LECTURER: Nghia Duong-Trung

MACHINE LEARNING

MACHINE LEARNING
TOPIC OUTLINE

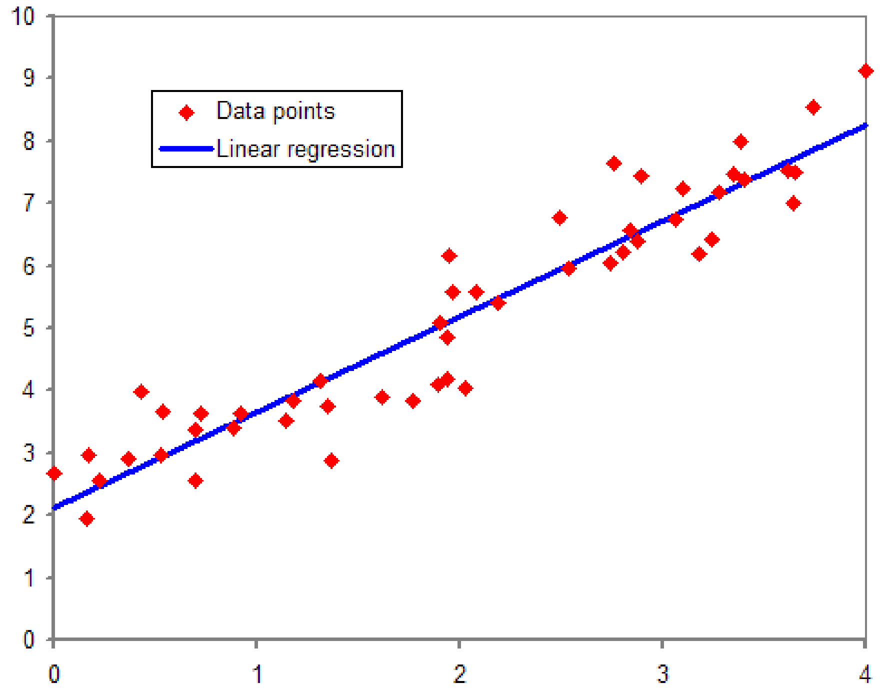
UNIT 3

REGRESSION

STUDY GOALS

- Know the definitions and terms used for regression
- Comprehend common applications of regression analysis
- Understand different methods for regression analysis
- Understand regularization for regression analysis
- Implement regression methods in Python

- Regression is a **supervised** learning approach for:
  - Estimating the **relationships** between the **dependent** variable and **independent** variable(s).



An example of Regression

Source of the text: Fernandez, 2020; Zöller, 2022.
Source of image: File: Normdist_regression.png (2023).

Regression applications

- Capital Asset Pricing
- Stock prediction
- Trajectory prediction
- Weather Forecasting
- Student marks prediction
- Diabetes prediction
- Crop yields prediction

Regression methods

- LINEAR
- NON-LINEAR
- LOGISTIC
- QUANTILE

Source of the text: Kumari, 2020; Brian, 2022; Zöller, 2022.

- **Linear regression**: the relationship is a expressed as a straight-line equation

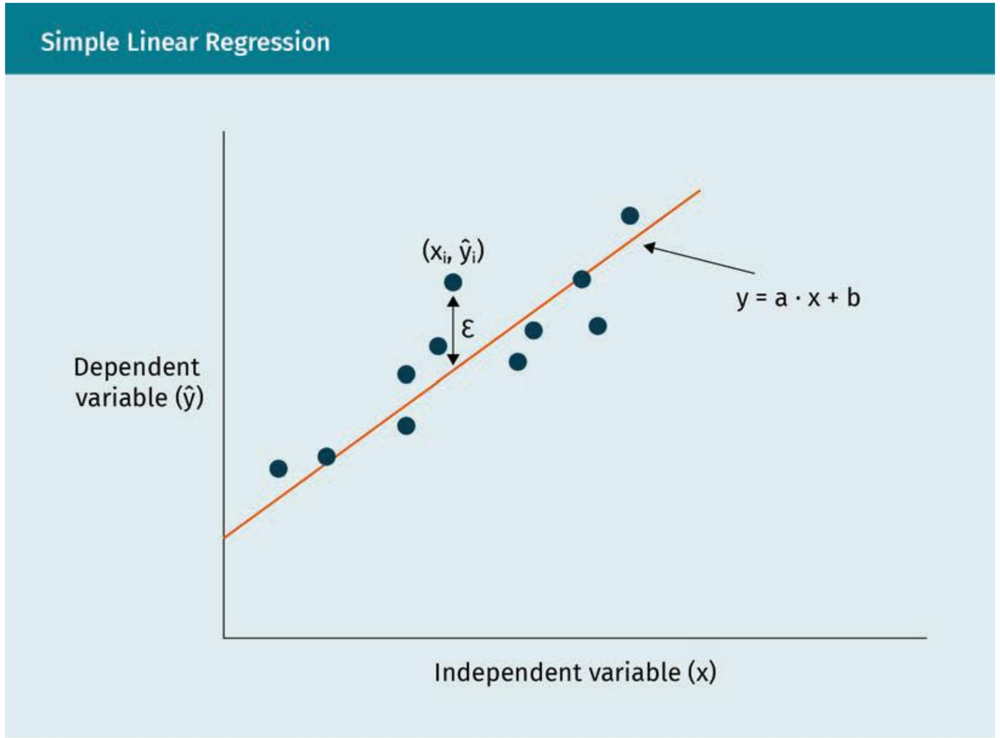
$$y = a_1 \cdot x_1 + a_2 \cdot x_2 + \cdots + a_n \cdot x_n + b$$

Where:

$\{x_1, x_2, \cdots, x_n\}$ - independent variables

$\{a_1, a_2, \cdots, a_n\}$ - coefficients or weights

$\{b\}$ – constant or bias



An example of Linear Regression

Source of the text: Kumari, 2020; Zöller, 2022.
Source of the image: Zöller (2022).

- **Nonlinear regression**: the relationship is a nonlinear equation (e.g., poliminal, exponential)

  $y = f(x, \alpha, b)$

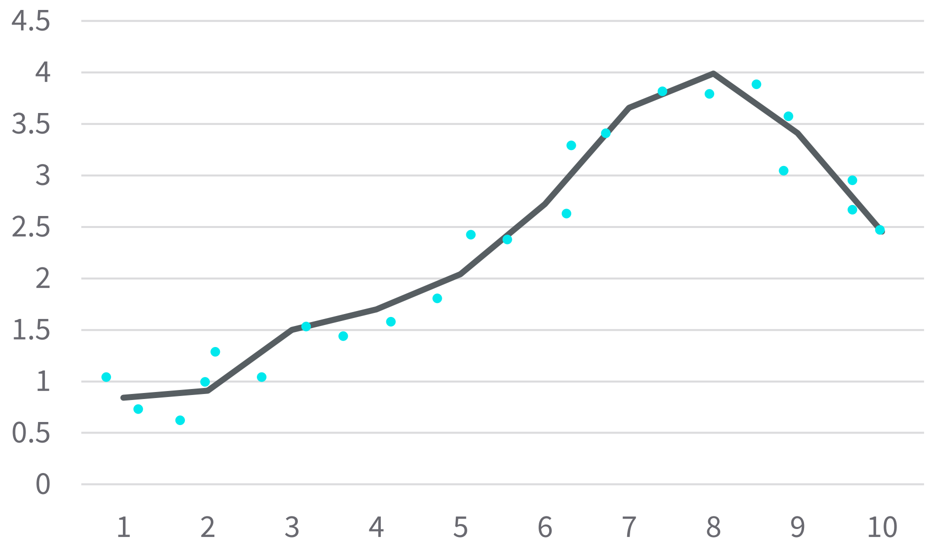
  Where:

  $\mathrm{x} = \{x_1, x_2, \cdots, x_n\}$ - variables

  $\alpha = \{a_1, a_2, \cdots, a_n\}$ - coefficients or weights

  $\{b\}$ - coefficients or weights

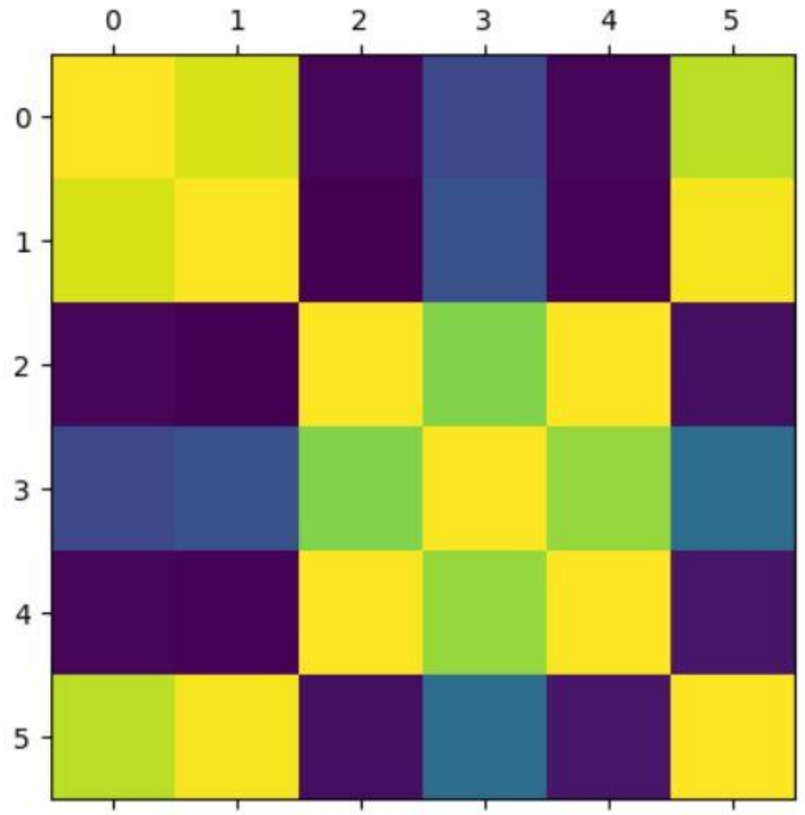
- **Linearization**: approximation of a nonlinear function to linear equation



An example of Nonlinear Regression

Source of the text: Zöller, 2022.

- Regression **steps**:
  - Selection: choice of model
  - Fitting: finding of unknown coefficients
  - Prediction: estimation of the target variable
  - Evaluation: checking difference between model's predictions and the desired values
- **Correlation** analysis: Describes how strong are the relationships between the variables

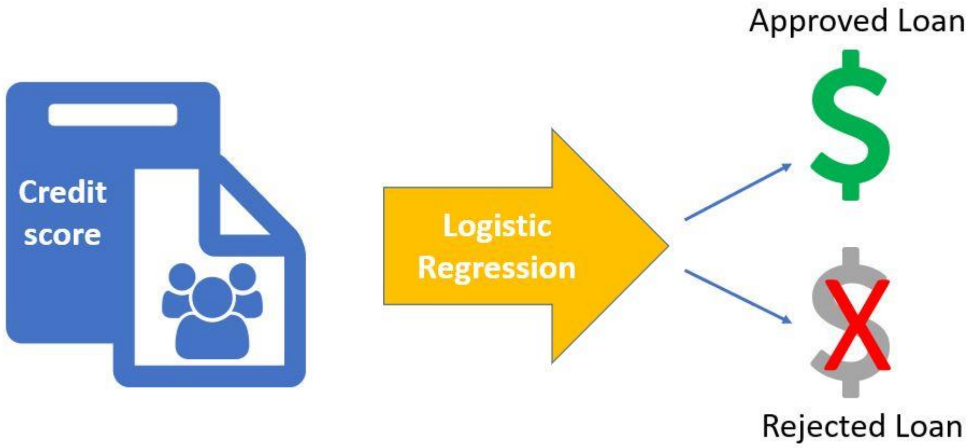


An example of Correlation analysis

Source of the text: Zöller, 2023.

- **Logistic regression**: uses a logistic function to model a binary dependent variable
- Equation:

Where: $p(y)$ - probability that $y = 1$



An application of Logistic Regression

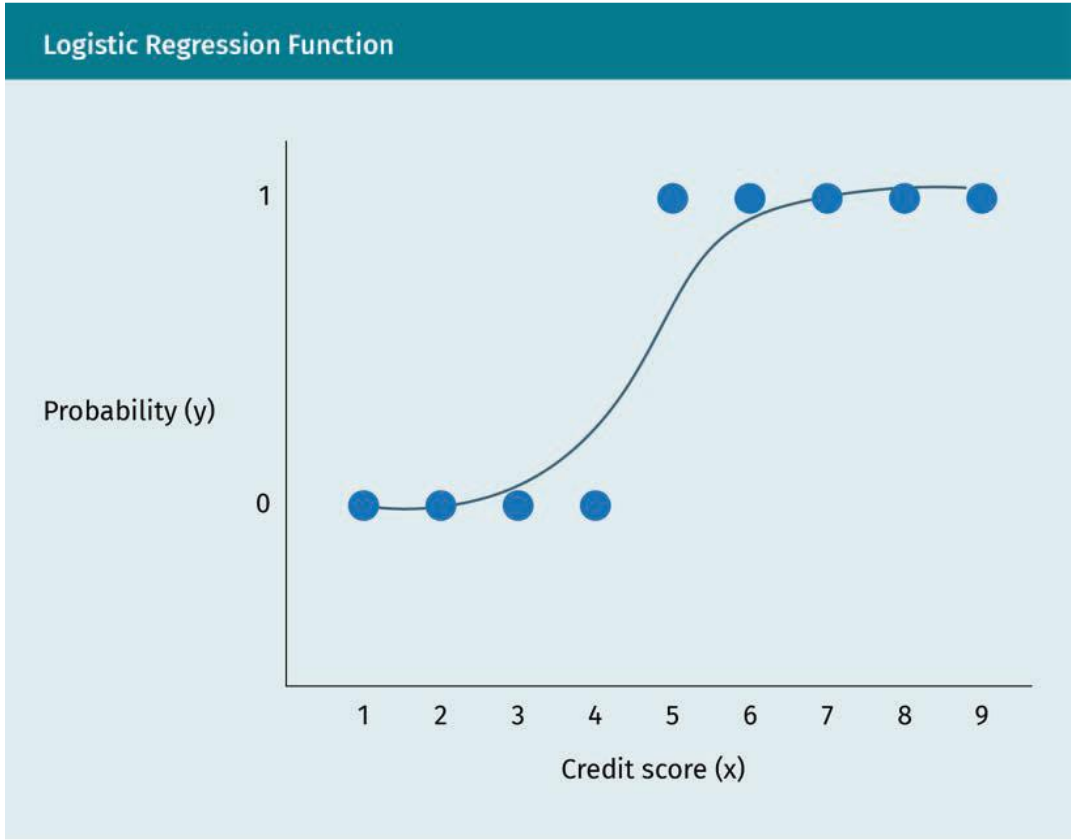
Source of the text: Bhattacharyya, 2018; Zöller, 2022.

- Logistic regression **equation**:

  $$p(y) = \frac{1}{1+e^{-(a \cdot x+b)}}$$

  Where: $p(y)$ - probability that $y = 1$

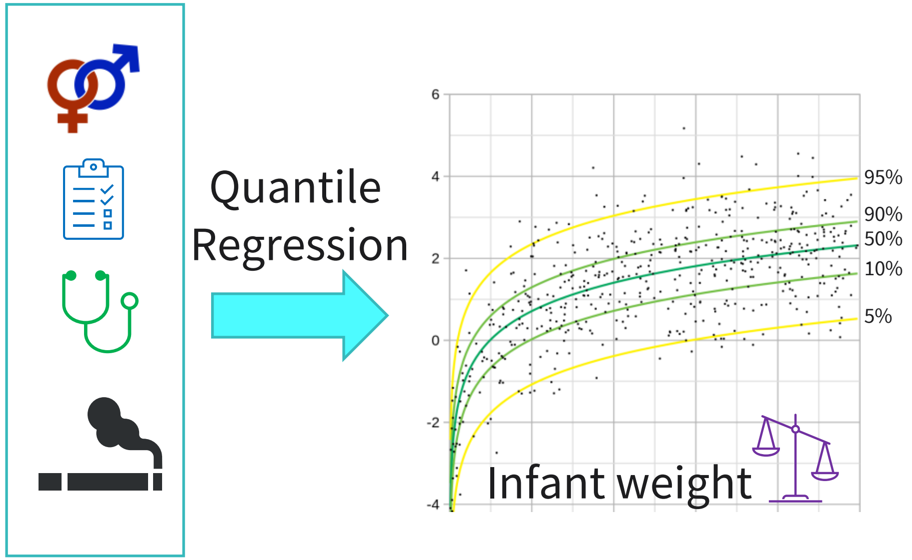
- **Maximum likelihood** $\log(\frac{p}{1-p})$ is used to identify the Regression curve:



An example of Logistic Regression Curve

Source of the text: Bhattacharyya, 2018; Zöller, 2022.
Source of the image: Zöller (2022).

- **Quantile regression:** an extension of linear regression
- Use case: the conditions of linear regression are not met
- Method: **divide** the dependent variable **into segments** (i.e., quantiles) and develop a linear regression for each quantile



Quantile regression analysis of infant's weight based on the knowledge of infant's gender, mother's marital status, pregnancy care, and smoking status

Source of the text: Dye, 2020; Koenker, & Hallock, 2001; Zöller, 2022.
Source of the image: File:Quantilsregression.svg. (n.d.)

- **Steps**:
  - Calculate the regression coefficients of the quantiles.
  - minimize a “weighted” sum of the absolute errors at each quantile:

$$\min\left(\tau \sum_{\substack{segments \\ above\ \tau}} |\hat{y} - y| + (1-\tau) \sum_{\substack{segments \\ below\ \tau}} |\hat{y} - y|\right)$$

where: τ – quantile level

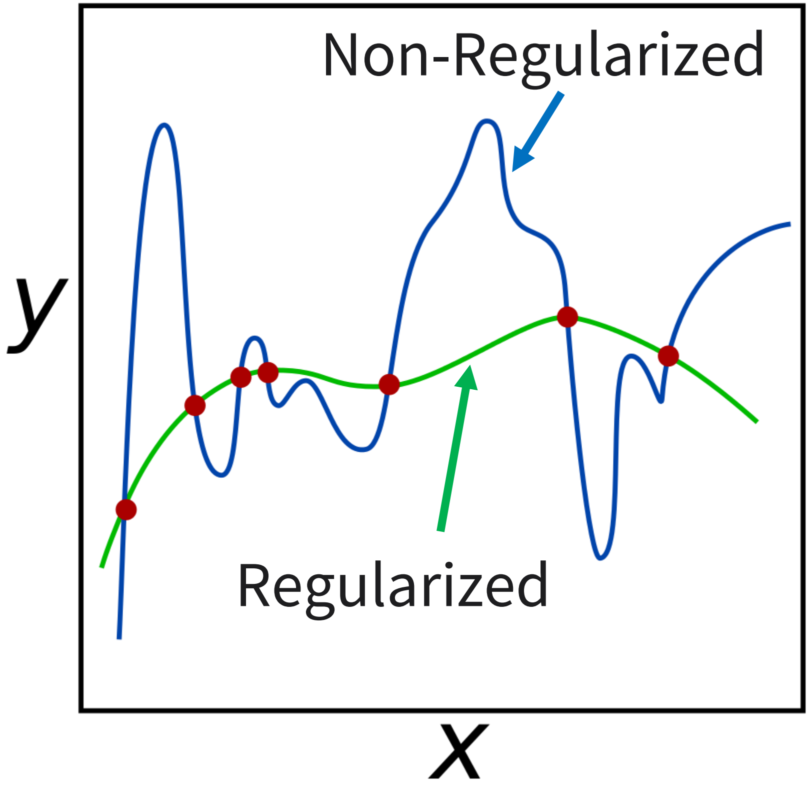
Quantile Regression Example

Coeffi-cient	Basic model	$\tau = 5\,\%$	$\tau = 10\,\%$	$\tau = 50\,\%$	$\tau = 90\,\%$	$\tau = 95\,\%$
a1	3224	2353	2608	3252	3856	4031
a2	161.1	227	171	149	141	165
a3	115.9	28	84	121	142	142
a4	–227	–536	–418	–164	–111	–57
b	–200.9	–255	–226	–190	–177	–199

An example of Quantile regression coefficients

Source of the text: Zöller, (2022).
Source of the image: Zöller (2022).

- Regularization: a ML process to
  - avoid **overfitting**
  - have more **robust** model
- Method: adding a **penalty** term to the regression model



An example of Regularization

Source of the text: Zöller, 2022.
Source of the image: File:Regularization.svg . (n.d.)

Ridge regression:

- L_2 regularization
- Penalty term in lost function: square of model coefficients

$$E = \sum(\hat{y} - y)^2 + \boxed{\lambda \sum W^2}$$

Where:

λ – constant controlling penalty level

W - model coefficients

- mainly punishes the largest coefficients

Lasso regression:

- L_1 regularization
- Penalty term in lost function: sum of absolute values of model coefficients

$$E = \sum(\hat{y} - y)^2 + \boxed{\lambda \sum |W|}$$

Where:

λ – constant controlling penalty level

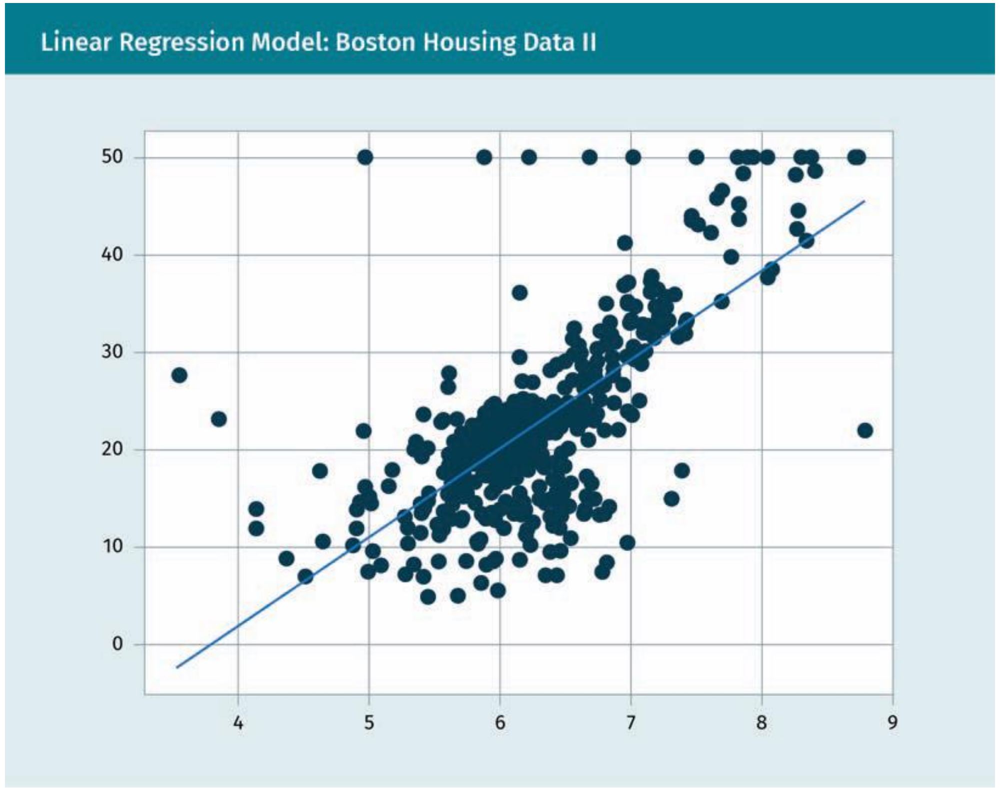
W - model coefficients

- Punishes both large and small coefficients

Source of the text: Zöller, 2022.

3.5 REGRESSION ANALYSIS IN PYTHON

```
>>> # Linear regression with Python
>>> import pandas as pd
>>> import matplotlib.pyplot as plt
>>> plt.style.use('ggplot')
>>> from sklearn import datasets
>>> from sklearn import linear_model
>>> import numpy as np
>>> # Load dataset
>>> bostonData = datasets.load_boston() % built-in dataset
>>> yb = bostonData.target.reshape(-1, 1)
>>> Xb = bostonData['data'][:,5].reshape(-1, 1)
>>> plt.scatter(Xb,yb)
>>> plt.ylabel('value of house /1000 ($)')
>>> plt.xlabel('number of rooms')
>>> plt.show()
>>> regr = linear_model.LinearRegression() # Create the model
>>> regr.fit( Xb, yb) # Train the model
>>> plt.scatter(Xb, yb, color='black‘)
>>> plt.plot(Xb, regr.predict(Xb), color='blue', linewidth=3)
>>> plt.show()
```

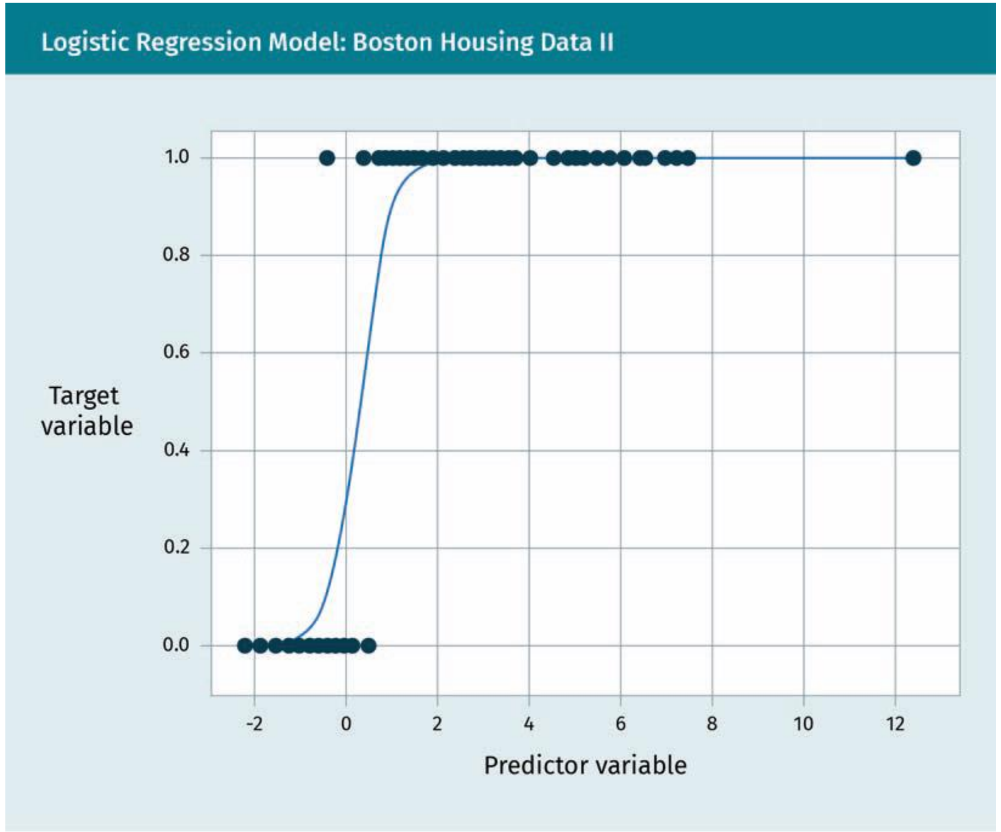


Linnear regression analysis

Source of the text: Zöller, 2022.
Source of the image: Zöller (2022).

3.5 REGRESSION ANALYSIS IN PYTHON

```
>>> # Logistic regression with Python
>>> import pandas as pd
>>> import matplotlib.pyplot as plt
>>> plt.style.use('ggplot')
>>> from sklearn import datasets, linear_model
>>> import numpy as np
>>> X1 = np.random.normal(size=150)
>>> y1 = (X1 > 0).astype(np.float)
>>> X1[X1 > 0] *= 4
>>> X1 += .3 * np.random.normal(size=150)
>>> X1= X1.reshape(-1, 1)
>>> plt.scatter(X1,y1);  plt.ylabel('y1‘); plt.xlabel('X1‘); plt.show()
>>> lm_log = linear_model.LogisticRegression()
>>> lm_log.fit(X1, y1)
>>> X1_ordered = np.sort(X1, axis=0)
>>> plt.scatter(X1.ravel(), y1, color='black', zorder=20 , alpha = 0.5)
>>> plt.plot(X1_ordered, lm_log.predict_proba(X1_ordered)[:,1], color='blue' ,
linewidth = 3)
>>> plt.ylabel('target variable‘);  plt.xlabel('predictor variable')
>>> plt.show()
```

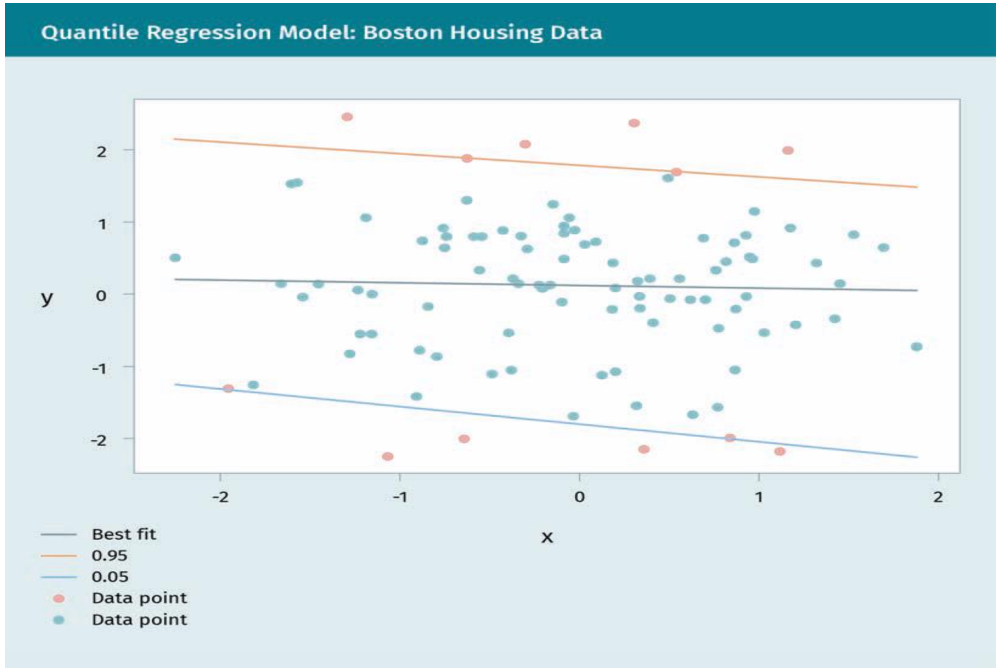


Logistic regression analysis

Source of the text: Zöller, 2022.
Source of the image: Zöller (2022).

3.5 REGRESSION ANALYSIS IN PYTHON

```
>>> # Quantile regression with Python
>>> X1 = np.random.normal(size=150)
>>> import numpy as np
>>> import pandas as pd
>>> import matplotlib.pyplot as plt
>>> import statsmodels.formula.api as smf
>>> df = pd.DataFrame(np.random.normal(0, 1, (100, 2)))
>>> df.columns = ['x', 'y']; >>> x = df['x'];  y = df['y']
>>> fit = np.polyfit(x, y, deg=1)
>>> _x = np.linspace(x.min(), x.max(), num=len(y))
>>> model = smf.quantreg('y ~ x', df)
>>> quantiles = [0.05, 0.1, 0.25, 0.5, 0.75, 0.95]
>>> fits = [model.fit(q=q) for q in quantiles]
>>> _y_005 = fits[0].params['x'] * _x + fits[0].params['Intercept']
>>> _y_095 = fits[5].params['x'] * _x + fits[5].params['Intercept']
>>> p = np.column_stack((x, y))
>>> a = np.array([_x[0], _y_005[0]]) #first point of 0.05 quantile fit line
>>> b = np.array([_x[-1], _y_005[-1]]) #last point of 0.05 quantile fit line
>>> a_ = np.array([_x[0], _y_095[0]])
>>> b_ = np.array([_x[-1], _y_095[-1]])
>>> mask = lambda p, a, b, a_, b_: (np.cross(p-a, b-a) > 0) | (np.cross(p-
a_, b_-a_) < 0)
>>> mask = mask(p, a, b, a_, b_)
>>> figure, axes = plt.subplots()
>>> axes.scatter(x[mask], df['y'][mask], facecolor='r',
edgecolor='none', alpha=0.3, label='data point')
>>> axes.scatter(x[~mask], df['y'][~mask], facecolor='g',
edgecolor='none', alpha=0.3, label='data
>>> axes.plot(x, fit[0] * x + fit[1], label='best fit', c='lightgrey')
>>> axes.plot(_x, _y_095, label=quantiles[5], c='orange')
>>> axes.plot(_x, _y_005, label=quantiles[0], c='lightblue')
>>> axes.legend() ;  axes.set_xlabel('x');  axes.set_ylabel('y')
>>> plt.show()
```



Quantile regression analysis

Source of the text: Zöller, 2022.
Source of the image: Zöller (2022).

REVIEW STUDY GOALS

- Know the definitions and terms used for regression
- Comprehend common applications of regression analysis
- Understand different methods for regression analysis
- Understand regularization for regression analysis
- Implement regression methods in Python

SESSION 3

TRANSFER TASK

1. Create the dataset using the following code:

```
>>> from sklearn import datasets
>>> X, y = datasets.load_diabetes(return_X_y=True)
```

Implement linear regression analysis

2. Create the dataset using the following code:

```
>>> import numpy as np
>>> X = np.random.normal(size=50)
>>> y = (X > 0).astype(np.float)
>>> X[X > 0] *= 2
>>> X += .3 * np.random.normal(size=50)
>>> X= X.reshape(-1, 1)
```

Implement logistic regression analysis

3. Create the dataset using the following code:

```
>>> import numpy as np
>>> rng = np.random.RandomState(30)
>>> x = np.linspace(start=0, stop=10, num=100)
>>> X = x[:, np.newaxis]
>>> y_true_mean = 10 + 0.5 * x
```

Implement quantile regression analysis

TRANSFER TASK
PRESENTATION OF THE RESULTS

Please present your results.

The results will be discussed in plenary.

1. Which of the following features is true about regularized regression?
   a) It cannot help with model selection.
   b) It cannot help with variance trade-off.
   c) It can help with bias variance trade-off.
   d) All of these are true.

2. The grade a student earns in a competitive exam in relation to study time and other related factors can be estimated using a __________ regression model.?

a) Linear
b) Multilinear
c) Logistic
d) One-dimensional Polynomial

LEARNING CONTROL QUESTIONS

3. Logistic regression is used to predict __________ valued output.

a) Discrete
b) Continuous
c) Maximum
d) Minimum

4. In _________ regression, there is ________ dependent variable and ________ independent variable(s).

   a) Simple linear, multiple, one
   b) Simple linear, one, multiple
   c) Multiple, multiple, multiple
   d) Multiple, one, multiple

5. Which of the following types of cost functions is used for univariate linear regression?

a) Squared error
b) Simple error
c) Logarithmic error
d) F-score

LIST OF SOURCES

Text:

Zöller, T. (2022). Course Book – Machine Learning. *IU International University of Applied Science.*

Fernandez, J. (2020). Introduction to Regression Analysis. https://towardsdatascience.com/introduction-to-regression-analysis-9151d8ac14b3

Brian, B. (2022). What is regression? Definition, Calculation and Example. https://www.investopedia.com/terms/r/regression.asp

Kumari, R. (2020). Simple Linear Regression. Application, Limitation & Example. https://www.analyticssteps.com/blogs/simple-linear-regression-applications-limitations-examples

Bhattacharyya, S. (2018). Logit of logistic regression: Understanding the fundamentals. https://towardsdatascience.com/logit-of-logistic-regression-understanding-the-fundamentals-

Dye, S. (2020). Quantile Regression. https://towardsdatascience.com/quantile-regression-ff2343c4a03F384152a33d1.

Koenker, R., & Hallock, K. F. (2001). Quantile regression. Journal of Economic Perspectives, 15(4), 143—156. http://doi.org/10.1257/jep.15.4.143

Image:

Zöller (2022).

File:Normdist_regression.png. (2023, January 22). *In Wikimedia Commons, the free media repository*. Retrieved, January 29, 2023, from https://en.wikipedia.org/wiki/Regression_analysis

File:Quantilsregression.svg. (2022, November 23). *In Wikimedia Commons, the free media repository*. Retrieved, January 30, 2023, from https://en.wikipedia.org/wiki/Quantile_regression

File:Regularization.svg. (2023, January 15). *In Wikimedia Commons, the free media repository*. Retrieved, January 30, 2023. https://en.wikipedia.org/wiki/Regularization_%28mathematics%29

File:Combotrans.svg. (2023, January 30). In *Wikimedia Commons, the free media repository*. Retrieved, January 30, 2023. https://en.wikipedia.org/wiki/Gender