Econometrics

Course Calendar	
Week	Main Content
Week 1	**Introduction to Simple Regression**
Week 2	**Simple Regression**
Week 3	**Simple Regression: r^2 & Hands-on-Exercise**
Week 4	**Central Limit Theorem, Probability and Probability Density Function (PDF)**
Week 5	**Hypothesis Testing: Basics**
Week 6	**Simple Regression: Testing of Hypothesis**

Econometrics
Lecture 6. Simple Regression: Testing of Hypothesis

Geetha Rani Prakasam, Ph.D.
Professor

Recap

- Estimation of intercept and slope coefficients
- Estimation of the precision – SE of intercept and slope coefficients
- Normality assumption of U, intercept and slope
- Pre-requisites for Testing of Hypothesis – CLT, Prob, Prob Distribution, etc

Outline

- Example of Estimation
- Interval Estimation and Hypothesis Testing
- Interval estimation: Some basic Ideas
- Confidence Intervals for Regression coefficients
- Hypothesis Testing: General Comments
- Hy. Testing: The test of significance approach
- Testing the significance of σ^2 : The χ^2 Test

Outline

- Hypothesis Testing: Some practical aspects
- Regression Analysis and Analysis of Variance
- Application of Regression Analysis: Problem of Prediction
- Reporting the results of regression analysis
- Evaluating the results of regression analysis
- Summary and Conclusions

Mean hourly wage by Education

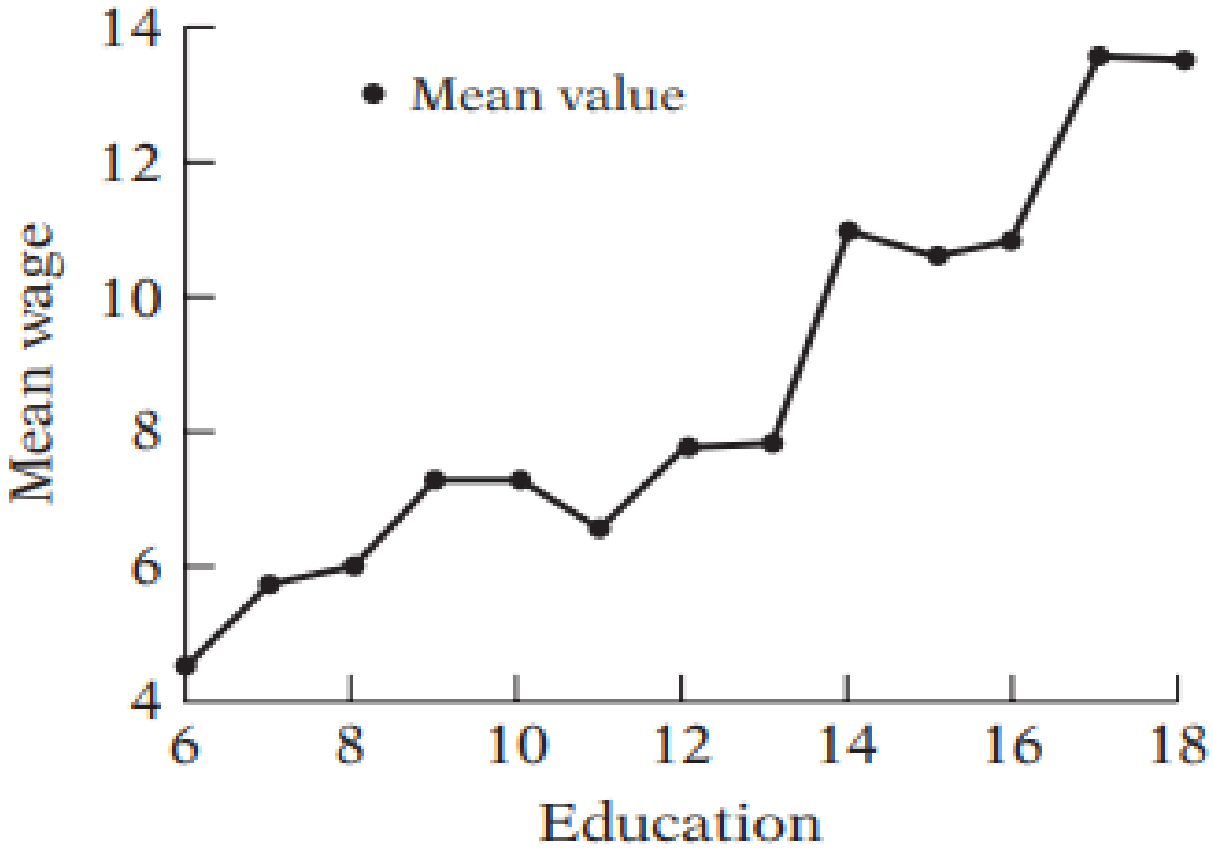
MEAN HOURLY WAGE BY EDUCATION

Years of schooling	Mean wage, $	Number of people
6	4.4567	3
7	5.7700	5
8	5.9787	15
9	7.3317	12
10	7.3182	17
11	6.5844	27
12	7.8182	218
13	7.8351	37
14	11.0223	56
15	10.6738	13
16	10.8361	70
17	13.6150	24
18	13.5310	31
	Total	528

Source: Arthur S. Goldberger, *Introductory Econometrics,* Harvard University Press, Cambridge, Mass., 1998, Table 1.1, p. 5 (adapted).

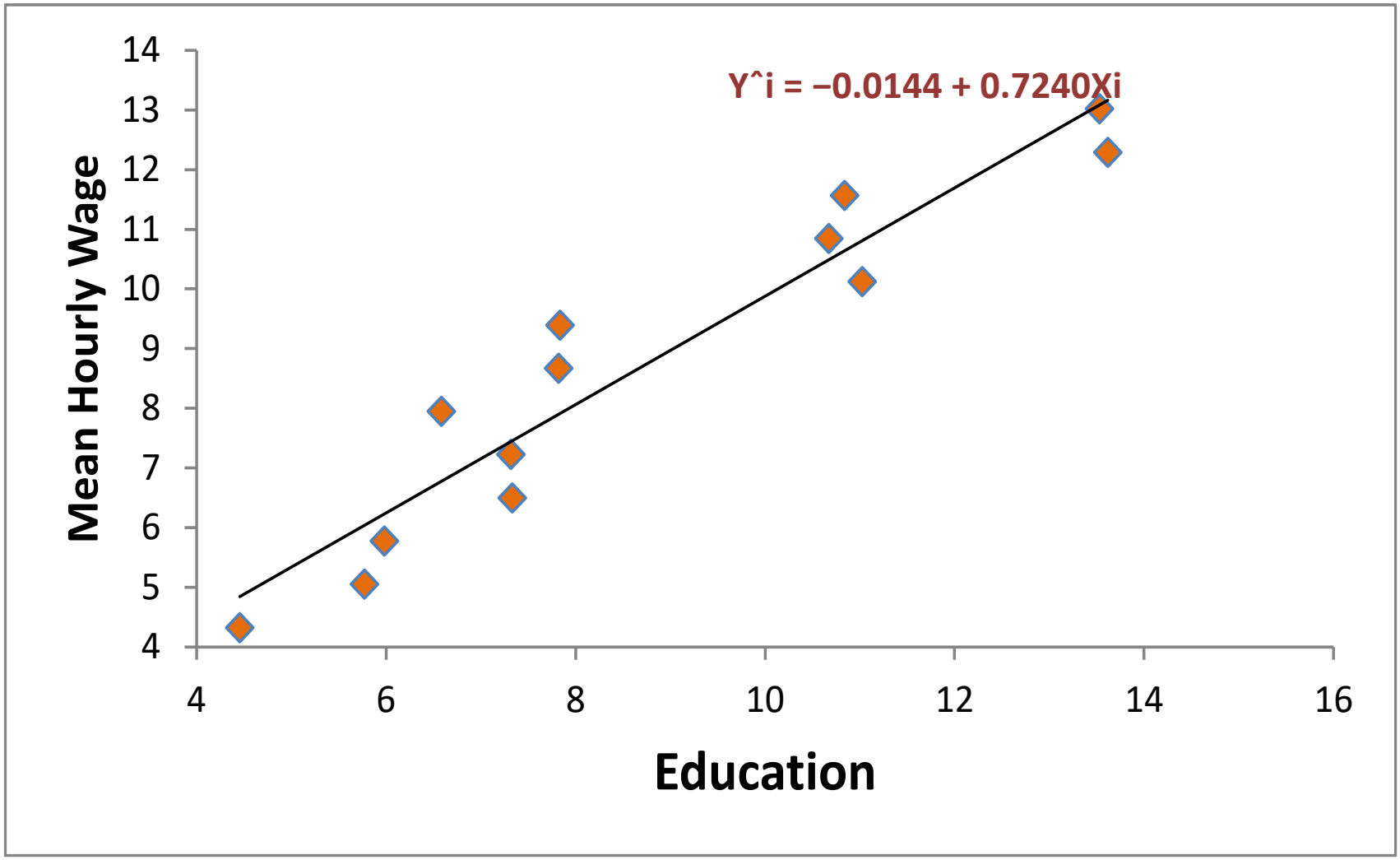
Source: Basic Econometrics by Domodar Gujarati, P.51

Relationship between mean wages & education



Source: Basic Econometrics by Domodar Gujarati, P.51

Figure 3.11.Estimated Regression Line



Ex: Mean hourly wage and Education

- *Yˆi = −0.0144 + 0.7240Xi*
- Geometrically, the estimated regression line is as shown in Figure 3.11.
- Each point on the regression line gives an estimate of the mean value of *Y* corresponding to the chosen *X value, that is, Yˆ i is an estimate of E(Y|Xi).*
- *The value of* βˆ2 = 0.7240, which measures the slope of the line, shows that, within the sample range of *X* between 6 and 18 years of education, as *X increases by 1, the estimated increase in mean* hourly wages is about 72 cents.
- That is, each additional year of schooling, on average, increases hourly wages by about 72 cents.

Ex: Mean hourly wage and Education

- The value of $\beta \hat{}1 = -0.0144$, which is the intercept of the line, indicates the average level of wages when the level of education is zero.
- Such literal interpretation of the intercept in the present case does not make any sense.
- The *r2 value of about 0.926 suggests that education explains about 93 percent of the variation* in hourly wage.
- The coefficient of correlation, *r = 0.9265, shows that wages and education are* highly positively correlated

Interval Estimation and Hypothesis Testing

- Wage equation ex. shows that the estimated average increase in mean hourly wage related to a one-year increase in schooling ($\hat{\beta}2$) is 0.7240, which is a one number (point) estimate of the unknown population value $\beta 2$.
- How reliable is this estimate?
- As noted earlier, because of sampling fluctuations, a single estimate is likely to differ from the true value, although in repeated sampling its mean value is expected to be equal to the true value. [Note: $E(\hat{\beta}2) = \beta 2$.]

Interval Estimation and Hypothesis Testing

- In statistics, the reliability of a point estimator is measured by its standard error.
- Therefore, instead of relying on the point estimate alone, we may construct an interval around the point estimator, say within two or three standard errors on either side of the point estimator, such that this interval has, say, 95 percent probability of including the true parameter value.
- This is roughly the idea behind **interval estimation.**

5-2. Interval estimation: Some basic Ideas

- To be more specific, assume that we want to find out **how "close" is, say, $\hat{\beta}_2$ to β_2 ?**
- For this purpose we try to find out two positive numbers δ and α, the latter lying between 0 and 1, such that the probability that the **random interval $(\hat{\beta}_2 - \delta, \hat{\beta}_2 + \delta)$ contains the true β_2** is $1 - \alpha$.
- Symbolically,

$$\Pr(\hat{\beta}_2 - \delta \leq \beta_2 \leq \hat{\beta}_2 + \delta) = 1 - \alpha \qquad (5.2.1)$$

5-2. Interval estimation: Some basic Ideas

- **Such a Random interval $\hat{\beta}_2 - \delta \leq \beta_2 \leq \hat{\beta}_2 + \delta$, if exits, known as confidence interval.**
- And *α $(0 < \alpha < 1)$ is known as the* ***level of significance*** **- a**lso known as the **probability of committing a Type I error.**
- **Type I error consists in** rejecting a true hypothesis, whereas a Type II error consists in accepting a false hypothesis.

5-2. Interval estimation: Some basic Ideas

- The symbol α is also known as the **size of the (statistical) test.**
- **$(1 - \alpha)$ is confidence coefficient; $0 < \alpha < 1$ is significance level**
- **$\beta\hat{}_2 - \delta$ is lower confidence limit; $\beta\hat{}_2 + \delta$ is upper confidence limit**
- In practice α and $1 - \alpha$ are often expressed in percentage forms as 100α and $100(1 - \alpha)$ percent.

5-2. Interval estimation: Some basic Ideas

- Equation 5.2.1 shows that an **interval estimator, in contrast to a point estimator, is an** interval constructed in such a manner that it has a specified probability $1 - \alpha$ of including within its limits the true value of the parameter.
- For example, if $\alpha = 0.05$, or 5 percent, Eq. (5.2.1) would read: The probability that the (random) interval shown there includes the true β_2 is 0.95, or 95 percent.
- The interval estimator thus gives a range of values within which the true β_2 may lie.

5-2. Interval estimation: Some basic Ideas

- **Equation (5.2.1) does not mean that the Pr of β_2 lying between the given limits is (1 - α), but the Pr of constructing an interval that contains β_2 is (1 - α)**
- **($\hat{\beta}_2 - \delta$, $\hat{\beta}_2 + \delta$) is random interval**
- Since the confidence interval is random, the probability statements attached to it should be understood in the long-run sense, i.e, repeated sampling

5-2. Interval estimation: Some basic Ideas

- In repeated sampling, the intervals will enclose, in (1 - α)*100 of the cases, the true value of the parameters
- For a specific sample, we *can not* say that the probability is (1 - α) that *a given fixed interval* includes the true β_2
- If the sampling or probability distributions of the estimators are known, one can make confidence interval statement like (5.2.1)

5-3.Confidence Intervals for Regression coefficients

- With the normality assumption for *ui, the OLS* estimators $\beta\hat{}_1$ *and* $\beta\hat{}_2$ *are also normally distributed with means and variances given* therein.
- $Z = (\beta\hat{}_2 - \beta_2)/se(\beta\hat{}_2) = (\beta\hat{}_2 - \beta_2) / \sqrt{\sum x^2_i}/\sigma \sim N(0,1)$
  - (5.3.1)
- As noted in Eq. (4.3.6), Z is a standardized normal variable.
- **=>** we can use the normal distribution to make probabilistic statements about β_2 *provided the true population* variance σ^2 *is known.*

5-3.Confidence Intervals for Regression coefficients

- If σ^2 is known, an important property of a normally distributed variable with mean μ and variance σ^2 is that the area under the normal curve between
  - **$\mu \pm \sigma$ is about 68 percent,**
  - **between the limits $\mu \pm 2\sigma$ is about 95 percent,**
  - ***between $\mu \pm 3\sigma$ is about 99.7 percent.***
- But, we do not know σ and have to use $\sigma\hat{}$, *so Eq. 5.3.1 may be written as:*
- **$t = (\hat{\beta}_2 - \beta_2)/se(\hat{\beta}_2) = (\hat{\beta}_2 - \beta_2)\sqrt{\sum x_i^2}/\hat{\sigma} \sim t(n-2)$**
  - (5.3.2)
- Estimator − Parameter / Estimated SE of estimator

5-3.Confidence Intervals for Regression coefficients

- It can be shown that this *t variable follows the t distribution with n − 2 df.* [Note the difference between Eqs. (5.3.1) and (5.3.2).]
- **=> Interval for β_2**
- **$Pr[-t_{\alpha/2} \leq t \leq t_{\alpha/2}] = 1-\alpha$ (5.3.3)**
- where the t value is the t value given by Eq 5.3.2 and where $t_{\alpha/2}$ is the value of the t variable obtained from the t distribution for α/2 level of significance and n−2 df;
- This is called the **critical t value at α/2 level of** significance

5-3.Confidence Intervals for Regression coefficients

- Substitution of Eq. (5.3.2) into Eq. 5.3.3 yields Eq. 5.3.4. Rearranging Eq. 5.3.4, we obtain, which is **confidence interval for β_2 :**
- **$\Pr[\hat{\beta}_2 - t_{\alpha/2}\,se(\hat{\beta}_2) \leq \beta_2 \leq \hat{\beta}_2 + t_{\alpha/2}\,se(\hat{\beta}_2)] = 1-\alpha$** **(5.3.5)**
- Eq. 5.3.5 provides a 100(1 – *α) percent* ***confidence interval for Β2, which can be*** written more compactly as
- $\beta_2 \pm t_{\alpha/2}\,se(\hat{\beta}_2)$ **(5.3.6)**

5-3.Confidence Intervals for Regression coefficients

- Now using Eqs. (4.3.1) and (4.3.2), we can write **Confidence Interval for β_1 as:**
- **$Pr\ [\hat{\beta}_1 - t_{\alpha/2} se(\hat{\beta}_1) \leq \beta_1 \leq \hat{\beta}_1 + t_{\alpha/2} se(\hat{\beta}_1)] = 1 - \alpha$ (5.3.7)**
- or, more compactly, 100(1 – α)% confidence interval for β1:
- $\beta_1 \pm t_{\alpha/2} se(\hat{\beta}_1)$ **(5.3.8)**

5-3.Confidence Intervals for Regression coefficients

- Imp. feature of the confidence intervals given in Eq 5.3.6 and 5.3.8: In both cases the width of the confidence interval is proportional to the SE of the estimator.
- That is, the larger the SE, the larger is the width of the confidence interval.
- Larger the SE of the estimator, the greater is the uncertainty of estimating the true value of the unknown parameter.
- Thus, the SE of an estimator - measure of the **precision of the** estimator (i.e., how precisely the estimator measures the true population value).

Ex: Wage Education Eq.

- Mean hourly wages (Y) on education (X), recall that $\hat{\beta}2 = 0.7240$; se $(\hat{\beta}2) = 0.0700$.
- Since there are 13 observations, the df are 11. Assuming $\alpha = 5\%$, i.e, a 95% confidence coefficient, then the t table shows that for 11 df the **critical** $t\alpha/2 = 2.201$.
- Substituting these values in Eq. (5.3.5), we can verify that the 95 percent confidence interval for $\beta 2$ is as follows:
- $0.5700 \leq \beta 2 \leq 0.8780$ **(5.3.9)**
- Or, using Eq. (5.3.6), it is
- $0.7240 \pm 2.201(0.0700)$ **=>** 0.7240 ± 0.1540 **(5.3.10)**

Ex: interpretation of confidence interval

- **Given the confidence coefficient of** 95 percent, in 95 out of 100 cases intervals like Eq. 5.3.9 will contain the true β2.
- But, as warned earlier, we cannot say that the probability is 95 percent that the specific interval in Eq. (5.3.9) contains the true β2 because this interval is now fixed and no longer random; therefore β2 either lies in it or it does not:
- The probability that the specified fixed interval includes the true β2 is therefore 1 or 0.

Ex: interpretation of confidence interval

- Following Eq. (5.3.7), and the estimates that we derived we can easily verify that the 95 percent confidence interval for β1 for our example is
- $-1.8871 \leq \beta 1 \leq 1.8583$ **(5.3.11)**
- Again one should be careful in interpreting this confidence interval.
- In 95 out of 100 cases, intervals like Eq 5.3.11 will contain the true β1; the probability that this particular fixed interval includes the true β1 is either 1 or 0.

5-5. Hypothesis Testing: General Comments

- The problem of statistical hypothesis testing may be stated simply as follows: Is a given observation or finding compatible with some stated hypothesis or not?
- The word “compatible,” as used here, means “sufficiently” close to the hypothesized value so that we do not reject the stated hypothesis.
- Thus, if some theory or prior experience leads us to believe that the true slope coefficient β2 of the wages-education example is unity, is the observed ˆ β2 0.724 obtained from the sample of Table 3.2 consistent with the stated hypothesis?
- If it is, we do not reject the hypothesis; otherwise, we may reject it.

5-5. Hypothesis Testing: General Comments

- In the language of statistics, the stated hypothesis is known as the **null hypothesis and** is denoted by the symbol H0.
- The null hypothesis is usually tested against an **alternative hypothesis (also known as maintained hypothesis) denoted by H1, which may state, for** example, that true β2 is different from unity.
- The alternative hypothesis may be **simple or composite.**
- **For example, H1: β2 = 1.5 is a simple hypothesis, but H1: β2 = not 1.5 is a composite** hypothesis.

5-5. Hypothesis Testing: General Comments

- The theory of hypothesis testing - developing rules or procedures for deciding whether to reject or not reject the H0.
- 2 mutually complementary approaches - for devising such rules, **confidence interval and test of significance.**
- **Both these approaches predicate that the variable (statistic or estimator) under** consideration has some probability distribution and that hypothesis testing involves making statements or assertions about the value(s) of the parameter(s) of such distribution.
- For ex., we know that with the normality assumption $\hat{}$ β_2 is normally distributed with mean equal to β_2 and variance given by Eq. (4.3.5).
- If we hypothesize that $\beta_2 = 1$, we are making an assertion about one of the parameters of the normal distribution, i,.e., the mean.

5-5. Hypothesis Testing: General Comments

- Most of the statistical hypotheses will be of this type—making assertions about one or more values of the parameters of some assumed probability distribution such as the normal, F, t, or $\chi 2$.
- The stated hypothesis is known as the null hypothesis: H_o
- The H_o is tested against and alternative hypothesis: H_1
- 5-6. Hypothesis Testing: The confidence interval approach

One-sided or one-tail Test

$H_0: \beta_2 \leq \beta^*$ versus $H_1: \beta_2 > \beta^*$

Two-sided or two-tail Test

$$H_0: \beta_2 = \beta^* \quad \text{versus} \quad H_1: \beta_2 \# \beta^*$$

$\hat{\beta}_2 - t_{\alpha/2} se(\hat{\beta}_2) \leq \beta_2 \leq \hat{\beta}_2 + t_{\alpha/2} se(\hat{\beta}_2)$ values of β_2 lying in this interval are plausible under H_o with $100*(1-\alpha)\%$ confidence.

- If β_2 lies in this region we do not reject H_o (the finding is statistically insignificant)
- If β_2 falls outside this interval, we reject H_o (the finding is statistically significant)

5-7. Hy. Testing: The test of significance approach

- A test of significance is a procedure by which sample results are used to verify the truth or falsity of a null hypothesis
- Testing the significance of regression coefficient: The t-test

$$\Pr[\hat{\beta}_2 - t_{\alpha/2}\,se(\hat{\beta}_2) \le \beta_2 \le \hat{\beta}_2 + t_{\alpha/2}\,se(\hat{\beta}_2)] = 1-\alpha \quad (5.7.2)$$

- which gives the interval in which $\hat{\beta}_2$ *will fall with* $1-\alpha$ *probability, given* $\beta_2 = \beta_2^*$.
- 100(1 – *α)% confidence interval established in Eq.* 5.7.2 is known as the **region of acceptance (of the null hy.) and the *region(s)*** outside the confidence interval is (are) called the **region(s) of rejection (of *H0) or the* critical region(s).**

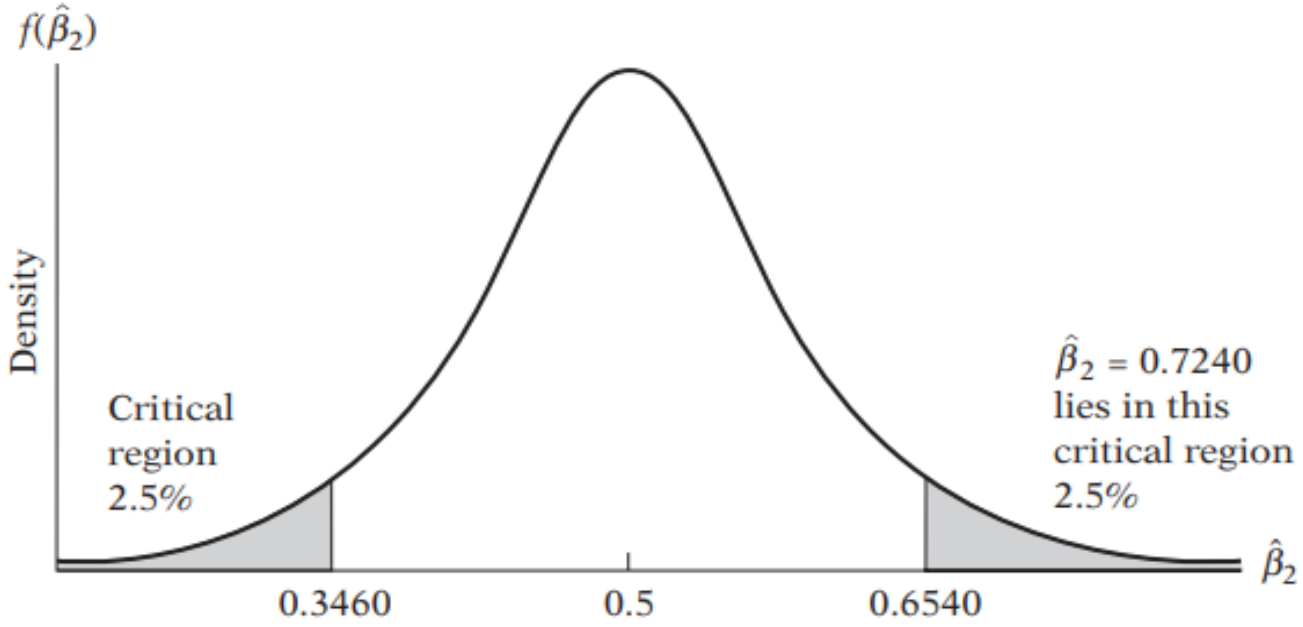
5-7. Hy. Testing: The test of significance approach

- **The confidence limits, the endpoints of the confidence** interval, are called **critical values.**
- There is an intimate connection between the confidence-interval and test-of-significance approaches to hypothesis testing.
- In the confidence-interval procedure we try to establish a range or an interval that has a certain probability of including the true but unknown β_2, whereas in the test-of-significance approach we hypothesize some value for β_2 and try to see whether the computed $\hat{\beta}_2$ lies within reasonable (confidence) limits around the hypothesized value.

EX: Wage Education Eq.

- We know that $\hat{\beta}_2$ = 0.7240, se ($\hat{\beta}_2$) = 0.0700, and df = 11. If we assume α = 5%, tα/2 = 2.201.
- If we assume H0: $\beta_2 = \beta^*_2$ = 0.5 and H1: β_2 = 0.5, then Eq. (5.7.2) becomes Pr (0.3460 ≤ ˆ β2 ≤ 0.6540) **(5.7.3)**
- as shown diagrammatically in Figure 5.3.

FIG 5.3: The 95% confidence interval for βˆ 2 under the hypothesis that β2 = 0.5

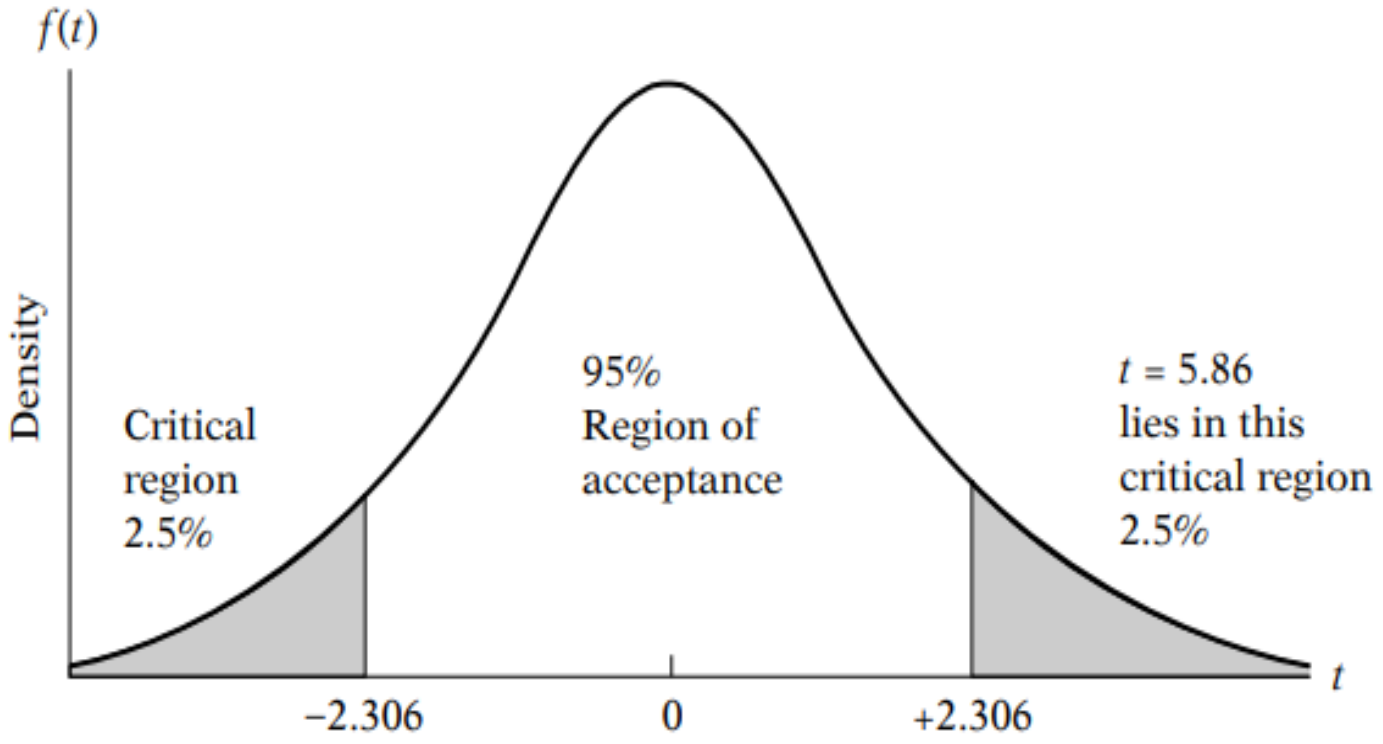


Source: Basic Econometrics by Domodar Gujarati, P.130

EX: Wage Education Eq.

- In practice, there is no need to estimate Eq. (5.7.2) explicitly.
- One can compute the t value in the middle of the double inequality given by Eq. (5.7.1) and see whether it lies between the critical t values or outside them.
- For our example,
- $t = 0.7240 - 0.5 / 0.0700 = 3.2$ ***(5.7.4)***
- which clearly lies in the critical region of Figure 5.4.
- The conclusion remains the same; namely, we reject *H0.*

FIG 5.4 The 95% confidence interval for t(8 df)

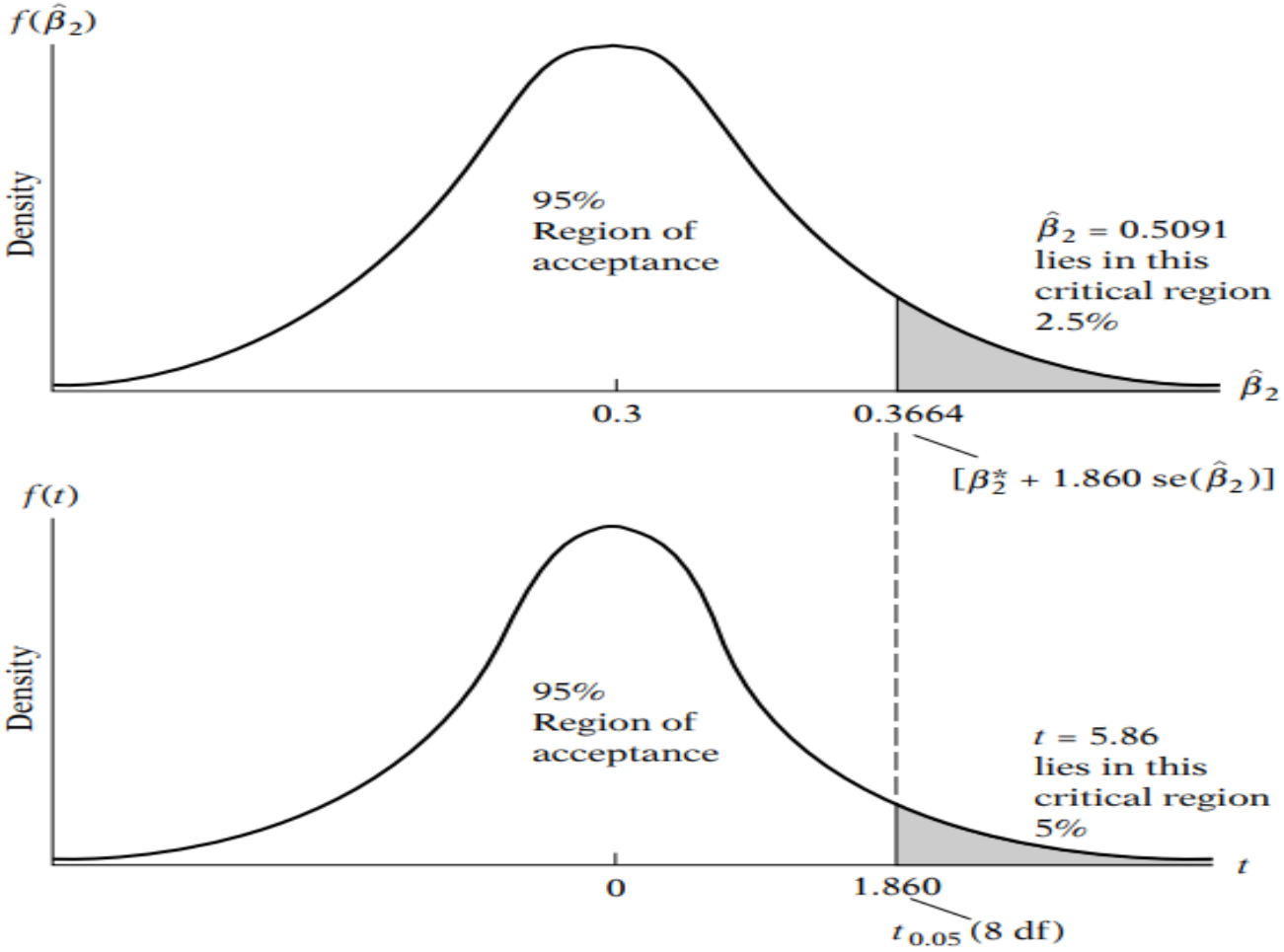


Source: Basic Econometrics by Domodar Gujarati, P.131

5-7. Hy. Testing: The test of significance approach

- Since we use the *t distribution, this is called* the t test.
- In the language of significance tests, a statistic is said to be statistically significant if the value of the test statistic lies in the critical region.
- In this case the null hypothesis is rejected. Also, a test is said to be statistically insignificant if the value of the test statistic lies in the acceptance region.
- One tail test in fig 5.5

FIG 5.5 One-tail test of significance



Source: Basic Econometrics by Domodar Gujarati, P.132

Table 5-1: Decision Rule for t-test of significance

Type of Hypothesis	H_0	H_1	Reject H_0 if
Two-tail	$\beta_2 = \beta_2^*$	$\beta_2 \# \beta_2^*$	$\lvert t \rvert > t_{\alpha/2,df}$
Right-tail	$\beta_2 \leq \beta_2^*$	$\beta_2 > \beta_2^*$	$t > t_{\alpha,df}$
Left-tail	$\beta_2 \geq \beta_2^*$	$\beta_2 < \beta_2^*$	$t < -t_{\alpha,df}$

5-7. Testing the significance of σ^2 : The χ^2 Test

- Under the Normality assumption we have:

$$\chi^2 = (n-2)\frac{\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{(n-2)} \qquad (5.4.1)$$

- This test procedure is called the **chi-square test of significance.**
- **The *χ2 test of significance approach to hypothesis testing*** is summarized in Table 5.2.

Table 5-2: A summary of the χ^2 Test

H_0	H_1	Reject H_0 if
$\sigma^2 = \sigma^2_0$	$\sigma^2 > \sigma^2_0$	$Df.(\sigma\text{^}2)/\ \sigma^2_0 > \chi^2_{\alpha,df}$
$\sigma^2 = \sigma^2_0$	$\sigma^2 < \sigma^2_0$	$Df.(\sigma\text{^}2)/\ \sigma^2_0 < \chi^2_{(1-\alpha),df}$
$\sigma^2 = \sigma^2_0$	$\sigma^2 \# \sigma^2_0$	$Df.(\sigma\text{^}2)/\ \sigma^2_0 > \chi^2_{\alpha/2,df}$ or $< \chi^2_{(1-\alpha/2),\ df}$

5-8. Hypothesis Testing: Some practical aspects

1) The meaning of "Accepting" or "Rejecting" a Hypothesis:

- It is always preferable to say that we *may accept the null hypothesis rather than we (do) accept it.*
- . . just as a court pronounces a verdict as "not guilty" rather than "innocent," so the conclusion

2) **The Null Hypothesis and the '2-t' Rule of Thumb:**

"If the number of degrees of freedom is 20 or more and if α, the level of significance, is set at 0.05, then the null hypothesis $\beta_2 = 0$ can be rejected if the t value [= $\hat{\beta}_2 / se(\hat{\beta}_2)$] computed from Eq. (5.3.2) exceeds 2 in absolute value".

5-8. Hypothesis Testing: Some practical aspects

3. Forming the Null and Alternative Hypotheses:

- theoretical expectations or prior empirical work or both can be relied upon to formulate hypotheses. But no matter how the hypotheses are formed, it is extremely important that the researcher establish these hypotheses before carrying out the empirical investigation.

4. **Choosing α, the Level of Significance:**

- whether we reject or do not reject the null hypothesis depends critically on α, the level of significance or the probability of committing a **Type I error—the probability of rejecting the true hypothesis.**
- **Type II error (**the probability of accepting the false hypothesis) . Classical statistics generally concentrates on a Type I error.
- Fortunately, the dilemma of choosing the appropriate value of α can be avoided by using what is known as the **p value of the test statistic.**

5-8. Hypothesis Testing: Some practical aspects

5) **The Exact Level of Significance: The p-Value**

- p value (i.e., probability value), also known as the observed or exact level of significance or the exact probability of committing a Type I error.
- More technically, the p value is defined as the lowest significance level at which a null hypothesis can be rejected.
- If the researcher wants to choose a p value of about 0.02 percent and not take a chance of being wrong more than 2 out of 10,000 times.
- In an application the *p value of a test statistic happens to be, say, 0.145, or 14.5 percent, and if* the reader wants to reject the null hypothesis at this (exact) level of significance, so be it.
- Nothing is wrong with taking a chance of being wrong 14.5 percent of the time if you reject the true null hypothesis.

5-8. Hypothesis Testing: Some practical aspects

6. Statistical Significance versus Practical Significance:

- one should not confuse statistical significance with practical, or economic, significance.
- As sample size becomes very large, issues of statistical significance become much less important but issues of economic significance become critical.

7. The Choice between Confidence-Interval and Test-of-Significance Approaches to Hypothesis Testing

5-9. Regression Analysis and Analysis of Variance

- Regression analysis from the point of view of the analysis of variance (ANOVA) and complementary way of looking at the statistical inference problem
- TSS = ESS + RSS
- F =[MSS of ESS]/[MSS of RSS] =

$= \beta_2 \text{^} 2 \sum x_i^2 / \sigma \text{^} 2$ (5.9.1)

- If u_i are normally distributed; H_0: $\beta_2 = 0$ then F follows the F distribution with 1 and n-2 degree of freedom

5-9. Regression Analysis and Analysis of Variance

- What use can be made of the preceding *F ratio?*
- F provides a test statistic to test the null hypothesis that true β_2 is zero by comparing this F ratio with the F-critical obtained from F tables at the chosen level of significance, or obtain the p-value of the computed F statistic to make decision.

Table 5-3. ANOVA for two-variable regression model

Source of Variation	Sum of square (SS)	Degree of Freedom -(Df)	Mean sum of square (MSS)
ESS (due to regression)	$\sum \hat{y}_i^2 = \hat{\beta}_2^2 \sum x_i^2$	1	$\hat{\beta}_2^2 \sum x_i^2$
RSS (due to residuals)	$\sum \hat{u}_i^2$	n-2	$\sum \hat{u}_i^2 /(n-2)=\hat{\sigma}^2$
TSS	$\sum y_i^2$	n-1	

EX. ANOVA for two-variable regression model: Wage-Edn Eq.

Source of Variation	Sum of square (SS)	Degree of Freedom - (Df)	Mean sum of square (MSS)
ESS (due to regression)	95.4255	**1**	95.4255
RSS (due to residuals)	9.6928	**11**	0.8811
TSS	105.1183	**12**	

F = 95.4255 /0.8811 = 108.3026

Wage-Edn Eq.'s F statistics

- The p value of this F statistic corresponding to 1 and 11 df can be obtained from the F table given in Appendix D, but by using electronic statistical tables it can be shown that the p value is 0.0000001, an extremely small probability indeed.
- If we decide to choose the level-of significance approach to hypothesis testing and fix α at 0.01, or a 1 percent level, we can see that the computed F of 108.3026 is obviously significant at this level.
- Therefore, if we reject the null hypothesis that $\beta_2 = 0$, the probability of committing a Type I error is very small.
- For all practical purposes, our sample could not have come from a population with zero β_2 value and we can conclude with great confidence that X, education, does affect Y, average wages.

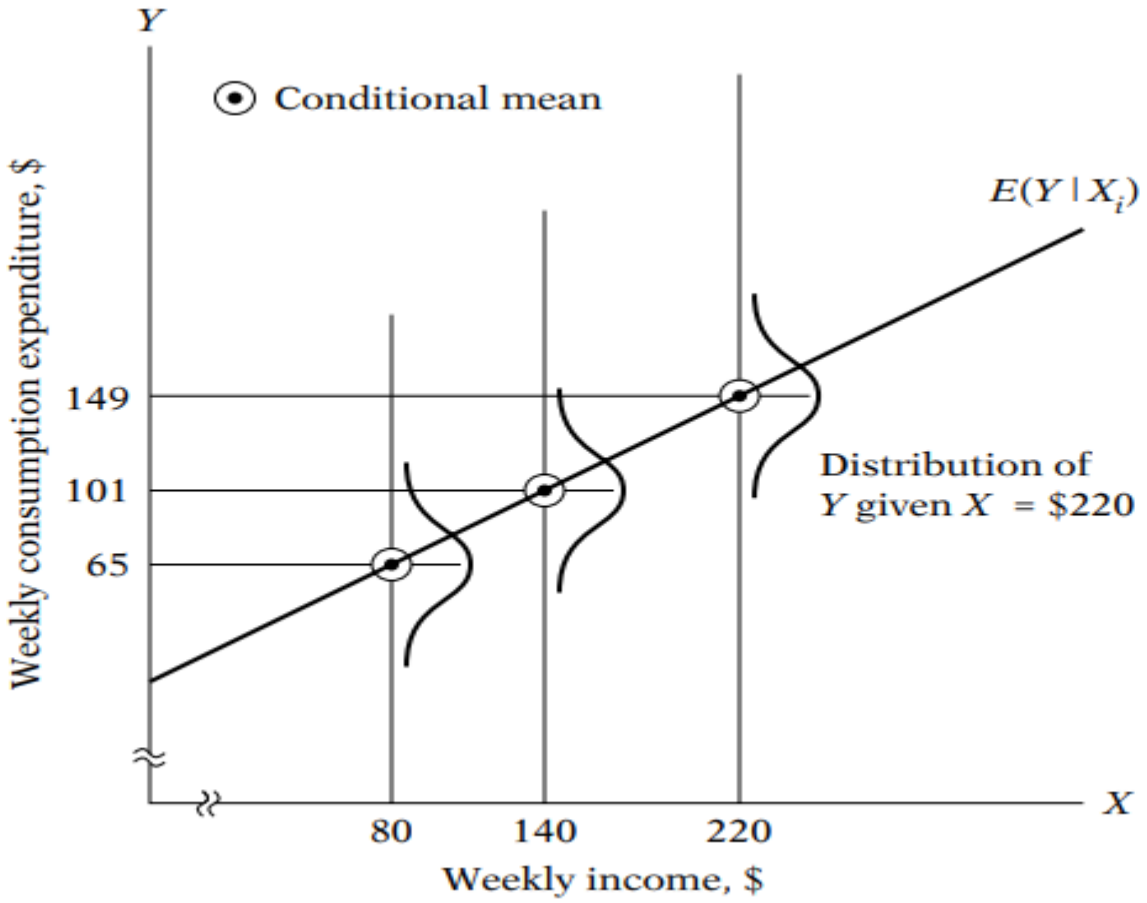
5-10. Application of Regression Analysis: Problem of Prediction

- By the data of Table 3-2, we obtained the sample regression (3.6.2) :
  $\hat{Y}_i = 24.4545 + 0.5091X_i$, where
  $\hat{Y}_i$ is the estimator of true $E(Y_i)$
- There are two kinds of prediction: Mean Prediction and Individual prediction
- Mean prediction: Prediction of the conditional mean value of Y corresponding to a chosen X, say X_0, that is the point on the population regression line itself
- Individual prediction: Prediction of an individual Y value corresponding to X_0

5-10. Application of Regression Analysis: Mean Prediction

- On the basis of the sample data of Table 3.2 we obtained the following sample regression:
- $\hat{Y}_i = -0.0144 + 0.7240X_i$ **(3.6.1)**
- where $\hat{Y}_i$ is the estimator of true $E(Y_i)$ corresponding to given X.
- Mean Prediction: prediction of the conditional mean value of Y corresponding to a chosen X, say, X0, that is the point on the population regression line itself (see Figure 2.2).

FIG 2.2 Population regression line (data of Table 2.1)



Source: Basic Econometrics by Domodar Gujarati, P.40

5-10. Application of Regression Analysis: Mean Prediction

- To fix the ideas, assume that X0 = 20 and we want to predict E(Y | X0 = 20).
- Now it can be shown that the historical regression in Eq. (3.6.1) provides the point estimate of this mean prediction as follows:
- $\hat{Y}_0 = \hat{\beta}_1 + \hat{\beta}_2 X_0$
- = −0.0144 + 0.7240(20) **(5.10.1)**
- = 14.4656
- where $\hat{Y}_0$ = estimator of $E(Y|X_0)$. It can be proved that this point predictor is a best linear unbiased estimator (BLUE).
- Since $\hat{Y}_0$ is an estimator, it is likely to be different from its true value. The difference between the two values will give some idea about the prediction or forecast error.
- To assess this error, we need to find out the sampling distribution of $\hat{Y}_0$.

5-10. Application of Regression Analysis: Mean Prediction

- $\hat{Y}_0$ in Equation 5.10.1 is normally distributed with mean $(\beta_1 + \beta_2 X_0)$ and the variance is given by the following formula:
- **$\text{var}(\hat{Y}_0) = \sigma^2 [1/n + (X_0 - \bar{X})^2 / \sum x_i^2]$ (5.10.2)**
- By replacing the unknown σ2 by its unbiased estimator $\hat{\sigma}2$, we see that the variable
- **$t = \hat{Y}_0 - (\beta_1 + \beta_2 X_0) / \text{se}(\hat{Y}_0)$ (5.10.3)**
- follows the t distribution with n − 2 df. The t distribution can be used to derive confidence intervals for the true $E(Y_0 | X_0)$ and test hypotheses about it in the usual manner,

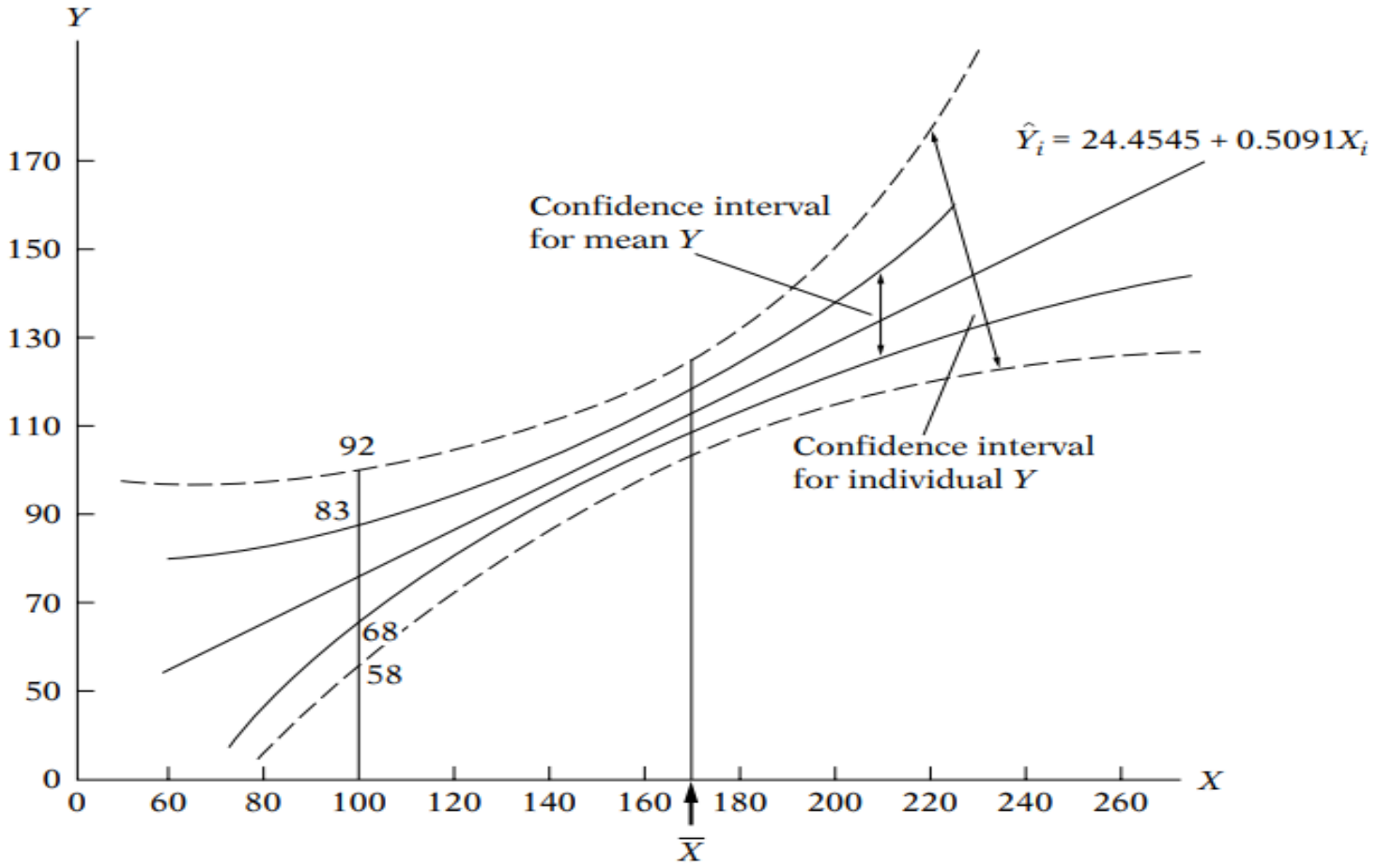
$\Pr [\hat{\beta}_1 + \hat{\beta}_2 X_0 - t_{\alpha/2} \text{se}(\hat{Y}_0) \leq \beta_1 + \beta_2 X_0 \leq \hat{\beta}_1 + \hat{\beta}_2 X_0 + t_{\alpha/2} \text{se}(\hat{Y}_0)] = 1 - \alpha$

- **(5.10.4)**
- where $\text{se}(\hat{Y}_0)$ is obtained from Eq. (5.10.2).
- For our data (see Table 3.2), $\text{var}(\hat{Y}0) = 0.8936 [1/13 + (20-12)^2/182]$
- $= 0.3826$ *and* $\text{se}(\hat{Y}_0) = 0.6185$

5-10. Application of Regression Analysis: Mean Prediction

- Therefore, the 95 percent confidence interval for true $E(Y/X_0) = \beta_1 + \beta_2 X_0$ is given by
- $14.4656 - 2.201(.6185) \leq E(Y_0 / X = 20) \leq 14.4656 + 2.20(0.6185)$
- that is, $13.1043 \leq E(Y \mid X = 20) \leq 15.8260$ (5.10.5)
- Thus, given $X_0 = 100$, in repeated sampling, 95 out of 100 intervals like Equation 5.10.5 will include the true mean value; the single best estimate of the true mean value is of course the point estimate 14.4656.
- If we obtain 95 percent confidence intervals like Eq. (5.10.5) for each of the X values given in Table 3.2, we obtain what is known as the **confidence interval, or confidence band, for the population regression function, which is shown in Figure 5.6.**

FIG 5.6 Confidence intervals (bands) for mean Y and individual Y values



Source: Basic Econometrics by Domodar Gujarati, P.144

5-10. Application of Regression Analysis: Individual Prediction

- If our interest lies in predicting an individual Y value, Y_0, corresponding to a given X value, say, X_0, then estimate var(Y_0 - $Y\hat{}_0$) and t stat and Pr
- Notice an important feature of the confidence bands shown in Figure 5.6. The width of these bands is smallest when X_0 = X bar (Why?)
- However, the width widens sharply as X_0 moves away from Xbar. (Why?) This change would suggest that the predictive ability of the historical sample regression line falls markedly as X_0 departs progressively from X bar.
- Therefore, one should exercise great caution in "extrapolating" the historical regression line to predict $E(Y| X_0)$ or Y_0 associated with a given X_0 that is far removed from the sample mean X bar.

5-13. Summary and Conclusions

1. Estimation and Hypothesis testing constitute the two main branches of classical statistics
2. Hypothesis testing answers this question: Is a given finding compatible with a stated hypothesis or not?
3. There are two mutually complementary approaches to answering the preceding question: Confidence interval and test of significance.

5-13. Summary and Conclusions

4. Confidence-interval approach has a specified probability of including within its limits the true value of the unknown parameter. If the null-hypothesized value lies in the confidence interval, H_0 is not rejected, whereas if it lies outside this interval, H_0 can be rejected
5. significance test procedure develops a test statistic which follows a well-defined probability distribution (like normal, t, F, or Chi-square). Once a test statistic is computed, its p-value can be easily obtained.

   The p-value The p-value of a test is the lowest significance level, at which we would reject H_0. It gives exact probability of obtaining the estimated test statistic under H_0. If p-value is small, one can reject H_0, but if it is large one may not reject H_0.

5-13. Summary and Conclusions

6. Type I error is the error of rejecting a true hypothesis. Type II error is the error of accepting a false hypothesis. In practice, one should be careful in fixing the level of significance α, the probability of committing a type I error (at arbitrary values such as 1%, 5%, 10%). It is better to quote the p-value of the test statistic.
7. This chapter introduced the normality test to find out whether u_i follows the normal distribution. Since in small samples, the t, F, and Chi-square tests require the normality assumption, it is important that this assumption be checked formally

Reference

Chapter 5: TWO-VARIABLE REGRESSION: Interval Estimation and Hypothesis Testing, Basic Econometrics by Domodar Gujarati

What Next?

- Reporting and evaluating the results of estimated regression equation
- Examples and Estimating the t statistic in Excel and its interpretation
- Multiple Linear Regression