Linear and logistic regression

Predicting numeric values and probabilities

Learning goals

1. Learn the idea, limitations and applicability of linear regression in predicting values of numerical variables.
2. Learn to carry out linear regression analysis in Python.
3. Learn to apply a logistic regression model to predict an outcome of an event.
4. Learn to carry out the logistic regression analysis in Python environment.

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

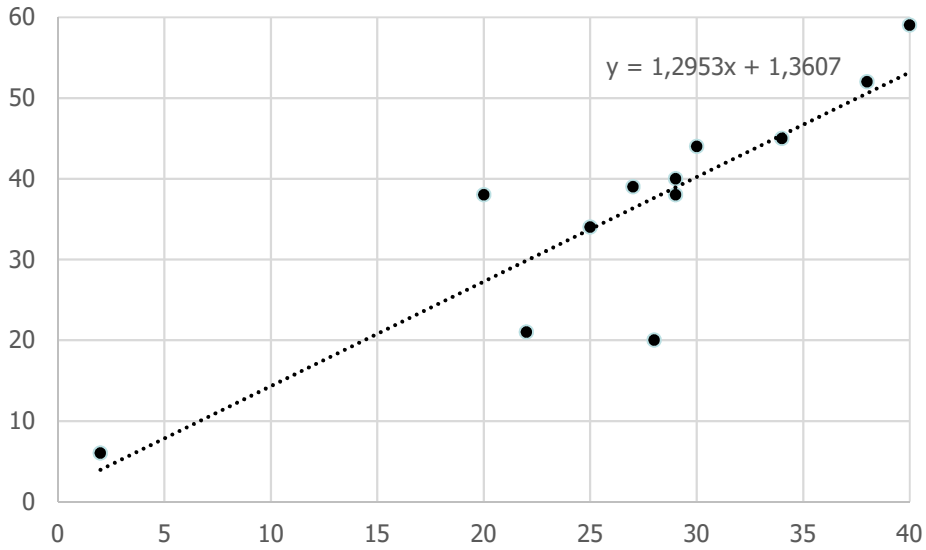
- For linear regession, categorical variables should be re-encoded as multiple dummy variables.
- Example: eye colour with four values: blue, brown, grey, green.
- For n categories, there are two options:
  - n dummy variables (one per category, option 1 above)
  - $n-1$ dummy variables (one per gategory,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



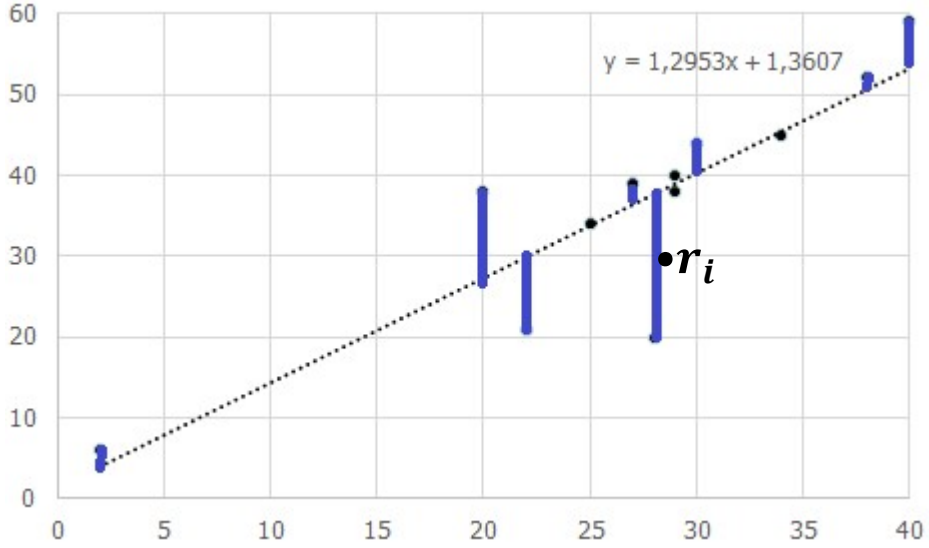
- This is an example of simple linear regression (one explanatory variable).

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*).
- For observation i:

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

- 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.

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

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

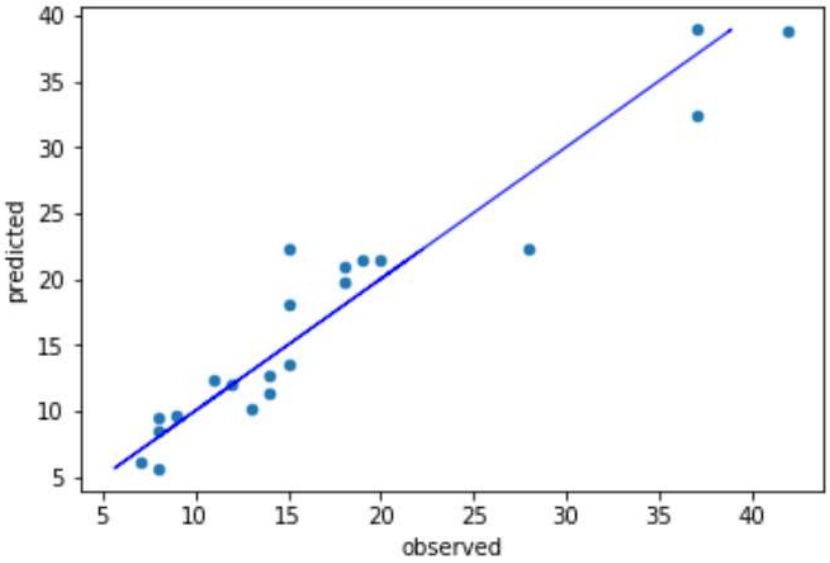
- Mean squared error.
- Measures the average of squared differences between the observed (y_i) and estimated ($\hat{y}_i$) values.
  - Lower values are better.
  - 0 indicates that the response variable values can be predicted from the explanatory variables without any error.
  - No fixed upper limit.

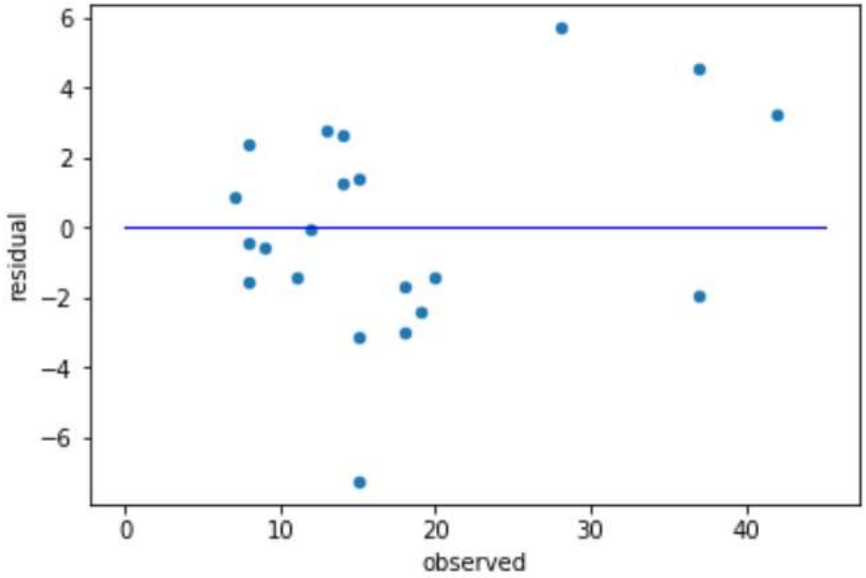
Option 2: R^2

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

- Coefficient of multiple determination.
- In the equation:
  - SS_{res} is the sum of squares of the residuals (differences between the observed values y_i and the estimated values $\hat{y}_i$).
  - SS_{tot} is the total sum of squares (sum of squared distances from the mean $\bar{y}$).
- 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 (see **stack loss** demo)

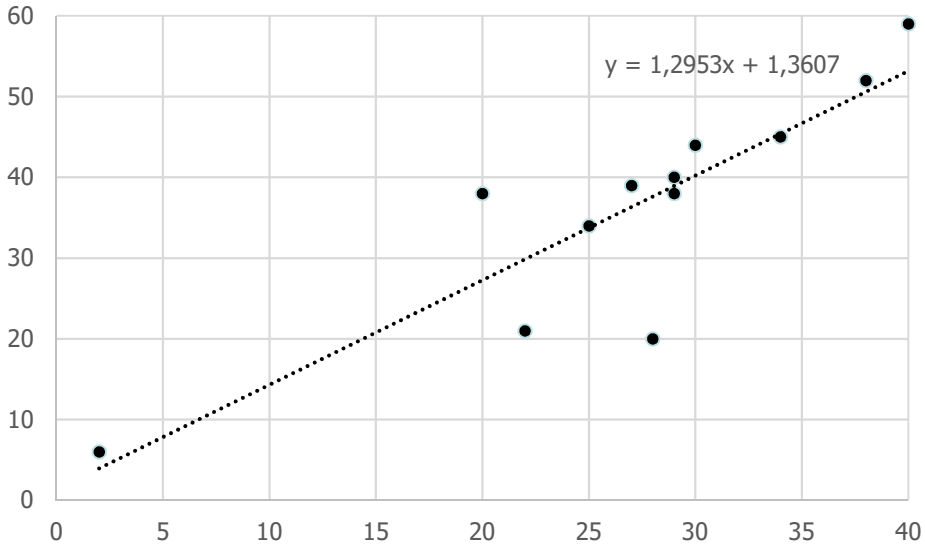
```
> stackloss
   Air.Flow Water.Temp Acid.Conc. stack.loss
1        80         27         89         42
2        80         27         88         37
3        75         25         90         37
4        62         24         87         28
5        62         22         87         18
6        62         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.

From linear to logistic regression

- Earlier we focused on linear regression.
  - In linear regression, the response variable is a continuous variable that can ideally vary in the interval $]-\infty, \infty[$.
  - In practice, the applicability of the model is limited.



From linear to logistic regression

- In logistic regression, the response variable is not a numerical variable but a binary class variable (yes/no).
- The goal of logistic regression analysis is to predict whether an event occurs or no.
- Technically the target of prediction is the probability p of an event.
  - Example: predict the probability for a subject experiencing a stroke, or, whether he/she will buy a car.

Logistic regression as a classifier

- As a consequence of estimating the probability of an event, logistic regression model can be used as a binary classifier.
  - Rule: If the probability of an event is estimated to be greater than 0.5, classify as ”yes”; otherwise ”no”.
  - Such binary classification based on the outcomes loses information on the uncertainty.

Probability as a response variable

- Recall that the equation in a generalized linear model is of form:

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

- The probability p of an event would be a bad choice for a response variable y, since its value is always in the interval $[0,1]$.
- A linear model is designed to provide predictions within an unlimited range.
- Key question: how could the response variable be transformed in such a way that the range becomes $[0,1]$?

Odds

- Odds describes the ratio between an event and its complement.
  - Denote the probability of an event by p.
  - Odds is then defined as $\frac{p}{1-p}$.
- Let's assume that the probability of a persion winning a running contest is 0,15.
- The odds are $\frac{0,15}{1-0,15} = \frac{0,15}{0,85} \approx 0,176$.
  - The probability of a win is 0,176 times as big as that of a loss.
  - Or, expressed in reverse terms, the probability of a loss is appxroximately
  - 5,67-fold in comparison to that of a win.
- The range of odds is $[0, \infty[$.
  - It is still constrained from the lower edge.
  - In addition, the values of interest are often "packed" close to the zero.

Logit

- If we take a natural logarithm of the odds, we have made a logit transformation for the probability p.
- Thus, logit transformation is expressed as:

  $\ln \frac{p}{1-p}$

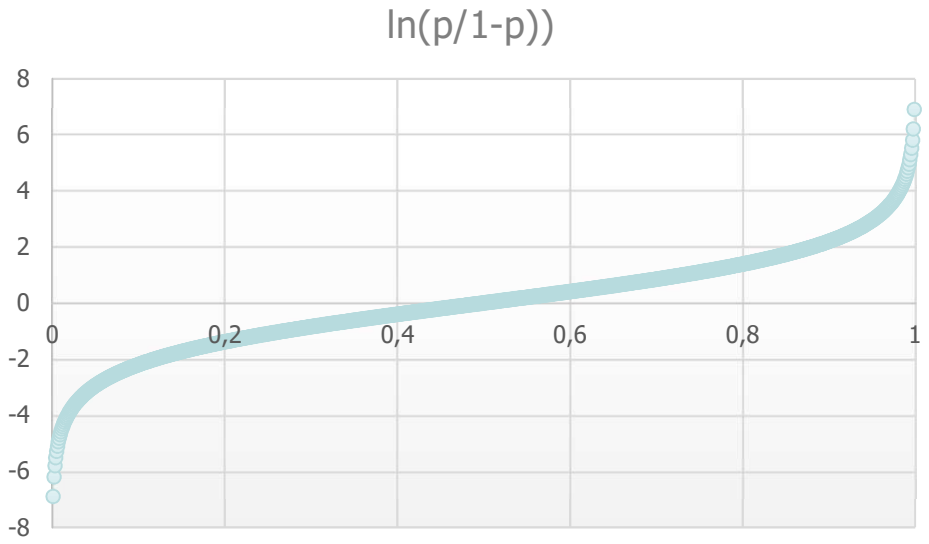
- E.g. the probability $p = 0{,}15$ of a win corresponds to the logit value:
- $\ln \frac{0{,}15}{1-0{,}15} \approx -1{,}735$

Logit

p	1-p	ln(p/1-p))
0,001	0,999	-6,90675
0,002	0,998	-6,21261
0,003	0,997	-5,80614
0,004	0,996	-5,51745
0,005	0,995	-5,2933
0,006	0,994	-5,10998
0,007	0,993	-4,95482
0,008	0,992	-4,82028
0,009	0,991	-4,70149
0,01	0,99	-4,59512
0,011	0,989	-4,4988
0,012	0,988	-4,41078
0,013	0,987	-4,32972

0,498	0,502	-0,008
0,499	0,501	-0,004
0,5	0,5	0
0,501	0,499	0,004
0,502	0,498	0,008

0,998	0,002	6,212606
0,999	0,001	6,906755



- Note that the value of the logit function is >0, when p>0,5.

Logistic regression model

- A logistic regression model is of form:

$$\ln \frac{p_i}{1-p_i} = \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \varepsilon_i$$

- The transformed response variable can vary between $]-\infty, \infty[$.
  - The corresponding probabilities p are in range $[0,1]$.
- If the predicted value of a logit response variable is positive:
  - The prediction for $p > 0{,}5$
  - The event is predicted to happen (prediction ”1”)
- Likewise, a negative predicted value of the logit response variable corresponds to predicted non-occurrence of an event (predicted ”0”).

Limitations

- Many of the limitations of linear regression can be relaxed for logistic regression. For example:
  - The relationships needs not be linear.
  - The residuals don't need to be normally distributed.
  - The variances of the variables need not be constant.
- However:
  - The response variable needs to be binary.
  - The observations should be independent of each other.
  - Theres should not be much collinearity (dependence between variables).
  - The log odds of the response variable should be linearly related to the explanatory variables.

Limitations

- There is no simple way to calculate the relative importances of the variables.
  - Regression coefficients from normalized data don't provide the answer.
  - This is due to the inherent non-linearities in the model

`sklearn` specifics

- Logistic regression can also be used as a multiclass classifier.
- By default **sklearn** uses one-vs-rest (OvR) scheme.
  - A binary classifier is built for each value of the categorical response variable.
  - For each binary classifier, all remaining values of the response variable are lumped together.
  - Finally, for each observation, the classifier that provides a classification with a highest confidence score, outputs the final class.

One-vs-Rest classification

	Classifier-specific response variables			
Orig. resp. variable	C1_class	C2_class	C3_class	C4_class
1	1	0	0	0
2	0	1	0	0
3	0	0	1	0
4	0	0	0	1
2	0	1	0	0
2	0	1	0	0
1	1	0	0	0
4	0	0	0	1
3	0	0	1	0

- For logistic regression, **sklearn** implements the OvR schema automatically, under the hood.

Example (see **stroke** demo)

	ID	Weight	Smoking	Exercise	Cholesterol	Income	Exphappiness	Birthyear	Sex	Stroke
0	1	117	1	2	8.0	1080	27	1913	M	1
1	2	62	0	8	5.5	2120	55	1949	M	0
2	3	74	0	6	4.8	3170	65	1976	M	0
3	4	77	0	5	4.2	4740	61	1973	F	0
4	5	67	0	8	4.5	1900	53	1929	M	0
5	6	76	0	6	6.2	3410	72	1959	F	0
6	7	63	0	7	4.1	3640	71	1979	F	0
7	8	75	0	5	5.2	2500	99	1960	F	0
8	9	70	0	6	4.9	2110	48	1922	F	0
9	10	82	0	5	5.8	2560	34	2007	F	1

- In **stroke** data set, the data consists of people that have experienced a stroke.
- The goal is to predict whether a second stroke will occur.