Walmart Statistical Computation

Terrence Josiah

6/18/2022

1. Read the data

```
getwd()

## [1] "C:/Users/terre/Desktop/Personal Projects/Retail analysis with Walmart
sales data/Programming files"

setwd("C:/Users/terre/OneDrive/Documents")

walmartS = read.csv("WalmartStatisticComp.csv", header = TRUE)

head(walmartS)

##   Weekly_Sales      CPI Unemployment Fuel_Price Time
## 1      1643691 211.0964        8.106      2.572    1
## 2      1641957 211.2422        8.106      2.548    2
## 3      1611968 211.2891        8.106      2.514    3
## 4      1409728 211.3196        8.106      2.561    4
## 5      1554807 211.3501        8.106      2.625    5
## 6      1439542 211.3806        8.106      2.667    6

str(walmartS)

## 'data.frame':    48 obs. of  5 variables:
##  $ Weekly_Sales: num  1643691 1641957 1611968 1409728 1554807 ...
##  $ CPI         : num  211 211 211 211 211 ...
##  $ Unemployment: num  8.11 8.11 8.11 8.11 8.11 ...
##  $ Fuel_Price  : num  2.57 2.55 2.51 2.56 2.62 ...
##  $ Time        : int  1 2 3 4 5 6 7 8 9 10 ...

summary(walmartS)

##   Weekly_Sales          CPI         Unemployment     Fuel_Price
##  Min.   :1345454   Min.   :210.3   Min.   :7.787   Min.   :2.514
##  1st Qu.:1429059   1st Qu.:211.1   1st Qu.:7.787   1st Qu.:2.625
##  Median :1494366   Median :211.4   Median :7.808   Median :2.691
##  Mean   :1526642   Mean   :211.3   Mean   :7.861   Mean   :2.697
##  3rd Qu.:1552446   3rd Qu.:211.6   3rd Qu.:7.838   3rd Qu.:2.762
##  Max.   :2387950   Max.   :212.0   Max.   :8.106   Max.   :2.943
##       Time
##  Min.   : 1.00
##  1st Qu.:12.75
```

```
##  Median :24.50
##  Mean   :24.50
##  3rd Qu.:36.25
##  Max.   :48.00
```

2. Make the linear model

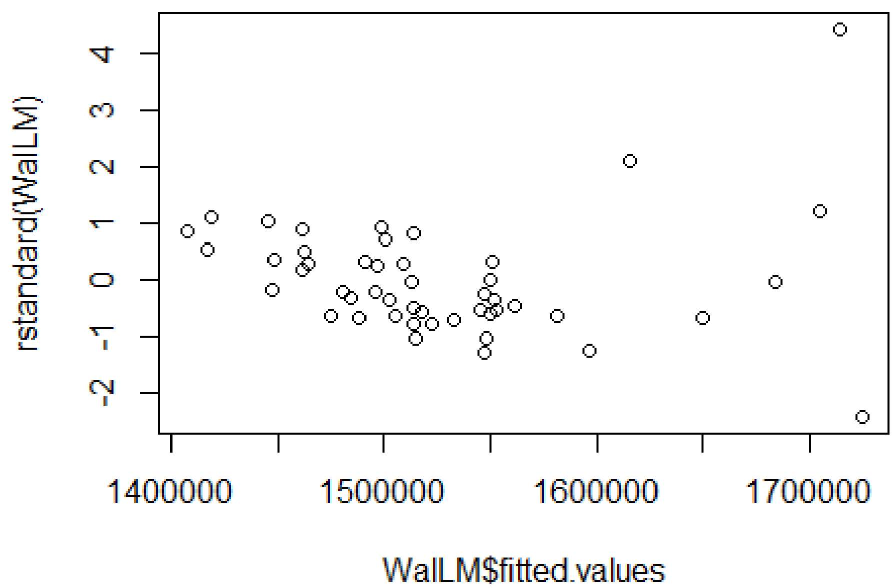
```
WalLM = lm(walmartS$Weekly_Sales ~., data = walmartS)
WalLM

## 
## Call:
## lm(formula = walmartS$Weekly_Sales ~ ., data = walmartS)
## 
## Coefficients:
##  (Intercept)          CPI  Unemployment    Fuel_Price          Time  
##     31928955      -172742        744673          1524          9845

summary(WalLM)

## 
## Call:
## lm(formula = walmartS$Weekly_Sales ~ ., data = walmartS)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -356990  -96628  -30584   65444  673605 
## 
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)  
## (Intercept)  31928955   28574626   1.117   0.2700  
## CPI           -172742     139848  -1.235   0.2235  
## Unemployment   744673     352276   2.114   0.0404 *
## Fuel_Price       1524     420089   0.004   0.9971  
## Time             9845       5067   1.943   0.0586 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 166200 on 43 degrees of freedom
## Multiple R-squared:  0.177,  Adjusted R-squared:  0.1004 
## F-statistic: 2.312 on 4 and 43 DF,  p-value: 0.07292
```

3. Testing the 4 assumptions of linear regression

1. Testing the Independent Errors/Residuals (Autocorrelation)

- Looking at the residual plot

```
plot(WalLM$fitted.values, rstandard(WalLM))
```



CONCLUSION: It appears that there are random patterns in the residual plot.

- Conduct the Durbin-Watson test

```
library(lmtest)

## Loading required package: zoo

##
## Attaching package: 'zoo'

## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric

dwtest(WalLM)

##
##  Durbin-Watson test
##
## data:  WalLM
```

```
## DW = 1.9124, p-value = 0.1885
## alternative hypothesis: true autocorrelation is greater than 0
```

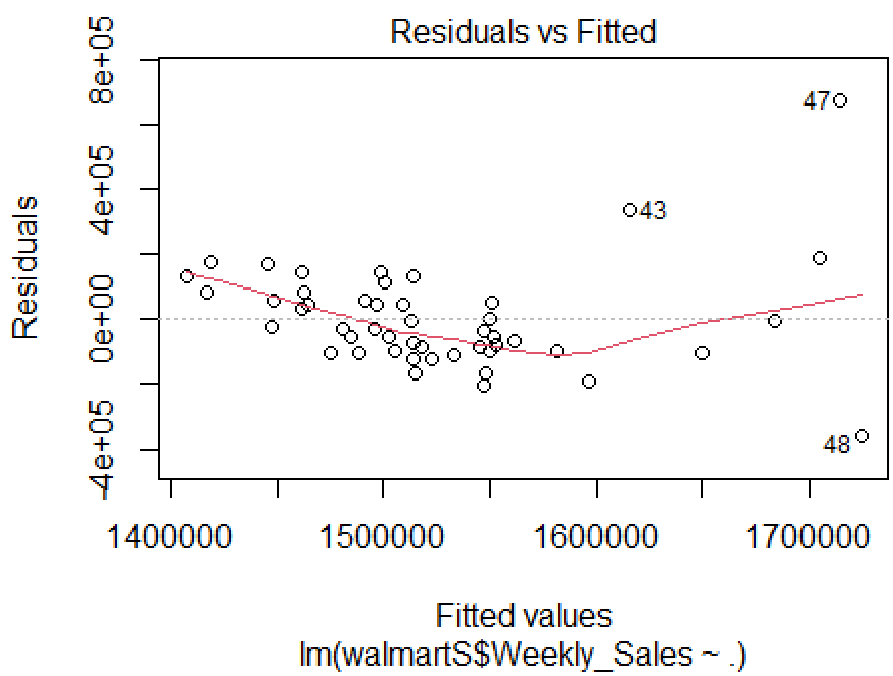
CONCLUSION: The p-value of the Durbin-Watson test (0.1885) is above 0.05, which implies that the Errors/Residuals are independent.

Therefore, there are no Autocorrelation between the errors/residuals.

2. Linearity test (Linear relationship between dependent and independent variables)

- Looking at the plot

```
plot(WalLM, which = 1)
```

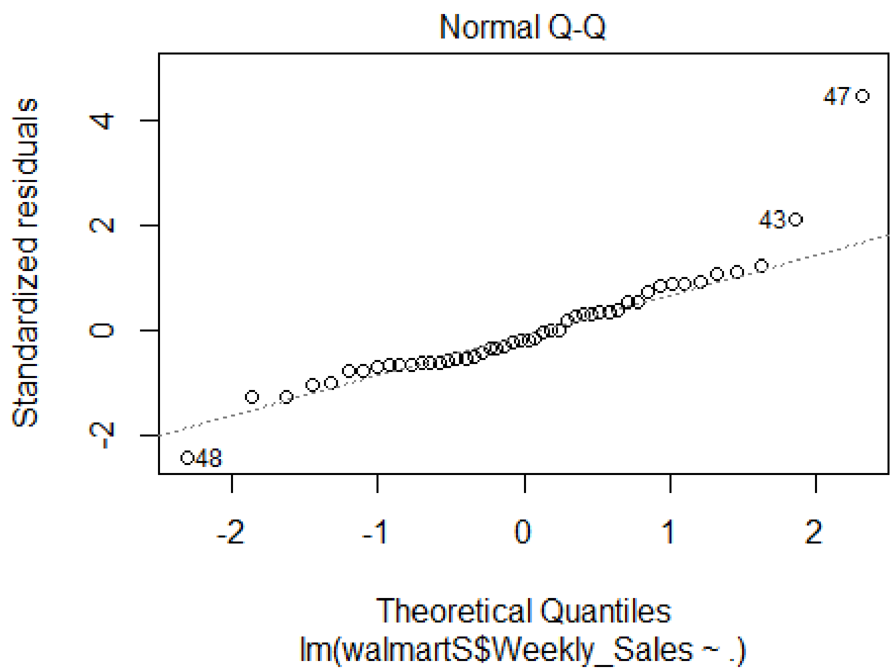


CONCLUSION: Linearity doesn't seem to be hold as the red line has a pattern and thus not close to the dashed line (residual = 0)

3. Normality test (See if the residuals are normally distributed)

- Looking at the normality plot

```
plot(WalLM, which = 2)
```



CONCLUSION: The residuals don't follow normal distribution since Point 48, 43, and 47 are are far from the diagonal dashed line.

- Conduct the Shapiro-Wilk normality test on the residuals

```
shapiro.test(WalLM$residuals)

##
##  Shapiro-Wilk normality test
##
## data:  WalLM$residuals
## W = 0.88117, p-value = 0.0001637
```

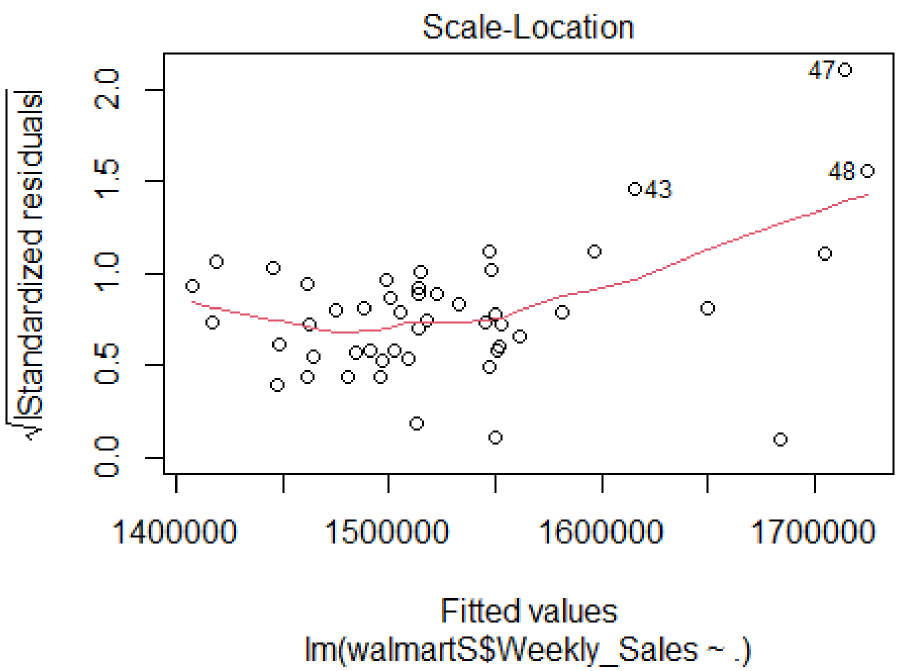
CONCLUSION: The p-value of the test is below 0.05, indicating that the residuals don't follow normal distribution.

Thus, we need to remove the data to fulfill the normality assumption.

4. Equal Variance test

- Observe the plot

```
plot(WalLM, which = 3)
```



CONCLUSION: The spread of the residuals varies much around the red line, which implies that the variance of the residuals in nor constant.

- Conduct the Breusch-Pagan test

```
bptest(WalLM)

##
##  studentized Breusch-Pagan test
##
## data:  WalLM
## BP = 13.061, df = 4, p-value = 0.01098
```

CONCLUSION: The p-value of the test is less than 0.05, indicating that the residuals are not distributed with equal variance.

ASSUMPTIONS CONCLUSION: The linear regression model might not be the best model to predict the weekly sales as the only assumption that is fulfilled is only the independent errors/residuals.

However in this case, we still want to determine the best possible linear regression model to predict weekly sales.

4. Make an analysis of the model

1. The full model with all variables included

- Looking at the Adjusted R-Squared, AIC, and BIC values

```
print(paste("Adjusted R-squared =", summary(WalLM)$adj.r.squared))

## [1] "Adjusted R-squared = 0.100440114649426"

print(paste("AIC =", AIC(WalLM)))

## [1] "AIC = 1296.95790891306"

print(paste("BIC =", BIC(WalLM)))

## [1] "BIC = 1308.18511497851"
```

2. CPI as the independent variable

- Looking at the Adjusted R-Squared, AIC, and BIC values

```
CPI.lm = lm(Weekly_Sales ~ CPI, data = walmartS)
print(paste("Adjusted R-squared =", summary(CPI.lm)$adj.r.squared))

## [1] "Adjusted R-squared = -0.0169478435910482"

print(paste("AIC =", AIC(CPI.lm)))

## [1] "AIC = 1300.08255356192"

print(paste("BIC =", BIC(CPI.lm)))

## [1] "BIC = 1305.69615659465"
```

3. Unemployment as the independent variable

- Looking at the Adjusted R-Squared, AIC, and BIC values

```
Unemploy.lm = lm(Weekly_Sales ~ Unemployment, data = walmartS)
print(paste("Adjusted R-squared =", summary(Unemploy.lm)$adj.r.squared))

## [1] "Adjusted R-squared = -0.0199728776266452"

print(paste("AIC =", AIC(Unemploy.lm)))

## [1] "AIC = 1300.22512341155"

print(paste("BIC =", BIC(Unemploy.lm)))

## [1] "BIC = 1305.83872644427"
```

4. Fuel Price as the independent variable

- Looking at the Adjusted R-Squared, AIC, and BIC values

```
fuel.price.lm = lm(Weekly_Sales ~ Fuel_Price, data = walmartS)
print(paste("Adjusted R-squared =", summary(fuel.price.lm)$adj.r.squared))

## [1] "Adjusted R-squared = 0.0406308928402247"

print(paste("AIC =", AIC(fuel.price.lm)))

## [1] "AIC = 1297.2848629165"

print(paste("BIC =", BIC(fuel.price.lm)))

## [1] "BIC = 1302.89846594922"
```

5. Time as the independent variable

- Looking at the Adjusted R-Squared, AIC, and BIC values

```
Time.lm = lm(Weekly_Sales ~ Time, data = walmartS)
print(paste("Adjusted R-squared =", summary(Time.lm)$adj.r.squared))

## [1] "Adjusted R-squared = 0.0358412399766361"

print(paste("AIC =", AIC(Time.lm)))

## [1] "AIC = 1297.52390682798"

print(paste("BIC =", BIC(Time.lm)))

## [1] "BIC = 1303.13750986071"
```

6. Time and Unemployment as the independent variables

- Looking at the Adjusted R-Squared, AIC, and BIC values

```
Time.Unemploy.lm = lm(Weekly_Sales ~ Time + Unemployment, data = walmartS)
print(paste("Adjusted R-squared =", summary(Time.Unemploy.lm)$adj.r.squared))

## [1] "Adjusted R-squared = 0.0608276087629132"

print(paste("AIC =", AIC(Time.Unemploy.lm)))

## [1] "AIC = 1297.20858727641"

print(paste("BIC =", BIC(Time.Unemploy.lm)))

## [1] "BIC = 1304.69339132004"
```

7. Time and Fuel Price as the independent variables

- Looking at the Adjusted R-Squared, AIC, and BIC values

```
Time.Fuel.lm = lm(Weekly_Sales ~ Time + Fuel_Price, data = walmartS)
print(paste("Adjusted R-squared =", summary(Time.Fuel.lm)$adj.r.squared))

## [1] "Adjusted R-squared = 0.0500213353009017"

print(paste("AIC =", AIC(Time.Fuel.lm)))

## [1] "AIC = 1297.75773000592"

print(paste("BIC =", BIC(Time.Fuel.lm)))

## [1] "BIC = 1305.24253404955"
```

8. CPI and Fuel Price as the independent variables

- Looking at the Adjusted R-Squared, AIC, and BIC values

```
CPI.Fuel.lm = lm(Weekly_Sales ~ CPI + Fuel_Price, data = walmartS)
print(paste("Adjusted R-squared =", summary(CPI.Fuel.lm)$adj.r.squared))

## [1] "Adjusted R-squared = 0.0460913430318487"

print(paste("AIC =", AIC(CPI.Fuel.lm)))

## [1] "AIC = 1297.9558928853"

print(paste("BIC =", BIC(CPI.Fuel.lm)))

## [1] "BIC = 1305.44069692893"
```

9. Unemployment and Fuel Price as the independent variables

- Looking at the Adjusted R-Squared, AIC, and BIC values

```
Unemploy.Fuel.lm = lm(Weekly_Sales ~ Unemployment + Fuel_Price, data = walmar
tS)
print(paste("Adjusted R-squared =", summary(Unemploy.Fuel.lm)$adj.r.squared))

## [1] "Adjusted R-squared = 0.0331855095517596"

print(paste("AIC =", AIC(Unemploy.Fuel.lm)))

## [1] "AIC = 1298.60095131653"

print(paste("BIC =", BIC(Unemploy.Fuel.lm)))

## [1] "BIC = 1306.08575536016"
```

10. CPI and Time as the independent variables

- Looking at the Adjusted R-Squared, AIC, and BIC values

```
CPI.Time.lm = lm(Weekly_Sales ~ CPI + Time, data = walmartS)
print(paste("Adjusted R-squared =", summary(CPI.Time.lm)$adj.r.squared))
```

```
## [1] "Adjusted R-squared = 0.0273934577256756"
print(paste("AIC =", AIC(CPI.Time.lm)))
## [1] "AIC = 1298.88765472907"
print(paste("BIC =", BIC(CPI.Time.lm)))
## [1] "BIC = 1306.3724587727"
```

11. CPI and Unemployment as the independent variables

- Looking at the Adjusted R-Squared, AIC, and BIC values

```
CPI.Unemploy.lm = lm(Weekly_Sales ~ CPI + Unemployment, data = walmartS)
print(paste("Adjusted R-squared =", summary(CPI.Unemploy.lm)$adj.r.squared))
## [1] "Adjusted R-squared = -0.037682081606"
print(paste("AIC =", AIC(CPI.Unemploy.lm)))
## [1] "AIC = 1301.99638013002"
print(paste("BIC =", BIC(CPI.Unemploy.lm)))
## [1] "BIC = 1309.48118417365"
```

12. CPI, Unemployment, and Time as the independent variables

- Looking at the Adjusted R-Squared, AIC, and BIC values

```
CPI.Unemploy.Time.lm = lm(Weekly_Sales ~ CPI + Unemployment + Time, data = wa
lmartS)
print(paste("Adjusted R-squared =", summary(CPI.Unemploy.Time.lm)$adj.r.squar
ed))
## [1] "Adjusted R-squared = 0.120884388597244"
print(paste("AIC =", AIC(CPI.Unemploy.Time.lm)))
## [1] "AIC = 1294.95792359515"
print(paste("BIC =", BIC(CPI.Unemploy.Time.lm)))
## [1] "BIC = 1304.31392864969"
```

13. Unemployment, Fuel Price, and Time as the independent variables

- Looking at the Adjusted R-Squared, AIC, and BIC values

```
Unemploy.Fuel.Time.lm = lm(Weekly_Sales ~ Unemployment + Fuel_Price + Time, d
ata = walmartS)
print(paste("Adjusted R-squared =", summary(Unemploy.Fuel.Time.lm)$adj.r.squa
red))
```

```
## [1] "Adjusted R-squared = 0.0896917327257885"

print(paste("AIC =", AIC(Unemploy.Fuel.Time.lm)))

## [1] "AIC = 1296.63153394559"

print(paste("BIC =", BIC(Unemploy.Fuel.Time.lm)))

## [1] "BIC = 1305.98753900013"
```

14. CPI, Unemployment, and Fuel Price as the independent variables

- Looking at the Adjusted R-Squared, AIC, and BIC values

```
CPI.Unemploy.Fuel.lm = lm(Weekly_Sales ~ CPI + Unemployment + Fuel_Price, dat
a = walmartS)
print(paste("Adjusted R-squared =", summary(CPI.Unemploy.Fuel.lm)$adj.r.squar
ed))

## [1] "Adjusted R-squared = 0.0437006240615653"

print(paste("AIC =", AIC(CPI.Unemploy.Fuel.lm)))

## [1] "AIC = 1298.99734457199"

print(paste("BIC =", BIC(CPI.Unemploy.Fuel.lm)))

## [1] "BIC = 1308.35334962653"
```

15. CPI, Fuel Price, and Time as the independent variables

- Looking at the Adjusted R-Squared, AIC, and BIC values

```
CPI.Fuel.Time.lm = lm(Weekly_Sales ~ CPI + Fuel_Price + Time, data = walmartS
)
print(paste("Adjusted R-squared =", summary(CPI.Fuel.Time.lm)$adj.r.squared))

## [1] "Adjusted R-squared = 0.0295275939509904"

print(paste("AIC =", AIC(CPI.Fuel.Time.lm)))

## [1] "AIC = 1299.7035182062"

print(paste("BIC =", BIC(CPI.Fuel.Time.lm)))

## [1] "BIC = 1309.05952326074"
```

CONCLUSION: - Highest Adjusted R-Squared: Model 12 (CPI + Unemployment + Time) (0.1209) ` - Lowest AIC: Model 12 (CPI + Unemployment + Time) (1294.9579) - Lowest BIC: Model 4 (Fuel Price) (1302.8985)

- Model 12 can explain 12.09% of the variance.

Therefore, Model 12 with CPI, Unemployment, and Time as the independent variables is the best model to predict the dependent variable, weekly sales. We don’t need variable fuel price, which means that the independent variable fuel price has no impact on predicting the weekly sales.