STAT 435
SPRING QUARTER 2018

Homework # 2
Due Friday, April 13, 2018 at 12:00 PM (Noon)
Online Submission Via Canvas

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, for the problems that involve coding, you must also provide written answers: you will receive no credit if you submit code without written answers. You might want to use Rmarkdown to prepare your assignment.

1. Suppose we have a quantitative response Y, and a single feature $X \in \mathbb{R}$. Let RSS_1 denote the residual sum of squares that results from fitting the model

$$Y = \beta_0 + \beta_1 X + \epsilon$$

using least squares. Let RSS_{12} denote the residual sum of squares that results from fitting the model

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \epsilon$$

using least squares.

   (a) Prove that $RSS_{12} \leq RSS_1$.

   (b) Prove that the R^2 of the model containing just the feature X is no greater than the R^2 of the model containing both X and X^2.

2. Describe the null hypotheses to which the p-values in Table 3.4 of the textbook correspond. Explain what conclusions you can draw based on these p-values. Your explanation should be phrased in terms of `sales`, `TV`, `radio`, and `newspaper`, rather than in terms of the coefficients of the linear model.

3. Consider a linear model with just one feature,

$$Y = \beta_0 + \beta_1 X + \epsilon.$$

Suppose we have n observations from this model, $(x_1, y_1), \ldots, (x_n, y_n)$. The least squares estimator is given in (3.4) of the textbook. Furthermore, we saw

in class that if we construct a $n \times 2$ matrix $\tilde{\mathbf{X}}$ whose first column is a vector of 1's and whose second column is a vector with elements $x_1, \ldots, x_n$, and if we let $\mathbf{y}$ denote the vector with elements $y_1, \ldots, y_n$, then the least squares estimator takes the form

$$\begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{pmatrix} = \left(\tilde{\mathbf{X}}^T \tilde{\mathbf{X}}\right)^{-1} \tilde{\mathbf{X}}^T \mathbf{y}. \quad (1)$$

Prove that (1) agrees with equation (3.4) of the textbook, i.e. $\hat{\beta}_0$ and $\hat{\beta}_1$ in (1) equal $\hat{\beta}_0$ and $\hat{\beta}_1$ in (3.4).

4. This question involves the use of multiple linear regression on the `Auto` data set, which is available as part of the `ISLR` library.

   (a) Use the `lm()` function to perform a multiple linear regression with `mpg` as the response and all other variables except `name` as the predictors. Use the `summary()` function to print the results. Comment on the output. For instance:

      i. Is there a relationship between the predictors and the response?

      ii. Which predictors appear to have a statistically significant relationship to the response?

      iii. Provide an interpretation for the coefficient associated with the variable `year`.

      Make sure that you treat the qualitative variable `origin` appropriately.

   (b) Try out some models to predict `mpg` using functions of the variable `horsepower`. Comment on the best model you obtain. Make a plot with `horsepower` on the x-axis and `mpg` on the y-axis that displays both the observations and the fitted function (i.e. $\hat{f}$(`horsepower`)).

   (c) Now fit a model to predict `mpg` using `horsepower`, `origin`, and an interaction between `horsepower` and `origin`. Make sure to treat the qualitative variable `origin` appropriately. Comment on your results. Provide a careful interpretation of each regression coefficient.

5. Consider fitting a model to predict credit card `balance` using `income` and `student`, where `student` is a qualitative variable that takes on one of three values: `student`$\in$ {`graduate`, `undergraduate`, `not student`}.

   (a) Encode the student variable using two dummy variables, one of which equals 1 if `student=graduate` (and 0 otherwise), and one of which equals 1 if `student=undergraduate` (and 0 otherwise). Write out an expression for a linear model to predict `balance` using `income` and `student`, using this coding of the dummy variables. Interpret the coefficients in this linear model.

   (b) Now encode the student variable using two dummy variables, one of which equals 1 if `student=not student` (and 0 otherwise), and one of which

equals 1 if `student=graduate` (and 0 otherwise). Write out an expression for a linear model to predict `balance` using `income` and `student`, using this coding of the dummy variables. Interpret the coefficients in this linear model.

(c) Using the coding in (a), write out an expression for a linear model to predict `balance` using `income`, `student`, and an interaction between `income` and `student`. Interpret the coefficients in this model.

(d) Using the coding in (b), write out an expression for a linear model to predict `balance` using `income`, `student`, and an interaction between `income` and `student`. Interpret the coefficients in this model.

(e) Using simulated data for `balance`, `income`, and `student`, show that the fitted values (predictions) from the models in (a)–(d) do not depend on the coding of the dummy variables (i.e. the models in (a) and (b) yield the same fitted values, as do the models in (c) and (d)).

6. **Extra Credit.** Consider a linear model with just one feature,

$$Y = \beta_0 + \beta_1 X + \epsilon,$$

with $\mathrm{E}(\epsilon) = 0$ and $\mathrm{Var}(\epsilon) = \sigma^2$. Suppose we have n observations from this model, $(x_1, y_1), \ldots, (x_n, y_n)$. We assume that $x_1, \ldots, x_n$ are fixed, so the only randomness in the model comes from $\epsilon_1, \ldots, \epsilon_n$. Use (3.4) in the textbook — or, if you prefer, the matrix algebra formulation in (1) of this homework assignment — in order to derive the expressions for $\mathrm{Var}(\hat{\beta}_0)$ and $\mathrm{Var}(\hat{\beta}_1)$ given in (3.8) of the textbook.