Applied Statistics Project

2021/2022

Group #10

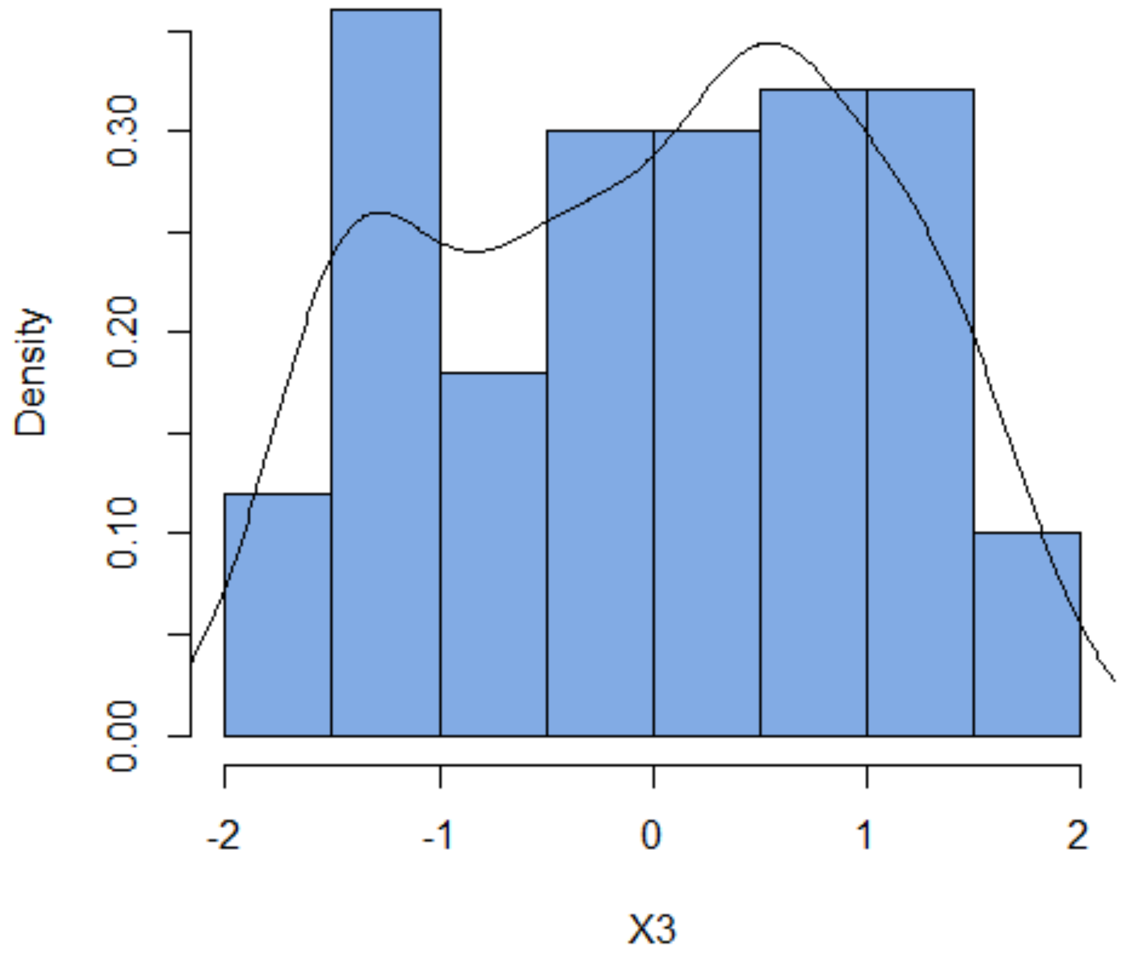
“All models are wrong, but some are useful”

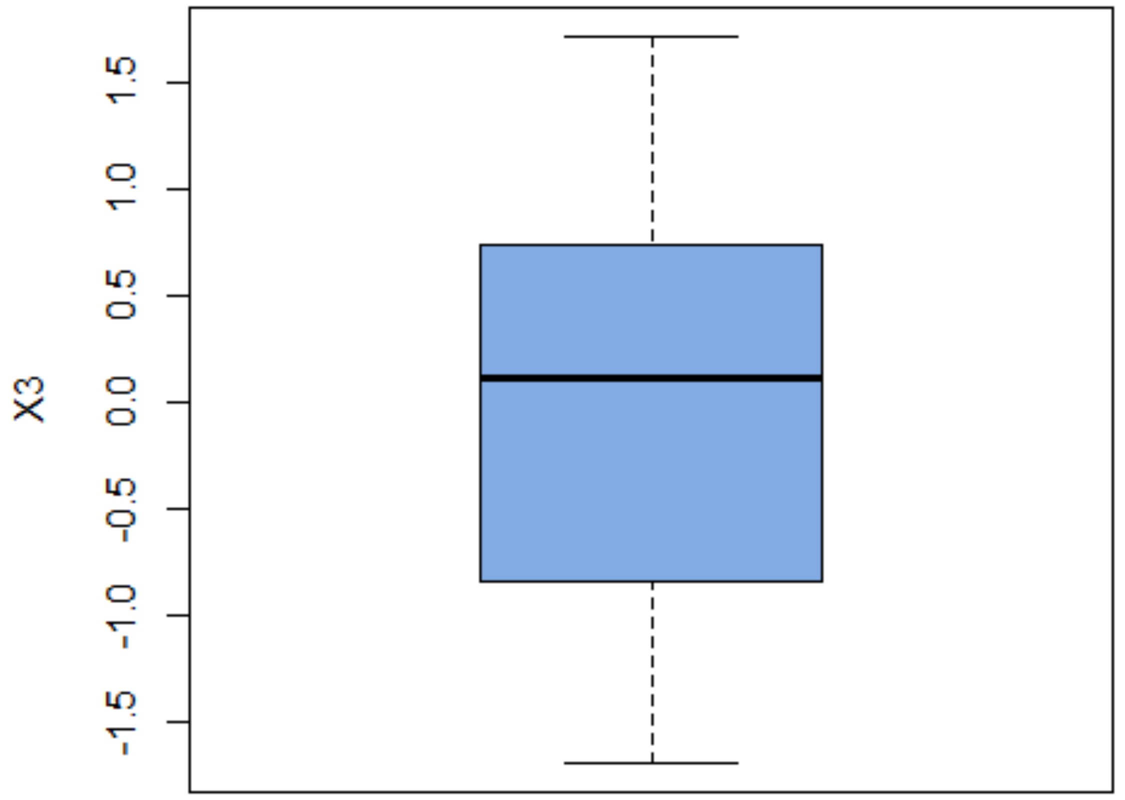
George E. P. Box

1. Data analysis

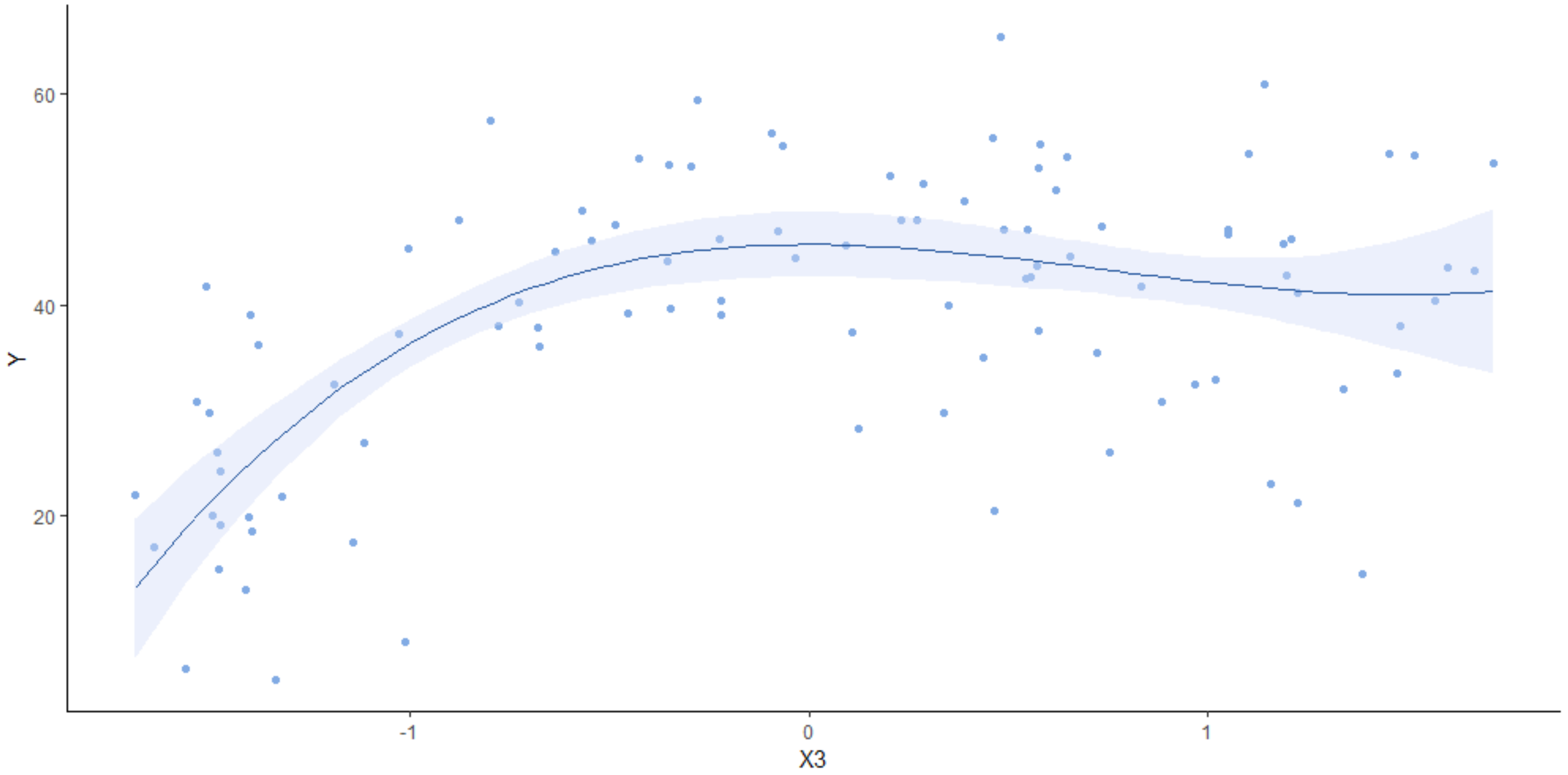
1. Data Analysis

- Variable analysis:
  - Histograms
  - Box-plots





- Polynomial regression analysis:
  - Scatter-Plots
  - ANOVA

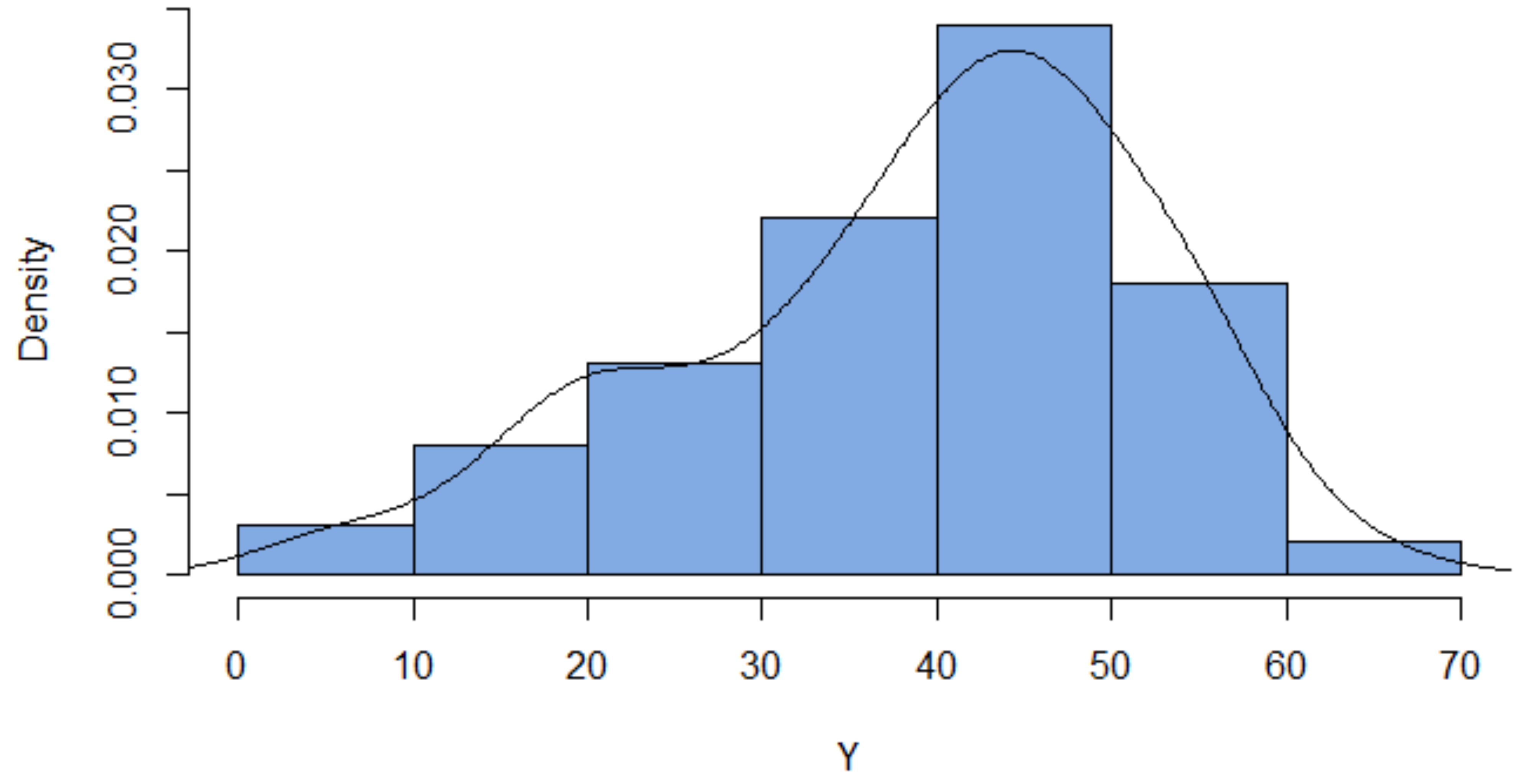


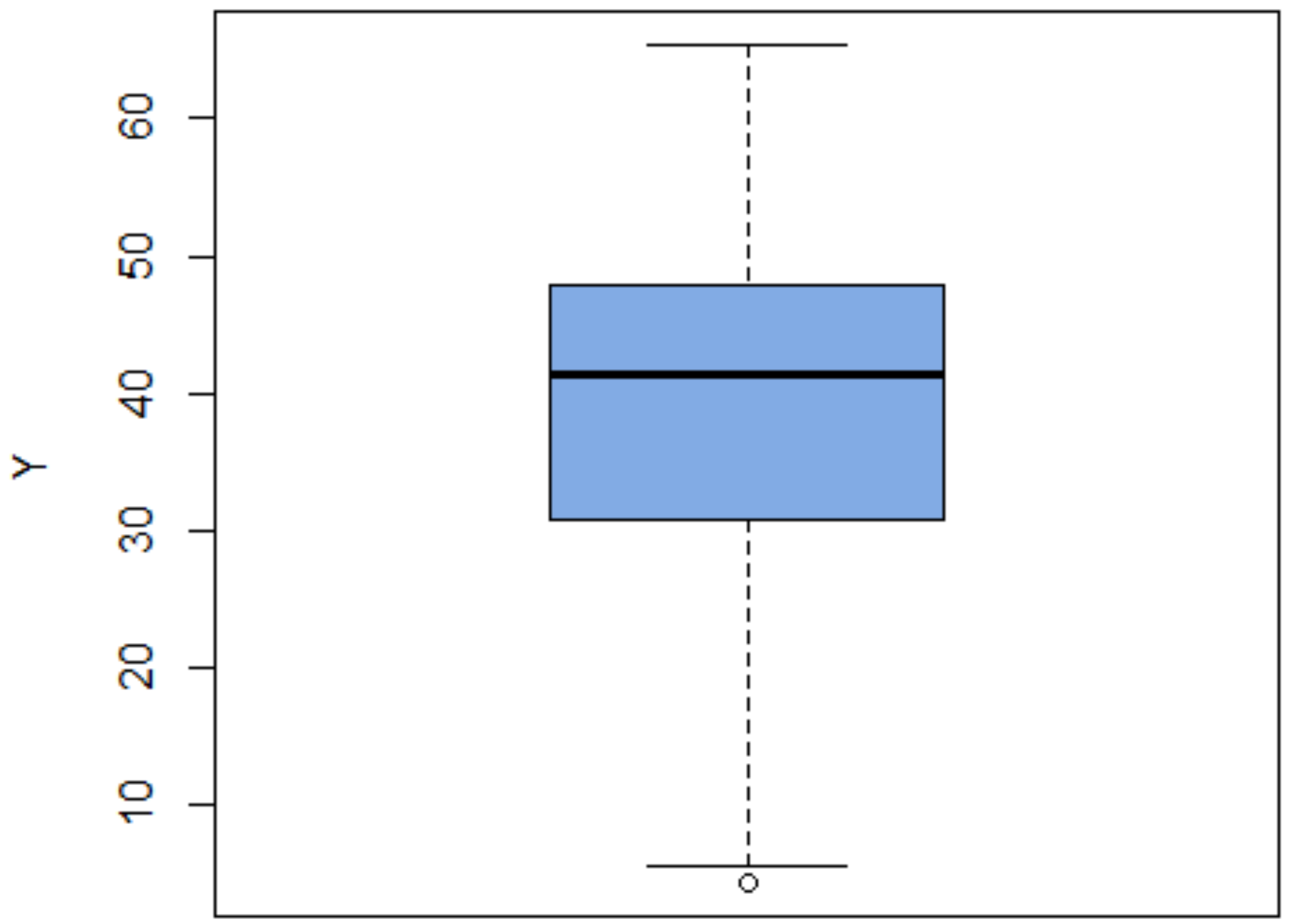
1.0 Response Variable

Y describes the performance of a calculation software

```
> summary(dataset$y)

Min.   1st Qu. Median Mean   3rd Qu. Max.
4.454 30.819 41.413 38.957 47.687 65.336
```



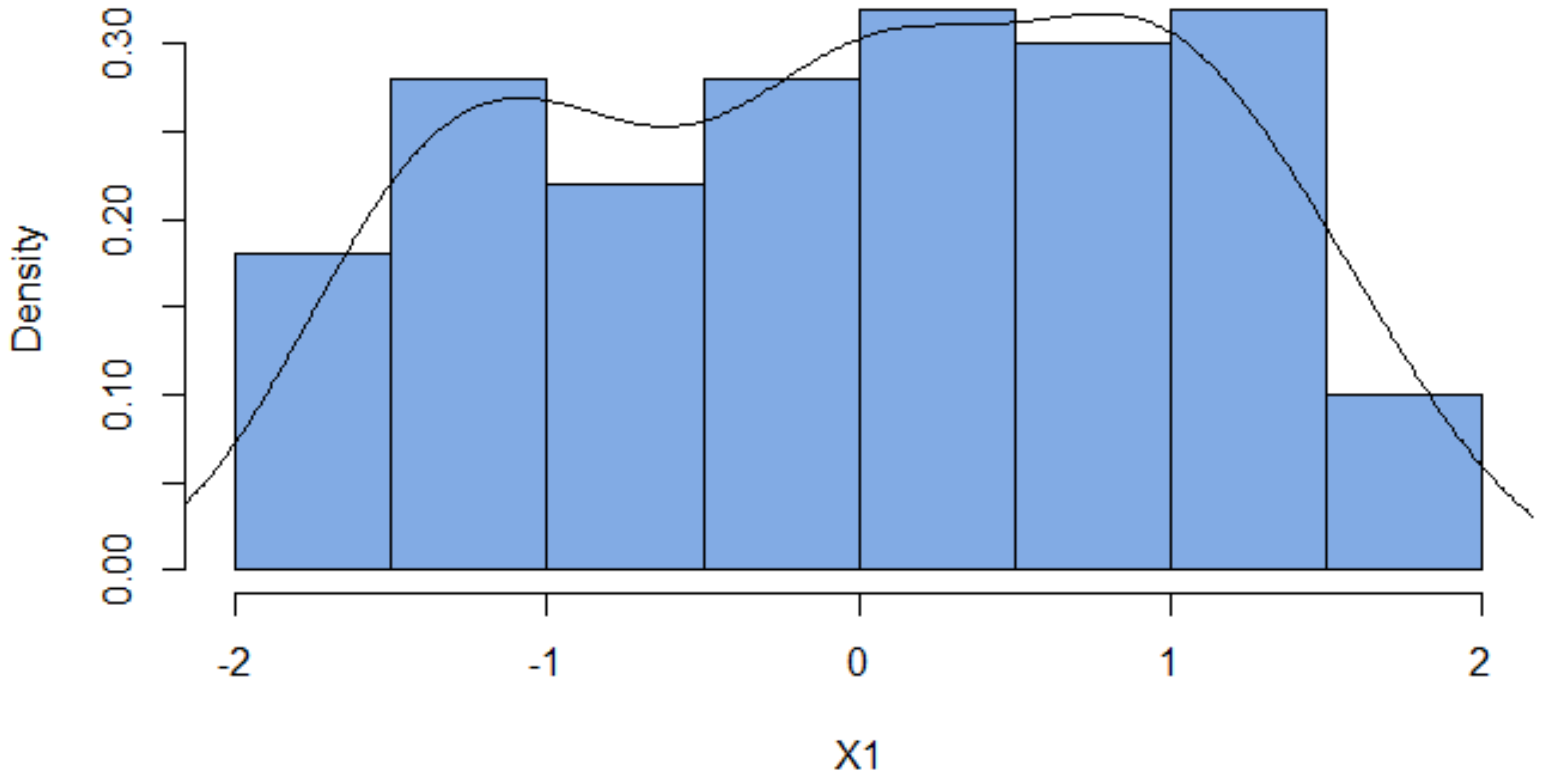


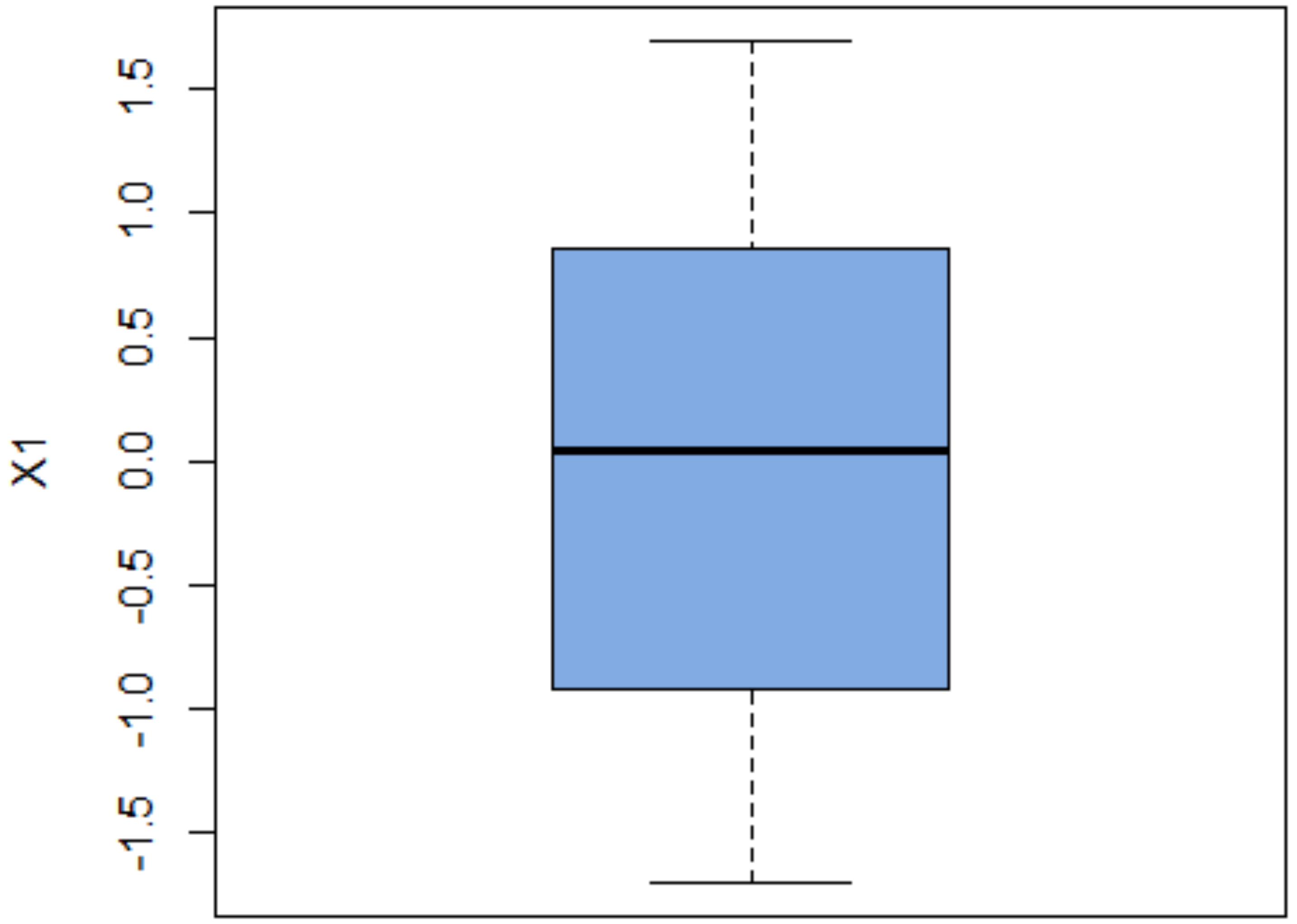
1.1.1 CPU Speed

Data analysis

```
> summary(dataset$x1)

 Min.   1st Qu. Median Mean 3rd Qu. Max.
-1.700 -0.912  0.046  0.000 0.842 1.695
```





1.1.2 CPU Speed

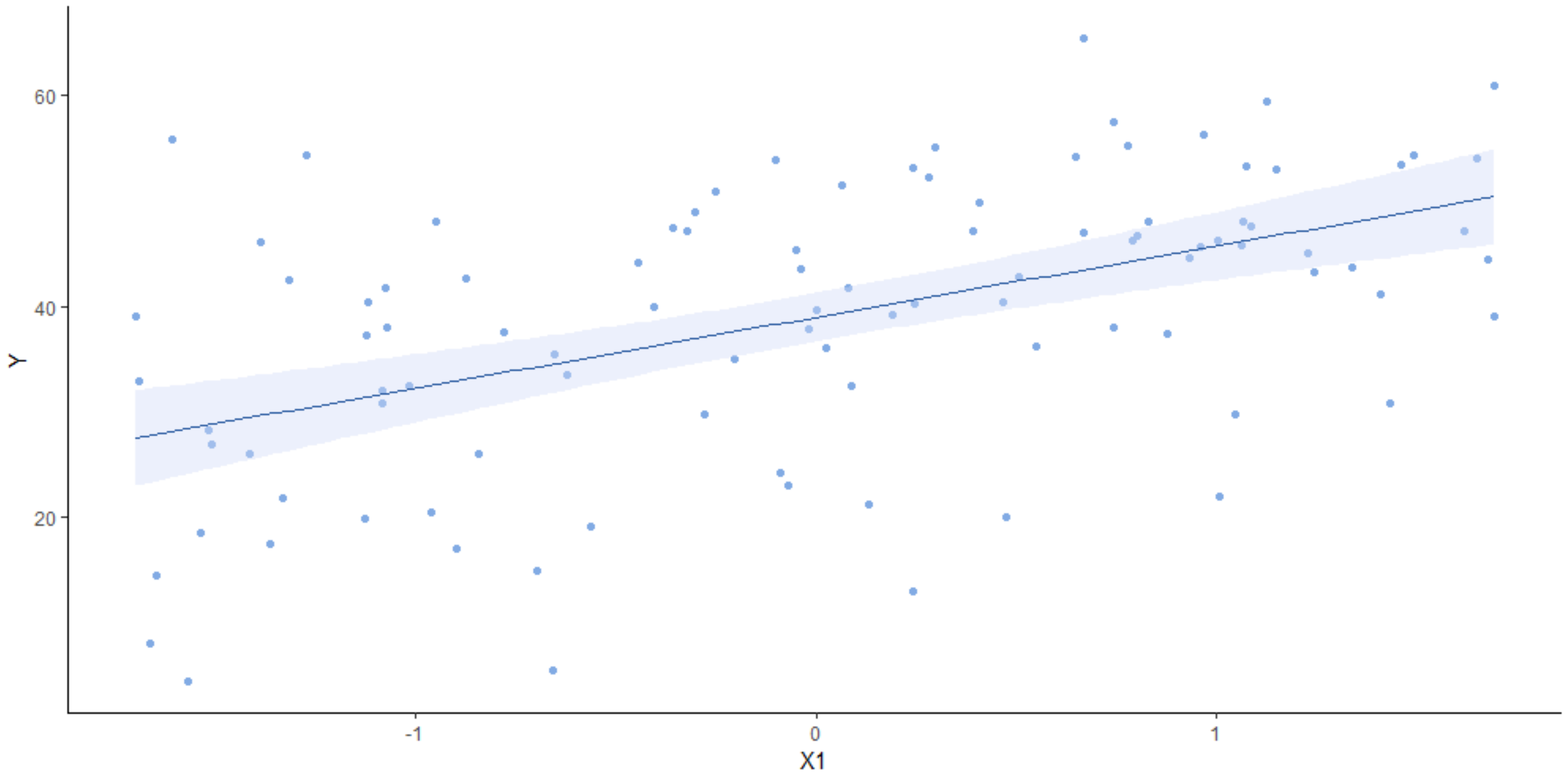
Polynomial regression

$$Y = \beta_0 + \beta_1 X_1 + \varepsilon$$

```
> lm(formula = y ~ x1, data = dataset)

Coefficients:
             Estimate   Std. Error   t value    Pr(>|t|)
(Intercept)  38.957     1.147        33.962     < 2e-16  ***
x1           6.728      1.153        5.836      6.92e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 11.47 on 98 degrees of freedom
Multiple R-squared:  0.2579, Adjusted R-squared:  0.2503
F-statistic: 34.06 on 1 and 98 DF,  p-value: 6.924e-08
```




```
> cor(dataset$x1, dataset$y)
0.5078409
```

1.1.3 CPU Speed

F Test

P-Values lower than the significance level means we can reject the null hypothesis

```
> summary(model_cpu_ftest)

              Df   Sum Sq  Mean Sq   F value    Pr(>F)
x1            1     4481    4481      34.06     6.92e-08 ***
Residuals     98   12895     132
```

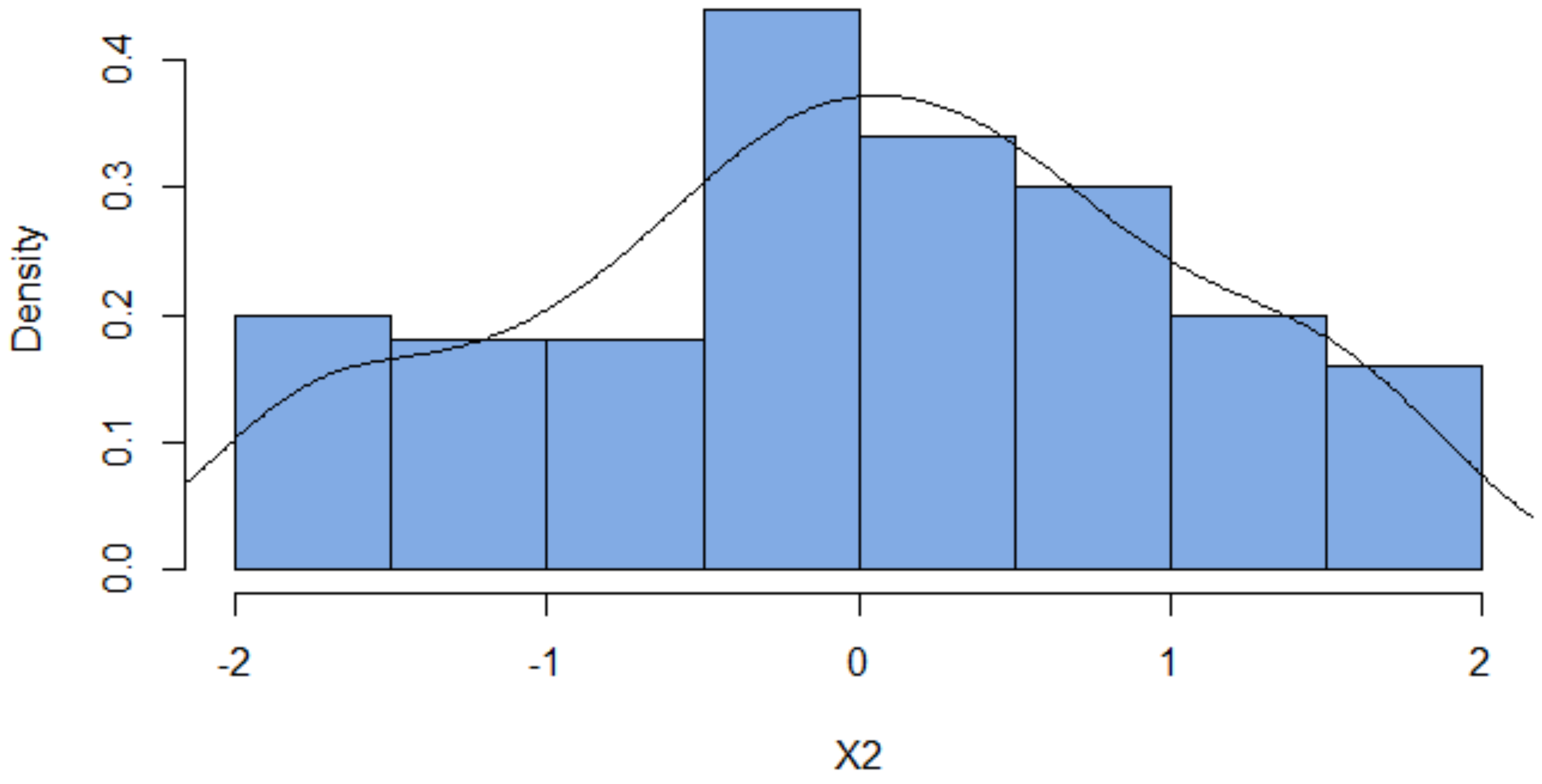
$$\begin{cases} H_0 : \beta_1 = 0 \\ H_A : \beta_1 \neq 0 \end{cases}$$

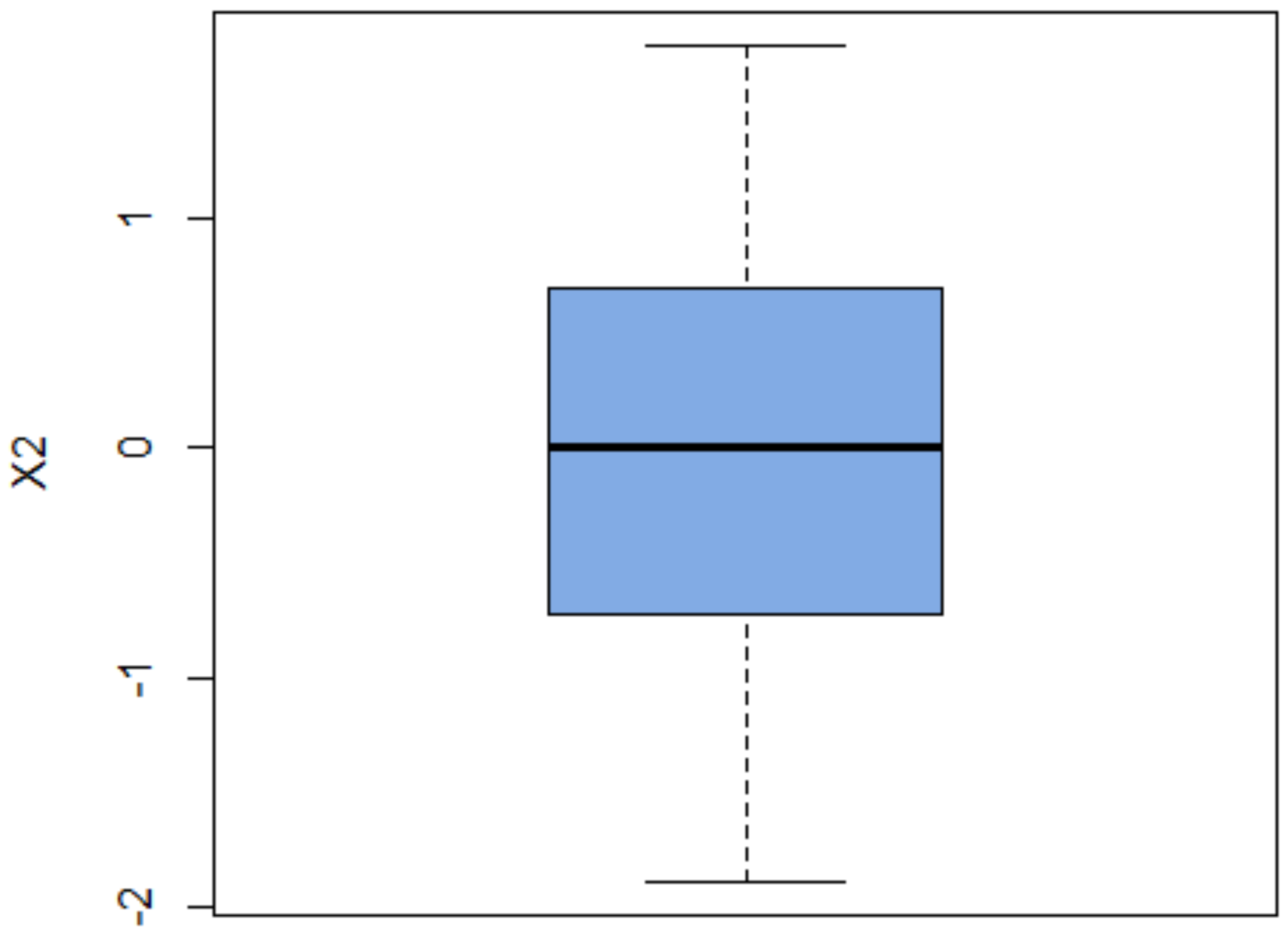
1.2.1 HDD Capacity

Data analysis

```
> summary(dataset$x2)

 Min.   1st Qu. Median Mean 3rd Qu. Max.
-1.889 -0.720  0.001  0.000 0.687 1.754
```





1.2.2 HDD Capacity

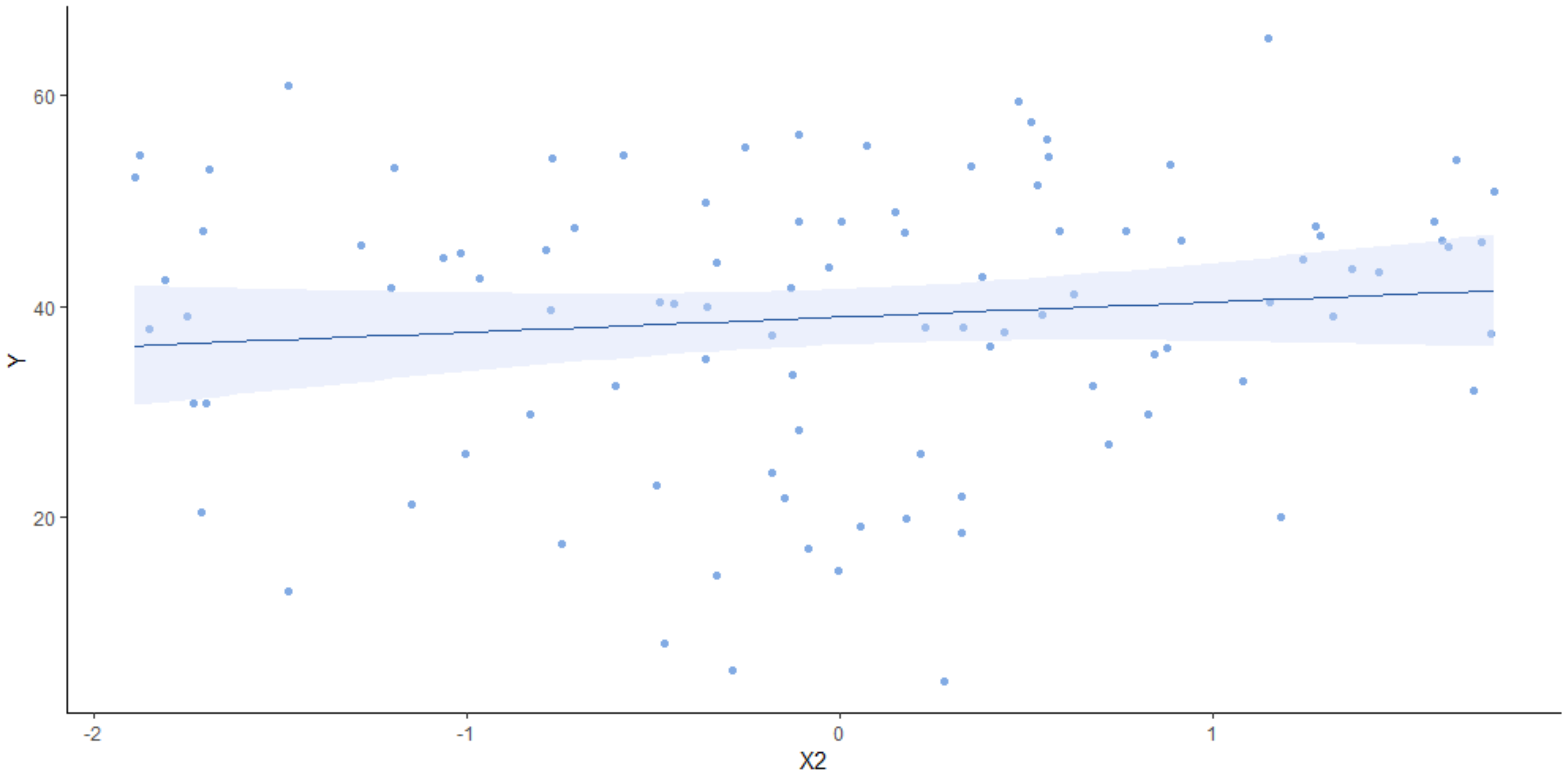
Polynomial regression

$$Y = \beta_0 + \beta_1 X_2 + \varepsilon$$

```
> lm(formula = y ~ x2, data = dataset)

Coefficients:
             Estimate   Std. Error   t value    Pr(>|t|)
(Intercept)  38.957     1.324        29.429     <2e-16     ***
x2            1.433     1.330         1.077      0.284
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 13.24 on 98 degrees of freedom
Multiple R-squared:  0.01169, Adjusted R-squared:  0.001609
F-statistic:  1.16 on 1 and 98 DF,  p-value: 0.2842
```




```
> cor(dataset$x2, dataset$y)
0.1081392
```

1.2.3 HDD Capacity

F Test

P-Values higher than the significance level means we fail to reject the null hypothesis

```
> summary(model_hdd_ftest)

              Df    Sum Sq  Mean Sq    F value    Pr(>F)
x2            1      203     203.2      1.16      0.284
Residuals     98    17173    175.2
```

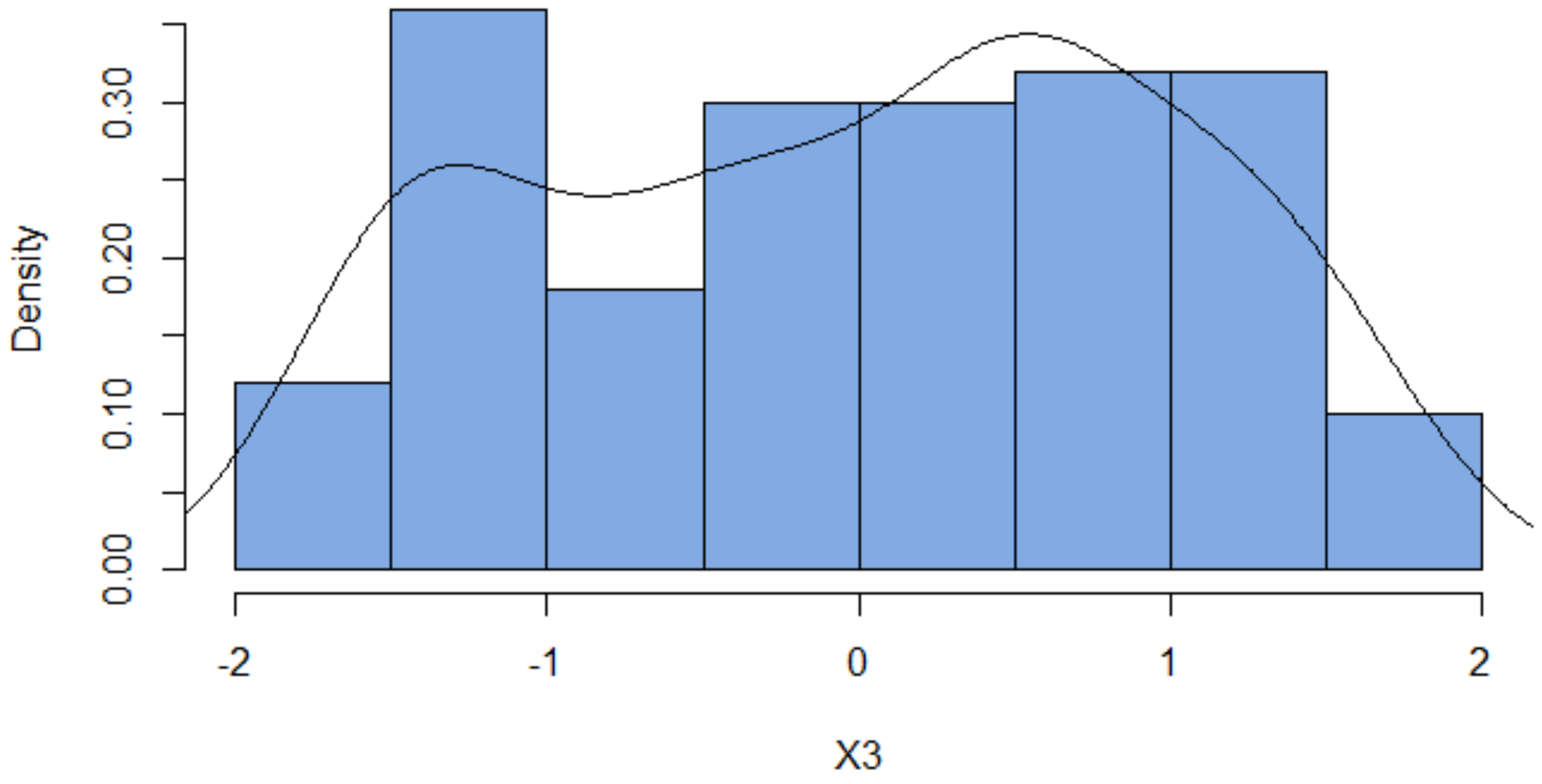
$$\begin{cases} H_0 : \beta_1 = 0 \\ H_A : \beta_1 \neq 0 \end{cases}$$

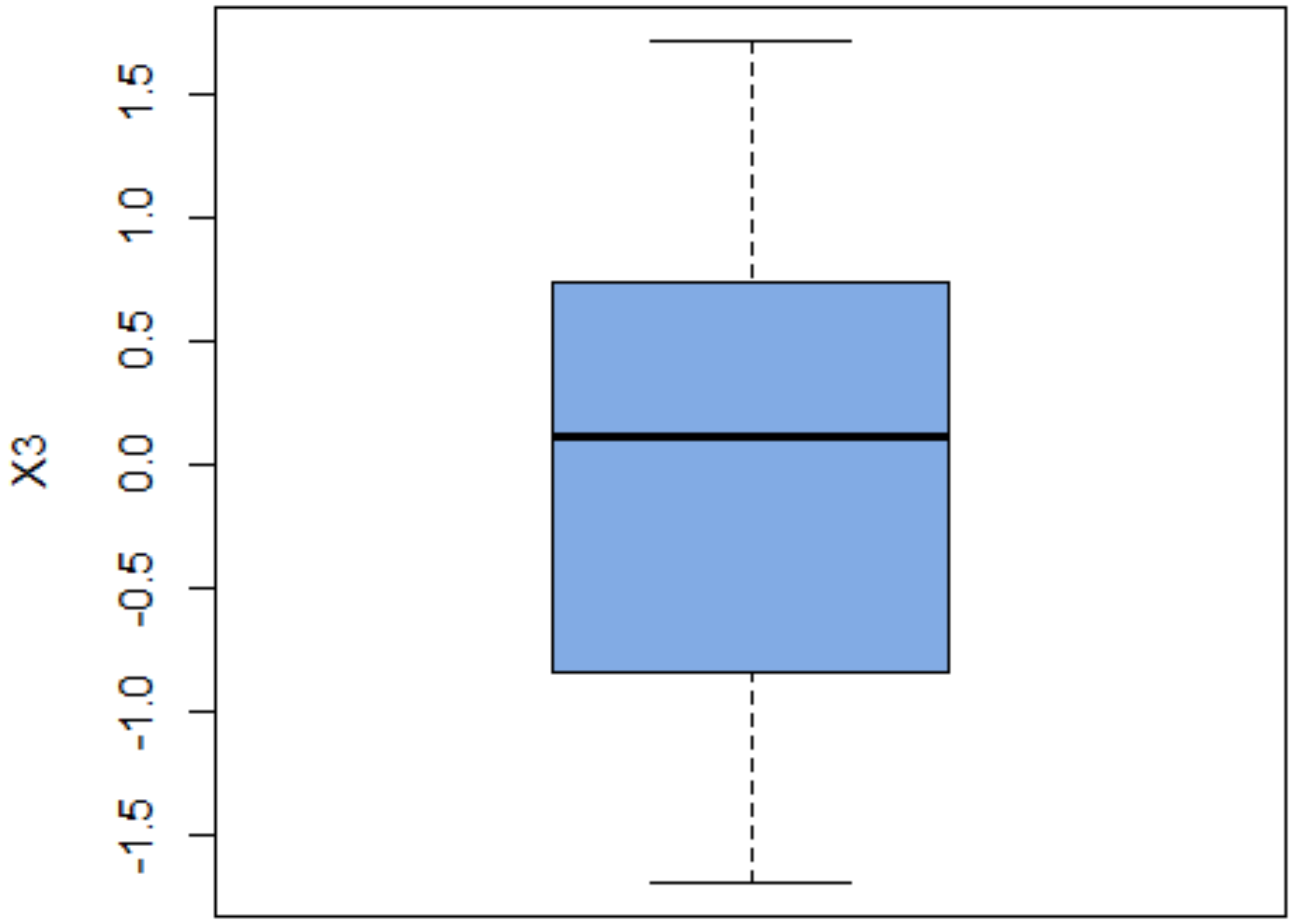
1.3.1 Number of tasks

Data analysis

```
> summary(dataset$x3)

 Min.   1st Qu. Median Mean 3rd Qu. Max.
-1.691 -0.818  0.116  0.000 0.739 1.716
```





1.3.2 Number of tasks

Polynomial regression

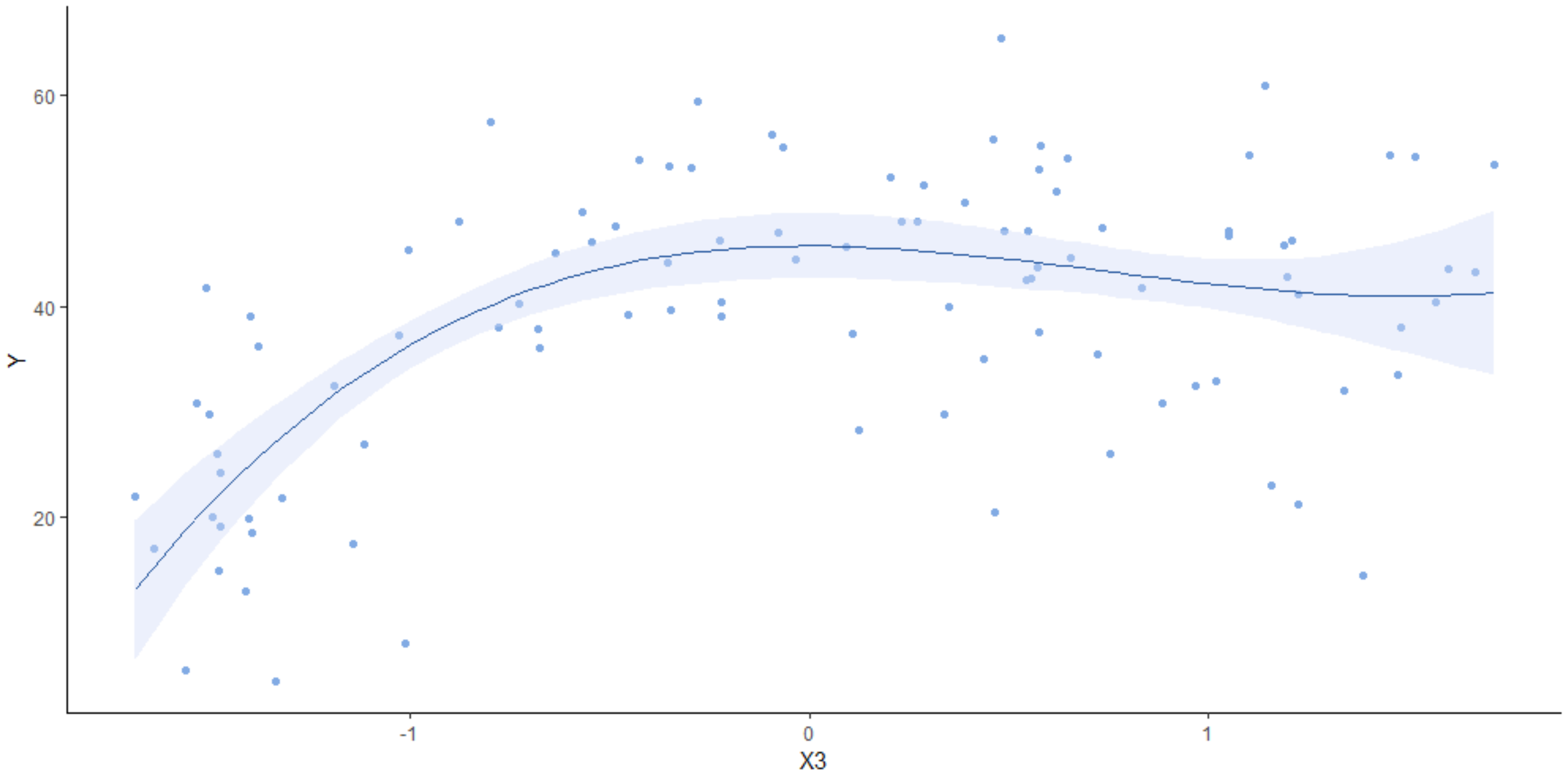
$$Y = \beta_0 + \beta_1 X_3 + \beta_2 X_3^2 + \beta_3 X_3^3 + \varepsilon$$

```
> lm(formula = y ~ x3 + I(x3^2) + I(x3^3),
data = dataset)
```

Coefficients:

	Estimate	Std. Error	t value	Pr(>ltl)	
(Intercept)	45.788	1.565	29.255	< 2e-16	***
x3	-1.123	2.734	-0.411	0.6821	
I(x3^2)	-6.508	1.174	-5.544	2.59e-07	***
I(x3^3)	3.441	1.413	2.435	0.0168	*

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1




```
> cor(dataset$x3, dataset$y)
0.4348121
```

1.3.3 Number of tasks

$$Y = \beta_0 + \beta_1 X_3^2 + \beta_2 X_3^3 + \varepsilon$$

```
> lm(formula = y ~ I(x3^2) + I(x3^3), data
= dataset)

Coefficients:
              Estimate   Std. Error   t value   Pr(>|t|)
(Intercept)   45.6951    1.5421       29.631    < 2e-16    ***
I(x3^2)       -6.4751    1.1661       -5.553    2.45e-07   ***
I(x3^3)        2.9046    0.5387        5.392    4.90e-07   ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 10.28 on 97 degrees of freedom
Multiple R-squared: 0.4098, Adjusted R-squared: 0.3976
F-statistic: 33.67 on 2 and 97 DF, p-value: 7.832e-12
```

P-Values lower than than significance level

means we can reject the null hypothesis

```
> summary(model_proc_ftest)

            Df   Sum Sq  Mean Sq   F value   Pr(>F)
I(x3^2)      1    4046     4046     38.27    1.47e-08 ***
I(x3^3)      1    3074     3074     29.08    4.90e-07 ***
Residuals   97   10256      106
```

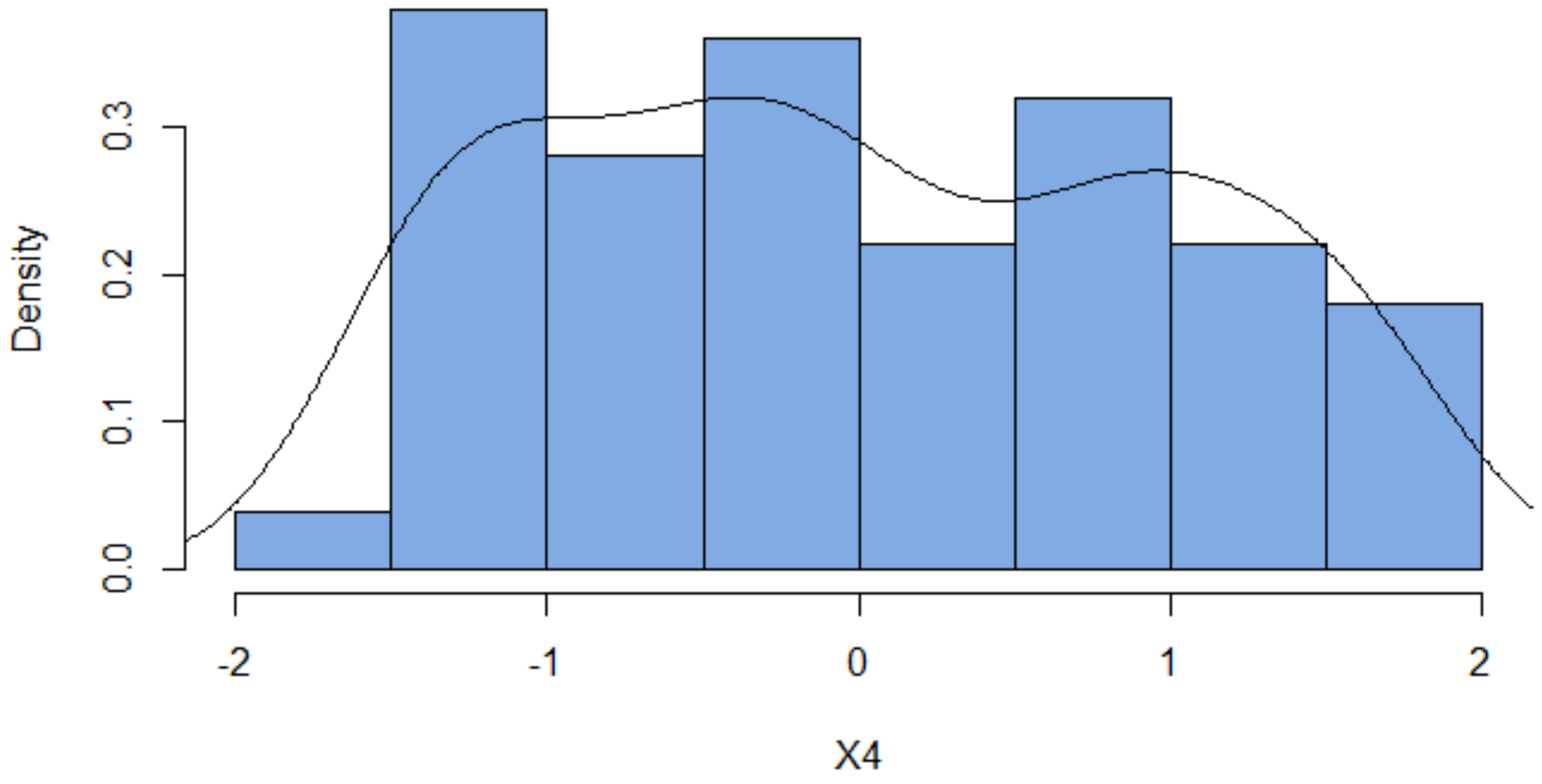
$$\begin{cases} H_0 : \beta_1 = \beta_2 = 0 \\ H_A : \exists \beta_i \neq 0 \end{cases}$$

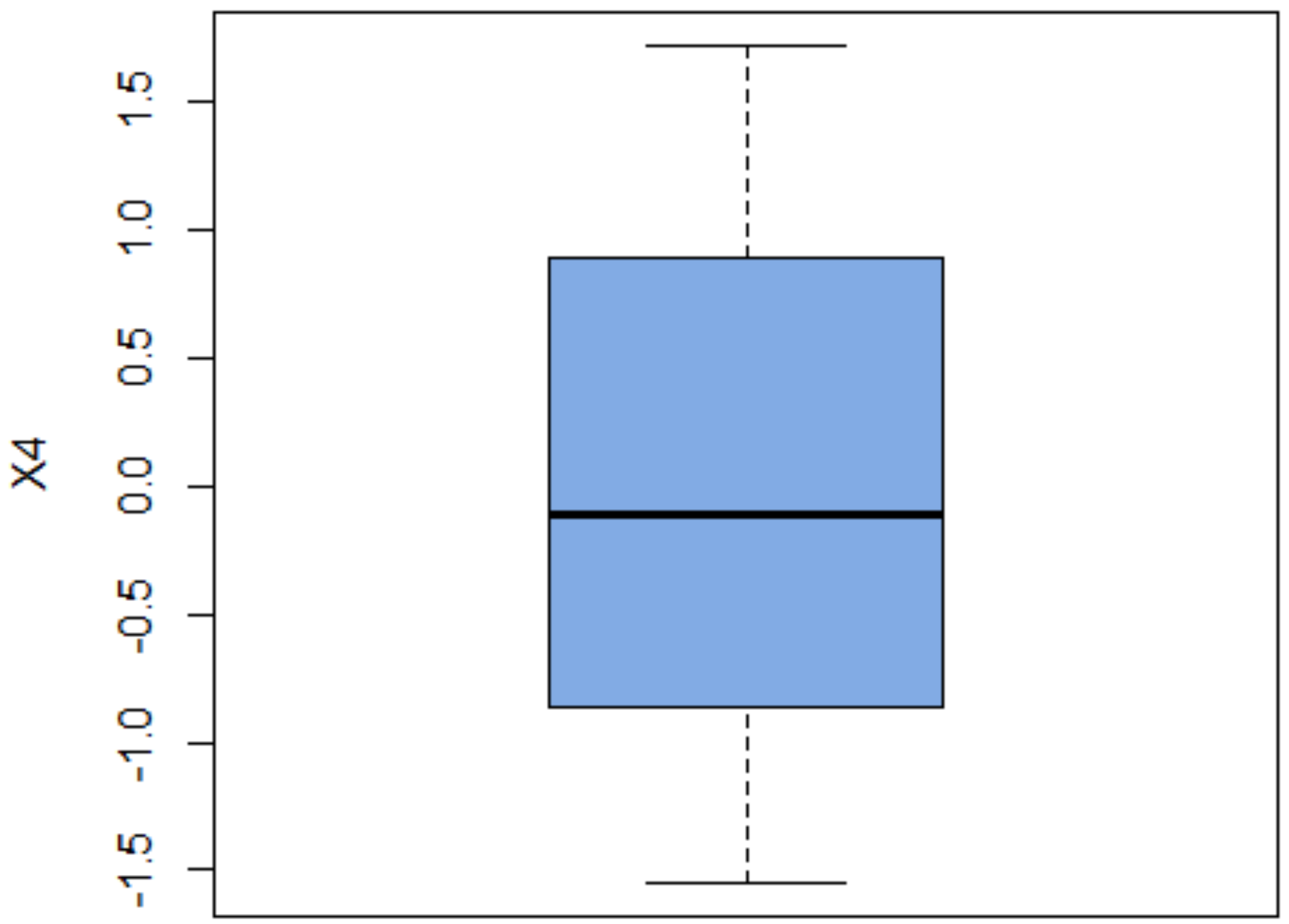
1.4.1 Aging

Data analysis

```
> summary(dataset$x4)

 Min.   1st Qu. Median Mean 3rd Qu. Max.
-1.546 -0.840  -0.109 0.000 0.890 1.715
```





1.4.2 Aging

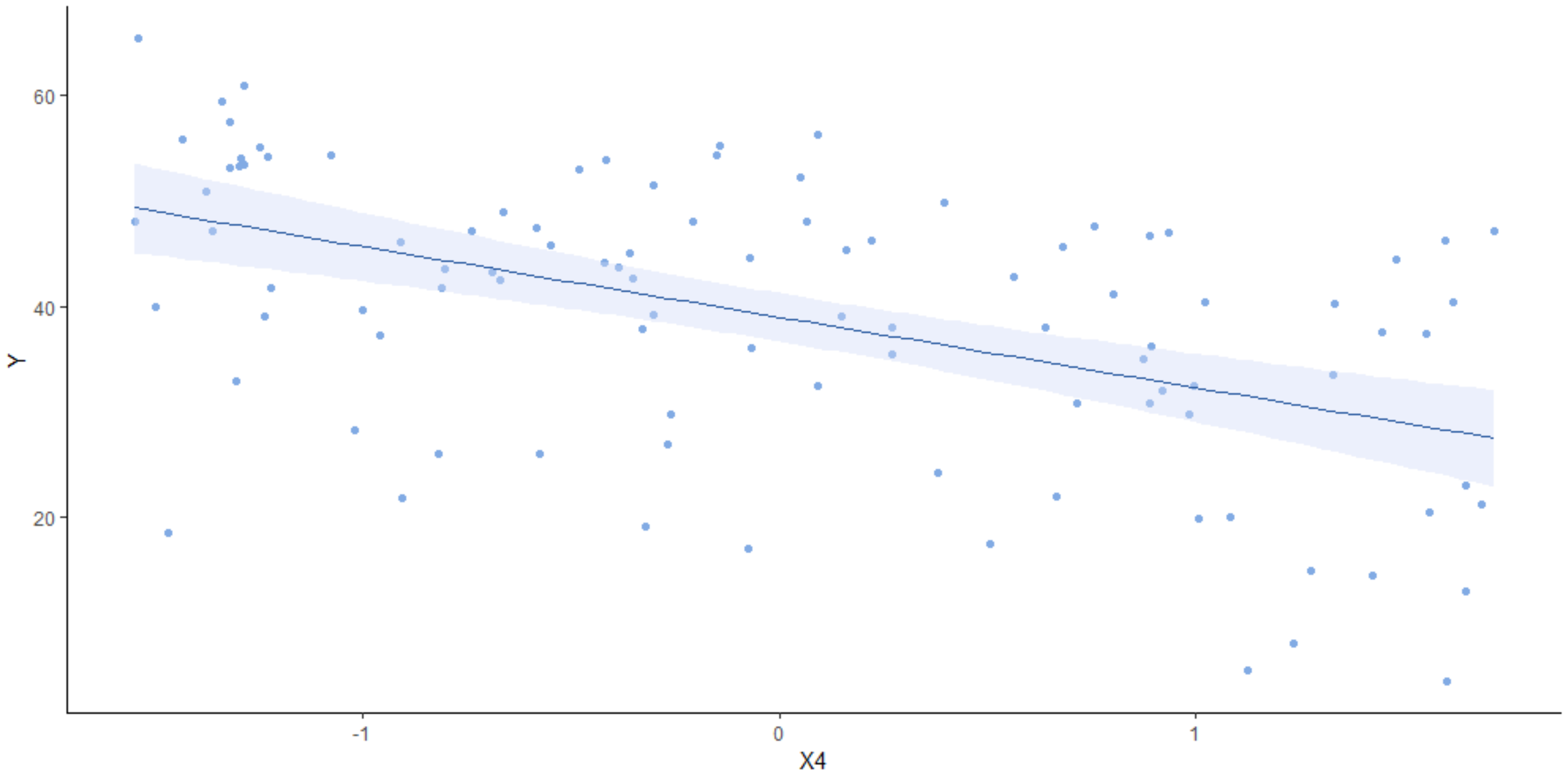
Polynomial regression

$$Y = \beta_0 + \beta_1 X_4 + \varepsilon$$

```
> lm(formula = y ~ x4, data = dataset)

Coefficients:
             Estimate   Std. Error   t value    Pr(>|t|)
(Intercept)  38.957     1.150        33.888     < 2e-16   ***
x4           -6.686     1.155        -5.787     8.62e-08  ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 11.5 on 98 degrees of freedom
Multiple R-squared:  0.2547, Adjusted R-squared:  0.2471
F-statistic: 33.48 on 1 and 98 DF,  p-value: 8.621e-08
```




```
> cor(dataset$x4, dataset$y)
-0.5046424
```

1.4.3 Aging

F Test

P-Values lower than the significance level means we can reject the null hypothesis

```
> summary(model_aging_ftest)

              Df   Sum Sq  Mean Sq   F value    Pr(>F)
x4             1    4425     4425    33.48     8.62e-08 ***
Residuals     98   12951      132
```

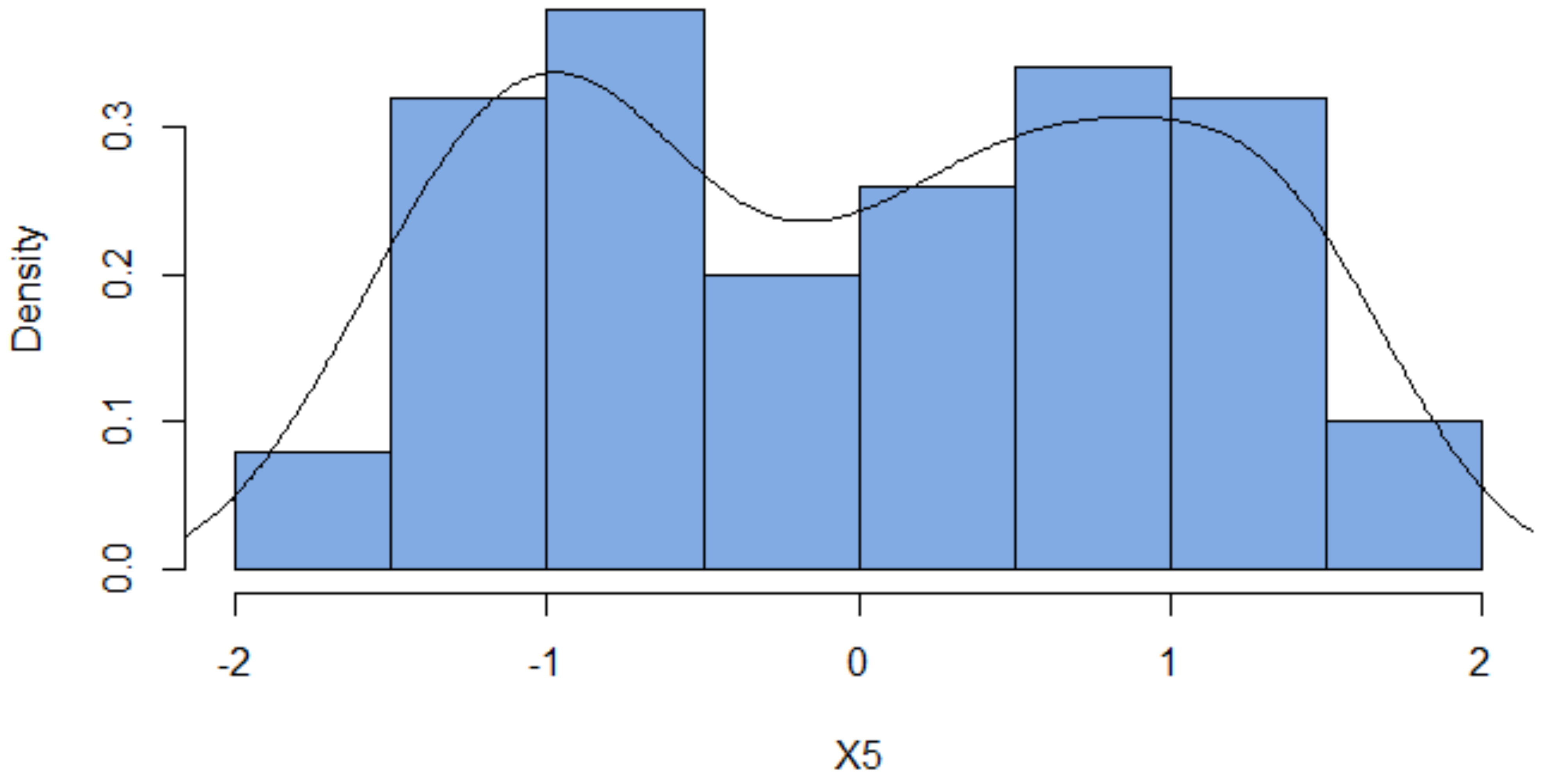
$$\begin{cases} H_0 : \beta_1 = 0 \\ H_A : \beta_1 \neq 0 \end{cases}$$

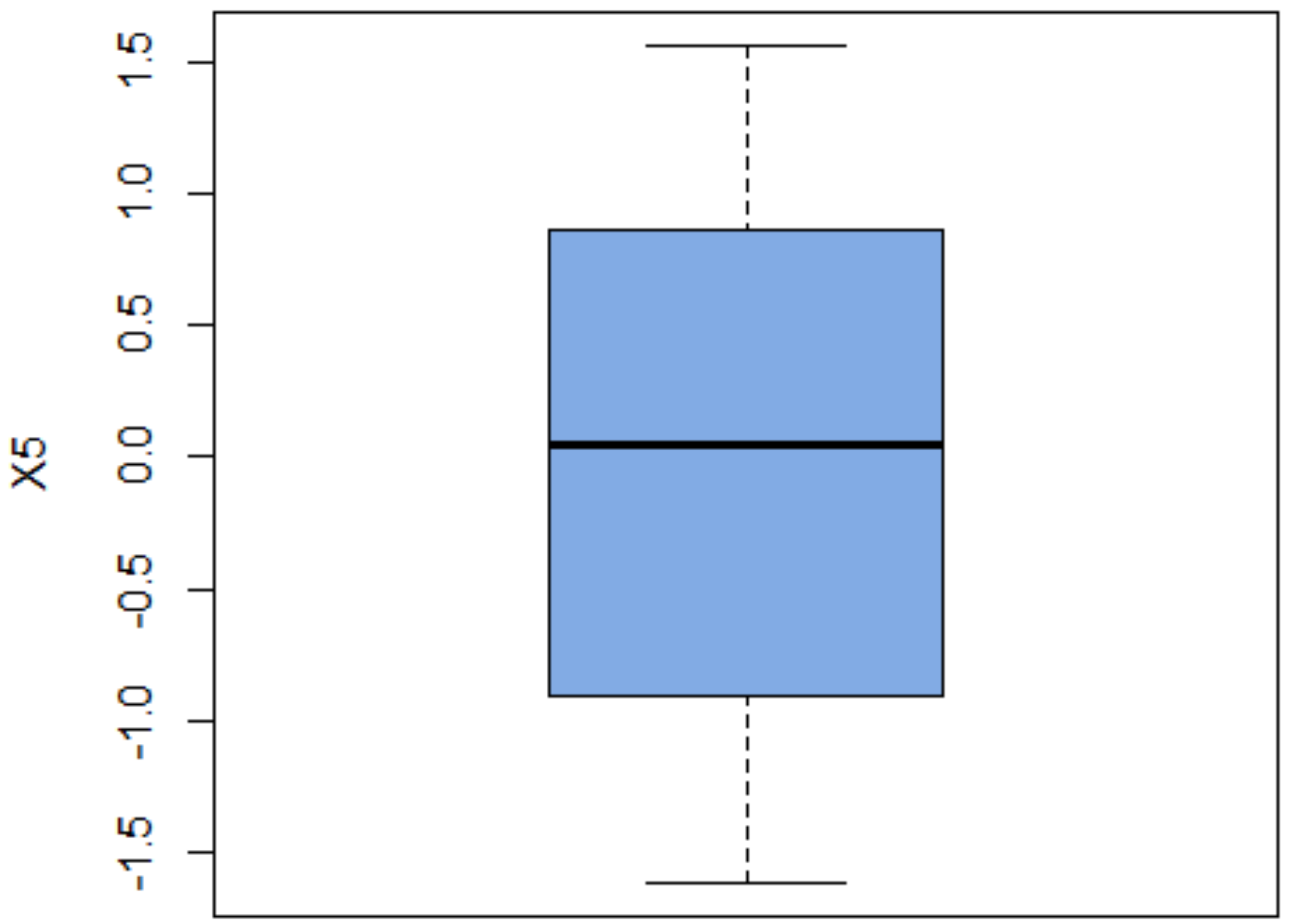
1.5.1 Sound card performance

Data analysis

```
> summary(dataset$x5)

 Min.   1st Qu. Median Mean 3rd Qu. Max.
-1.614 -0.902  0.051  0.000 0.851 1.562
```





1.5.2 Sound card performance

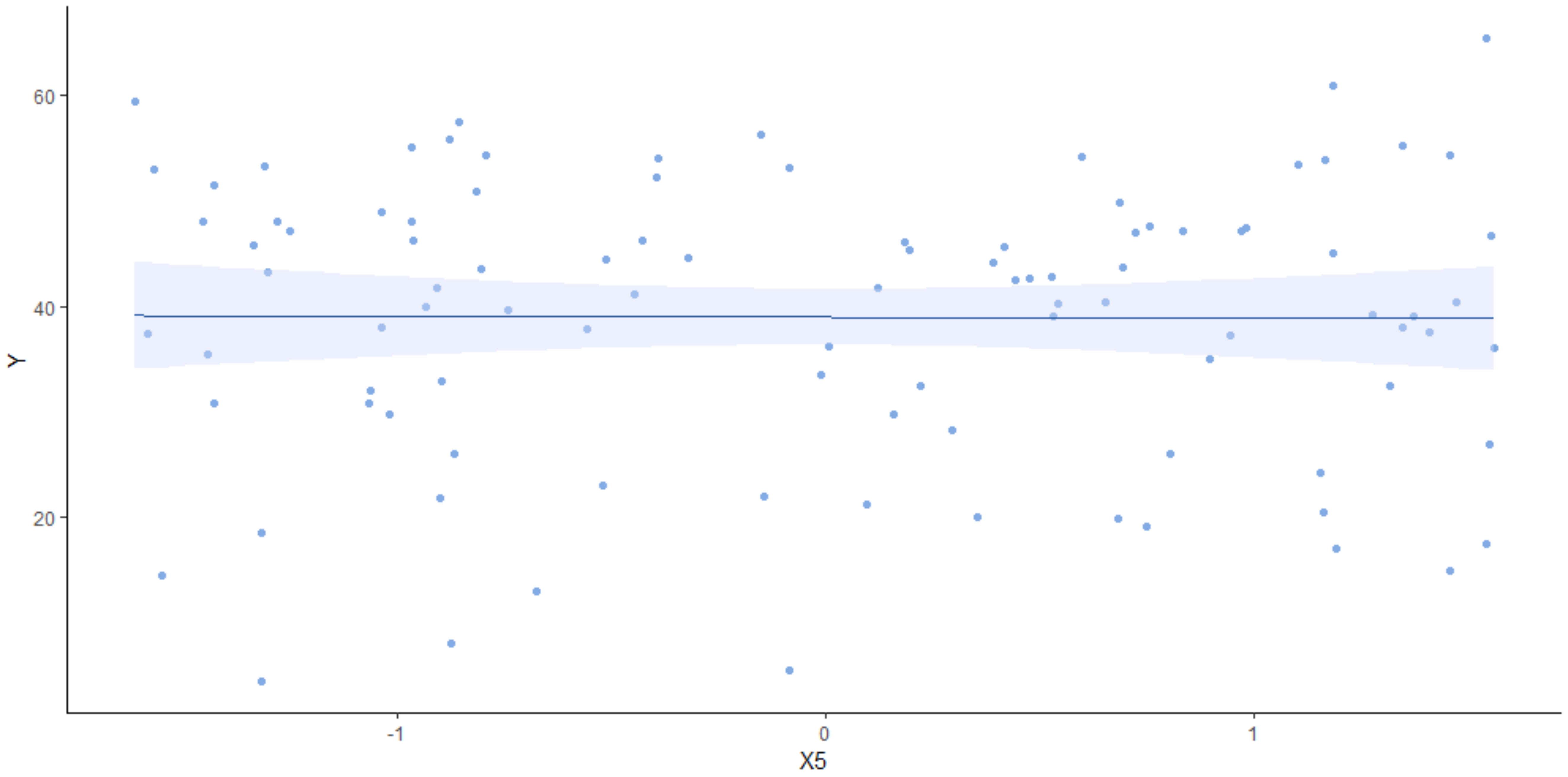
Polynomial regression

$$Y = \beta_0 + \beta_1 X_5 + \varepsilon$$

```
> lm(formula = y ~ x5, data = dataset)

Coefficients:
             Estimate   Std. Error  t value    Pr(>ltl)
(Intercept)  38.95731   1.33153     29.26      <2e-16    ***
x5           -0.09379   1.33824     -0.07      0.944
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 13.32 on 98 degrees of freedom
Multiple R-squared:  5.012e-05, Adjusted R-squared: -0.01015
F-statistic: 0.004912 on 1 and 98 DF,  p-value: 0.9443
```




```
> cor(dataset$x5, dataset$y)
-0.007079301
```

1.5.3 Sound card performance

F Test

P-Values higher than the significance level means we fail to reject the null hypothesis

```
> summary(model_audio_ftest)

               Df   Sum Sq  Mean Sq   F value   Pr(>F)
x5              1        1     0.87     0.005    0.944
Residuals      98    17375   177.30
```

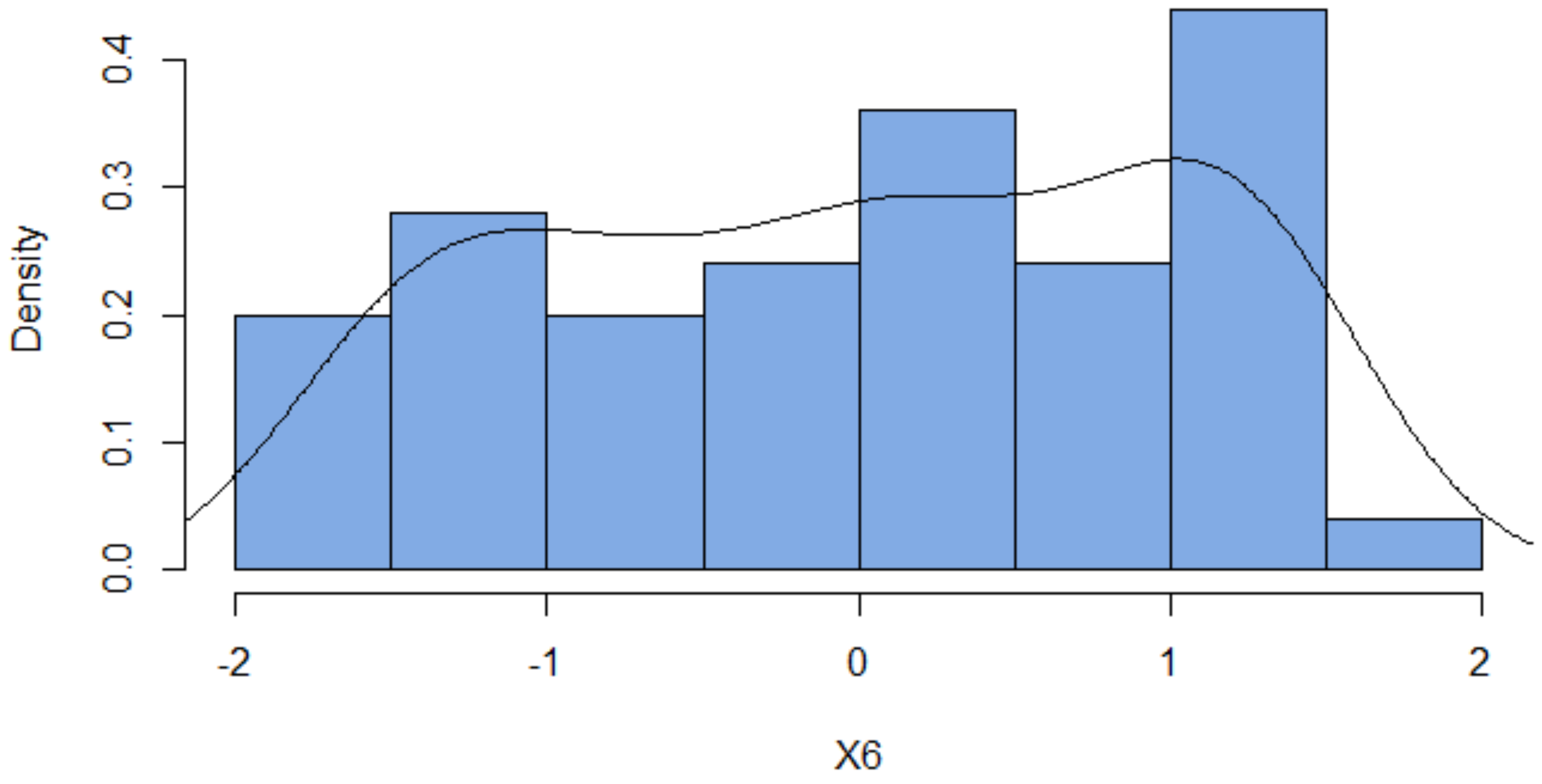
$$\begin{cases} H_0 : \beta_1 = 0 \\ H_A : \beta_1 \neq 0 \end{cases}$$

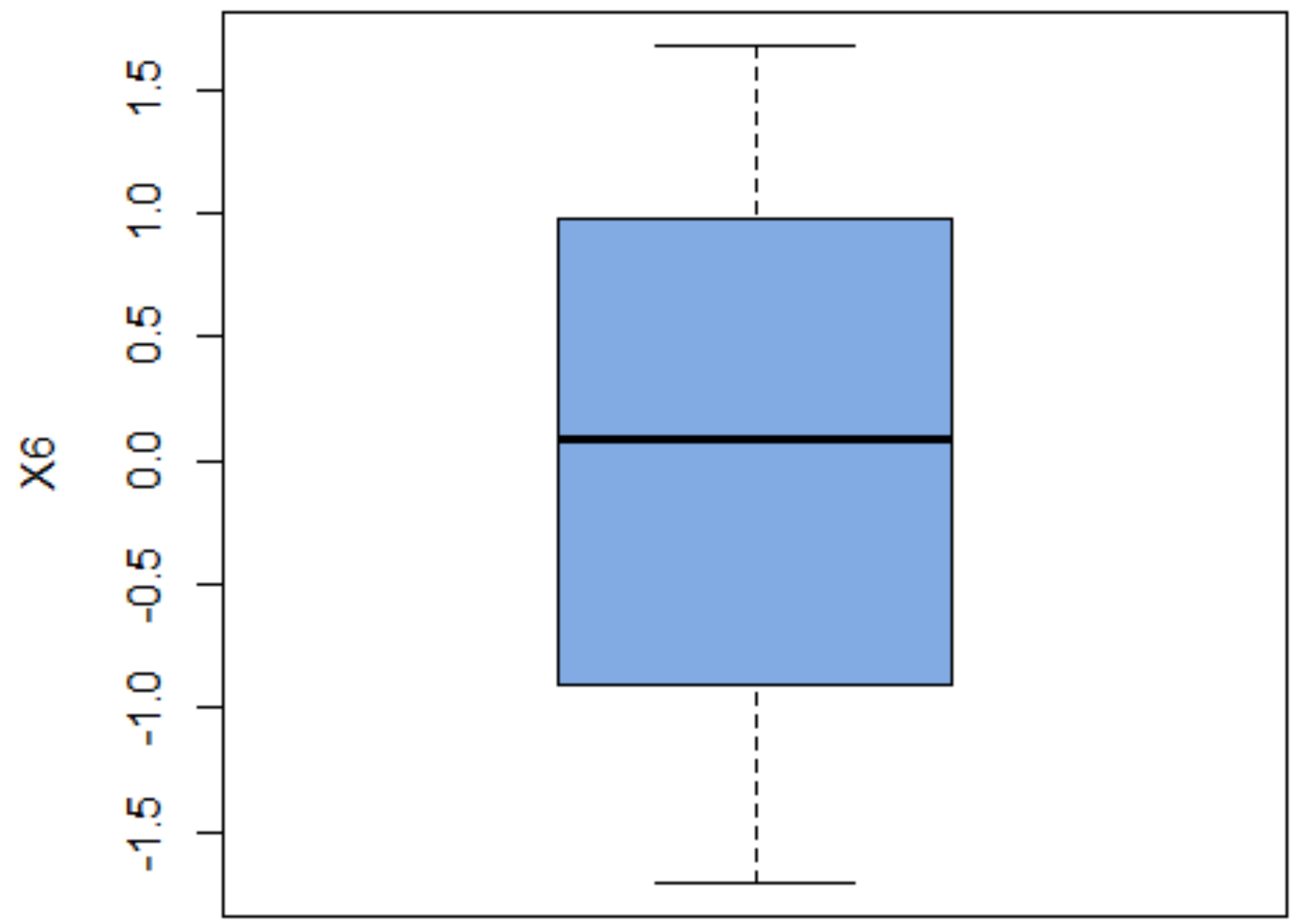
1.6.1 RAM Performance

Data analysis

```
> summary(dataset$x6)

 Min.   1st Qu. Median Mean 3rd Qu. Max.
-1.700 -0.901  0.091  0.000 0.982 1.675
```





1.6.2 RAM Performance

Polynomial regression

$$Y = \beta_0 + \beta_1 X_6 + \beta_2 X_6^2 + \varepsilon$$

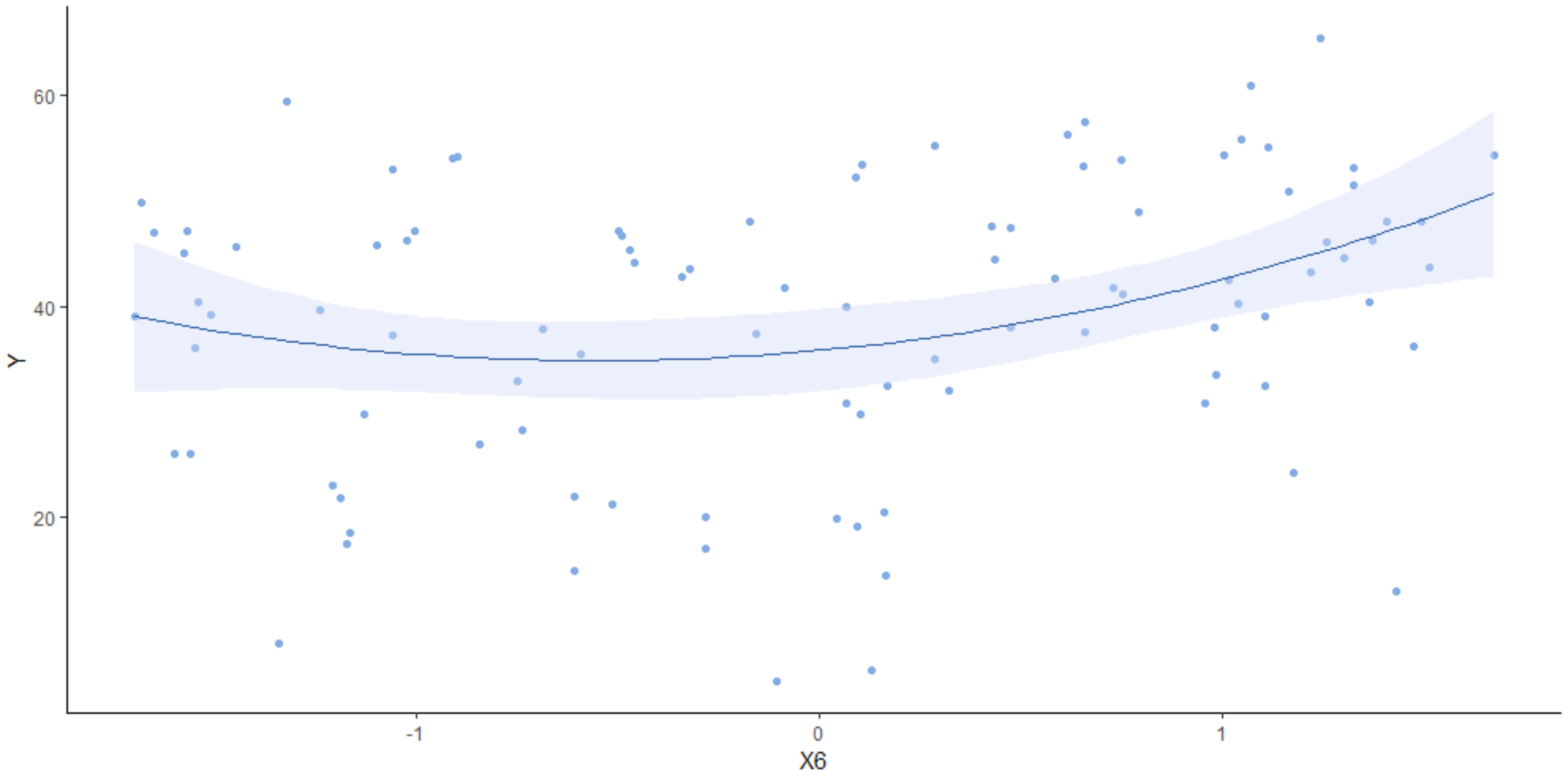
```
> lm(formula = y ~ x6 + I(x6^2), data =
dataset)
```

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	35.819	1.960	18.275	< 2e-16	***
x6	3.526	1.290	2.734	0.00744	**
I(x6^2)	3.170	1.507	2.104	0.03797	*

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 12.71 on 98 degrees of freedom
Multiple R-squared: 0.09781, Adjusted R-squared: 0.07921
F-statistic: 5.258 on 2 and 97 DF, p-value: 0.00679




```
> cor(dataset$x6, dataset$y)
0.2379949
```

1.6.3 RAM Performance

F Test

P-Values lower than the significance level means we can reject the null hypothesis

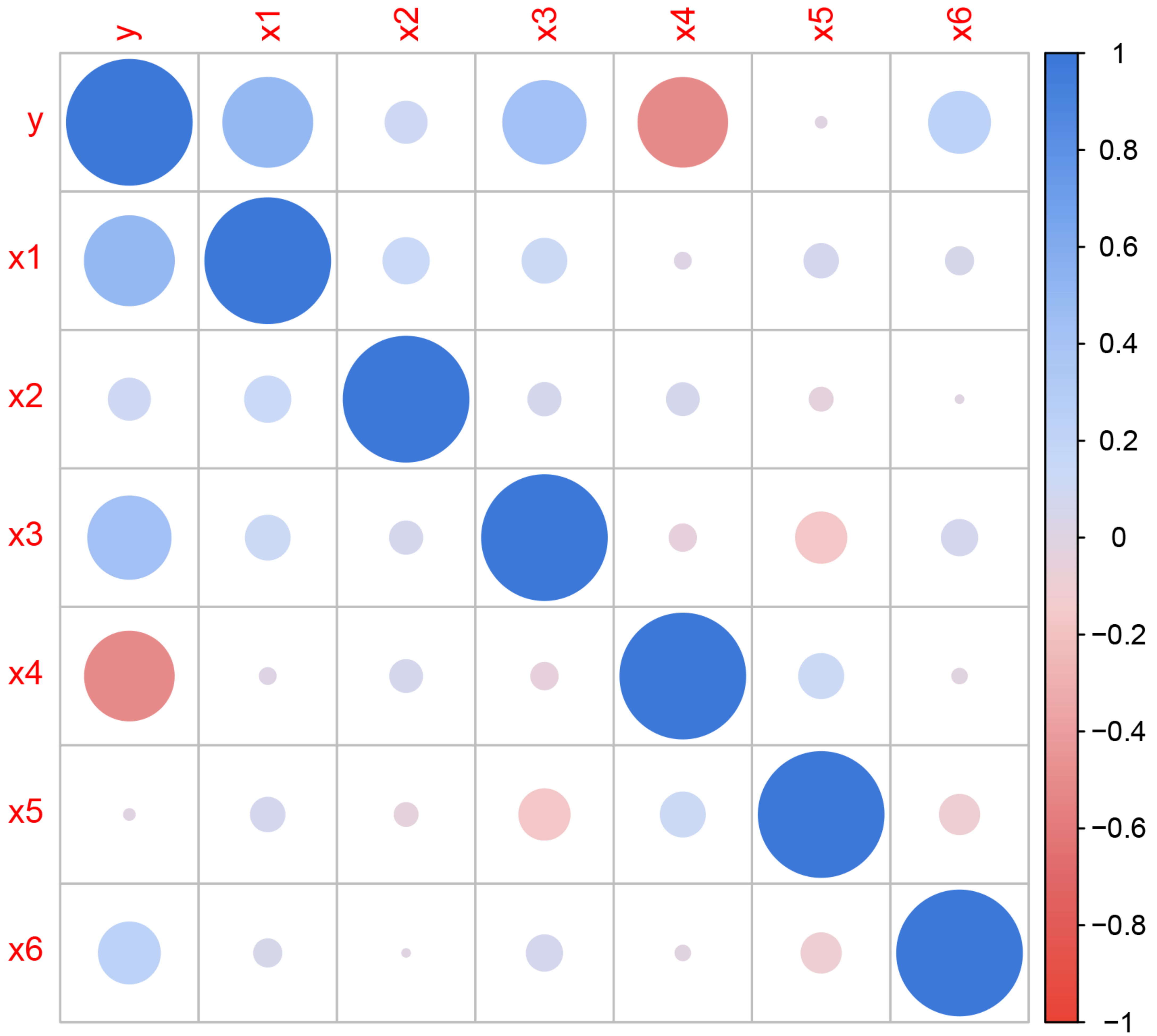
```
> summary(model_ram_ftest)

            Df  Sum Sq  Mean Sq  F value  Pr(>F)
x6           1     984    984.2    6.090  0.0153 *
I(x6^2)      1     715    715.4    4.427  0.0380 *
Residuals   97   15676    161.6
```

$$\begin{cases} H_0 : \beta_1 = \beta_2 = 0 \\ H_A : \exists \beta_i \neq 0 \end{cases}$$

1.7 Correlation analysis

```
> cor(dataset)
         y    x1     x2     x3     x4     x5     x6
y    1.000 0.508  0.108  0.435 -0.505 -0.007  0.238
x1   0.508 1.000  0.131  0.123  0.016  0.071  0.047
x2   0.108 0.131  1.000  0.066  0.065 -0.033 -0.004
x3   0.435 0.123  0.066  1.000 -0.044 -0.162  0.080
x4  -0.505 0.016  0.065 -0.044  1.000  0.124 -0.013
x5  -0.007 0.071 -0.033 -0.162  0.124  1.000 -0.099
x6   0.238 0.047 -0.004  0.080 -0.013 -0.099  1.000
```



2. Regression analysis

2.0 Stepwise regression

P-value criteria

- Stepwise regression with forward selection:
    - Starting from null model, we add a variable to the model
    - We perform linear regression on this model and we check the T-test results
    - We remove any variables with p-values greater than the significance level (α = 0.05)

2.0.1 T test

$$\begin{cases} H_0: \;\; \beta_i = 0 \\ H_A: \;\; \beta_i \neq 0 \end{cases}$$

$$t = \frac{\hat{\beta}_i}{SE(\hat{\beta}_i)} \sim T_{n-k-1}$$

$$\begin{cases} p < \alpha & \text{reject } H_0 \\ p > \alpha & \text{cannot reject } H_0 \end{cases}$$

$$SE(\hat{\beta}_i) = \hat{\sigma}\sqrt{v_i}$$

2.1 Two-predictors model

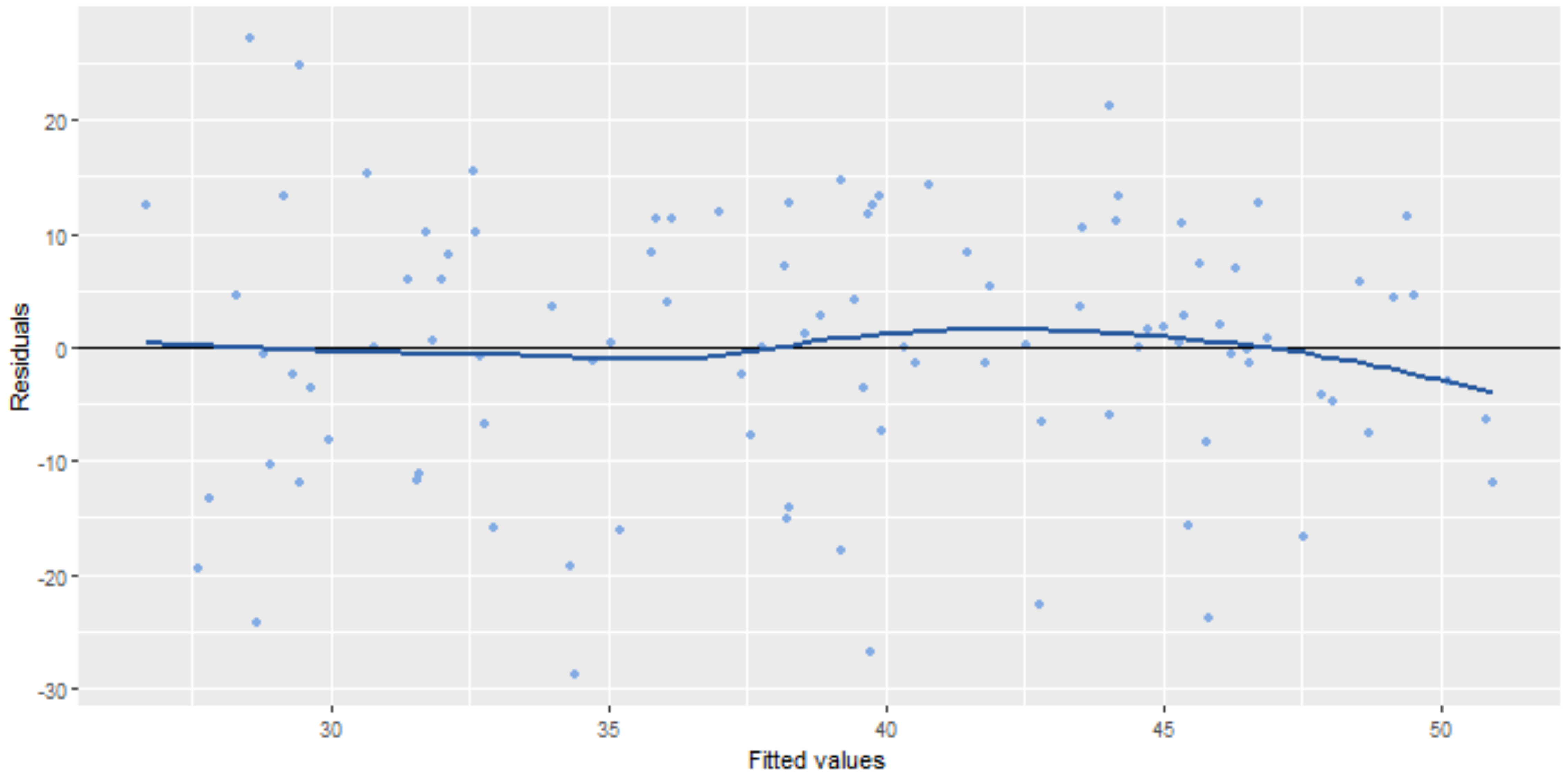
```
> lm(formula = y ~ x1 + x2, data = dataset)
```

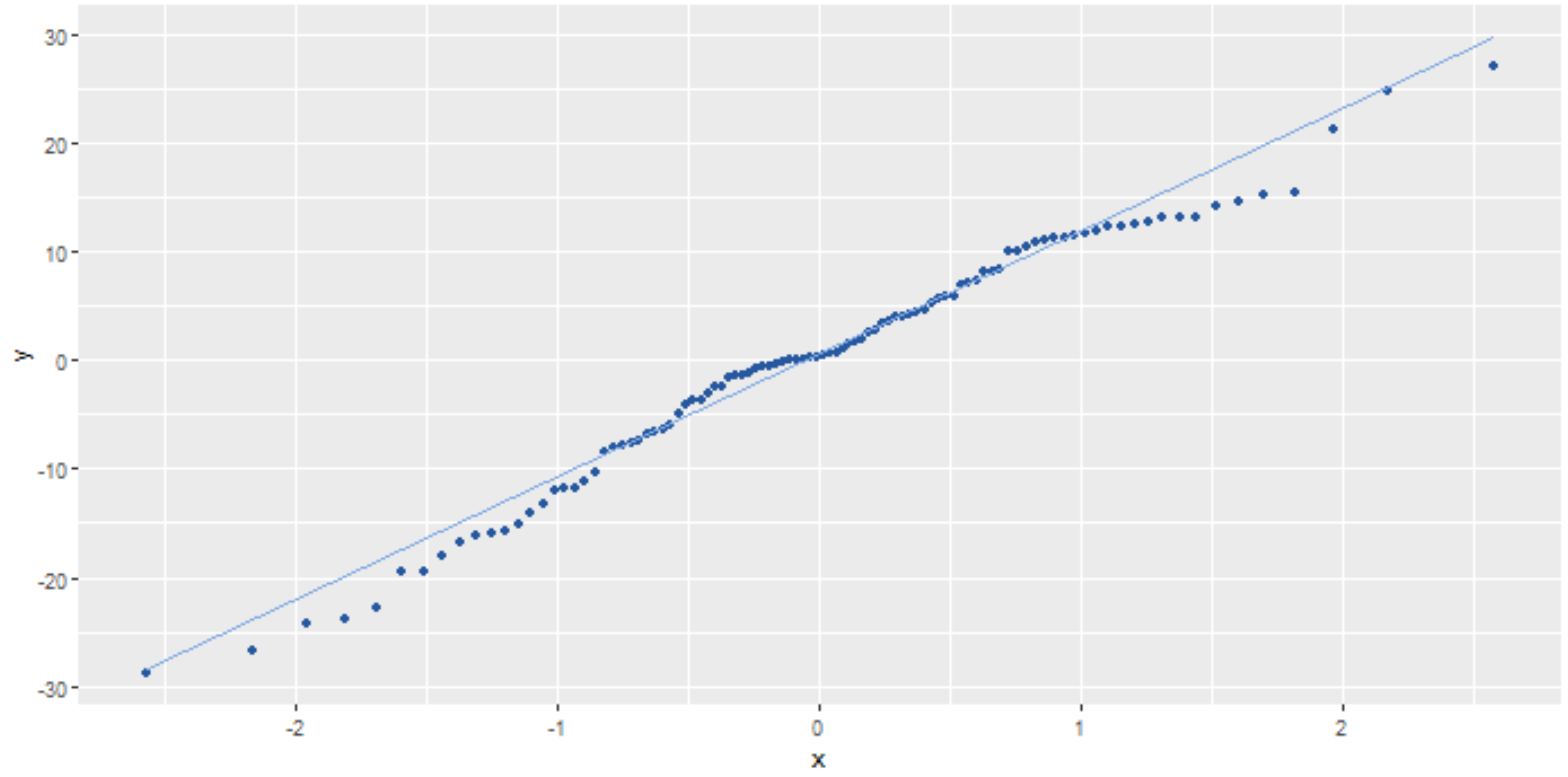
Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	38.9573	1.1516	33.828	< 2e-16	***
x1	6.6546	1.1675	5.700	1.29e-07	***
x2	0.5588	1.1675	0.479	0.633	

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 11.52 on 97 degrees of freedom
Multiple R-squared: 0.2597,   Adjusted R-squared: 0.2444
F-statistic: 17.01 on 2 and 97 DF, p-value: 4.652e-07





2.1.1 Diagnostics

- Residual analysis is used to show whether the assumptions made in OLS estimation are valid
- Residuals true mean needs to be zero
- Our model should not be affected by heteroschedasticity
- Error term should be normally distributed

2.2 Four-predictors model

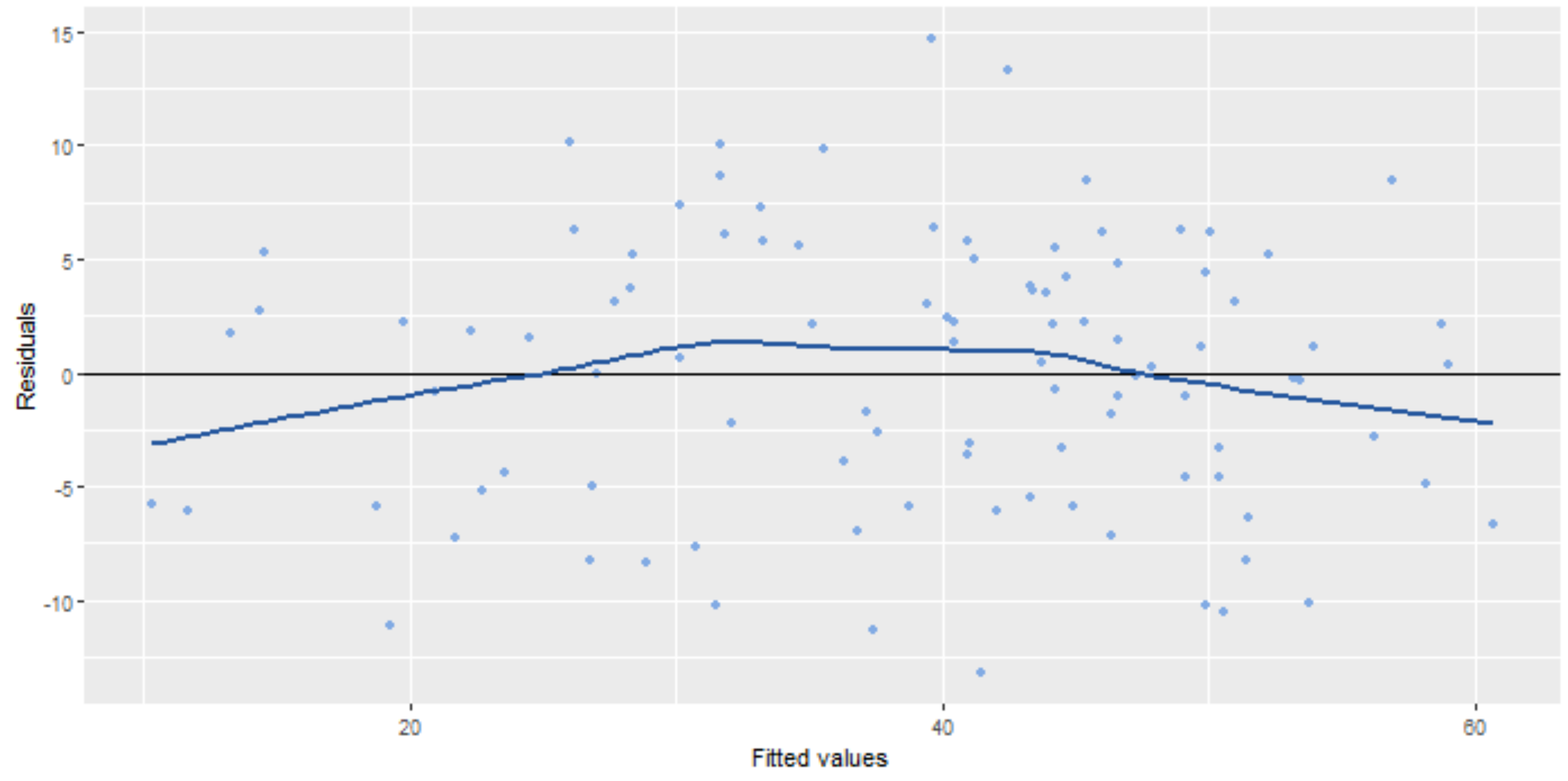
```
> lm(formula = y ~ x1 + x4 + I(x3^2) +
I(x3^3), data = dataset)
```

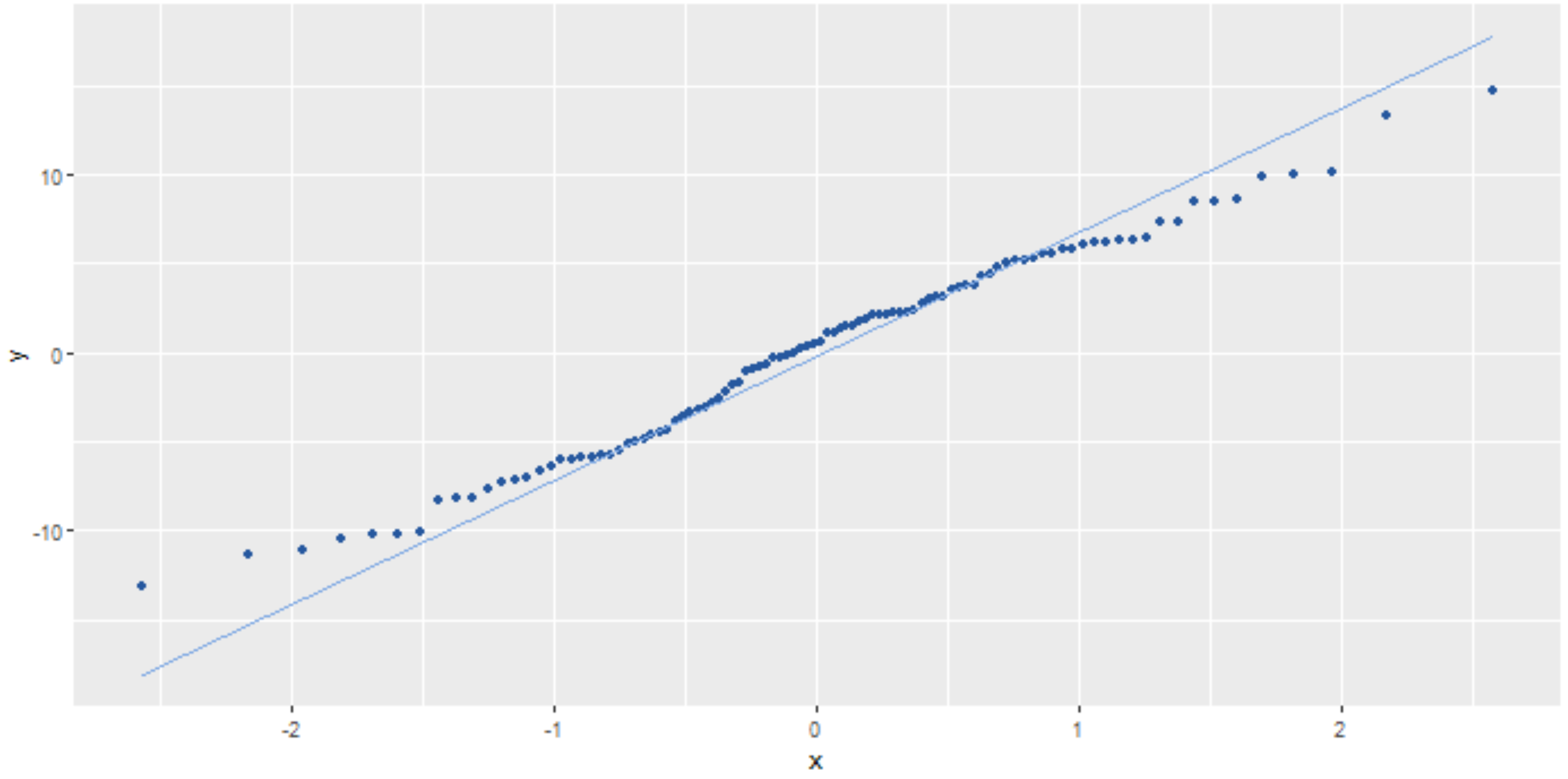
Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	44.8061	0.9145	48.994	< 2e-16	***
x1	6.0777	0.6166	9.857	3.34e-16	***
x4	-5.8706	0.6161	-9.529	1.67e-15	***
I(x3^2)	-5.6451	0.6934	-8.141	1.52e-12	***
I(x3^3)	2.3070	0.3217	7.171	1.61e-10	***

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6.073 on 95 degrees of freedom
Multiple R-squared: 0.7984, Adjusted R-squared: 0.7899
F-statistic: 94.05 on 4 and 95 DF, p-value: < 2.2e-16





2.3 Final model

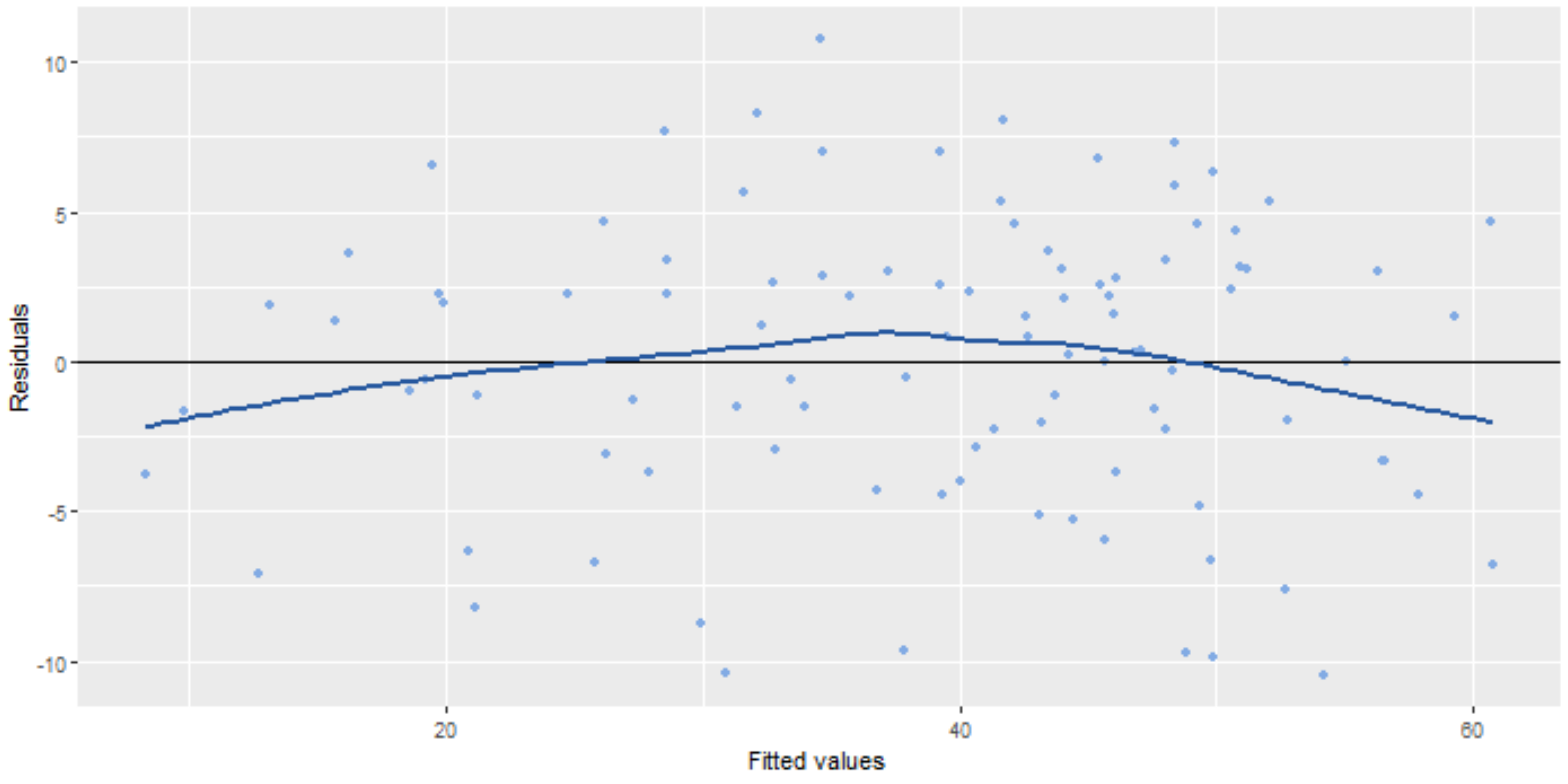
```
> lm(formula = y ~ x1 + x4 + I(x3^2) +
I(x3^3) + x6 + x1:x6 + x5, data = dataset)
```

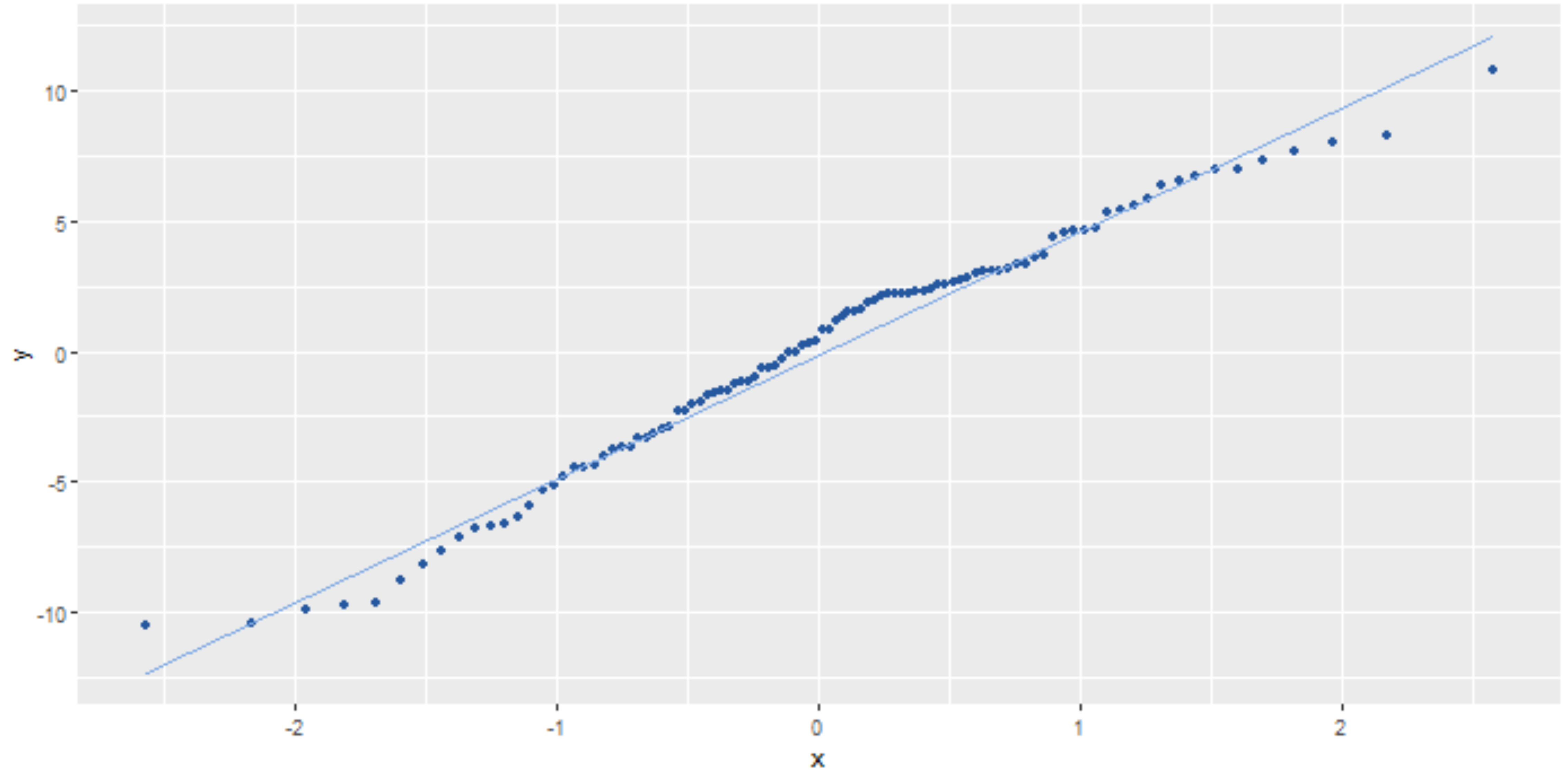
Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	44.6421	0.7490	59.605	< 2e-16	***
x1	5.8433	0.5046	11.579	< 2e-16	***
x4	-6.1328	0.5053	-12.138	< 2e-16	***
I(x3^2)	-5.3854	0.5730	-9.399	4.20e-15	***
I(x3^3)	2.2342	0.2650	8.432	4.52e-13	***
x6	2.7491	0.5026	5.469	3.86e-07	***
x5	1.3506	0.5123	2.636	0.00984	**
x1:x6	-2.1690	0.5056	-4.290	4.41e-05	***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.945 on 92 degrees of freedom
Multiple R-squared: 0.8705,   Adjusted R-squared: 0.8607
F-statistic: 88.38 on 7 and 92 DF, p-value: < 2.2e-16





2.4 R^2

$$R^2 = \frac{SQR}{SQTOT} = 1 - \frac{SQE}{SQTOT}$$

Step 1	Step 2	Step 3	Step 4	Step 5	Step 6	Step 7	Step 8
0.2579023	0.259651	0.605671	0.7983896	0.8077924	0.8449553	0.8705436	0.8731806

2.5 BIC

$$BIC = -2\log L + k\log n$$

Step 1	Step 2	Step 3	Step 4	Step 5	Step 6	Step 7	Step 8
783.5433	787.9126	729.5241	667.0445	666.8735	654.5976	636.5607	639.1078

2.6 AIC

$$AIC = -2\log L + 2k$$

Step 1	Step 2	Step 3	Step 4	Step 5	Step 6	Step 7	Step 8
775.7278	777.4919	716.4983	651.4134	648.6373	631.151	613.1141	613.0561

//Step command outputs a model with non-significant predictors

2.7 Overfitting

- Is our final model overfitted?
- Overfitting happens when fitted values of a model are too close to the original ones
  - Overfitted models predictions are not reliable
- How to check for overfitting:
  - Cross-validation
  - Train-test splitting

2.8 Train test splitting

- We split dataset in two parts:
  - Training set (70 observations)
  - Test set (30 observations)

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \overline{y_i})^2$$

- We compare MSE across different models
  - This gives an idea of the prediction error

2.9 *Should we include x5 then?*

It depends.

- Variable x5 describes the sound card performances; so we are led to believe that x5 doesn't explain Y
- Despite this, we overlooked what the variables describes, so we attempted to fit all the predictors
- It is quite counterintuitive to think of a relation between sound card performances and software performances
- Why does experimental evidence suggests otherwise?
  - Spurious correlation
  - P-value deflation

Thanks for your attention!