Ch. 3 Regression - Problem Bank Questions

August 24, 2020

1. Suppose we have a data set 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 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? Explain why other statements are incorrect.

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.

(d) Interpret $\hat{\beta}_0$, $\hat{\beta}_3$, $\hat{\beta}_4$, $\hat{\beta}_5$.

2. I collect a set of data (n = 100 observations) containing a single predictor and a quantitative response. I then fit a linear regression model to the data, as well as a separate cubic regression, i.e. $Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \epsilon$.

(a) Suppose that the true relationship between X and Y is linear, i.e. $Y = \beta_0 + \beta_1 X + \epsilon$. Consider the training residual sum of squares (RSS) for the linear regression, and also the training RSS for the cubic regression. Would we expect one to be lower than the other, would we expect them to be the same, or is there not enough information to tell? Justify your answer.

(b) Answer (a) using test rather than training RSS.

(c) Suppose that the true relationship between X and Y is not linear, but we don't know how far it is from linear. Consider the training RSS for the linear regression, and also the training RSS for the cubic regression. Would we expect one to be lower than the other, would we expect them to be the same, or is there not enough information to tell? Justify your answer.

(d) Answer (c) using test rather than training RSS.

3. (a) Interpret R^2.

(b) Theoretically, Is there a lower limit to R2? If so, what is it?

(c) Practically, an R^2 of zero is equivalent to predicting what value for all data points?

(d) Based on the formula for R^2, what can you say about the predicted y_i values for a regression fit where $R^2 = 1$?

~~(e) It is claimed in the text that in the case of simple linear regression of Y onto X, the R^2 statistic (3.17 ISL) is equal to the square of the correlation between X and Y (3.18 ISL). Prove that this is the case. For simplicity, you may assume that $\bar{x} = \bar{y} = 0$.~~

(f) Why is R^2 not a good measure of model fit for multiple regression?

4. Let's say you and your friend are predicting y values for the same dataset.

(a) Your friend recommends an insanely flexible neural network. Why might you recommend linear regression instead?

(b) Your friend sees your point and decides to throw all the predictors they can into a linear regression model, saying if the variable doesn't matter its coefficient will just go to zero so it's worth putting everything in. What could go wrong?

(c) Realizing you're right again, your friend settles on a linear regression model with interaction terms. They notice that main effects aren't significant and so they want to exclude them. Why might you want to include these insignificant variables?

5. Show that $\hat{\beta}_0$ and $\hat{\beta}_1$ minimize the residual sum of squares.

6. Mark each statement below as TRUE or FALSE, and justify why in 1-2 sentences. If FALSE, reword the incorrect statement into a true one.

(a) T/F: Confidence intervals are always wider than prediction intervals.

(b) T/F: Residual plots can be a good method of diagnosing model fit.

(c) T/F: Individual t-statistic p-values should be used in multiple regression to determine what variables to include in a model.

(d) T/F: ϵ in the multiple regression model is an example of irreducible error.

7. Model selection

(a) Given a model with p predictors, how many total models contain subsets of p variables?

(b) Is the previous method of variable selection practical or efficient, and why?

(c) Describe two other different variable selection methods.

(d) Which method in (c) cannot be used if $p > n$, and why?

(e) What are some possible issues that might arise with using forward or backwards selection to determine important variables in a multiple regression setting? Are you guaranteed to always find the most important and relevant variables?

8. In section 3.3.3, ISL covers 6 potential problems with linear regression models. Explain each of them, and describe how to identify and address the problem.

(a) Non-linearity of the response-predictor relationships

(b) Correlation of error terms

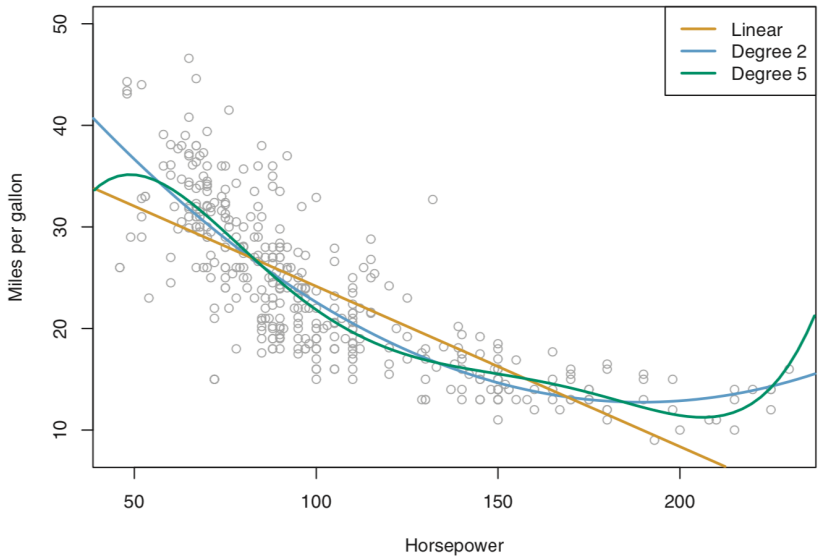
(c) Non-constant variance of error terms

(d) Outliers

(e) High-leverage points

(f) Collinearity

9. In the following plot, three regression fits are shown. The orange line shows a regular linear regression fit. The blue line shows a linear regression model that includes horsepower^2. Finally, the linear regression fit for a model that includes all polynomials of horsepower up to fifth-degree is shown in green.



Based on this plot (and your knowledge about linear regression), which model seems like the best choice? Why?

10. Mark each statement below as TRUE or FALSE, and justify why in 1-2 sentences. If FALSE, reword the incorrect statement into a true one.

Are these the possible reasons that you might want to perform variable selection?

(a) T/F: To remove unimportant variables from your model.

(b) T/F: To increase the R^2 value of your model.

(c) T/F: To improve interpretability of the model.