# Basic Business Statistics
## 11th Edition

# Chapter 15

## Multiple Regression Model Building

# Learning Objectives

**In this chapter, you learn:**

- To use quadratic terms in a regression model
- To use transformed variables 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

- 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

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

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

# Quadratic Regression Model

Model form:

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

- where:

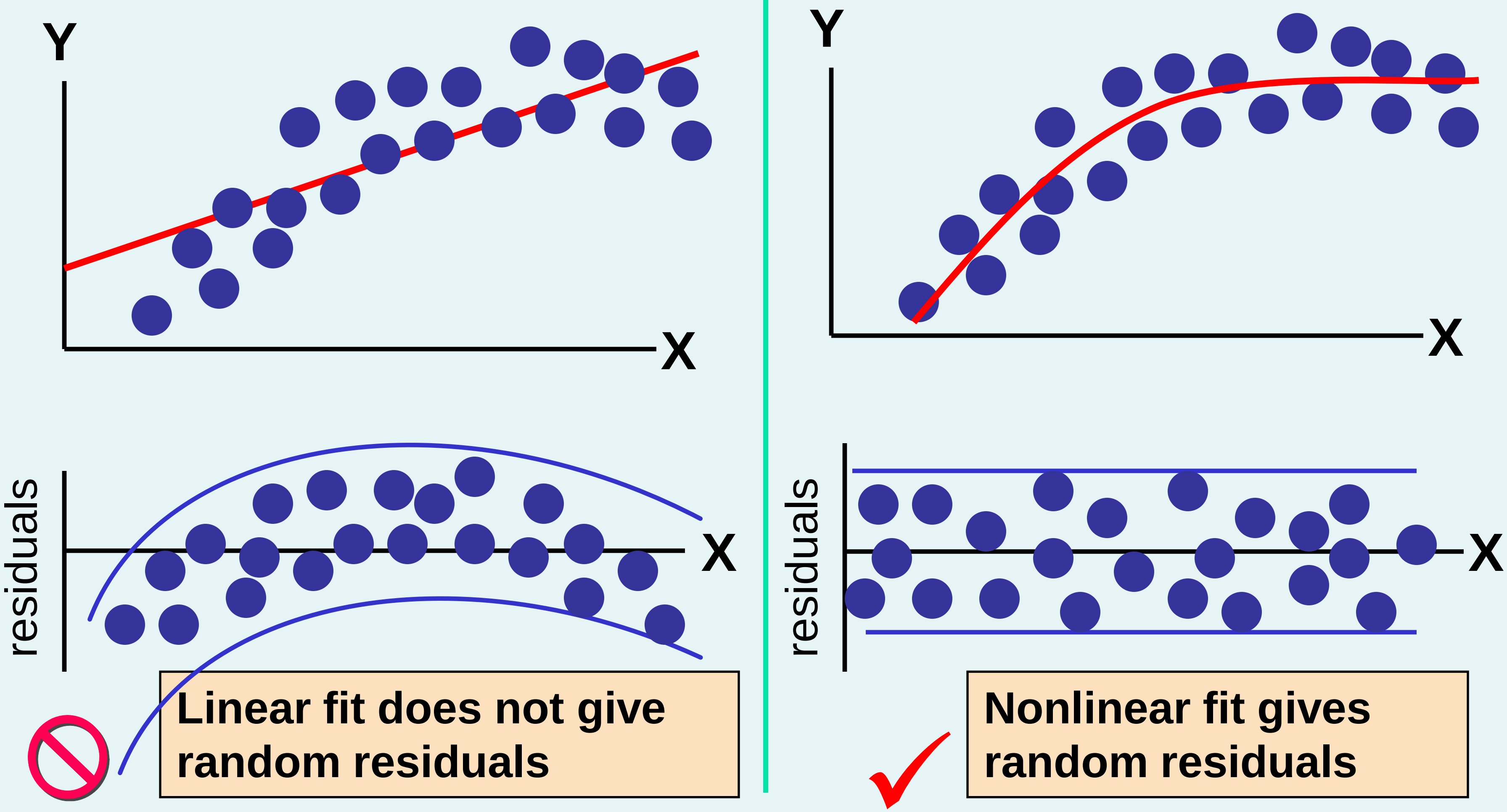
  $\beta_0$ = Y intercept

  $\beta_1$ = regression coefficient for linear effect of X on Y

  $\beta_2$ = regression coefficient for quadratic effect on Y

  $\varepsilon_i$ = random error in Y for observation i

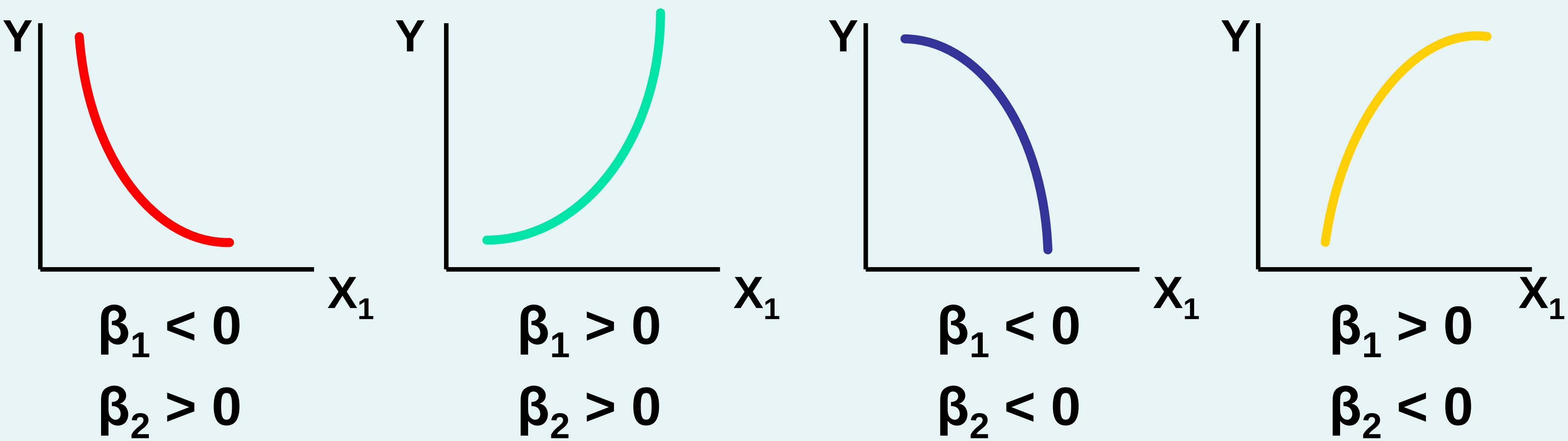
# Linear vs. Nonlinear Fit



# Quadratic Regression Model

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

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




$\beta_1$ = the coefficient of the linear term

$\beta_2$ = the coefficient of the squared term

# Testing the Overall Quadratic Model

- 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)

$H_1$: $\beta_1$ and/or $\beta_2 \neq 0$ (there is a relationship between X and Y)

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

# Testing for Significance: Quadratic Effect

- 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)*

- Testing the Quadratic Effect
  - Consider the quadratic regression equation

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

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)*

- 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

$\beta_2$ = hypothesized slope (zero)

$S_{b_2}$ = standard error of the slope

# Testing for Significance: Quadratic Effect

*(continued)*

- Testing the Quadratic Effect

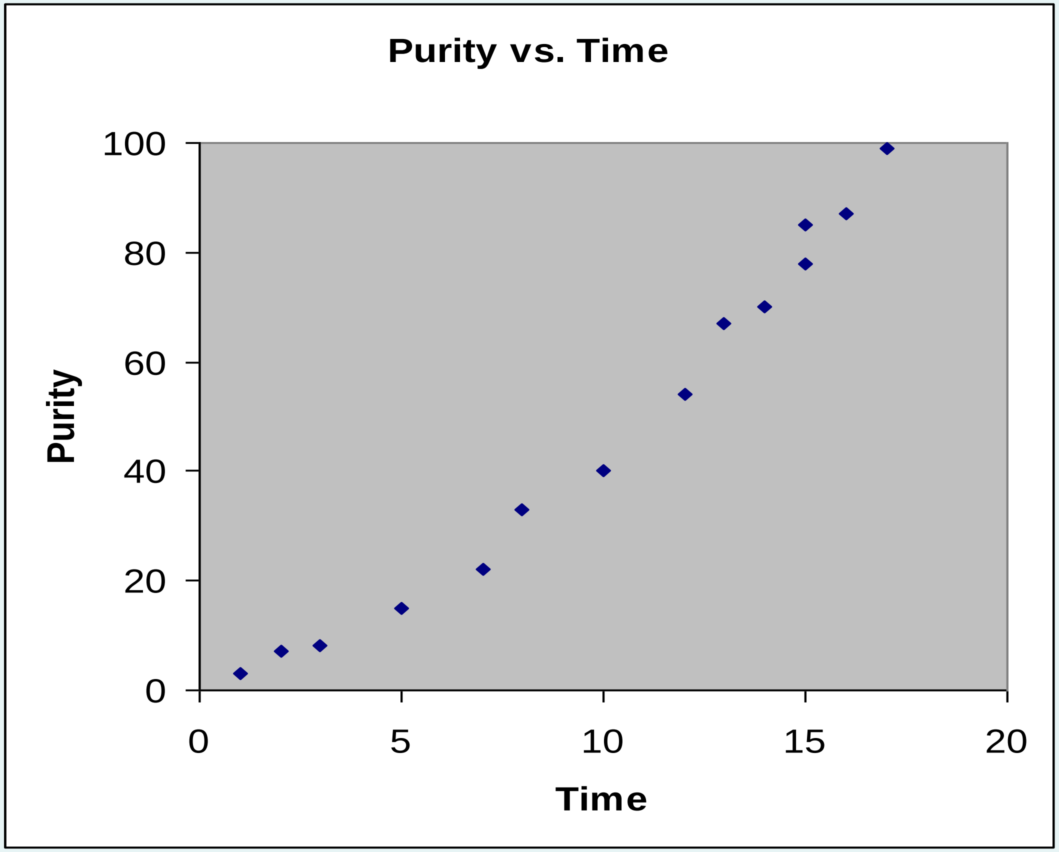
Compare $r^2$ from simple regression to adjusted $r^2$ from the quadratic model

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

# Example: Quadratic Model

| 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: Quadratic Model

*(continued)*

- Simple regression results:

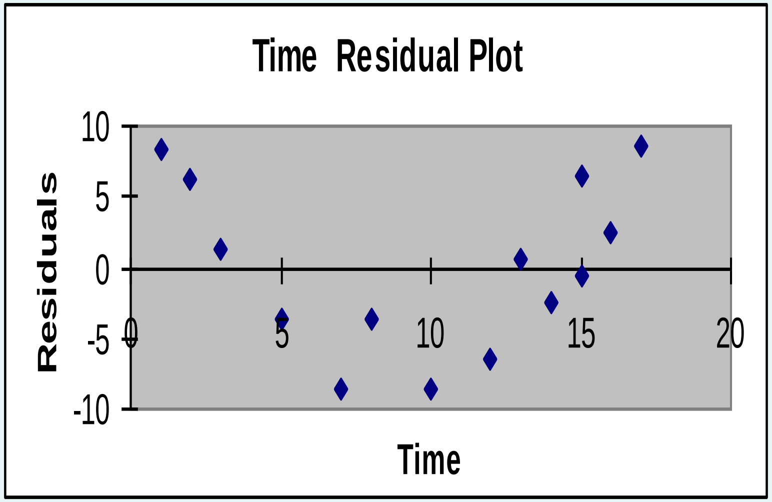
$$\hat{Y} = -11.283 + 5.985 \text{ Time}$$

| | *Coefficients* | *Standard Error* | *t Stat* | *P-value* |
|---|---|---|---|---|
| Intercept | -11.28267 | 3.46805 | -3.25332 | 0.00691 |
| Time | 5.98520 | 0.30966 | **19.32819** | 2.078E-10 |

t statistic, F statistic, and $r^2$ are all high, but the residuals are not random:

| *Regression Statistics* | |
|---|---|
| R Square | **0.96888** |
| Adjusted R Square | 0.96628 |
| Standard Error | **6.15997** |

| *F* | *Significance F* |
|---|---|
| **373.57904** | 2.0778E-10 |



# Example: Quadratic Model in Excel

*(continued)*

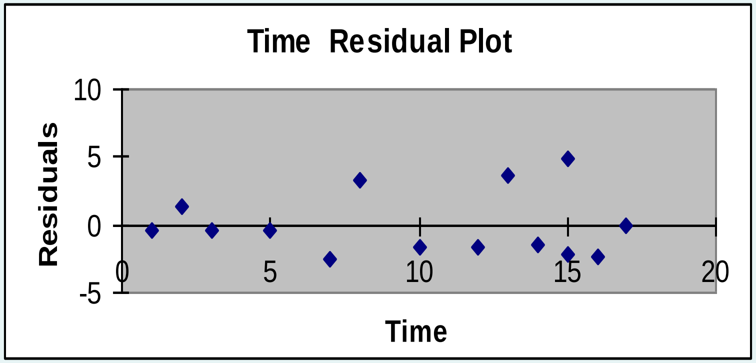
- Quadratic regression results:

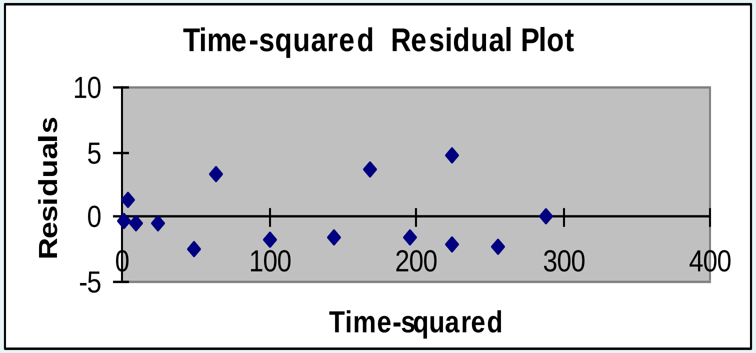
$$\hat{Y} = 1.539 + 1.565\ \text{Time} + 0.245\ (\text{Time})^2$$

| | ***Coefficients*** | ***Standard Error*** | ***t Stat*** | ***P-value*** |
|---|---|---|---|---|
| Intercept | 1.53870 | 2.24465 | 0.68550 | 0.50722 |
| Time | 1.56496 | 0.60179 | 2.60052 | 0.02467 |
| Time-squared | 0.24516 | 0.03258 | **7.52406** | 1.165E-05 |

| ***Regression Statistics*** | |
|---|---|
| R Square | 0.99494 |
| Adjusted R Square | **0.99402** |
| Standard Error | **2.59513** |

| ***F*** | ***Significance F*** |
|---|---|
| **1080.7330** | 2.368E-13 |




The quadratic term is significant and improves the model: adj. $r^2$ is higher and $S_{YX}$ is lower, residuals are now random

# Example: Quadratic Model in Minitab

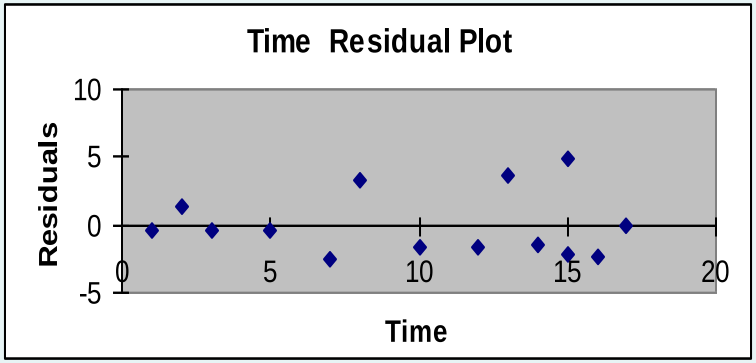
## Quadratic regression results:

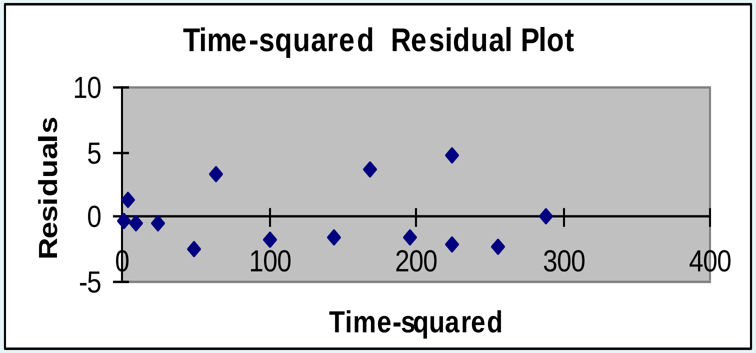
$$Y = 1.539 + 1.565\ \text{Time} + 0.245\ (\text{Time})^2$$

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 quadratic term is significant and improves the model: adj. $r^2$ is higher and $S_{YX}$ is lower, residuals are now random

# Using Transformations in Regression Analysis

Idea:

- non-linear models can often be transformed to a linear form
  - Can be estimated by least squares if transformed
- transform X or Y or both to get a better fit or to deal with violations of regression assumptions
- Can be based on theory, logic or scatter plots

# The Square Root Transformation

- The square-root transformation

$$Y_i = \beta_0 + \beta_1 \sqrt{X_{1i}} + \varepsilon_i$$

- Used to
  - overcome violations of the constant variance assumption
  - fit a non-linear relationship

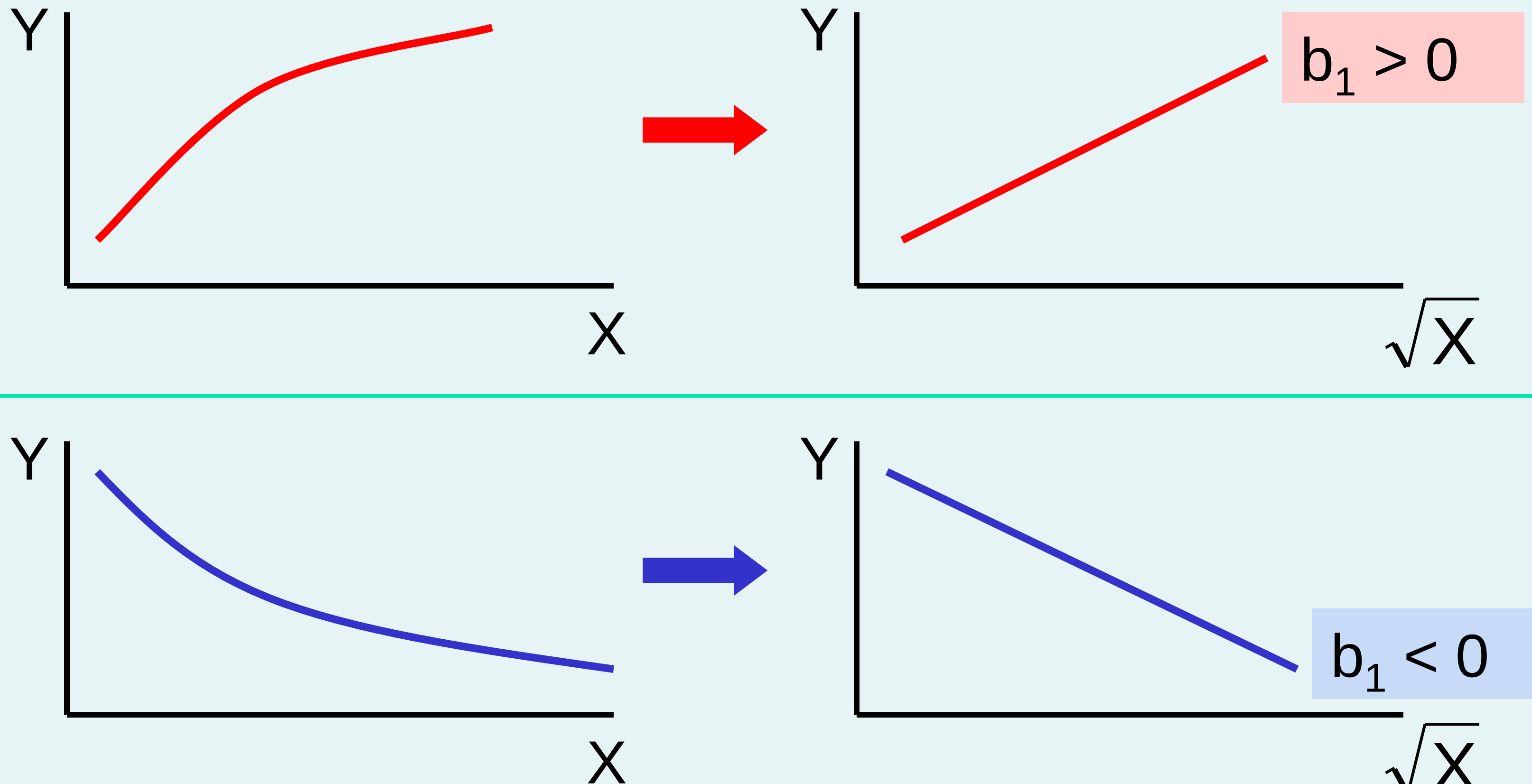
# The Square Root Transformation

*(continued)*

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

$$Y_i = \beta_0 + \beta_1 \sqrt{X_{1i}} + \varepsilon_i$$

- Shape of original relationship
- Relationship when transformed



# The Log Transformation

## The Multiplicative Model:

- Original multiplicative model

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

- Transformed multiplicative model

$$\log Y_i = \log \beta_0 + \beta_1 \log X_{1i} + \log \varepsilon_i$$

## The Exponential Model:

- Original multiplicative model

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

- Transformed exponential model

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

# Interpretation of coefficients

For the multiplicative model:

$$\log Y_i = \log \beta_0 + \beta_1 \log X_{1i} + \log \varepsilon_i$$

- When both dependent and independent variables are logged:
  - The coefficient of the independent variable $X_k$ can be interpreted as : a 1 percent change in $X_k$ leads to an estimated $b_k$ percentage change in the average value of Y. Therefore $b_k$ is the elasticity of Y with respect to a change in $X_k$ .

# Collinearity

- 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)*

- 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

- 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)

$VIF_j$ 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 |

Recall the multiple regression equation of chapter 14:

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

# Detecting Collinearity in Excel using PHStat

PHStat / regression / multiple regression …

Check the "variance inflationary factor (VIF)" box

| Regression Analysis | |
|---|---|
| Price and all other X | |
| *Regression Statistics* | |
| Multiple R | 0.030438 |
| R Square | 0.000926 |
| Adjusted R Square | -0.075925 |
| Standard Error | 1.21527 |
| Observations | 15 |
| **VIF** | **1.000927** |

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

- 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

- 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

- 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

# Alternative Best Subsets Criterion

- 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)*

- 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

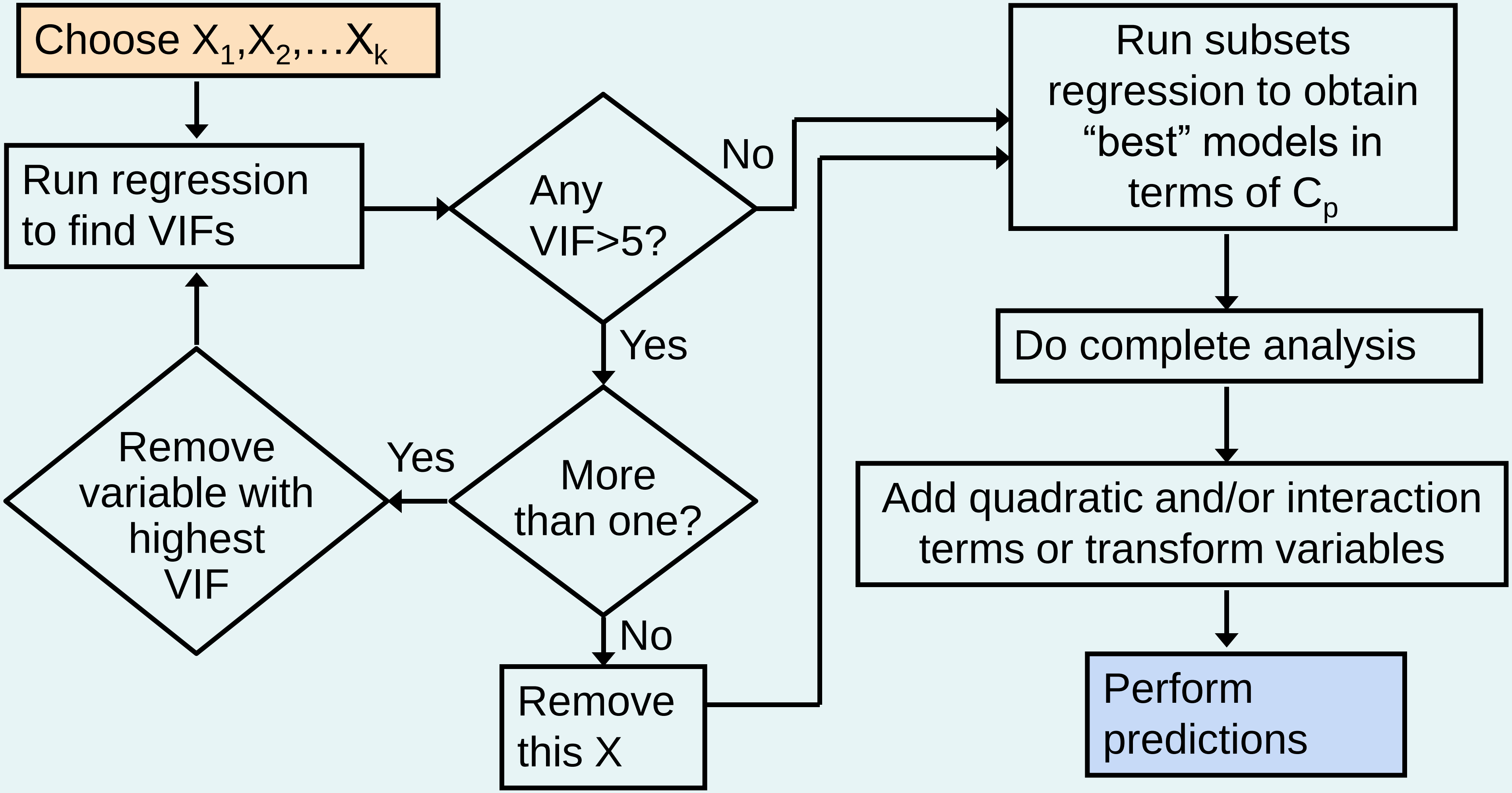
1. Compile a listing of all independent variables under consideration
2. Estimate full model and check VIFs
3. Check if any VIFs > 5
   - If no VIF > 5, go to step 4
   - If one VIF > 5, remove this variable
   - If more than one, eliminate the variable with the highest VIF and go back to step 2
4. Perform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

# Steps in Model Building

*(continued)*

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 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 Flowchart



# 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

- Developed the quadratic regression model
- Discussed using transformations in regression models
  - The multiplicative model
  - The exponential model
- Described collinearity
- Discussed model building
  - Stepwise regression
  - Best subsets
- Addressed pitfalls in multiple regression and ethical considerations