Chapter 15

Multiple Regression
Model Building

Learning Objectives

In this chapter, you learn:

- To use quadratic terms in a regression model
- To measure the correlation among the independent variables
- To build a regression model using either the stepwise or best-subsets approach
- To avoid the pitfalls involved in developing a multiple regression model

Nonlinear Relationships

DCOVA

- The relationship between the dependent variable and an independent variable may not be linear
- Can review the scatter plot to check for non-linear relationships
- Example: Quadratic model

$$\boxed{Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{1i}^2 + \varepsilon_i}$$

  - The second independent variable is the square of the first variable

Quadratic Regression Model

DCOVA

Model form:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{1i}^2 + \varepsilon_i$$

- where:

  β_0 = Y intercept

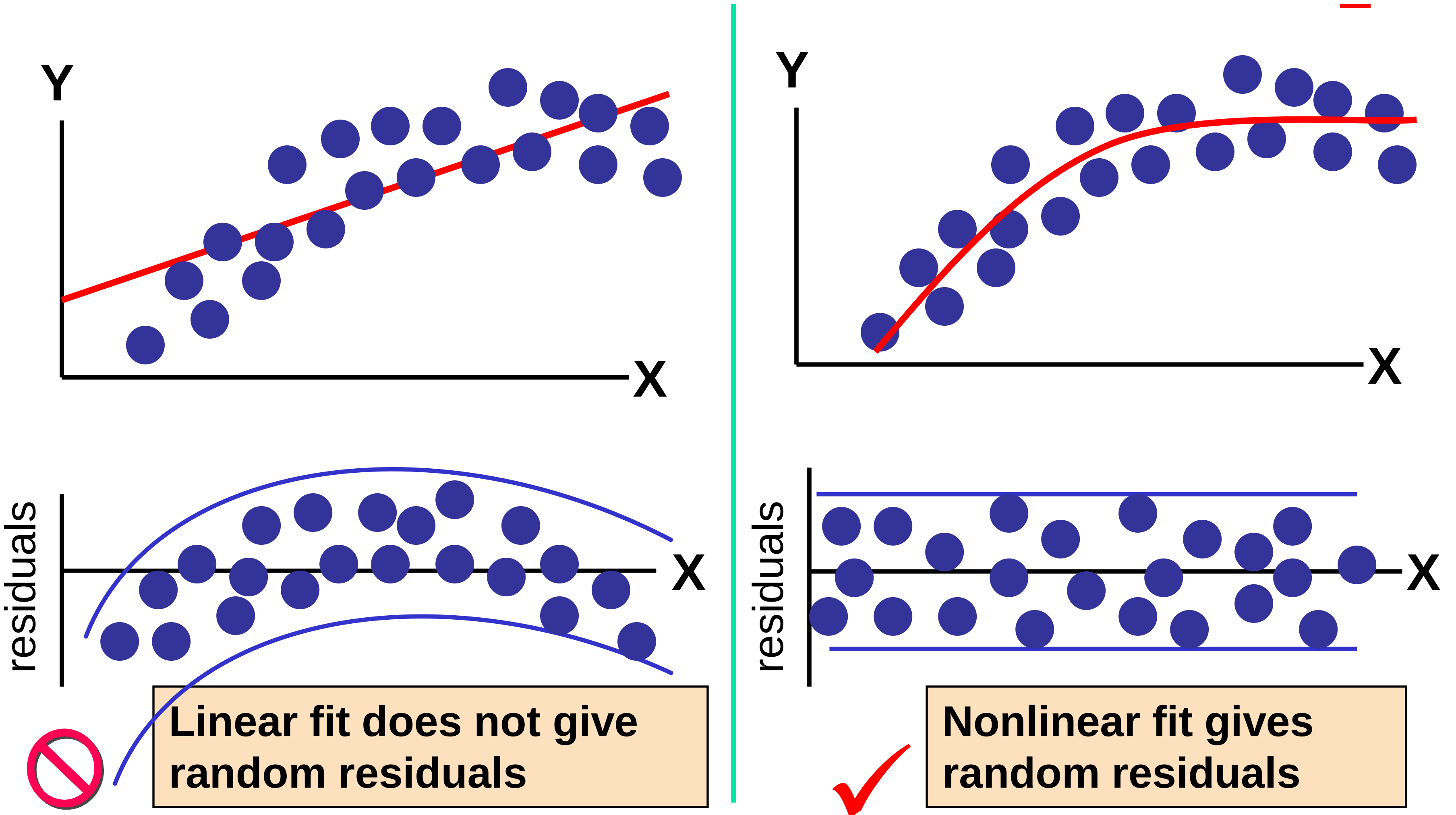
  β_1 = regression coefficient for linear effect of X on Y

  β_2 = regression coefficient for quadratic effect on Y

  ε_i = random error in Y for observation i

Linear vs. Nonlinear Fit

DCOVA



Linear fit does not give random residuals

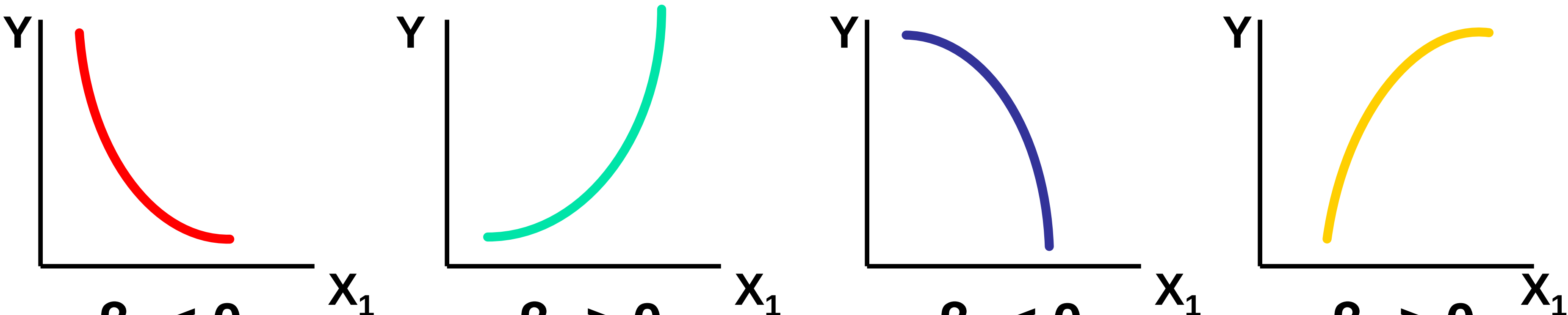
Nonlinear fit gives random residuals

Quadratic Regression Model

DCOVA

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{1i}^2 + \varepsilon_i$$

Quadratic models may be considered when the scatter plot takes on one of the following shapes:



$\beta_1 < 0$	$\beta_1 > 0$	$\beta_1 < 0$	$\beta_1 > 0$
$\beta_2 > 0$	$\beta_2 > 0$	$\beta_2 < 0$	$\beta_2 < 0$

β_1 = the coefficient of the linear term

β_2 = the coefficient of the squared term

Testing the Overall Quadratic Model

DCOVA

- Estimate the quadratic model to obtain the regression equation:

$$\hat{Y}_i = b_0 + b_1 X_{1i} + b_2 X_{1i}^2$$

- Test for Overall Relationship

H_0: $\beta_1 = \beta_2 = 0$ (no overall relationship between X and Y)

H_1: β_1 and/or $\beta_2 \neq 0$ (there is a relationship between X and Y)

- $$F_{STAT} = \frac{MSR}{MSE}$$

Testing for Significance: Quadratic Effect

DCOVA

- Testing the Quadratic Effect
  - Compare quadratic regression equation

$$Y_i = b_0 + b_1 X_{1i} + b_2 X_{1i}^2$$

with the linear regression equation

$$Y_i = b_0 + b_1 X_{1i}$$

Testing for Significance: Quadratic Effect

(continued)

DCOVA

- Testing the Quadratic Effect
  - Consider the quadratic regression equation

$$\boxed{Y_i = b_0 + b_1 X_{1i} + b_2 X_{1i}^2}$$

Hypotheses

H_0: $\beta_2 = 0$ (The quadratic term does not improve the model)

H_1: $\beta_2 \neq 0$ (The quadratic term improves the model)

Testing for Significance: Quadratic Effect

(continued)

DCOVA

- Testing the Quadratic Effect

  Hypotheses

  $H_0: \beta_2 = 0$ (The quadratic term does not improve the model)

  $H_1: \beta_2 \neq 0$ (The quadratic term improves the model)

- The test statistic is

$$t_{STAT} = \frac{b_2 - \beta_2}{S_{b_2}}$$

$$d.f. = n - 3$$

where:

b_2 = squared term slope coefficient

β_2 = hypothesized slope (zero)

S_{b_2} = standard error of the slope

Testing for Significance: Quadratic Effect

(continued)

DCOVA

- Testing the Quadratic Effect

Compare adjusted r^2 from simple regression to adjusted r^2 from the quadratic model

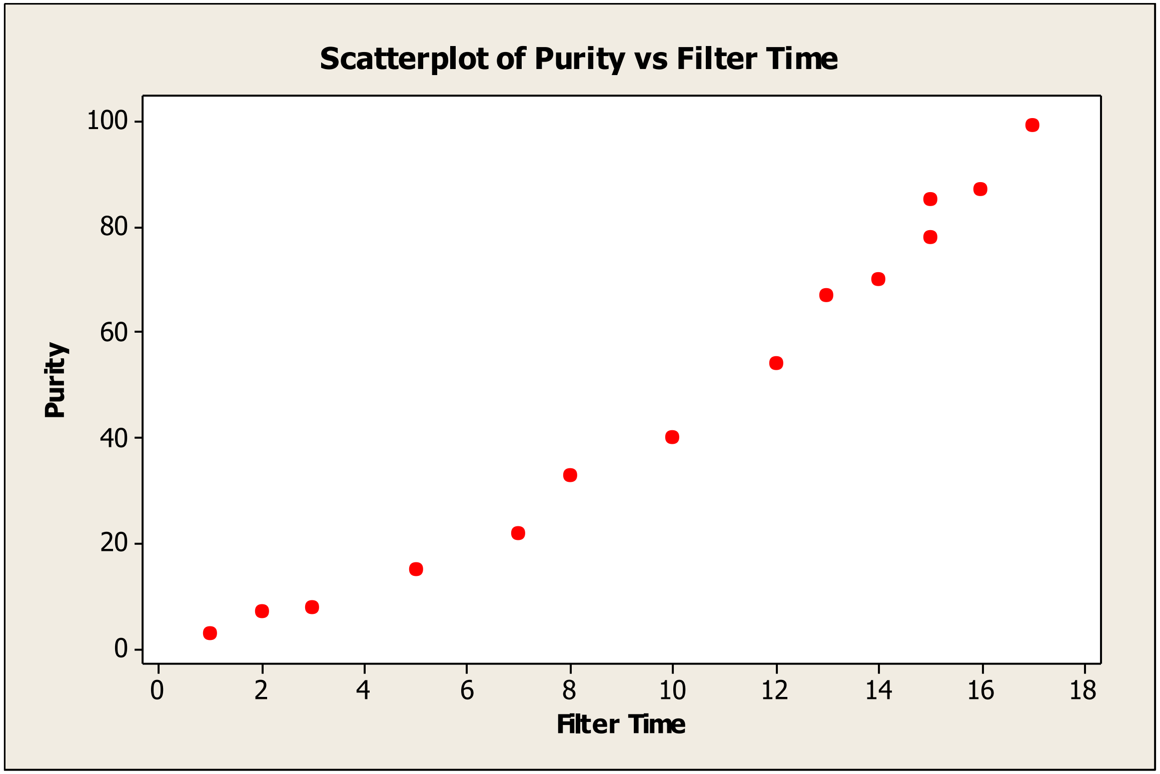
- If adj. r^2 from the quadratic model is larger than the adj. r^2 from the simple model, then the quadratic model is likely a better model

Example 1: Quadratic Model

DCOVA

Purity	Filter Time
3	1
7	2
8	3
15	5
22	7
33	8
40	10
54	12
67	13
70	14
78	15
85	15
87	16
99	17

Purity increases as filter time increases:



Example 1: Quadratic Model

- Simple regression results:

(continued)

DCOVA

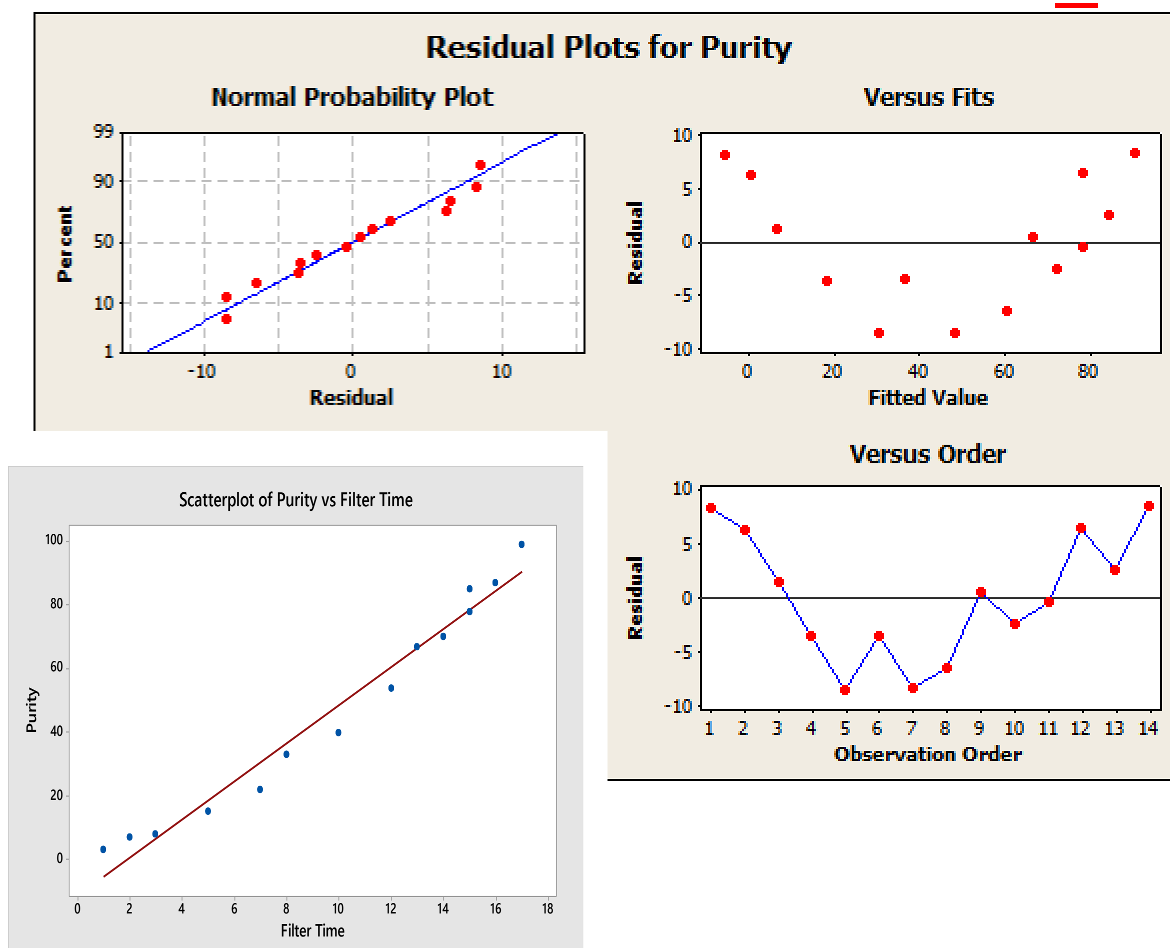
The regression equation is
Purity = - 11.3 + 5.99 Filter Time

Predictor	Coef	SE Coef	T	P
Constant	-11.283	3.468	-3.25	0.007
Filter Time	5.9852	0.3097	19.33	0.000

S = 6.15996   R-Sq = 96.9%   R-Sq(adj) = 96.6%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	1	14176	14176	373.58	0.000
Residual Error	12	455	38		
Total	13	14631			



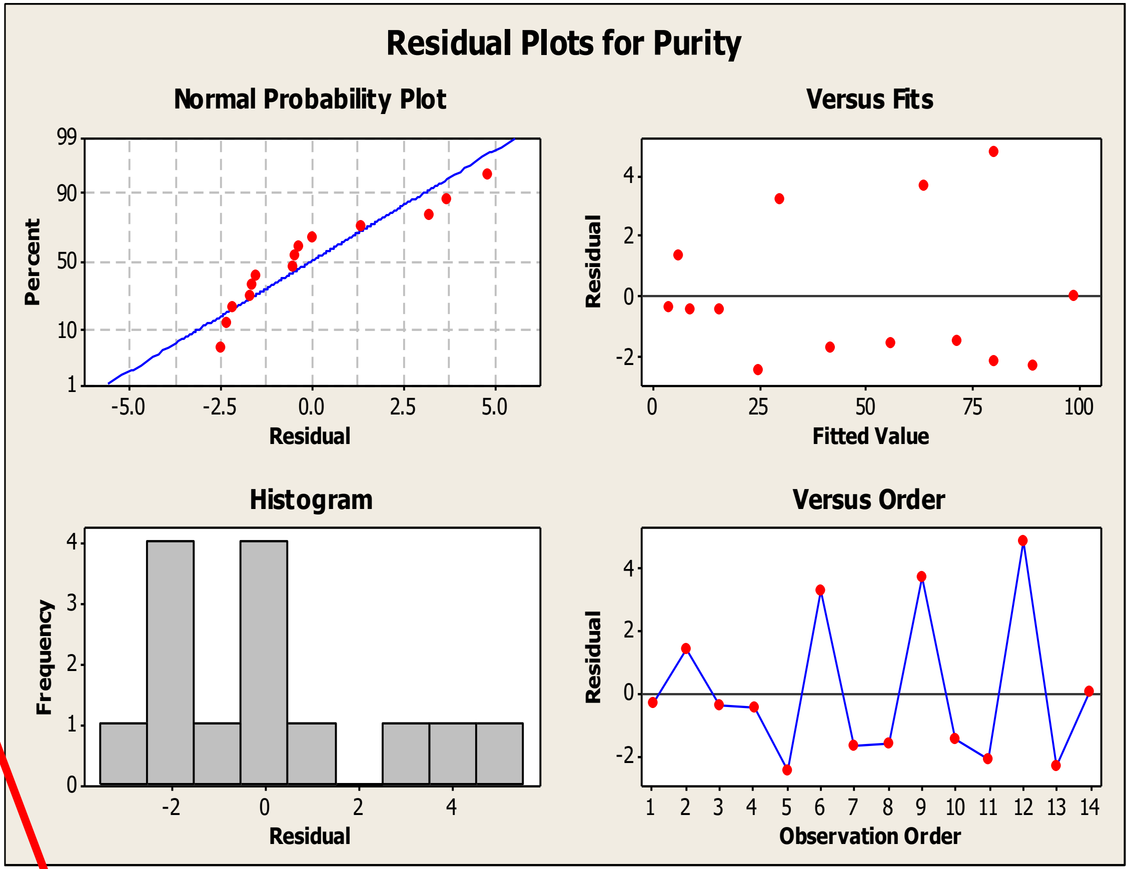
Example 1: Quadratic Model in Minitab

- Quadratic regression results:

The regression equation is
Purity = 1.54 + 1.56 Time + 0.245 Time Squared

Predictor	Coef	SE Coef	T	P
Constant	1.5390	2.24500	0.69	0.507
Time	1.5650	0.60180	2.60	0.025
Time Squared	0.24516	0.03258	7.52	**0.000**

S = 2.59513 R-Sq = 99.5% R-Sq(adj) = 99.4%



The quadratic term is statistically significant (p-value very small)

Example 1: Quadratic Model in Minitab

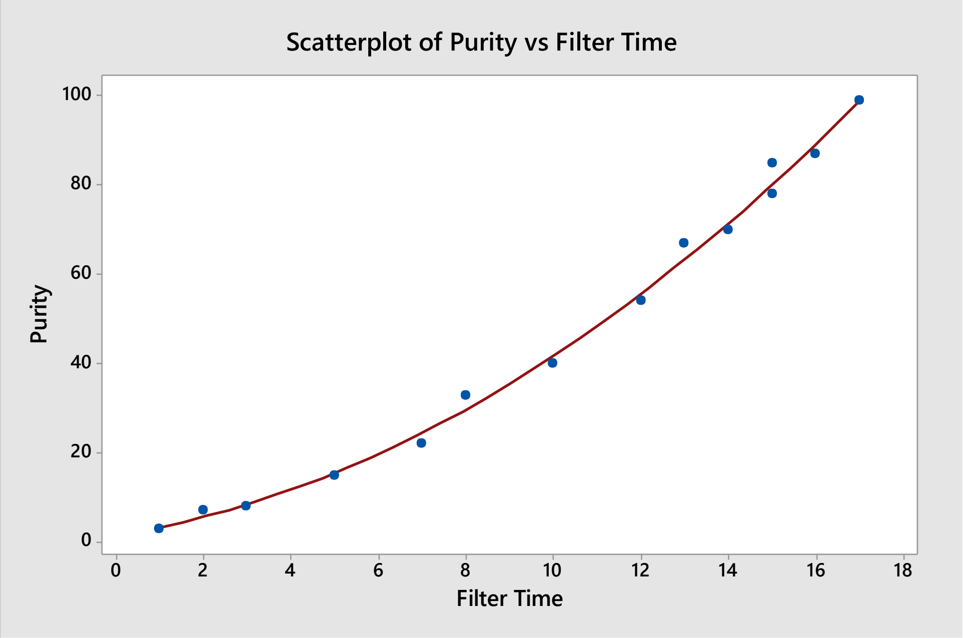
- Quadratic regression results:

$$\hat{Y} = 1.539 + 1.565\ \text{Time} + 0.245\ (\text{Time})^2$$

The regression equation is
Purity = 1.54 + 1.56 Time + 0.245 Time Squared

Predictor	Coef	SE Coef	T	P
Constant	1.5390	2.24500	0.69	0.507
Time	1.5650	0.60180	2.60	0.025
Time Squared	0.24516	0.03258	7.52	0.000

S = 2.59513   R-Sq = 99.5%   R-Sq(adj) = 99.4%

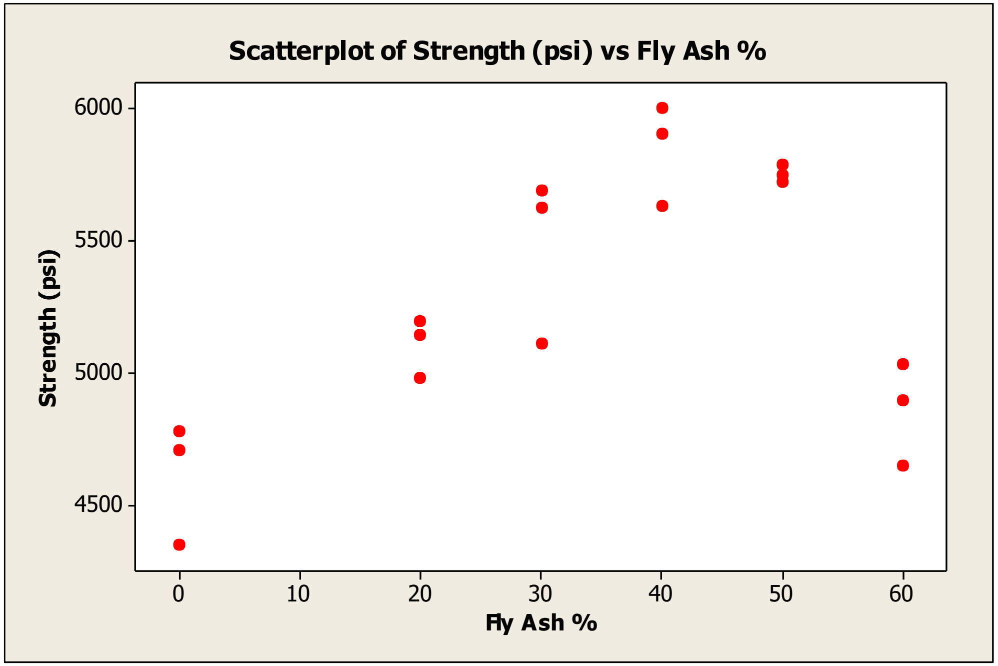


The adjusted r^2 of the quadratic model is higher than the adjusted r^2 of the simple regression model. The quadratic model explains 99.4% of the variation in Y.

Example 2: Quadratic Model

Fly Ash %	Strength (psi)
0	4779
0	4706
0	4350
20	5189
20	5140
20	4976
30	5110
30	5685
30	5618
40	5995
40	5628
40	5897
50	5746
50	5719
50	5782
60	4895
60	5030
60	4648

To illustrate the quadratic regression model, consider a study that examined the business problem facing a concrete supplier of how adding fly ash affects the strength of concrete. Batches of concrete were prepared in which the percentage of fly ash ranged from 0% to 60%. Data were collected from a sample of 18 batches and organized in the following table:



Example 2: Quadratic Model

- Simple regression results:

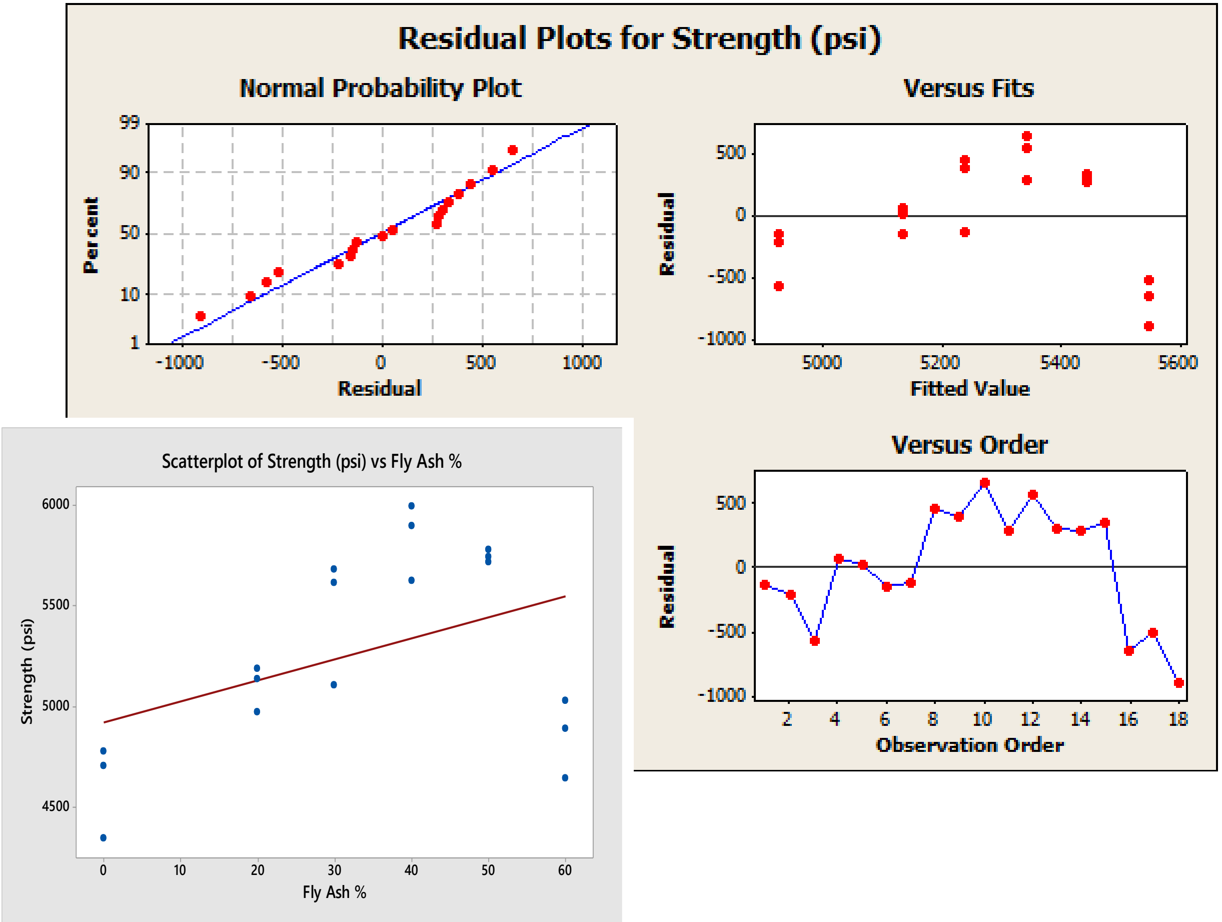
The regression equation is
Strength (psi) = 4925 + 10.4 Fly Ash %

Predictor	Coef	SE Coef	T	P
Constant	4924.6	213.3	23.09	0.000
Fly Ash %	10.417	5.507	1.89	0.077

S = 460.779   R-Sq = 18.3%   R-Sq(adj) = 13.2%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	1	759618	759618	3.58	0.077
Residual Error	16	3397072	212317		
Total	17	4156690			



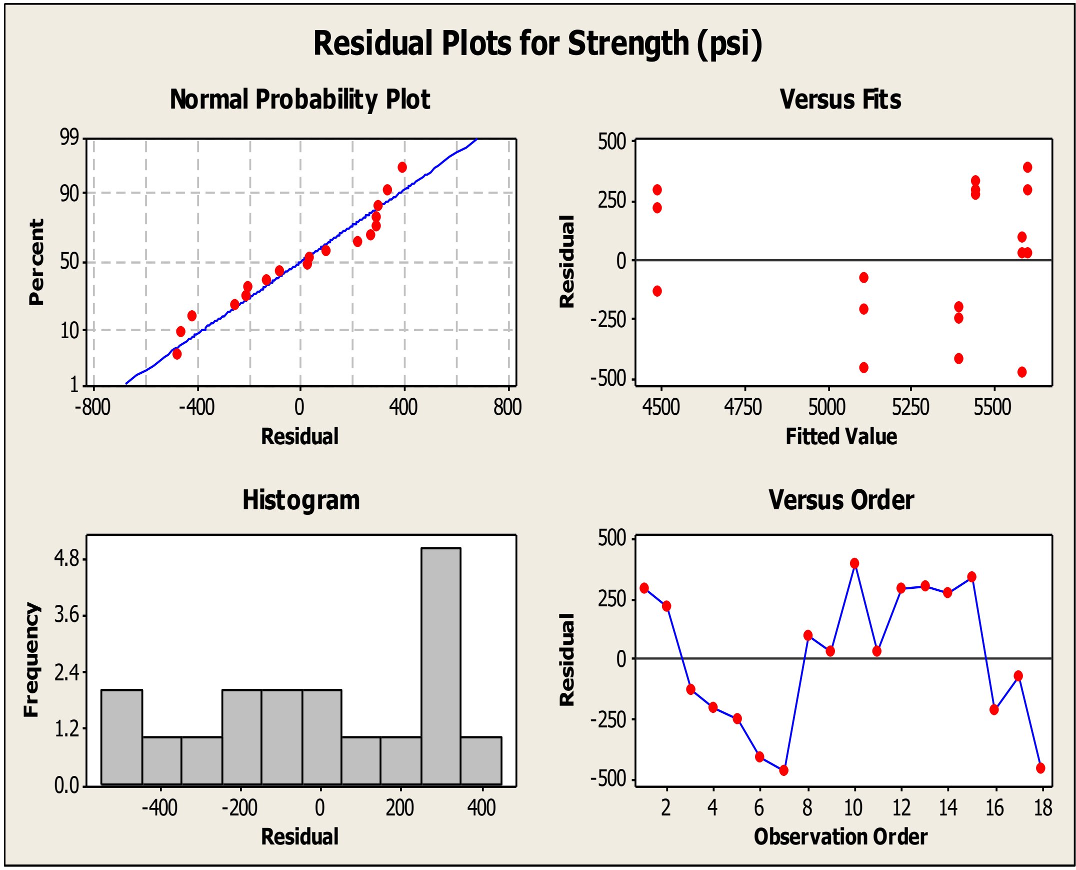
Example 2: Quadratic Model in Minitab

- Quadratic regression results:

The regression equation is
Strength (psi) = 4486 + 63.0 Fly Ash %
- 0.876 (Fly Ash %)^2

Predictor	Coef	SE Coef	T	P
Constant	4486.4	174.8	25.67	0.000
Fly Ash %	63.01	12.37	5.09	0.000
(Fly Ash %)^2	-0.8765	0.1966	-4.46	**0.000**

S = 312.113   R-Sq = 64.8%   R-Sq(adj) = 60.2%



The quadratic term is statistically significant (p-value very small)

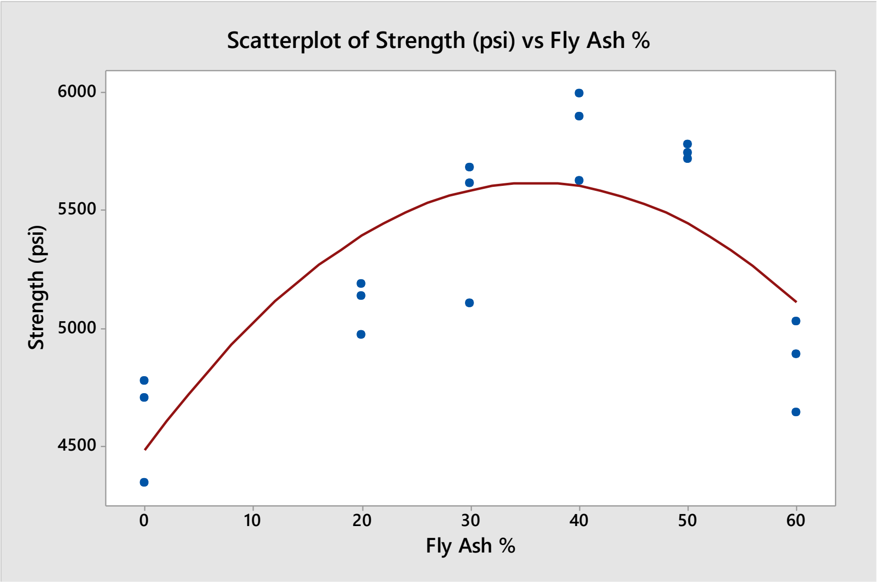
Example 2: Quadratic Model in Minitab

- Quadratic regression results:

The regression equation is
Strength (psi) = 4486 + 63.0 Fly Ash % - 0.876 (Fly Ash %)^2

Predictor	Coef	SE Coef	T	P
Constant	4486.4	174.8	25.67	0.000
Fly Ash %	63.01	12.37	5.09	0.000
(Fly Ash %)^2	-0.8765	0.1966	-4.46	0.000

S = 312.113 R-Sq = 64.8% R-Sq(adj) = 60.2%



The adjusted r^2 of the quadratic model is higher than the adjusted r^2 of the simple regression model. The quadratic model explains 60.2% of the variation in Y.

Collinearity

DCOVA

- Collinearity: High correlation exists among two or more independent variables
- This means the correlated variables contribute redundant information to the multiple regression model

Collinearity

(continued)

DCOVA

- Including two highly correlated independent variables can adversely affect the regression results
  - No new information provided
  - Can lead to unstable coefficients (large standard error and low t-values)
  - Coefficient signs may not match prior expectations

Some Indications of Strong Collinearity

DCOVA

- Incorrect signs on the coefficients
- Large change in the value of a previous coefficient when a new variable is added to the model
- A previously significant variable becomes non-significant when a new independent variable is added
- The estimate of the standard deviation of the model increases when a variable is added to the model

Detecting Collinearity (Variance Inflationary Factor)

DCOVA

VIF_j is used to measure collinearity:

$$VIF_j = \frac{1}{1 - R_j^2}$$

where R^2_j is the coefficient of determination of variable X_j with all other X variables

If $VIF_j > 5$, X_j is highly correlated with the other independent variables

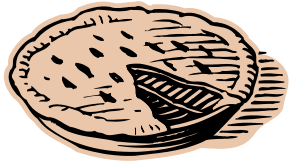
Example: Pie Sales

Week	Pie Sales	Price ($)	Advertising ($100s)
1	350	5.50	3.3
2	460	7.50	3.3
3	350	8.00	3.0
4	430	8.00	4.5
5	350	6.80	3.0
6	380	7.50	4.0
7	430	4.50	3.0
8	470	6.40	3.7
9	450	7.00	3.5
10	490	5.00	4.0
11	340	7.20	3.5
12	300	7.90	3.2
13	440	5.90	4.0
14	450	5.00	3.5
15	300	7.00	2.7

DCOVA

Recall the multiple regression equation of chapter 14:

$$\widehat{\text{Sales}} = b_0 + b_1\,(\text{Price}) + b_2\,(\text{Advertising})$$



Detecting Collinearity in Excel using PHStat

DCOVA

PHStat / regression / multiple regression …

Check the "variance inflationary factor (VIF)" box

Regression Analysis	
Price and all other X	
Regression Statistics	
Multiple R	0.0304
R Square	0.0009
Adjusted R Square	-0.0759
Standard Error	1.2153
Observations	15
VIF	**1.0009**

Output for the pie sales example:

- Since there are only two independent variables, only one VIF is reported
  - VIF is < 5
  - There is no evidence of collinearity between Price and Advertising

Detecting Collinearity in Minitab

Predictor	Coef	SE Coef	T	P	VIF
Constant	306.50	114.3	2.68	0.020	
Price	- 24.98	10.83	-2.31	0.040	1.001
Advertising	74.13	25.97	2.85	0.014	1.001

- Output for the pie sales example:
  - Since there are only two independent variables, the VIF reported is the same for each variable
    - VIF is < 5
    - There is no evidence of collinearity between Price and Advertising

Model Building

DCOVA

- Goal is to develop a model with the best set of independent variables
  - Easier to interpret if unimportant variables are removed
  - Lower probability of collinearity
- Stepwise regression procedure
  - Provide evaluation of alternative models as variables are added and deleted
- Best-subset approach
  - Try all combinations and select the best using the highest adjusted r^2 and lowest standard error

Stepwise Regression

DCOVA

- Idea: develop the least squares regression equation in steps, adding one independent variable at a time and evaluating whether existing variables should remain or be removed

- The coefficient of partial determination is the measure of the marginal contribution of each independent variable, given that other independent variables are in the model

Best Subsets Regression

DCOVA

- Idea: estimate all possible regression equations using all possible combinations of independent variables

- Choose the best fit by looking for the highest adjusted r^2 and lowest standard error

Stepwise regression and best subsets regression can be performed using PHStat or Minitab

Alternative Best Subsets Criterion

DCOVA

- Calculate the value C_p for each potential regression model

- Consider models with C_p values close to or below k + 1
  - k is the number of independent variables in the model under consideration

Alternative Best Subsets Criterion

(continued)

DCOVA

- The C_p Statistic

$$C_p = \frac{(1 - R_k^2)(n - T)}{1 - R_T^2} - (n - 2(k + 1))$$

Where k = number of independent variables included in a particular regression model

T = total number of parameters to be estimated in the full regression model

R_k^2 = coefficient of multiple determination for model with k independent variables

R_T^2 = coefficient of multiple determination for full model with all T estimated parameters

Steps in Model Building

DCOVA

1. Compile a listing of all independent variables under consideration
2. Estimate full model and check VIFs
3. Check if any VIFs > 5
   - If no VIF > 5, go to step 4
   - If one VIF > 5, remove this variable
   - If more than one, eliminate the variable with the highest VIF and go back to step 2
4. Perform best subsets regression with remaining variables …

Steps in Model Building

(continued)

DCOVA

5. List all models with C_p close to or less than (k + 1)
6. Choose the best model
   - Consider parsimony
   - Do extra variables make a significant contribution?
7. Perform complete analysis with chosen model, including residual and influence analysis
8. Transform the model if necessary to deal with violations of linearity or other model assumptions
9. Use the model for prediction and inference

Model Building Example (File: Standy)

DCOVA

Develop a model to predict Standby Hours using four independent variables (Total Staff Present, Remote Hours, Dubner Hours, and Total labor Hours)

Week	Standby Hours (Y)	Total Staff Present (X_1)	Remote Hours (X_2)	Dubner Hours (X_3)	Total Labor Hours (X_4)
1	245	338	414	323	2,001
2	177	333	598	340	2,030
3	271	358	656	340	2,226
4	211	372	631	352	2,154
5	196	339	528	380	2,078
6	135	289	409	339	2,080
7	195	334	382	331	2,073
8	118	293	399	311	1,758
9	116	325	343	328	1,624
10	147	311	338	353	1,889
11	154	304	353	518	1,988
12	146	312	289	440	2,049
13	115	283	388	276	1,796

(continued)

First Check For Collinearity

DCOVA

	A	B	C	D	E	F	G
1	**Standby Hours Analysis**						
2							
3	*Regression Statistics*						
4	Multiple R	0.7894					
5	R Square	0.6231					
6	Adjusted R Square	0.5513					
7	Standard Error	31.8350					
8	Observations	26					
9							
10	**ANOVA**						
11		*df*	*SS*	*MS*	*F*	*Significance F*	
12	Regression	4	35181.7937	8795.4484	8.6786	0.0003	
13	Residual	21	21282.8217	1013.4677			
14	Total	25	56464.6154				
15							
16		*Coefficients*	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
17	Intercept	-330.8318	110.8954	-2.9833	0.0071	-561.4514	-100.2123
18	Total Staff	1.2456	0.4121	3.0229	0.0065	0.3887	2.1026
19	Remote	-0.1184	0.0543	-2.1798	0.0408	-0.2314	-0.0054
20	Dubner	-0.2971	0.1179	-2.5189	0.0199	-0.5423	-0.0518
21	Total Labor	0.1305	0.0593	2.2004	0.0391	0.0072	0.2539

	A	B
1	**Durbin-Watson Calculations**	
2		
3	Sum of Squared Difference of Residuals	47241.6126
4	Sum of Squared Residuals	21282.8217
5		
6	Durbin-Watson Statistic	2.2197

	A	B	C	D	E
1	**Variance Inflationary Factor (*VIF*) Calculations**				
2		Regression Model			
3		Total Staff and all other X	Remote and all other X	Dubner and all other X	Total Labor and all other X
4	R Square	0.4143	0.1891	0.3147	0.4998
5	VIF	1.7074	1.2333	1.4592	1.9993

Regression Analysis: Standby versus Total Staff, ...

```
The regression equation is
Standby = - 331 + 1.25 Total Staff - 0.118 Remote
          - 0.297 Dubner + 0.131 Total Labor

Predictor        Coef   SE Coef      T      P    VIF
Constant       -330.8     110.9  -2.98  0.007
Total Staff    1.2456    0.4121   3.02  0.006  1.707
Remote       -0.11842   0.05432  -2.18  0.041  1.233
Dubner        -0.2971    0.1179  -2.52  0.020  1.459
Total Labor   0.13053   0.05932   2.20  0.039  1.999

S = 31.8350   R-Sq = 62.3%   R-Sq(adj) = 55.1%

Analysis of Variance
Source          DF     SS    MS     F      P
Regression       4  35182  8795  8.68  0.000
Residual Error  21  21283  1013
Total           25  56465

Source       DF  Seq SS
Total Staff   1   20667
Remote        1    6995
Dubner        1    2612
Total Labor   1    4907

Durbin-Watson statistic = 2.21971
```

VIF’s are small indicating little evidence of collinearity

Stepwise Results For Excel & Minitab

DCOVA

A	B	C	D	E	F	G	H
1	Stepwise Analysis for Standby Hours						
2	Table of Results for General Stepwise						
3							
4	Total Staff entered.						
5							
6		df	SS	MS	F	Significance F	
7	Regression	1	20667.3980	20667.3980	13.8563	0.0011	
8	Residual	24	35797.2174	1491.5507			
9	Total	25	56464.6154				
10							
11		Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
12	Intercept	-272.3816	124.2402	-2.1924	0.0383	-528.8008	-15.9625
13	Total Staff	1.4241	0.3826	3.7224	0.0011	0.6345	2.2136
14							
15							
16	Remote entered.						
17							
18		df	SS	MS	F	Significance F	
19	Regression	2	27662.5429	13831.2714	11.0450	0.0004	
20	Residual	23	28802.0725	1252.2640			
21	Total	25	56464.6154				
22							
23		Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
24	Intercept	-330.6748	116.4802	-2.8389	0.0093	-571.6322	-89.7175
25	Total Staff	1.7649	0.3790	4.6562	0.0001	0.9808	2.5490
26	Remote	-0.1390	0.0588	-2.3635	0.0269	-0.2606	-0.0173
27							
28							
29	No other variables could be entered into the model. Stepwise ends.						

Stepwise Regression: Standby versus Total Staff, Remote, ...

```
 Alpha-to-Enter: 0.05  Alpha-to-Remove: 0.05

Response is Standby on 4 predictors, with N = 26

Step              1       2
Constant     -272.4  -330.7

Total Staff    1.42    1.76
T-Value        3.72    4.66
P-Value       0.001   0.000

Remote               -0.139
T-Value               -2.36
P-Value               0.027

S              38.6    35.4
R-Sq          36.60   48.99
R-Sq(adj)     33.96   44.56
Mallows Cp     13.3     8.4
```

Stepwise stops with a two variable model (Total Staff & Remote Hours.) Remaining independent variables are not significant at the 0.05 level and therefore do not enter the model.

Best-Subsets Results For Excel & Minitab

DCOVA

Best-Subsets Analysis for Standby Hours

Intermediate Calculations	
$R2T$	0.6231
$1 - R2T$	0.3769
n	26
T	5
n - T	21

Model	Cp	k+1	R Square	Adj. R Square	Std. Error
X1	13.3215	2	0.3660	0.3396	38.6206
X1X2	8.4193	3	0.4899	0.4456	35.3873
X1X2X3	7.8418	4	0.5362	0.4729	34.5029
X1X2X3X4	5.0000	5	0.6231	0.5513	31.8350
X1X2X4	9.3449	4	0.5092	0.4423	35.4921
X1X3	10.6486	3	0.4499	0.4021	36.7490
X1X3X4	7.7517	4	0.5378	0.4748	34.4426
X1X4	14.7982	3	0.3754	0.3211	39.1579
X2	33.2078	2	0.0091	-0.0322	48.2836
X2X3	32.3067	3	0.0612	-0.0205	48.0087
X2X3X4	12.1381	4	0.4591	0.3853	37.2608
X2X4	23.2481	3	0.2238	0.1563	43.6540
X3	30.3884	2	0.0597	0.0205	47.0345
X3X4	11.8231	3	0.4288	0.3791	37.4466
X4	24.1846	2	0.1710	0.1365	44.1619

Best Subsets Regression: Standby versus Total Staff, Remote, ...

Response is Standby

Vars	R-Sq	R-Sq(adj)	Mallows Cp	S	Total Staff	Remote	Dubner	Total Labor
1	36.6	34.0	13.3	38.621	X			
1	17.1	13.7	24.2	44.162				X
1	6.0	2.1	30.4	47.035			X	
2	49.0	44.6	8.4	35.387	X	X		
2	45.0	40.2	10.6	36.749	X		X	
2	42.9	37.9	11.8	37.447			X	X
3	53.8	47.5	7.8	34.443	X		X	X
3	53.6	47.3	7.8	34.503	X	X	X	
3	50.9	44.2	9.3	35.492	X	X		X
4	62.3	55.1	5.0	31.835	X	X	X	X

Best-subsets often yields numerous candidate models. Both adjusted r-squared and C_p are used to pick a model to use.

Residual Analysis Should Be Done On The Chosen Model

DCOVA

- Utilizing the model with all four independent variables residual analysis (shown on the next three slides) reveals:
  - No autocorrelation (based on the Durbin-Watson test)
  - No apparent patterns in residuals versus the four independent variables
  - No evidence of unequal variance
  - Only a moderate departure from normality
  - No overly influential observations

Check For Independence

DCOVA

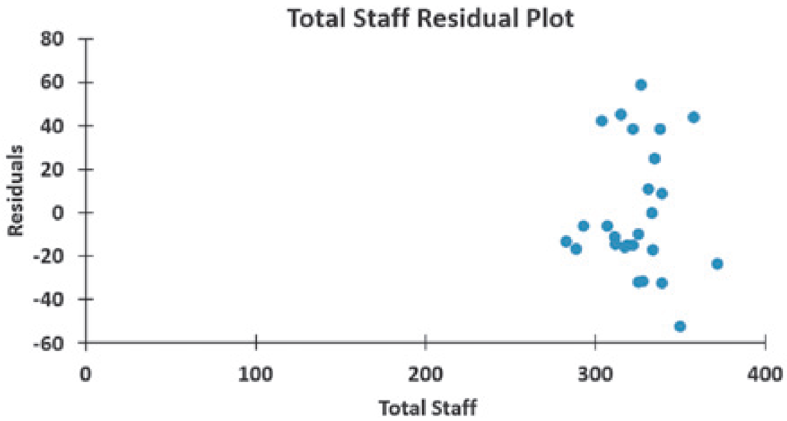
Regression Analysis: Standby versus Total Staff, ...

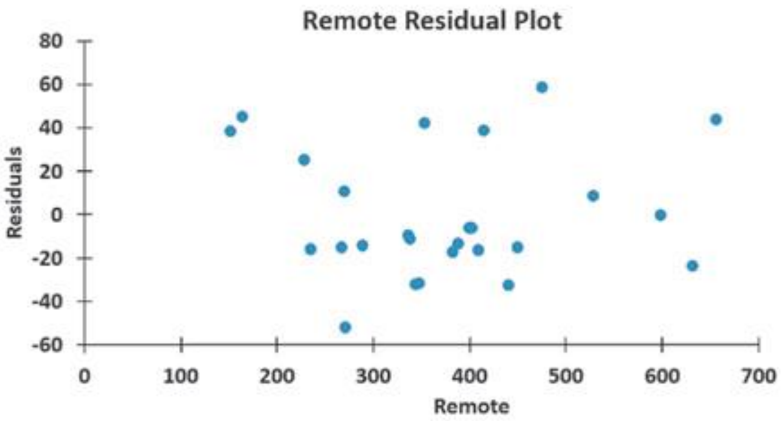
```
The regression equation is
Standby = - 331 + 1.25 Total Staff - 0.118 Remote
          - 0.297 Dubner + 0.131 Total Labor

Predictor       Coef   SE Coef      T      P    VIF
Constant      -330.8     110.9  -2.98  0.007
Total Staff   1.2456    0.4121   3.02  0.006  1.707
Remote      -0.11842   0.05432  -2.18  0.041  1.233
Dubner       -0.2971    0.1179  -2.52  0.020  1.459
Total Labor  0.13053   0.05932   2.20  0.039  1.999

S = 31.8350   R-Sq = 62.3%   R-Sq(adj) = 55.1%

Analysis of Variance
Source          DF     SS    MS     F      P
Regression       4  35182  8795  8.68  0.000
Residual Error  21  21283  1013
Total           25  56465

Source       DF  Seq SS
Total Staff   1   20667
Remote        1    6995
Dubner        1    2612
Total Labor   1    4907

Durbin-Watson statistic = 2.21971
```

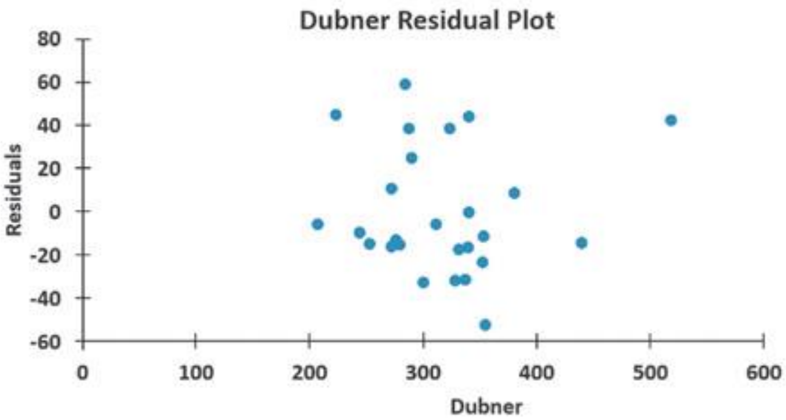
Durbin-Watson is > 2 indicating no positive autocorrelation

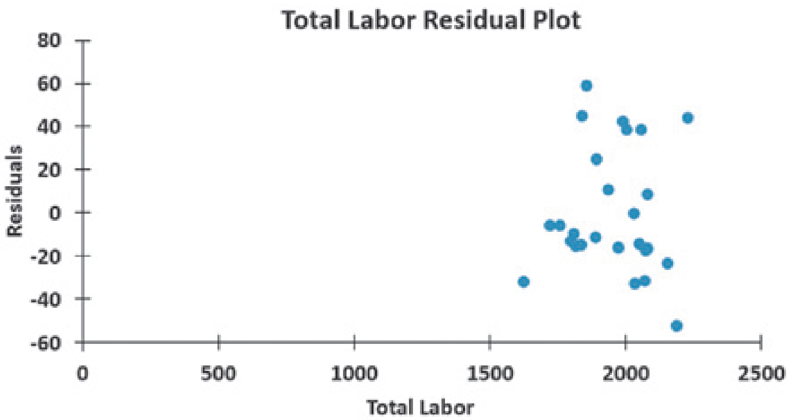
Check For Correct Functional Form

DCOVA





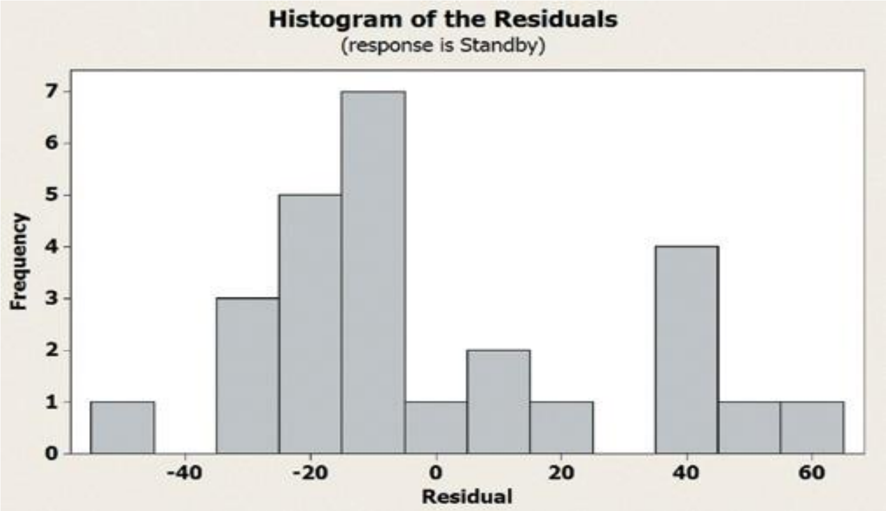


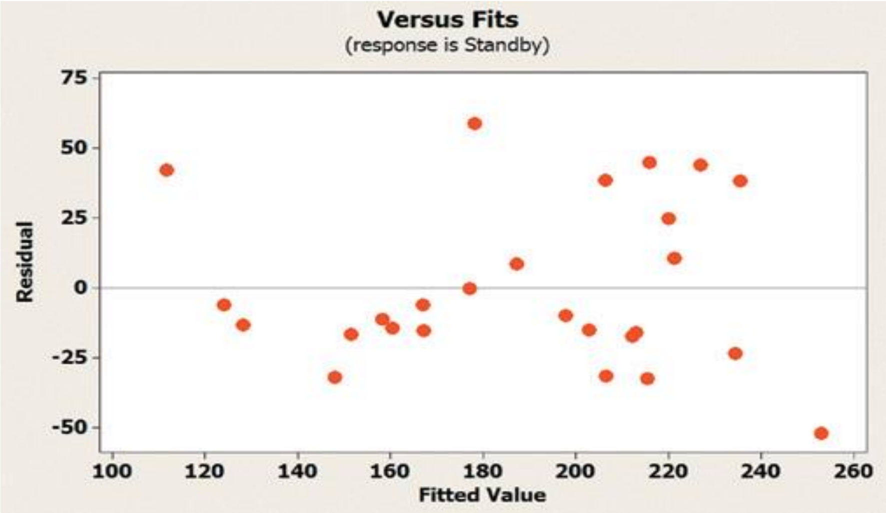


No apparent patterns on any of these plots

Checks For Normality & Unequal Variance

DCOVA





Histogram shows slight departure from normality.

No evidence of unequal variance present in this residual plot

Check For Influential Observations

DCOVA

↓	C1	C2	C3	C4	C5	C6	C7	C8
	Standby	Total Staff	Remote	Dubner	Total Labor	TRES1	HI1	COOK1
1	245	338	414	323	2001	1.26648	0.057851	0.019147
2	177	333	598	340	2030	-0.00450	0.159009	0.000001
3	271	358	656	340	2226	1.74109	0.308626	0.246769
4	211	372	631	352	2154	-0.88635	0.317663	0.073903
5	196	339	528	380	2078	0.28517	0.117963	0.002275
6	135	289	409	339	2080	-0.66415	0.404904	0.061665
7	195	334	382	331	2073	-0.55135	0.066692	0.004493
8	118	293	399	311	1758	-0.20251	0.177819	0.001859
9	116	325	343	328	1624	-1.38665	0.453697	0.305928
10	147	311	338	353	1889	-0.36082	0.080160	0.002367
11	154	304	353	518	1988	2.06732	0.521708	0.806606
12	146	312	289	440	2049	-0.50740	0.239542	0.016814
13	115	283	388	276	1796	-0.47510	0.267786	0.017142
14	161	307	402	207	1720	-0.20841	0.219149	0.002554
15	274	322	151	287	2056	1.42189	0.241543	0.122799
16	245	335	228	290	1890	0.84801	0.155157	0.026771
17	201	350	271	355	2187	-2.00716	0.240144	0.222548
18	183	339	440	300	2032	-1.06250	0.073308	0.017752
19	237	327	475	284	1856	2.10290	0.101265	0.085690
20	175	328	347	337	2068	-1.02770	0.071640	0.016257
21	152	319	449	279	1813	-0.49356	0.105621	0.005969
22	188	325	336	244	1808	-0.31659	0.107193	0.002515
23	188	322	267	253	1834	-0.48404	0.100514	0.005434
24	197	317	235	272	1973	-0.52597	0.123908	0.008105
25	261	315	164	223	1839	1.63790	0.193025	0.118818
26	232	331	270	272	1935	0.34618	0.094110	0.002599

Even though a few observations are flagged by the studentized residual and h_i criteria, Cooks D does not flag any observations as overly influential.

The Final Model Building Step Is Model Validation

DCOVA

- Models can be validated via multiple methods
  - Collect new data and compare the results
  - Compare the results of the regression model to previous results
  - If the data set is large enough, split the data into two parts and cross-validate the results
    - To do this you split the data prior to building the model and use one half of the data to build the model and the other half to validate the model

Pitfalls and Ethical Considerations

To avoid pitfalls and address ethical considerations:

- Understand that interpretation of the estimated regression coefficients are performed holding all other independent variables constant
- Evaluate residual plots for each independent variable
- Evaluate interaction terms

Additional Pitfalls and Ethical Considerations

(continued)

To avoid pitfalls and address ethical considerations:

- Obtain VIFs for each independent variable before determining which variables should be included in the model
- Examine several alternative models using best-subsets regression
- Use other methods when the assumptions necessary for least-squares regression have been seriously violated

Chapter Summary

In this chapter we discussed

- The quadratic regression model
- Collinearity
- Model building
  - Stepwise regression
  - Best subsets
- The pitfalls & ethical considerations in multiple regression