

IIT KHARAGPUR



NPTEL ONLINE
CERTIFICATION COURSES

Business Analytics for Management Decision

Regression Analysis

Rudra P. Pradhan
Vinod Gupta School of Management
IIT KHARAGPUR

NPTEL

Regression Analysis

- Regression Analysis
- Simple Linear Regression
- Residual Analysis and Regression Assumptions
- Multiple Linear Regression
- Building Good Regression Models
- Regression with Categorical Independent Variables
- Regression Models with Nonlinear Terms

NPTEL

Regression Analysis

- Regression analysis is a tool for building statistical models that characterize relationships among a dependent variable and one or more independent variables, all of which are numerical.
- Simple linear regression involves a single independent variable.
- Multiple regression involves two or more independent variables.

NPTEL

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.

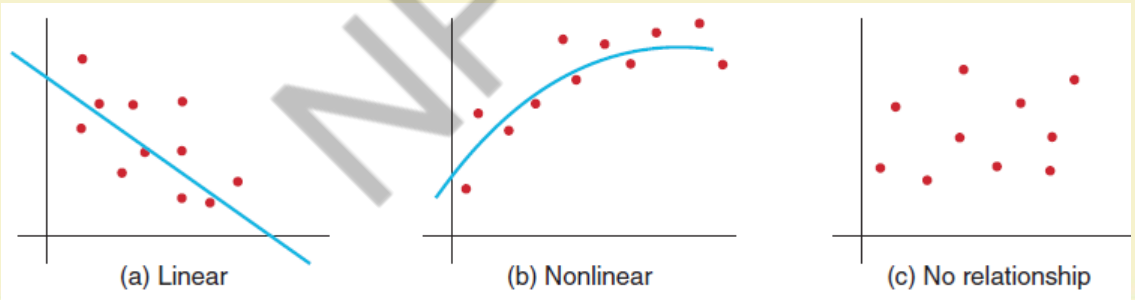


Figure 9.1

NPTEL

Simple Linear Regression

Example 9.1

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

Figure 9.2

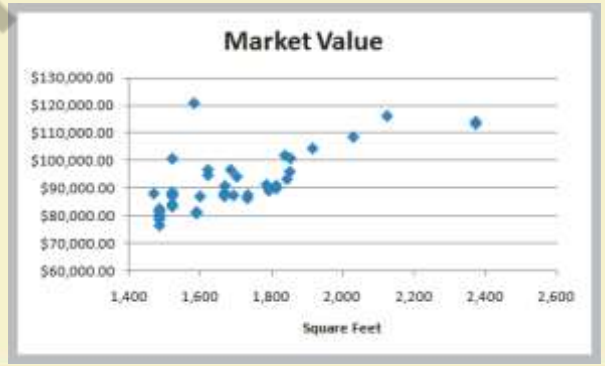


Figure 9.3

NPTEL

Copyright © 2013 Pearson Education, Inc. publishing as Prentice Hall

Simple Linear Regression

Finding the Best-Fitting Regression Line

- Two possible lines are shown below.
- Line A is clearly a better fit to the data.
- We want to determine the best regression line.

$\hat{Y} = b_0 + b_1X$

where

b_0 is the intercept

b_1 is the slope

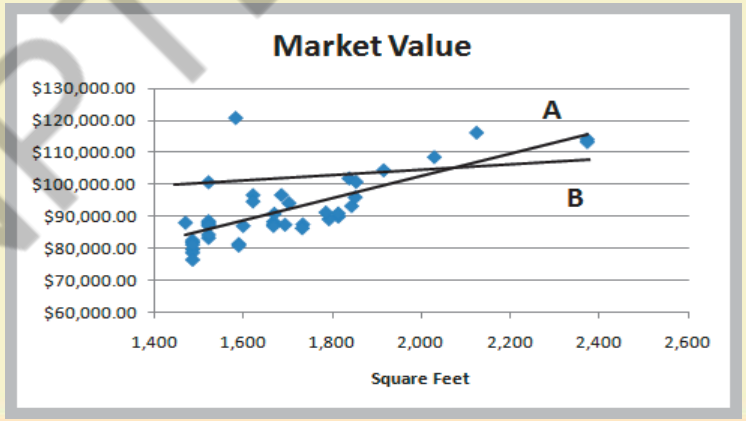


Figure 9.4

NPTEL

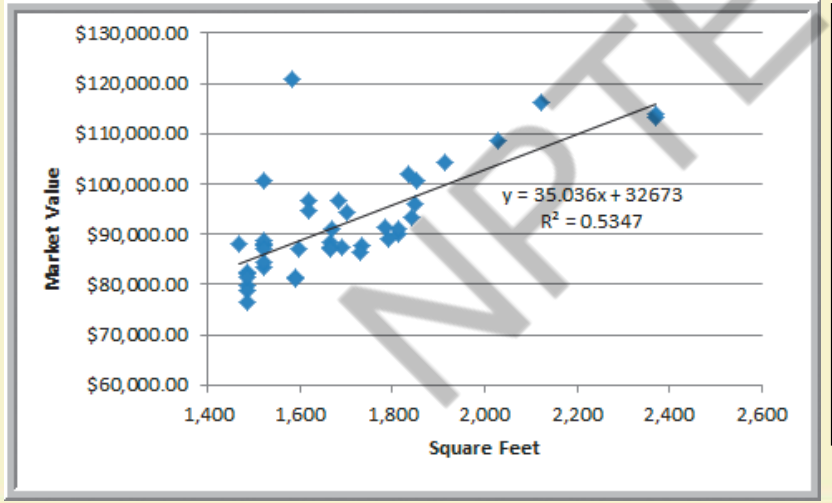
Copyright © 2013 Pearson Education, Inc. publishing as Prentice Hall

Simple Linear Regression

Example 9.2

Using Excel to Find the Best Regression Line

- Market value = 32673 + 35.036(square feet)



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.

Figure 9.5

NPTEL

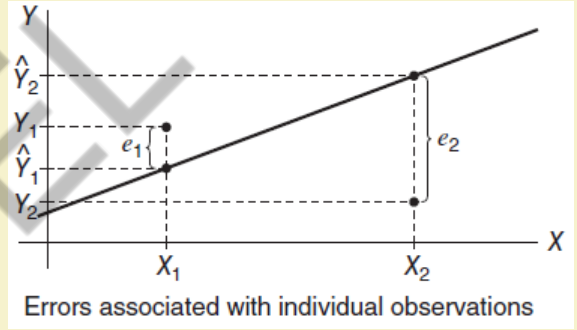
Simple Linear Regression

Least-Squares Regression

Regression analysis finds the equation of the best-fitting line that minimizes

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

the sum of the squares of the observed errors (residuals).



Errors associated with individual observations

Figure 9.6

Using calculus we can solve for the slope and intercept of the least-squares regression line.

NPTEL

Simple Linear Regression

Least-Squares Regression Equations

Slope $b_1 = \dfrac{\sum_{i=1}^{n} X_i Y_i - n\overline{X}\,\overline{Y}}{\sum_{i=1}^{n} X_i^2 - n\overline{X}^2}$

b_1 =SLOPE*(known ys, known xs)*

Intercept $b_0 = \overline{Y} - b_1\overline{X}$

b_0 =INTERCEPT*(known ys, known xs)*

Predict *Y* for specified *X* values: $\hat{Y} = b_0 + b_1 X$

$\hat{Y}$ =TREND*(known ys, known xs, new xs)*

NPTEL

Simple Linear Regression

Example 9.3 Using Excel Functions to Find Least-Squares Coefficients

	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

Figure 9.2

- Slope = b_1 = 35.036
  =SLOPE(C4:C45, B4:B45)
- Intercept = b_0 = 32,673
  =INTERCEPT(C4:C45, B4:B45)
- Estimate Y when X = 1800 square feet
  Y = 32,673 + 35.036(1800) = $95,737.80
  =TREND(C4:C45, B4:B45, 1800)

NPTEL

Simple Linear Regression

Excel *Regression* tool

Data

Data Analysis

Regression

- *Input Y Range*
- *Input X Range*
- *Labels*

Excel outputs a table with many useful regression statistics.

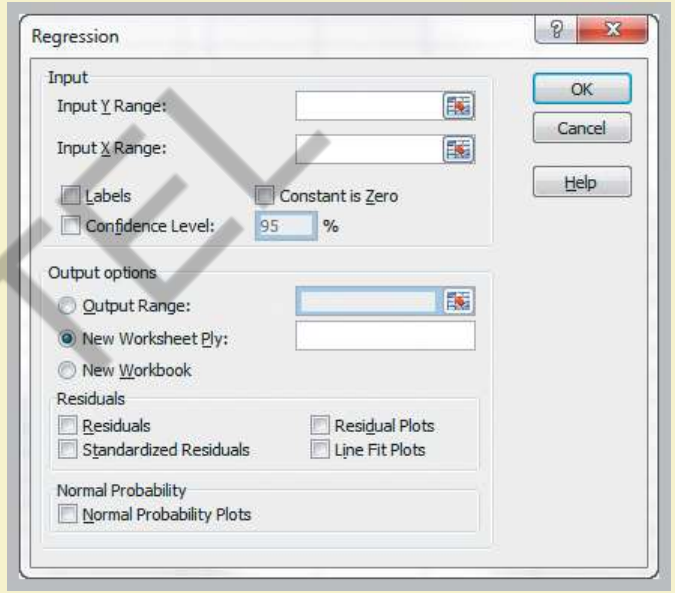


Figure 9.7

NPTEL

Simple Linear Regression

Regression Statistics in Excel's Output

- Multiple R
  | r | where r is the sample correlation coefficient
  r varies from -1 to +1 (r is negative if slope is negative)
- R Square
  coefficient of determination, R^2
  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 & predicted Y variables

NPTEL

Simple Linear Regression

Example 9.4 Interpreting Regression Statistics for Simple Linear Regression (*Home Market Value*)

	A	B	C	D	E	F	G
1	Home Market Value						
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.

Figure 9.8

NPTEL

Simple Linear Regression

Regression 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 ≠ 0

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

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

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			

From Figure 9.8

NPTEL

Simple Linear Regression

Example 9.5 Interpreting Significance of Regression

$H_0: \beta_1 = 0$ Home size is not a significant variable

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

p-value = 3.798×10^{-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.

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			

From Figure 9.8

NPTEL

Simple Linear Regression

Testing Hypotheses for Regression Coefficients

- An alternate method for testing $H_0: \beta_1 = 0$, $H_1: \beta_1 \neq 0$

  is to use a t-test: $t = \dfrac{b_1 - 0}{\text{standard error}}$

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

16		Coefficients	Standard Error	t Stat	P-value
17	Intercept	32673.2199	8831.950745	3.699434116	0.000649604
18	Square Feet	35.03637258	5.16738385	6.780292234	3.79802E-08

From Figure 9.8

NPTEL

Simple Linear Regression

Example 9.6 Interpreting Hypothesis Tests for Regression Coefficients (*Home Market Value*)

- *p*-value for test on the intercept = 0.000649
- *p*-value for test on the slope = 3.798×10^{-8}
- Both tests reject their null hypotheses.
- Both the intercept and slope coefficients are significantly different from zero.

16		*Coefficients*	*Standard Error*	*t Stat*	*P-value*
17	Intercept	32673.2199	8831.950745	3.699434116	0.000649604
18	Square Feet	35.03637258	5.16738385	6.780292234	3.79802E-08

From Figure 9.8

NPTEL

Simple Linear Regression

Example 9.7 Interpreting Confidence Intervals for Regression Coefficients (*Home Market Value*)

- 95% confidence interval estimates

  Intercept is between $14,823 and $50,523

  Slope is between 24.59 and 45.48 $/sq.ft.

- Lower extreme: $\hat{Y} = 14{,}823 + 24.59X$
- Upper extreme: $\hat{Y} = 50{,}523 + 45.48X$

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

From Figure 9.8

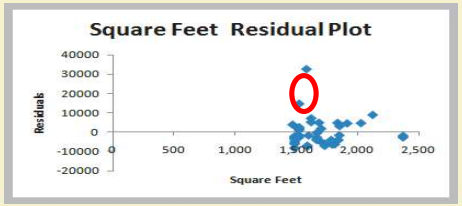
NPTEL

Residual Analysis and Regression Assumptions

Residual Analysis

- Residuals are observed errors.
- 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

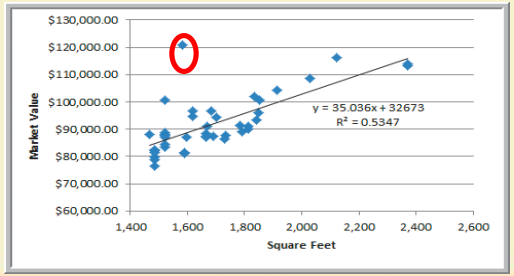


Residual Analysis and Regression Assumptions

Example 9.8 Interpreting Residual Output

- None of the residuals in the table of 5 homes shown below appear to be outliers.
- In the full data set of 42 homes, there is a standardized residual larger than 4.
- This small home may have a pool or unusually large piece of land.

	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



NPTEL

Residual Analysis and Regression Assumptions

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
- *Independence of Errors*
  - successive observations should not be related

NPTEL

Residual Analysis and Regression Assumptions

Example 9.9 Checking Regression Assumptions for the *Home Market Value* Data

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

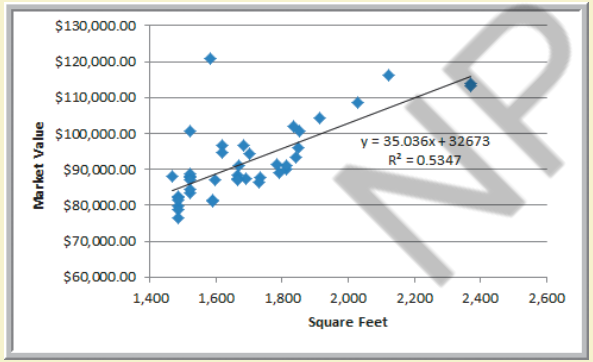


Figure 9.3

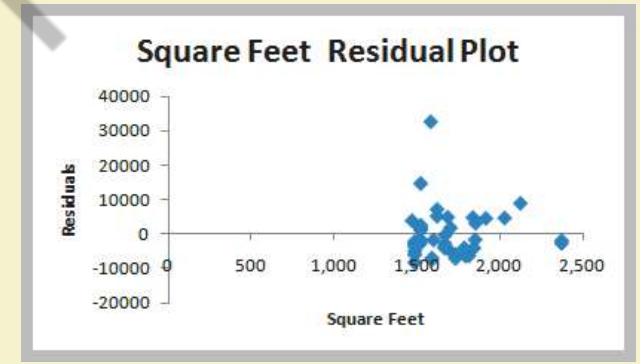


Figure 9.10

NPTEL

Residual Analysis and Regression Assumptions

Example 9.9 (continued) Checking Regression Assumptions for the *Home Market Value* Data

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

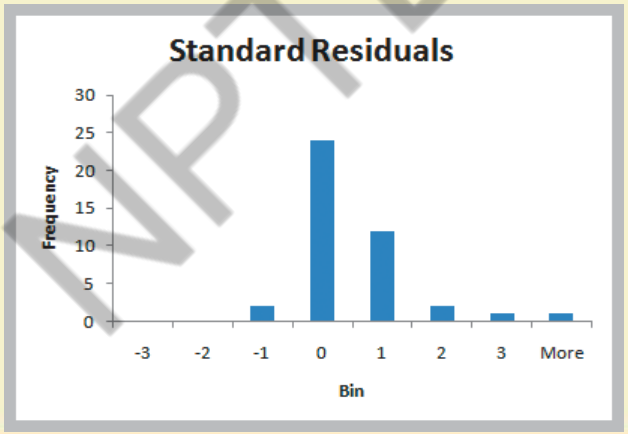


Figure 9.11

NPTEL

Residual Analysis and Regression Assumptions

Example 9.9 (continued) Checking Regression Assumptions for the *Home Market Value* Data

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

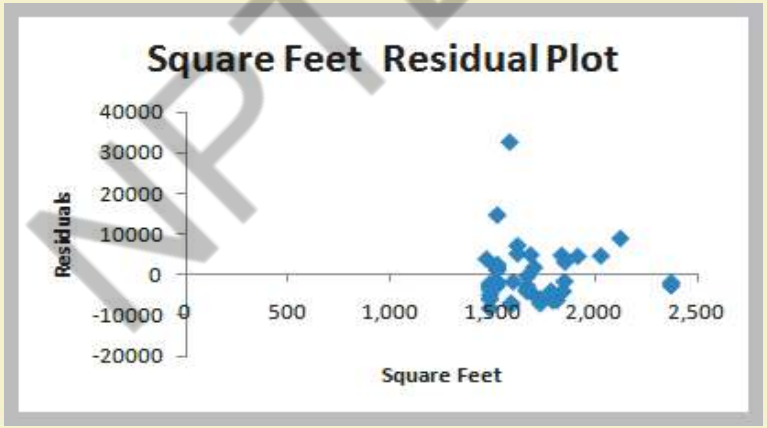


Figure 9.10

NPTEL

Residual Analysis and Regression Assumptions

Example 9.9 (continued) Checking Regression Assumptions for the *Home Market Value* Data

- Independence of Errors – Because the data is cross-sectional, we can assume this assumption holds.
- All 4 regression assumptions are reasonable for the *Home Market Value* data.

NPTEL

Multiple Linear Regression

Multiple Regression has more than one independent variable.

- The multiple linear regression equation is:

$$\hat{Y} = b_0 + b_1X_1 + b_2X_2 + \cdots + b_kX_k$$

- The ANOVA test for significance of the entire model is:

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

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

- One can also test for significance of individual regression coefficients.

NPTEL

Multiple Linear Regression

Example 9.10 Interpreting Regression Results for the *Colleges and Universities* Data

- Colleges try to predict student graduation rates using a variety of characteristics, such as:
  1. Median SAT
  2. Expenditures/student
  3. Acceptance rate
  4. Top 10% of HS class

	A	B	C	D	E	F	G
1	Colleges and Universities						
2							*Y*
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

Figure 9.12

NPTEL

Multiple Linear Regression

Example 9.10 (continued) Interpreting Regression Results for the *Colleges and Universities* Data

	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

Graduation% = 17.92 + 0.072 SAT – 24.859 ACCEPTANCE – 0.000136 EXPENDITURES – 0.163 TOP10% HS

Figure 9.13

NPTEL

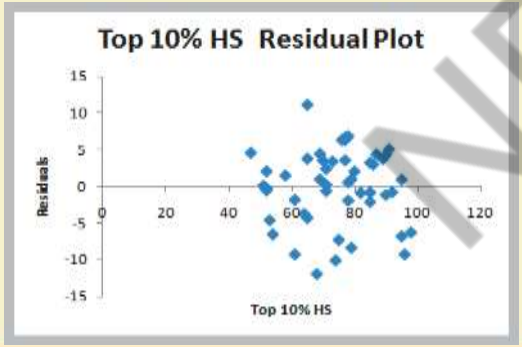
Multiple Linear Regression

Example 9.10 (continued) Interpreting Regression Results for the *Colleges and Universities* Data

16		Coefficients	Standard Error	t Stat	P-value
17	Intercept	17.92095587	24.55722367	0.729763108	0.469402466
18	Median SAT	0.072006285	0.017983915	4.003927007	0.000236106
19	Acceptance Rate	-24.8592318	8.315184822	-2.989618672	0.004559569
20	Expenditures/Student	-0.00013565	6.59314E-05	-2.057438385	0.045600176
21	Top 10% HS	-0.162764489	0.079344518	-2.051364015	0.046213846

From Figure 9.13

All of the slope coefficient *p*-values are < 0.05.



The residual plots (only one shown here) show random patterns about 0.

Normal probability plots (not shown) also validate assumptions.

NPTEL

Multiple Linear Regression



Analytics in Practice:
Using Linear Regression and Interactive Risk Simulators to Predict Performance at ARAMARK

- ARAMARK, located in Philadelphia, is an award-winning provider of professional services
- They developed an on-line tool called “interactive risk simulators” (shown on next slide) that allows users to change various business metrics and immediately see the results.
- The simulators use linear regression models.

NPTEL

Multiple Linear Regression

Analytics in Practice: (ARAMARK continued)

Risk metrics are adjusted using sliders.

Allows users (managers and directors) to see the impact of these risks on the business.

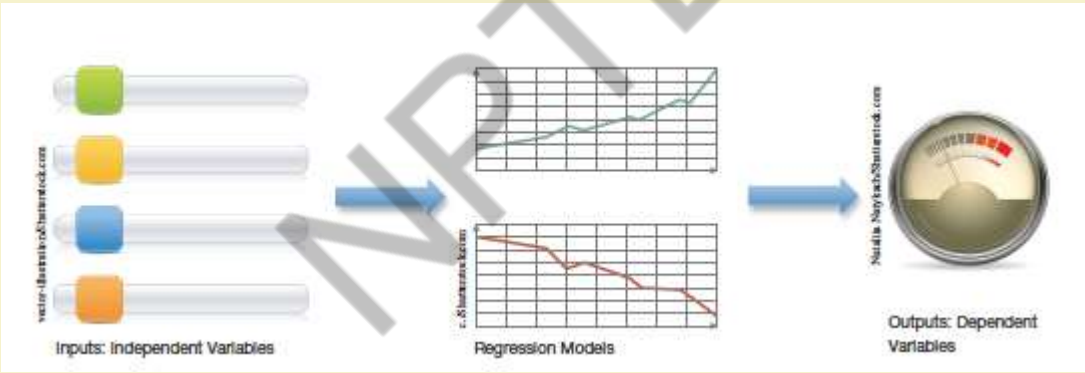


Figure 9.15

NPTEL

Building Good Regression Models

- All of the independent variables in a linear regression model are not always significant.
- We will learn how to build good regression models that include the "best" set of variables.
- *Banking Data* includes demographic information on customers in the bank's current market.

	A	B	C	D	E	F
1	Banking Data					
2						
3	Median Age	Median Years Education	Median Income	Median Home Value	Median Household Wealth	Average Bank Balance
4						
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

Figure 9.16

Building Good Regression Models

Predicting Average Bank Balance using 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	0.013613178	-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

Home Value and Education are not significant.

Figure 9.17

NPTEL

Building Good Regression Models

Systematic Approach to Building Good Multiple Regression Models

1. Construct a model with all available independent variables and check for significance of each.
2. Identify the largest *p*-value that is greater than α.
3. Remove that variable and evaluate adjusted R^2.
4. Continue until all variables are significant.

→ Find the model with the highest adjusted R^2.

(Do not use unadjusted R^2 since it always increases when variables are added.)

NPTEL

Building Good Regression Models

Example 9.11

Identifying the Best Regression Model

- 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	-19812.99569	-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

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

Figure 9.18

NPTEL

Building Good Regression Models

Multicollinearity

- occurs when there are strong correlations among the independent variables
- makes it difficult to isolate the effects of independent variables
- signs of slope coefficients may be opposite of the true value and *p*-values can be inflated

- Correlations exceeding ± 0.7 are an indication that multicollinearity might exist.
- Variance Inflation Factors are a better indicator.
- Parsimony is an age-old principle that applies here.

NPTEL

Building Good Regression Models

Example 9.12

Identifying Potential Multicollinearity

- *Colleges and Universities* (full model)

	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

Full model
Adjusted R^2 = 0.4921

NPTEL

Figure 9.13

Building Good Regression Models

Example 9.12 (continued)

Identifying Potential Multicollinearity

- Correlation Matrix (*Colleges and Universities* data)

	A	B	C	D	E
1		*Median SAT*	*Acceptance Rate*	*Expenditures/Student*	*Top 10% HS*
2	Median SAT	1			
3	Acceptance Rate	-0.60190196	1		
4	Expenditures/Student	0.572741729	-0.284254415	1	
5	Top 10% HS	0.503467995	-0.609720972	0.505782049	1

From Figure 9.19

- All of the correlations are within ± 0.7

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

- Signs of the coefficients are questionable for Expenditures and Top 10%.

Building Good Regression Models

Example 9.12 (continued)

Identifying Potential Multicollinearity

- *Colleges and Universities* (reduced model)

	A	B	C	D	E	F	G	H	I
1	SUMMARY OUTPUT								
2									
3	*Regression Statistics*								
4	Multiple R	0.699928026							
5	R Square	0.489899242							
6	Adjusted R Square	0.455892525							
7	Standard Error	5.494299164							
8	Observations	49							
9									
10	ANOVA								
11		*df*	*SS*	*MS*	*F*	*Significance F*			
12	Regression	3	1304.631676	434.8772252	14.4059551	1.02877E-06			
13	Residual	45	1358.429549	30.18732331					
14	Total	48	2663.061224						
15									
16		*Coefficients*	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
17	Intercept	3.340739129	24.32962712	0.137311563	0.89139707	-45.66164531	52.34312357	-45.6616453	52.34312357
18	Median SAT	0.072676027	0.018610858	3.905033628	0.00031304	0.035191834	0.110160219	0.035191834	0.110160219
19	Acceptance Rate	-16.3480775	7.458545966	-2.19185852	0.03360495	-31.37036024	-1.325794823	-31.3703602	-1.325794823
20	Expenditures/Student	-0.00018843	6.28314E-05	-2.99890365	0.0044027	-0.000314974	-6.18764E-05	-0.00031497	-6.18764E-05

Dropping Top 10%
Adjusted R^2 drops to 0.4559

NPTEL

Building Good Regression Models

Example 9.12 (continued)

Identifying Potential Multicollinearity

- *Colleges and Universities* (reduced model)

	A	B	C	D	E	F	G	H	I
1	SUMMARY OUTPUT								
2									
3	*Regression Statistics*								
4	Multiple R	0.699739346							
5	R Square	0.489635152							
6	Adjusted R Square	0.455610829							
7	Standard Error	5.49572124							
8	Observations	49							
9									
10	ANOVA								
11		*df*	*SS*	*MS*	*F*	*Significance F*			
12	Regression	3	1303.928387	434.6427956	14.3907389	1.04057E-06			
13	Residual	45	1359.132838	30.20295195					
14	Total	48	2663.061224						
15									
16		*Coefficients*	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
17	Intercept	43.22114366	22.0069424	1.963977679	0.05572803	-1.103113604	87.54540093	-1.1031136	87.54540093
18	Median SAT	0.053875251	0.016230195	3.319445778	0.00179405	0.021185961	0.086564541	0.021185961	0.086564541
19	Acceptance Rate	-29.4513889	8.292781736	-3.5514487	0.00091191	-46.15390871	-12.7488691	-46.1539087	-12.7488691
20	Top 10% HS	-0.22646469	0.075633402	-2.9942418	0.00445903	-0.378798185	-0.0741312	-0.37879818	-0.0741312

Dropping Expenditures
Adjusted R^2 drops to 0.4556

NPTEL

Building Good Regression Models

Example 9.12 (continued)

Identifying Potential Multicollinearity

- *Colleges and Universities* (reduced model)

	A	B	C	D	E	F	G	H	I
1	SUMMARY OUTPUT								
2									
3	*Regression Statistics*								
4	Multiple R	0.622859249							
5	R Square	0.387953644							
6	Adjusted R Square	0.361342932							
7	Standard Error	5.9525605							
8	Observations	49							
9									
10	ANOVA								
11		*df*	*SS*	*MS*	*F*	*Significance F*			
12	Regression	2	1033.144305	516.5721525	14.5788529	1.2475E-05			
13	Residual	46	1629.916919	35.43297651					
14	Total	48	2663.061224						
15									
16		*Coefficients*	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
17	Intercept	35.44644156	23.6698037	1.497538467	0.14108426	-12.19840214	83.09128526	-12.1984021	83.09128526
18	Median SAT	0.043396966	0.017165857	2.528097904	0.01496524	0.008843889	0.077950044	0.008843889	0.077950044
19	Acceptance Rate	-18.4146051	8.046076985	-2.28864391	0.02674178	-34.61051804	-2.21869213	-34.610518	-2.21869213

Dropping Expenditures and Top 10%
Adjusted R^2 drops to 0.3613

Which of the 4 models would you choose?

NPTEL

Building Good Regression Models

Example 9.12 (continued)

- *Banking* Data (full model)

	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	0.013613178	-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

Full Model
Adjusted R^2 = 0.9441
Education and Home Value
are not significant.

NPTEL

Figure 9.17

Building Good Regression Models

Example 9.12 (continued)
Identifying Potential Multicollinearity

- Correlation matrix for the *Banking* data

	A	B	C	D	E	F
1		*Age*	*Education*	*Income*	*Home Value*	*Wealth*
2	Age	1				
3	Education	0.17340715	1			
4	Income	0.47714740	0.57539402	1		
5	Home Value	0.38649311	0.75352107	0.79535516	1	
6	Wealth	0.46809179	0.46941304	0.94666545	0.69847779	1

From Figure 9.20

	Coefficients
Intercept	-10710.6428
Age	318.6649626
Education	621.8603472
Income	0.146323453
Home Value	0.009183067
Wealth	0.074331533

From Figure 9.17

- Some of the correlations exceed 0.7 for Home Value and Wealth.
- Signs of the coefficients for predicting bank balance are as expected (positive).

NPTEL

Building Good Regression Models

Example 9.12 (continued)

- *Banking* Data (reduced model)

	A	B	C	D	E	F	G	H	I
1	SUMMARY OUTPUT								
2									
3	*Regression Statistics*								
4	Multiple R	0.9604652							
5	R Square	0.92249341							
6	Adjusted R Square	0.92012076							
7	Standard Error	2458.25514							
8	Observations	102							
9									
10	ANOVA								
11		*df*	*SS*	*MS*	*F*	*Significance F*			
12	Regression	3	7048628348	2349542783	388.802855	2.89264E-54			
13	Residual	98	592215796.6	6043018.33					
14	Total	101	7640844145						
15									
16		*Coefficients*	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
17	Intercept	-9539.94542	4423.05947	-2.15686574	0.03346194	-18317.363	-762.527849	-18317.363	-762.527849
18	Age	332.500072	72.33548609	4.59663838	1.2816E-05	188.9526538	476.04749	188.952654	476.04749
19	Education	288.655512	300.533758	0.96047617	0.33917941	-307.74395	885.054973	-307.74395	885.054973
20	Income	0.3870509	0.01748407	22.1373457	6.4906E-40	0.35235433	0.42174746	0.35235433	0.42174746

Dropping Wealth and Home Value
Adjusted R^2 drops to 0.9201
Education is not significant.

NPTEL

Building Good Regression Models

Example 9.12 (continued)
Identifying Potential Multicollinearity

- Re-ordered Correlation matrix for *Banking* data

	A	B	C	D	E	F
1		*Age*	*Education*	*Wealth*	*Income*	*Home Value*
2	Age	1				
3	Education	0.17340715	1			
4	Wealth	0.46809179	0.46941304	1		
5	Income	0.47714740	0.57539402	0.94666545	1	
6	Home Value	0.38649311	0.75352107	0.69847779	0.79535516	1

From Figure 9.20

- By re-ordering the variables, we can see the correlations for Age, Education, and Wealth are all within ± 0.7.
- Let's try a reduced model with the Age, Education, and Wealth variables.

NPTEL

Building Good Regression Models

Example 9.12 (continued)

- *Banking* Data (reduced model) ** best model

	A	B	C	D	E	F	G
1	Regression Analysis						
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	-25276.72737	-10188.17547
18	Age	367.8214086	64.59823831	5.693985134	1.2977E-07	239.6283103	496.0145069
19	Education	1300.308712	249.9731413	5.201793703	1.08292E-06	804.2451615	1796.372263
20	Wealth	0.116467903	0.004679827	24.88722652	3.75813E-44	0.10718094	0.125754866

Dropping Income and Home Value.
Adjusted R^2 = 0.9345.
All variables are significant.
Multicollinearity is not a problem.

Figure 9.21

Regression with Categorical Variables

Dealing with Categorical Variables

- Must be coded numeric using *dummy variables.*
- For variables with 2 categories, code as 0 and 1.
- For variables with $k \geq 3$ categories, create $k-1$ binary (0,1) variables.

Interaction Terms

- A dependence between two variables is called interaction.
- Test for interaction by adding a new term to the model, such as $X_3 = X_1X_2$

NPTEL

Regression with Categorical Variables

Example 9.13 A Model with Categorical Variables

- *Employee Salaries* provides data for 35 employees
- Predict *Salary* using *Age* and *MBA* (yes=1, no=0)

	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

Figure 9.22

NPTEL

Regression with Categorical Variables

Example 9.13 (continued)

Salary = 893.59 + 1044(Age) for those without MBA

Salary =15,660.82 + 1044(Age) for those with MBA

	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	-2822.950618	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

Adjusted R^2 = 0.949858

NPTEL

Figure 9.23

Regression with Categorical Variables

Example 9.14 Incorporating Interaction Terms in a Regression Model

- Define an interaction between Age and MBA and include in the regression model.
- Interaction = (Age)(MBA)

	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

Figure 9.24

NPTEL

Regression with Categorical Variables

Example 9.14 (continued) Incorporating Interaction Terms in a Regression Model

	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	1176.908399	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

MBA is now insignificant so we will drop it from the model.

Adjusted R^2 = 0.976701

Figure 9.25

Regression with Categorical Variables

Example 9.14 (continued)

Salary = 3,323 + 984(Age) for those without MBA

Salary = 3,323 + 1410(Age) for those with MBA

	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	882.1441051	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

Adjusted R^2 = 0.976727
(a slight improvement)

Figure 9.26

Regression with Categorical Variables

Example 9.15 A Regression Model with Multiple Levels of Categorical Variables

- *Surface Finish* data provides measurements for 35 parts produced on a lathe.

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

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

Tool type A: $Y = \beta_0 + \beta_1 X_1 + \varepsilon$
Tool type B: $Y = \beta_0 + \beta_1 X_1 + \beta_2 + \varepsilon$
Tool type C: $Y = \beta_0 + \beta_1 X_1 + \beta_3 + \varepsilon$
Tool type D: $Y = \beta_0 + \beta_1 X_1 + \beta_4 + \varepsilon$

	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

Figure 9.27

NPTEL

Regression with Categorical Variables

Example 9.15 (continued)
A Regression Model with Multiple Levels of Categorical Variables

- Tool Type (A,B,C,D) is now coded as 3 dummy 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

Figure 9.28

NPTEL

Regression with Categorical Variables

Example 9.15 (continued) A Regression Model with Multiple Levels of Categorical Variables

	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	19.4432239	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

Tool A: Surf. Finish = 24.5 + 0.098 RPM
Tool B: Surf. Finish = 11.2 + 0.098 RPM
Tool C: Surf. Finish = 4.0 + 0.098 RPM
Tool D: Surf. Finish = -1.6 + 0.098 RPM

NPTEL

Figure 9.29

Regression Models with Nonlinear Terms

Curvilinear Regression

- 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.

NPTEL

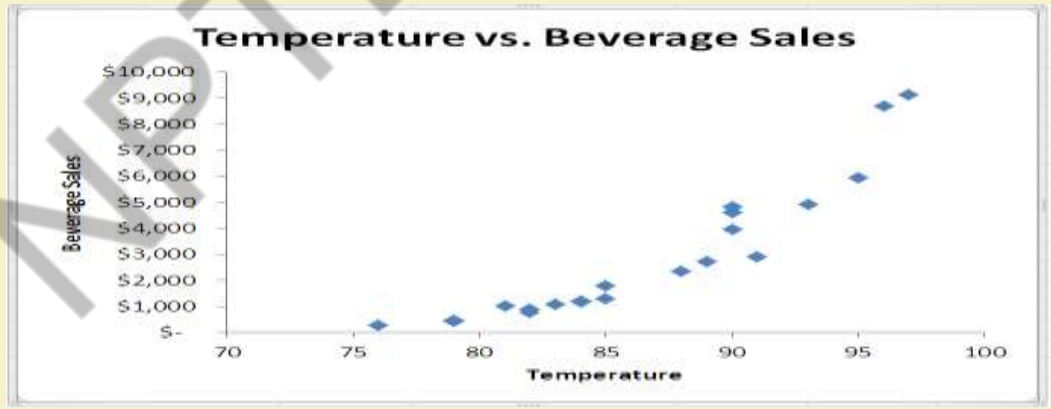
Regression Models with Nonlinear Terms

Example 9.16 Modeling Beverage Sales Using Curvilinear Regression

- Sales of cold beverages increase when it is hotter outside.

	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

Figure 9.30

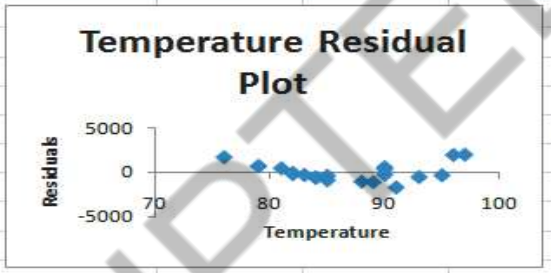


NPTEL

Regression Models with Nonlinear Terms

Example 9.16 (continued) Modeling Beverage Sales Using Curvilinear Regression

	A	B	C	D	E	F	G	H	I
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%*	*Lower 95.0%*	*Upper 95.0%*
17	Intercept	-32511.24671	3408.723477	-9.53766034	1.12197E-08	-39645.78693	-25376.7065	-39645.78693	-25376.7065
18	Temperature	408.6026284	39.26555335	10.40613447	2.7611E-09	326.4188809	490.7863759	326.4188809	490.7863759



U-shape residual plot

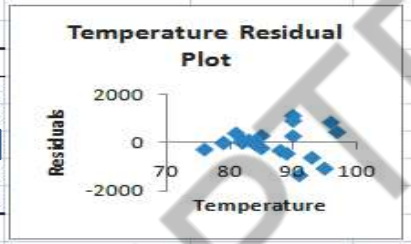
Figure 9.31

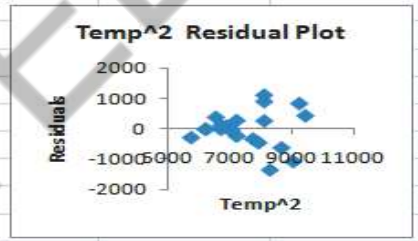
NPTEL

Regression Models with Nonlinear Terms

Example 9.16 (continued) Modeling Beverage Sales Using Curvilinear Regression

	A	B	C	D	E	F	G	H	I
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%*	*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





Residual pattern is more random

Sales = 142,850
−3643(temperature)
+ 23.3(temperature)^2

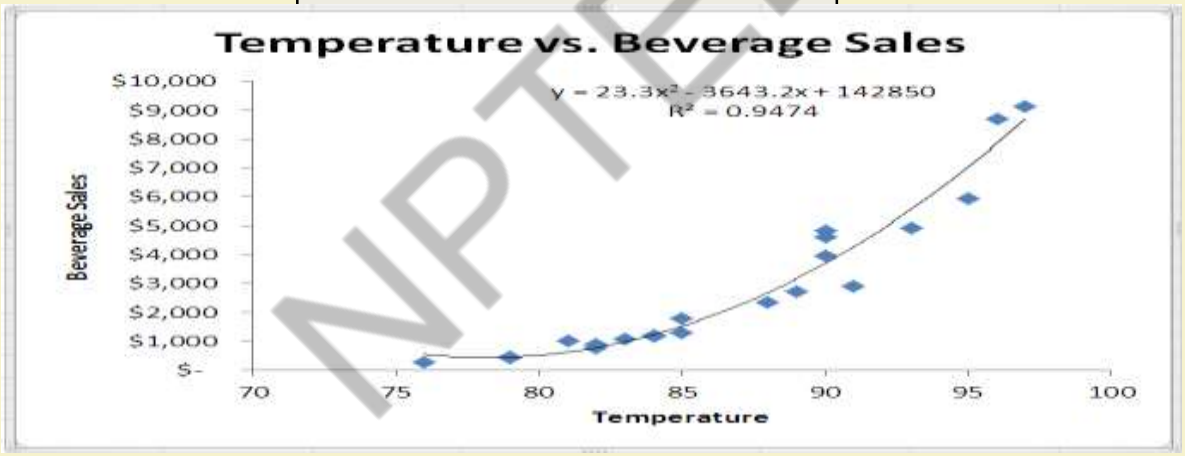
Figure 9.32

NPTEL

Regression Models with Nonlinear Terms

Example 9.16 (continued) Modeling Beverage Sales Using Curvilinear Regression

Second Order Polynomial Trendline



NPTEL

Keywords

- Autocorrelation
- Coefficient of determination
- Coefficient of multiple determination
- Curvilinear regression model
- Dummy variables
- Homoscedasticity
- Interaction
- Least-squares regression
- Mulitcollinearity
- Multiple correlation coefficient

NPTEL

Keywords (continued)

- Multiple linear regression
- Parsimony
- Partial regression coefficient
- Regression analysis
- Residuals
- Significance of regression
- Simple linear regression
- Standard error of the estimate
- Standard residuals

NPTEL

Case Study: Performance Lawn Equipment (9)

- Recall that PLE produces lawnmowers and a medium size diesel power lawn tractor.
- Predict what might have happened if PLE never implemented the 2009 defect reduction initiative.
- Determine the effect of education, GPA, and age when hired on employee retention.
- Investigate the rate of learning following the implementation of the new production technology.
- Write a formal report summarizing your results.

NPTEL

Books to read

- Business Analytics by J R Evans
- Business Analytics for Managers by H N Laursen

NPTEL

Thank You!!

NPTEL