# Business Analytics minor

UC Leuven Limburg
MOVING MINDS

Week 5: Regression

# Agenda

- Quick Review
  - Classification Techniques
- Regression
  - Regression Models
  - Evaluation Metrics
  - Compare to Classification
- Feature Selection
- Exercises

# Agenda

- Quick Review
  - Classification Techniques
- Regression
  - Regression Models
  - Evaluation Metrics
  - Compare to Classification
- Feature Selection
- Exercises

# Quick Review

## Classification Techniques

Naive Bayes and k-Nearest Neighbors

Given a set of predictor variables, $x_1$, $x_2$, …, $x_{p,}$ try to determine class y (discrete).

For example, suppose a bank is evaluating an an applicant for a house loan: The applicant is 32 years old, has a salary of 2500 euros per month, is unmarried, and has an existing loan for a car in the amount of 8750 euros which is due to be paid in 30 months. Is the applicant a good customer (likely to pay back the home loan?) or a bad customer (unlikely to pay back the home loan)?

# Quick Review

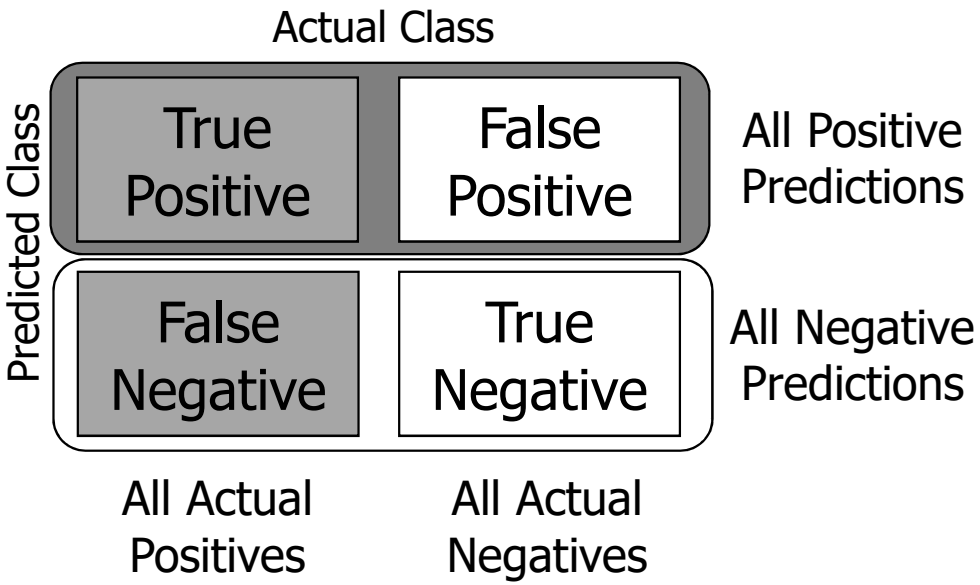
## Classification Process

1. Gather all the data you have on the banks' loan customers.
2. Split that data into a training set and a test set.
3. Build a model on the training set that shows which of them repay their loans and which of them default.
4. Use the test set to test the accuracy of the model.
5. Now use the model to classify the new applicant as "good" or "bad".

# Quick Review

## Evaluation of Classification Models

### Confusion Matrix




Precision/Positive Predictive Value:

$$\frac{\text{True Positive}}{\text{All Positive Predictions}}$$

Recall/Sensitivity/True Positive Rate:

$$\frac{\text{True Positive}}{\text{All Actual Positives}}$$

Specificity/True Negative Rate:

$$\frac{\text{True Negative}}{\text{All Actual Negatives}}$$

# Agenda

- Quick Review
  - Classification Techniques
- Regression
  - Regression Models
  - Evaluation Metrics
  - Compare to Classification
- Feature Selection
- Exercises

# Regression

## Techniques

### Linear Regression

Given a set of predictor variables, $x_1$, $x_2$, …, $x_{p,}$ try to determine value y (continuous).

For example, suppose a bank is evaluating an an applicant for a house loan: The applicant is 32 years old, has a salary of 2500 euros per month, is unmarried, and has an existing loan for a car in the amount of 8750 euros which is due to be paid in 30 months. How many months will the customer continue paying their loan?

# Regression

## Simple Linear Regression

Simple: 1 predictor variable X, try to predict the value of Y

$$Y = \beta 0 + \beta 1 * X + \varepsilon$$

ε represents “noise”

Use the training data set to estimate β0 and β1.

Estimate the noise or error constant based on the standard deviation estimate in output

# Regression

## Multiple Linear Regression

Multiple: a set of predictor variable $X_1$, $X_2$, …, $X_p$, try to predict the value of Y

$$Y = \beta 0 + \beta 1 \cdot X1 + \ldots + \beta p \cdot Xp + \varepsilon$$

ε represents "noise"

Use the training data set to estimate β0, β1, …, βp.

Estimate the noise or error constant based on the standard deviation estimate in output

# Regression

## Multiple Linear Regression

What predictor variables are important? Which ones influence the target value?

- Find the coefficients significantly different from 0:
  - Be aware, this may be a coincidence!
- Calculate the p-value:
  - Probability that that this value occurred by coincidence
  - Low p-value = significant
  - Find the p-value in the Coefficient output

# Regression

## Prediction Process

1. Gather all the data you have on the banks' loan customers.
2. Split that data into a training set and a test set.
3. Build a model on the training set that shows how long each customer continues to repay their loans.
4. Use the test set to test the accuracy of the model.
5. Now use the model to predict how many months the new applicant will continue paying back their loan.

# Regression

Evaluation of Regression Models

How good is the model?

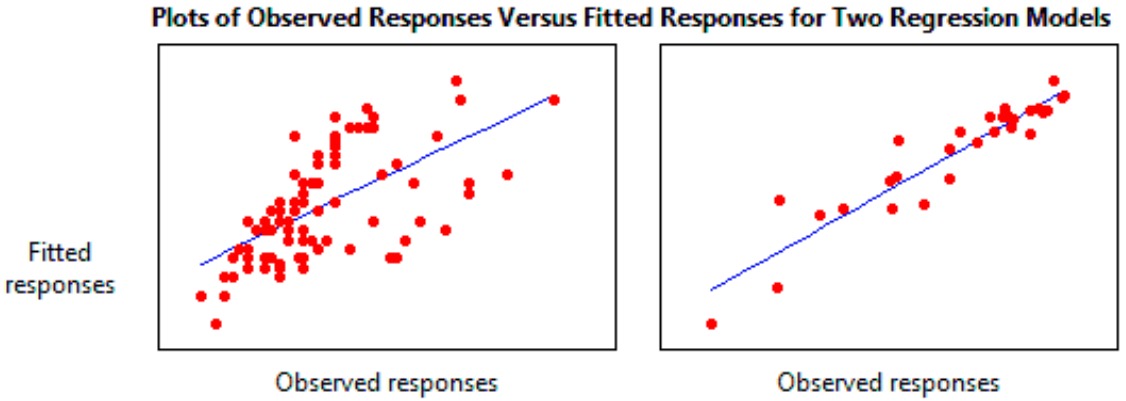
This depends on what you're using it for:

Describing the data vs Predicting outcomes

Choosing between Descriptive and Predictive

- For data mining, emphasis is usually on predictive power

# Regression




## Purpose: Description

- Goodness of fit
- Calculations on training data
- Calculate $R^2$
  - measures how close the data is to the regession line
  - Available as output

## Purpose: Prediction

- Accuracy measures
- Calculations on test or validation data
  - Available as Output

# Regression

## Evaluation of Regression Models: Residuals

Compare the predicted value with the observed value

$$e_i = y_i - \hat{y}_i$$

$e_i$ : error

$y_i$ : actual value

$\hat{y}_i$ : predicted value

# Regression

## Evaluation Metrics

Mean Absolute Error:
(add all the errors, divide by the number)

$$\frac{\sum_{i=1}^{n} |e_i|}{n}$$

Mean Absolute Percentage Error:
Absolute value of ((actual – predicted) / actual)
Add them all up and multiply by 100/number

$$\frac{100}{n} \sum_{t=1}^{n} \left| \frac{A_t - F_t}{A_t} \right|$$

Mean Squared Error:
(square all the errors, add them up, divide by the number)

$$\frac{1}{n} \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$$

Total Sum of Squared Errors:
(square all the errors, add them up)

$$\sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$$

# Agenda

- Quick Review
  - Classification Techniques
- Prediction
  - Regression Models
  - Evaluation Metrics
  - Compare to Classification
- Feature Selection
- Exercises

# Feature Selection

**Optimal number + choice of predictor variables?**

- Too many variables: possibility of overfitting!
  - perhaps low predictive power
- Preferably don't take in variables that don't contribute to the prediction.
  - leads to larger dispersion in the predictions
- Preference for not removing variables that 'effectively' contribute to the prediction.
  - leads to a higher average error in the predictions
- Beware of predictor variables that are strongly correlated!
  - can falsely represent coefficients
  - track down correlations ('matrix plot' or 'correlation matrix')
- Be aware of outliers!
- Rule of thumb: number of observations *n* in training data equals at least 5(p+2)

# Feature Selection

Methods to choose the best subset of predictor variables

- first: reduce number of predictor variables by means of domain knowledge
- then: use algorithms
  - 'Exhaustive search': try all predictor variables subsets
  - 'Forward selection': start with 1 predictor variable, add each time the most significant one
  - 'Backward selection': start with all predictor variables, remove each time the least significant one

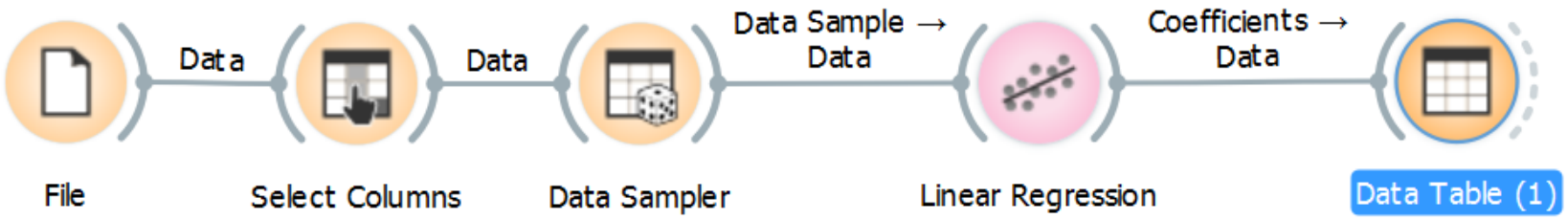
# Agenda

- Quick Review
  - Classification Techniques
- Prediction
  - Regression Models
  - Evaluation Metrics
  - Compare to Classification
- Feature Selection
- Exercises

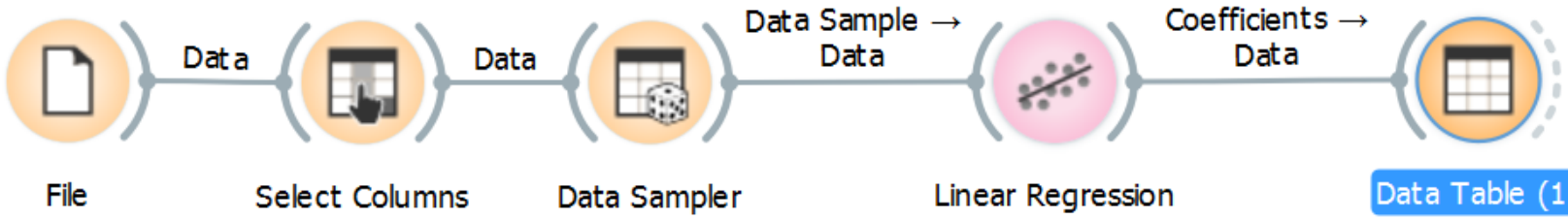
UC Leuven Limburg
MOVING MINDS

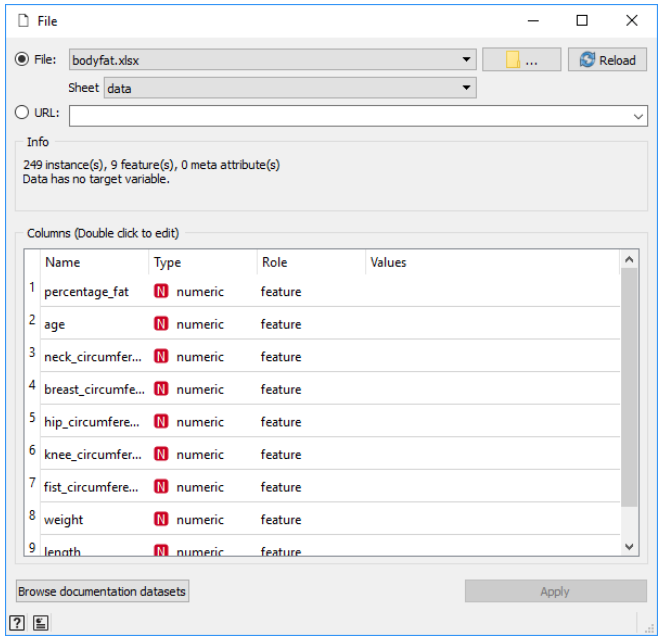
# Exercise 6.1

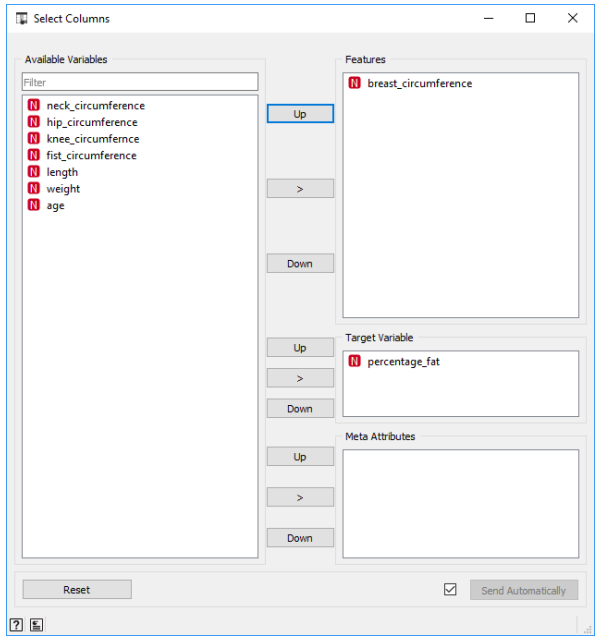
1. Separate the data into a training and test set (60/40). Make sure the sampling is stratified and replicable.
2. Choose the fat percentage as Y -the class to predict- and the breast circumference as X. (select columns)
3. Build a linear regression model based on the training data (linear regression with default settings)

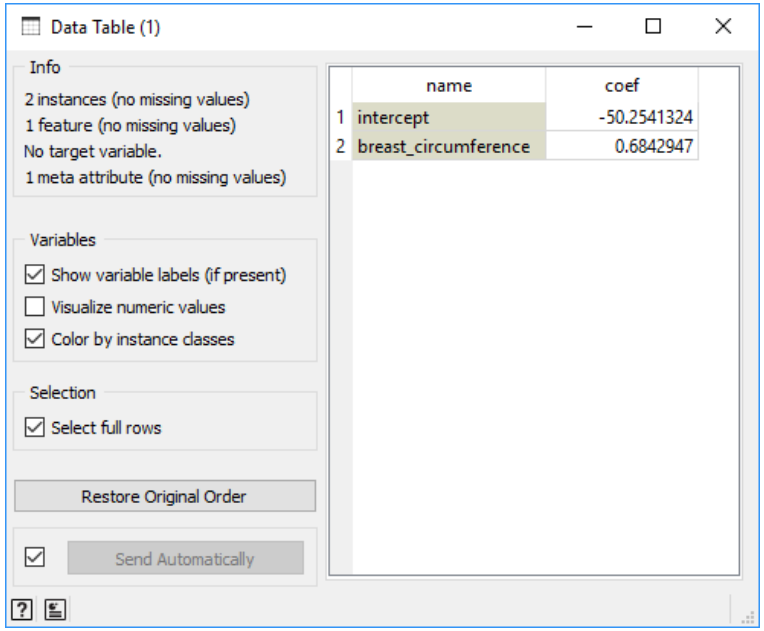


# Exercise 6.1


Data
Data
Data Sample → Data
Coefficients → Data
File
Select Columns
Data Sampler
Linear Regression
Data Table (1)


File
File: bodyfat.xlsx
Reload
Sheet data
URL:
Info
249 instance(s), 9 feature(s), 0 meta attribute(s)
Data has no target variable.
Columns (Double click to edit)
Name Type Role Values
1 percentage_fat numeric feature
2 age numeric feature
3 neck_circumfer... numeric feature
4 breast_circumfe... numeric feature
5 hip_circumfere... numeric feature
6 knee_circumfer... numeric feature
7 fist_circumfere... numeric feature
8 weight numeric feature
9 length numeric feature
Browse documentation datasets
Apply


Select Columns
Available Variables
Filter
neck_circumference
hip_circumference
knee_circumfernce
fist_circumference
length
weight
age
Up
>
Down
Features
breast_circumference
Target Variable
percentage_fat
Meta Attributes
Reset
Send Automatically

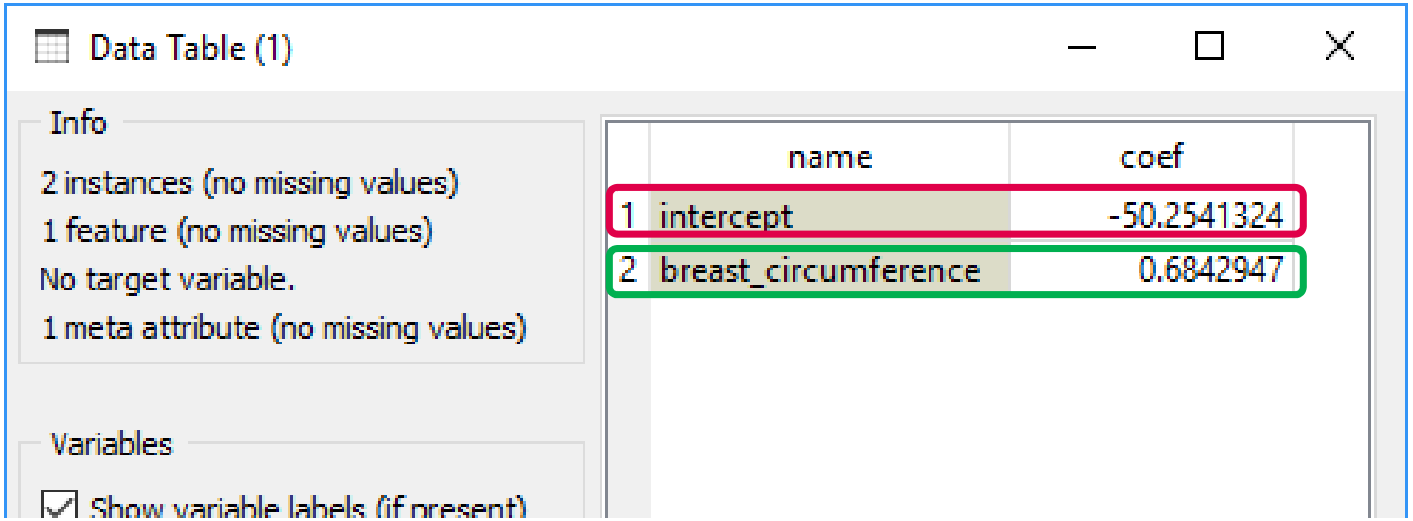

Data Table (1)
Info
2 instances (no missing values)
1 feature (no missing values)
No target variable.
1 meta attribute (no missing values)
Variables
Show variable labels (if present)
Visualize numeric values
Color by instance classes
Selection
Select full rows
Restore Original Order
Send Automatically
name coef
1 intercept -50.2541324
2 breast_circumference 0.6842947

# Exercise 6.1

a) What is the regression equation? (data table)

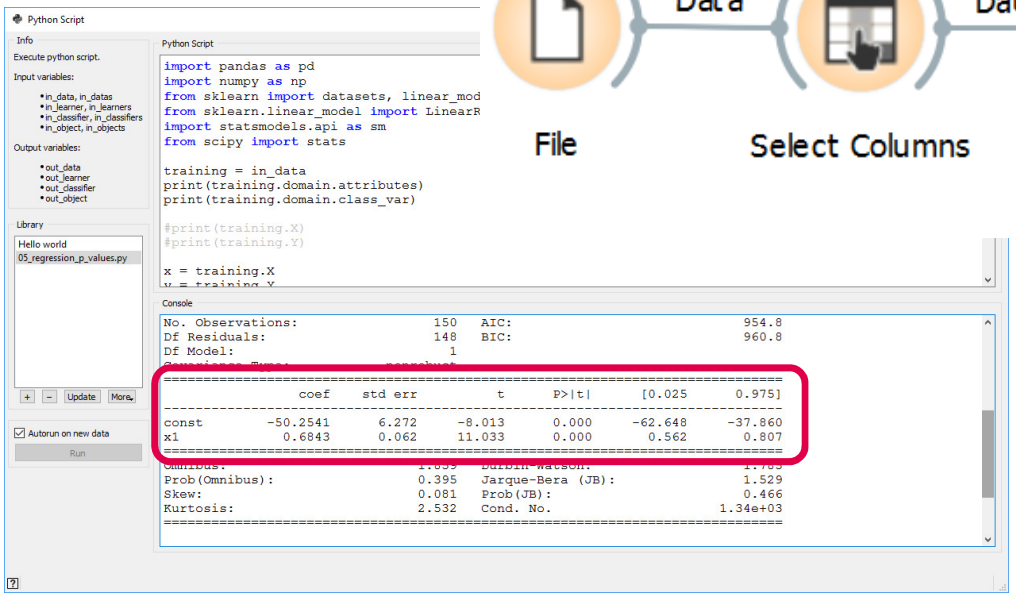
Intercept: -50.25 Coefficient: 0.684

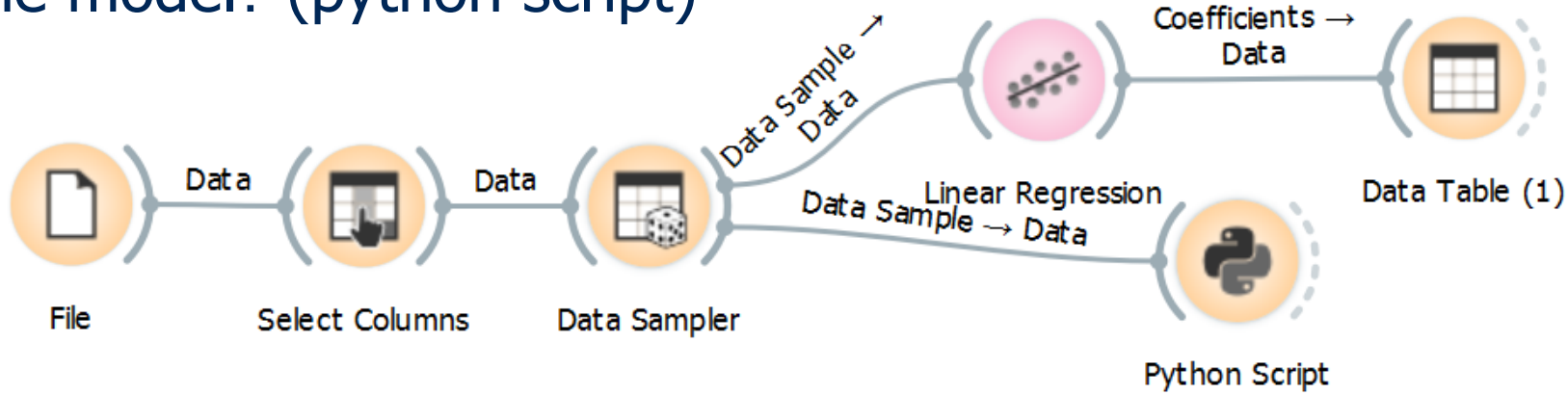
Equation: Y = 0.684*X – 50.25



# Exercise 6.1

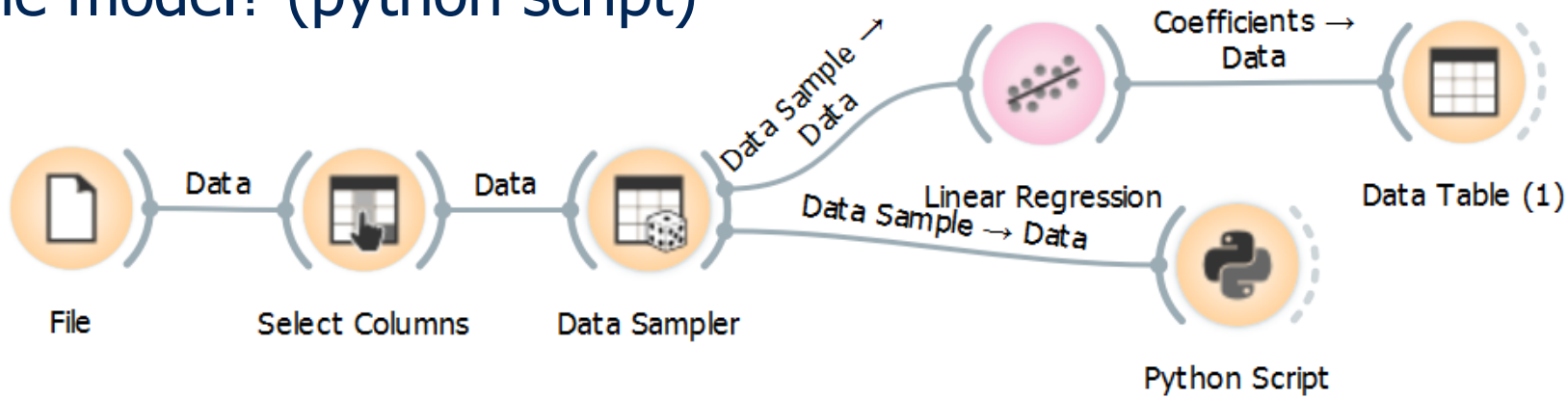
b) Are the intercept and X variable statistically significant for the model? (python script)



# Exercise 6.1

b) Are the intercept and X variable statistically significant for the model? (python script)




Covariance Type: nonrobust

| | coef | std err | t | P>\|t\| | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| const | -50.2541 | 6.272 | -8.013 | 0.000 | -62.648 | -37.860 |
| x1 | 0.6843 | 0.062 | 11.033 | 0.000 | 0.562 | 0.807 |

| | | | |
|---|---|---|---|
| Omnibus: | 1.859 | Durbin-Watson: | 1.785 |
| Prob(Omnibus): | 0.395 | Jarque-Bera (JB): | 1.529 |
| Skew: | 0.081 | Prob(JB): | 0.466 |
| Kurtosis: | 2.532 | Cond. No. | 1.34e+03 |

Intercept: (const)

p-value 0.000 => significant

Coefficient: (x1)

p-value 0.000 => significant

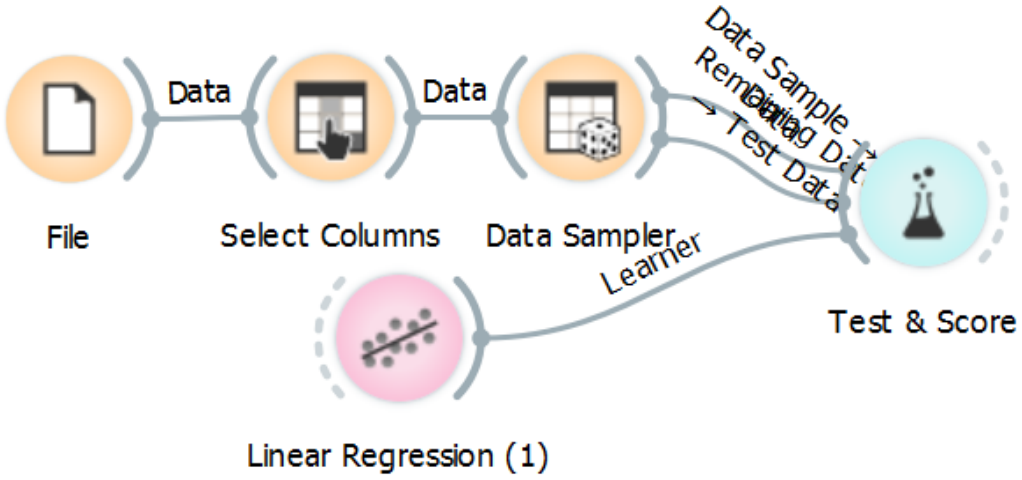
# Exercise 6.1

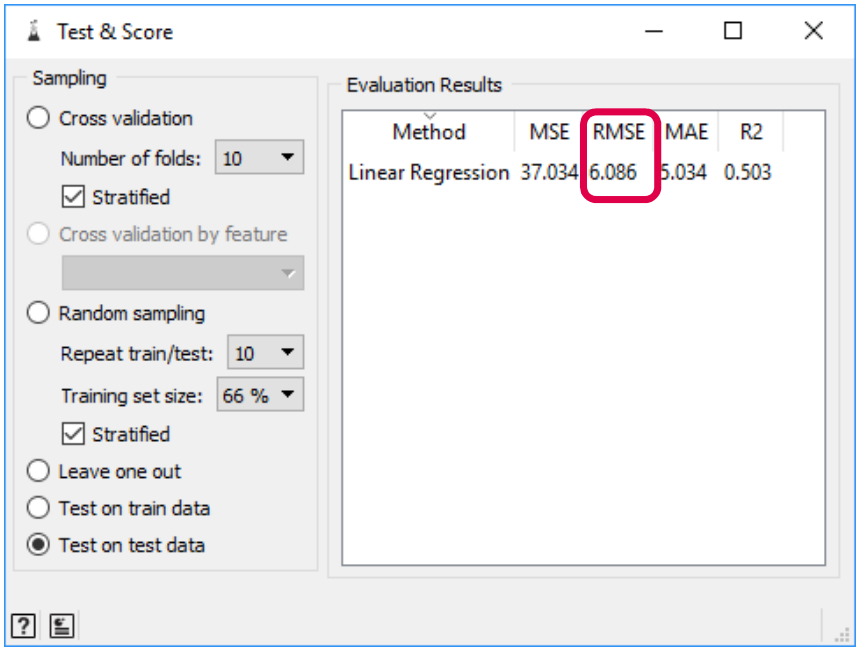
c) What is the Root Mean Squared Error?

i. Give the formula.

$$\mathrm{RMSD} = \sqrt{\frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}{n}}.$$

ii. How much is it? 6.086



# Exercise 6.1

d) What is the Mean Squared Error?

i. What is the relationship between MSE and RMSE?

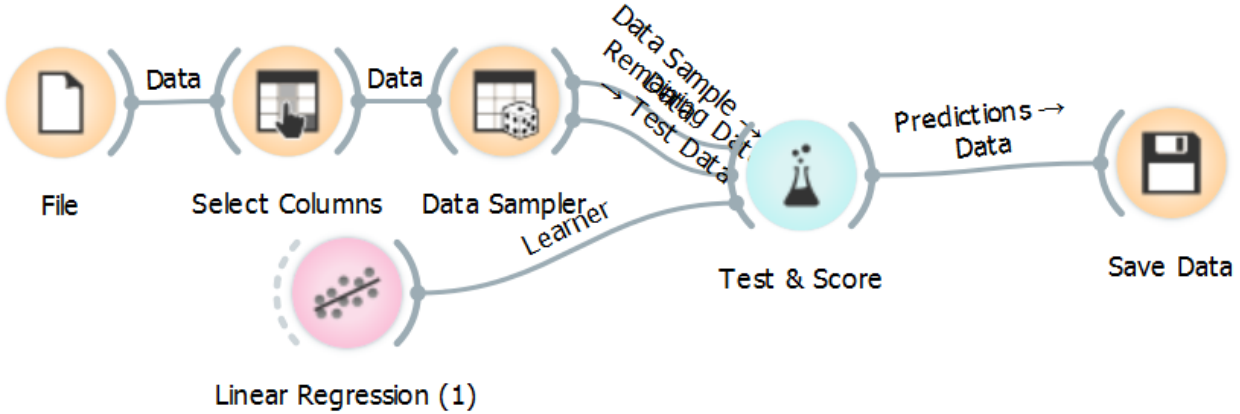
RMSE is the square root of MSE.

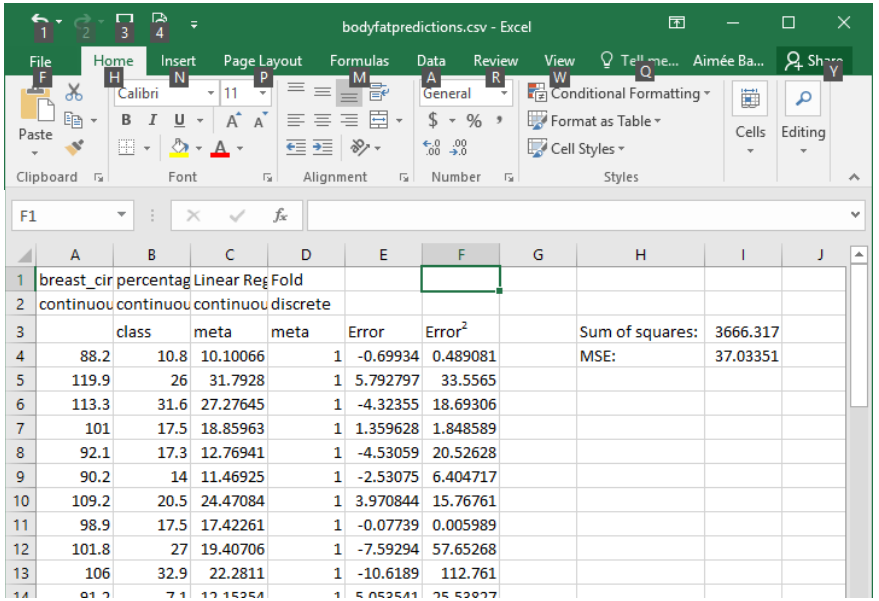
ii. Calculate the MSE for yourself.

Export Data using Save Data.

Use Excel for calculations.

MSE = 37.03



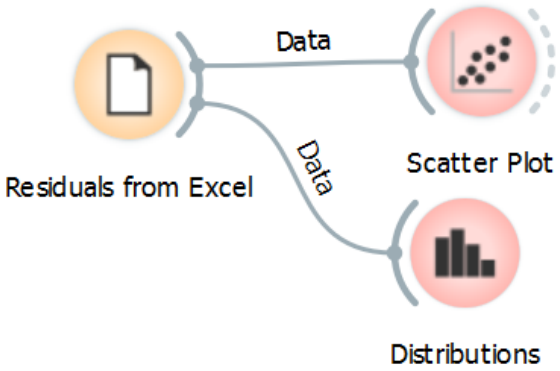




| | A | B | C | D | E | F | G | H | I | J |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | breast_cir | percentag | Linear Reg | Fold | | | | | | |
| 2 | continuou | continuou | continuou | discrete | | | | | | |
| 3 | | class | meta | meta | Error | Error$^2$ | | Sum of squares: | 3666.317 | |
| 4 | 88.2 | 10.8 | 10.10066 | 1 | -0.69934 | 0.489081 | | MSE: | 37.03351 | |
| 5 | 119.9 | 26 | 31.7928 | 1 | 5.792797 | 33.5565 | | | | |
| 6 | 113.3 | 31.6 | 27.27645 | 1 | -4.32355 | 18.69306 | | | | |
| 7 | 101 | 17.5 | 18.85963 | 1 | 1.359628 | 1.848589 | | | | |
| 8 | 92.1 | 17.3 | 12.76941 | 1 | -4.53059 | 20.52628 | | | | |
| 9 | 90.2 | 14 | 11.46925 | 1 | -2.53075 | 6.404717 | | | | |
| 10 | 109.2 | 20.5 | 24.47084 | 1 | 3.970844 | 15.76761 | | | | |
| 11 | 98.9 | 17.5 | 17.42261 | 1 | -0.07739 | 0.005989 | | | | |
| 12 | 101.8 | 27 | 19.40706 | 1 | -7.59294 | 57.65268 | | | | |
| 13 | 106 | 32.9 | 22.2811 | 1 | -10.6189 | 112.761 | | | | |

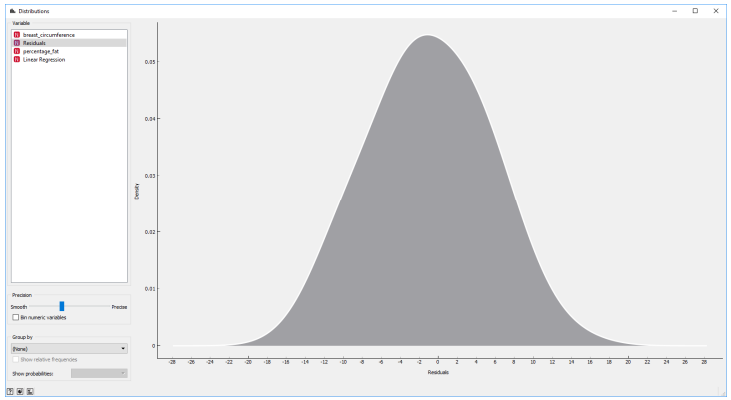
# Exercise 6.1

d) What is the Mean Squared Error?

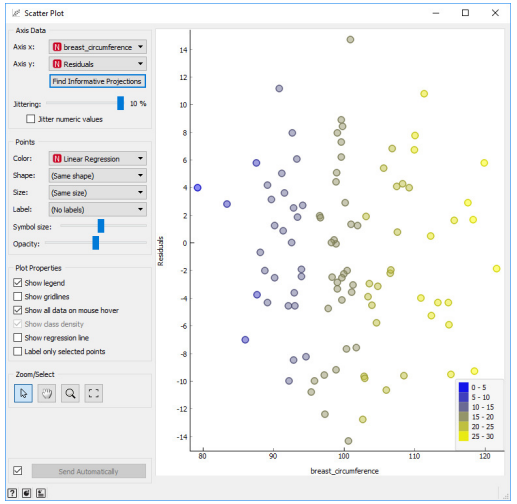
i. Do the residuals fulfill the conditions of independency and homoscedasticity? (Are they normally distributed and independent of X values?) Use a Scatterplot.




Load the Excel file (residuals)



Distribution appears normal



Residuals appear with the errors independent of breast circumference

UC Leuven Limburg

MOVING MINDS

# Any questions?