Linear Regression
HW 4
Due 9/21 at 11:59pm

Directions: Submit a .pdf file containing your responses. The .pdf can be converted from a Latex file, pictures of your handwritten solutions, word files, markdown files, etc. If there are coding problems, upload a separate notebook for Python code.

Written Questions

1. Suppose we have fit a MLR model between response variable Y and predictors $X_1, ..., X_{p-1}$. using a data of size n. The global F-test aka omnibus test considers the hypotheses:

$$H_0: \; y_i = \beta_0 + \epsilon_i \quad \text{vs.} \quad H_1: y_i = \beta_0 + \beta_1 X_{1i} + ... + \beta_{p-1} X_{(p-1)i} + \epsilon_i$$

   using the statistic $F = \dfrac{MSR}{MSE}$ where MSR and MSE are calculated under the full model. Show that this definition is equivalent to the alternative formulation of the F statistic as:

$$F_{alt} = \frac{\dfrac{SSE_{H_0} - SSE_{H_1}}{df_{SSE_{H_0}} - df_{SSE_{H_1}}}}{\dfrac{SSE_{H_1}}{df_{SSE_{H_1}}}}.$$

2. If a predictor variable is categorical with six states and we want to include it in a regression model, how many dummy variables do we need to use?

3. Suppose a predictor variable is categorical with three states "C1", "C2", "C3". When we include it in a regression model and the individual t tests used "C1" as reference level, and showed "C2" is significant and "C3" is not. Would you conclude that "we should drop C3 and fit a new model"? Why or why not?

   Coding Questions

4. This question will help you to understand the calculation of ANOVA in MLR using an example. For the dataset KelleyBlueBookData.csv, response= Price against the following predictors: Mileage, Liter, Cylinder (in this order). Treat Cylinder as a quantitative variable.

   (a) Run the sequential ANOVA for the fitted model. Report null and alternative hypothesis, the F stat, and the p-value for the F-test for dropping or including the 'Cylinder' predictor. What is the conclusion of this test?

   (b) Manually run the test in part (a) yourself: 1. fit the null model in python and extract SSE and degrees of freedom of this SSE; then 2. fit the alternative model in python and extract SSE and degrees of freedom of this SSE. Plug in the numbers to

$$F_{alt} = \frac{\dfrac{SSE_{H_0} - SSE_{H_1}}{df_{SSE_{H_0}} - df_{SSE_{H_1}}}}{\dfrac{SSE_{H_1}}{df_{SSE_{H_1}}}}.$$

   Does the value you calculated match the F-statistic from part (a)?

(c) Run the partial ANOVA (typ=2) for the fitted model. Does the F-test for 'Cylinder' match the F-test from part (a)? Why or why not?

(d) From the partial ANOVA (typ=2) table in (c), report the null and alternative hypothesis, the F stat, and the p-value for the F-test for dropping or including the 'Mileage' predictor. Interpret the result of this test.

(e) Manually run the test in part (d) yourself: 1. fit the null model in python and extract SSE and degrees of freedom of this SSE; then 2. fit the alternative model in python and extract SSE and degrees of freedom of this SSE. Plug in the numbers to

$$F_{alt} = \frac{\dfrac{SSE_{H_0} - SSE_{H_1}}{df_{SSE_{H_0}} - df_{SSE_{H_1}}}}{\dfrac{SSE_{H_1}}{df_{SSE_{H_1}}}}.$$

Does the value you calculated match the F-statistic from part (d)?

5. This question will help you to understand the calculation of R^2 and R^2_{adj} in MLR using an example. For the dataset KelleyBlueBookData.csv:

(a) Fit Model 1: a model which considers Price as the response and regresses it against the predictors Mileage and Cylinder. Report the R^2 and R^2_{adj} values from the summary table.

(b) Calculate R^2 and R^2_{adj} for the model in part (a) yourself. Obtain the SSE and SST of the model in part (a), then plug in the formulas: $R^2 = 1 - \frac{SSE}{SST}$ and $R^2_{adj} = 1 - \dfrac{SSE/n-p}{SST/n-1}$ Do the values match with the python output in (a)?

(c) Fit Model 2: a model which considers Price as the response and regresses it against the predictors Mileage, Liter and Cylinder. Report the $R^2$and R^2_{adj} values from the summary table. Which model is preferable according to R^2_{adj} between Model 1 and Model 2? Why?

(d) Open question: Consider simultaneously the t-test results, ANOVA, R^2_{adj} and any other concepts we have covered so far (e.g. diagnostics). Which model would you choose, Model 1 or Model 2? Argue for your model in terms of these statistics and also the real life meaning of the problem.