VIRGINIA COMMONWEALTH UNIVERSITY

STATISTICAL ANALYSIS & MODELING

A1a: CONSUMPTION PATTERN OF ANDHRA PRADESH USING PYTHON AND R

Kirthan Shaker Iyangar
V01072700

Date of Submission: 23/06/2023

CONTENTS

Content:	Page no:
INTRODUCTION	3
OBJECTIVE	3
BUSINESS SIGNIFICANC	3-4
RESULTS AND INTERPRETATIONS (Python and R)	4-9
CODES	10-14

Multiple Regression Analysis and Interpretation on NSSO Data and Regression Analysis on IPL Player Data, with Player Salary and Performance Relation. (Both R and Python have been used for the analysis)

INTRODUCTION

The focus of this study is to use regression analysis , from the NSSO data,. In the process, we manipulate and clean the dataset to get the required data to analyze. To facilitate this analysis, we have gathered a dataset containing consumption-related information, including data on rural and urban sectors, as well as district-wise variations. In this case we take the dependent variable and see if this has a correlation on the independent variable. The dataset has been imported into R, a powerful statistical programming language renowned for its versatility in handling and analyzing large datasets.

Our objectives include identifying missing values, addressing outliers, standardizing district and sector names, summarizing consumption data regionally and district-wise, and testing the significance of mean differences. The findings from this study can inform policymakers and stakeholders, fostering targeted interventions and promoting equitable development across the state.

OBJECTIVES

a) Perform Multiple regression analysis, carry out the regression diagnostics, and explain your findings. Correct them and revisit your results and explain the significant differences you observe. [data "NSSO68.csv"]Check for outliers and describe the outcome of your test and make suitable amendments.
b) Using IPL data, establish the relationship between the player's performance and payment he receives and discuss your findings. * Use the data sets [data "Cricket_data.csv"]
c) Analysing the Relationship Between Salary and Performance Over the Last Three Years (Regression Analysis)

BUSINESS SIGNIFICANCE

The focus of this study NSSO Data and regression analyse to see if one variable has a correlation on another consumption patterns from NSSO data holds significant implications for businesses and policymakers. By identifying the top and bottom three consuming to also find out ipl player performance and salry given will help team owners make decision , the study provides valuable insights for market entry, resource allocation, Player budget allocation and targeted interventions. Through Regression - Notation and Assumptions, Goodness of Fit - Concept of R2 - Multiple coefficients of correlation and determination, Adjusted R square; Panel data regression.

RESULTS AND INTERPRETATION

a) Perform Multiple regression analysis, carry out the regression diagnostics, and explain your findings. Correct them and revisit your results and explain the significant differences you observe. [data “NSSO68.csv”].

#Identifying the missing values.

Code and Result:

```
Call:
lm(formula = foodtotal_q ~ MPCE_MRP + MPCE_URP + Age + Meals_At_Home +
    Possess_ration_card + Education, data = subset_data)
# Fit the regression model
model <- lm(foodtotal_q~
MPCE_MRP+MPCE_URP+Age+Meals_At_Home+Possess_ration_card+Education, data =
subset_data)

# Print the regression results
print(summary(model))


library(car)
# Check for multicollinearity using Variance Inflation Factor (VIF)
vif(model) # VIF Value more than 8 its problematic

# Extract the coefficients from the model
coefficients <- coef(model)

# Construct the equation
equation <- paste0("y = ", round(coefficients[1], 2))
for (i in 2:length(coefficients)) {
  equation <- paste0(equation, " + ", round(coefficients[i], 6), "*x", i-1)
}
# Print the equation
print(equation)




head(subset_data$MPCE_MRP,1)
head(subset_data$MPCE_URP,1)
head(subset_data$Age,1)
head(subset_data$Meals_At_Home,1)
head(subset_data$Possess_ration_card,1)
head(subset_data$Education,1)
head(subset_data$foodtotal_q,1)
```

Results :

```
Residuals:
    Min      1Q  Median      3Q     Max 
-68.609  -3.971  -0.654   3.291 239.668 

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)    
(Intercept)         1.138e+01  8.243e-01  13.811  < 2e-16 ***
MPCE_MRP            1.140e-03  5.659e-05  20.152  < 2e-16 ***
MPCE_URP            9.934e-05  3.422e-05   2.903  0.00372 ** 
Age                 9.884e-02  9.613e-03  10.282  < 2e-16 ***
Meals_At_Home       5.079e-02  6.420e-03   7.911 3.27e-15 ***
Possess_ration_card -2.187e+00  3.025e-01  -7.229 5.79e-13 ***
Education           2.458e-01  3.564e-02   6.898 6.11e-12 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7.667 on 4028 degrees of freedom
  (59 observations deleted due to missingness)
Multiple R-squared:  0.202,	Adjusted R-squared:  0.2008 
F-statistic: 169.9 on 6 and 4028 DF,  p-value: < 2.2e-16
```

Interpretation:

- **Dependent Variable**: `foodtotal_q` (total food consumption)
- **Independent Variables**: `MPCE_MRP`, `MPCE_URP`, `Age`, `Meals_At_Home`, `Possess_ration_card`, `Education`

Residuals

- **Min**: -68.609
- **1Q**: -3.971
- **Median**: -0.654
- **3Q**: 3.291
- **Max**: 239.668

The residuals represent the differences between the observed and predicted values of the dependent variable (foodtotal_q). The range of the residuals indicates how far off the predictions can be from the actual values.

Coefficients and Their Interpretation

1. **Intercept**: 11.38
    - This is the expected value of `foodtotal_q` when all independent variables are zero.
2. **MPCE_MRP (Monthly Per Capita Expenditure at Market Prices)**: 0.00114
    - For each unit increase in MPCE_MRP, the `foodtotal_q` increases by 0.00114 units, holding other variables constant. This is highly significant ($p < 2e-16$).
3. **MPCE_URP (Monthly Per Capita Expenditure at Uniform Retail Prices)**: 0.00009934
    - For each unit increase in MPCE_URP, the `foodtotal_q` increases by 0.00009934 units, holding other variables constant. This is also statistically significant ($p = 0.00372$).
4. **Age**: 0.09884
    - For each additional year of age, the `foodtotal_q` increases by 0.09884 units, holding other variables constant. This is highly significant ($p < 2e-16$).
5. **Meals_At_Home**: 0.05079
    - For each additional meal eaten at home, the `foodtotal_q` increases by 0.05079 units, holding other variables constant. This is highly significant ($p = 3.27e-15$).
6. **Possess_ration_card**: -2.187
    - If the household possesses a ration card, the `foodtotal_q` decreases by 2.187 units, holding other variables constant. This is highly significant ($p = 5.79e-13$).
7. **Education**: 0.2458
    - For each additional level of education, the `foodtotal_q` increases by 0.2458 units, holding other variables constant. This is highly significant ($p = 6.11e-12$).

Model Fit

- **Residual Standard Error**: 7.667 on 4028 degrees of freedom
    - This measures the typical distance between the observed values and the model's predicted values.
- **Multiple R-squared**: 0.202
    - This indicates that approximately 20.2% of the variability in `foodtotal_q` is explained by the model.
- **Adjusted R-squared**: 0.2008
    - This adjusted measure accounts for the number of predictors in the model, providing a more accurate measure of fit when multiple predictors are used.
- **F-statistic**: 169.9 on 6 and 4028 DF, p-value: < 2.2e-16
    - This tests whether at least one of the predictors is significantly related to the dependent variable. A highly significant p-value indicates that the model is a good fit for the data.

Conclusion

The regression analysis indicates that all the predictors (MPCE_MRP, MPCE_URP, Age, Meals_At_Home, Possess_ration_card, and Education) are significantly associated with food consumption. The positive coefficients for MPCE_MRP, MPCE_URP, Age, Meals_At_Home, and Education suggest that increases in these variables are associated with an increase in food consumption. Conversely, possessing a ration card is associated with a decrease in food consumption. The model explains about 20.2% of the variation in food consumption, which is a moderate level of explanatory power.

b) Using IPL data, establish the relationship between the player's performance and payment he receives and discuss your findings. * Use the data sets [data "Cricket_data.csv"]

#Codes

import pandas as pd
from sklearn.linear_model import LinearRegression
from sklearn.metrics import r2_score, mean_absolute_percentage_error
X = df_merged[['runs_scored']] # Independent variable(s)
y = df_merged['Rs'] # Dependent variable
# Split the data into training and test sets (80% for training, 20% for testing)
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)
# Create a LinearRegression model
model = LinearRegression()
# Fit the model on the training data
model.fit(X_train, y_train)

```
LinearRegression()
In a Jupyter environment, please rerun this cell to show the HTML representation or
trust the notebook.
On GitHub, the HTML representation is unable to render, please try loading this
page with nbviewer.org.
```

```
import pandas as pd
from sklearn.model_selection import train_test_split
import statsmodels.api as sm

# Assuming df_merged is already defined and contains the necessary columns
X = df_merged[['runs_scored']] # Independent variable(s)
y = df_merged['Rs'] # Dependent variable

# Split the data into training and test sets (80% for training, 20% for testing)
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2,
random_state=42)

# Add a constant to the model (intercept)
X_train_sm = sm.add_constant(X_train)

# Create a statsmodels OLS regression model
model = sm.OLS(y_train, X_train_sm).fit()

# Get the summary of the model
summary = model.summary()
print(summary)
```

Interpretation of OLS Regression Results:-

1. **Model Summary:**
    - **R-squared: 0.069**: This indicates that only 6.9% of the variance in the dependent variable (Rs) is explained by the independent variable (runs_scored). This is a relatively low value, suggesting that runs scored alone is not a strong predictor of payment.

    - **Adj. R-squared: 0.068**: The adjusted R-squared is slightly lower, accounting for the number of predictors in the model. It also indicates a low explanatory power.
    - **F-statistic: 44.66**: This value and its corresponding p-value (5.38e-11) suggest that the model is statistically significant. Despite the low R-squared, the predictor (runs_scored) is significantly related to the response variable (Rs).
2. **Coefficients:**
    - **Intercept (const): 461.4442**: This is the estimated payment when no runs are scored. It represents the baseline payment regardless of performance.
    - **runs_scored: 0.7218**: For each additional run scored, the payment increases by approximately 0.7218 units. This coefficient is positive and statistically significant (p-value = 0.000), indicating a positive relationship between runs scored and payment.
3. **Statistical Significance:**
    - Both the intercept and the runs_scored coefficient have p-values of 0.000, meaning they are highly statistically significant.
4. **Diagnostics:**
    - **Omnibus: 46.540, Prob(Omnibus): 0.000**: These values indicate that the residuals are not normally distributed.
    - **Durbin-Watson: 1.951**: This value is close to 2, suggesting that there is no strong autocorrelation in the residuals.
    - **Jarque-Bera (JB): 55.585, Prob(JB): 8.51e-13**: These values also suggest non-normality of the residuals.
    - **Skew: 0.736, Kurtosis: 2.767**: These statistics indicate slight skewness and kurtosis in the residuals, confirming non-normality.

The regression analysis suggests that there is a statistically significant but weak positive relationship between the number of runs scored by a player and their payment. Specifically, the model indicates that on average, each additional run scored by a player increases their payment by approximately 0.7218 units.

However, the low R-squared value implies that runs scored explain only a small portion of the variation in payments. This suggests that other factors not included in the model likely play a significant role in determining player payments. These factors could include player experience, marketability, overall performance metrics, team budget, and more.

c) Analysing the Relationship Between Salary and Performance Over the Last Three Years (Regression Analysis)

Code:

```
import pandas as pd

from sklearn.model_selection import train_test_split

import statsmodels.api as sm


# Assuming df_merged is already defined and contains the necessary columns

X = df_merged[['wicket_confirmation']] # Independent variable(s)
```

y = df_merged['Rs'] # Dependent variable

Split the data into training and test sets (80% for training, 20% for testing)

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

Add a constant to the model (intercept)

X_train_sm = sm.add_constant(X_train)

Create a statsmodels OLS regression model

model = sm.OLS(y_train, X_train_sm).fit()

Get the summary of the model

summary = model.summary()

print(summary)

Result :

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                     Rs   R-squared:                       0.074
Model:                            OLS   Adj. R-squared:                  0.054
Method:                 Least Squares   F-statistic:                     3.688
Date:                Sun, 23 Jun 2024   Prob (F-statistic):             0.0610
Time:                        17:56:23   Log-Likelihood:                -360.96
No. Observations:                  48   AIC:                             725.9
Df Residuals:                      46   BIC:                             729.7
Df Model:                           1
Covariance Type:            nonrobust
=======================================================================================
                          coef    std err          t      P>|t|      [0.025      0.975]
---------------------------------------------------------------------------------------
const                 396.6881     91.270      4.346      0.000     212.971     580.405
wicket_confirmation    17.6635      9.198      1.920      0.061      -0.851      36.179
==============================================================================
Omnibus:                        6.984   Durbin-Watson:                   2.451
Prob(Omnibus):                  0.030   Jarque-Bera (JB):                6.309
Skew:                           0.877   Prob(JB):                       0.0427
Kurtosis:                       3.274   Cond. No.                         13.8
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

Interpretation of OLS Regression Results

1. **Model Summary:**
   - **R-squared: 0.074**: This indicates that 7.4% of the variance in salary (Rs) is explained by the number of wickets confirmed (wicket_confirmation). This is a relatively low value, suggesting that this performance metric alone does not explain much of the variation in salary.
   - **Adj. R-squared: 0.054**: The adjusted R-squared, which adjusts for the number of predictors in the model, is slightly lower. It also indicates a low explanatory power.
   - **F-statistic: 3.688**: This value and its corresponding p-value (0.0610) suggest that the overall model is marginally significant at the 10% level but not at the 5% level.
2. **Coefficients:**
   - **Intercept (const): 396.6881**: This is the estimated salary when no wickets are confirmed. It represents the baseline salary regardless of performance.
   - **wicket_confirmation: 17.6635**: For each additional wicket confirmed, the salary increases by approximately 17.6635 units. This coefficient has a p-value of 0.061, indicating it is marginally significant at the 10% level but not at the 5% level. The confidence interval (-0.851, 36.179) includes zero, which further suggests uncertainty about the precise effect of wickets confirmed on salary.
3. **Statistical Significance:**
   - The intercept is highly significant (p-value = 0.000).
   - The coefficient for wicket_confirmation is marginally significant (p-value = 0.061).
4. **Diagnostics:**
   - **Omnibus: 6.984, Prob(Omnibus): 0.030**: These values indicate that the residuals are not perfectly normally distributed.
   - **Durbin-Watson: 2.451**: This value is close to 2, suggesting that there is no significant autocorrelation in the residuals.
   - **Jarque-Bera (JB): 6.309, Prob(JB): 0.0427**: These values also suggest non-normality of the residuals.
   - **Skew: 0.877, Kurtosis: 3.274**: These statistics indicate some skewness and kurtosis in the residuals, confirming non-normality.

The regression analysis suggests a positive but weak relationship between the number of wickets confirmed by a player and their salary. Specifically, the model indicates that, on average, each additional wicket confirmed by a player increases their salary by approximately 17.6635 units. However, this relationship is only marginally significant, indicating some uncertainty about its strength and precision.

The low R-squared value implies that the number of wickets confirmed explains only a small portion of the variation in salaries. This suggests that other factors not included in the model likely play a significant role in determining player salaries. These factors could include other performance metrics (e.g., runs scored, economy rate), player experience, marketability, and team-specific factors

CODES

Python :

```
setwd('C:\\Users\\SERVICE POINT\\Desktop\\SCMA\\A1A')
getwd()
library(dplyr)
library(readr)
library(readxl)
library(tidyr)
install.packages("ggplot2")
library(ggplot2)

#READING THE FILE INTO R
data=read.csv("4. NSSO68 data set.csv")

#FILTERING FOR AP
df=data%>%
filter(state_1=="AP")
names(df)
head(df)
dim(df)

#FINDING MISSING VALUES
is.na(df)
any(is.na(df))
sum(is.na(df))

sort(colSums(is.na(df)),decreasing=T)

# SUBSETIING
apnew = df%>%

select(state_1,District,Region,Sector,State_Region,Meals_At_Home,ricepds_v,Wheatpds_q,chicken
_q,pulsep_q,wheatos_q,No_of_Meals_per_day)
fix(apnew)
```

```
any(is.na(apnew))
sum(is.na(apnew))
head(apnew)
sort(colSums(is.na(apnew)),decreasing=T)


#IMPUTING THE VALUES i.e REPLACING MISSING VALUES WITH MEAN
apnew=apnew%>%
  mutate(across(all_of(c("Meals_At_Home")), ~ifelse(is.na(.), mean(., na.rm = TRUE), .)))
any(is.na(apnew))
fix(apnew)

# FINDING OUTLIERS AND MAKING AMENDMENTS
boxplot(apnew$ricepds_v)
boxplot(apnew$Wheatpds_q)
boxplot(apnew$chicken_q)
boxplot(apnew$pulsep_q)
boxplot(apnew$No_of_Meals_per_day)

# Calculate quartiles and IQR
Q1 <- quantile(apnew$ricepds_v, 0.25)
Q3 <- quantile(apnew$ricepds_v, 0.75)
IQR <- Q3 - Q1

# Define outlier thresholds
lower_threshold <- Q1 - (1.5 * IQR)
upper_threshold <- Q3 + (1.5 * IQR)

apnew = subset(apnew,apnew$ricepds_v>=lower_threshold & apnew$ricepds_v<=upper_threshold)
fix(apnew)
boxplot(apnew$ricepds_v)

Q1 <- quantile(apnew$chicken_q, 0.25)
Q3 <- quantile(apnew$chicken_q, 0.75)
```

```
IQR <- Q3 - Q1

# Define outlier thresholds
lower_threshold <- Q1 - (1.5 * IQR)
upper_threshold <- Q3 + (1.5 * IQR)

apnew = subset(apnew,apnew$chicken_q>=lower_threshold &
apnew$chicken_q<=upper_threshold)
fix(apnew)
boxplot(apnew$chicken_q)

#Renaming the districts as well as the sector, viz. rural and urban.

apnew$District <- ifelse(apnew$District == 5, "East Godavari",
                  ifelse(apnew$District == 10, "West Godavari",
                  ifelse(apnew$District == 6, "Nellore",
                  ifelse(apnew$District == 3, "Anantapur", apnew$Dist))))
fix(apnew)



apnew$Sector <- ifelse(apnew$Sector == 2, "URBAN",
                ifelse(apnew$Sector == 1, "RURAL",apnew$Sector))
fix(apnew)

# Summarize the critical variables in the data set region wise and district wise and indicate the top
three districts and the bottom three districts of consumption.
# 1. Districts
apnew$total_consumption=
apnew$ricepds_v+apnew$Wheatpds_q+apnew$chicken_q+apnew$pulsep_q+apnew$wheatos_q
apnew%>%
  group_by(District)%>%
  summarise(total=sum(total_consumption))%>%
  arrange(total,District)
#TOP 3 Consuming distircts are Anantapur, (3), District 23, Nellore(6)
```

```
# 2. Region
apnew%>%
  group_by(Region)%>%
  summarise(total=sum(total_consumption))%>%
  arrange(-total,Region)
# Region 3,1 and 5 are the top 3 consuming regions.

#e) Test whether the differences in the means are significant or not.

#H0: There is no difference in consumption between urban and rural.

#H1: There is difference in consumption between urban and rural.

rural=apnew%>%
  select(Sector,total_consumption)%>%
  filter(Sector=="RURAL")
fix(rural)

urban=apnew%>%
  select(Sector,total_consumption)%>%
  filter(Sector=="URBAN")
fix(urban)

cons_rural=rural$total_consumption
cons_urban=urban$total_consumption



length(cons_rural)
length(cons_urban)

install.packages("BSDA")
library(BSDA)
```

```
z.test(cons_rural,
       cons_urban,
       alternative="two.sided",
       mu=0,
       sigma.x = 2.56,sigma.y=2.34,
       conf.level = 0.95)

# P value is <0.05, Therefore we reject the null hypothesis.
#There is difference between mean consumptions of urban and rural.
```

R Studio :

```
# Fit the regression model
model <- lm(foodtotal_q~ MPCE_MRP+MPCE_URP+Age+Meals_At_Home+Possess_ration_card+Education, data = subset_data)

# Print the regression results
print(summary(model))


library(car)
# Check for multicollinearity using Variance Inflation Factor (VIF)
vif(model) # VIF Value more than 8 its problematic

# Extract the coefficients from the model
coefficients <- coef(model)

# Construct the equation
equation <- paste0("y = ", round(coefficients[1], 2))
for (i in 2:length(coefficients)) {
  equation <- paste0(equation, " + ", round(coefficients[i], 6), "*x", i-1)
}
# Print the equation
print(equation)
```

```
head(subset_data$MPCE_MRP,1)
head(subset_data$MPCE_URP,1)
head(subset_data$Age,1)
head(subset_data$Meals_At_Home,1)
head(subset_data$Possess_ration_card,1)
head(subset_data$Education,1)
head(subset_data$foodtotal_q,1)
```