The Stata Journal (2002)
2, Number 3, pp. 296–300

Least likely observations in regression models for categorical outcomes

Jeremy Freese
University of Wisconsin–Madison

Abstract. This article presents a method and program for identifying poorly fitting observations for maximum-likelihood regression models for categorical dependent variables. After estimating a model, the program `leastlikely` will list the observations that have the lowest predicted probabilities of observing the value of the outcome category that was actually observed. For example, when run after estimating a binary logistic regression model, `leastlikely` will list the observations with a positive outcome that had the lowest predicted probabilities of a positive outcome and the observations with a negative outcome that had the lowest predicted probabilities of a negative outcome. These can be considered the observations in which the outcome is most surprising given the values of the independent variables and the parameter estimates and, like observations with large residuals in ordinary least squares regression, may warrant individual inspection. Use of the program is illustrated with examples using binary and ordered logistic regression.

Keywords: st0022, outliers, predicted probabilities, categorical dependent variables, logistic regression

1 Overview

After estimating a linear regression model with a continuous outcome, data analysts will commonly calculate the residuals for each observation and examine those with the largest residuals more carefully to check if there is some discernible reason why the parameter estimates fit these observations so poorly. The commonsense understanding of observations with large residuals is that they are the observations for which the values of the dependent variable are most "surprising" ("unexpected", "weird"), given the regression coefficients and the values of the independent variables. In regression models for categorical outcomes, such as those discussed in Long (1997), residuals are not as readily conceptualized, although attempted extensions from models for continuous outcomes have been made (Pregibon 1981). Stata estimates many of these models using maximum likelihood, and, in many cases, the parameter estimates can be thought of as the set of estimates that maximizes the joint probability of observing the values of the outcome categories that were actually observed. Accordingly, for the observations within each outcome category, the most "surprising" are those that have the lowest predicted probabilities of observing that outcome, and these may warrant closer inspection by analysts for precisely the reason that observations with large residuals do in the more familiar linear regression model.

When run after various models for categorical outcomes, the command `leastlikely` will list the least likely observations. In other words, after a binary model, `leastlikely` will list both the observations with the lowest $\widehat{\Pr}(y = 0 \mid y = 0)$ and $\widehat{\Pr}(y \neq 0 \mid y \neq 0)$. `leastlikely` may be used after many models for binary outcomes in which the option `p` after `predict` generates the predicted probabilities of a positive outcome (e.g., `logit`, `probit`, `cloglog`, `scobit`, `hetprob`) and after many models for ordered or nominal outcomes in which the option `outcome(#)` after `predict` generates the predicted probability of outcome # (e.g., `ologit`, `oprobit`, `mlogit`). `leastlikely` is not appropriate for models in which the probabilities produced by `predict` are probabilities within groups or panels or for "blocked" data, and it will produce an error if executed after `blogit`, `bprobit`, `clogit`, `glogit`, `gprobit`, `nlogit`, or `xtlogit`.

2 Syntax

`leastlikely` [*varlist*] [`if` *exp*] [`in` *range*] [`, n(#) generate(`*varname*`)`
[`no`]`display nolabel noobs doublespace`]

where *varlist* contains any variables whose values are to be listed in addition to the observation numbers and probabilities.

3 Options

`n(#)` specifies the number of observations to be listed for each outcome. The default is 5. In the case of multiple observations with identical predicted probabilities, all will be listed.

`generate(`*varname*`)` specifies that the probabilities of observing the outcome value that was observed should be stored in *varname*. If not specified, the variable name `Prob` will be created but dropped after the output is produced.

The remaining options are standard options available after `list`.

`[no]display` forces the format into `display` or tabular (`nodisplay`) format. If you do not specify one of these two options, then Stata chooses one based on its judgment of which would be most readable.

`nolabel` causes the numeric codes rather than the label values to be displayed.

`noobs` suppresses printing of the observation numbers.

`doublespace` produces a blank line between each observation in the listing when in `nodisplay` mode; it has no effect in `display` mode.

4 Examples

Using data from Mroz (1987; see Long and Freese 2001, Chapter 4), I estimate a logistic regression model of the effects of several independent variables on a woman's probability of being in the labor force (`lfp`), and then I use `leastlikely` to list the predicted probabilities and observation numbers of the least likely observations.

```
. use binlfp2, clear
(PSID 1976 / T Mroz)
. logit lfp k5 k618 age wc hc lwg inc
Iteration 0:   log likelihood =  -514.8732
Iteration 1:   log likelihood = -454.32339
Iteration 2:   log likelihood = -452.64187
Iteration 3:   log likelihood = -452.63296
Iteration 4:   log likelihood = -452.63296
Logit estimates                                   Number of obs   =        753
                                                  LR chi2(7)      =     124.48
                                                  Prob > chi2     =     0.0000
Log likelihood = -452.63296                       Pseudo R2       =     0.1209
```

lfp	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
k5	-1.462913	.1970006	-7.43	0.000	-1.849027	-1.076799
k618	-.0645707	.0680008	-0.95	0.342	-.1978499	.0687085
age	-.0628706	.0127831	-4.92	0.000	-.0879249	-.0378162
wc	.8072738	.2299799	3.51	0.000	.3565215	1.258026
hc	.1117336	.2060397	0.54	0.588	-.2920969	.515564
lwg	.6046931	.1508176	4.01	0.000	.3090961	.9002901
inc	-.0344464	.0082084	-4.20	0.000	-.0505346	-.0183583
_cons	3.18214	.6443751	4.94	0.000	1.919188	4.445092

```
. leastlikely
Outcome: 0 (NotInLF)
          Prob
 60.  .1231792
172.  .1490344
221.  .1470691
235.  .1666356
252.  .1088271
Outcome: 1 (inLF)
          Prob
338.  .1760865
534.  .0910262
568.   .178205
635.  .0916614
662.  .1092709
```

Of respondents not in the labor force (`lfp`=0), observation #252 had the lowest predicted probability of not being in the labor force. For women in the labor force (`lfp`=0), observation #534 has the lowest predicted probability of being there.

Using data from Long and Freese (2001, Chapter 5; see also Long 1997), I estimate an ordered logistic regression model of the probability of agreement with the proposition that working mothers can establish as warm a relationship with their children as mothers who do not work (`warm`).

```
. use ordwarm2, clear
(77 & 89 General Social Survey)
. ologit warm yr89 male white age ed prst
Iteration 0:   log likelihood = -2995.7704
Iteration 1:   log likelihood = -2846.4532
Iteration 2:   log likelihood = -2844.9142
Iteration 3:   log likelihood = -2844.9123
Ordered logit estimates                           Number of obs   =       2293
                                                  LR chi2(6)      =     301.72
                                                  Prob > chi2     =     0.0000
Log likelihood = -2844.9123                       Pseudo R2       =     0.0504
```

warm	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
yr89	.5239025	.0798988	6.56	0.000	.3673037	.6805013
male	-.7332997	.0784827	-9.34	0.000	-.8871229	-.5794766
white	-.3911595	.1183808	-3.30	0.001	-.6231815	-.1591374
age	-.0216655	.0024683	-8.78	0.000	-.0265032	-.0168278
ed	.0671728	.015975	4.20	0.000	.0358624	.0984831
prst	.0060727	.0032929	1.84	0.065	-.0003813	.0125267
_cut1	-2.465362	.2389126	(Ancillary parameters)			
_cut2	-.630904	.2333155				
_cut3	1.261854	.2340179				

I use the option `n(3)` after `leastlikely` to restrict output to only the three most unlikely observations within each of the four outcome categories. Specifying `age` and `male` tells Stata to list values of these variables along with the observation numbers and probabilities.

```
. leastlikely age male, n(3)
Outcome: 1 (SD)
           Prob         age        male
 167.  .0401364          37       Women
 222.  .0449925          29       Women
 271.  .0407333          20       Women
Outcome: 2 (D)
           Prob         age        male
 563.  .1072643          41       Women
 803.  .1028648          30       Women
1001.  .1307181          32       Women
Outcome: 3 (A)
           Prob         age        male
1344.  .1559092          72         Men
1449.  .1358758          71         Men
1729.  .1283106          81         Men
Outcome: 4 (SA)
           Prob         age        male
1963.  .0387174          64         Men
2107.  .0413501          69         Men
2138.  .0393529          57         Men
```

As we would expect given the direction of the `ologit` coefficients, the respondents who strongly disagreed with the proposition but were least likely to do so were all younger women, while the least likely respondents who nonetheless strongly agreed with the proposition were all older males.

5 References

Long, J. S. 1997. *Regression Models for Categorical and Limited Dependent Variables*. Thousand Oaks, CA: Sage Publications.

Long, J. S. and J. Freese. 2001. *Regression Models for Categorical Dependent Variables using Stata*. College Station, TX: Stata Press.

Mroz, T. A. 1987. The sensitivity of an empirical model of married women's hours of work to economic and statistical assumptions. *Econometrica* 55: 765–799.

Pregibon, D. 1981. Logistic regression diagnostics. *Annals of Statistics* 9: 705–724.

About the Author

Jeremy Freese is Assistant Professor of Sociology at the University of Wisconsin–Madison.