Prof. Yiping Lu
IEMS 304: Statistical Learning for Data Analysis
May 10, 2025

Homework 4

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. The data in the HW4P1 of `HW4_data.xls` are the GPA data that you analyzed in HW2.

(a) Plot the residuals from the fitted linear model $Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$ versus x_1 and x_2. Do these plots suggest the need to fit a quadratic model $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1^2 + \beta_4 x_2^2 + \beta_5 x_1 x_2 + \varepsilon$?

(b) Plot the residuals from the quadratic fit versus x_1 and x_2. Plot the standardized residuals versus $\hat{y}$. Suggest a suitable transformation to remove the non-constant variance exhibited in the plot of residuals versus $\hat{y}$.

Question 2. The data in HW4P2 of `HW4_Data.xls` are the following variables for 19 stocks: profit margin (x_1), growth rate (x_2), type of industry (x_3), and price to earnings (P/E) ratio (y). The type of industry is a nominal categorical variable coded as 1 = oil, 2 = drug/health, and 3 = computers/electronics.

(a) Explain why the above coding of the type of industry is not appropriate. Suppose your software could not automatically handle categorical/factor predictors (the lm() function and most R functions can). Explain how you would recode the data by defining indicator variables for the type of industry, using the computers/electronic industry as a baseline.

(b) Fit a linear model to predict the P/E ratio as a function of the profit margin, growth rate, and type of industry, the latter as a categorical factor. Which predictors appear to have a statistically significant effect on the P/E ratio? Interpret the coefficients of the dummy variables.

R Hints: Suppose the data frame is called STOCK. Then (don't copy paste the following latex rendering)
STOCK$Industry←as.factor(STOCK$Industry)
converts the numeric Industry variable to a factor, and
levels(STOCK$ Industry)←c("OIL", "DRUG", "COMP")
gives the correct sector names to the three factor levels. In the preceding, "OIL" is the first level of the Industry factor, because it corresponds to the first level 1 before renaming the factor levels. By default, when there is a predictor variable that is a factor, R's lm*() function treats the first level of the factor ("OIL" in this case) as the base category. You can change the first level of the factor to "COMP" using the* relevel() *function. Type* ?relevel *to see how to use this function.*

(c) Suppose we choose the drug/health care industry as the baseline. How will the coefficients in the new equation be related to those fitted in (b)? You can determine this entirely from the results in (b), without even refitting the model. Also refit the model with drug/health as the baseline to confirm.

(d) Conduct a partial F-test to determine whether there is any interaction between the type of industry and the growth rate.

(e) Referring to part (d), suppose there were a statistically significant interaction between growth rate and the type of industry. Provide an interpretation of the interaction.

Question 3. The `HW4P3` worksheet in `HW4_data.xls` contains average times (in seconds or minutes, depending on the event) for eight different track events for 55 different countries. Treat this as 55 observations of 8 different variables. View the marathon times as the response variable and the times for the seven other events as 7 predictor variables. For all parts of this problem, work with the standardized predictors and standardized response (i.e., use the scale command to standardize the variables before fitting a regression model to them).

(a) Conduct forward stepwise regression using R's step command. Discuss what criterion R uses to decide whether to add/remove a predictor at each iteration. Write out the model recommended by stepwise regression.

(b) Conduct best subsets regression using R's leaps command in the leaps package. Use C_p as the criterion for selecting the best model (which is the same as one of the versions if AIC), and write out the form of the model with the lowest C_p?