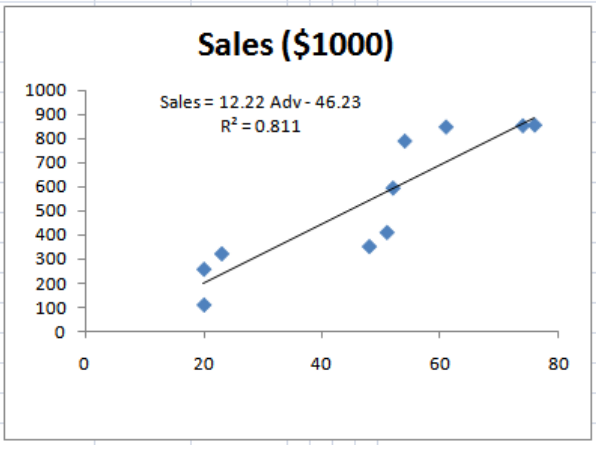
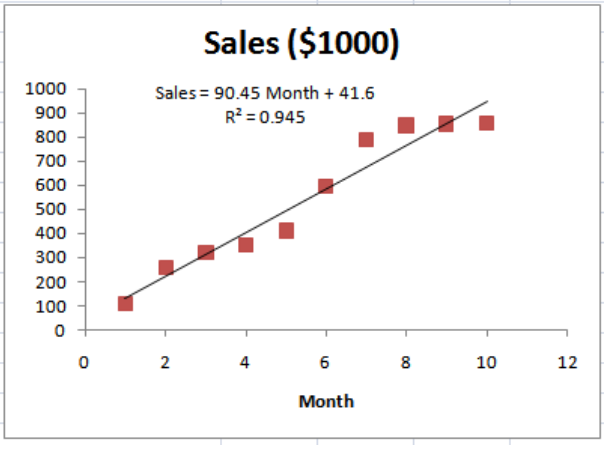
Simple and Multiple Regression Analysis

Example: Explore the relationships among Month, Adv.$ and Sales $:

1. Prepare a scatter plot of these data. The scatter plots for Adv.$ versus Sales, and Month versus Sales are given in the Figures below with Excel@ Insert/Scatter.
   a. Do the data appear to be stationary or nonstationary? The data appear to be nonstationary, it is not random, but with clear linear trend upward.
   b. Do the data appear to have a trend? Yes, the data have clear up trend, that is as the Adv.$ or Month increase, the Sales increase as well.
   c. If we want to fit a straight line to the data, how many lines could we possibly fit? We can fit infinite number of straight lines to the data. Each line is represented with a different set of b_0 (Y intercept), b_1 (Slope for Month) and b_2 (Slope for Adv.$) for this case.
   d. Compute the coefficient of correlation *r* between Month, Adv.$ and Sales, respectively, with =CORREL(Array1,Array2) and interpret the meanings. *r*(Adv versus Sales) = 0.901 and *r*(Month versus Sales) = 0.9722 indicate strong positive correlation between the Adv and Sales, and Month and Sales, respectably.

	A	B	C	H	I	J	K	L
1	Month	Adv ($1000)	Sales ($1000)	Linear Relationship between Adv$ and Sales $:				
2	1	20	112	Coef. of correlation r b/t Adv$ & Sales $	0.9010	=CORREL(B2:B11,C2:C11)		
3	2	20	259	Coef. of correlation r b/t Month & Sales $	0.9722	=CORREL(A2:A11,C2:C11)		
4	3	23	323					
5	4	48	353					
6	5	51	411					
7	6	52	594					
8	7	54	788					
9	8	61	846					
10	9	74	851					
11	10	76	854					





(Regression.xls/Reg0)

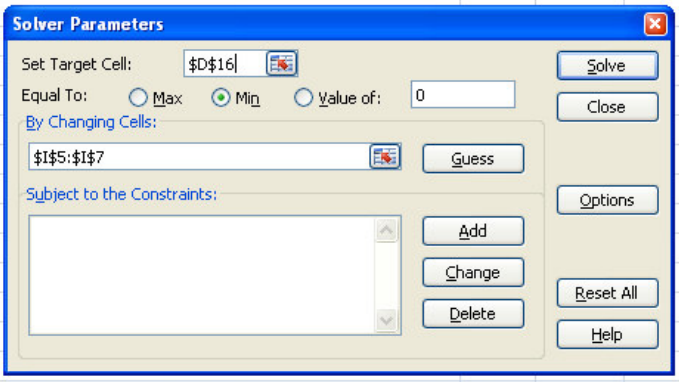
2. What is the general linear model to be used to model linear trend? (Write out the model)

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \varepsilon_i \text{ or}$$
$$Sales_i = \beta_0 + \beta_1 Month_i + \beta_2 Adv_i + \varepsilon_i$$

where Y_i is the Sales in Month I with the amount of Adv.$ given in Month *I*, β_0 is the *Y* intercept, or the Sales at Month =0 and Adv.$ = 0, β_1 is the slope of the regression line drawn with Month as independent variable (X_1) and Sales as dependent variable (Y), it shows the marginal change (increase or decrease) in Sales when the variable Month changes one unit (increase or decrease) while keep no change for all of other variables, β_2 is the slope of the regression line drawn with Adv$ as independent variable (X_2) and Sales as dependent variable (Y), it shows the marginal change (increase or decrease) in Sales (Y) when the amount of the variable Adv$ (X_2) incrementally changes (increases or decreases ONE unit) while keep no change for all of other variables.

3. Use FIVE possible ways in Excel@ to find b_0, b_1 and b_2 in the linear regression model for Adv, Month and Sales data set, and predict Sales in Months 11 to 13.
   a. Use Excel@ Solver to Minimize ESS or SSE in order to get optimal values of b_0, b_1 and b_2. 1) to assign arbitrary values for b_0, b_1 and b_2 first, 2) compute Sales = $b_0 + b_1$ (Month) $+b_2$ (Adv), 3) compute SSE with =SUMXMY2(SalesRange,FcstRagne), 4) use Excel@ Solver to minimize SSE to get the optimal values of b_0, b_1 and b_2.

	A	B	C	D	E	F	G	H	I	J	K	L
1	Month	Adv ($1000)	Sales ($1000)	Est Sales ($1000)				Linear Relationship between Adv$ and Sales $:				
2	1	20	112	350.000				Coef. of correlation r b/t Adv$ & Sales $	0.9010	=CORREL(B2:B11,C2:C11)		
3	2	20	259	400.000				Coef. of correlation r b/t Month & Sales $	0.9722	=CORREL(A2:A11,C2:C11)		
4	3	23	323	480.000								
5	4	48	353	780.000				Arbitrary b_0	100			
6	5	51	411	860.000				Arbitrary b_1 (Month)	50.000			
7	6	52	594	920.000				Arbitrary b_2 (Adv)	10.000			
8	7	54	788	990.000								
9	8	61	846	1110.000				b0 Adv only	-46.239	=INTERCEPT(C2:C11,B2:B11)		
10	9	74	851	1290.000				b1 Adv only	12.220	=SLOPE(C2:C11,B2:B11)		
11	10	76	854	1360.000								
12	11							b0 Month only	41.6	=INTERCEPT(C2:C11,A2:A11)		
13	12							b1 Month only	90.455	=SLOPE(C2:C11,A2:A11)		
14	13								b_2	b_1	b_0	
15								b_0, b_1 & b_2 with =LINEST() w/F2, CTRL+SHIFT+ENTER	-6.862	135.828	120.743	
16	SSE			1150637								
17												
18												
19												
20												
21												
22												
23												
24												
25												
26												
27												
28												

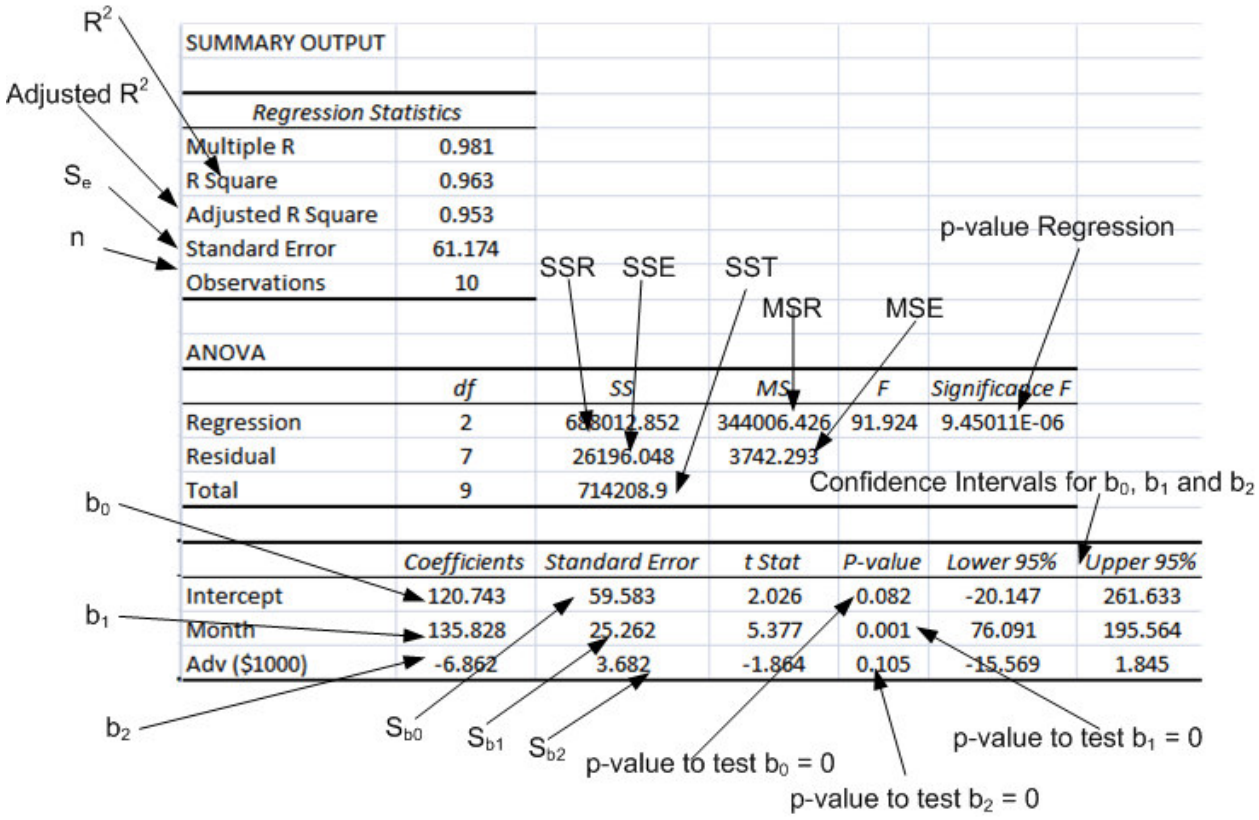


(Regression.xls/Reg1)

Use Excel@ Solver to get optimal values of b_0, b_1 and b_2 that will minimize SSE
Objective Function: SSE
Changing Cells: I5:I7

b. Use Excel@ Data/Data Analysis/Regression to get the Summary Output for the data and print a copy of it, find values of b_0, b_1, and b_2 in the Summary Output. The values of b_0, b_1, and b_2 are labeled in the Summary Output below.

Meanings of Regression Summary Output



SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.981
R Square	0.963
Adjusted R Square	0.953
Standard Error	61.174
Observations	10

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	2	688012.852	344006.426	91.924	9.45011E-06
Residual	7	26196.048	3742.293		
Total	9	714208.9			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	120.743	59.583	2.026	0.082	-20.147	261.633
Month	135.828	25.262	5.377	0.001	76.091	195.564
Adv ($1000)	-6.862	3.682	-1.864	0.105	-15.569	1.845

(Regression.xls/Reg1SOa)

c. Use Excel@ =LINEST(ArrayY, ArrayXs) to get b_0, b_1 and b_2 simultaneously. Use Excel@ =LINEST(C2:C11,A2:B11) as in Regression.xls/Reg1. Note, Highlight the I15:K15, type =LINEST(C2:C11,A2:B11), then CTRL+SHIFT+ENTER.

	H	I	J	K	L	M
14		b_2	b_1	b_0		
15	b_0, b_1 & b_2 with =LINEST() w/F2, CTRL+SHIFT+ENTER	-6.862	135.828	120.743	=LINEST(C2:C11,A2:B11)	

(Regression.xls/Reg1)

d. =INTERCEPT(Y-RANGE,X-RANGE) for b_0 and =SLOPE(Y-RANGE,X-RANGE) for b_1 when only single X variable is considered each time.

	G	H	I	J	K	L
9		b0 Adv only	-46.239	=INTERCEPT(C2:C11,B2:B11)		
10		b1 Adv only	12.220	=SLOPE(C2:C11,B2:B11)		
11						
12		b0 Month only	41.6	=INTERCEPT(C2:C11,A2:A11)		
13		b1 Month only	90.455	=SLOPE(C2:C11,A2:A11)		

(Regression.xls/Reg1)

e. Click any data point on the scatter plots for Month and Sales, or Adv and Sales, select Add Trendline / Display equations & Display R-Squared value on the charts. The Y and Xs are renamed to Month, Adv and Sales, respectively, for the regression lines.

4. What are the values of b_0, b_1, and b_2, and what is the estimated regression function? The values of b_0, b_1, and b_2 are 120.7428, 135.8275 and -6.862, respectively as given in the Table above.
5. What are the meaning of b_0, b_1, and b_2? When in Month=0, and Adv=0, the Sales = b_0=\$120.7428, b_1=\$135.8275 shows the marginal change (increase or decrease) of \$135.8275 in Sales when the variable Month changes one unit (increase or decrease) while keep no change for Adv., b_2 = –6.862 shows the marginal change (decrease or increase) of –\$6.862 in Sales (Y) when the amount of the variable Adv\$ (X_2) incrementally changes (increases or decreases ONE unit) while keep no change for Month.
6. Use Excel@ =RSQ(Array Y,Array X) to compute the coefficient of Determination R^2 of the regression line for the data, and interpret the meaning of R^2 for the data?

	H	I	J	K
17	R^2 or R Square (Sales & Month)	0.945	=RSQ(C2:C11,A2:A11)	
18	R^2 or R Sqaure (Sales & Adv.)	0.8118	=RSQ(C2:C11,B2:B11)	

(Regression.xls/Reg1)

For the regression line for Month versus Sales, R^2 = 94.5% means 94.5% of the total variations in Sales are counted for or explained and 5.5% of the total variations are not counted for or not explained by the regression line between Month and Sales. For the regression line for Adv versus Sales, R^2 = 81.18% means 81.18% of the total variations in Sales are counted for or explained and 18.82%% of the total variations are not counted for or not explained by the regression line between the Adv and Sales.

7. For the Summary Table from Data/Data Analysis, answer the following questions:
   a. R^2, Adjusted R^2, Number of Observations, b_0, b_1, p-value for b_0, p-value for b_1. The values are as labeled in the above table from Regression.xls/Reg1SO.
   b. Use the p-value approach to test the population parameters β_0, β_1 and β_2 with the p-values from the Summary Output of Data Analysis/Regression, and state your conclusion. Assume the significance coefficient α = 0.05.

	A	B	C	D	E	F	G
16		*Coefficients*	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
17	Intercept	120.7428	59.5825	2.0265	0.0823	-20.1474	261.6330
18	Month	135.8275	25.2624	5.3767	0.0010	76.0914	195.5637
19	Adv (\$1000)	-6.8621	3.6823	-1.8635	0.1047	-15.5694	1.8453

(Regression.xls/Reg1SOa)

Hypothesis Test for β_0:

i. What are the H_0 and H_1? H_0: $\beta_0 = 0$ and H_1: $\beta_0 \neq 0$
ii. What are the decision rules?

Decision Rules with p-value Approach:

If p-value $\geq \alpha$ (significance coefficient), then conclude H_0 or $\beta_0 = 0$;

Otherwise, if p-value $< \alpha$, then conclude H_a, or $\beta_0 \neq 0$.

iii. What is the conclusion? The p-value for β_0 = 0.082 as given in the Summary Output, it is greater than α = 0.05, therefore, we conclude H_0: $\beta_0 = 0$ or fail to reject H_0, i.e., we should not include the Y intercept term in the regression model for Sales.

Hypothesis Test for β_1(Month):

i. What are the H_0 and H_1? H_0: $\beta_1 = 0$ and H_1: $\beta_1 \neq 0$

ii. What are the decision rules?

Decision Rules with p-value Approach:
If p-value $\geq \alpha$ (significance coefficient), then conclude H_0 or $\beta_1 = 0$;
Otherwise, if p-value $< \alpha$, then conclude H_a, or $\beta_1 \neq 0$.

iii. What is the conclusion? The p-value for $\beta_1 = 0.001$ as given in the Summary Output, it is less than $\alpha = 0.05$, therefore, we conclude H_1: $\beta_1 \neq 0$ or reject reject H_0, i.e., we should include the variable Month in the regression model for Sales.

Hypothesis Test for β_2(Adv):

i. What are the H_0 and H_1? H_0: $\beta_2 = 0$ and H_1: $\beta_2 \neq 0$

ii. What are the decision rules?

Decision Rules with p-value Approach:
If p-value $\geq \alpha$ (significance coefficient), then conclude H_0 or $\beta_2 = 0$;
Otherwise, if p-value $< \alpha$, then conclude H_a, or $\beta_2 \neq 0$.

iii. What is the conclusion? The p-value for $\beta_2 = 0.105$ as given in the Summary Output, it is greater than $\alpha = 0.05$, therefore, we conclude H_0: $\beta_0 = 0$ or fail to reject H_0, i.e., we should not include the variable Adv in the regression model for Sales.

Therefore the final regression model for Sales becomes: Sales = b_1 (Month). We have to go through additional procedures below to find out the value of b_1 when the Y intercept b_0 is zero.

For reference:
Decision Rules with Confidence Interval Approach:
If the given CI spans zero (with zero as part of CI), conclude H_0
Otherwise, if the given CI does not span zero, then conclude H_a

8. What are the forecasts for the next two years (11 to 12) with the regression line? Because the hypothesis tests reveal only β_1 is significant and should be included in the model, we further run the models with Adv only and Month only with the results in the following two tables.

	A	B	C	D	E	F	G
16		*Coefficients*	*Standard*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
17	Intercept	41.6	47.81414	0.870036	0.409626	-68.6596	151.8596
18	Month	90.4545455	7.705946	11.73828	2.54E-06	72.6846	108.2245

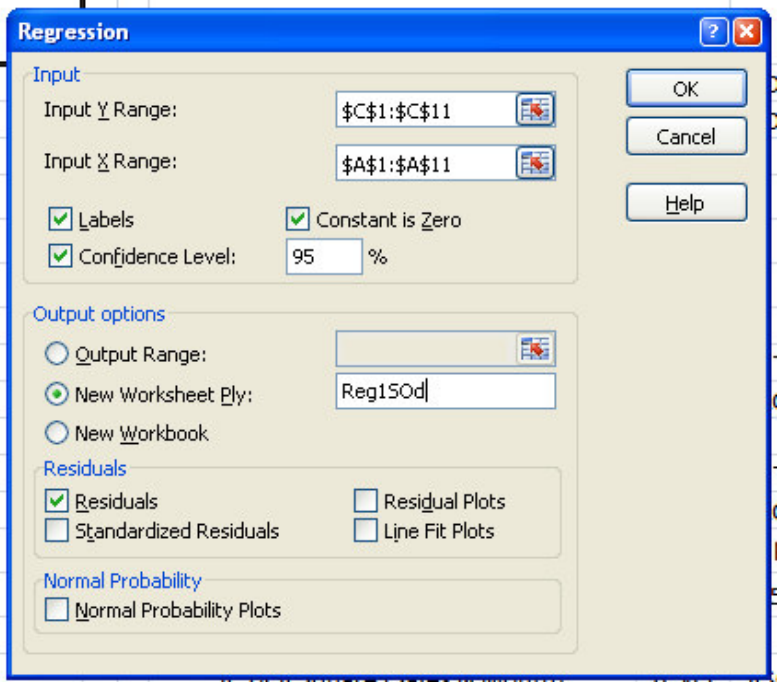
(Regression.xls/Reg1SOb)

	A	B	C	D	E	F	G
16		*Coefficients*	*Standard Err*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
17	Intercept	-46.2387648	107.728473	-0.42922	0.679091	-294.661	202.1835
18	Adv ($100	12.22001597	2.07990955	5.875263	0.000372	7.423736	17.0163

(Regression.xls/Reg1SOc)

The results reveal that the Y intercept terms on both cases are not significant, thus should not be included in the model, the variables Month and Adv, each is significant to model the Sales by itself. We therefore decide to use Month only as recommended in the procedure 7.b above. To find out the value of b_1 without b_0 with the variable Month only, we need to rerun the Data/Data Analysis/Regression with the option of Constant is Zero as given below.

	A	B	C	D
1	Month	Adv ($1000)	Sales ($1000)	Est Sales ($1000)
2	1	20	112	350.000
3	2	20	259	400.000
4	3	23	323	480.000
5	4	48	353	780.000
6	5	51	411	860.000
7	6	52	594	920.000
8	7	54	788	990.000
9	8	61	846	1110.000
10	9	74	851	1290.000
11	10	76	854	1360.000
12	11			
13	12			
14				
15				
16	SSE			1150637



(Regression.xls/Reg1)

The final Summary Output Table is given in Regression.xls/Reg1SOd)

	A	B	C	D	E	F	G
16		*Coefficients*	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
17	Intercept	0	#N/A	#N/A	#N/A	#N/A	#N/A
18	Month	96.3974	3.518665848	27.39601	5.58E-10	88.43763	104.3572

(Regression.xls/Reg1SOd)

a. Manually compute the forecasts with the b_0 ,b_1 and b_2 from the previous results

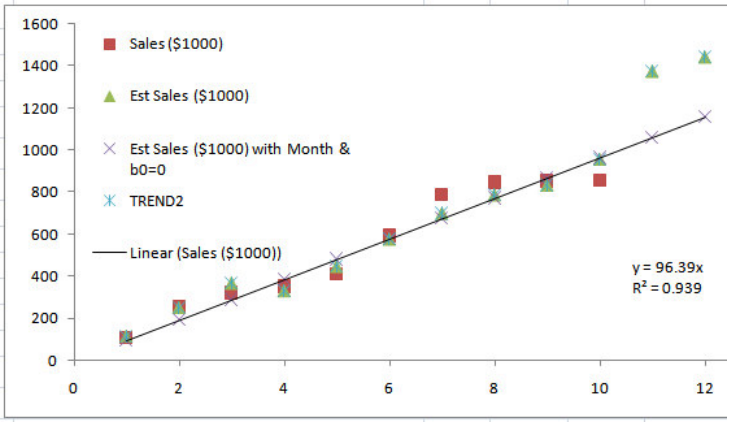
Sales (Month=11) = 96.3974 * 11 (Month) = $1060.37 in Excel@ =Reg1SOd!$B$18*'Reg1'!A12
Sales (Month=12) = 96.3974 * 12 (Month) = $1156.77 in Excel@ =Reg1SOd!$B$18*'Reg1'!A13

In this case, Excel@ = TREND() cannot be used to forecast future Sales. The following procedures are used to show how to use =TREND() to forecast Sales when b_0, b_1, and b_2 are all included in the model for Sales. We use TREND2 to represent the forecasts developed with =TREND() with both Month and Adv in the model for Sales. Assume the Adv = 35 for Month = 11, and Adv = 45 for Month = 12. Please note the use of Absolute Address in =TREND().

b. Use Excel@ =TREND(Y-RANGE,X-RANGE,X-VALUE)

F2 =TREND($C$2:$C$11,$A$2:$B$11,A2:B2)

	A	B	C	D	E	F
1	Month	Adv ($1000)	Sales ($1000)	Est Sales ($1000)	Est Sales ($1000) with Month & b_0=0	TREND2
2	1	20	112	119.330	96.397	119.328
3	2	20	259	255.157	192.795	255.156
4	3	23	323	370.397	289.192	370.397
5	4	48	353	334.674	385.590	334.672
6	5	51	411	449.915	481.987	449.914
7	6	52	594	578.879	578.384	578.879
8	7	54	788	700.982	674.782	700.983
9	8	61	846	788.775	771.179	788.775
10	9	74	851	835.395	867.577	835.396
11	10	76	854	957.498	963.974	957.499
12	11	35		1374.665	1060.371	1374.672
13	12	45		1441.872	1156.769	1441.879



(Regression.xls/Reg1)

c. What is the assumption you made when you develop forecasts for the next two years? The crucial assumption made for using linear regression is that the linear trend for Sales is going to continue in Months 11 and 12 with the $b_2 = 96.3974$. Thus any forecasts made outside out the original ranges of independent variable Xs in the historical data may not be valid.

9. What is the difference between standard error (S_e) and the standard prediction error (S_p)?

Standard Error of Estimate (S_{YX} *or* S_e):

$$S_{YX} = S_e = \sqrt{\frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{n-k-1}} = \sqrt{\frac{SSE}{n-k-1}} = \sqrt{MSE} = RMSE = 69.0412$$

S_e measures the variation of the actual data around the estimated regression line, where k is the number of independent variables in the model.

Standard prediction error (S_p): thus the S_p is always larger than S_e.

$$S_p = S_e\sqrt{1 + \frac{1}{n} + \frac{(X_{ih} - \bar{X})^2}{\sum_{i=1}^{n}(X_{ih} - \bar{X})^2}}$$

(1–α)% Prediction Interval for individual response Y:

$$\hat{Y}_{ih} = b_0 + b_1 X_{ih} \qquad \text{and} \qquad \hat{Y}_{ih} \pm t_{(1-\frac{\alpha}{2};n-2)} S_p$$

	A	B	C	D	E	F	G
1	SUMMARY OUTPUT						
2							
3	*Regression Statistics*						
4	Multiple R	0.9941					
5	R Square	0.9882					
6	Adjusted R Square	0.8770					
7	Standard Error	69.0412					
8	Observations	10					
9							
10	ANOVA						
11		*df*	*SS*	*MS*	*F*	*Significance F*	
12	Regression	1	3577596.80	##########	750.5413	0.0000	
13	Residual	9	42900.20	4766.69			
14	Total	10	3620497				
15							
16		*Coefficients*	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
17	Intercept	0	#N/A	#N/A	#N/A	#N/A	#N/A
18	Month	96.3974	3.5187	27.3960	0.0000	88.4376	104.3572
19							
20							
21							
22	RESIDUAL OUTPUT						
23							
24	*Observation*	*icted Sales ($1*	*Residuals*				
25	1	96.3974026	15.6025974				
26	2	192.7948052	66.20519481				

(Regression.xls/Reg1SOd)

The following results are in Regsssion.xls/Reg2.

10. What is the margin of error for an approximate 95% prediction interval individual response for Month=11? The margin of prediction error for Month=11 is 252.981
11. What is the 95% mean prediction interval for Month=11?

939.603 Lw Lmt of 95% Mean Pred CI for Month=11
1181.140 Up Lmt of 95% Mean Pred CI for Month=11

12. What is the 95% prediction intervals individual response for the Month = 11?

807.391 Lw Lmt of 95% Pred CI for Month=11
1313.352 Up Lmt of 95% Pred CI for Month=11

867.577 Appro Lw Lmt of 95% Pred CI for Month = 11
1253.17 Appro Up Lmt of 95% Pred CI for Month = 11

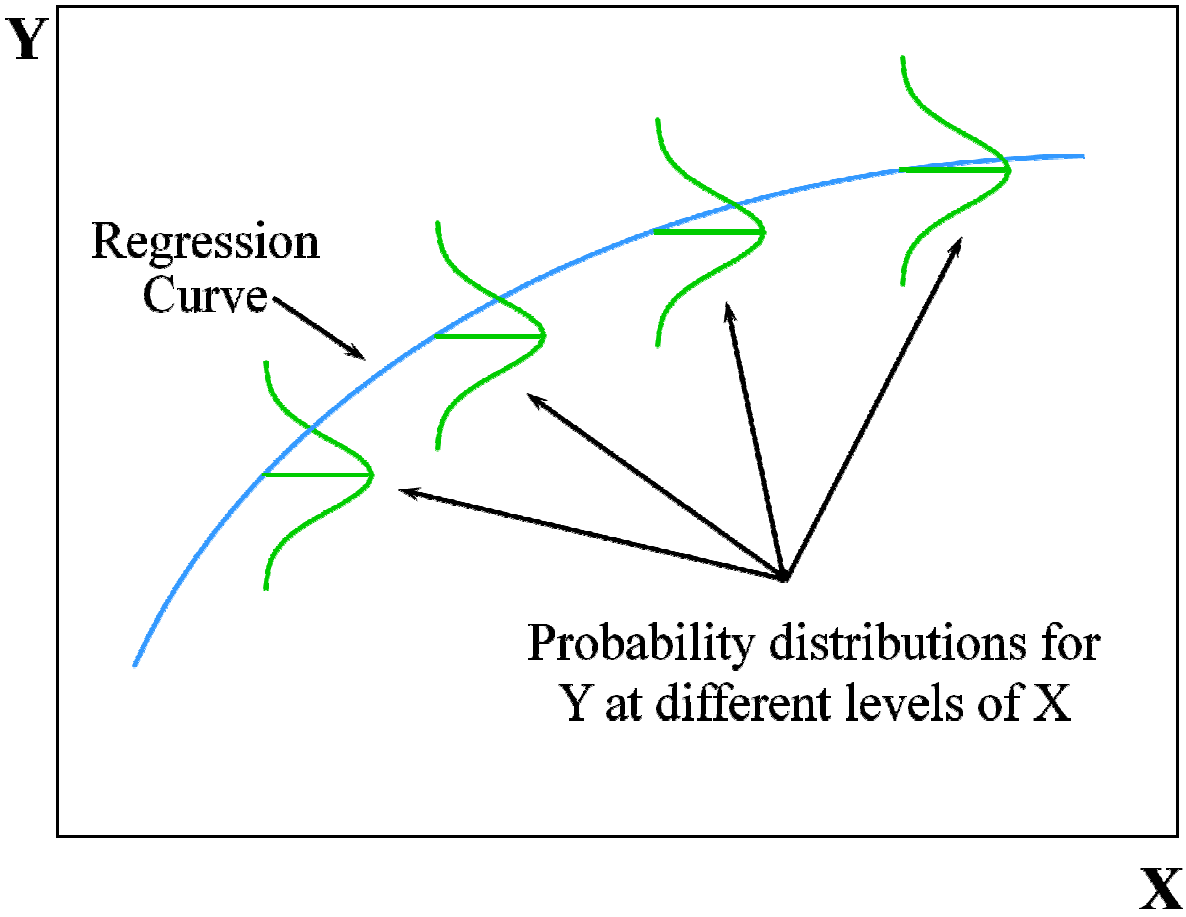
	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	Month	dv ($1000)	Sales ($1000)		Est Sales ($1000) with Month & b_0=0		(Xih-Xbar)^2	Std Pred. Error Sp	Margin of Pred Error = t * Sp	Std Mean Error Sa	Margin of Mean Error = t Sa				
2	1	20	112		96.397		20.3	106.938	246.599	46.295	106.757	SteYx or Se (Month only & b0=0)	96.3974	=Reg1SOd!$B$18	
3	2	20	259		192.795		12.3	104.671	241.373	40.788	94.057	Sample size n	10	=COUNT(C2:C11)	
4	3	23	323		289.192		6.3	102.939	237.377	36.110	83.270				
5	4	48	353		385.590		2.3	101.767	234.676	32.621	75.224	t((1-α/2),n-k-1) two tailed & k=1	2.306		
6	5	51	411		481.987		0.3	101.177	233.314	30.728	70.860				
7	6	52	594		578.384		0.3	101.177	233.314	30.728	70.860	Approximated t value to be used	2		
8	7	54	788		674.782		2.3	101.767	234.676	32.621	75.224				
9	8	61	846		771.179		6.3	102.939	237.377	36.110	83.270				
10	9	74	851		867.577		12.3	104.671	241.373	40.788	94.057				
11	10	76	854		963.974		20.3	106.938	246.599	46.295	106.757				
12	11	35			1060.371		30.3	109.705	252.981	52.372	120.769				
13	12	45			1156.769		42.3	112.936	260.432	58.840	135.686				
14															
15					Average X		5.5		807.391			Lw Lmt of 95% Pred CI for Month=11			
16	SSE				SSEX		155		1313.352			Up Lmt of 95% Pred CI for Month=11			
17															
18											939.603	Lw Lmt of 95% Mean Pred CI for Month=11			
19											1181.140	Up Lmt of 95% Mean Pred CI for Month=11			
20															
21									867.577			Appro Lw Lmt of 95% Pred CI for Month = 11			
22									1253.17			Appro Up Lmt of 95% Pred CI for Month = 11			

(Regression.xls/Reg2)

Topics to be covered:

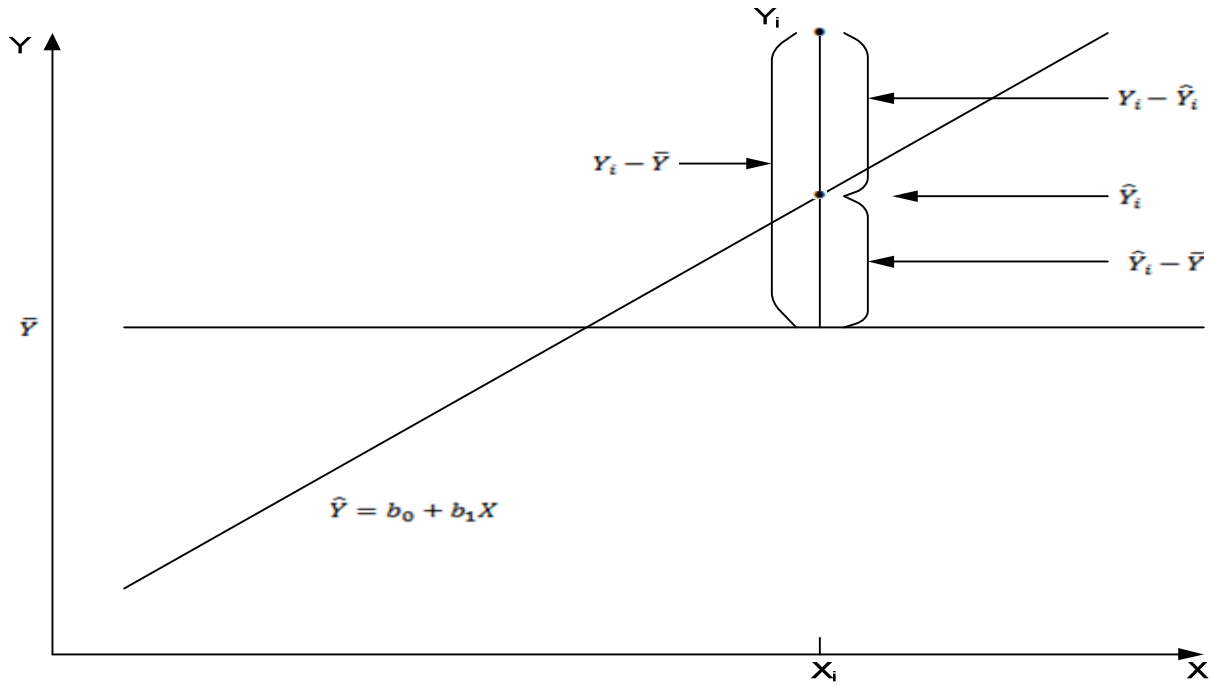
1. Regression as a method in business analytics $Y = f(X_1, X_2, \cdots, X_k) + \varepsilon$
   a. Simple Linear Regression (SLP)
      $Y = f(X) + \varepsilon$ or $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$ and b_0 and $b_1$as estimates for $\beta_0, and\ \beta_1$
      $\hat{Y}_i = b_0 + b_1 X_i$ to min
      $ESS = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n} [Y_i - (b_0 + b_1 X_i)]^2$ and the method of least squares
   b. Simple Linear Regression with Time as Independent Variable
      $Y = f(T) + \epsilon$ or $Y_i = \beta_0 + \beta_1 T_i + \varepsilon_i$
   c. Multiple Regression $Y = f(X_1, X_2, \cdots, X_k) + \varepsilon$, where f(.) describes systematic variations and ε describers unsystematic variations of the system.

2. Assumptions for Regression



3. Using Excel@ to do Regression analysis

$\hat{Y} = b_0 + b_1 X$ $\bar{Y}$ • X_i $Y_i - \bar{Y}$ $Y_i - \hat{Y}_i$ $\hat{Y}_i$ $\hat{Y}_i - \bar{Y}$



Decomposition of the Total Error: $Y_i - \bar{Y} = (Y_i - \hat{Y}_i) + (\hat{Y}_i - \bar{Y})$

$$\sum_{i=1}^{n}(Y_i - \bar{Y})^2 = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 + \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2$$

TSS = ESS + RSS

Or Total Sum of Squared Errors (TSS) = Error Sum of Squares (ESS) + Regression Sum of Squares (RSS)

$R^2 = \frac{RSS}{TSS} = 1 - \frac{ESS}{TSS}$, and $0 \le R^2 \le 1$

R^2 refers to the percentage of the total variation of Y around its mean that is explained or counted for by the estimated regression line or how well the regression line fits the data.

1-R^2 is the percentage of the total variation of Y around its mean that is unexplained or uncounted for by the regression line.

Equations to compute b_0 and b_1:

$$b_1 = \frac{SSXY}{SSX} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2} = \frac{\sum_{i=1}^{n} X_i Y_i - \frac{[\sum_{i=1}^{n} X_i][\sum_{i=1}^{n} Y_i]}{n}}{\sum_{i=1}^{n} X_i^2 - \frac{[\sum_{i=1}^{n} X_i]^2}{n}}$$

$$b_0 = \bar{Y} - b_1 \bar{X}$$

Standard error of estimate versus Standard prediction error:

Standard Error of Estimate (S_{YX} *or* S_e), the Rule of Thumb and Standard Prediction Error (S_p):

$$S_{YX} = S_e = \sqrt{\frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{n-k-1}} = \sqrt{\frac{SSE}{n-k-1}} = \sqrt{MSE} = RMSE$$

Se measures the variation of the actual data around the estimated regression line, where k is the number of independent variables in the model.
The rule of thumb: 68%, 95% and 99.7% of the observations are within ±1 S_e, ±2 S_e and ±3 S_e or the approximate 95% Confidence Interval for individual $\hat{Y}_i$ can be roughly estimated as $\hat{Y}_i \pm 2\ S_e$

Standard prediction error (S_p): thus the S_p is always larger than S_e.

$$S_p = S_e\sqrt{1 + \frac{1}{n} + \frac{(X_{ih} - \bar{X})^2}{\sum_{i=1}^{n}(X_{ih} - \bar{X})^2}}$$

(1–α)% Prediction Interval for individual response Y:

$\hat{Y}_{ih} = b_0 + b_1 X_{ih}$ and $\hat{Y}_{ih} \pm t_{(1-\frac{\alpha}{2};n-2)} S_p$

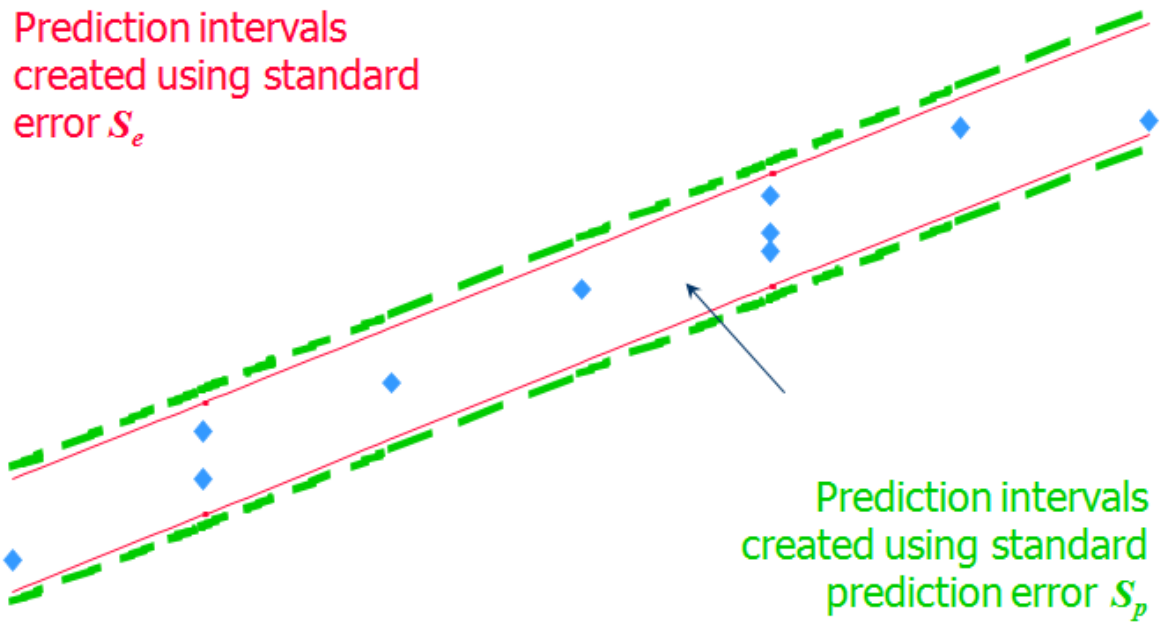
For α = 0.10, 1–α/2 = 0.95, the $t_{(1-\frac{\alpha}{2};n-2)}$ = TINV(1-α,n-2)=TINV(0.10,n-2). Please note, TINV assumes two tailed case.

(1–α)% Confidence Interval for the mean of Y at the given point of X:

$$\hat{Y}_{ih} \pm t_{(1-\frac{\alpha}{2};n-2)} S_e\sqrt{\frac{1}{n} + \frac{(X_{ih} - \bar{X})^2}{\sum_{i=1}^{n}(X_{ih} - \bar{X})^2}}$$

For multiple Regressions:

$$R_d^2 = 1 - \left(\frac{ESS}{TSS}\right)\left(\frac{n-1}{n-k-1}\right)$$



Statistical Tests for $\beta_0, \beta_1, \cdots, \beta_k$ with F-Statistic, p-value and t-statistic:

- How to get the regression line in Excel@?
  - =INTERCEPT(Y-RANGE,X-RANGE) and =SLOPE(Y-RANGE,X-RANGE)
  - =TREND(Y-RANGE,X-RANGE,X-VALUE)
  - In Excel@ Data/Data Analysis/Regression
  - In Excel@, insert/scatter plot, Click any data points in the scatter plot, select Add Trendline / Display equations & Display R-Squared value on chart
  - Use Excel Solver to Minimize ESS

What are the meanings of the slope($\beta_1, \ldots \beta_n$) and the intercept(β_0)? Meanings of b_0 and b_1:

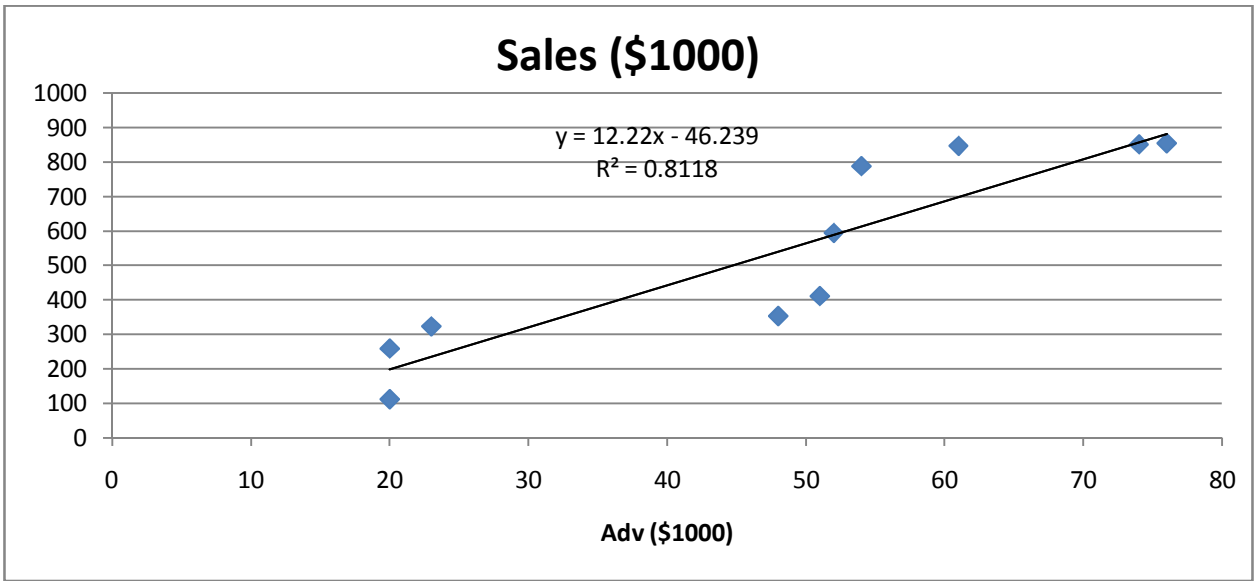
- b_0 is the intercept or Y value as X = 0 (e.g. Fixed cost)
- b_1 is the slope or marginal change in Y with unit change in X
- b_1 is similar to the slope. However, since it is calculated with the variability of the data in mind, its formulation is not as straight-forward as our usual notion of slope.

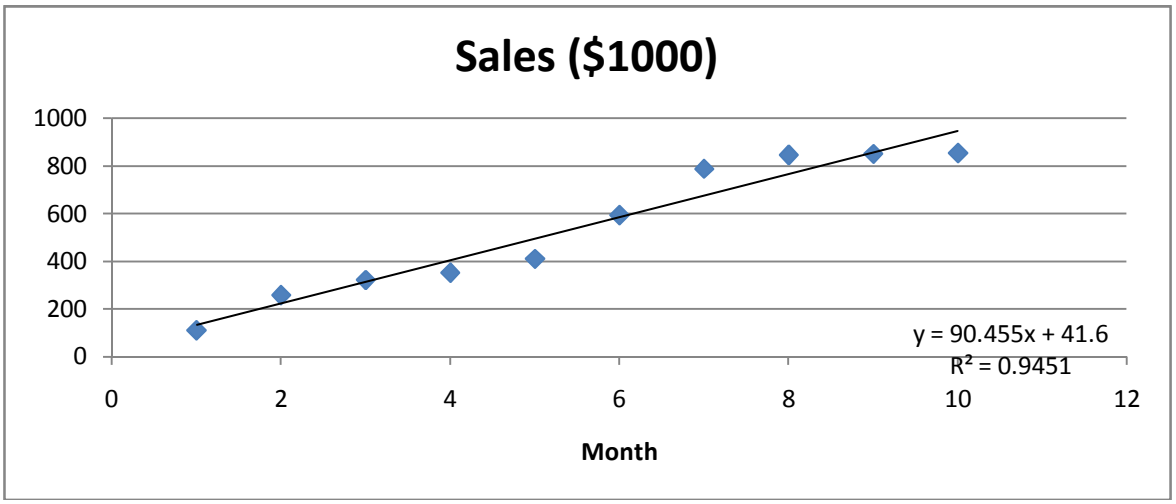
- How to statistically test whither $\beta_0, \beta_1, \ldots \beta_n$ equals zero (in the Null Hypothesis H_0)?
  - Use the p-value approach because p-values will be in Summary Output. Decision Rule is: if p-value is less than α value (Type I error, not the one in Exponential smoothing forecasting), then conclude $\beta_0, \beta_1, \ldots \beta_n$ not equal to zero.
  - Only include $b_0, b_1, \ldots, b_n$ in the equation if $\beta_0, \beta_1, \ldots \beta_n$ not equal to zero. Exclude the zero ones
- How to interpret R^2? Percentage of total variations explained by the regression line. (1-R2) is the percentage of total variations not explained by the regression line.
- How to develop forecasts with given $b_0, b_1, \ldots, b_n$?
  - $\hat{Y}_{ih} = b_0 + b_1 X_{ih} + b_2 X_{iv}$ for each value of X and could be $X_1, X_2, \ldots$, and X_n.

- Predictions made using an estimated regression function may have little or no validity for values of the independent variables that are substantially different from those represented in the sample.
- Avoid multicollinearity when more than one independent variables are in the model.

- How to estimate the prediction confidence intervals?
  - The approximate prediction confidence interval, $\hat{Y}_i \pm 2\ S_e$, where S_e is the standard error given in the Summary Output beneath the Adjusted R^2.
  - The prediction confidence interval for the mean response is given by $\hat{Y}_i \pm t\ S_e const$
  - The prediction confidence interval for individual response is given by $\hat{Y}_i \pm t\ S_p$, where S_p is the standard prediction error and S_p is always larger than S_e.
- How to develop forecasting with Trend, Seasonal and Random components with the multiplicative model Y = T.S.I?
  - Use regression or centered moving average to take out the trend component – deTrend
  - Use S.I = Y/T to get the percentage of Trend
  - Compute Seasonal Relatives by average the percentages of Trend in the same season
  - Adjust the Seasonal Relatives to the whole number (Quarterly seasonality in a year will be 4, Weekly seasonality will be 7, etc.) to get the Seasonal Index.
  - Attached the Seasonal Index to each season
  - Develop the trend forecast with regression or other methods
  - Use the Seasonal Index to multiply the previous forecasts to develop the final seasonally adjusted forecast.

	A	B	C	D	E	F	G	H	I
1			Adv only	Equations				Month only	
2		Intercept =	-46.239	-46.239	=INTERCEPT(C6:C15,B6:B15) for Adv and Sales			41.600	
3		Slope =	12.220	12.220	=SLOPE(C6:C15,B6:B15) for Adv and Sales			90.455	
4				Use Adv only	se Adv on	Use Adv only	Use Adv only	Use Month only	Use Month only
5	Month	Adv ($1000)	Sales ($1000)	Est Sales ($1000)	Error	Squared Error	Est Sales ($1000)	Est Sales ($1000)	Est Sales ($1000)
6	1	20	112	198.161	-86.161	7423.798	198.162	132.055	132.055
7	2	20	259	198.161	60.839	3701.328	198.162	222.509	222.509
8	3	23	323	234.821	88.179	7775.449	234.822	312.964	312.964
9	4	48	353	540.322	-187.322	35089.449	540.322	403.418	403.418
10	5	51	411	576.982	-165.982	27549.962	576.982	493.873	493.873
11	6	52	594	589.202	4.798	23.022	589.202	584.327	584.327
12	7	54	788	613.642	174.358	30400.765	613.642	674.782	674.782
13	8	61	846	699.182	146.818	21555.547	699.182	765.236	765.236
14	9	74	851	858.042	-7.042	49.591	858.042	855.691	855.691
15	10	76	854	882.482	-28.482	811.230	882.482	946.145	946.145
16									
17		"=TREND($C$6:$C$15,$B$6:$B$15,B6)"			ESS=	134380.140			
18									





Simple Linear Regression with Time as Independent Variable

$Y = f(T) + \epsilon$ or $Y_i = \beta_0 + \beta_1 T_i + \varepsilon_i$ and $\hat{Y}_i = b_0 + b_1 t_i$

	J	K	L	M	N	O	P	Q	R
63	SUMMARY OUTPUT								
64									
65	*Regression Statistics*								
66	Multiple R	0.972							
67	R Square	0.945							
68	Adjusted R Square	0.938							
69	Standard Error	69.993							
70	Observations	10							
71									
72	ANOVA								
73		*df*	*SS*	*MS*	*F*	*Significance F*			
74	Regression	1	675017.0455	675017.045	137.79	2.53573E-06			
75	Residual	8	39191.85455	4898.98182					
76	Total	9	714208.9						
77									
78		*Coefficients*	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
79	Intercept	41.600	47.814	0.870	0.410	-68.660	151.860	-68.660	151.860
80	Month	90.455	7.706	11.738	0.000	72.685	108.224	72.685	108.224
81									

	J	K	L	M	N	O	P	Q	R
83	SUMMARY OUTPUT								
84									
85	*Regression Statistics*								
86	Multiple R	0.901025843							
87	R Square	0.81184757							
88	Adjusted R Square	0.788328516							
89	Standard Error	129.6052373							
90	Observations	10							
91									
92	ANOVA								
93		*df*	*SS*	*MS*	*F*	*Significance F*			
94	Regression	1	579828.7596	579828.76	34.519	0.000371987			
95	Residual	8	134380.1404	16797.5175					
96	Total	9	714208.9						
97									
98		*Coefficients*	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
99	Intercept	-46.2387648	107.7284729	-0.4292158	0.6791	-294.6610685	202.183539	-294.661069	202.1835388
100	Adv ($1000)	12.22001597	2.079909549	5.87526317	0.0004	7.423735951	17.016296	7.423735951	17.01629598

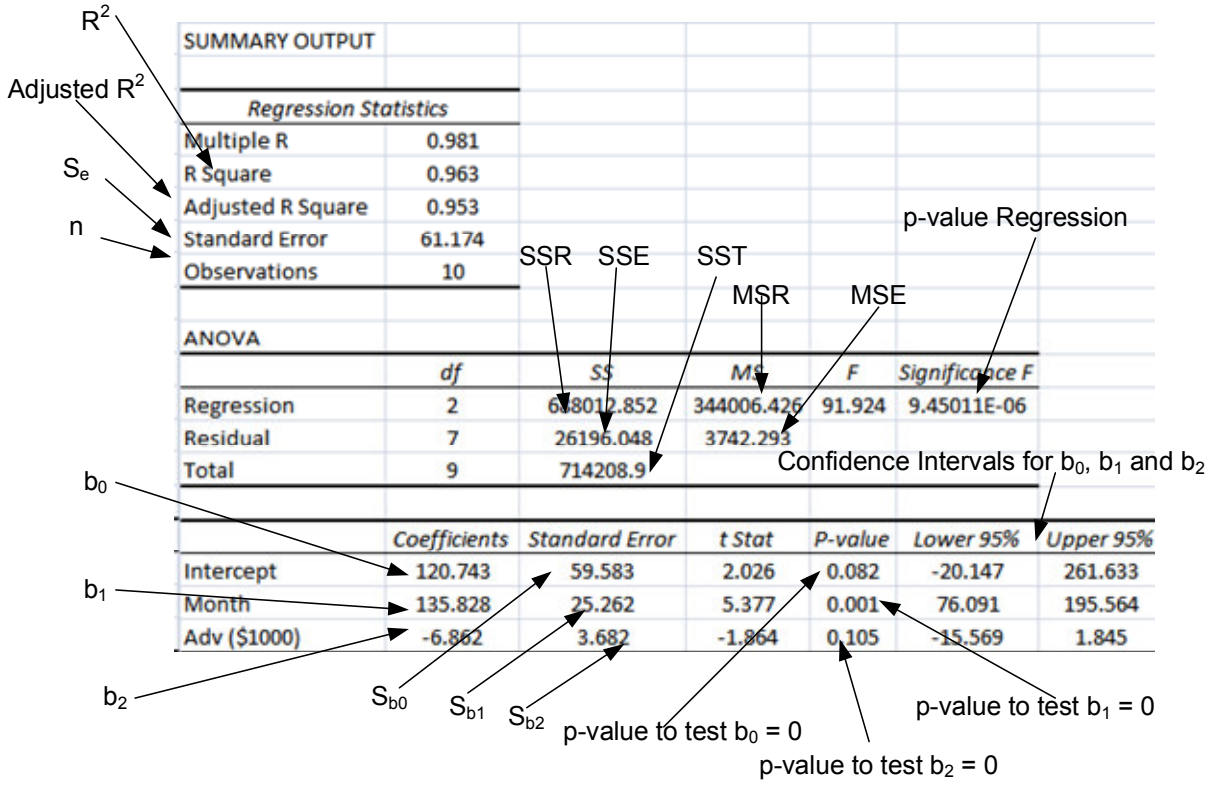
	A	B	C	D	E	F	G
19		Adv and Month					
20		Intercept =	120.739		ESS=	26196.048	
21		Slope (Adv) =	-6.862				
22		Slope (Month) =	135.826				
23	Month	Adv ($1000)	Sales ($1000)	Est Sales ($1000)	Error	Squared Error	Est Sales ($1000)
24	1	20	112	119.328	-7.328	53.706	119.328
25	2	20	259	255.154	3.846	14.789	255.156
26	3	23	323	370.395	-47.395	2246.257	370.397
27	4	48	353	334.675	18.325	335.815	334.672
28	5	51	411	449.915	-38.915	1514.388	449.914
29	6	52	594	578.879	15.121	228.639	578.879
30	7	54	788	700.981	87.019	7572.233	700.983
31	8	61	846	788.774	57.226	3274.760	788.775
32	9	74	851	835.397	15.603	243.468	835.396
33	10	76	854	957.499	-103.499	10711.994	957.499
34							
35		TREND($C$24:$C$33,$A$24:$B$33,A24:B24)					

	J	K	L	M	N	O	P	Q	R
26	SUMMARY OUTPUT								
27									
28	*Regression Statistics*								
29	Multiple R	0.981							
30	R Square	0.963							
31	Adjusted R Square	0.953							
32	Standard Error	61.174							
33	Observations	10							
34									
35	ANOVA								
36		*df*	*SS*	*MS*	*F*	*Significance F*			
37	Regression	2	688012.852	344006.426	91.924	9.45011E-06			
38	Residual	7	26196.048	3742.293					
39	Total	9	714208.9						
40									
41		*Coefficients*	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
42	Intercept	120.743	59.583	2.026	0.082	-20.147	261.633	-20.147	261.633
43	Month	135.828	25.262	5.377	0.001	76.091	195.564	76.091	195.564
44	Adv ($1000)	-6.862	3.682	-1.864	0.105	-15.569	1.845	-15.569	1.845

-

	A	B	C	D	E	F	G
1				Equations			
2		Intercept =	-46.2387648878136	=INTERCEPT(C6:C15,B6:B15)	ESS=	=SUM(F6:F15)	
3		Slope =	12.2200113449996	=SLOPE(C6:C15,B6:B15)			
4							
5	Obs	Adv ($1000)	Sales ($1000)	Est Sales ($1000)	Error	Squared Error	Est Sales ($1000)
6	1	20	112	=$C$2+$C$3*B6	=C6-D6	=E6*E6	=TREND($C$6:$C$15,$B$6:$B$15,B6)
7	2	20	259	=$C$2+$C$3*B7	=C7-D7	=E7*E7	=TREND($C$6:$C$15,$B$6:$B$15,B7)
8	3	23	323	=$C$2+$C$3*B8	=C8-D8	=E8*E8	=TREND($C$6:$C$15,$B$6:$B$15,B8)
9	4	48	353	=$C$2+$C$3*B9	=C9-D9	=E9*E9	=TREND($C$6:$C$15,$B$6:$B$15,B9)
10	5	51	411	=$C$2+$C$3*B10	=C10-D10	=E10*E10	=TREND($C$6:$C$15,$B$6:$B$15,B10)
11	6	52	594	=$C$2+$C$3*B11	=C11-D11	=E11*E11	=TREND($C$6:$C$15,$B$6:$B$15,B11)
12	7	54	788	=$C$2+$C$3*B12	=C12-D12	=E12*E12	=TREND($C$6:$C$15,$B$6:$B$15,B12)
13	8	61	846	=$C$2+$C$3*B13	=C13-D13	=E13*E13	=TREND($C$6:$C$15,$B$6:$B$15,B13)
14	9	74	851	=$C$2+$C$3*B14	=C14-D14	=E14*E14	=TREND($C$6:$C$15,$B$6:$B$15,B14)
15	10	76	854	=$C$2+$C$3*B15	=C15-D15	=E15*E15	=TREND($C$6:$C$15,$B$6:$B$15,B15)

Meanings of Regression Summary Output



SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.981
R Square	0.963
Adjusted R Square	0.953
Standard Error	61.174
Observations	10

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	2	688012.852	344006.426	91.924	9.45011E-06
Residual	7	26196.048	3742.293		
Total	9	714208.9			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	120.743	59.583	2.026	0.082	-20.147	261.633
Month	135.828	25.262	5.377	0.001	76.091	195.564
Adv ($1000)	-6.862	3.682	-1.864	0.105	-15.569	1.845

Decision Rules with p-value Approach:

If p-value $\geq \alpha$ (significance), then conclude H_0 or b_0, b_1 or b_2 is zero
Otherwise, if p-value $< \alpha$, then conclude H_a or b_0, b_1 or b_2 is not zero

Decision Rules with Confidence Interval Approach:

If the given CI spans zero (with zero as part of CI), conclude H_0
Otherwise, if the given CI does not span zero, then conclude H_a