LATVIJAS UNIVERSITATE
ANNO 1919

UNIVERSITY OF LATVIA

Python library statmodels

statsmodels

Done by:
Mg. oec. Alexey Leontyev (al19087)

Rīga 2020

General description: *statsmodels* is a Python module that provides classes and functions for the estimation of many different statistical models, as well as for conducting statistical tests, and statistical data exploration

Installation by Anaconda:

```
conda install -c conda-forge statsmodels
```

Installation by PyPI (pip):

```
pip install statsmodels
```

Links:

The online documentation – statsmodels.org

Git repository – https://github.com/statsmodels/statsmodels

Statsmodels dependencies

The current minimum dependencies are:

- Python version >= 3.6
- NumPy version >= 1.15
- SciPy version >= 1.2
- Pandas version >= 0.23
- Patsy version >= 0.5.1

Optional dependencies:

- Matplotlib version >= 2.2

Data upload

- □ To get gathered statistical data to work with it used Pandas (from a local storage, this way I'll be doing in examples);
- □ Statsmodels also support online data from

  R-datasets by using

  sm.datasets.get_rdatase(*dataname*, *package='datasets'*, *cache=False*)

It's also utilizes nice feature of saving downloaded datasets in STATSMODELS_DATA folder in user home directory to avoid downloading dataset again and again (by setting cache=True).

Categories of analysis

- Linear Regression Models
- Discrete Choice Models
- Nonparametric Statistics
- Generalized Linear Models
- Robust Regression
- Generalized Estimating Equations
- Time Series Analysis
- State space models
- Forecasting
- Multivariate Methods

Why do we need regression analysis?

- Is it possible to prognose academic performance at the university knowing academic performance at the high school?
- Is it possible to predict test results if results of one test are available?
- How good is to use IQ results to prognose academic performance?
- Can we prognose economic changes of one index based on another?

Regression analysis

- Ways of prognosing values of one variable based on the value of another is a statistical method called regression analysis.

- While there are many types of regression analysis, at their core they all examine the influence of one or more independent variables on a dependent variable.

Regression analysis – simple linear regression

- □ X – independent variable (predictor, exogenous variable), which has values of $x_1, x_2, x_3, \dots, x_n$;
- □ Y – dependent variable (endogenous variable), which has values of $y_1, y_2, y_3, \dots, y_n$;

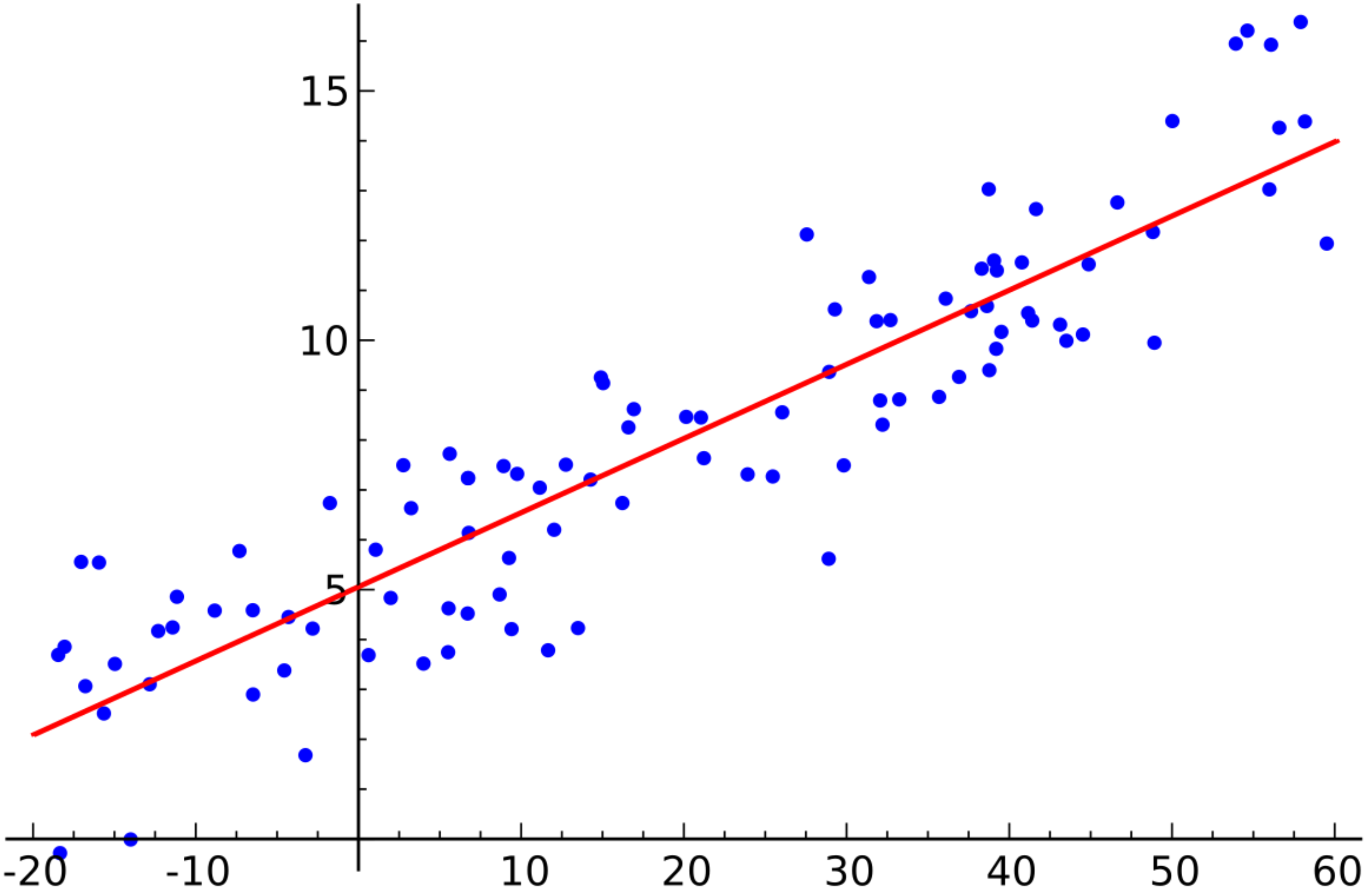
$$\hat{Y} = b_0 + b_1 * X, \qquad b_1 \neq 0,$$

where:

- □ b_0 – constant (or intercept)
- □ b_1 – β-coefficient

Y	X
5.1	24.4
0.6	20.8
5.7	23.7
3.3	26
-0.3	23.8
-4.3	20.8
-5.5	17.6
-9.1	15.3
-7.3	12.6
-3.2	12.2
0.4	11.6
-0.2	11.5
0	11.4

Ordinary least squares (OLS)



Regression analysis – higher order regression (polynomial regression)

If the straight line is not enough to describe the model it’s possible to use different shape of curve:

Polynomial regression for 2 degrees:

$$\hat{Y} = b_0 + b_1 * X + b_2 X^2$$

Polynomial regression for 3 degrees:

$$\hat{Y} = b_0 + b_1 * X + b_2 X^2 + b_3 X^3$$

Multiple linear regression

- Sometimes it’s not enough to try to explain changes in Y using only one X. In such cases it’s possible to use Multiple regression models which explain changes of Y using several independent variables ($X_{1,}X_{2,…,}X_m$).

Regression equation reflecting all independent variables and their β-coefficients:

$$\hat{Y} = \ b_0 + \ b_1 * X_1 + \ b_2X_2 +, \dots, +b_mX_m$$

Into the practice!

Models and topics to be covered:

- Simple linear regression model (by OLS)
- Higher order regression (polynomials)
- Basic forecasts
- Multiple linear regression
- Data standardization and standardized regression for Multiple linear regression

Thank you for your attention!