**STAT 3502 – Elle Nguyen**

**Question 1:**

A screenshot of a computer code

Description automatically generated

a. Do the predictors have significant effects on the response? Comment on the signs of estimated regression coefficients.

A screenshot of a computer

Description automatically generated

🡪 All three predictors have relatively small p-values compared to the alpha level of 0.05, suggesting that all predictors have significant effects on the response.

🡪 : x1 has a positive estimated regression coefficient, meaning that x1 has a positive effect on y and for every 1-unit increase of x1, y will increase by a value of 2.0164 while holding x2 and x3 constant.

🡪 : x2 has a negative estimated regression coefficient, meaning that x2 has a negative effect on y and for every 1-unit increase of x2, y will decrease by a value of 6.3309 while holding x1 and x3 constant.

🡪 : x3 has a positive estimated regression coefficient, meaning that x3 has a positive effect on y and for every 1-unit increase of x3, y will increase by a value of 2.4702 while holding x1 and x2 constant.

b. The first column of the data set denotes the week of the year. Plot the residual from the model above v.s. the variable week. Do you think we should add this variable to the model? Do you expect week to have a positive or a negative effect on the cases sold of 18-packs?



A graph with dots and lines

Description automatically generated

🡪 There is a moderate linear trend within the plot between the residuals and the predictor Week that is currently not in the model. It appears that adding the predictor Week to the model might help to explain some of the remaining variability in the response.

🡪 I expect the predictor Week to have a positive effect on the cases sold of 18-packs due to the fact that as week increases, the residuals also increase which suggests that higher weeks are associated with larger observed values for the response variable.

c. Denote week as x4. Fit a multiple linear regression with y as the response and x1, x2, x3, x4 as the predictors. Perform a formal test to see if x4 has a positive effect on y.

A white background with green text

Description automatically generated

A screenshot of a computer

Description automatically generated

🡪 Adding the predictor Week as x4 helps increase R-squared value from 0.9291 to 0.9465, which supports our previous conjecture that adding such variable helps explaining some of the remaining variability in the response. Its small p-value (0.000174) also suggests that the predictor Week is a statistically significant in our new multiple linear regression model.

A screenshot of a computer

Description automatically generatedA table with numbers and symbols

Description automatically generated

🡪 By conducting an ANOVA test, we can reject the null hypothesis (p-value = 0.0001737) and conclude that x4 has a significant contribution on the response. To test whether x4 has a positive effect on y, we fit a simple linear regression model with y as the response and x4 as the only predictor. The fitted model obtains a positive estimated regression coefficient for x4 (0.028583), meaning that x4 has a positive effect on y.

**Question 2:**

a. Draw the scatter plot between mpg (miles per gallon) v.s. weight (in pounds). Do you think linear regression with mpg as the response and weight as the predictor will be a good fit? Why or why not?

A graph of weight and weight

Description automatically generated with medium confidence

🡪 The scatterplot indicates a slight exponential trend between two variables mpg and weight, hence building a linear regression model using mpg as the response and weight as the predictor is not effective. It is essential to consider applying a log transformations before building an appropriate model.

b. Create a new variable y (gallons per 100 miles) and another new variable x1 (weight in 1000 pounds). Draw the scatter plot between y and x1. Do you think linear regression will be a good fit?

A white background with green text

Description automatically generated

A graph of dots and numbers

Description automatically generated with medium confidence

🡪 This scatterplot also displays a slight exponential trend between x1 and y instead of a linear trend, suggesting that linear regression at this point is not a good fit. As previously explained, it is beneficial to consider a log transformation.

c. Create a new variable x2 that corresponds to the model year of the auto-mobile. Fit a multiple linear regression with y as the response, x1 and x2 as the predictors. Write the fitted regression model, comment on the signs of the estimated regression coefficients, and explain the meanings of ˆβ1 and ˆβ2.

A screenshot of a computer program

Description automatically generated

🡪 Fitted regression model:

🡪 : x1 has a positive estimated regression coefficient, meaning that x1 has a positive effect on y and for every 1-unit increase of x1, y will increase by a value of 1.544516 while holding x2 constant.

🡪 : x2 has a negative estimated regression coefficient, meaning that x2 has a negative effect on y and for every 1-unit increase of x2, y will decrease by a value of 0.141650 while holding x1 constant.

d. Create a new variable ̃y (gallons per mile). Fit a multiple linear regression with ̃y as the response, x1 and x2 as the predictors. Compare the R2 of this model with the model in part c). Then compare the estimated regression coefficients of this model with the model in part c).

A screenshot of a computer program

Description automatically generated

🡪 Both models have the same R-squared value of 0.8733 but different estimated regression coefficients with the model in part d) having much smaller estimated regression coefficients. was previously 1.544516 from the model in part c) and is now 0.01544516 from the model in part d) and similarly, was previously -0.141650 from the model in part c) and is now -0.001417 from the model in part d).

🡪 While both models have the same performance, this difference occurs due to the fact that we use gallons per 100 miles for the first model, which is 100 times gallons per mile for the second model. Scaling down the response range by 100 times results in scaling down the coefficients by 100 times.

**Question 3:**

A close-up of a number

Description automatically generated

a. What transformations should we use before we fit a linear regression model? Implement the transformation(s) needed and then fit a linear regression model. Comment on the estimated regression coefficient. Here length should be the length from the nose to the end of the tail.

A computer code with text

Description automatically generated with medium confidenceA screenshot of a computer code

Description automatically generated

🡪 We should use the log transformation to both weight and length before fitting a linear regression model (as shown in the code block above).

🡪 Here, the transformed variable of length has a positive effect on the transformed variable of weight, meaning that for every 1-unit increase in length, weight will increase by a value of 3.17242.

b. Create an indicator variable z such that z = 1 if the fish species is smelt or pike, and z = 0 otherwise. Fit a multiple linear regression model with z as an additional predictor to the simple linear regression model in part a). Formally test if z has a significant effect on the response.

A screenshot of a computer code

Description automatically generated

🡪 By adding z as a second predictor, the R-squared value increased from 0.9458 to 0.9919 suggesting that the variable z is statistically significant in the new model.

A computer screen shot of a table

Description automatically generated

🡪 By conducting ANOVA test, we obtain a relatively small p-value, indicating that z has a significant effect on the response when being added to the model as a predictor.

c. Denote the continuous predictor in part a) as x. Write the population level interaction model where we have the interaction between x and z. Do you expect the interaction effect to be significant? Why or why not? Perform an analysis to formally test your conjecture.

A screenshot of a computer

Description automatically generated

🡪 Population level interaction model:

with z = 1 as the baseline category

A black text on a white background

Description automatically generated A comparison of a graph

Description automatically generated with medium confidence

🡪 By conducting ANOVA test, adding the interaction between x and z to the model does not increase the R-square value (remained as 0.9919) so I expect the interaction effect to be insignificant. This aligns with a large p-value of 0.4545, which is greater than the alpha of 0.05.

🡪 This conclusion is also supported by the plots between the residuals v.s. fitted response of two models with/without the interaction term. There is no update with the model on the right when the interaction term is added.