WORK SAMPLE

By

Suhasini Singh (1811068)

Department of Commerce

CHRIST (Deemed to be University)

Bangalore

2019-2020

Business Stats (COH233)

Stepwise Linear Regression Analysis:

U.S. Housing Market Factors

SUMMARY:

The document "Linear Regression Models Practical (STA551)" focuses on a stepwise linear regression analysis of factors affecting the U.S. housing market. It uses a dataset from FRED to examine various factors like house price index, stock price index, consumer price index, population, unemployment rate, real GDP, mortgage rate, and real disposable income. The analysis includes initial regression modeling and then refines it using stepwise regression to identify the most significant predictors. The study reveals insights into the relationships between these economic factors and housing prices, providing a detailed statistical analysis of their impact on the U.S. housing market.

OBJECTIVES:

1. Choose a dataset.
2. Do a complete regression analysis and then stepwise.
3. Apply the suitable multiple linear regression model and analyze the data.
4. Write a report on the same.

DATA DESCRIPTION

Source: https://www.kaggle.com/datasets/faryarmemon/usa-housing-market-factors

The data in this dataset is collected from FRED.

I decided to create this dataset while reading the research paper Factors Affecting House Prices in Cyprus: 1988-2008 by Panos Pashardes & Christos S. Savva. This research paper is extremely informative and covers a lot of details regarding the macroeconomics involved in real estate market. So I would recommend you all to go through it once.

General Defintions:

House_Price_Index: House price change according to the index base period set (you can check the date at which this value is 100).

Stock_Price_Index: Stock price change according to the index base period set (you can check the date at which this value is 100).

Consumer_Price_Index: The Consumer Price Index measures the overall change in consumer prices based on a representative basket of goods and services over time.

Population: Population of USA (unit: thousands).

Unemployment_Rate: Unemployment rate of USA (unit: percentage).

Real_GDP: GDP with adjusted inflation (Annual version unit: billions of chain 2012 dollars in, Monthly version unit: Annualised change).

Mortgage_Rate: Interest charged on mortgages (unit: percentage).

Real_Disposable_Income (Real Disposable Personal Income): Money left from salary after all the taxes are paid (unit: billions of chain 2012 dollars).

Inflation: Decline in purchasing power over time (unit: percentage). [Forgot to remove this column in Annual version since CPI is one of the measures used to determine inflation].

Dataset Head, Summary and Structure:

```
library(corrplot)

## corrplot 0.92 loaded

library(dplyr)

##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union

library(car)

## Loading required package: carData

##
## Attaching package: 'car'

## The following object is masked from 'package:dplyr':
##
##     recode

library(DescTools)

##
## Attaching package: 'DescTools'
## The following object is masked from 'package:car':
##
##     Recode
```

```
data<-read.csv("Annual_Macroeconomic_Factors.csv")
head(data)
```

```
##         Date House_Price_Index Stock_Price_Index Consumer_Price_Index
## 1 1975-01-01           61.0900          67.14653             65.30488
## 2 1976-01-01           65.5250          79.96264             69.05653
## 3 1977-01-01           73.4350          78.82540             73.54636
## 4 1978-01-01           83.7450          78.84679             79.15866
## 5 1979-01-01           95.1325          85.63207             88.06755
## 6 1980-01-01          102.6675         100.00000            100.00000
##   Population Unemployment_Rate Real_GDP Mortgage_Rate Real_Disposable_Income
## 1    0.98599           8.46667 5648.462       9.04712                  19908
## 2    0.95022           7.71667 5952.809       8.86585                  20346
## 3    1.00577           7.06667 6228.076       8.84519                  20780
## 4    1.05957           6.06667 6572.819       9.64173                  21497
## 5    1.10358           5.83333 6780.924      11.20365                  21672
## 6    0.95959           7.14167 6763.514      13.74212                  21584
```

```
str(data)
```

```
## 'data.frame':    47 obs. of 9 variables:
##  $ Date                  : chr "1975-01-01" "1976-01-01" "1977-01-01" "1978-01-01" ...
##  $ House_Price_Index     : num 61.1 65.5 73.4 83.7 95.1 ...
##  $ Stock_Price_Index     : num 67.1 80 78.8 78.8 85.6 ...
## $ Consumer_Price_Index  : num  65.3 69.1 73.5 79.2 88.1 ...
##  $ Population            : num 0.986 0.95 1.006 1.06 1.104 ...
##  $ Unemployment_Rate     : num 8.47 7.72 7.07 6.07 5.83 ...
##  $ Real_GDP              : num 5648 5953 6228 6573 6781 ...
##  $ Mortgage_Rate         : num  9.05 8.87 8.85 9.64 11.2 ...
## $ Real_Disposable_Income: num 19908 20346 20780 21497 21672 ...
```

```
summary(data)
```

```
##      Date           House_Price_Index Stock_Price_Index Consumer_Price_Index
##  Length:47          Min.    : 61.09    Min. : 67.15 Min. : 65.3
##  Class :character   1st Qu.:140.79    1st Qu.: 209.90  1st Qu.:135.4
##  Mode  :character   Median :211.46    Median : 756.56 Median :197.8
##                     Mean   :240.15    Mean : 743.13 Mean :198.6
##                     3rd Qu.:339.35    3rd Qu.:1114.17  3rd Qu.:262.9
##                     Max.   :523.26    Max.   :2255.84   Max.   :328.8
##    Population     Unemployment_Rate    Real_GDP     Mortgage_Rate
##  Min.   :0.1184   Min.   :3.667     Min. : 5648 Min. : 2.958
##  1st Qu.:0.8627   1st Qu.:5.167     1st Qu.: 8374  1st Qu.: 4.863
##  Median :0.9459   Median :5.992    Median :12046   Median : 7.440
##  Mean  :0.9352   Mean   :6.310   Mean   :12140   Mean   : 7.781##
3rd Qu.:1.0816  3rd Qu.:7.442    3rd Qu.:15646    3rd Qu.: 9.886##
Max. :1.3869 Max. :9.708 Max. :19427 Max. :16.642
##  Real_Disposable_Income
## Min. :19908
## 1st Qu.:25432
## Median :31712
##   Mean  :32041
##  3rd Qu.:38235
## Max. :48219
```

```
model<-lm(House_Price_Index~Consumer_Price_Index+Population+Unemployment_Rate+
Real_GDP+Mortgage_Rate+Real_Disposable_Income,data=data)
```

ASSUMPTIONS OF LINEAR REGRESSION:

BASIC TESTS OF NORMALITY:

```
durbinWatsonTest(model)
```

```
## lag Autocorrelation D-W Statistic p-value
##    1       0.6194733     0.6697391       0
## Alternative hypothesis: rho != 0
```

THE DATA IS NOT AUTOCORRELATED.

```
shapiro.test(data$House_Price_Index)
```

```
##
##  Shapiro-Wilk normality test
##
## data: data$House_Price_Index
## W = 0.95093, p-value = 0.04709
```

```
shapiro.test(residuals(model))
```

```
##
##  Shapiro-Wilk normality test
##
## data: residuals(model)
## W = 0.95746, p-value = 0.0853
```

```
round(sum(data$House_Price_Index-mean(data$House_Price_Index)))
```

```
## [1] 0
```

```
round(sum(residuals(model)))
```

```
## [1] 0
```

```
sum(fitted.values(model))
```

```
## [1] 11286.84
```

```
sum(data$House_Price_Index)
```

```
## [1] 11286.84
```

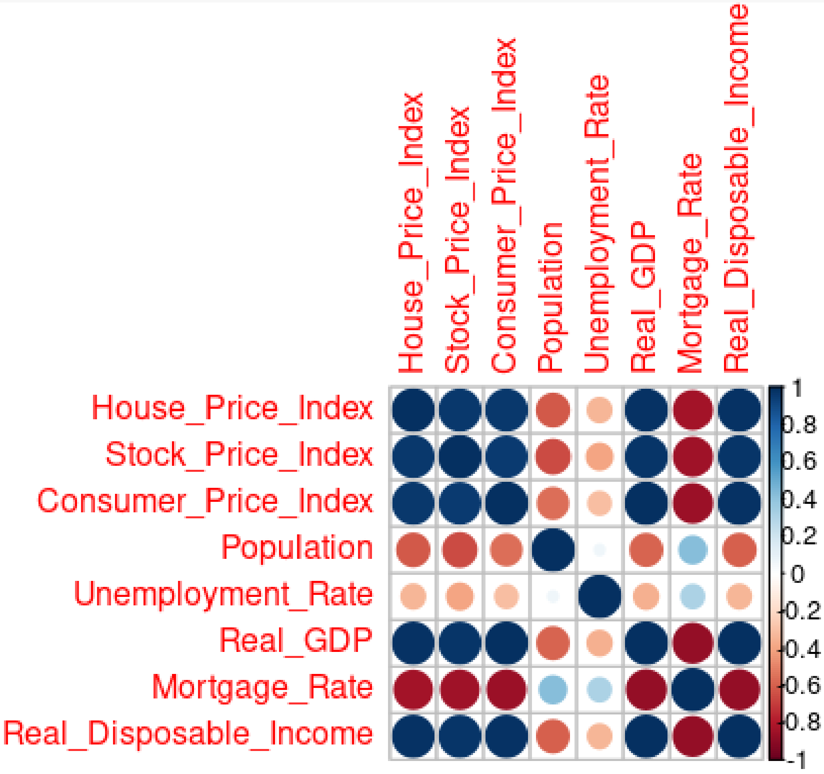
The sum of deviation from mean is zero.

The sum of residuals is zero.

The sum of fitted values is equal to observed values.

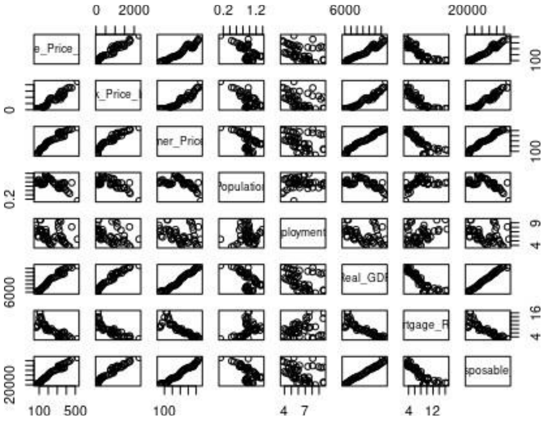
CORRELATION PLOT AND PAIRS PLOT FOR LINEARITY AND STRENGTH OF LINEARITY:

```
data2<-select(data,-Date)
mtrix<-cor(data2)
corrplot(mtrix)
```



The correlation plot showcases that all the variables have a strong correlation, other than Unemployment Rate. The Mortgage Rate shows negative correlation because of inverse relationship with buying sentiment and the same trend is shown by population as prices soar.

```
pairs(data2)
```



The pairs plot showcases that almost all the variables have a linear trend, other thanpopulation and employement, population and market price, population and stock price , population and House price, Population and GDP, Population and Mortgage Rate andPopulation and Disposable Income.

```
data$Date<-as.factor(data$Date)

model<-lm(House_Price_Index~.,data=data2)
model

##
## Call:
## lm(formula = House_Price_Index ~ ., data = data2)
##
## Coefficients:
##          (Intercept)      Stock_Price_Index    Consumer_Price_Index
##           -373.82332              0.01517              -0.73859
##           Population      Unemployment_Rate                Real_GDP
##              9.62930              1.08338               0.00853
##        Mortgage_Rate  Real_Disposable_Income
##              5.32877              0.01837

summary(model)
```

```
##
## Call:
## lm(formula = House_Price_Index ~ ., data = data2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -38.099 -7.610 -0.874  6.949  41.825
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)           -3.738e+02 7.916e+01  -4.722 3.00e-05 ***
## Stock_Price_Index       1.517e-02  2.579e-02  0.588  0.55970
## Consumer_Price_Index -7.386e-01 3.770e-01  -1.959  0.05726 .
## Population              9.629e+00 1.932e+01 0.499 0.62093
## Unemployment_Rate       1.083e+00  2.204e+00  0.492  0.62573
## Real_GDP                8.530e-03  9.455e-03  0.902  0.37248
## Mortgage_Rate           5.329e+00  1.955e+00  2.726  0.00955 **
## Real_Disposable_Income 1.837e-02 4.105e-03  4.474 6.47e-05 ***
## ---
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 18.58 on 39 degrees of freedom
## Multiple R-squared: 0.9798, Adjusted R-squared: 0.9762
## F-statistic: 270.3 on 7 and 39 DF,  p-value: < 2.2e-16
```

The minimum value of residual is -38.099 and maximum value is 41.825. The range of residuals is not very high which suggests that the deviation of observed values from expected values is low.

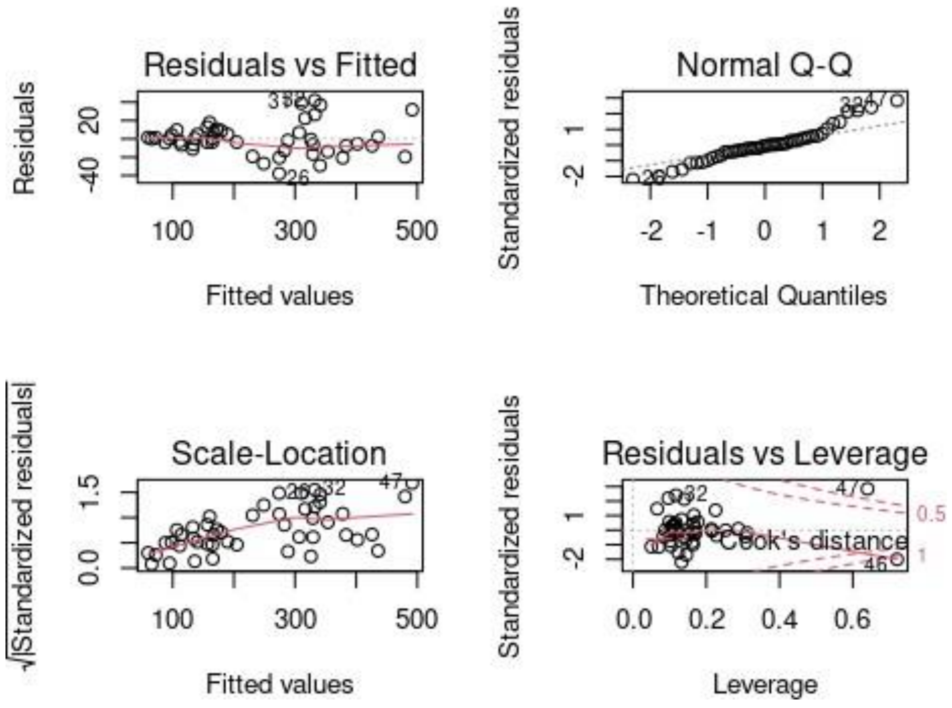
The p values in the case of intercept is less than 0.05, thus we may accept the null hypothesis at 5% significance level and state that the intercept does not have a significant impact of the regression model.

The p value for mortgage rate and Disposable Income are less than 0.05, thus we may reject the null hypothesis at 5% significance level and state that they have a significant impact.

The adjusted R squared value is 0.9762 which implies that the regression model is a very good fit to the data as 97.62% of variation in the price is successfully explained.

Further, the p value for the overall model is less than 0.05 which implies that the regression model is significant.

```
par(mfrow=c(2,2))
plot(model)
```



In statistics, a Q-Q (**quantile-quantile**) plot is a probability plot, which is a graphical method for comparing two probability distributions by plotting their quantiles against each other. Normal Q-Q graph shows a slight change from Ideal Normal plot and is tailed, or can be termed as left skewed in precise explanation.

Our plot is tailed in nature because of the deviations at both ends in the QQ Plot and the residuals are normally distributed.

The residual vs Fitted Graph showcases that the variance is not constant but y is an increasing function as its creating some sort of funnel shape.

```
confint(model)

##                              2.5 %        97.5 %
## (Intercept)           -533.94717358 -213.69946870
## Stock_Price_Index        -0.03699599    0.06734500
## Consumer_Price_Index     -1.50109419    0.02391262
## Population              -29.44137973  48.69997142
## Unemployment_Rate        -3.37385393    5.54061328
## Real_GDP                 -0.01059358    0.02765403
## Mortgage_Rate             1.37502255    9.28251705
## Real_Disposable_Income    0.01006398    0.02667046

sampledata<-data.frame("Stock_Price_Index"=67.14564,"Consumer_Price_Index"=65.30488
,"Population"=0.98599,"Unemployment_Rate"=8.46667,"Real_GDP"=5648.462,"Mortgage_
Rate"=9.04712,"Real_Disposable_Income"=19908)
predict(model,sampledata)

##       1
## 59.6763
```

The expected answer was 58.6 and hence we can conclude that the model is a good fit and can be used for prediction.

Now Use the stepwise regression model selection method. n statistics, stepwise regression is a method of fitting regression models in which the choice of predictive variables is carried out by an automatic procedure. In each step, a variable is considered for addition to or subtraction from the set of explanatory variables based on some prespecified criterion. Usually, this takes the form of a forward, backward, or combined sequence of F-tests or t-tests.

```
stepmodel<-step(model,direction="both")

## Start:  AIC=281.93
## House_Price_Index ~ Stock_Price_Index + Consumer_Price_Index +
##     Population + Unemployment_Rate + Real_GDP + Mortgage_Rate +
##     Real_Disposable_Income
```

```
##
##                          Df Sum of Sq   RSS    AIC
## - Unemployment_Rate       1      83.5 13553 280.22
## - Population              1      85.8 13556 280.23
## - Stock_Price_Index       1     119.5 13589 280.34
## - Real_GDP                1     281.1 13751 280.90
## <none>                                13470 281.93
## - Consumer_Price_Index    1    1325.8 14796 284.34
## - Mortgage_Rate           1    2566.8 16037 288.13
## - Real_Disposable_Income  1    6914.3 20384 299.40
##
## Step: AIC=280.22
## House_Price_Index ~ Stock_Price_Index + Consumer_Price_Index +
##     Population + Real_GDP + Mortgage_Rate + Real_Disposable_Income
##
##                          Df Sum of Sq   RSS    AIC
## - Population              1      35.0 13588 278.34
## - Stock_Price_Index       1      76.5 13630 278.48
## - Real_GDP                1     207.2 13760 278.93
## <none>                                13553 280.22
## + Unemployment_Rate       1      83.5 13470 281.93
## - Consumer_Price_Index    1    1262.3 14816 282.40
## - Mortgage_Rate           1    2484.7 16038 286.13
## - Real_Disposable_Income  1    7343.9 20897 298.57
##
## Step: AIC=278.34
## House_Price_Index ~ Stock_Price_Index + Consumer_Price_Index +
##     Real_GDP + Mortgage_Rate + Real_Disposable_Income
##
##                          Df Sum of Sq   RSS    AIC
## - Stock_Price_Index       1      56.1 13644 276.53
## - Real_GDP                1     191.8 13780 277.00
## <none>                                13588 278.34
## + Population              1      35.0 13553 280.22
## + Unemployment_Rate       1      32.6 13556 280.23
## - Consumer_Price_Index    1    1243.4 14832 280.46
## - Mortgage_Rate           1    2555.1 16143 284.44
## - Real_Disposable_Income 1     7328.6 20917 296.61
##
## Step: AIC=276.53
## House_Price_Index ~ Consumer_Price_Index + Real_GDP + Mortgage_Rate +
##     Real_Disposable_Income
##
##                         Df Sum of Sq  RSS     AIC
## - Real_GDP               1     328.0 13972 275.65
## <none>                               13644 276.53
## + Stock_Price_Index       1      56.1 13588 278.34
## + Unemployment_Rate       1      20.1 13624 278.46
## + Population             1      14.6 13630 278.48
## - Consumer_Price_Index    1    2070.7 15715 281.18
## - Mortgage_Rate           1    2961.8 16606 283.77
## - Real_Disposable_Income  1    9708.9 23353 299.79
##
## Step: AIC=275.65
## House_Price_Index ~ Consumer_Price_Index + Mortgage_Rate +
Real_Disposable_Income
##
##                         Df Sum of Sq   RSS    AIC
## <none>                               13972 275.65
## + Real_GDP               1     328.0 13644 276.53
## + Stock_Price_Index       1     192.3 13780 277.00
## + Unemployment_Rate       1       7.2 13965 277.63
## + Population             1       0.0 13972 277.65
## - Consumer_Price_Index    1    1789.1 15761 279.31
## - Mortgage_Rate           1    2636.9 16609 281.78
## - Real_Disposable_Income  1   26244.1 40216 323.34
```

```
stepmodel
```

```
##
## Call:
## lm(formula = House_Price_Index ~ Consumer_Price_Index + Mortgage_Rate +
##     Real_Disposable_Income, data = data2)
##
## Coefficients:
##          (Intercept)  Consumer_Price_Index         Mortgage_Rate
##            -382.6446               -0.5614                4.6188
## Real_Disposable_Income
##               0.0218
```

The model hence chosen was with Consumer Price Index, Mortgage rate and Real disposable income. Every other variable has been termed unfit and insignificant.

```
summary(stepmodel)
```

```
##
## Call:
## lm(formula = House_Price_Index ~ Consumer_Price_Index + Mortgage_Rate +
##     Real_Disposable_Income, data = data2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -34.472  -9.838  -2.159  5.903  45.611
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)            -3.826e+02  4.424e+01  -8.648 5.84e-11 ***
## Consumer_Price_Index -5.614e-01 2.393e-01  -2.346  0.02363 *
## Mortgage_Rate           4.619e+00 1.621e+00 2.849 0.00671 **
## Real_Disposable_Income  2.179e-02  2.425e-03  8.987 1.99e-11 ***
## ---
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 18.03 on 43 degrees of freedom
## Multiple R-squared: 0.979, Adjusted R-squared: 0.9776
## F-statistic: 669.8 on 3 and 43 DF,  p-value: < 2.2e-16
```

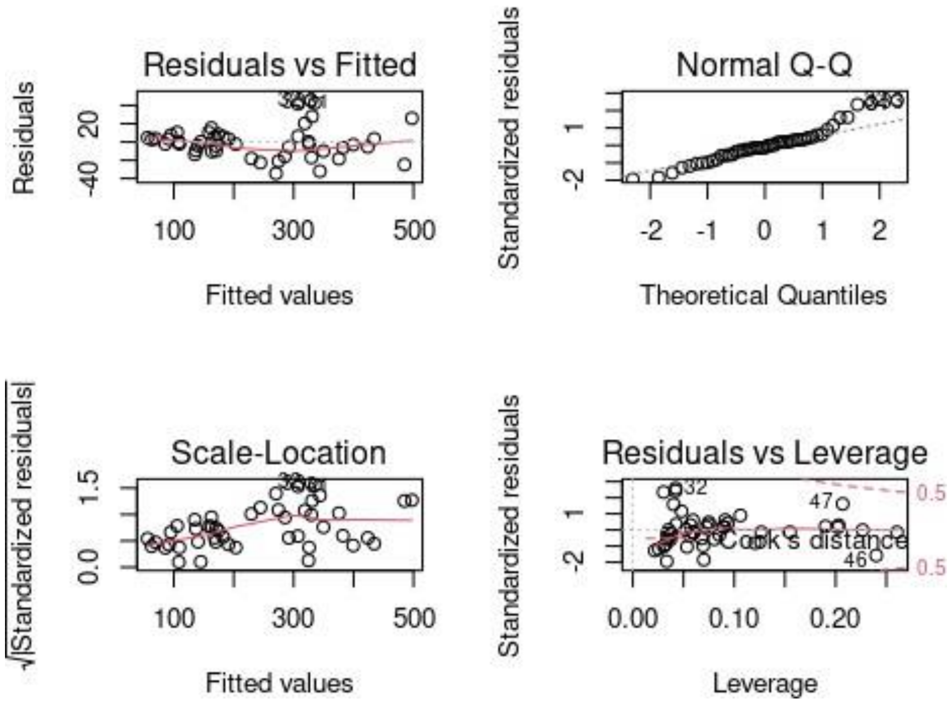
The minimum value of residual is -34.472 and maximum value is 45.611. The range of residuals is not very high which suggests that the deviation of observed values from expected values is low.

The p values in the case are significant and all the factors are significant.

The adjusted R squared value is 0.9776 which implies that the regression model is a very good fit to the data as 97.76% of variation in the price is successfully explained vs 97.62% from previous model.

Further, the p value for the overall model is less than 0.05 which implies that the regression model is significant.

```
par(mfrow=c(2,2))
plot(stepmodel)
```



In statistics, a Q-Q (**quantile-quantile**) plot is a probability plot, which is a graphical method for comparing two probability distributions by plotting their quantiles against each other. Normal Q-Q graph shows a slight change from Ideal Normal plot and is tailed, or can be termed as left skewed in precise explanation.

Our plot is tailed in nature because of the deviations at both ends in the QQ Plot and the residuals are normally distributed.

The residual vs Fitted Graph showcases that the variance is not constant but y is an increasing function as its creating some sort of funnel shape.

```
confint(stepmodel)

##                               2.5 %        97.5 %
## (Intercept)            -471.8727667 -293.41646394
## Consumer_Price_Index     -1.0439330   -0.07890170
## Mortgage_Rate             1.3490262    7.88865605
## Real_Disposable_Income    0.0169044    0.02668617

sampledata2<-data.frame("Consumer_Price_Index"=65.30488,"Mortgage_Rate"=9.04712,"Real_Disposable_Income"=19908)
predict(stepmodel,sampledata2)

##        1
## 56.37987
```

The expected answer was 58.6 and hence we can conclude that the model is a good fit and can be used for prediction.

CONCLUSION:

1. The pairs plot showcases that almost all the variables have a linear trend, other than population and employement, population and market price, population and stock price , population and House price, Population and GDP, Population and Mortgage Rate and Population and Disposable Income.

2. MODEL 1:

   The minimum value of residual is -38.099 and maximum value is 41.825. The range of residuals is not very high which suggests that the deviation of observed values from expected values is low. The p values in the case of intercept is less than 0.05, thus we may accept the null hypothesis at 5% significance level and state that the intercept does not have a significant impact of the regression model.

   The p value for mortgage rate and Disposable Income are less than 0.05, thus we may reject the null hypothesis at 5% significance level and state that they have a significant impact.

   The adjusted R squared value is 0.9762 which implies that the regression model is a very good fit to the data as 97.62% of variation in the price is successfully explained.Further, the p value for the overall model is less than 0.05 which implies that the regression model is significant.

   In statistics, a Q-Q (quantile-quantile) plot is a probability plot, which is a graphical method for comparing two probability distributions by plotting their quantiles against each other. Normal Q-Q graph shows a slight change from Ideal Normal plot and is tailed, or can be termed as left skewed in precise explanation. Our plot is tailed in nature because of the deviations at both ends in the QQ Plot and the residuals are normally distributed.

   The residual vs Fitted Graph showcases that the variance is not constant but y is an increasing function as its creating some sort of funnel shape.

   The expected answer was 58.6 and hence we can conclude that the model is a good fit and can be used for prediction.

3. Now Use the stepwise regression model selection method. n statistics, stepwise regression is a method of fitting regression models in which the choice of predictive variables is carried out by an automatic procedure. In each step, a variable is

considered for addition to or subtraction from the set of explanatory variables based on some prespecified criterion. Usually, this takes the form of a forward, backward, or combined sequence of F-tests or t-tests.

4. MODEL 2:

The model hence chosen was with Consumer Price Index, Mortgage rate and Real disposable income. Every other variable has been termed unfit and insignificant.

The minimum value of residual is -34.472 and maximum value is 45.611. The range of residuals is not very high which suggests that the deviation of observed values from expected values is low. The p values in the case are significant and all the factors are significant.

The adjusted R squared value is 0.9776 which implies that the regression model is a very good fit to the data as 97.76% of variation in the price is successfully explained vs 97.62% from previous model. Further, the p value for the overall model is less than 0.05 which implies that the regression model is significant.

5. Our plot is tailed in nature because of the deviations at both ends in the QQ Plot and the residuals are normally distributed.

The residual vs Fitted Graph showcases that the variance is not constant but y is an increasing function as its creating some sort of funnel shape. The expected answer was 58.6 and hence we can conclude that the model is a good fit and can be used for prediction.