**DSCI 5180: INTRODUCTION TO BUSINESS DECISION PROCESS**

**FINAL PROJECT**

Name: Kiran Kumar Gontyala

Student ID: 11718445

The Main objective of this project is to apply the Statistical models I’ve acquired in my course work to practical situations. In this project I will be actively engaging with the concepts like Normal distribution, Mean estimation and Standard deviation, Confidence interval and the Hypothesis Testing. The dataset I have used for the project is a secondary dataset, meaning it was not collected by the current user but was previously obtained from Kaggle. please use the provided link to access this dataset on Kaggle's website.

Link: <https://www.kaggle.com/datasets/elikplim/forest-fires-data-set>

**Description of the Data Set:**

The Data set I have used for the project is Forest Fires dataset of Montesinho park, Darlington. It is used to predict the burned area of forest fires using meteorological and other data. It consists of 13 variables and 517 Data points. I will be using the Temperature as working variable for first 3 questions & at the end I will be performing the Multiple regression Analysis.

|  |  |
| --- | --- |
| S.NO | VARIABLES & DESCRIPTION |
| 1. | X - x-axis spatial coordinate within the Montesinho park map: 1 to 9 |
| 2. | Y - y-axis spatial coordinate within the Montesinho park map: 2 to 9 |
| 3. | month - month of the year: "jan" to "dec" |
| 4. | day - day of the week: "mon" to "sun" |
| 5. | FFMC – FFMC (Fine Fuel Moisture Code) index from the FWI (Fire Weather Index) system: 18.7 to 96.20 |
| 6. | DMC – DMC (Duff Moisture Code) index from the FWI system: 1.1 to 291.3 |
| 7. | DC – DC (Drought Code) index from the FWI system: 7.9 to 860.6 |
| 8. | ISI - ISI (Initial Spread index) from the FWI system: 0.0 to 56.10 |
| 9. | temp - temperature in Celsius degrees: 2.2 to 33.30 |
| 10. | RH - Relative humidity in %: 15.0 to 100 |
| 11. | wind - wind speed in km/h: 0.40 to 9.40 |
| 12. | rain - outside rain in mm/m2: 0.0 to 6.4 |
| 13. | area - the burned area of the forest (in hectares): 0.00 to 1090.84 (this output variable is very skewed towards 0.0, thus it may make sense to model with the logarithm transform). |

Link of the Dataset (including Analysis) I have used:



**Question-1 (Module-1):**

For the Population of Forest fires data of Montesinho park, where the Temperature is normally distributed with the mean Temperature of 18.8892 degrees Celsius and Standard deviation of 5.8066 degree Celsius. What is the Probability that a randomly selected Temperature is more than 22 degrees Celsius. Round your answer to four decimal places.

**Answer:**

μ = 18.8892; σ= 5.8066; X= 22

To find out the probability:

Z = X – μ/σ = (22-18.8892) /5.8066 = 0.5357= 0.54

Probability (>22) = 1-0.7054 = 0.2946

The Probability of randomly selected temperature more than 22 degree Celsius is 0.2946.

**Question-2 (Module-2):**

Assuming the Temperature of Forest fires data of Montesinho park is Normally distributed with population standard deviation of 5.8066 degree Celsius. Construct the 95% Confidence interval for the Mean Temperature of Forest Fires Data, based on the sample of 50 temperature values. The sample Mean Temperature is 18.1041 degree Celsius, and sample standard deviation is 5.1138 degree Celsius.

**Step-1 of 2:** Find theCritical value that should be used in constructing the confidence interval for the mean. Round your answer to three decimal places.

**Step-2 of 2:** Construct the 95% Confidence interval for the mean. Round your answer to four decimal places.

**Answer**:

X= 18.1041; σ= 5.8066; n=50

Since σ is known we use Z-value. Z-value for the confidence interval of 95% is 1.96.

The margin of Error:

E = Zα/2\*(σ/√n) = 1.96\*(5.8066/√50) = 1.609

The formula for the The confidence interval is:

Lower endpoint = x ̅− E = 18.1041-1.609= 16.4951

Upper Endpoint = x ̅+ E = 18.1041+1.609= 19.7131

Rounded to four decimal places, the mean values for 95% Confidence interval are 16.4951& 19.7131.

**Question-3: (Module-3):**

According to the results, the average Temperature of the Forest fires data of Montesinho park is 18.8892 degrees Celsius. But the Forest officer believes that the average Temperature of the Montesinho park is lesser than the noted value. To test his claim, he selected the 64 Temperature variables randomly with the sample mean of 18.5547 degree Celsius and assumed the sample standard deviation as 4.7110 degree Celsius. Is there enough evidence to test the Forest officer’s claim at the significance level of 0.05.

**Step-1 of 2:** State the Null and Alternative Hypothesis for the above Criteria.

**Step-2 of 2:** Perform the Hypothesis Testing and give the hypothesis decision whether to reject or fail to reject the Null Hypothesis.

**Answer:**

Hο =18.8892--- Null Hypothesis

Hɑ <18.8892 ----- Alternate Hypothesis

μ = 18.8892; n= 64; s= 4.7110; α = 0.05; degrees of freedom(df) = n-1 = 64-1 = 63.

t-Critical value= -1.6694

By substituting the values, we get

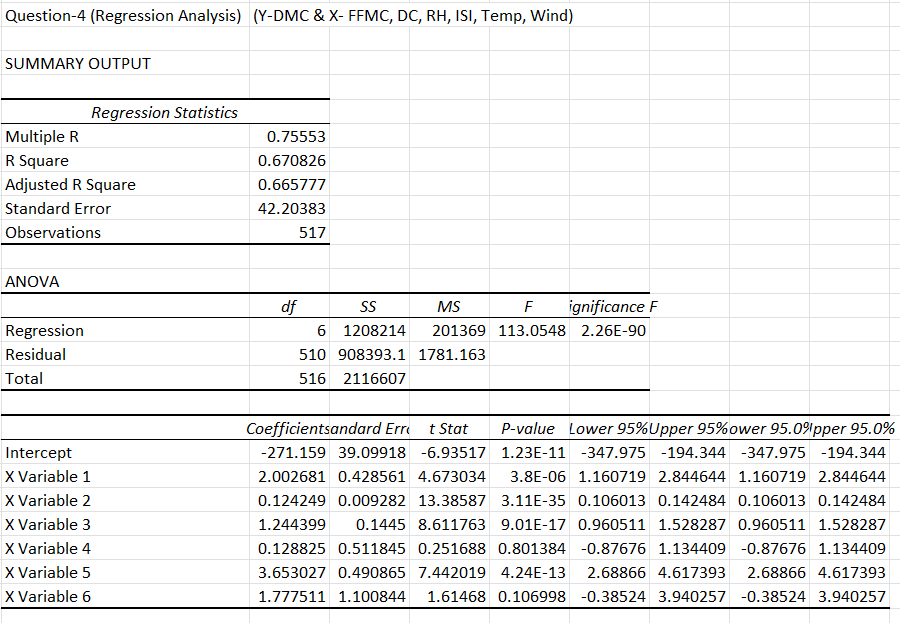
t = (x̄-µ)/(s/√n) = (18.5547- 18.8892)/ (4.7110/√64) = -0.3345/0.5888= -0.5681

Since it is left-sided test, t-value> t-critical value (t-value is not in the rejection region).

Therefore, we fail to reject the null hypothesis. There isn't enough evidence at a 95% confidence level to support the claim that the average temperature of Montesinho park is less than the 18.8892 degree Celsius.

**Question-4: (Module-4):**

Analyse the dataset and predict the Duff Moisture Code (DMC) as a function of FFMC, DC, RH, ISI, Temp & Wind using regression analysis and write down the regression equation. Also, find the solution for the following questions.



**Answer:**

Multiple Linear Regression Equation:

Duff Moisture Code (DMC)= -271.159+2.0027×(FFMC)+0.1242×(Drought Code)+1.2444×(Relative Humidity)+0.1288×(Initial Spread Index)+3.6530×(Temperature)+1.7775×(Wind)

**1.What is the coefficient of determination for this model, R^2? Round your answer to four decimal places.**

**Answer**: 0.6708

**2. What is the adjusted coefficient of determination for this model, Ra^2? Round your answer to four decimal places.**

**Answer:** 0.6658

**3.How many independent variables are used in this model?**

**Answer**: 6

**4.Which statistic is most appropriate for the above scenario to determine the usefulness of the regression model and why?**

**Answer**: Ra^2 since the model is a multiple linear regression model.

**5. What is the estimated variance of error?**

**Answer:** 42.2^2= 1780.84

**6.What is the value of Dependent variable (Duff Moisture Code)) when FFMC, DC, RH, ISI, Temp, Wind are 90, 520, 65, 12, 14 & 3?**

**Answer**:

Duff Moisture Code (DMC)

= -271.159+2.0027×(90)+0.1242×(520)+1.2444×(65)+0.1288×(12)+3.6530×(14)+1.7775×(3)

= 112.5741

Certainly! In a dataset of 517 observations, a multiple linear regression analysis was conducted & the model explains about 67.08% of the variability in the dependent variable, Duff Moisture Code. The adjusted R-square, which accounts for the number of predictors, is 66.58%, indicating a good reasonable fit of the model. The coefficients for the independent variables (FFMC, DC, RH, ISI, Temp, Wind) signify their respective contributions to predicting the Duff Moisture Code.