STAT 231 — Linear Models — Fall 2021

Homework 7 (R) – Multiple Linear Regression

NAME: _____

Directions: Create a R markdown file in order to write up your solutions. You must show all of your work to receive full credit. Put a box around your final answer, whenever possible. Put the following command between each problem's writeup:

```
\pagebreak
```

1. (10 pts) Detailed interviews were conducted with over 1,000 street vendors in the city of Puebla, Mexico, in order to study the factors influencing vendors' incomes (World Development, February 1998). Vendors were defined as individuals working in the street, and included vendors with carts and stands on wheels and excluded beggars, drug dealers, and prostitutes. The researchers collected data on gender, age, hours worked per day, annual earnings, and education level. These data can be found in `STREETVN.Rdata`.

   (a) Use R to fit a first-order model for mean annual earnings, $E(y)$, as a function of age (x_1) and hours worked (x_2). Summarize the results of the model fit.

   (b) Interpret the estimated β coefficients in your model separately.

   (c) Use the output to comment on the global utility of the model (at $\alpha = .01$). Interpret the result.

   (d) Use the output to report and interpret the value of R_a^2.

   (e) Is age (x_1) a statistically useful predictor of annual earnings? Test using $\alpha = .01$.

   (f) Find a 95% confidence interval for β_2. Interpret the interval in the words of the problem.

2. (10 pts) Refer to the previous problem of street vendors' earnings (y).

   (a) Use R to compute a 95% confidence interval for $E(y)$ for a 45-year-old vendor who works 10 hours a day (i.e., for $x_1 = 45$ and $x_2 = 10$). Interpret the confidence interval in the words of the problem.

   (b) Use R to compute a 95% prediction interval for annual earnings for a 45-year-old vendor who works 10 hours a day (i.e., for $x_1 = 45$ and $x_2 = 10$). Interpret the prediction interval in the words of the problem.

3. (10 pts) Environmental Science and Technology (January 2005) reported on a study of the reliability of a commercial kit to test for arsenic in groundwater. The field kit was used to test a sample of 328 groundwater wells in Bangladesh. In addition to the arsenic level (micrograms per liter), the latitude (degrees), longitude (degrees), and depth (feet) of each well was measured. These data can be found in `ASWELLS.Rdata`.

   (a) Use R to fit a first-order model for arsenic level, $E(y)$, as a function of latitude, longitude, and depth. Summarize the results of the model fit.

   (b) Interpret the estimated β coefficients in your model separately.

   (c) Use the output to comment on the global utility of the model (at $\alpha = .05$). Interpret the result.

(d) Use the output to report and interpret the value of R_a^2.

(e) Based on the results of the previous parts, would you recommend using the model to predict arsenic level (y)? Explain.

4. (10 pts) Refer to the previous problem of arsenic level (y) in groundwater. Using the data in the `ASWELLS.Rdata` file, you fit a first-order model for arsenic level as a function of latitude, longitude, and depth. Based on the model statistics, the researchers concluded that the arsenic level is highest at a low latitude, high longitude, and low depth. Do you agree?

   (a) Use R to compute a 95% confidence interval for $E(y)$ for the lowest latitude, highest longitude, and lowest depth that are within the range of the sample data. Interpret the confidence interval in the words of the problem.

   (b) Use R to compute a 95% prediction interval for arsenic level for the lowest latitude, highest longitude, and lowest depth that are within the range of the sample data. Interpret the prediction interval in the words of the problem.

5. (10 pts) Refer to Problem 1 regarding street vendors' earnings (y). Recall that the vendors' mean annual earnings, $E(y)$, was modeled as a first-order function of age (x_1) and hours worked (x_2). Now, consider the interaction model $E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2$.

   (a) Use R to fit a second-order interaction model for mean annual earnings, $E(y)$. Summarize the results of the model fit.

   (b) What is the estimated slope relating annual earnings (y) to age (x_1) when number of hours worked (x_2) is 10? Interpret the result.

   (c) What is the estimated slope relating annual earnings (y) to hours worked (x_2) when age (x_1) is 40? Interpret the result.

   (d) Give the null hypothesis for testing whether age (x_1) and hours worked (x_2) interact.

   (e) Report the p-value of the interaction test. Give the appropriate conclusion in the words of the problem.

6. (10 pts) Refer to Problem 3 regarding arsenic level. Write a model for arsenic level (y) that includes first-order terms for latitude, longitude, and depth, as well as terms for interaction between latitude and depth and interaction between longitude and depth.

   (a) Use R to fit the interaction model for arsenic level, $E(y)$. Summarize the results of the model fit.

   (b) Conduct a test (at $\alpha = .05$) to determine whether latitude and depth interact to effect arsenic level.

   (c) Conduct a test (at $\alpha = .05$) to determine whether longitude and depth interact to effect arsenic level.

   (d) Practically interpret the results of the tests, parts b and c.