5.7: Multiple Logistic Regression

Learning Objectives

- To use multiple logistic regression when you have one nominal variable and two or more measurement variables, and you want to know how the measurement variables affect the nominal variable. You can use it to predict probabilities of the dependent nominal variable, or if you're careful, you can use it for suggestions about which independent variables have a major effect on the dependent variable.

When to use it

Use multiple logistic regression when you have one nominal and two or more measurement variables. The nominal variable is the dependent (Y) variable; you are studying the effect that the independent (X) variables have on the probability of obtaining a particular value of the dependent variable. For example, you might want to know the effect that blood pressure, age, and weight have on the probability that a person will have a heart attack in the next year.

Heart attack vs. no heart attack is a binomial nominal variable; it only has two values. You can perform multinomial multiple logistic regression, where the nominal variable has more than two values, but I'm going to limit myself to binary multiple logistic regression, which is far more common.

The measurement variables are the independent (X) variables; you think they may have an effect on the dependent variable. While the examples I'll use here only have measurement variables as the independent variables, it is possible to use nominal variables as independent variables in a multiple logistic regression; see the explanation on the multiple linear regression page.

Epidemiologists use multiple logistic regression a lot, because they are concerned with dependent variables such as alive vs. dead or diseased vs. healthy, and they are studying people and can't do well-controlled experiments, so they have a lot of independent variables. If you are an epidemiologist, you're going to have to learn a lot more about multiple logistic regression than I can teach you here. If you're not an epidemiologist, you might occasionally need to understand the results of someone else's multiple logistic regression, and hopefully this handbook can help you with that. If you need to do multiple logistic regression for your own research, you should learn more than is on this page.

The goal of a multiple logistic regression is to find an equation that best predicts the probability of a value of the Y variable as a function of the X variables. You can then measure the independent variables on a new individual and estimate the probability of it having a particular value of the dependent variable. You can also use multiple logistic regression to understand the functional relationship between the independent variables and the dependent variable, to try to understand what might cause the probability of the dependent variable to change. However, you need to be very careful. Please read the multiple regression page for an introduction to the issues involved and the potential problems with trying to infer causes; almost all of the caveats there apply to multiple logistic regression, as well.

As an example of multiple logistic regression, in the 1800s, many people tried to bring their favorite bird species to New Zealand, release them, and hope that they become established in nature. (We now realize that this is very bad for the native species, so if you were thinking about trying this, please don't.) Veltman et al. (1996) wanted to know what determined the success or failure of these introduced species. They determined the presence or absence of 79 species of birds in New Zealand that had been artificially introduced (the dependent variable) and 14 independent variables, including number of releases, number of individuals released, migration (scored as 1 for sedentary, 2 for mixed, 3 for migratory), body length, etc. Multiple logistic regression suggested that number of releases, number of individuals released, and migration had the biggest influence on the probability of a species being successfully introduced to New Zealand, and the logistic regression equation could be used to predict the probability of success of a new introduction. While hopefully no one will deliberately introduce more exotic bird species to new territories, this logistic regression could help understand what will determine the success of accidental introductions or the introduction of endangered species to areas of their native range where they had been eliminated.

Null hypothesis

The main null hypothesis of a multiple logistic regression is that there is no relationship between the X variables and the Y variable; in other words, the Y values you predict from your multiple logistic regression equation are no closer to the actual Y values than you would expect by chance. As you are doing a multiple logistic regression, you'll also test a null hypothesis for each X variable, that adding that X variable to the multiple logistic regression does not improve the fit of the equation any more than

expected by chance. While you will get P values for these null hypotheses, you should use them as a guide to building a multiple logistic regression equation; you should not use the P values as a test of biological null hypotheses about whether a particular X variable causes variation in Y.

How it works

Multiple logistic regression finds the equation that best predicts the value of the Y variable for the values of the X variables. The Y variable is the probability of obtaining a particular value of the nominal variable. For the bird example, the values of the nominal variable are "species present" and "species absent." The Y variable used in logistic regression would then be the probability of an introduced species being present in New Zealand. This probability could take values from 0 to 1. The limited range of this probability would present problems if used directly in a regression, so the odds, $Y/(1-Y)$, is used instead. (If the probability of a successful introduction is 0.25, the odds of having that species are $0.25/(1-0.25)=1/3$. In gambling terms, this would be expressed as "3 to 1 odds against having that species in New Zealand.") Taking the natural log of the odds makes the variable more suitable for a regression, so the result of a multiple logistic regression is an equation that looks like this:

$$\ln\left[\frac{Y}{1-Y}\right]=a+b_1X_1+b_2X_2+b_3X_3+\ldots \qquad (5.7.1)$$

You find the slopes (b_1, b_2, etc.) and intercept (a) of the best-fitting equation in a multiple logistic regression using the maximum-likelihood method, rather than the least-squares method used for multiple linear regression. Maximum likelihood is a computer-intensive technique; the basic idea is that it finds the values of the parameters under which you would be most likely to get the observed results.

You might want to have a measure of how well the equation fits the data, similar to the R^2 of multiple linear regression. However, statisticians do not agree on the best measure of fit for multiple logistic regression. Some use deviance, D, for which smaller numbers represent better fit, and some use one of several pseudo-R^2 values, for which larger numbers represent better fit.

Using nominal variables in a multiple logistic regression

You can use nominal variables as independent variables in multiple logistic regression; for example, Veltman et al. (1996) included upland use (frequent vs. infrequent) as one of their independent variables in their study of birds introduced to New Zealand. See the discussion on the multiple linear regression page about how to do this.

Selecting variables in multiple logistic regression

Whether the purpose of a multiple logistic regression is prediction or understanding functional relationships, you'll usually want to decide which variables are important and which are unimportant. In the bird example, if your purpose was prediction it would be useful to know that your prediction would be almost as good if you measured only three variables and didn't have to measure more difficult variables such as range and weight. If your purpose was understanding possible causes, knowing that certain variables did not explain much of the variation in introduction success could suggest that they are probably not important causes of the variation in success.

The procedures for choosing variables are basically the same as for multiple linear regression: you can use an objective method (forward selection, backward elimination, or stepwise), or you can use a careful examination of the data and understanding of the biology to subjectively choose the best variables. The main difference is that instead of using the change of R^2 to measure the difference in fit between an equation with or without a particular variable, you use the change in likelihood. Otherwise, everything about choosing variables for multiple linear regression applies to multiple logistic regression as well, including the warnings about how easy it is to get misleading results.

Assumptions

Multiple logistic regression assumes that the observations are independent. For example, if you were studying the presence or absence of an infectious disease and had subjects who were in close contact, the observations might not be independent; if one person had the disease, people near them (who might be similar in occupation, socioeconomic status, age, etc.) would be likely to have the disease. Careful sampling design can take care of this.

Multiple logistic regression also assumes that the natural log of the odds ratio and the measurement variables have a linear relationship. It can be hard to see whether this assumption is violated, but if you have biological or statistical reasons to expect a

non-linear relationship between one of the measurement variables and the log of the odds ratio, you may want to try data transformations.

Multiple logistic regression does not assume that the measurement variables are normally distributed.

Example

> Some obese people get gastric bypass surgery to lose weight, and some of them die as a result of the surgery. Benotti et al. (2014) wanted to know whether they could predict who was at a higher risk of dying from one particular kind of surgery, Roux-en-Y gastric bypass surgery. They obtained records on $81,751$ patients who had had Roux-en-Y surgery, of which 123 died within 30 days. They did multiple logistic regression, with alive vs. dead after 30 days as the dependent variable, and 6 demographic variables (gender, age, race, body mass index, insurance type, and employment status) and 30 health variables (blood pressure, diabetes, tobacco use, etc.) as the independent variables. Manually choosing the variables to add to their logistic model, they identified six that contribute to risk of dying from Roux-en-Y surgery: body mass index, age, gender, pulmonary hypertension, congestive heart failure, and liver disease.
>
> Benotti et al. (2014) did not provide their multiple logistic equation, perhaps because they thought it would be too confusing for surgeons to understand. Instead, they developed a simplified version (one point for every decade over 40, 1 point for every 10 BMI units over 40, 1 point for male, 1 point for congestive heart failure, 1 point for liver disease, and 2 points for pulmonary hypertension). Using this RYGB Risk Score they could predict that a 43-year-old woman with a BMI of 46 and no heart, lung or liver problems would have an 0.03% chance of dying within 30 days, while a 62-year-old man with a BMI of 52 and pulmonary hypertension would have a 1.4% chance.

Graphing the results

Graphs aren't very useful for showing the results of multiple logistic regression; instead, people usually just show a table of the independent variables, with their P values and perhaps the regression coefficients.

Similar tests

If the dependent variable is a measurement variable, you should do multiple linear regression.

There are numerous other techniques you can use when you have one nominal and three or more measurement variables, but I don't know enough about them to list them, much less explain them.

How to do multiple logistic regression

Spreadsheet

I haven't written a spreadsheet to do multiple logistic regression.

Web page

There's a very nice web page for multiple logistic regression. It will not do automatic selection of variables; if you want to construct a logistic model with fewer independent variables, you'll have to pick the variables yourself.

R

Salvatore Mangiafico's *R Companion* has a sample R program for multiple logistic regression.

SAS

You use PROC LOGISTIC to do multiple logistic regression in SAS. Here is an example using the data on bird introductions to New Zealand.

```
DATA birds;
INPUT species $ status $ length mass range migr insect diet clutch
broods wood upland water release indiv;
DATALINES;
Cyg_olor 1 1520 9600 1.21 1 12 2 6 1 0 0 1 6 29
Cyg_atra 1 1250 5000 0.56 1 0 1 6 1 0 0 1 10 85
```

Cer_nova 1 870 3360 0.07 1 0 1 4 1 0 0 1 3 8
Ans_caer 0 720 2517 1.1 3 12 2 3.8 1 0 0 1 1 10
Ans_anse 0 820 3170 3.45 3 0 1 5.9 1 0 0 1 2 7
Bra_cana 1 770 4390 2.96 2 0 1 5.9 1 0 0 1 10 60
Bra_sand 0 50 1930 0.01 1 0 1 4 2 0 0 0 1 2
Alo_aegy 0 680 2040 2.71 1 . 2 8.5 1 0 0 1 1 8
Ana_plat 1 570 1020 9.01 2 6 2 12.6 1 0 0 1 17 1539
Ana_acut 0 580 910 7.9 3 6 2 8.3 1 0 0 1 3 102
Ana_pene 0 480 590 4.33 3 0 1 8.7 1 0 0 1 5 32
Aix_spon 0 470 539 1.04 3 12 2 13.5 2 1 0 1 5 10
Ayt_feri 0 450 940 2.17 3 12 2 9.5 1 0 0 1 3 9
Ayt_fuli 0 435 684 4.81 3 12 2 10.1 1 0 0 1 2 5
Ore_pict 0 275 230 0.31 1 3 1 9.5 1 1 1 0 9 398
Lop_cali 1 256 162 0.24 1 3 1 14.2 2 0 0 0 15 1420
Col_virg 1 230 170 0.77 1 3 1 13.7 1 0 0 0 17 1156
Ale_grae 1 330 501 2.23 1 3 1 15.5 1 0 1 0 15 362
Ale_rufa 0 330 439 0.22 1 3 2 11.2 2 0 0 0 2 20
Per_perd 0 300 386 2.4 1 3 1 14.6 1 0 1 0 24 676
Cot_pect 0 182 95 0.33 3 . 2 7.5 1 0 0 0 3 .
Cot_aust 1 180 95 0.69 2 12 2 11 1 0 0 1 11 601
Lop_nyct 0 800 1150 0.28 1 12 2 5 1 1 1 0 4 6
Pha_colc 1 710 850 1.25 1 12 2 11.8 1 1 0 0 27 244
Syr_reev 0 750 949 0.2 1 12 2 9.5 1 1 1 0 2 9
Tet_tetr 0 470 900 4.17 1 3 1 7.9 1 1 1 0 2 13
Lag_lago 0 390 517 7.29 1 0 1 7.5 1 1 1 0 2 4
Ped_phas 0 440 815 1.83 1 3 1 12.3 1 1 0 0 1 22
Tym_cupi 0 435 770 0.26 1 4 1 12 1 0 0 0 3 57
Van_vane 0 300 226 3.93 2 12 3 3.8 1 0 0 0 8 124
Plu_squa 0 285 318 1.67 3 12 3 4 1 0 0 1 2 3
Pte_alch 0 350 225 1.21 2 0 1 2.5 2 0 0 0 1 8
Pha_chal 0 320 350 0.6 1 12 2 2 2 1 0 0 8 42
Ocy_loph 0 330 205 0.76 1 0 1 2 7 1 0 1 4 23
Leu_mela 0 372 . 0.07 1 12 2 2 1 1 0 0 6 34
Ath_noct 1 220 176 4.84 1 12 3 3.6 1 1 0 0 7 221
Tyt_alba 0 340 298 8.9 2 0 3 5.7 2 1 0 0 1 7
Dac_nova 1 460 382 0.34 1 12 3 2 1 1 0 0 7 21
Lul_arbo 0 150 32.1 1.78 2 4 2 3.9 2 1 0 0 1 5
Ala_arve 1 185 38.9 5.19 2 12 2 3.7 3 0 0 0 11 391
Pru_modu 1 145 20.5 1.95 2 12 2 3.4 2 1 0 0 14 245
Eri_rebe 0 140 15.8 2.31 2 12 2 5 2 1 0 0 11 123
Lus_mega 0 161 19.4 1.88 3 12 2 4.7 2 1 0 0 4 7
Tur_meru 1 255 82.6 3.3 2 12 2 3.8 3 1 0 0 16 596
Tur_phil 1 230 67.3 4.84 2 12 2 4.7 2 1 0 0 12 343
Syl_comm 0 140 12.8 3.39 3 12 2 4.6 2 1 0 0 1 2
Syl_atri 0 142 17.5 2.43 2 5 2 4.6 1 1 0 0 1 5
Man_mela 0 180 . 0.04 1 12 3 1.9 5 1 0 0 1 2
Man_mela 0 265 59 0.25 1 12 2 2.6 . 1 0 0 1 80
Gra_cyan 0 275 128 0.83 1 12 3 3 2 1 0 1 1 .
Gym_tibi 1 400 380 0.82 1 12 3 4 1 1 0 0 15 448
Cor_mone 0 335 203 3.4 2 12 2 4.5 1 1 0 0 2 3
Cor_frug 1 400 425 3.73 1 12 2 3.6 1 1 0 0 10 182
Stu_vulg 1 222 79.8 3.33 2 6 2 4.8 2 1 0 0 14 653

```
Acr_tris 1 230 111.3 0.56 1 12 2 3.7 1 1 0 0 5 88
Pas_dome 1 149 28.8 6.5 1 6 2 3.9 3 1 0 0 12 416
Pas_mont 0 133 22 6.8 1 6 2 4.7 3 1 0 0 3 14
Aeg_temp 0 120 . 0.17 1 6 2 4.7 3 1 0 0 3 14
Emb_gutt 0 120 19 0.15 1 4 1 5 3 0 0 0 4 112
Poe_gutt 0 100 12.4 0.75 1 4 1 4.7 3 0 0 0 1 12
Lon_punc 0 110 13.5 1.06 1 0 1 5 3 0 0 0 1 8
Lon_cast 0 100 . 0.13 1 4 1 5 . 0 0 1 4 45
Pad_oryz 0 160 . 0.09 1 0 1 5 . 0 0 0 2 6
Fri_coel 1 160 23.5 2.61 2 12 2 4.9 2 1 0 0 17 449
Fri_mont 0 146 21.4 3.09 3 10 2 6 . 1 0 0 7 121
Car_chlo 1 147 29 2.09 2 7 2 4.8 2 1 0 0 6 65
Car_spin 0 117 12 2.09 3 3 1 4 2 1 0 0 3 54
Car_card 1 120 15.5 2.85 2 4 1 4.4 3 1 0 0 14 626
Aca_flam 1 115 11.5 5.54 2 6 1 5 2 1 0 0 10 607
Aca_flavi 0 133 17 1.67 2 0 1 5 3 0 1 0 3 61
Aca_cann 0 136 18.5 2.52 2 6 1 4.7 2 1 0 0 12 209
Pyr_pyrr 0 142 23.5 3.57 1 4 1 4 3 1 0 0 2 .
Emb_citr 1 160 28.2 4.11 2 8 2 3.3 3 1 0 0 14 656
Emb_hort 0 163 21.6 2.75 3 12 2 5 1 0 0 0 1 6
Emb_cirl 1 160 23.6 0.62 1 12 2 3.5 2 1 0 0 3 29
Emb_scho 0 150 20.7 5.42 1 12 2 5.1 2 0 0 1 2 9
Pir_rubr 0 170 31 0.55 3 12 2 4 . 1 0 0 1 2
Age_phoe 0 210 36.9 2 2 8 2 3.7 1 0 0 1 1 2
Stu_negl 0 225 106.5 1.2 2 12 2 4.8 2 0 0 0 1 2
;
PROC LOGISTIC DATA=birds DESCENDING;
MODEL status=length mass range migr insect diet clutch broods wood upland
water release indiv / SELECTION=STEPWISE SLENTRY=0.15 SLSTAY=0.15;
RUN;
```

In the MODEL statement, the dependent variable is to the left of the equals sign, and all the independent variables are to the right. SELECTION determines which variable selection method is used; choices include FORWARD, BACKWARD, STEPWISE, and several others. You can omit the SELECTION parameter if you want to see the logistic regression model that includes all the independent variables. SLENTRY is the significance level for entering a variable into the model, if you're using FORWARD or STEPWISE selection; in this example, a variable must have a P value less than 0.15 to be entered into the regression model. SLSTAY is the significance level for removing a variable in BACKWARD or STEPWISE selection; in this example, a variable with a P value greater than 0.15 will be removed from the model.

Summary of Stepwise Selection

Step	Effect Entered	Effect Removed	DF	Number In	Score Chi-Square	Wald Chi-Square	Pr > ChiSq
1	release		1	1	28.4339		<.0001
2	upland		1	2	5.6871		0.0171
3	migr		1	3	5.3284		0.0210

The summary shows that "release" was added to the model first, yielding a P value less than 0.0001. Next, "upland" was added, with a P value of 0.0171. Next, "migr" was added, with a P value of 0.0210. SLSTAY was set to 0.15, not 0.05, because you might want to include a variable in a predictive model even if it's not quite significant. However, none of the other variables have a P value less than 0.15, and removing any of the variables caused a decrease in fit big enough that P was less than 0.15, so the stepwise process is done.

Analysis of Maximum Likelihood Estimates

Standard Wald
Parameter DF Estimate Error Chi-Square Pr > ChiSq

Intercept 1 -0.4653 1.1226 0.1718 0.6785
migr 1 -1.6057 0.7982 4.0464 0.0443
upland 1 -6.2721 2.5739 5.9380 0.0148
release 1 0.4247 0.1040 16.6807 <.0001

The "parameter estimates" are the partial regression coefficients; they show that the model is:

$$\ln\left[\frac{Y}{1-Y}\right] = -0.4653 - 1.6057(migration) - 6.2721(upland) + 0.4247(release) \qquad (5.7.2)$$

Power analysis

You need to have several times as many observations as you have independent variables, otherwise you can get "overfitting"—it could look like every independent variable is important, even if they're not. A frequently seen rule of thumb is that you should have at least 10 to 20 times as many observations as you have independent variables. I don't know how to do a more detailed power analysis for multiple logistic regression.

References

Benotti, P., G.C. Wood, D.A. Winegar, A.T. Petrick, C.D. Still, G. Argyropoulos, and G.S. Gerhard. 2014. Risk factors associated with mortality after Roux-en-Y gastric bypass surgery. Annals of Surgery 259: 123-130.

Veltman, C.J., S. Nee, and M.J. Crawley. 1996. Correlates of introduction success in exotic New Zealand birds. American Naturalist 147: 542-557.