. The critical value at $\alpha = .05$ with 3 df and 574 df is 2.62. Therefore, we fail to reject that $H_0 : \beta_3 = \beta_4 = \beta_5 = 0$.

(i) Fortunately, Stata has a F-test function after estimating regression model:

```
. test year wtsupp serial

 ( 1)  year = 0
 ( 2)  wtsupp = 0
 ( 3)  serial = 0

       F(  3,   574) =    0.32
            Prob > F =    0.8123
```

9. The assumptions of OLS Regression Models:

   (a) *Linearity*: The expectation of the error–that is, the average value of ε_i given the value of $x's$–is zero: $E(\varepsilon_i|x_1, x_2, \cdots, x_k) = 0$. Equivalently, the expected value of the response variable is a linear function of the explanatory variable.

   (b) *Constant variance* or *Homoscedasticity*: The variance of the errors is the same regardless of the value of $x's$. That is, $V(\varepsilon_i|x_1, x_2, \cdots, x_k) = \sigma^2_{\varepsilon_i}$. Because the distribution of the errors is the same as the distribution of the response variable around the population regression line, constant error variance implies constant conditional variance of y, given $x's$.

   $$V(y|x_1, x_2, \cdots, x_k) = E(y_i - \{\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \cdots + \beta_k x_{ki}\})^2 = E(\varepsilon_i^2) = \sigma^2_{\varepsilon_i}$$

   Note that because the mean of ε_i is 0, its variance is simply $E(\varepsilon_i^2)$. In other words, the variance of ε_i is equal to $\frac{\sum(\varepsilon_i - \bar{\varepsilon}_i)^2}{n-k}$ where $\bar{\varepsilon}_i = 0$.

   The violation of homoscedasticity assumption means *heteroscedasticity*. Serious violations in homoscedasticity may result in overestimating the goodness of fit (i.e., high R^2). Violations of homoscedasticity also make it difficult to gauge the true standard deviation of the errors, usually resulting in confidence intervals that are too wide or too narrow. Heteroscedasticity may also have the effect of giving too much weight to small subset of the data (namely the subset where the error variance was largest) when estimating coefficients. However, the coefficients estimated are still unbiased, thus, The violation of homoscedasticity assumption is not a so serious problem.

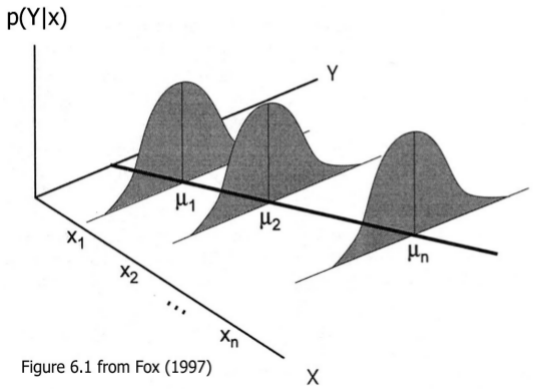


Figure 6.1 from Fox (1997)

   (c) *Normality*: The errors are normally distributed: $\varepsilon_i \sim N(0, \sigma^2_{\varepsilon_i})$. Equivalently, the conditional distribution of the response variable is normal: $y_i \sim N(\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \cdots + \beta_k x_{ki}, \sigma^2_{\varepsilon_i})$. Note that there is no normality assumption for independent variables.

   Violations of normality compromise the estimation of coefficients and the calculation of confidence intervals. Sometimes the error distribution is "skewed" by the presence of a few large outliers. Since parameter estimation is based on the minimization of squared error, a few extreme observations can exert a disproportionate

influence on parameter estimates. Calculation of confidence intervals and various significance tests for coefficients are all based on the assumptions of normally distributed errors. If the error distribution is significantly non-normal, confidence intervals may be too wide or too narrow. Nonetheless, the coefficients estimated are still unbiased. Therefore, the severity of the violation of the normality assumption is not so great.

An easy fix of the violation of the normality assumption if the dependent variable is rightly skewed (e.g., income) is to transform the dependent variables by taking natural log.

(d) *Independence*: The observations are sampled independently. Any pair of errors ε_i and ε_j (or equivalently, of conditional response-variable values y_i and y_j) are independent for $i \neq j$. The assumption of independence needs to be justified by the procedures of data collection. As long as it is randomly sampled, the independence assumption is fulfilled for cross-sectional data.

Note that sometimes it is called an *i.i.d.* assumption: A sequence or other collection of random variables is independent and identically distributed (i.i.d.) if each random variable has the same probability distribution as the others and all are mutually independent.

The independence assumption is usually violated with time serious data, which typically causes the problem of autocorrelation.

(e) *Fixed $x's$, or $x's$ measured without error and independent of the error*: In observational studies (i.e., surveys), we assume that the explanatory variables are measured without error and the explanatory variable and the error are independent in the population. That is, the error has the same distribution, $N(0, \sigma_\varepsilon^2)$, for every value of x. This is the most problematic assumption in OLS, and there are various statistical models that can be applied when this assumption is violated.

If $x's$ and error terms are correlation, then, the coefficients estimated are biased.

(f) *x's are not invariant*: All independent variables should not be constant.

(g) *No multicollinearity*: No x is a perfect linear function of the others. When explanatory variables in regression are invariant or perfectly collinear, the least-square coefficients are not uniquely defined.

10. Gauss-Markov Theorem
    Properties of the Least-Squares Estimators: Best Linear Unbiased Estimators (BLUE)

    (a) The Gauss–Markov theorem, named after Carl Friedrich Gauss and Andrey Markov, states that in a linear regression model in which the errors have expectation zero and are uncorrelated and have equal variances, **the best linear unbiased estimator (BLUE)** of the coefficients is given by the ordinary least squares estimator.

    In other words, within the class of linear unbiased estimates of β_0 and β_1, $\beta_2, \cdots, \beta_k$, the least square estimator, $\hat{\beta}_0$ and $\hat{\beta}_1$, $\hat{\beta}_2, \cdots, \hat{\beta}_k$ has **minimum variance (is most efficient)**.

    (b) Here "best" means giving the lowest possible mean squared error of the estimate (i.e., the smallest $\sum \varepsilon^2$, and, thus, the smallest mean squared error (MSE)). The errors need not be normal. The Gauss-Markov theorem requires the assumptions of linearity, homoscedasticity, and independence. That is, the theorem does not depend on any assumption of normality and, thus, any other particular shape of the distribution of the error term.

    (c) In plain English, the Gauss-Markov Theorem states that the OLS estimates ($\hat{\beta}_0$ and $\hat{\beta}_1$, $\hat{\beta}_2, \cdots, \hat{\beta}_k$) are the best guesses of the true population parameters. In other words, the OLS estimates are more likely to be close to the true population intercept and slope. (Read p.103. Check also pp.196–197 if you can understand Matrix Allegra).

11. Other Properties of the Least-Squares Estimators

    (a) Under normality, the least-square estimators are the most efficient among all unbiased estimators, not just among linear estimators. When the error distribution is heavier tailed than normal (i.e., rightly or leftly skewed), for example, the least-squares may be much less efficient than certain robust-regression estimators, which are not linear functions of the data.

    (b) Under the assumption of normality, the least-squares coefficients are the maximum-likelihood estimators (I'm not sure but if we have time, we will discuss the maximum-likelihood estimators later.)