DUE NOV 15TH 2024, FRIDAY, 11:59PM. SEE THE SUBMISSION INSTRUCTIONS ON CANVAS.

(1) (6 points) Suppose you are given a dataset where the response variable Y is modeled by a simple linear regression of the form:

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i,$$

where X_i is the predictor, and ϵ_i is the error term.

In each of the scenarios below, the variance of the error term ϵ_i is non-constant (heteroscedastic). For each case, determine the appropriate weight w_i to assign if Weighted Least Squares (WLS) is used.

(a) **Quadratic Variance in X:** The variance of the error term is proportional to X_i^2:

$$\mathrm{Var}(\epsilon_i) = \sigma^2 X_i^2.$$

What weight w_i should you use?

(b) **Linear Variance in X:** The variance of the error term is proportional to X_i:

$$\mathrm{Var}(\epsilon_i) = \sigma^2 X_i.$$

What weight w_i should you use?

(c) **Inverse Variance in X:** The variance of the error term is inversely proportional to X_i:

$$\mathrm{Var}(\epsilon_i) = \frac{\sigma^2}{X_i}.$$

What weight w_i should you use?

(d) **Exponential Variance in X:** The variance of the error term grows exponentially with X_i:

$$\mathrm{Var}(\epsilon_i) = \sigma^2 e^{X_i}.$$

What weight w_i should you use?

(e) **Square Root Variance in X:** The variance of the error term is proportional to the square root of X_i:

$$\mathrm{Var}(\epsilon_i) = \sigma^2 \sqrt{X_i}.$$

What weight w_i should you use?

(2) (6 points) You need to upload the dataset file "HW9Data1.csv" into R to solve this problem.

(a) Use the provided R code below to compute Cook's Distance values for each data point.

```
# Load dataset
data <- read.csv("HW9Data1.csv")

# Fit a linear regression model
model <- lm(y ~ x1 + x2 + x3, data = data)

# Calculate Cook's Distance
cooksD <- cooks.distance(model)

# Identify influential points
threshold <- 4 / nrow(data)
influential_points <- which(cooksD > threshold)

# Output results
influential_points
cooksD[influential_points]
```

(b) According to standard criteria, a data point with a Cook's Distance greater than

$$\frac{4}{n}$$

(where n is the total number of observations) is considered influential. Use this threshold to identify influential points. Which data points are identified as influential based on Cook's Distance?

(c) Set up a second regression model with the influential points removed. Compute the RSE for the two models, what conclusion can you make?

3. (6 points)

   (a) Load the dataset `HW9Data2.csv` in R and use the VIF (Variance Inflation Factor) method to identify which variables are highly collinear.

   (b) Interpret the VIF values to determine which variables might be problematic.

````
```{r}

# Load necessary library
library(car)

# Load the dataset
HW9Data2 <- read.csv("HW9Data2.csv")

# Fit a linear model
model <- lm(y ~ x1 + x2 + x3 + x4 + x5, data =
HW9Data2)

# Calculate VIF values
vif_values <- vif(model)
print(vif_values)

```
````

Note: Below is the interpretation of the VIF values.

- $VIF = 1$: There is no correlation between the predictor and the other variables, indicating no multicollinearity.
- $1 < VIF < 5$: Moderate correlation; multicollinearity is not a severe issue.
- $VIF \geq 5$: High correlation, suggesting the presence of multicollinearity. The variable may be problematic and its inclusion in the model could lead to inflated standard errors and unstable coefficient estimates.
- $VIF \geq 10$: This typically indicates severe multicollinearity. In such cases, the predictor is highly correlated with one or more other variables, and it's advisable to reconsider its inclusion or apply remedial measures.

4. (6 points) After identifying multicollinearity in the model using VIF in Problem 3, examine the output of `summary(model)` and interpret what it reveals about the model. Specifically:

   - Which types of information in the `summary()` output may be unreliable due to multicollinearity?
   - How might multicollinearity impact the coefficients, p-values, and overall interpretability of the model?

5. (6 points)

(a) Suppose Y denote whether a student get admitted to attend Illinois Tech. While

- X_1 denote the math scores (0-100)
- X_2 denote the reading scores (0-100)

of a particular standardize exam. Suppose we collected n-data and run a logistic regression to obtain

$$\ln\left(\frac{\pi}{1-\pi}\right) = 0.13 + 0.0456X_1 + 0.032X_2$$

where

$$\pi = \Pr(Y = 1|X_1 = x_1, X_2 = x_2).$$

Holding the reading scores at a fixed value, show that the odds of being admitted to Illinois Tech is 4.67% higher for a one-unit increase in the math scores. That is show that

$$\frac{\text{Odds}(x_0+1) - \text{Odds}(x_0)}{\text{Odds}(x_0)} \times 100\% = 4.67\%$$

(b) In Part (a), the logistic regression model is used to predict the probability of a student being admitted to Illinois Tech ($Y = 1$) based on their math (X_1) and reading (X_2) scores.

Why is logistic regression particularly suitable for modeling this type of outcome, where the dependent variable is binary (i.e., a student is either admitted or not admitted)? Specifically, explain how logistic regression addresses the challenges posed by modeling probabilities that must lie between 0 and 1.