Homework 9 A20545853

QC1) 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.

This question in the simple Linear Regression is hetroscedasticity. For each case, determine the appropriate weight w_i to assign if Weighted Least squares (WLS) is used.

The WLS is to assign the weights of observation so the constant variance will be for the transformed errors.

For SLR, the variance of errors is $var(\varepsilon_i) = \sigma^2 f(x_i)$

and weight $w_i = \dfrac{1}{var(\varepsilon_i)}$

So, we can effectively normolize the influence of each observation on bases of the error variance.

(a) Quadratic Variance in X :

The variance of error term is proportional to.

$$var(\varepsilon_i) = \sigma^2 x_i^2$$

The weight can be shown as

$$w_i = \frac{1}{var(\varepsilon_i)} = \frac{1}{\sigma^2 x_i^2}$$

→ weights is inversely proportional to the square of x_i.

(b) Linear Variance in X :

The variance of the error term is proportional to x_i

$$var(\varepsilon_i) = \sigma^2 x_i$$

The weight can be shown as $w_i = \frac{1}{var(\varepsilon_i)} = \frac{1}{\sigma^2 x_i}$

→ weight are inversely proportional to x_i

(c) Inverse Variance in X :

The variance of the error term is inversely proportional to x_i

$$var(\varepsilon_i) = \frac{\sigma^2}{x_i}$$

The weight can be shown as $w_i = \frac{1}{var(\varepsilon_i)} = \frac{1}{\sigma^2/x_i} = \frac{x_i}{\sigma^2}$

→ here, weights are directly proportional to x_i,

(d) Exponential Variance in X :

The variance of the error term grows exponentially with x_i

$$var(\varepsilon_i) = \sigma^2 e^{x_i}$$

The weight can be shown as $w_i = \frac{1}{var(\varepsilon_i)} = \frac{1}{\sigma^2 e^{x_i}}$

→ here, the weight is inversely proportional to exponential function of x_i

(e) Square Root Variance in X :

The variance of the error term is proportional to the square root of X_i $var(\varepsilon_i) = \sigma^2 \sqrt{x_i}$

The weight can be shown as $w_i = \frac{1}{var(\varepsilon_i)} = \frac{1}{\sigma^2 \sqrt{x_i}}$

→ here, weight are inversely proportional to the square root of x_i

Q(4) 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.

※ The following are the types of information in the summary() output may be unreliable due to multicollinearity.

Coefficient : the estimated coefficient for predictor with high multicollinearity is unstable, where the small change in the data is leading to the coefficient value.

Standard error : the high multicollinearity means the standard error where making the it harder to determine the significance of predictor which means the inflation is leading to wider confidence interval.

P value : Due to the inflated standard errors, the p values for the predictors is Misleading. the high p-value could incorrectly suggest a predictors is not significant, which might be important.

* The following might multicollinearity impact the coefficients, p-values, and overall interpretability of the model

coefficient : the coefficient may not say the true relationship between the predictors & the response variables. this means that the small changes in the data, leading to poor model interpretability.

p-value : the inflated standard errors from multicollinearity may result in having the high p-value, which may be causing of the mistakenly concluding that a variable is not statistically significant when it is this will make to ignore the important variables from the model.

overall : the multicollinearity makes it difficult to asses the individual, and contribution of each predictor to the response variable which predictors is highly correlated, which becomes the challenging to determine which variable is truly driving changes in the outcome reducing the clarity & usefulness of the model.

Q(5) (a) Given that, suppose Y denote whether a student 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

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

where, $\pi = \Pr(Y=1 \mid X_1 = x_1, X_2 = x_2)$

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

Show that

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

→ Odds of $X_1 = x_0$

The odds of admission of $X_1 = x_0$ & $X_2 = x_2$ is

$$\text{Odds}(x_0) = \frac{\pi}{1-\pi} = e^{0.13 + 0.0456 x_1 + 0.032 x_2}$$

→ Odds of $x_1 = x_0 + 1$

The odds of admission of $x_1 = x_0 + 1$ & $x_2 = x_2$ is

$$\text{Odds}(x_0) = e^{0.13 + 0.0456(x_0+1) + 0.032x_2}$$

$$= e^{0.13 + 0.0456x_0 + 0.0456 + 0.032x_2}$$

$$= e^{0.1756 + 0.0456x_0 + 0.032x_2}$$

The change in order of increasing x_1 by 1 is:

$$\text{odd}(x_0+1) - \text{odd}(x_0) = e^{0.1756 + 0.0456x_0 + 0.032x_2} - e^{0.13 + 0.0456x_0 + 0.032x_2}$$

factoring the common term $e^{0.0456x_0 + 0.032x_2}$

$$\text{odd}(x_0+1) - \text{odd}(x_0) = e^{0.0456x_0 + 0.032x_2}\left[e^{0.1756} - e^{0.13}\right]$$

→ The percentage of increase in odds is

$$\frac{\text{odd}(x_0+1) - \text{odd}(x_0)}{\text{odd}(x_0)} \times 100\% = \left(\frac{e^{0.1756} - e^{0.13}}{e^{0.13}}\right) \times 100\%$$

$$\left(e^{0.1753 - 0.13} - 1\right) \times 100\% = \left(e^{0.0456} - 1\right) \times 100\%$$

using the approximation $e^x \approx 1 + x$ for small x

$$e^{0.0456} \approx 1 + 0.0456 \Rightarrow e^{0.0456} - 1 \approx 0.0456$$

$$(0.0456) \times 100\% \approx 4.56\%$$

Since, 4.56% is nearly to 4.67%, where we can say that the odds increase by approximately 4.67% with increase of one unit in math scores.

(b) Logistic regression is particularly suitable for binary outcomes for several reasons.

Probability Range: Logistic regression, changes to the linear which is combination of predictors through the logistic function, making sure that all the predicted probabillities are from 0 to 1. The logical function $\sigma(z) = \frac{1}{1+e^{-z}}$ which maps to a real value number z to the interval $(0,1)$ which is crutial from probabilities cannot be more than boundaries

odds interpretation: Logistic regression which gives the straightforward way to intercept coefficients in terms of all odds. where each coefficient represent the log odds of the outcome with occurs to allow us to understand the influence of predictor variables on the likelihood of an event.

Non Linear Relationship: Logistic regression can accommmodation for the linear relationship between predictors & the log odds of the dependents variable, which is valueable because the predictors are often having a non linear relationship with the probability of an outcome like admission.

Max likelihood estimation: Logistic regression relies on max likelihood estimation to identify the best fitting model. This estimation is well suited for binary outcome variables, providing efficient & unbiased estimates.