

Project 11 - Simple Linear Regression

Statistics 401: Fall 2006
Due Wednesday, December 6

Use R to complete this project. Attach all R commands used to complete this project in an appendix. Annotate with the problem number.

1. Do problem 5.4 on page 194 of your textbook.
2. Do problem 5.18 on page 206.

The rest of the questions in this project refer to the following:

A large university is interested in determining the relationship between the number of hours that its students spend each week studying, and three different explanatory variables: the number of hours spent watching TV; fastfood consumption; and coffee consumption. A simple random sample of 28 students were asked the following four questions:

- About how many hours of TV do you watch each week?
- About how many fast food meals do you eat each week?
- Do you drink coffee? If so, about how many cups each week?
- About how many hours do you study for academics each week?

A file containing the data can be found on the STAT 401 website.

3. Display a scatterplot of the number of hours spent studying each week (y) versus coffee consumption (x).
4. Calculate the sample correlation coefficient r . What parameter does r estimate? You may use R for the calculation.
5. Give the form, direction and strength of the relationship between x and y . For each, indicate what output you are using.
6. Give the SLR model which describes the number of hours spent studying each week as a linear function of coffee consumption. (The model consists of the parameters β_0 , β_1 and σ , NOT the estimates for these parameters!)
7. Fit a SLR model to describe the number of hours spent studying each week as a function of coffee consumption. In your report, give the least-squares regression line.
8. Use the output from R's `lm()` and `summary()` functions to fill in the following table. Label it and reference it from the text of your report.

	Estimate	SE	t	p-value
(Intercept)				
coffee				

9. Use the output from R's `lm()` and `anova()` functions to fill in the following ANOVA table. Label it and reference it from the text of your report.

Source	DF	SS	MS	F	p -value
Model					
Residual					
Total					

10. Give an unbiased estimate for σ^2 and interpret this value in terms of the problem.
11. Give the value of r^2 and interpret it in the context of this problem.
12. Perform each step below to determine if there is a significant linear relationship between the number of hours spent studying each week and coffee consumption.
 - (a) State the hypotheses.
 - (b) Check the Assumptions. Even if you have concern with one of the assumptions, continue to answer the rest of the questions anyways.
 - i. Display a plot of the residuals versus fitted values. Which assumptions can you use this plot to check? Are there any serious problems? Comment on each assumption separately.
 - ii. Display a normal probability plot of the residuals. Is there strong evidence against the assumption that the residuals are normally distributed? Explain.
 - (c) Give the t and F test statistic values. What is the relationship between t and F ?
 - (d) Give the distribution of each test statistic when H_0 is true, and give the p -value.
 - (e) Make a decision regarding H_0 at $\alpha = .05$.
 - (f) Give a conclusion in terms of the problem.
13. Calculate a 95% confidence interval for the slope of the population regression line. Interpret the confidence interval in the context of this problem.
14. Among all students who drink 6 cups of coffee a week, predict how many hours on average that they study.
15. Use R to calculate a 95% CI for the mean number of hours studied by students who drink 6 cups of coffee a week. Interpret this CI in terms of the problem.
16. For a student who drinks 6 cups of coffee a week, predict how many hours that person studies.
17. Use R to calculate a 95% PI for the number of hours studied by a student who drinks 6 cups of coffee a week. Interpret this PI in terms of the problem.
18. Fit the other two SLR models - (1) hours studied versus TV and (2) hours studied versus fastfood.
 - (a) For each of these two models, comment on whether the predictor is significant (at $\alpha = .05$) for predicting number of hours studied. Indicate what output justifies your answer.
 - (b) Give the r^2 values for each of these two SLR's.
 - (c) Based on your answers to (a) and (b) above, which of TV, fastfood or coffee is the best single predictor of number of hours studied? Give at least two reasons why your answer is correct.