STAT 331: Assignment 3

Due: Wednesday July 10, 2019 by 11:59pm

Submission

Your assignment must be submitted by the due date listed at the top of this document, and it must be submitted electronically via *Crowdmark*. This means that your responses for different questions should be on separate pages.

For non-computational questions, you may upload solutions produced by LaTeX (or some other equation editing software), or handwritten solutions that have been scanned or photographed. Please note that if you scan or photograph your solutions it is your duty to ensure they are readable and of high quality. If the marking team cannot read your solutions, they simply will not be marked.

For computational questions, I highly recommend you produce your solutions as a nicely formatted .pdf file with R Markdown. R Markdown facilitates the seamless combination of written text, R code and output, and LaTeX equations. Your submission for these questions should include the code, the corresponding output, and the interpretations where appropriate.

Question 1 [8 points]

(a) [5 points] Consider two hypothetical models: Model 1 and Model 2, where Model 2 contains all of the explanatory variables in Model 1 plus extra ones. Thus, Model 1 is *nested* within Model 2. Show that the coefficient of determination for Model 2 will always be greater than the coefficient of determination for Model 1. In other words, prove that $R_2^2 > R_1^2$.

(b) [1 point] Explain, in your own words, why the quantity R_{adj}^2 is more appropriate than R^2 as a model selection criterion.

(c) [2 points] By considering

$$\lim_{n\to\infty} R_{adj}^2$$

show that if the sample size is large enough (and the number of explanatory variables is fixed), then there is no difference between R_{adj}^2 and R^2 in which case either is appropriate.

Question 2 [23 points]

Candy Crush is experimenting with three different versions of in-game "boosters": the lollipop hammer, the jelly fish, and the colour bomb. Users in each booster condition receive (for free) 5 boosters corresponding to their condition and interest lies in evaluating the effect of these different boosters on the length of time a user plays the game. In particular interest lies in testing the following hypothesis:

$$\mathrm{H}_0: \mu_1 = \mu_2 = \mu_3 \text{ vs. } \mathrm{H}_\mathrm{A}: \mu_j \neq \mu_k \text{ for some } j \neq k \qquad [1]$$

where μ_i represents the expected length of game play (in minutes) associated with booster $i = 1, 2, 3$. The data are shown below. (Note that this is a small experiment with only 3 users in each booster condition).

Time (y)	Booster (x)
12	Lollipop Hammer
11	Lollipop Hammer
10	Lollipop Hammer
10	Jelly Fish
7	Jelly Fish
13	Jelly Fish
15	Colour Bomb
14	Colour Bomb
16	Colour Bomb

(a) [4 points] Hypothesis [1] may be tested with an appropriately defined linear regression model and the *additional sum of squares* principle. Let

$$x_{i1} = \begin{cases} 1 & \text{if user } i \text{ is in the lollipop hammer condition} \\ 0 & \text{otherwise} \end{cases}$$

$$x_{i2} = \begin{cases} 1 & \text{if user } i \text{ is in the jelly fish condition} \\ 0 & \text{otherwise} \end{cases}$$

$$x_{i3} = \begin{cases} 1 & \text{if user } i \text{ is in the colour bomb condition} \\ 0 & \text{otherwise} \end{cases}$$

and define the linear regression model

$$y_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \varepsilon_i$$

Interpret β_1, β_2 and β_3 and explain why hypothesis [1] is equivalent to the following:

$$\mathrm{H}_0: \beta_1 = \beta_2 = \beta_3 \text{ vs. } \mathrm{H}_\mathrm{A}: \beta_j \neq \beta_k \text{ for some } j \neq k \qquad [2]$$

(b) [7 points] By hand (show your work), calculate the least squares estimates $\hat{\beta}_1$, $\hat{\beta}_2$, $\hat{\beta}_3$ using the formula $\widehat{\boldsymbol{\beta}} = (X^T X)^{-1} X^T \boldsymbol{y}$. Note that you will have to appropriately define $\boldsymbol{y}$ and X.

(c) [2 points] By hand (show your work), calculate the residual sum of squares (i.e. SSE) for this model.

(d) [1 point] State the matrix A associated with the additional sum of squares test of hypothesis [2], and state the reduced model that results when H_0 is true.

(e) [3 points] By hand (show your work), calculate the residual sum of squares (i.e. SSE) associated with the reduced model from part (d).

(f) [3 points] Using your results from parts (c) and (e), calculate the additional sum of squares test statistic associated with hypothesis [2].

(g) [3 points] Using R, calculate the *p*-value associated with hypothesis [2] and state whether, at a 5% level of significance, you reject or fail to reject H_0. Yes or no, is the expected length of game play the same for each booster condition?

Question 3 [32 points]

In this question you will consider the baseball dataset found in the file `hitters.csv`. This dataset records the salary of $n = 263$ Major League Baseball players during the 1987 season as well as $p = 19$ statistics associated with the performance of each player during the previous season. Specifically, the dataset contains observations from the following variables:

- `AtBat`: Number of times at bat in 1986
- `Hits`: Number of hits in 1986
- `HmRun`: Number of home runs in 1986
- `Runs`: Number of runs in 1986
- `RBI`: Number of runs batted in in 1986
- `Walks`: Number of walks in 1986
- `Years`: Number of years in the major leagues
- `CAtBat`: Number of times at bat during his career
- `CHits`: Number of hits during his career
- `CHmRun`: Number of home runs during his career
- `CRuns`: Number of runs during his career
- `CRBI`: Number of runs batted in during his career
- `CWalks`: Number of walks during his career
- `League`: A categorical variable with levels `A` (for American) and `N` (for National) indicating the player's league at the end of 1986
- `Division`: A categorical variabke with levels `E` (for East) and `W` (for West) indicating the player's division at the end of 1986

- `PutOuts`: Number of put outs in 1986
- `Assists`: Number of assists in 1986
- `Errors`: Number of errors in 1986
- `Salary`: 1987 annual salary on opening day in thousands of dollars
- `NewLeague`: A categorical variable with levels `A` and `N` indicating the player's league at the beginning of 1987

Interest lies in developing a model that relates a player's annual salary to their previous performance.

(a) [3 points] Fit a multiple linear regression model relating `Salary` to the nineteen explanatory variables listed above and calculate and interpret that value of R^2.

(b) [3 points] Construct the ANOVA table for this model and these data. Note that the table should have just three rows: one for variation the model explains, one for variation the model doesn't explain and one for total variation.

(c) The following questions concern the test of *overall significance* of the model.
   i. [1 point] State the hypothesis associated the test of overall significance of the model.
   ii. [1 point] State the value of the test statistic associated with the hypothesis in i.
   iii. [1 point] State the null distribution of the test statistic in ii.
   iv. [1 point] Calculate the *p*-value associated with the hypothesis in i.

(d) [4 points] Fit a reduced model relating `Salary` to `AtBat`, `Hits`, `Walks`, `CRuns`, `CWalks`, `Division`, `PutOuts`, and `Assists`. Calculate and interpret that value of R^2. By commenting on the R^2 value computed here versus in part (a) and relevant *p*-values, comment on whether `HmRun`, `Runs`, `RBI`, `Years`, `CAtBat`, `CHits`, `CHmRun`, `CRBI`, `League`, `Errors`, and `NewLeague` seem to significantly influence a player's salary.

(e) [1 point] State the null hypothesis that gives rise to the reduced model from part (d).

(f) [4 points] Test the hypothesis from (e) using the additional sum of squares principle. Be sure to state the value of the test statistic, the *p*-value, and whether you reject or fail to reject the null hypothesis at a 5% level of significance.

(g) In the context of the reduced model from part (d), construct the following residual plots. Be sure to include plot titles and axis labels.
   i. [1 point] Studentized Residuals vs. Index
   ii. [1 point] Studentized Residuals vs. Fitted Values
   iii. [1 point] Histogram of Studentized Residuals
   iv. [1 point] QQ-plot of Studentized Residuals

(h) Based on the plots in part (g), answer "Yes" or "No" to the following questions and give a one sentence justification.

i. [2 points] Do the residuals appear to be independent?
ii. [2 points] Do the residuals appear to have constant variance?
iii. [2 points] Do the residuals appear to be normally distributed?
iv. [2 points] Do the residuals suggest the existence of an outlier?

(i) [1 point] Suppose that any one of the residual assumptions is not satisfied. Indicate, from the list below, which inferences would *no longer be valid*.

- Parameter estimates (i.e., $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p$ and $\hat{\sigma}^2$).
- Hypothesis tests (i.e., $H_0: \beta_j = 0$ vs. $H_A: \beta_j \neq 0$).
- Confidence intervals (i.e., for $\beta_0, \beta_1, \ldots, \beta_p$ or μ_0).
- Prediction intervals for y_0.