Chapter 9 – Multiple Regression

Section 9.1

1. House prices.

a) $\widehat{Price} = 20,986.09 - 7483.10\,Bedrooms + 93.84\,LivingArea$
$= 20,986.09 - 7483.10(2) + 93.84(1000) = 99,859.89$

According to the multiple linear regression model, we would expect an Upstate New York home with 2 bedrooms with 1000 square feet of living space to have a price of approximately \$99,859.89.

b) The residual is \$135,000 – \$99,859.89 = \$35,140.11.

c) The house sold for about \$35,100 more than our estimate.

2. Candy sales.

a) $\widehat{Calories} = 28.4 + 11.37\,Fat(g) + 2.91Sugar(g) = 28.4 + 11.37(15) + 2.91(20) = 257.15$; According to the multiple linear regression model, we would expect a chocolate bar with 15g of fat and 20g of sugar to have approximately 257.15 calories.

b) The residual is 227 – 257.15 = –30.15 calories.

c) Her candy has about 30 fewer calories than expected from the model.

Section 9.2

3. Movie profits.

a) $\widehat{USGross} = -52.3692 + 0.9723\,Budget + 0.3872\,RunTime + 0.6403\,CriticsScore$

b) After allowing for the effects of *RunTime* and *CriticsScore*, each million dollars spent making a film yields about 0.9723 million dollars in gross revenue.

4. Movie profits again.

The manager is incorrectly interpreting the coefficient causally. The model says that longer films had larger gross incomes (after allowing for *Budget* and *CriticsScore*), but it doesn't say that making a movie longer will increase its gross. In fact, lengthening arbitrarily would, for example, probably reduce the *CriticsScore* rating.

Section 9.3

5. More movies profits.

a) Linearity: The plot is reasonably linear with no bends.

b) Equal spread: The plot fails the Equal Spread condition. It is much more spread out to the right than on the left.

c) Normality: A scatterplot of two of the variables doesn't tell us anything about the distribution of the residuals.

6. Movie residuals.

a) Linearity: A histogram doesn't show whether a relationship is linear or not.

b) Nearly Normal: The histogram is unimodal and slightly skewed to the right. But it does seem to have an outlier on the right, which would violate the Nearly Normal condition**.**

c) Equal Spread: A histogram doesn't show whether there is equal spread in different parts of the data; we need a scatterplot for that.

Section 9.4

7. Movie profits once more.

a) The partial regression plot is for US Gross residual and Budget residual, so the slope of the line is the Budget coefficient from the multiple regression, 0.9723.

b) *Avatar* has a budget and US gross that is influential in the regression, and pulls the line toward it. If *Avatar*'s point were removed, the slope of the line would decrease.

8. Hopkins Forest wind.

a) $\widehat{MaxWind} = 22.1975 - 0.113 AvgTemp + 2.013 Precip$

b) After allowing for *AvgTemp*, a 1-inch increase in rainfall is associated with an increase in *MaxWind* speed of 2.013 mph. (Perhaps when it rains, the winds pick up too.)

c) The partial regression plot is for the Maximum Wind residual and the Precipitation residual, so the slope of the line is the Precipitation coefficient from the multiple regression, 2.013.

d) The Precipitation and Maximum Wind speed for the date of Hurricane Irene was influential on the regression. If that point were removed, the slope of the line would decrease and possibly become very close to zero.

Section 9.5

9. Indicators.

a) Use an indicator. Code Male = 0, Female = 1 (or the reverse).

b) Treat it as a quantitative predictor.

c) Use an indicator. Code older than 65 = 1, Younger than 65 = 0.

10. More indicators.

a) Treat it as quantitative predictor.

b) Use an indicator. Code 1 for buildings with elevators, 0 otherwise.

c) Use two indicators. One could be called *Child* and would be 1 for children and 0 otherwise. The other could be called *Senior* and would be 1 for those over 65 and 0 for others. You can't use three indicators, but must leave out one of the set. Other choices are possible.

11. Interpretations.

a) There are two problems with this interpretation. First, the other predictors are not mentioned. Secondly, the prediction should be stated in terms of a mean, not a precise value.

b) This is a correct interpretation.

c) This interpretation attempts to predict in the wrong direction. This model cannot predict *lotsize* from *price.*

d) R^2 concerns the fraction of variability accounted for by the regression model, not the fraction of data values.

12. More interpretations.

a) Regression predicts mean values, not a precise value.

b) *Sales* increases should be stated in terms of a mean, not a precise value. In addition, we cannot assume causality. The word "makes" could be replaced by the phrase "is associated with".

c) Changes in one predictor variable are not necessarily associated with changes in another predictor variable.

d) This is a correct interpretation.

13. Predicting final exams.

a) $\widehat{Final} = -6.7210 + 0.2560(Test1) + 0.3912(Test2) + 0.9015(Test3)$

b) $R^2 = 77.7\%,$ which means that 77.7% of the variation in *Final* grade is accounted for by the multiple regression model.

c) According to the multiple regression model, each additional point on *Test3* is associated with an average increase of 0.9015 points on the final, for students with given *Test1* and *Test2* scores.

d) Test scores are probably collinear. If we are only concerned about predicting the final exam score, *Test1* may not add much to the regression. However, we would expect it to be associated with the final exam score.

14. Scottish hill races.

a) $\widehat{Time} = -10.3723 + 4.04204(Distance) + 0.034227(Climb)$; Men's record *Time* is associated with both *Distance* and *Climb.* An increase of one kilometer in *Distance* is associated with an average increase of 4.04 minutes in the men's record *Time* for races with a given *Climb*. An increase in one meter of *Climb* is associated with an average increase of 0.0342 minutes in the men's record *Time* for races with a given *Distance*.

b) $R^2 = 98\%;$ which means that 98% of the variation in men's record *Time* is accounted for by the multiple regression model.

c) According to the multiple regression model, an increase in one meter of *Climb* is associated with an average increase of 0.034 minutes in the men's record *Time* for races with a given distance.

15. Attendance 2016.

a) *Won* and *Runs* are probably correlated. Including *Won* in the model is then very likely to change the coefficient of *Runs*, which now must be interpreted after allowing for the effects of *Won*.

b) The Indians' actual *Attendance* was less than we would predict for the number of *Runs* they scored.

16. Candy bars: calories.

a) On average, a serving of candy bars has 3.9 fewer calories per gram of sugar.

b) After accounting for *Protein/serving*, the model predicts candy bars have on average of 2.18 fewer calories/serving per gram of sugar/serving.

c) The change in the *Sugar* coefficient when *Protein* is added to the model means that *Protein* and *Sugar* are likely correlated for these data.

17. Home prices.

a) $\widehat{Price} = -152,037 + 9530 Bathrooms + 139.87 LivingArea$

b) According to the multiple regression model, the asking *Price* increases, on average, by about $139.87 for each additional square foot, for homes with the same number of bathrooms.

c) The number of bathrooms is probably correlated with the size of the house, even after considering the square footage of the bathroom itself. This correlation may account for the coefficient of *Bathrooms* not being discernibly different from 0. Moreover, the regression model does not predict what will happen when a house is modified, for example, by converting existing space into a bathroom.

18. More hill races.

a) The two models are similar. In both models, additional *Distance* and additional *Climb* are associated with increases in average race men's record *Time*.

b) The residuals appear to fan out as the predicted value increases. This violates the "does the plot thicken?" condition. There may be two high outliers as well, but re-expressing *Time* may improve both problems.

19. Predicting finals II.

Straight enough condition: The plot of residuals versus fitted values looks curved, rising in the middle, and falling on both ends. This is a potential difficulty.
Randomization condition: It is reasonable to think of this class as a representative sample of all classes.
Nearly Normal condition: The Normal probability plot and the histogram of residuals suggest that the highest five residuals are extraordinarily high.
Does the plot thicken? condition: The spread is not consistent over the range of predicted values.
These data may benefit from a re-expression.

20. Home prices II.

Straight enough condition: The plot of residuals versus predicted values looks curved. Multiple regression is not appropriate.
Randomization condition: A random sample of homes for sale in the area was chosen from the Internet.
Nearly Normal condition: The Normal probability plot is not straight, and the histogram of residuals is skewed to the right. Multiple regression is not appropriate.
Does the plot thicken? condition: There may be a slight increase in variability as the fitted values increase.

21. Admin performance.

a) $\widehat{Salary} = 9.788 + 0.11Service + 0.053Education + 0.071Score + 0.004Speed + 0.065Dictation$

b) $\widehat{Salary} = 9.788 + 0.11(120) + 0.053(9) + 0.071(50) + 0.004(60) + 0.065(30) \approx \$29,200$

c) Although *Age* and *Salary* are positively correlated, after removing the effects of years of education and months of service from *Age*, what is left is years not spent in education or service. Those non-productive years may well have a negative effect on salary.

22. GPA and SATs.

a) $\widehat{GPA} = 0.574968 + 0.001394(SATV) + 0.001978(SATM)$

b) $\widehat{GPA} = 0.574968 + 0.001394(500) + 0.001978(550) \approx 2.3599$

c) **Straight enough condition:** The scatterplots of the response versus each predicted value should be reasonably straight.
Randomization condition: Hopefully, this section of statistics is representative of all sections, and can be considered a random sample. Otherwise, the regression residuals plots should show no pattern.
Nearly Normal condition: The Normal probability plot should be straight, and the histogram of residuals should be unimodal and symmetric.
Does the plot thicken? condition: The spread of the residual plots should be constant.

23. Body fat revisited.

a) According to the linear model, each pound of *Weight* is associated with a 0.189 increase in *%Body Fat*.

b) After removing the linear effects of *Waist* and *Height*, each pound of *Weight* is associated, on average, with a decrease of 0.10% in *%Body Fat*. The change in coefficient and sign is a result of including the other predictors. We expect *Weight* to be correlated with both *Waist* and *Height*. It may be collinear with them.

c) We should examine the partial regression plot for *Weight*. It would have an *x*-axis that is *Weight* with the effects of *Waist* and *Height* removed, so we can understand that better.

24. Cereals.

a) $\widehat{Calories} = -0.879994 + 3.60495 Protein + 8.56877 Fat + 0.309180 Fiber + 4.13996 Carbo + 4.00677 Sugars$

b) We should examine a scatterplot of residuals vs. predicted values and a Normal probability plot of the residuals.

c) After allowing for the linear effects of *protein, fiber, carbo,* and *sugars*, the multiple regression model predicts that each gram of *fat* is, on average, associated with an increase of 8.56877 *calories.*

25. Cereals again.

a) $\widehat{Calories} = 83.0469 + 0.05721 Sodium - 0.01933 Potassium + 2.38757 Sugars$

b) We should examine a scatterplot of residuals vs. predicted values and a Normal probability plot of the residuals.

c) No, adding *Potassium* wouldn't necessarily lower the *Calories* in a breakfast cereal. A regression model doesn't predict what would happen if we change a value of a predictor. It only models the data as they are.

26. Grades.

Answers may vary. The regression of *Final* on *Midterm 1* and *Midterm 2* has a strong R^2 and significant *t*-statistics on both coefficients. The conditions appear to be satisfied. *Homework* is another promising predictor in some regression models, but it is skewed and has a low outlier.

27. Hand dexterity.

a) For children of the same *Age*, their *Dominant Hand* was faster on this task by about 0.304 seconds on average.

b) Yes. The relationship between *Speed* and *Age* is straight and the lines for dominant and non-dominant hands are very close to parallel.

28. Candy bars with nuts.

a) Among candy bars with the same *Protein* and *Sugar* content per serving, those with nuts tend to have 0.1556 calories fewer per serving.

b) Adding the predictor *Nuts* did not improve the regression. The predicted change in calories is very small and R^2 is only slightly larger.

29. Scottish hill races, men and women.

a) The change in *Time* with *Distance* is different for men and women, so a different slope is needed in the model.

b) After accounting for *Distance*, the increase in race time with distance is greater for men than for women. Said another way, women's times increase less rapidly for longer races than do men's.

30. Scottish hill races, men and women climbing.

a) The change in *Time* with *Climb* is different for men and women, so a different slope is needed in the model.

b) After accounting for the *Climb*, the increase in race time is greater, on average, for men than for women.