743- Regression and Time Series

Mamikon S. Ginovyan

Analysis of Variance and Correlation

Goodness of Fit

Regression residuals can provide a useful **measure of the fit** between the **estimated regression line** and the **data**.

- A **good regression equation** is one which **allows explain a large proportion of the variance** of $\boldsymbol{Y}$.

- **Large residuals imply a poor fit**, while **small residuals imply a good fit**.

Thus, residuals (more precisely, the **SSE**) can be used as a **measure of goodness of fit**.

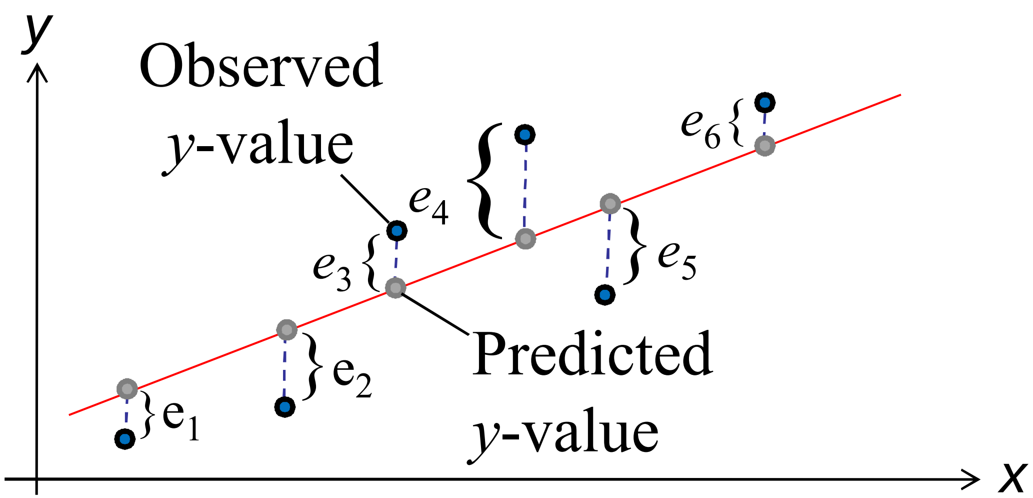
Goodness of Fit

Recall Residuals

- The **difference** between the **observed *y*-value** and the **predicted *y*-value** for a given ***x*-value** on the line.

For a given *x*-value,

e_i = (observed *y*-value) – (predicted *y*-value)



Goodness of Fit

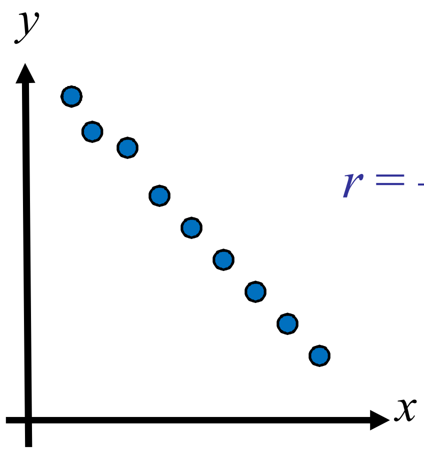
Note: **SSE** can be used as a **measure of goodness of fit** (explaining y-variation).

(a) Perfect Fit	**(b) Moderate Fit**	**(c) Poor Fit**
All variation explained.	Most variation explained.	Regression line does not fit the data, little variation explained.
SSE=0	***SSE*** is small	***SSE*** is large

Goodness of Fit

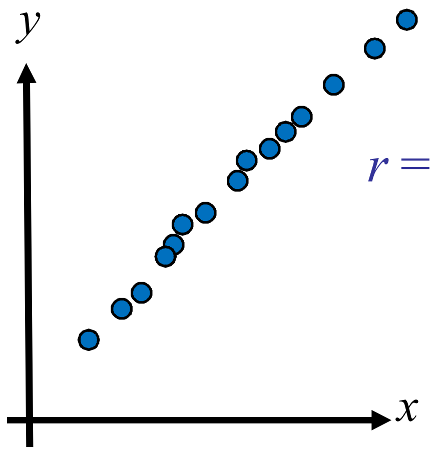
- **Example 1.**

Consider the following **scatter plots** (explaining y-variation).



$r = -0.9985$

Strong negative correlation



$r = 0.9876$

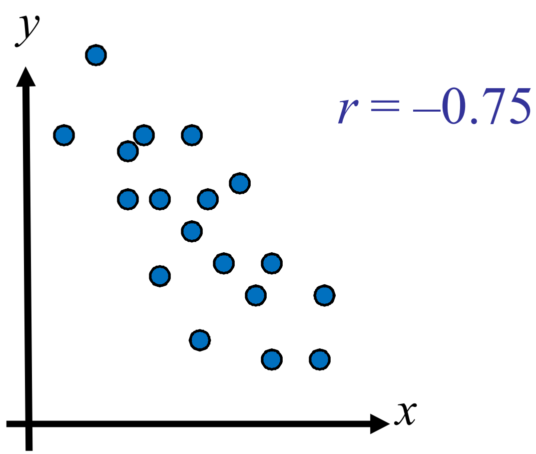
Strong positive correlation

(a) Perfect Fit

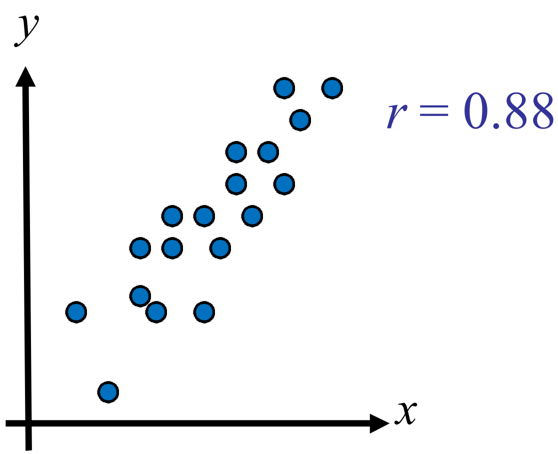
All variation explained.

$SSE=0$

Goodness of Fit



Strong negative correlation



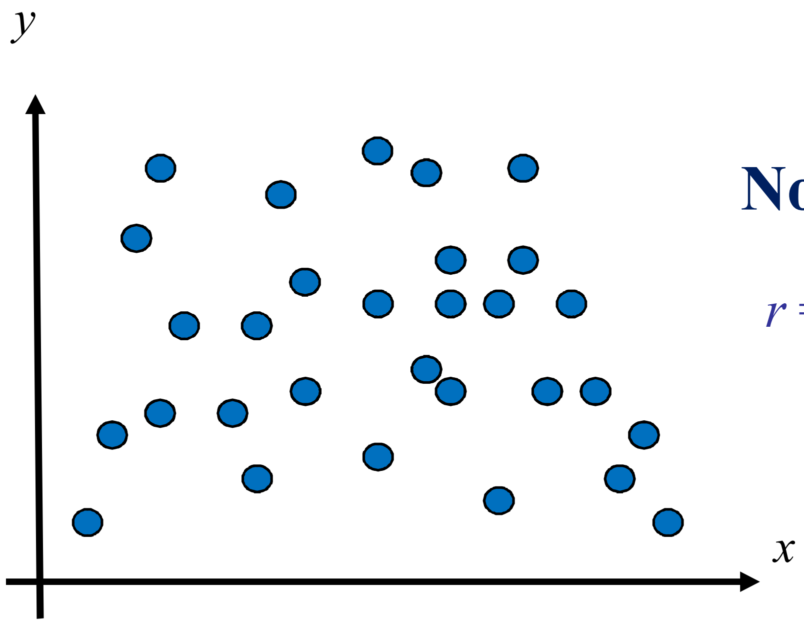
Strong positive correlation

(b) Moderate Fit

Most variation explained.

SSE **is small**

Goodness of Fit



Nonlinear Correlation

$r = 0.07$

(c) Poor Fit

Regression line does not fit the data, little variation explained.

***SSE* is large**

Goodness of Fit

A **quantitative** measure of the **total amount of variation** in observed y -values is given by the **total sum of squares**:

$$Variation\ (y) = SST = S = \sum_{i=1}^{n} (y_i - \bar{y})^2 \qquad (1)$$

the sum of squared deviation about the sample mean of the observed y -values.

Let $\hat{y} = \hat{a} + \hat{b}x$ be the **estimated regression line**, then for all observations we have

$$y_i - \bar{y} = (y_i - \hat{y}_i) + (\hat{y}_i - \bar{y}) \qquad (2)$$

Goodness of Fit

From (1) and (2) we obtain the following

ANOVA Identity for Regression

$$SST = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

$$= \sum_{i=1}^{n} \left[(y_i - \hat{y}_i) + (\hat{y}_i - \bar{y}) \right]^2$$

$$= \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 + \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$$

$$= SSE + SSR. \qquad (3)$$

(**the cross-product is equal to 0**), where

Goodness of Fit

$$SSE = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$$ is the **Error Sum of squares**

(equivalently, the **Residual Sum of Squares**),

$$SSR = \sum_{i=1}^{n} (\hat{y}_i - \overline{y})^2$$ is the **Regression Sum of Squares**,

$$SST = \sum_{i=1}^{n} (Y_i - \overline{y})^2$$ is the **Total Sum of Squares**.

Goodness of Fit

Observe that

- **SSE measures the unexplained variation of the data,** while
- **SSR measures variation explained by the linear relationship**, that is, **SSR is the amount of total variation that is explained by the model**.

Observe that (by (3))

$$SSE < SST$$

(**unless** the horizontal line $\hat{Y} = \overline{Y}, \hat{b} = 0$, is the **least squares line**).

Goodness of Fit

The **ratio**

$$\frac{SSE}{SST}$$

is the **proportion of total variation that cannot be explained by the simple regression model**, and

$$1 - \frac{SSE}{SST} = \frac{SSR}{SST}$$

is the **proportion of the observed y -variation explained by the model**.

This leads to the following definition.

Coefficient of Determination

Definition.

The number

$$R^2 = 1 - \frac{SSE}{SST} = \frac{SSR}{SST}$$

is called the **coefficient of determination**, and describes the proportion of observed $\boldsymbol{y}$ -variation that can be explained by the **simple linear regression model** (attributed to an approximate linear relationship between $\boldsymbol{y}$ and $\boldsymbol{x}$).

AVOVA

It is occasionally useful to summarize the breakdown of the ***y*** - variation in terms of an **ANOVA**.
Taking into account the **ANOVA Identity**:

$$SST = \sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$$

$$= \quad SSE \quad + \quad SSR,$$

SST (df =n-1)

SSE (df =n-2) **SSR (df =1)**

we can construct the following **ANOVA table**, which is also useful for hypothesis testing problem.

AVOVA

ANOVA Table for Simple Linear Regression

Source of Variation	*df*	Sum of Squares	Mean Squares	F
Regression	1	SSR	$MSR = SSR / 1$	$\dfrac{SSR}{SSE / (n-2)}$
Error	n -2	SSE	$MSE = s^2 = \dfrac{SSE}{n-2}$	
Total	n -1	SST		

ANOVA

Remark.
The hypothesis

$$H_0 : b = 0 \quad vs. \quad H_a : b \neq 0$$

can be tested using the **ANOVA Table** and the **Decision Rule**:

Reject H_0 if $F \geq F_{a,1,(n-2)}$.

Thus, the ***F*** -test gives exactly the same result as the **model utility *t* -test** because

$$T^2 = F \quad \text{and} \quad t^2_{a/2,(n-2)} = F_{a,1,(n-2)}.$$

Correlation vs. Coefficient of Determination

Recall that the correlation $r(X,Y)$ is defined by

$$r(X,Y) = \frac{Cov(X,Y)}{\sqrt{Var(X)} \cdot \sqrt{Var(Y)}}, \qquad \text{and}$$

$$r_{XY} = \hat{r}_{XY} = \frac{S_{XY}}{S_X S_Y} = \frac{\frac{1}{n-1}\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\frac{1}{n-1}\sum (X_i - \bar{X})^2} \cdot \sqrt{\frac{1}{n-1}\sum (Y_i - \bar{Y})^2}}, \qquad (1)$$

$$S_X^2 = S_{XX}$$

is the **point estimator** for $r(X,Y)$, and

$$r_{xy} = \hat{r}_{xy} = \frac{S_{xy}}{S_x S_y}$$ is the **point estimate** for $r(X,Y)$.

Correlation vs. Coefficient of Determination

Remark.

Taking into account that

$$\hat{b} = \frac{S_{XY}}{S_{xx}} \qquad (S_X^2 = S_{XX}). \qquad (2)$$

we obtain the following useful equality.

$$r_{XY} = \hat{r}_{XY} = \frac{S_{XY}}{S_X S_Y} = \frac{S_{XY}}{S_{xx}} \cdot \frac{S_x}{S_Y} = \hat{b} \cdot \frac{S_x}{S_Y}. \qquad (3)$$

Correlation vs. Coefficient of Determination

Recall also that,
for any two RV's ***X*** and ***Y***,

$$(a) \quad |r(X,Y)| \leq 1$$

$$(b) \quad |r(X,Y)| = 1 \Leftrightarrow Y = aX + b.$$

Interpreting $r_{XY} = \hat{r}_{XY}$.

The above result is **not sufficient to provide a useful interpretation** of $r_{XY} = \hat{r}_{XY}$.

Correlation vs. Coefficient of Determination

For example, what does it mean to say that the sample correlation coefficient is .73, .55 or .51?

One way to answer such a question focuses on the **square** of $r_{XY} = \hat{r}_{XY}$ rather than $r_{XY} = \hat{r}_{XY}$ on itself.

Now we show that

$$r_{XY}^2 = \hat{r}_{XY}^2 = R^2, \qquad (4)$$

where $R^2 = \dfrac{SSR}{SST}$ is the **coefficient of determination**.

Correlation vs. Coefficient of Determination

Indeed, first observe that

$$SSE=\sum_{i=1}^{n}(Y_i-\hat{Y}_i)^2=\sum_{i=1}^{n}(Y_i-\hat{a}-\hat{b}x_i)^2 \quad (\text{since} \quad \hat{a}=\overline{Y}-\hat{b}\,\overline{x})$$

$$=\sum_{i=1}^{n}(Y_i-\overline{Y}+\hat{b}\,\overline{x}-\hat{b}x_i)^2=\sum_{i=1}^{n}\left[(Y_i-\overline{Y})-\hat{b}(x_i-\overline{x})\right]^2$$

$$=\sum_{i=1}^{n}(Y_i-\overline{Y})^2+\hat{b}^2\sum_{i=1}^{n}(x_i-\overline{x})^2-2\hat{b}\sum_{i=1}^{n}(Y_i-\overline{Y})(x_i-\overline{x})$$

$$=S_{YY}+\hat{b}^2S_{XX}-2\hat{b}S_{XY} \quad (\text{since} \quad S_{XY}=\hat{b}S_{XX})$$

$$=S_{YY}+\hat{b}^2S_{XX}-2\hat{b}^2S_{XX}=S_{YY}-\hat{b}^2S_{XX}.$$

Correlation vs. Coefficient of Determination

Thus

$$SSE = S_{YY} - \hat{b}^2 S_{XX}. \qquad (5)$$

Now by definition of R^2 and (5) we have

$$R^2 = 1 - \frac{SSE}{SST} = 1 - \frac{S_{YY} - \hat{b}^2 S_{XX}}{S_{YY}}$$

$$= 1 - 1 + \hat{b}^2 \frac{S_{XX}}{S_{YY}} = \hat{b}^2 \frac{S_{XX}}{S_{YY}} \qquad (by(3))$$

$$= \hat{r}^2_{XY} = r^2_{XY}.$$

Correlation vs. Coefficient of Determination

Remark 1.

The equality (4) we proved under **linear relationship** between $\boldsymbol{X}$ and $\boldsymbol{Y}$. **In general**, it is **not true**.

More precisely, if $\boldsymbol{X}$ and $\boldsymbol{Y}$ are in **non-linear relationship**, then

$R^2 \geq r_{XY}^2$, and the difference $R^2 - r_{XY}^2$

is a **measure of non-linearity.**

Correlation vs. Coefficient of Determination

Remark 2.
The definition of the coefficient of determination $R = R_{YX}$ is based on the **Conditional Variance Formula**:

$$Var(Y) = Var(E[Y|X]) + E[(Var(Y|X)],$$

and is defined to be

$$R_{YX}^2 = 1 - \frac{E[Var(Y|X)]}{Var(Y)}.$$

Correlation vs. Coefficient of Determination

Properties of the **Coefficient of Determination** R_{YX}^2

1. $0 \le R_{YX}^2 \le 1$, and $R_{YX}^2 = 1$ **if and only if**

$$E[Y - E(Y|x = x)]^2 = 0 \Leftrightarrow P(Y = E[Y|X = x]) = 1,$$

that is, there is a **functional relationship** between $\boldsymbol{X}$ and $\boldsymbol{Y}$.

2. In general, R_{YX}^2 is **not symmetric**, that is, $R_{YX}^2 \ne R_{XY}^2$.

3. $R_{YX}^2 = r_{YX}^2$ if $Y = aX + b$, otherwise $R_{YX}^2 > r_{YX}^2$.

4. If $E[Y|X = x]$ is **independent** of $\boldsymbol{x}$, then $R_{YX} = 0$.
This is, the case when the RV's $\boldsymbol{X}$ and $\boldsymbol{Y}$ are **independent**.

Hypothesis Test for Correlation

- To test the null hypothesis of **no linear association**:

$$\boxed{H_0:\ \rho = 0}$$

we use the **test statistic:**

$$\boxed{T = \frac{r}{\sqrt{\dfrac{1-r^2}{n-2}}} = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} \sim t_{(n-2)}}$$

which follows the **Student's t-distribution** with $\boldsymbol{(n-2)}$ degrees of freedom:

Remark: Hypothesis Test for Correlation

- The test of the hypothesis H_0: $\rho = 0$ of **no linear association** may also be based on the **Fisher statistic:**

$$\boxed{z = \frac{Z - m_Z}{s_Z} = (Z - m_Z)\sqrt{n-3} \sim N(0,1),}$$

where

$Z = \frac{1}{2} \cdot \ln \frac{1+r}{1-r}$ is the **Fisher Z- transformation,**

$m_Z = \frac{1}{2} \cdot \ln \frac{1+r}{1-r}$; $s_Z = \frac{1}{\sqrt{n-3}}$ and $z(\text{obs}) = Z(\text{obs})\sqrt{n-3}$.

Observe that under H_0: $\rho = 0$, we have $m_Z = (.5) \cdot \ln 1 = 0.$

Example: Hypothesis Test for Correlation

Age and Price of Orions. The data on age and price for a sample of **11** of Orions are given in the following table.

Age (yr) x	5	4	6	5	5	5	6	6	2	7	7
Price ($100) y	85	103	70	82	89	98	66	95	169	70	48

At the **5%** significance level, do the data provide sufficient evidence to conclude that age and price of Orions are **negatively linearly correlated?**

Example: Hypothesis Test for Correlation

Solution:

Step 1. State the null and alternative hypotheses.

Let ρ denote the population **linear correlation coefficient** for the variables age and price of Orions. **Then**

H_0: $\rho = 0$ (**age and price are linearly uncorrelated**)

H_1: $\rho < 0$ (**age and price are linearly correlated**)

Step 2. Compute the **observed value** of **test statistic:**

$$t_0 = \frac{r}{\sqrt{\dfrac{1-r^2}{n-2}}} = \frac{-0.924}{\sqrt{\dfrac{1-(-0.924)^2}{11-2}}} = -7.249.$$

Example: Hypothesis Test for Correlation

Calculations: We have

$$\sum_{i=1}^{11} x_i = 58; \quad \sum_{i=1}^{11} y_i = 975; \quad \sum_{i=1}^{11} x_i y_i = 4732;$$

$$\sum_{i=1}^{11} x_i^2 = 326; \quad \sum_{i=1}^{11} y_i^2 = 96129;$$

$$r = r_{XY} = \hat{r}_{XY} = \frac{S_{XY}}{S_X S_Y} = \frac{SS_{xy}}{\sqrt{SS_{xx} SS_{yy}}} = -0.924.$$

<u>Example:</u> Hypothesis Test for Correlation

<u>**Step 4.**</u> Decide on the significance level: a **= .05.**

The **critical value** for ***df*** **=** ***n*****-2 = 11-2 =9** and a **= .05** is

$$-t_{a,-n-2} = -t_{.05,9} = \text{-}1.833,$$

The **critical region** is: $CR = \{T < \text{-}1.833\}$.

<u>**Step 5.**</u> Since $t_0 = -7.249 < \text{-}1.833 = \mathrm{t}_a$, we **reject** $\boldsymbol{H_0: \rho = 0}$**,** and conclude that the test results are **statistically significant** at the **5%** level of significance.

Inferences about the mean

Inferences about the mean $m_{Y|X} = E[Y|X = x]$.

In addition to statistical inference about the **regression parameters** (***unknown***) a, b and s^2, we will sometimes find it also helpful to draw inferences about the **regression line**

$$m_{Y|x_0} = E[Y|X = x_0] = a + b\,x_0, \qquad (1)$$

where x_0 is a specified value of **independent** variable $\boldsymbol{x}$.

Inferences about the mean

Once the point estimators $\hat{a}$ and $\hat{b}$ have been calculated,

$$\hat{Y} = \hat{a} + \hat{b}\, x_0 \qquad (2)$$

can be regarded either as:

- a **point estimate** of $m_{Y|x_0}$, or
- a **prediction** of the $\boldsymbol{Y}$ value that will result from a single observation made when $x = x_0$.

Inferences about the mean

Observe that the point estimate or prediction by itself gives **no information** concerning how precisely $m_{Y|x_0}$ has been **estimated** or $\boldsymbol{Y}$ has been **predicted.**

This can be remedied by **developing CI's** for $m_{Y|x_0}$ and a **prediction interval (PI)** for a single $\boldsymbol{Y}$ value.

Thus, as a point estimator for **regression line** $m_{Y|x_0}$ given by (1) we consider the statistic $\hat{Y}$ given by (2).

Inferences about the mean

Properties of point estimator $\hat{Y} = \hat{a} + \hat{b}\, x_0$.

1. $\hat{Y}$ is an **unbiased estimator** for $m_{Y|x_0}$, that is,

$$E[\hat{Y}] = a + b\, x_0. \qquad (3)$$

Indeed, since $\hat{a}$ and $\hat{b}$ are unbiased estimators for a and b, respectively, we have

$$E[\hat{Y}] = E[\hat{a} + \hat{b}\, x_0]$$

$$= E[\hat{a}] + x_0 E[\hat{b}] = a + b\, x_0.$$

Properties of Point Estimator

2. The **variance** $Var(\hat{Y})$ of $\hat{Y}$ is given by

$$s_{\hat{Y}}^2 = Var(\hat{Y}) = s^2\left[\frac{1}{n} + \frac{(x_0 - \overline{x})^2}{S_{xx}}\right]. \qquad (4)$$

2'. The **estimated variance** $s_{\hat{Y}}^2$ is given by

$$s_{\hat{Y}}^2 = Va\hat{r}(\hat{Y}) = s^2\left[\frac{1}{n} + \frac{(x_0 - \overline{x})^2}{S_{xx}}\right],$$ where

$$S_{XX} = \sum_{i=1}^{n}(x_i - \overline{x})^2. \qquad (5)$$

Properties of Point Estimator

Proof of (4).

We have

$$Var(\hat{Y}) = Var(\hat{a} + \hat{b}\, x_0) \qquad (\text{since} \quad \hat{a} = \overline{Y} - \hat{b}\, \overline{x})$$

$$= Var(\overline{Y} - \hat{b}\, \overline{x} + \hat{b}\, x_0)$$

$$= Var(\overline{Y} + \hat{b}\, (x_0 - \overline{x}))$$

(since $\overline{Y}$ and $\hat{b}$ are **independent**)

$$= Var(\overline{Y}) + Var[\hat{b}\, (x_0 - \overline{x})]$$

$$= Var(\overline{Y}) + (x_0 - \overline{x})^2 Var(\hat{b})$$

$$= \frac{s^2}{n} + (x_0 - \overline{x})^2 \cdot \frac{s^2}{s_{xx}} = s^2 \left[\frac{1}{n} + \frac{(x_0 - \overline{x})^2}{s_{xx}} \right].$$

Properties of Point Estimator

3. $\hat{Y}$ has a **normal distribution**: $\hat{Y} \sim N(a + b x_0, s_{\hat{Y}}^2)$.

Proof.

Taking into account that

$$S_{xY} = \sum_{i=1}^{n} (x_i - \overline{x})(Y_i - \overline{Y}) = \sum_{i=1}^{n} (x_i - \overline{x}) Y_i, \qquad (6)$$

$$\text{(since} \quad \sum_{i=1}^{n} (x_i - \overline{x})\overline{Y} = \overline{Y} \sum_{i=1}^{n} (x_i - \overline{x})$$

$$= \overline{Y}(n\overline{x} - n\overline{x}) = 0.)$$

we obtain,

Properties of Point Estimator

$$\hat{Y} = \hat{a} + \hat{b}x_0 = \overline{Y} + \hat{b}(x_0 - \overline{x}) \quad (\text{since} \quad \hat{a} = \overline{Y} - \hat{b}\overline{x}))$$

$$= \overline{Y} + \frac{S_{xY}}{S_{xx}}(x_0 - \overline{x}) \qquad\qquad (\text{since} \quad \hat{b} = \frac{S_{xY}}{S_{xx}})$$

$$= \frac{1}{n}\sum_{i=1}^{n} Y_i + \sum_{i=1}^{n}(x_i - \overline{x})Y_i \frac{(x_0 - \overline{x})}{S_{xx}} \qquad (by \quad (6))$$

$$= \sum_{i=1}^{n}\left[\frac{1}{n} + \frac{(x_0 - \overline{x})(x_i - \overline{x})}{S_{xx}}\right]Y_i = \sum_{i=1}^{n} a_i Y_i. \qquad (7)$$

Properties of Point Estimator

Thus, $\hat{Y} = \sum_{i=1}^{n} a_i Y_i,$ where the coefficients $a_i, i = \overline{1, n}$

involve the x_i's and $x_0,$ all of which are **fixed.**

Because Y_i are **normally distributed** and **independent**, $\hat{Y}$ as a **linear combination** of Y_i's will also be **normally distributed**.

A Confidence Interval for regression line

A Confidence Interval Regression line:

$$m_{Y|x_0} = E[Y \mid X = x_0].$$

Using **Properties 1-3** we can state

Theorem 1. The statistic

$$T = \frac{\hat{Y} - (a + b\,x_0)}{s_{\hat{Y}}} \sim t_{(n-2)} \qquad (8)$$

has a **Student *t* -distribution** with (***n*** -2) df.

A Confidence Interval for regression line

Proof.

The statistic $\boldsymbol{T}$ can be represented as follows

$$T = \frac{\hat{Y} - (a + b\,x_0)}{s_{\hat{Y}}} \cdot \left[\sqrt{\frac{(n-2)s^2}{s^2} \Big/ (n-2)} \right]^{-1} .$$

Since $\dfrac{\hat{Y} - (a + b\,x_0)}{s_{\hat{Y}}} \sim N(0,1)$ and

$$\frac{(n-2)s^2}{s^2} \sim c^2(n-2),$$

the result follows from the **definition** of $\boldsymbol{t}$ **-distribution.**

A Confidence Interval for regression line

Now using **standard *t* -procedure** with (***n*** -2) df, we can write, for given a $(0 \le a \le 1)$

$$P(-t_{a/2,(n-2)} \le T \le t_{a/2,(n-2)}) = 1 - a\,. \qquad (9)$$

Substituting, ***T*** from (8) into (9) we obtain

$$P(-t_{a/2,(n-2)} \le \frac{\hat{Y} - m_{Y|x_0}}{s_{\hat{Y}}} \le t_{a/2,(n-2)}) = 1 - a\,.$$

Solving the inside inequalities for $m_{Y|x_0}$ we get the **desired CI** for $m_{Y|x_0}$.

Summing up we can state the following result.

A Confidence Interval for regression line

Theorem 2.
For given a $(0 \le a \le 1)$ a $100(1-a)\%$ **CI** for **regression line**

$$m_{Y|x_0} = E[Y \mid X = x_0]$$

is the interval

$$\hat{Y} \pm t_{a/2,(n-2)} \cdot s_{\hat{Y}}, \qquad (10)$$

where $s_{\hat{Y}}$ is given by

$$s_{\hat{Y}}^2 = V\hat{a}r(\hat{Y}) = s^2\left[\frac{1}{n} + \frac{(x_0 - \overline{x})^2}{S_{xx}}\right].$$

A Prediction Interval for a Future Value of Y

Similar to **CI** (10) for $m_{Y|x_0}$, we frequently wishes to obtain an **interval of plausible values for the value** of $\boldsymbol{Y}$ associated with some **future observation** when the independent variable $\boldsymbol{x}$ has value $\boldsymbol{x_0}$, that is,

- the value of $\boldsymbol{Y}$ of a **single future observation** to be recorded at some given level of $\boldsymbol{x = x_0}$.

Prediction Interval for a Future Value of Y

Remark 1.

A **CI** refers to a **parameter** (= population characteristic), whose value is **fixed but unknown** to us.

In contrast, a **future value** of Y is not a parameter but instead is a **RV**.

For this reason we refer to an interval of plausible values for a future Y as a **Prediction Interval (PI)** rather than **CI**.

Prediction Interval for a Future Value of *Y*

- For the **CI** we use the **error of estimation**:

$$(a + b x_0) - (\hat{a} + \hat{b} x_0) =$$

= a **difference** between
a **fixed (but unknown) quantity** and
a **RV**.

- The **error of prediction** is

$$Error = Y - \hat{Y} = Y - (\hat{a} + \hat{b} x_0)$$

$$= (a + b x_0 + e) - (\hat{a} + \hat{b} x_0)$$

= a **difference** between **two RV's**.

Prediction Interval for a Future Value of Y

With the additional random term e, there is **more uncertainty in prediction** than in **estimation**, so a **PI** will be **wider** than a **CI.**

Let $(x_1, Y_1), (x_2, Y_2), \ldots, (x_n, Y_n)$ be a set of n observations that satisfy the **Model Assumptions**, and let (x_0, Y) be a hypothetical **future observation**, where Y is **independent** of the Y_i 's, $i = 1, \ldots, n$.

Definition: A **prediction interval** is a **range** of numbers that contains Y with a specified **probability** $= (1 - \alpha)$.

Prediction Interval for a Future Value of Y

Consider the difference $Y - \hat{Y}$ = **error of prediction.**

- For the **expectation** of this error we have

$$E[Y - \hat{Y}] = E[Y] - E[\hat{Y}] = (a + b x_0) - (a + b x_0) = 0 \qquad (1)$$

- For **variance** of the $Y - \hat{Y}$ we have

$$Var(Y - \hat{Y}) = Var(Y) + Var(\hat{Y})$$

(since $\hat{Y}$ and Y are **independent**)

$$= s^2 \left[\frac{1}{n} + \frac{(x_0 - \overline{x})^2}{S_{xx}} \right] + s^2$$

$$= s^2 \left[1 + \frac{1}{n} + \frac{(x_0 - \overline{x})^2}{S_{xx}} \right]. \qquad (12)$$

Prediction Interval for a Future Value of *Y*

As in the **CI case**, we can show that the RV

$$Z = \frac{Y - \hat{Y}}{\sqrt{Var(\hat{Y} - Y)}} \sim N(0,1).$$

Therefore the RV

$$T = \frac{Y - \hat{Y}}{s\sqrt{1 + \frac{1}{n} + \frac{(x_0 - \overline{x})^2}{S_{xx}}}} \sim t_{(n-2)} \qquad (13)$$

has a **Student *t* -distribution** with (***n*** -2) df.

Prediction Interval for a Future Value of $\boldsymbol{Y}$

Therefore, for given a $(0 \le a \le 1)$,

$$P(-t_{a/2,(n-2)} \le T \le t_{a/2,(n-2)}) = 1 - a\,. \qquad (14)$$

Substituting $\boldsymbol{T}$ from (13) into (14), and solving the inside inequalities for $\boldsymbol{Y}$ we get the following result.

Prediction Interval for a Future Value of Y

Theorem 3 (PI for Y).

For given α $(0 \le \alpha \le 1)$ a $100(1-\alpha)\%$ **PI** for a future observation Y when $x = x_0$ is the interval

$$PI = (\hat{a} + \hat{b}\,x_0) \pm t_{\alpha/2,(n-2)} \cdot s\sqrt{1 + \frac{1}{n} + \frac{(x_0 - \overline{x})^2}{S_{xx}}}$$

$$= (\hat{a} + \hat{b}\,x_0) \pm t_{\alpha/2,(n-2)} \cdot \sqrt{s^2 + s_{\hat{Y}}^2}$$

$$= \hat{Y} \pm t_{\alpha/2,(n-2)} \cdot \sqrt{s^2 + s_{\hat{Y}}^2}.$$

Prediction Interval for a Future Value of Y

Remark.

Notice that the **length** of a **CI** for $E[Y]$ when $x = x_0$ is given by

$$2t_{\alpha/2,(n-2)} \cdot s\sqrt{\frac{1}{n} + \frac{(x_0 - \overline{x})^2}{S_{xx}}}, \qquad (15)$$

whereas the **length** of a **PI** for an actual value of Y when $x = x_0$ is given by

$$2t_{\alpha/2,(n-2)} \cdot s\sqrt{1 + \frac{1}{n} + \frac{(x_0 - \overline{x})^2}{S_{xx}}}. \qquad (16)$$

Thus, we observe that PI for the actual value of Y is longer than CI's for expectation $E[Y]$ if both are determined for the same value of $x = x_0$.

Testing the Equality of Two Slopes

We often are interested in **comparison of two linear** ***XY*** **- relationships**. In such situations we test the hypothesis

$$H_0 : b_1 = b_2,$$

where b_1 and b_2 are the **true slopes** associated with the two regressions.

If the data points taken on the two regressions are **all independent**, a **two-sample *t* -test** can be set up based on the properties proved in **Theorem 2 and 3** (for $\hat{a}$ and $\hat{b}$).

The following result is true.

Testing the Equality of Two Slopes

Theorem 1.

Let $(x_1, Y_1), (x_2, Y_2), \ldots, (x_n, Y_n)$ and $(x_1{*}, Y_1{*}), (x_2{*}, Y_2{*}), \ldots, (x_n{*}, Y_n{*})$ be two **independent** observations, each satisfying the **Model Assumptions:**

$$E[Y \mid x_0] = a_1 + b_1 x_0, \quad E[Y{*} \mid x{*}] = a_2 + b_2 x{*}.$$

(a) Let $$T = \frac{(\hat{b}_1 - \hat{b}_2) - (b_1 - b_2)}{s\sqrt{\dfrac{1}{\sum_{i=1}^{n}(x_i - \overline{x})^2} + \dfrac{1}{\sum_{i=1}^{m}(x_i{*} - \overline{x}{*})^2}}},$$

where

Theorem 1.

$$s = \frac{1}{\sqrt{n+m-4}} \sqrt{\sum_{i=1}^{n} [Y_i - (\hat{a}_1 + \hat{b}_1 x_i)]^2 + \sum_{i=1}^{m} [Y_i * - (\hat{a}_2 + \hat{b}_2 x_i *)]^2} .$$

Then $T \sim t_{(n-m-4)}$ = ***t* -distribution** with (***n*** + ***m*** - 4) df.

Theorem 1.

(b) To test the hypothesis

$$H_0 : b_1 = b_2 \quad vs. \quad H_1 : b_1 \neq b_2$$

at the a level of significance,

Reject H_0 if either $T_{obs} \leq -t_{a/2,(n+m-4)}$ or $T_{obs} \geq t_{a/2,(n+m-4)}$, where

$$T_{obs} = \frac{(\hat{b}_1 - \hat{b}_2)}{s\sqrt{\dfrac{1}{\sum_{i=1}^{n}(x_i - \overline{x})^2} + \dfrac{1}{\sum_{i=1}^{m}(x_i^* - \overline{x}^*)^2}}}.$$

Theorem 1.

Remark.

One-sided tests are defined in the usual way by replacing the **critical values**

$$\pm t_{a/2,(n+m-4)}$$

by $t_{a/2,(n+m-4)}$ for **upper-tailed** test $(H_1 : b_1 > b_2)$, or

by $-t_{a,(n+m-4)}$ for **upper-tailed** test $(H_1 : b_1 < b_2)$.

Summary Example

Suppose we have a data set consisting of $\boldsymbol{n}=5$ points.
The data set is given by the following table.

$\boldsymbol{x}$	-2	-1	0	1	2
$\boldsymbol{y}$	0	0	1	1	3

(i) Use the method of least squares to fit a straight line to given points.

(ii) Find the variance of the estimates $\hat{a}$ and $\hat{b}$ obtained in part (i).

(iii) Estimate s^2 from the data.

(iv) Calculate a 95% CI for the slope parameter b.

Summary Example

(v) Do the data present sufficient evidence to indicate that the slope differs from 0? **(Model Utility Test)**.
Test the hypothesis at $\boldsymbol{\alpha} = \mathbf{.05}$ level of significance using

(a) Critical-value method

(b) $\boldsymbol{P}$ -value method

(c) CI-method.

(vi) Find a 90% CI for the regression line $m_{y|x_0} = E\left[Y \middle| X = x_0\right]$, when $\boldsymbol{x_0} = \mathbf{1}$**.**

(vii) Suppose that the experiment that generated the data in Part (i) is to be run again with $\boldsymbol{x} = \boldsymbol{x_0} = \mathbf{2}$**.**
Predict the particular value of $\boldsymbol{Y}$ with $\boldsymbol{\alpha} = \mathbf{.1}$**,** that is, construct 90% PI for $m_{Y|x=2} = E[Y|x = 2]$.

Summary Example

Solution.

(i) Use the method of least squares to fit a straight line to given points.

First we construct the following **Calculation Table.**

x_i	y_i	$x_i y_i$	x_i^2	y_i^2
-2	0	0	4	0
-1	0	0	1	0
0	1	0	0	1
1	1	1	1	1
2	3	6	4	9
$\sum_{i=1}^{n} x_i = 0$	$\sum_{i=1}^{n} y_i = 5$	$\sum_{i=1}^{n} x_i y_i = 7$	$\sum_{i=1}^{n} x_i^2 = 10$	$\sum_{i=1}^{n} y_i^2 = 11$

Summary Example-Solution

Compute

$$S_{xy} = \sum_{i=1}^{n} x_i y_i - \frac{1}{n}\sum_{i=1}^{n} x_i y_i = 7 - \frac{1}{5}(0)(5) = 7,$$

$$S_{xx} = \sum_{i=1}^{n} x_i^{\,2} - \frac{1}{n}(\sum_{i=1}^{n} x_i)^2 = 10 - \frac{1}{5}(0)^2 = 10.$$

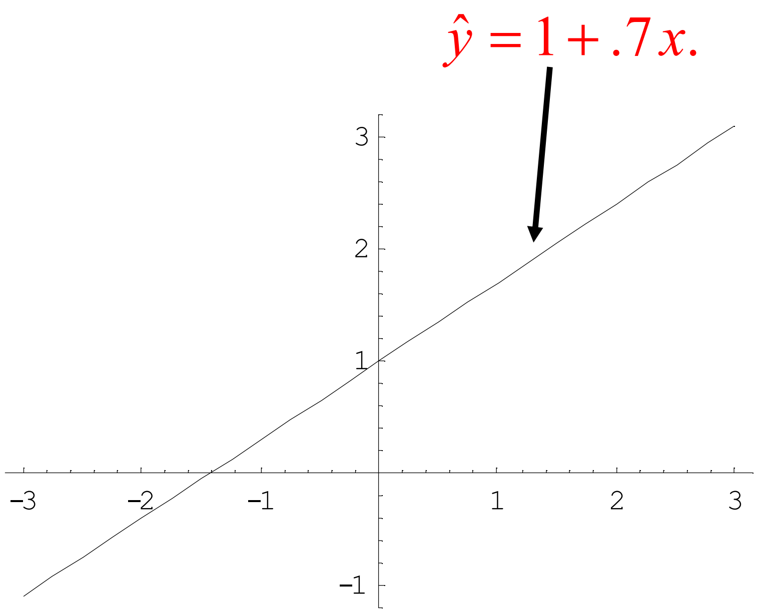
Thus,

$$\hat{b} = \frac{S_{xy}}{S_{xx}} = \frac{7}{10} = .7,$$

$$\hat{a} = \overline{y} - \hat{b}\,\overline{x} = \frac{5}{5} - (.7)(0) = 1.$$

Therefore the **fitted line** is:

$$\hat{y} = \hat{a} + \hat{b}\,x = 1 + .7x.$$



Summary Example-Solution

(ii) Find the variance of the estimates $\hat{a}$ and $\hat{b}$ obtained in Part (i).

Solution. We have

$$Var(\hat{a}) = \frac{s^2 \sum_{i=1}^{n} x_i^2}{nS_{xx}} = \frac{s^2(10)}{5(10)} = \frac{1}{5}s^2$$

and

$$Var(\hat{b}) = \frac{s^2}{S_{xx}} = \frac{s^2}{10}.$$

Remark. Notice that in this example $\sum x_i = 0$. Hence $Cov(\hat{a}, \hat{b}) = 0$, since

$$Co\hat{v}(\hat{a}, \hat{b}) = -\frac{\overline{x}s^2}{S_{xx}}, \quad \text{where } s^2 = s_{LS}^2, \text{ and } \overline{x} = \frac{1}{n}\sum x_i = 0.$$

Summary Example-Solution

(iii) Estimate s^2 from the data.

Solution. We have

$$\sum_{i=1}^{5} y_i = 5, \quad \sum_{i=1}^{5} y_i^2 = 11, \quad S_{xy} = 7, \quad \hat{b} = .7, \quad \overline{y} = 1.$$

Hence

$$S_{yy} = \sum_{i=1}^{5} (y_i - \overline{y})^2 = \sum_{i=1}^{5} y_i^2 - 5(\overline{y}^2) = 11 - 5(1)^2 = 6,$$

and

$$SSE = S_{yy} - \hat{b}\, S_{xy} = 6 - (.7)(7) = 1.1.$$

Therefore

$$s^2 = \frac{SSE}{n-2} = \frac{1.1}{5-2} = \frac{1.1}{3} = .367.$$

Summary Example-Solution

(iv) Calculate a 95% CI for the slope parameter b.

Solution.

We have that for $0 \le a \le 1$, a $100(1-a)\%$ CI for b is the interval

$$\hat{b} \pm t_{a/2,(n-2)} s_{\hat{b}} = \hat{b} \pm t_{a/2,(n-2)} \frac{s}{\sqrt{S_{xx}}} \qquad (1)$$

Now we compute the quantities in (1).

Since

$$n = 5 \Rightarrow n-2 = 3, a = .05 \Rightarrow \frac{a}{2} = .025.$$

So from $\boldsymbol{t}$ -table we find

$$t_{a/2,(n-2)} = t_{.025,3} = 3.182. \qquad (2)$$

Summary Example-Solution

From Part (iii) we have

$$s = \sqrt{\frac{SSE}{n-2}} = \sqrt{.367} = .606. \qquad (3)$$

From Part (i) we have

$$\hat{b} = .7 \quad \text{and} \quad S_{xx} = 10. \qquad (4)$$

Substituting (2) - (4) into (1) we get the desired 95% CI for ***β***

$$\hat{b} \pm t_{a/2,(n-2)} \frac{s}{\sqrt{S_{xx}}}$$

$$= .7 \pm (3.182)(.606)\sqrt{.1} = .7 \pm .61 = (.09, 1.31).$$

Summary Example-Solution

Remark. If we wish to estimate β correct to within .15 unit, it is obvious that the CI is too wide and that the sample size (n =5) must be increased.

(v) Do the data present sufficient evidence to indicate that the slope differs from 0? (Model Utility Test).

Test the hypothesis at $\alpha = .05$ level of significance using

(a) Critical-value method

(b) P -value method

(c) CI-method.

Summary Example-Solution

Solution.

(a) **Step 1.** Set up the hypothesis:

$$H_0 : b = 0 \quad vs. \quad H_1 : b \neq 0. \quad (b_0 = 0)$$

Step 2. Specify test statistic and its distribution under H_0

$$TS = T = \frac{\hat{b} - b}{s_{\hat{b}}}.$$

under

$$H_0 : b = 0, T = \frac{\hat{b} - 0}{s_{\hat{b}}} = \frac{\hat{b}}{s_{\hat{b}}} \sim t_{(n-2)}.$$

Summary Example-Solution

Step 3. Compute the observed value of TS:

$$T_0 = T_{ob} = \frac{\hat{b} - 0}{s_{\hat{b}}} = \frac{.7 - 0}{.192} = 3.65,$$

Since

$$s_{\hat{b}} = \frac{s}{\sqrt{S_{xx}}} = (.606)(\sqrt{.1}) = .192.$$

Step 4. Compute the Critical Values and set up the Rejection Region (RR).

Since the test is two-sided, the Critical Values are $\pm t_{\alpha/2} = \pm t_{\alpha/2,(n-2)}$.

For ***n* = 5** and **α =.05** from ***t*** -table we find

$$\pm t_{\alpha/2,(n-2)} = \pm t_{.025,3} = \pm 3.182.$$

So the **RR** is: $RR = \{t : \ |T| \geq 3.182\}.$

Summary Example-Solution

Step 5. **Decision** (Use **Decision Rule** based on CV method).

Since **the observed value** of TS: $T_{ob} = 3.65$
falls the RR (3.65>3.182!),

we **reject** H_0 at $a = .05$ level of significance.

Step 6. **Conclusion** (The answer of question)
Yes, the data provide sufficient evidence to indicate that the slope *β* differs from 0.

Summary Example-Solution

(b) P-value approach

Steps 1-3 are the same as in CV approach.

Step 4' Compute the ***P*** -value corresponding to the observed value of $TS : T_{ob} = 3.65.$

Since the test is two-sided the ***P*** -value is

$$P - value = 2.P(T > |T_{ob}|) = 2.P(T > 3.65),$$

where $T \sim t_{(n-2)} = t_3.$

Using ***t*** -table we find that

$$.01 < P(T > 3.65) < .025.$$

Therefore

$$.02 < P - value < .05.$$

Summary Example-Solution

Remark.

In contrast to $\boldsymbol{Z}$ -table, $\boldsymbol{t}$ -table is not reach, by this reason we can find only limits for $\boldsymbol{P}$ -values.

Step 5' Decision

(Use **Decision Rule** based on $\boldsymbol{P}$ -value approach)

Since given significance level $a = .05 > P - value$,
we **reject** H_0.

Step 6' (**Conclusion**) is the same as **Step 6** in CV approach.

Summary Example-Solution

(c) CI-approach

Step1. Compute $100(1-a)\% = 100(1-.05)\% = 95\%$ CI for b.
In Part (iv) we have constructed this **CI:**

$$\hat{b} \pm t_{a/2,(n-2)} \cdot s_{\hat{b}} = (.09, 1.31).$$

Step 2. Decision (Use **Decision Rule** based on CI approach)
Since the specified value $b_0 = 0$ **does not fall** the **95% CI**, we **reject** H_0.

Summary Example-Solution

(vi) Find a 90% CI for the regression line $m_{y|x_0} = E\left[Y \middle| X = x_0\right]$, **when** $x_0 = 1$.

Solution.

We know that a $100(1-a)\%$ CI for $m_{y|x_0}$ is

$$\hat{Y} \pm t_{a/2,(n-2)} \cdot s_{\hat{Y}} = \hat{a} + \hat{b}\, x_0 \pm t_{a/2,(n-2)} \cdot s_{\hat{Y}}, \qquad (1)$$

where

$$s_{\hat{Y}} = s\sqrt{\frac{1}{n} + \frac{(x_0 - \overline{x})^2}{S_{xx}}}.$$

Summary Example-Solution

We have $\hat{a} = 1, \hat{b} = .7, s = .606.$ (2)

For $n = 5, \overline{x} = 0, S_{xx} = 10,$ we find

$$\frac{1}{n} + \frac{(x_0 - \overline{x})^2}{S_{xx}} = \frac{1}{5} + \frac{(1-0)^2}{10} = .3. \qquad (3)$$

For $n = 5, a = .10, \frac{a}{2} = .05,$ from ***t*** -table we find

$$t_{a/2,(n-2)} = t_{.05,3} = 2.353. \qquad (4)$$

Summary Example-Solution

Substituting (2) - (4) into (1) we find the desired 90% CI for $m_{y|x=1} = E[Y|x=1]$:

$$\hat{a} + \hat{b}x_0 \pm t_{a/2,(n-2)} \cdot s_{\hat{Y}}$$
$$= \left[1 + (.7)(1)\right] \pm (2.353)(.606)\sqrt{.3}$$
$$= 1.7 \pm .781 = (.919, 2.481).$$

Thus, a **90% CI** for $m_{y|x=1} = E[Y|x=1]$ is the interval **(.919, 2.481)**.

Summary Example-Solution

Interpretation of CI.

We are 90% confident that, when the independent variable takes on the value $\boldsymbol{x}=1$, the conditional mean value $m_{y|x=1} = E[Y|x=1]$ of the dependent variable $\boldsymbol{Y}$ is between **.918** and **2.481.**

Obviously, the CI is very wide, but it is not surprising, because it is based on very small ($\boldsymbol{n}=5$) sample size.

Summary Example-Solution

(vii) Suppose that the experiment that generated the data in Part (i) is to be run again with $x = x_0 = 2$.
Predict the particular value of Y with $\alpha = .1$, that is, construct 90% PI for $m_{Y|x=2} = E[Y|x=2]$.

Solution.

We know that for given $a\ (0 \le a \le 1)$, a $100(1-a)\%$ PI for $m_{Y|x_0}$ is

$$\hat{Y} \pm t_{a/2,(n-2)} \cdot s\sqrt{1 + \frac{1}{n} + \frac{(x_0 - \overline{x})^2}{S_{xx}}}. \qquad (1)$$

Summary Example-Solution

From Part (i), we have

$$\hat{a} = 1, \quad \text{and} \quad \hat{b} = .7.$$

Hence the **predicted value** of $\boldsymbol{Y}$ with $\boldsymbol{x} = 2$ is

$$\hat{Y} = \hat{a} + \hat{b}\,x_0 = 1 + (.7)(2) = 1 + 1.4 = 2.4. \qquad (2)$$

From Part (vi), we have for $n = 5, a = .10,$

$$t_{a/2,(n-2)} = t_{.05,3} = 2.353 \qquad (3)$$

From Part (iv), we have $s = .606,$
and from Part (i) $S_{xx} = 10$ and $\overline{x} = 0.$

Summary Example-Solution

Hence for $x_0 = 2$,

$$1+\frac{1}{n}+\frac{(x_0-\overline{x})^2}{S_{xx}}=1+\frac{1}{5}+\frac{(2-0)^2}{10}=1.6. \qquad (4)$$

Substituting (2) - (4) into (1) we obtain

$$\hat{Y}\pm t_{a/2,(n-2)}\cdot s\sqrt{1+\frac{1}{n}+\frac{(x_0-\overline{x})^2}{S_{xx}}}$$

$$=2.4\pm(2.353)(.606)\sqrt{1.6}=2.4\pm1.804=(.596,\ \ 4.204).$$

Thus, a 90% PI for $m_{Y|x=2}$ is the interval **(.596, 4.204).**

Summary Example-Solution

Remark.

If we construct a **90% PI** for $m_{Y|x=1}$ (instead of $\boldsymbol{x}=2$), we obtain

$$PI = 1.7 \pm 1.63 = (.07, 3.33).$$

Comparing this **PI** with the **90% CI** for $m_{Y|x=1}$ obtained in Part (iv):

$$CI = 1.7 \pm .781 = (.919, 2.481),$$

we found that **PI** is **wider** than **CI.**