

Chapter 8: Multiple Regression Models

Quadratic Regression

Residual Standard Deviations

Indicator Variables

Several (Continuous) Predictor Variables

Residual Plots

Chap. 8 - page 1

Quadratic Regression (page 208)

Observed Response
= Fitted Value + Residual

$$\hat{y} = b_0 + b_1x + b_2x^2$$

b_0 is the **constant term**

b_1 is the **linear term**

b_2 is the **quadratic term**
The least squares regression quadratic curve

Chap. 8 - page 3

Review from Chapter 7

A **regression model** has two parts:

- A mathematical curve that summarizes the general tendency of the relationship between the variables
- A measure of the amount of variation in the data around that mathematical curve

Chap. 8 - page 2

Choose b_0 , b_1 , and b_2 to minimize

$$\sum_{i=1}^n \left[y_i - (b_0 + b_1x_i + b_2x_i^2) \right]^2$$

We will let Minitab (or other statistical software) find the solution for the best values of b_0 , b_1 , and b_2 with particular data sets.. We simply tell the software that x and x^2 are the predictor variables.

There are no simple equations for b_0 , b_1 , and b_2 unless you use matrix and vector notation. See the Appendix in the textbook, page 622.

Chap. 8 - page 4

Example: Average Heights of Young Girls at Various Ages. (Exercise 8.4D, page 215)

“All models are wrong; some models are useful!” George Box

Example 7.6C and 8.2C (page 211) shows an example where neither a straight line nor a quadratic curve give an acceptable model.

Chap. 8 - page 5

Degrees of Freedom (page 213)

The **degrees of freedom for residuals** is given by the number of observations minus the number of coefficients in the fitted regression model.

For a simple straight-line model this gives $n-2$ since we are fitting a slope and an intercept.

For a quadratic curve model this gives $n-3$ since we are fitting three coefficients.

Chap. 8 - page 7

Residual Standard Deviation s (page 213)

- Compute the deviations from the fitted values $y_1 - \hat{y}_1, y_2 - \hat{y}_2, \dots, y_n - \hat{y}_n$, (these are the residuals!)
- Square the deviations (residuals) and sum them
- Divide the sum by the degrees of freedom for the residuals
- Finally, take the positive square root to get s

In symbols

$$s = \sqrt{\frac{\sum (y_i - \hat{y}_i)^2}{df}}$$

Chap. 8 - page 6

Comparison of Models (page 214)

The smaller the value of the residual standard deviation, the better the regression curve fits the data.

We can use the value of s to compare the fit of different regression models—the model with the smaller s is the better the model.

Chap. 8 - page 8

Several Straight Lines:

Indicator Variables (page 215)

Can use categorical variables in regression models—especially in conjunction with continuous variables.

An **indicator variable** is a binary variable that is equal to 1 for an observation that belongs to one particular category and is equal to 0 for an observation that does not belong to that category.

An indicator variable indicates group membership.

Chap. 8 - page 9

For group 1 the model reduces to

$$\hat{y} = (b_0 + b_2) + b_1x$$

while for group 2 the model reduces to

$$\hat{y} = b_0 + b_1x$$

These two lines have the same slope b_1 but they have different constant terms ($b_0 + b_2$) and b_0 .

Therefore they represent parallel lines.

Chap. 8 - page 11

Parallel-Lines Models (page 215)

Let x be a continuous predictor variable and let z be an indicator variable for two groups.

$z = 1$ for group 1 and $z = 0$ for group 2.

Consider the regression model

$$\hat{y} = b_0 + b_1x + b_2z$$

Chap. 8 - page 10

Example:

City Gallons per mile versus car weight and transmission type—automatic or manual. ($n = 72$)

The regression line without using the transmission type information is

$$GP100M = -1.26 + 0.00253 \text{ Weight}$$

and the residual standard deviation is

$$s = 0.6673$$

Chap. 8 - page 12

Using the transmission type information in a parallel-lines model gives

Fitted Model:

$$GP100M = -1.03 + 0.00229 \text{ Weight} + 0.691 \text{ Trans}$$

or

$$GP100M = -1.03 + 0.00229 \text{ Weight} + 0.691 \\ = -0.339 + 0.00229 \text{ Weight} \\ \text{for automatic}$$

$$GP100M = -1.03 + 0.00229 \text{ Weight} \\ \text{for manual}$$

The residual standard deviation is

$$s = 0.6401$$

Chap. 8 - page 13

Now let's do the prediction using the parallel-lines model that uses the transmission type.

We have

$$GP100M = -1.03 + 0.00229 \text{ Weight} + 0.691 \text{ Trans} \\ = -1.03 + 0.00229 (3145) + 0.691(1) \\ = -0.339 + 7.020205 \\ = 6.9$$

In fact for the Cirrus

$$GP100M = 100(1/14) = 7.1$$

so the parallel lines model predicts better.

Chap. 8 - page 15

Prediction for a New Car

Chrysler Cirrus

weighs 3145 pounds (and has an automatic transmission)

Suppose we first ignore the transmission type and just use the straight line model based on weight.

The regression line is

$$GP100M = -1.26 + 0.00253 \text{ Weight}$$

so the prediction is

$$GP100M = -1.26 + 0.00253 (3145) \\ = -1.26 + 7.95685 \\ = 6.7 \text{ gallons per 100 miles}$$

Chap. 8 - page 14

Categorical Variables with Several Categories (page 218)

If a categorical variable has k different values, we need $k-1$ different indicator variables to account for the categorical variable in a parallel lines model. (page 219)

Examples:

For male-female you need **one** indicator variable.

For construction grades of low, medium and high, you need **two** indicator variables:

- One to indicate whether the house is low grade or not
- One to indicate whether the house is medium grade or not
- If a house is neither low nor medium grade, it is high grade by default

Chap. 8 - page 16

Low = 1 for low grade construction and Low = 0 otherwise

Medium = 1 for medium grade construction and Low = 0 otherwise

Then model might be

$$\widehat{Market} = b_0 + b_1 Sqft + b_2 Low + b_3 Medium$$

Chap. 8 - page 17

Fitting Planes and Curved Surfaces (page 221)

A Regression Plane

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2$$

Chap. 8 - page 18

The General Linear Model (page 230)

Fitted Regression Equation

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2 + \dots + b_k x_k$$

$$\sum (y_i - \bar{y})^2 = \text{Total SS}$$

$$\sum (\hat{y}_i - \bar{y})^2 = \text{Regression SS}$$

$$\sum (y_i - \hat{y}_i)^2 = \text{Error SS}$$

(Error SS is just Residual SS)

Total SS = Regression SS + Error SS

Chap. 8 - page 19

Multiple Coefficient of Determination (page 234)

Also called R-Squared and denoted R^2

$$R^2 = 100 \left[1 - \frac{\text{Error SS}}{\text{Total SS}} \right]$$

Adjusted Multiple Coefficient of Determination

$$\text{adjusted } R^2 = 100 \left[1 - \frac{s^2}{s_y^2} \right]$$

Chap. 8 - page 20

Residual Plots (page 225)

When there are several predictor variables, the most basic residual plot is to graph residuals versus the corresponding fitted values.

As before, residual plots containing nonrandom patterns indicate that the model can be improved.

Example: Gallons per 100 Miles versus weight and transmission type.

Chap. 8 - page 21

Created variables:

Points per minute
= points per game \times games/
minutes

Minutes per game = Minutes/games

Assists per game = Assists/games

Rebounds per minute
= rebounds \times games / minutes

Chap. 8 - page 23

Case Study:

The Effectiveness of NBA Guards.

Data on all NBA guards for 92-93 season.

Source: *The Pro Basketball Bible*, by Jordan Cohn, 1994.

Variables:

Player's name
Player's height (in centimeters)
Number of games played in
Total minutes played in season
Player's age
Point scored per game
Assists per game
Rebounds per game
Field goal percentage
Free throw percentage

Chap. 8 - page 22

Regression Relationships:

Model 1:

$\text{Pts/Min} = 0.242 + 0.00669 \text{ Min/Game}$

with $s = 0.08367$,

$R^2 = 32.1\%$, and $R^2(\text{adj}) = 31.4\%$

Model 2:

$\text{Pts/Min} = -0.477 + 0.00663 \text{ Min/Game}$
 $+ 0.00379 \text{ Height}$

with $s = 0.08015$,

$R^2 = 38.4\%$, and $R^2(\text{adj}) = 37.0\%$

Chap. 8 - page 24

Model 3:

$$\text{Pts/Min} = -0.781 + 0.00563 \text{ Min/Game} \\ + 0.00412 \text{ Height} + 0.00340 \% \text{FT}$$

with $s = 0.07612$,

$$R^2 = 45.0\%, \text{ and } R^2(\text{adj}) = 43.2\%$$

Prediction for Michael Jordan:

$$\text{Pts/Min} = -0.781 + 0.00563 (39.3205) \\ + 0.00412 (198) + 0.00340 (83.7) \\ = 0.54$$

In fact, Michael Jordan's Pts/Min was 0.83.

This is $(0.83 - 0.54)/0.07612 = 3.8$ residual standard deviations above the prediction of the model.

Truly, outstanding performance for someone of his height, his playing time, and his free throw percentage.

Prediction for B. J. Armstrong:

$$\text{Pts/Min} = -0.781 + 0.00563 (30.7654) \\ + 0.00412 (188) + 0.00340 (86.1) \\ = 0.46$$

BJ's actual Pts/Min was 0.40 so he is $(0.40 - 0.46)/0.07612 = -0.08$ residual standard deviations below the prediction of the model.

For someone of his height, his playing time, and his free throw percentage, BJ's performance, is just barely below average.