UNIVERSIDADE FEDERAL FLUMINENSE

Programa de Mestrado e Doutorado em Engenharia de Produção

Multivariate Data Analysis

Multiple Linear Regression

Professor: Valdecy Pereira. D. Sc.

email: valdecy.pereira@gmail.com

Outline

1. Definition

2. MLR

3. Residuals

4. Bibliography

MVDA - *Multiple Linear Regression*

The purpose of **MLR** (Multiple Linear Regression) to analyze the relationship between metric (or binary) independent variables (predictors) and a metric the dependent variable (response variable), with the following formulation:

$$Y = B_0 + B_1 X_1 + \cdots + B_i\, X_i$$

Where:

Y= Dependent Variable;

X_i= Independent Variable;

B_0= Intercept;

B_i= *Slopes*.

MVDA - *Multiple Linear Regression*

What is the optimal number of predictors? The suggested rules are:

- Evan rule (conservative): $\frac{n}{k} \geq 10$ $\rightarrow$ at least 10 observations (n) for predictor (k)

- Doane rule (relaxed): $\frac{n}{k} \geq 5$ $\rightarrow$ at least 5 observations (n) for predictor (k)

Categorical variables can be included as dummy variables (1 = one belongs to category; 0 = it does not belong to category). It is not necessary to encode all categories because the last one is identified when all the others have a zero value. This method prevents the occurrence of collinearity, and allows the design matrix to be invertible. Dummy variables have the same statistical treatment of the independent variables.

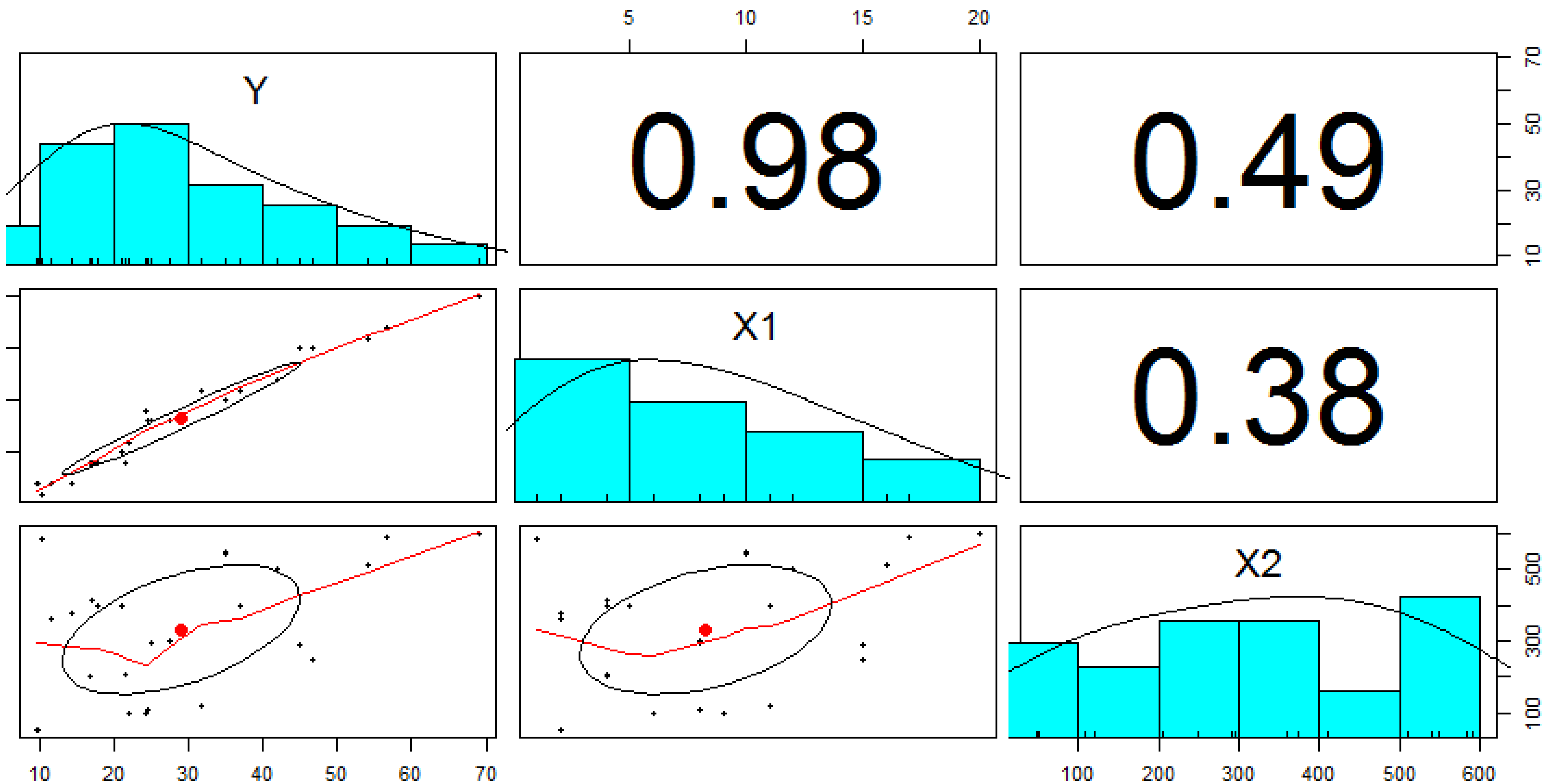
MVDA - *Multiple Linear Regression*

Observation	Y	X1	X2
1	9.6	2	52
2	9.95	2	50
3	10.3	1	585
4	11.66	2	360
5	14.38	2	375
6	16.86	4	200
7	17.08	4	412
8	17.89	4	400
9	21.15	5	400
10	21.65	4	205
11	22.13	6	100
12	24.35	9	100
13	24.45	8	110
14	25.02	8	295
15	27.5	8	300
16	31.75	11	120
17	34.93	10	540
18	35	10	550
19	37	11	400
20	41.95	12	500
21	44.88	15	290
22	46.59	15	250
23	54.12	16	510
24	56.63	17	590
25	69	20	600

In order to explain a **MLR** approach, the following dataset will be used: The simulated dataset of 25 observations and 2 independent Variables X_1 and X_2.

```
# Graph
library (psych)
pairs.panels(my_data)
```

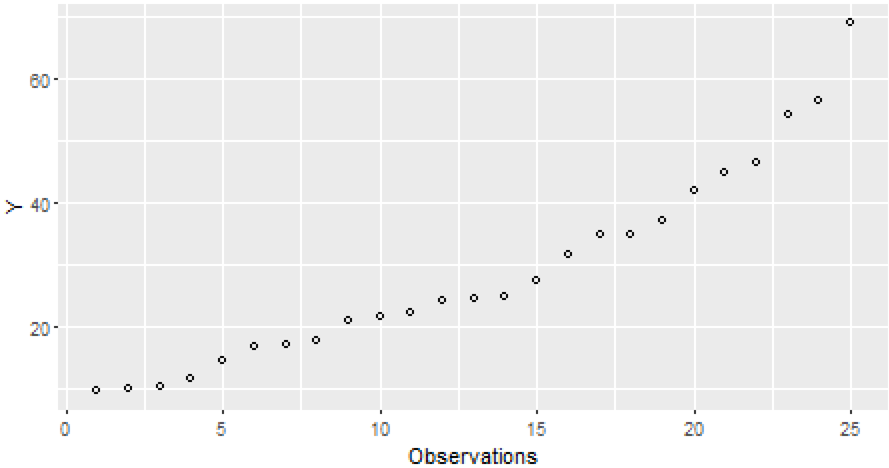
MVDA - *Multiple Linear Regression*

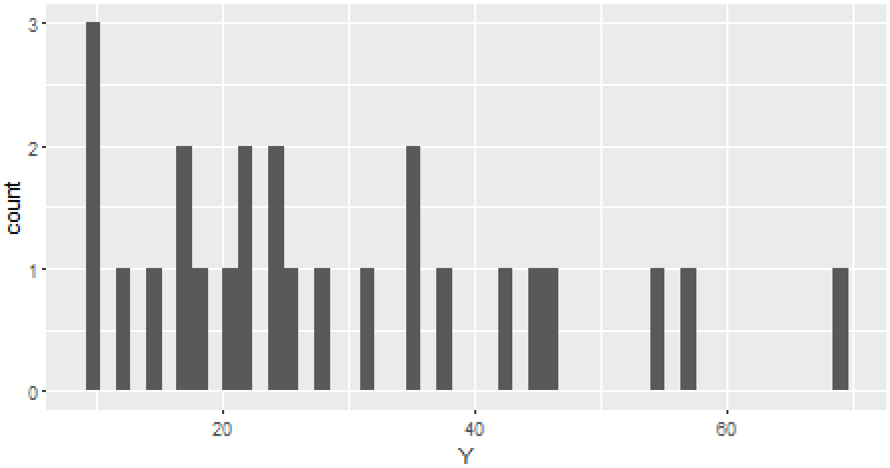


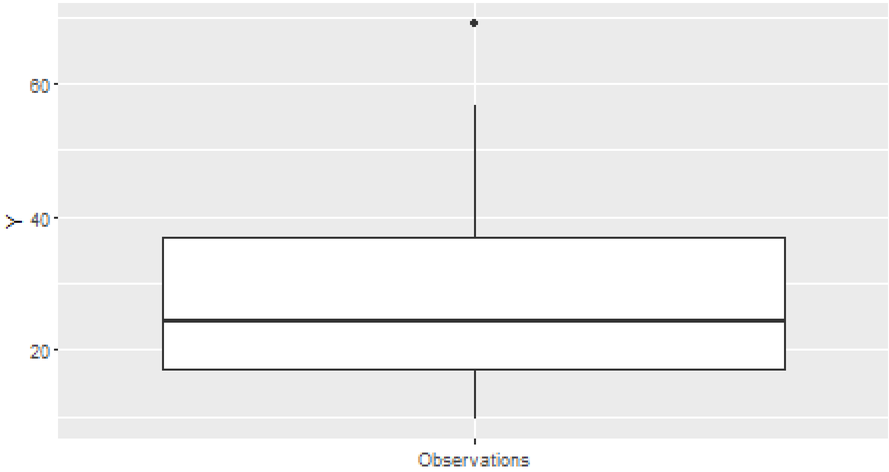
MVDA - *Multiple Linear Regression*

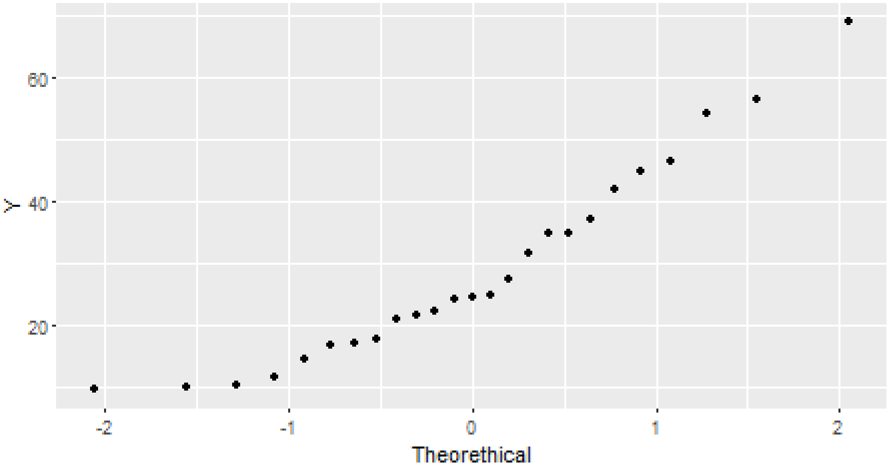
```
library(ggplot2)
ggplot(data = my_data, aes(x = 1:25, y = my_data$Y)) + geom_point(shape = 1) + labs(x = "Observations", y =
"Y")
ggplot(data = my_data, aes(x = "Observations", y = my_data$Y)) + geom_boxplot() + theme(axis.title.x =
element_blank()) + labs(y = "Y")
ggplot(my_data, aes(my_data$Y)) + geom_histogram(bins = 50) + labs(x = "Y")
ggplot(data = my_data, aes(x = 1:25, y = my_data$X1)) + geom_point(shape = 1) + labs(x = "Observations", y =
"X1")
ggplot(data = my_data, aes(x = "Observations", y = my_data$X1)) + geom_boxplot() + theme(axis.title.x =
element_blank()) + labs(y = "X1")
ggplot(my_data, aes(my_data$X1)) + geom_histogram(bins = 50) + labs(x = "X1")
ggplot(data = my_data, aes( sample = my_data$X1)) + stat_qq()+ xlab("Theorethical") + ylab("X1")
ggplot(data = my_data, aes(x = 1:25 y = my_data$X2)) + geom_point(shape = 1) + labs(x = "Observations", y =
"X2")
ggplot(data = my_data, aes(x = "Observations", y = my_data$X2)) + geom_boxplot() + theme(axis.title.x =
element_blank()) + labs(y = "X2")
ggplot(my_data, aes(my_data$X2)) + geom_histogram(bins = 50) + labs(x = "X2")
ggplot(data = my_data, aes( sample = my_data$X2)) + stat_qq()+ xlab("Theorethical") + ylab("X2")
```

MVDA - *Multiple Linear Regression*

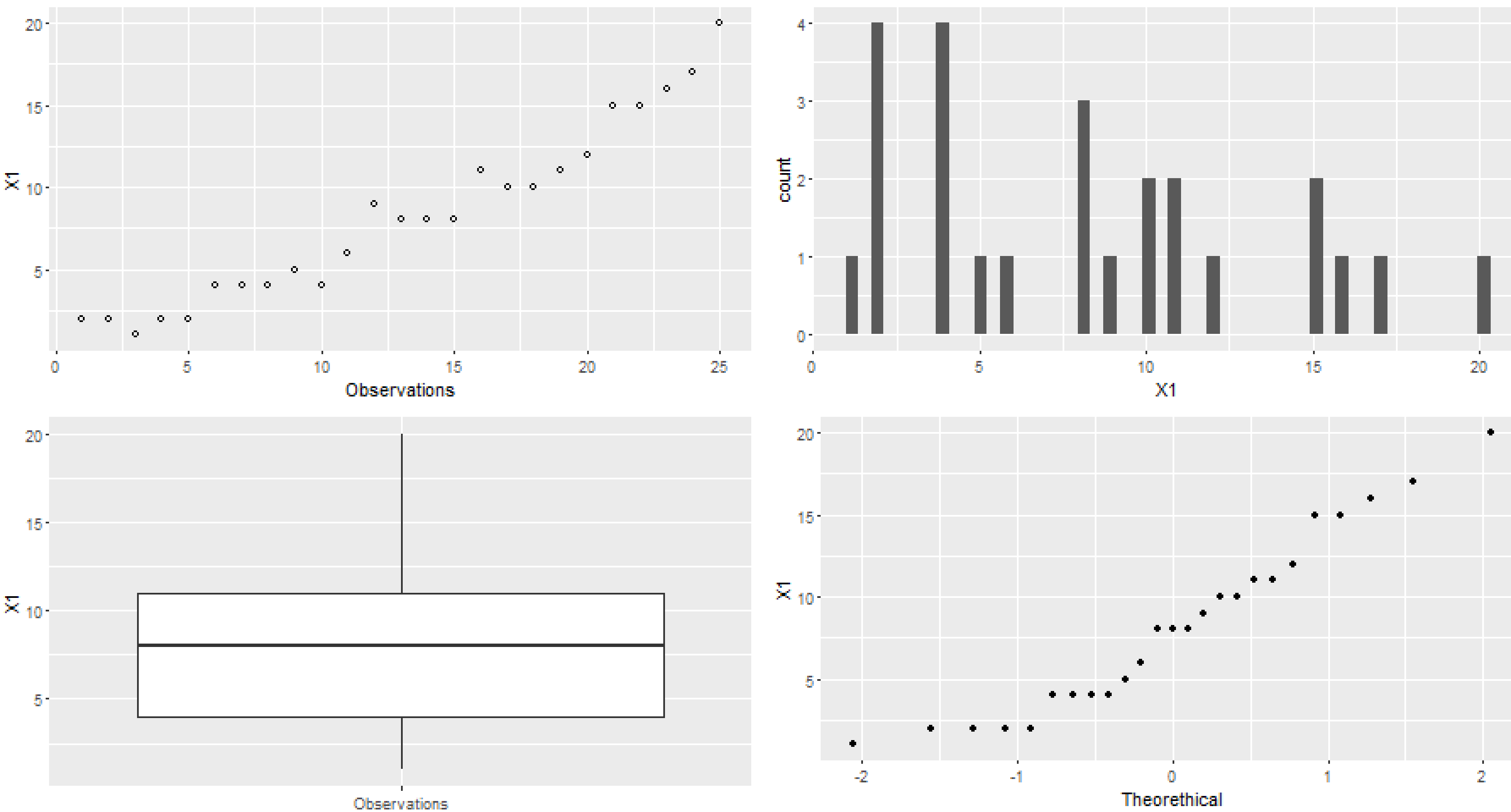




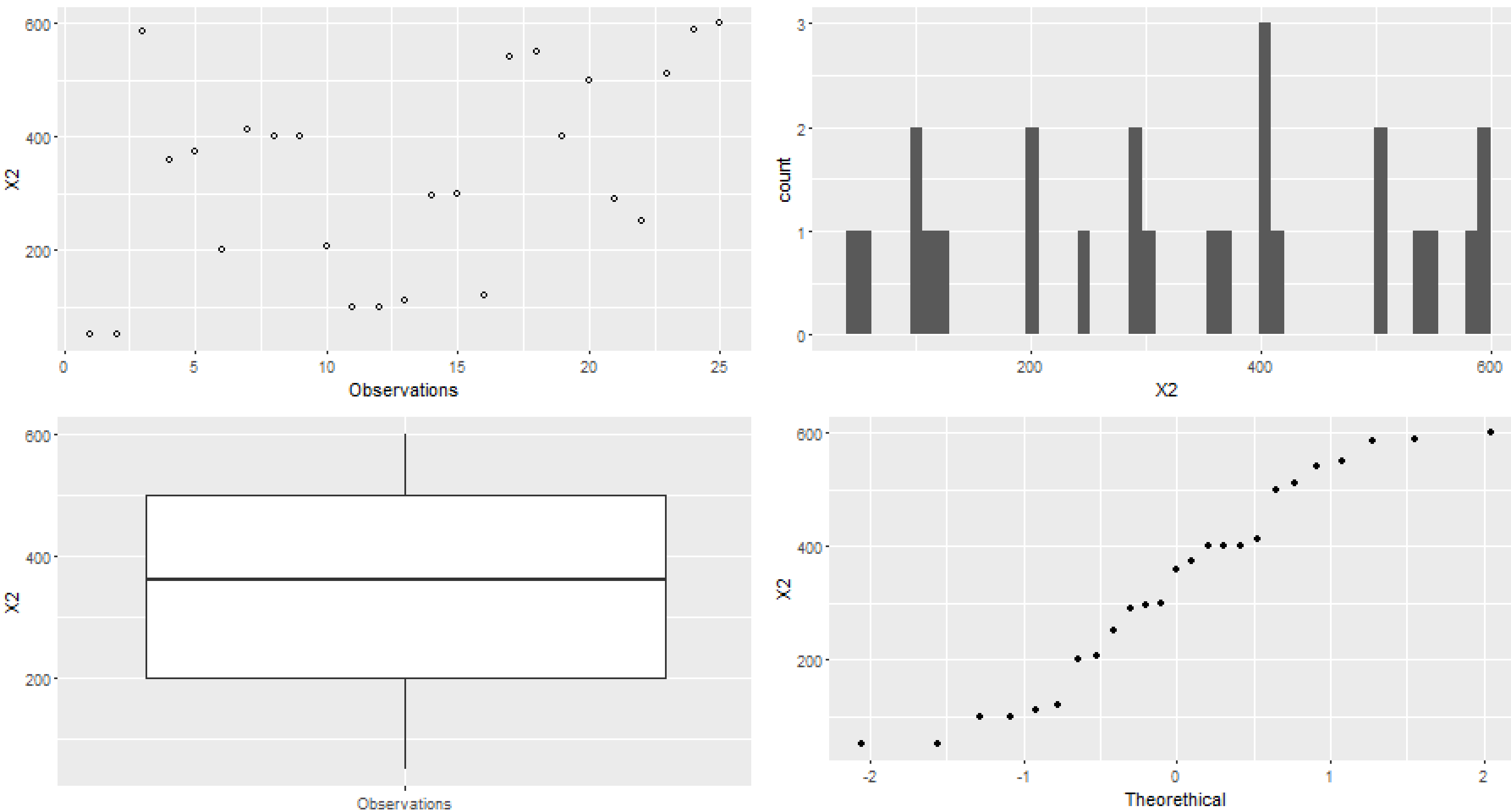




MVDA - *Multiple Linear Regression*



MVDA - *Multiple Linear Regression*



Multiple Linear Regression

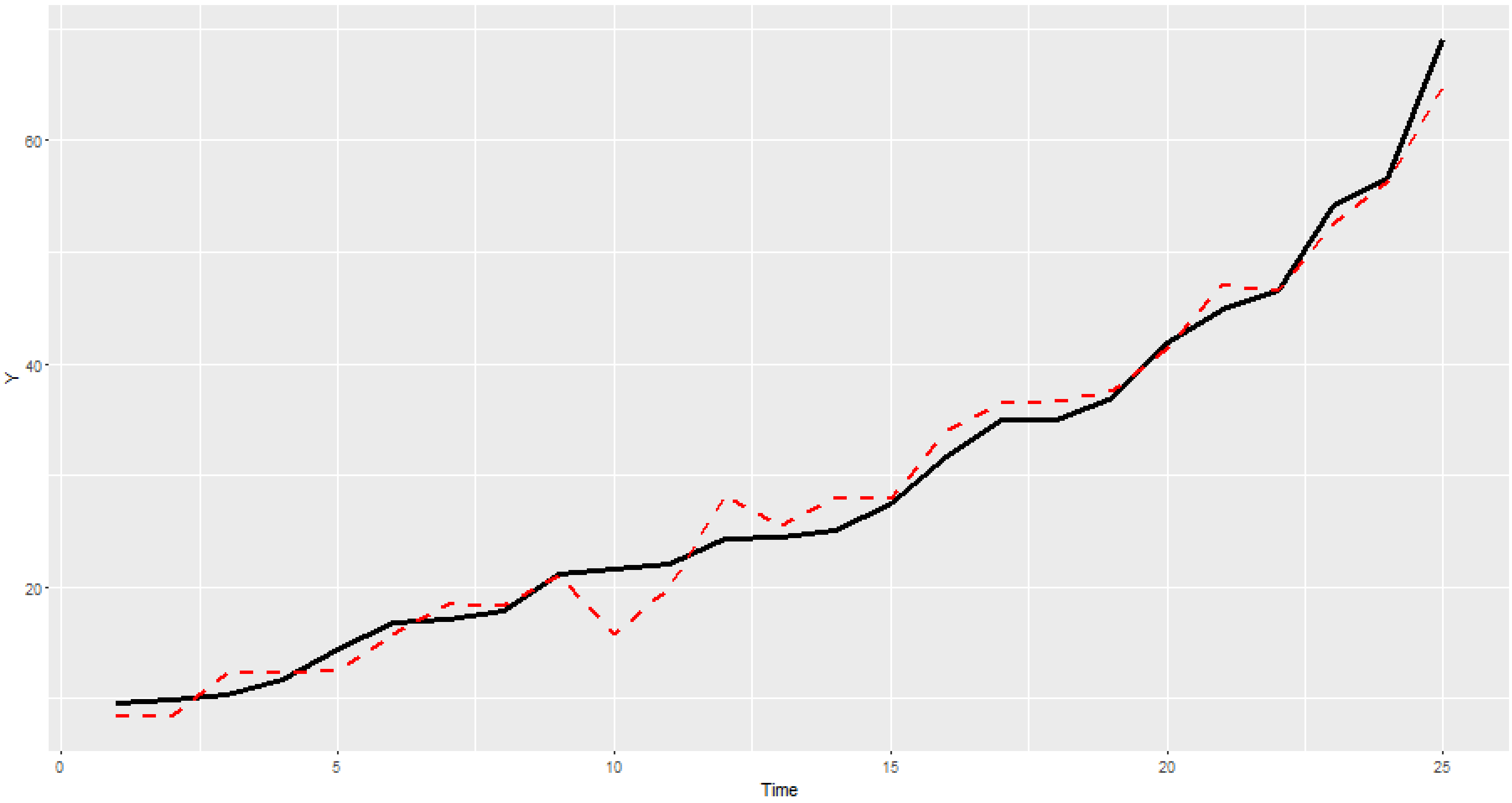
MVDA - *Multiple Linear Regression*

```
# MLR
mlr <- lm(Y ~ ., data = my_data)
summary(mlr)

Call:
lm(formula = Y ~ ., data = my_data)

Residuals:
   Min        1Q    Median      3Q      Max
-3.865    -1.542    -0.362   1.196    5.841

Coefficients:
             Estimate   Std. Error   t value    Pr(>|t|)
(Intercept)  2.263791   1.060066      2.136   0.044099 *
X1           2.744270   0.093524     29.343    < 2e-16 ***
X2           0.012528   0.002798      4.477   0.000188 ***
---
Signif. codes:  0   '***' 0.001   '**' 0.01   '*' 0.05   '.' 0.1   1

Residual standard error: 2.288 on 22 degrees of freedom
Multiple R-squared:  0.9811
Adjusted R-squared:  0.9794
F-statistic: 572.2 on 2 and 22 DF,  p-value: < 2.2e-16
```

MVDA - *Multiple Linear Regression*

Observation	Y	X1	X2	$\hat{y}$
1	9.6	2	52	8.40
2	9.95	2	50	8.38
3	10.3	1	585	12.34
4	11.66	2	360	12.26
5	14.38	2	375	12.45
6	16.86	4	200	15.75
7	17.08	4	412	18.40
8	17.89	4	400	18.25
9	21.15	5	400	21.00
10	21.65	4	205	15.81
11	22.13	6	100	19.98
12	24.35	9	100	28.21
13	24.45	8	110	25.60
14	25.02	8	295	27.91
15	27.5	8	300	27.98
16	31.75	11	120	33.95
17	34.93	10	540	36.47
18	35	10	550	36.60
19	37	11	400	37.46
20	41.95	12	500	41.46
21	44.88	15	290	47.06
22	46.59	15	250	46.56
23	54.12	16	510	52.56
24	56.63	17	590	56.31
25	69	20	600	64.67

```
# Predicted Y
mlr$fitted.values

ggplot() + geom_line(data = my_data, aes(x = 1:25, y = Y), colour = "black", size = 1.2)  + xlab("Time") + ylab("Y") + geom_line(data = my_data, aes(x = 1:25, y = mlr$fitted.values), colour = "red", size = 1, linetype = 2)
```



60
40
20
Y
0
5
10
15
20
25
Time

MVDA - *Multiple Linear Regression*

The model standard error is calculated by.

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-(k+1)}; \hat{\sigma}^2 = 5.235$$

$$\hat{\sigma} = 2.288$$

Where:

k = Total number of independent variables;

$\hat{y}$ = Estimated value of the dependent variable;

```
summary(mlr)

Residual standard error: 2.288 on 22 degrees of freedom
Multiple R-squared:  0.9811
Adjusted R-squared:  0.9794
F-statistic: 572.2 on 2 and 22 DF,  p-value: < 2.2e-16
```

MVDA - *Multiple Linear Regression*

The following hypothesis test for B_i can then be done:

$$H_0: B_i = 0 \ (There\ is\ not\ a\ linear\ relation\ between\ x_i\ e\ y)$$
$$H_1: B_i \neq 0 \ (There\ is\ a\ linear\ relation\ between\ x_i\ e\ y)$$

```
summary(mlr)
Call:

Coefficients:
             Estimate   Std. Error   t value   Pr(>|t|)
(Intercept)  2.263791   1.060066     2.136     0.044099 *
X1           2.744270   0.093524     29.343      < 2e-16 ***
X2           0.012528   0.002798     4.477     0.000188 ***
---
Signif. codes:  0   '***' 0.001   '**' 0.01   '*' 0.05   '.' 0.1   1
```

MVDA - *Multiple Linear Regression*

To evaluate the adequacy of the model with the data, the following hypothesis is made to test the model adequacy:

$$H_0: The\ model\ is\ not\ adequate$$
$$H_1: The\ model\ is\ adequate$$

```
summary(mlr)

Residual standard error: 2.288 on 22 degrees of freedom
Multiple R-squared:  0.9811
Adjusted R-squared:  0.9794
F-statistic: 572.2 on 2 and 22 DF,  p-value: < 2.2e-16
```

MVDA - *Multiple Linear Regression*

The r^2 (r-squared or coefficient of determination) measures the strength of the relationship, indicating that the model explains a percentage of the variance of the dependent variable. For example, a r^2 equal to 0.80 means that 80% of the variance of the dependent variable comes from its relation with the independent variables. The $(r^2)_a$ - adjusted r^2 - is na indicator that adjusts the r^2 based on the number of k independent variables and it is useful to penalize models that uses too many independent variables.

```
summary(mlr)

Residual standard error: 2.288 on 22 degrees of freedom
Multiple R-squared:  0.9811
Adjusted R-squared:  0.9794
F-statistic: 572.2 on 2 and 22 DF,  p-value: < 2.2e-16
```

MVDA - *Multiple Linear Regression*

The collinearity (correlation between two predictors) or multicollinearity (correlation between multiple predictors) can be harmful for the model because:

- Estimates may be unstable;
- Standard errors can be unreliable;
- Confidence intervals can become very large;
- The coefficient of determination can be high, even though the T-tests are insignificant.

To detect this problem is necessary to calculate the **VIF** (***Variance Inflator Factor***) for each predictor. If its value is > 5 the predictor is highly correlated indicating collinearity:

```
# VIF (Variance Inflation Factors)
library("car")
vif(mlr)

      X1             X2
1.167128       1.167128
```

MVDA - *Multiple Linear Regression*

Suggest interpretation:

VIF	**Interpretation**
$VIF_j = 1.00$	Insignifiant
$VIF_j = 2.00$	Medium
$VIF_j = 10.00$	Strong
$VIF_j = 100.00$	Severe

Residuals

MVDA - *Multiple Linear Regression*

Once verified the adequacy of the estimated model, it is also necessary to validate the errors (residuals) of the model. Supposedly, the residuals must be:

- Normally distributed;
- Homoscedastic;
- Independent (uncorrelated within a series of time).

The studentized residuals are most indicated approach to proceed with the evaluation. The studentized residual is the quotient resulting from the division of a residual by an estimate of its standard deviation, usually ranging from -3 to +3.

```
# Studentized Residuals
library("MASS")
sresid <- studres(mlr)
```

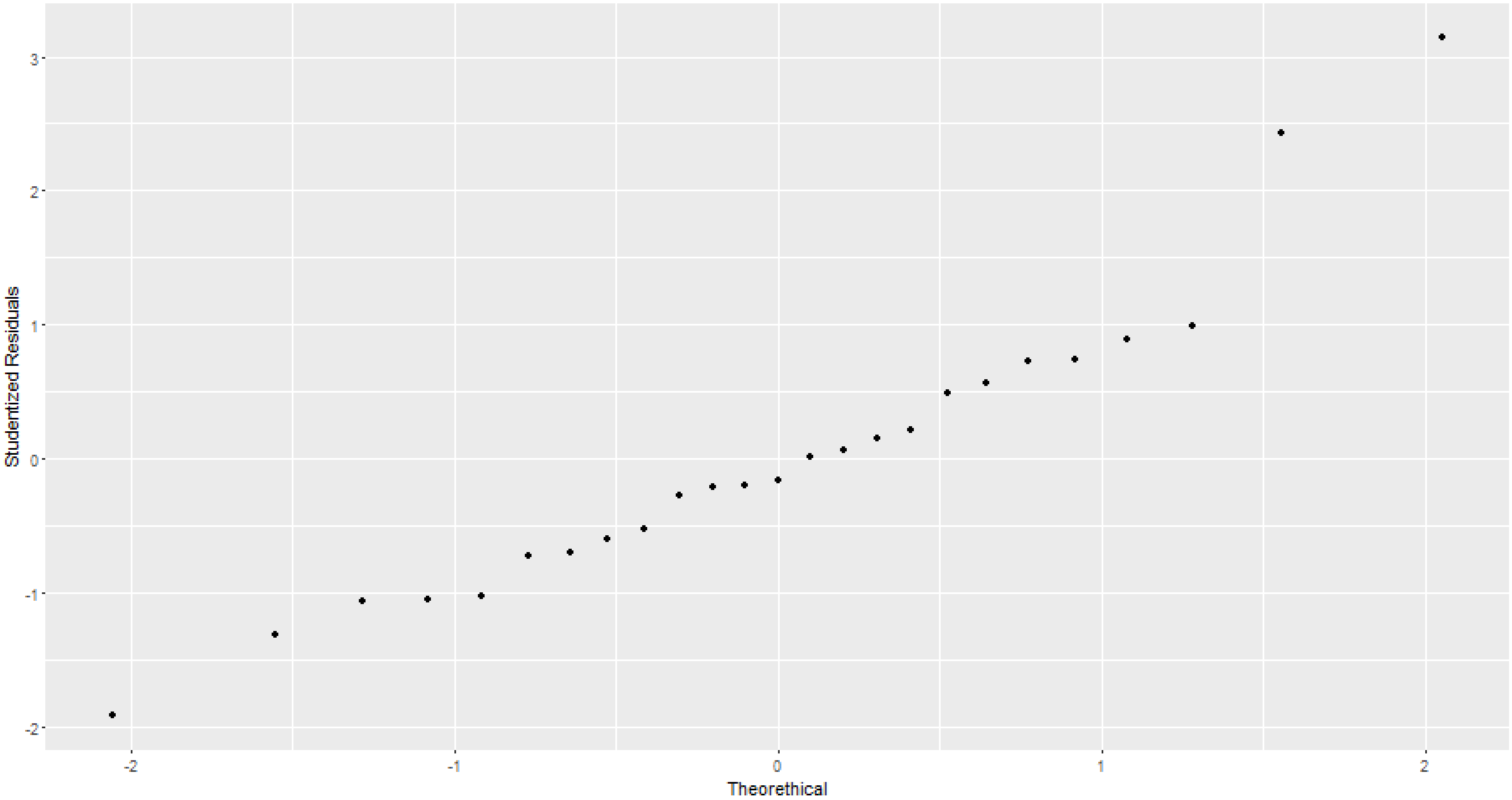
MVDA - *Multiple Linear Regression*

- Normally Distributed

Violation of this assumption is considered mild and can make the confidence intervals unreliable. Large samples, logarithmic transformations in the dependent and independent variables or removal of outliers, can avoid this violation.

```
# QQ Plot
ggplot(data = my_data, aes( sample = sresid)) + stat_qq()+ xlab("Theorethical") + ylab("Studentized Residuals")

# Univariate Normality
shapiro.test(sresid)

Shapiro-Wilk normality test

data:  sresid
W = 0.93094, p-value = 0.09136
```

MVDA - *Multiple Linear Regression*



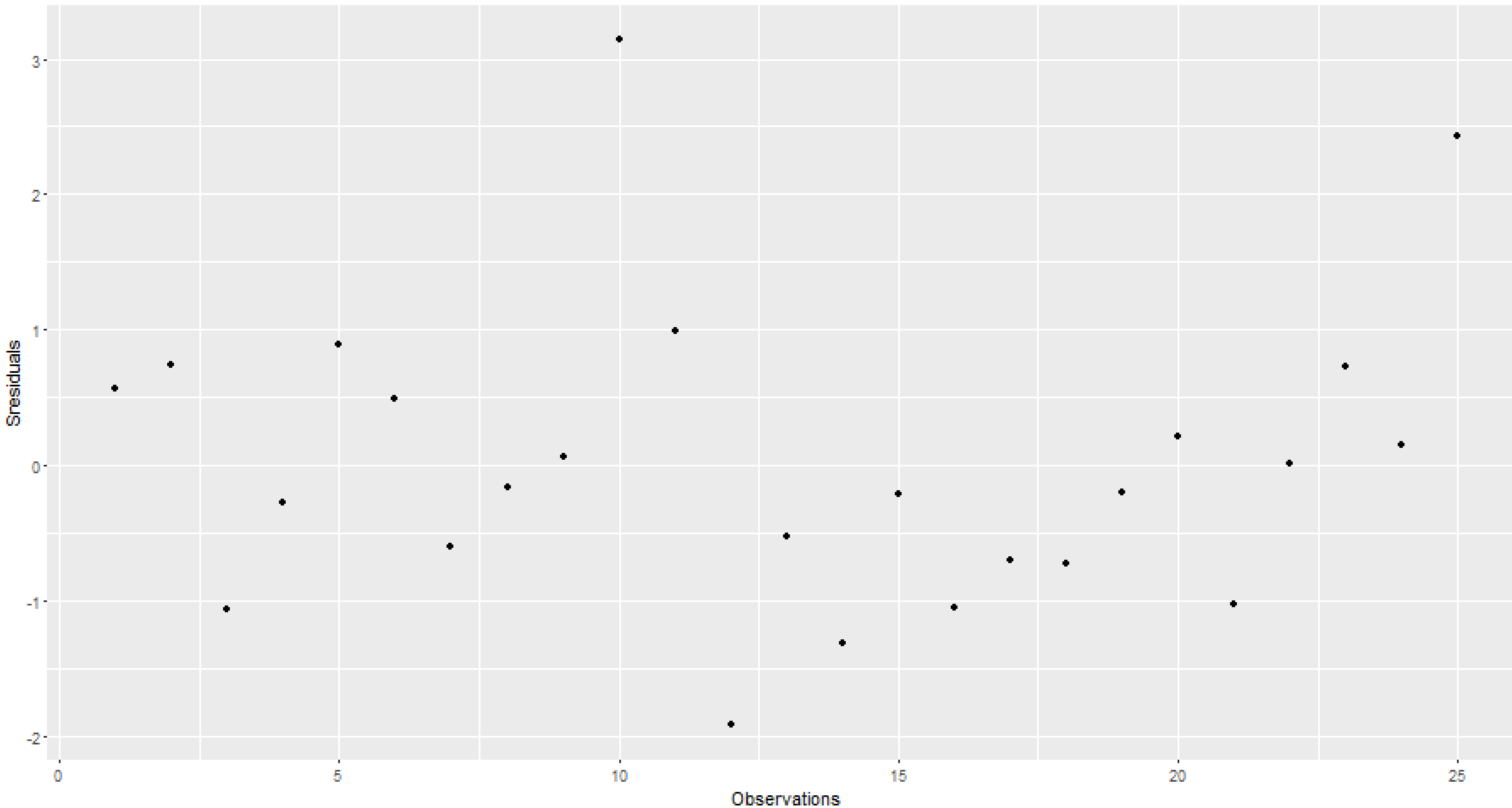
MVDA - *Multiple Linear Regression*

- Homocedastic

Violation of this assumption is considered severe and can increase the range of confidence intervals and make the model unfeasible. Large samples, logarithmic transformations in the dependent and independent variables or removal of outliers, can avoid this violation. To verify that the errors are homoscedastic (constant variance) the residual graphic, ideally, will have no patterns.

```
# Homoscedascity
ggplot(data = my_data, aes(x = 1:25, y = sresid)) + geom_point() + labs(x = "Observations", y = "Sresiduals")
```

MVDA - *Multiple Linear Regression*



MVDA - *Multiple Linear Regression*

The Breusch-Pagan Test can be done to verify the homoscedasticity, with the following hypothesis:

$$H_0: The\ model\ is\ homoscedastic$$
$$H_1: The\ model\ is\ not\ homoscedastic$$

```
# Breusch-pagan test
library("car")
ncvTest(mlr)

Non-constant Variance Score Test
Variance formula: ~ fitted.values
Chisquare = 0.04524206    Df = 1     p = 0.8315596
```

MVDA - *Multiple Linear Regression*

- Independency

Violation of this assumption is considered mild and can increase the range of the confidence intervals. This assumption is only considered in time series.

A differentiation of the first order or the removal of outliers, can avoid this violation.

Prediction

MVDA - *Multiple Linear Regression*

```
# Prediction
new_data <- as.data.frame(cbind(2, 50))
colnames(new_data) <- c("X1", "X2")
predict(mlr, new_data)
```

MVDA

https://github.com/Valdecy/Multivariate_Data_Analysis

##
##

Created by: Prof. Valdecy Pereira. D.Sc.
UFF - Universidade Federal Fluminense (Brazil)
email: valdecypereira@yahoo.com.br
Course: Multivariate Data Analysis
Lesson: Multiple Linear Regression

Citation:

PEREIRA. V. (2016). Project: Multivariate Data Analysis. File: R-MVDA-07-MLR.pdf. GitHub repository: <https://github.com/Valdecy/Multivariate_Data_Analysis>

##
##

Bibliography

CORRAR, L.J.; PAULO, E.; DIAS FILHO, J. M. **Análise Multivariada para Cursos de Administração, Ciências Contábeis e Economia**. ATLAS, 2009.

FÁVERO, L. P.; BELFIORE, P.; SILVA, F. L.; CHAN, B. **Análise de Dados: Modelagem Multivariada para Tomada de Decisões**. CAMPUS, 2009.

HAIR, J. F.; BLACK, W. C.; BABIN, B. J.; ANDERSON, R. E.; TATHAM, R. L. **Análise Multivariada de Dados**. BOOKMAN, 2009.

LATTIN, J.; CARROLL, J. D.; GREEN, P. E. **Análise de Dados Multivariados**. CENGAGE Learning, 2011.

LEVINE, D. M.; STEPHAN, D. F.; KREHBIEL, T. C.; BERENSON, M. L. **Estatística - Teoria e Aplicações - Usando Microsoft Excel**. LTC, 2012.