Linear regression

Part 5: Multiple linear regression

By: Noureddin Sadawi, PhD

University of London

Multiple linear regression

- When there is more than one independent variable, the equation is extended to accommodate them:

$$Y = A + B_1 X_1 + B_2 X_2 + ... + B_n X_n + \epsilon$$

- We now have a linear model rather than a line.
- An extension of simple linear regression that allows us to study the effect of multiple independent variables (X_1, X_2, …) on a single dependent variable Y.
- The relationship between each coefficient and its independent variable is linear.

Multiple linear regression

- When there is more than one independent variable, the equation is extended to accommodate them:

$$Y = A + B_1 X_1 + B_2 X_2 + \ldots + B_n X_n + \epsilon$$

- B_1, B_2, … are called partial regression coefficients.
- A is the mean value of Y when all X's = 0.
- B_i is the average amount by which Y changes for a unit change in X_i (while keeping all other X's constant).

Significance tests and CIs

- It is possible to compute a multiple correlation coefficient, R, and perform a statistical significance test of the fit of the model.
- H_0 says: *Y* is totally unrelated to X's.
- C.I. values can be computed for *A*, *B*'s and R.
- Here's how to use the C.I. of any *B*:
  - If the C.I. contains 0 (i.e. *B* is not significantly different from 0) then the corresponding *X* can be deleted (which means the equation becomes simpler).
  - Bear in mind that when removing one input variable, coefficients of other input variables will change.
- Use existing tools.

Dummy variables

- We have learnt the one-hot encoding technique to transform categorical variables into numeric representation.
- This means we can use multiple linear regression even if one or more independent variables are categorical.
- Please refer to Part 1 in Topic 2.

city	population
Tripoli	2
London	8
Tripoli	3
London	10
Sydney	3

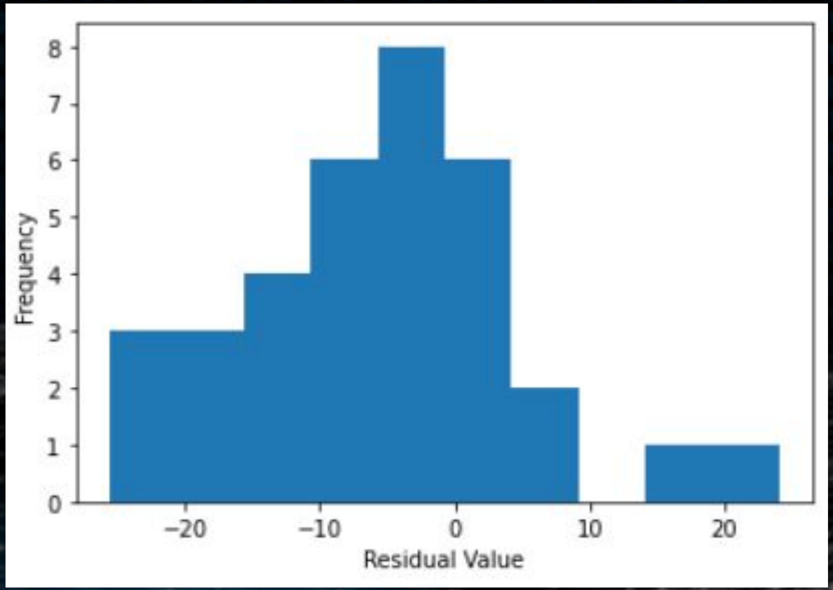
population	London	Sydney
2	0	0
8	1	0
3	0	0
10	1	0
3	0	1

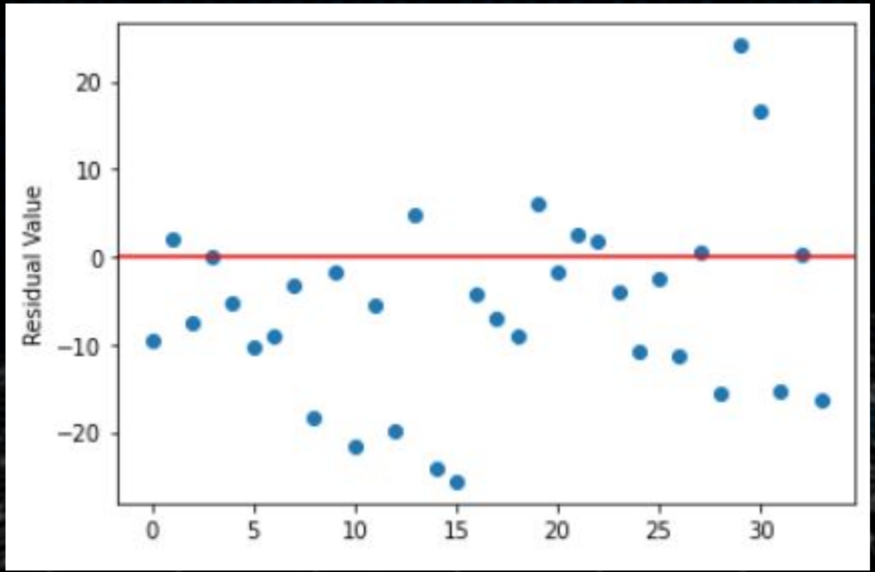
Dummy encoding

Model validity

- In simple linear regression, it is easy to examine whether the relationship between the independent variable and dependent variable is linear by using a scatter plot of the two of them (points should generally be evenly scattered on both sides of the straight line).
- In multiple linear regression, it is not easy to visualise the relationship.
- Therefore, we visualise the **residuals** of the fitted model.
- We plot a histogram of the residuals and expect them to follow a normal distribution with a mean of zero.
- If that is not the case, then the model may not be an adequate description of the data.

Residual plots





residuals
-9.444894
2.107716
-7.430005
0.079699
-5.210715
-10.241177
-9.016965
-3.264300
-18.243700
-1.625193
-21.678809
-5.531273
-19.834242
4.670380
-24.063590
-25.639270
-4.152754
-7.047979
-9.131655
6.107624
-1.626371
2.632234
1.806113
-4.021244
-10.741431

Correlated variables

- If some of the independent variables are highly correlated, then it may be difficult to interpret the model.
  Hence, before we perform multiple linear regression, we should examine that correlation.
- If there are highly correlated independent variables, then there might be reason(s) to select only one of them to be used in the model.
- Highly correlated independent variables cause **multicollinearity**.
- Multicollinearity makes model interpretation difficult because it is not possible to change one variable and keep the others constant if they are highly correlated.

Model Selection and Stepwise Regression

- ‘In some problems, many variables could be used as predictors in a regression.
- For example, to predict house value, additional variables such as the basement size or year built could be used.
- Adding more variables, however, does not necessarily mean we have a better model.
- Statisticians use the principle of Occam’s razor to guide the choice of a model: all things being equal, a simpler model should be used in preference to a more complicated model.’

(Bruce and Bruce *Practical statistics for data scientists*, second edition, 2020).

Model Selection and Stepwise Regression

- ‘Including additional variables always reduces *RMSE* and increases R^2 for the training data.
- Hence, these are not appropriate to help guide the model choice.
- One approach to including model complexity is to use the adjusted R^2:

$$R^2_{adj} = 1 - (1 - R^2)\frac{n-1}{n-P-1}$$

Here, n is the number of records and P is the number of variables in the model.’

(Bruce and Bruce *Practical statistics for data scientists*, second edition, 2020).

Model Selection and Stepwise Regression

- ‘In the 1970s, Hirotugu Akaike, the eminent Japanese statistician, developed a metric called AIC (Akaike’s Information Criteria) that penalizes adding terms to a model.
- In the case of regression, AIC has the form:

$$\mathrm{AIC} = 2P + n \log(\mathrm{RSS}/n)$$

where *P* is the number of variables and *n* is the number of records. **The goal is to find the model that minimizes AIC**; models with *k* more extra variables are penalized by *2k*.’

- RSS is the same as SSE (error, or residual, sum of squares).

(Bruce and Bruce *Practical statistics for data scientists*, second edition, 2020).

Variants to AIC

'There are several variants to AIC:

- **AICc:** A version of AIC corrected for small sample sizes.
- **BIC or Bayesian information criteria:** Similar to AIC, with a stronger penalty for including additional variables to the model.
- **Mallows Cp:** A variant of AIC developed by Colin Mallows.

These are typically reported as in-sample metrics (i.e., on the training data), and data scientists using holdout data for model assessment do not need to worry about the differences among them ...'

(Bruce and Bruce *Practical statistics for data scientists*, second edition, 2020).

Model Selection and Stepwise Regression

- ‘How do we find the model that minimizes AIC or maximizes adjusted R2? One way is to search through all possible models, an approach called all subset regression.
- This is computationally expensive and is not feasible for problems with large data and many variables.
- An attractive alternative is to use stepwise regression.
- It could start with a full model and successively drop variables that don’t contribute meaningfully. This is called backward elimination.
- Alternatively one could start with a constant model and successively add variables (forward selection).’

(Bruce and Bruce *Practical statistics for data scientists*, second edition, 2020).