Handout #8

Title: Foundations of Econometrics — Fall/2015
Course: Econ 367 — Instructor: Dr. I-Ming Chiu

Linear Regression Model

So far we have focused mostly on the study of a single random variable, its corresponding theoretical distribution, and sampling scheme. However, very often we are more interested in bivariate or even multivariate relationships between/among random variables. We'll begin with a bivariate case where X and Y are theoretically related using coin tossing example. We'll show that the "conditional mean" of Y on X can be formed using a deterministic linear function in X. After that we'll introduce the simple linear regression model where Y can be linearly dependent on X empirically. In the study of economics, we are often interested in whether one variable Y can be explained by the other variable X. For example, is inflation caused by the over-injection of money supply in the long run? Are the households spending governed by their disposable income? Is employment status for a female worker dependent on the number of children she has? All of the above questions can be answered and examined using simple linear regression model. Be noticed that the causal relationship is established using economic theory and empirical linear regression model is used to examine the validity of economic theory. Among these three examples the only difference is, in the third case, the Y variable is categorical. We'll study the third case later using Probit or Logit model, an extension of linear regression model. As you should find out by now, the data type introduced in Handout#1 plays an important role to decide how we choose an appropriate model to study the data.

*Consider an experiment where a fair coin is tossed four times; sample space $\Omega = (0, 1)^4$
X = # of heads obtained on the first three tosses, Y = # of heads obtained on all four tosses
Table 8.1 Joint Distribution

X\Y	0	1	2	3	4	f(X)
0	1/16	1/16	0	0	0	1/8
1	0	3/16	3/16	0	0	3/8
2	0	0	3/16	3/16	0	3/8
3	0	0	0	1/16	1/16	1/8
g(Y)	1/16	1/4	3/8	1/4	1/16	1

Table 8.2 Simulation outcome based on tossing a coin four times and repeat it 100 times. (I did't set the seed number, so the outcome from another experiment will be different)

x/y	0	1	2	3	4
0	0.07	0.04	0	0	0
1	0	0.17	0.22	0	0
2	0	0	0.21	0.17	0
3	0	0	0	0.05	0.07

What is the conditional mean function of Y given X based on the above joint distribution?
Answer: It's E(Y|X); expectation of Y given X.

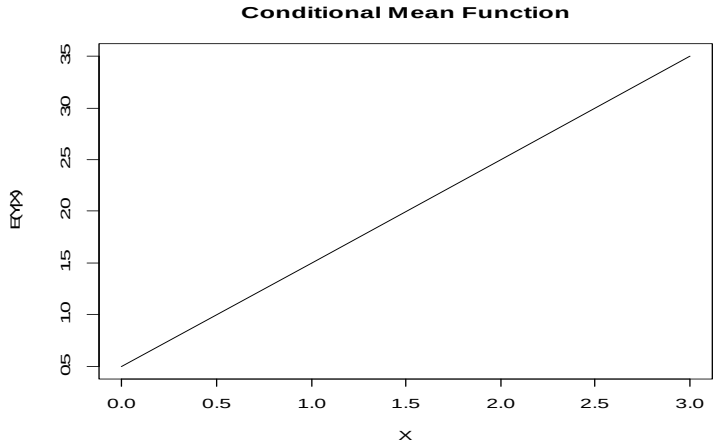
Table 8.3 Conditional Distribution

Y	0	1	2	3	4
g(Y|X=0)	1/2	1/2	0	0	0
g(Y|X=1)	0	1/2	1/2	0	0
g(Y|X=2)	0	0	1/2	1/2	0
g(Y|X=3)	0	0	0	1/2	1/2

$E(Y|X=0) = \frac{1}{2}$, $E(Y|X=1) = \frac{3}{2}$, $E(Y|X=2) = \frac{5}{2}$, $E(Y|X=3) = \frac{7}{2}$

If we plot E(Y|X) against X in a scatter diagram, it looks like the following:

Figure 8.1



There is an "exact" (i.e., deterministic) linear relationship between E(Y|X) & X, so we can write the following equation:

$E(Y|X) = \beta_0 + \beta_1 * X$

How do we find β_0 and β_1?
Answer:

$$\beta_1 = \frac{\Delta Y}{\Delta X} = \frac{E(Y|X=1) - E(Y|X=0)}{1-0} = 1$$

When $X = 0 \Rightarrow \beta_0 = E(Y|X=0) = \frac{1}{2}$

$$E(Y|X) = \frac{1}{2} + X$$

The slope and intercept of the above conditional mean function are known constants given that we know how X and Y are related (i.e., knowing their joint distribution function).

Linear Regression Model: regression analysis is concerned with the study of the relationship between one variable called the **explained**, or **dependent**, variable and one or more other variables called **independent**, or **explanatory**, variables.

$$Y = \beta_0 + \beta_1 * X + \varepsilon \qquad (1)$$

Equation (1) is termed **"simple"** (one X) **"linear"** (linearity in X) **regression model**. Y is the dependent variable, X is the independent variable, and ε is the error term. In practice, it's unlikely the relationship between X and Y is an exact straight line like the coin tossing example we studied earlier. Therefore, the error term is added to represent uncertainty and all other potential factors that may contribute to the variation of Y (i.e., the capture-all effect). We will make assumptions about the error term later for inference purpose.

Objectives of linear regression model:

(1) To estimate the mean value of the dependent variable, given the value of the independent variable(s). In other words, we assume the conditional mean function is linear: $E(Y|X) = \beta_0 + \beta_1 * X$

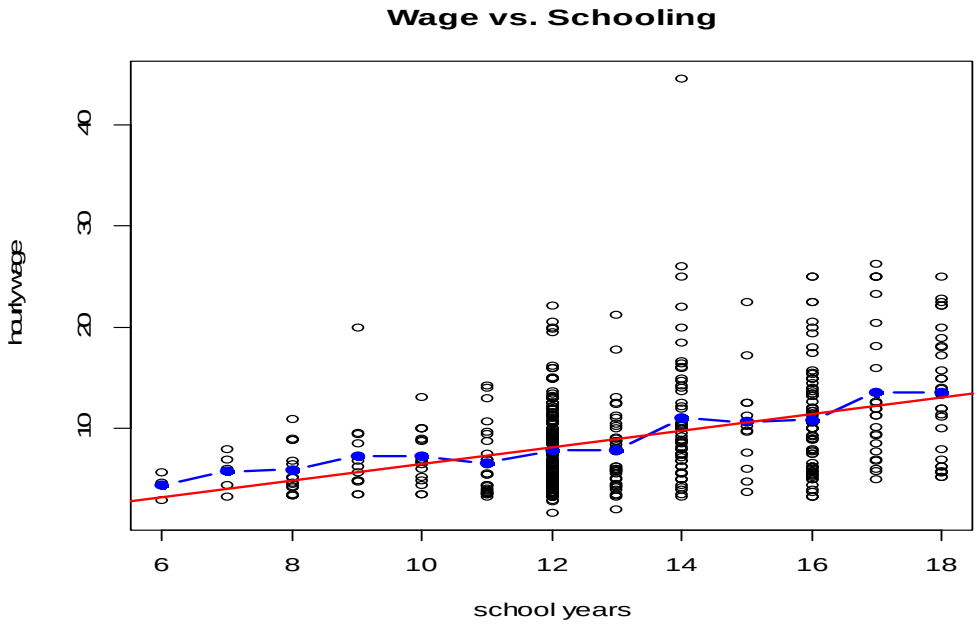
(2) To test hypotheses about the nature of the dependence (i.e., β_1). The size and magnitude of beat(s) (if there are more than one independent variables) tell us how the changes in Xs affect Y. This is called marginal effect.

(3) To predict, or forecast, the mean value of the dependent variable, given the value(s) of the independent variable(s).

For example, we are interested in finding the relationship between wage and schooling in a small market economy. If the population data is available, equation (1) is called linear population regression function (PRF). However, very often it is too costly to get the population data. Therefore, a small sample is drawn from the population and our goal is to uncover the unknown parameters in the linear sample regression function (SRF). Let's use Y to denote the hourly wage and X education background (measured in schooling years). The plot of Y against X is shown in the following **scatter** diagram (Fig. 8.2). As you may notice the actual conditional mean function is not a straight line. However, a linear regression line seems an appropriate approximation for describing the relationship between wage and education.

Be noticed that, one sample is obtained from a model (simulated population) where people's wages are a linear function of schooling. By using a simulated data, we know all of the corresponding parameters in the population and the corresponding sampling scheme. There is an advantage of using simulated data; first, we can see how the SRF is different from the PRF. Secondly, we can visualize the consequences of assumption violations by changing the model assumption one at a time (i.e., generate a new population but with a different model assumption). Thirdly, we can examine the usefulness of model predictability.

Fig. 8.2



In Fig. 8.2, the sample conditional mean of wage (the blue line) tends to increase with schooling years. Although it's not an exact straight line, it can be approximated using a linear equation (the red line) as following:

$E(\text{Wage}|\text{Schooling}) = \beta_0 + \beta_1 * \text{Schooling} + \varepsilon$, where $\varepsilon \sim N(0, \sigma^2)$

Where ε is a deviation term that captures other factors that may affect wage. Usually we assume that it is normally distributed with mean zero and variance σ^2. This assumption is required for statistical inference purpose.

Linear regression model: Find an "estimator" that best describes the linear relationship between Wage (dependent variable) and Schooling (independent variable). In other words, we need a method to uncover unknown parameters "β_0", "β_1", and "σ^2".

There are usually two approaches to do that; MLE (maximum likelihood estimator) and OLS (ordinary least square). We'll adopt the OLS estimator because it has a nice "BLUE"[1] property. I'll give more information about this "BLUE" property later.

OLS:

$$Y_i = \beta_0 + \beta_1 * X_i + \varepsilon_i \ (i = 1 \ldots n) \qquad (2)$$

Choosing β_0 & β_1 to minimize $\sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i)^2$ (3)

[1] It stands for "Best Linear Unbiased Estimator".

Apply differential calculus[2] on equation (3), we can find

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(X_i - \overline{X}) * (Y_i - \overline{Y})}{\sum_{i=1}^{n}(X_i - \overline{X})^2} \quad \begin{matrix} \nearrow \text{SXY} \\ \longrightarrow \text{SXX} \end{matrix} \qquad (1.1)$$

$$\hat{\beta}_0 = \overline{Y} - \hat{\beta}_1 * \overline{X} \qquad (1.2)$$

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 * X \ (\hat{Y} : \text{predicted value}) \qquad (1.3)$$

$$\hat{e} = Y - \hat{Y} \ (\hat{e} : \text{residual; an estimator for } \varepsilon) \qquad (1.4)$$

The numerator part is called residual sum of squares; RSS.

$$\hat{\sigma}^2 = \frac{\sum \hat{e}_i^{\,2}}{n-2} \qquad (1.5)$$

$\hat{e}_i$ is termed "residual" which is obtained using "$Y_i - (\hat{\beta}_0 + \hat{\beta}_1 * X_i)$".

$$\text{Var}(\hat{\beta}_0) = \hat{\sigma}^2 * (\frac{1}{n} + \frac{\overline{X}^2}{\text{SXX}}) \qquad (1.6)$$

$$\text{Var}(\hat{\beta}_1) = \hat{\sigma}^2 * \frac{1}{\text{SXX}} \qquad (1.7)$$

Analysis of Variance (ANOVA)

$$\Sigma(Yi - \overline{Y})^2 = \Sigma(\hat{Y}_i - \overline{Y})^2 + \Sigma \hat{e}_i^{\,2} \qquad (1.8)$$

$$\text{TSS} = \text{ESS} + \text{RSS}$$

$$R^2 \text{ (coefficient of determination)} = \frac{\text{ESS}}{\text{TSS}} \qquad (1.9)$$

$$R^2 = \frac{\hat{\beta}_1^{\,2} * \text{SXX}}{\text{SYY}} = \frac{(\frac{\text{SXY}}{\text{SXX}})^2 * \text{SXX}}{\text{SYY}} = \frac{(\text{SXY})^2}{\text{SXX} * \text{SYY}} \qquad (1.10)$$

$$R = \frac{\text{SXY}}{\sqrt{\text{SXX}}\sqrt{\text{SYY}}} \qquad (1.11)$$

[2] We need to solve two equations simultaneously. Taking the first derivative with respect to β_0 and β_1 respectively, we can obtain: $-2*\Sigma (Y_i - \beta_0 - \beta_1*X_i) = 0$ & $-2 X_i * \Sigma (Y_i - \beta_0 - \beta_1*X_i) = 0$. The above two equations imply that $\Sigma \hat{e}_i = 0$ & $\Sigma X_i * \hat{e}_i = 0$

Hypothesis Testing:

a) Null vs. Alternative

H_0: $\beta_1 = \beta^*$ (usually the null is to examine whether β_1 is zero or not (i.e., $\beta^* = 0$), but it can be other parameter rather than zero, which is specified in the economic model)

H_A: $\beta_1 \neq 0$

b) Test statistics (TS) = $\dfrac{\hat{\beta}_1 - \beta^*}{\sigma_{\hat{\beta}_1}} \sim t_{n-p}$ ($\sigma\hat{\beta}_1$: $\sqrt{\text{Var}(\hat{\beta}_1)}$ same in (1.7), n: number of observations; p: number of explanatory variables need to be estimated)

c) Rejection region

It is decided by data analyst (YOU), usually 1%, 5% or 10% is chosen. It can be explained as the probability we allow for the Type I error. In other words, the probability we reject the null hypothesis when it is true.

d) P-Value

The probability to obtain a value that is as extreme as the test statistic (TS).

e.g. Study time and exam score

Time (X)	2	5	1	3	8	2	0	6	3	1
Score (Y)	65	69	64	75	90	75	49	77	74	58

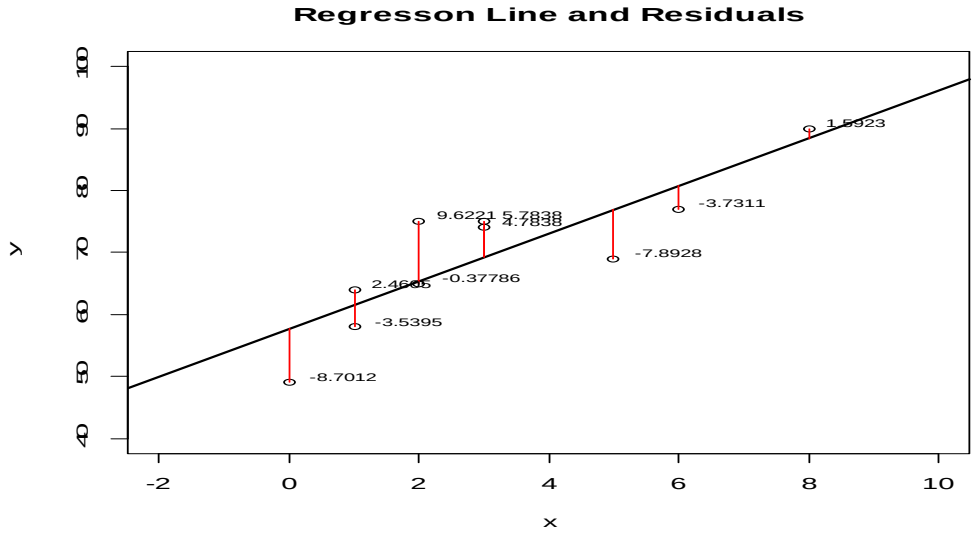
Let's compute $\hat{\beta}_0$, $\hat{\beta}_1$, and $\hat{\sigma}^2$ manually and report the outcomes as follows:

$\hat{\beta}_1$ = 3.8383 (score/hour) …. the marginal score increases by 3.8383 if the study hour increases by one unit.

$\hat{\beta}_0$ = 57.7013 (score)

$\hat{\sigma}^2$ = 322.1125/(10-2) = 40.264

Fig. 8.3



In Fig 8.3, it shows that the regression line we get will yield the smallest squared aggregate deviations; $\sum_{i=1}^{n}(Y_i - \beta_0 - \beta_1 X_i)^2$. In other words, the total of $(-8.7012)^2 + \ldots + (1.5923)^2$ will be the smallest given the OLS estimator $\hat{\beta}_0$ and $\hat{\beta}_1$.

Hypothesis Testing on β_0 and β_1

Suppose we are interested in whether β_1 equals certain value, say β_1*.

H_0: $\beta_1 = \beta_1$*
H_A: $\beta_1 \neq \beta_1$*

Assume the error term, ε, is NIID[3] $(0, \sigma^2)$
$\hat{\beta}_1 \sim N(\beta_1^*, \sigma_{\hat{\beta}_1}^2)$

$$TS = \frac{\hat{\beta}_1 - \beta_1^*}{\sigma_{\hat{\beta}_1}} \sim t_{n-2}$$ (Why is the degree of freedom "n-2"?)

Let's use the numerical example from page six to examine whether β_1 equals zero (does study time affect exam score?) In other words, we want to test

H_0: $\beta_1 = 0$
H_A: $\beta_1 \neq 0$

$$\sigma_{\hat{\beta}_1}^2 (= Var(\hat{\beta}_1)) = \hat{\sigma}^2 * \frac{1}{SXX} = \frac{40.264}{56.9} = 0.7076$$

$\hat{\sigma}^2 = 40.264$
SXX = 56.9

$$TS = \frac{3.8383 - 0}{\sqrt{0.7076}} = 4.56$$ (table value, $t_8 = 2.306$ at 5% level of significance)

The above TS indicates that we can not reject the null hypothesis at 5% (or even 1%) level of significance using two-tail test.

Alternatively, we can calculate the p-value to decide whether we can reject the null. A p-value is a **measure of how much evidence we have against the null hypothesis**. I'll show you how to calculate p-value using Stata. The p-value for TS = 4.56 with 8 df is about $9.2*10^{-4}$

[3] "Normally", "Identically", "Independently", "Distributed". The common assumptions about the error tem are a) $E(\varepsilon | X) = 0$, b)$E(\varepsilon^2 | X) = \sigma^2$. Meaning, the error tem has a mean equals zero and it is "homoscedastic" (i.e., constant variance).

What can we conclude in our numerical example?
If students study one more hour, his (her) exam score will be 3.8383 "significantly" higher.

The Analysis of Variance (Again; you should see the pattern when F test is needed)
Let's consider the conditional mean of Y: $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 * X$

$$Y_i = \hat{Y}_i + \hat{e}_i \Rightarrow Y_i - \overline{Y} = (\hat{Y}_i - \overline{Y}) + \hat{e}_i$$
$$\Sigma(Y_i - \overline{Y})^2 = \Sigma(\hat{Y}_i - \overline{Y})^2 + \Sigma \hat{e}_i^2 \qquad (1.8)$$
$$TSS[4] = ESS + RSS$$
$$ESS = \Sigma(\hat{Y}_i - \overline{Y})^2 = \Sigma(\hat{\beta}_0 + \hat{\beta}_1 * X_i - (\hat{\beta}_0 + \hat{\beta}_1 * \overline{X}))^2 = \hat{\beta}_1^2 * \Sigma(X_i - \overline{X})^2$$
$$R^2 \text{ (coefficient of determination)} = \frac{ESS}{TSS} \qquad (1.9)$$

We will apply equation (1.8) and (1.9) using matrix operations in Stata to obtain coefficient of determination in the next session.

Table 1 **ANOVA**

Source	df	SS	MS	F	p-value
Regression	1	SSreg	SSreg/1	MSreg[5]/ $\hat{\sigma}^2$	
Residual	n-2	RSS	$\hat{\sigma}^2$ =RSS/n-2		
Total	n-1	SYY			

Notice: $t^2 = F$ (see pp. 6, Handout #6)

The Residuals
The residuals can be used for "diagnostic check". This examines whether the model assumptions are violated. Knowing whether the assumptions are violated affects the hypothesis testing results. In the next handout we will show that how to "modify" OLS estimator given that we detect either "multicollinearity", "heteroscedascity", and/or "autocorrelation" problems.

Predictions
$\hat{Y}_\bullet = \hat{\beta}_0 + \hat{\beta}_1 * X_\bullet$
Where "$X_\bullet$" is a chosen value (vector). For example, given "certain study time, $X_\bullet$" what should be the expected exam score ($\hat{Y}_\bullet$)?

[4] Denote this term by SYY.
[5] SSreg/1 = MSreg.

Standard error of prediction = $\hat{\sigma} * (1 + \frac{1}{n} + \frac{(X_\bullet - \overline{X})^2}{SXX})^{1/2}$

Standard error of prediction in matrix form = $\{\hat{\sigma}^2 * [1 + X_\bullet^T (X^T X)^{-1} X_\bullet]\}^{1/2}$

Linear Regression Model using Matrix Operations

ALTERNATIVELY, we can write equation (1) on page 3 into a matrix form as following:

$$Y = X*\varphi + \varepsilon \qquad (2.1)$$

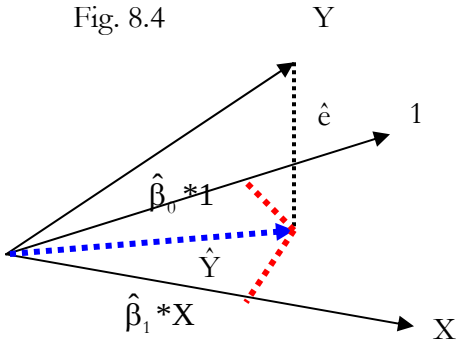
Where Y is a nx1 matrix, X is a nx2 matrix and φ is a 2x1 vector. Let's ignore "ε" temporarily.

$$Y = \begin{bmatrix} y_1 \\ y_2 \\ \bullet \\ \bullet \\ \bullet \\ y_n \end{bmatrix}, X = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \bullet & \bullet \\ \bullet & \bullet \\ \bullet & \bullet \\ 1 & x_n \end{bmatrix}, \varphi = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}.$$

Applying matrix differentiation according to equation (3)[6] on page 4, we can obtain

$$\hat{\varphi} = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{bmatrix} = (X^T X)^{-1} (X^T Y) \qquad (2.2)^{7}$$

Fig. 8.4



Projection (blue arrow) of Y on (1, X) vector space

[6] In matrix form eq. (3) is equivalent to select both beta estimators to minimize $\varepsilon^T\varepsilon$.

[7] The derivation of equation (2.2) is same as the one I show in footnote 2 on page 5. I will briefly explain matrix differentiation in our meeting. As you can see (2.2) is a generalization of obtaining beta estimators. In other words, it can be applied easily to linear multiple (i.e., more than one regressors) regression model.

What does Fig. 8.4 mean mathematically?

$$\begin{bmatrix}\hat{y}_1\\ \hat{y}_2\\ \bullet\\ \bullet\\ \bullet\\ \hat{y}_n\end{bmatrix} = \hat{\beta}_0 * \begin{bmatrix}1\\ 1\\ \bullet\\ \bullet\\ \bullet\\ 1\end{bmatrix} + \hat{\beta}_1 * \begin{bmatrix}x_1\\ x_2\\ \bullet\\ \bullet\\ \bullet\\ x_n\end{bmatrix}$$

(We want to assign a "weight" to 1 and X vector and thus results in a value (distance) that is closet to Y vector).

Small Sample Property of OLS Estimator

Estimated β_0 and β_1

$$E\left(\begin{bmatrix}\hat{\beta}_0\\ \hat{\beta}_1\end{bmatrix}\right) = \begin{bmatrix}\beta_0\\ \beta_1\end{bmatrix} \ldots \text{OLS estimator is unbiased} \qquad (2.3)$$

$$\text{Var}\left(\begin{bmatrix}\hat{\beta}_0\\ \hat{\beta}_1\end{bmatrix}\right) \text{ has the smallest variance among all the linear} \qquad (2.4)$$

estimators … OLS estimator is efficient.

Estimated Variances of β_0 and β_1

$$\text{Var}(\hat{\varphi})^{8} = \begin{bmatrix}\text{Var}(\hat{\beta}_0) & \text{Cov}(\hat{\beta}_0,\hat{\beta}_1)\\ \text{Cov}(\hat{\beta}_0,\hat{\beta}_1) & \text{Var}(\hat{\beta}_1)\end{bmatrix} = \hat{\sigma}^2 * (X^T X)^{-1} = \hat{\sigma}^2 * \left(\begin{bmatrix}n & \sum X_i\\ \sum X_i & \sum X_i^2\end{bmatrix}\right)^{-1}$$

$$= \hat{\sigma}^2 * \frac{1}{n * \sum X_i^2 - (\sum X_i)^2} * \begin{bmatrix}\sum X_i^2 & -\sum X_i\\ -\sum X_i & n\end{bmatrix} \qquad (2.5)^{9}$$

$$SXX = \sum (X_i - \overline{X})^2 = \sum (X_i^2 - 2X_i * \overline{X} + \overline{X}^2) = \sum X_i^2 - 2 * \overline{X} * \sum X_i + \sum \overline{X}^2$$

$$= \sum X_i^2 - 2 * \frac{\sum X_i}{n} * \sum X_i + n * \overline{X}^2 = \sum X_i^2 - 2 * \frac{(\sum X_i)^2}{n} + n * \left(\frac{\sum X_i}{n}\right)^2$$

---

[8] $\text{Var}(\varphi) = \sigma^2 (X^T X)^{-1}$; σ^2 needs to be estimated since it's unknown. The estimator is $\hat{\sigma}^2$.

[9] This variance-covariance matrix is the most important estimator for statistical inference purpose.

$$= \sum X_i^2 - 2*\frac{(\sum X_i)^2}{n} + \frac{(\sum X_i)^2}{n} = \sum X_i^2 - \frac{(\sum X_i)^2}{n}$$

Let's focus on the estimated variance of βs:

$$\text{Var}(\hat{\beta}_0) = \hat{\sigma}^2 * \frac{\sum X_i^2}{n*\sum X_i^2 - (\sum X_i)^2} = \hat{\sigma}^2 * \frac{SXX + (\sum X_i)^2/n}{n*SXX} = \hat{\sigma}^2 * \frac{SXX + n*\overline{X}^2}{n*SXX}$$

$$= \hat{\sigma}^2 * (\frac{1}{n} + \frac{\overline{X}^2}{SXX}) \qquad (2.6)$$

$$\text{Var}(\hat{\beta}_1) = \hat{\sigma}^2 * \frac{n}{n*\sum X_i^2 - (\sum X_i)^2} = \hat{\sigma}^2 * \frac{1}{\sum X_i^2 - (\sum X_i)^2/n} = \hat{\sigma}^2 * \frac{1}{SXX} \qquad (2.7)$$

$$\text{Cov}(\hat{\beta}_0, \hat{\beta}_1) = \hat{\sigma}^2 * \frac{-\sum X_i}{n*\sum X_i^2 - (\sum X_i)^2} = -\hat{\sigma}^2 * \frac{\sum X_i/n}{\sum X_i^2 - (\sum X_i)^2/n} = -\hat{\sigma}^2 * \frac{\overline{X}}{SXX} \qquad (2.8)$$

How do we find the estimate of correlation coefficient between β_0 & β_1 (i.e., $\rho(\hat{\beta}_0, \hat{\beta}_1)$)?

From here, all the routines and concepts are the same as I show earlier. We rely on mainly two equations to get the job done. Please go over the Mata routine at least once to learn how we obtain all the important outcomes in a linear regression model.

In general if we have p regressors, then $\hat{\varphi} = [\hat{\beta}_0, \hat{\beta}_1, ..., \hat{\beta}_{p-1}, \hat{\beta}_p]^T$

$$\hat{\varphi} = (X^T X)^{-1}(X^T Y)$$

$$\text{Var}(\hat{\varphi}) = \hat{\sigma}^2 * (X^T X)^{-1}$$