

Chapter 6: Regression Analysis



Statistics, Data Analysis, and
Decision Modeling, Fifth Edition
James R. Evans



Regression Analysis

- Building models that characterize the relationships between a dependent variable and one (single) or more (multiple) independent variables, all of which are numerical.
- Regression analysis can be used for:
 - Cross-sectional data
 - Time series data (forecasting)



Example – Simple Regression

	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

$$\text{Market Value} = a + b * \text{Square Feet}$$



Example – Multiple Regression

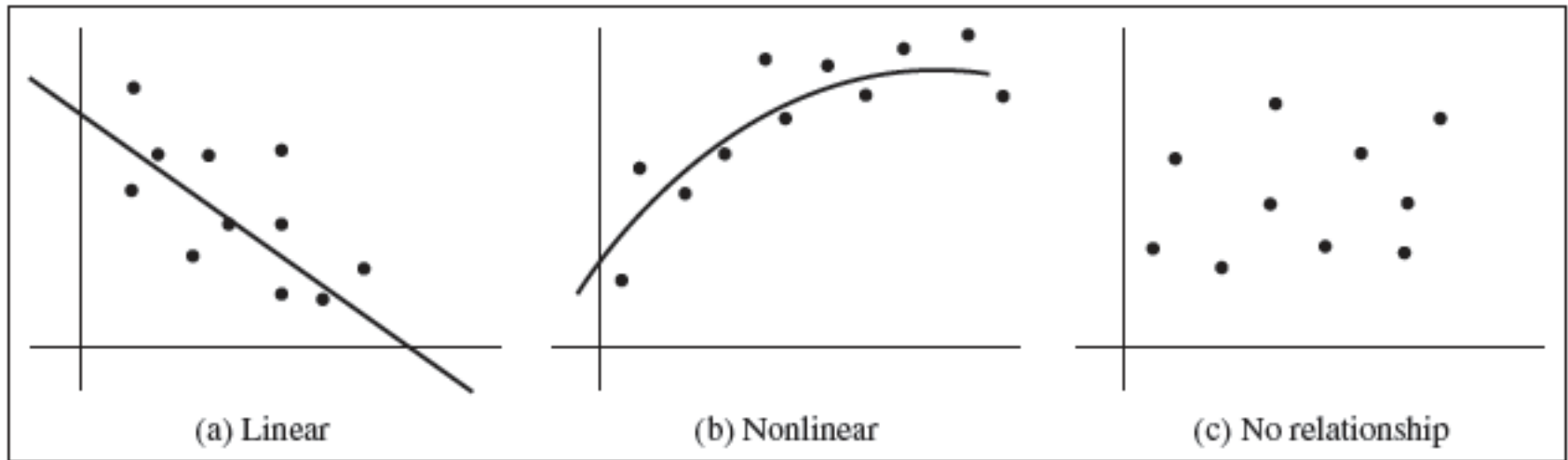
	A	B	C	D	E	F	G
1	Colleges and Universities						
2							
3	School	Type	Median SAT	Acceptance Rate	Expenditures/Student	Top 10% HS	Graduation %
4	Amherst	Lib Arts	1315	22%	\$ 26,636	85	93
5	Barnard	Lib Arts	1220	53%	\$ 17,653	69	80
6	Bates	Lib Arts	1240	36%	\$ 17,554	58	88
7	Berkeley	University	1176	37%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

$$\text{Graduation\%} = a + b * \text{Median SAT} + c * \text{Acceptance Rate} + d * \text{Expenditures/Student} + e * \text{Top 10\% HS}$$

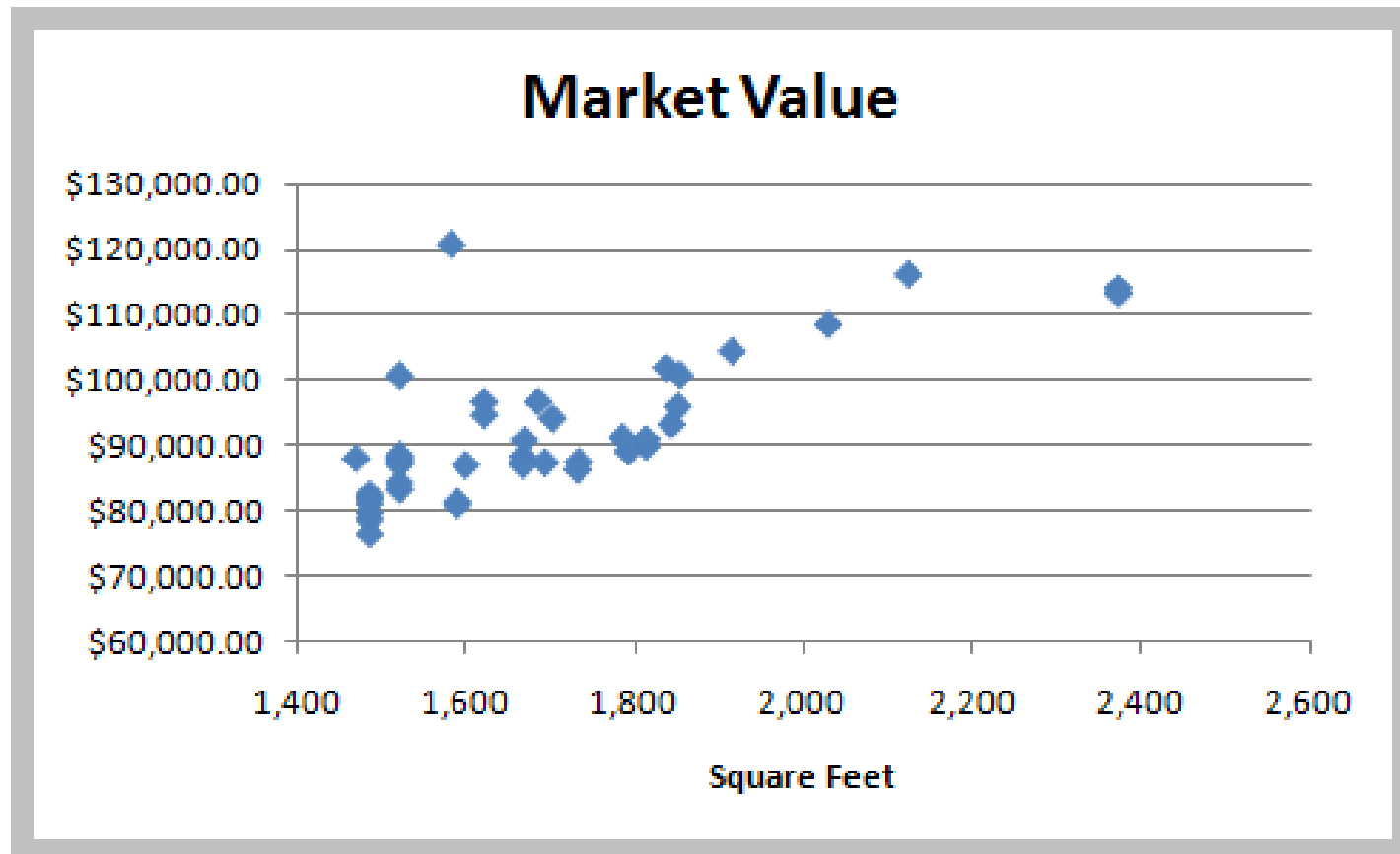


Simple Linear Regression

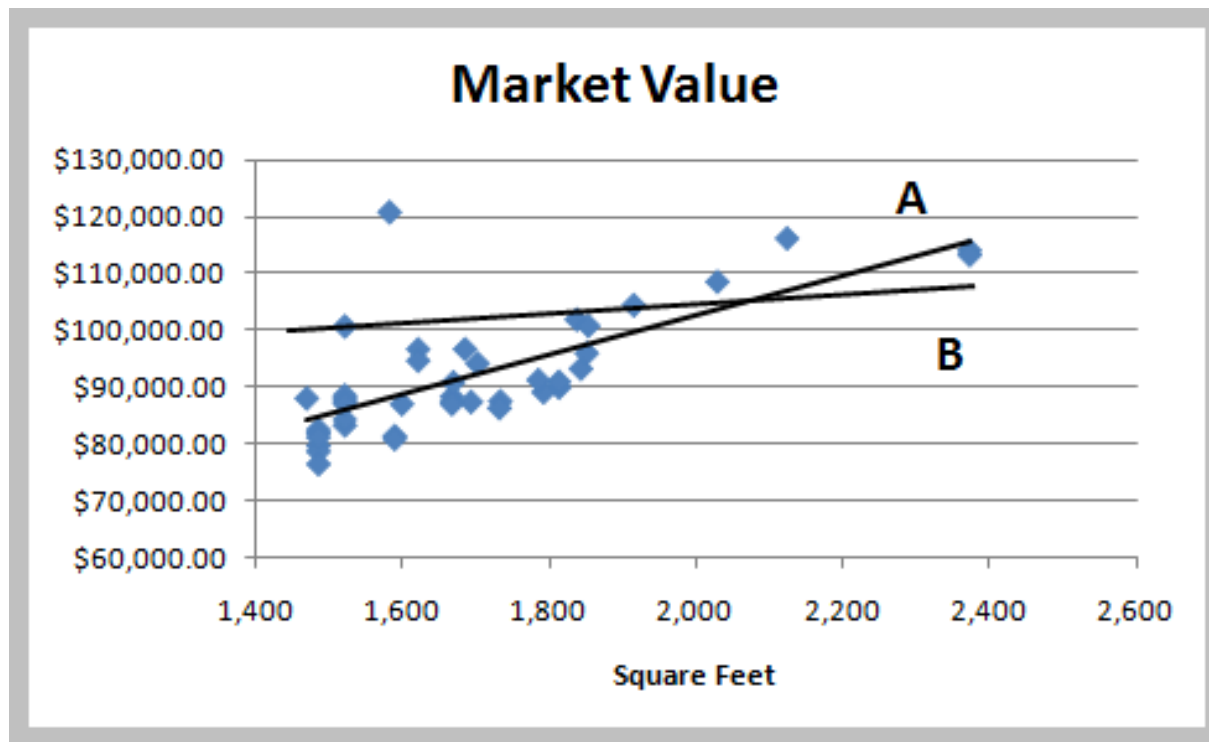
- Single independent variable
- Linear relationship



Example: *Home Market Value* Data



Two Possible Regression Lines



Which one is better and why?



Least-Squares Regression

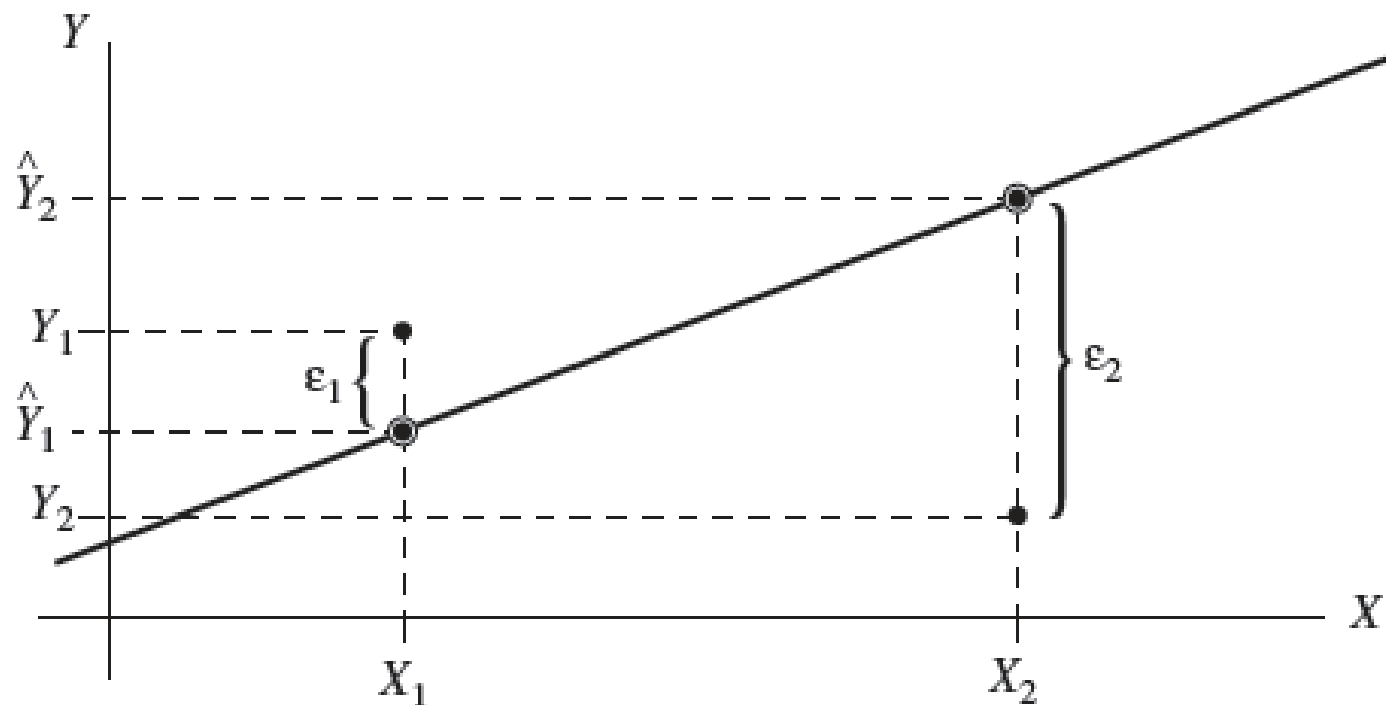
- Simple linear regression model:

$$Y = \beta_0 + \beta_1 X + \varepsilon \quad (6.1)$$

- Least-squares regression estimates β_0 and β_1 by b_0 and b_1 by minimizing the sum of squares of the residuals:

$$\hat{Y} = b_0 + b_1 X \quad (6.2)$$

Errors (Residuals)



Errors associated with individual observations



Least Squares Regression

Minimize $\sum_{i=1}^n e_i^2 = \sum_{i=1}^n \left(Y_i - \hat{Y}_i \right)^2$ (6.3)

$$b_1 = \frac{\sum_{i=1}^n X_i Y_i - n \bar{X} \bar{Y}}{\sum_{i=1}^n X_i^2 - n \bar{X}^2} \quad (6.4)$$

$$b_0 = \bar{Y} - b_1 \bar{X} \quad (6.5)$$



Home Market Value Example

$$\hat{Y} = 32,673 + 35.036X$$

For every additional square foot, the market value increases by \$35.036.

Thus, for a house with 1,750 square feet, the estimated market value is

$$32,673 + 35.036(1,750) = \$93,986.$$

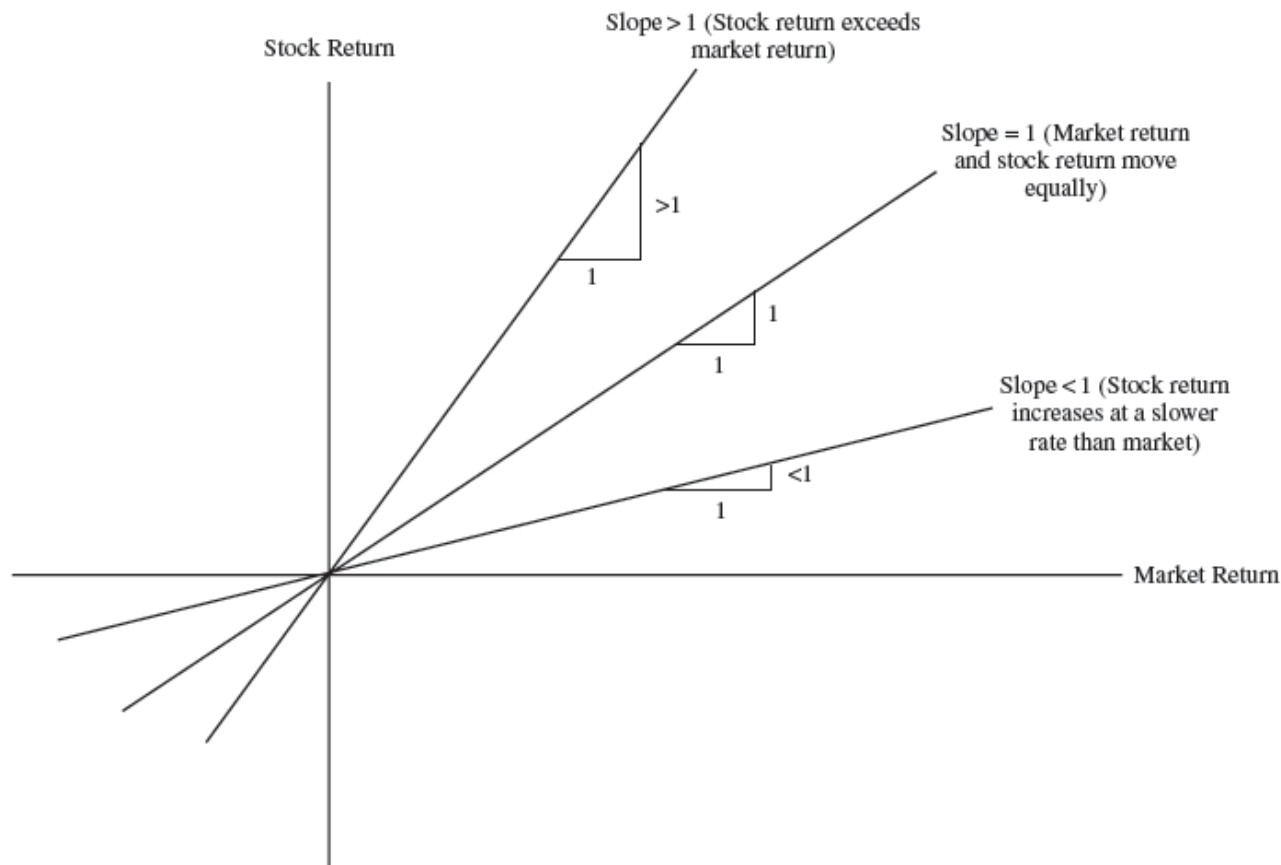


Regression and Investment Risk

- Systematic risk – variation in stock price explained by the market
 - Measured by beta
 - Beta = 1: perfect match to market movements
 - Beta < 1: stock is less volatile than market
 - Beta > 1: stock is more volatile than market

Systematic Risk (Beta)

Beta is the slope of the regression line



Simple Linear Regression Excel Output

	A	B	C	D	E	F	G
1	Home Market Value						
2							
3	<i>Regression Statistics</i>						
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	<i>ANOVA</i>						
11		<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>Significance F</i>	
12	Regression	1	2441633669	2441633669	45.97236277	3.79802E-08	
13	Residual	40	2124436093	53110902.32			
14	Total	41	4566069762				
15							
16		<i>Coefficients</i>	<i>Standard Error</i>	<i>t Stat</i>	<i>P-value</i>	<i>Lower 95%</i>	<i>Upper 95%</i>
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



Coefficient of Determination

Regression Statistics	
Multiple R	0.731255223
R Square	0.534734202
Adjusted R Square	0.523102557
Standard Error	7287.722712
Observations	42

- “Multiple R” = R^2 = **coefficient of determination**: the proportion of variation explained by the independent variable (regression model)

$$0 \leq R^2 \leq 1$$

- The square root of R^2 is the **sample correlation coefficient**, r (where the sign of r is the same as the slope of the fitted line)



Standard Error of the Estimate

<i>Regression Statistics</i>	
Multiple R	0.731255223
R Square	0.534734202
Adjusted R Square	0.523102557
Standard Error	7287.722712
Observations	42

- “Standard Error” = S_{YX} = an unbiased estimate of the variance of the errors about the regression line
- Measures the spread of data about the line

Regression as ANOVA

- Testing for significance of regression $H_0: \beta_1 = 0$
 $H_1: \beta_1 \neq 0$
- If $F > \text{critical value}$ (not provided in output), it is likely that $\beta_1 \neq 0$, or that the regression line is significant
- *Significance F* is the p-value associated with this test

	A	B	C	D	E	F	G
1	Home Market Value						
2							
3	<i>Regression Statistics</i>						
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	<i>ANOVA</i>						
11		<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>Significance F</i>	
12	Regression	1	2441633669	2441633669	45.9723627	3.79802E-08	
13	Residual	40	2124436093	53110902.32			
14	Total	41	4566069762				
15							
16		<i>Coefficients</i>	<i>Standard Error</i>	<i>t Stat</i>	<i>P-value</i>	<i>Lower 95%</i>	<i>Upper 95%</i>
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

t-tests for Slope

$$t = \frac{b_1 - 0}{\text{Standard Error}} \quad \text{with } n-2 \text{ degrees of freedom}$$

	A	B	C	D	E	F	G
1	Home Market Value						
2							
3	<i>Regression Statistics</i>						
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	<i>ANOVA</i>						
11		<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>Significance F</i>	
12	Regression	1	2441633669	2441633669	45.97236277	3.79802E-08	
13	Residual	40	2124436093	53110902.32			
14	Total	41	4566069762				
15							
16		<i>Coefficients</i>	<i>Standard Error</i>	<i>t Stat</i>	<i>P-value</i>	<i>Lower 95%</i>	<i>Upper 95%</i>
17	Intercept	32673.2139	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

Confidence Intervals

	A	B	C	D	E	F	G
1	Home Market Value						
2							
3	<i>Regression Statistics</i>						
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		<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>Significance F</i>	
12	Regression	1	2441633669	2441633669	45.97236277	3.79802E-08	
13	Residual	40	2124436093	53110902.32			
14	Total	41	4566069762				
15							
16		<i>Coefficients</i>	<i>Standard Error</i>	<i>t Stat</i>	<i>P-value</i>	<i>Lower 95%</i>	<i>Upper 95%</i>
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

Tighter confidence intervals provide more accuracy

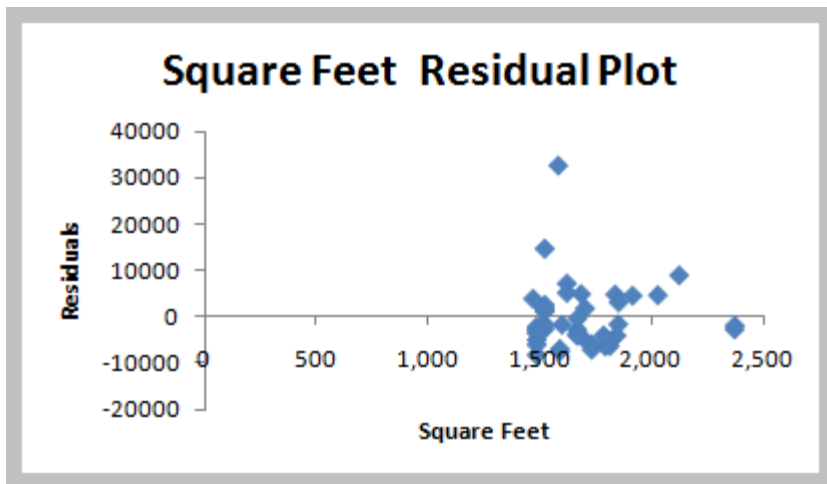
Confidence and Prediction Intervals

- A confidence interval for the mean value of Y quantifies the uncertainty about the population of Y-values for a given value of the independent variable.
- A prediction interval quantifies the uncertainty in Y for a single future observation.

	A	B
1	Confidence Interval Estimate	
2		
3	Data	
4	X Value	1750
5	Confidence Level	95%
6		
7	Intermediate Calculations	
8	Sample Size	42
9	Degrees of Freedom	40
10	t Value	2.021075
11	XBar, Sample Mean of X	1695.262
12	Sum of Squared Differences from XBar	1989034
13	Standard Error of the Estimate	7287.723
14	h Statistic	0.025316
15	Predicted Y (YHat)	93986.87
16		
17	For Average Y	
18	Interval Half Width	2343.533
19	Confidence Interval Lower Limit	91643.3
20	Confidence Interval Upper Limit	96330.4
21		
22	For Individual Response Y	
23	Interval Half Width	14914.31
24	Prediction Interval Lower Limit	79072.6
25	Prediction Interval Upper Limit	108901

Residuals

	A	B	C	D
22	RESIDUAL OUTPUT			
23				
24	<i>Observation</i>	<i>Predicted Market Value</i>	<i>Residuals</i>	<i>Standard Residuals</i>
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



Standard residuals are residuals divided by their standard error, expressed in units independent of the units of the data.



Assumptions Underlying Regression

- **Linearity**
 - Check with scatter diagram of the data or the residual plot
- **Normality of errors** for each X with mean 0 and constant variance
 - Examine histogram of standardized residuals or use goodness-of-fit tests
- **Homoscedasticity** – constant variance about the regression line for all values of the independent variable
 - Examine by plotting residuals and looking for differences in variances at different values of X
- **Independence of errors**. Residuals should be independent for each value of the independent variable. Important if the independent variable is time (e.g., forecasting models).



Autocorrelation

- Important when using regression for forecasting

- Durbin-Watson statistic

$$D = \frac{\sum_{i=2}^n (e_i - e_{i-1})^2}{\sum_{i=1}^n e_i^2}$$

- $D < 1$ suggest autocorrelation
 - $D > 1.5$ suggest no autocorrelation
 - $D > 2.5$ suggest negative autocorrelation
- *PHStat* tool calculates this statistic



Multiple Linear Regression

- Multiple linear regression model:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_k X_k + \varepsilon$$

- Predicted model:

$$\hat{Y} = b_0 + b_1 X_1 + b_2 X_2 + \dots + b_k X_k$$

- The b 's are called **partial regression coefficients**.

Example: *Colleges and Universities Data*

	A	B	C	D	E	F	G
1	Regression Analysis						
2							
3	<i>Regression Statistics</i>						
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		<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>Significance F</i>	
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		<i>Coefficients</i>	<i>Standard Error</i>	<i>t Stat</i>	<i>P-value</i>	<i>Lower 95%</i>	<i>Upper 95%</i>
17	Intercept	17.92095587	24.55722367	0.729763108	0.469402466	-31.57087575	67.4127875
18	Median SAT	0.072006285	0.017983915	4.003927007	0.000236106	0.035762085	0.108250484
19	Acceptance Rate	-24.8592318	8.315184822	-2.989618672	0.004559569	-41.61738544	-8.10107817
20	Expenditures/Student	-0.00013565	6.59314E-05	-2.057438385	0.045600176	-0.000268526	-2.77379E-06
21	Top 10% HS	-0.162764489	0.079344518	-2.051364015	0.046213846	-0.322672855	-0.002856122

$$\text{Graduation\%} = 17.92 + 0.072 \text{ SAT} - 24.859 \text{ Acceptance} \\ - 0.000136 \text{ Expenditures} - 0.163 \text{ Top 10\% HS}$$



Interpreting Results

- Regression statistics similar to single independent variable case
 - R Square (**coefficient of multiple determination**)
 - The value .534 indicates that about 53% of the variation in graduation rate can be explained by the variation in the independent variables.
 - Adjusted R^2 accounts for sample size and number of independent variables. It is useful for comparing models with different sets of independent variables.

ANOVA Results

- Significance of regression

- $H_0: \beta_1 = \beta_2 = \dots = \beta_k = 0$
- $H_1: \text{at least one } \beta_j \text{ is not } 0$

Note: df for residual is $n - k - 1$;
df for regression is k

	A	B	C	D	E	F	G
1	Regression Analysis						
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	-31.57087575	67.4127875
18	Median SAT	0.072006285	0.017983915	4.003927007	0.000236106	0.035762085	0.108250484
19	Acceptance Rate	-24.8592318	8.315184822	-2.989618672	0.004559569	-41.61738544	-8.10107817
20	Expenditures/Student	-0.00013565	6.59314E-05	-2.057438385	0.045600176	-0.000268526	-2.77379E-06
21	Top 10% HS	-0.162764489	0.079344518	-2.051364015	0.046213846	-0.322672855	-0.002856122



Multicollinearity

- **Multicollinearity** – when two or more independent variables contain high levels of the same information.
- The independent variables predict each other better than the dependent variable, making it difficult to interpret the regression coefficients and lead to poor statistical conclusions.
- Effects: Estimates of the regression coefficients are unstable depending on which variables are present, signs may be opposite of expectations, and p-values can be inflated



Correlation Matrix

	A	B	C	D	E	F
1		<i>Median SAT</i>	<i>Acceptance Rate</i>	<i>Expenditures/Student</i>	<i>Top 10% HS</i>	<i>Graduation %</i>
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

There are potential multicollinearity issues. However, multicollinearity is best measured by computing variance inflation factors (VIFs).



Measuring Multicollinearity

- Variance Inflation Factor, $VIF = \frac{1}{1 - r_j^2}$
- Option in *PHStat* routine (be sure to check the box).
- If no multicollinearity, $VIF = 1$
- Researchers suggest that VIF should be no greater than 5

VIF Results

	A	B	C	D	E
1	Regression Analysis			Regression Analysis	
2	Median SAT and all other X			Expenditures/Student and all other X	
3	<i>Regression Statistics</i>			<i>Regression Statistics</i>	
4	Multiple R	0.733444235		Multiple R	0.659705397
5	R Square	0.537940446		R Square	0.435211211
6	Adjusted R Square	0.507136476		Adjusted R Square	0.397558625
7	Standard Error	44.0015595		Standard Error	12002.17094
8	Observations	49		Observations	49
9	VIF	2.164223188		VIF	1.770573388
10					
11	Regression Analysis			Regression Analysis	
12	Top 10% HS and all other X			Acceptance Rate and all other X	
13	<i>Regression Statistics</i>			<i>Regression Statistics</i>	
14	Multiple R	0.701553357		Multiple R	0.724669621
15	R Square	0.492177112		R Square	0.525146059
16	Adjusted R Square	0.458322253		Adjusted R Square	0.49348913
17	Standard Error	9.973219924		Standard Error	0.095165693
18	Observations	49		Observations	49
19	VIF	1.969190488		VIF	2.10591071



Building Good Models

- Include only significant independent variables. Use the fewest necessary to permit adequate interpretation of the dependent variable.
- 10 variables has potentially $2^{10} = 1024$ models!
- As you add more explanatory variables to a model, R^2 increases (even if the variables are irrelevant). However, the Adjusted R^2 could either increase or decrease, thus providing information about the value of additional variables.



Modeling 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 from the model and evaluate adjusted R^2
4. Continue until all variables are significant.



Example: *Banking Data* File

	A	B	C	D	E	F
1	Banking Data					
2						
3	Median	Median Years	Median	Median	Median Household	Average Bank
4	Age	Education	Income	Home Value	Wealth	Balance
5	35.9	14.8	\$91,033	\$183,104	\$220,741	\$38,517
6	37.7	13.8	\$86,748	\$163,843	\$223,152	\$40,618
7	36.8	13.8	\$72,245	\$142,732	\$176,926	\$35,206
8	35.3	13.2	\$70,639	\$145,024	\$166,260	\$33,434
9	35.3	13.2	\$64,879	\$135,951	\$148,868	\$28,162
10	34.8	13.7	\$75,591	\$155,334	\$188,310	\$36,708

Regression Results

	A	B	C	D	E	F	G
1	SUMMARY OUTPUT						
2							
3	<i>Regression Statistics</i>						
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		<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>Significance F</i>	
12	Regression	5	7235179873	1447035975	342.4394584	1.5184E-59	
13	Residual	96	405664271.9	4225669.499			
14	Total	101	7640844145				
15							
16		<i>Coefficients</i>	<i>Standard Error</i>	<i>t Stat</i>	<i>P-value</i>	<i>Lower 95%</i>	<i>Upper 95%</i>
17	Intercept	-10710.64278	4260.976308	-2.513659314	0.013613178	-19168.6137	-2252.671867
18	Age	318.6649626	60.98611242	5.225205378	1.01152E-06	197.6084892	439.721436
19	Education	621.8603472	318.9595184	1.949652891	0.054135369	-11.26927724	1254.989972
20	Income	0.146323453	0.040781001	3.588029937	0.000526666	0.065373808	0.227273099
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	0.052121018	0.096542049

Model After Dropping Home Value

	A	B	C	D	E	F	G
1	SUMMARY OUTPUT						
2							
3	<i>Regression Statistics</i>						
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		<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>Significance F</i>	
12	Regression	4	7232255152	1808063788	429.2386497	9.68905E-61	
13	Residual	97	408588992.5	4212257.655			
14	Total	101	7640844145				
15							
16		<i>Coefficients</i>	<i>Standard Error</i>	<i>t Stat</i>	<i>P-value</i>	<i>Lower 95%</i>	<i>Upper 95%</i>
17	Intercept	-12432.45673	3718.674319	-3.343249681	0.001177705	-19812.99569	-5051.917773
18	Age	325.0652837	60.40284468	5.381622098	5.1267E-07	205.1823604	444.9482071
19	Education	773.3800418	261.4330936	2.958233142	0.003886994	254.5077323	1292.252351
20	Income	0.159747379	0.037393587	4.272052794	4.52422E-05	0.085531461	0.233963297
21	Wealth	0.072988791	0.011054665	6.602532898	2.16051E-09	0.051048341	0.094929241



Stepwise Regression

- Stepwise regression is a search process that adds or deletes variables at each step until no changes can improve the model.

Stepwise Regression Results – Forward Selection

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	
1	Banking Data General Stepwise																
2	Table of Results for General Stepwise																
3																	
4	Income entered.																
5																	
6																	
7	Regression	1	6920338342	6920338342	960.4833602	4.43329E-53											
8	Residual	100	720505802.5	7205058.025													
9	Total	101	7640844145														
10																	
11																	
12	Intercept	4020.251548	723.8864893	5.553704355	2.31287E-07	2584.081409	5456.421688										
13	Income	0.427519104	0.013794647	30.99166598	4.43329E-53	0.400150917	0.454887291										
14																	
15																	
16	Wealth entered.																
17																	
18																	
19	Regression	2	7088382281	3544191140	635.1115723	3.37651E-57											
20	Residual	99	552461863.7	5580422.866													
21	Total	101	7640844145														
22																	
23																	
24	Intercept	6279.906804	758.5622772	8.278696414	6.04778E-13	4774.754713	7785.058894										
25	Income	0.231792518	0.037676923	6.152108514	1.62772E-08	0.157033331	0.306551705										
26	Wealth	0.066901189	0.012191467	5.487542179	3.13968E-07	0.042710674	0.091091703										
27																	
28	Age entered.																
29																	

Best Subsets Regression

- Evaluates all possible models or those containing a fixed number of independent variables to identify the best.
- Selects appropriate models based on C_p
- *PHStat* output

10	Model	Cp	k+1	R Square	Adj. R Square	Std. Error
11	X1	1132.021	2	0.319753	0.312950268	7209.482
12	X2	1153.467	2	0.307893	0.300971473	7272.06
13	X3	72.5069	2	0.905703	0.90476041	2684.224
14	X4	648.154	2	0.587349	0.583222723	5615.158
15	X5	82.72217	2	0.900054	0.899054494	2763.462
16	X1X2	744.972	3	0.534911	0.525515617	5991.298
17	X1X3	45.46647	3	0.921764	0.920183276	2457.293
18	X1X4	496.0324	3	0.672584	0.665969731	5026.931
19	X1X5	50.6053	3	0.918922	0.917283904	2501.526
20	X2X3	74.36336	3	0.905783	0.903879382	2696.611
21	X2X4	648.0156	3	0.588532	0.58021935	5635.354
22	X2X5	56.89208	3	0.915445	0.913736844	2554.599
23	X3X4	74.06647	3	0.905947	0.904046887	2694.26
24	X3X5	34.73949	3	0.927696	0.926235544	2362.292
25	X4X5	46.7294	3	0.921065	0.919470723	2468.237
26	X1X2X3	46.14721	4	0.922493	0.920120756	2458.255
27	X1X2X4	497.0943	4	0.673103	0.663095931	5048.509
28	X1X2X5	20.88464	4	0.936465	0.93451958	2225.695
29	X1X3X4	47.11486	4	0.921958	0.919569227	2466.727
30	X1X3X5	11.4155	4	0.941701	0.939916675	2131.999
31	X1X4X5	20.80774	4	0.936507	0.934563412	2224.95
32	X2X3X4	76.06601	4	0.905947	0.903068036	2707.968
33	X2X3X5	31.56207	4	0.93056	0.928433803	2326.826
34	X2X4X5	46.41904	4	0.922343	0.919965823	2460.638
35	X3X4X5	30.37914	4	0.931214	0.929108031	2315.84
36	X1X2X3X4	48.13093	5	0.922502	0.919306637	2470.75
37	X1X2X3X5	4.692132	5	0.946526	0.944320547	2052.379
38	X1X2X4X5	16.87396	5	0.939789	0.937305731	2177.83
39	X1X3X4X5	7.801146	5	0.944806	0.942530244	2085.113
40	X2X3X4X5	31.30277	5	0.931809	0.928997006	2317.652
41	X1X2X3X4X5	6	6	0.946908	0.944143263	2055.643





Art of Model Building

- Independent variables should make sense in the model
- Logic and theory should guide model development
- Sampling error can affect results
- Good models are **parsimonious** – as simple as possible



Models with Categorical Independent Variables

- Examples

- Gender (male, female)
- College graduate (no, 2-year degree, 4-year degree, postgraduate degree)
- Own home (yes, no)



Example: *Employee Salaries*

- How do age and MBA degree affect employee salaries?

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

where

Y = salary, X_1 = age, X_2 = MBA indicator (0 = No; 1 = Yes)

	A	B	C	D
1	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
9	6	\$ 57,119	57	No
10	7	\$ 48,907	45	No
11	8	\$ 34,301	32	No
12	9	\$ 31,104	25	No
13	10	\$ 60,054	57	No

Initial Regression Model

	A	B	C	D	E	F	G
1	SUMMARY OUTPUT						
2							
3	<i>Regression Statistics</i>						
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		<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>Significance F</i>	
12	Regression	2	5591651177	2795825589	323.0353318	6.05341E-22	
13	Residual	32	276955521.7	8654860.054			
14	Total	34	5868606699				
15							
16		<i>Coefficients</i>	<i>Standard Error</i>	<i>t Stat</i>	<i>P-value</i>	<i>Lower 95%</i>	<i>Upper 95%</i>
17	Intercept	893.5875971	1824.575283	0.489751015	0.627650922	-2822.950618	4610.125812
18	Age	1044.146043	42.14128238	24.77727265	1.8878E-22	958.3070603	1129.985026
19	MBA	14767.23159	1351.801764	10.92411031	2.49752E-12	12013.70151	17520.76166



Model

- $\text{Salary} = 893.59 + 1044.15 \text{ Age} + 14767.23 \text{ MBA}$
 - No MBA: $\text{Salary} = 893.59 + 1044.15 \text{ Age}$
 - MBA: $\text{Salary} = 15660.82 + 1044.15 \text{ Age}$
- The models suggest that the rate of salary increase for age is the same for both groups. However, individuals with MBAs might earn relatively higher salaries as they get older. In other words, the slope of *Age* may depend on the value of *MBA*. Such a dependence is called an **interaction**.



Interaction Model

- $Y = b_0 + b_1\text{Age} + b_2\text{MBA} + b_3\text{Age}*\text{MBA} + e$

	A	B	C	D	E
1	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
8	5	\$ 36,807	32	0	0

Results With Interaction Term

	A	B	C	D	E	F	G
1	SUMMARY OUTPUT						
2							
3	<i>Regression Statistics</i>						
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		<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>Significance F</i>	
12	Regression	3	5743939086	1914646362	476.098288	5.31397E-26	
13	Residual	31	124667613.2	4021535.91			
14	Total	34	5868606699				
15							
16		<i>Coefficients</i>	<i>Standard Error</i>	<i>t Stat</i>	<i>P-value</i>	<i>Lower 95%</i>	<i>Upper 95%</i>
17	Intercept	3902.509386	1336.39766	2.920170772	0.006467654	1176.908399	6628.110372
18	Age	971.3090382	31.06887722	31.26308786	5.23658E-25	907.9436456	1034.674431
19	MBA	-2971.080074	3026.24236	-0.98177202	0.333812767	-9143.142034	3200.981887
20	Interaction	501.8483604	81.55221742	6.153705887	7.9295E-07	335.5215171	668.1752038

Final Model

	A	B	C	D	E	F	G
1	SUMMARY OUTPUT						
2							
3	<i>Regression Statistics</i>						
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		<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>Significance F</i>	
12	Regression	2	5740062823	2870031411	714.4720368	2.80713E-27	
13	Residual	32	128543876.4	4016996.136			
14	Total	34	5868606699				
15							
16		<i>Coefficients</i>	<i>Standard Error</i>	<i>t Stat</i>	<i>P-value</i>	<i>Lower 95%</i>	<i>Upper 95%</i>
17	Intercept	3323.109564	1198.353141	2.773063675	0.009184278	882.1441051	5764.075022
18	Age	984.2455409	28.12039088	35.00113299	4.40388E-27	926.9661794	1041.524902
19	Interaction	425.5845915	24.81794165	17.14826304	1.08793E-17	375.0320988	476.1370841



Model Results

- $\text{Salary} = 3323.11 + 984.25 \text{ Age} + 425.58 \text{ MBA} * \text{Age}$
 - No MBA: $\text{Salary} = 3323.11 + 984.25 \text{ Age} + 425.58 (0) * \text{Age}$
$$= 3323.11 + 984.25 \text{ Age}$$
 - MBA: $\text{Salary} = 3323.11 + 984.25 \text{ Age} + 425.58 (1) * \text{Age}$
$$= 3323.11 + 1409.83 \text{ Age}$$



Categorical Variables With More Than Two Levels

- For $k > 2$ levels, add $k-1$ additional variables.
- Example: The Excel file *Surface Finish.xls* 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. The engineer who collected the data is interested in predicting the surface finish as a function of RPM and type of tool.



Model

- $Y = b_0 + b_1X_1 + b_2X_2 + b_3X_3 + b_4X_4 + e$

where

Y = surface finish

X1 = RPM

X2 = tool type B

X3 = tool type C

X4 = tool type D

Tool Type	X ₂	X ₃	X ₄
A	0	0	0
B	1	0	0
C	0	1	0
D	0	0	1

Tool Type A: $Y = \beta_0 + \beta_1X_1 + \varepsilon$

Tool Type B: $Y = \beta_0 + \beta_1X_1 + \beta_2 + \varepsilon$

Tool Type C: $Y = \beta_0 + \beta_1X_1 + \beta_3 + \varepsilon$

Tool Type D: $Y = \beta_0 + \beta_1X_1 + \beta_4 + \varepsilon$



Data Matrix

	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

Regression Results

	A	B	C	D	E	F	G
1	SUMMARY OUTPUT						
2							
3	<i>Regression Statistics</i>						
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		<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>Significance F</i>	
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		<i>Coefficients</i>	<i>Standard Error</i>	<i>t Stat</i>	<i>P-value</i>	<i>Lower 95%</i>	<i>Upper 95%</i>
17	Intercept	24.49437244	2.473298088	9.903526211	5.73134E-11	19.4432239	29.54552099
18	RPM	0.097760627	0.010399996	9.400064035	1.89415E-10	0.076521002	0.119000251
19	Type B	-13.31056756	0.487142953	-27.32374035	9.37003E-23	-14.30544619	-12.31568893
20	Type C	-20.487	0.487088553	-42.06011387	3.12134E-28	-21.48176753	-19.49223247
21	Type D	-26.03674519	0.596886375	-43.62094073	1.06415E-28	-27.25574979	-24.81774059



Results

- Surface Finish = $24.49 + 0.098 \text{ RPM} - 13.31 \text{ Type B} - 20.49 \text{ Type C} - 26.04 \text{ Type D}$
 - Tool A: Surface Finish = $24.49 + 0.098 \text{ RPM} - 13.31(0) - 20.49(0) - 26.04(0) = 24.49 + 0.098 \text{ RPM}$
 - Tool B: Surface Finish = $24.49 + 0.098 \text{ RPM} - 13.3(1) - 20.49(0) - 26.04(0) = 11.18 + 0.098 \text{ RPM}$
 - Tool C: Surface Finish = $24.49 + 0.098 \text{ RPM} - 13.31(0) - 20.49(1) - 26.04(0) = 4.00 + 0.098 \text{ RPM}$
 - Tool D: Surface Finish = $24.49 + 0.098 \text{ RPM} - 13.3(0) - 20.49(0) - 26.04(1) = -1.55 + 0.098 \text{ RPM}$



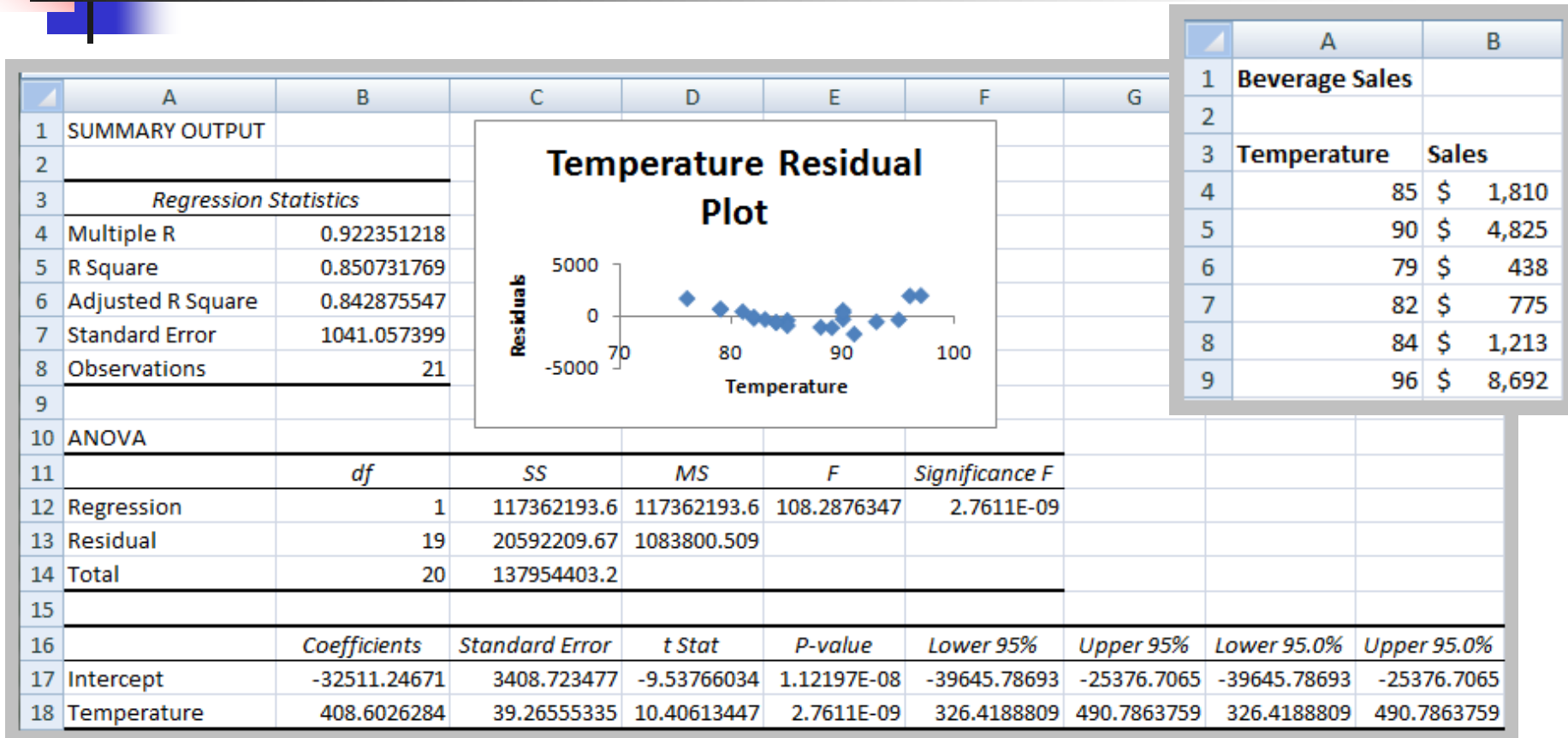
Models with Nonlinear Terms

- Interaction terms ($X_1 * X_2$) or nonlinear variables (X_2^2) do not make a model nonlinear; linear regression still applies because the model is linear in the parameters:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + \beta_4 X_2^2 + \varepsilon$$

- However, if the parameters are nonlinear ($Y = aX^b$), then you must try to transform the model or use a nonlinear regression technique.

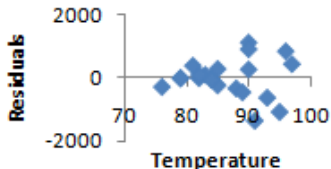
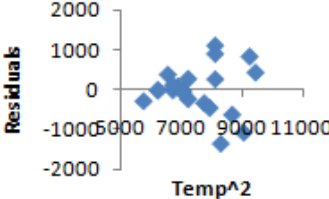
Example: *Beverage Sales*

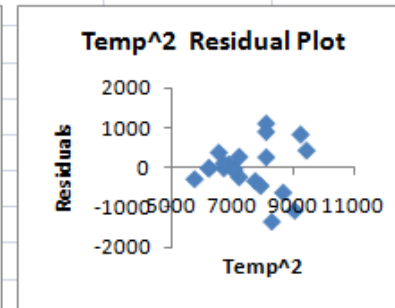
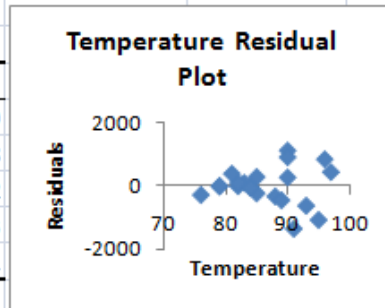


Residual plot suggests nonlinearity

Curvilinear Regression Model

- 2nd order polynomial: $Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \varepsilon$

	A	B	C	D	E	F	G	H	I
1	SUMMARY OUTPUT		<div><div>Temperature Residual Plot</div></div> <div><div>Temp^2 Residual Plot</div></div>						
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%	Lower 95.0%	Upper 95.0%
17	Intercept	142850.3406	30575.70155	4.672021683	0.000189738	78613.17542	207087.5058	78613.17542	207087.5058
18	Temperature	-3643.171723	705.2304165	-5.165931075	6.492E-05	-5124.805846	-2161.5376	-5124.805846	-2161.5376
19	Temp^2	23.30035581	4.053196314	5.748637374	1.89343E-05	14.78490636	31.81580527	14.78490636	31.81580527





Underlying Theory

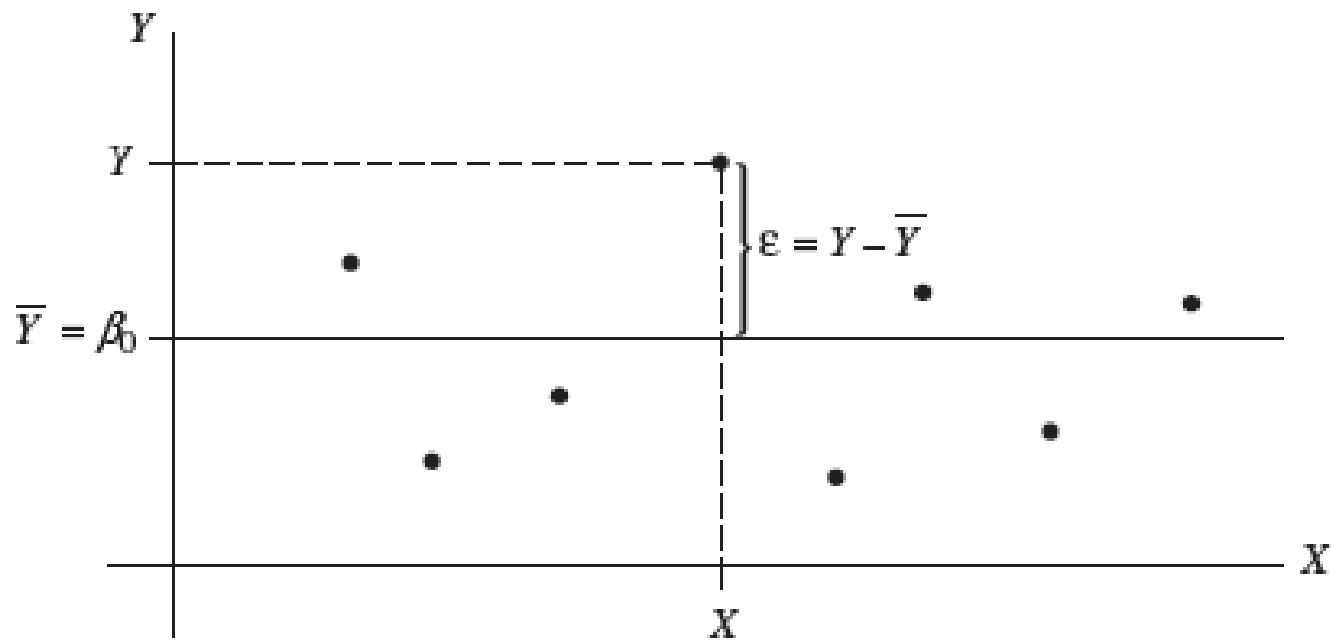
Three measures of variation:

1. Variation between the observations and the mean: $(Y - \bar{Y})$. The sum of squares of these terms is SST.
2. Variation between the predicted values using the regression line and the mean, which is explained by the regression line: $(\hat{Y} - \bar{Y})$. The sums of squares of these terms is SSR.
3. Variation between the individual observations and the predicted values, which is the remaining unexplained variation: $(Y - \hat{Y})$. The sums of squares of these terms is SSE.

Key fact: $SST = SSR + SSE$



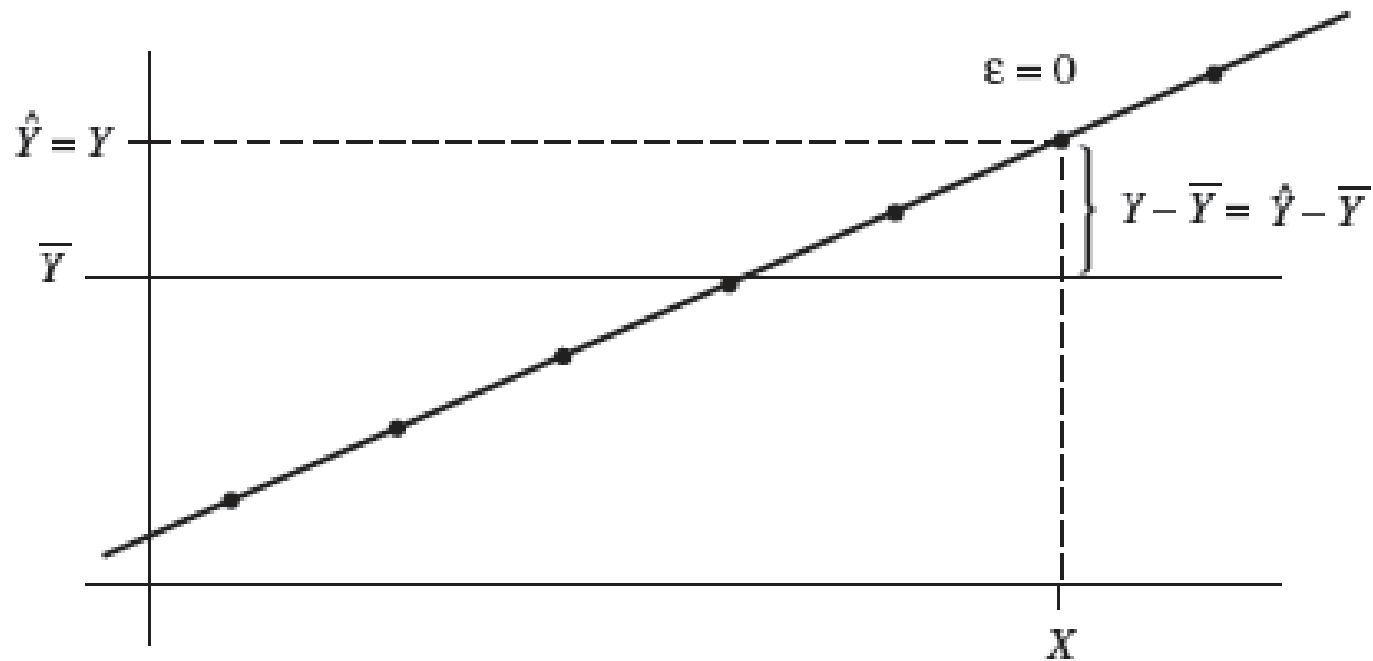
SST = SSE



(a) No relationship with independent variable

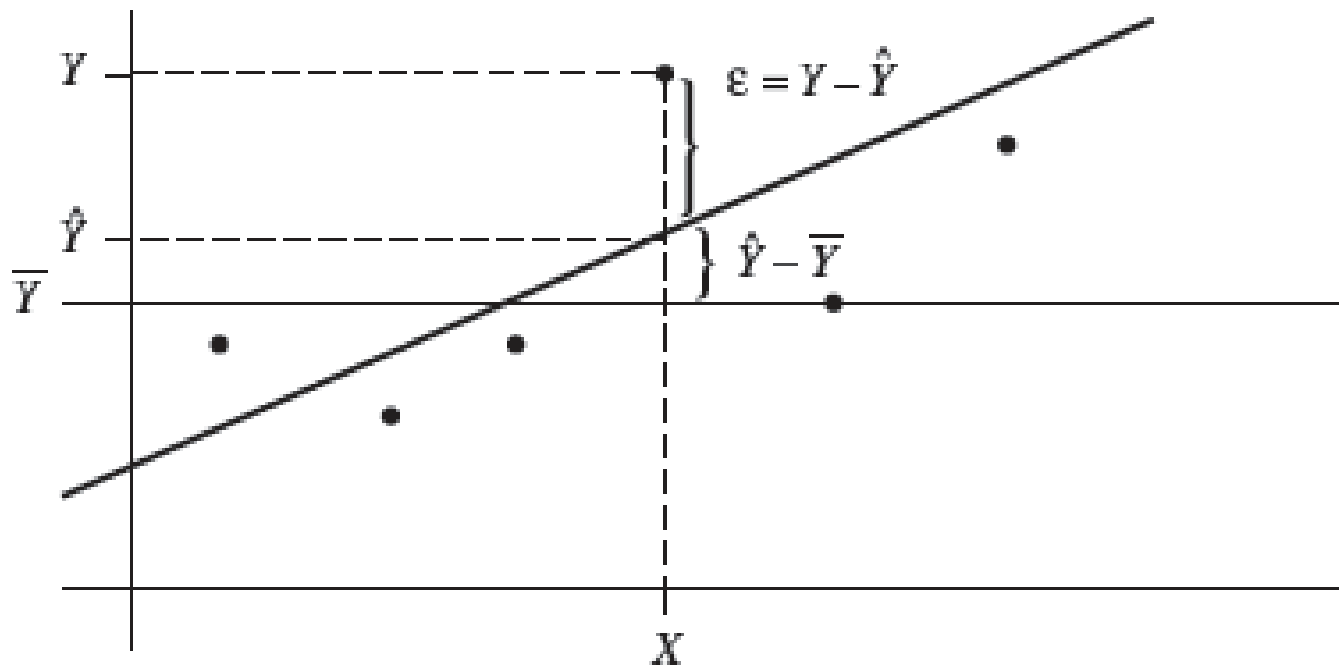


SST = SSR



(b) Perfect relationship


$$SST = SSR + SSE$$



(c) Imperfect relationship