SUMMARY OF FORMULAS

Simple Linear Regression (SLR)

SLR Regression Model:

SLR Regression Equation:

Estimate SLR Equation:

Least Square Method

Slope of the Estimated Regression Equation:

y-Intercept for the Estimated Regression Equation:

Least Square Criterion:

Coefficient of Correlation:

Coefficient of Determination:

Sum of Squares due to Error (SSE):

Sum of Squares due to Regression (SSR):

Total Sum of Squares (SST= SSyy):

Mean Square Error (MSE):

Standard Error of the Estimate (s):

Standard Deviation of the Slope b_1:

Confidence Interval for β_1:

Covariance

Covariance for samples s_{xy} :

Correlation coefficient for samples:

Standard Deviation equation:

Coefficient of Determination using covariance:

Confidence Intervals and Prediction Intervals

Confidence Interval Estimate:

Estimated Standard Deviation for Confidence Interval:

Prediction Interval Estimate:

Estimated Standard Deviation for Prediction Interval:

PROBLEM # 10.1 Explain each term in the linear regression model,

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i.$$

PROBLEM # 10.2 What assumptions are required in using the linear regression model?

The linear regression model requires the following assumptions:

a. For any given value of x, the y values _____ with a mean that is on the regression line.

b. Regardless of the value of x, the standard deviation of the distribution of y values about the regression line is _____; this is the assumption of homoscedasticity.

c. Each value of y is statistically _____.

These assumptions can be restated in terms of the error term:

a. For any given value of x, the error terms will be _____ with a mean of 0.

b. The standard deviation for the distribution of error terms is _____.

c. The values of the error term are statistically _____.

PROBLEM # 10.3 In the linear regression equation, $\hat{y} = b_0 + b_1x_1$, why is the term at the left given as y hat instead of simply y?

PROBLEM # 10.4 What is the least-squares criterion, and what does it have to do with obtaining a regression line for a given set of data?

PROBLEM # 10.5 A scatter diagram includes the data points (x=2,y=10), (x=3, y=12), (x=4, y=20), and (x=5, y=16). Two regression lines are proposed: (1) $\hat{y} = 10 + x$, and (2) $\hat{y} = 8 + 2x$. Using the least-squares criterion, which of these regression lines is the better fit to the data? Why?

x	**y**	$\hat{y}$	$(y - \hat{y})$	$(y - \hat{y})^2$

x	**y**	$\hat{y}$	$(y - \hat{y})$	$(y - \hat{y})^2$

PROBLEM # 10.6 A scatter diagram includes the data points (x=3,y=8), (x=5, y=18), (x=7, y=30), and (x=9, y=32). Two regression lines are proposed: (1) $\hat{y}$ = 5 + 3x, and (2) $\hat{y}$ = -2 + 4x. Using the least-squares criterion, which of these regression lines is the better fit to the data? Why?

x	**y**	$\hat{y}$	$(y - \hat{y})$	$(y - \hat{y})^2$

x	**y**	$\hat{y}$	$(y - \hat{y})$	$(y - \hat{y})^2$

PROBLEM # 10.7 For a sample of 8 employees, a personnel director has collected the following data on ownership of company stock versus years with the firm.

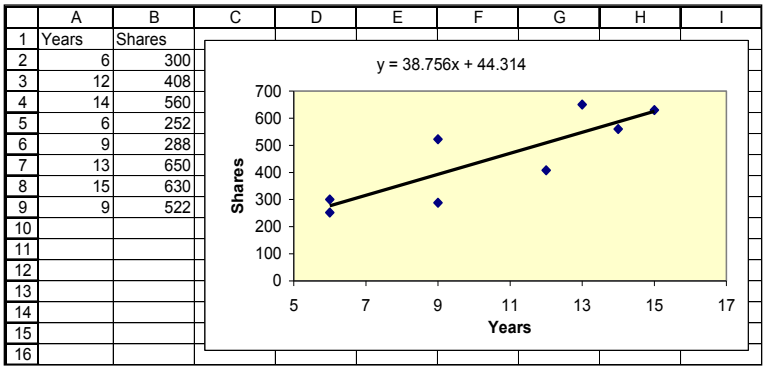
X= years	**Y = shares**
6	300
12	408
14	560
6	252
9	288
13	650
15	630
9	522

a. Determine the least-squares regression line and interpret its slope.

x	y	xy	x^2
6.00	300.00	1800.00	36.00
12.00	408.00	4896.00	
14.00	560.00		
6.00	252.00	1512.00	36.00
9.00	288.00		81.00
13.00	650.00	8450.00	
15.00	630.00	9450.00	225.00
9.00	522.00		81.00
Σ x =	Σ y =	Σ xy =	Σx^2 =

b. For an employee who has been with the firm 10 years, what is the predicted number of shares of stock owned?

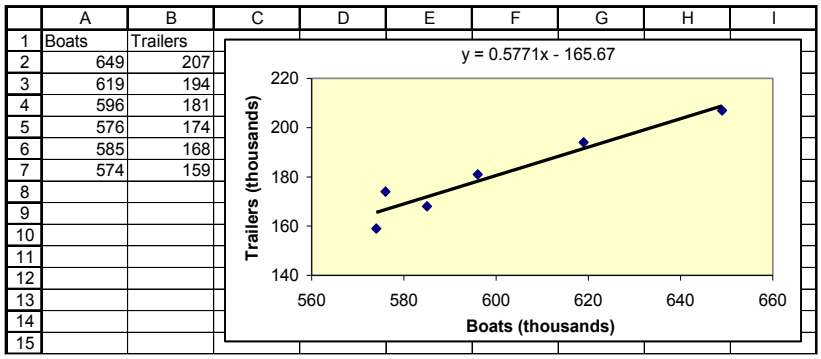
	C	D	E	F	G	H	I
1	SUMMARY OUTPUT						
2							
3	*Regression Statistics*						
4	Multiple R	0.849					
5	R Square	0.720					
6	Adjusted R Square	0.673					
7	Standard Error	91.479					
8	Observations	8					
9							
10	ANOVA						
11		*df*	*SS*	*MS*	*F*	*Significance F*	
12	Regression	1	129173.13	129173.13	15.436	0.008	
13	Residual	6	50210.37	8368.40			
14	Total	7	179383.50				
15							
16		*Coefficients*	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
17	Intercept	44.3140	108.51	0.408	0.697	-221.197	309.825
18	Years	38.7558	9.86	3.929	0.008	14.618	62.893



PROBLEM # 10.8 The following data represent x= boat sales and y = boat trailer sales from 1995 through 2000. Source: Bureau of the Census, Statistical Abstract of the United States 1999, p269 and Statistical Abstract 2002, p 755.

Year	Boat Sales (Thousands)	Boat Trailer Sales (Thousands)		
1995	649	207		
1996	619	194		
1997	596	181		
1998	576	174		
1999	585	168		
2000	574	159		
	$\Sigma x =$	$\Sigma y =$	$\Sigma xy =$	$\Sigma x^2 =$

a. Determine the least- squares regression line and interpret its slope.
b. Estimate, for a year during which 500, 000 boats are sold, the number of boat trailers that would be sold.
c. What reasons might explain why the number of boats sold per year is less than the number of boats sold per year?



PROBLEM # 10.9 For a set of data, the total variation or sum of squares for y is SST = 143.0, and error sum of squares is SSE = 24.0. What proportion of the variation in y is explained by the regression equation?

PROBLEM # 10.10 In a regression analysis, the sum of the squared deviations between y and $\bar{y}$ is SST = 120.0. If the coefficient of correlation of r = 0.7, what are the values of SSE and SSR?

PROBLEM # 10.11 What is the standard error of estimate, what role does it play in simple linear regression and correlation analysis?

The standard error of the estimate is a measure which describes

_____.

The standard error may be used as a measure of how well the regression line fits the data. The standard error is used in calculating _____ for the mean value of y given a specific value of x, and in calculating the _____for an individual y observation.

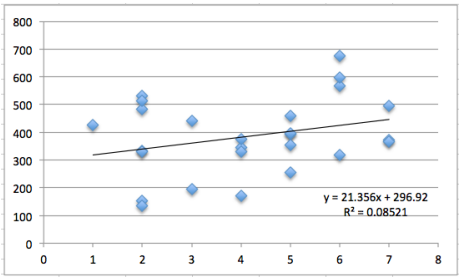
PROBLEM # 10.12 Is it possible for the standard error of estimate to be equal to zero? If so, under what circumstances?

PROBLEM # 10.13 For a set of 8 data points, the sum of the squared differences between observed and estimated values of y is 34.72. Given this information what is the standard error of estimate?

PROBLEM # 10.14 A computer analysis of 30 pairs of observations results in the least-squares regression equation $\hat{y}$ = 14.0+ 5.0 x, and the standard deviation of the slope is listed as s_{b1} =2.25.

a. At the 0.05 level of significance, can we conclude that no linear relationship exists within the population of x and y values?
b. Construct the 95% confidence interval for the population slope, β_1.
c.

PROBLEM # 10.15 At 5% level significance. The manager of Colonial Furniture has been reviewing weekly advertising expenditures. During the past 6 months, all advertisements for the store have appeared in the local newspaper. The number of ads per week has varied from one to seven. The store's sales staff has been tracking the number of customers who enter the store each week. The number of ads and the number of customers per week for the past 26 weeks were recorded.



a. Determine the sample regression line
b. Interpret the coefficients
c. Can the manager infer that the larger the number of ads, the larger the number of customers? REFER TO POWERPOINT SLIDES LESSON 10/11.
d. Find and interpret the coefficient of determination.
e. In your opinion, is it worthwhile exercise to use the regression equation to predict the number of customers who will enter the store, given that Colonial intends to advertise five times in the newspaper? If so, find the 95% prediction interval. If not, explain why not.

Ads	Customer
5	353
6	319
3	440
2	332
4	172
2	331
4	344
2	483
4	329
2	532
7	496
5	393
4	376
7	372
2	512
5	254
5	459
2	153
1	426
6	566
6	596
5	395
6	676
3	194
2	135
7	367

PROBLEM # 10.16 At 5% level of significance. The president of a company that manufactures car seats has been concerned about the number and cost of machine breakdowns. The problem is that the machines are old and becoming quite unreliable. However, the cost of replacing them is quite high, and the president is not certain that the cost can be made up in today's slow economy. To help make a decision about replacement, he gathered data about last month's costs for repairs and the ages (in months) of the plant's 20 welding machines.

Age	Repairs
110	327.67
113	376.68
114	392.52
134	443.14
93	342.62
141	476.16
115	324.74
115	338.98
115	433.45
142	526.37
96	362.42
139	448.76
89	335.27
93	350.94
91	291.81
109	467.80
138	474.48
83	354.15
100	420.11
137	416.04

a. Find the sample regression line. Run the values through Excel.
b. Interpret the coefficients.
c. Determine the coefficient of determination, and discuss what this statistic tells you.
d. Conduct a test to determine whether the age of a machine and its monthly cost of repair are linearly related.

e. Is the fit of the simple linear model good enough to allow the president to predict the monthly repair cost of a welding machine that is 120 months old? If so, find a 95% prediction interval. If not explain why not.

PROBLEM # 10.17 What happens to the width of a prediction interval for y as the x value on which the interval estimate is based gets farther away from the mean of x? Why?

PROBLEM # 10.18 For n=6 data points, the following quantities have been calculated:

$\sum x = 40 \quad \sum y = 76 \quad \sum xy = 400$

$\sum x^2 = 346 \quad \sum y^2 = 1160 \quad \sum (y - \hat{y})^2 = 52.334$

a. Determine the least-squares regression line.
b. Determine the standard error of estimate.
c. Construct the 95% confidence interval for the mean of y when x=7.0
d. Construct the 95% confidence interval for the mean of y when x=9.0
e. Compare the width of the confidence interval obtained in part (c) with the obtained in part (d). Which is wider and why?

PROBLEM # 10.19 For the summary data provided in Problem #10.18, construct a 95% prediction interval for an individual y value whenever

a. x=2
b. x=3
c. x=4

PROBLEM # 10.20 Differentiate between a confidence interval and a prediction interval.

The point estimate for the mean will fall on the regression line.

PROBLEM # 10.21 Differentiate between the coefficients of correlation and determine. What information does each one offer that the other does not?

PROBLEM # 10.22 Attempting to analyze the relationship between advertising and sales, the owner of a furniture store recorded the monthly advertising budget ($ thousands) and the sales ($ millions) for a sample of 12 months. The data are listed here:

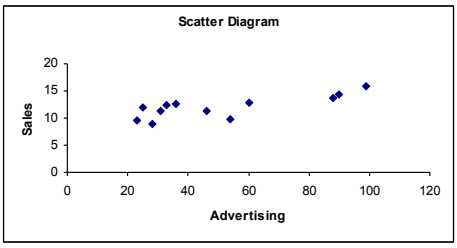
Advertising	23	46	60	54	28	33
Sales	9.6	11.3	12.8	9.8	8.9	12.5
Advertising	25	31	36	88	90	99
Sales	12.0	11.4	12.6	13.7	14.4	15.9

a. Does it appear that advertising and sales are linearly related?

$\sum_{i=1}^{n} x_i = 613 \quad \sum_{i=1}^{n} y_i = 144.9 \quad \sum_{i=1}^{n} x_i^2 = 39{,}561 \quad \sum_{i=1}^{n} x_i y_i = 7{,}882.2$

b. Calculate the least squares line and interpret the coefficients.
c. Determine the standard error of estimate
d. Is there evidence of a linear relationship between advertising and sales?
e. Estimate β_1 with 95% confidence.

f. Compute the coefficient of determination and interpret the value.



PROBLEM # 10.24 In a regression analysis, the sum of the squared deviations between y and $\bar{y}$ is SST = 200.0. If the sum of the squared deviations about the regression line is SSE= 40.0, what is the coefficient of determination?

PROBLEM # 10.25 For households in a community, many different variables can be observed or measured. If y= monthly mortgage payment, x= annual income would probably be directly related to y. For y= monthly mortgage payment, provide an example of an x variable likely to be (a) directly related to y, (b) inversely related to y, and (c) unrelated to y.

a. The monthly mortgage payment would likely be directly related to the market value of the house, the interest rate, the size of the house, or the monthly taxes and insurance, among other variables.

b. The monthly mortgage payment would likely be inversely related to the age of the house, among other variables.

c. The monthly mortgage payment would likely be unrelated to the amount of chocolate consumed by the owners, and a wide variety of other variables.

PROBLEM # 10.26 A Tire Company has carried out tests in which rolling resistance (pounds) and inflation pressure (pounds per square inch, or psi) have been measured for psi values ranging from 20 to 45. The regression analysis is summarized in the following MiniTab printout:
Regression Analysis
The regression equation is ROLRESIS = 9.45 – 0.0811 PSI

Predictor	Coef	StDev	T	P
Constant	9.450	1.228	7.69	0.000
PSI	-0.08113	0.03416	-2.38	0.029

S= 0.8808 R-Sq = 23.9% R-Sq (adj) = 19.6%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	1	4.3766	4.3766	5.64	0.029
Error	18	13.9657	0.7759		
Total	19	18.3422			

a. To the greatest number of decimal places in the print-out, what is the least-squares regression line?
b. What proportion of the variation in rolling resistance is explained by the regression line?
c. At what level of significance does the slope of the line differ from zero? What type of test did Minitab use in reaching this conclusion?

d. At what level of significance does the coefficient of correlation differ from zero? Compare this with the level found in part (c) and explain either why they are different or why they are the same.
e. Construct the 95% confidence interval for the slope of the population regression line.

Understanding the Basics: Suggested Problems from the Book.

In **Bold** are the Suggested Problems, in Green are the problems on Connect and the book.

	Chapter 13									
13.1	The Simple Linear Regression Model and the Least Sc	**13.01**	**13.02**	**13.03**	**13.04**	**13.05**	**13.06**			
13.2	Simple Coefficients of Determination and Correlation	**13.07**	**13.08**	**13.09**	**13.10**	13.11	13.12			
13.3	Model Assumptions and the Standard Error	**13.13**	**13.14**	**13.15**	13.16	**13.17**	**13.18**			
13.4	Testing the Significance of the Slope and y-intercept	**13.19**	13.20	**13.21**	13.22	**13.23**	13.24	13.25	**13.26**	13.27
13.5	Confidence and Prediction Intervals	**13.28**	**13.29**	**13.30**	**13.31**	**13.32**	**13.33**	**13.34**		
13.6	Testing the Significance of the Population Correlation (	**13.36**	**13.37**	13.38	**13.39**					
	Supplementary	13.50	13.51	13.52	**13.53**	13.54				

	Chapter 3				
3.7	Decision Trees	**3.54**	**3.55**	3.56	3.57
3.8	Cluster Analysis and Multidimensional Scaling	**3.58**	**3.59**	3.60	3.61
3.9	Factor Analysis	**3.62**	**3.63**	3.64	3.65

This statistical workbook is compiled from the following books:

- Keller, G. (2012). *Statistics for management and economics*. Mason: Cengage Learning.
- McClave, J. T., Benson, G. P., & Sincich, T. (2008). *Statistics for Business and Economics*. New Jersey: Prentice Hall.
- Weiers, R. M. (2011). *Introduction to Business Statistics*. Mason: Cengage Learning.
- (GMAC), F. t. (Ed.). (2005). *GMAT -Quantitative Review.* Oxford, UK: Blackwell.
- Bowerman, B. L., O'Connell, R. T., Murphree, E., Huchendorf, S. C., & Porter, D. C. (2003). *Business statistics in practice*(pp. 728-730). New York: McGraw-Hill/Irwin.