Prof. Yiping Lu
IEMS 304: Statistical Learning for Data Analysis
April 26, 2025

Homework 3

This homework is to give a brief reminder of R, RStudio, and statistical topics covered in IEMS 303.

Note: The homework is scored out of 100 points. The problems add up to 90 points, while the remaining ten points will be graded according to a writing rubric, given at the end of the assignment.

R/RStudio installation If you have not installed R and RStudio, follow the installation instructions outlined in https://posit.co/download/rstudio-desktop/. You are strongly encouraged to use R Markdown to integrate text, code, images and mathematics or you can you use the latex code we provide.

Question 1. GPA Data The `HW2_data.xls` worksheet HW2P2 shows the GPA of college students, along with their college entrance verbal and mathematical test scores. The objective is to predict the GPA (`y`) of matriculating freshmen based on their verbal (`x1`) and mathematical (`x2`) test scores.

a) For fitting a multiple linear regression model $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$ to the data `HW2_data.xls`. write down response vector Y, data matrix X. In addition, use R to do the matrix calculations and report $X^\top X$,$X^\top Y$, $(X^\top X)^{-1}X^\top Y$ and the estimated coefficients $\hat{\beta}$

(b) Fitting a multiple linear regression model $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$. Provide an interpretation of the coefficients for the verbal and mathematics test scores.

c) In the `summary()` command applied to the linear model object fitted in part (a), a P-value is reported for each of the coefficients. State the null and alternative hypotheses that are associated with each of these P-values. Explain what conclusions can be drawn for each of these P-values. Your explanation should be in terms of the variables (GPA, verbal score, and math score), and not in terms of the parameters.

d) For a matriculating freshman having a verbal score of 75 and a mathematical score of 90, what is their predicted college GPA?

e) Caculate 95% confidence intervals for the coefficients for the verbal and mathematics test scores. Provide an interpretation of these confidence intervals. Also, write out the equations that were used to calculate the confidence intervals. For the student in part (d), calculate and interpret a 95% prediction interval on the response. Calculate and interpret a 95% confidence interval on the response mean.

Quadratic model for GPA. We fit a linear model. Consider the same GPA data but for a quadratic model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1^2 + \beta_4 x_2^2 + \beta_5 x_1 x_2 + \epsilon$$

a) Fit the above quadratic model and display the results using the `summary()` command. Do the P -values produced by the summary command provide a reliable indication of whether the additional terms $\beta_3 x_1^2$, $\beta_4 x_2^2$, and $\beta_5 x_1 x_2$ should be included in the model?

b) Use a partial sum of squares F-test to test whether the quadratic model offers an improvement over the linear model. Formally state what hypotheses you are testing. Do your conclusions from the F -test agree with your conclusions from part (a)?

Rubric (10)

- The text is laid out cleanly, with clear divisions between problems and sub-problems. The writing itself is well-organized, free of grammatical and other mechanical errors, and easy to follow.
- Questions which ask for a plot or table are answered with both the figure itself and the command (or commands) use to make the plot. Plots are carefully labeled, with informative and legible titles, axis labels.
- All quantitative and mathematical claims are supported by appropriate derivations, included in the text, or calculations in code. Numerical results are reported to appropriate precision.
- Code is either properly integrated with a tool like R Markdown or included as a separate R file. In the former case, both the knitted and the source file are included. In the latter case, the code is clearly divided into sections referring to particular problems. In either case, the code is indented, commented, and uses meaningful names.
- All parts of all problems are answered with actual coherent sentences, and never with raw computer code or its output. For full credit, all code runs, and the Markdown file knits (if applicable).