BAN 602: Quantitative Fundamentals

Spring, 2020 Lecture Slides – Week 7

CAL STATE
EAST BAY

Agenda

- Simple Linear Regression
  - Simple Linear Regression Model
  - Least Squares Method
  - Coefficient of Determination
  - Model Assumptions
  - Testing for Significance
  - Estimation and Prediction
  - Residual Analysis
- Multiple Regression
  - Multiple Regression Model
  - Least Squares Method
  - Multiple Coefficient of Determination
  - Model Assumptions
  - Testing for Significance
  - Estimation and Prediction
  - Categorical Independent Variables
  - Residual Analysis

Simple Linear Regression

- Managerial decisions often are based on the relationship between two or more variables.
- Regression analysis can be used to develop an equation showing how the variables are related.
- The variable being predicted is called the dependent variable and is denoted by y.
- The variables being used to predict the value of the dependent variable are called the independent variables and are denoted by x.
- 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.

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 the parameters of the model.
ε is a random variable called the error term.

- The simple linear regression equation is

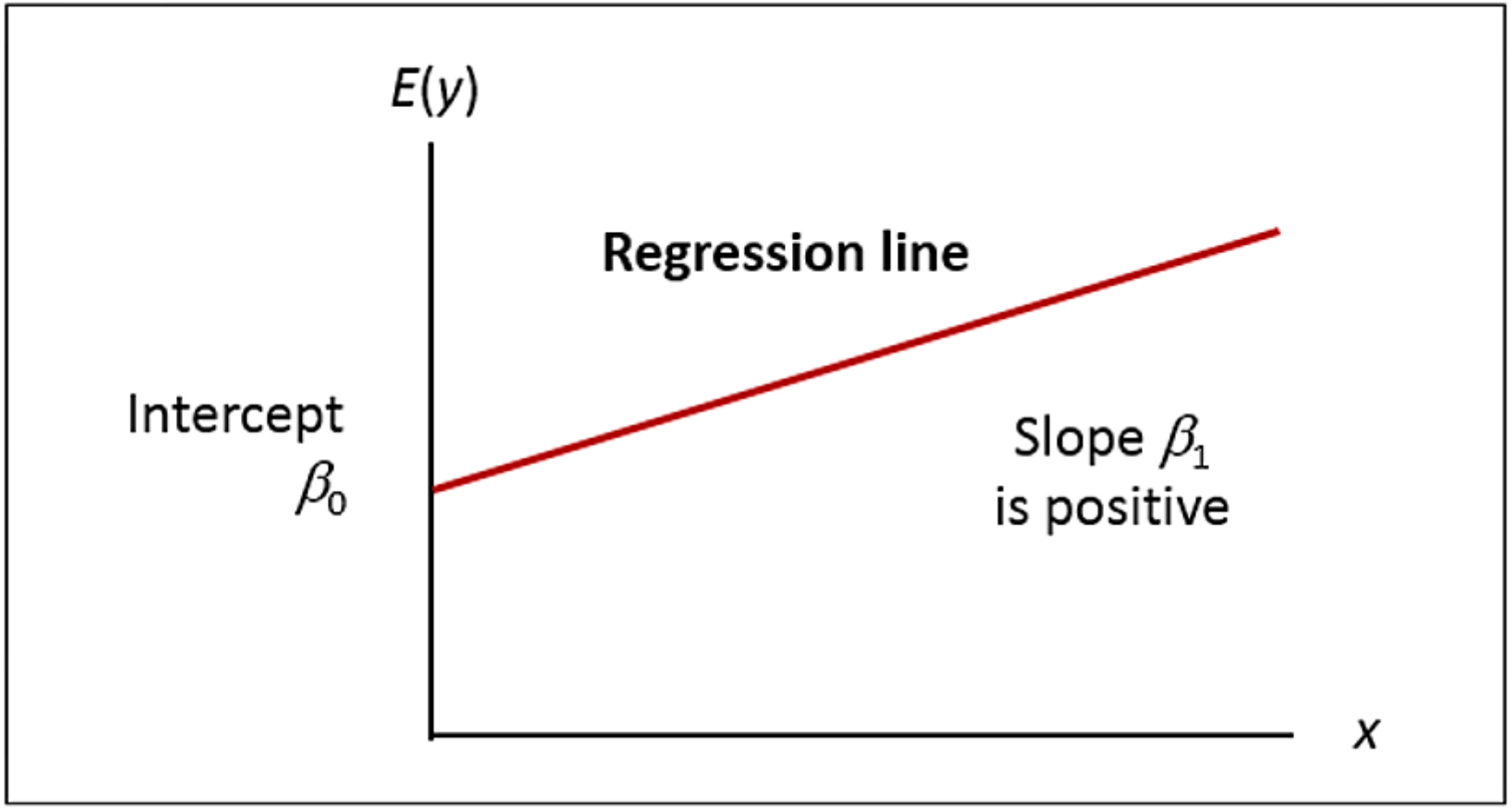
$$E(y) = \beta_0 + \beta_1 x$$

where: β_0 is the y intercept of the regression line.
β_1 are the slope of the regression line.
$E(y)$ is the expected value of y for a given x value.

The graph of the regression equation is a straight line.

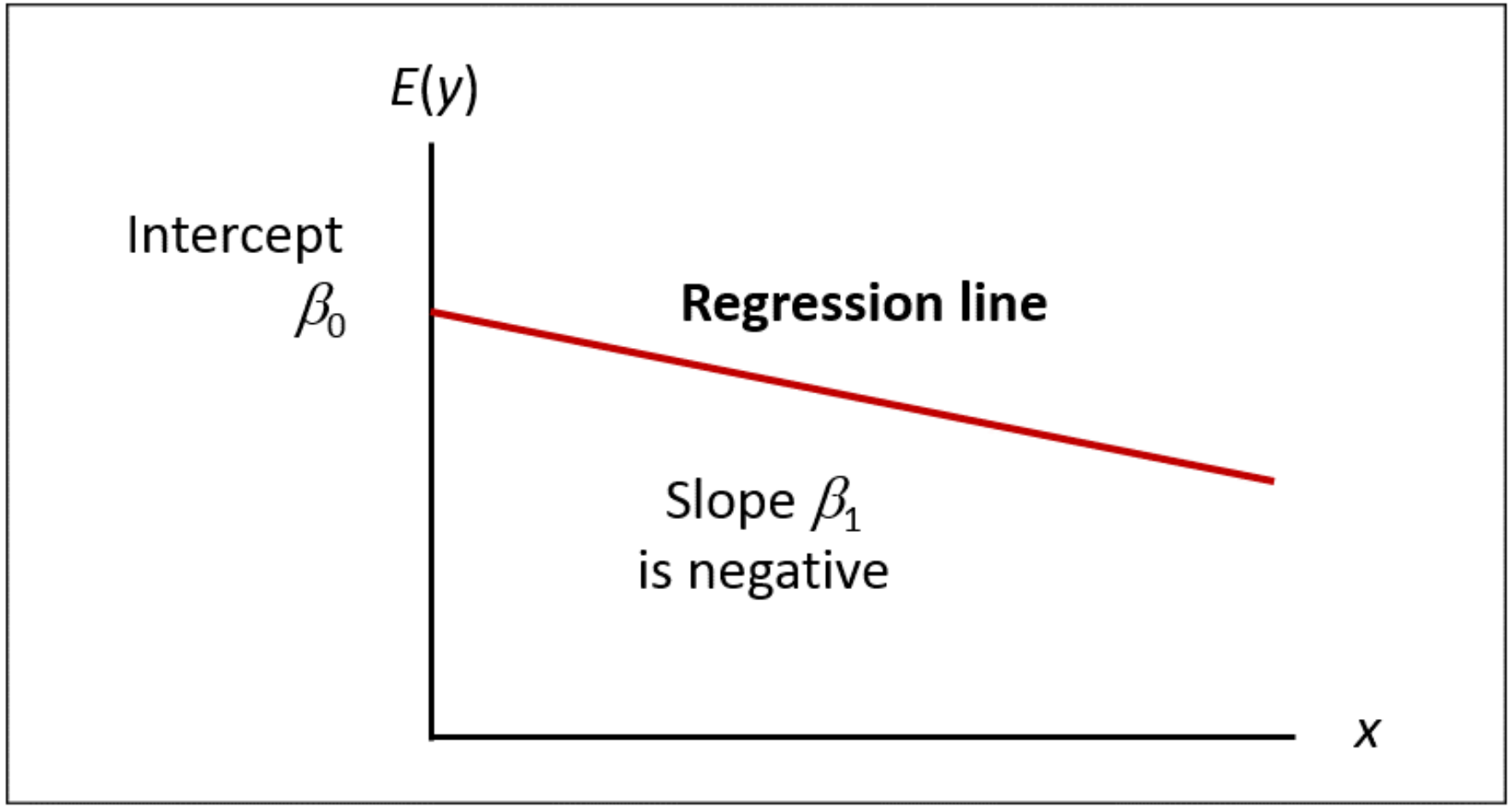
Simple Linear Regression Equation

Positive Linear Relationship



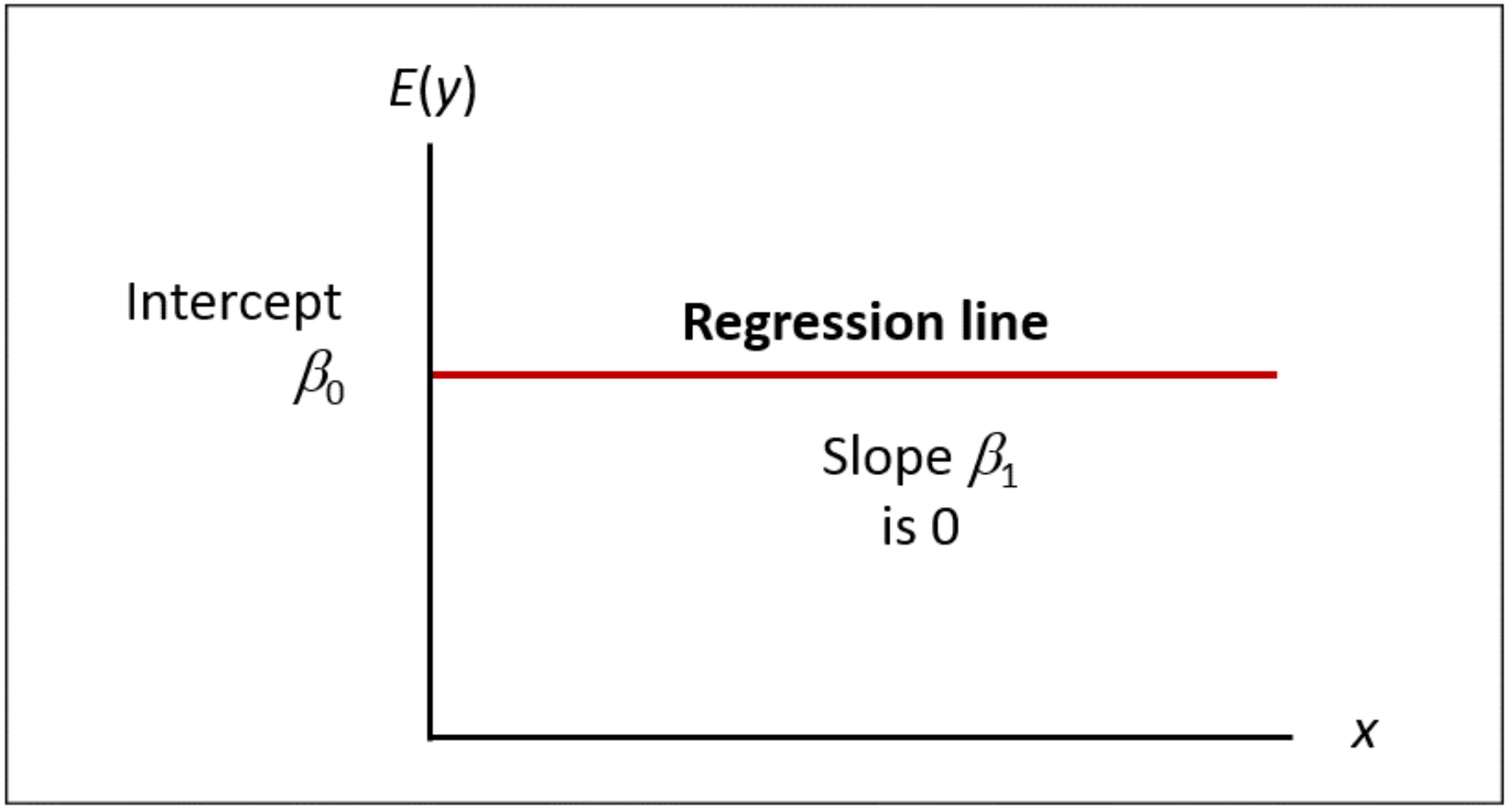
Simple Linear Regression Equation

Negative Linear Relationship



Simple Linear Regression Equation

No Relationship



Estimated Simple Linear Regression Equation

The estimated simple linear regression equation:

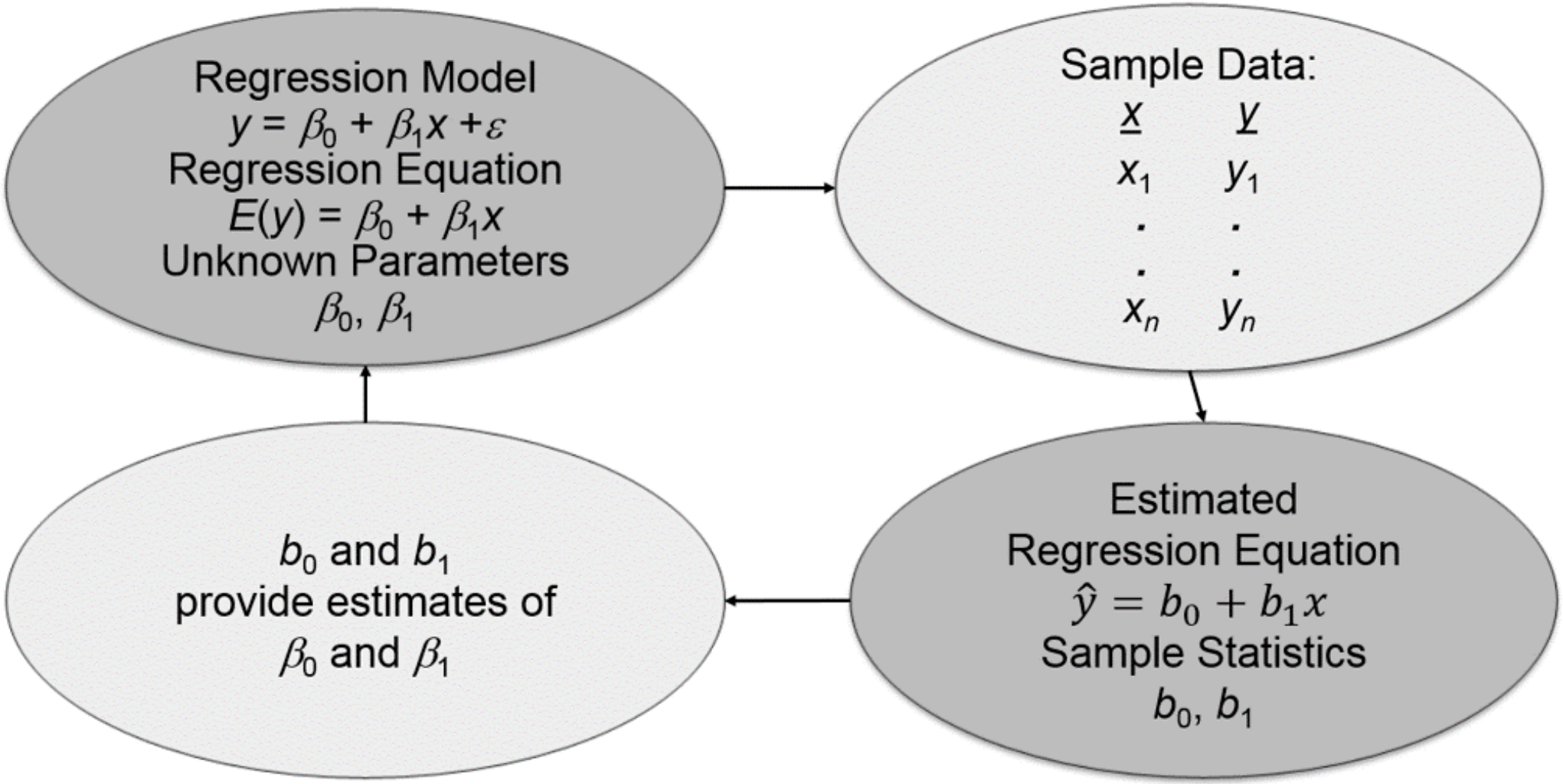
$$\hat{y} = b_0 + b_1 x$$

where: b_0 is the y intercept of the regression line.

b_1 are the slope of the regression line.

$\hat{y}$ is the estimated value of y for a given x value.

Estimation Process



Least Squares Method

Least Squares Criterion

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

where:

y_i = the observed value of the dependent variable for the i^{th} observation

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

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 = the value of the independent variable for the i^{th} observation

y_i = the value of the dependent variable for the i^{th} observation

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

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

y-intercept for the estimated regression equation: $b_0 = \bar{y} - b_1\bar{x}$

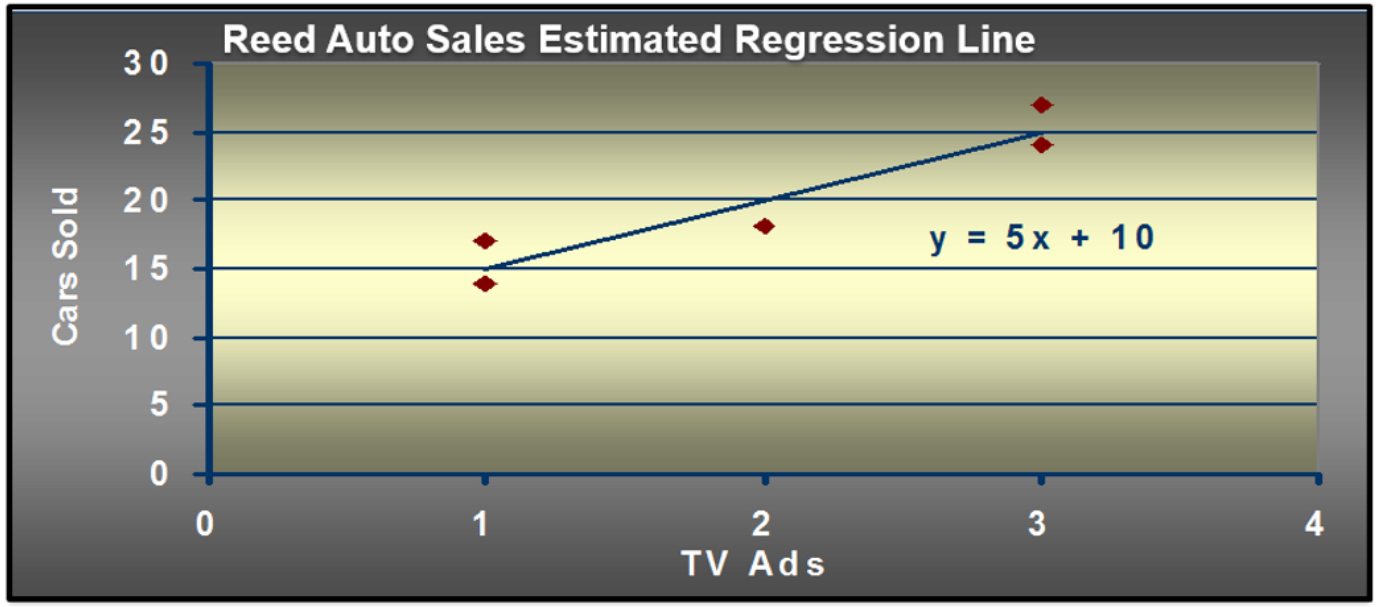
Simple Linear Regression

Reed Auto periodically has a special week-long sale. As part of the advertising campaign Reed runs one or more television commercials during the weekend preceding the sale. Here are the data from a sample of 5 previous sales:

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



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

The coefficient of determination is: $r^2 = \frac{SSR}{SST}$

where:

SSR = sum of squares due to regression

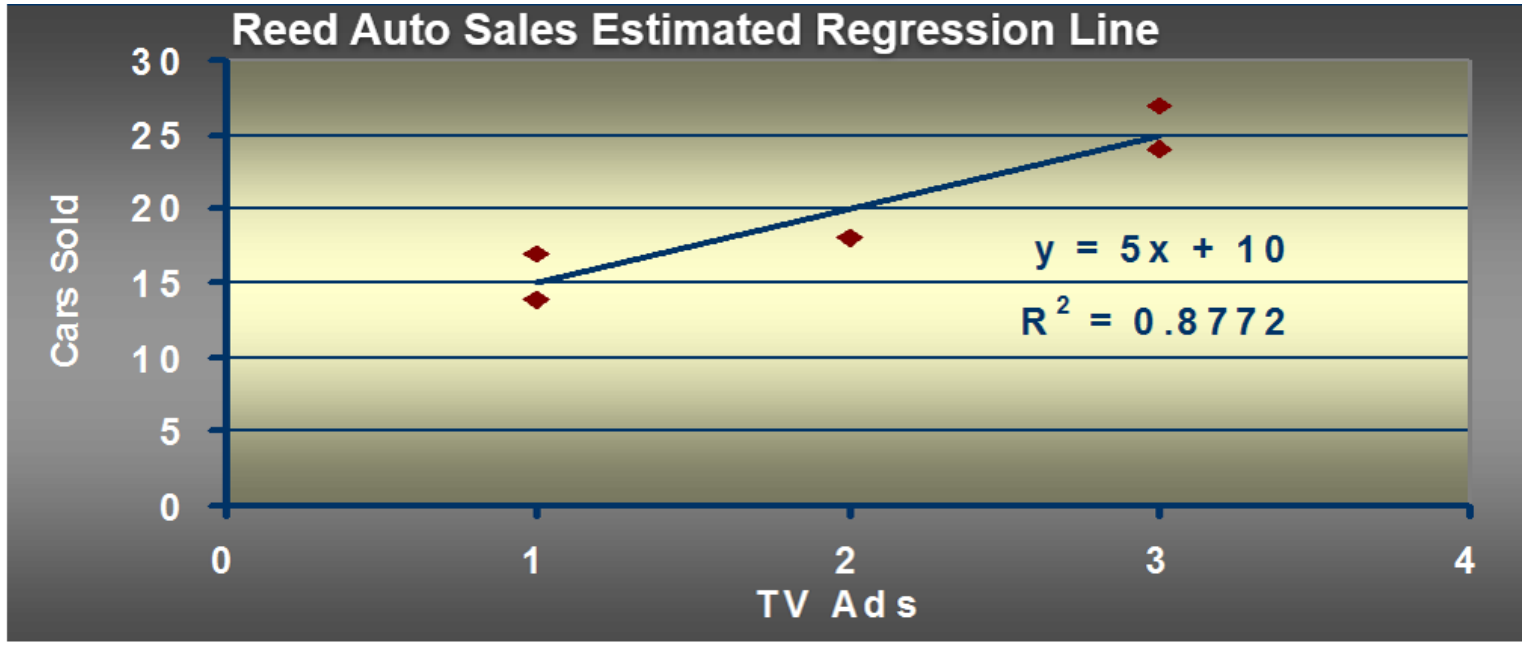
SST = total sum of squares

$$r^2 = \frac{SSR}{SST} = \frac{100}{114} = 0.8772$$

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

Adding r^2 Value to Scatter Diagram



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

where: b_1= the slope of the estimated regression equation

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

The sign of b_1 in the equation $\hat{y} = 10 + 5x$ is positive, so

$$r_{xy} = \sqrt{0.8772} = \mathbf{0.9366}$$

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 are the 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.

An Estimate of σ^2: The mean square error (MSE) provides the estimate of σ^2, and the notation s^2 is also used.

$$s^2 = \text{MSE} = \frac{SSE}{n-2}$$

where:

$$SSE = \sum (y_i - \hat{y}_i)^2 = \sum (y_i - b_0 - b_1 x_i)^2$$

Testing for Significance

To estimate σ, we take the square root of s^2.

The resulting s is called the standard error of the estimate.

$$s^2 = \text{MSE} = \frac{SSE}{n-2}$$

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

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

Test Statistic: $t = \frac{b_1}{s_{b_1}}$ where $s_{b_1} = \frac{s}{\sqrt{\sum(x_i - \bar{x})^2}}$

Rejection Rule: Reject H_0 if p –value $\leq \alpha$

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

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. $\alpha = 0.05$

3. Select the test statistic. $t = \dfrac{b_1}{s_{b_1}}$

4. State the rejection rule. Reject H_0 if the p-value ≤ 0.05 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.

$t = 4.4541$ provides an area of 0.01 in the upper tail. Hence the p-value < 0.02.

Also, $t = 4.63 > 3.182$, so 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.

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

where

b_1 = the point estimator,

$t_{\alpha/2}(s_{b_1})$ = the margin of error, and

$t_{\alpha/2}$ = 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

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

$$5 \pm 3.182(1.08)$$

$$5 \pm 3.44$$

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: $F = \frac{MSR}{MSE}$
- Rejection Rule: Reject H_0 if the p-value ≤ 0.05 or if $F \geq F_\alpha$
  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.

Example:

1. Determine the hypotheses. $H_0: \beta_1 = 0$, $H_a: \beta_1 \neq 0$
2. Specify the level of significance. $\alpha = 0.05$
3. Select the test statistic. $F = \frac{MSR}{MSE}$
4. State the rejection rule. Reject H_0 if the p-value ≤ 0.05 or $F \geq 10.13$ with 1 *df* in the numerator and 3 *df* in the denominator.

Testing for Significance: F Test

5. Compute the value of the test statistic. $F = \dfrac{MSR}{MSE} = \dfrac{100}{4.667} = 21.43$

6. Determine whether to reject H_0.

   $F = 17.44$ provides an area of 0.025 in the upper tail. Thus, the p-value corresponding to $F = 21.43$ is less than 0.025. 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.

Cautions:

- 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.

Using the Estimated Regression Equation for Estimation & Prediction

- A confidence interval is an interval estimate of the *mean value of* y for a given value of x.
- A prediction interval is used whenever we want to *predict an individual value of* y for a new observation corresponding to a given value of x.
- The margin of error is larger for a prediction interval.
- Confidence Interval Estimate of $E(y^*)$:

$$\hat{y}^* \pm t_{\alpha/2}\left(s_{\hat{y}^*}\right)$$

- Prediction Interval Estimate of $E(y^*)$:

$$\hat{y}^* \pm t_{\alpha/2}\left(s_{\text{pred}}\right)$$

where $1-\alpha$ is the confidence coefficient and $t_{\alpha/2}$ is based on a t distribution with $n-2$ degrees of freedom.

Point and Interval Estimations

If 3 TV ads are run prior to a sale, we expect the mean number of cars sold to be: $\hat{y} = 10 + 5(3) = 25$ cars

Estimate of the Standard Deviation of $\hat{y}^*$: $s_{\hat{y}^*} = s\sqrt{\frac{1}{n} + \frac{(x^* - \bar{x})^2}{\sum(x_i - \bar{x})^2}}$

$$s_{\hat{y}^*} = 2.16025\sqrt{\frac{1}{5} + \frac{(3-2)^2}{(1-2)^2 + (3-2)^2 + \cdots + (3-2)^2}} \qquad s_{\hat{y}^*} = 2.16025\sqrt{\frac{1}{5} + \frac{1}{4}} = \mathbf{1.4491}$$

The 95% confidence interval estimate of the mean number of cars sold when 3 TV ads are run is:

$$\hat{y}^* \pm t_{\alpha/2}(s_{\hat{y}^*})$$

$$25 \pm 3.1824(1.4491)$$

$$25 \pm 4.61$$

20.39 to 29.61 cars

Prediction Interval for y^*

Estimate of the Standard Deviation of an Individual Value of y^*

$$s_{\text{pred}} = s\sqrt{1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{\sum(x_i - \bar{x})^2}}$$

$$s_{\text{pred}} = 2.16025\sqrt{1 + \frac{1}{5} + \frac{1}{4}}$$

$$s_{\text{pred}} = 2.16025(1.20416) = \mathbf{2.6013}$$

The 95% prediction interval estimate of the number of cars sold in one particular week when 3 TV ads are run.

$$\hat{y}^* \pm t_{\alpha/2}(s_{\text{pred}})$$

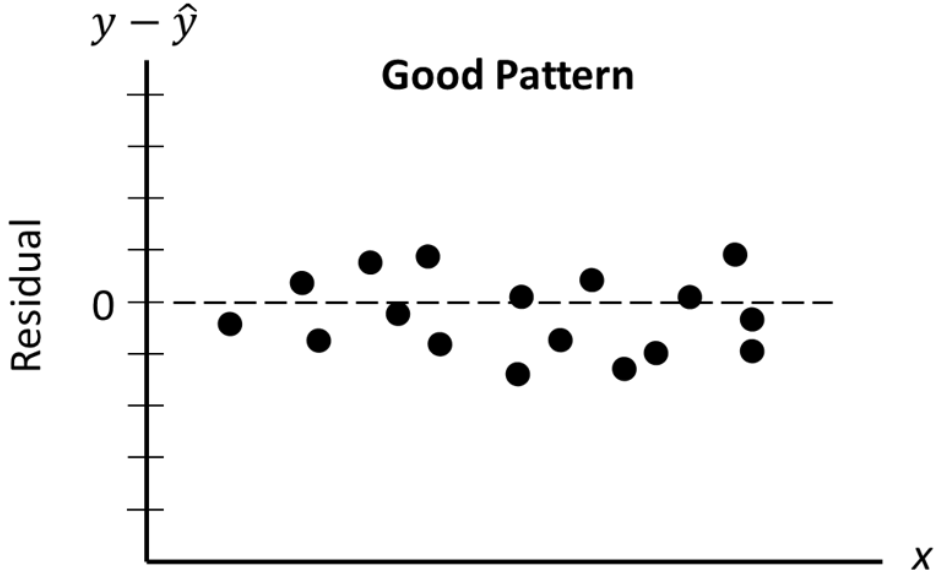
$$25 \pm 3.1824(2.6013)$$

$$25 \pm 8.28$$

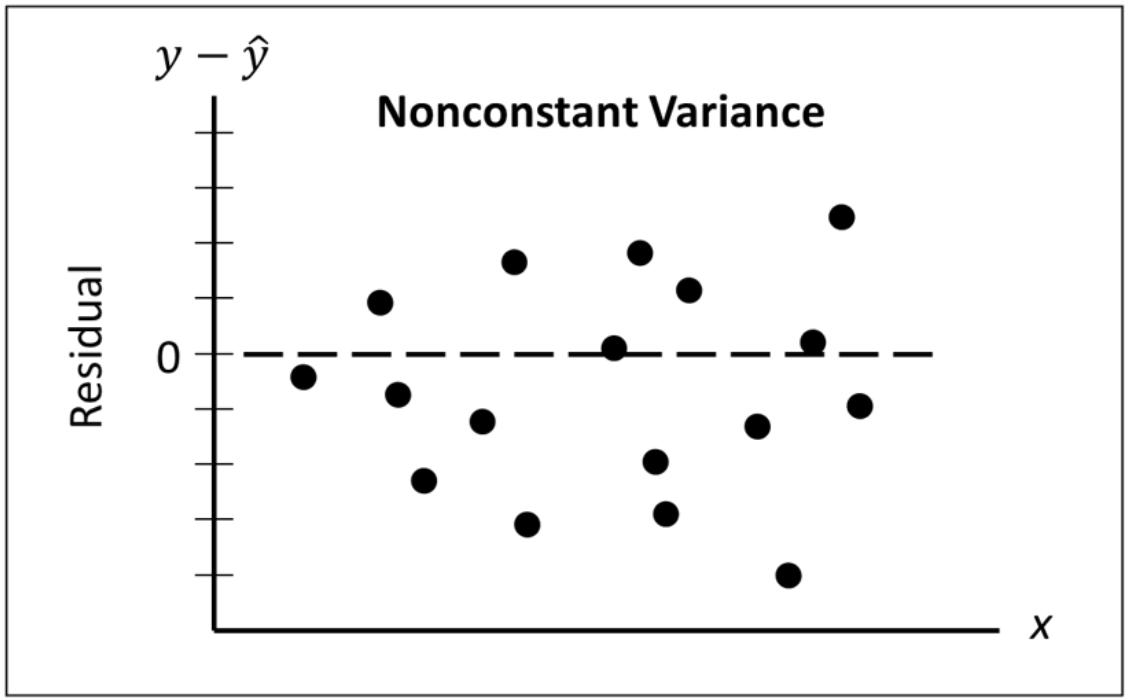
16.72 to 33.28 cars

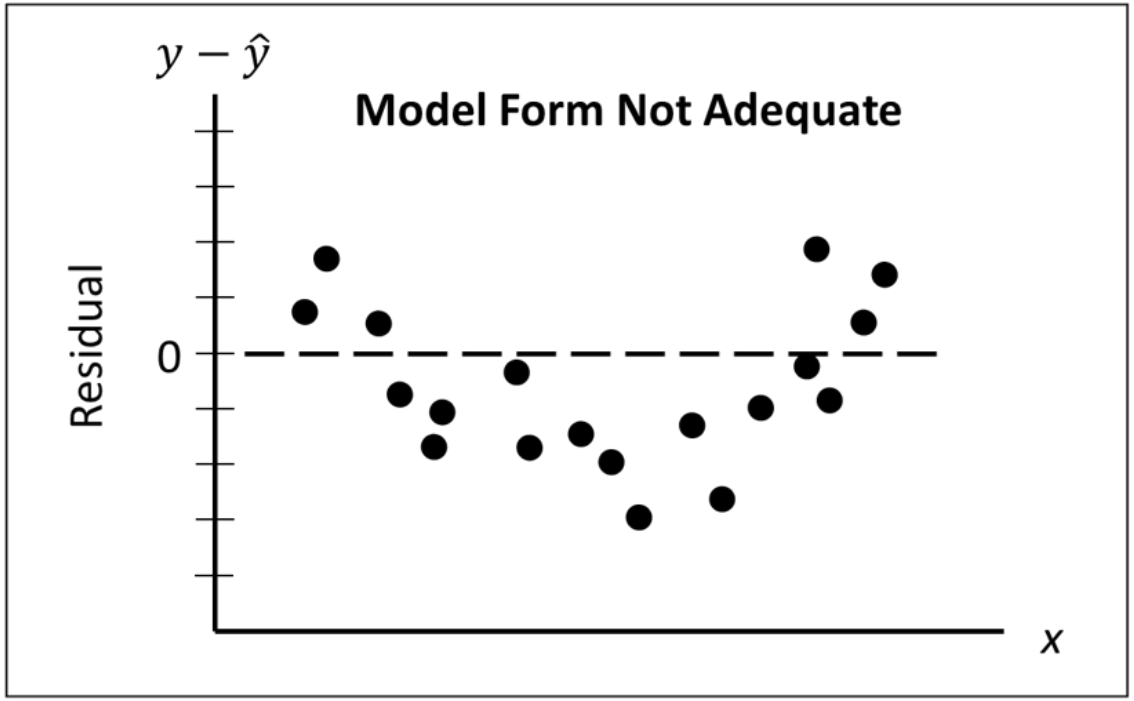
Residual Analysis

- If the assumptions about the error term ε appear questionable, the hypothesis tests about the significance of the regression relationship and the interval estimation results may not be valid.
- The residuals provide the best information about ε.
- Residual for observation i: $y_i - \hat{y}_i$
- Much of the residual analysis is based on an examination of graphical plots.
- If the assumption that the variance of ε is the same for all values of x is valid, and the assumed regression model is an adequate representation of the relationship between the variables, then the residual plot should give an overall impression of a horizontal band of points.



Residual Plot Against x

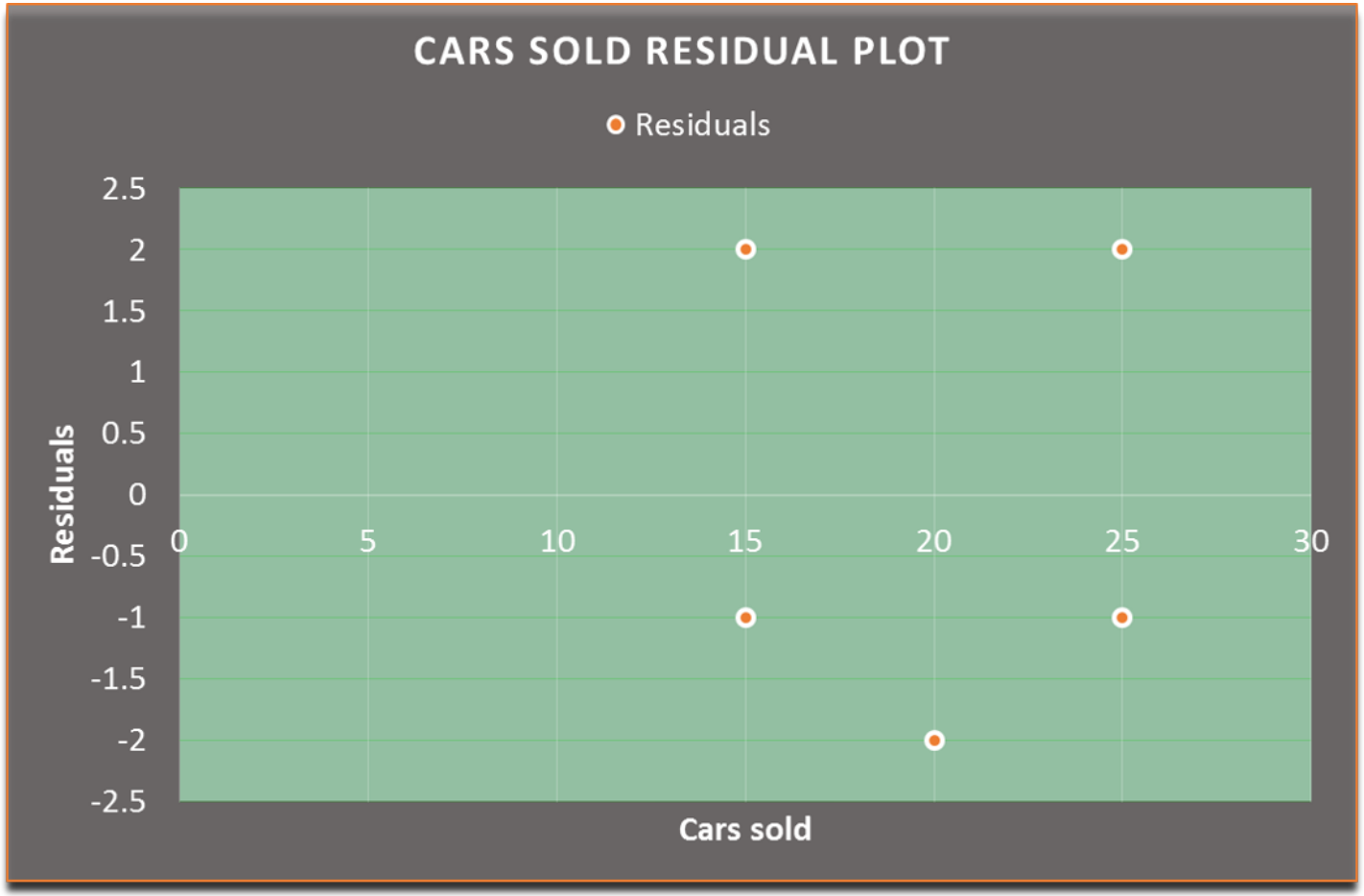




Residual Plot Against x

Residuals:

Observation	Predicted Cars Sold	Residuals
1	15	-1
2	25	-1
3	20	-2
4	15	2
5	25	2



Standardized Residuals & Standardized Residual Plot

Standardized Residual for Observation i:

$$\frac{y_i - \hat{y}_i}{s_{y_i - \hat{y}_i}}$$

$$s_{y_i - \hat{y}_i} = s\sqrt{1 - h_i}$$

Where:

$$h_i = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\Sigma(x_i - \bar{x})^2}$$

Standardized Residual Plot:

The standardized residual plot can provide insight about the assumption that the error term ε has a normal distribution.

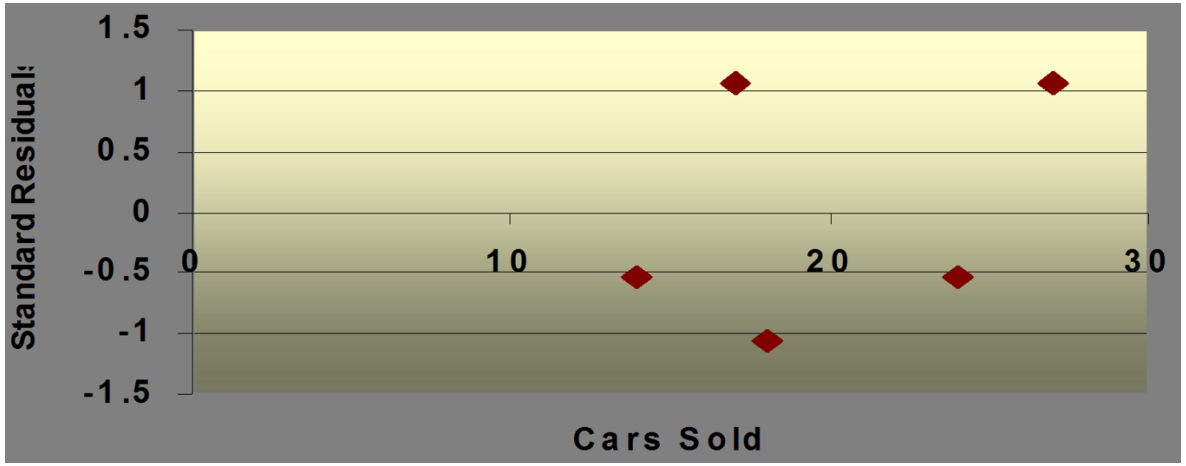
If this assumption is satisfied, the distribution of the standardized residuals should appear to come from a standard normal probability distribution.

Standardized Residual Plot

Standardized Residuals:

Observation	Predicted y	Residual	Standardized Residual
1	15	-1	-0.5345
2	25	-1	-0.5345
3	20	-2	-1.0690
4	15	2	1.0690
5	25	2	1.0690

Standardized Residual Plot: All of the standardized residuals are between –1.5 and +1.5 indicating that there is no reason to question the assumption that ε has a normal distribution.



Outliers and Influential Observations

Detecting Outliers:

- An outlier is an observation that is unusual in comparison with the other data.
- We classify an observation as an outlier if its standardized residual value is < -2 or $> +2$.
- This standardized residual rule sometimes fails to identify an unusually large observation as being an outlier.
- This rule's shortcoming can be circumvented by using studentized deleted residuals.
- The $|i^{th}$ studentized deleted residual$|$ will be larger than the $|i^{th}$ studentized residual$|$.

Multiple Regression

- In this chapter we continue our study of regression analysis by considering situations involving two or more independent variables.
- This subject area, called multiple regression analysis, enables us to consider more factors and thus obtain better estimates than are possible with simple linear regression.

The equation that describes how the dependent variable y is related to the independent variables $x_1, x_2, \ldots, x_p$ and an error term is:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \varepsilon$$

where $\beta_0, \beta_1, \beta_2, \ldots, \beta_p$ are the parameters, and ε is a random variable called the error term.

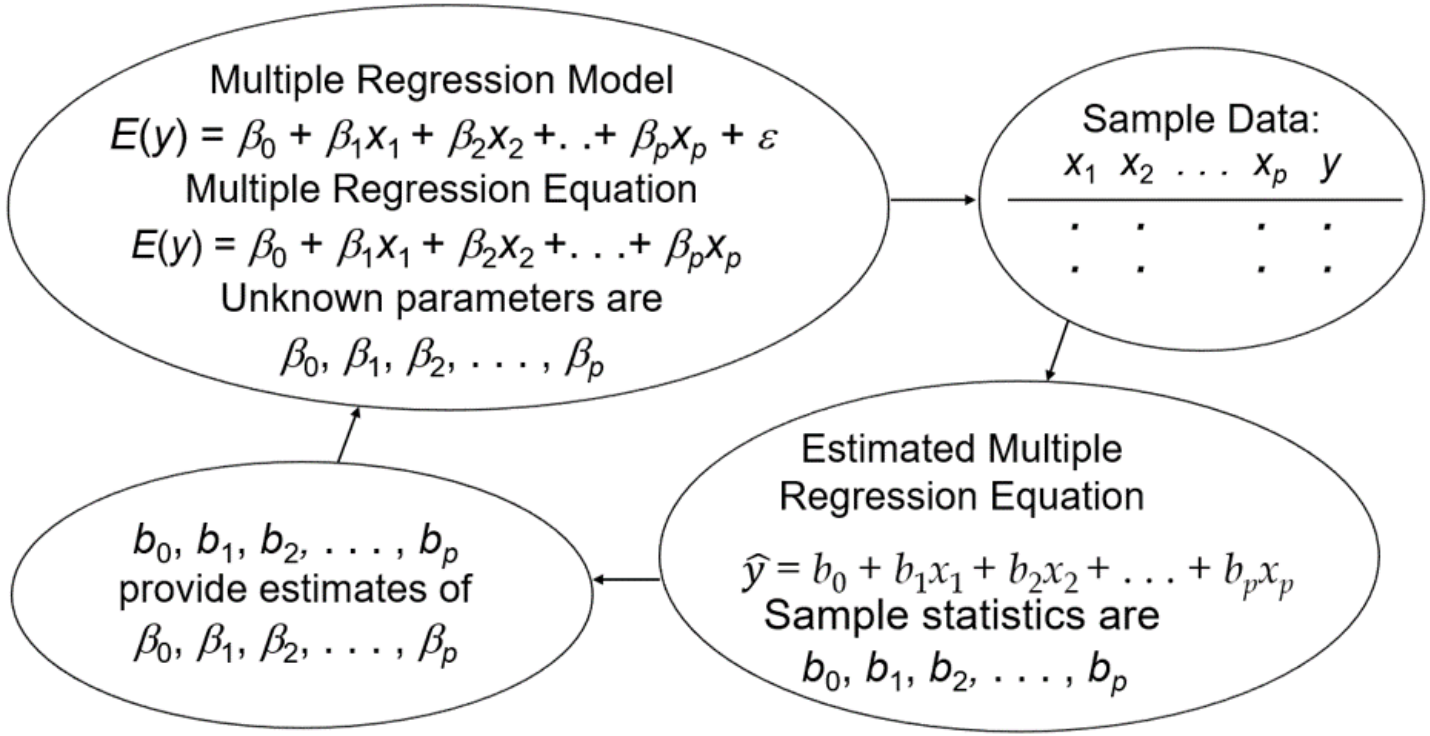
The equation that describes how the mean value of y is related to $x_1, x_2, \ldots, x_p$ is:

$$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$$

Estimated Multiple Regression Equation

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_p x_p$$

A simple random sample is used to compute the sample statistics b_0, b_1, b_2,…, b_p that are used as the point estimators of the parameters β_0, β_1, β_2,…, β_p.



Least Squares Method

- Least Squares Criterion: $min \sum (y_i - \hat{y}_i)^2$

- Computation of Coefficient Values:

The formulas for the regression coefficients b_0, b_1, b_2,…, b_p involve the use of matrix algebra. We will rely on computer software packages to perform the calculations.

The emphasis will be on how to interpret the computer output rather than on how to make the multiple regression computations.

Multiple Regression Model

Example: Programmer Salary Survey: A software firm collected data for a sample of 20 computer programmers. A suggestion was made that regression analysis could be used to determine if salary was related to the years of experience and the score on the firm's programmer aptitude test.

The years of experience, score on the aptitude test, and corresponding annual salary ($1000s) for a sample of 20 programmers is shown on the next slide.

Experience (Yrs.)	Test score	Salary ($1000s)
4	78	24.0
7	100	43.0
1	86	23.7
5	82	34.3
10	84	38.0
0	75	22.2
1	80	23.1
6	83	30.0
6	91	33.0

Multiple Regression Model

Suppose we believe that salary (y) is related to the years of experience (x_1) and the score on the programmer aptitude test (x_2) by the following regression model:

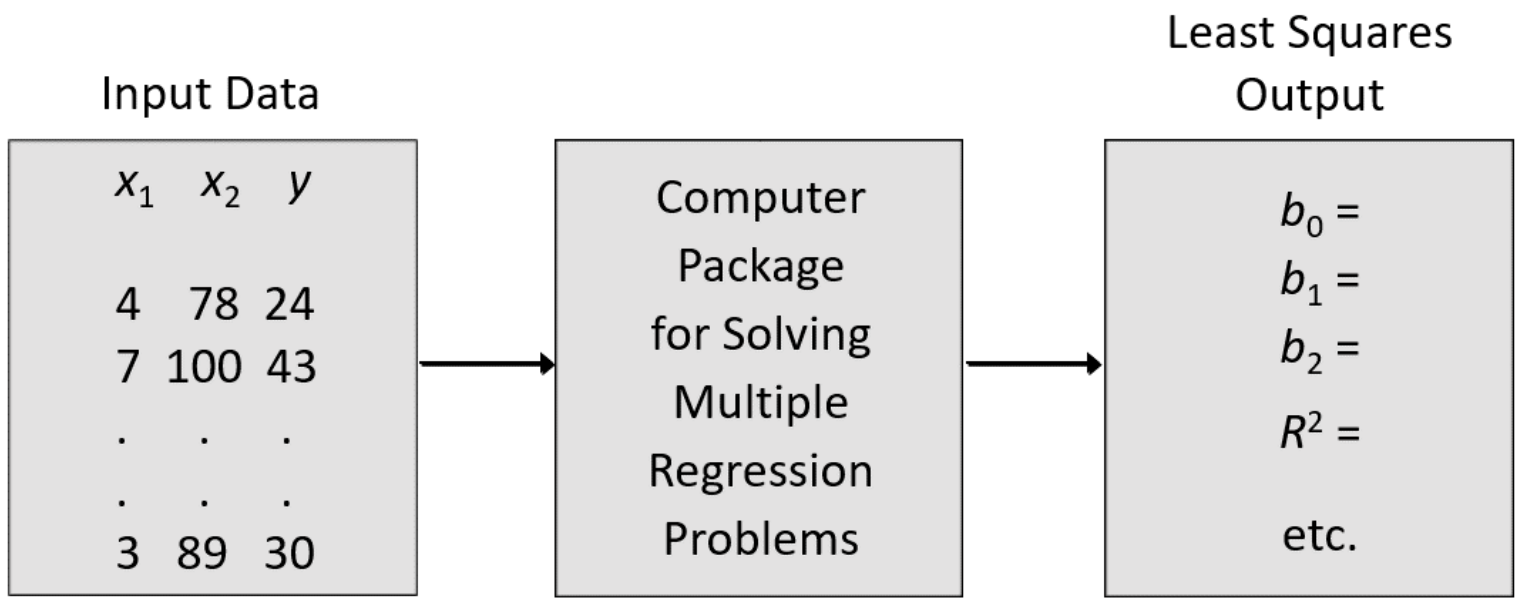
$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$$

where

y = annual salary ($1000s)

x_1 = years of experience

x_2 = score on programmer aptitude test



Solving for the Estimates of β_0, β_1, β_2

Regression Equation Output:

Predictor	Coef	SE Coef	T	p
Constant	3.17394	6.15607	0.5156	0.61279
Experience	1.4039	0.19857	7.0702	1.9E-06
Test Score	0.25089	0.07735	3.2433	0.00478

The estimated regression equation is:

SALARY = 3.174 + 1.404(EXPERIENCE) + 0.251(SCORE)

(Note: Predicted salary will be in thousands of dollars.)

Interpreting the Coefficients

In multiple regression analysis, we interpret each regression coefficient as follows:

b_1 represents an estimate of the change in y corresponding to one unit increase in x_1 when all other independent variables are held constant.

$b_1 = 1.404$ Salary is expected to increase by \$1,404 for each additional year of experience (when the variable *score on programmer attitude test* is held constant).

$b_2 = 0.251$ Salary is expected to increase by \$251 for each additional point scored on the programmer aptitude test (when the variable *years of experience* is held constant).

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

Multiple Coefficient of Determination

ANOVA Output:

Analysis of Variance

SOURCE	DF	SS	MS	F	P
Regression	2	500.3285	250.164	42.76	0.000
Residual Error	17	99.45697	5.850		
Total	19	599.7855			

$$R^2 = \frac{SSR}{SST} = \frac{500.3285}{599.7855} = 0.83418$$

Adjusted Multiple Coefficient of Determination

- Adding independent variables, even ones that are not statistically significant, causes the prediction errors to become smaller, thus reducing the sum of squares due to error, SSE.

- Because SSR = SST – SSE, when SSE becomes smaller, SSR becomes larger, causing r^2 = SSR/SST to increase.

- The adjusted multiple coefficient of determination compensates for the number of independent variables in the model.

$$R_\alpha{}^2 = 1 - (1 - R^2)\left(\frac{n-1}{n-p-1}\right)$$

$$R_\alpha{}^2 = 1 - (1 - 0.834179)\left(\frac{20-1}{20-2-1}\right) = 0.814671$$

Assumptions About the Error Term ε

- The error ε is a random variable with mean of zero.
- The variance of ε, denoted by σ^2, is the same for all values of the independent variables.
- The values of ε are independent.
- The error ε is a normally distributed random variable reflecting the deviation between the y value and the expected value of y given by $\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$.

Testing for Significance

- In simple linear regression, the F and t test provide the same conclusion.
- In multiple regression, the F and t test have different purposes.
- The F test is referred to as the test for overall significance.

- If the F test shows an overall significance, the t test is used to determine whether each of the individual independent variables is significant.
- A separate t test is conducted for each of the independent variables in the model.
- We refer to each of these t tests as a test for individual significance.

Testing for Significance: F Test

- Hypotheses: H_0: $\beta_1 = \beta_2 = \cdots = \beta_p = 0$

  H_a: One or more of the parameters is not equal to zero.

- Test Statistic: $F = \frac{MSR}{MSE}$

- Rejection Rule: Reject H_0 if the p-value $\leq \alpha$ or if $F \geq F_\alpha$

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

F Test for Overall Significance

- Hypotheses: H_0: $\beta_1 = \beta_2 = 0$

  H_a: One or more of the parameters is not equal to zero.

- Rejection Rule: For $\alpha = 0.05$ and $df = 2, 17$; $F_{0.05} = 3.59$

  Reject H_0 if the p-value ≤ 0.05 or $F \geq 3.59$

- ANOVA Output:

Analysis of Variance

SOURCE	DF	SS	MS	F	P
Regression	2	500.3285	250.164	42.76	0.000
Residual Error	17	99.45697	5.850		
Total	19	599.7855			

p-value used to test for overall significance

F Test for Overall Significance

- Test Statistics: $F = \frac{MSR}{MSE} = \frac{250.16}{5.85} = 42.76$

- Conclusion: p-value ≤ 0.05 so we can reject H_0.

  Also, $F = 42.76 \geq 3.59$

Testing for Significance: t Test

- Hypotheses: $H_0: \beta_i = 0$
  $H_a: \beta_i \neq 0$

- Test Statistic: $t = \frac{b_i}{s_{b_i}}$

- Rejection Rule: Reject H_0 if the p-value $\leq \alpha$ or if $t \leq -t_{\alpha/2}$ or if $t \geq t_{\alpha/2}$
  where $t_{\alpha/2}$ is based on a t distribution with $n - p - 1$ degrees of freedom.

- Hypotheses: $H_0: b_i = 0$
  $H_a: b_i \neq 0$

- Rejection Rule: For $\alpha = 0.05$ and $df = 17$; $t_{0.025} = 2.11$
  Reject H_0 if the p-value ≤ 0.05 or $t \leq -2.11$ or if $t \geq 2.11$

t Test for Significance of Individual Parameters

- Hypotheses: $H_0: b_i = 0$

  $H_a: b_i \neq 0$

- Rejection Rule: For $\alpha = 0.05$ and $df = 17$; $t_{0.025} = 2.11$

  Reject H_0 if the p-value ≤ 0.05 or $t \leq -2.11$ or if $t \geq 2.11$

Regression Output:

Predictor	Coef	SE Coef	T	p
Constant	3.17394	6.15607	0.5156	0.61279
Experience	1.4039	0.19857	7.0702	1.9E-06
Test Score	0.25089	0.07735	3.2433	0.00478

t Test for Significance of Individual Parameters

Regression Equation Output:

Predictor	Coef	SE Coef	T	p
Constant	3.17394	6.15607	0.5156	0.61279
Experience	1.4039	0.19857	7.0702	1.9E-06
Test Score	0.25089	0.07735	3.2433	0.00478

t statistic and *p*-value used to test for the individual significance of "Test Score"

t Test for Significance of Individual Parameters

- Test Statistics:

$$t = \frac{b_1}{s_{b_1}} = \frac{1.4039}{0.1986} = 7.07$$

$$t = \frac{b_2}{s_{b_2}} = \frac{0.25089}{0.07735} = 3.24$$

- Conclusions

Reject both H_0: $\beta_1 = 0$ and H_0: $\beta_2 = 0$.

Both independent variables are significant.

Testing for Significance: Multicollinearity

- The term multicollinearity refers to the correlation among the independent variables.
- When the independent variables are highly correlated (say, $|r| > 0.7$), it is not possible to determine the separate effect of any particular independent variable on the dependent variable.
- If the estimated regression equation is to be used only for predictive purposes, multicollinearity is usually not a serious problem.
- Every attempt should be made to avoid including independent variables that are highly correlated.

Estimated Regression Equation for Estimation & Prediction

- The procedures for estimating the mean value of y and predicting an individual value of y in multiple regression are similar to those in simple regression.

- We substitute the given values of $x_1, x_2, \ldots, x_p$ into the estimated regression equation and use the corresponding value of $\hat{y}$ as the point estimate.

- The formulas required to develop interval estimates for the mean value of $\hat{y}$ and for an individual value of y are beyond the scope of the textbook.
- Software packages for multiple regression will often provide these interval estimates.

Residual Analysis

- For simple linear regression the residual plot against $\hat{y}$ and the residual plot against x provide the same information.

- In multiple regression analysis it is preferable to use the residual plot against $\hat{y}$ to determine if the model assumptions are satisfied.

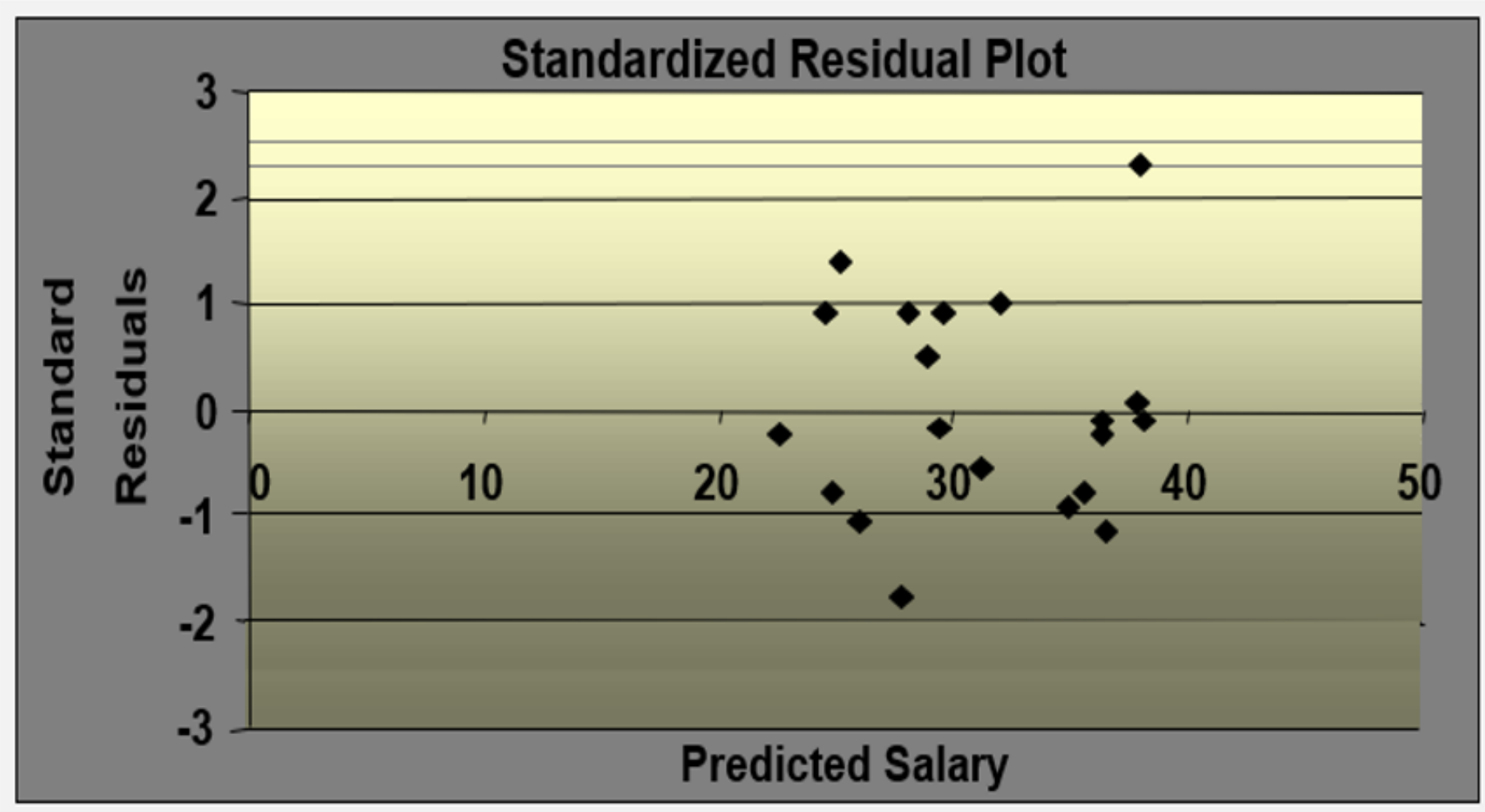
Standardized Residual Plot Against $\hat{y}$

- Standardized residuals are frequently used in residual plots for purposes of:
  - Identifying outliers (typically, standardized residuals < –2 or > +2)
  - Providing insight about the assumption that the error term ε has a normal distribution
- The computation of the standardized residuals in multiple regression analysis is too complex to be done by hand.

Residual Output:

Observation	Predicted Y	Residuals	Standard Residuals
1	27.89626	–3.89626	–1.771707
2	37.95204	5.047957	2.295406
3	26.02901	–2.32901	–1.059048
4	32.11201	2.187986	0.992921
5	36.34251	0.53251	–0.246689

Standardized Residual Plot Against $\hat{y}$



Categorical Independent Variables

- In many situations we must work with categorical independent variables such as gender (male, female, other), method of payment (cash, check, credit card), etc.
- For example, x_2 might represent gender where $x_2 = 0$ indicates male, $x_2 = 1$ indicates female, and $x_2 = 2$ indicates other.
- In this case, x_2 is called a dummy or indicator variable.

Example: Programmer Salary Survey

As an extension of the problem involving the computer programmer salary survey, suppose that management also believes that the annual salary is related to whether the individual has a graduate degree in computer science or information systems.

The years of experience, the score on the programmer aptitude test, whether the individual has a relevant graduate degree, and the annual salary ($1000) for each of the sampled 20 programmers are shown on the next slide.

Categorical Independent Variables

Experience (Yrs.)	Test Score	Degree	Salary ($1000)
4	78	No	24.0
7	100	Yes	43.0
1	86	No	23.7
5	82	Yes	34.3
8	86	Yes	35.8
10	84	Yes	38.0
0	75	No	22.2
1	80	No	23.1
6	83	No	30.0
6	91	Yes	33.0

Categorical Independent Variables

Regression Equation: $\hat{y} = b_0 + b_1 x_1 + b_2 x_2 + b_3 x_3$
where:

$\hat{y}$ = annual salary ($1000)

x_1 = years of experience

x_2 = score on programmer aptitude test

x_3 = 0 if individual does not have a graduate degree, 1 if they do (x_3 is a dummy variable)

Analysis of Variance

SOURCE	DF	SS	MS	F	P
Regression	3	507.8960	269.299	29.48	0.000
Residual Error	16	91.8895	5.743		
Total	19	599.7855			

$R^2 = \frac{507.896}{599.7855} = \mathbf{0.8468}$ Previously, $R^2 = 0.8342$

$R_\alpha{}^2 = 1 - (1 - 0.8468)\frac{20-1}{20-3-1} = \mathbf{0.8181}$ Previously, the Adjusted $R^2 = 0.815$

Categorical Independent Variables

Regression Equation Output:

Predictor	Coef	SE Coef	T	p
Constant	7.945	7.382	1.076	0.298
Experience	1.148	0.298	3.856	0.001
Test Score	0.197	0.090	2.191	0.044
Grad. Degr.	2.280	1.987	1.148	0.268

Not significant

More Complex Categorical Variables

- If a categorical variable has k levels, $k - 1$ dummy variables are required, with each dummy variable being coded as 0 or 1.
- For example, a variable with levels A, B, and C could be represented by x_1 and x_2 values of (0, 0) for A, (1, 0) for B, and (0, 1) for C.
- Care must be taken in defining and interpreting the dummy variables.
- For example, a variable indicating level of education could be represented by x_1 and x_2 values as follows:

High Degree	x_1	x_2
Bachelor's	0	0
Master's	1	0
Ph.D.	0	1

Modeling Curvilinear Relationships

Example: Sales of Laboratory Scales

A manufacturer of laboratory scales wants to investigate the relationship between the length of employment of their salespeople and the number of scales sold.

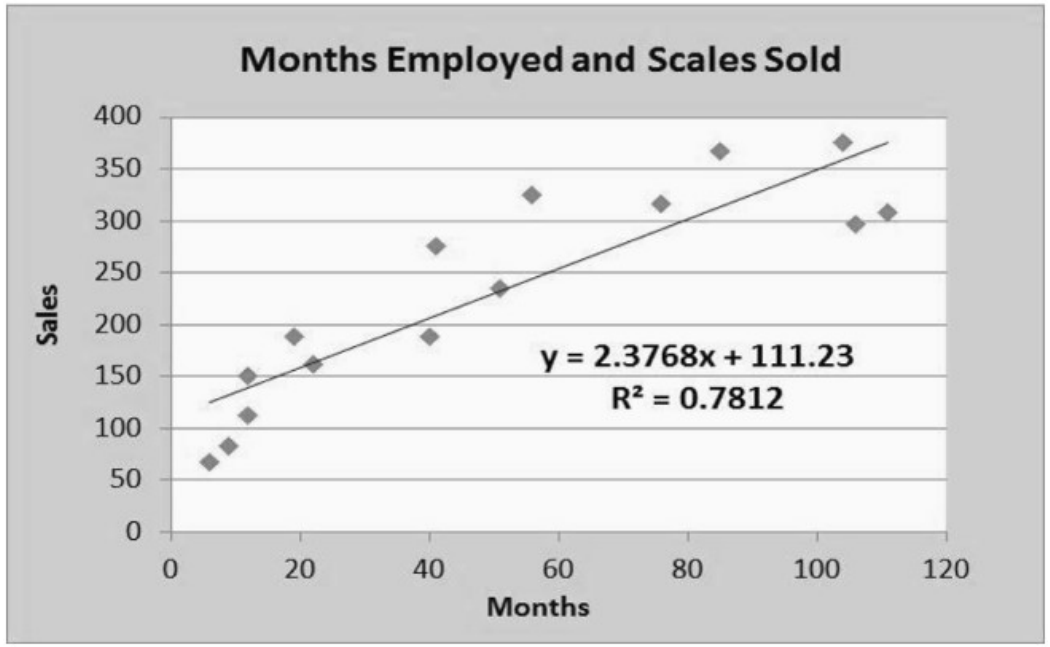
The table gives the number of months each salesperson has been employed by the firm (x) and the number of scales sold (y) by 15 randomly selected salespersons.

Months	Sales
41	275
106	296
76	317
104	376
22	162
12	150
85	367
111	308

Months	Sales
40	189
51	235
9	83
12	112
6	67
56	325
56	325

Modeling Curvilinear Relationships

- A scatter diagram can be developed and a straight line could be fit to bivariate data.
- The estimated regression equation and the coefficient of determination for simple linear regression can also be developed.

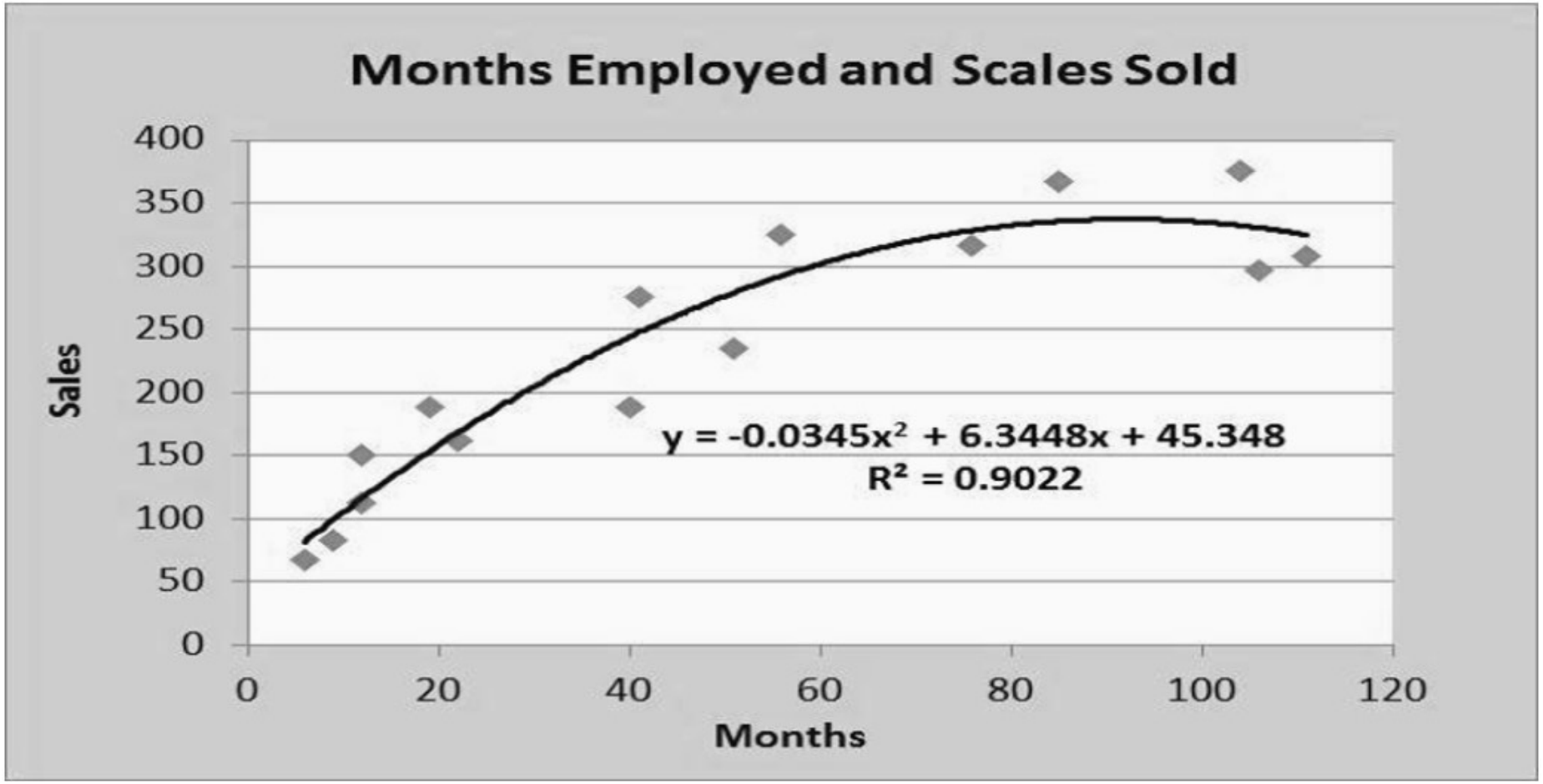


The scatter diagram indicates a possible curvilinear relationship between the length of time employed and the number of scales sold. So, we develop a multiple regression model with two independent variables: x and x^2.

$y = b_0 + b_1x + b_2x^2 + \varepsilon$ This model is often referred to as a second-order polynomial or a quadratic model.

Modeling Curvilinear Relationships

Quadratic Model:



Modeling Curvilinear Relationships

- To build the regression model we will treat the values of x^2 as a second independent variable (called MonthSq).

Months	MonthsSq	Sales
41	1681	275
106	11236	296
76	5776	317
104	10816	376
22	484	162
12	144	150
85	7225	367
111	12321	308

Months	MonthsSq	Sales
40	1600	189
51	2601	235
9	81	83
12	144	112
6	36	67
56	3136	325
56	361	325

Modeling Curvilinear Relationships

Regression Result:

Regression Statistics	
Multiple R	0.9498469
R Square	0.9022091
Adjusted R Square	0.8859107
Standard Error	34.4527668
Observations	15

We should be pleased with the fit provided by the estimated multiple regression equation.

ANOVA

	df	SS	MS	F	Signif. F
Regression	2	131413.016	65706.51	55.355	8.7456E-07
Residual	12	14243.918	1186.99		
Total	14	145656.933			

The overall model is significant (p-value for the F test is 8.75E-07)

Modeling Curvilinear Relationships

Regression Equation Output:

	Coefficients	Std. Error	t Stat	P-value
Intercept	45.3475789	22.7746541	1.99114	0.06973
Month	6.3448071	1.05785144	5.99782	6.24E-05
MonthSq	-0.0344856	0.00894828	-3.85388	0.00229

We can conclude that adding MonthsSq to the model is significant.