Homework 2

Your Name

March 6, 2019 at 11:59pm

Homework Guidelines:

Please read Homework Guidelines. You must follow these guidelines.

Turning the Homework in:

Please turn the homework in through canvas. You may use a pdf, html or word doc file to turn the assignment in.

PHP 1511 Assignment Link

PHP 2511 Assignment Link

For the R Markdown Version of this assignment: HW2.Rmd

Part 1: Multiple Linear Regression (PHP 1511-2511 Both Complete)

The Data

The goal of this study was to identify risk factors associated with giving birth to a low birth weight baby (weighing less than 2500 grams). Data were collected on 189 women, 59 of which had low birth weight babies and 130 of which had normal birth weight babies. Four variables which were thought to be of importance were age, weight of the subject at her last menstrual period, race, and the number of physician visits during the first trimester of pregnancy.

Low birth weight is an outcome that has been of concern to physicians for years. This is due to the fact that infant mortality rates and birth defect rates are very high for low birth weight babies. A woman's behavior during pregnancy (including diet, smoking habits, and receiving prenatal care) can greatly alter the chances of carrying the baby to term and, consequently, of delivering a baby of normal birth weight.

The variables identified in the code sheet given in the table have been shown to be associated with low birth weight in the obstetrical literature. The goal of the current study was to ascertain if these variables were important in the population being served by the medical center where the data were collected. This data is from Hosmer et al. , 2013.

Variable Name	Description
id	Identification Code
low	0 = Birthweight $\geq$ 2500g
	1=Birthweight$<$ 2500g
age	Age of mother in years
lwt	Weight in Pounds at last menstrual period
race	1 = white
	2 = black
	3 = other
ptl	History of Premature Labor (0=none, 1= One, ...)
ht	History of hyptertension
ui	Presence of Uterine Irritability
	0 = No

Variable Name	Description
	1 = Yes
ftv	Number of Physician visits during first trimester (0=none, 1=One, . . .)
btw	Birth weight in grams

You can read the data in with the command below.

```
low.weight <- read.table("https://drive.google.com/uc?export=download&id=0B8CsRLdwqzbzMzJyVkt5QkdvVnM",
```

Model Building

1. Your goal will be to build a model to predict birth weight. Begin by using number summaries and graphs to start to explore relationships of variables in this data set and `bwt`.
2. The variables of `low`, `race` and `ui` are categorical variables but they are not yet factors. Code them in R to be factors in the data. Then make sure they have correct level names.
3. Start your model building by looking at simple linear regressions for each of the 8 predictor variables. Display and Examine relevant plots. Summarize the simple linear regression results using a table (hide the intercepts when combining your `tidy()` commands).

```
fit1 <- lm(bwt ~ age, low.weight)
fit2 <- lm(bwt ~ lwt, low.weight)
fit3 <- lm(bwt ~ factor(race), low.weight)
fit4 <- lm(bwt ~ smoke, low.weight)
fit5 <- lm(bwt ~ ptl, low.weight)
fit6 <- lm(bwt ~ ht, low.weight)
fit7 <- lm(bwt ~ ui, low.weight)
fit8 <- lm(bwt ~ ftv, low.weight)
```

4. Comment of the significance of the 8 variables. What variables do you think would best be used in a multiple linear regression?
5. Explore the possibility of interaction between smoking and race. Display a graph that would allow you to explore this and then run a regression with the interaction term. Interpret the results of this model.
6. Build a multiple regression model with what you have found in problems 4 and 5. Do the coefficients change from the simple regressions? Comment on both direction and magnitude changes.
7. Use the plots we have identified to check the model fit.
    a. Are the assumptions of linear regression met by this?
    b. How does this model fit?
    c. Comment on if you see any possible outliers or collinearity.

Part 2: More Advanced Data Cleaning (PHP 2511 Only)

The Data

This data comes from a study which sought to determine if significant sex differences existed between subjects 65 years of age and older, with regard to calcium, phosphorus, and alkaline phosphatase levels. A retrospective chart review of laboratory procedures performed in six different physician practices. The data consisted of 178 subjects representing 92 males and 86 females over the age of 65. Patient data were obtained from the charts housed in a cardiac care center. Subjects, with charts preceding them, were referred by a physician to the cardiac center. The laboratory procedures had been performed previously in the laboratories of the referring physicians. Boyd et al. , 1998.

The data contains:

Variable Name	Description
obsno	Patient observation number
age	Age in years
sex	1=Male, 2=Female
alkphos	Alkaline Phosphatase IU/L
Lab	1=Metpath
	2=Deyor
	3=St. Elizabeth's
	4=CB Rouche
	5=Youngstown Osteopathic Hospital
	6=Horizon
calcium	Calcium mmol/L
phos	Inorganic Phosphorus mmol/L
agegroup	1=65-69
	2=70-74
	3=75-79
	4=80-84
	5=85-89 Years

You can read the data in with the command below.

```
calcium <- read.table("https://drive.google.com/uc?export=download&id=0B8CsRLdwqzbzbW1oTFlwUlRuSmM", hea
```

Cleaning Data

8. The first task of the assignment is to check the validity of the data. Determine if this is a "messy" dataset with variable values that appear incorrect. Attempt to recover the correct values by looking up the true values from the actual data records. Copies of these can be found on bigtable.htm. Be sure to catalogue the problem values in the data and the changes that were made to clean the dataset. Include a paragraph detailing the steps taken to clean the dataset.

9. Compare the mean and standard deviation of age, alkphos, cammol and phosmmol from the messy dataset with the mean and standard deviation from your cleaned dataset. Does cleaning the data make a difference? Explain.