

FR 3218 Spring 2008

Assignment 1 Solutions

1. The first part of the question asked that you calculate the average, standard deviation, coefficient of variation, and 90% confidence interval of the hunter success data for Oracle Junction and Pinnacle Peak.

Start with calculation of the following statistics:

Statistic	Oracle Junction	Pinnacle Peak
n	7	7
$\sum x$	29.32	21.57
$\sum x^2$	138.3528	72.5043

Use n , $\sum x$, and $\sum x^2$ to calculate the following statistics:

Statistic	Formula	Oracle Junction	Pinnacle Peak
Average or sample mean	$\bar{x} = \frac{\sum x}{n}$	4.19 quail/trip	3.08 quail/trip
Standard deviation	$s = \sqrt{\frac{\sum x^2 - \frac{(\sum x)^2}{n}}{n-1}}$	1.61 quail/trip	1.00 quail/trip
Coefficient of variation	$CV = \frac{s}{\bar{x}} * 100$	38.43 %	32.55 %
Standard error of the mean	$s_{\bar{x}} = \sqrt{\frac{s^2}{n}}$	0.61 quail/trip	0.38 quail/trip
Confidence interval	$\bar{x} \pm t * s_{\bar{x}}$; where t is 1.943 from Appendix Table 6	3.01 to 5.37 quail/trip	2.34 to 3.81 quail/trip

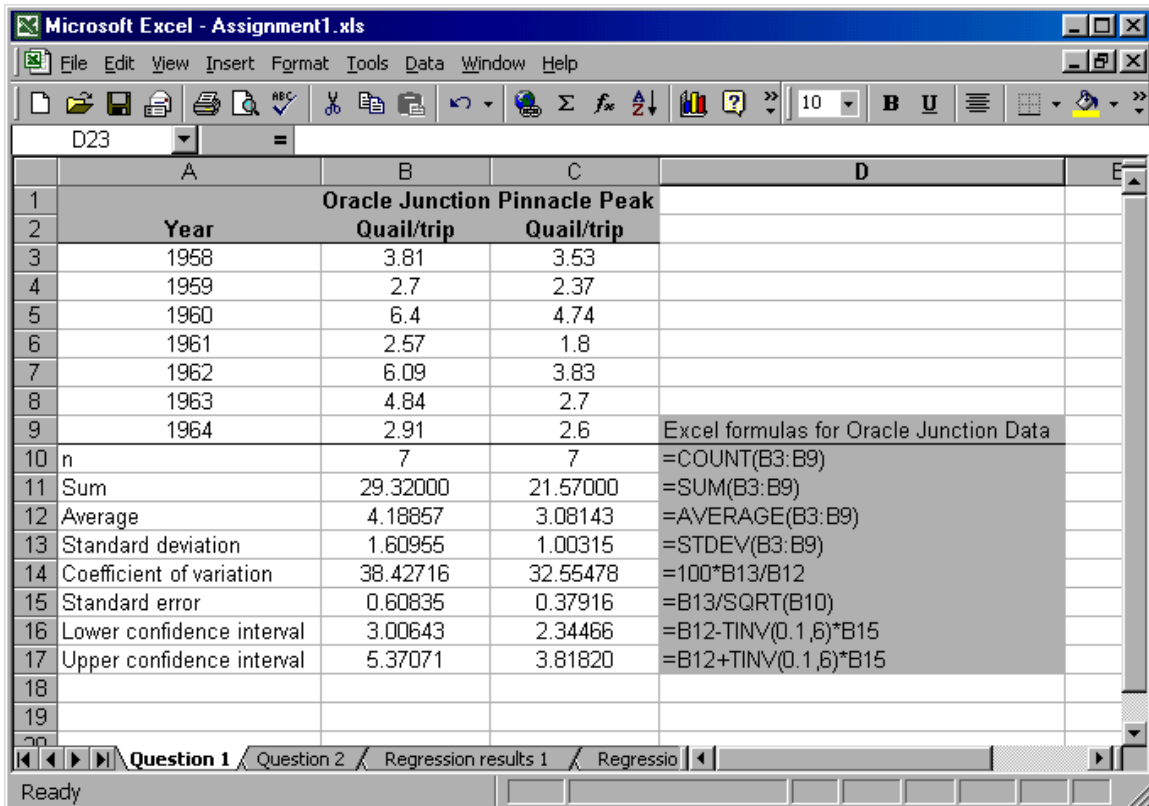
All significant digits were carried throughout the calculations. Rounding occurred at the end when the values were reported.

A t-value of 1.943 was used in the calculation of the confidence interval. Degrees of freedom (df) were 6 (sample size $n - 1$). The probability was 0.1 (1 - 0.90). The value was found in Appendix Table 6.

The second part of the question dealt with estimating mean success and standard error for two hunters in Oracle Junction. The textbook in section 2-16 *Expansion of Means and Standard Errors* gives direction on how to do the calculations. "The rule to remember is that expansion of sample means must be accompanied by a similar expansion of standard errors." (p. 21) Multiply the mean and standard error for one hunter by two to estimate the mean and standard error of hunting parties (2 hunters).

Statistic	Hunting parties in Oracle Junction
Mean	8.38 quail/trip
Standard error	1.22 quail/trip

The calculations done by hand were checked using Excel. Formulas used to calculate Oracle Junction statistics are shown in column D. Excel functions used include COUNT, SUM, AVERAGE, STDEV, and TINV. The TINV function provides t-values for a given probability and degrees of freedom.



2. The first part of the question asked for calculation by hand of the intercept and slope, correlation coefficient, and standard error about the regression of visitor hours (Y) on traffic counter reading (X). Start by calculating the following 5 statistics:

Statistic	Value
n	16
$\sum X$	7591
$\sum X^2$	5773217
$\sum Y$	11421
$\sum Y^2$	15681127
$\sum XY$	9263698

Use n , $\sum X$, $\sum X^2$, $\sum Y$, $\sum Y^2$, and $\sum XY$ to calculate the corrected sum of squares and cross products.

Statistic	Formula	Value
Corrected sum of squares for Y	$SS_y = \sum Y^2 - \frac{(\sum Y)^2}{n}$	7528674.4375
Corrected sum of squares for X	$SS_x = \sum X^2 - \frac{(\sum X)^2}{n}$	2171761.9375
Corrected sum of cross products	$SP_{xy} = \sum XY - \frac{(\sum X)(\sum Y)}{n}$	3845147.3125

Regression estimates are then estimated using the follow formulas:

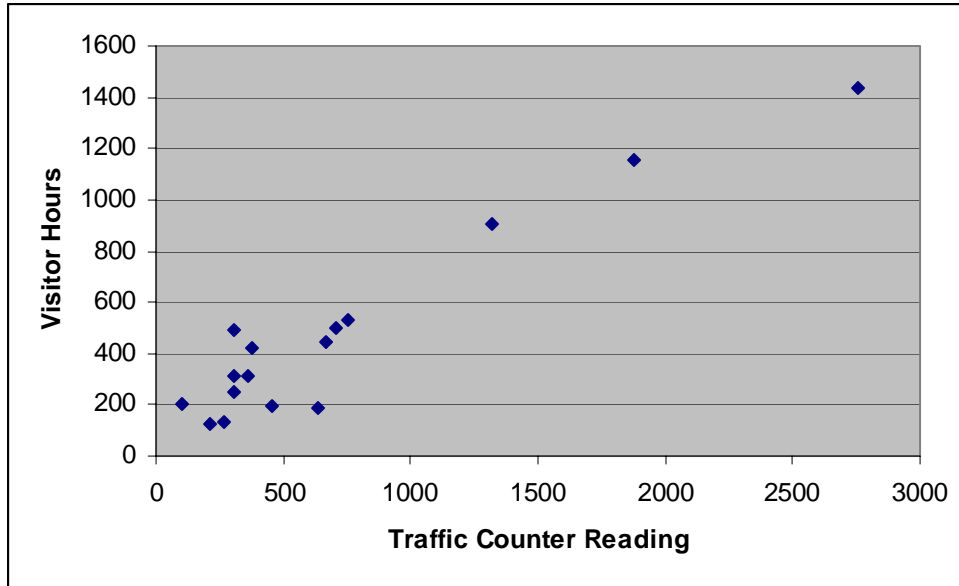
Regression estimates	Formula	Value
Slope b_1	$b_1 = \frac{SP_{xy}}{SS_x}$	1.77
Intercept b_0	$b_0 = \frac{\sum Y}{n} - b_1 \frac{\sum X}{n}$	-126.19
Coefficient of determination (correlation coefficient)	$r^2 = \frac{(SP_{xy})^2}{SS_x SS_y}$	0.90
Standard error about the regression	$S_{yx} = \sqrt{\frac{SS_y - \frac{(SP_{xy})^2}{SS_x}}{n - 2}}$	226.90

The fitted equation is

Visitor hours = -126.19 + 1.77 traffic reading counter

Visitor hours for a traffic reading counter of 350 is -126.19 + 1.77*350 = 493 visitor hours.

These calculations were checked against Excel regression results. First a scatter plot of visitor hours versus traffic counter reading was made. This is generally a good first step when exploring the relationship between two variables.



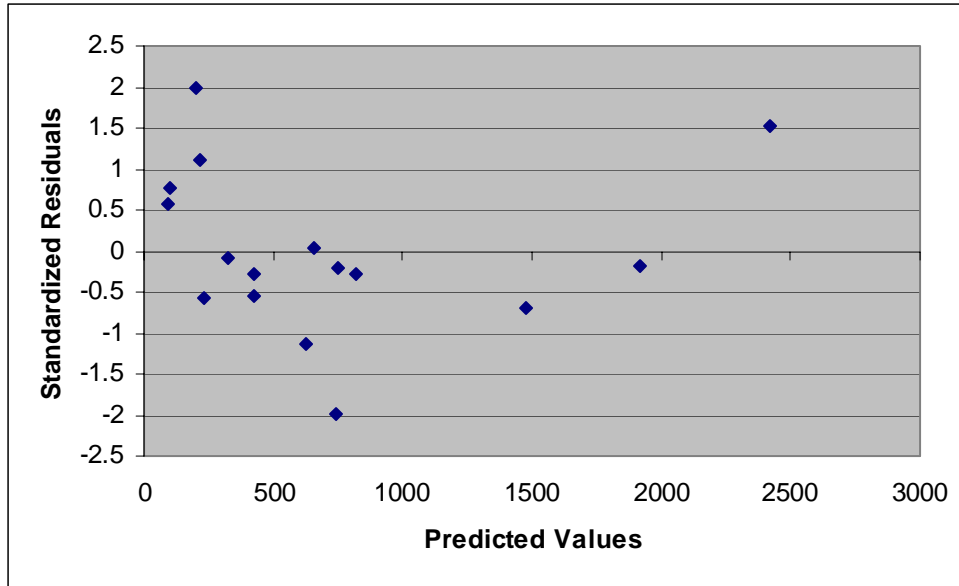
The assumption of a linear relationship between visitor hours and traffic counter reading appears to be valid. More data points in excess of 1000 traffic counter reading would be beneficial to increase confidence that a linear model is appropriate.

The regression of visitor hours on traffic counter reading in Excel produced the following results:

SUMMARY OUTPUT					
<i>Regression Statistics</i>					
Multiple R	0.950928				
R Square	0.904264				
Adjusted R Square	0.897426				
Standard Error	226.8992				
Observations	16				
ANOVA					
	<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>Significance F</i>
Regression	1	6807909	6807909	132.2354	1.61E-08
Residual	14	720765.3	51483.24		
Total	15	7528674			
	<i>Coefficients</i>	<i>Standard Error</i>	<i>t Stat</i>	<i>P-value</i>	
Intercept	-126.188	92.48595	-1.36441	0.193971	
Traffic Counter	1.77052	0.153967	11.49937	1.61E-08	

Estimates for the slope, intercept, standard error about the regression, and coefficient of determination are consistent with the estimates calculated by hand (a good sign!).

Standardized residuals versus predicted values are graphed below:



The residuals range between -2 and 2 ; a good sign as a very large residual is indicative of an outlier. More data points would be beneficial in testing the assumptions of linearity and constant variance.

The second part of the question asked that the simple linear regression assuming a slope of zero be calculated. The slope is calculated from the following equation:

$$b_1 = \frac{\sum XY}{\sum X^2} = 1.60$$

Running the regression assuming an intercept of zero in Excel produces the following results:

SUMMARY OUTPUT					
<i>Regression Statistics</i>					
Multiple R	0.944211				
R Square	0.891534				
Adjusted R Square	0.824867				
Standard Error	233.3248				
Observations	16				
ANOVA					
	<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>Significance F</i>
Regression	1	6712068	6712068	123.2919	2.52E-08
Residual	15	816606.8	54440.45		
Total	16	7528674			
	<i>Coefficients</i>	<i>Standard Error</i>	<i>t Stat</i>	<i>P-value</i>	
Intercept	0	#N/A	#N/A	#N/A	
Traffic Counter	1.604599	0.097107	16.52398	4.92E-11	

Number of visitor hours for a traffic reading counter of 350 is $1.60 \times 350 = 562$ visitor hours. The first model (with intercept) estimated 493 visitor hours, 69 visitor hours less than the second model (intercept of 0).

Standard error about the regression was slightly lower in the first model (with intercept) than the second model (intercept of 0), which suggests that the first model is a better fit. However, closer inspection of the first model reveals that the intercept is not significant. The models are similar enough that it is difficult to distinguish one as "better".