Regression Analysis

Assignment 2A

Problem 3.1 on page 132. This exercise uses the MOVIES dataset and covers the multiple linear regression model (in particular, see page 85 for part (a) and pages 90-91 for parts (b) and (c)).

Problem 3.1. The **MOVIES** data file contains data on 25 movies from "The Internet Movie Database" (www.imdb.com). Based on this dataset, we wish to investigate whether all-time U.S. box office receipts (*Box*, in millions of U.S. dollars unadjusted for inflation) are associated with any of the following variables:

Rate = Internet Movie Database user rating (out of 10)
User = Internet Movie Database users rating the movie (in thousands)
Meta = "metascore" based on 35 critic reviews (out of 100)
Len = runtime (in minutes)
Win = award wins
Nom = award nominations

Theatrical box office receipts (movie ticket sales) may include theatrical re-release receipts, but exclude video rentals, television rights, and other revenues.

a. Write out an equation (like the equation at the bottom of page 85) for a multiple linear regression model for predicting response *Box* from *just three predictors: Rate, User, and Meta*.

   Hint: This question is asking you to write an equation for E(Box).

b. Use statistical software to fit this model [computer help #31] and write out the estimated multiple linear regression equation [i.e., replace the b's in part (a) with numbers].

   Hint: This question is asking you to write an equation for Box-hat.

c. Interpret the estimated regression parameter for *Rate* in the context of the problem [i.e., put the appropriate number from part (b) into a meaningful sentence, remembering to include the correct units for any variables that you use in your answer].

   Hint: See the bottom of page 125 for an example of the type of sentence expected.

[5 points]

Problem 3.3 on page 133. This exercise also uses the MOVIES dataset and covers the nested model F-test (pages 104-109), the regression standard error (pages 92-94), and adjusted R-squared (pages 95-99). Skip part (d) if the statistical software you are using cannot do a nested model test directly.

Problem 3.3. Consider the **MOVIES** data file from Problem 3.1 again.

a. Use statistical software to fit the following (complete) model for *Box* as a function of all six predictor variables [computer help #31]:

   $$E(Box) = b_0 + b_1 Rate + b_2 User + b_3 Meta + b_4 Len + b_5 Win + b_6 Nom.$$

   Write down the residual sum of squares for this model.

b. Use statistical software to fit the following (reduced) model [computer help #31]:

$$E(Box) = b_0 + b_1 Rate + b_2 User + b_3 Meta.$$

[This is the model from Problem 3.1 part (b)]. Write down the residual sum of squares for this model.

c. Using the results from parts (a) and (b) together with the nested model test F-statistic formula on page 105, test the null hypothesis NH: $b_4 = b_5 = b_6 = 0$ in the complete model, using significance level 5%. Write out all the hypothesis test steps and interpret the result in the context of the problem.

Hint: To solve this part you may find the following information useful. The 95th percentile of the F-distribution with 3 numerator degrees of freedom and 18 denominator degrees of freedom is 3.160.

d. Check your answer for part (c) by using statistical software to do the nested model test directly [computer help #34]. State the values of the F-statistic and the p-value, and draw an appropriate conclusion.

e. Another way to see whether we should prefer the reduced model for this example is to see whether the regression standard error (s) is smaller for the reduced model than for the complete model and whether adjusted R^2 is higher for the reduced model than for the complete model. Confirm whether these relationships hold in this example (i.e. compare the values of s and adjusted R^2 in the reduced and complete models).

[10 points]