Madeline Kuehn
Math 634

Reflection #7: Confidence interval, p-value, and multiple linear regression

1. You have run some simulations of sampling distributions and observed how they produced confidence intervals. So, what does 95% confidence interval really mean? After teaching confidence interval in my Statistics class last year, I came across a poll that reports that a 95% confidence interval for the mean ideal weight for adult American women is 139 ± 1.4 pounds. When I asked my students to explain the meaning of this confidence interval, two students said the following:

   Student #1: "95% of all adult American women would say that their ideal weight is between 137.6 and 140.4 pounds."

   Student #2: "We can be 95% confident that future samples of adult American women will say their mean ideal weight is between 137.6 and 140.4 pounds."

   What do you think about the explanations of the two students? Are they correct? Explain.

Student 1 uses the confidence interval to describe individual responses, and Student 2 uses it to describe future samples. Both students have some pieces correct, but neither is exactly correct. What a 95% confidence interval means is that if all possible samples of a given size are taken from a population, 95% of the samples would produce intervals that have captured the true population mean and 5% will not.

So, for this poll, the interval is given as 137.6 to 140.4 pounds. The 95% confidence interval means that we are 95% confident that the interval contains the true mean ideal weight of all adult American women.

2. a) Discuss your understanding of the concept of a p-value in a layperson's language.

The p-value helps us to determine if the result of an experiment is significant. It tells us how likely it is that the results we got could have happened by change.

For example, if we are testing a drug, we assume that the drug has no effect. After we do trials of the drug, we calculate the p-value. If the p-value is low, it means that getting results like we did by chance is very unlikely. This means that our results are significant. If the p-value is high, it means that the results could happen by chance, so the results of our experiment are not significant, so we can't say for sure if the drug works.

   b) Discuss how to find a p-value using the z table.

First, we need to have a z-score.

Find the row in the z-table up to the first decimal and the column with the second decimal. The intersection point is the area from $-\infty$ and z. The value that we find will then help us find our p-value.

Next, we need to know what type of test we are using, a two-tailed test, a right-tailed test, or a left-tailed test.

Left-tailed test:

Madeline Kuehn
Math 634

- We use the cumulative probability given to us by the z-table.

Right-tailed test:

- We find the p-value by taking 1-the answer given to us by the z-table

Two-tailed test:

- Using the area from the z-table, we are given 1 tail. To find the p-value, we take 2 * the cumulative probability from the z-table.

c) Discuss how to find a p-value using the t table.

First, we need to have a t-score. Then we need to find the degrees of freedom.

We find the row matching with the degrees of freedom, and find the value closest to the t-score. We then find the header of the column we are in.

Next, we need to know what type of test we are using, a two-tailed test, a right-tailed test, or a left-tailed test.

Left-tailed test:

- The p-value is given by taking 1 – the alpha value given in the header.

Right-tailed test:

- We find the p-value by taking the alpha value given in the header.

Two-tailed test:

- To find the p-value, we take 2 * the alpha value given in the header.

Part 2: Multiple Linear Regression

The data set in the excel spread sheet named "Reflection #6_Bear Weights" represents the weight of bears. A group of researchers who are interested in the prediction of bear weight, over a period of time sampled 54 bears and recorded information on the following: Length (in inches of body), Chest (distance in inches around the chest), Neck (distance in inches around the neck), Age (in months), Headlen (length in inches, of head), Headwth (width in inches of head) and Sex (1 = male, 2 = female), Weight (measured weight in pounds). Data collected by the group can be accessed by clicking on the attached link Reflection #7_Bear Weight Use any software of your choice for this analysis and use the result of your analysis to answer the following questions.

a) Identify the dependent variable and the independent variables

Because we are trying to predict what might influence the weight of the bears, our dependent variable is weight.

The variables that possibly influence weight are: Age, Sex, Headlen, Headwth, Neck, Length, and Chest. These are the independent variables.

Madeline Kuehn
Math 634

b) Write the regression model that relates the dependent variable to the independent variables.

The regression model that relates the dependent variable (Weight) to the independent variables is:
$Weight = \beta_0 + \beta_1 \cdot Age + \beta_2 \cdot Sex + \beta_3 \cdot Headlen + \beta_4 \cdot Headwth + \beta_5 \cdot Neck + \beta_6 \cdot Length + \beta_7 \cdot Chest + \varepsilon$

c) Carefully interpret the error term in the model.

The error term in the regression model, ε, represents the difference between the actual observed weight of the bear and the weight predicted by the models based on the independent variables.

d) Write the regression equation that expresses the dependent variable with the independent variables.

To find the regression equation, I put the data set into Minitab and ran the regression analysis.

$$\widehat{Weight} = -180.0 + 0.558Age - 13.1Sex - 9.20Headlen - 0.06Headwidth + 5.16Neck + 0.72Length + 9.18Chest$$

e) Verify the regression assumptions

We assume that the error is normally distributed with a mean of zero and a standard deviation of σ_e (The errors are independent of each other). We could validate these assumptions by doing residual analysis.

f) Does the equation seem suitable for predicting the weight of a bear? Justify your answer using only relevant information. Do not cut and paste a computer printout.

First, using Minitab's analysis , the Coefficient of determination, $R^2 = 0.9469 = 94.69\%$. This means that approximately 94.69% of the variance in bear weight is explained by the independent variables.

Additionally, the Adjusted R^2, $R_\alpha^2 = 0.9388 = 93.88\%$. This is slightly lower than the R^2 value because this takes into account the 7 independent variables. However, since it is still high, this equation does seem like an excellent predictor for bear weight.

g) Write the relevant hypotheses to be tested in this problem.

$$H_0: \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = \beta_6 = \beta_7 = 0$$

$$H_a: \beta_1 \neq 0, or\ \beta_2 \neq 0, or\ \beta_3 \neq 0, or\ \beta_4 \neq 0, or\ \beta_5 \neq 0, or\ \beta_6 \neq 0, or\ \beta_7 \neq 0.$$

h) Test the overall significance of the model at α = .05. Interpret the result of this test.

Using the analysis from Minitab, the F-statistic is 117.08 and the associated p-value is 0.000. Because $0.000 < 0.05$, we reject the null hypothesis. This means that the overall regression model is statistically significant. The model is an excellent way to predict the weight of the bear.

i) Which variables seem to be important in the model? Explain.

Looking at the p-values for each of the independent variables will tell us if they are important.

Age: $p - value = 0.015 < 0.05$, which means that this variable is significant.

Madeline Kuehn
Math 634

Sex: $p - value = 0.255 > 0.05$, which means that this variable is not significant.

Headlen: $p - value = 0.106 > 0.05$, which means that this variable is not significant.

Headwth: $p - value = 0.990 > 0.05$, which means that this variable is not significant.

Neck: $p - value = 0.059 > 0.05$, which means that this variable is not significant.

Length: $p - value = 0.533 > 0.05$, which means that this variable is not significant.

Chest: $p - value = 0.000 < 0.05$, which means that this variable is significant.

This means that the variables that are important to the model are **Age and Chest.**

j) Compute 95% confidence intervals for age and chest. Carefully interpret each interval.

The 95% confidence interval is found by the equation $b_i \pm t_{\alpha/2\,,n-k-1} * s_{b_i}$. Because we are dealing with an $\alpha = 0.05$ we first should find the critical value for the t-distribution for 0.025 with 46 degrees of freedom $(n - k - 1 = 54 - 7 - 1 = 46)$. Thus, $t_{0.025,46} = 2.013$

The 95% confidence interval for the variable age:

$$b_i = 0.558, and\ s_{b_i} = 0.222$$

$$0.558 \pm 2.013(0.222)$$

$$(0.111, 1.005)$$

The 95% confidence interval for the variable chest:

$$b_i = 9.18, and\ s_{b_i} = 1.42$$

$$9.18 \pm 2.013(1.42)$$

$$(6.322, 12.038)$$

k) Remove variables that are not important and redo the analysis. Did removal of unimportant variables improve the quality of the model? Explain.

After removing the variables, the regression equation:

$$\widehat{Weight} = -242.3 + 0.460Age + 11.361Chest$$

For the new equation, Coefficient of determination, $R^2 = 0.9356 = 93.56\%$. This means that approximately 93.56% of the variance in bear weight is explained by the independent variables.

Additionally, the Adjusted R^2, $R_a^2 = 0.9331 = 93.31\%$. This is slightly lower than the R^2 value because this takes into account the 2 independent variables.

Removing the other independent variables lowered the R_a^2 slightly, but overall, it did not really affect the quality of the model. However, it is easier to use, because it only has two variables to input.