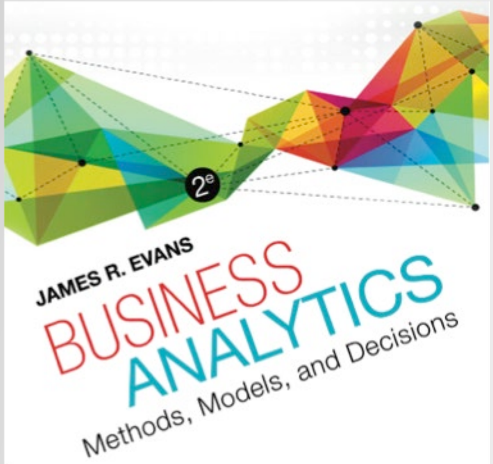


Chapter 8 Trendlines and Regression Analysis

Modeling Relationships and Trends in Data

- Create charts to better understand data sets.
- For cross-sectional data, use a scatter chart.
- For time series data, use a line chart.

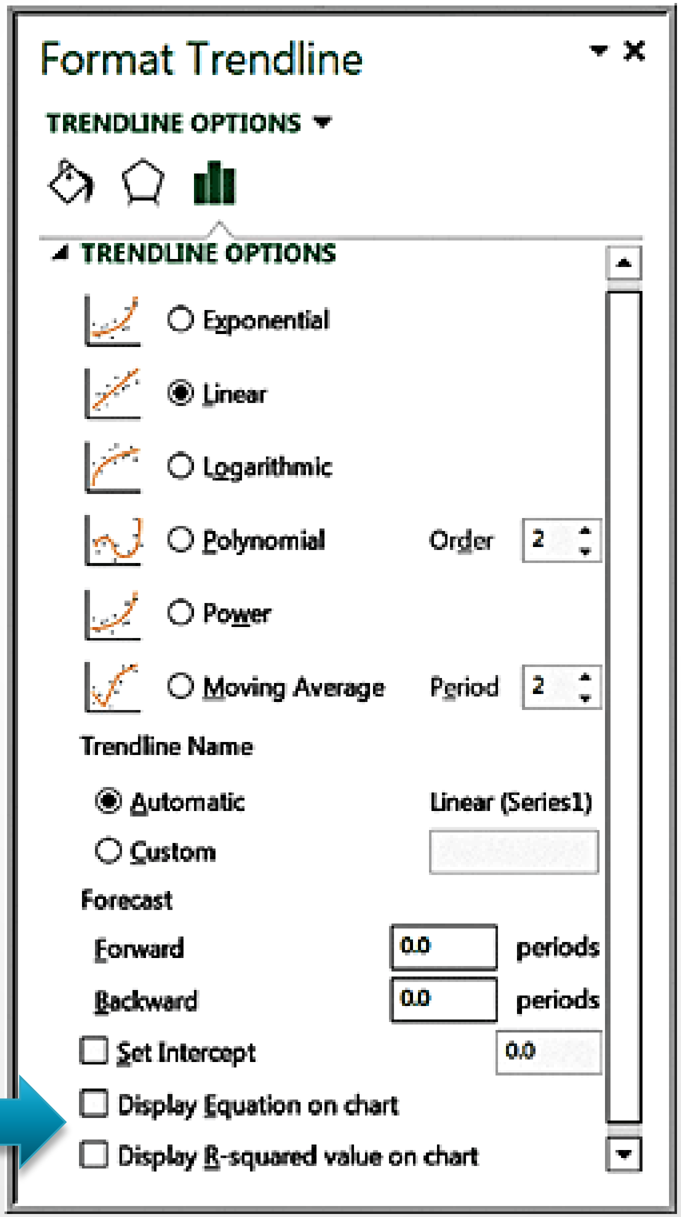
Common Mathematical Functions Used n Predictive Analytical Models

Linear	$y = a + bx$
Logarithmic	$y = ln(x)$
Polynomial (2nd order)	$y = ax^2 + bx + c$
Polynomial (3rd order)	$y = ax^3 + bx^2 + dx + e$
Power	$y = ax^b$
Exponential	$y = ab^x$

(the base of natural logarithms, *e* = 2.71828…is often used for the constant *b*)

Excel *Trendline* Tool

- Right click on data series and choose *Add trendline* from pop-up menu
- Check the boxes *Display Equation on chart* and *Display R-squared value on chart*



R^2

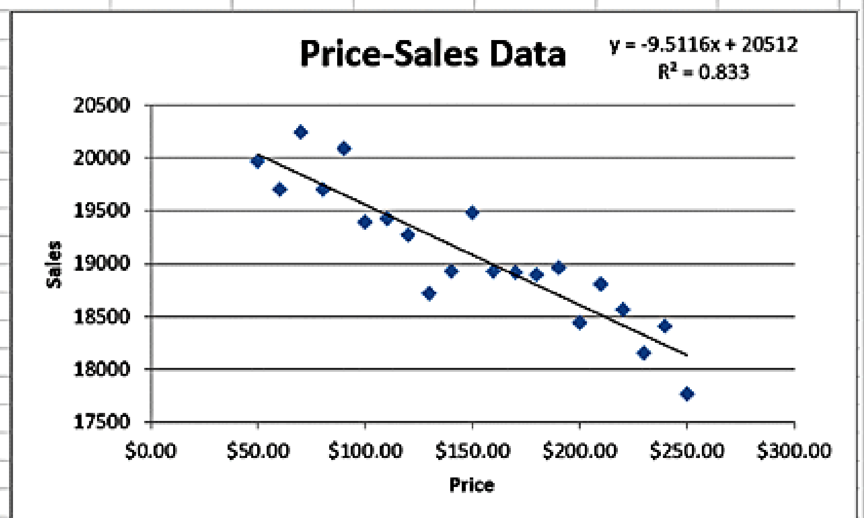
- ***R^2* (*R-squared*)** is a measure of the “fit” of the line to the data.
  - The value of R^2 will be between 0 and 1.
  - A value of 1.0 indicates a perfect fit and all data points would lie on the line; the larger the value of R^2 the better the fit.

Example 8.1: Modeling a Price-Demand Function

Linear demand function:
Sales = 20,512 - 9.5116(price)

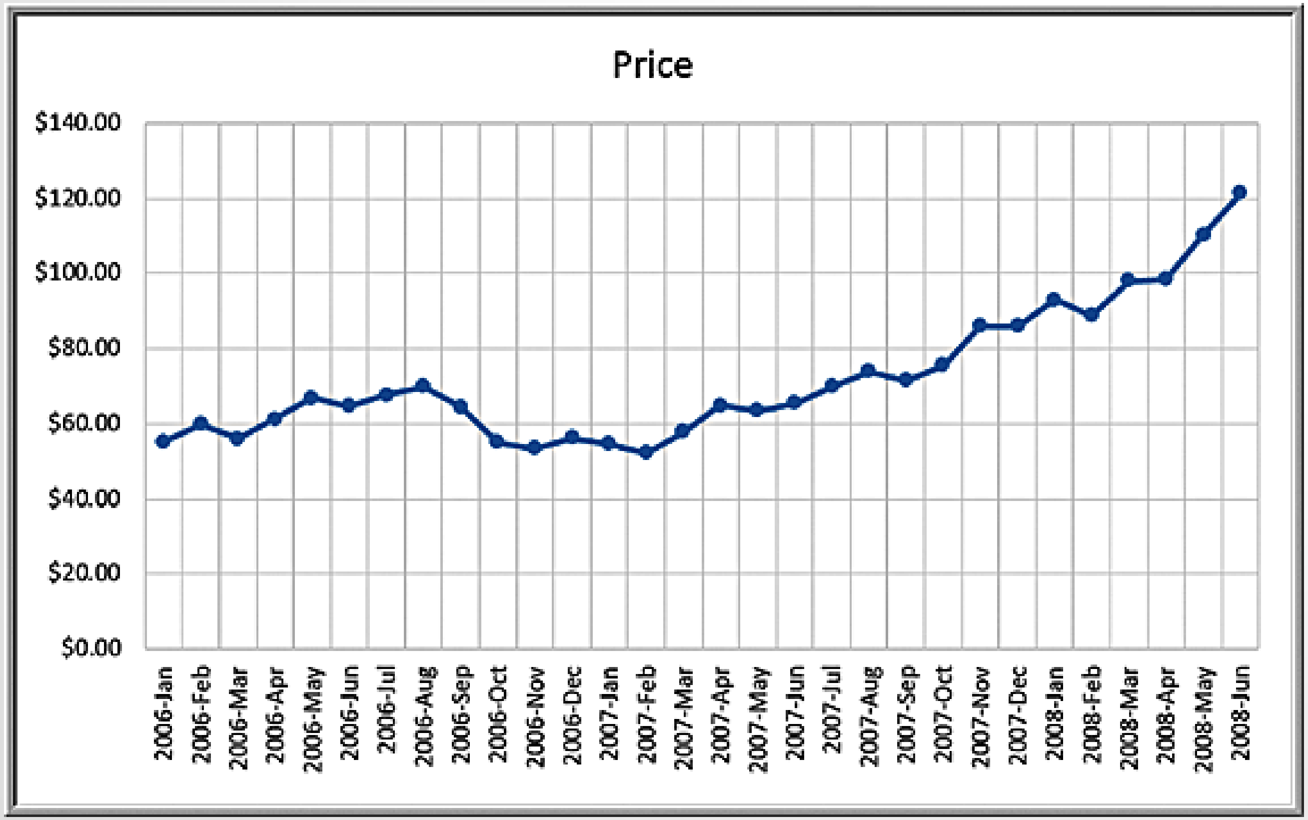
Price-Sales Data

Price	Demand
$50.00	19964.09
$60.00	19706.85
$70.00	20240.83
$80.00	19698.81
$90.00	20095.81
$100.00	19390.99
$110.00	19430.07
$120.00	19273.69
$130.00	18716.38
$140.00	18925.36
$150.00	19484.78
$160.00	18934.88
$170.00	18915.77
$180.00	18893.37
$190.00	18961.62
$200.00	18443.29
$210.00	18811.98
$220.00	18561.92
$230.00	18158.62
$240.00	18412.56
$250.00	17771.39



Example 8.2: Predicting Crude Oil Prices

- Line chart of historical crude oil prices



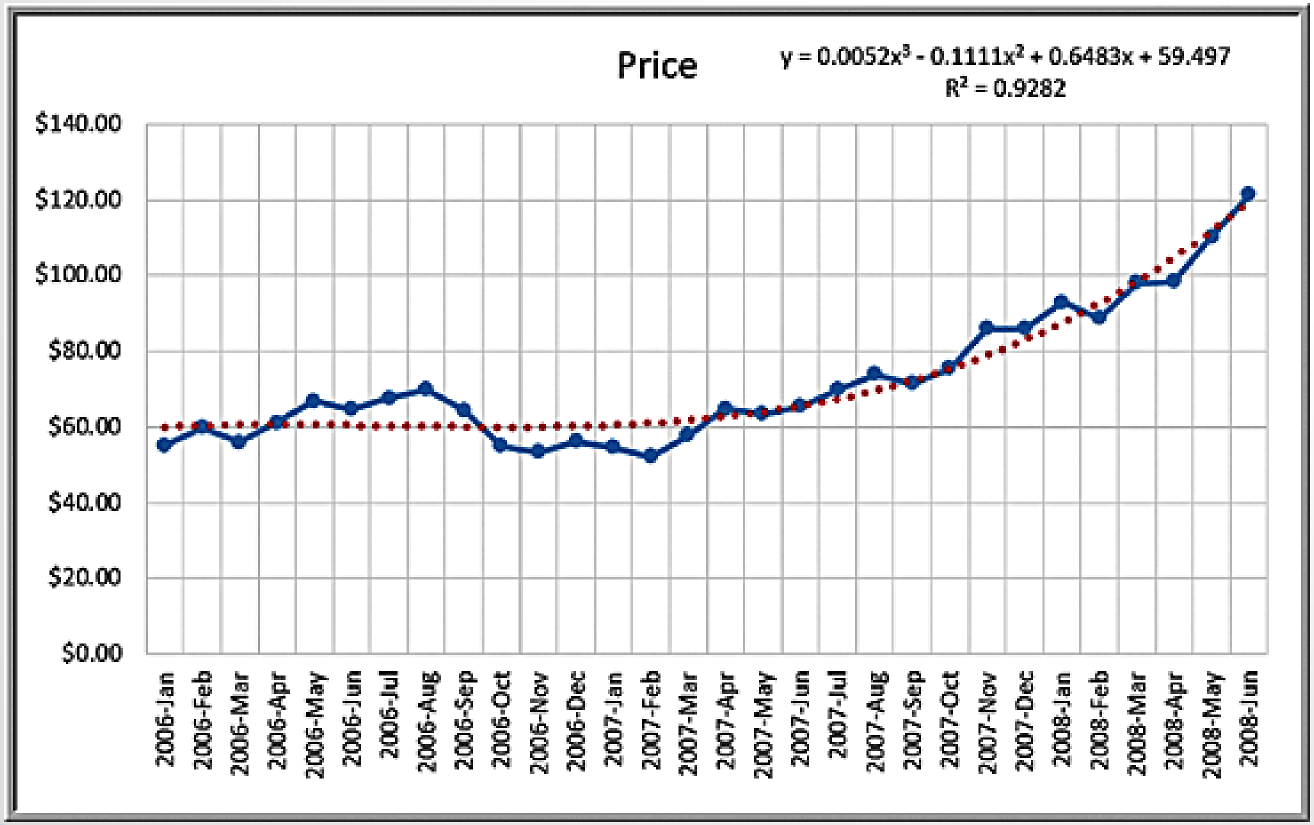
Example 8.9 Continued

- Excel's *Trendline* tool is used to fit various functions to the data.

Exponential	$y = 50.49e^{0.021x}$	$R^2 = 0.664$
Logarithmic	$y = 13.02\ln(x) + 39.60$	$R^2 = 0.382$
Polynomial 2°	$y = 0.13x^2 - 2.399x + 68.01$	$R^2 = 0.905$
Polynomial 3°	$y = 0.005x^3 - 0.111x^2 + 0.648x + 59.497$	$R^2 = 0.928$ *
Power	$y = 45.96x^{0.0169}$	$R^2 = 0.397$

Example 8.2 Continued

- Third order polynomial trendline fit to the data



Caution About Polynomials

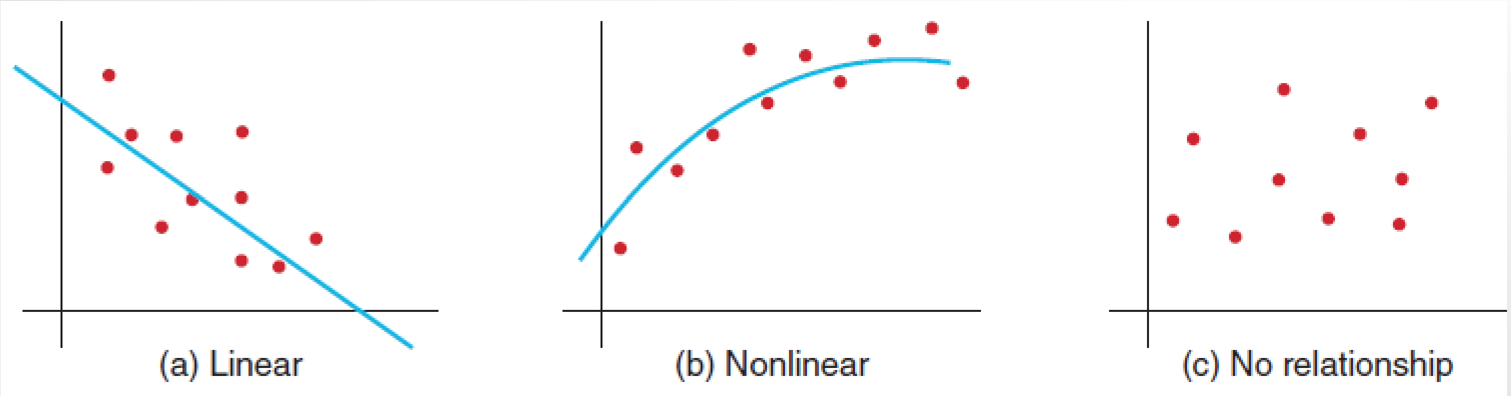
- The R^2 value will continue to increase as the order of the polynomial increases; that is, a 4th order polynomial will provide a better fit than a 3rd order, and so on.
- Higher order polynomials will generally not be very smooth and will be difficult to interpret visually.
    - Thus, we don't recommend going beyond a third-order polynomial when fitting data.
- Use your eye to make a good judgment!

Regression Analysis

- **Regression analysis** is a tool for building mathematical and statistical models that characterize relationships between a dependent (ratio) variable and one or more independent, or explanatory variables (ratio or categorical), all of which are numerical.
- **Simple linear regression** involves a single independent variable.
- **Multiple regression** involves two or more independent variables.

Simple Linear Regression

- Finds a linear relationship between:
  - one independent variable X and
  - one dependent variable Y
- First prepare a scatter plot to verify the data has a linear trend.
- Use alternative approaches if the data is not linear.



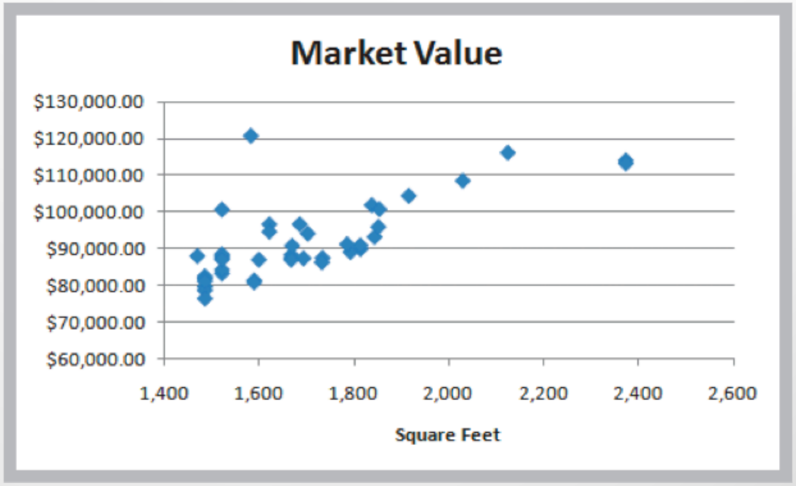
Example 8.3: *Home Market Value* Data

Size of a house is typically related to its market value.

X = square footage
Y = market value ($)

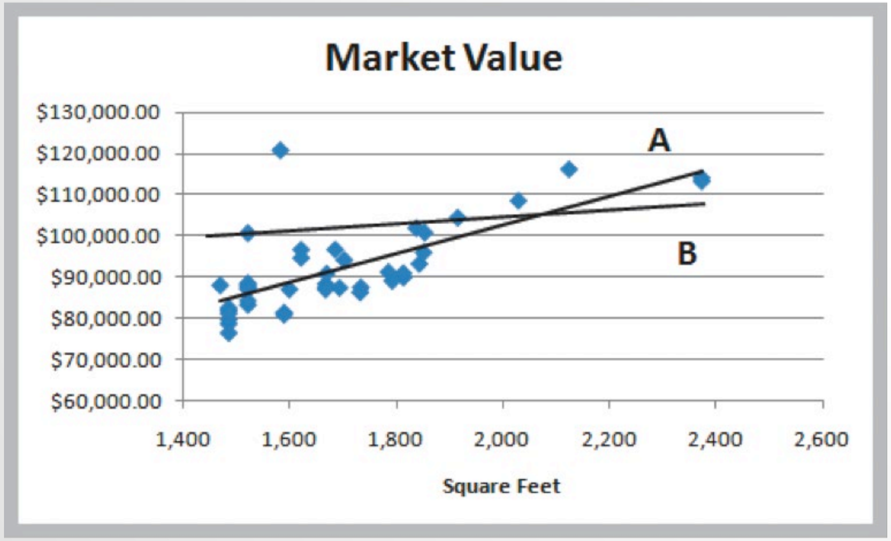
The scatter plot of the full data set (42 homes) indicates a linear trend.

	A	B	C
1	Home Market Value		
2			
3	House Age	Square Feet	Market Value
4	33	1,812	$90,000.00
5	32	1,914	$104,400.00
6	32	1,842	$93,300.00
7	33	1,812	$91,000.00
8	32	1,836	$101,900.00
9	33	2,028	$108,500.00
10	32	1,732	$87,600.00



Finding the Best-Fitting Regression Line

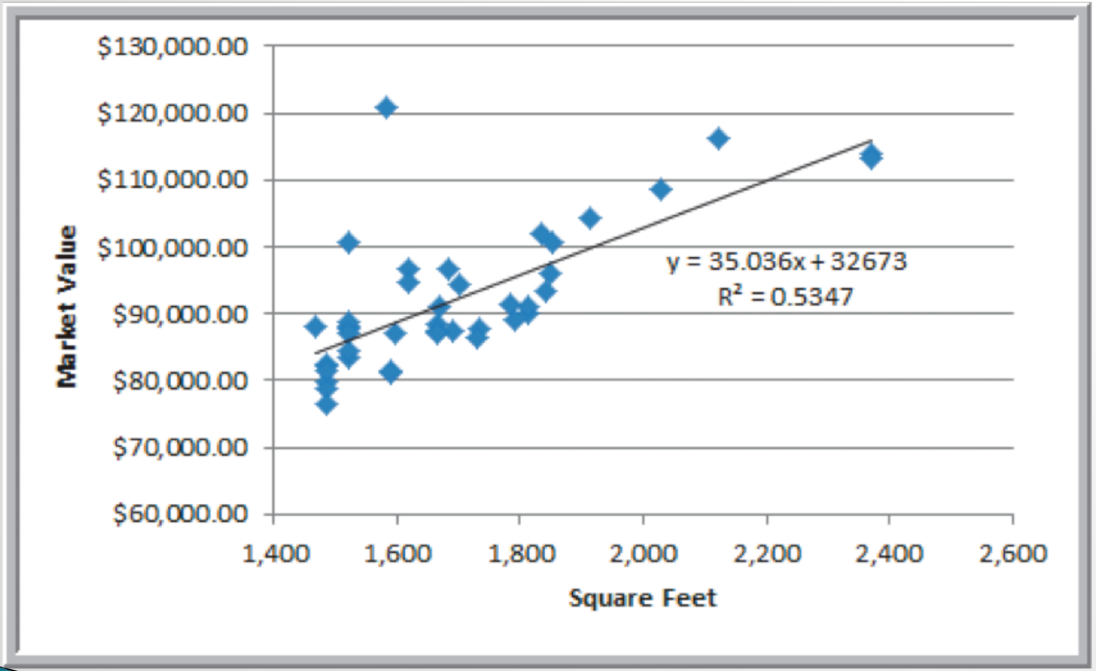
- Market value = $a + b \times$ square feet
- Two possible lines are shown below.



- Line A is clearly a better fit to the data.
- We want to determine the best regression line.

Example 8.4: Using Excel to Find the Best Regression Line

- Market value = 32,673 + $35.036 × square feet
  - The estimated market value of a home with 2,200 square feet would be: market value = $32,673 + $35.036 × 2,200 = $109,752



The regression model explains variation in market value due to size of the home.

It provides better estimates of market value than simply using the average.

Least-Squares Regression

- Simple linear regression model:

$$Y = \beta_0 + \beta_1 X + \varepsilon \tag{8.1}$$

- We estimate the parameters from the sample data:

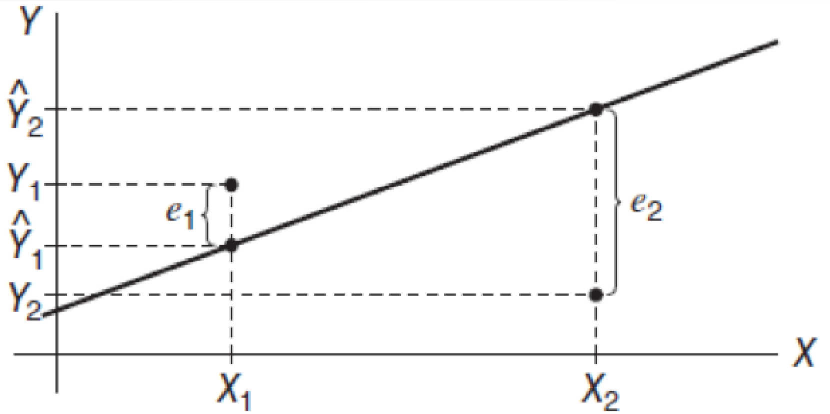
$$\hat{Y} = b_0 + b_1 X \tag{8.2}$$

- Let X_i be the value of the independent variable of the i^{th} observation. When the value of the independent variable is X_i, then $\hat{Y}_i = b_0 + b_1 X_i$ is the estimated value of Y for X_i.

Residuals

- Residuals are the observed errors associated with estimating the value of the dependent variable using the regression line:

$$e_i = Y_i - \hat{Y}_i \qquad (8.3)$$



Errors associated with individual observations

Least Squares Regression

- The best-fitting line minimizes the sum of squares of the residuals.

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n} (Y_i - [b_0 + b_1 X_i])^2 \qquad \textbf{(8.4)}$$

$$b_1 = \frac{\sum_{i=1}^{n} X_i Y_i - n\overline{X}\,\overline{Y}}{\sum_{i=1}^{n} X_i^2 - n\overline{X}^2} \qquad \textbf{(8.5)}$$

$$b_0 = \overline{Y} - b_1\overline{X} \qquad \textbf{(8.6)}$$

- Excel functions:
  - =INTERCEPT(*known_y's, known_x's*)
  - =SLOPE(*known_y's, known_x's*)

Example 8.5: Using Excel Functions to Find Least-Squares Coefficients

- Slope = b_1 = 35.036
  =SLOPE(C4:C45, B4:B45)
- Intercept = b_0 = 32,673
  =INTERCEPT(C4:C45, B4:B45)

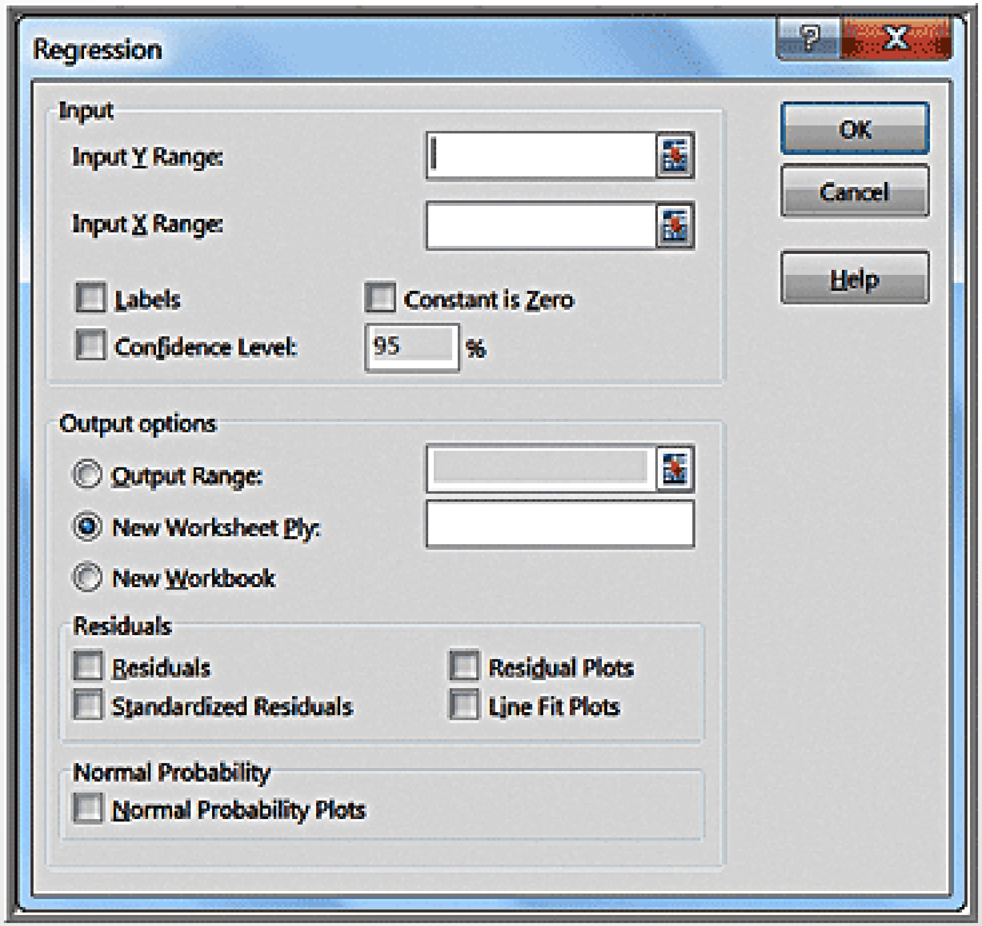
	A	B	C
1	**Home Market Value**		
2			
3	**House Age**	**Square Feet**	**Market Value**
4	33	1,812	$90,000.00
5	32	1,914	$104,400.00
6	32	1,842	$93,300.00
7	33	1,812	$91,000.00
8	32	1,836	$101,900.00
9	33	2,028	$108,500.00
10	32	1,732	$87,600.00

- Estimate *Y* when *X* = 1750 square feet
  $\hat{Y}$ = 32,673 + 35.036(1750) = $93,986
  =TREND(C4:C45, B4:B45, 1750)

Simple Linear Regression With Excel

Data > Data Analysis > Regression
Input Y Range (with header)
Input X Range (with header)
Check *Labels*

Excel outputs a table with many useful regression statistics.



Home Market Value Regression Results

	A	B	C	D	E	F	G
1	Regression Analysis						
2							
3	*Regression Statistics*						
4	Multiple R	0.731255223					
5	R Square	0.534734202					
6	Adjusted R Square	0.523102557					
7	Standard Error	7287.722712					
8	Observations	42					
9							
10	ANOVA						
11		*df*	*SS*	*MS*	*F*	*Significance F*	
12	Regression	1	2441633669	2441633669	45.97236277	3.79802E-08	
13	Residual	40	2124436093	53110902.32			
14	Total	41	4566069762				
15							
16		*Coefficients*	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
17	Intercept	32673.2199	8831.950745	3.699434116	0.000649604	14823.18178	50523.25802
18	Square Feet	35.03637258	5.16738385	6.780292234	3.79802E-08	24.59270036	45.48004481

Regression Statistics

- **Multiple R** - | *r* |, where *r* is the sample correlation coefficient. The value of *r* varies from -1 to +1 (*r* is negative if slope is negative)
- **R Square** - coefficient of determination, R^2, which varies from 0 (no fit) to 1 (perfect fit)
- **Adjusted R Square** - adjusts R^2 for sample size and number of *X* variables
- **Standard Error** - variability between observed and predicted *Y* values. This is formally called the **standard error of the estimate**, $\boldsymbol{S_{YX}}$.

Example 8.6: Interpreting Regression Statistics for Simple Linear Regression

	A	B	C	D	E	F	G
1	Regression Analysis						
2							
3	*Regression Statistics*						
4	Multiple R	0.731255223					
5	R Square	0.534734202					
6	Adjusted R Square	0.523102557					
7	Standard Error	7287.722712					
8	Observations	42					
9							
10	ANOVA						
11		*df*	*SS*	*MS*	*F*	*Significance F*	
12	Regression	1	2441633669	2441633669	45.97236277	3.79802E-08	
13	Residual	40	2124436093	53110902.32			
14	Total	41	4566069762				
15							
16		*Coefficients*	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
17	Intercept	32673.2199	8831.950745	3.699434116	0.000649604	14823.18178	50523.25802
18	Square Feet	35.03637258	5.16738385	6.780292234	3.79802E-08	24.59270036	45.48004481

53% of the variation in home market values can be explained by home size.

The standard error of $7287 is less than standard deviation (not shown) of $10,553.

Regression as Analysis of Variance

ANOVA conducts an *F*-test to determine whether variation in *Y* is due to varying levels of *X*.

ANOVA is used to test for *significance of regression:*

H_0: population slope coefficient = 0
H_1: population slope coefficient $\neq$ 0

Excel reports the *p*-value (*Significance F*).

Rejecting H_0 indicates that *X* explains variation in *Y*.

Example 8.7: Interpreting Significance of Regression

$H_0: \beta_1 = 0$ Home size is <u>not</u> a significant variable

$H_1: \beta_1 \neq 0$ Home size is a significant variable

- *p*-value = 3.798 x 10^{-8}
  - Reject H_0: The slope is not equal to zero. Using a linear relationship, home size is a significant variable in explaining variation in market value.

	A	B	C	D	E	F	G
1	Regression Analysis						
2							
3	*Regression Statistics*						
4	Multiple R	0.731255223					
5	R Square	0.534734202					
6	Adjusted R Square	0.523102557					
7	Standard Error	7287.722712					
8	Observations	42					
9							
10	ANOVA						
11		*df*	*SS*	*MS*	*F*	*Significance F*	
12	Regression	1	2441633669	2441633669	45.97236277	3.79802E-08	
13	Residual	40	2124436093	53110902.32			
14	Total	41	4566069762				
15							
16		*Coefficients*	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
17	Intercept	32673.2199	8831.950745	3.699434116	0.000649604	14823.18178	50523.25802
18	Square Feet	35.03637258	5.16738385	6.780292234	3.79802E-08	24.59270036	45.48004481

Testing Hypotheses for Regression Coefficients

- An alternate method for testing whether a slope or intercept is zero is to use a t-test:

$$t = \frac{b_1 - 0}{\text{standard error}} \quad (8.8)$$

- Excel provides the *p*-values for tests on the slope and intercept.

	A	B	C	D	E	F	G
1	Regression Analysis						
2							
3	*Regression Statistics*						
4	Multiple R	0.731255223					
5	R Square	0.534734202					
6	Adjusted R Square	0.523102557					
7	Standard Error	7287.722712					
8	Observations	42					
9							
10	ANOVA						
11		*df*	*SS*	*MS*	*F*	*Significance F*	
12	Regression	1	2441633669	2441633669	45.97236277	3.79802E-08	
13	Residual	40	2124436093	53110902.32			
14	Total	41	4566069762				
15							
16		*Coefficients*	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
17	Intercept	32673.2199	8831.950745	3.699434116	0.000649604	14823.18178	50523.25802
18	Square Feet	35.03637258	5.16738385	6.780292234	3.79802E-08	24.59270036	45.48004481

Example 8.8: Interpreting Hypothesis Tests for Regression Coefficients

$$t = \frac{b_1 - 0}{\text{standard error}} \tag{8.8}$$

- Use *p*-values to draw conclusion

	A	B	C	D	E	F	G
1	Regression Analysis						
2							
3	*Regression Statistics*						
4	Multiple R	0.731255223					
5	R Square	0.534734202					
6	Adjusted R Square	0.523102557					
7	Standard Error	7287.722712					
8	Observations	42					
9							
10	ANOVA						
11		*df*	*SS*	*MS*	*F*	*Significance F*	
12	Regression	1	2441633669	2441633669	45.97236277	3.79802E-08	
13	Residual	40	2124436093	53110902.32			
14	Total	41	4566069762				
15							
16		*Coefficients*	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
17	Intercept	32673.2199	8831.950745	3.699434116	0.000649604	14823.18178	50523.25802
18	Square Feet	35.03637258	5.16738385	6.780292234	3.79802E-08	24.59270036	45.48004481

- Neither coefficient is statistically equal to zero.

Confidence Intervals for Regression Coefficients

- Confidence intervals (*Lower 95%* and *Upper 95%* values in the output) provide information about the unknown values of the true regression coefficients, accounting for sampling error.
- We may also use confidence intervals to test hypotheses about the regression coefficients.
  - To test the hypotheses

$$H_0: \beta_1 = B_1$$
$$H_1: \beta_1 \neq B_1$$

check whether B_1 falls within the confidence interval for the slope. If it does, reject the null hypothesis.

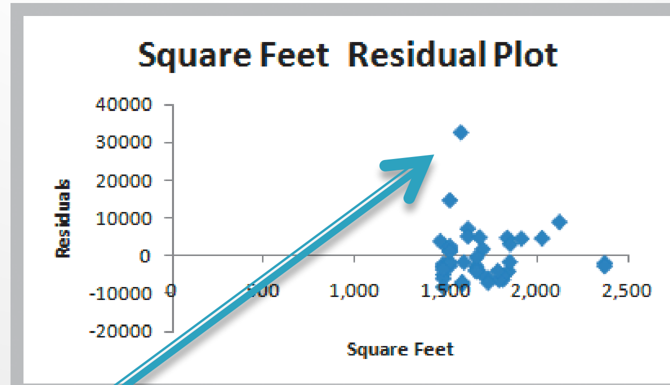
Example 8.9: Interpreting Confidence Intervals for Regression Coefficients

- For the Home Market Value data, a 95% confidence interval for the intercept is [14,823, 50,523], and for the slope, [24.59, 45.48].
- Although we estimated that a house with 1,750 square feet has a market value of 32,673 + 35.036(1,750) =$93,986, if the true population parameters are at the extremes of the confidence intervals, the estimate might be as low as 14,823 + 24.59(1,750) = $57,855 or as high as 50,523 + 45.48(1,750) = $130,113.

Residual Analysis and Regression Assumptions

- **Residual** = Actual *Y* value – Predicted *Y* value
- **Standard residual** = residual / standard deviation
- Rule of thumb: Standard residuals outside of ± 2 or ± 3 are potential outliers.
- Excel provides a table and a plot of residuals.

	A	B	C	D
22	RESIDUAL OUTPUT			
23				
24	*Observation*	*Predicted Market Value*	*Residuals*	*Standard Residuals*
25	1	96159.12702	-6159.127018	-0.855636403
26	2	99732.83702	4667.162978	0.64837022
27	3	97210.2182	-3910.218196	-0.543214164
28	4	96159.12702	-5159.127018	-0.716714702
29	5	96999.99996	4900.00004	0.680716341



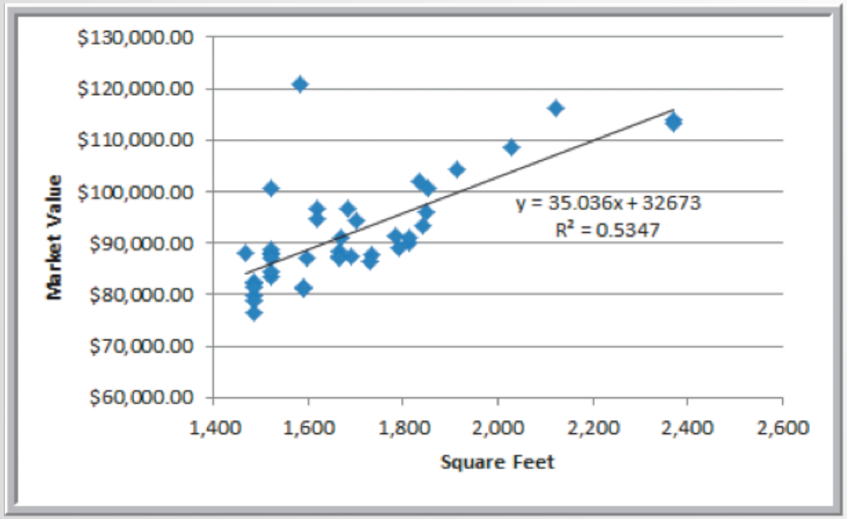
This point has a standard residual of 4.53

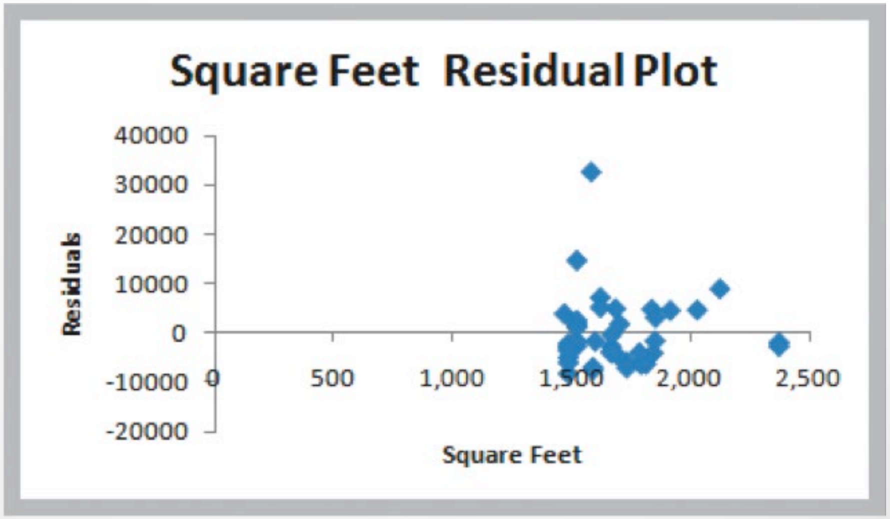
Checking Assumptions

- *Linearity*
  - examine scatter diagram (should appear linear)
  - examine residual plot (should appear random)
- *Normality of Errors*
  - view a histogram of standard residuals
  - regression is robust to departures from normality
- *Homoscedasticity:* variation about the regression line is constant
  - examine the residual plot
- *Independence of Errors*: successive observations should not be related.
  - This is important when the independent variable is time.

Example 8.11: Checking Regression Assumptions for the *Home Market Value* Data

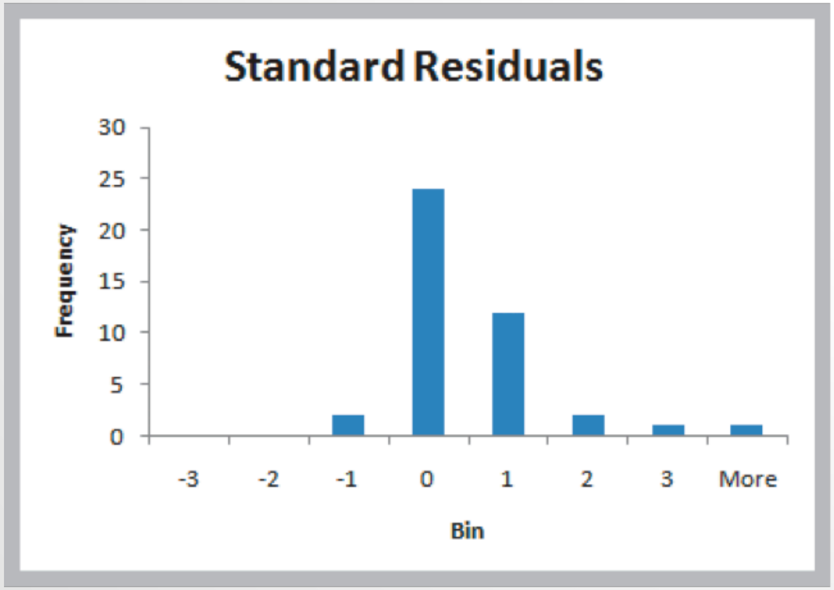
- Linearity - linear trend in scatterplot
  - no pattern in residual plot





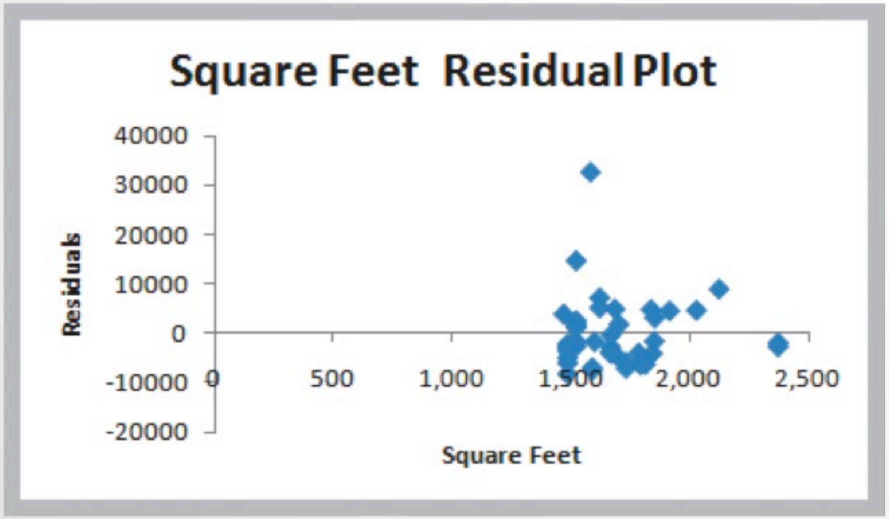
Example 8.11 Continued

Normality of Errors – residual histogram appears slightly skewed but is not a serious departure



Example 8.11 Continued

- Homoscedasticity – residual plot shows no serious difference in the spread of the data for different *X* values.



Example 8.11 Continued

- Independence of Errors – Because the data is cross-sectional, we can assume this assumption holds.

Multiple Linear Regression

- A linear regression model with more than one independent variable is called a **multiple linear regression model**.

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k + \varepsilon \qquad \textbf{(8.10)}$$

where

Y is the dependent variable,
$X_1, \ldots, X_k$ are the independent (explanatory) variables,
β_0 is the intercept term,
$\beta_1, \ldots, \beta_k$ are the regression coefficients for the independent variables,
ε is the error term

Estimated Multiple Regression Equation

- We estimate the regression coefficients—called **partial regression coefficients** — b_0, b_1, b_2, … b_k, then use the model:

$$\hat{Y} = b_0 + b_1X_1 + b_2X_2 + \cdots + b_kX_k \qquad \textbf{(8.11)}$$

- The partial regression coefficients represent the expected change in the dependent variable when the associated independent variable is increased by one unit *while the values of all other independent variables are held constant*.

Excel Regression Tool

- The independent variables in the spreadsheet must be in contiguous columns.
  - So, you may have to manually move the columns of data around before applying the tool.
- Key differences:
- **Multiple R** and **R Square** are called the **multiple correlation coefficient** and the **coefficient of multiple determination**, respectively, in the context of multiple regression.
- ANOVA tests for significance of the entire model. That is, it computes an F-statistic for testing the hypotheses:

$$H_0: \beta_1 = \beta_2 = \cdots = \beta_k = 0$$

H_1: at least one β_j is not 0

ANOVA for Multiple Regression

- ANOVA tests for significance of the entire model. That is, it computes an F-statistic for testing the hypotheses:

$$H_0: \beta_1 = \beta_2 = \cdots = \beta_k = 0$$

$$H_1: \text{at least one } \beta_j \text{ is not } 0$$

- The multiple linear regression output also provides information to test hypotheses about *each* of the individual regression coefficients.
  - If we reject the null hypothesis that the slope associated with independent variable *i* is 0, then the independent variable *i* is significant and improves the ability of the model to better predict the dependent variable. If we cannot reject H0, then that independent variable is not significant and probably should not be included in the model.

Example 8.12: Interpreting Regression Results for the *Colleges and Universities* Data

- Predict student graduation rates using several in

	A	B	C	D	E	F	G
1	**Colleges and Universities**						
2							
3	**School**	**Type**	**Median SAT**	**Acceptance Rate**	**Expenditures/Student**	**Top 10% HS**	**Graduation %**
4	Amherst	Lib Arts	1315	22%	$ 26,636	85	93
5	Barnard	Lib Arts	1220	53%	$ 17,653	69	80
6	Bates	Lib Arts	1240	36%	$ 17,554	58	88
7	Berkeley	University	1176	37%	$ 23,665	95	68
8	Bowdoin	Lib Arts	1300	24%	$ 25,703	78	90
9	Brown	University	1281	24%	$ 24,201	80	90

Example 8.12 Continued

- Regression model

	A	B	C	D	E	F	G
1	SUMMARY OUTPUT						
2							
3	*Regression Statistics*						
4	Multiple R	0.731044486					
5	R Square	0.534426041					
6	Adjusted R Square	0.492101135					
7	Standard Error	5.30833812					
8	Observations	49					
9							
10	ANOVA						
11		*df*	*SS*	*MS*	*F*	*Significance F*	
12	Regression	4	1423.209266	355.8023166	12.62675098	6.33158E-07	
13	Residual	44	1239.851958	28.1784536			
14	Total	48	2663.061224				
15							
16		*Coefficients*	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
17	Intercept	17.92095587	24.55722367	0.729763108	0.469402466	-31.57087643	67.41278818
18	Median SAT	0.072006285	0.017983915	4.003927007	0.000236106	0.035762085	0.108250485
19	Acceptance Rate	-24.8592318	8.315184822	-2.989618672	0.004559569	-41.61738567	-8.101077939
20	Expenditures/Student	-0.00013565	6.59314E-05	-2.057438385	0.045600178	-0.000268526	-2.77379E-06
21	Top 10% HS	-0.162764489	0.079344518	-2.051364015	0.046213848	-0.322672857	-0.00285612

$$\text{Graduation\%} = 17.92 + 0.072\text{ SAT} - 24.859\text{ ACCEPTANCE} - 0.000136\text{ EXPENDITURES} - 0.163\text{ TOP10\% HS}$$

- The value of R^2 indicates that 53% of the variation in the dependent variable is explained by these independent variables.
- All coefficients are statistically significant.

Model Building Issues

- A good regression model should include only significant independent variables.
- However, it is not always clear exactly what will happen when we add or remove variables from a model; variables that are (or are not) significant in one model may (or may not) be significant in another.
  - Therefore, you should not consider dropping all insignificant variables at one time, but rather take a more structured approach.
- Adding an independent variable to a regression model will always result in R^2 equal to or greater than the R^2 of the original model.
- ***Adjusted R^2*** reflects both the number of independent variables and the sample size and may either increase or decrease when an independent variable is added or dropped. An increase in adjusted R^2 indicates that the model has improved.

Systematic Model Building Approach

1. Construct a model with all available independent variables. Check for significance of the independent variables by examining the p-values.
2. Identify the independent variable having the largest p-value that exceeds the chosen level of significance.
3. Remove the variable identified in step 2 from the model and evaluate adjusted R^2.

   (Don't remove all variables with p-values that exceed a at the same time, but remove only one at a time.)
4. Continue until all variables are significant.

Example 8.13: Identifying the Best Regression Model

- *Banking Data*

Home value has the largest p-value; drop and re-run the regression.

	A	B	C	D	E	F	G
1	SUMMARY OUTPUT						
2							
3	*Regression Statistics*						
4	Multiple R	0.97309221					
5	R Square	0.946908448					
6	Adjusted R Square	0.944143263					
7	Standard Error	2055.64333					
8	Observations	102					
9							
10	ANOVA						
11		*df*	*SS*	*MS*	*F*	*Significance F*	
12	Regression	5	7235179873	1447035975	342.4394584	1.5184E-59	
13	Residual	96	405664271.9	4225669.499			
14	Total	101	7640844145				
15							
16		*Coefficients*	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
17	Intercept	-10710.64278	4260.976308	-2.513659314	0.013613179	-19168.61391	-2252.671659
18	Age	318.6649626	60.98611242	5.225205378	1.01152E-06	197.6084862	439.721439
19	Education	621.8603472	318.9595184	1.949652891	0.054135377	-11.26929279	1254.989987
20	Income	0.146323453	0.040781001	3.588029937	0.000526666	0.065373806	0.227273101
21	Home Value	0.009183067	0.011038075	0.831944635	0.407504891	-0.012727338	0.031093473
22	Wealth	0.074331533	0.011189265	6.643111131	1.84838E-09	0.052121017	0.096542049

Example 8.13 Continued

- Bank regression after removing *Home Value*

	A	B	C	D	E	F	G
1	SUMMARY OUTPUT						
2							
3	*Regression Statistics*						
4	Multiple R	0.97289551					
5	R Square	0.946525674					
6	Adjusted R Square	0.944320547					
7	Standard Error	2052.378536					
8	Observations	102					
9							
10	ANOVA						
11		*df*	*SS*	*MS*	*F*	*Significance F*	
12	Regression	4	7232255152	1808063788	429.2386497	9.68905E-61	
13	Residual	97	408588992.5	4212257.655			
14	Total	101	7640844145				
15							
16		*Coefficients*	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
17	Intercept	-12432.45673	3718.674319	-3.343249681	0.001177705	-19812.99587	-5051.917589
18	Age	325.0652837	60.40284468	5.381622098	5.1267E-07	205.1823574	444.9482101
19	Education	773.3800418	261.4330936	2.958233142	0.003886994	254.5077194	1292.252364
20	Income	0.159747379	0.037393587	4.272052794	4.52422E-05	0.085531459	0.233963298
21	Wealth	0.072988791	0.011054665	6.602532898	2.16051E-09	0.051048341	0.094929242

Adjusted R^2 improves slightly.
All X variables are significant.

Alternate Criterion

- Use the t-statistic.
- If | t | < 1, then the standard error will decrease and adjusted R^2 will increase if the variable is removed. If | t | > 1, then the opposite will occur.
- You can follow the same systematic approach, except using t-values instead of p-values.

Multicollinearity

- **Multicollinearity** occurs when there are strong correlations among the independent variables, and they can predict each other better than the dependent variable.
  - When significant multicollinearity is present, it becomes difficult to isolate the effect of one independent variable on the dependent variable, the signs of coefficients may be the opposite of what they should be, making it difficult to interpret regression coefficients, and *p*-values can be inflated.
- Correlations exceeding ± 0.7 may indicate multicollinearity
- The **variance inflation factor** is a better indicator, but not computed in Excel.

Example 8.14: Identifying Potential Multicollinearity

- *Colleges and Universities* correlation matrix; none exceed the recommend threshold of ±0.7

	A	B	C	D	E	F
1		*Median SAT*	*Acceptance Rate*	*Expenditures/Student*	*Top 10% HS*	*Graduation %*
2	Median SAT	1				
3	Acceptance Rate	-0.601901959	1			
4	Expenditures/Student	0.572741729	-0.284254415	1		
5	Top 10% HS	0.503467995	-0.609720972	0.505782049	1	
6	Graduation %	0.564146827	-0.55037751	0.042503514	0.138612667	1

- *Banking Data* correlation matrix; large correlations exist

	A	B	C	D	E	F	G
1		*Age*	*Education*	*Income*	*Home Value*	*Wealth*	*Balance*
2	Age	1					
3	Education	0.173407147	1				
4	Income	0.4771474	0.57539402	1			
5	Home Value	0.386493114	0.753521067	0.795355158	1		
6	Wealth	0.468091791	0.469413035	0.946665447	0.698477789	1	
7	Balance	0.565466834	0.55488066	0.951684494	0.766387128	0.948711734	1

Example 8.14 Continued

- If we remove Wealth from the model, the adjusted R^2 drops to 0.9201, but we discover that Education is no longer significant.
- Dropping Education and leaving only Age and Income in the model results in an adjusted R^2 of 0.9202.
- However, if we remove Income from the model instead of Wealth, the Adjusted R^2 drops to only 0.9345, and all remaining variables (Age, Education, and Wealth) are significant.

	A	B	C	D	E	F	G
1	SUMMARY OUTPUT						
2							
3	*Regression Statistics*						
4	Multiple R	0.967710981					
5	R Square	0.936464543					
6	Adjusted R Square	0.93451958					
7	Standard Error	2225.695322					
8	Observations	102					
9							
10	ANOVA						
11		*df*	*SS*	*MS*	*F*	*Significance F*	
12	Regression	3	7155379617	2385126539	481.4819367	1.71667E-58	
13	Residual	98	485464527.3	4953719.667			
14	Total	101	7640844145				
15							
16		*Coefficients*	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
17	Intercept	-17732.45142	3801.662822	-4.664393517	9.79978E-06	-25276.72757	-10188.17528
18	Age	367.8214086	64.59823831	5.693985134	1.2977E-07	239.6283071	496.0145102
19	Education	1300.308712	249.9731413	5.201793703	1.08292E-06	804.2451489	1796.372276
20	Wealth	0.116467903	0.004679827	24.88722652	3.75813E-44	0.107180939	0.125754866

Practical Issues in Trendline and Regression Modeling

- Identifying the best regression model often requires experimentation and trial and error.
- The independent variables selected should make sense in attempting to explain the dependent variable
  - Logic should guide your model development. In many applications, behavioral, economic, or physical theory might suggest that certain variables should belong in a model.
- Additional variables increase R^2 and, therefore, help to explain a larger proportion of the variation.
  - Even though a variable with a large p-value is not statistically significant, it could simply be the result of sampling error and a modeler might wish to keep it.
- Good models are as simple as possible (the principle of **parsimony**).

Overfitting

- **Overfitting** means fiting a model too closely to the sample data at the risk of not fitting it well to the population in which we are interested.
  - In fitting the crude oil prices in Example 8.2, we noted that the R^2-value will increase if we fit higher-order polynomial functions to the data. While this might provide a better mathematical fit to the sample data, doing so can make it difficult to explain the phenomena rationally.
- In multiple regression, if we add too many terms to the model, then the model may not adequately predict other values from the population.
- Overfitting can be mitigated by using good logic, intuition, theory, and parsimony.

Regression with Categorical Variables

- Regression analysis requires numerical data.
- Categorical data can be included as independent variables, but must be coded numeric using *dummy variables.*
- For variables with 2 categories, code as 0 and 1.

Example 8.15: A Model with Categorical Variables

- *Employee Salaries* provides data for 35 employees

	A	B	C	D
1	Employee Salary Data			
2				
3	Employee	Salary	Age	MBA
4	1	$ 28,260	25	No
5	2	$ 43,392	28	Yes
6	3	$ 56,322	37	Yes
7	4	$ 26,086	23	No
8	5	$ 36,807	32	No

- Predict *Salary* using *Age* and *MBA* (code as yes=1, no=0)

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$$

where

Y = salary

X_1 = age

X_2 = MBA indicator (0 or 1)

Example 8.15 Continued

- Salary = 893.59 + 1044.15 × Age + 14767.23 × MBA
  - If MBA = 0, salary = 893.59 + 1044 × Age
  - If MBA = 1, salary =15,660.82 + 1044 × Age

	A	B	C	D	E	F	G
1	SUMMARY OUTPUT						
2							
3	*Regression Statistics*						
4	Multiple R	0.976118476					
5	R Square	0.952807278					
6	Adjusted R Square	0.949857733					
7	Standard Error	2941.914352					
8	Observations	35					
9							
10	ANOVA						
11		*df*	*SS*	*MS*	*F*	*Significance F*	
12	Regression	2	5591651177	2795825589	323.0353318	6.05341E-22	
13	Residual	32	276955521.7	8654860.054			
14	Total	34	5868606699				
15							
16		*Coefficients*	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
17	Intercept	893.5875971	1824.575283	0.489751015	0.627650922	-2822.950634	4610.125828
18	Age	1044.146043	42.14128238	24.77727265	1.8878E-22	958.3070599	1129.985026
19	MBA	14767.23159	1351.801764	10.92411031	2.49752E-12	12013.7015	17520.76168

Interactions

- An **interaction** occurs when the effect of one variable is dependent on another variable.
- We can test for interactions by defining a new variable as the product of the two variables, X3 = X1 × X2 , and testing whether this variable is significant, leading to an alternative model.

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon$$

Example 8.16: Incorporating Interaction Terms in a Regression Model

- Define an interaction between Age and MBA and re-run the regression.

	A	B	C	D	E
1	Employee Salary Data				
2					
3	Employee	Salary	Age	MBA	Interaction
4	1	$ 28,260	25	0	0
5	2	$ 43,392	28	1	28
6	3	$ 56,322	37	1	37
7	4	$ 26,086	23	0	0

	A	B	C	D	E	F	G
1	SUMMARY OUTPUT						
2							
3	*Regression Statistics*						
4	Multiple R	0.989321416					
5	R Square	0.978756863					
6	Adjusted R Square	0.976701076					
7	Standard Error	2005.37675					
8	Observations	35					
9							
10	ANOVA						
11		*df*	*SS*	*MS*	*F*	*Significance F*	
12	Regression	3	5743939086	1914646362	476.098288	5.31397E-26	
13	Residual	31	124667613.2	4021535.91			
14	Total	34	5868606699				
15							
16		*Coefficients*	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
17	Intercept	3902.509386	1336.39766	2.920170772	0.006467654	1176.908389	6628.110383
18	Age	971.3090382	31.06887722	31.26308786	5.23658E-25	907.9436454	1034.674431
19	MBA	-2971.080074	3026.24236	-0.98177202	0.333812767	-9143.142058	3200.981911
20	Interaction	501.8483604	81.55221742	6.153705887	7.9295E-07	335.5215164	668.1752044

The MBA indicator is not significant; drop and re-run.

Example 8.16 Continued

- Adjusted R^2 increased slightly, and both age and the interaction term are significant. The final model is

salary = 3,323.11 + 984.25 × age + 425.58 × MBA × age

	A	B	C	D	E	F	G
1	SUMMARY OUTPUT						
2							
3	*Regression Statistics*						
4	Multiple R	0.98898754					
5	R Square	0.978096355					
6	Adjusted R Square	0.976727377					
7	Standard Error	2004.24453					
8	Observations	35					
9							
10	ANOVA						
11		*df*	*SS*	*MS*	*F*	*Significance F*	
12	Regression	2	5740062823	2870031411	714.4720368	2.80713E-27	
13	Residual	32	128543876.4	4016996.136			
14	Total	34	5868606699				
15							
16		*Coefficients*	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
17	Intercept	3323.109564	1198.353141	2.773063675	0.009184278	882.1440943	5764.075033
18	Age	984.2455409	28.12039088	35.00113299	4.40388E-27	926.9661791	1041.524903
19	Interaction	425.5845915	24.81794165	17.14826304	1.08793E-17	375.0320986	476.1370843

Categorical Variables with More Than Two Levels

- When a categorical variable has $k > 2$ levels, we need to add $k - 1$ additional variables to the model.

Example 8.17: A Regression Model with Multiple Levels of Categorical Variables

- The Excel file *Surface Finish* provides measurements of the surface finish of 35 parts produced on a lathe, along with the revolutions per minute (RPM) of the spindle and one of four types of cutting tools used.

	A	B	C	D
1	Surface Finish Data			
2				
3	Part	Surface Finish	RPM	Cutting Tool
4	1	45.44	225	A
5	2	42.03	200	A
6	3	50.10	250	A
7	4	48.75	245	A
8	5	47.92	235	A
9	6	47.79	237	A
10	7	52.26	265	A
11	8	50.52	259	A
12	9	45.58	221	A
13	10	44.78	218	A
14	11	33.50	224	B
15	12	31.23	212	B
16	13	37.52	248	B
17	14	37.13	260	B
18	15	34.70	243	B

Example 8.17 Continued

- Because we have k = 4 levels of tool type, we will define a regression model of the form

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \varepsilon$$

where

$$Y = \textit{surface finish}$$
$$X_1 = \textit{RPM}$$
$$X_2 = 1 \text{ if tool type is B and 0 if not}$$
$$X_3 = 1 \text{ if tool type is C and 0 if not}$$
$$X_4 = 1 \text{ if tool type is D and 0 if not}$$

Example 8.17 Continued

- Add 3 columns to the data, one for each of the tool type variables

	A	B	C	D	E	F
1	Surface Finish Data					
2						
3	Part	Surface Finish	RPM	Type B	Type C	Type D
4	1	45.44	225	0	0	0
5	2	42.03	200	0	0	0
6	3	50.10	250	0	0	0
7	4	48.75	245	0	0	0
8	5	47.92	235	0	0	0
9	6	47.79	237	0	0	0
10	7	52.26	265	0	0	0
11	8	50.52	259	0	0	0
12	9	45.58	221	0	0	0
13	10	44.78	218	0	0	0
14	11	33.50	224	1	0	0
15	12	31.23	212	1	0	0
16	13	37.52	248	1	0	0
17	14	37.13	260	1	0	0
18	15	34.70	243	1	0	0
19	16	33.92	238	1	0	0
20	17	32.13	224	1	0	0
21	18	35.47	251	1	0	0
22	19	33.49	232	1	0	0
23	20	32.29	216	1	0	0
24	21	27.44	225	0	1	0
25	22	24.03	200	0	1	0
26	23	27.33	250	0	1	0
27	24	27.20	245	0	1	0
28	25	27.10	235	0	1	0
29	26	27.30	237	0	1	0
30	27	28.30	265	0	1	0
31	28	28.40	259	0	1	0
32	29	26.80	221	0	1	0
33	30	26.40	218	0	1	0
34	31	21.40	224	0	0	1
35	32	20.50	212	0	0	1
36	33	21.90	248	0	0	1
37	34	22.13	260	0	0	1
38	35	22.40	243	0	0	1

Example 8.17 Continued

- Regression results

	A	B	C	D	E	F	G
1	SUMMARY OUTPUT						
2							
3	*Regression Statistics*						
4	Multiple R	0.994447053					
5	R Square	0.988924942					
6	Adjusted R Square	0.987448267					
7	Standard Error	1.089163115					
8	Observations	35					
9							
10	ANOVA						
11		*df*	*SS*	*MS*	*F*	*Significance F*	
12	Regression	4	3177.784271	794.4460678	669.6973322	7.32449E-29	
13	Residual	30	35.58828875	1.186276292			
14	Total	34	3213.37256				
15							
16		*Coefficients*	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
17	Intercept	24.49437244	2.473298088	9.903526211	5.73134E-11	19.44322388	29.54552101
18	RPM	0.097760627	0.010399996	9.400064035	1.89415E-10	0.076521002	0.119000252
19	Type B	-13.31056756	0.487142953	-27.32374035	9.37003E-23	-14.3054462	-12.31568893
20	Type C	-20.487	0.487088553	-42.06011387	3.12134E-28	-21.48176754	-19.49223246
21	Type D	-26.03674519	0.596886375	-43.62094073	1.06415E-28	-27.25574979	-24.81774059

Surface finish = 24.49 + 0.098 RPM - 13.31 type B - 20.49 type C - 26.04 type D

Regression Models with Nonlinear Terms

- Curvilinear models may be appropriate when scatter charts or residual plots show nonlinear relationships.
- A second order polynomial might be used

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \varepsilon$$

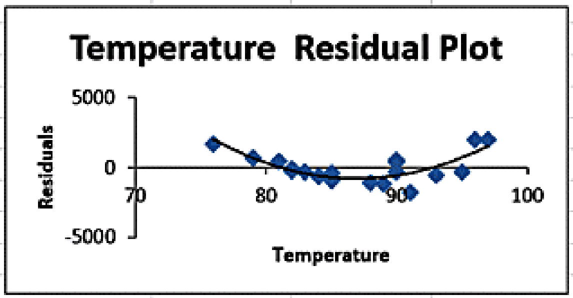
- Here β_1 represents the linear effect of X on Y and β_2 represents the curvilinear effect.
- This model is linear in the β parameters so we can use linear regression methods.

Example 8.18: Modeling Beverage Sales Using Curvilinear Regression

- The U-shape of the residual plot (a second-order polynomial trendline was fit to the residual data) suggests that a linear relationship is not appropriate.

	A	B
1	**Beverage Sales**	
2		
3	**Temperature**	**Sales**
4	85	$ 1,810
5	90	$ 4,825
6	79	$ 438
7	82	$ 775
8	84	$ 1,213
9	96	$ 8,692

	A	B	C	D	E	F	G
1	SUMMARY OUTPUT						
2							
3	*Regression Statistics*						
4	Multiple R	0.922351218					
5	R Square	0.850731769					
6	Adjusted R Square	0.842875547					
7	Standard Error	1041.057399					
8	Observations	21					
9							
10	ANOVA						
11		*df*	*SS*	*MS*	*F*	*Significance F*	
12	Regression	1	117362193.6	117362193.6	108.2876347	2.7611E-09	
13	Residual	19	20592209.67	1083800.509			
14	Total	20	137954403.2				
15							
16		*Coefficients*	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
17	Intercept	-32511.24671	3408.723477	-9.53766034	1.12197E-08	-39645.78695	-25376.70648
18	Temperature	408.6026284	39.26555335	10.40613447	2.7611E-09	326.4188807	490.786376

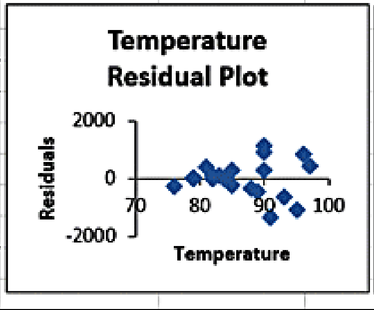


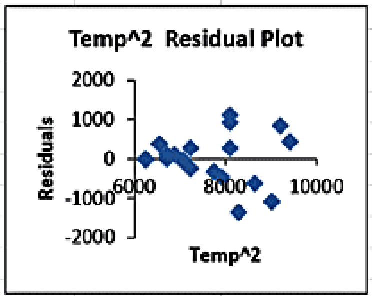
Example 8.18 Continued

- Add a variable for temperature squared.
- The model is:

sales = 142,850 - 3,643.17 × temperature + 23.3 × temperature^2

	A	B	C	D	E	F	G
1	SUMMARY OUTPUT						
2							
3	*Regression Statistics*						
4	Multiple R	0.973326989					
5	R Square	0.947365428					
6	Adjusted R Square	0.941517142					
7	Standard Error	635.1365123					
8	Observations	21					
9							
10	ANOVA						
11		*df*	*SS*	*MS*	*F*	*Significance F*	
12	Regression	2	130693232.2	65346616.12	161.9902753	3.10056E-12	
13	Residual	18	7261171.007	403398.3893			
14	Total	20	137954403.2				
15							
16		*Coefficients*	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
17	Intercept	142850.3406	30575.70155	4.672021683	0.000189738	78613.17532	207087.5059
18	Temperature	-3643.171723	705.2304165	-5.165931075	6.492E-05	-5124.805849	-2161.537598
19	Temp^2	23.30035581	4.053196314	5.748637374	1.89343E-05	14.78490634	31.81580528





Advanced Techniques for Regression Modeling using *XLMiner*

- The regression analysis tool in *XLMiner* has some advanced options not available in Excel's *Descriptive Statistics* tool.
- **Best-subsets regression** evaluates either all possible regression models for a set of independent variables or the best subsets of models for a fixed number of independent variables.

Evaluating Best Subsets Models

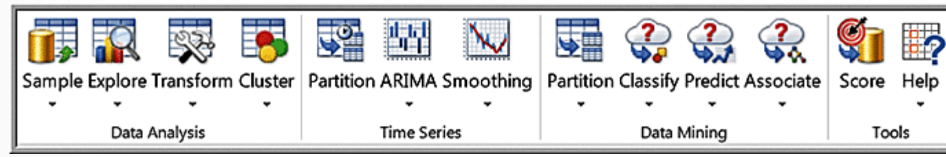
- Best subsets evaluates models using a statistic called *Cp*, (the Bonferroni criterion).
  - *Cp* estimates the bias introduced in the estimates of the responses by having an *underspecified model* (a model with important predictors missing).
  - If *Cp* is much greater than (the number of independent variables plus 1), there is substantial bias. The full model always has $Cp = k + 1$.
  - If all models except the full model have large *Cp*s, it suggests that important predictor variables are missing. Models with a minimum value or having *Cp* less than or at least close to are good models to consider.

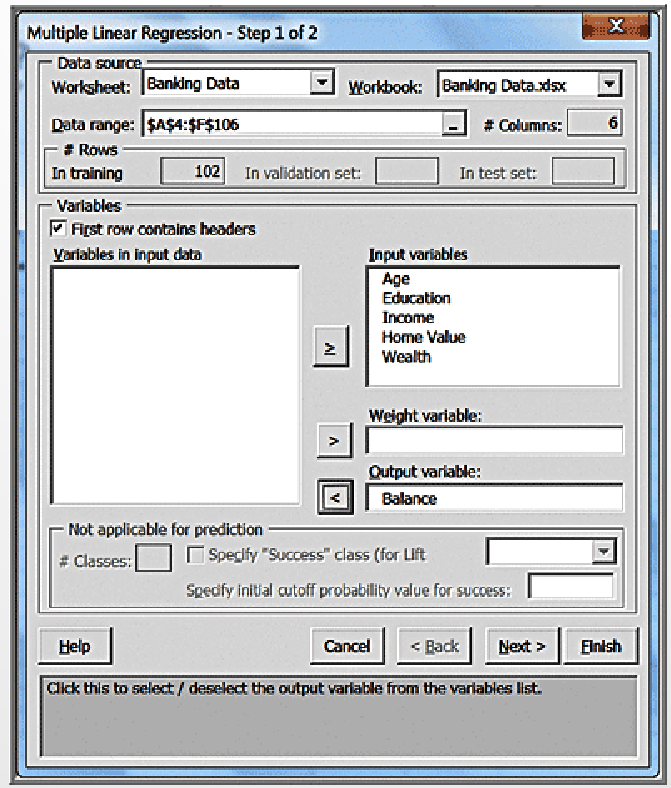
Best-Subsets Procedures

- Backward Elimination begins with all independent variables in the model and deletes one at a time until the best model is identified.
- Forward Selection begins with a model having no independent variables and successively adds one at a time until no additional variable makes a significant contribution.
- Stepwise Selection is similar to Forward Selection except that at each step, the procedure considers dropping variables that are not statistically significant.
- Sequential Replacement replaces variables sequentially, retaining those that improve performance. These options might terminate with a different model.
- Exhaustive Search looks at all combinations of variables to find the one with the best fit, but it can be time consuming for large numbers of variables.

Example 8.19: Using *XLMiner* for Regression

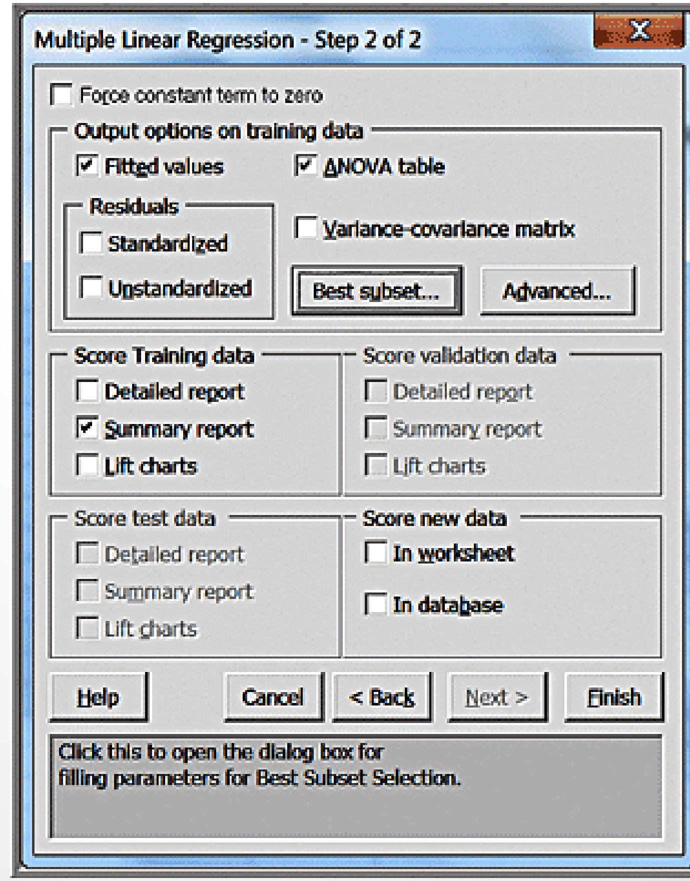
- Click the *Predict* button in the *Data Mining* group and choose *Multiple Linear Regression.*
- Enter the range of the data (including headers)
- Move the appropriate variables to the boxes on the right.

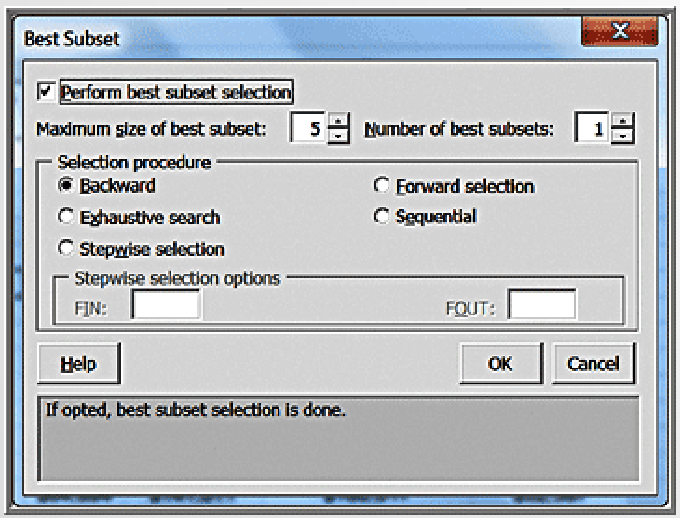




Example 8.19 Continued

- Select the output options and check the *Summary report* box. Before clicking Finish, click on the *Best subsets* button.
- Select the best subsets option:





Example 8.19 Continued

- View results from the "Output Navigator" links.

XLMiner : Multiple Linear Regression

Output Navigator				
Inputs	Train. Score - Summary	Valid. Score - Summary	Test Score - Summary	Database Score
Elapsed Time	Train. Score - Detailed Rep.	Valid. Score - Detailed Rep.	Test Score - Detailed Rep.	New Score - Detailed Rep.
ANOVA	Training Lift Charts	Validation Lift Charts	Test Lift Charts	Subset selection
Reg. Model	Fitted Values	Var. Covar. Matrix	Collinearity Diagnostics	

Example 8.19 Continued

- Regression output (all variables)

The Regression Model

Input variables	Coefficient	Std. Error	p-value	SS
Constant term	-10710.64063	4260.976074	0.01361319	63179490000
Age	318.6649475	60.98611069	0.00000101	2443181000
Education	621.8602905	318.9594727	0.05413537	1643993000
Income	0.14632344	0.040781	0.00052667	2961454000
Home Value	0.00918307	0.01103808	0.40750477	68818.42188
Wealth	0.07433154	0.01118927	0	186482700

Residual df	96
R-squared	0.946908442
Std. Dev. estimate	2055.643311
Residual SS	405664300

ANOVA

Source	df	SS	MS	F-statistic	p-value
Regression	5	7235179518	1447035904	342.4394179	1.51841E-59
Error	96	405664300	4225669.792		
Total	101	7640843818			

- Best subsets results

Best subset selection

	#Coeffs	RSS	Cp	R-Squared	Adj. R-Squared	Probability	Model (Constant present in all models)					
							1	2	3	4	5	6
Choose Subset	2	720505856	72.5069046	0.90570337	0.904760403	0	Constant	Income	*	*	*	*
Choose Subset	3	552461888	34.73949051	0.927696223	0.926235541	0.00000139	Constant	Income	Wealth	*	*	*
Choose Subset	4	445451072	11.41549969	0.941701327	0.939916674	0.01116341	Constant	Age	Income	Wealth	*	*
Choose Subset	5	408588992	4.69212961	0.946525674	0.944320547	0.40748432	Constant	Age	Education	Income	Wealth	*
Choose Subset	6	405664288	6.00000191	0.946908446	0.944143261	1	Constant	Age	Education	Income	Home Value	Wealth

If you click “Choose Subset,” XLMiner will create a new worksheet with the results for this model.

Interpreting *XLMiner* Output

- Typically choose the model with the highest adjusted R^2.
- Models with a minimum value of Cp or having Cp less than or at least close to k + 1 are good models to consider.
- *RSS* is the residual sum of squares, or the sum of squared deviations between the predicted probability of success and the actual value (1 or 0).
- *Probability* is a quasi-hypothesis test that a given subset is acceptable; if this is less than 0.05, you can rule out that subset.

Google Analytics Project 4 Email me (albert.kalim@asbury.edu) your answers by Sunday, 6/19,11:59 p.m. ET (10 points total)

Log in to your Google Analytics dashboard here and click Sign in to Analytics.

After you logged in, focus on the Acquisition tools on the left-hand side. Pick a channel with data (all traffic or social) and **list two business recommendations on how to improve the acquisition and elaborate on them**. You will be graded based on your familiarity with the tools and how you read/interpret the data.