Instructions

- This homework assignment is worth 84 points.
- Please submit a **.ipynb** file to Blackboard.
- **Please strive for clarity and organization.**
- **Due Date: October 27, 2023 by 11:59 pm.**

Exercise 1

(4 points) Which of the following statements is **true** about outliers in linear regression?

(a) Linear regression is sensitive to outliers.

(b) Linear regression is not sensitive to outliers.

(c) It depends.

(d) None of the above.

Exercise 2

(4 points) Suppose you create a scatter-plot between the residuals and the fitted values in a linear regression task, and you found that there is a relationship between them. Which of the following conclusion do you make about this situation?

(a) Since the there is a relationship the assumption of linearity is not met. So the model is not good.

(b) Since the there is a relationship the assumption of linearity is met. So the model is good.

(c) Since the there is a relationship the assumption of constant variance is not met. So the model is not good.

(d) Since the there is a relationship the assumption of constant variance is met. So the model is good.

(e) None of the above.

Exercise 3

Suppose we have a data set with three predictors, X_1 = GPA, X_2 = IQ, X_3 = Gender (1 for female and 0 for male). The response variable is the starting salary after graduation (in thousands of dollars). Suppose we use the least squares to fit the liner model and we obtain the following:

$$\hat{\beta}_0 = 50, \quad \hat{\beta}_1 = 20, \quad \hat{\beta}_2 = 0.07, \quad \hat{\beta}_3 = -5$$

(a) (4 points) For a fixed value of GPA and IQ, Do males earn more money than females? Explain.

(b) (3 points) Predict the salary of a female with GPA of 4.0 and IQ of 110.

(c) (3 points) Predict the salary of a male with GPA of 4.0 and IQ of 110.

Exercise 4

(4 points) The least squares regression line minimizes the sum of the

(a) Differences between actual and predicted Y values.

(b) Absolute deviations between actual and predicted Y values.

(c) Absolute deviations between actual and predicted X values.

(d) Squared differences between actual and predicted Y values.

(e) Squared differences between actual and predicted X values.

Exercise 5

Consider the `respiratory.csv` datafile. Researchers are interested in the differences in the respiratory rate for young children. With a goal to develop a model to predict the respiratory rate based on the age of the young child the researchers looked at a random sample of over 600 children in the US all between the ages of 0.1 months and 36 months. For your reference the researchers used the following definition of the respiratory rate: Number of breaths taken within one minute. This type of problem is of specific interest to convey typical expectations of the relationships between age and respiratory rate to medical practitioners. Unusual results from a young child may help doctors diagnose a chronic diseases such as asthma much earlier in development. **In Python**, answer the following:

(a) (3 points) Using the `pandas` library, read the csv file and create a data-frame called `respiratory`.

(b) (4 points) Create a scatter-plot of the data (Age in Months on the x-axis and the Respiratory Rate on the y-axis). Describe the relationship between age and respiratory rate based on this scatterplot.

(c) (5 points) Let x_i represents the Age in months of children i and y_i represents Respiratory Rate of children i. Consider the following simple linear regression model:

$$y_i = \beta_0 + \beta_1 x_i \quad \text{for } i = 1, 2, \ldots, 618$$

Find the least square estimates of β_0 and β_1.

(d) (5 points) Using the appropriate plot, check the linearity assumption. Comment on the plot.

(e) (5 points) We now consider a log-linear model for the data set. That is,

$$\log(y_i) = \beta_0 + \beta_1 x_i \quad \text{for } i = 1, 2, \ldots, 618$$

Find the least square estimates of β_0 and β_1. Notice that you can compute the log of a number using the `log()` function from the `numpy` library.

(f) (4 points) Using the estimated model of part (e), what is the predicted respiratory rate for a child who is 5 months old?

Exercise 6

(6 points) A study was designed to compare Red Bull energy drink commercials. Each participant was shown the commercials, A and B, in random order and asked to select the better one. There were 140 women and 130 men who participated in the study. Commercial A was selected by 65 women and by 67 men. Find the odds of selecting Commercial A for the men. Do the same for women.

Exercise 7

(4 points) Select the correct statement. In binary (0-1) logistic regression:

(a) The target variable is continuous. That is, the target variable could be any number from a range of values.

(b) The target variable is divided into two equal subcategories.

(c) The target variable takes the values of 0 or 1.

(d) There is no target variable.

(e) All of the above.

(f) None of the above.

Exercise 8

Consider the `Default.csv` datafile presented in class. In this data set, we are interested in predicting whether an individual will default on his or her credit card payment, on the basis of annual income and monthly credit balance. **In Python**, answer the following:

(a) (3 points) Using the `pandas` library, read the csv file and called `credit_default`.

(b) (4 points) Using the `where` function from the `numpy` library, create a new variables in the `credit_default` data-frame called `default_numb` that takes the value of 1 when `default = Yes` and 0 when `default = No`.

(c) (5 points) Using the `statsmodels.formula.api`, build a logistic regression model in which `default_numb` is the target variable, `balance` and `income` are input variables.

(d) (4 points) Using the model from part (c), predict the likelihood of default of a customer with `balance` = $\$2,000$ and `income` = $\$70,000$.

(e) (5 points) Using the `statsmodels.formula.api`, build a logistic regression model in which `default_numb` is the target variable, `balance`, `income` and `student` are input variables. Notice that when the input variable is categorical, you need to included with `C()` in the model. That is, you need to enter `C(student)` as input in the logistic model.

(f) (5 points) Using the model from part (e), predict the likelihood of default of a customer with `balance` = $\$2,500$ and `income` = $\$20,000$ and `student = Yes`.