Homework 9: Sections 9.1 - 10.3

STA209-04: Applied Statistics

Assigned: 04/26/2019
Due: 05/03/2019

From the Book:

Questions: 9.21, 9.27, 9.55, 10.27, 10.51, 10.63

9.21 The FloridaLakes dataset, introduced in Data 2.4, includes data on 53 lakes in Florida. Two of the variables recorded are *pH* (acidity of the lake water) and *AvgMercury* (average mercury level for a sample of fish from each lake). We wish to use the pH of the lake water (which is easy to measure) to predict average mercury levels in fish, which is harder to measure. A scatterplot of the data is shown in Figure 2.49(a) on page 109 and we see that the conditions for fitting a linear model are reasonable met. Computer output for the regression analysis is shown below.

The regression equation is AvgMercury =1.53 - 0.152 pH

Predictor	Coef	SE Coef	T	P
Constant	1.5309	0.2035	7.52	0.000
pH	-0.15230	0.03031	-5.02	0.000

S = 0.281645 R-Sq = 33.1% R-Sq(adj) = 31.8%

a) Use the fitted model to predict the average mercury level in fish for a lake with a pH of 6.0.

b) What is the slope in the model? Interpret the slope in context.

c) What is the test statistic for a test of the slope? What is the p-value? What is the conclusion of the test, in context?

d) Compute and interpret a 95% confidence interval for the slope.

e) What is R^2? Interpret it in context.

9.27 A random sample of 50 countries is stored in the dataset SampCountries. Two variables in teh dataset are life expectancy (*LifeExpectancy*) and percentage of government expenditure spent on health care (*Health*) for each country. We are interested in whether or not the percent spent on health care can be used to effectively predict life expectancy.

a) What are the cases in this model?

b) Create a scatterplot with a regression line and use it to determine whether we should have any serious concerns about the conditions being met for using a linear model with these data.

c) Run the simple linear regression, and report and interpret the slope.

d) Find and interpret a 95% confidence interval for the slope.

e) Is the percentage of government expenditure on health care a significant predictor of life expectancy?

f) The population slope (for all countries) is 0.467. Is this captured in your 95% CI from part d)?

g) Find and interpret R^2 for this linear model.

9.55 The dataset HomesForSaleCA contains a random sample of 30 houses for sale in California. We are interested in whether we can use number of bathrooms *Baths* to predict number of bedrooms *Beds* in houses in California. Use technology to answer the following questions:

a) What is the fitted regression equation? Use the regression equation to predict the number of bedrooms in a house with three bathrooms.

b) Give the t-statistic and the p-value for the t-test for slope in the regression equation. State the conclusion of the test.

c) Give the F-statistics and the p-value from an ANOVA for regression for this model. State the conclusion of the test.

d) Give and interpret R^2 for this model.

10.27 Categorical variables with only two categories (such as male/female or yes/no) can be used in a multiple regression model if we code the answers with numbers. Exercise 2.143 on page 102 introduces a study examining years playing football, brain size, and percentile score on a cognitive skills test. We show computer output below for a model to predict *Cognition* score based on *Years* playing football and a categorical variable *Concussion*. The variable *Concussion* is coded 1 if the player has ever been diagnosed with a concussion and is coded 0 if he has not been diagnosed with a concussion.

Regression Equation
Cognition = 100.6 - 3.07 Years - 2.70 Concussion

Coefficients

Term	Coef	SE Coef	T	P
Constant	100.6	16.9	5.97	0.000
Years	-3.07	1.62	-1.90	0.064
Concussion	-2.70	9.49	-0.29	0.777

S = 25.7829 R-Sq = 13.56% R-Sq(adj) = 9.35%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	2	4277.3	2138.63	3.22	0.050
Residual Error	41	27255.0	664.76		
Total	43	31532.2			

a) One of the participants in the study played football for 9 years, had never been diagnosed with a concussion, and scored a 74 on the cognitive skills test. What is his predicted cognition score? What is the residual for this prediction?

b) Another one of the participants played football for 7 years, had been diagnosed with a concussion, and scored a 42 on the cognitive skills test. What is his predicted cognition score? What is the residual for this prediction?

c) What is the coefficient of *Years* in this model? Interpret it in context.

d) What is the coefficient of *Concussion* in this model? Interpret it in context. (Pay attention to how the variable is coded.)

e) At a 10% level, is the overall model effective at predicting cognition scores? What value in the computer output are you basing your answer on?

f) There are two variables in this model. How many of them are significant at the 10% level? How many are significant at the 5% level?

g) Which of the two variables is most significant in this model?

h) How many football players were included in the analysis?

i) What is R^2 Interpret it in context.

10.51 The data in CommuteAtlanta show information on both the commute distance (in miles) and time (in minutes) for a sample of 500 Atlanta commuters. Suppose that we want to build a model for predicting the commute time based on the distance.

a) Fit the simple linear model, $Time = \beta_0 + \beta_1 Distance + e$, for the sample of Atlanta commuters and write down the prediction equation.

b) What time (in minutes) does the fitted model predict for a 20-mile commute?

c) Produce a scatterplot of the relationship between *Time* and *Distance* and comment on any interesting patterns in the plot.

d) Produce a dotplot or histogram to show the distribution of residuals for this model. Comment on whether the normality condition is reasonable.

e) Produce a plot of the residuals vs the fitted values. Comment on what this plot says about the simple linear model conditions in this situation.

10.63 Baseball is played at a fairly leisurely pace - in fact, sometimes too slow for some sports fans. What contributes to the length of a major league baseball game? The file BaseballTimes contains information from a sample of 30 games to help build a model for the time of a game (in minutes). Potential predictors include:

Runs	Total runs scored by both teams
Margin	Difference between the winner's and loser's scores
Hits	Total base hits for both teams
Errors	Total number of errors charged to both teams
Pitchers	Total number of pitchers used by both teams
Walks	Total number of walks issued by pitchers from both teams

a) Use technology to find the correlation between each of the predictors and the response variable *Time*. Identify predictors that appear to be potentially useful based on these correlations.

b) Try different models and combinations of predictors to help explain the game times. Try to get a good R^2 and a good ANOVA p-value, but also have significant predictors. Decide on a final model and briefly indicate why you chose it.