UTM
UNIVERSITI TEKNOLOGI MALAYSIA

SECI 2143 – PROBABILITY & STATISTICAL DATA ANALYSIS

SECTION 6

PROJECT 2 REPORT

NAME : YASMEEN BINTI ABU BAKAR

MATRIC NUMBER : A19EC0212

LECTURER : DR CHAN WENG HOWE

TABLE OF CONTENTS

INTRODUCTION

Meat consumption has increases gradually in the last few years and has even increased dramatically in some countries. Although it has increased globally, in some developed countries especially the Western countries, meat consumption has fallen slightly over the years. Assumptions are made that people are starting to adopt plant-based diet which is being vegetarian or vegan for the sake of the agricultural economy. However, customers still place a higher value of buying and eating meat. Some would say meat is a luxury since some countries are not privileged to obtain animal proteins easily. With the meat demand rising and grabbing global attention, it is only necessary to study on this particular topic for better understanding. The dataset presented for this project consists of a few variables that may influence the demand for beef in the United States. It provides an example of the influence of inflation as well as providing some statistical features for models in regression.

OBJECTIVE

The study is conducted with the purpose of acknowledging the influences that the variables have on one another. Moreover, it can be observed what type of food the citizens consume during a certain year by comparing the inflation-adjusted DPI of the year with its mean. If it is greater than the mean, the United States citizens eat more exotic food. If it is less than the mean, the citizens consume more affordable food like pasta, beans, and rice. Nevertheless, the beef consumption decreases by year.

CONTENT

FOCUS ON TOPIC

Correlation

A correlation analysis was conducted to know the inflation adjusted beef retail price affects the beef consumption among the US citizens. Besides that, it is also to assess the strength of the linear relationship between the two variables. With 0.05 of significance level, the hypothesis claims are as below:

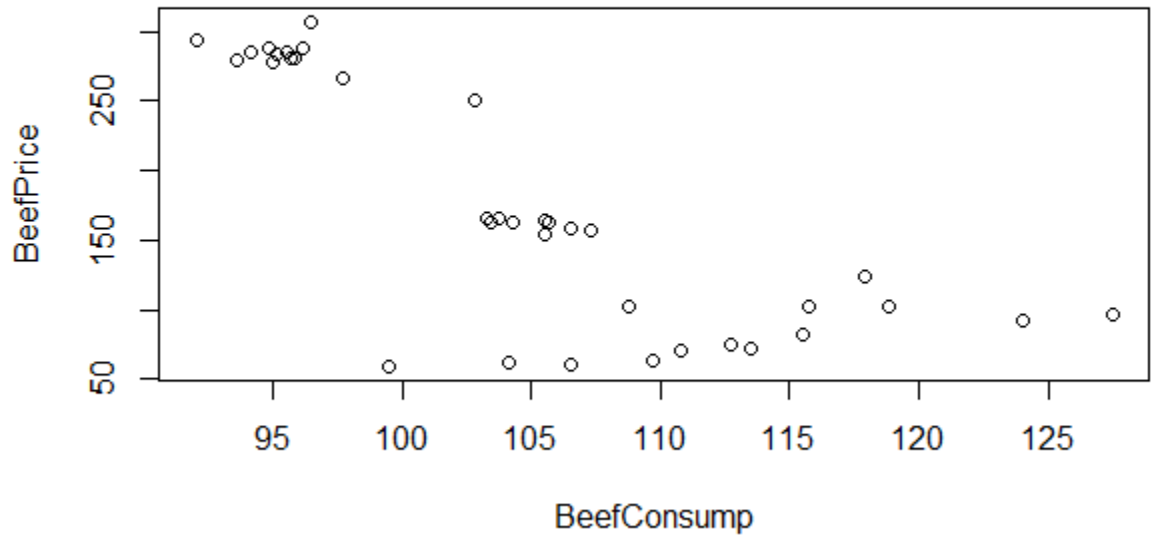
$H_0 : \rho = 0$ (no linear correlation)

$H_1 : \rho \neq 0$ (linear correlation exists)

```
> #calculate correlation coeeficient (Pearson's)
> cor.test(BeefDemand$BeefPrice, BeefDemand$BeefConsump)

        Pearson's product-moment correlation

data:  BeefDemand$BeefPrice and BeefDemand$BeefConsump
t = -7.6061, df = 34, p-value = 7.747e-09
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.8900856 -0.6291245
sample estimates:
       cor
-0.7936258

> plot(BeefPrice~BeefConsump,data=BeefDemand)
```

Using the R-studio to calculate the correlation coefficient, Pearson's method became the default method to conduct the test. As shown in the picture, the correlation coefficient or also known as r is -0.7936258 which is within the 95% confidence interval [-0.8900856, -0.6291245]. The result shows that the p-value = 7.747×10^{-9} is less than $\alpha = 0.05$. Thus, this rejects the null hypothesis, showing that there is significant evidence that there is true correlation between the two variables.

With the value of r obtained, it is known that it is a strong negative linear relationship. However, it is clearer if it was drawn. Attached below is a scatter plot of the two variables to convey the strong negative linear relationship between the two variables.



The summary of this correlation test proves that the strength of association between the variables is high correlation coefficient, r = -0.7936258. It is also proved that there is indeed a significant correlation between the variables as the p-value7.747×10^{-9} is less than the significance level, $\alpha = 0.05$. To conclude, there was a strong negative correlation between the beef price and the beef consumption by the citizens with information obtained below.

p-value = $7.747 \times 10^{-9} < \alpha = 0.05$

t = -7.606 df = 34

confidence interval = [-0.8900856, -0.6291245]

correlation coefficient, r = -0.7936258

Regression

A simple linear regression was made to predict the inflation-adjusted chicken retail price (dependent variable) based on the chicken price (independent variable). With significance level of 0.05, the hypothesis claims are as stated:

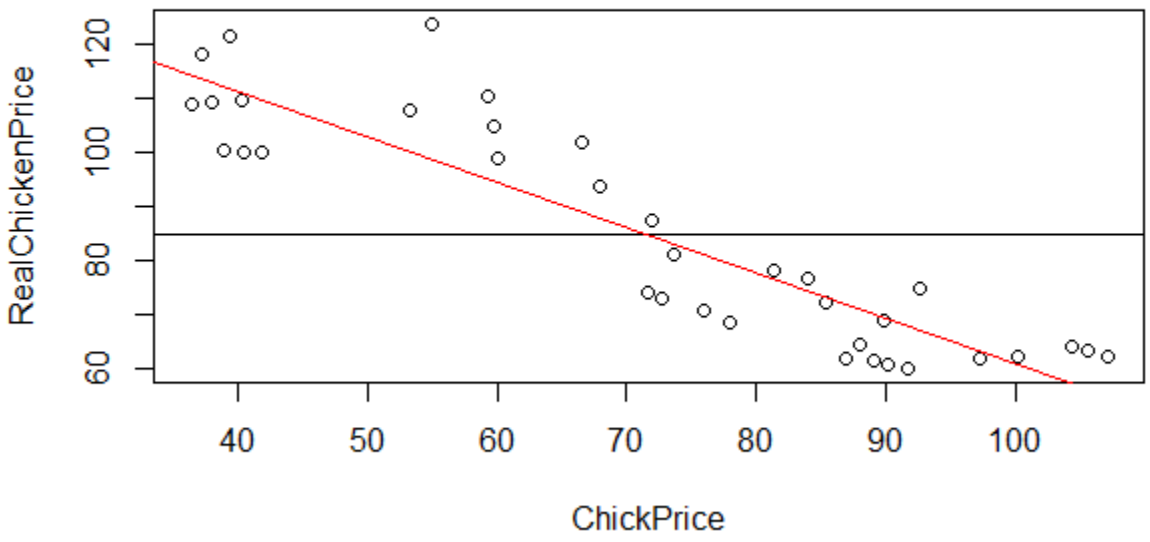
$H_0 : \beta_1 = 0$ (no linear relationship)

$H_1 : \beta_1 \neq 0$ (linear relationship exists)

```
> #simple linear regression
> x1 <- BeefDemand$ChickPrice
> y1 <- BeefDemand$RealChickenPrice
> plot(RealChickenPrice~ChickPrice,data=BeefDemand)
> #calculate mean of inflation-adjusted chicken price
> mean.y1<-mean(y1,na.rm=T)
> mean.y1
[1] 84.98087
> abline(h=mean.y1)
> #lm to draw the model line for regression
> model<-lm(y1~x1)
> model

Call:
lm(formula = y1 ~ x1)

Coefficients:
(Intercept)           x1
   145.1917      -0.8427

> abline(model,col="red")
```

```
> summary(model)

Call:
lm(formula = y1 ~ x1)

Residuals:
     Min       1Q   Median       3Q      Max
-11.9358  -8.5088  -0.8558   6.8811  24.9879

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 145.19168    5.14334   28.23  < 2e-16 ***
x1           -0.84266    0.06881  -12.25 5.11e-14 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 9.066 on 34 degrees of freedom
Multiple R-squared:  0.8152,    Adjusted R-squared:  0.8098
F-statistic:   150 on 1 and 34 DF,  p-value: 5.11e-14
```

If the test statistics were to be implemented in the Rstudio, it will give a summary output as shown in the picture. From this data, we can see that the p-value = 5.11 x 10^{-14} is much less than 0.05. Thus, we reject the null hypothesis that $\beta_1 = 0$. Hence, there is a significant relationship between the variables in the linear regression model of the data set.



Based on the plotted graph, a black line parallel to the x-axis can be observed to indicate the mean of the inflation-adjusted chicken price, 84.98087. Another red line can be observed to present the model which shows that the linear relationship is negative. The model can be drawn with the formula, $\hat{y}$ = 145.19168 – 0.84266x. Thus, the inflation-adjusted chicken price is affected by the chicken price.

To sum it up, a simple linear regression was calculated to predict the inflation-adjusted chicken price based on the chicken retail price. According to the summary presented by the Rstudio, a significant regression equation was found ($F_{(1,34)}$ = 150, p = 5.11 x 10^{-14} < 0.005), with an R^2 of 0.8152. The inflation-adjusted chicken price is equal to 145.19168 – 0.84266 of the chicken price ($\hat{y}$ = 145.19168 – 0.84266x).

SUPPORT ON TOPIC

Hypothesis Testing (One Sample)

A one sample hypothesis testing was made based on the hypothesis stating that the mean of the inflation-adjusted Disposable Personal Income per capita is less than USD10,000. The sample mean and standard deviation were calculated with the help of Microsoft Excel where it records it as 11,061.09932 and 1,845.63057 respectively. 0.05 will be used as the significance level.

$H_0 : \mu = 10000$

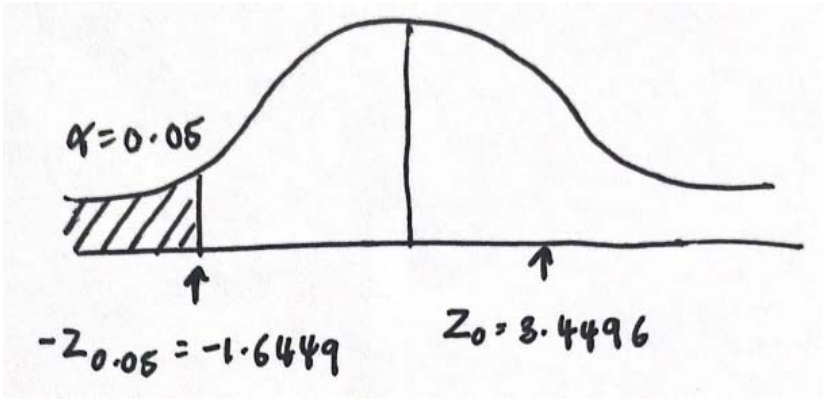
$H_1 : \mu < 10000$

Using the R-studio, a result was obtained to convey that the null hypothesis we initially made could be accepted. Two methods were shown, the traditional method which is calculating the critical value with z test statistics and the p-value method.

```
> #hypothesis testing (one sample) for RealDPI
> n=36
> alpha=0.05
> xbar=11061.09932
> sd=1845.63057
> mu=10000
> #Traditional Method
> z=(xbar-mu)/(sd/sqrt(n))
> z.alpha=qnorm(1-alpha)
> z
[1] 3.449551
> -z.alpha
[1] -1.644854
> #P-value method
> pval=pnorm(z)
> pval          #lower tail p-value
[1] 0.9997192
```

By using the traditional method, the z test statistics and the critical value were obtained to compare. The z test statistics, $Z_0 = 3.449551$ is greater than the critical value, $-Z_{(0.05)} = -1.644854$. As the z test statistics fails to fall into the rejection region, we fail to reject the null hypothesis that the mean of the inflation-adjusted Disposable Personal Income per capita can be USD10,000 at 0.05 significance level.

If p-value was used, the lower tail p-value would be 0.9997. Since the p-value is greater than the significance value, $\alpha = 0.05$, we fail to reject the null hypothesis. Thus, there is significant evidence that the mean of inflation-adjusted Disposable Personal Income per capita is equals to USD10,000.



p-value = 0.9997 $<$ α = 0.05

$-Z_{(0.05)} = -1.644854 < Z_0 = 3.449551$

One-Way ANOVA

A one-way ANOVA was conducted to compare the effect of the independent variables, chicken price, beef price and consumer price index on the dependent variables, the inflation-adjusted of both chicken and beef price. It is also used to determine the equality between the variables. Thus, R-studio was used to calculate and obtain the significant differences between the variables mentioned. The hypothesis claims are as stated:

$H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5$

H_1 : at least one of the means is different

```
> #ANOVA
> x2<-BeefDemand$BeefPrice
> x3<-BeefDemand$CPI
> y2<-BeefDemand$RealBeefPrice
> Combined_Groups <- data.frame(cbind(x1, x2, x3, y1, y2))
> summary(Combined_Groups)
       x1               x2               x3               y1               y2
 Min.   : 36.39   Min.   : 59.50   Min.   : 31.50   Min.   : 60.15   Min.   :143.6
 1st Qu.: 54.54   1st Qu.: 95.67   1st Qu.: 48.08   1st Qu.: 64.48   1st Qu.:172.4
 Median : 73.25   Median :162.18   Median : 98.05   Median : 77.49   Median :186.7
 Mean   : 71.45   Mean   :174.43   Mean   : 94.74   Mean   : 84.98   Mean   :185.4
 3rd Qu.: 89.25   3rd Qu.:279.71   3rd Qu.:137.22   3rd Qu.:102.73   3rd Qu.:198.3
 Max.   :107.12   Max.   :306.40   Max.   :172.20   Max.   :123.84   Max.   :231.5
> StackedGroups <- stack(Combined_Groups)
> Anova <- aov(values~ind, data=StackedGroups)
> summary(Anova)
             Df Sum Sq Mean Sq F value Pr(>F)
ind           4 411562  102891   44.91 <2e-16 ***
Residuals   175 400956    2291
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

As observed from the summary of the ANOVA, the p-value is way less than the the significance level set. P-value = $2x10^{-16}$ is less than $\alpha = 0.05$. Thus, there were no statistically significant differences between group means as determined by the one-way ANOVA. Thus, we fail to reject the null hypothesis. The means of the 5 variables mentioned are equal.

$$F_{(4,175)} = 44.91 , \qquad p = 2x10^{-16} < \alpha = 0.05$$

CONCLUSION

Overall, inferential statistical analysis was used to draw conclusions for the data regarding the beef demand in the United States. Since correlation and regression were the main focus of this topic, it can be concluded that there is true correlation between variables based on the data presented. In this case, the beef demand in the United States along with the beef price were tested. There is a significant relationship between the two variables and the relationship itself is a strong negative linear relationship. The p-value obtained was 7.747×10^{-9} which was less than the significance level $\alpha = 0.05$. The summary of this correlation test indicates that the strength of association between the variables is high ($r = -0.7936258$), and that there is a significant correlation between the variables. In short, there is a strong negative correlation between the beef price and the beef demand.

As for regression, the concept is quite similar to correlation, that is to determine whether one variable affects the other. However, a prediction was done for the dependent variable. According to the data acquired with the help of Rstudio, a prediction was done for the inflation-adjusted chicken price based on the chicken retail price. It results to the inflation-adjusted chicken price being equal to 145.19168 – 0.84266 of the chicken price ($\hat{y} = 145.19168 – 0.84266x$). With this information, the population regression model can be inserted into the scatter plot. Thus, the predicted inflation-adjusted chicken price based on the chicken price is as the mentioned formula.

With the two tests, we can simply conclude that most of the variables influences and links to one another. A variable may affect another variable with the information obtained from calculating or programming as proof.

REFERENCE

Kopcso, D., 2020. [online] Jse.amstat.org. Available at: http://jse.amstat.org/v22n1/kopcso/BeefDemandDoc.txt

Howe, C., (2020). *R Programming Tutorial*. [PowerPoint Slides] Retrieved from http://elearning.utm.my/19202/course/view.php?id=740

UTSSC, (2014, February 24). *Correlation in RStudio* [YouTube video] from URL https://www.youtube.com/watch?v=xsL4yLBNyDg&list=WL&index=4&t=0s

Scherber, C., (2013, September 5). *Statistics with R(1) – Linear Regression* [YouTube video] from URL https://www.youtube.com/watch?v=Xh6Rex3ARjc&list=WL&index=2

Statisticsfun, (2014, October 19). *How to Calculate Anova Using R* [YouTibe video] from URL https://www.youtube.com/watch?v=fT2No3Io72g&list=WL&index=1