Jonathan Lindahl
Student ID: 2178005
I have completed this work independently. The solutions given are entirely my own work.
1/25/2024
DSC 423 Homework 1

Question 1

A. I would say that this is a good regression model but there is still more information I would need. The closer R-squared is to 1, the better the model. With an R-squared of 0.69, that means that 69% of the variance in our dependent variable can be explained by the model.

I would still need other information such as the p-values for coefficients to make a final conclusion on whether this is a good model or not. I would also want to look at the independent variables to see if they are significant or not, and make sure the data set is large enough.

B. The “Regression Fallacy” is the mistaken belief that a regression model’s results imply a causal relationship between independent and independent variables. This fallacy is caused by overlooked variables or may be due to chance.

An example of this is people who wear sunscreen and cases of skin cancer. A regression analysis could show a significant correlation between these two variables, making people think that sunscreen causes skin cancer.

In reality, though, there are other variables to consider such as sun exposure. Sun exposure is the real cause of skin cancer, and people who go outside more often are more likely to wear sunscreen.

People who don’t go outside won’t use sunscreen as often, but also won’t have as much sun exposure either. People could also be applying sunscreen wrong or not enough which would also impact their chance of getting skin cancer.

Question 2

I would say that the smaller standard deviation between the two would be the 100 graduate students studying Data Science. Data Science is a technical degree which requires students to have knowledge in Python, R, Linear Algebra, Calculus, Statistics, Machine Learning and other related subjects.

Although the programs may vary slightly, the material would be very similar, making the costs similar across the 100 students, leading to a smaller standard deviation.

On the other hand, 100 graduate students at Depaul University would vary greatly. A Law degree requires alot of reading, and the textbooks used are big and usually expensive.

Comparing this to someone studying a performance art such as dance, or acting, this student would not have to buy nearly as many textbooks, which would make the standard deviation far greater.

Question 3

A. The 68-95-99.7 rule means that 68% of the students will fall between 1 standard deviation of the mean age of 28. Since the standard deviation is 4, we subtract 4 from 28 and add 4 to 28 to get out range. 28 - 4 is 24, and 28 + 4 is 32, meaning approximately 68% of the students ages will fall between ages 24 to 32.

B. As for the 95-99.7 part of the empirical rules, this means that 95% of students will fall between 2 standard deviations of the mean age of 28. Since the standar deviation is 4, we multiply 4 by 2 and get 8. We then subtract 8 from 28 which is 20, and add 8 to 28 which is 36.

So 95% of students with fall between the ages of 20 and 36. Since we are looking for the percentage of students older than 36, we would subtract 100 - 95 which is 5%. But this percentage includes students who are younger than 20, so we need to divide this percentage by 2, which is 2.5%, meaning approximately 2.5% of students are older than 36 years old.

Question 4

First we need to find the Z -score for the 99th percentile, which is 2.33. Next, we need to convert the Z-score to an actual value. Since our standard deviation is 35,000 and our mean is $150,000, we can plug these numbers into the formula Actual Value = (Z-score * Standard Deviation) + Mean.

After we plug these numbers in, we get Actual Vale = (2.33 * 35,000) + 150,000. 2.33 * 35,000 = 81,550. 81,550 + 150,000 is equal to 231,550. This shows that the top 1% monthly sale figure is approximately $231,550.

Question 5

First lets start by defining the Null Hypothesis(H0) and the Alternative Hypthesis(H1). The null Hypothesis is there was no reduction in the mean number of intrusions, which would remain at 45. The Alternative Hypothesis is that the mean number of intrusions was reduced, making it less than 45.

Since we are looking for a decrease in the mean number of intrusions, we will be looking at the left side of the mean, making it a left-tailed test.

Next, we will determine the parameters.
The mean before the change (population mean, μ) = 45 intrusions per day. The sample mean (x̄) after the change = 42 intrusions per day. The sample standard deviation (s) = 15.5. The number of days (sample size, n) = 35.

Next we will calculate the t-statistic by using the formula:
t-statistic = (x̄ - μ) / (s / sqrt(n))
When we plug in the numbers we get (42 - 45) / (15.5 / sqrt(35))
After we perform this calculation we get a t-statistic of approximately -1.145.

Next we calculate the p-value:
First we find the degrees of free(DF) which is N -1, so 35 - 1 is 34. With a t-statistic of -1.145 and a DF of 34, we can use a t table or use the R function pt(t_statistic, df, lower.tail = TRUE) to find the value that corresponds to this. After we plug in the correct numbers p_value <- pt(-1.145, 34, lower.tail = TRUE), the p-value is s approximately 0.130.

Since the p-value of 0.130 is less than the significance level of 0.05 we determine that we do not have enough evidence to reject the null hypothesis. Therefore, we fail to reject the null hypothesis, meaning we cannot conclude that the change in firewall settings significantly reduced the number of intrusions.

Question 6

A.

```
setwd("/Users/jonathanlindahl/Desktop/bears/QUASAR1.txt")
quasar_data <- read.table("/Users/jonathanlindahl/Desktop/bears/QUASAR1.txt", header =
TRUE, sep = "\t")
head(quasar_data)

model_x1 <- lm(Y1 ~ X1, data = quasar_data)
summary(model_x1)
```

```
model_x2 <- lm(Y1 ~ X2, data = quasar_data)
summary(model_x2)

model_x3 <- lm(Y1 ~ X3, data = quasar_data)
summary(model_x3)

model_x4 <- lm(Y1 ~ X4, data = quasar_data)
summary(model_x4)

model_x5 <- lm(Y1 ~ X5, data = quasar_data)
summary(model_x5)
```

B.

Model with Redshift (X1):

While looking at the model with Redshift as our variable, we see an R-squared of just 0.005073. That's really low, meaning Redshift alone isn't telling us much—only about half a percent of the variance in the Rest frame Equivalent Width is explained.

And when you spot a p-value of 0.735 for Redshift, it's clear we don't have a statistically significant predictor on our hands.

Model with Line Flux (X2):

The Line Flux model is a little better with an R-squared of 0.04365, suggesting that 4.365% of the variation in our dependent variable is accounted for.

But when we look at the p-value for Line Flux it is only 0.316, which is not low enough to clear the usual threshold of 0.05 for statistical significance.

Model with Line Luminosity (X3):

Line Luminosity has an R-squared of 0.03611, which is also quite low. And with an adjusted R-squared that's negative, it's stil not looking good.

The p-value here is 0.363, so once again, we don't have a significant relationship between our variables.

Model with AB1450 Magnitude (X4):

As for AB1450 Magnitude we achieve an R-squared of 0.3024 meaning about 30.24% of the Rest frame Equivalent Width's variance is explained, which is the best so far.

The p-value for AB1450 Magnitude is 0.0044, which is below our significance level, this suggests a potentially meaningful relationship here.

Model with Absolute Magnitude (X5):

The Absolute Magnitude model has the highest R-squared of 0.3724, so about 37.24% of the variance in the Rest frame Equivalent Width is explained by this model.

The p-value for Absolute Magnitude is 0.001197, which is very significant statistically.

Best Model Analysis:
If I had to pick one model, it would be the model with Absolute Magnitude(X5). It is the only model with both a high R-squared and a low p-value, making it a clear winner for explaining the Rest frame Equivalent Width of quasars.

We should still consider exploring more complex models or bring in additional variables to further analyze this data.