Statistical Analysis

Johns Hopkins
Carey Business School

Simple Regression Analysis

Simple Regression Analysis

1. Regression Analysis Introduction
2. Scatter Plots
3. Simple Regression
4. Ordinary Least Squares (OLS) Method
5. Measures of Goodness of Fit
6. Coefficient of Determination
7. Standard Error of the Estimate
8. Excel-Data Analysis Add-On
9. Model Assumptions and the Standard Error

Regression Analysis Introduction

Regression Analysis is the quantitative expression of the relationship between the independent variables and a dependent variable.

Examples

- Hours studying and test grade
- Education level and income
- Interest rates and houses sold
- Cigarette smoking and lung cancer

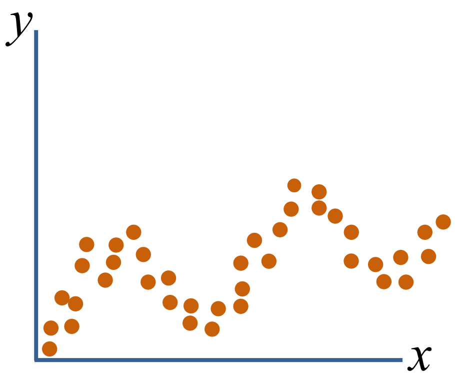
Regression Analysis Introduction *continued*

- A good model can predict your dependent variable
- Our dependent variable y is a function of the independent variables $x_1, x_2, x_3, x_4,\ldots$

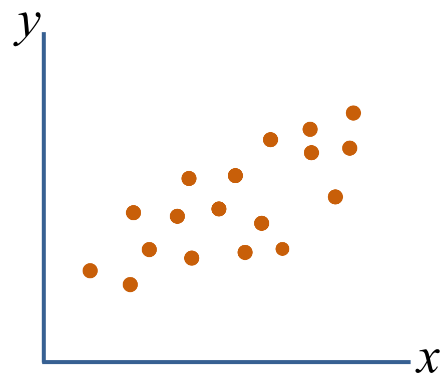
$$Y = f(x_1, x_2, x_3, x_4,\ldots)$$

- Criteria
  - Scatter plot indicates a linear relationship or trend
  - Historical data or data collection is available for building the model

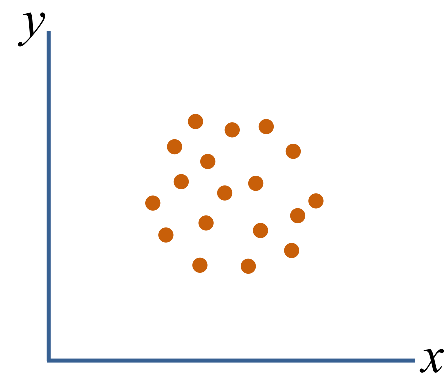
Scatter Plots



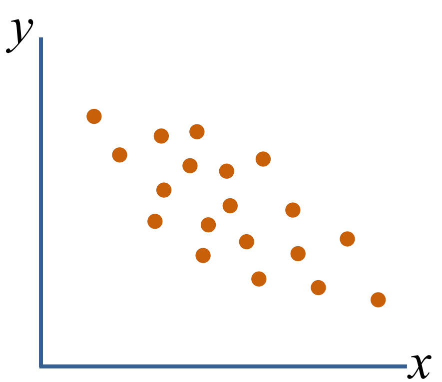
(a) cyclic



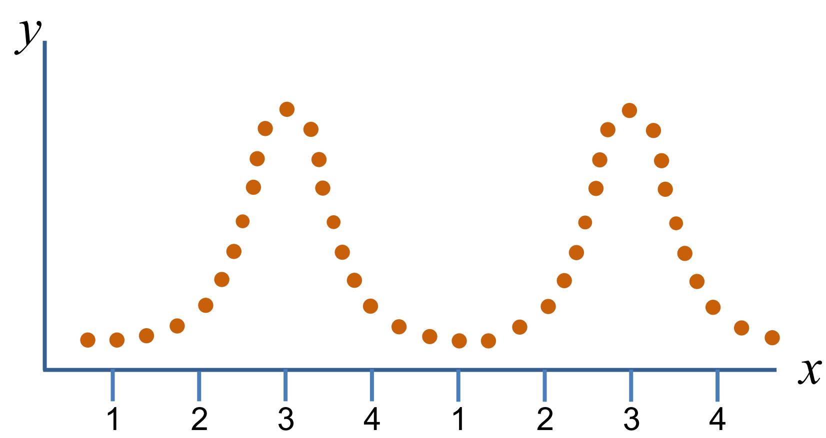
(b) Positive trend



(c) Random



(d) Negative trend



(e) yearly quarters
seasonal affect

Simple Regression

- **Regression analysis** is a statistical technique that uses observed data to relate the dependent variable to one or more independent variables.
- The **dependent (or response) variable** is the variable we wish to understand or predict.
- The **independent (or predictor) variable** is the variable we will use to understand or predict the dependent variable
    - The objective is to build a regression model that can describe, predict and control the dependent variable based on the independent variable.
    - The method we will discuss uses least squares point estimates.

Adapted from Bowerman, B. L., O'Connell, R. T., & Murphree, E. S., (2010). Business Statistics in Practice (6th Ed.), Copyright © McGraw-Hill Education

Regression Model Formulation

$$\hat{y} = \beta_0 + \beta_1 x + \varepsilon$$

$\hat{y}$ is the predicted dependent variable

β_0 represents the y intercept

β_1 represents the slope and the quantitative relationship between x_1 and y

Slope is defined as $\dfrac{rise}{run} = \dfrac{vertical}{horizontal}$

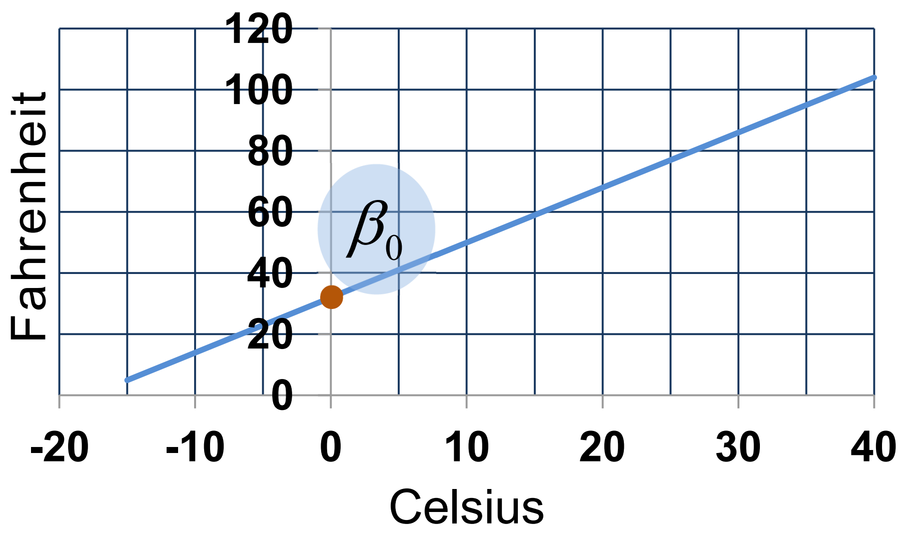
ε is an error term that describes the effect on y of all factors other than x

Simple Regression *example*

How to convert temperature from Celsius (x) to Fahrenheit (y)?

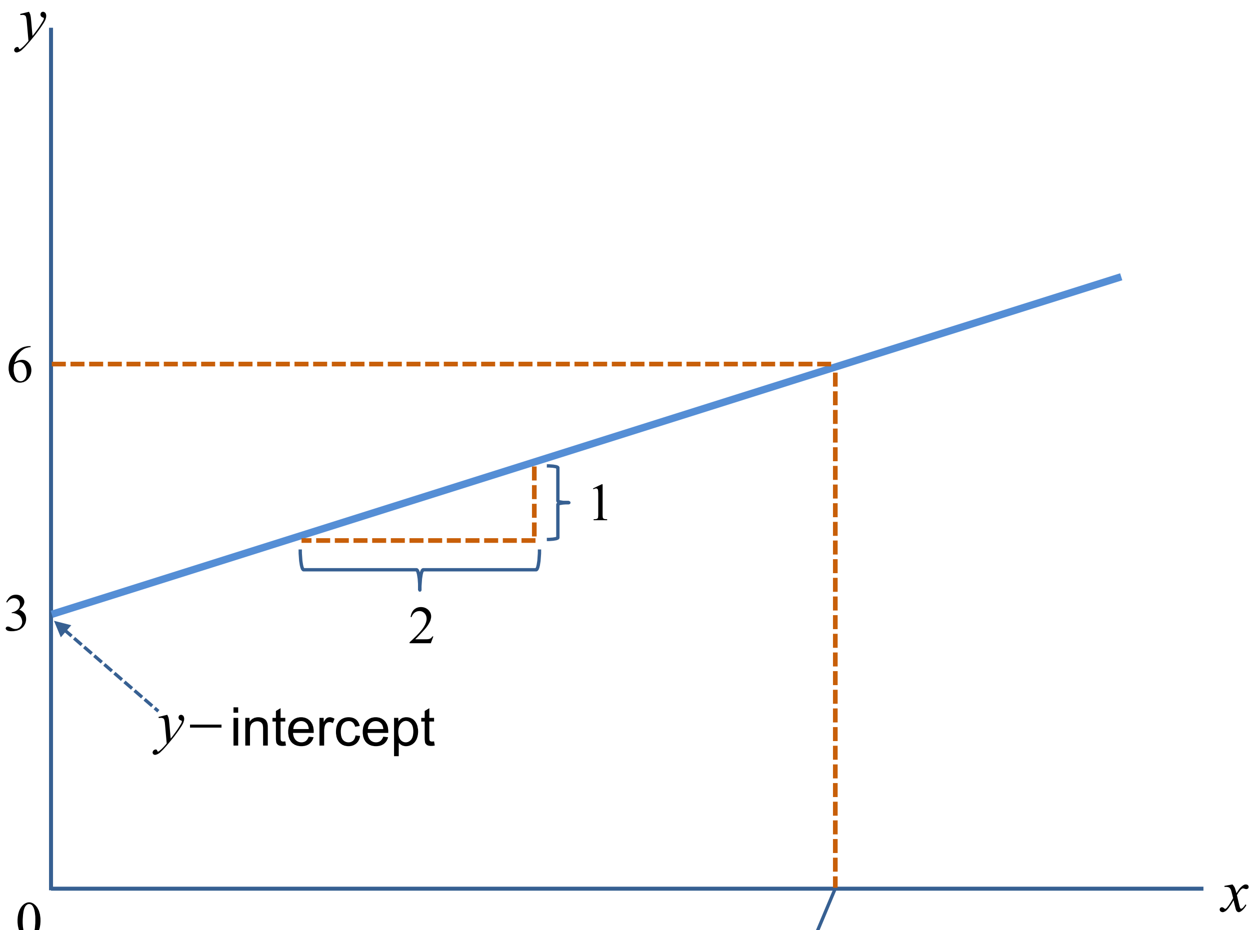
$y = 32 + (9/5)\,x$ or $y = \beta_0 + \beta_1 x$
where β_0 is called the intercept
and β_1 is the slope.

- y : dependent variable
- x : independent variable



- In general when we are predicting y using different values of x and believe the relationship to be linear, it may not be perfectly linear and y may be slightly above or below the line.
- This will be represented by $y=\beta_0+\beta_1 x+\varepsilon$

Simple Regression Model Illustrated



$x_0 =$ A specific value of the independent variable x

Straight line defined by the equation

$$\mu_{y|x} = \beta_0 + \beta_1 x$$

$$\hat{y} = 3 + \frac{1}{2}x$$

where $x = 6$

$$\hat{y} = 3 + \frac{1}{2}6$$

$$\hat{y} = 6$$

Ordinary Least Squares (OLS) Method

$$SSxy = \sum x_i y_i - \frac{\left(\sum x_i\right)\left(\sum y_i\right)}{n}$$

$$SSxx = \sum x_i^2 - \frac{\left(\sum x_i\right)^2}{n}$$

$$SSyy = \sum y_i^2 - \frac{\left(\sum y_i\right)^2}{n}$$

$$b_1 = \frac{SSxy}{SSxx}$$

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

Flat Screen TV *example*

Observation (months)	Advertising (in $1,000's) (X)	Flat Screen TV (in 1,000's) (Y)
1	10	15
2	12	17
3	8	13
4	17	23
5	10	16
6	15	21
7	10	14
8	14	20
9	19	24
10	10	17
11	11	16
12	13	18
13	16	23
14	10	15
15	12	16
	187	**268**

Ordinary Least Squares (OLS) Method

$$SSxy = \sum x_i y_i - \frac{\left(\sum x_i\right)\left(\sum y_i\right)}{n}$$

$$=3,490 - \frac{(187)(268)}{15} = 148.9333$$

$$SSxx = \sum x_i^2 - \frac{\left(\sum x_i\right)^2}{n}$$

$$= 2,469 - \frac{(187)^2}{15} = 137.7333$$

$$SSyy = \sum y_i^2 - \frac{\left(\sum y_i\right)^2}{n}$$

$$= 4,960 - \frac{(268)^2}{15} = 171.7333$$

$$b_1 = \frac{SSxy}{SSxx} = \frac{148.9333}{137.7333} = 1.08$$

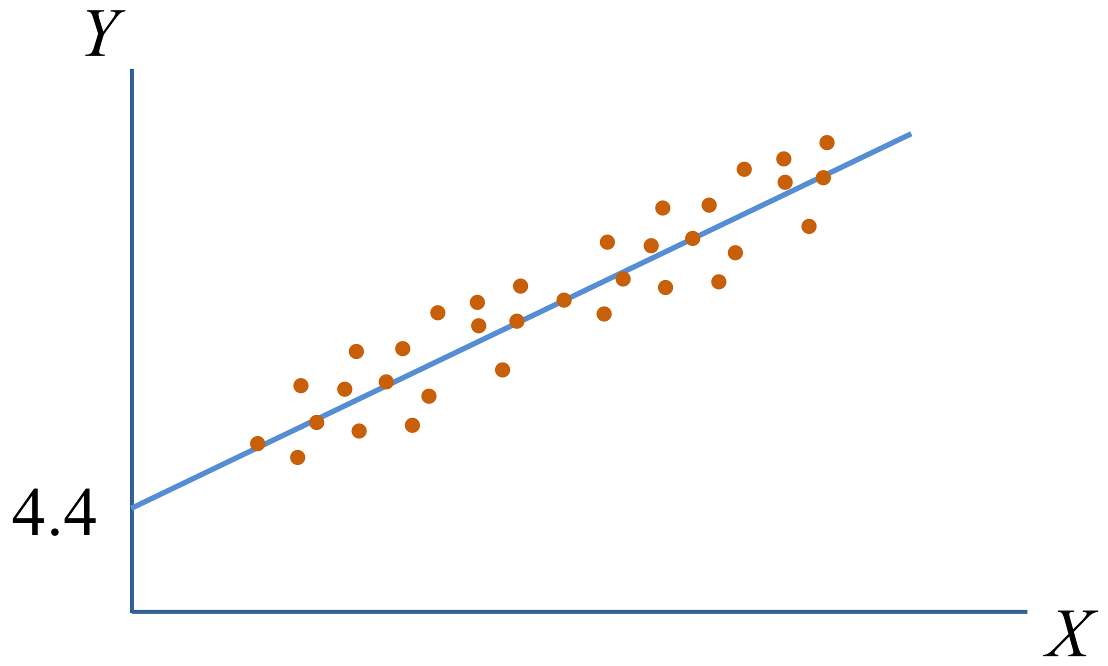
$$b_0 = \bar{y} - b_1 \bar{x}$$
$$= 17.8666 - (1.08)(12.4666)$$
$$= 4.4$$

$$\hat{Y} = 4.4 + 1.08x$$

Relationship of Advertising Dollars to Sales

$$\hat{Y} = 4.4 + 1.08x$$

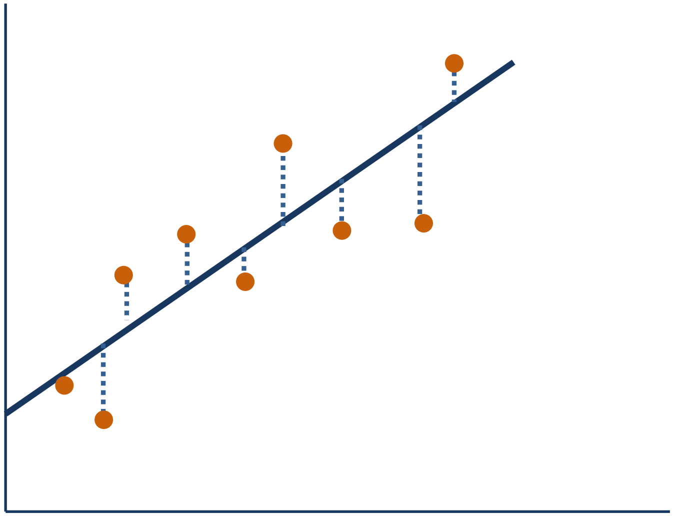
OLS method fits the line to the data points so as to minimize the distance from each point to the line



Minimizing Error

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

We want to find the line such that $\sum \varepsilon^2$ is minimum.

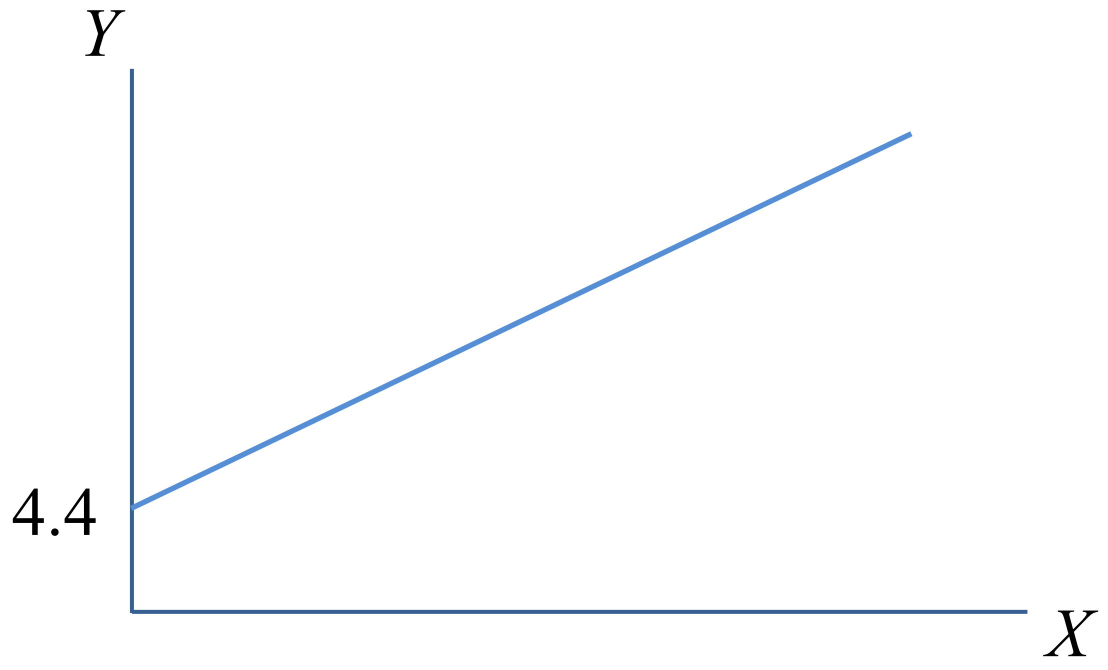


Suppose the line is $\hat{y} = b_0 + b_1 x$ or $\bar{y} = b_0 + b_1 \bar{x}$

Relationship of Advertising Dollars to Sales

$$\hat{Y} = 4.4 + 1.08x$$

For each one unit change in x, y increases by 1.08 units. In other words, if an additional thousand dollars is spent in advertising, 1,080 more TVs are sold.



Flat Screen TV *example*

Observation (months)	Advertising (in $1,000's) (X)	Flat Screen TV (in $1,000's) (Y)
1	**10**	**15**
2	12	17
3	8	13
4	17	23
5	**10**	**16**
6	15	21
7	**10**	**14**
8	14	20
9	19	24
10	**10**	**17**
11	11	16
12	13	18
13	16	23
14	**10**	**15**
15	12	16
	187	**268**

Relationship of Advertising Dollars to Sales

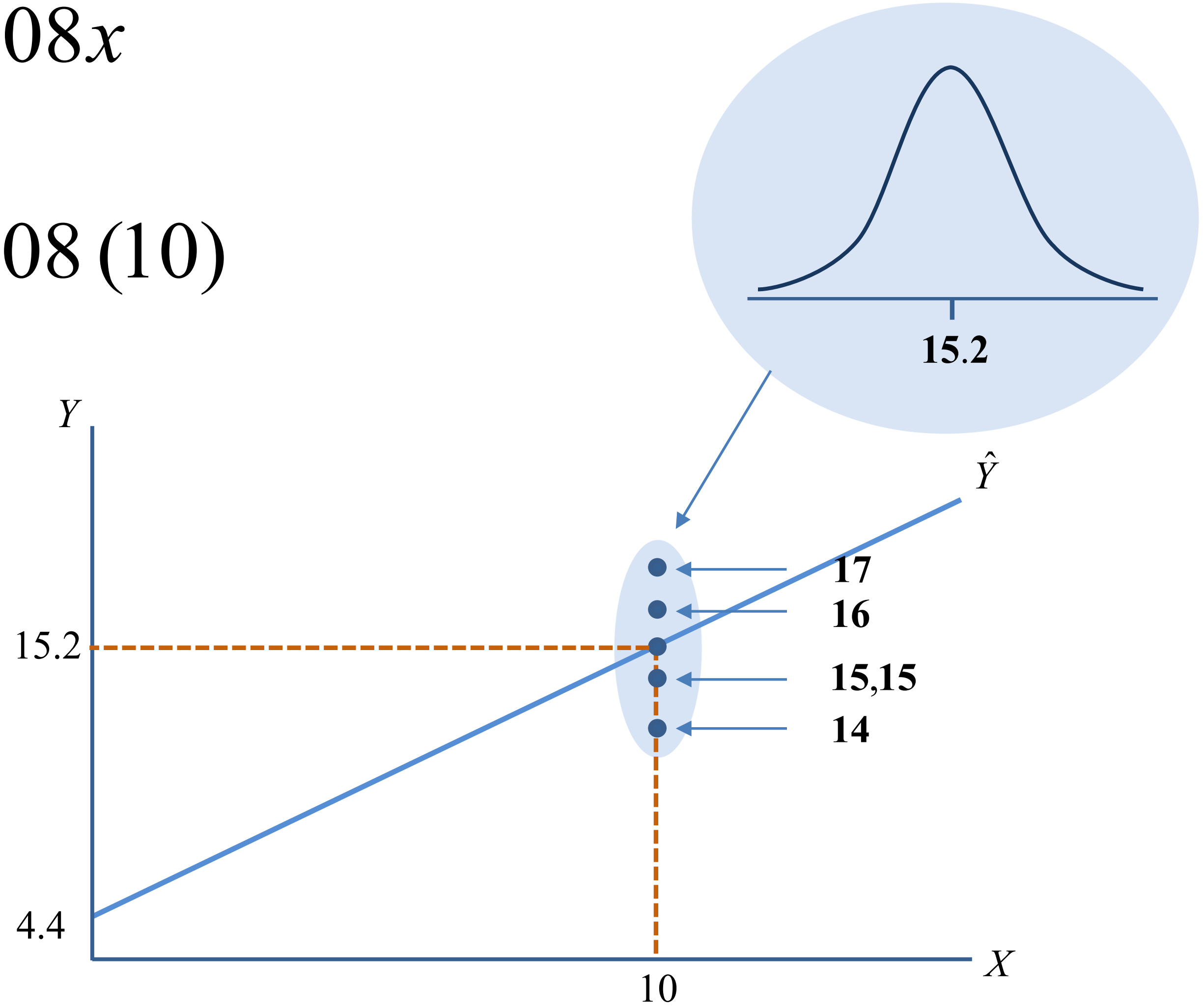
$$\hat{Y} = 4.4 + 1.08x$$

$$x = 10$$

$$\hat{Y} = 4.4 + 1.08\,(10)$$

$$\hat{Y} = 15.2$$

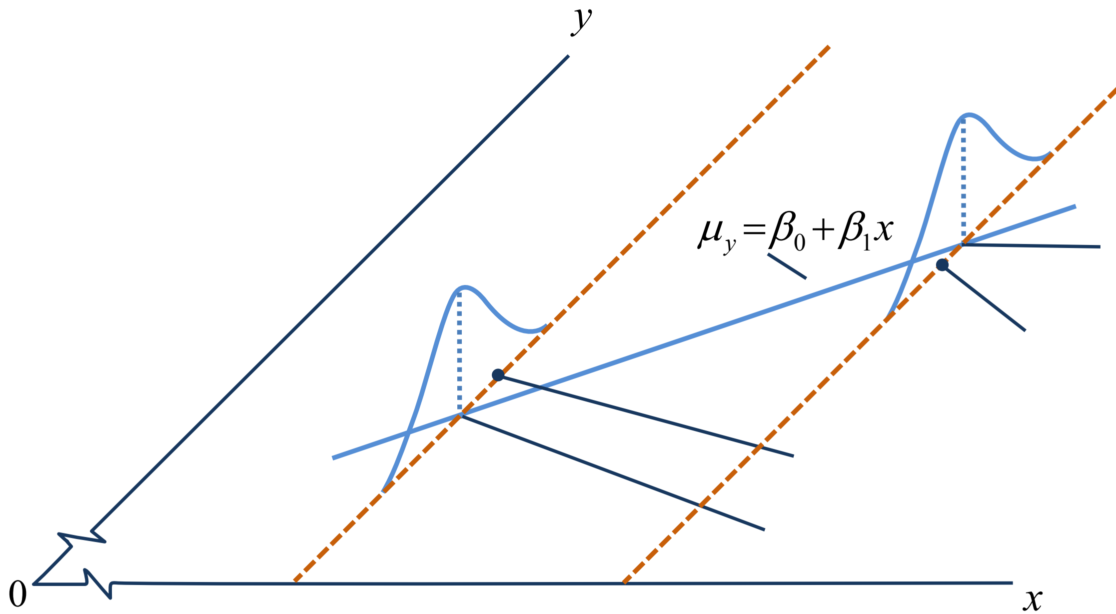
where $x = 10$ thousand



Individual y values and the Regression Line

Individual y values are normally distributed around the regression line.

Shortly we will learn to calculate the measure of goodness of fit, the standard error of the estimate which is the standard deviation in relationship to the regression line.

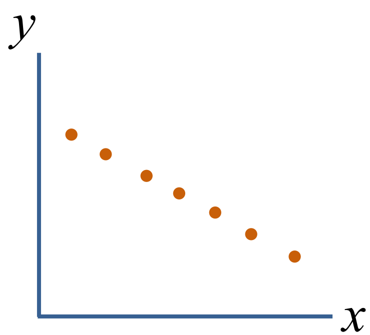


Adapted from Bowerman, B. L., O'Connell, R. T., & Murphree, E. S., (2010). Business Statistics in Practice (6th Ed.), Copyright © McGraw-Hill Education

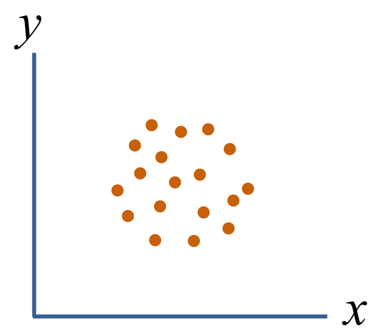
Measures of Goodness of Fit

- Coefficient of Correlation, r

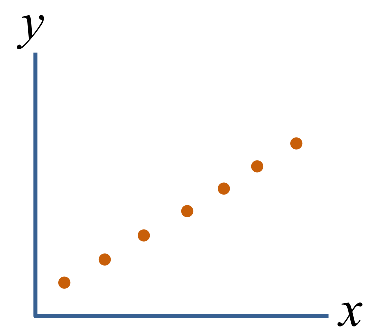
$-1\ldots$ $\ldots 0 \ldots$ $\ldots +1$



(a) r = -1:
perfect negative correlation



(b) r near 0 :
little linear relationship



(c) r = 1:
perfect positive correlation

- Coefficient of Determination, r^2
  r^2 ranges from 0 to +1
- Standard Error of the Estimate, Se
  The standard deviation for the regression line

Correlation Coefficient r

- The simple correlation coefficient measures the strength of the linear relationship between y and x and is denoted by r

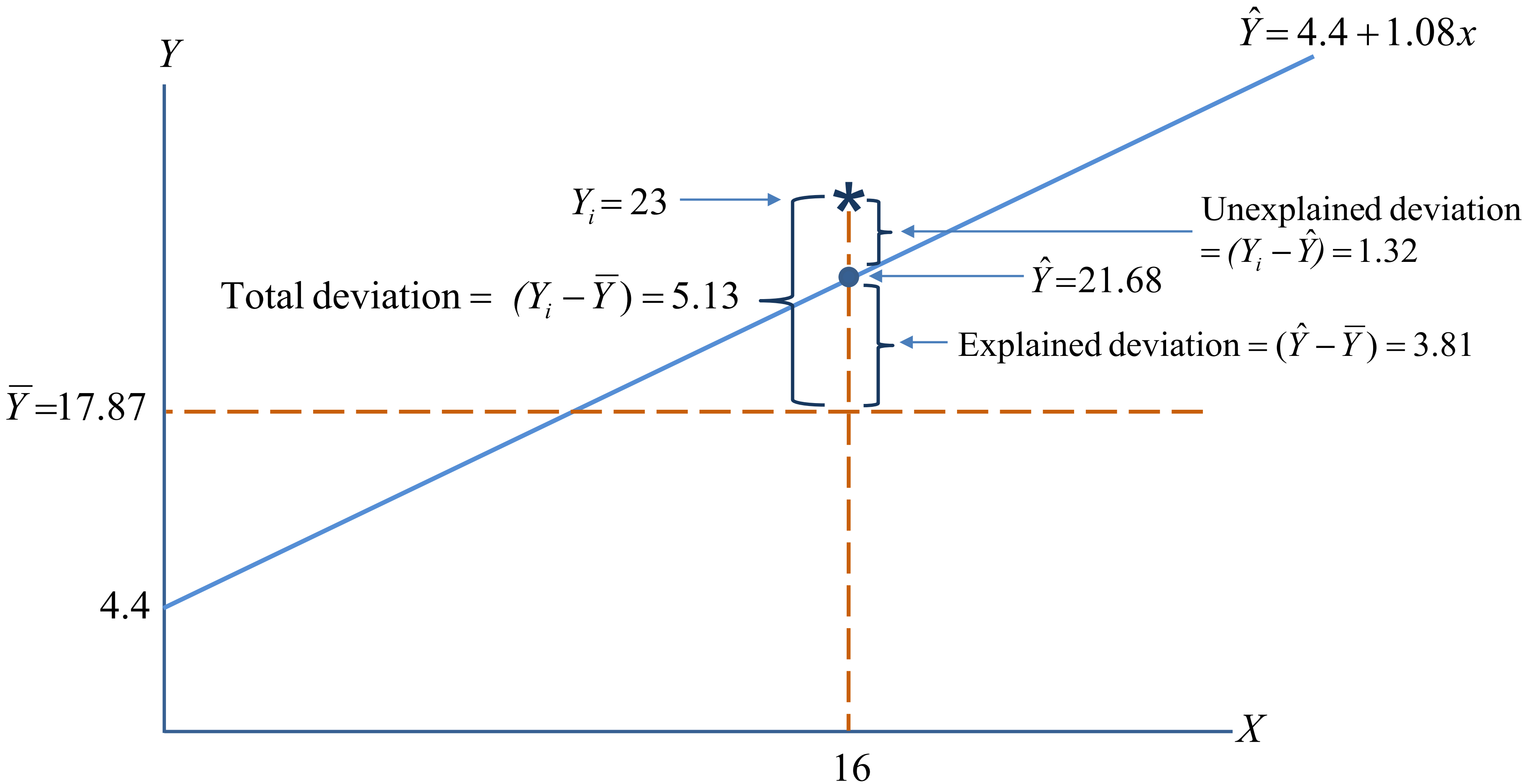
$$r = +\sqrt{r^2} \text{ if } b_1 \text{ is positive, and}$$
$$r = -\sqrt{r^2} \text{ if } b_1 \text{ is negative}$$

- Where b_1 is the slope of the least squares line

Coefficient of Determination

- Quantitative measure of the goodness of fit
- Ranges from zero to 1
- Explained deviation divided by the total deviation

Deviations for Flat Screen TV



Adapted from Webster, Allen L., (1992). Applied Statistics for Business and Economics (2nd Ed.), Copyright © Irwin

Deriving r^2

Step 1: Express in words

Total deviation = Explained deviation + Unexplained deviation

Step 2: Change words to mathematical symbols

$$(Y_i - \overline{Y}) = (\hat{Y} - \overline{Y}) + (Y_i - \hat{Y})$$

Step 3: Square each expression and sum

$$\sum (Y_i - \overline{Y})^2 = \sum (\hat{Y} - \overline{Y})^2 + \sum (Y_i - \hat{Y})^2$$

Step 4: Replace with notation

$$SSTo = SSR + SSE$$

Step 5: Explanation of r^2

$$r^2 = \frac{SSR}{SSTo} = \frac{(SSxy)^2}{(SSxx)(SSyy)} = \frac{148.9333^2}{(137.7333)(171.7333)} = 0.94$$

Standard Error of the Estimate (Se)

Standard deviation for the regressor line and follows the same empirical rules

$$Se = \sqrt{\frac{\sum (Y_i - \hat{Y})^2}{n-2}}$$

Shortcut method

$$SSE = SSyy - \frac{(SSxy)^2}{SSxx}$$

$$MSE = \frac{SSE}{n-2}$$

$$Se = \sqrt{MSE}$$

Standard Error of the Estimate (Se)

Shortcut method

$$SSE = SSyy - \frac{(SSxy)^2}{SSxx} \qquad SSE = 171.7333 - \frac{(148.9333)^2}{137.7333} = 10.6893$$

$$MSE = \frac{SSE}{n-2} \qquad MSE = \frac{10.6893}{15-2} = 0.82226$$

$$Se = \sqrt{MSE} \qquad Se = \sqrt{0.82226} = 0.907$$

Interpretation

Within one standard error of the regressor line lies 68.4% of the data points.

For this example $3(0.907) = 2.721$

Therefore 99.7% of the data points lie within 2.721 units around the regressor.

Relationship of Population to Restaurant Revenue *example*

Population to Revenue for China Noodle House Restaurant

x : population size in thousand

y : revenue in thousand

$n = 10$

Week	y	x	x²	xy
1	527.1	20.8	432.64	10963.68
2	548.7	27.5	756.25	15089.25
3	767.2	32.3	1043.29	24780.56
4	722.9	37.2	1383.84	26891.88
5	826.3	39.6	1568.16	32721.48
6	810.5	45.1	2034.01	36553.55
7	1040.7	49.9	2490.01	51930.93
8	1033.6	55.4	3069.16	57261.44
9	1090.3	61.7	3806.89	67271.51
10	1235.8	64.6	4173.16	79832.68
Total	**8603.1**	**434.1**	**20757.41**	**403296.96**

$$SSxy = \sum (x_i y_i) - \frac{\sum (x_i) \sum (y_i)}{n}$$

$$= 29836.389$$

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

$$= 1913.129$$

$$b_1 = \frac{SSxy}{SSxx} = 15.596$$

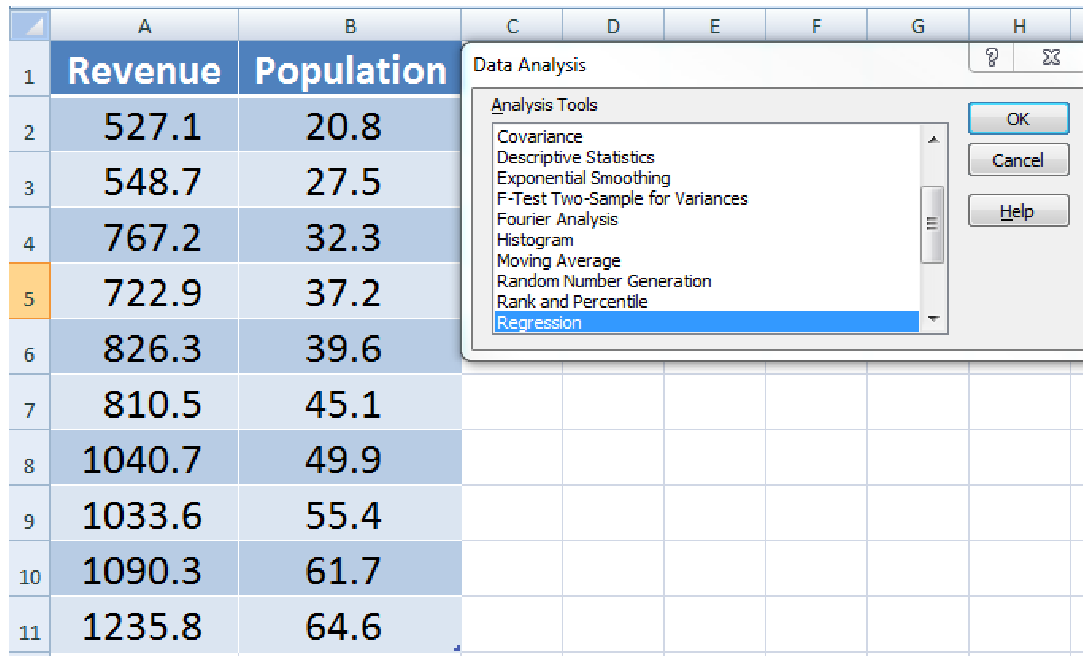
$$b_0 = \bar{y} - b_1 \bar{x}$$

$$= 183.31$$

Excel- Data Analysis Add-On

Enter data in Excel

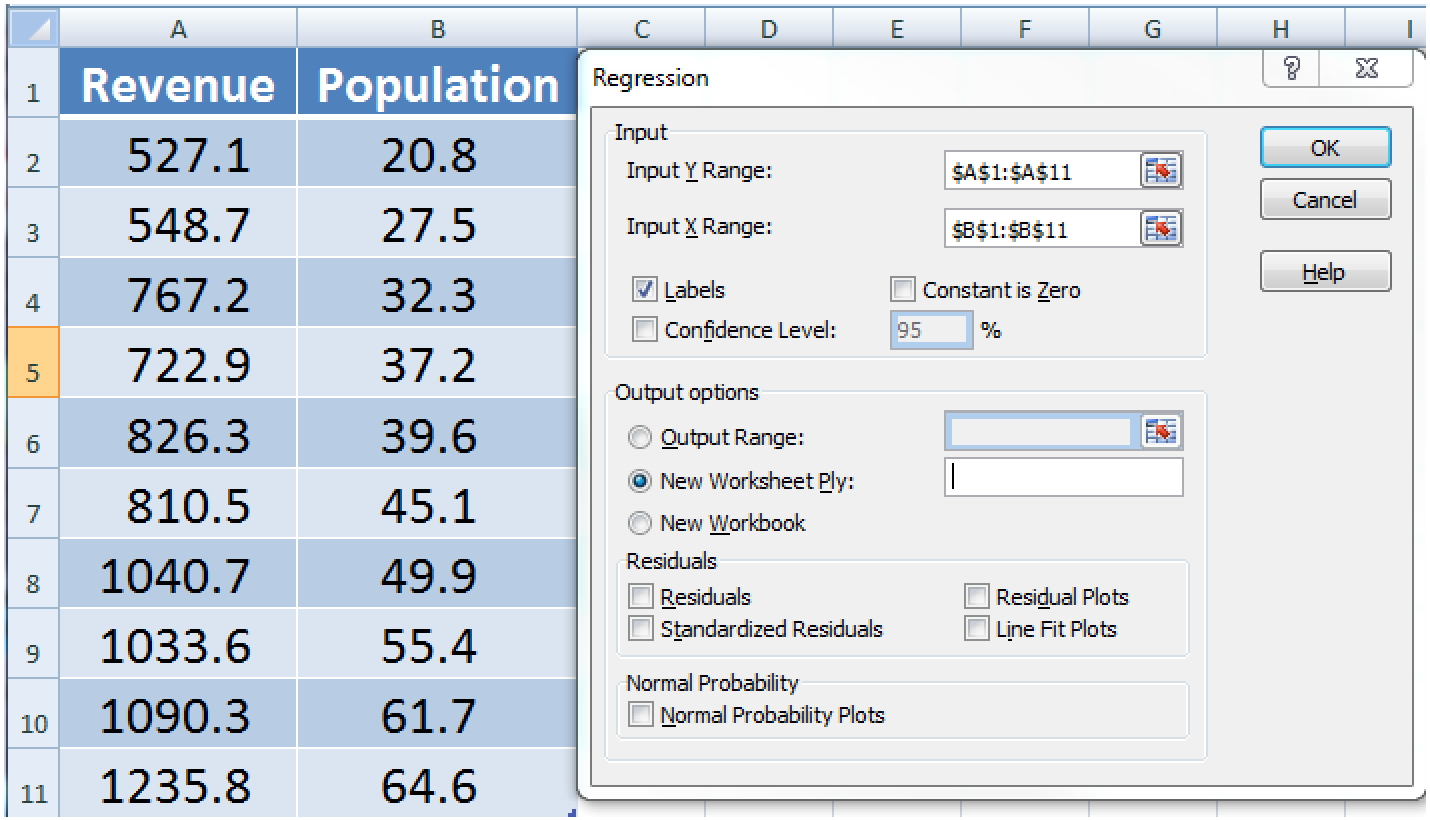
1) Revenue – dependent variable (y)
2) Population – independent variable (x)
3) Open data tab
4) Click on Data Analysis
5) Select Regression



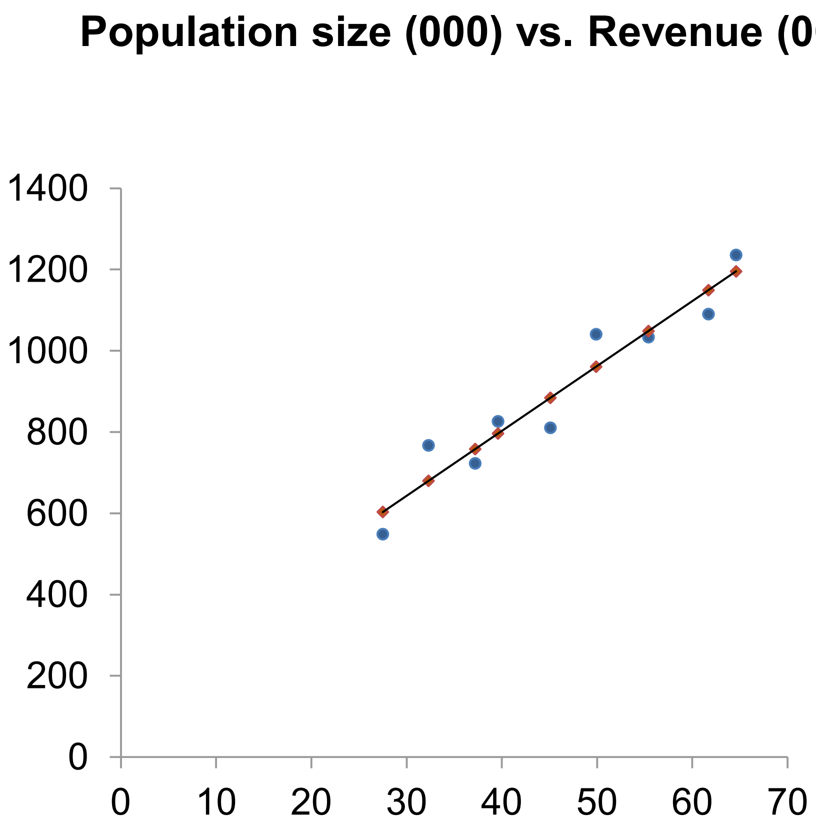
Revenue	Population
527.1	20.8
548.7	27.5
767.2	32.3
722.9	37.2
826.3	39.6
810.5	45.1
1040.7	49.9
1033.6	55.4
1090.3	61.7
1235.8	64.6

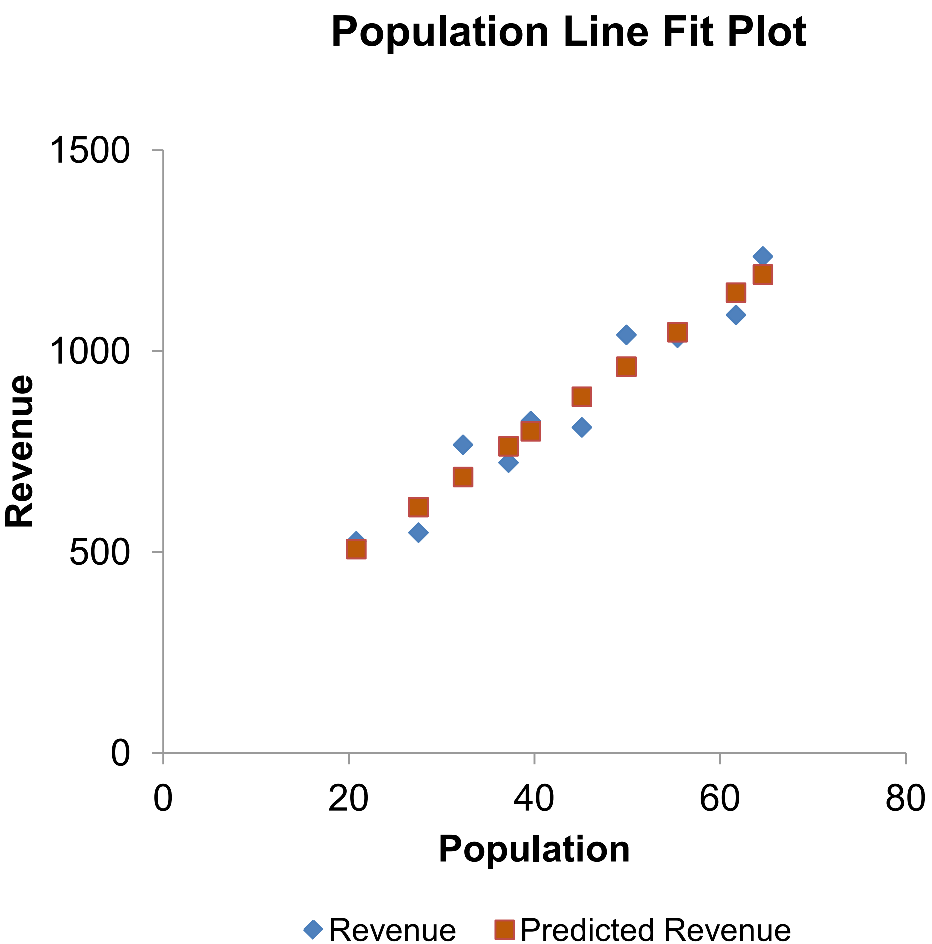
Example

A new work sheet is produced for the output:



Example





$$y = 15.5956x + 183.305$$

$$r^2 = 0.9386$$

Excel Output for China Noodle House

Regression Statistics	
Multiple R	0.96879338
R^2	0.93856061
Adjusted R^2	0.93088068
Std. Error	61.7051543
Observations	**10**

Residual Output		
Obs.	*Predicted Y*	*Residuals*
1	507.693551	19.40645
2	612.184051	-63.4841
3	687.042917	80.15708
4	763.461342	-40.5613
5	800.890775	25.40922
6	886.666559	-76.1666
7	961.525425	79.17458
8	1047.30121	-13.7012
9	1145.55347	-55.2535
10	**1190.7807**	**45.0193**

ANOVA					
	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	465316.3	465316.3	122.2096	3.99584E-06
Residual	8	30460.21	3807.526		
Total	**9**	**495776.5**			

Model Assumptions and the Standard Error

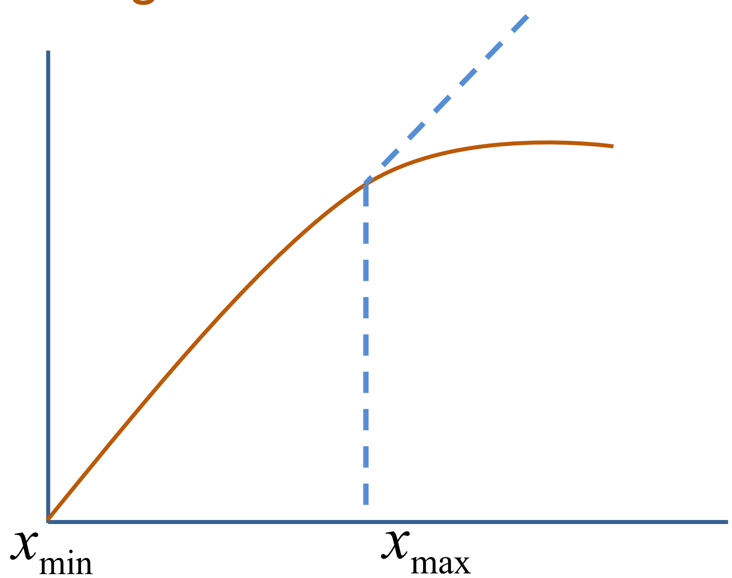
The term "ε" is called *the standard error* and the regression model is based on several assumptions regarding the standard error.

1. **Mean of Zero**
   At any given value of x, the population of potential error term values has a mean equal to zero.
2. **Constant Variance Assumption**
   At any given value of x, the population of potential error term values has a variance that does not depend on the value of x.
3. **Normality Assumption**
   At any given value of x, the population of potential error term values has a normal distribution.
4. **Independence Assumption**
   Any one value of the error term ε is statistically independent of any other value of ε.

Bowerman, B. L., O'Connell, R. T., & Murphree, E. S., (2010). Business Statistics in Practice (6th Ed.), Copyright © McGraw-Hill Education

General Comments

- The estimated value of the dependent variable (y) is $\hat{y}$.
- $\hat{y} = b_0 + b_1 x$ is sometimes called the *least square prediction equation.* The predicted value represents the *point estimate.*
- When the regression line is based on the values of x between $x_{\min}$ and $x_{\max}$, one may not want to extrapolate beyond this *experimental region.*



General Comments *continued*

- Every time we collect a sample, we will get slightly different values of the slope and the intercept resulting in a different value of y for a given x. Therefore we may want to get an interval estimation on y.

- r^2 is a statistic that will give some information about the goodness of fit of a model. In regression, the r^2 *coefficient of determination* is a statistical measure of how well the regression line approximates the real data points. An r^2 of 1.0 indicates that the regression line perfectly fits the data.

Multiple Regression Analysis

Multiple Regression Analysis

1. Multiple Regression Introduction
2. Building a Multiple Regression Model
3. Testing the Significance of the Slope and y-intercept
4. Confidence and Prediction Intervals
5. Testing the Significance of the Population Correlation Coefficient

Multiple Regression Introduction

- Multiple regression uses two or more independent variables to describe the dependent variable
  a) This allows multiple regression models to handle more complex situations
  b) There is no limit to the number of independent variables a model can use
- Multiple regression has only one dependent variable.
- Independent variables are called explanatory variables.
- The linear regression model relating y to $x_1, x_2, \ldots x_k$ is
  $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots \beta_k x_k + \varepsilon$
- $\beta_0, \beta_1, \beta_2 \ldots \beta_k$ are unknown regression parameters
- ε is an error term that describes the effects on y of all factors other than the independent variables $x_1, x_2, \ldots x_k$

Building a Multiple Regression Model

Predicting number of houses sold- dependent variable y, dependent on:
First independent variable being interest rates.

Independent variable	x_i	r^2	r
$\hat{y}=b_0+b_1x_1$ Interest rate	x_1	.80	- 0.89
$\hat{y}=b_0+b_1x_1+b_2x_2$ Employment rate	x_2	.84	.918
$\hat{y}=b_0+b_1x_1+b_2x_2+b_3x_3$ GDP	x_3	.86	.93
$\hat{y}=b_0+b_1x_1+b_2x_2+b_3x_3+b_4x_4$ Population growth rate	x_4	.92	.96
$\hat{y}=b_0+b_1x_1+b_2x_2+b_3x_3+b_4x_4+b_5x_5$ Average salary	x_5	.95	.975

Building a Multiple Regression Model

Variable	B-coeff	Std err	Beta	t value	p value
Rain (x_1)	0.2273	0.1588	0.2508	1.4309	0.1802
Fert (x_2)	1.1524	0.2772	0.7714	4.1571	0.0016
Acid (x_3)	-0.1113	0.1093	-0.0935	-1.0181	0.3304
BO Intercept	3.2946				
Critical t:	2.2010				
C.O.D. (R-Squared):					
Adjusted C.O.D. (R-Squared):					0.9283
Multiple Correlation Coefficient:					0.9087 0.9635
Standard Error Estimate:					6.4987

$$\hat{y} = 3.3 + 0.23x_1 + 1.15x_2 - 0.11x_3$$

Use of Dummy Variables

- So far, we assumed quantitative variables in a regression model. However, we may wish to include descriptive qualitative data as well.
- We can model the effects of different levels of a *qualitative variable* by using what are called *dummy variables* (also known as *indicator variables*).
- When the qualitative variable has two categories (such as defective /non defective, male / female), the corresponding dummy variable will be either 0 or 1 (arbitrarily assigned).
- If we have k categories, we need to assign $k-1$ dummy variables for a single qualitative variable. These dummy variables are assigned to $k-1$ categories and the $k\ th$ category is called the reference category.

Federal Reserve Rates and Discount Rates

Question: Can the discount rate y be predicted from the Federal Reserve rate x ?

Data:

$$SSxx = 4.9166667$$

$$SSyy = 6.72917$$

$$SSxy = 3.20833$$

$$\bar{Y} = 6.21$$

$$n = 12$$

$$b_1 = 0.6525$$

$$b_0 = 1.6949$$

$$\hat{y} = 1.69 + 0.653x$$

Federal Reserve Rates and Discount Rates *continued*

$$\hat{y} = 1.69 + 0.653x$$

$$r^2 = \frac{(3.20833)^2}{(4.92)(6.73)}$$

$$= 0.3111$$

$$Se = 0.6808$$

$$r = 0.56$$

Interpretation

A test of the significance of the correlation coefficient would prove useful at this point. Set the level of confidence at 95%. With 10 degrees of freedom the critical value for t is therefore ± 2.228

Testing the Significance of the Slope and y-Intercept

- A regression model is not likely to be useful unless there is a significant relationship between x and y
- To test significance, we use the null hypothesis:

$$H_0: \beta_1 = 0$$

- Versus the alternative hypothesis:

$$H_a: \beta_1 \neq 0$$

Testing the Significance of the Slope and y-Intercept

A test of the significance of the sample regression coefficient of $b_1 = 0.6525424$ is also wise. The test will be conducted at the 99% level. With 10 degrees of freedom the critical t-value ± 3.169 .

$$H_0 : \beta_1 = 0$$

$$H_a : \beta_1 \neq 0$$

Reject H_0 if $\mathrm{t} < -3.169$ or $\mathrm{t} > 3.169$.

Do not reject H_0 if $-3.169 < \mathrm{t} < 3.169$.

The test requires that t_α, $t_{\alpha/2}$ and p-values are based on $n-2$ degrees of freedom

$$t = \frac{b_1}{s_{b_1}} \quad where\ s_{b_1} = \frac{Se}{\sqrt{SSxx}} \qquad = 0.681/\sqrt{4.92} = 0.307$$

$$= \frac{0.652542}{0.307} \qquad = 2.126$$

Interpretation
Since the value 2.126 lies between ± 3.169, the hypothesis that $\beta_1 = 0$ cannot be rejected.

Confidence Interval

- The point on the regression line corresponding to a particular value of x_0 of the independent variable x is $\hat{y} = b_0 + b_1 x_0$
- It is unlikely that this value will equal the mean value of y when x equals x_0
- Therefore, we need to place bounds on how far the predicted value might be from the actual value
- We can do this by calculating a confidence interval mean for the value of y and a prediction interval for an individual value of y

Bowerman, B. L., O'Connell, R. T., & Murphree, E. S., (2010). Business Statistics in Practice (6th Ed.), Copyright © McGraw-Hill Education

Confidence Interval *example*

People employed in banking and finance would be interested in an interval estimate for the mean value of the discount rate if the federal funds rate was held constant for several months at 7% . This is, of course, an interval estimate of the conditional mean of the discount rate:

$$\text{C.I. for } \mu_{y|x} = \hat{Y} \pm tS_Y$$

$$S_Y = Se\sqrt{\frac{1}{n} + \frac{(X - \overline{X})^2}{SSxx}}$$

$$= 0.681\sqrt{\frac{1}{12} + \frac{(7 - 6.9167)^2}{4.92}}$$

$$= 0.1982$$

$$\hat{Y} = b_0 + b_1 X$$

$$= 1.6949 + 0.6525424(7)$$

$$= 6.2627$$

Confidence Interval *continued*

If the interval is calculated at 95% level of confidence, the critical t–value is $t_{.05,n-2} = \pm 2.228$. We then have

$$\text{C.I. for } \mu_{\mathrm{y}|x} = \hat{Y} \pm tS_Y$$

$$= 6.2627 \pm (2.228)(0.1982)$$

$$5.82 < \mu_{\mathrm{y}|x} < 6.70$$

Interpretation

The banker can be 95% confident that if the federal funds rate is 7%,the mean discount rate they must pay to borrow money from the fed will fall between 5.82% and 6.70%. Their plans and policies can be formulated according to this expectation.

Prediction Interval

A prediction interval for a single of value of y for a given x. If we were to set x to sum amount just one time we would get a resulting value of y. We can be 95% certain that the single value of y falls within the specified interval.

$$\text{C.I. for } Y_x = \hat{Y} \pm tSy_i$$

$$S_{yi} = Se\sqrt{1 + \frac{1}{n} + \frac{(X - \overline{X})^2}{SSxx}} = 0.70927$$

Since $\hat{Y} = 6.2627$, we have

$$\text{C.I. for } Y_x = 6.2627 \pm (2.228)(0.70927)$$

$$4.68 < Y_x < 7.85$$

Prediction Interval *continued*

$$4.68 < Y_x < 7.85$$

Interpretation

The banker could formulate plans for next month's operations on the realization that he or she could be 95% confident that if the federal funds rate was 7%, the discount rate would fall between 4.68% and 7.85% percent. This is a wider range than that found for the conditional mean of the discount rate.

Which to Use?

- The prediction interval is useful if it is important to predict an individual value of the dependent variable
- A confidence interval is useful if it is important to estimate the mean value
- The prediction interval will always be wider than the confidence interval

Testing the Significance of the Population Correlation Coefficient

- The simple correlation coefficient (r) measures the linear relationship between the observed values of x and y from the sample
- The population correlation coefficient (ρ) measures the linear relationship between all possible combinations of observed values of x and y
- r is an estimate of ρ

Testing ρ

The hypotheses are: $H_0: \rho = 0$

$H_a: \rho \neq 0$

Reject H_0 if $t < -2.228$ or if $t > 2.228$

Do Not Reject H_0 if $t - 2.228 < t < 2.228$

$$t = \frac{r}{S_r} = \frac{r}{\sqrt{(1-r^2)/(n-2)}} = \frac{0.56}{\sqrt{(1-0.31)/10}} = \frac{0.56}{0.2627} = 2.13$$

The null hypothesis cannot be rejected.

Interpretation

It would certainly appear that the statement concerning the use of the federal funds rate to estimate or predict the discount rate is questionable. The r^2 is rather low, and the tests for significance of ρ and β_1 suggest that the hypotheses $\rho = 0$ and $\beta_1 = 0$ cannot be rejected at any acceptable levels of significance.