Dr. Saeed A. Dobbah Alghamdi

Department of Statistics
Faculty of Sciences
Building 90, Office 26F41
King Abdulaziz University
http://saalghamdy.kau.edu.sa

Dr. Saeed's Website

Statistics
for
Business & Economics

David R. Anderson
Dennis J. Sweeney
Thomas A. Williams

Simple Linear Regression

Chapter 12

Learning Objectives

LO1 Define the terms used in correlation analysis.

LO2 Calculate, test, and interpret the relationship between two variables using the *correlation* coefficient.

LO3 Apply regression analysis to estimate the linear relationship between two variables

LO4 Interpret the regression analysis.

LO5 Calculate and interpret confidence and prediction intervals.

Simple Linear Regression Analysis

Simple Linear Regression analysis is a statistical procedure develop a linear equation showing how the variables are related. In regression terminology, the variable being predicted is called the dependent variable. The variable being used to predict the value of the dependent variable is called the independent variable.

Examples

1. Is there a relationship between the amount Healthtex spends per month on advertising and its sales in the month?
2. Can we base a predict of the cost to heat a home in January on the number of square feet in the home?
3. Modeling the relationship between the miles per gallon achieved by large pickup trucks and the size of the engine?
4. Modeling the relationship between the number of hours that students studied for an exam and the score earned?

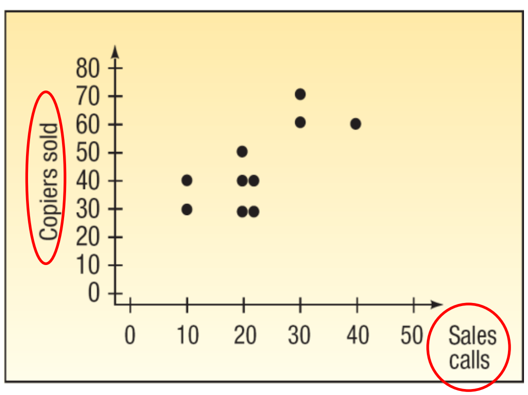
Regression Analysis

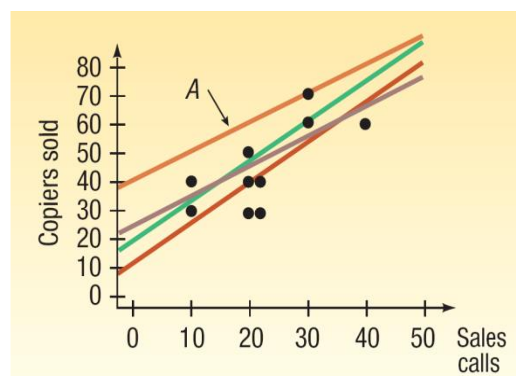
In regression analysis we use the independent variable (*X*) to estimate the dependent variable (*Y*).

- The relationship between the variables is linear.
- Both variables must be at least interval scale.
- The least squares criterion is used to determine the equation.

REGRESSION EQUATION An equation that expresses the linear relationship between two variables.

LEAST SQUARES PRINCIPLE Determining a regression equation by minimizing the sum of the squares of the vertical distances between the actual *Y values and* the predicted values of *Y*.





Regression Model and Regression Equation

SIMPLE LINEAR REGRESSION MODEL

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

β_0 (the y-intercept of the regression line) and β_1 (the slope) are referred to as the parameters of the model, and ε (the Greek letter epsilon) is a random variable referred to as the error term. The error term accounts for the variability in *y* that cannot be explained by the linear relationship between *x* and *y*.
In practice, the parameter values are not known and must be estimated using sample data. Sample statistics (denoted b_0 and b_1) are computed as estimates of the population parameters β_0 and β_1.

ESTIMATED SIMPLE LINEAR REGRESSION EQUATION

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

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

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

Regression Equation - Example

The sales manager gathered information on the number of sales calls made and the number of copiers sold for a random sample of 10 sales representatives. Use the least squares method to determine a linear equation to express the relationship between the two variables.

What is the expected number of copiers sold by a representative who made 20 calls?

Sales Representative	Number of Sales Calls	Number of Copiers Sold
Tom Keller	20	30
Jeff Hall	40	60
Brian Virost	20	40
Greg Fish	30	60
Susan Welch	10	30
Carlos Ramirez	10	40
Rich Niles	20	40
Mike Kiel	20	50
Mark Reynolds	20	30
Soni Jones	30	70

The regression equation is:

$$\hat{Y} = a + bX$$

$$\hat{Y} = 18.9476 + 1.1842X$$

$$\hat{Y} = 18.9476 + 1.1842(20)$$

$$\hat{Y} = 42.6316$$

Regression Equation - Example

Regression Analysis

r^2	0.576
r	0.759
Std. Error	9.901
n	10
k	1
Dep. Var.	**Copiers**

The value of the Pearson linear correlation coefficient

ANOVA table

Source	*SS*	*df*	*MS*	*F*	*p-value*
Regression	1,065.7895	1	1,065.7895	10.87	.0109
Residual	784.2105	8	98.0263		
Total	1,850.0000	9			

The value of the regression intercept "a"

The value of the regression slope "b"

Regression output

					confidence interval	
variables	*coefficients*	*std. error*	*t (df=8)*	*p-value*	*95% lower*	*95% upper*
Intercept	18.9474					
Calls	1.1842	0.3591	3.297	.0109	0.3560	2.0124

The predicated value when x=10

Predicted values for: Copiers

		95% Confidence Interval		*95% Prediction Interval*		
Calls	*Predicted*	*lower*	*upper*	*lower*	*upper*	*Leverage*
20	42.632	35.224	50.039	18.629	66.635	0.105

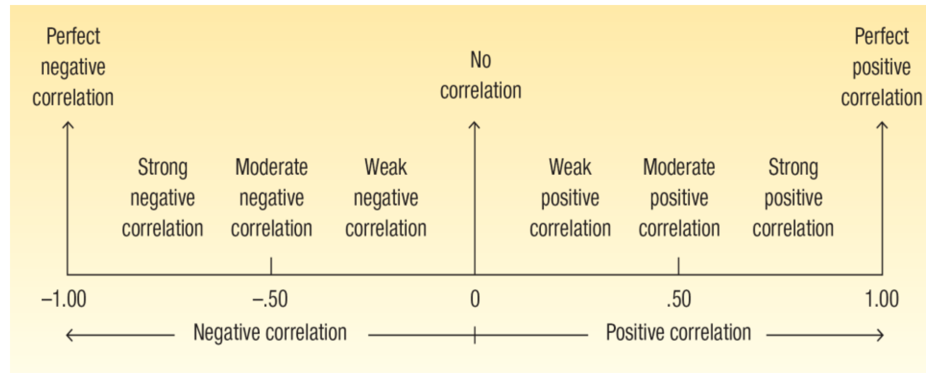
The equation of the fitted regression line is **$y' = 18.9474 + 1.1842\ x$**

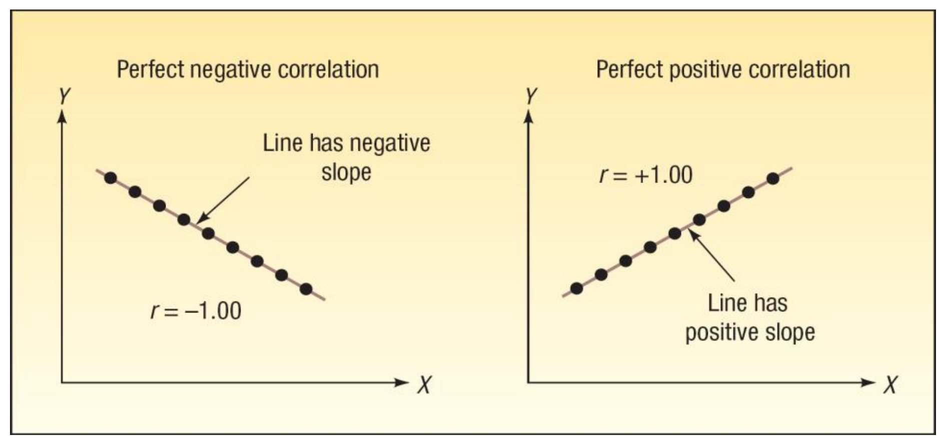
Hence, a representative who made 20 calls , we expect the number of copiers sold by him to be 42 on average.

The Coefficient of Correlation, r

The **Coefficient of Correlation** (r) a descriptive measure of the strength of linear association between two variables, x and y.

- It shows the direction and strength of the linear relationship between two interval or ratio-scale variables
- It can range from -1.00 to +1.00.
- Values close to 0.0 indicate weak correlation.
- Negative values indicate an **inverse** relationship and positive values indicate a **direct** relationship.
- Values of -1.00 or +1.00 indicate perfect and strong correlation.





The Coefficient of Determination, r^2

The coefficient of determination (r^2) is the proportion of the total variation in the dependent variable (Y) that is explained or accounted for by the variation in the independent variable (X). It is the square of the coefficient of correlation.

- It ranges from 0 to 1.
- It does not give any information on the direction of the relationship between the variables.

Example

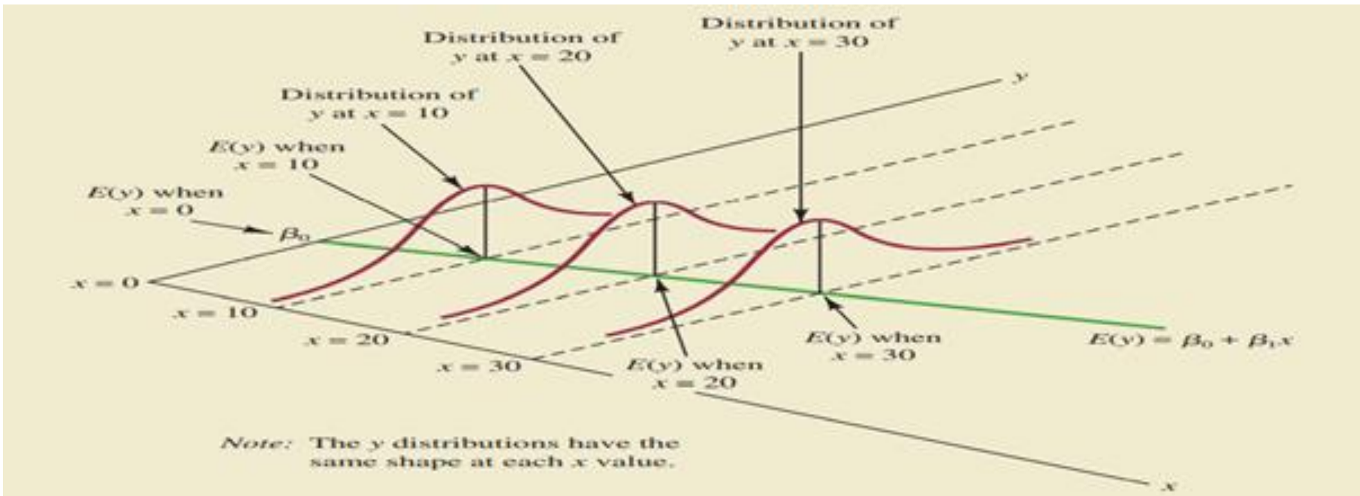
In the pervious example, compute and interpret the correlation coefficient and the coefficient of determination.

Regression Analysis		
	r^2	0.576
	r	0.759

The correlation coefficient is 0.759 which indicates strong direct relationship between number of calls and number of copiers sold. The coefficient of determination is 0.576 which means that 57.6% of the variation in the number of copiers sold is explained or accounted for by the variation in the number of calls.

Assumptions Underlying Linear Regression

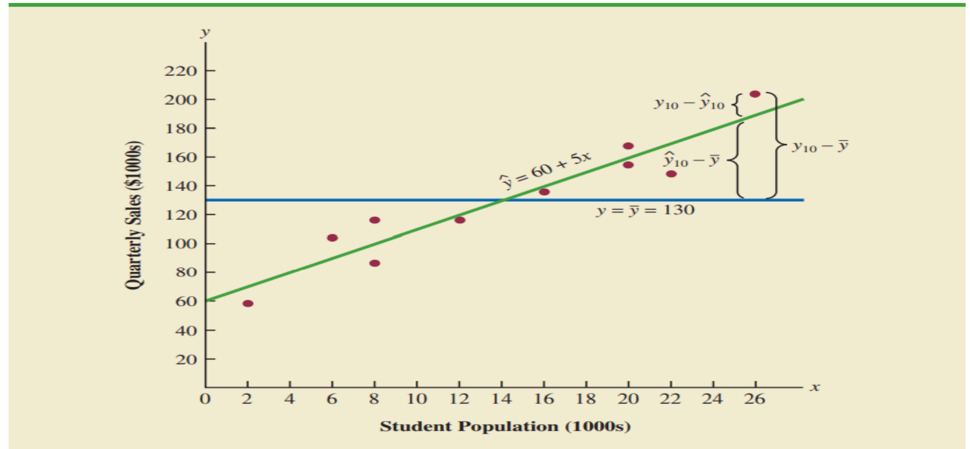
- For each value of X, there is a group of Y values, and these Y values are ***normally distributed***.



- The ***means*** of these normal distributions of Y values all lie on the straight line of regression.
- The ***standard deviations*** of these normal distributions are **equal**, and the best estimate are the standard error of the estimates $s_{y.x}$.
- The **Y values are statistically independent**. This means that in the selection of a sample, the Y values chosen for a particular X value do not depend on the Y values for any other X values.

The Standard Error of the Estimate

The **standard error of the estimate** measures the scatter, or dispersion, of the observed values around the line of regression.



Example

In the previous example, we found

Regression Analysis					
	r^2	0.576			
	r	0.759			
	Std. Error	9.901			
	n	10			
	k	1			
	Dep. Var.	Copiers			
ANOVA table					
Source	SS	df	MS	F	p-value
Regression	1,065.7895	1	1,065.7895	10.87	.0109
Residual	784.2105	8	98.0263		
Total	1,850.0000	9			

STANDARD ERROR OF THE ESTIMATE

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

The value of the standard error of the estimate

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.

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

If the value of β_1 is not equal to zero, H_0 is rejected, we would conclude that the two variables are related.

Example

In the previous example, test the significance of the regression at 5% significance level

Regression output					confidence interval	
variables	coefficients	std. error	t (df=8)	p-value	95% lower	95% upper
Intercept	18.9474					
Calls	1.1842	0.3591	3.297	.0109	0.3560	2.0124

The p-value for testing the significance of the regression relationship

The p-value 0.0109 is less than 0.05 indicating the rejection of the null hypothesis, thus the two variables are related.

Confidence Interval for β_1

The form of a confidence interval for β_1 is as follows:

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

The point estimate of β_1 is b_1 and the margin of error is $t_{\alpha/2} s_{b_1}$

Example

In the previous example, find the 95% confidence interval estimate of the slop β_1.

Regression output					confidence interval	
					95%	95%
variables	coefficients	std. error	t (df=8)	p-value	lower	upper
Intercept	18.9474					
Calls	1.1842	0.3591	3.297	.0109	0.3560	2.0124

Thus, the 95% confidence interval estimate of β_1 is (0.3560, 2.0124). Note that the confidence interval estimate dose not include zero which indicates that the slop value is significantly different form zero.

Confidence and Prediction Interval Estimates of y.

- A **confidence interval** reports the *mean* value of *Y* for a given *X*.

$$\hat{Y} \pm t(s_{y \cdot x})\sqrt{\frac{1}{n} + \frac{(X - \bar{X})^2}{\Sigma(X - \bar{X})^2}}$$

- *A* **prediction interval** reports the *range of values* of *Y* for a *particular* value of *X*.

$$\hat{Y} \pm ts_{y \cdot x}\sqrt{1 + \frac{1}{n} + \frac{(X - \bar{X})^2}{\Sigma(X - \bar{X})^2}}$$

where

$\hat{Y}$ is the predicted value for any selected X value.

X is any selected value of X.

$\bar{X}$ is the mean of the Xs, found by $\Sigma X/n$.

n is the number of observations.

$s_{y \cdot x}$ is the standard error of estimate.

t is the value of t from Appendix B.2 with $n - 2$ degrees of freedom.

Confidence and Predication Interval - Example

Example

In the previous example, find the 95% confidence interval for the average number of copiers sold and the 95% predication interval for the number of copiers sold when number of calls is 20.

Predicted values for: Copiers						
		95% Confidence Interval		95% Prediction Interval		
Calls	Predicted	lower	upper	lower	upper	Leverage
20	42.632	35.224	50.039	18.629	66.635	0.105

Thus, the 95% confidence interval for the average number of copiers sold is between 35.224 and 50.039 when the number of calls is 20.

The 95% prediction interval for the number of copiers sold is between 18.629 and 66.635 when the number of calls is 20.

Summary

Dependent variable The variable that is being predicted or explained.

Independent variable The variable that is doing the predicting or explaining.

Simple linear regression Regression analysis involving one independent variable and one dependent variable in which the relationship between the variables is approximated by a straight line.

Regression model The equation that describes how y is related to x and an error term.

Regression equation The equation that describes how the mean or expected value of the dependent variable is related to the independent variable.

Estimated regression equation The estimate of the regression equation developed from sample data by using the least squares method.

Least squares method A procedure used to develop the estimated regression equation.

Summary

Coefficient of determination A measure of the goodness of fit of the estimated regression equation. It can be interpreted as the proportion of the variability in the dependent variable y that is explained by the estimated regression equation.

Correlation coefficient A measure of the strength of the linear relationship between two variables (previously discussed in Chapter 3).

Standard error of the estimate The square root of the mean square error, denoted by s. It is the estimate of σ, the standard deviation of the error term.

ANOVA table The analysis of variance table used to summarize the computations associated with the F test for significance.

Confidence interval The interval estimate of the mean value of y for a given value of x.

Prediction interval The interval estimate of an individual value of y for a given value of x.