BIOST 546
WINTER QUARTER 2023

Homework # 1
Due Via Online Submission to Canvas: Wed, 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_2.RData` (introduced in class), which you can find on Canvas.

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

   (b) If any, remove the missing data using the command `na.omit`.

   (c) 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.

   (d) 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).

- 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 29 and 31.5.
- Compute the predicted difference in charges between a smoker with `bmi` 31.5 and one with `bmi` 29. Do the same for non-smokers. Comment on the results.

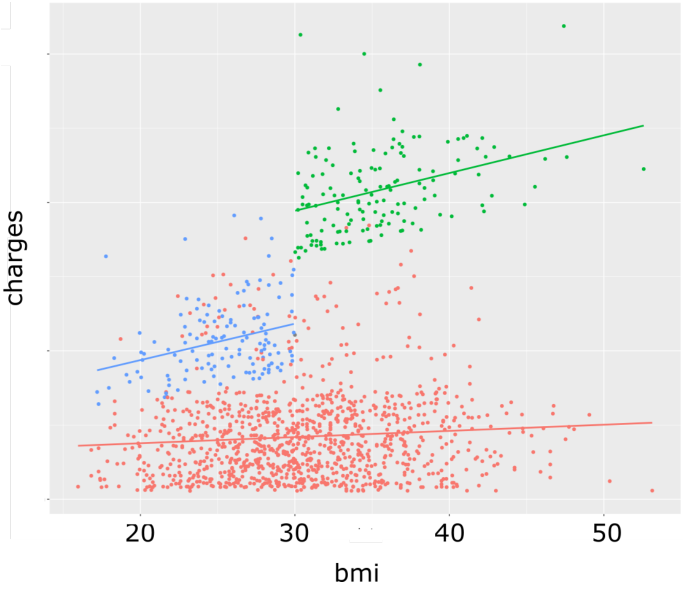

Figure 1: Scatter plot with BMI in the x-axis and charges in the y-axis. Red dots denote non-smokers, while blue (green) dots denote smokers whose BMI is below (above) 30. A straight line has been fitted, in the least-squares sense, to each of the described three groups.

(e) 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 use 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.
- According to this newly defined model, compute the predicted difference in charges between a smoker with `bmi` 31.5 and one with `bmi` 29.

Do the same for non-smokers. Compare with the analogous results in point (d) and comment on the results.

2. This problem has to do with the notion of bias-variance trade-off. For (a) and (b), it's okay to submit hand-sketched plots: 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, 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".

3. This problem has to do with numerical explorations of the bias-variance trade-off phenomenon. You will generate simulated data, and will use these data to perform **linear regression**. Set the seed with `set.seed(0)` before you begin.

   (a) Use the `rnorm()` function to generate a predictor vector X of length $n = 30$, and use `runif()` to generate a noise vector ϵ of length $n = 30$.

   (b) Generate a response vector Y of length $n = 30$ according to the model

   $$Y = f^{\text{true}}(X) + \epsilon,$$

   with $f^{\text{true}}(X) = 3 + 2X + 3 * X^3$

   (c) Fit the model $Y = f(X) + \epsilon$ to the data (using the `lm()` function), for the following choices of f:

   - $f(X) = \beta_0 + \beta_1 * X$
   - $f(X) = \beta_0 + \beta_1 * X + \beta_2 * X^2$
   - $f(X) = \beta_0 + \beta_1 * X + \beta_2 * X^2 + \beta_3 * X^3 + \beta_4 * X^4$
   - $f(X) = \beta_0 + \beta_1 * X + \beta_3 * X^3$

   (d) For each of the models above compute the training mean squared error (MSE). Comment on the results.

   (e) Now generate 10K (new) **test** observations following steps 3(a) and 3(b). Compute the test MSE of the models fitted in 3(c) on these **test** observations. Report and comment on the results.

   (f) Compute training and test MSEs of the true regression function f^{true}. Compare to those of the models fitted in 3(c). Comment on the results.

   (g) (**Extra Credit**) Now we want to estimate the bias and variance of each one of the three models in 3(c). To this purpose, we generate 40 datasets of 30 observations repeating points 1(a) and 1(b) 40 times. For **each** model

   - $f(X) = \beta_0 + \beta_1 * X$

- $f(X) = \beta_0 + \beta_1 * X + \beta_2 * X^2$
- $f(X) = \beta_0 + \beta_1 * X + \beta_2 * X^2 + \beta_3 * X^3 + \beta_4 * X^4$
- $f(X) = \beta_0 + \beta_1 * X + \beta_3 * X^3$

do the following:

- Estimate its coefficients fitting a separate model to each of the 40 datasets and compute the associated 40 predictions at $x_0 = 0.3$, i.e.,

  $$\hat{f}(x_0), \text{ with } x_0 = 0.3.$$

- Approximate the bias and variance as follows

  $$\Big(\text{Ave}[\hat{f}(x_0)] - f^{\text{true}}(x_0)\Big)^2$$

  and

  $$\text{Var}[\hat{f}(x_0)],$$

  where Ave and Var are the average and variance computed across the 40 datasets.

Comment on the results and relate to the test MSEs computed in 3(e).