Chapter 24 – Multiple Regression Wisdom

Section 24.1

1. **Combining and separating.**

   a) This is filtering, you are focusing on the subset of items containing meat.

   b) This is merging, you are combining data together.

2. **New cases and variables.**

   a) This is appending, you are joining additional cases to the data.

   b) This is feature extraction, you are constructing a new ratio variable from *fat* and *carbohydrates*.

Section 24.2

3. **Residual, leverage, influence.**

   a) likely high leverage

   b) likely outlier

4. **Residual, leverage, influence, 2.**

   a) Likely influential. It has high leverage for the unusual combination of SAT scores and an unusually high GPA**.**

   b) Likely large (negative) residual. We aren't told that the country is otherwise unusual.

Section 24.3

5. **Significant coefficient?**

   No. The predictors are almost certainly collinear, which would affect the coefficient of *Age*.

6. **Better model?**

   The smaller model is likely better. The boss could check the adjusted R^2 statistic, which is likely to be larger for the smaller regression.

Chapter Exercises.

7. **Climate change 2016.**

   a) CO_2 and *Year* are clearly collinear. That will make the *t*-ratios small even though the R^2 is large. It would not be valid to conclude that CO_2 has no effect on *Avg Global temp*.

   b) The data were filtered to select these years. Without a good reason for selecting these particular years, we might be suspicious of the analysis.

8. **Climate change 2016 again.**

   a) Although the coefficients are now statistically significant, a regression on observed data like these can't prove a causal relationship between CO_2 and Global temperature.

   b) If the pattern of the relationships among these variables changed at some point, then it might be more appropriate to filter the data to the recent years for which the relationship was consistent. (Answers may vary.)

9. **Healthy breakfast, sick data.**

   a) The slope of a partial regression plot is the coefficient of the corresponding predictor, in this case, –1.020.

   b) Quaker oatmeal makes the slope more strongly negative. It appears to have substantial influence on this slope.

9. (continued)

c) Not surprisingly, omitting Quaker oatmeal changes the coefficient of *fiber*. It is now positive (although not significantly different from 0). This second regression model has a higher R^2, suggesting that it fits the data better. Without the influential point, the second regression is probably the better model.

d) The coefficient of *Fiber* is not discernibly different from zero. We have no evidence that it contributes significantly to *Calories* after allowing for the other predictors in the model.

10. Fifty states.

a) Yes, they should be influential. Points that have both large leverage and large Studentized residuals are bound to be influential.

b) The *t*-ratios for indicator variables are *t*-tests of whether those cases fit the regression model established by the other cases. Both of the indicators have *t*-ratios that are large enough to be significant at the 0.05 level.

c) The *t*-ratios for *illiteracy* and *income* are not very large. The coefficient for *income* is near zero. Either predictor might be considered for removal from the regression model.

11. Cereals.

a) After allowing for the effects of *sodium* and *sugars*, the model predicts a decrease of 0.019 calories for each additional gram of *potassium*.

b) Those points pull the slope of the relationship down. Omitting them should increase the value of the coefficient of *potassium*. It would likely become positive, since the remaining points show a positive slope in the partial regression plot.

c) These appear to be influential points. They have both high leverage and large residuals, and the partial regression plot shows their influence.

d) If our goal is to understand the relationships among these variables, then it might be best to omit these cereals because they seem to behave as if they are from a separate subgroup.

12. Scottish hill races 2008.

a) The scatterplot of the residuals against the predicted values shows a fan shape with the spread increasing for longer races, and two races have large residuals. The Lairig Ghru race has high leverage.

b) Both races may be outliers. Their larger residuals have inflated the residual standard deviation and may have reduced the value of R^2, but we can't tell if they have affected the coefficients without examining partial regression plots.

c) The partial regression plots show that these races have had little effect on the coefficients other than on the intercept, which may have been increased. A scatterplot of *Distance* vs. *Climb* shows that the Lairig Ghru race is unusually long and has very little climb (for these races), accounting for its large leverage.

13. World data.

a) The distributions of *Population*, *Birthrate*, *GDP/capita*, *Imports/Capita*, *Exports/capita*, and possibly *Airports/capita* are skewed to the right and might benefit from re-expression. The distributions of *Birthrate* and *Obesity%* may be bimodal. (Answers may vary.)

b) The re-expressions have improved the symmetry of variables. After re-expression, *Obesity%* seems more clearly bimodal. (Answers may vary.)

14. Pizza.

a) According to the multiple regression model, the *pizza score* is higher by 15.6 points for cheese pizza than for pepperoni, after allowing for the effects of *Calories* and *Fat*.

b) We should plot the residuals against predicted values or against each predictor, looking for pattern or outliers. We should check for evidence that the residuals are nearly Normal with a Normal probability plot or histogram.

15. World data correlated.

a) Most show reasonably straight relationships. *LogPopulation* and *Education* seem to be only weakly associated with other variables. (Answers may vary.)

b) The variables most closely (linearly) associated with *Life Expectancy* are *LogBirthrate*, *LogGDP/capita*, *LogImports/capita*, and √*Airports/capita.* The scatterplots in the plot matrix suggest that any of these is linearly associated with *Life Expectancy*, but several are also linearly associated with each other, so we must beware of collinearity if we choose several of them as predictors.

16. Gourmet pizza.

a) Reggio's and Michelina's received lower scores than we would otherwise have expected from the model.

b) The *t*-ratio for the indicator variable for *Michelina's* is –4.03, which is large. We can reject the null hypothesis that Michelina's fits the regression model, and we can conclude that it is an outlier.

17. World data regression.

The regression seems like a good model. Seventy-three point two percent of the variation in *Life Expectancy* is accounted for by the model and all *t*-ratios are significant. The model says that, after allowing for *LogGDP/capita*, *LogBirthrate* is negatively associated with life expectancy—higher birthrates go with lower life expectancy. After allowing for *LogBirthrate*, higher GPD/capita is associated with a higher life expectancy. However, the residuals show 5 low outliers that deserve attention and explanation.

18. Another slice of pizza.

a) Cheese and pepperoni pizzas do not appear to be described by the same model.

b) The slope of taste *score* on *calories*, after allowing for the linear effects of fat and removing the influence of the two outlying pizzas, is estimated to be 1.92 points per gram for pepperoni pizzas and 1.92 – 0.4615 = 1.45 points per gram for cheese pizzas.

c) This should be a better regression model. We have identified a consistent difference between pepperoni and cheese pizzas and incorporated it into the model. All coefficients are significantly different from zero, and both the R^2 and adjusted R^2 are higher than the model in Exercise 16.

19. World data diagnosis

All five outlying countries are in southern Africa. We would like to find something they have in common that could depress life expectancy, after allowing for GDP/capita and birthrate.

20. The final slice.

a) The coefficient for the indicator for Weight Watchers is not significantly discernible from zero at the 0.05 level, but with a P-value of 0.09, it may still improve the model. This model has a slightly higher R^2 and adjusted R^2, but it is not enough improved to be grounds for choosing between the models. But the *t*-ratios are larger and P-values smaller for a number of the coefficients. That is a sign of improvement.

b) Looking at the other coefficients in the model (and especially at the coefficient for *Calories*—not too surprising, considering the identity of the newly isolated pizza), the addition of an indicator for the Weight Watcher's pizza has made several of them more significantly different from zero. This seems to be a cleaner model and one that might lead to a better understanding.

c) The tasters score cheese pizzas substantially higher than pepperoni pizzas. Even after allowing for that, additional fat lowers scores, but higher calories pizzas score a bit better.

21. World data with HIV.

This model has a higher R^2 and higher adjusted R^2. The coefficients of the previously used predictors have not changed very much and both are still statistically significant. The coefficient of *LogHIV* says that after allowing for the effects of the other predictors, higher HIV incidence is associated with lower life expectancy—as we would expect.