Homework 4: Linear regression

Complete this homework as a jupyter notebook titled “LASTNAME_Homework4.ipynb” posted on your GitHub account. Turn in your assignment by emailing a link to the notebook to timothyv@andrew.cmu.edu by no later than 3pm on Thursday March 8th, 2018.

This homework covers material covered in Chapter 3 of *Introduction to Statistical Learning with Applications in R (ISLR)* and the chapter *A Simple, Linear, Mixed-effects Model*. Some questions come directly from the book exercises, others are new questions.

Conceptual: Short answer questions and proofs. Be concise. Show stepwise solutions to proofs where appropriate

Remember that:

- $E[X^2] = Var[X] + E[X]^2$
- $E[XY] = Cov[XY] + E[X]E[Y]$
- $E[aX] = aE[X]$

1. Using the mean squared error (MSE) as your objective function allows for a closed form solution to finding the maximum likelihood estimate (MLE) of your model parameters in linear regression. Let’s consider the simple, single predictor variable model $Y = b_0 + b_1X$.
   (a) Use algebra to show how you can expand out the MSE to get from i to ii below.
      (i) $MSE(b_0, b_1) = E[(Y - (b_0 + b_1X))^2]$
      (ii) $MSE(b_0, b_1) = E[Y^2] - 2b_0E[Y] - 2b_1Cov[X, Y] - 2b_1E[X]E[Y] + b^2{}_0 + 2b_0b_1E[X] + b^2{}_1Var[X] + b^2{}_1(E[X])^2$
   (b) Prove that the MLE of b_0 is $E[Y] - b_1E[X]$ by taking the derivative of ii above, with respect to b_0, setting the derivative to zero, and solving for b_0.
   (c) Prove that the MLE for b_1 is $Cov[X, Y]/Var[X]$ by taking the derivative of ii above, with respect to b_1. setting the derivative to zero, and solving for b_1

2. Three methods for evaluating a regression model’s performance are R^2, RSE, and the F Test. Show the functional form of each method and describe conceptually what it is telling you about your model.

3. Suppose we have a dataset with five predictors, X_1 = GPA, X_2 = IQ, X_3 =Gender (1 for Female and 0 for Male), X_4 = Interaction between GPA and IQ, and X_5 = Interaction between GPA and Gender. The response variable is starting salary after graduation (in thousands of dollars). Suppose we use least squares to fit the model, and get $\hat{\beta}_0 = 50$, $\hat{\beta}_1 = 20$, $\hat{\beta}_2 = 0.07$, $\hat{\beta}_3 = 35$, $\hat{\beta}_4 = 0.01$, $\hat{\beta}_5 = -10$.

(a) Which answer is correct, and why?

i. For a fixed value of IQ and GPA,males earn more on average than females.

ii. For a fixed value of IQ and GPA, females earn more on average than males.

iii. For a fixed value of IQ and GPA, males earn more on average than females provided that the GPA is high enough.

iv. For a fixed value of IQ and GPA, females earn more on average than males provided that the GPA is high enough.

(b) Predict the salary of a female with IQ of 110 and a GPA of 4.0.

(c) True or false: Since the coefficient for the GPA/IQ interaction term is very small, there is very little evidence of an interaction effect. Justify your answer.

4. A linear mixed effect model has the form $y = X\beta + Z\upsilon + \varepsilon$.

(a) What are the assumptions about Z and υ that make this different from using a regular control variable in a simple linear regression model?

(b) What is the objective function for this model and how is it different than the objective function for the ordinary least squares model?

(c) What advantage does including a random effects term in your model give you with regards to understanding the fixed effects relationship?

Applied: Submit all R code, along with plots and written responses.

Do applied exercises 3.10 and 3.13 from *ISLR* as well as question 5 below.

5. Load the *cbpp* dataset from the *lme4* library. Run *?cbpp* after loading the *lme4* library to get information on the dataset.

(a) Plot the relationship between herd size and CBPP incidence

(b) Use the *lm* function to model the effect of herd size (predictor) on CBPP incidence (response). How does herd size impact disease rate? Report the results of the model.

(c) Use the *lmer* function to include the herd identifier as a random effect. How does this impact the fixed effect of herd size on CBPP incidence?

(d) Compare the simple linear model (i.e., the fixed effects only model) with the mixed effect model using AIC (https://en.wikipedia.org/wiki/Akaike_information_criterion). Does adding the random effect to the model improve or change the fixed effect?