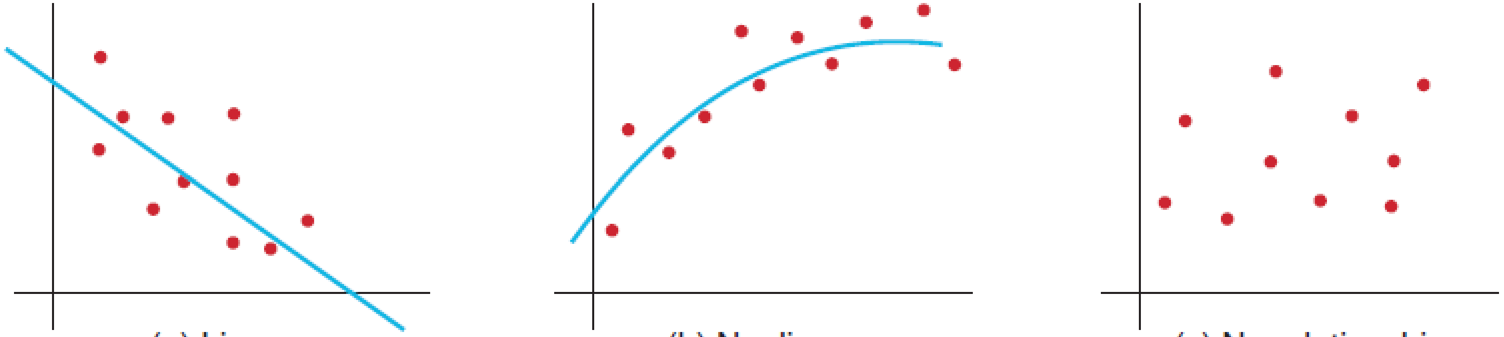
Regression Analysis

Regression Analysis

- It is a tool for building mathematical and statistical models that characterize relationships between
  - A dependent variable (ratio)
  - And one or more independent variables (ratio or categorical)
- The relationship may be linear, non linear or no relationship at all.



(a) Linear

(b) Nonlinear

(c) No relationship

Modeling Relationships and Trends in Data

- Step 1: Create a chart of the data to understand it
  - For cross sectional data – scatter chart
  - For time series data – line chart
- Step 2: To choose the appropriate type of functional relationship to incorporate into an analytical model.

Mathematical Functions

- Linear function: $y = a + bx$
  - Shows steady increase or decrease over the range of x
  - Simplest type of function, easy to understand
- Logarithmic function: $y = \ln(x)$
  - Rate of change of variable increases or decreases quickly and then levels out
  - Marketing model
- Polynomial function: $y=ax^2 + bx + c$
  - Revenue models that incorporate price elasticity are often polynomial function

Mathematical Functions

- Power Function: $y = ax^b$
  - Define phenomena that increase at a specific rate
  - Learning curves that express improving time in performing a task
- Exponential Function: $y=ab^x$
  - Y rises or falls at constantly increasing rates
  - The perceived brightness of a lightbulb grows at a decreasing rate as the wattage increases.

Modeling Price –Demand Model

A market research study has collected data on sales volumes for different levels of pricing of a particular product.

Using Price Sales Data Excel file.

The resulting model is sales = 20,512 – 9.5116 x price

This model can be used as the demand function in other marketing or financial analyses.

Predicting Crude Oil Prices

Crude Oil Prices Excel file shows a chart of historical data on crude oil prices on the first Friday of each month from January 2008.

Try to fit it to:

Exponential, logarithmic, polynomial (second order), polynomial (third order), power

Find the best fitting model.

Types of Regression Models

- Based on type of independent variable
  - Regression model for cross sectional data
  - Regression model for time series data
- Based on number of independent variable
  - Simple regression involves a single independent variable.
  - Multiple regression involves two or more independent variable.

Simple Linear Regression

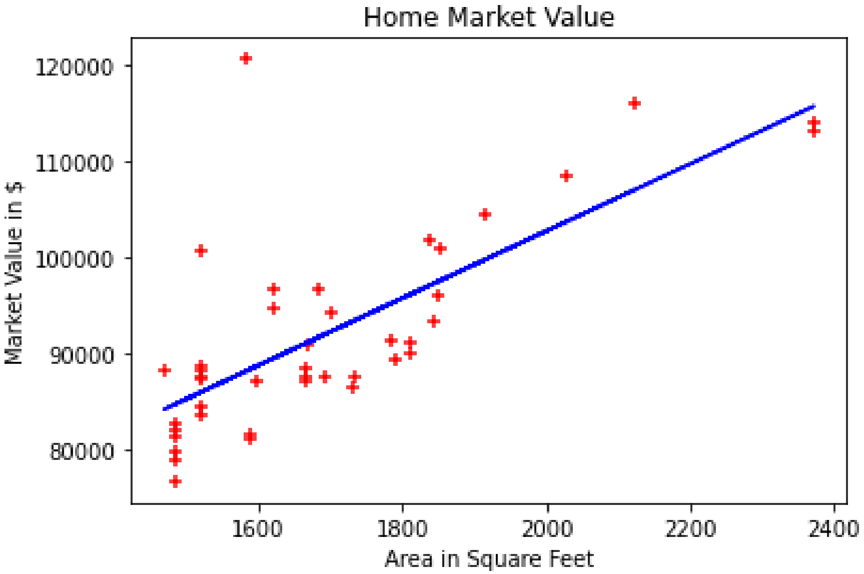
- It is a statistical technique used for finding a linear relationship between one independent variable and one dependent variable.
  - X: independent / explanatory/ predictor variable or feature.
  - Y: dependent / response / outcome variable.
- We can establish that change in the value of the outcome variable (Y) is associated with change in the value of feature X.

Home Market Value

```
import pandas as pd
import matplotlib.pyplot as plt
from sklearn import linear_model

df=pd.read_excel("F:\python\Home Market Value.xlsx",header=2)
print(df)
plt.title('Home Market Value')
plt.xlabel('Area in Square Feet')
plt.ylabel('Market Value in $')
plt.scatter(df['Square Feet'],df['Market Value'],color='red',marker='+')

reg=linear_model.LinearRegression()
reg.fit(df[['Square Feet']].values,df[['Market Value']])
print(reg.predict([[1800]]))
print(reg.coef_)
print(reg.intercept_)
plt.plot(df['Square Feet'],reg.predict(df[['Square Feet']]),color='blue')
```

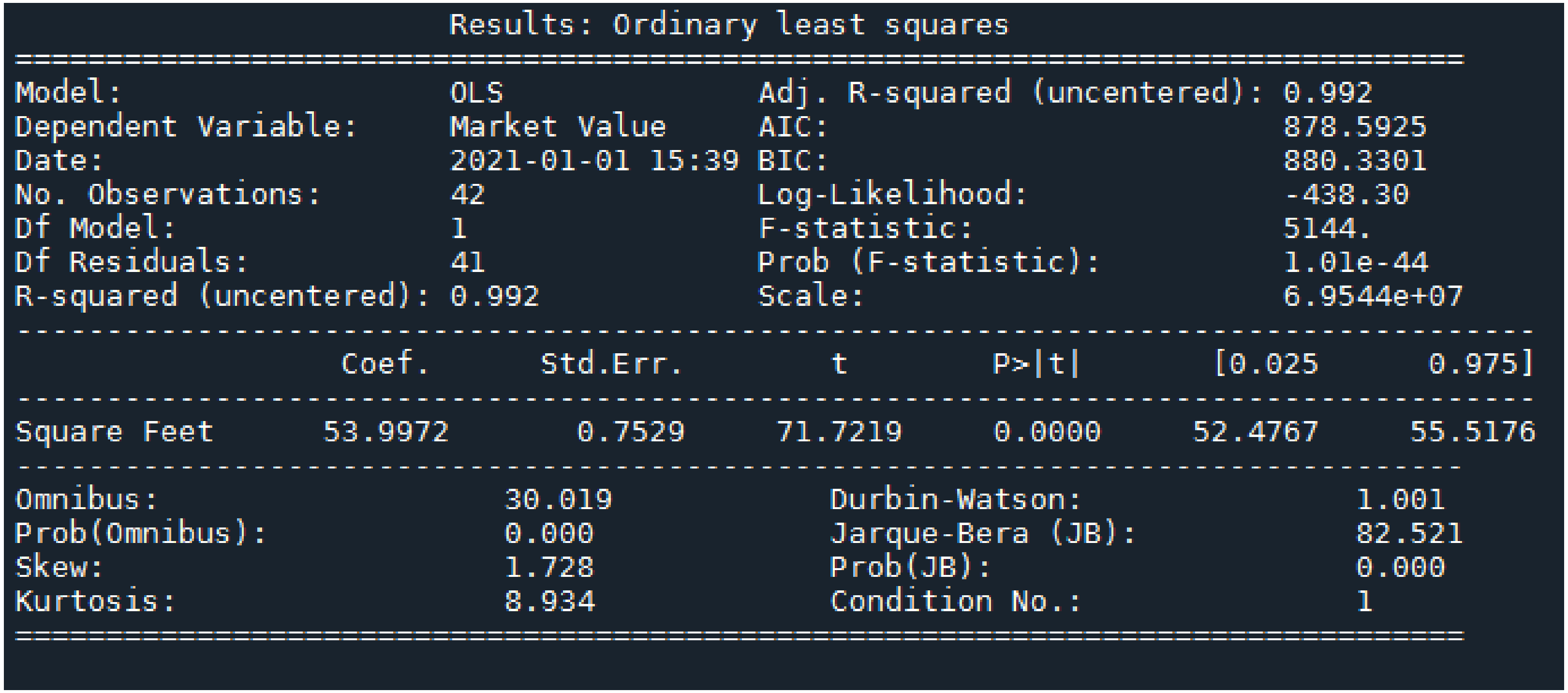


Least Squares Regression

- The mathematical basis for the best fitting regression line.
- e_i is the observed errors (residual) associated with estimating the value of dependent variable.
- The best fitting line should minimize this error.
- Excel function INTERCEPT(known y’s, known x’s) and SLOPE(known y’s, known x’s) are used to find least squares coefficients.
- TREND(known y’s, known x’s,new x’s) is used to estimate Y for any value of X.
- Example: Home Market Value excel

Regression Summary

```
import statsmodels.api as sm
model=sm.OLS(df['Market Value'],df['Square Feet']).fit()
print(model.summary2())
```



```
                        Results: Ordinary least squares
================================================================================
Model:                  OLS              Adj. R-squared (uncentered): 0.992
Dependent Variable:     Market Value     AIC:                         878.5925
Date:                   2021-01-01 15:39 BIC:                         880.3301
No. Observations:       42               Log-Likelihood:              -438.30
Df Model:               1                F-statistic:                 5144.
Df Residuals:           41               Prob (F-statistic):          1.01e-44
R-squared (uncentered): 0.992            Scale:                       6.9544e+07
------------------------------------------------------------------------------------
                  Coef.      Std.Err.        t        P>|t|       [0.025      0.975]
------------------------------------------------------------------------------------
Square Feet      53.9972       0.7529     71.7219     0.0000     52.4767     55.5176
--------------------------------------------------------------------------------
Omnibus:                   30.019            Durbin-Watson:               1.001
Prob(Omnibus):             0.000             Jarque-Bera (JB):            82.521
Skew:                      1.728             Prob(JB):                    0.000
Kurtosis:                  8.934             Condition No.:               1
================================================================================
```

Regression Statistics

- Multiple R: sample correlation coefficient r. Values of r ranges from -1 to 1.
- R-squared (R^2): coefficient of determination. The value of R^2 is between 0 and 1.
- Adjusted R square: by incorporating the sample size and the number of explanatory variables in the model.
- Standard error: standard error of estimate. the variability of the observed Y values from the predicted values.

Analysis of Variance

- ANOVA is commonly applied to regression to test for significance of regression.
- It is hypothesis test whether the regression coefficient β_1 is zero:
  - H_0: $\beta_1=0$
  - H_1: $\beta_1<>0$
- If null hypothesis is rejected, slope is not zero and independent variable is statistically significant.
- Example: ANOVA test for Home Market Value.

Confidence Interval for Regression Coefficient

- Confidence intervals (Lower 95% and Upper95% values in the output)are used to test hypotheses about the regression coefficient.
- We can use them to test the hypotheses that the regression coefficient equal some value other than zero.
  - H_0: β_1=B_1
  - H_1: β_1<>B_1
- If B_1 falls within the confidence interval for the slope then we reject the null hypothesis otherwise we fail to reject.

Residual Analysis

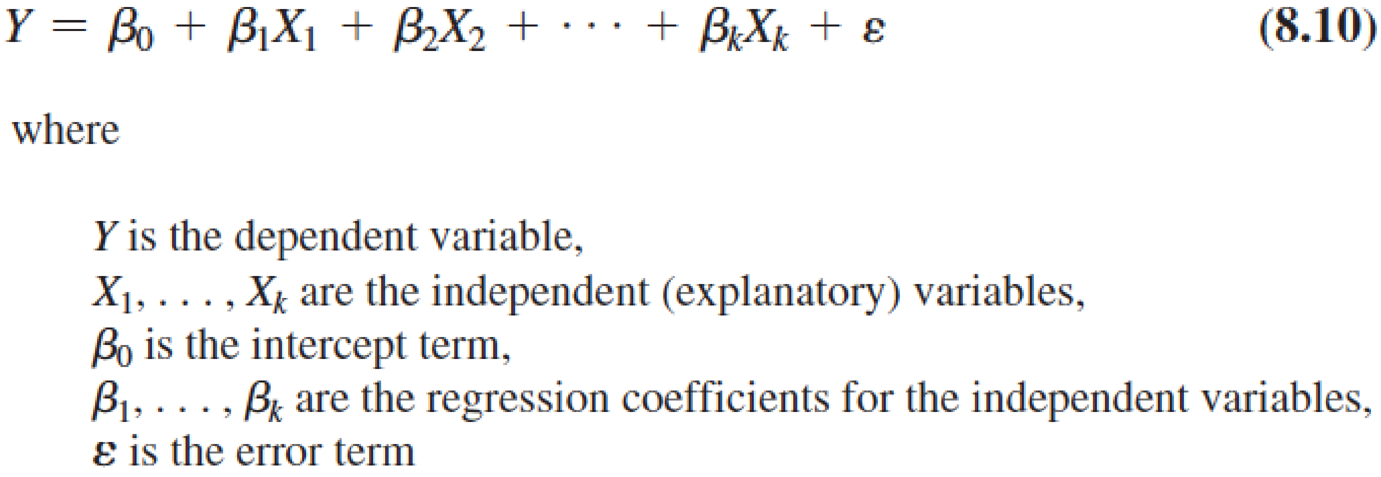
- The residual output includes for each observation, the predicted value, the residual and standard residual.
  - The residual is the difference between the actual value and predicted value.
  - Standard residuals are residuals divided by their standard deviation.
- Standard residuals are useful in checking regression assumptions and to detect outliers that may bias the result.

Checking Assumptions

- Linearity: examined by residual plot or scatter diagram.
- Normality of Errors: standard residual is normally distributed.
- Homoscedasticity: variation about the regression line is constant for all values of the independent variable.
- Independence of errors: residuals should be independent for each value of the independent variable.

Multiple Linear Regression

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



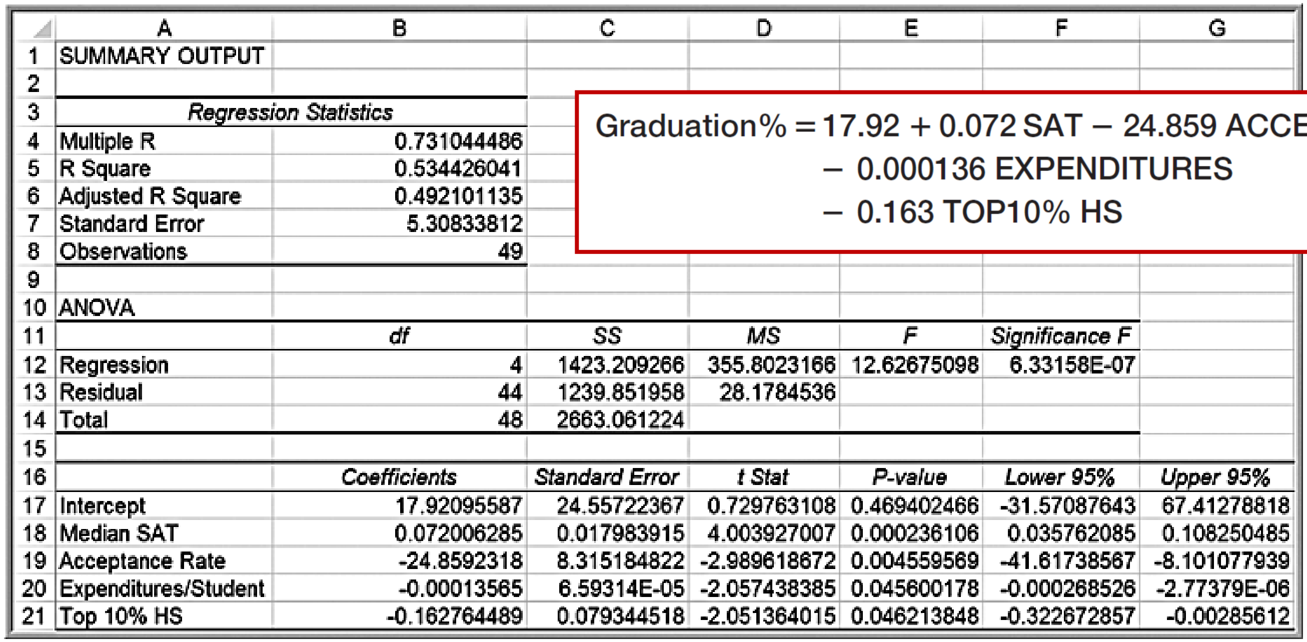
$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k + \varepsilon \qquad \textbf{(8.10)}$$

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

- Colleges and Universities Excel File

Example



	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

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

Building Good Regression Model

- Construct a model with all available independent variables. Check the significance of the independent variables by examining the p-value.
- Identify the independent variable having the largest p-value that exceeds the chosen level of significance.
- Remove the variable identified in step2 from the model and evaluate adjusted R^2.
- Continue until all variables are significant.

Correlation and Multicollinearity

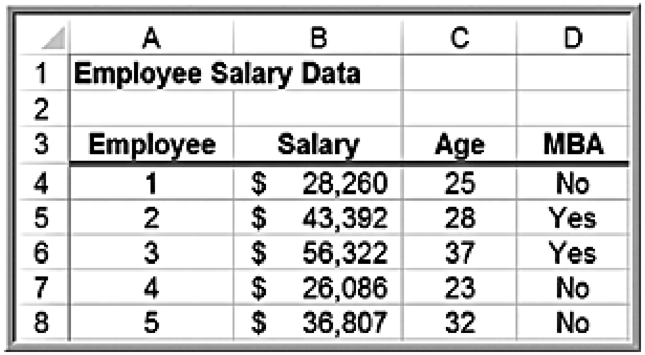
- The strong correlation between dependent and independent variables represent strong linear relationship between them.
- However strong correlations among the independent variables can be problematic.
- Multicollinearity is a condition occurring when two or more independent variables in the same regression model contain high levels of the same information and consequently are strongly correlated with one another and can predict each other better than the dependent variable.

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.

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

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

where

Y = salary

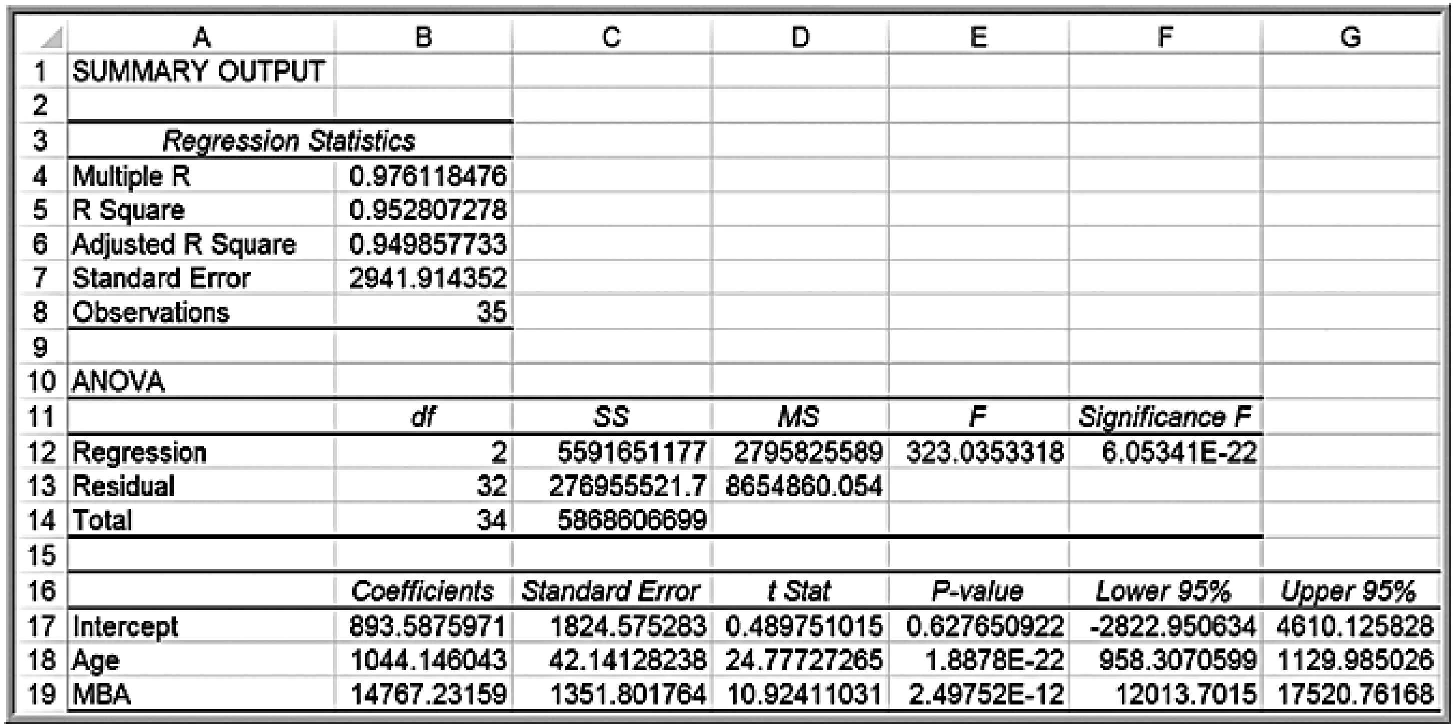
X_1 = age

X_2 = MBA indicator (0 or 1)

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

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

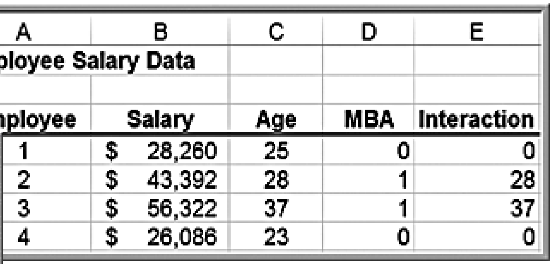
Interactions

- 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.

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

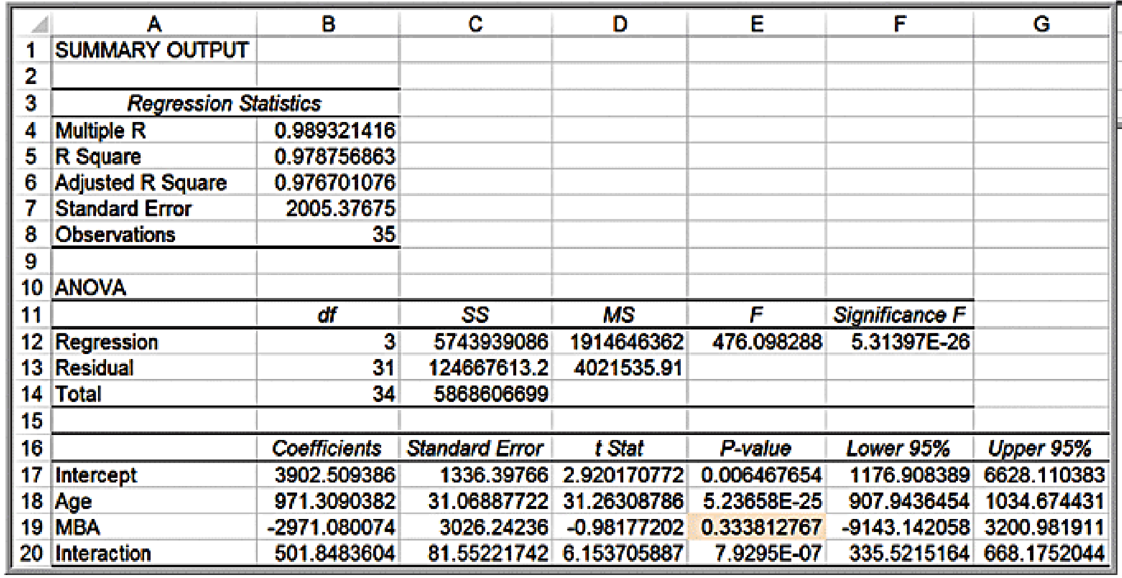
Incorporating Interaction Terms in a Regression Model

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



Employee Salary Data

Employee	Salary	Age	MBA	Interaction
1	$ 28,260	25	0	0
2	$ 43,392	28	1	28
3	$ 56,322	37	1	37
4	$ 26,086	23	0	0



SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.989321416
R Square	0.978756863
Adjusted R Square	0.976701076
Standard Error	2005.37675
Observations	35

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	3	5743939086	1914646362	476.098288	5.31397E-26
Residual	31	124667613.2	4021535.91		
Total	34	5868606699			

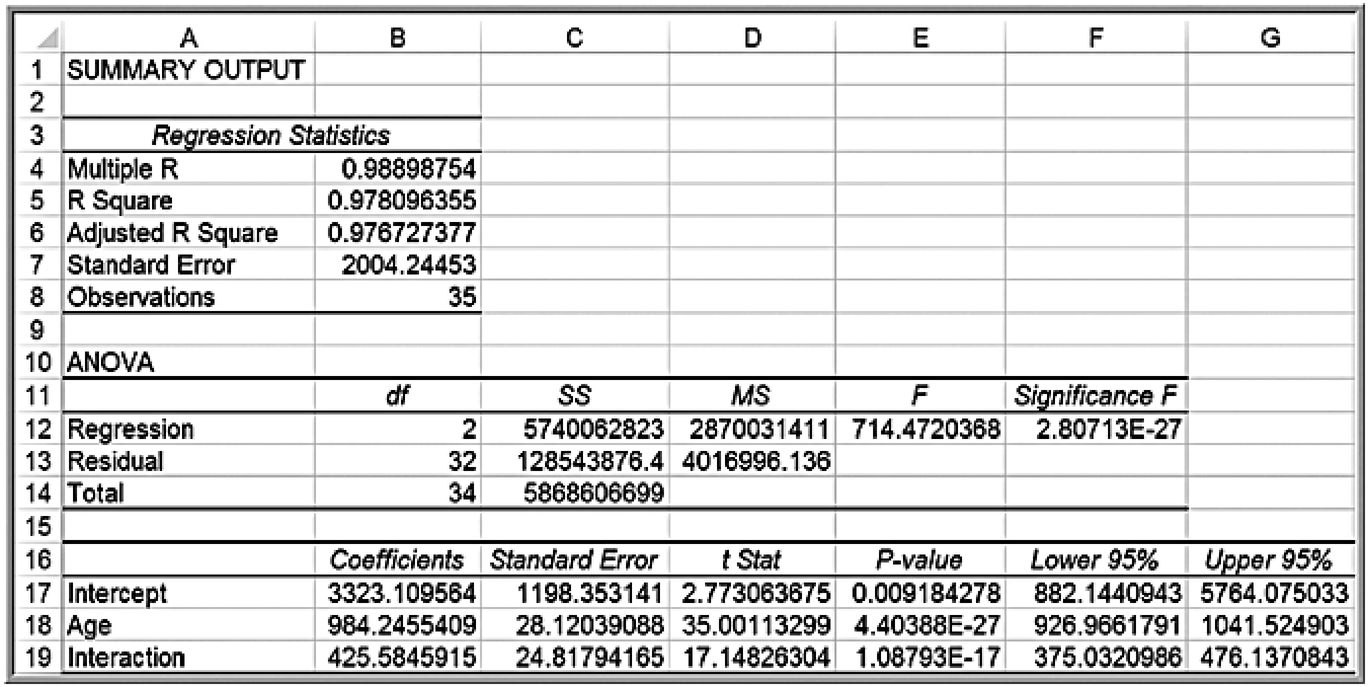
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	3902.509386	1336.39766	2.920170772	0.006467654	1176.908389	6628.110383
Age	971.3090382	31.06887722	31.26308786	5.23658E-25	907.9436454	1034.674431
MBA	-2971.080074	3026.24236	-0.98177202	0.333812767	-9143.142058	3200.981911
Interaction	501.8483604	81.55221742	6.153705887	7.9295E-07	335.5215164	668.1752044

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

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



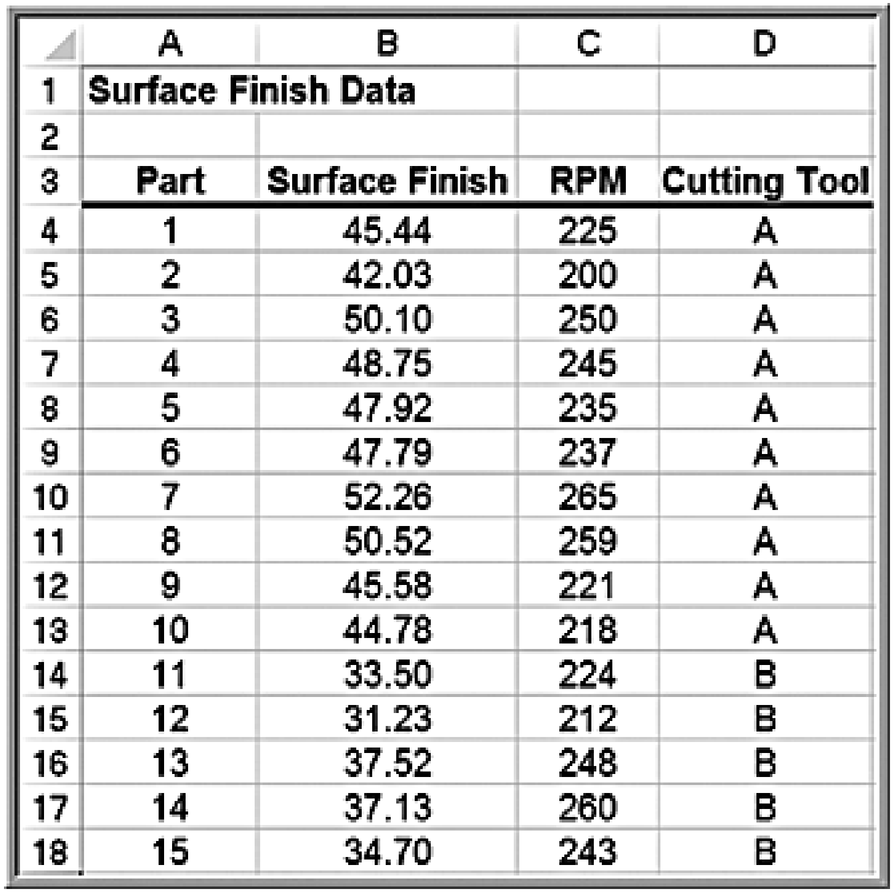
	A	B	C	D	E	F	G
1	SUMMARY OUTPUT						
2							
3	*Regression Statistics*						
4	Multiple R	0.98898754					
5	R Square	0.978096355					
6	Adjusted R Square	0.976727377					
7	Standard Error	2004.24453					
8	Observations	35					
9							
10	ANOVA						
11		*df*	*SS*	*MS*	*F*	*Significance F*	
12	Regression	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.

A Regression Model with Multiple Levels of Categorical Variables

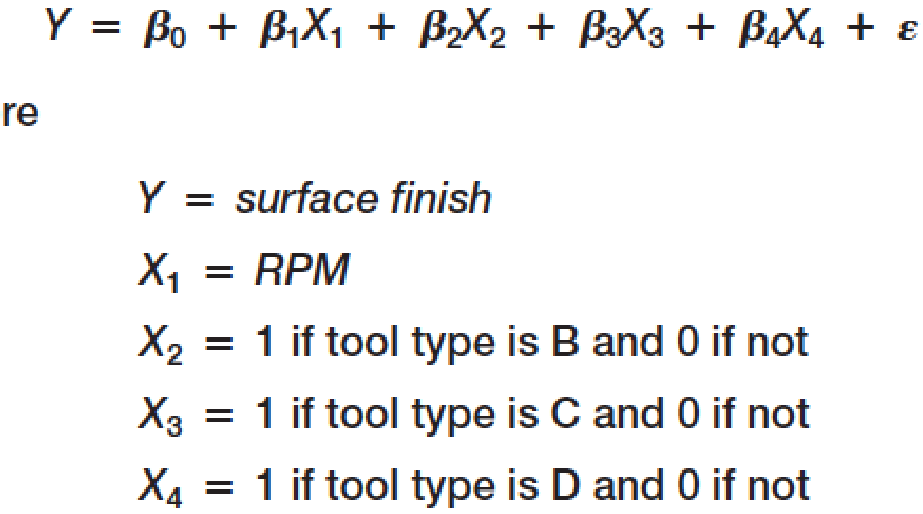
- The Excel file *Surface Finish* provides measurements of the surface finish of 35 parts produced on a lathe, along with the revolutions per minute (RPM) of the spindle and one of four types of cutting tools used.



	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

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 = *surface finish*

X_1 = *RPM*

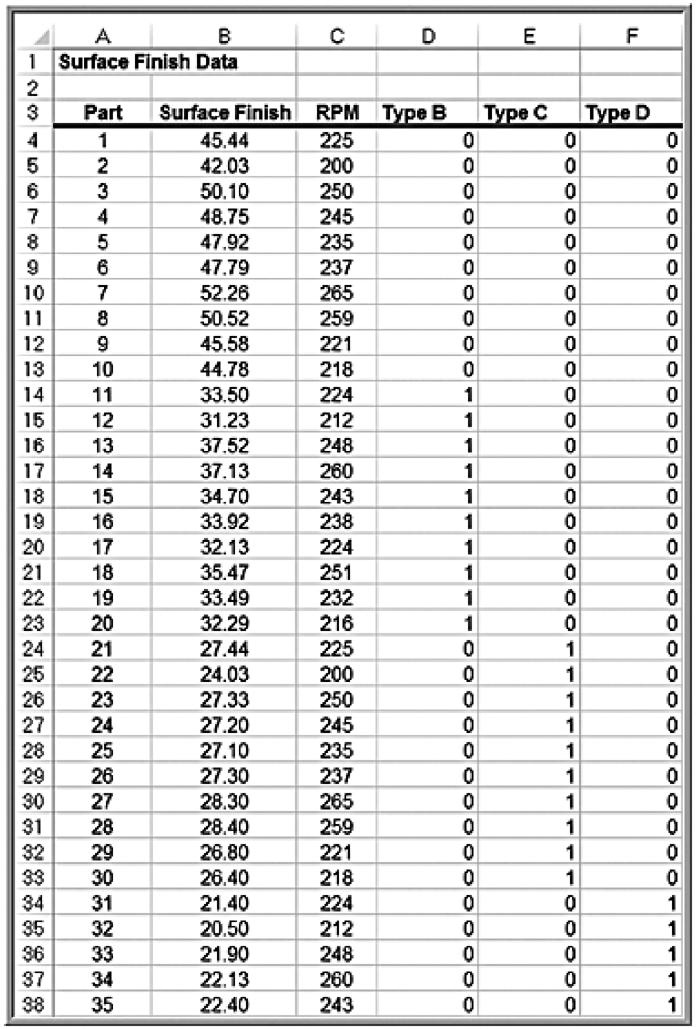
X_2 = 1 if tool type is B and 0 if not

X_3 = 1 if tool type is C and 0 if not

X_4 = 1 if tool type is D and 0 if not

Continued

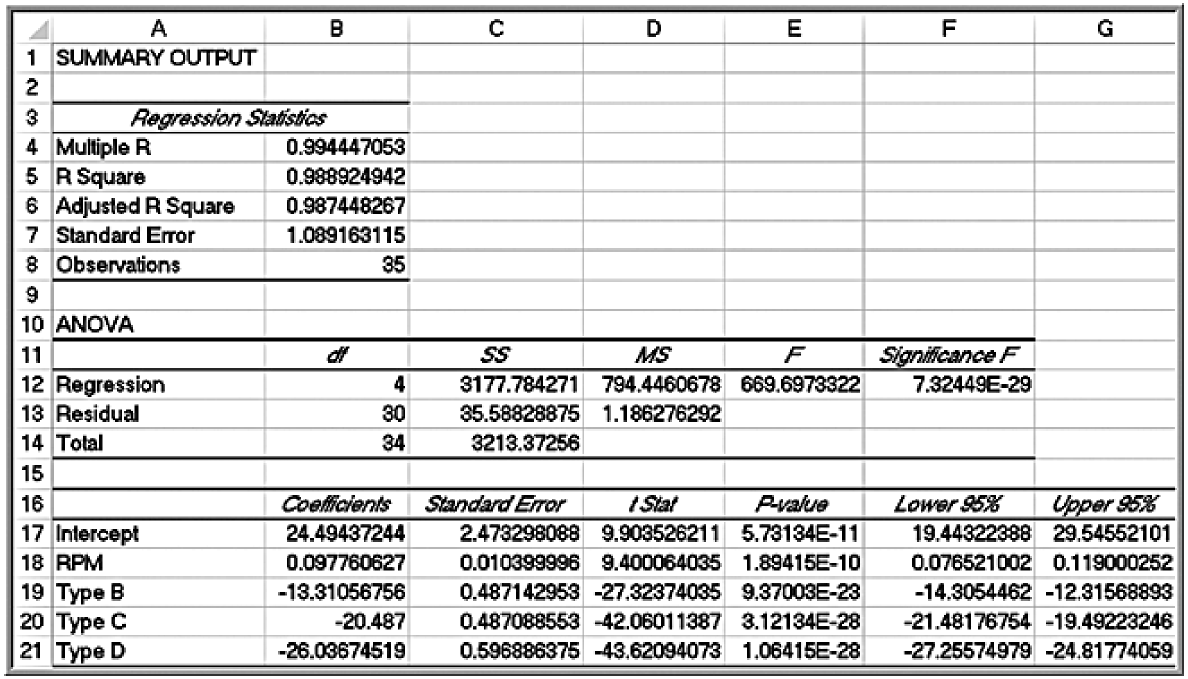
- Add 3 columns to the data, one for each of the tool type variables



	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

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

- Linear regression models are not appropriate for:
  - A scatter chart of the data might show a nonlinear relationship.
  - The residuals for a linear fit might result in a non linear pattern.
- A non linear model, for example, second order polynomial is used.

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

Curvilinear Regression Model

- In this model, β_1 represents the linear effect of X on Y
- β_2 represents the curvilinear effect.
- Though not linear regression model, it is still linear in the parameters.
- Least square to estimate the regression.

Beverages Sales Example

- The excel fil Beverages Sales provides data on the sales of cold beverages at a small restaurant with a large outdoor patio during the summer month.
- The owner has observed that sales tend to increase on hotter days.
- Residual plot: The U shape of the residual plot suggests that a linear relationship is not appropriate.
- Curvilinear Regression Model: add a column by squaring the temperature.

$$\text{sales} = 142{,}850 – 3{,}643.17 \text{ x temperature} + 23.3 \text{ x temperature}^2$$

Summary

- Regression is one of the most popular supervised learning algorithms in predictive analytics.
- A regression model requires the knowledge of both the outcome and the feature variable in the training set.