Kakhramon Yusupov

June 15th, 2017
9:00am – 10:30am
Session 1

Fundamentals of Regression Analysis

Regression

Regression analysis is concerned with the study of the ***dependence*** of one variable, the ***dependent variable***, on one or more other variables, the ***explanatory variables***, with a view of *estimating* and/or *predicting* the population *mean or average* values of the former in terms of the *known* or *fixed* (in repeated sampling) values of the latter.

Dataset

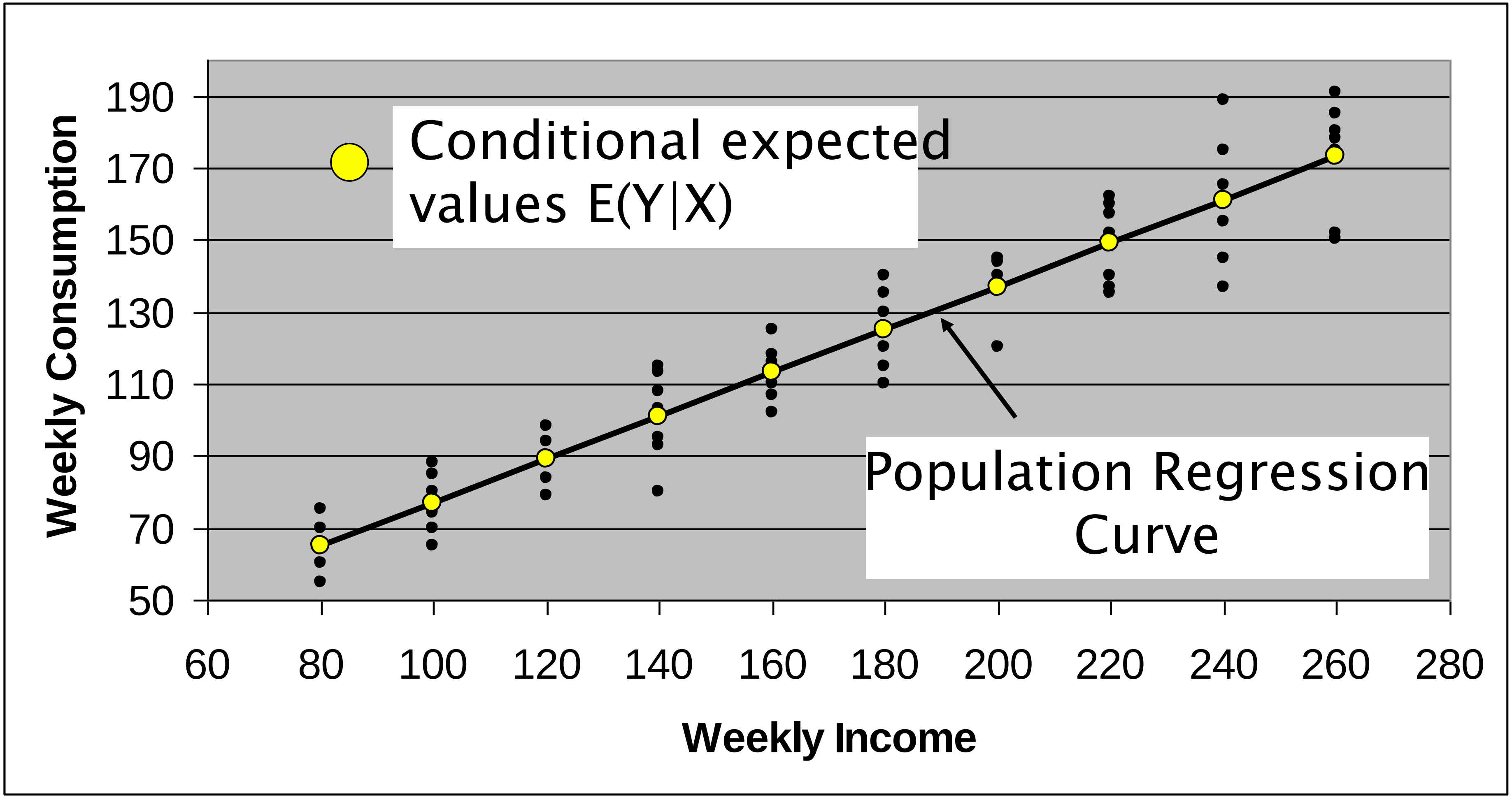
Terminology and Notation

Dependent Variable	**Independent variable**
Explained variable	Independent variable
Predictand	Predictor
Regressand	Regressor
Response	Stimulus
Endogenous	Exogenous
Outcome	Covariate
Controlled variable	Control variable

Conditional Mean

	Income	**80**	**100**	**120**	**140**	**160**	**180**	**200**	**220**	**240**	**260**
Consumption		55	65	79	80	102	110	120	135	137	150
		60	70	84	93	107	115	136	137	145	152
		65	74	90	95	110	120	140	140	155	175
		70	80	94	103	116	130	144	152	165	178
		75	85	98	108	118	135	145	157	175	180
			88		113	125	140		160	189	185
					115				162		191
Total		325	462	445	707	678	750	685	1043	966	1211
Conditional mean		**65**	**77**	**89**	**101**	**113**	**125**	**137**	**149**	**161**	**173**

Simple Regression



A population regression curve is simply the locus of the conditional means of the dependent variable for the fixed values of the explanatory variable(s).

Dataset

Simple Regression

$$E(Y \mid X_i) = f(X_i)$$

Conditional Expectation Function (CEF)

Population Regression

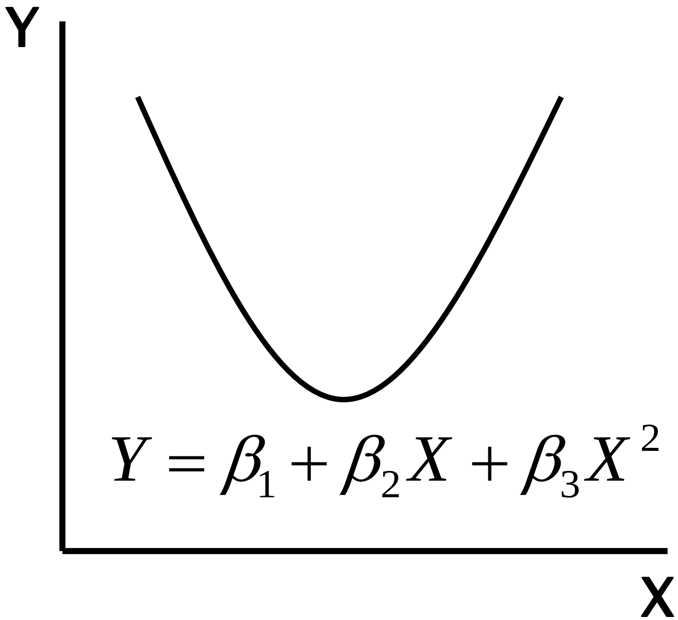
Population Regression Function (PRF)

$$E(Y \mid X_i) = \beta_1 + \beta_2 X_i$$

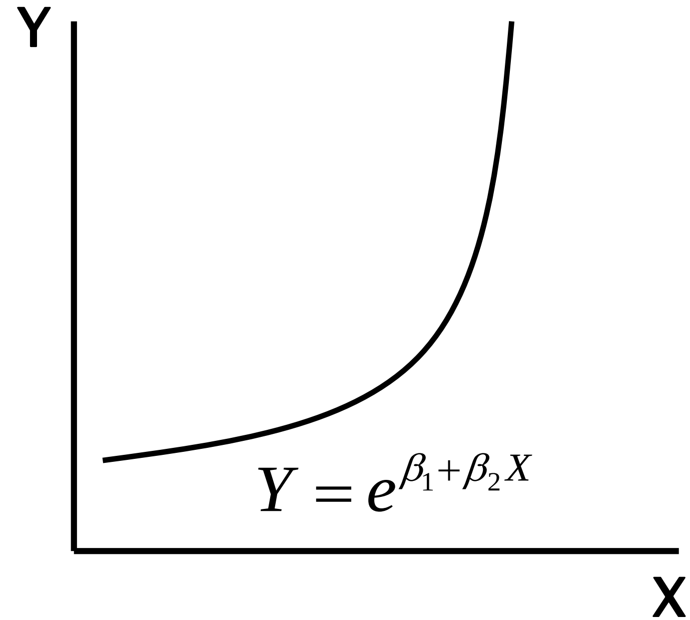
Linear Population Regression Function

Regression Coefficients

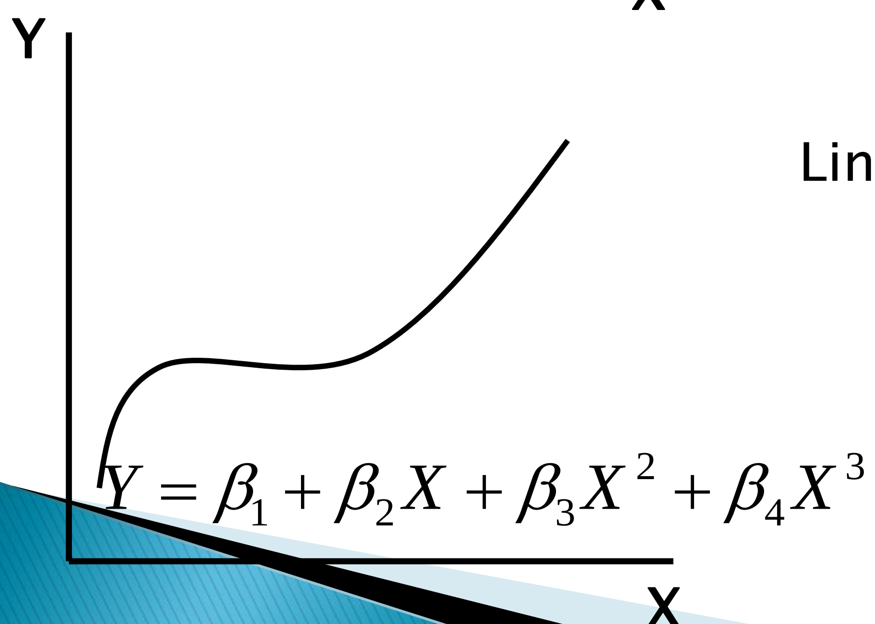
Linear



$$Y = \beta_1 + \beta_2 X + \beta_3 X^2$$



$$Y = e^{\beta_1 + \beta_2 X}$$



$$Y = \beta_1 + \beta_2 X + \beta_3 X^2 + \beta_4 X^3$$

Linear in parameter functions

$$E(Y \mid X_i) = \beta_1 + \beta_2{}^2 X_i$$

Non-linear in parameter function

Stochastic specification

$$u_i = Y_i - E(Y \mid X_i)$$ Stochastic error term

$$Y_i = E(Y \mid X_i) + u_i$$

Nonsystematic component

Systematic component

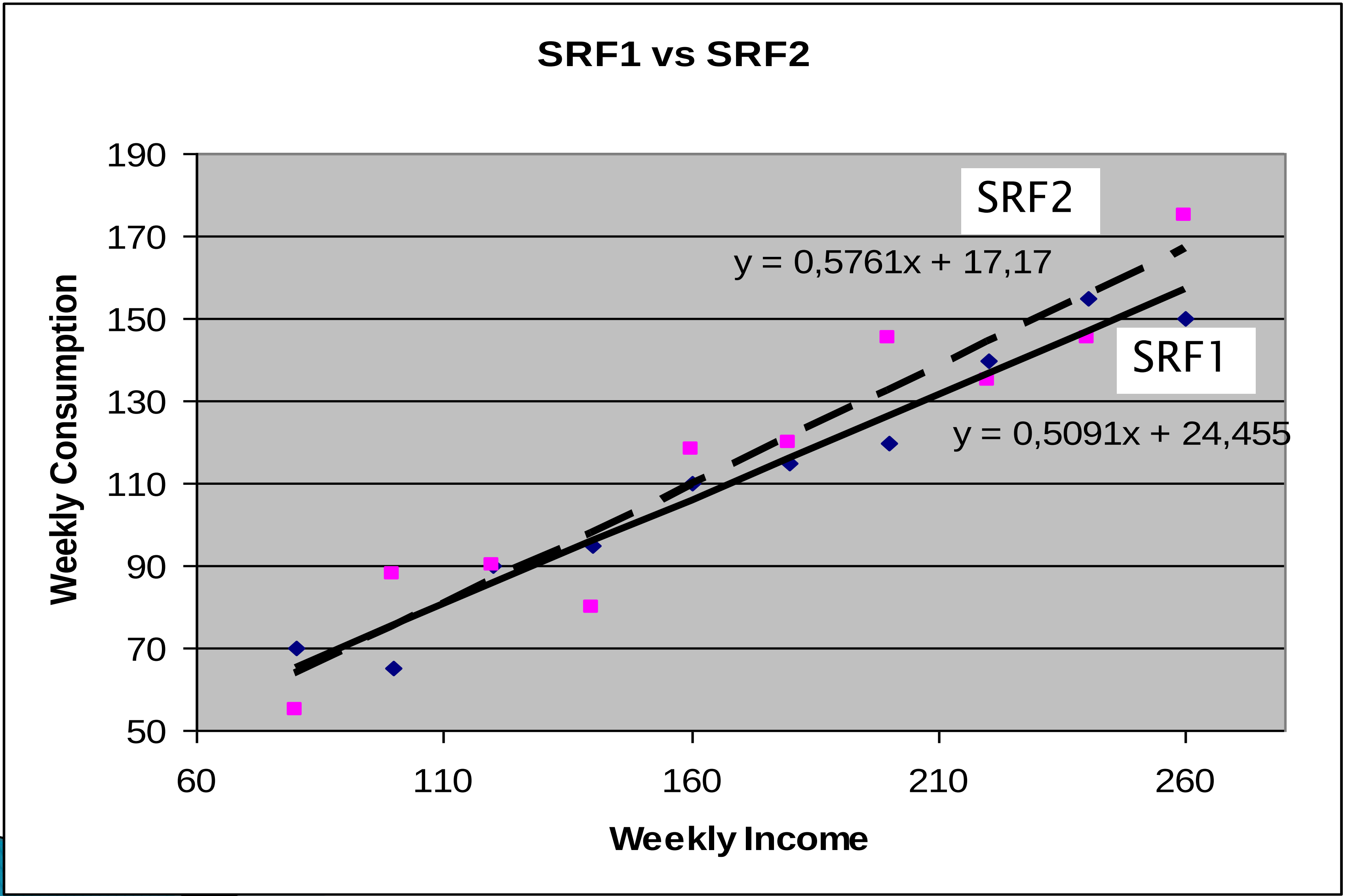
$$Y_i = \beta_1 + \beta_2 X_i + u_i$$

$$E(Y_i \mid X_i) = E[E(Y \mid X_i)] + E(u_i \mid X_i)$$
$$= E(Y \mid X_i)] + E(u_i \mid X_i)$$

$$E(u_i \mid X_i) = 0$$

Dataset

Sample Regression Function



Sample Regression Function

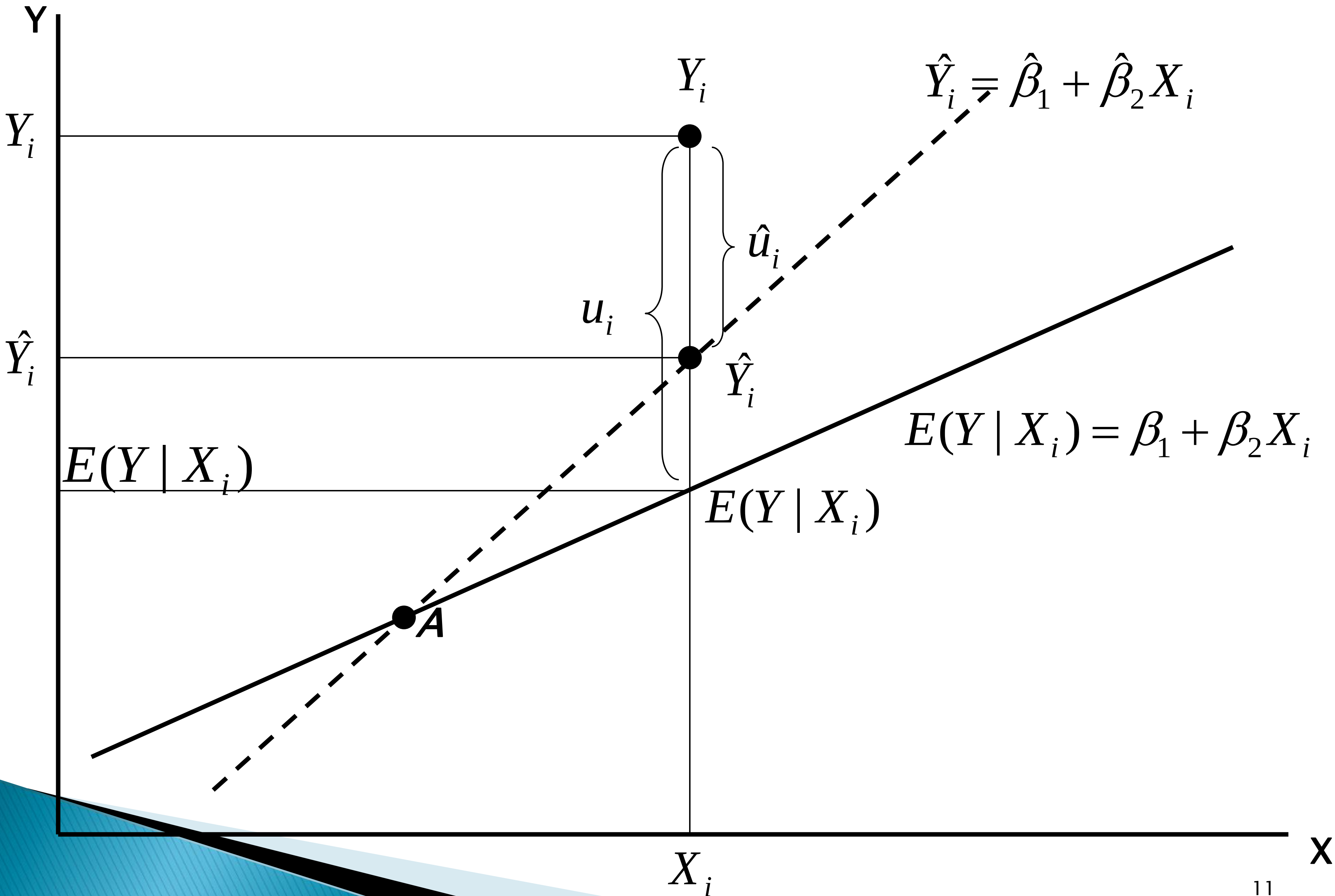
$$E(Y \mid X_i) = \beta_1 + \beta_2 X_i$$ PRF

$$\hat{Y}_i = \hat{\beta}_1 + \hat{\beta}_2 X_i$$ SRF

Estimate

$$Y_i = \hat{\beta}_1 + \hat{\beta}_2 X_i + \hat{u}_i$$

Sample Regression Function



Assumptions.

- **Linearity.** The relationship between independent and dependent variable is linear.
- **Full Rank**. There is no exact relationship among any independent variables.
- **Exogeneity of independent variables**. The error term of the regression is not a function of independent variables.
- **Homoscedastisity and no Autocorrelation**. Error term of the regression is independently and normally distributed with zero means and constant variance.
- Normality of Error term

Ordinary Least Squares

$$Y_i = \hat{\beta}_1 + \hat{\beta}_2 X_i + \hat{u}_i = \hat{Y}_i + \hat{u}_i$$

$$u_i = Y_i - \hat{Y}_i = Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_i$$

$$\sum u_i = \sum (Y_i - \hat{Y}_i)$$

$$\sum u_i^{\,2} = \sum (Y_i - \hat{Y}_i)^2 = \sum (Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_i)^2$$

$$\sum \hat{u}_i^2 = f(\hat{\beta}_1, \hat{\beta}_2)$$

Ordinary Least Squares

$$\sum u_i^2 = \sum Y_i^2 + n\hat{\beta}_1^2 + \hat{\beta}_2^2 \sum X_i^2 - 2\hat{\beta}_1 \sum Y_i - 2\hat{\beta}_2 \sum X_i Y_i + 2\hat{\beta}_1 \hat{\beta}_2 \sum X_i$$

$$\frac{\partial(\sum u_i^2)}{\partial \hat{\beta}_1} = 0 \quad \Rightarrow \quad 2n\hat{\beta}_1 - 2\sum Y_i + 2\hat{\beta}_2 \sum X_i = 0$$

$$n\hat{\beta}_1 = \sum Y_i - \hat{\beta}_2 \sum X_i$$

$$\hat{\beta}_1 = \overline{Y} - \hat{\beta}_2 \overline{X}$$

Ordinary Least Squares

$$\sum u_i^2 = \sum Y_i^2 + n\hat{\beta}_1^2 + \hat{\beta}_2^2 \sum X_i^2 - 2\hat{\beta}_1 \sum Y_i - 2\hat{\beta}_2 \sum X_i Y_i + 2\hat{\beta}_1 \hat{\beta}_2 \sum X_i$$

$$\frac{\partial(\sum u_i^2)}{\partial \hat{\beta}_2} = 0 \quad \Rightarrow \quad 2\hat{\beta}_2 \sum X_i^2 - 2\sum X_i Y_i + 2\hat{\beta}_1 \sum X_i = 0$$

$$\hat{\beta}_2 \sum X_i^2 - \sum X_i Y_i + \hat{\beta}_1 \sum X_i = 0$$

$$\hat{\beta}_2 \sum X_i^2 - \sum X_i Y_i + (\bar{Y} - \hat{\beta}_2 \bar{X}) \sum X_i = 0$$

$$\hat{\beta}_2 \sum X_i^2 - \sum X_i Y_i + (\bar{Y} - \hat{\beta}_2 \bar{X}) n\bar{X} = 0$$

$$\hat{\beta}_2 \left(\sum X_i^2 - n\bar{X}^2 \right) = \sum X_i Y_i - n\bar{X}\bar{Y}$$

$$\hat{\beta}_2 \left(\frac{1}{n} \sum X_i^2 - \bar{X}^2 \right) = \frac{1}{n} \sum X_i Y_i - \bar{X}\bar{Y}$$

$$\hat{\beta}_2 \mathrm{Var}(X) = \mathrm{Cov}(X, Y)$$

$$\boxed{\hat{\beta}_2 = \frac{\mathrm{Cov}(X, Y)}{\mathrm{Var}(X)}}$$

Assumptions

Linear Regression Model

$$Y_i = \beta_1 + \beta_2 X_i + u_i$$

X is assumed to be *nonstochastic*.

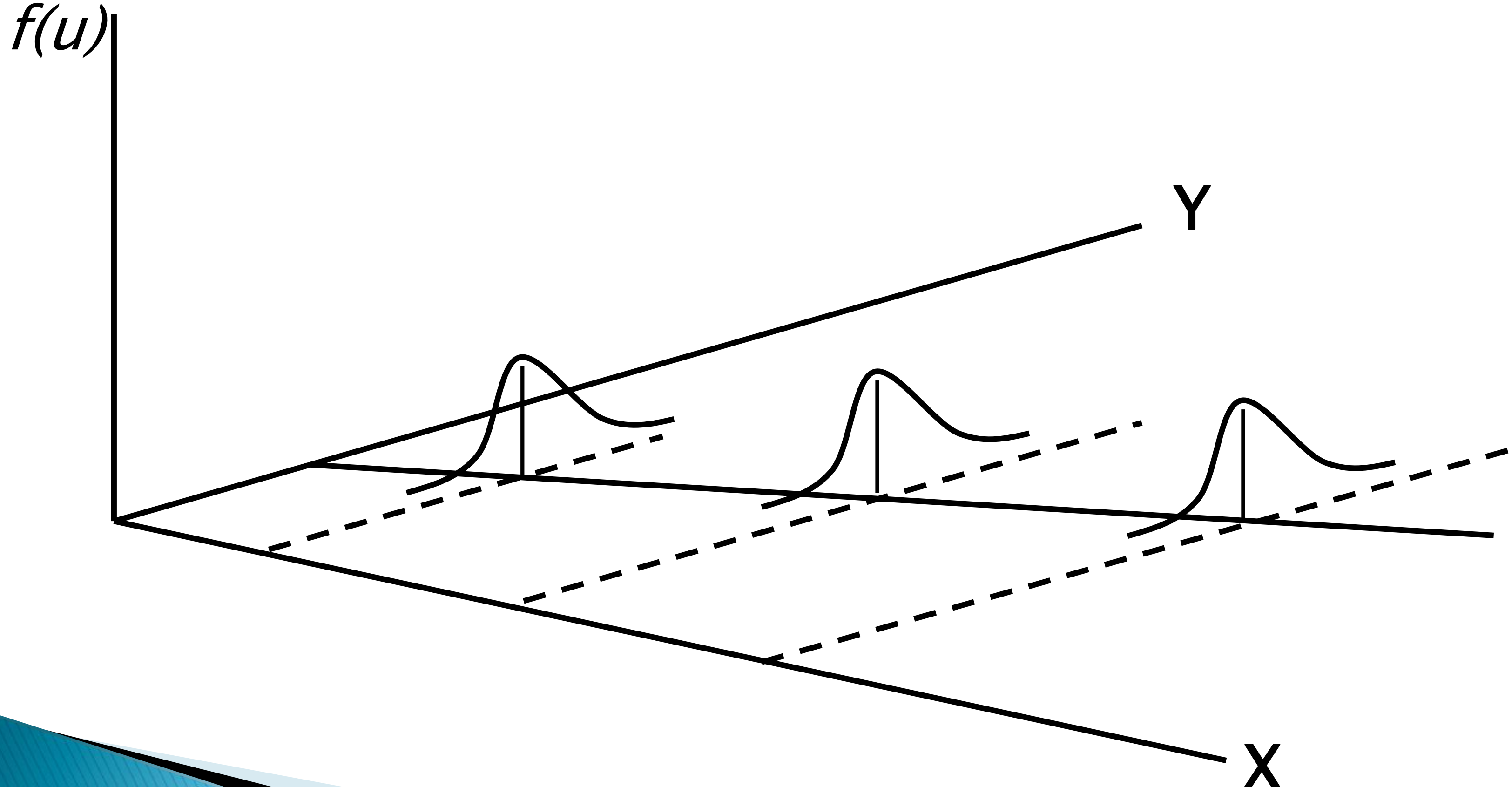
Zero mean values of disturbance u_i

$$E(u_i \mid X_i) = 0$$

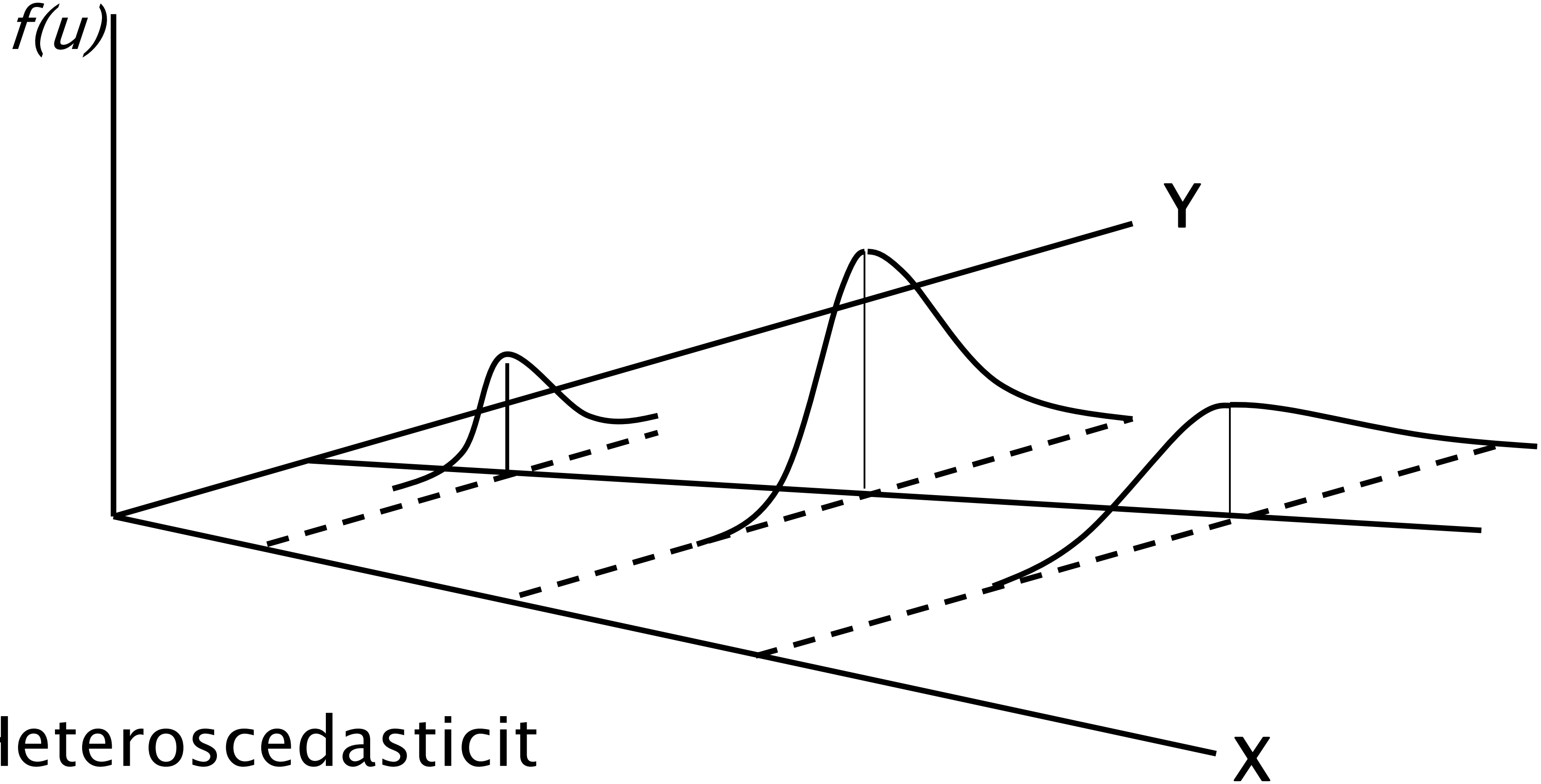
Assumptions

Homoscedasticity or equal variance of u_i

$$\operatorname{var}(u_i \mid X_i) = E[u_i - E(u_i \mid X_i)] = E(u_i^2 \mid X_i) = \sigma^2$$



Assumptions



Heteroscedasticit
y

$$\operatorname{var}(u_i \mid X_i) = \sigma_i^2$$

Assumptions

No autocorrelation between the disturbances

$$\operatorname{cov}(u_i, u_j \mid X_i, X_j) = E\{[u_i - E(u_i)] \mid X_i\}\{[u_j - E(u_j)] \mid X_i\}$$

$$= E(u_i \mid X_i)(u_j \mid X_j) = 0$$

Exogeneity. Zero covariance between X_i and u_i

$$\operatorname{cov}(X_i, u_i) = 0$$

Assumptions

The number of observations *n* should be greater than the number of parameters to be estimated *k*.

Variability in *X* values

The regression model is correctly specified.

There is no perfect multicollinearity.

Coefficient moments

Estimator

$$\hat{\beta}_2 = \frac{\sum_{i=1}^{n}\left(X_i - \bar{X}\right)\left(Y_i - \bar{Y}\right)}{\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2} = \sum_{i=1}^{n} W_i\left(Y_i - \bar{Y}\right) = \sum_{i=1}^{n} W_i Y_i$$

True value

$$Y_i = \beta_1 + \beta_2 X_i + u_i$$

$$W_i = \frac{\left(X_i - \bar{X}\right)}{\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2}$$

Note that $\sum_{i=1}^{n} W_i = 0$

$$\hat{\beta}_2 = \sum_{i=1}^{n} W_i Y_i = \sum_{i=1}^{n} W_i(\beta_1 + \beta_2 X_i + u_i) = \sum_{i=1}^{n} W_i \beta_1 + \sum_{i=1}^{n} W_i \beta_2 X_i + \sum_{i=1}^{n} W_i u_i =$$

$$= \beta_2 + \sum_{i=1}^{n} W_i u_i$$

$$\sum_{i=1}^{n} W_i X_i = \frac{\sum_{i=1}^{n}\left(X_i - \bar{X}\right)X_i}{\sum_{j=1}^{n}\left(X_j - \bar{X}\right)^2} = 1$$

Coefficient moments

$$E(\hat{\beta}) = E(\beta) + E(\sum_{i=1}^{n} W_i u_i)$$

$$E(\sum_{i=1}^{n} W_i u_i) = 0$$

According to our "Exogenity" assumption. (Error term is independent from X variable.

Thus, OLS estimator is unbiased estimator.

Coefficient moments

$$\hat{\beta} = \beta + \sum_{i=1}^{n} W_i u_i$$

$$Var(\hat{\beta}) = E\left((\hat{\beta} - \beta)^2\right) = E\left((\beta + \sum_{i=1}^{n} W_i u_i - \beta)^2 \right) =$$

$$E\left((\sum_{i=1}^{n} W_i u_i)^2 \right) = E\left((\sum_{i=1}^{n} W_i{}^2 u^2{}_i + 2\sum_{i<j}^{n} W_i W_j u_i u_j) \right) =$$

$$\sum_{i=1}^{n} W_i{}^2 E(u^2{}_i) + 2\sum_{i<j}^{n} W_i W_j E(u_i u_j)$$

According to Homoscedasticity and no auto-correlation assumptions.

Coefficient moments

$$E(u_i^2) = \sigma^2$$

$$E(u_i u_j) = 0$$

According to Homoscedasticity and no auto-correlation assumptions.

$$\sum_{i=1}^{n} W_i^2 = \frac{1}{\sum_{i=1}^{n} \left(X_i - \bar{X}\right)^2}$$

$$Var(\hat{\beta}) = \frac{\sigma^2}{\sum_{i=1}^{n} \left(X_i - \bar{X}\right)^2}$$

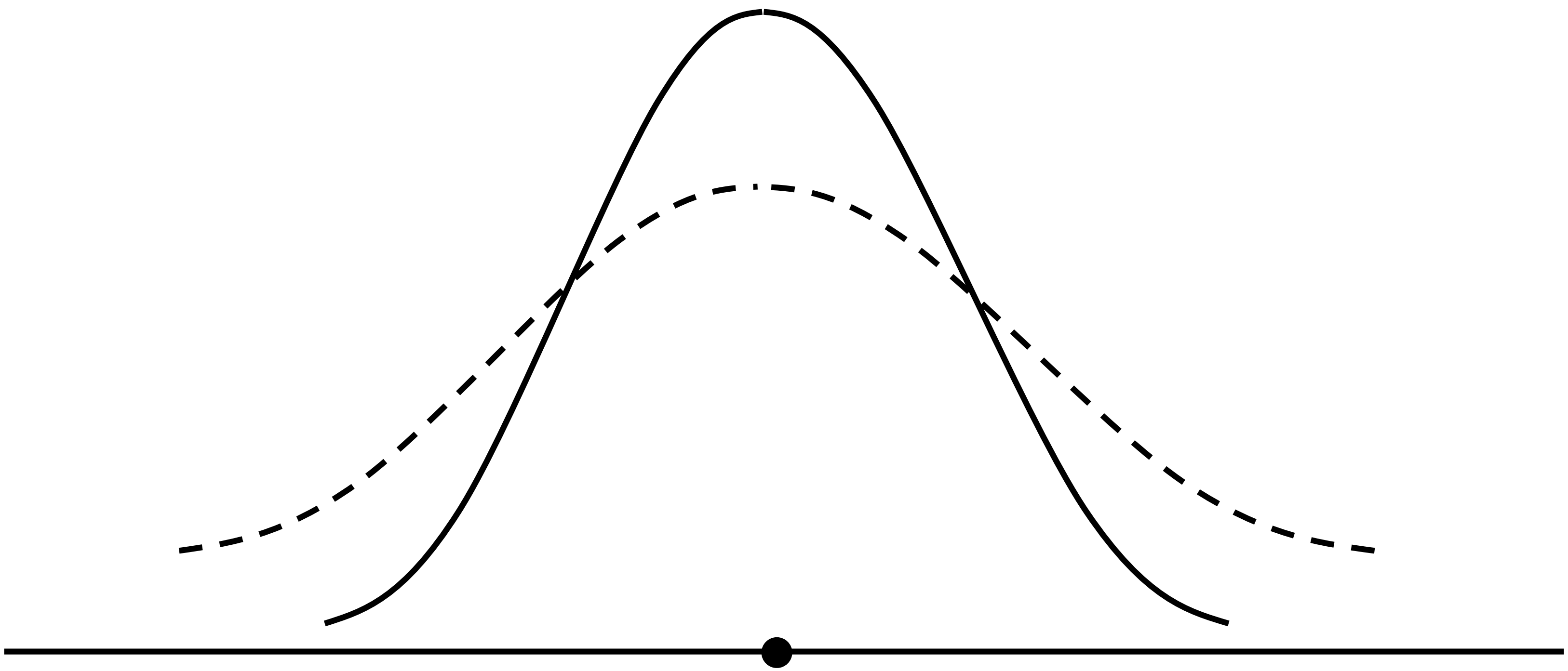
Using similar argument

$$\mathrm{var}(\hat{\beta}_2) = \frac{\sigma^2}{\sum (X_i - X)^2} \qquad STDEV(\hat{\beta}_2) = \sqrt{\frac{\sigma^2}{\sum (X_i - X)^2}}$$

$$STDEV(\hat{\beta}_1) = \sqrt{\frac{\sum X_i^2}{n\sum (X_i - X)^2}\sigma^2} \qquad \mathrm{var}(\hat{\beta}_1) = \frac{\sum X_i^2}{n\sum (X_i - X)^2}\sigma^2$$

$$\mathrm{cov}(\hat{\beta}_1, \hat{\beta}_2) = -\bar{X}\left(\frac{\sigma^2}{\sum (X_i - \bar{X})^2}\right)$$

BLUE estimator



$E(\hat{\beta}_2) = \beta_2$

Sampling distribution of $\hat{\beta}_2$

Goodness of Fit

$$\sum (Y_i - \bar{Y})^2 = \sum (\hat{Y}_i - \bar{Y})^2 + \sum \hat{u}_i^2$$

$$TSS = ESS + RSS$$

$$1 = \frac{ESS}{TSS} + \frac{RSS}{TSS}$$

$$R^2 = \frac{\sum (\hat{Y}_i - \bar{Y})^2}{\sum (Y_i - \bar{Y})^2}$$

Goodness of Fit

$$\sum (y_i - \bar{y})^2 = \sum (\hat{y}_i - \bar{y})^2 + \sum \hat{u}_i^2$$

$$TSS = ESS + RSS$$

$$1 = \frac{ESS}{TSS} + \frac{RSS}{TSS}$$

$$R^2 = \frac{\sum (\hat{y}_i - \bar{y})^2}{\sum (y_i - \bar{y})^2}$$

Goodness of Fit

$$TSS = ESS + RSS$$

$$1 = \frac{ESS}{TSS} + \frac{RSS}{TSS}$$

$$R^2 = 1 - \frac{RSS}{TSS}$$

The *R squared* **increases if a regressor is added to the model. Why? Hint: Consider sum of squared residuals.**

$$\overline{R}^2 = 1 - \frac{RSS/(N-k)}{TSS/(N-1)}$$

$$\overline{R}^2 = 1 - (1 - R^2)\left(\frac{N-1}{N-k}\right)$$

Confidence intervals

$$\hat{\beta}_1 \pm t_{\alpha/2} se(\hat{\beta}_1) \qquad \hat{\beta}_2 \pm t_{\alpha/2} se(\hat{\beta}_2)$$

OLS estimates have *t*-distribution with n-k **df, where** k **is the number of parameters.**

$$\Pr[\hat{\beta}_2 - t_{\alpha/2} se(\hat{\beta}_2) \le \beta_2 \le \hat{\beta}_2 + t_{\alpha/2} se(\hat{\beta}_2)] = 1 - \alpha$$

Hypothesis Testing

In this lecture we will consider two cases

1. **Hypotheses involving 1 coefficient (t-test)**

2. **Hypotheses involving 2 or more coefficients (F-test)**

Hypothesis Testing

1. Hypotheses involving 1 coefficient

Examples:

For regression coefficient the hypothesis can be

$$H_0 : \beta_2 = b$$
$$H_a : \beta_2 \neq b$$

which is a two-sided test,

or it can be

$$H_0 : \beta_2 \geq b$$
$$H_a : \beta_2 < b$$

which is a one-sided test

Interpret these tests for $b = 0$

Hypothesis Testing

Test statistics *t* which has *t*-distribution with df n-k

$$|t| = \left| \frac{\hat{\beta}_k - b}{se(\hat{\beta}_k)} \right|$$

If the *Null hypothesis* **is not true, then** *t-statistic* **is likely to have a large absolute value.**

If the absolute value of *t-statistic* **is greater than its critical value we reject the** *Null hypothesis,* **otherwise we cannot.**

The critical value can be looked up from the table for t-distribution's critical values.

For two-sided test it is $t_{\alpha/2}$

For one-sided test it is t_{α}

Hypothesis Testing

Hypothesis involving 2 or more coefficients

The need for these kinds of tests arise when we estimate multiple regression models

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots + \beta_k X_{ki} + u_i$$

For example, the wage function

$$wage_i = \beta_0 + \beta_1 educ_i + \beta_2 \exp_i + \beta_3 gender_i + \beta_4 region_i + u_i$$

Hypothesis Testing

Hypothesis involving 2 or more coefficients

$$H_0 : \beta_3 = 0, \beta_4 = 0$$

$$Ha : \text{They are not both } 0$$

Technically we need to compare two models

$$wage_i = \beta_0 + \beta_1 educ_i + \beta_2 \exp_i + \beta_3 gender_i + \beta_4 region_i + u_i$$

which says that gender and region are important factors of wage

and $$wage_i = \beta_0 + \beta_1 educ_i + \beta_2 \exp_i + u_i$$

which says that they are not

Hypothesis Testing

Hypothesis involving 2 or more coefficients

Denote RSS_0 as the sum of squared errors for the first model (if the Null hypothesis is true)

Denote RSS_1 as the sum of squared errors for the second model (if the Null hypothesis is not true)

$$F = \frac{(RSS_0 - RSS_1)/m}{RSS_1/(N-k)}$$

This F statistic has F-distribution with m df for the numerator and N-k df of denominator, where m is the number of restrictions (the difference in the number of independent variables)

If $F > F_c$ then reject your Null hypothesis

Hypothesis testing

Example 1.

Suppose you want to test that the marginal propensity to consume is less than 0.65

Based on 30 observations you estimated the simple regression model by OLS and obtained the results

$$Cons_i = 97.05 + 0.6 Inc_i + \hat{u}_i$$

Let the standard error of the slope be

$$se(\hat{\beta}_2) = 0.02$$

Hypothesis testing

Example 1 (continued)
Your hypothesis test is set up as

$$H_0 : \beta_2 \geq 0.65$$
$$H_a : \beta_2 < 0.65$$

t–statistic is equal to

$$|t| = \left| \frac{0.6 - 0.65}{0.02} \right| = 2.5$$

Compare that with $t_{0.05} = 1.701$ with df = 28
The conclusion is that we reject the null.

Hypothesis testing

Example 2:
Suppose that you want to test that gender and age do not affect wage
Based on 50 observations you estimated the following wage function

Your hypothesis is set up as

$$Ha : \text{Both are not equal to zero}$$

$$wage_i = 125.85 + 20.47 educ_i + 12.39 \exp_i + $$
$$+ 36.08 gender_i - 8.54 age_i + 45.12 region_i + \hat{u}_i$$

$$H_0 : \beta_3 = 0, \beta_4 = 0$$

Hypothesis testing

Example 2 (continued)

Suppose that *RSS* of the unrestricted model (Null is not true) is 33.85 and *RSS* of the restricted model (Null is true) is 42.52, then F – statistic is

$$F = \frac{(42.52 - 33.85)/2}{33.85/44} = 5.64$$

The critical value for F – statistic with df 2 and 44 and 5% significance level is 3.34

Since our F – statistic is greater than the critical value then we reject the null.