Linear regression

Predicting numeric values

Learning goals

1. Learn the idea, limitations and applicability of multiple linear regression in predicting values of numerical variables.
2. Learn to carry out multiple linear regression analysis in Python.

Foundations

- Linear regression analysis is a method to describe how the values in a variable of interest depend on on other variables.
- E.g. electricity consumption depends on the size of the flat, number of inhabitants, number of refridgerators etc.
- The variable of interest is called response variable.
  - In the example, it is the energy consumption.
  - The response variable must be of interval or ratio scale.
- The remaining variables are explanatory variables.
  - They must be numeric and of at least interval scale.

Categorical to numeric variables

Original encoding		*Re-encoding option 1*					*Re-encoding option 2*		
EyeColour		**Blue**	**Brown**	**Grey**	**Green**		**Blue**	**Brown**	**Grey**
Blue		1	0	0	0		1	0	0
Brown		0	1	0	0		0	1	0
Grey	→	0	0	1	0	or	0	0	1
Green		0	0	0	1		0	0	0
Brown		0	1	0	0		0	1	0
Blue		1	0	0	0		1	0	0

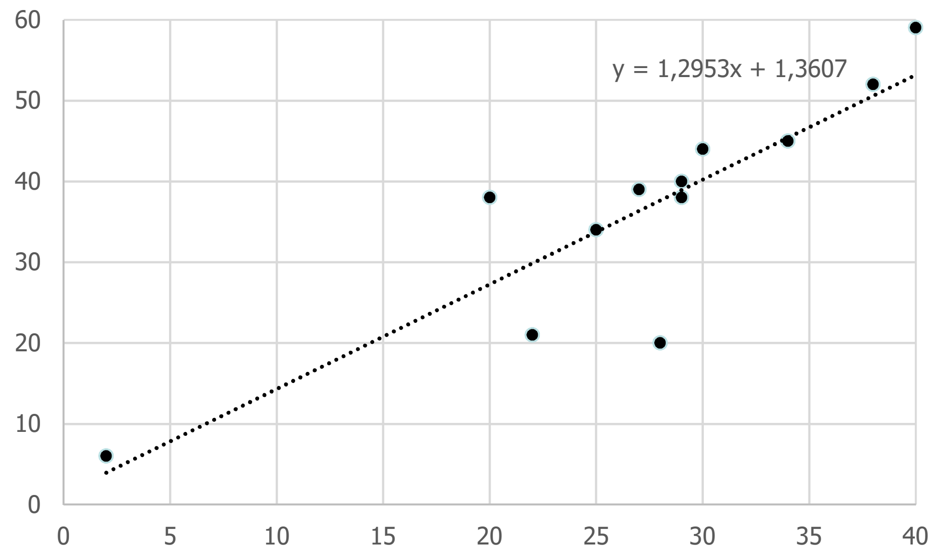
- For linear regession, categorical variables should be re-encoded as multiple dummy variables.
- Example: eye colour with four values: blue, brown, grey, green.
- For n categories, there are two options:
  - n dummy variables (one per category, option 1 above)
  - $n-1$ dummy variables (one per gategory, omitting one category, option 2 above)

Model construction and predicting

- After data preprocessing, the first step is to build a model.
  - To do this, we need a training set that contains the values of both the explanatory variables and the response variable.
  - In the example, the electricity company could use historical data: for a large number of households, the actual consumption may be known as well as the values of the explanatory variables (square meters etc.)
- Next, the constructed model can be used in prediction.
  - For a new customer, it is straightforward to ask the values of the explanatory variables.
  - The energy consumption can then be predicted using the model.
    - Getting an idea of the consumption by other means could be difficult, as the consumption has not yet happened.

Example

- Let's examine how the course points obtained from exercises (max. 40) predict the points obtained from an exam (max. 60).
- First, plot the observations as a scatterplot.
- It seems that the points are located near a straight line.
- This straight line is called a regression line.
  - The regression line in the example is included in the image, as is its equation.



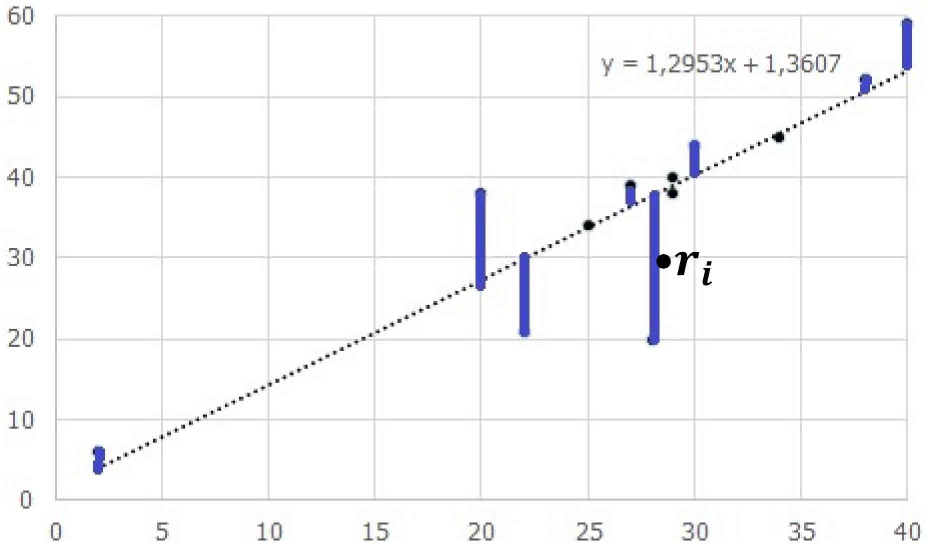
- This is an example of simple linear regression (one explanatory variable).

Equation of a regression line

- The general equation of regression line is $y = ax + b$.
  - Here, y is the response variable (exam points).
  - Likewise, x is the explanatory variable (exercise points).
  - Constants a and b are called regression coefficients.
- The equation of the straight line predicts the value of the response variable.
  - Example: a student scores 15 points in the exercises. The exam points is predicted to be $1{,}2953 \times 15 + 1{,}3607 \ \approx 21$.
- The challenge is to find the values of the regression coefficients a and b in such a way that the straight line matches the observations in the training set as well as possible.
- To achieve this, the least-sum-of-squares method is applied.

Least-sum-of-squares method

- If the regression line was known, it would be possible to compute the distance of each response variable value from the value predicted by the regression line.
  - These are vertical distances r_i.
  - The goodness of fit of the entire data set to the regression line can be measured by the sum of their squares: $\sum_i r_i^2$.
- The remaining problem is to find a straight line that minimizes the sum of squares.
  - It can be done analytically by means of matrix calculus.
  - Machine learning and statistical software provide means for finding the equation.



Many explanatory variables

- In the example before, there was just one explanatory variable
- The method generalizes to many explanatory variables (MLR, *multiple linear regression*):

$$y_i = \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \varepsilon_i$$

- In the example:
  - y_i is the value of the response variable in observation i.
  - $x_{i1}, \ldots, x_{ip}$ are the values of variables $x_1, \ldots, x_p$ in observation i.
  - $\beta_1, \ldots, \beta_p$ are the regression coefficents to be found out.
  - ε_i is an offset constant that specifies where the regression line cuts the y axis.
- Technically, if we have 1 response variable and p explanatory variables, instead of a regression line we have a p-dimensional plane in a $p + 1$ dimensional space.
  - Due to the high number of dimensions, it is no longer possible to produce a single visualization of the observations and the regression plane.

The assumptions of a linear model

- The general assumptions for making a linear model include that:
  1. The relationship between the variables in indeed linear,
  2. The explanatory variables are not correlated.
  3. The variance of error terms is constant throughout the values of explanatory variables.

- The violation of assumption 2 is called multicollinearity.
  – It makes the interpretation of the model more difficult, even though the model may still be usable for prediction.
  – Revealed by a correlation matrix.

On applicability

- In traditional statistical analysis a linear model is tailored by stringently analysing each variable.
- In machine learning the starting point is often the inclusion of all potential variables.
  - Unnecessary variables that contribute little to the outcome can then be pruned.
- The interpretation of the constructed model(s) requires caution.
  - Consider the exercise/exam points example: how much can we really say anything about students who have less than 20 exercise points? Can we safely extrapolate?

Estimation error

- Measures for estimation error (aka. prediction error):
  - MSE
  - R^2
- The estimation error is usually higher for the scoring set than for the training set.
  - Danger of model overfitting.
  - Consider validation.
    - Use training set for model building.
    - Use testing set for model evaluation.

Option 1: MSE

- $MSE = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n}$

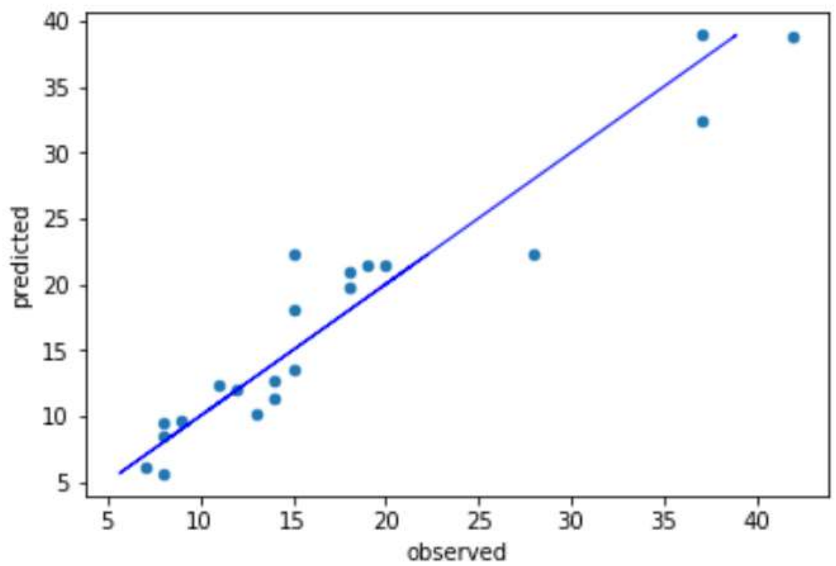
- Mean squared error.
- Measures the average of squared differences between the observed (y_i) and estimated ($\hat{y}_i$) values.
  - Lower values are better.
  - 0 indicates that the response variable values can be predicted from the explanatory variables without any error.
  - No fixed upper limit.

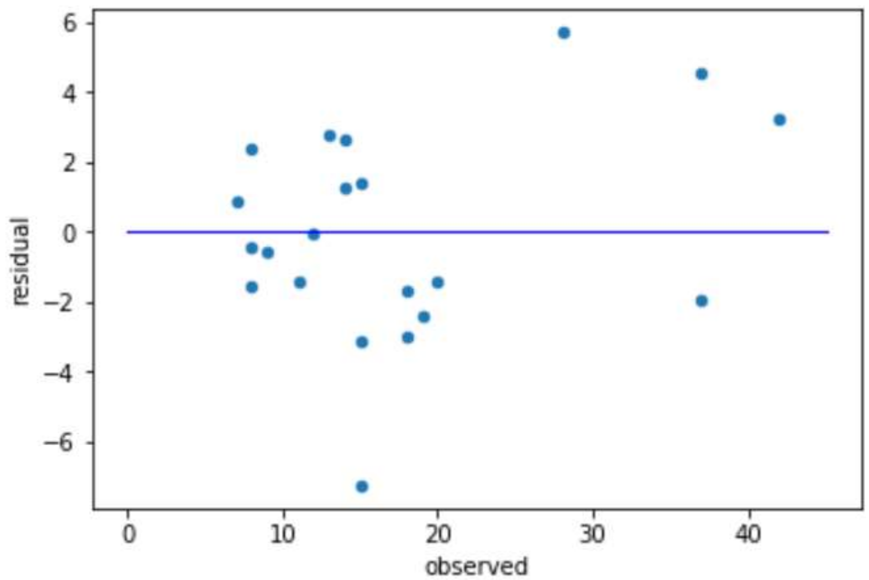
Option 2: R^2

- $\boldsymbol{R^2 = 1 - \frac{SS_{res}}{SS_{tot}} = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$

- Coefficient of multiple determination.
- In the equation:
  - $\boldsymbol{SS_{res}}$ is the sum of squares of the residuals (differences between the observed values $\boldsymbol{y_i}$ and the estimated values $\boldsymbol{\hat{y}_i}$).
  - $\boldsymbol{SS_{tot}}$ is the total sum of squares (sum of squared distances from the mean $\boldsymbol{\bar{y}}$).
- Describes the proportion of variance in response variable that is explained by the explanatory variables.
  - Higher values are better.
  - The upper limit of 1 is reached when the response variable fully depends on the explanatory variables.
  - 0 indicates the response variable's full independence of explanatory variables.

Residual plots





- Observed vs predicted values
- Observed values vs residuals

Residuals

- For an observation, the distance between the observed and predicted response variable value value is called a residual.
- The applicability of a linear model to a data set can be examined by looking at the residuals.
- Ideally:
  1. The residuals should be independent from each other.
  2. They should be normally distributed.
  3. The variance of the residuals should stay constant as the response variable values change.
- To check these, produce a scatterplot where the observed values are on the horizontal axis and the residuals are on the vertical axis.
  – The scatterplot should be symmetrical to the horizontal axis.
  – The vertical axis values should not change as the values on the horizontal axis change.

Variable importance

- Some variables tend to be more important in relation to the model than others.
- Note that for MLR to produce a correct model, standardization is not necessary.
- For non-standardized data, the importance can not be directly inferred from the regression coefficients, as the variables' standard deviation varies.
- Solution: standardize the data first to have:
  - A mean of zero
  - A standard deviation of one
- After standardization, a regression coefficient directly tells how many standard deviations it is away from zero.
  - The higher the absolute value, the more valuable the explanatory variable is to the model.

Example

```
> stackloss
   Air.Flow Water.Temp Acid.Conc. stack.loss
1        80         27         89         42
2        80         27         88         37
3        75         25         90         37
4        62         24         87         28
5        62         22         87         18
6        62         23         87         18
7        62         24         93         19
8        62         24         93         20
9        58         23         87         15
10       58         18         80         14
11       58         18         89         14
12       58         17         88         13
13       58         18         82         11
14       58         19         93         12
15       50         18         89          8
16       50         18         86          7
17       50         19         72          8
18       50         19         79          8
19       50         20         80          9
20       56         20         82         15
21       70         20         91         15
```

- **stackloss** is a small (n=21) demonstration data set.
- The data is for a chemical factory.
- Variable **stack.loss** is the amount of lost product due to conditions.
- The goal is to estimate it based on the other variables.

•Data source: Brownlee, K. A. (1960, 2nd ed. 1965) *Statistical Theory and Methodology in Science and Engineering*. New York: Wiley. pp. 491–500.

Example

- The Jupyter Notebook file for the stackloss example can be found at **Documents/Methods/Data/Stack loss (demo)**.