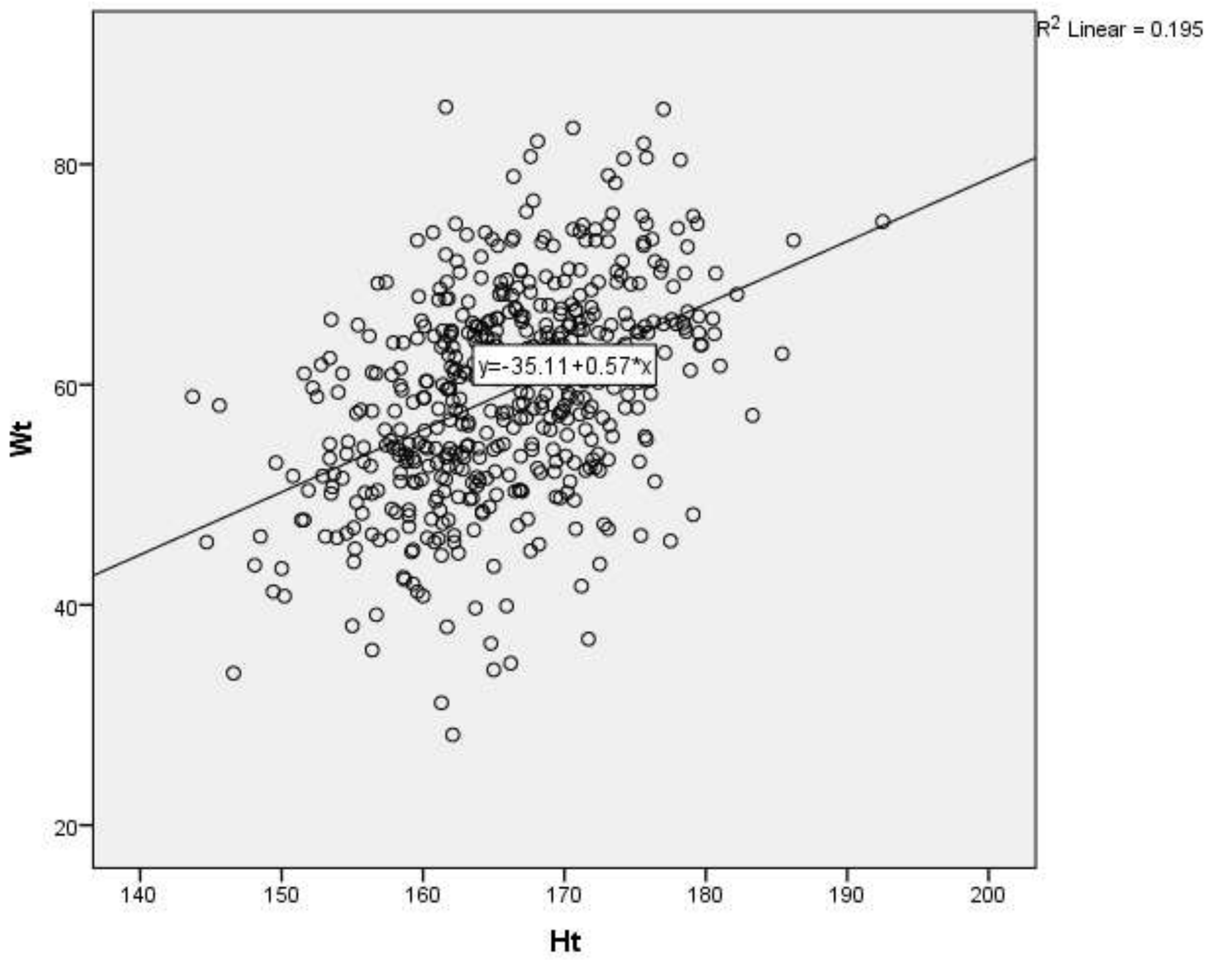
1. **Scatterplot of htwt_Chicago for 500 females**




<u>According to the scatterplot, 19.5% of the variance in weight can by explained by height.</u>

**Correlation of the two variables**

```
CORRELATIONS
  /VARIABLES=ht wt
  /PRINT=TWOTAIL NOSIG
  /MISSING=PAIRWISE.
```

**GGraph**
**Correlations**

**Correlations**

| | | ht | wt |
|---|---|---|---|
| ht | Pearson Correlation | 1 | .441** |
| | Sig. (2-tailed) | | .000 |
| | N | 500 | 500 |
| wt | Pearson Correlation | .441** | 1 |

| | | | |
|---|---|---|---|
| | Sig. (2-tailed) | .000 | |
| | N | 500 | 500 |

**. Correlation is significant at the 0.01 level (2-tailed).

Yes, it makes sense to go for a regression line for the data. This is because as per the scatterplot, there looks like a linear dependency among the two variables of height and weight. Also, as per the correlation table, the correlation between the two variables is significant (equivalent to .000). To further accurately estimate the relationship between height and weight, we use regression.

2. **Regression Model**

```
REGRESSION
/MISSING LISTWISE
/STATISTICS COEFF OUTS R ANOVA
/CRITERIA=PIN(.05) POUT(.10)
/NOORIGIN
/DEPENDENT wt
/METHOD=ENTER ht.
```

**Regression**

**Variables Entered/Removed[a]**

| Model | Variables Entered | Variables Removed | Method |
|---|---|---|---|
| 1 | ht[b] | . | Enter |

a. Dependent Variable: wt
b. All requested variables entered.

**Model Summary**

| Model | R | R Square | Adjusted R Square | Std. Error of the Estimate |
|---|---|---|---|---|
| 1 | .441[a] | .195 | .193 | 8.5411 |

a. Predictors: (Constant), ht

**ANOVA[a]**

| Model | | Sum of Squares | df | Mean Square | F | Sig. |
|---|---|---|---|---|---|---|
| 1 | Regression | 8778.792 | 1 | 8778.792 | 120.340 | .000[b] |
| | Residual | 36328.974 | 498 | 72.950 | | |

| | | | | | | |
|---|---|---|---|---|---|---|
| Total | 45107.767 | 499 | | | | |

a. Dependent Variable: wt

b. Predictors: (Constant), ht

**Coefficients**[a]

| Model | | Unstandardized Coefficients | | Standardized Coefficients | t | Sig. |
|---|---|---|---|---|---|---|
| | | B | Std. Error | Beta | | |
| 1 | (Constant) | -35.106 | 8.614 | | -4.075 | .000 |
| | ht | .569 | .052 | .441 | 10.970 | .000 |

a. Dependent Variable: wt

DESCRIPTIVES VARIABLES=ht wt
/STATISTICS=MEAN STDDEV VARIANCE MIN MAX.

Sensitivity of Height on Weight

So, y = a + bx + e

y= -35.106 + 0.569 x + e

Estimate of Slope: The slope of 0.569 means that for each increase of 1 unit X (Education), we predict the average of Y (Income) to increase by an estimate of 0.569

Estimate of Intercept: Theoretically, If independent variable (height)= 0, then the dependent variable (weight) is equal to -35.106

3. **Descriptives**

**Descriptive Statistics**

| | N | Minimum | Maximum | Mean | Std. Deviation | Variance |
|---|---|---|---|---|---|---|
| ht | 500 | 143.7 | 192.5 | 165.908 | 7.3714 | 54.337 |
| wt | 500 | 28.2 | 85.2 | 59.297 | 9.5077 | 90.396 |
| Valid N (listwise) | 500 | | | | | |

$$\text{Estimate of sigma square or variance} = \frac{SSE}{n-2} = \frac{36328.974}{498} = 72.94$$

The **t**-value measures the size of the difference relative to the variation in the sample data
The p value for the t-distribution is 0.000 which is statistically significant. This means that the variation of weight (Dependent variable) can be explained by the variation of height (Independent variable)

4. **Regression**

```
REGRESSION
  /MISSING LISTWISE
  /STATISTICS COEFF OUTS R ANOVA
  /CRITERIA=PIN(.05) POUT(.10)
  /NOORIGIN
  /DEPENDENT wt
  /METHOD=ENTER ht
  /SCATTERPLOT=(*ZRESID ,*ZPRED).
```

**Variables Entered/Removed[a]**

| Model | Variables Entered | Variables Removed | Method |
|---|---|---|---|
| 1 | ht[b] | . | Enter |

a. Dependent Variable: wt
b. All requested variables entered.

**Model Summary[b]**

| Model | R | R Square | Adjusted R Square | Std. Error of the Estimate |
|---|---|---|---|---|
| 1 | .441[a] | .195 | .193 | 8.5411 |

a. Predictors: (Constant), ht
b. Dependent Variable: wt

**ANOVA[a]**

| Model | | Sum of Squares | df | Mean Square | F | Sig. |
|---|---|---|---|---|---|---|
| 1 | Regression | 8778.792 | 1 | 8778.792 | 120.340 | .000[b] |
| | Residual | 36328.974 | 498 | 72.950 | | |
| | Total | 45107.767 | 499 | | | |

a. Dependent Variable: wt
b. Predictors: (Constant), ht

**Coefficients[a]**

| Model | | Unstandardized Coefficients | | Standardized Coefficients | t | Sig. |
|---|---|---|---|---|---|---|
| | | B | Std. Error | Beta | | |
| 1 | (Constant) | -35.106 | 8.614 | | -4.075 | .000 |
| | ht | .569 | .052 | .441 | 10.970 | .000 |

a. Dependent Variable: wt

F ratio is 120.340. This value is high. This implies that the variation among group means is more than you would expect to see by chance.
p-value for the F test = .000 (<5%) which is statistically significant.

Numerical Relationship between F and t-value: F = t^2 (t squared). Yes, the two test have the same p value of 0.000. This is because the relationship between the two variables is highly statistically significant. Hence, the p value is so low for both the test nearly approximating to zero.

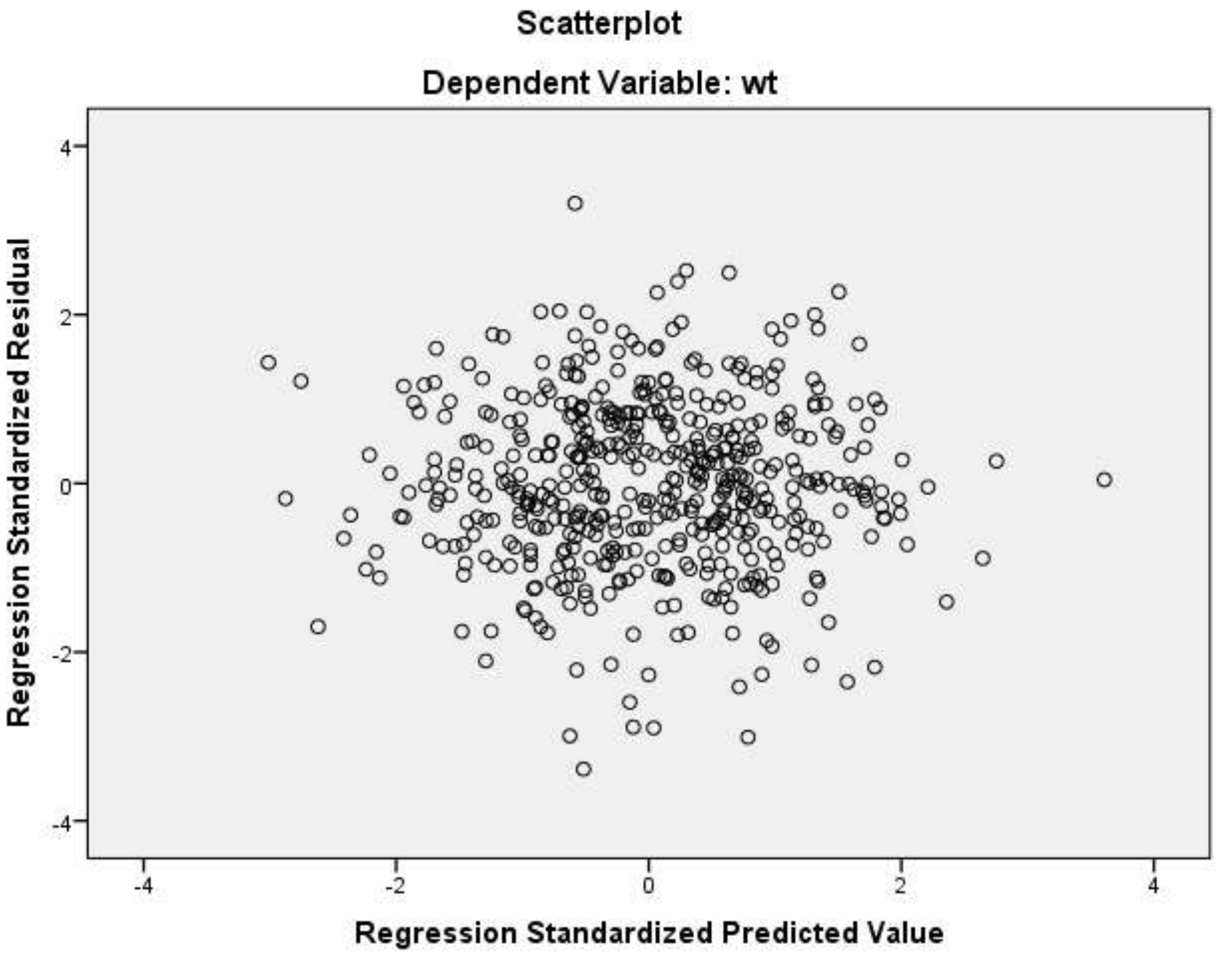
**5. Residual Plot**

**Residuals Statistics[a]**

| | Minimum | Maximum | Mean | Std. Deviation | N |
|---|---|---|---|---|---|
| Predicted Value | 46.661 | 74.428 | 59.297 | 4.1944 | 500 |
| Residual | -28.9305 | 28.3540 | .0000 | 8.5325 | 500 |
| Std. Predicted Value | -3.013 | 3.607 | .000 | 1.000 | 500 |
| Std. Residual | -3.387 | 3.320 | .000 | .999 | 500 |

a. Dependent Variable: wt

**Charts**




<u>Conclusion: When points are randomly dispersed across the horizontal axis, then a linear regression model is appropriate.</u>

**QQ Plots**

```
PPLOT
/VARIABLES=ht wt
/NOLOG
/NOSTANDARDIZE
/TYPE=Q-Q
/FRACTION=BLOM
/TIES=MEAN
/DIST=NORMAL.
```

**PPlot**

**Model Description**

| | | |
|---|---|---|
| Model Name | | MOD_1 |
| Series or Sequence | 1 | ht |
| | 2 | wt |
| Transformation | | None |
| Non-Seasonal Differencing | | 0 |

| | | |
|---|---|---|
| Seasonal Differencing | | 0 |
| Length of Seasonal Period | | No periodicity |
| Standardization | | Not applied |
| Distribution | Type | Normal |
| | Location | estimated |
| | Scale | estimated |
| Fractional Rank Estimation Method | | Blom's |
| Rank Assigned to Ties | | Mean rank of tied values |

Applying the model specifications from MOD_1

**Case Processing Summary**

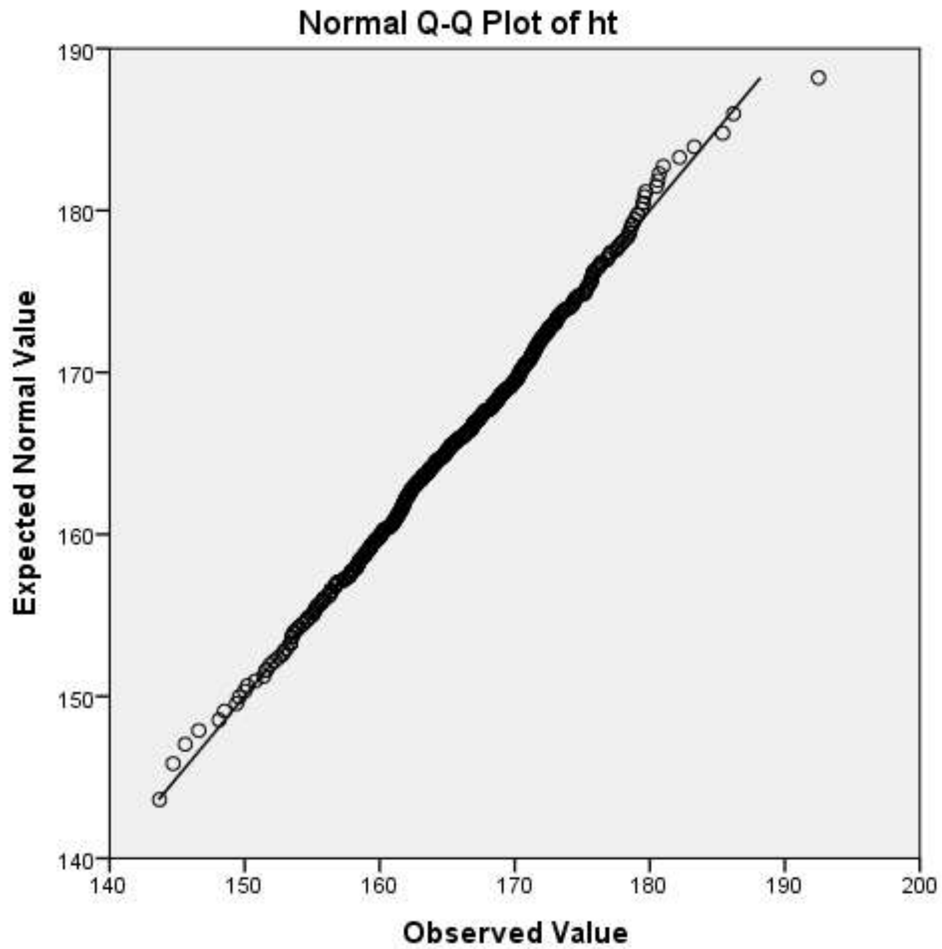
| | | ht | wt |
|---|---|---|---|
| Series or Sequence Length | | 500 | 500 |
| Number of Missing Values in the Plot | User-Missing | 0 | 0 |
| | System-Missing | 0 | 0 |

The cases are unweighted.

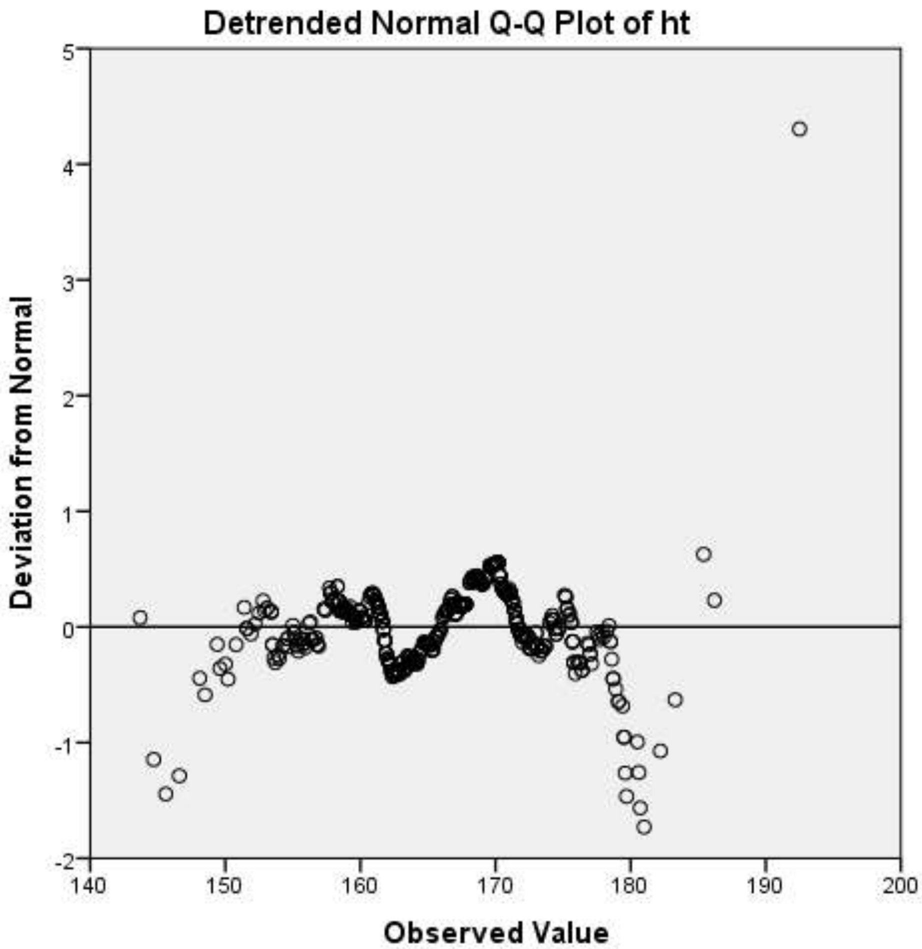
**Estimated Distribution Parameters**

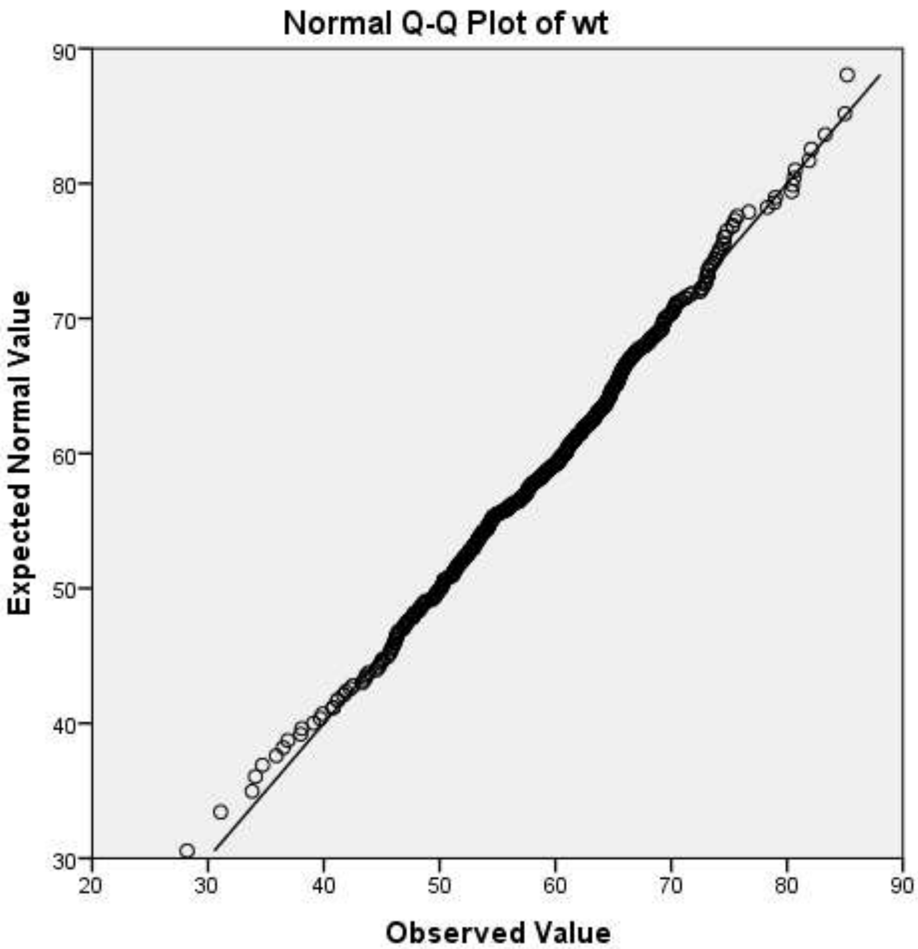
| | | ht | wt |
|---|---|---|---|
| Normal Distribution | Location | 165.908 | 59.297 |
| | Scale | 7.3714 | 9.5077 |

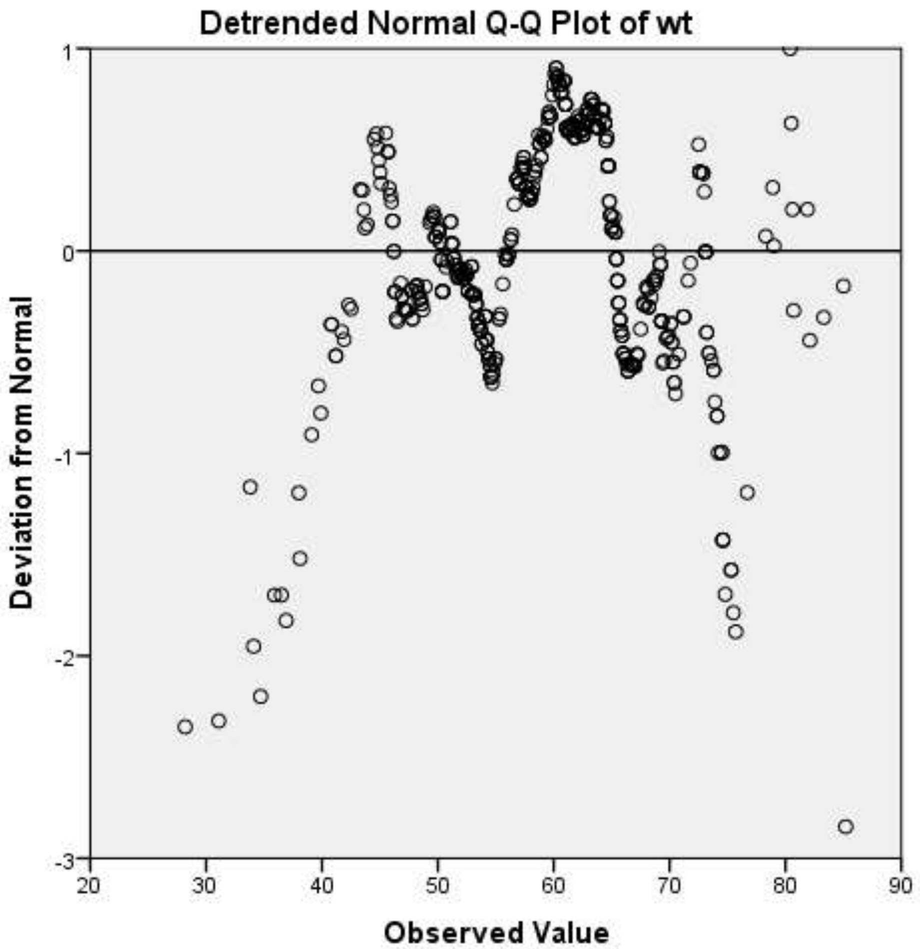
The cases are unweighted.



Normal Q-Q Plot of ht
Expected Normal Value
Observed Value
190
180
170
160
150
140
140
150
160
170
180
190
200


Detrended Normal Q-Q Plot of ht
Deviation from Normal
Observed Value
5
4
3
2
1
0
-1
-2
140
150
160
170
180
190
200


Normal Q-Q Plot of wt
Expected Normal Value
Observed Value
90
80
70
60
50
40
30
20
30
40
50
60
70
80
90




Conclusion: Normal QQ Plot with light tailed distribution. Linearity of points suggest that data is normally distributed.

**6.** Beta (B) = 0.569

Beta* (B*) = 0.43

$$T(n-2) = \frac{B - B}{S.D(B\hat{})}$$

$$= \frac{0.569 - 0.43}{8.54 * 0.509}$$

P value for t-distribution 1 tail at 0.05 % = 0.0001 (highly significant).

This implies that the sensitivities are statistically different.

**7. Regression Output for the htwt_NY.txt dataset**

**Model Summary**

| Model | R | R Square | Adjusted R Square | Std. Error of the Estimate |
|---|---|---|---|---|
| 1 | .409[a] | .167 | .166 | 8.543 |

a. Predictors: (Constant), ht

**ANOVA[a]**

| Model | | Sum of Squares | df | Mean Square | F | Sig. |
|---|---|---|---|---|---|---|
| 1 | Regression | 7299.106 | 1 | 7299.106 | 100.014 | .000[b] |
| | Residual | 36344.614 | 498 | 72.981 | | |
| | Total | 43643.720 | 499 | | | |

a. Dependent Variable: wt

b. Predictors: (Constant), ht

**Coefficients[a]**

| Model | | Unstandardized Coefficients | | Standardized Coefficients | t | Sig. |
|---|---|---|---|---|---|---|
| | | B | Std. Error | Beta | | |
| 1 | (Constant) | -24.125 | 8.409 | | -2.869 | .004 |
| | ht | .519 | .052 | .409 | 10.001 | .000 |

a. Dependent Variable: wt

B = 0.569

B* = 0.43

$$T(n-2) = \frac{B - B^*}{S.D(B)}$$

$$\frac{0.569 - 0.43}{8.54\,(0.569)}$$