Introduction to Business Decision Process (5180)

Final Project

Introduction to Business Decision Process (5180)

Final Project

Introduction to Business Decision Process (5180)

Final Project

**Introduction to Business Decision Process**

**Final Project**

By Jeevan Deep Borugadda

11740913

***Introduction to the dataset:***

Life Expectancy (WHO):

The World Health Organization (WHO) maintains a database called the Global Health Observatory (GHO), which tracks health information for all countries. This data is publicly available for health analysis. Specifically, for this project, we've gathered data on life expectancy for 193 countries from the WHO repository, and economic data from the United Nations. Our focus, however, is on 10 Asian countries: Afghanistan, Bangladesh, Bhutan, India, Indonesia, Iran, Iraq, Kazakhstan, Myanmar, and Nepal. We're analyzing data from 2000 to 2015 for these countries, resulting in a final dataset of 22 columns and 161 rows.

Data Source:

[Life Expectancy (WHO)](https://www.kaggle.com/datasets/kumarajarshi/life-expectancy-who/data) - [LINK](https://www.kaggle.com/datasets/lashagoch/life-expectancy-who-updated)

# ***Applying lessons learnt from:***

## ***Module 1:***

*Question:*

The mean lifespan of population in 10 Asian countries from 2000-2015 is 66.8 years with a standard deviation of 4.60, then find the probability that the sample mean for countries Iran, Iraq, Kazakhstan and Nepal combined would be less than 66 years?

*Solution:*

Given μ = 66.8, sd = 4.60, x bar = 66, n = 64.

* p(z<66) = (65-66.8)/(4.60/sqrt(64))
* p(z<66) = -1.39
* z = 0.0820

Therefore, the probability that the sample mean would be less than 66 years is 0.0820 (8.20%)

***Module 2:***

***Question:***

**With reference to the data given in Module 1, construct a confidence interval of 97% for the mean lifespan of people for the same 4 countries.**

***Solution:***

Given x bar = 66, sd = 4.60, n = 64, Alpha = 0.03, Critical Value = 2.17

Formula = x bar ± E, where E is the error calculated by z alpha/2 \* (sd/sqrt n)

* 66 ± 1.25

Therefore, Lower Limit = 64.75. Upper Limit = 67.25

***Module 3:***

*Question:*

The researcher set claims the mean life expectancy of the 10 Asian countries from 2000 - 2015 is 66.8. From a random sample of 45 with mean and standard deviation as 68 and 4.60, perform hypothesis test at alpha = 0.03 level of significance that the average life expectancy is greater than 66.8 and conclude whether there is sufficient evidence to reject the null hypothesis. Assume the population is normally distributed.

***Solution:***

Given μ = 66.8, x bar = 68, s = 4.60, n = 45, Alpha = 0.03

Null hypothesis: Ho: μ = 66.8

Alternative Hypothesis: Ha: μ > 66.8

It is a right tailed test.

Formula for Test Statistic = (68-66.8)/(4.60/sqrt(45)) = 1.75

Test Statistic t = 1.75, which is less than the critical value. (i.e., 1.75 < 1.88)

Also, Since the p-value is greater than alpha = 0.03 (i.e., 0.04 > 0.03), we have sufficient evidence to not reject the null hypothesis and conclude that the average life expectancy is not greater than 66.8.

***Module 4:***

*Question:*

From the dataset of 10 countries, 2 samples were collected with 5 countries in each. If both are approximately normal and have equal population variances, test the claim that there is a difference in the mean life expectancy between these samples with 0.05 level of significance. And the 2 samples have the following values respectively:

Sample1: n1 = 80, x1 bar = 68.33, s1 = 4.27, alpha = 0.05; Sample2: n2 =80, x2 bar = 65.33, s2 = 4.45

***Solution:***

Given:

Null hypothesis: Ho: μ1 = μ2

Alternative Hypothesis: Ha: μ1 <> μ2

Formula for s^2 = (n1-1)(s1^2)+(n2-1)(s2^2) / (n1+n2-2) = 19.02

Formula for Test statistics = (x1bar-x2bar)-(μ1-μ2) / SQRT (s^2\*(1/n1 + 1/n2)) = 4.35

Critical t-value for 0.05 = 1.97

P value = 2.42441E-05

Hence on comparing p value(2.42441E-05) with alpha(0.05) we can reject null hypothesis and conclude that mean life expectancy of the two samples are not equal.

***Module 5:***

Data Source: Body Fat Prediction - [LINK](https://www.kaggle.com/datasets/fedesoriano/body-fat-prediction-dataset)

The data is for estimating the percentage of body fat determined by underwater weighing and various body circumference measurements for 252 men using 14 independent variables.

*Question:*

A screenshot of a computer

Description automatically generatedAnalyze the dataset and predict the BodyFat percentage as a function of Age, Weight(lbs), Circumferences of Neck(cm), Abdomen(cm), Thigh(cm), forearm(cm) and Wrist(cm) using Multiple regression analysis and write down the regression equation at 0.05 significance level.

*Solution:*

Null hypothesis: Ho: μ1=μ2=μ3…=μn

Alternative Hypothesis: Ha: At least one mean is different.

|  |  |  |  |
| --- | --- | --- | --- |
| *All Tested Models* | | | |
| *Model* | *R^2* | *MSE* | *Considered “x” values* |
| 1 | 0.7379 | 18.72833 | x6, x2, x13, x12, x4 |
| 2 | 0.7445 | 18.40922 | x1, x2, x4, x6, x8, x12, x13 |
| 3 | 0.5240 | 33.74 | x1, x2, x3 |

After testing all the models for the highest R^2 value and least error, I was able to select the second model displayed in the table above.

It has R^2 = 0.7445 and MSE as 18.40922. This model was run based on the 7 variables, x1, x2, x4, x6, x8, x12, x13 which are the representations of Age, Weight(lbs), Circumferences of Neck(cm), Abdomen(cm), Thigh(cm), forearm(cm) and Wrist(cm) respectively.

*Output from SAS:*

A screenshot of a computer

Description automatically generatedA screenshot of a computer

Description automatically generatedA screenshot of a computer

Description automatically generated

**Report/Interpretation from the result:**

Regression equation:

* Y hat = -33.257991 + 0.0681658\*x1 – 0.1194405\*x2 – 0.4038021\*x4 + 0.917885\*x6 + 0.2219598\*x8 + 0.5531394\*x12 – 1.5324011\*x13
* This implies that for the coefficient values of the independent variables, with every 1 unit increase in x1, x6, x8, x12, there will be an increase of 0.0681658%, 0.917885%, 0.2219598%, 0.5531394% of body fat respectively and with every 1 unit increase in x2, x4 and x13 there would be a decrease of 0.1194405%, 0.4038021% and 1.5324011% respectively.

To simplify this, let me explain it with an example here where I calculate body fat percentage of a person.

(x1)Age = 25yrs, (x2)Weight = 224.5 lbs, (x4)Neck = 41.1 cm, (x6)Abdomen = 99.2 cm, (x8)Thigh= 61.7 cm, (x12)Forearm = 30.8 cm, (x13)Wrist = 20.4 cm

* Y hat = -33.257991 + 0.0681658\*25 – 0.1194405\*224.5– 0.4038021\*41.1 + 0.917885\*99.2 + 0.2219598\*61.7 + 0.5531394\*30.8 – 1.5324011\*20.4
* Y hat = 15.60
* This implies that 15.60% out of 224 lbs of the person’s bodyweight is bodyfat.

Also, the model is significant as the P value is less than alpha (0.05), and we can conclude that at least one or more body fat percentages recorded in the data set are not equal, rejecting Null Hypothesis.

**APPENDIX**

***Complete SAS code and Outputs:***

Code: [LINK](https://drive.google.com/file/d/1CPZaIvkb2T6aoUeEHCAL7HxSJsmC1AN4/view?usp=sharing)

Results/Output: [LINK](https://drive.google.com/file/d/1LTGHlAGePtDQC3lvwRxlGG3_aTLsslxt/view?usp=sharing)

***Data Sources (Kaggle and Excel):***

DataSet 1 (Original): [LINK](https://www.kaggle.com/datasets/lashagoch/life-expectancy-who-updated)

Sample Excel used from DataSet1: [LINK](https://docs.google.com/spreadsheets/d/1HDIzaHyE1E4a_4qMlVpT61Aqm0yY8cxK/edit?usp=sharing&ouid=109387509359970175892&rtpof=true&sd=true)

DataSet 2 Original(Used for Module 5 only): [LINK](https://www.kaggle.com/datasets/fedesoriano/body-fat-prediction-dataset)

Sample Excel used: [LINK](https://drive.google.com/file/d/12zLe8mHfZ5iSDxdVkLdvB6wf0F_A0Awt/view?usp=sharing)