BIOST 546
WINTER QUARTER 2021

Homework # 1
Due Via Online Submission to Canvas: Monday, Jan 25 at 10 AM

Instructions: You may discuss the homework problems in small groups, but you must write up the final solutions and code yourself. Please turn in your code for the problems that involve coding. However, code without written answers will receive no credit. To receive credit, you must explain your answers and show your work. All plots should be appropriately labeled and legible, with axis labels, legends, etc., as needed.

1. In this problem, we will make use of the dataset `Medical_Cost.RData` (introduced in class), which you can find attached on Canvas.

   (a) Load the dataset with the command `load` and check if there are missing data.

   (b) We decide to focus on the outcome variable `charges` (individual medical costs billed by health insurance) and the predictors `bmi` (body mass index), and `smoker` (whether the subjects is a smoker or not). Make a scatterplot with `bmi` on the x-axis, `charges` on the y-axis, and with the color of each dot representing whether the subject is a smoker or not.

   (c) Fit a least-squares linear model, with intercept, in order to predict

   - `charges` using `bmi` as the only predictor;
   - `charges` using `bmi` and `smoker` as predictors;
   - `charges` using `bmi` and `smoker` as in the previous model; but allowing for an interaction term between the variables `bmi` and `smoker`;

   For each of the three models

   - Present your results in the form of a table where you report the estimated regression coefficients and their interpretation (be careful with the dummy variables).
   - Report the 95% confidence interval for the coefficient of the variable `bmi`, and provide a sentence explaining the meaning of this confidence interval.
   - Draw (can be hand-sketched) the regression line(s) of the model on the scatter plot produced in point (b) (See also Figure 1 for an example).

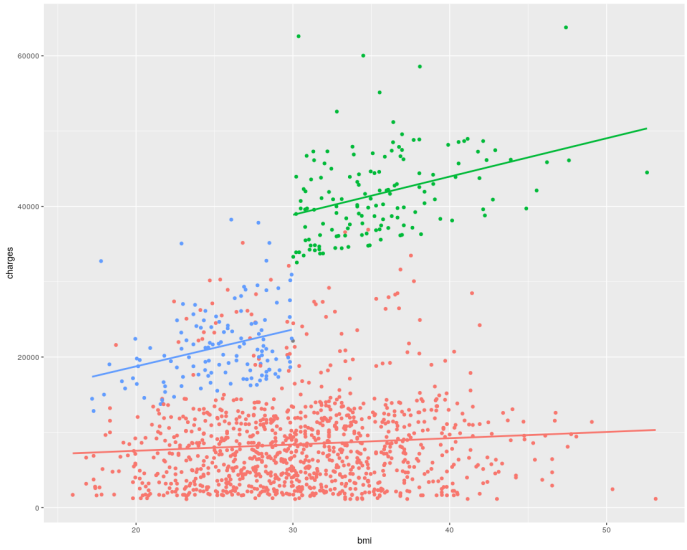

Figure 1: Scatter plot with BMI in the x-axis and charges in the y-axis. Red dots denote non-smokers, while blue (green) dots denote smokers whose BMI is below (above) 30. A straight line has been fitted, in the least-squares sense, to each of the described three groups.

- Report the (training set) mean squared error of the model.
- Predict the medical costs billed by the health insurance company to a smoker with a `bmi` that is 32.
- According to the model with interaction, on average, by how much would the charges change if the subject were to lower their bmi to 28.

(d) Now define and add to the dataset a new boolean variable `smoker_bmi30p` that is `True` only if the subject is a smoker **and** has a `bmi` greater than 30. Use this newly defined variable, together with `bmi` and `smoker`, to fit the linear model represented in Figure 1 by carefully defining the interaction terms (allow each of the three straight lines to have their own intercept and slope, but us the command `lm` only once).

- Present your results in the form of one table where you report the estimated coefficients of the model.
- For each predictor, comment on whether you can reject the null hypothesis that there is no (linear) association between that predictor and `charges`, conditional on the other predictors in the model.
- Explain the interpretation of the non-significant variables in the model ($p > 0.05$) and explain how Figure 1 would change if we were to discard those variables, i.e. perform variable selection.

2. For this problem, you will come up with some examples of statistical learning for biomedical or public health applications possibly related to your field.

(a) Provide three examples of regression problems motivated by biomedical

research or public health. In each example, describe Y and $X_1, \ldots, X_p$ as well as the scientific task and whether the problem is high-dimensional or low-dimensional.

(b) Provide three examples of classification problems motivated by biomedical research or public health. In each example, describe Y and $X_1, \ldots, X_p$ as well as the scientific task and whether the problem is high-dimensional or low-dimensional.

(c) Provide three examples of unsupervised learning motivated by biomedical research or public health. In each example, describe $X_1, \ldots, X_p$ as well as the scientific task and whether the problem is high-dimensional or low-dimensional.

3. This problem has to do with the bias-variance trade-off and related ideas. For (a) and (b), it's okay to submit hand-sketched plots: you are not supposed to actually compute the quantities referred to below on data; instead, this is a conceptual exercise.

(a) Make a plot, like the one we saw in class, with "flexibility" on the x-axis. Sketch the following curves: squared bias, variance, irreducible error, reducible error, expected prediction error. Be sure to label each curve. Indicate which level of flexibility is "best".

(b) Make a plot with "flexibility" on the x-axis. Sketch curves corresponding to the training error and the test error. Be sure to label each curve. Indicate which level of flexibility is "best".

(c) Explain where on the x-axis would a simple regression ($p = 1$) and a K-nearest neighbors approach (with $K = 1$) fall. Justify your answer.

4. Suppose that you are interested in performing regression on a particular dataset, and need to decide whether to pick a linear model or a non-parametric model (e.g. KNN).

- Describe how the issue of interpretability of the model might drive your choice between a linear model and a non-parametric model.
- What properties of the dataset would lead you to *definitely* use a linear model as opposed to a non-parametric approach?

5. Consider using only the `region` variable to predict `charges` on the `Medical_Cost.RData` data set. In this problem, we will explore the coding of this qualitative variable.

(a) First, code the `region` variable using three dummy (indicator) variables, with `northwest` as the default value. Write out the equation of the linear model conditional on all the different values of `region` and explain the meaning of the intercept and dummy variables.

(b) Now, code the `region` variable using three variables that take on values of $+\frac{1}{2}$ or $-\frac{1}{2}$. Write out the equation of the linear model conditional on all the different values of `region` and explain the meaning of the intercept and dummy variables.