MachineLearnAthon - Microlecture
Regression analysis quide

MachineLearnAthon
A project Co-funded by the Erasmus+ programme of the European Union

Steps of regression analysis:

1. **Choose a model:**

Select a model that best describes the relationship between inputs (independent variables) and the output (dependent variable).

- **Linear Regression**
  Assumes a straight-line relationship: An increase in the input leads to an increase or decrease in the output in a constant proportion.

- **Non-Linear Regression**
  Captures relationships where the effect of input on output changes depending on the input level.Used when patterns cannot be adequately described by a straight line.

Choose a model

Estimate regression function

"Check it out." regression function

Steps of regression analysis:

2. **Estimate the regression function:**

- There is no single statistical test definitively says “use linear” or “use polynomial” regression.
- There are diagnostic tools and model comparison tests that help you decide whether a linear model is sufficient or if a polynomial (nonlinear) regression fits the data better.

Choose a model

Estimate
regression function

"Check it out."
regression function

Steps of regression analysis:

3. **Check the quality of the estimate:**

   - How well does the found model fit to explain the "response" (target variable)?
   - What contribution do the different factors of influence have in explaining the response?

   Look for:

   - R-squared
   - P-values
   - AIC/BIC

Choose a model

Estimate regression function

"Check it out." regression function

Example: margarine shop

- The sales manager of a large margarine manufacturer wonders whether the turnover achieved is due to the level of advertising costs!?

- The source of the data used are: Backhaus/Erichson/Plinke/Weiber - Multivariate Analysemethoden, Springer Verlag 2000

Quantity per year	Price per carton	Advertising per year	Customer visits per year
2 585	12,50	2 000	109
1 819	12,00	550	107
1 647	9,95	1 000	99
1 496	11,50	800	70
921	12,00	0	81
2 278	10,00	1 500	102
1 810	8,00	800	110
1 987	9,00	1 200	92
1 612	9,50	1 100	87
1 913	12,50	1 300	79

Example: margarine shop

Choose a model

$$\hat{Y} = b_0 + b_1 \times X$$

Estimate regression function

$$Y_i = b_0 + b_1 - X_i + e_i$$

The observed value of Y_i consists of:

- The proportion of $b_0 + b_1 - X_i$ quantity sold that can be explained by the regression model.
- Residual variation in the quantity sold e_i, which cannot be explained by the model (Residuum).

"Check it out." regression function

R-squared: 0.829					Adj. R-squared: 0.808			
	x_i	b_i	b_i [low]	b_i [high]	s_i	|t|	P
Constant	f(x1, x2)	1036.373	717.339	1355.407	138.349	7.491	0.0
Advertising per year	X1	0.752	0.474	1.030	0.121	6.235	0.0

b_0 b_1

technische universität dortmund
TU Delft
Co-funded by the Erasmus+ programme of the European Union
Machine LearnAthon

Example: margarine shop

Choose a model

Estimate
regression function

"Check it out."
regression function

	B = R-squared: 0.829				Adj. R-squared: 0.808		
	x_i	b_i	b_i [low]	b_i [high]	s_i	|t|	P
Constant	f(x1, x2)	1036.373	717.339	1355.407	138.349	7.491	0.0
Advertising per year	X1	0.752	0.474	1.030	0.121	6.235	0.0

B ("certainty factor") = R-Sq (= R²)

$$s = \sqrt{\frac{\sum_{i=1}^{n}\left(Y-\hat{Y}\right)^2}{n-2}}$$

$$R^2 = \frac{\sum_{i=1}^{n}\left(\hat{Y}-\bar{Y}\right)^2}{\sum_{i=1}^{n}(Y-\bar{Y})^2}$$

Dummy Variable Added

- **Reason:** A "dummy" variable (binary variable) was added to capture whether the advertising budget is above or below the median value.
- **Definition:**
  - HighAdvertising = 1 → Advertising budget is above the median.
  - HighAdvertising = 0 → Advertising budget is below or equal to the median.
- **Purpose:** The dummy variable was included to account for possible non-linear effects in the relationship between advertising budget and sales.

Assessment of the model

Basic Model Statistics

	Metric	Value
0	Coefficient of Determination (R)	93.28%
1	Adjusted Coefficient of Determination (R*)	87.90%
2	Residual Variance (s^2)	24418.68
3	Residual Standard Deviation (s)	156.26

Test for Independence of Predictor (F-test)

	Significance Level α	Critical Value	Test Statistic	Result
0	5.0%	5.190000	17.340000	Reject H0
1	1.0%	11.390000	17.340000	Reject H0
2	0.1%	31.090000	17.340000	Do not reject H0

Ramsey RESET Test

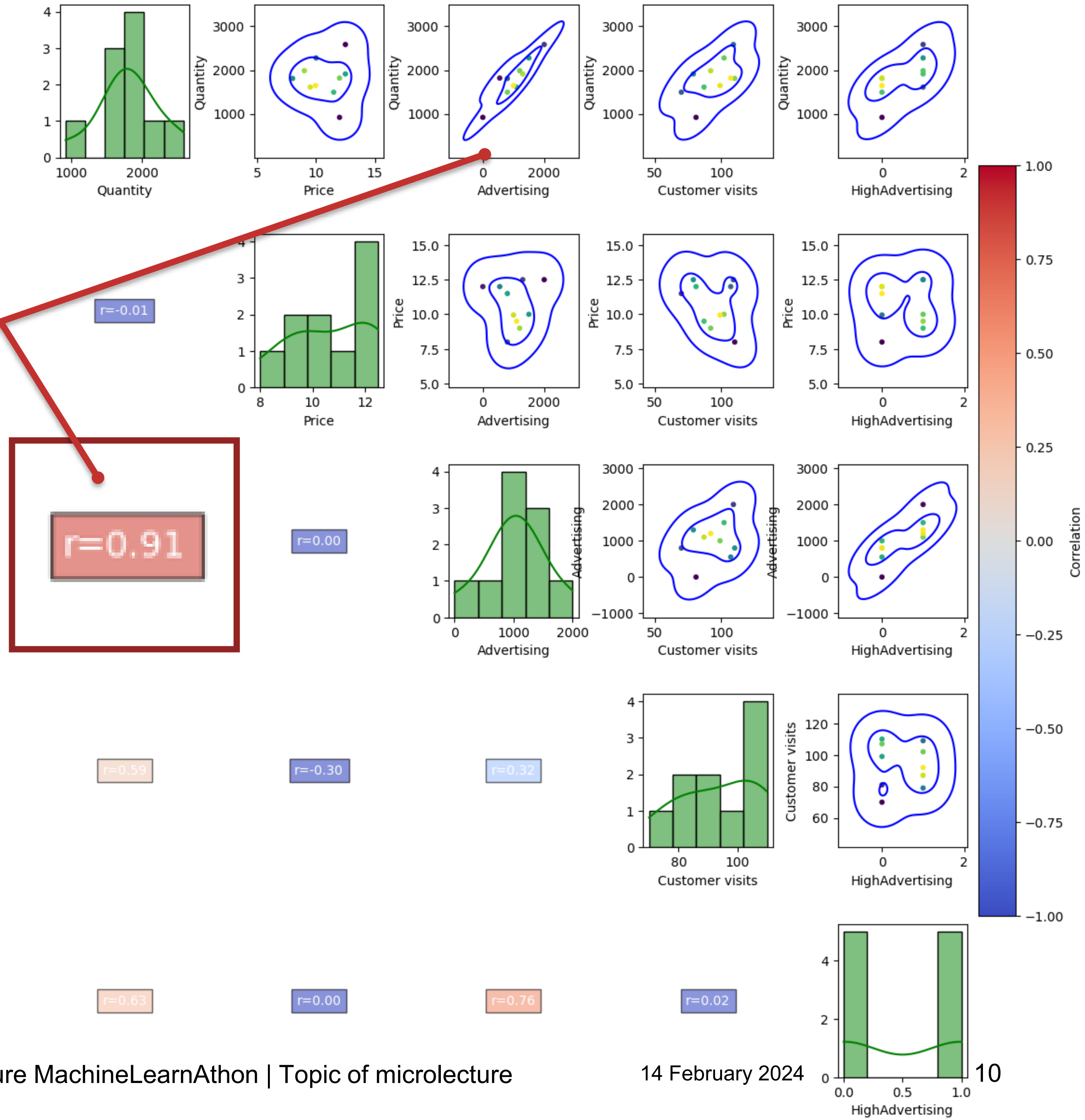
	Significance Level α	Critical Value	Test Statistic	Result
0	5.0%	5.190000	0.760000	Do not reject H0
1	1.0%	11.390000	0.760000	Do not reject H0
2	0.1%	31.090000	0.760000	Do not reject H0

Conclusions

	Test	Test Statistic	p-value	Conclusion
0	Test for Independence of Predictor	17.340000	0.003900	We reject the null hypothesis at least at one of the tested significance levels.
1	Ramsey RESET Test	0.760000	0.382520	We cannot reject the null hypothesis at any of the tested significance levels.

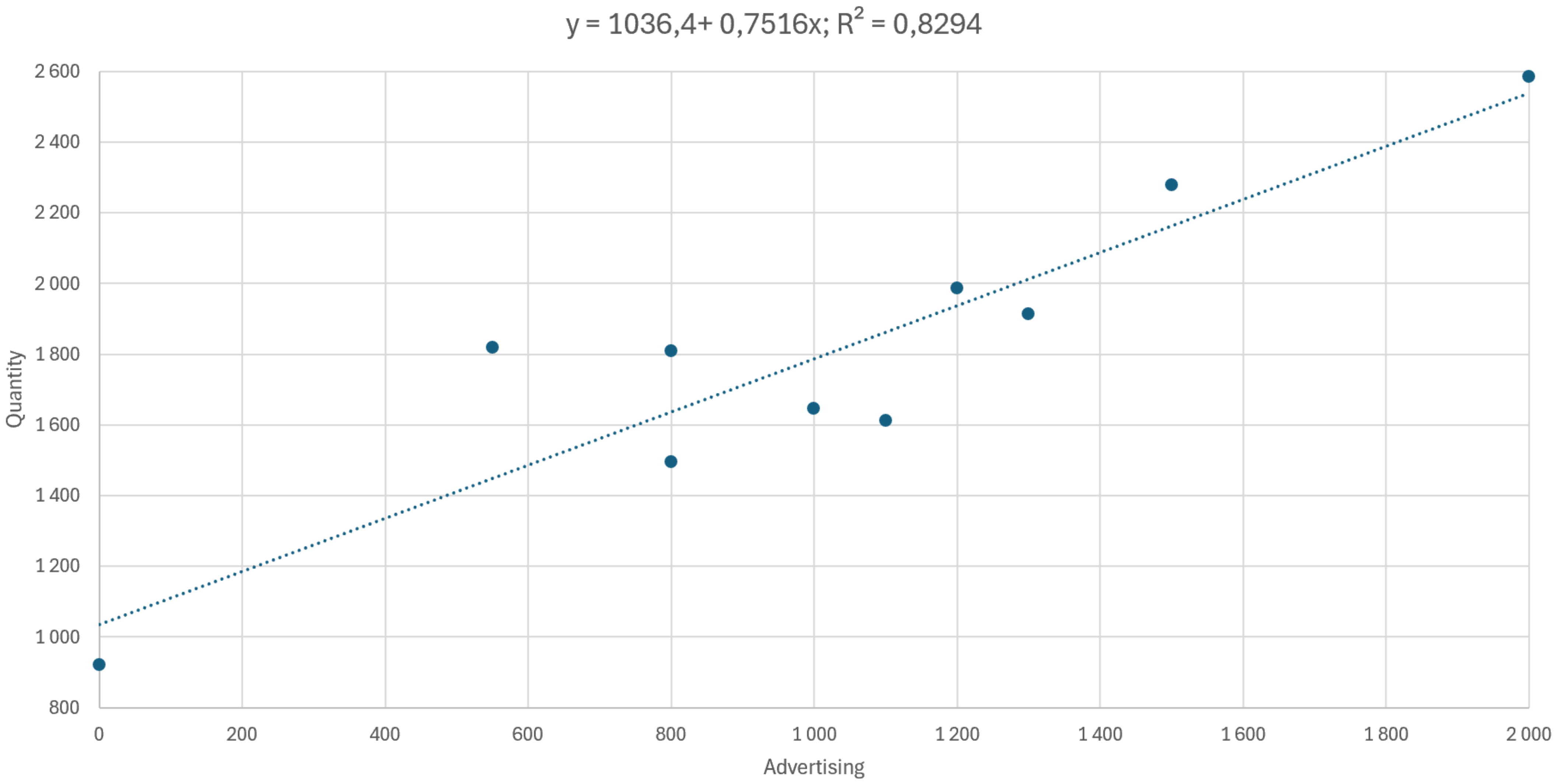
Example: margarine shop

There could be a dependency

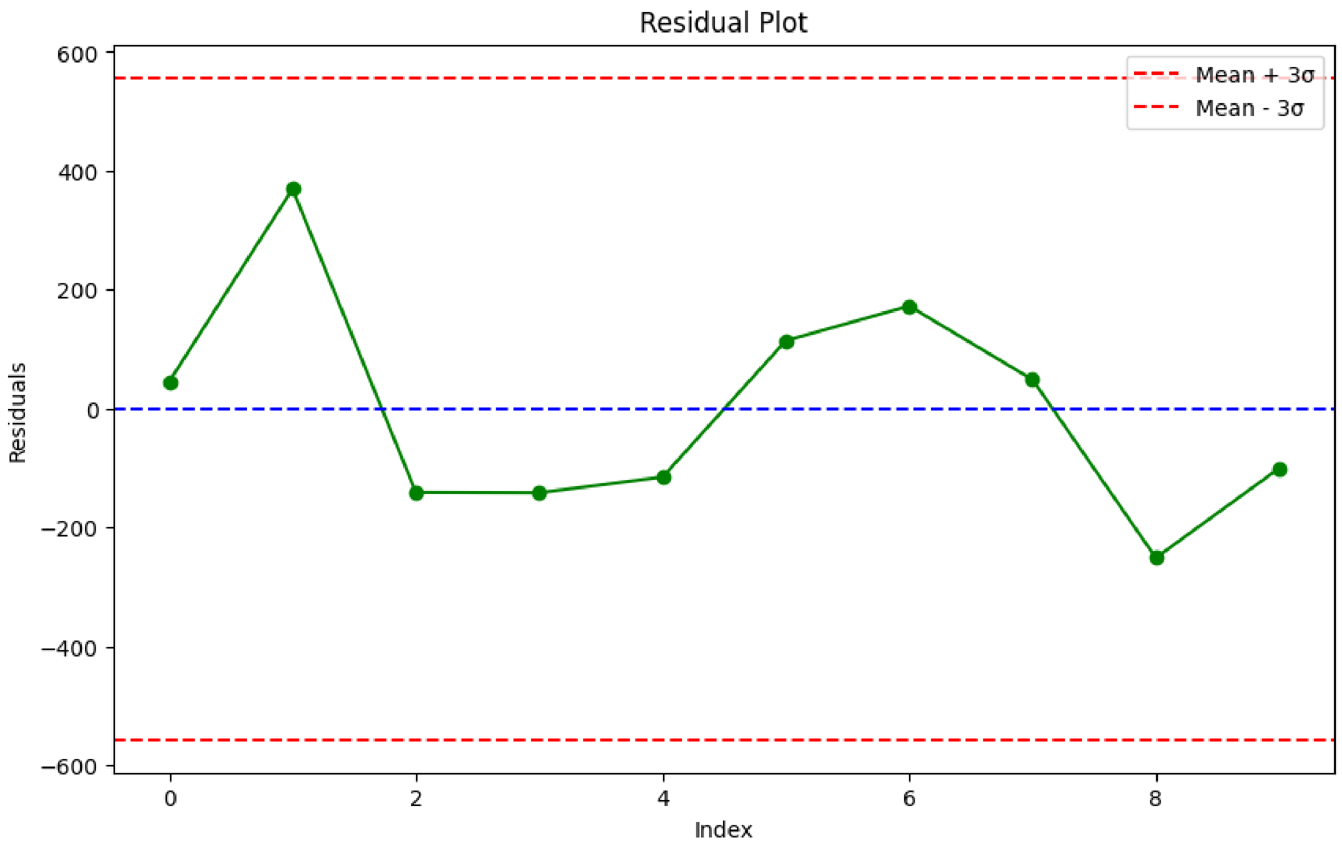


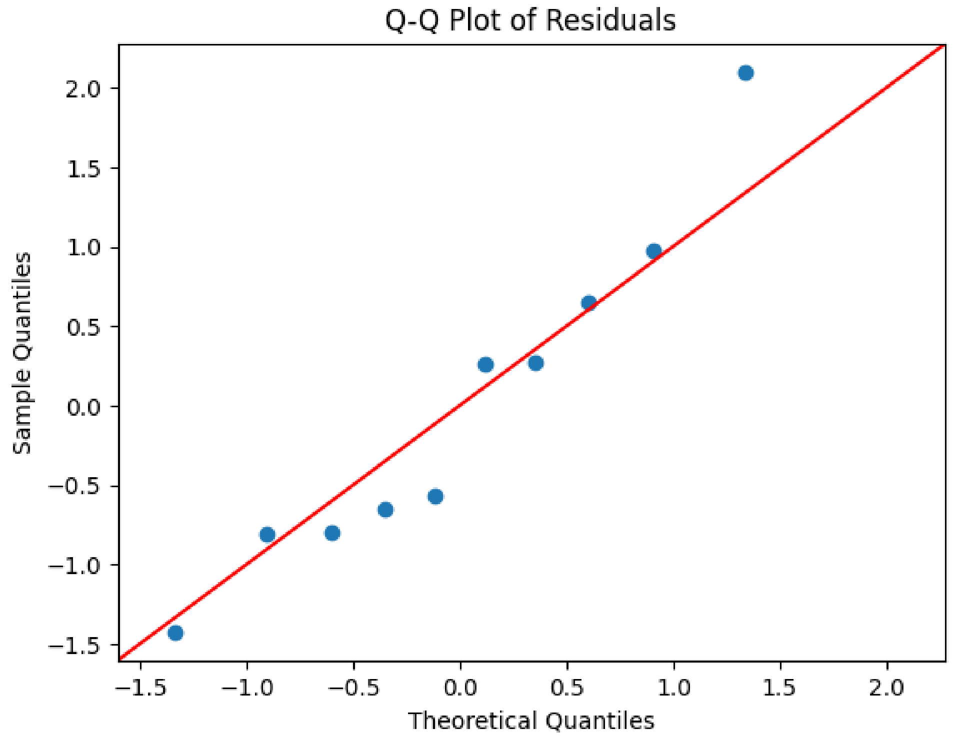
technische universität dortmund
TU Delft
Co-funded by the Erasmus+ programme of the European Union
Machine LearnAthon

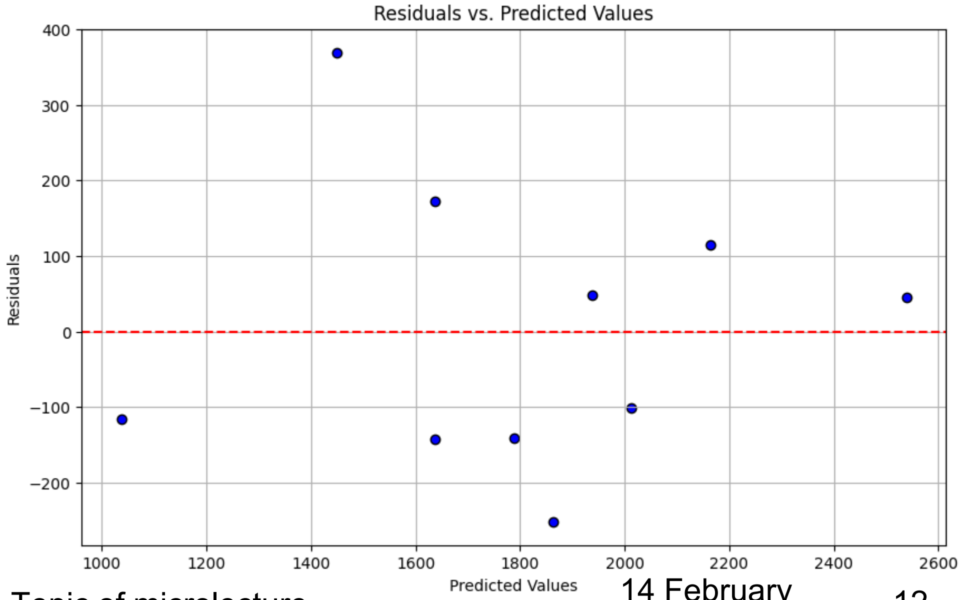
Quantity sold = f(advertising costs)



Residuum







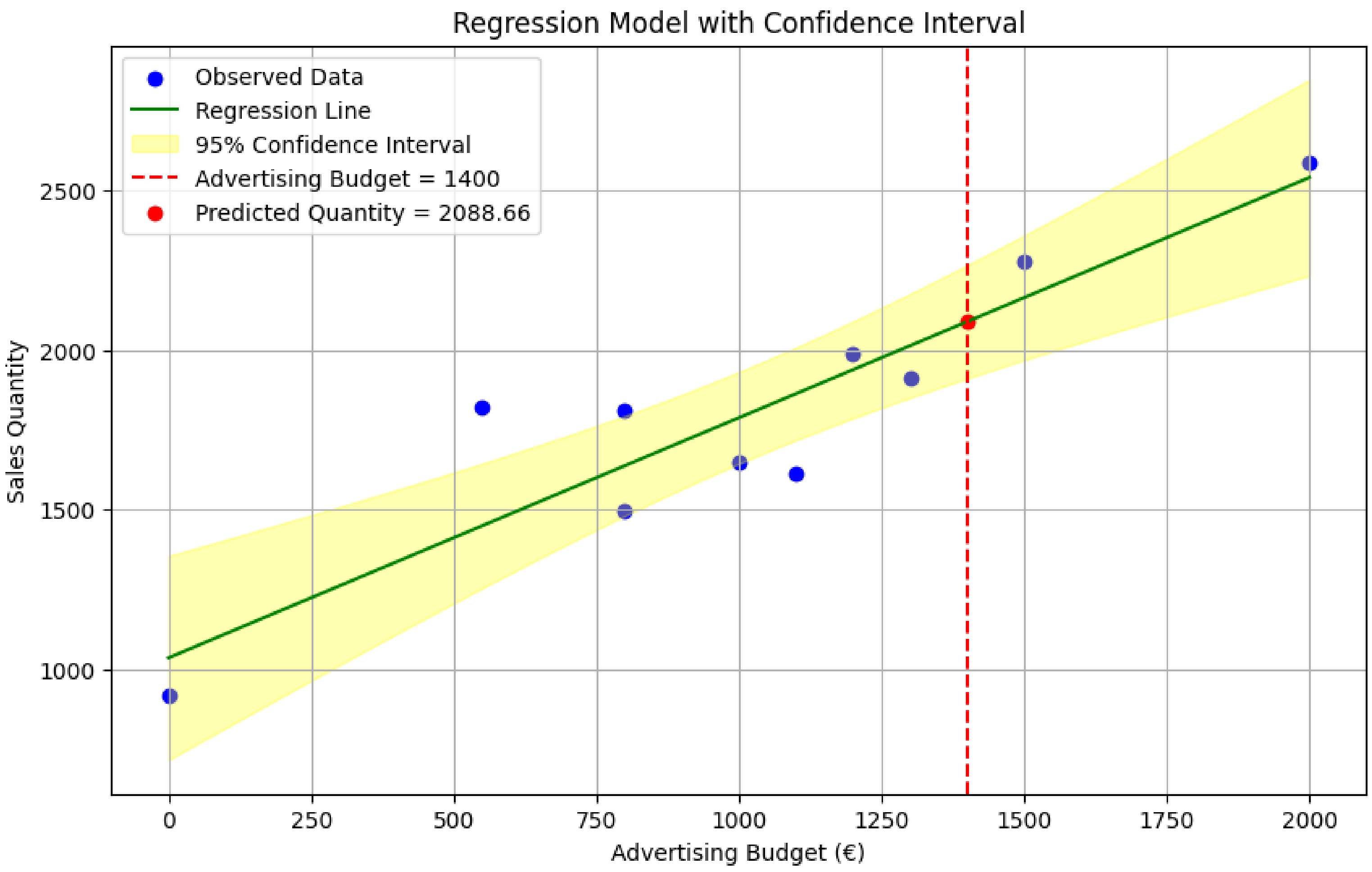
Co-funded by the Erasmus+ programme of the European Union

Machine LearnAthon

Forecast of Y values

- What value the response will have with a particular set of influence factors
- The marketing department is given a margarine advertising budget of 1,400 per month, what will be the expected turnover?

Best practice: Use interval estimates to predict the expected range of turnover.



Thank you very much for your attention!

Literature

- Backhaus, K., Erichson, B., Plinke, W., & Weiber, R. (2000). Multivariate Analysemethoden: Eine anwendungsorientierte Einführung (9th ed.). Springer. https://doi.org/10.1007/978-3-662-08893-7
- Burnham, K. P., & Anderson, D. R. (2002). Model selection and multimodel inference: A practical information-theoretic approach (2nd ed.). Springer. https://doi.org/10.1007/b97636
- Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). Applied multiple regression/correlation analysis for the behavioral sciences (3rd ed.). Routledge. https://doi.org/10.4324/9780203774441
- Hastie, T., Tibshirani, R., & Friedman, J. (2009). The elements of statistical learning: Data mining, inference, and prediction (2nd ed.). Springer. https://doi.org/10.1007/978-0-387-84858-7
- James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). An introduction to statistical learning: With applications in R (2nd ed.). Springer. https://doi.org/10.1007/978-1-0716-1418-1
- Lumley, T., & Diehr, P. (2002). The importance of the normality assumption in large public health data sets. Annual Review of Public Health, 23(1), 151–169. https://doi.org/10.1146/annurev.publhealth.23.100901.140546

Literature (cont.)

- Montgomery, D. C., Peck, E. A., & Vining, G. G. (2012). Introduction to linear regression analysis (5th ed.). Wiley. https://doi.org/10.1002/9781118625590
- Rencher, A. C., & Christensen, W. F. (2012). Methods of multivariate analysis (3rd ed.). Wiley. https://doi.org/10.1002/9781118391686
- Seber, G. A. F., & Lee, A. J. (2003). Linear regression analysis (2nd ed.). Wiley. https://doi.org/10.1002/9780471722199
- Weisberg, S. (2005). Applied linear regression (3rd ed.). Wiley. https://doi.org/10.1002/0471704091

The European Commission support for the production of this publication does not constitute an endorsement of the contents which reflects the views only of the authors, and the Commission cannot be held responsible for any use which may be made of the information contained therein.