Simple Linear Regression

Simple Linear Regression

- Simple Linear Regression Model
- Least Squares Method
- Coefficient of Determination
- Model Assumptions
- Testing for Significance

Simple Linear Regression

- Managerial decisions often are based on the relationship between two or more variables.
- From correlation coefficient r, we know if two variables are linearly related and the strength of the relationship.
- But we do not know the exact relationship.
  Mere knowledge of r is inadequate for prediction purpose.
- Regression analysis can be used to develop an equation showing how the variables are related.

Simple Linear Regression

- Simple linear regression involves one independent variable and one dependent variable.
- The relationship between the two variables is approximated by a straight line.
- Regression analysis involving two or more independent variables is called multiple regression.
- The variable being predicted is called the dependent or response variable and is denoted by y.
- The variables being used to predict the value of the dependent variable are called the independent or predictor or or regressor or explanatory variables and are denoted by x.

Simple Linear Regression Model

- The equation that describes how y is related to x and an error term is called the regression model.
- The simple linear regression model is:

$$y = \beta_0 + \beta_1 x + \varepsilon$$

where: β_0 and β_1 are called parameters of the model,

ε called the random error term.

For a fixed x, ε is a random variable with $E[\varepsilon] = 0$ and its variance is called error variance.

y is a random variable since ε is random .

The value x of the regressor variable is not random and, in fact, is measured with negligible error.

Simple Linear Regression Equation

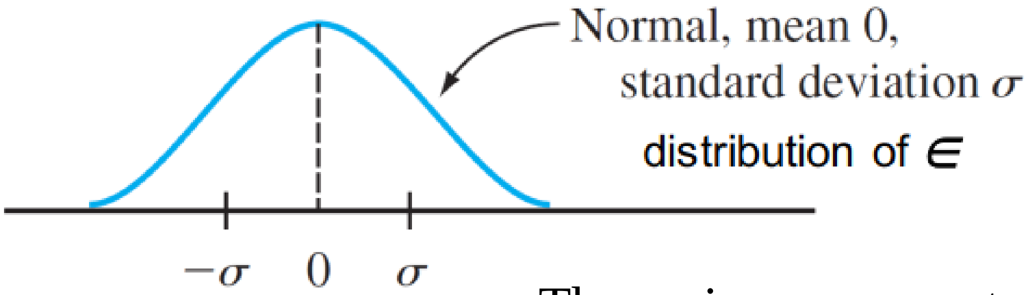
- The simple linear regression equation is:

$$E(y) = \beta_0 + \beta_1 x$$

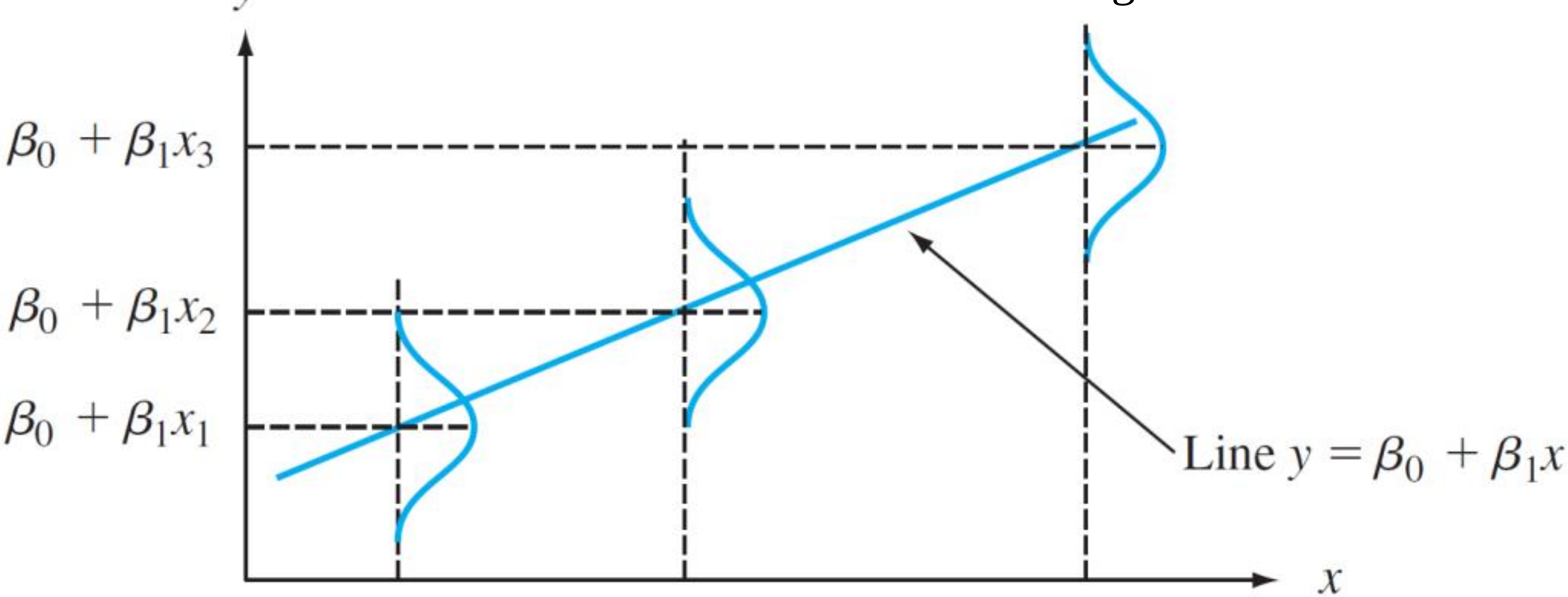
  - Graph of the regression equation is a straight line.
  - β_0 is the y intercept of the regression line.
  - β_1 is the slope of the regression line.
  - $E(y)$ is the expected value of y for a given x value.

We assume the variance (amount of variability) of the distribution of Y values to be the same at each different value of fixed x. (i.e. homogeneity of variance assumption)

When errors are normally distributed...



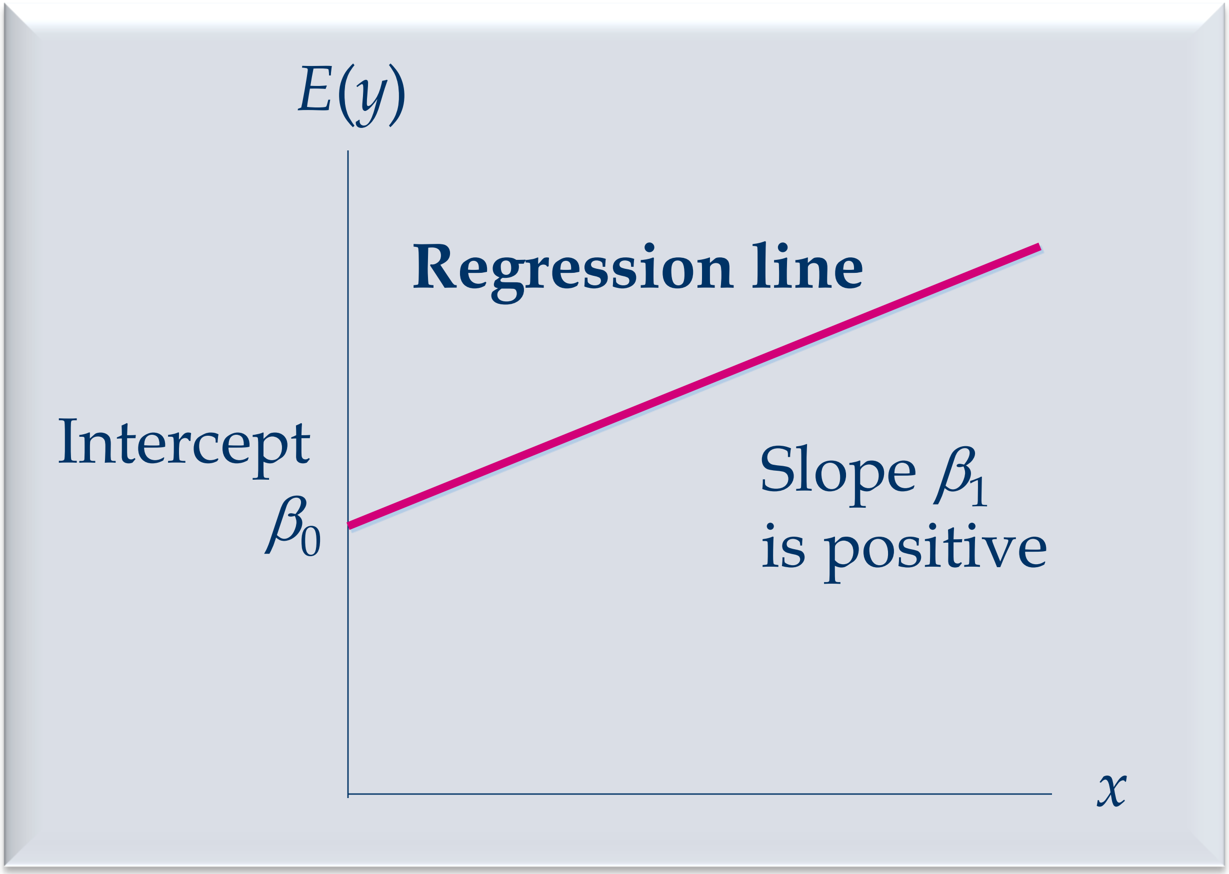
The variance parameter σ² determines the extent to which each normal curve spreads about the regression line. This variance can be used to tell us how good the linear fit is.



Distribution of Y for different values of x

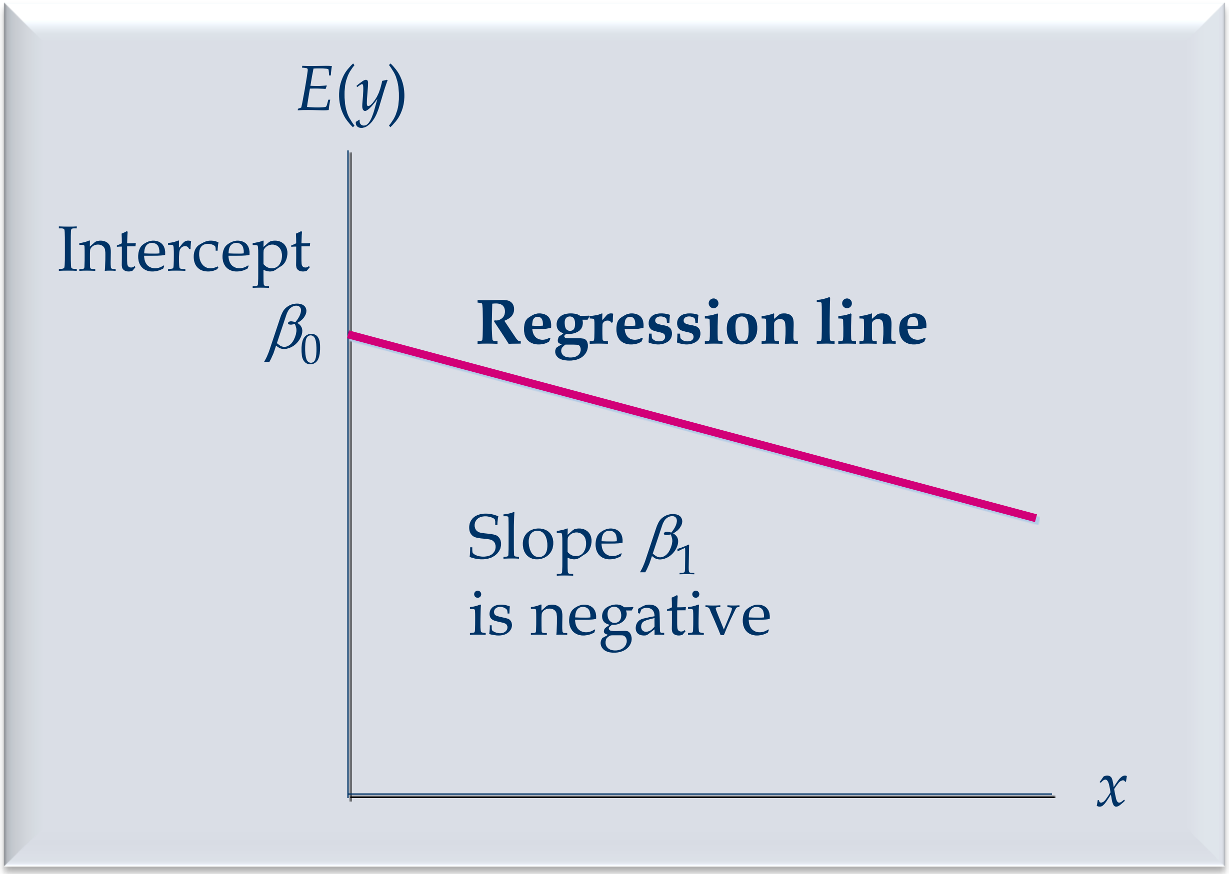
Simple Linear Regression Equation

- Positive Linear Relationship



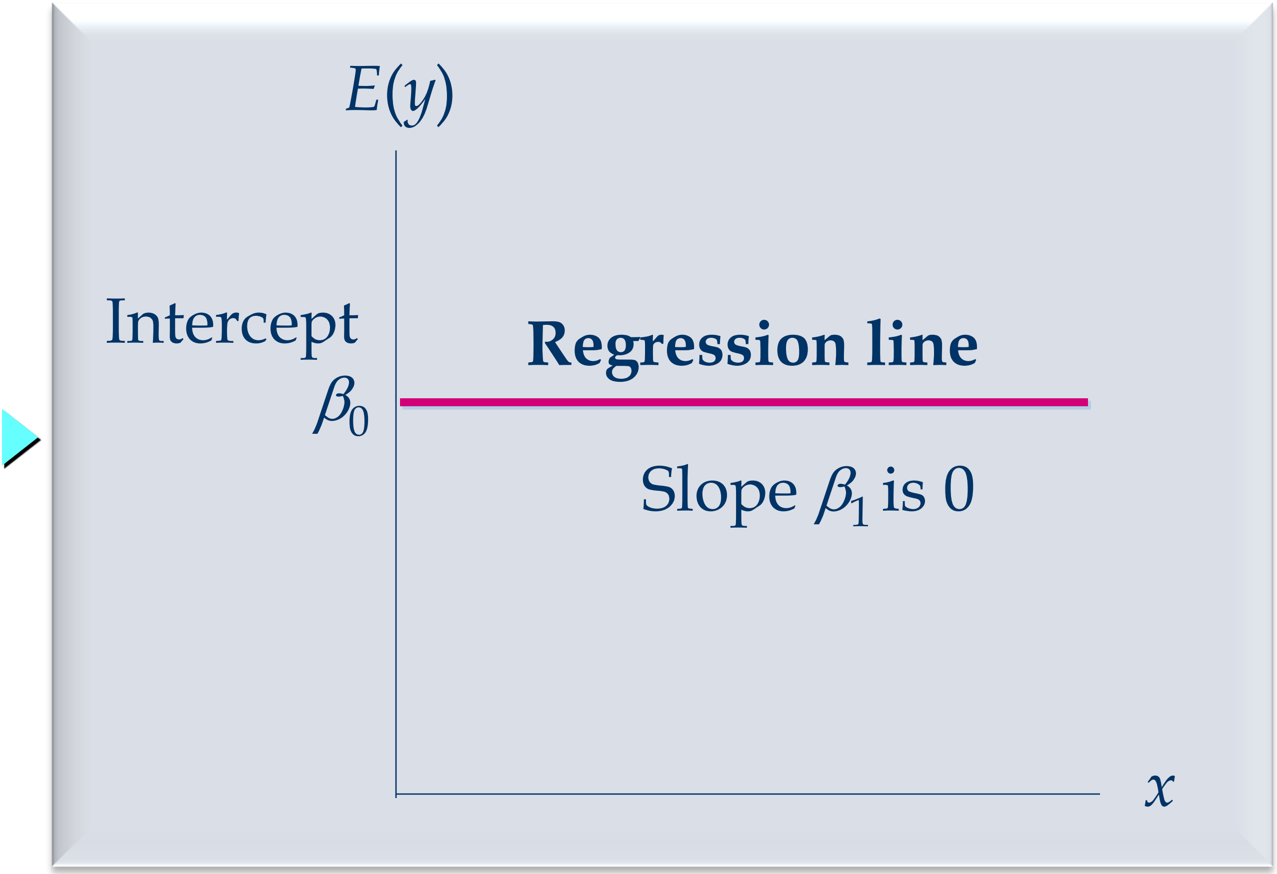
Simple Linear Regression Equation

- Negative Linear Relationship



Simple Linear Regression Equation

- No Relationship



Estimated Simple Linear Regression Equation

- The estimated simple linear regression equation

$$\hat{y} = b_0 + b_1 x$$

  - The graph is called the estimated regression line.
  - b_0 is the y intercept of the line.
  - b_1 is the slope of the line.
  - $\hat{y}$ is the estimated value of y for a given x value.

Estimation Process

Regression Model

$y = \beta_0 + \beta_1 x + \varepsilon$

Regression Equation

$E(y) = \beta_0 + \beta_1 x$

Unknown Parameters

β_0, β_1

Sample Data:

x	y
x_1	y_1
.	.
.	.
x_n	y_n

Estimated

Regression Equation

$\hat{y} = \mathbf{b}_0 + \mathbf{b}_1 x$

Sample Statistics

b_0, b_1

b_0 and b_1 provide estimates of β_0 and β_1

Least Squares Method

- Least Squares Criterion

$$\min \sum (y_i - \hat{y}_i)^2$$

where:

y_i = observed value of the dependent variable for the ith observation

$\hat{y}_i$ = estimated value of the dependent variable for the ith observation

$(y_i - \hat{y}_i)$ is called residual or error

$\sum (y_i - \hat{y}_i)^2$ is called residual or error sum of squares (SSE)

Least Squares Method

- Slope for the Estimated Regression Equation

$$b_1 = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2}$$

where:

x_i = value of independent variable for ith observation

y_i = value of dependent variable for ith observation

$\bar{x}$ = mean value for independent variable

$\bar{y}$ = mean value for dependent variable

- y-Intercept for the Estimated Regression Equation

$$b_0 = \bar{y} - b_1 \bar{x}$$

Simple Linear Regression

- Example: Auto Sales

An Auto sales company, as part of the advertising campaign runs one or more television commercials during the weekend preceding the sale. Data from a sample of 5 previous sales are.

Number of TV Ads (x)	Number of Cars Sold (y)
1	14
3	24
2	18
1	17
3	27
$\Sigma x = 10$	$\Sigma y = 100$
$\bar{x} = 2$	$\bar{y} = 20$

Estimated Regression Equation

- Slope for the Estimated Regression Equation

$$b_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} = \frac{20}{4} = 5$$

- y-Intercept for the Estimated Regression Equation

$$b_0 = \bar{y} - b_1 \bar{x} = 20 - 5(2) = 10$$

- Estimated Regression Equation

$$\hat{y} = 10 + 5x$$

Another formula

$$b_1 = \frac{n \sum x_i y_i - \sum x_i \sum y_i}{n \sum x_i{}^2 - (\sum x_i)^2}$$

Sum of Squares

- The Error sum of squares (SSE):

The line of "best" fit was that line with the smallest sum of squared residuals. This is also called the residual sum of squares.

$$SSE = \sum (y_i - \hat{y}_i)^2$$

- The regression sum of squares (SSR):

It measures the variability in y that is predicted by the model, i.e., the variability in $\hat{y}$.

$$SSR = \sum (\hat{y}_i - \bar{y})^2$$

- The total sum of squares (SST):

It measures the observed variability in y.

$$SST = \sum (y_i - \bar{y})^2$$

Explaining Variation

- One goal of regression is to "explain" the variation in y.
- For example, if x were height and y were weight, how would we explain the variation in weight?
  - That is, why do some people weigh more than others?
- Or if x were the hours spent studying for a math test and y were the score on the test, how would we explain the variation in scores?
  - That is, why do some people score higher than others?

Explaining Variation

- A certain amount of the variation in y can be explained by the variation in x.
  - Some people weigh more than others because they are taller.
  - Some people score higher on math tests because they studied more.
- But that is never the full explanation.
  - Not all taller people weigh more.
  - Not everyone who studies more scores higher.

Explaining Variation

- High degree of correlation between x and y
  $\Rightarrow$ variation in x explains most of the variation in y.
- Low degree of correlation between x and y
  $\Rightarrow$ variation in x explains only a little of the variation in y.
- In other words, the amount of variation in y that is explained by the variation in x should be related to r.

- Statisticians consider the predicted variation SSR to be the amount of variation in *y* (i.e., SST) that is *explained* by the model.
- The remaining variation in *y*, i.e., the *residual* variation SSE, is the amount that is *not explained* by the model.

Explaining Variation

SST = SSE + SSR

Explaining Variation

SST = SSR + SSE

Total variation in *y*
(to be explained)

Explaining Variation

$$\boxed{SST = SSR + SSE}$$

Total variation in *y*
(to be explained)

Variation in *y* that is
explained by the model

Explaining Variation

Variation in *y* that is
explained by the model

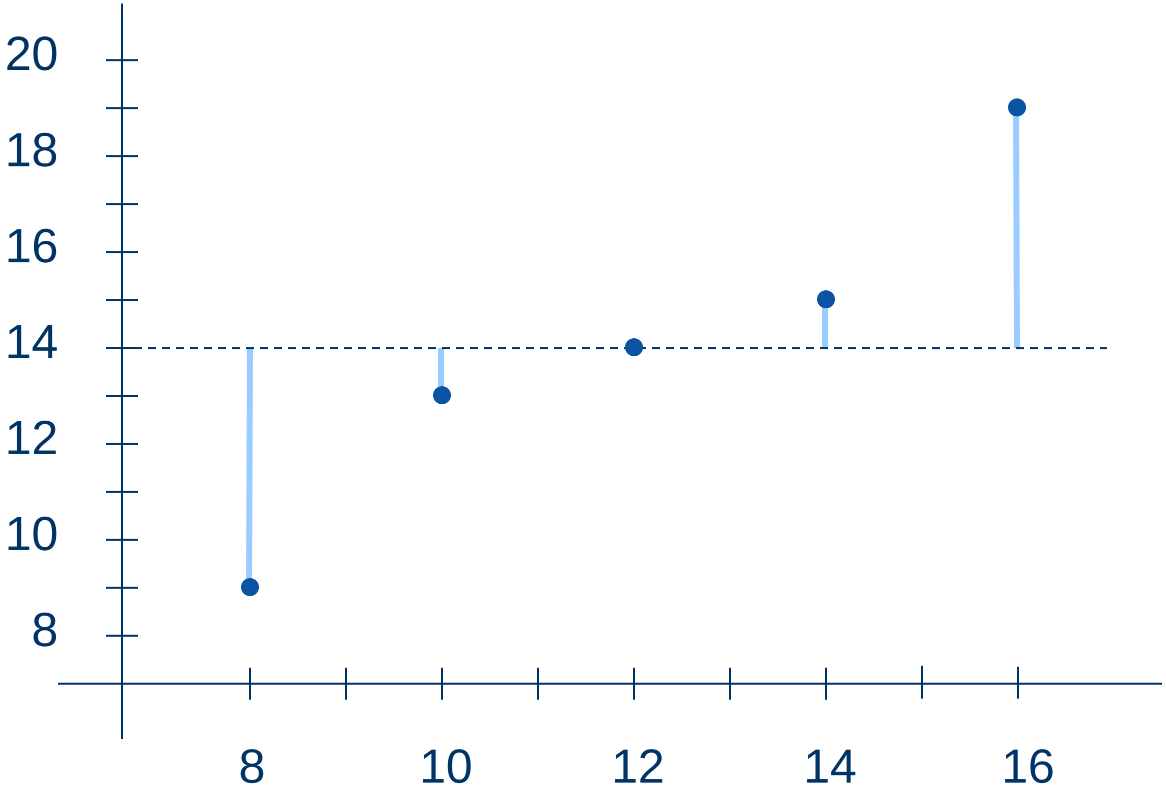
$$SST = SSR + SSE$$

Total variation in *y*

Variation in *y* that is
unexplained by the model

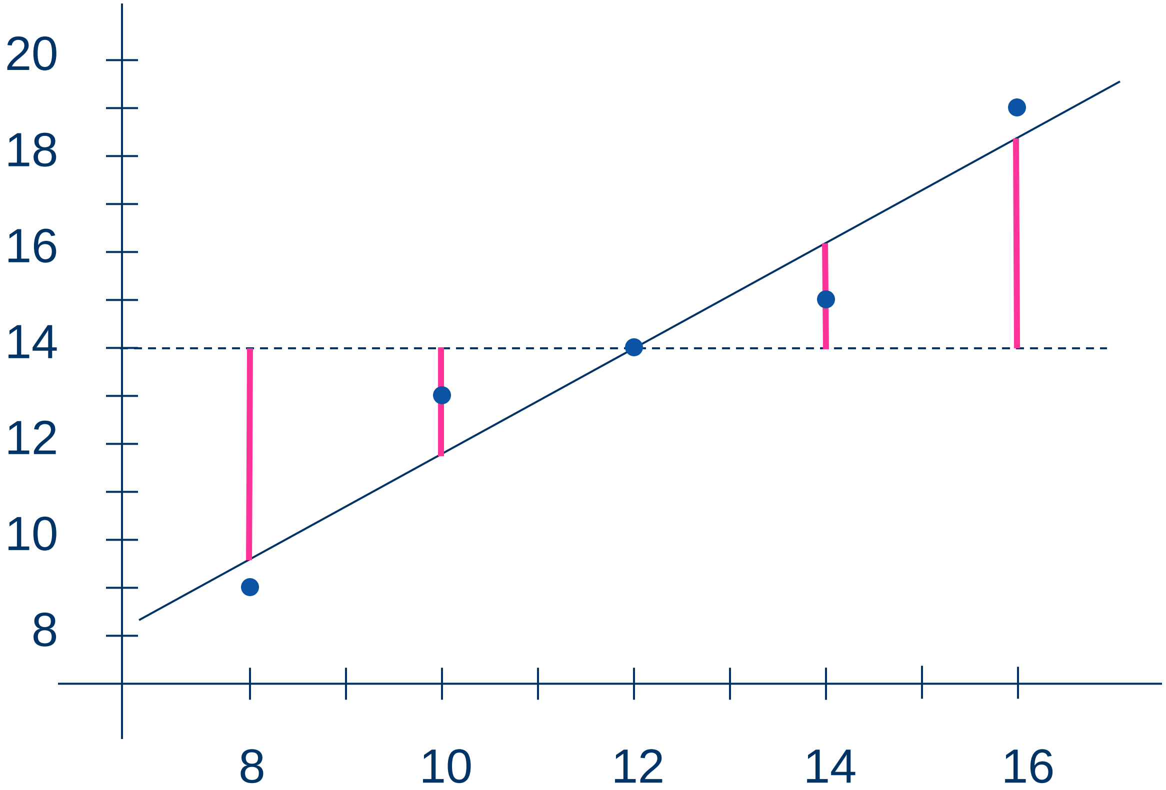
Example – SST, SSR, and SSE

- The total (observed) variation in y.



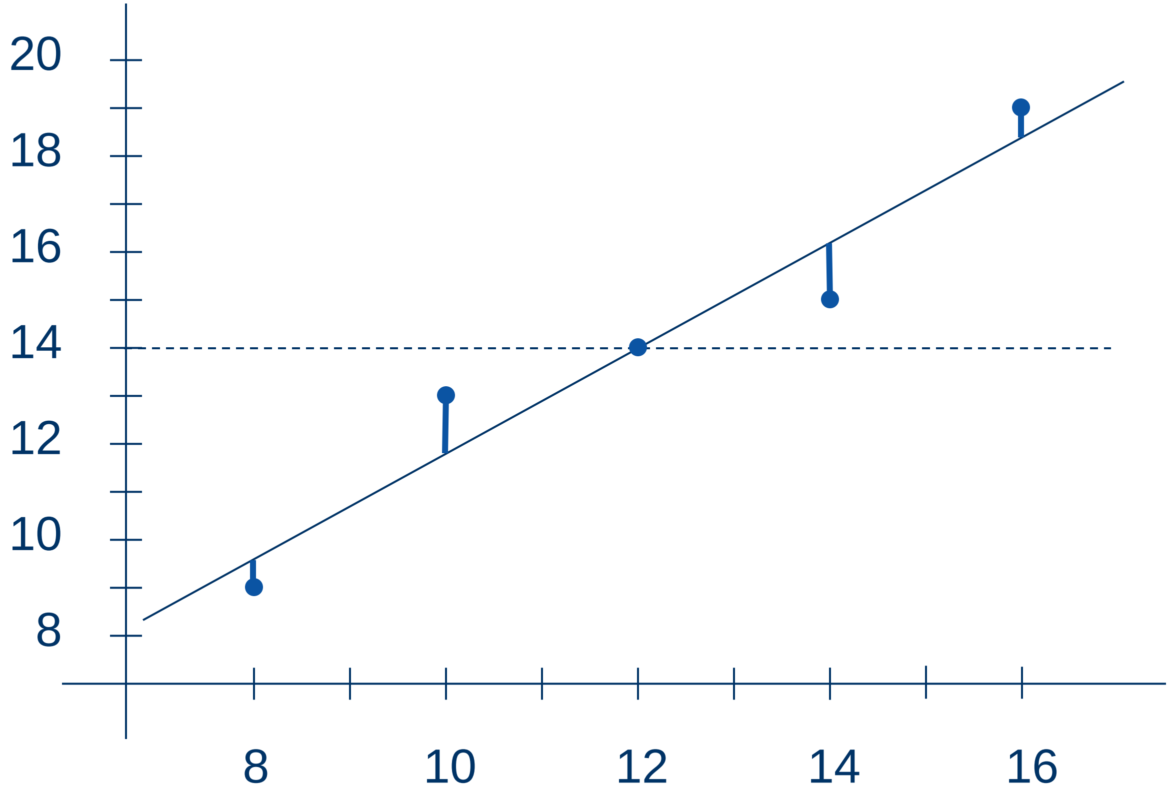
Example – SST, SSR, and SSE

- The variation in y that is explained by the model (i.e., due to the variation in x).



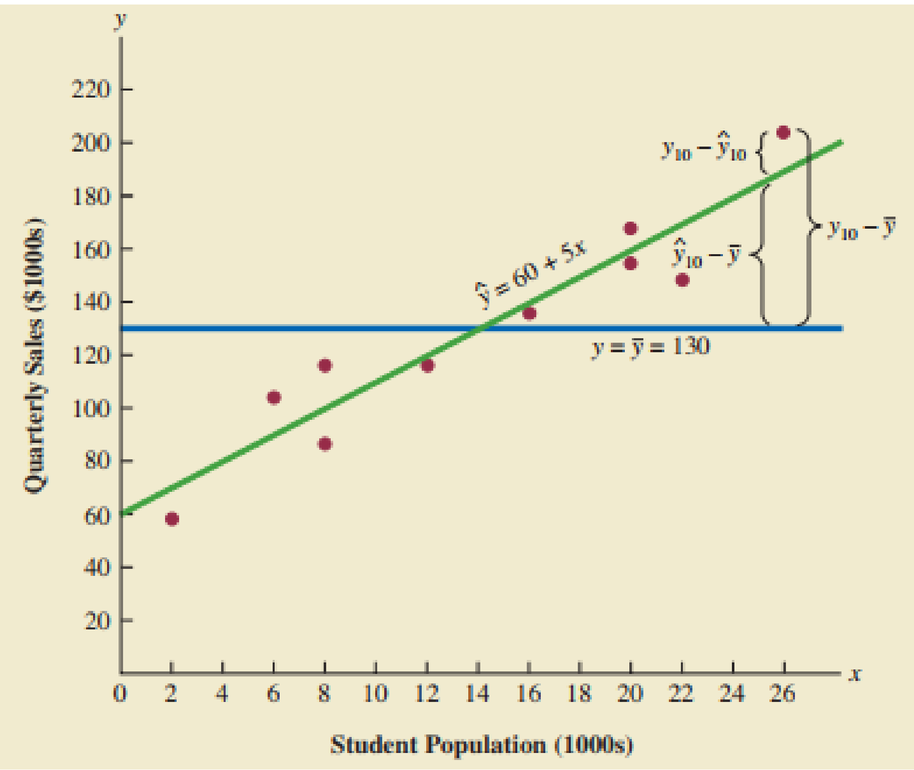
Example – SST, SSR, and SSE

- The variation in y that is not explained by the model (i.e., "random" variation).



Coefficient of Determination

- Relationship Among SST, SSR, SSE



SST = SSR + SSE

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

where:

SST = total sum of squares

SSR = sum of squares due to regression

SSE = sum of squares due to error

Coefficient of Determination

- Therefore,

$$SSR = r^2 \times SST$$

$$SSE = (1 - r^2) \times SST$$

- r^2 is the proportion of variation in y that is explained by the model and $1 - r^2$ is the proportion that is not explained by the model.

Coefficient of Determination

- The coefficient of determination is:
  - $r^2 = SSR/SST$ $r^2 = 1- SSE/SST$

  where:

  SSR = sum of squares due to regression

  SST = total sum of squares

- *In the Auto sales problem:* r^2 = SSR/SST = 100/114 = .8772
- The regression relationship is very strong; 87.72% of the variability in the number of cars sold can be explained by the linear relationship between the number of TV ads and the number of cars sold.

Sample Correlation Coefficient

$$r_{xy} = (\text{sign of } b_1)\sqrt{\text{Coefficient of Determination}}$$

$$r_{xy} = (\text{sign of } b_1)\sqrt{r^2}$$

where:

b_1 = the slope of the estimated regression equation $\hat{y} = b_0 + b_1 x$

The sign of b_1 in the equation $\hat{y} = 10 + 5x$ is "+".

- $r_{xy} = +\sqrt{.8772}$
- $r_{xy} = +.9366$

Assumptions About the Error Term ε

1. The error ε is a random variable with mean of zero.
2. The variance of ε, denoted by σ^2, is the same for all values of the independent variable.
3. The values of ε are independent.
4. The error ε is a normally distributed random variable.

Testing for Significance

- To test for a significant regression relationship, we must conduct a hypothesis test to determine whether the value of β_1 is zero.

- Two tests are commonly used:

  t Test and *F* Test

- Both the *t* test and *F* test require an estimate of σ^2, the variance of ε in the regression model.

Testing for Significance

- An Estimate of σ^2

The mean square error (MSE) provides the estimate of σ^2. The notation s_e^2 is also used for MSE.

$$s_e^2 = \text{MSE} = \text{SSE}/(n - 2)$$

- An Estimate of σ
  - To estimate σ we take the square root of σ^2.
  - The resulting s_e is called the standard error of the estimate.

$$s_e = \sqrt{\text{MSE}} = \sqrt{\frac{SSE}{n-2}}$$

Testing for Significance: *t* Test

- Hypotheses

$$H_0: \beta_1 = 0$$
$$H_1: \beta_1 \neq 0$$

- Test Statistic

$$t = \frac{b_1}{s_{b1}}$$

where

$$s_{b1} = \frac{s_e}{\sqrt{\sum (x_i - \bar{x})^2}}$$

Testing for Significance: t Test

- Rejection Rule

Reject H_0 if p-value $\leq \alpha$
or $t \leq -t_{\alpha/2}$ or $t \geq t_{\alpha/2}$

where:

$t_{\alpha/2}$ is based on a t distribution
with n - 2 degrees of freedom

Testing for Significance: t Test

1. Determine the hypotheses.

$$H_0: \beta_1 = 0$$

$$H_a: \beta_1 \neq 0$$

2. Specify the level of significance.

$\alpha = .05$

3. Select the test statistic.

$$t = \frac{b_1}{s_{b_1}}$$

4. State the rejection rule.

Reject H_0 if p-value $\leq .05$ or $|t| > 3.182$ (with 3 degrees of freedom)

Testing for Significance: t Test

- 5. Compute the value of the test statistic.

$$t = \frac{b_1}{s_{b_1}} = \frac{5}{1.08} = 4.63$$

- 6. Determine whether to reject H_0.

$t = 4.63 > 3.182$. We can reject H_0.

Confidence Interval for β_1

- We can use a 95% confidence interval for β_1 to test the hypotheses just used in the t test.
- H_0 is rejected if the hypothesized value of β_1 is not included in the confidence interval for β_1.

Confidence Interval for β_1

- The form of a confidence interval for β_1 is:

$$b_1 \pm t_{\alpha/2} s_{b_1}$$

b_1 is the point estimator

$t_{\alpha/2}\, s_{b_1}$ is the margin of error

where $t_{\alpha/2}$ is the t value providing an area of $\alpha/2$ in the upper tail of a t distribution with n - 2 degrees of freedom

Confidence Interval for β_1

- Rejection Rule

  Reject H_0 if 0 is not included in the confidence interval for β_1.

- 95% Confidence Interval for β_1

  $$b_1 \pm t_{\alpha/2} s_{b_1} = 5 +/- 3.182(1.08) = 5 +/- 3.44$$

  or 1.56 to 8.44

- Conclusion

  0 is not included in the confidence interval. Reject H_0

Testing for Significance: F Test

- Hypotheses

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

- Test Statistic

F = MSR/MSE

- Rejection Rule

Reject H_0 if
p-value $\leq \alpha$
or $F \geq F_\alpha$

where:

F_α is based on an F distribution with
1 degree of freedom in the numerator and
n - 2 degrees of freedom in the denominator

Testing for Significance: F Test

1. Determine the hypotheses.

$$H_0: \beta_1 = 0$$
$$H_1: \beta_1 \neq 0$$

2. Specify the level of significance.

$\alpha = .05$

3. Select the test statistic.

F = MSR/MSE

4. State the rejection rule.

Reject H_0 if p-value $\leq$ .05 or $F \geq 10.13$ (with 1 d.f. in numerator and 3 d.f. in denominator)

Testing for Significance: F Test

- 5. Compute the value of the test statistic.

$$F = \text{MSR}/\text{MSE} = 100/4.667 = 21.43$$

- 6. Determine whether to reject H_0.

F = 17.44 provides an area of .025 in the upper tail. Thus, the p-value corresponding to F = 21.43 is less than .025. Hence, we reject H_0.

The statistical evidence is sufficient to conclude that we have a significant relationship between the number of TV ads aired and the number of cars sold.

Some Cautions about the Interpretation of Significance Tests

- Rejecting H_0: $\beta_1 = 0$ and concluding that the relationship between x and y is significant does not enable us to conclude that a cause-and-effect relationship is present between x and y.

- Just because we are able to reject H_0: $\beta_1 = 0$ and demonstrate statistical significance does not enable us to conclude that there is a linear relationship between x and y.