Logistic Regression

Samantha Estrada PhD

ORSSP Research Design & Data Analysis Lab Consultant
University of Texas at Tyler

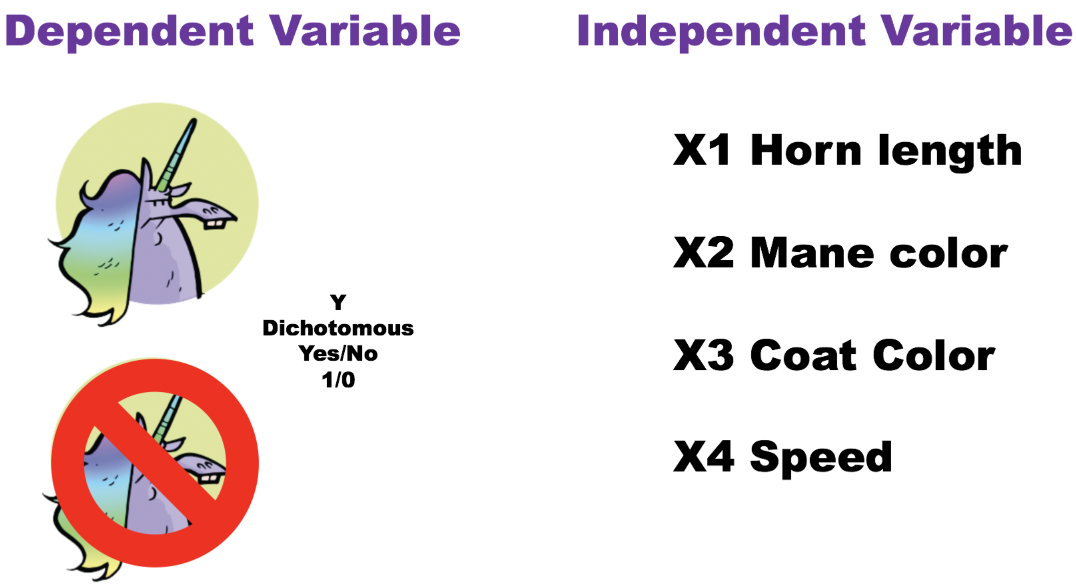
10/21/2022

Contents

In simple and multiple linear regression outcome and predictor variable(s) were continuous data. In logistic regression our outcome and dependent variable will be a binary predictor.

Potential applications of logistic regression

- Retention studies
  - i.e., want to examine factors which predict whether college students will or will not stay in school
- Marriage/family studies



- e.g., might look at variables which predict which couples will or will not divorce or factors which predict

- Medical research
  - Factors distinguishing between those who will and will not survive (e.g., surgery, a particular illness, etc.)
- Since logistic regression is nonparametric, you have more flexibility with variables because there are no normality assumptions.

- The outcome variable is categorical. The predictor variables can be a mix of categorical or continuous variables
- Logistic regression is all about predicting the *odds* that a given outcome will occur.
  - Odds are different than probabilities.
  - Probabilities range from 0-1
  - Odds can range from negative infinity to positive infinity.
  - Positive odds means a thing is more likely to occur, and negative odds mean a thing is less likely to occur

Brief Probability Review

- Probabilities are simply the likelihood that something will happen; a probability of .20 of rain means that there is a 20% chance of rain.
- If there is a 20% chance of rain, then there is an 80% chance of no rain; the odds, then, are:

$Odds = \frac{prob(rain)}{prob(norain)} = \frac{20}{80} = \frac{1}{4} = .25$

- Remember that probability can range from 0 to 1. But the odds can be greater than 1.
  - For instance, a 50% chance of rain has odds of 1.

Odds Ratio

- Odds ratio (OR) is the effect size for logistic regression
- Odds ratios greater than 1 = increase of the odds of that outcome
- Odds ratios less than 1 = decrease in the odds of that outcome.
- The comparison group is the group coded as 0.
  - So if your odds ratio is greater than 1, you have an increase in the odds of being in the 1 group.
  - Less than 1 decrease in odds of the 1 group (or increase in the 0 group).

Example: Logistic Regression

Data

The data for this example is available in `logistic.omv` The dataset for this example contains N = 275 observations and seven variables.

In the following example we would like to predict heart attacks in males from the following data:

- `Nominal` DV: Heart Attack where 0=no heart attack and 1=heart attack.
- `Continuous` IV: AGE in years
- `Continuous` IV: Systolic blood pressure (SYSBP)
- `Continuous` IV: Diastolic blood pressure (DIABP)
- `Continuous` IV: Cholesterol (CHOLES)
- `Continuous` IV: Height (HT) height in inches
- `Continuous` IV: Weight (WT) weight in pounds

Research Question

Do body weight, height, blood pressure and age have an influence on the probability of having a heart attack (yes vs. no)?

Descriptives

Let's explore the dataset by running descriptives. In the `Analysis` tab, click on `Exploration` next select `Descriptives` from the menu.

Descriptives

Descriptives

	Heart Attack	Age	Weight	Systolic Blood Pressure	Diastolic Blood Pressure	Cholesterol	Height
N	275	275	275	275	275	275	275
Missing	0	0	0	0	0	0	0
Mean		45.0	168	124	83.0	297	68.5
Median		45.0	166	120	80.0	285	68.0
Minimum		23.0	108	90.0	55.0	135	62.0
Maximum		70.0	262	190	112	520	74.0

Frequencies

Frequencies of Heart Attack

Levels	Counts	% of Total	Cumulative %
No Heart Attack	175	63.6%	63.6%
Heart Attack	100	36.4%	100.0%

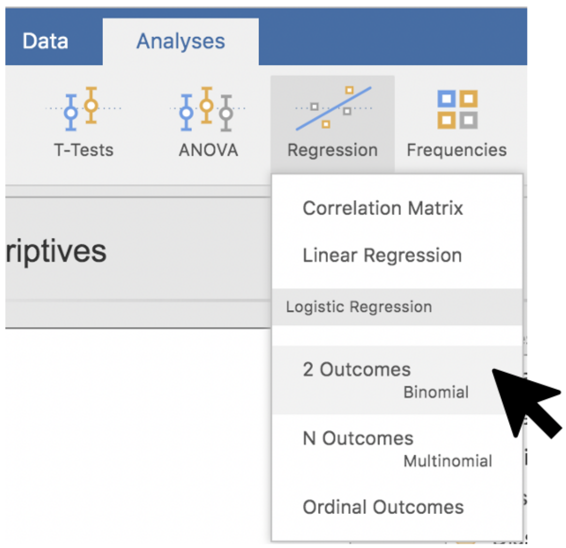
63.6% of the patients have not had a heart attack, and 36.4% of the patients have had one.

Sample Size Requirements

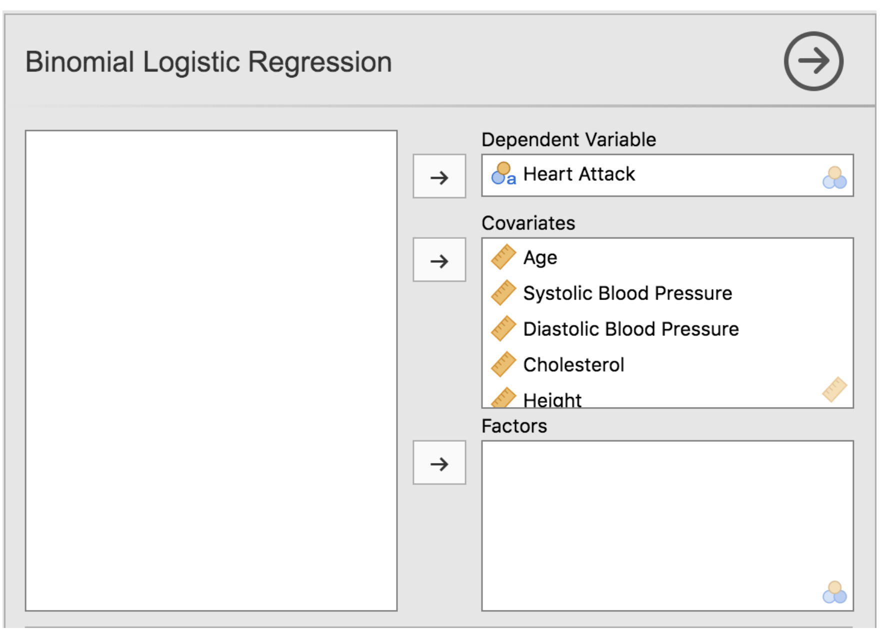
- In terms of the adequacy of sample sizes, the literature has not offered specific rules applicable to logistic regression (Peng et al., 2002).
- Several authors on multivariate statistics (Tabachnick & Fidell, 2019) have recommended:
  - A minimum ratio of 10 (observations) to 1 (variable), with a minimum sample size of 100 or 50

`jamovi` Logistic Regression

Under the `Analysis` tab select the `Regression` option, then select the `2 Outcomes Binomial` option.

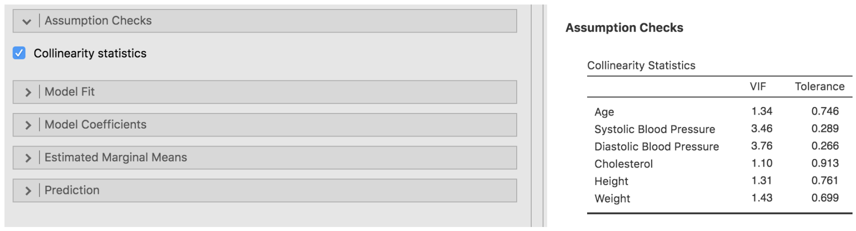


Place `Heart Attack` in the Dependent Variable option and the rest of the IVs in the covariates box.



Collinearity

- We don't want to have variables that explain the same thing in our regression, or that are too highly correlated.
- Logistic regression does not have to meet the assumptions of normality or heterogeneity of variance, but we do have to check for multicollinearty.

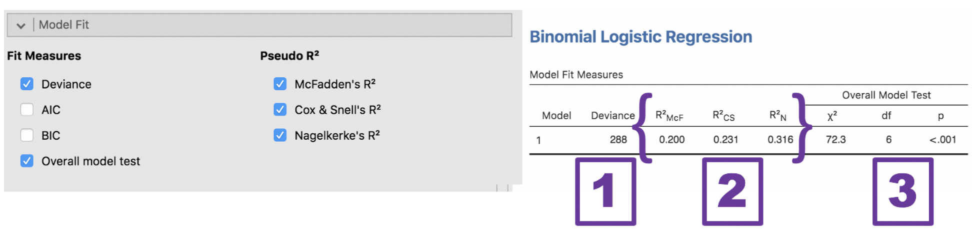


Everything looks good according to our rules of thumb VIF < 10 and Tolerance > .01

Output: Logistic Regression

Model Fit & Effect Size

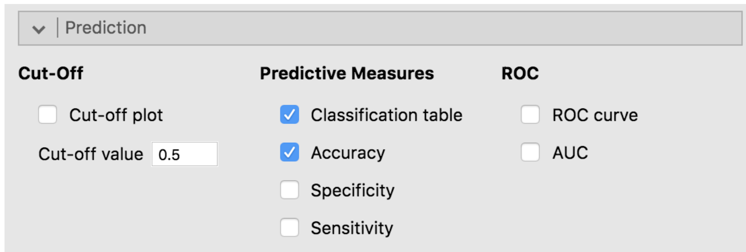
Under the `Model Fit` submenu select `Deviance`,`Overall model test`, and all the pseudo R^2



1. *Deviance*: This stat shows the predictive success of the model. The smaller the number, the better the model (in SPSS this is called 2 Log Likelihood in case you ever need to know).
2. Cox & Snell R^2 and Nagelkerke R^2*: These two numbers in the model summary box are similar to R^2 in multiple regression (a proportion of the variance in the DV accounted for by the variables in model). We will report both of them as "% of variance accounted for".
   - **Effect size notes**: Cox and Snell R^2 based on likelihoods and sample size BUT never can reach 1, even if you achieve perfect fit.
   - Use Nagelkerke R^2 which adjusts Cox and Snell so that the upper limit is 1 (most people report this type of effect size.)
3. *Overall Model Test* tests how well the model performs to predict the outcome. We want the omnibus chi-square for the model to be *significant*, and we will report it in our write-up.
   - Ho: the model with no independent variables fits the data as well as your model.
   - Ha: Your model fits the data better than the intercept-only model.

Under the `Prediction` submenu select `Classification Table`

- *Classification Table*: We will report the Accuracy as a percentage correct in our write-up. so .702 x 100% = 70.2%



Prediction

Classification Table – Heart Attack

	Predicted		
Observed	No Heart Attack	Heart Attack	% Correct
No Heart Attack	142	33	81.1
Heart Attack	49	51	51.0

Note. The cut-off value is set to 0.5

Predictive Measures

Accuracy
0.702

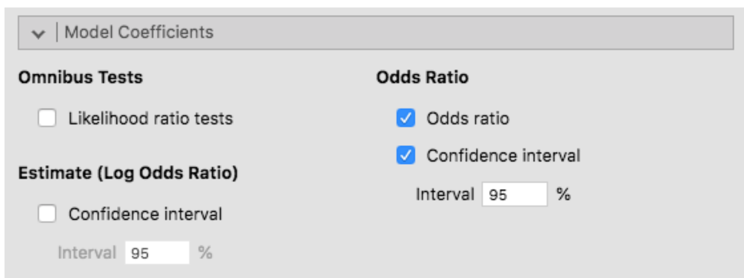
Note. The cut-off value is set to 0.5

- Accuracy: (142+51) / (142+33+49+51) = 70.2%

- *Sensitivity* is the percentage of cases that had the observed outcome was correctly predicted by the model (i.e., true positives).
  - Sensitivity: 142/(142+33)*100 = 81%
- *Specificity* is the percentage of observations that were also correctly predicted as not having the observed outcome (i.e., true negatives).
  - Specificity: 51 /(49+51)*100 = 51%

The classification table is shown. In this example, our prediction accuracy of those that didn't have a heart attack is high (81.1%) while our prediction accuracy of those that did have a heart attack is low (51.0%). Our overall accuracy rate is 70.2%.

Model Coefficients



β

Model Coefficients - Heart Attack

Predictor	Estimate	SE	Z	p	Odds ratio	95% Confidence Interval Lower	Upper
Intercept	−5.32860	5.07619	−1.050	0.294	0.00485	2.32e−7	101.55
Age	0.07229	0.01649	4.384	<.001	1.07496	1.041	1.11
Weight	0.02084	0.00677	3.079	0.002	1.02106	1.008	1.03
Height	−0.05316	0.07080	−0.751	0.453	0.94822	0.825	1.09
Cholesterol	0.00768	0.00239	3.212	0.001	1.00771	1.003	1.01
Diastolic Blood Pressure	−0.02911	0.02640	−1.103	0.270	0.97131	0.922	1.02
Systolic Blood Pressure	0.01285	0.01485	0.865	0.387	1.01293	0.984	1.04

Note. Estimates represent the log odds of "Heart Attack = Heart Attack" vs. "Heart Attack = No Heart Attack"

This table is similar to the coefficients table in multiple regression. We will report these results in two ways: make a regression table reporting the coefficients, standard error, significance; and we will highlight descriptively in the text of our results which predictors significantly contributed to predicting the DV. You can also select to report the 95% confidence interval (under the `Model Coefficients` menu).

Interpreting Odds Ratio

What if. . . ?

- **Scenario 1** Imagine `height` was significant and the odds ratio (OR) was .94. Then we would interpret the odds ratio like this:

The odds ratio indicates that for every unit increase in `height` *the odds of the outcome decrease by a factor of .94.*

Odds Ratio for Categorical Variables

- **Scenario 2** Imagine that `Weight` is a categorical variable coded as in Weight = 0 means "not overweight" and Weight = 1 is "overweight." Then we would interpret the odds ratio like this:

The odds that a person will experience the outcome are 1.02 times higher for those who are overweight than for those who are not.

Resources

- Research Design & Data Analysis Lab: https://www.uttyler.edu/research/ors-research-design-data-analysis-lab/
- Schedule a consultant appointment with me: https://www.uttyler.edu/research/ors-research-design-data-analysis-lab/ors-research-design-data-analysis-lab-consultants/
- Check out Lab Resources (including recording of this webinar): https://www.uttyler.edu/research/ors-research-design-data-analysis-lab/resources/

References

Peng, C.-Y. J., Lee, K. L., & Ingersoll, G. M. (2002). An introduction to logistic regression analysis and reporting. *The Journal of Educational Research*, *96*(1), 3–14.

Tabachnick, B. G., & Fidell, L. S. (2019). *Using multivariate statistics*. Pearson.