What population does the odds ratio from my logistic regression represent?

Frank Popham

28/04/2020

Working paper version 1, comments welcome (frank.popham at protonmail.com)

Introduction

Directly adjusting for confounders (C) using a regression model to estimate the effect of an exposure (X) on an outcome (Y) remains very popular in observational epidemiology (Hernán and Robins (2020)). There are alternative adjustment methods. For example, inverse probability weighting (IPW). IPW is a two stage process. Stage one, model X as a function of C. Stage two, model Y as a function of X weighted by the IPW from stage one where the IPW is the inverse of the probability of X. There are a number of notable advantages of IPW over an outcome regression, including controlling for confounding without the need to see outcome based results (this could limit positive result bias). Another advantage is that after stage one, the IPW can be used to assess confounder balance over exposure and to assess the average value that confounders are balanced at. The value confounders are balanced at is the population that your second stage estimate will represent. An outcome regression will balance confounders in the model over exposure but it is not clear the population it represents. Given effect modification by confounders, the magnitude and even direction of effect may matter. Say a drug has a positive treatment effect for women but a negative effect for men then the population effect is going to depend on the sex distribution.

Recent work shows how to derive the population for the effect of an outcome linear regression (Aronow and Samii 2016; Popham and Leyland 2018). Loosely, both methods show that an outcome linear regression is also two stage in that the population it represents can be derived without reference to the outcome. However it appears both methods are only approximate for other generalized linear regressions including logistic regression. In this working paper I empirically derive the population for a logistic regression but so far have not derived it algebraically. Ideas how to do so are welcome. Importantly when I refer to an outcome regression I exclude models with an interaction between X and C when modelling Y. Why? Well, normally researchers read their main effect directly from regressions output. Including X and C interaction would preclude this. So I recognize the model might be wrong. In fact checking balance can identify that the model may be sub optimal.

Table 1: Exposure and confounder relationship

C	X	N	N_C	X_C	R_X_C	C_R
0	0	58181	62583	0.0703386	-0.0703386	0.5385204
0	1	4402	62583	0.0703386	0.9296614	0.5385204
1	0	62610	66325	0.0560121	-0.0560121	0.4614796
1	1	3715	66325	0.0560121	0.9439879	0.4614796

Note:
N=count
N_C=count of C
X_C=probability of X given C
R_X_C=residual of X_C
C_R = C weighted by R_X_C

Methods and Results

The method of Aronow and Samii calculates the population of an outcome regression from the residual (R) of a regression of X as a function of C (Aronow and Samii (2016)). Table 1 displays for X and C, the count, the count of C and the probability of X given C and the resulting residual. The dataset is open access and included with the analysis code in the project directory of this working paper. All variables are binary. Using the residual as a weight, we find the average of C over X. So for both values of X, C==1 is balanced at 46%.

Table 2 displays the odds of Y by X and C as well as the value of the residual from Table 1. Table 3 displays the weighted odds by residual for X given C and the odds ratio (OR). This is very close but not quite the same as the OR from an outcome logistic regression (Y~X+C). This illustrates that the residual method is approximate for a logistic regression.

Table 2: Outcome odds by exposure and confounder

C	X	oddsY	C_R
0	0	0.6642639	0.5385204
0	1	1.3241816	0.5385204
1	0	0.9296678	0.4614796
1	1	1.9205975	0.4614796

Note:
oddsY= odds of the outcome
C_R = C weighted by residual of X given C

Table 3: Outcome odds and odds ratio by exposure weighted by residual

X	oddsY
0	0.7757302
1	1.5720681
OR	2.0265654
Model OR	2.0252372

Note:
oddsY = odds of the outcome
OR = Odds ratio from residual weighting
Model OR = Odds ratio from logistic regression (Y~X+C)

So what is the exact answer? Given the odds ratio from the outcome logistic regression and the confounder strata specific odds ratios for X on Y (Table 4), we can use the equation. log(OR Y ~ X | C==0) - log(OR Y~X+C) / log(OR Y ~ X | C==0) - log(OR Y ~ X | C==1)

As in Table 4 this suggests that the outcome logistic regression is balancing at 44% for C==1. Applying this new distribution of C to Table 2 gives the results in Table 5 and confirms this is the population the outcome logistic regression represents as the OR is the same as the model's OR.

Table 4: Strata specific odds ratios of the outcome and exact value of C for Y X+C

C	ORY_C	Cmean_ORY_C
0	1.993457	0.5568883
1	2.065897	0.4431117

Note:
ORY_C = Odds ratio Y ~ X for each strata of C
Cmean_ORY_C= Mean of C for outcome logistic regression

Table 5: Outcome odds and odds ratio by exposure weighted by Table 4 weight

X	oddsY
0	0.7709554
1	1.5613675
OR	2.0252372
Model OR	2.0252372

Note:
oddsY = odds of the outcome
OR = Odds ratio from Table 4 weighting
Model OR = Odds ratio from logistic regression (Y~X+C)

Table 6 reproduces Table 1 and adds the working residual. Th working residual is a normal residual divided by probability * (1-probability). So in Table 6 we calculate this as (X- X|C)/ (X|C * (1-X|C)). The working residual simplifies for X==1 to 1 / X | C and for X==0 to -1 / (1- X|C), in other words the (negative of) IPW. Concisely put the normal residual is additive while the working residual is multiplicative. If we work out the weighted odds of Y by X given C using the working residual we obtain the results in Table 7.

Table 6: Exposure and confounder combinations with working residual

C	X	N	N_C	X_C	R_X_C	WR_X_C	C_WR
0	0	58181	62583	0.0703386	-0.0703386	-1.075660	0.4854858
0	1	4402	62583	0.0703386	0.9296614	14.216947	0.4854858
1	0	62610	66325	0.0560121	-0.0560121	-1.059336	0.5145142
1	1	3715	66325	0.0560121	0.9439879	17.853297	0.5145142

Note:
N=count
N_C=count of C
X_C=probability of X given C
R_X_C=residual of X_C
WR_X_C=working residual of X_C
C_WR = C weighted by WR_X_C

We would expect to obtain the same result from a logistic regression model of Y ~ X with the working residual as a weight but we don't quite (2.01901448141945). This is because the IPW is a marginal model

Table 7: Outcome odds and odds ratio by exposure weighted by working residual

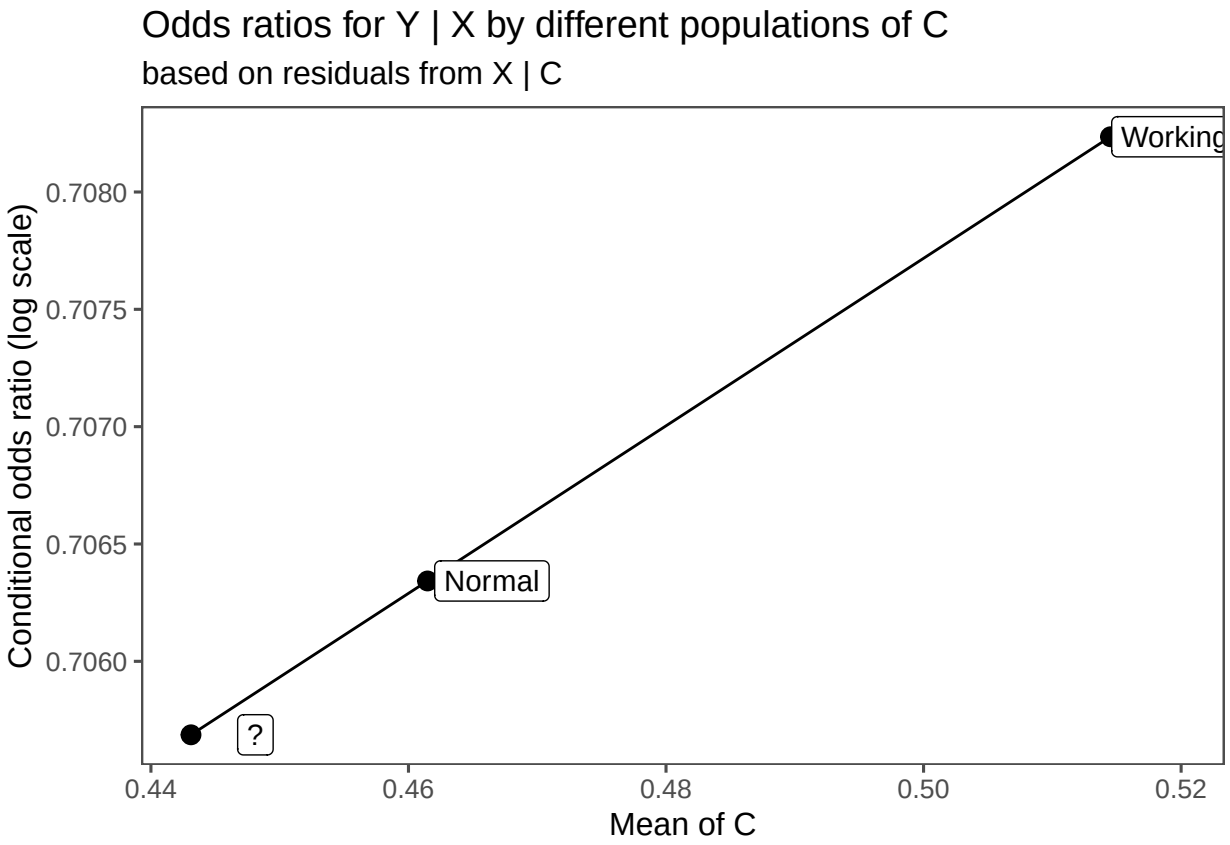
X	oddsY
0	0.7896835
1	1.6033777
OR	2.0304054

Note:
oddsY = odds of the outcome
OR = Odds ratio from working residual weighting
Model OR = Odds ratio from logistic regression (Y~X+C)

(unless you add a control variable) and in logistic regression results of models often differ whether they are conditional or marginal. Another way of thinking about the difference between marginal and conditional results is as follows. Say we model using logistic regression Y~ X * C, in other words we include an interaction between X and C. From the results we can predict the odds of Y or the probability of Y. To obtain an odds ratio for X we weight either the predicted odds or the predicted probability by the distribution of C in the population. The weighted average of the predicted odds is not equal to the weighted average of the predicted probability converted to odds after averaging. The conditional odds ratio comes from the weighted average of the predicted odds while the marginal odds ratio comes from the weighted average of predicted probabilities converted to odds.

Discussion

Figure 1 plots the conditional log odds ratio against the population (in terms of the confounder (C)) it represents. The working residual from the first stage model of an exposure as a function of confounding is the equivalent of the inverse probability weight. The working residual is a transformation of the normal residual (X - X | C). In a linear outcome regression setting the normal residual from the first stage model gives the population the effect the outcome regression model represents. This result is only approximate for a logistic regression outcome model. In figure 1 the true population (marked as ?) the outcome logistic regression represents is slightly different to that implied by the normal residual. At present I don't know what ? is and whether it is some form of transformation of the normal residual? I have experimented with transformation of the normal residual from a probability to odds scale but not had success. It may be that knowledge of the first stage model is not enough to derive this but it is sufficient for the working residual?



‘‘‘

References

Aronow, Peter M., and Cyrus Samii. 2016. “Does Regression Produce Representative Estimates of Causal Effects?” *American Journal of Political Science* 60 (1): 250–67. https://doi.org/10.1111/ajps.12185.

Hernán, MA, and J Robins. 2020. “Causal Inference Book.”

Popham, Frank, and Alastair H. Leyland. 2018. “Assessing Confounder Balance in Outcome Regressions.” *Epidemiology* 29 (5): e47–e48. https://doi.org/10.1097/ede.0000000000000871.