University of Limerick
Department of Sociology Working Paper Series

Working Paper WP2011-01
October 2011

Brendan Halpin

Department of Sociology

University of Limerick

Some notes on categorical data analysis, using simulation in Stata and R

Available at `http://www.ul.ie/sociology/pubs/wp2011-01.pdf`

Some notes on categorical data analysis, using simulation in Stata and R

Brendan Halpin
Dept of Sociology
University of Limerick*

Oct 2011

1 Introduction

Simple modelling of categorical data is not as simple as it seems. Standard formulas taught to generations of undergraduates are shown to be suboptimal, simple models are widely misunderstood, and high levels of controversy surround the suitability and interpretation of relatively standard models such as logistic regression. In this research note I discuss a number of these issues, using simple simulations in Stata and R to illuminate them.

2 Relative rates and odds ratios

A frequent theme in the medical statistics and epidemiological literature is that odds ratios (ORs) as effect measures for binary outcomes are counter intuitive and an impediment to understanding. Barros and Hirakata (2003), for instance, refer to the relative rate as the "measure of choice" and complain that the OR will "overestimate" the RR as the baseline probability rises. Clearly, ORs are less intuitive than relative rates (RRs), but in this note I take issue with the conclusion sometimes made, that models with relative-rate interpretations should be used instead of logistic regression and other OR models. This is because RRs are not measures of the size of the statistical association between a variable and an outcome (since they also vary inversely with the baseline probability), and because, under certain assumptions, ORs and related measures are. That is, RRs may feel more real but they are likely to be misleading.

While the argument is often cast in terms of rejecting logitistic in favour of log-binomial regression and other alternatives, let's look at some 2×2 tables

*threenotes.tex,v 1.3 2011/10/27 10:41:02 brendan Exp

and hand-calculated ORs and RRs. In the following two tabulations the OR is constant at 2.5, but the baseline probability (in class==0) is respectively 2% and 75%.[1]

```
. tab class outcome, matcell(n)

           |        outcome
     class |        No        Yes |     Total
-----------+----------------------+----------
  Class 0  |       980         20 |     1,000
  Class 1  |       951         49 |     1,000
-----------+----------------------+----------
     Total |     1,931         69 |     2,000

. scalar RR = (n[2,2]/(n[2,1]+n[2,2]))/(n[1,2]/(n[1,1]+n[1,2]))
. scalar OR = (n[2,2]/n[2,1])/(n[1,2]/n[1,1])
. scalar D  = n[2,2] - n[1,2]
. di  "RR " %6.3f RR "; OR " %6.3f OR "; N extra outcomes" %5.0f D
RR  2.450; OR  2.525; N extra outcomes   29
```

Here the RR is 2.45, the OR 2.53 and the number of extra outcomes in class 1 is 29, or 2.9%.

```
           |        outcome
     class |        No        Yes |     Total
-----------+----------------------+----------
  Class 0  |       270        730 |     1,000
  Class 1  |       129        871 |     1,000
-----------+----------------------+----------
     Total |       399      1,601 |     2,000
...
RR  1.193; OR  2.497; N extra outcomes  141
```

But when the baseline probability is high, the RR plummets (suggesting a 19% increase instead of 145%), despite the approximate substantive measure giving 141 or 14.1% extra cases. In this simple case, it seems RRs track substantive significance rather worse than ORs do.

But how do RRs and ORs compare in terms of estimating the size of the underlying statistical or causal association? There are many underlying causal structures possible, but let's use Stata to simulate a simple one.[2] Let the outcome of interest depend on an unobserved (and perhaps unobservable) interval variable. If this propensity is above a certain threshold, the outcome occurs, but let the threshold (and thus the proportion having the outcome) differ from time to time. Let the difference between the two

[1] Stata code at `http://teaching.sociology.ul.ie/catdat/ortab.do`.

[2] Stata code at `http://teaching.sociology.ul.ie/catdat/orsim.do`.

groups be that they have different distributions of the underlying propensity – normal, with the same variance but different means.[3] Conceptually, this inter-group difference is the source of effect we are trying to measure, while variation in the threshold is not related to the causal effect.

We run the simulation with a sample size of 10,000 and an inter-group difference of 0.2 standard deviations, and 2×2 tables are created for outcome probabilites. Here, for example, for 20% and 60% probabilities:

```
set obs 10000
gen class = _n <= 5000
gen propensity = invnorm(uniform()) + (class==1)*0.2
sort propensity
gen outcome20 = _n > (1 - 0.2)*_N
gen outcome60 = _n > (1 - 0.6)*_N
```

This yields the following:

class	outcome20 0	outcome20 1	Total
0	4,144	856	5,000
	82.88	17.12	100.00
1	3,856	1,144	5,000
	77.12	22.88	100.00
Total	8,000	2,000	10,000
	80.00	20.00	100.00

20% probability: RR: 1.34; OR: 1.44

class	outcome60 0	outcome60 1	Total
0	2,199	2,801	5,000
	43.98	56.02	100.00
1	1,801	3,199	5,000
	36.02	63.98	100.00
Total	4,000	6,000	10,000
	40.00	60.00	100.00

60% probability: RR: 1.14; OR: 1.39

Between 20% and 60% outcome probability, the OR drops but the RR drops rather more. Figure 1 show results for probabilities between 1% and 99%, replicated thirty times (lines for the average values, dots for the actual values). As can be seen, the ORs vary in a shallow U, but the RRs drop precipitously to a zero effect for high baseline probabilities.

Figure 2 repeats this exercise with logistic rather than normal propensity distributions, with the same variance (logistic distributions resemble normal but have higher kurtosis). Here the average OR is rather more stable. In fact, it can be shown mathematically that the OR is related directly to the difference in means, and is completely independent of the threshold.

Clearly, this causal backstory is simplistic.[4] The latent propensity may

[3] The attentive reader may recognise this as related to the latent variable justification of the logistic regression model, but for the moment please consider its plausibility as a simple causal model.

[4] It also suits only one-off outcomes – if the outcome is a result of exposure over time, the OR is as misleading as the RR, and an estimate of the hazard-rate ratio is needed.

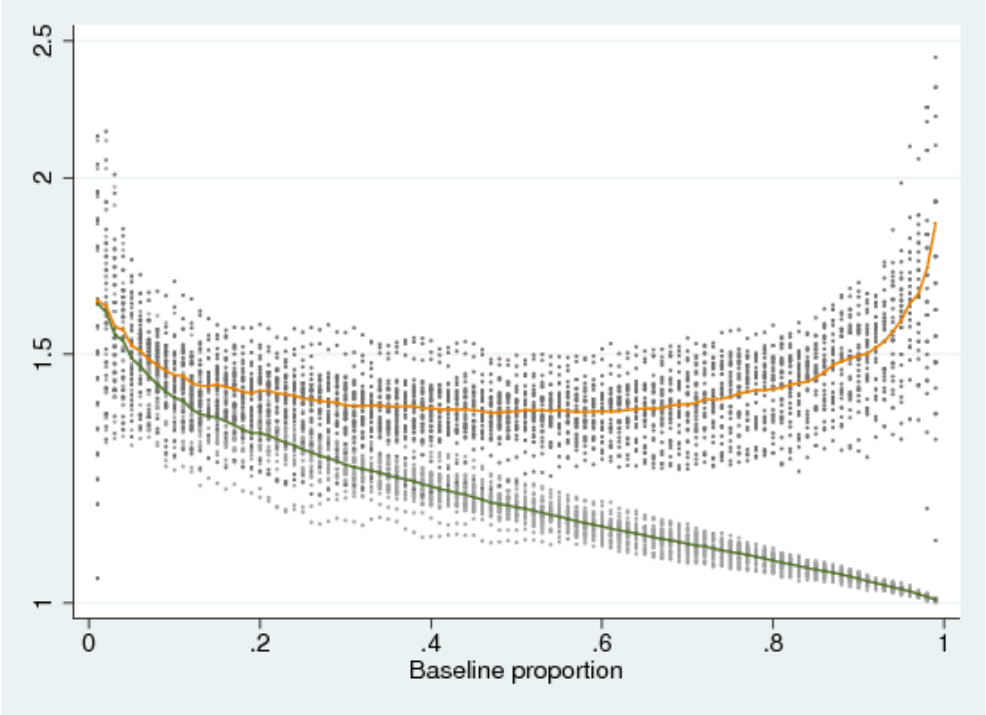

Figure 1: RRs and ORs: the points are the individual estimates and the lines the average across 30 replications, with a normal propensity distribution. The mean RR starts very close to the mean OR but drops to no effect (RR=1) in an almost linear fashion.

have other distributions, and the inter-class difference may be other than additive (though, note that log-normal distributions with a multiplicative difference are equivalent in effect to normal with an additive difference). If one class has greater variance, the causal effect will be non-linear (over-represented at both high and low propensity). However, in so far as it is approximately realistic, this story suggests that the odds ratio is a reasonably stable measure of an effect, while the relative rate is superficially intuitive but is not an effect measure.

While the OR and RR can be calculated by hand, the results from logistic (`logit outcome class`), poisson (`poisson outcome class`) and log-binomial regression (`glm outcome class, link(log) family(binomial)`) are exactly the same. The extension to probit regression is obvious. If the simulated distribution is normal, the mean probit estimate (not shown) is as flat as the OR is for the logistic distribution.[5]

When we are thinking in terms of models, rather than hand-calculated statistics, we can view the propensity distributions as conditional on the

[5] If you multiply the probit estimate by $\frac{\pi}{\sqrt{3}}$, it approximates the log of the odds ratio quite closely.

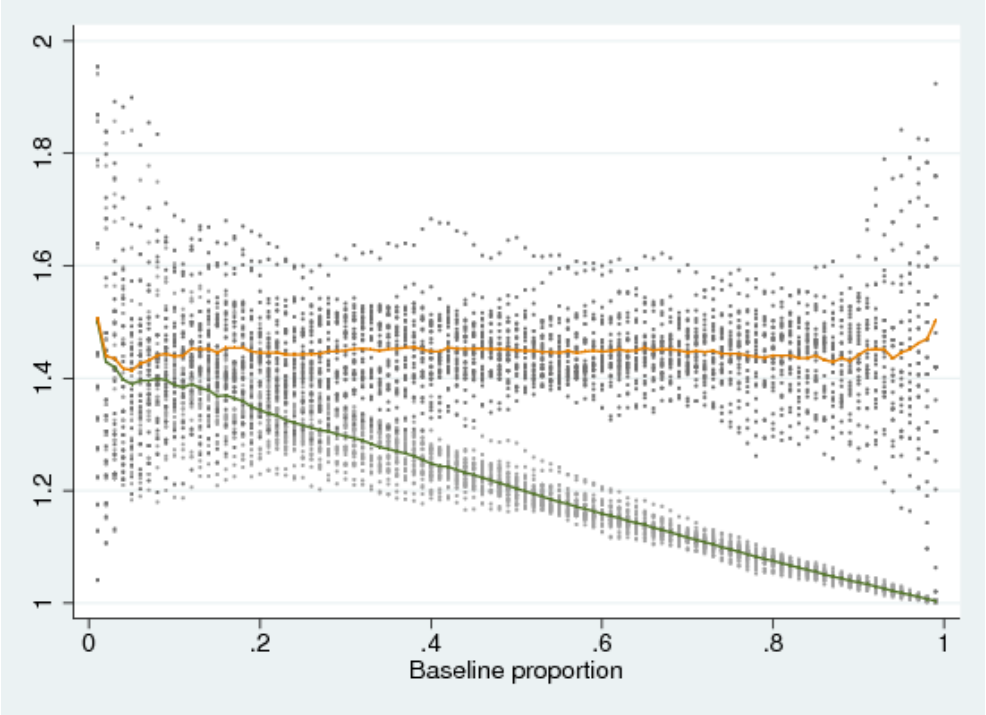

Figure 2: RRs and ORs with a logistic propensity distribution: the points are the individual estimates and the lines the average across 30 replications. Here the OR is much more stable.

other variables. The effect of these variables is analogous to shifting the threshold. In this case, the RR will be unreliable even if the average level of the outcome is stable, if there are other variables with large effects on the outcome. Thus, unless predicted probabilities in the data are all very low (say, under 10%) it seems unwise to base interpretations on RR models.

If it is important for the audience to see effects on probabilities, use `-margins-` to report marginal effects for different configurations of covariates. The fact that marginal effects vary with the values of the covariates is a feature, not a bug, reflecting the complexity of reality rather than being a wrong-headed consequence of an awkward model.

3 Relative rates, odds ratios and the complementary log-log model

In the previous section, I used Stata to simulate 2×2 tables of a one-off outcome. The simulation shows that odds ratios (ORs) are a much better estimate of the underlying causal effect or statistical association than relative rates are, given certain assumptions. One key assumption is that it is a one-off outcome, where it is reasonable to model the propensity for the event with

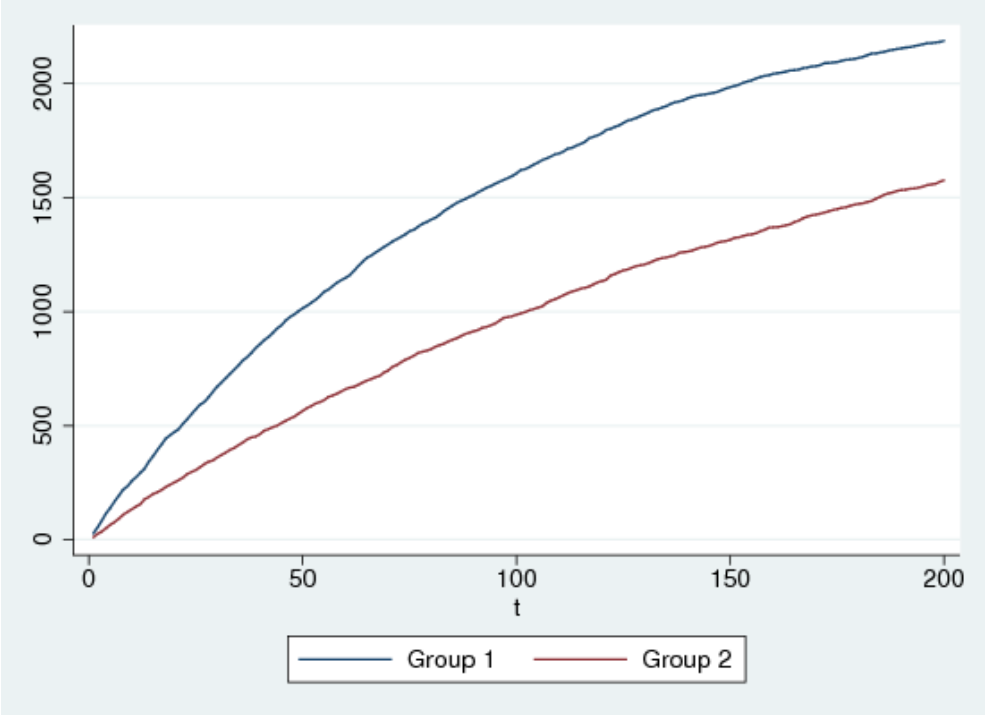

Figure 3: Cumulative infection rates in one run of the simulation

a normal or logistic distribution. Where the outcome is the result of potentially repeated exposure to a risk (such as being ever married or ever infected with a particular pathogen) the resulting propensity is not likely to be normal. That is, if you are exposed to many opportunities to marry, saying yes once means you become ever-married for ever after, and even if the propensity to marry at a specific opportunity is normally distributed, the combined distribution of propensity to be ever-married after an unknown number of opportunities is likely not to be well-described as normal.

I simulate this in terms of a epidemic: a population is exposed to a new pathogen, and I follow infection rates forward for a time. At each step of the simulation 100 individuals are chosen at random to be exposed to the pathogen, and they succumb at two separate rates: group 1 have a 0.25 probability, and group 2 a 0.50 probability. Once infected, you stay infected for the purposes of the summary. Individuals are likely to be exposed more than once (without consequence if they are already infected), though obviously they can't be exposed more than once in any single step. At each step, the 2×2 table is constructed, and the OR and RR calculated, and a complementary log-log model of the outcome is fitted. The code for the simulation is at `http://teaching.sociology.ul.ie/catdat/infection.do`.

Figure 3 shows the cumulative infection rates for the two groups, for one

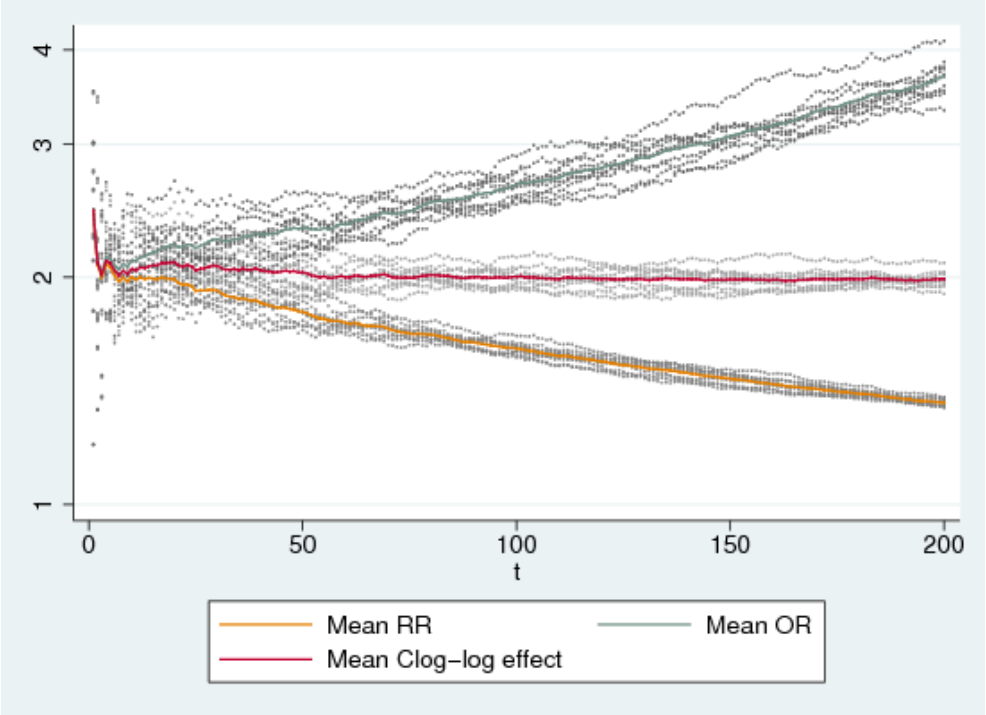

Figure 4: ORs, RRs and c-log-log estimates of the effect of smoking on ever becoming infected, calculated at each step of the simulation.

run of the simulation, with 200 iterations in a population of 5000, half in each group. Group 1 is much more susceptible, but is beginning to show signs of saturation, as the pool of uninfected subjects gets smaller. At time 200, about 85% of group 1 and 65% of group 2 have been infected, and this varies little across the multiple runs.

Figure 4 shows odds-ratios and relative rates (calculated arithmetically from the 2×2 table) and c-log-log regression parameters (exponentiated) for ten replications. RRs show the same sort of behaviour as with one-off outcomes: they begin at around the correct value (after a brief unstable period) but they head steadily towards the floor of no effect (RR=1) as the infection rate rises. ORs do the opposite, but to no better effect: they steadily deviate upwards from the correct value as infection increases. Only the exponentiated complementary-log-log estimate behaves well: it quickly settles very close to the ratio of 2.0 inherent in the simulation.

The fact that the complementary-log-log model generates consisent estimates of the effect suggest that it operates as a hazard model, since the ratio of 0.5/0.25 programmed into the simulation not a ratio of simple probabilities. That is, each probability is the probability of infection conditional on not yet being infected, and is thus a discrete hazard rate, not a probability, and

the ratio is a hazard rate ratio.

In passing I will note that these findings are in agreement with Pearce's robust defence of the OR versus the RR (2004), and in agreement with the detailed arguments about using ORs, RRs and other measures to make causal inferences, of Reichenheim and Coutinho (2010), while it runs counter to the support of log-binomial and other models in preference to logistic regression, of Barros and Hirakata (2003).

4 The benefit of non-linear models for non-linear problems – a simulation using R

Let's stay with logistic regression, but consider one aspect of the problem of interactions and logistic regression.[6]

A correspondent of Andrew Gelman's worries in passing about using Anova (i.e., linear regression) with binary dependent variables Gelman (2011). In particular, s/he worries that if there's a significant interaction in the Anova, but not in the logistic model, using the logistic model might be "losing" an interaction of substantive interest. However, it is more reasonable to think that if the dependent variable is binary, the linear model is mis-specified and that this mis-specification will cause problems with interactions. The intuition is that one variable (X_1) is high (and thereby leading to a greater propensity to have the outcome) the effect of changes in another variable (X_2) on the predicted probability will necessarily be smaller, than if X_1 is lower, even if there is no substantive interaction. There is a natural non-linearity in the relationship of the probability to the independent variables, which is dealt with by the log-odds transformation in logistic regression, but which can also be mathematically approximated by interactions in a linear model (though that might not be statistically defensible).

I present a simple simulation to demonstrate this. First, assume the outcome is more common as a continuous variable, X_1, rises, and more common for one value of a binary explanatory variable, X_2. This is simulated as a latent variable:

$$y^* \sim N(\beta_0 + \beta_1 X_1 + \beta_2 X_2, \sigma)$$

$$y = (y^* > 0)$$

X_1 and X_2 are correlated: X_1 has a higher mean for $X_2 = 1$.

We fit four models on the resulting data, a linear model on the unobserved scale variable y^*, a linear model on the observed binary y, and logit

[6]There's a possibly more interesting argument about interaction terms in logistic regression, started by Ai and Norton (2003), who claim that interactions in logistic regression cannot be interpreted directly. While their maths is impeccable, their claim rests on a definition of a "true interaction effect" that is quite contentious (see also Norton and Wang, 2004). But that's an argument for a different day.

and probit models on the binary. All models have the same linear predictor, containing the $X_1 \times X_2$ interaction.

The simulation counts the number of times each model return a t or z absolute value in excess of 2.0. See Figure 5 for the simulation code[7].

The first run sets the difference in the mean of X_1 to 1 across the two values of X_2:

```
> r <- simlogit(hilodiff=1,inteff=0,ncases=500,niter=1000)
Latent: 50; LPM: 329; Logit: 42; Probit: 38
Agreement: LPM:  0.671; Logit:  0.932; Probit:  0.938
```

The linear model on the latent variable yields a significant interaction about 5% of the time (the true null is rejected 5% of the time). The logit and probit show significant effects even less often, but the linear probability model (of the observed variable) estimates a significant interaction more that 30% of the time.

If we set `hilodiff` to a negative value (such that X_1 and X_2 have a negative correlation), significant interactions are found much more rarely. Where both variables have a positive effect on the probability, a positive correlation will cause the LPM to estimate spurious interactions, whereas the logit and probit approaches find them no more often than chance would suggest.

However, that's only one side of the issue. What about when there's a real interaction? How often will the LPM pick it up, while the logit fails to spot it?

```
> r <- simlogit(hilodiff=0,inteff=0.1,ncases=500,niter=1000)
Latent: 199; LPM: 37; Logit: 104; Probit: 99
Agreement: LPM:  0.810; Logit:  0.839; Probit:  0.838
```

In this example, with a positive interaction (the effect of X_1 is greater for X_2 high), the LPM performs poorly, finding about a fifth as many significant effects as the latent variable model, and less than half the logit and probit models. What if the interaction is negative?

```
> r <- simlogit(hilodiff=0,inteff=-0.1,ncases=500,niter=1000)
Latent: 186; LPM: 186; Logit: 110; Probit: 126
Agreement: LPM:  0.828; Logit:  0.832; Probit:  0.830
```

Oddly, in this case the LPM does very well. However, this is likely to be due to the LPM having an excessive tendency to report interactions under certain conditions, rather than a better ability to detect true interactions in these conditions – note that it reports many more significant interactions than

[7] R code at `http://teaching.sociology.ul.ie/catdat/g.R`

```
simlogit <- function(hilodiff = 0.5,
                     inteff = 0.0,
                     niter = 100,
                     ncases = 1000) {

results <- matrix(NA,niter,4)

for (iter in c(1:niter)) {

  ## Normally distributed x-var, differs by one SD in mean, male v high

  high <- c(rep(1,ncases/2) , rep(0,ncases/2))

  x1 <- rnorm(ncases,mean=0,sd=1) + hilodiff*high

  ## Dummy-var and x-var combine to create latent variable
  ystar <- 0 + (0.5 + inteff*high)*x1 + 0.5*high + rnorm(ncases,mean=0,sd=1)

  ## Latent variable reduced to binary
  y <- ystar>0

  ## Fit four models:
  ##  linear on latent variable,
  ##  linear on binary,
  ##  logit on binary,
  ##  probit on binary

  model1 <-  lm(ystar ~ as.factor(high)*x1)
  model2 <-  lm(y     ~ as.factor(high)*x1)
  model3 <- glm(y     ~ as.factor(high)*x1, family=binomial(link="logit"))
  model4 <- glm(y     ~ as.factor(high)*x1, family=binomial(link="probit"))

  ## save the z-stats for the interaction term
  results[iter,1] <- model1$coefficients[4] / sqrt(diag(vcov(model1)))[4]
  results[iter,2] <- model2$coefficients[4] / sqrt(diag(vcov(model2)))[4]
  results[iter,3] <- model3$coefficients[4] / sqrt(diag(vcov(model3)))[4]
  results[iter,4] <- model4$coefficients[4] / sqrt(diag(vcov(model4)))[4]

}

x = abs(results) > 2

cat(sprintf("Latent: %d; LPM: %d; Logit: %d; Probit: %d\n",
            sum(x[,1]),sum(x[,2]),sum(x[,3]),sum(x[,4])))

g12 <- table(x[,1],x[,2])
g13 <- table(x[,1],x[,3])
g14 <- table(x[,1],x[,4])

cat(sprintf("Agreement: LPM: %6.3f; Logit: %6.3f; Probit: %6.3f\n",
             sum(diag(g12))/sum(g12), sum(diag(g13))/sum(g13),
             sum(diag(g14))/sum(g14)))

return(results)

}
```

Figure 5: The simulation code

the logit or probit models, but does not agree better with the latent model than they do.

In general, what the simulations suggest clearly is that the LPM's response to mis-specification is to report interaction effects that are not present in the latent model: with a non-linear problem, one should believe the non-linear model over the linear approximation.

5 Agresti and Coull's formula for the variance of a proportion

Let's retreat from the complexities of linear and non-linear models, and consider one of the simplest of all categorical devices, the proportion. The formula for the standard deviation of a proportion, p, that has been taught to millions of undergraduates is:

$$\sqrt{p(1-p)}$$

It would seem that nothing could be simpler or better established, unless one were to go beyond textbooks and look into the literature. For the naïve (such as myself) it is then a little startling to see this formula contested, as for example by Agresti and Coull (1998). Rather than use the natural $p = \frac{x}{N}$ they propose a very odd $\tilde{p} = \frac{X+2}{N+4}$ in $\sigma = \sqrt{\tilde{p}(1-\tilde{p})}$. They decribe this as the "add two successess and two failures" approach, and recommend it strongly. Rather than go into the theory I run a simple simulation, comparing 1,000 samples of 100 for true values of p between 0.95 and 1, and count how many times the confidence interval around the observed rate contains the true rate, for the conventional approach and the add-two approach. Figure 6 shows how often add-two does worse (its interval doesn't contain the true value but the conventional interval does) and better (vice versa) than the conventional approach. As can be seen, the the Agresti-Coull estimate always performs better, dramatically so as p approaches 1.

6 Conclusion

These short notes have two things in common: a focus on relatively simple issues in categorical data analysis, and the use of simulation to resolve puzzles that lie behind the surface simplicity. Categorical data analysis is at once relatively transparent and dizzyingly complex. It is worth thinking carefully about the simple issues, and I hope the examples show both the value of simple simulations for helping one think, and the relative ease with which simulations can be conducted in Stata and R.

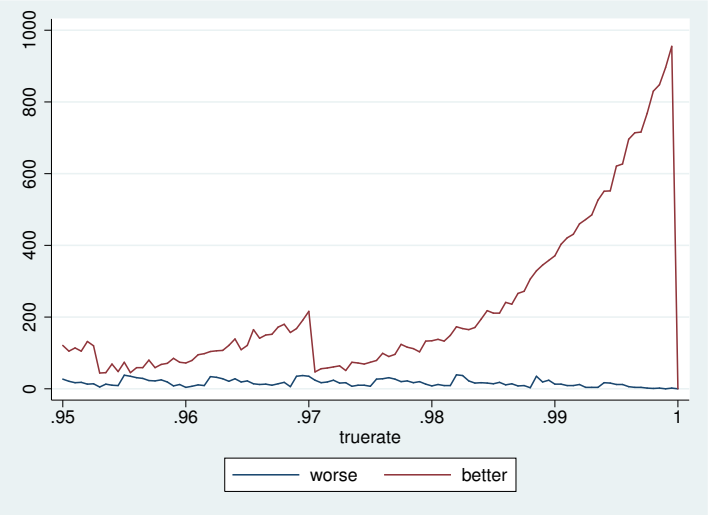

Figure 6: Add-two does better than the convention approach

Bibliography

Agresti, A. and Coull, B. A. (1998). Approximate is better than "exact" for interval estimation of binomial proportions. *The American Statistician*, 52(2):119–126.

Ai, C. and Norton, E. C. (2003). Interaction terms in logit and probit models. *Economics Letters*, 80(1):123–129.

Barros, A. J. and Hirakata, V. N. (2003). Alternatives for logistic regression in cross-sectional studies: An empirical comparison of models that directly estimate the prevalence ratio. *BMC Medical Research Methodology*, 3(21).

Gelman, A. (2011). A psychology researcher asks: Is anova dead? Blog post, *Statistical Modeling, Causal Inference, and Social Science*. `http://andrewgelman.com/2011/09/a-psychology-researcher-asks-is-anova-dead/`.

Norton, E. C. and Wang, H. (2004). Computing interaction effects and standard errors in logit and probit models. *The Stata Journal*, 4(2):154–167(14).

Pearce, N. (2004). Effect measures in prevalence studies. *Environmental Health Perspectives*, 112(10):1047–1050. doi: 10.1289/ehp.6927.

Reichenheim, M. E. and Coutinho, E. S. F. (2010). Measures and models for causal inference in cross-sectional studies: Arguments for the appropriateness of the prevalence odds ratio and related logistic regression. *BMC Medical Research Methodology*, 10(66).