Homework 9: Sections 9.1 - 10.3
KEY

STA209-04: Applied Statistics

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

Total Possible Points: 32

From the Book:

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

9.21 [5 pts] 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.

AvgMercury $= 1.53 - 0.152 * (6) = 0.618$. Our model's prediction for the average mercury level in fish for a lake with a pH of 6.0 is 0.618.

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

The sloe is -0.152. This means that for every unit increase in pH of the lake, we expect a 0.152 unit decrease in the average mercury level of fish residing in said lake.

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?

The value of the test statistic for the slope is -5.02, and the corresponding p-value is 0.000/ From this, we conclude that there is an association (i.e. non-zero regression slope) between lake pH and average mercury level in fish.

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

The 95% confidence interval for the slope is:

$$\hat{\beta} \pm t_{crit,df=53-2} SE = -0.152 \pm 2 * 0.03 = (-0.213, -0.091).$$

We are 95% confident that the average mercury level in fish decreases between 0.213 and 0.091 units with each unit increase in lake pH.

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

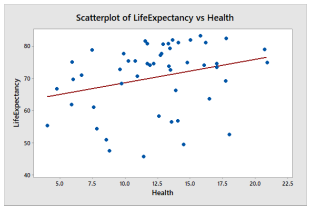
The R^2 value is 0.331. Our model using lake pH as a predictor explains 33.1% of the total variability observed in average mercury levels in fish. In other words, the change in lake pH helps us explain 33.1% of the observed changes in average mercury level in fish.

9.27 [7 pts] 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?

The countries.

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.



There does not appear to be any cause from concern based on the presented scatterplot.

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

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	61.32	4.70	13.04	0.000	
Health	0.729	0.364	2.00	0.051	1.00

The slope is 0.729. The life expectancy of a country is estimated to increase by 0.729 units for each percent increase in government health expenditure.

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

The 95% confidence interval for the slope is:

$$\hat{\beta} \pm t_{crit,df=50-2} SE = 0.729 \pm 2 * 0.364 = (-0.003, 1.461).$$

We are 95% confident that the average life expectancy changes by as little as a 0.003 unit decrease and a 1.461 unit increase with each percentage point increase in government health expenditure.

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

It is a marginally significant predictor, just barely outside the classic 0.05 threshold for statistical significance.

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

Yes.

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

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
10.1534	7.71%	5.79%	0.84%

The R^2 value is 0.0771. Our model using percent of government expenditure on health as a predictor explains 7.71% of the total variability observed in life expectancy.

9.55 [4 pts] 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.

Regression Equation

Beds = 1.367 + 0.746 Baths

The predicted number of bedrooms in a three bathroom house is:

$$1.367 + 0.746 * 3 = 3.605 \approx 4$$

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.

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	1.367	0.319	4.29	0.000	
Baths	0.746	0.117	6.39	0.000	1.00

Given that the test for the slope (i.e. Bath above) yielded a p-value of 0.000, we conclude that the slope of the regression of number of bedrooms on the number of bathrooms is non-zero.

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

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	1	19.782	19.7817	40.77	0.000
Baths	1	19.782	19.7817	40.77	0.000
Error	28	13.585	0.4852		
Lack-of-Fit	6	5.726	0.9543	2.67	0.042
Pure Error	22	7.859	0.3572		
Total	29	33.367			

As shown above, the F-statistic of 40.77 yields a p-value of 0.000. From this we would conclude that the additional variability explained by including the number of bathrooms in our model is greater than that expected by random chance. In other words, the regression model containing the number of baths as a predictor explains a greater proportion of the variability in our outcome than does the null, intercept-only model.

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

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
0.696547	59.29%	57.83%	51.21%

The R^2 value above indicates that inclusion of the number of bathrooms as a explanatory model results in a model which explains 59.29% of the observed variability in our outcome.

10.27 [9 pts] 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?

Given this information, the predicted cognition score is:

$$100.6 - 3.07 * 9 - 2.70 * 0 = 72.97.$$

The residual for this prediction is $74 - 72.97 = 1.03$.

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?

Given this information, the predicted cognition score is:

$$100.6 - 3.07 * 7 - 2.70 * 1 = 76.41.$$

The residual for this prediction is $42 - 76.41 = -34.41$.

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

The coefficient of *Years* is -3.07. For each additional year of playing football, assuming the concussion history is fixed, the cognition score is expected to decrease by 3.07 points.

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

The coefficient of *Concussion* is -2.70. Assuming the number of years played football is fixed, a decrease of 2.70 points in cognition score is expected for individuals who have been diagnosed with a concussion at least once before.

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?

Yes, the model is effective at predicting cognition scores (when using a 10% significance level). The p-value from the ANOVA table is what we use to make this determination.

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?

One is significant at the 10% level (*Years*), and none are significant at the 5% level.

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

Years is the most significant.

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

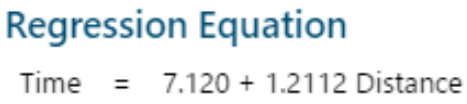
44 football players were included in the analysis. This was determined using the degrees of freedom for the Total Sum of Squares in the ANOVA table.

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

R^2 is 0.1356, or 13.56%. In modeling cognition score using the number of years played football and whether a player has ever been diagnosed with a concussion, we are able to explain 13.56% of the observed variability in cognition score.

10.51 [5 pts] 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.

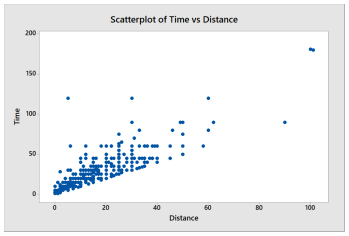


b) What time (in minutes) does the fitted model predict for a 20-mile commute?

The predicted time for a 20-mile commute is:

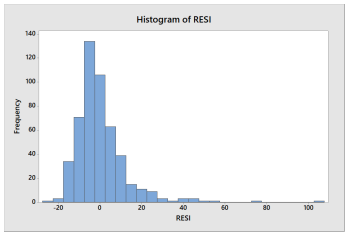
$$7.120 + 1.2112 * 20 = 31.344 \text{ minutes}$$

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



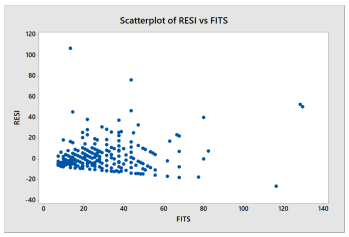
The provided scatterplot demonstrates a clear linear relationship between commute distance and time. However for larger distances, the amount of time taken seems to vary more than for shorter distances. This is evident by the more pronounced vertical spread in the data for larger values of distances and less pronounced spread for smaller values.

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



Based on the provided histogram, the normality condition would be unreasonable. The residual distribution is fairly right-skewed, which is not what would be expected if they were truly normally distributed.

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.



This scatterplot of the residuals vs fitted values confirms what was suspected from the scatterplot in c). There is a pattern in our residuals, which would indicate that they are not identically distributed. The specific pattern seen here is that the residuals increase in magnitude with increased value of commute distance. With this and previous plots, we can conclude that the simple linear model conditions are not met in this situation.

10.63 [2 pts] 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.

Correlations

	Time	Runs	Margin	Hits	Errors	Pitchers
Runs	0.504					
Margin	-0.116	0.192				
Hits	0.349	0.809	0.206			
Errors	-0.040	0.112	0.361	0.296		
Pitchers	0.721	0.496	-0.253	0.490	-0.185	
Walks	0.565	0.339	-0.177	0.098	-0.060	0.410

Cell Contents
Pearson correlation

Based on the above correlation matrix, predictors that may be useful include: *Ptchers*, *Walks*, *Runs*, and *Hits*. These each have correlations with our outcome, *Time*, that are at least 0.3 in magnitude.

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.

Various answers are possible here. The reasoning for choosing a model should mention considerations of model parsimony, adjusted R^2, and/or p-values. If students performed an exhaustive search of all models via best subsets selection, then the final model should be the one which uses *Walks* and *Pitchers* as the only two covariates. This model has the highest adjusted R^2 and has relatively few covariates (i.e. is a parsimonious model).

Best Subsets Regression: Time versus Runs, Margin, Hits, ... ers, Walks

Response is Time

Vars	R-Sq	R-Sq (adj)	R-Sq (pred)	Mallows Cp	S	Runs	Margin	Hits	Errors	Pitchers	Walks
1	52.0	50.2	43.7	5.2	14.485					X	
1	31.9	29.5	21.9	18.3	17.243						X
2	60.7	57.8	52.5	1.6	13.347					X	X
2	54.8	51.4	43.9	5.4	14.309	X				X	
3	62.1	57.7	49.6	2.6	13.352	X				X	X
3	61.5	57.1	48.3	3.0	13.451		X			X	X
4	63.1	57.2	47.4	4.0	13.441	X		X		X	X
4	62.5	56.5	47.5	4.4	13.545	X			X	X	X
5	64.5	57.1	46.1	5.1	13.446	X		X	X	X	X
5	63.5	55.8	42.4	5.8	13.646	X	X	X		X	X
6	64.6	55.4	41.3	7.0	13.713	X	X	X	X	X	X