Two-Variable Regression: Interval Estimation and Hypothesis Testing

The theory of estimation consists of

- **Point estimation**

  A single value used to approximate a population parameter

- **Interval estimation**

  a range of values estimating the parameter

  (may or may not contain the value of the parameter being estimated)

Confidence level of an interval estimate of a parameter:

the probability that the interval estimate will contain the parameter, assuming that a large number of samples are selected and that the estimation process on the same parameter is repeated

E.g. 90% Confidence level

95% Confidence level

99% Confidence level

Interval Estimation

- We want to find how "close", say, $\hat{\beta}_2$ is to β_2
- We try to find out δ and α such that the probability that the **random interval** $(\hat{\beta}_2 - \delta, \hat{\beta}_2 + \delta)$ contains the true β_2 is $1-\alpha$
- Confidence interval is

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

Interval Estimation

- $1-\alpha$ is known as confidence coefficient
- α is known as the level of significance

E.g. 95% confidence level

α = 1-0.95=0.05

α *is level of significance*

- The probability of committing a type I error
- A type I error consists in rejecting a true hypothesis

$$\alpha = 0.05$$

$$t_{\alpha/2, df=10-2} =$$

$$t_{\alpha, df=10-2} =$$

$$t_{\alpha/2, df=200-2} =$$

$$t_{\alpha, df=200-2} =$$

$$\alpha = 0.05$$

$$t_{\alpha/2, df=10-2} = 2.306$$

$$t_{\alpha, df=10-2} = 1.860$$

$$t_{\alpha/2, df=200-2} = 1.96$$

$$t_{\alpha, df=200-2} = 1.645$$

$$\alpha = 0.01$$

$$t_{\alpha/2,df=10-2} =$$

$$t_{\alpha,df=10-2} =$$

$$t_{\alpha/2,df=200-2} =$$

$$t_{\alpha,df=200-2} =$$

$$\alpha = 0.01$$

$$t_{\alpha/2, df=10-2} = 3.355$$

$$t_{\alpha, df=10-2} = 2.896$$

$$t_{\alpha/2, df=200-2} = 2.576$$

$$t_{\alpha, df=200-2} = 2.326$$

Confidence intervals for regression coefficients β_1 and β_2

- With the normality assumption for u_i
- The OLS estimators β_1 and β_2 are themselves normally distributed with means and variances
- Variance (σ^2) is known

$$Z = \frac{\hat{\beta}_2 - \beta_2}{se(\hat{\beta}_2)}$$

$$= \frac{(\hat{\beta}_2 - \beta_2)\sqrt{\sum x_i^2}}{\sigma}$$

Confidence intervals for regression coefficients β_1 and β_2

- σ^2 is rarely known, in practice it is determined by the unbiased estimator $\hat{\sigma}^2$

$$t = \frac{\hat{\beta}_2 - \beta_2}{se(\hat{\beta}_2)}$$

$$= \frac{(\hat{\beta}_2 - \beta_2)\sqrt{\sum x_i^2}}{\hat{\sigma}}$$

Confidence intervals for regression coefficients β_1 and β_2

$$\Pr(-t_{\alpha/2} \le t \le t_{\alpha/2}) = 1-\alpha$$

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

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

Confidence intervals for regression coefficients β_1 and β_2

$100(1-\alpha)\%$ *confidence interval for* β_2 :

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

TABLE 2.6
Mean Hourly Wage by Education

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

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

Confidence intervals for regression coefficients β_1 and β_2

Example: Mean hourly wages (Y) on education (X)

$\hat{\beta}_2 = 0.7240$

$se(\hat{\beta}_2) = 0.0700$

$n = 13, df = n - 2 = 11, \alpha = 0.05$

$1 - \alpha = 0.95$

$critical\ t_{\alpha/2} = 2.201$

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

Confidence intervals for regression coefficients β_1 and β_2

$$0.5700 \leq \beta_2 \leq 0.8780$$

or

$$0.7240 \pm 2.201(0.0700)$$

Given the confidence coefficient of 95 percent, the true parameter is between 0.5700 and 0.8780.

Confidence intervals for regression coefficients β_1 and β_2

The **correct interpretation** of this confidence interval is:

Given the confidence coefficient of 95 percent, in 95 out of 100 cases intervals will contain the true β_2

Confidence intervals for regression coefficients β_1 and β_2

The **wrong interpretation** of this confidence interval is:

The probability is 95 percent that the specific interval contains the true β_2

Since the interval is now fixed and no longer random; therefore β_2 either lies in it or it does not

Class exercise

Example: Mean hourly wages (Y) on education (X)

$\hat{\beta}_2 = 0.7240$

$se(\hat{\beta}_2) = 0.0700$

$n = 13, df = n - 2 = 11, \alpha = 0.01$

$1 - \alpha = 0.99$

$critical\ t_{\alpha/2} =$

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

Confidence intervals for regression coefficients β_1 and β_2

$$0.7240 \pm 3.106(0.0700)$$

Given the confidence coefficient of 99 percent, the true parameter is between 0.7240-3.106(0.0700) and 0.7240+3.106(0.0700).

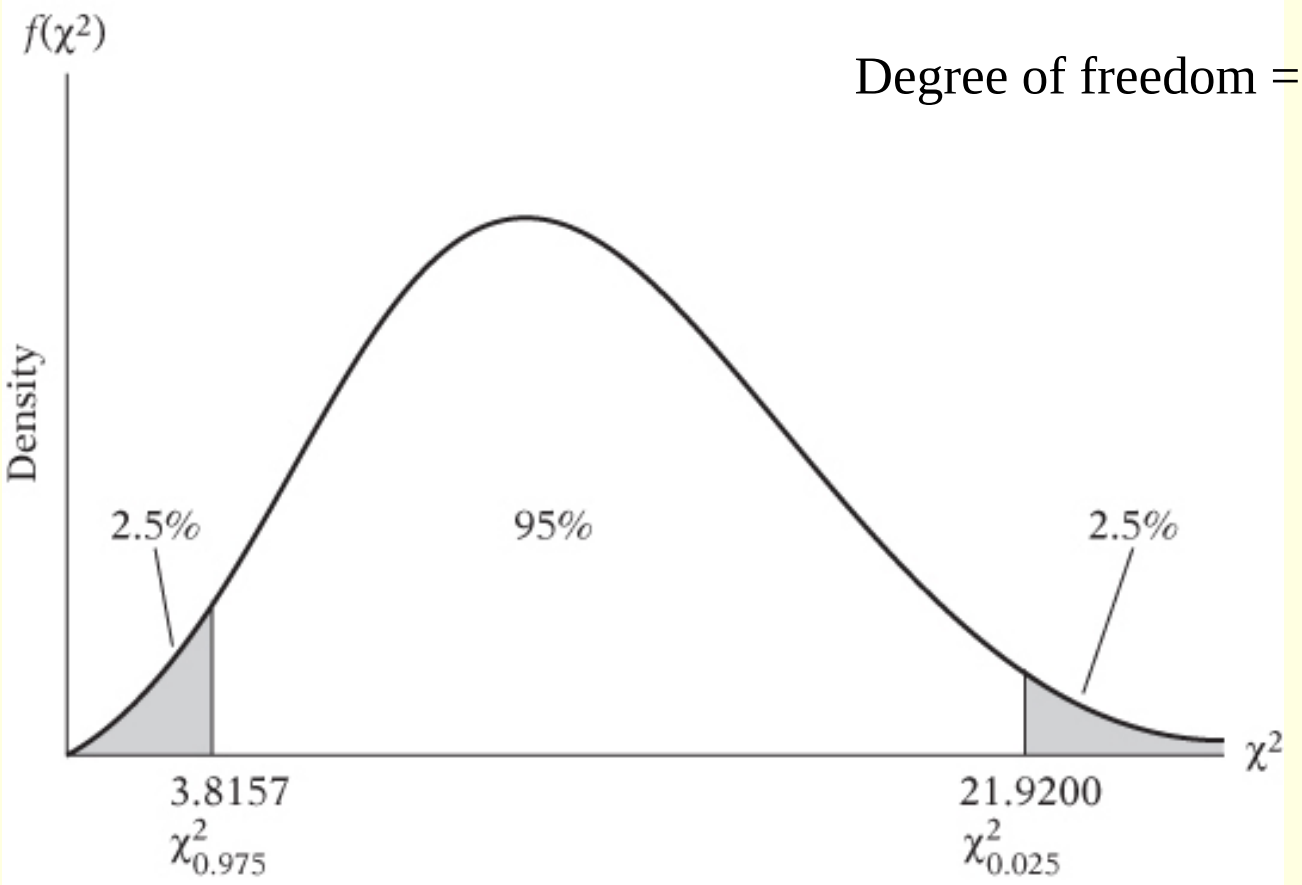
Confidence intervals for σ^2

Under the normality assumption, the variable follows the χ^2 distribution with n-2 df

$$\chi^2 = (n-2)\frac{\hat{\sigma}^2}{\sigma^2}$$

Confidence intervals for σ^2

Degree of freedom = n-2 =11



$$\alpha = 0.05$$

$$\chi_{1-\alpha/2,df=10-2} =$$

$$\chi_{\alpha/2,df=10-2} =$$

$$\chi_{1-\alpha/2,df=200-2} =$$

$$\chi_{\alpha/2,df=200-2} =$$

$$\alpha = 0.05$$

$$\chi_{1-\alpha/2,df=10-2} = 2.1797$$

$$\chi_{\alpha/2,df=10-2} = 17.5346$$

$$\chi_{1-\alpha/2,df=200-2} = 74.2219$$

$$\chi_{\alpha/2,df=200-2} = 129.561$$

$$\alpha = 0.01$$

$$\chi_{1-\alpha/2,df=10-2} =$$

$$\chi_{\alpha/2,df=10-2} =$$

$$\chi_{1-\alpha/2,df=200-2} =$$

$$\chi_{\alpha/2,df=200-2} =$$

$$\alpha = 0.01$$

$$\chi_{1-\alpha/2,df=10-2} = 1.3444$$

$$\chi_{\alpha/2,df=10-2} = 21.9550$$

$$\chi_{1-\alpha/2,df=200-2} = 67.3276$$

$$\chi_{\alpha/2,df=200-2} = 140.169$$

Confidence intervals for σ^2

$$\Pr(\chi^2_{1-\alpha/2} \le \chi^2 \le \chi^2_{\alpha/2}) = 1-\alpha$$

$$\Pr[(n-2)\frac{\hat{\sigma}^2}{\chi^2_{\alpha/2}} \le \sigma^2 \le (n-2)\frac{\hat{\sigma}^2}{\chi^2_{1-\alpha/2}}] = 1-\alpha$$

Which gives the $100(1-\alpha)\%$ confidence interval for σ^2

Confidence intervals for σ^2

Example: Wage-education

$$\hat{\sigma}^2 = 0.8936$$

$$\alpha = 0.05, df = n - 2 = 11$$

$$\chi^2_{0.025} = 21.9200, \chi^2_{0.975} = 3.8157$$

$$0.4484 \leq \sigma^2 \leq 2.5760$$

Given the confidence coefficient of 95 percent, the true parameter is between 0.4484 and 2.5760

Confidence intervals for σ^2

The interpretation of this interval is:

If we established 95 percent confidence limits on σ^2

And if we maintain a priori that these limits will

Include the true σ^2, we will be right in the long

run 95 percent of the time

Class exercise

Example: Wage-education

$$\hat{\sigma}^2 = 0.8936$$

$$\alpha = 0.01, df = n - 2 = 11$$

$$\chi^2_{0.025} = \_\_\_\_\_, \chi^2_{0.975} = \_\_\_\_\_$$

$$\_\_\_\_\_ \leq \sigma^2 \leq \_\_\_\_\_$$

Given the confidence coefficient of 99 percent, the true parameter is between _____ and _____

$$\Pr[(n-2)\frac{\hat{\sigma}^2}{\chi^2_{\alpha/2}} \le \sigma^2 \le (n-2)\frac{\hat{\sigma}^2}{\chi^2_{1-\alpha/2}}] = 1-\alpha$$

Example: Wage-education

$$\hat{\sigma}^2 = 0.8936$$

$$\alpha = 0.01, df = n-2 = 11$$

$$\chi^2_{0.005} = 26.7569, \chi^2_{0.995} = 2.6032$$

$$11\frac{0.8936}{26.7569} \le \sigma^2 \le 11\frac{0.8936}{2.6032}$$

Given the confidence coefficient of 99 percent, the true parameter is between _____ and _____

Hypothesis Testing

Hypothesis Testing

In statistics, a hypothesis is a claim or statement about a property of a population.

A hypothesis test (or test of significance) is a standard procedure for testing a claim about a property of a population.

Hypothesis Testing

1. Define population in study
2. State the hypothesis to be investigated
3. Give the desired significance level
4. Select a sample from population
5. Collect the data
6. Perform the calculations
7. Reach a conclusion

Hypothesis Testing

- **Null Hypothesis** (H_0):

  A Statistical hypothesis stating there is no difference between a parameter and a specific value
- We test the null hypothesis directly
- Either reject H_0 or fail to reject H_0

Hypothesis Testing

- **Alternative hypothesis**: (H_1):

  Stating the existence of a difference between a parameter and a specific value.

- The symbolic form of the alternative hypothesis must use one of these symbols: ≠, <, >.

TABLE 5.1
The *t* Test of Significance: Decision Rules

Type of Hypothesis	H_0: The Null Hypothesis	H_1: The Alternative Hypothesis	Decision Rule: Reject H_0 If
Two-tail	$\beta_2 = \beta_2^*$	$\beta_2 \neq \beta_2^*$	$\lvert t \rvert > t_{\alpha/2,\text{df}}$
Right-tail	$\beta_2 \leq \beta_2^*$	$\beta_2 > \beta_2^*$	$t > t_{\alpha,\text{df}}$
Left-tail	$\beta_2 \geq \beta_2^*$	$\beta_2 < \beta_2^*$	$t < -t_{\alpha,\text{df}}$

Notes: β_2^* is the hypothesized numerical value of β_2.
$|t|$ means the absolute value of t.
t_α or $t_{\alpha/2}$ means the critical t value at the α or $\alpha/2$ level of significance.
df: degrees of freedom, $(n - 2)$ for the two-variable model, $(n - 3)$ for the three-variable model, and so on.
The same procedure holds to test hypotheses about β_1.

Hypothesis Testing: The confidence-Interval Approach

Two-Sided or Two-Tail Test

$$H_0 : \beta_2 = 0.5$$

$$H_1 : \beta_2 \neq 0.5$$

Therefore, if β_2 under H_0 falls within the $100(1-\alpha)\%$ Confidence interval, we do not reject the null hypothesis; if it lies outside the interval, we may reject it

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

$-t_{\alpha/2}$ and $t_{\alpha/2}$ are the value of t (the critical t value)

The Test-of-Significance of Regression Coefficients: The t Test

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

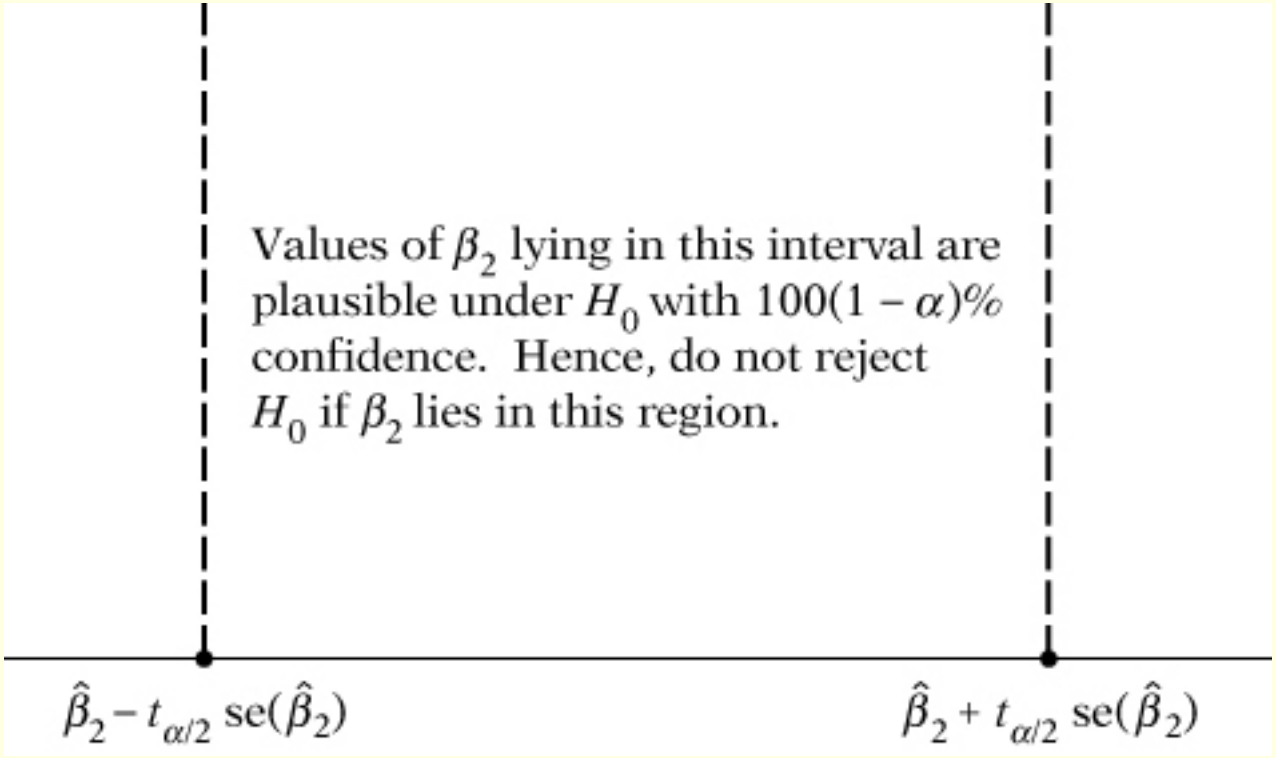
Region of acceptance

The $100(1-\alpha)\%$ confidence interval

Region of rejection

The regions outside the confidence interval

Hypothesis Testing: The confidence-Interval Approach



Hypothesis Testing: The confidence-Interval Approach

- When we reject the null hypothesis, we say that our finding is **statistically significant**
- When we do not reject the null hypothesis, we say that our finding is **not statistically significant**

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

The Test-of-Significance of Regression Coefficients: The t Test

Under the normality assumption the variable

$$t = \frac{\hat{\beta}_2 - \beta_2}{se(\hat{\beta}_2)}$$

$$= \frac{(\hat{\beta}_2 - \beta_2)\sqrt{\sum x_i^2}}{\hat{\sigma}} = \frac{(\hat{\beta}_2 - \beta_2)\sqrt{\sum (X_i - \bar{X})^2}}{\hat{\sigma}}$$

follows the t distribution with n-2 df

The Test-of-Significance of Regression Coefficients: The t Test

Example: Wage-education

$$\hat{\beta}_2 = 0.7240$$

$$se(\hat{\beta}_2) = 0.0700$$

$$df = 11, \alpha = 0.05, t_{\alpha/2} = 2.201$$

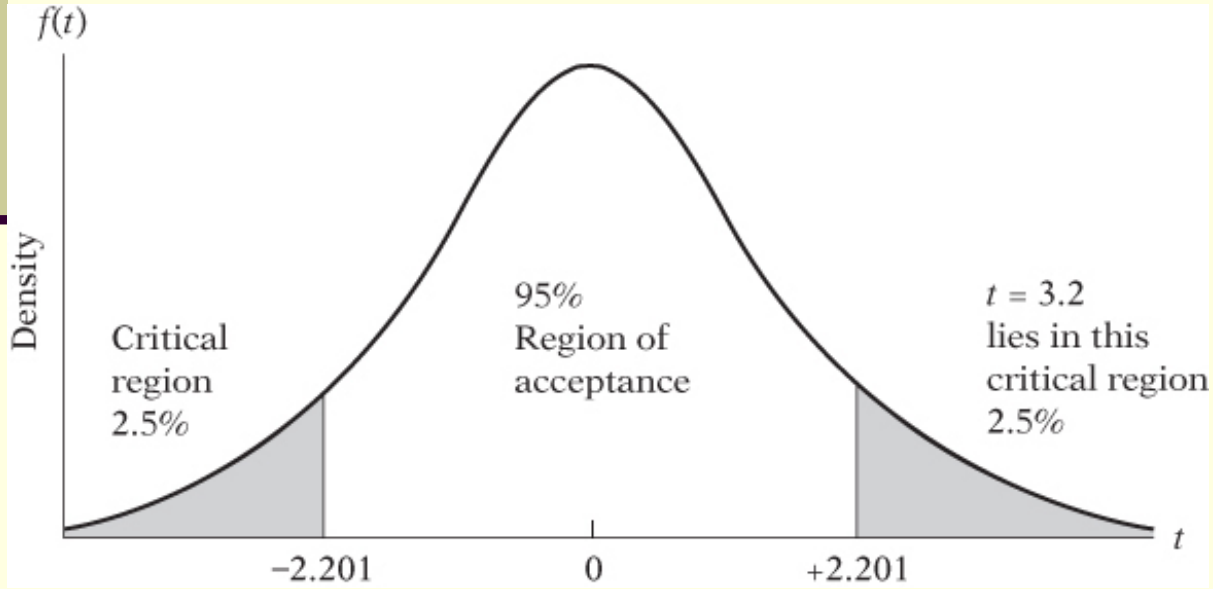
$$H_0 : \beta_2 = 0.5$$

$$H_1 : \beta_2 \neq 0.5$$

The Test-of-Significance of Regression Coefficients: The t Test

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

$$t = \frac{0.7240 - 0.5}{0.0700} = 3.2$$

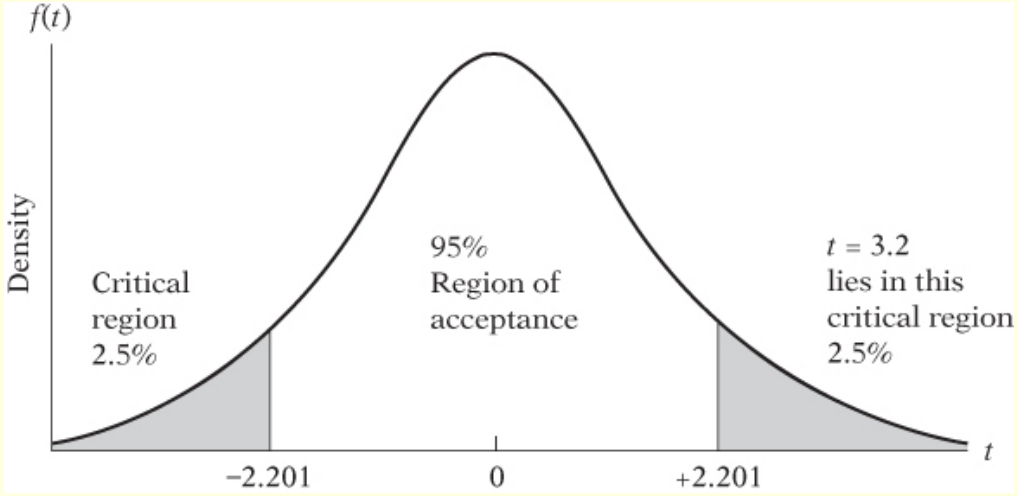


Reject H_0

The Test-of-Significance of Regression Coefficients: The t Test

$$t = \frac{\hat{\beta}_2 - \beta_2}{se(\hat{\beta}_2)}$$

$$= \frac{(\hat{\beta}_2 - \beta_2)\sqrt{\sum x_i^2}}{\hat{\sigma}} = \frac{(\hat{\beta}_2 - \beta_2)\sqrt{\sum (X_i - \bar{X})^2}}{\hat{\sigma}}$$



$$t = \frac{0.7240 - 0.5}{0.0700} = 3.2$$

Reject H_0

The Test-of-Significance of Regression Coefficients: The t Test

Significance tests

A statistic is said to be statistically significant if the value of the test statistic **lies in the critical region**

The null hypothesis is rejected

One tailed test

Y Sales	X Price
49	1
45	2
44	3
39	4
38	5
37	6
34	7
33	8
30	9
29	10

$$\hat{Y}_i = 49.667 - 2.1576 X_i$$

Right tail test

$\hat{\beta}_2 = -2.1576$

$se(\hat{\beta}_2) = 0.1204$

$df = 10 - 2 = 8, \alpha = 0.05, t_{0.05} = 1.860$

$$H_0 : \beta_2 \leq 0$$

$$H_1 : \beta_2 > 0$$

$$t = \frac{\hat{\beta}_2 - \beta_2}{se(\hat{\beta}_2)} = \frac{-2.1576 - 0}{0.1204} = -17.92$$

Not reject H_0 . There is not enough evidence to say that $\beta_2 > 0$.

Left tail test

$\hat{\beta}_2 = -2.1576$

$se(\hat{\beta}_2) = 0.1204$

$df = 10 - 2 = 8, \alpha = 0.05, t_{0.05} = -1.860$

$$H_0 : \beta_2 \geq 0$$

$$H_1 : \beta_2 < 0$$

$$t = \frac{\hat{\beta}_2 - \beta_2}{se(\hat{\beta}_2)} = \frac{-2.1576 - 0}{0.1204} = -17.92$$

Reject H_0. There is enough evidence to say that $\beta_2 < 0$

Decision Rule:

- Right tailed test

  If $t^* >$ critical (t_α) reject H_0

- Left tailed test

  If $t^* <$ - critical $(-t_\alpha)$ reject H_0

The Test-of-Significance of σ^2:The Test χ^2

The Test-of-Significance of σ^2:The Test χ^2

$$\chi^2 = (n-2)\frac{\hat{\sigma}^2}{\sigma^2}$$

TABLE 5.2
A Summary of the χ^2 Test

H_0: The Null Hypothesis	H_1: The Alternative Hypothesis	Critical Region: Reject H_0 If
$\sigma^2 = \sigma_0^2$	$\sigma^2 > \sigma_0^2$	$\frac{\text{df}(\hat{\sigma}^2)}{\sigma_0^2} > \chi^2_{\alpha,\text{df}}$
$\sigma^2 = \sigma_0^2$	$\sigma^2 < \sigma_0^2$	$\frac{\text{df}(\hat{\sigma}^2)}{\sigma_0^2} < \chi^2_{(1-\alpha),\text{df}}$
$\sigma^2 = \sigma_0^2$	$\sigma^2 \neq \sigma_0^2$	$\frac{\text{df}(\hat{\sigma}^2)}{\sigma_0^2} > \chi^2_{\alpha/2,\text{df}}$ or $< \chi^2_{(1-\alpha/2),\text{df}}$

Note: σ_0^2 is the value of σ^2 under the null hypothesis. The first subscript on χ^2 in the last column is the level of significance, and the second subscript is the degrees of freedom. These are critical chi-square values. Note that df is $(n - 2)$ for the two-variable regression model, $(n - 3)$ for the three-variable regression model, and so on.

The Test-of-Significance of σ^2 :The χ^2 Test

example

$\hat{\sigma}^2 = 0.8937, df = 11$

$H_0 : \sigma^2 = 0.6$

$H_1 : \sigma^2 \neq 0.6$

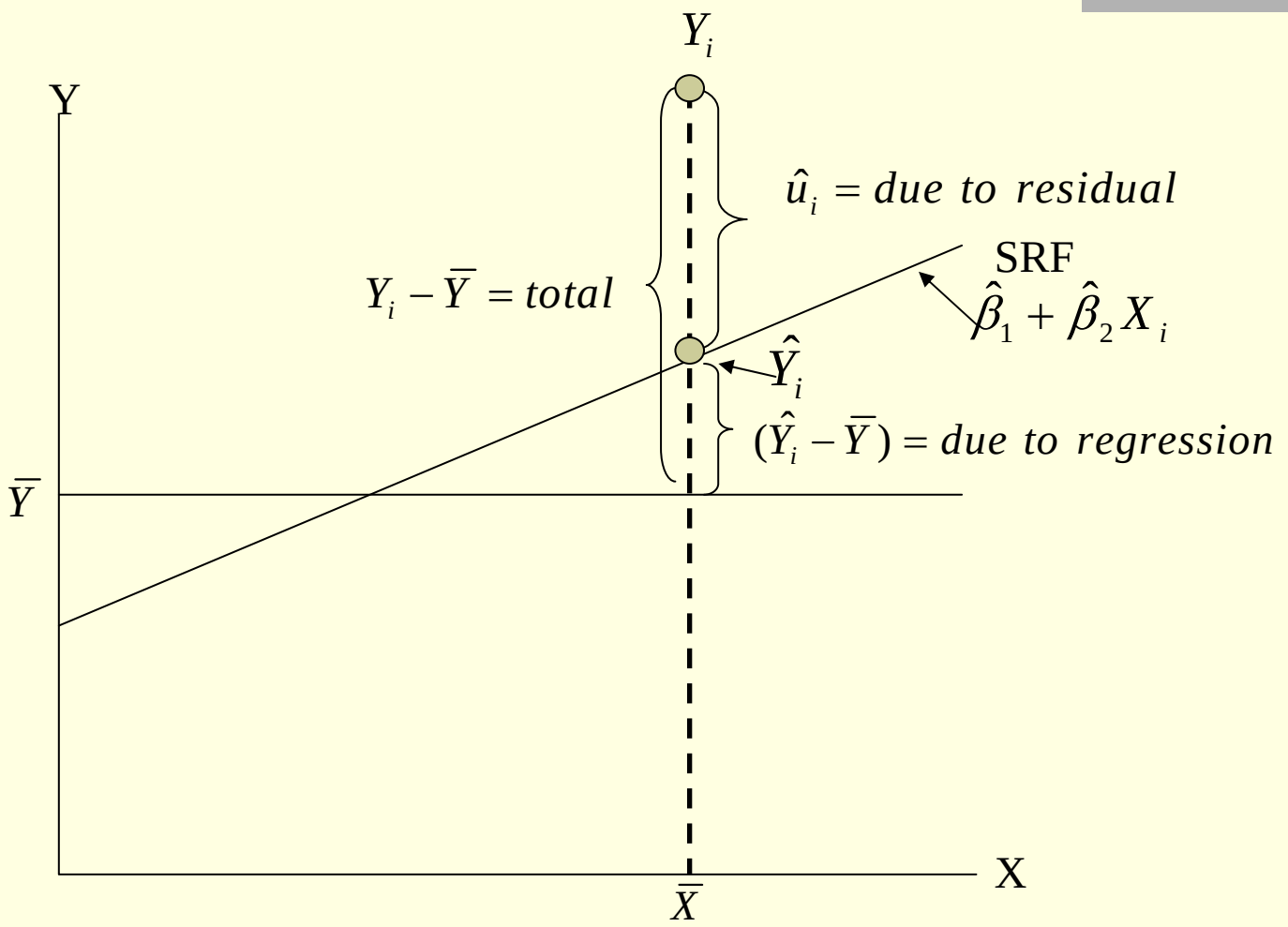
$\chi^2 = (n-2)\frac{\hat{\sigma}^2}{\sigma^2} = 16.3845$

$\alpha = 0.05, \textit{the critical } \chi^2 \textit{values are } \chi^2_{1-\alpha/2} = 3.81575 \textit{ and } \chi^2_{\alpha/2} = 21.9200$

We do not reject the null hypothesis.

There is not enough evidence to say that σ^2 is different from 0.6.

Analysis of Variance


Y_i
Y
$\hat{u}_i$ = due to residual
SRF
$\hat{\beta}_1 + \hat{\beta}_2 X_i$
$Y_i - \bar{Y}$ = total
$\hat{Y}_i$
$(\hat{Y}_i - \bar{Y})$ = due to regression
$\bar{Y}$
X
$\bar{X}$

$$TSS = ESS + RSS$$

$$\sum y_i^2 = \sum (Y_i - \bar{Y})^2$$ **(Total Sum of Squares, TSS)**

$$\sum \hat{y}_i^2 = \sum (\hat{Y}_i - \bar{Y})^2 = \hat{\beta}_2^2 \sum x_i^2$$

(Explained Sum of Squares, ESS)

$$\sum \hat{u}_i^2$$ **(Residual Sum of Squares, RSS)**

Analysis of Variance

$$\sum y_i^2 = \sum \hat{y}_i^2 + \sum \hat{u}_i^2 = \hat{\beta}_2^2 \sum x_i^2 + \sum \hat{u}_i^2$$

$$TSS = ESS + RSS$$

Analysis of Variance

TABLE 5.3
ANOVA Table for the Two-Variable Regression Model

Source of Variation	SS*	df	MSS†
Due to regression (ESS)	$\sum \hat{y}_i^2 = \hat{\beta}_2^2 \sum x_i^2$	1	$\hat{\beta}_2^2 \sum x_i^2$
Due to residuals (RSS)	$\sum \hat{u}_i^2$	$n - 2$	$\frac{\sum \hat{u}_i^2}{n-2} = \hat{\sigma}^2$
TSS	$\sum y_i^2$	$n - 1$	

*SS means sum of squares.
†Mean sum of squares, which is obtained by dividing SS by their df.

MSE (Mean of Explained Sum of Square)

MSR (Mean of Residual Sum of Square)

Analysis of Variance

$$F = \frac{MSS\ of\ ESS}{MSS\ of\ RSS}$$

$$= \frac{\hat{\beta}_2^2 \sum x_i^2}{\sum \hat{u}_i^2 /(n-2)}$$

$$= \frac{\hat{\beta}_2^2 \sum x_i^2}{\hat{\sigma}^2}$$

F distribution with 1 df in the numerator and

(n-2) df in the denominator

- If F* > Critical F value $(F_{\alpha;1,n-2})$ Reject H_0
- If F* < Critical F value $(F_{\alpha;1,n-2})$ Not reject H_0

$$\alpha = 0.05$$

$$F_{\alpha,1,5} =$$

$$F_{\alpha,1,8} =$$

$$F_{\alpha,2,5} =$$

$$F_{\alpha,3,5} =$$

$$\alpha = 0.05$$

$$F_{\alpha,1,5} = 6.61$$

$$F_{\alpha,1,8} = 5.32$$

$$F_{\alpha,2,5} = 5.79$$

$$F_{\alpha,3,5} = 5.41$$

$$\alpha = 0.01$$

$$F_{\alpha,1,5} =$$

$$F_{\alpha,1,8} =$$

$$F_{\alpha,2,5} =$$

$$F_{\alpha,3,5} =$$

$$\alpha = 0.01$$

$$F_{\alpha,1,5} = 16.3$$

$$F_{\alpha,1,8} = 11.3$$

$$F_{\alpha,2,5} = 13.3$$

$$F_{\alpha,3,5} = 12.1$$

Example: Wage-Education

TABLE 2.6
Mean Hourly Wage by Education

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

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

F-Test

$$H_0 : \beta_2 = 0$$

$$H_1 : \beta_2 \neq 0$$

TABLE 5.4
ANOVA Table for the Wages-Education Example

Source of Variation	SS	df	MSS	
Due to regression (ESS)	95.4255	1	95.4255	$F = \frac{95.4255}{0.8811}$
Due to residuals (RSS)	9.6928	11	0.8811	$= 108.3026$
TSS	105.1183	12		

- If F* > Critical F value $(F_{0.05;1,11})$ Reject H_0
- If F* < Critical F value $(F_{0.05;1,11})$ Not reject H_0

$$(F_{0.05;1,11}) = 4.84$$

Reject H_0

Application of Regression Analysis: The Problem of 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

TABLE 2.6
Mean Hourly Wage by Education

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

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

Mean Prediction

$$\hat{Y}_i = -0.0144 + 0.7240 X_i$$

Assume that $X_0 = 20$

$E(Y \mid X_0 = 20) = ?$

$$\hat{Y}_0 = -0.0144 + 0.7240 X_0$$
$$= -0.0144 + 0.7240(20)$$
$$= 14.4656$$

where $\hat{Y}_0 =$ *estimator of* $E(Y \mid X_0)$

Mean Prediction

$\hat{Y}_0$ *is normally distributed with mean* $(\beta_1 + \beta_2 X_0)$ *and*

$$\text{var}(\hat{Y}_0) = \sigma^2 \left[\frac{1}{n} + \frac{(X_0 - \bar{X})^2}{\sum x_i^2} \right]$$

By replace the unknown σ^2 by its unbiased estimators $\hat{\sigma}^2$

$$t = \frac{\hat{Y}_0 - (\beta_1 + \beta_2 X_0)}{se(\hat{Y}_0)}$$

follows the t distribution with n-2 df

Mean Prediction

The t distribution can therefore be used to derive confidence intervals for the true $E(Y_0 \mid X_0)$

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

where $se(\hat{Y}_0)$ is obtained from

$$\text{var}(\hat{Y}_0) = \sigma^2 \left[\frac{1}{n} + \frac{(X_0 - \bar{X})^2}{\sum x_i^2} \right]$$

$$\text{var}(\hat{Y}_0) = 0.8936 \left[\frac{1}{13} + \frac{(20-12)^2}{182} \right] = 0.3826$$

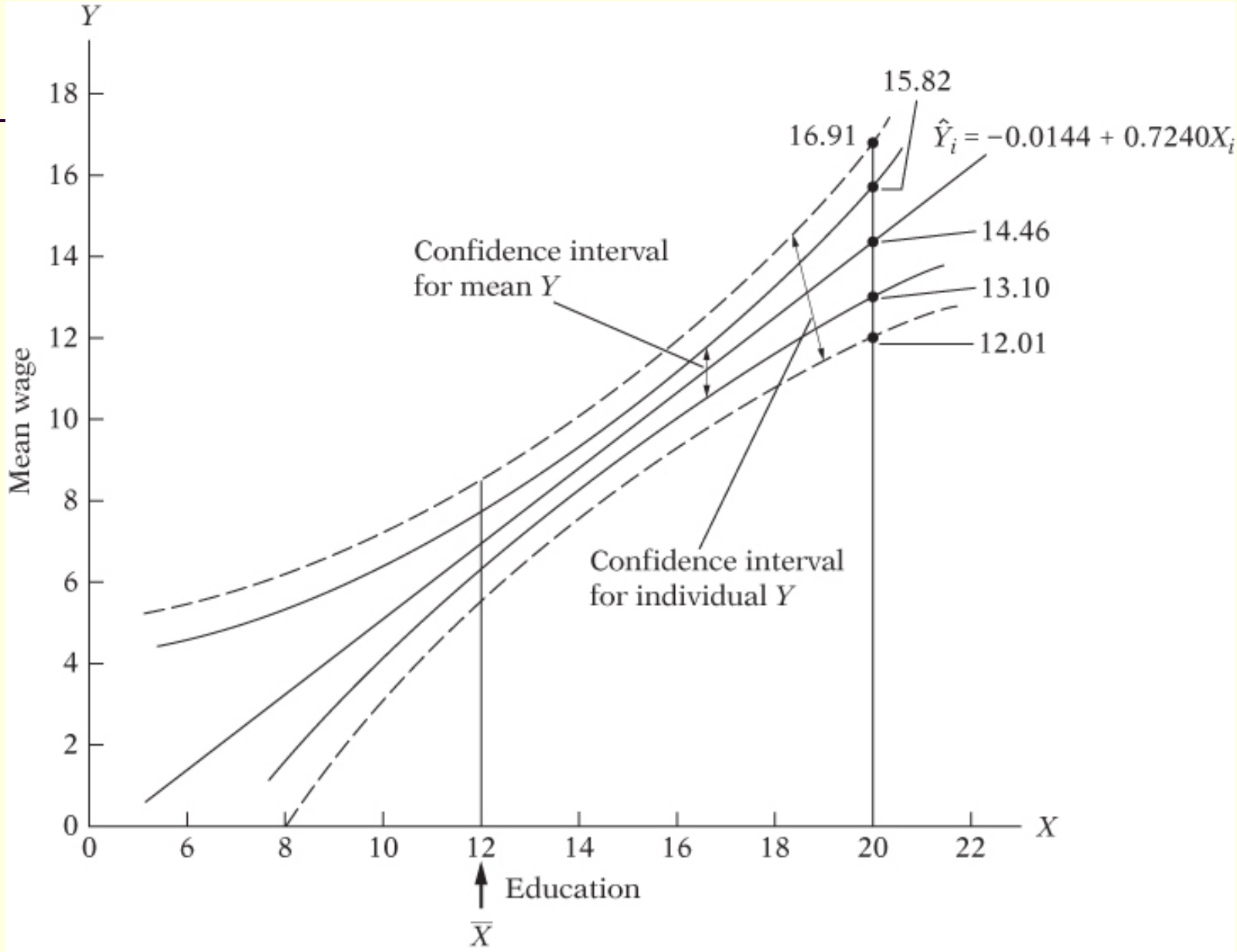
$$se(\hat{Y}_0) = 0.6185$$

Mean Prediction

Therefore, the 95 percent confidence interval for true $E(Y \mid X_0) = \beta_1 + \beta_2 X_0$ is given by

$$14.4656 - 2.201(0.6185) \leq E(Y_0 \mid X = 20) \leq 14.4656 + 2.20(0.6185)$$

$$13.1043 \leq E(Y_0 \mid X = 20) \leq 15.8260$$


Y
18
16
14
12
10
8
6
4
2
0
Mean wage
0
6
8
10
12
14
16
18
20
22
X
Education
$\overline{X}$
15.82
16.91
$\hat{Y}_i = -0.0144 + 0.7240X_i$
14.46
13.10
12.01
Confidence interval
for mean Y
Confidence interval
for individual Y

Individual Prediction

$$t = \frac{Y_0 - \hat{Y}_0}{se(Y_0 - \hat{Y}_0)}$$

$$\text{var}(Y_0 - \hat{Y}_0) = E[Y_0 - \hat{Y}_0]^2 = \sigma^2 \left[1 + \frac{1}{n} + \frac{(X_0 - \bar{X})^2}{\sum x_i^2} \right]$$

$$\hat{Y}_0 \pm t_{\alpha/2}\hat{\sigma}\{Y_0 - \hat{Y}_0\}$$

$$\hat{\sigma}\{Y_0 - \hat{Y}_0\} = \hat{\sigma}\sqrt{\left[1 + \frac{1}{n} + \frac{(X_0 - \bar{X})^2}{\sum (X_i - \bar{X})^2}\right]}$$

Therefore, the t distribution can be used to draw inferences about the true Y_0. The point prediction of Y_0 is 14.4656, the same as that of Y_0 , and its variance is 1.2357

$$(12.0190 \leq Y_0 \mid X_0 = 20 \leq 16.9122)$$