Logistic regression

Part 4: Multivariate logistic regression

By: Noureddin Sadawi, PhD

University of London

Multivariate logistic regression

- Usually more than one input variable could be included in the model.
- Purpose: Determine which variables result in the best model within the scientific context of the problem.
- Example diabetes data:
  - number of pregnancies (Preg)
  - plasma glucose concentration (Gluc)
  - diastolic blood pressure (BP)
  - triceps skin fold thickness (ST)
  - BMI
  - diabetes pedigree function (DPF)
  - age in years.

Multivariate logistic regression

Preg	Gluc	BP	ST	BMI	DPF	Age	Outcome
1	89.0	66	23	28.1	0.167	21.0	0
0	137.0	40	35	43.1	2.288	33.0	1
3	78.0	50	32	31.0	0.248	26.0	1
2	197.0	70	45	30.5	0.158	53.0	1
1	189.0	60	23	30.1	0.398	59.0	1
…	…	…	…	…	…	…	…
0	181.0	88	44	43.3	0.222	26.0	1
1	128.0	88	39	36.5	1.057	37.0	1
2	88.0	58	26	28.4	0.766	22.0	0
10	101.0	76	48	32.9	0.171	63.0	0
5	121.0	72	23	26.2	0.245	30.0	0

logit(probability of positive diabetes test) = -8.72 + 0.12*Preg + 0.03*gluc - 0.01 * BP + 0.01*ST + 0.09*BMI + 1.15*DPF + 0.03*Age

Multivariate logistic regression

logit(probability of positive diabetes test) = -8.72 + 0.12*Preg + 0.03*gluc - 0.01 * BP + 0.01*ST + 0.09*BMI + 1.15*DPF + 0.03*Age

- **Coefficients: effect on logit(p) for a unit change of one single predictor causes while keeping all the rest constant**

For example:

- A unit change in BMI **increases** the logit of the probability of diabetes by 0.09 while keeping the other variables constant.
- A unit change in blood pressure **decreases** the logit of the probability of diabetes by 0.01 while keeping the other variables constant.

Multivariate logistic regression

logit(probability of positive diabetes test) = -8.72 + 0.12*Preg + 0.03*gluc - 0.01 * BP + 0.01*ST + 0.09*BMI + 1.15*DPF + 0.03*Age

- **exp(coefficient): same as before BUT in relation to the actual odds of having a positive diabetes test rather than the logit of this.**

For example:

- The odds of having a positive diabetes test go up by a factor of 1.094 by a unit change in BMI.
- The odds of the same outcome go down by a factor of 1.01 by a unit change in DBP.

Values of coefficient

- Linear regression uses least squares to estimate coefficients for the best fit line that relates input variables to the outcome.
- Logistic regression uses **maximum likelihood estimation** (MLE) to obtain the model coefficients.
- This function is initially estimated, then the process is repeated until LL (log likelihood) does not change significantly.

Input variables and p-values

- We can obtain a p-value for each input variable in a logistic regression model.
- The p-value for each variable tests the null hypothesis that the variable has no effect on the outcome.
- A low p-value (< 0.05) indicates that we can reject the null hypothesis.
- In other words, a variable that has a low p-value is likely to be a meaningful addition to the model:
  - because changes in the variable's value are related to changes in the outcome.
- A larger (insignificant) p-value suggests that changes in the variable are not associated with changes in the outcome.
  - Hence we should consider removing that variable.

	Coef.	Std.Err.	z	P>|z|	[0.025	0.975]
gmat	-0.0262	0.0110	-2.3830	0.0172	-0.0477	-0.0046
gpa	3.9422	1.9641	2.0071	0.0447	0.0925	7.7918
work_experience	1.1983	0.4818	2.4871	0.0129	0.2540	2.1426

Goodness of fit

To what extent the fitted values under the model compare to the actual (i.e. observed) values.

- If the agreement between the observations and corresponding fitted values is good, the model may be acceptable.
- If not, the model is said to display ‘lack-of-fit’ and it needs to be revised.
- There are multiple diagnostic methods to measure the goodness of fit.

Variable importance

Example methods for measuring variable importance in logistic regression:

- If the input variables have the same scale, then coefficients can be used as a crude variable importance score.
- If the variables do not have the same scale, then a simple approach is to calculate variable importance as the magnitude of coefficient times the standard deviation of the corresponding variable in the data.
- The z score is also often used to determine variable importance.
  - It is the regression coefficient divided by the standard error.
- Wald chi-square value can be used to tank variables.

Model selection

Aim: find the simplest model that yields the best performance.

- Determine the smallest subset of input variables that produces the most accurate model.
- Multiple models can be created, the model with the lowest **Akaike information criterion** (**AIC**) is usually selected.
- Model building strategies:
  - forward selection
  - backward selection
  - stepwise selection.