Linear Regression
HW 3
Due 9/15 at 11:59pm

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

Textbook Questions:

1. P248: # 6.2
2. P248: # 6.3
3. P253: # 6.23
   (Hint: rewrite the matrix notation for X and β then it should lead to the same matrix notation for b)
4. P253: # 6.25
5. P254: # 6.27.
   Write down matrix X and vector y, and the formula to calculate each value before giving the answers. You can do the matrix calculation using Python or any calculator. Do NOT use Python to fit the model and read the output directly.
6. Consider the multiple linear regression model in the matrix form $Y = X\beta + \epsilon$ with $\mathrm{E}(\epsilon) = 0$ and $\mathrm{Var}(\epsilon) = \sigma^2 I_n$. Let $\hat{\beta}$ be the least squares estimator of β, $\hat{Y} = X\hat{\beta}$ be the fitted values, and $e = Y - \hat{Y}$ be the residuals. Express $\mathrm{Var}(\hat{\beta})$, $\mathrm{Var}(\hat{Y})$, and $\mathrm{Var}(e)$ in terms of σ^2 and X. Show your steps to obtain the expressions in detail.

Coding questions:

1. The dataset "insurance.csv" contains information on the individuals' medical costs billed by health insurance with the variables defined as below:

   - age: age of primary beneficiary
   - sex: insurance contractor gender, female, male
   - bmi: Body mass index, providing an understanding of weights that are relatively high or low relative to height, objective index of body weight (kg/m^2) using the ratio of height to weight
   - children: Number of children covered by health insurance / Number of dependents
   - smoker: Smoking
   - region: the beneficiary's residential area in the US, northeast, southeast, southwest, northwest.
   - charges: Individual medical costs billed by health insurance

   Regress "charges" against "age", "bmi" and "children" (use the order: charges $\sim$ age + bmi + childern) in Python.

(a) ONLY based on the model summary table (`model.summary()`), we have discussed several important things to look at when it comes to the model interpretation: Global F-test p-value, R^2 and R_a^2, and t-test for individual coefficient. Comment on each values above from your model summary, what kind of conclusion you have about your model fit and significance of predictors so far?

(b) Run `sm.stats.anova_lm(model,typ=1)` to obtain `sum_sq` of "bmi" in the ANOVA is a sequential sum of squares of "bmi" which is defined as $SeqSS(bmi) = SSE_{reduced} - SSE_{full}$.
What is the reduced model and full model here? How do we interpret the p-value of "bmi" in this F-test?

(c) Run `sm.stats.anova_lm(model,typ=2)` to obtain `sum_sq` of "bmi" in the anova is a partial sum of squares of "bmi" defined by $PartialSS(bmi) = SSE_{reduced} - SSE_{full}$.
What is the reduced model and full model here? How do we interpret the p-value of "bmi" in this F-test?

2. A commercial real estate company evaluates age (X_1), operating expenses (X_2, in thousands of dollars), vacancy rate (X_3), total square footage (X_4) and rental rates (Y, in thousands of dollars) for commercial properties in a large metropolitan area in order to provide clients with quantitative information upon which to make rental decisions. The data are taken from 81 suburban commercial properties. (The data can be found in `property.txt`; The first column is Y, followed by $X_1,\ X_2,\ X_3,\ X_4$.) Consider the multiple linear regression model

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{4i} + \epsilon_i,\ i = 1, \ldots, n, \text{ with } \epsilon_i \overset{iid}{\sim} N(0, \sigma^2).$$

(a) Fit the model in Python and produce the summary table.

(b) Compute the fitted values and residuals. Print the first 6 cases. Based on the residuals, give an estimate of σ^2.

(c) Test whether $\beta_2 = 0$ or not at 0.01 significance level. Interpret this coefficient in the real life terms of the problem.

(d) Test $H_0 : \beta_1 = \beta_2 = 0$. vs. H_a : either β_1 or β_2 not equal to 0.

(e) Consider a property with the following characteristics: $X_1 = 4, X_2 = 10, X_3 = 0.1, X_4 = 80,000$. Give a 90% prediction interval for the rental rate.