Better Predicted Probabilities from Linear Probability Models

With Applications to Multiple Imputation

Paul D. Allison, Richard A. Williams, and Paul von Hippel

July 2020

allison@statisticalhorizons.com

Linear models can be useful for binary outcomes

Logistic or probit preferred in most applications.

But linear models still have some attractions:

- Ease of interpretation.
- Not subject to convergence failures.
- Computational speed for intensive applications:
  - Huge data sets (e.g., https://j.mp/38NrmRW)
  - Variable selection with large pools of predictors (e.g., https://doi.org/10.1198/016214504000000287)
  - Multiple imputation of binary variables

For arguments in favor of linear models for binary outcomes, see Paul von Hippel:
https://statisticalhorizons.com/when-can-you-fit
https://statisticalhorizons.com/linear-vs-logistic

The linear probability model (LPM)

Ordinary least squares with a dummy (0,1) dependent variable produces unbiased estimates of the coefficients in a linear probability model:

$$p_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_k x_{ik}$$

where p_i is the probability that the dependent variable = 1.

But there are three well-known downsides:

- Inherent heteroscedasticity leads to
  - Standard error estimates that are not consistent and, hence, inaccurate *p*-values
  - Inefficient parameter estimates
- Non-normality of the dependent variable can also make *p*-values inaccurate.
- A linear model for a probability is inherently unrealistic.
  - With continuous *x*'s, there's always the possibility of implied probabilities greater than 1 or less than 0. Even if that doesn't happen, LPM doesn't do well in the vicinity of 1 or 0.

Can these defects be remedied?

- Heteroscedasticity is easily fixed with robust standard errors.
- Non-normality is a trivial problem with moderate to large size samples.
- The most intractable problem has been non-linearity, manifest by predicted probabilities greater than one or less than zero.
  - This is very common.
  - May not be an issue if the main interest is in testing hypotheses and estimating effects.
  - But there are many applications where getting valid predicted probabilities is essential, e.g.,
    - Inverse probability weighting
    - Propensity score methods
    - Expected loss
    - Discrete-time survival functions

Many authors state that invalid predicted probabilities are the principal disadvantage of LPMs (e.g., Westin 1973, Long 1997, Hellevik 2007, Wooldridge 2010, Greene 2017).

How to get predicted probabilities in the (0,1) interval?

My original problem: How to speed up multiple imputation for missing data?

- Fastest methods are based on the multivariate normal model (MVN), implying that variables with missing data are imputed by linear regression.
- But when imputing categorical variables, you often get probabilities greater than 1 or less than 0.

Amelia II, a popular R package for imputation based on MVN, truncates predicted values at 0 or 1 .

Me: There's got to be a better way.

A solution: The linear discriminant model (LDM)

R.A. Fisher (1936) proposed the linear discriminant function as a method for classifying units into one or the other of two categories, based on a linear function of the predictor variables. Here's the model:

- Let **x** be the vector of predictors, and let *y* have values of 1 or 0.
- Assume that within each category of *y*, **x** has a multivariate normal distribution, with different mean vectors for each category ($\mathbf{u}_1$ and $\mathbf{u}_0$), but a covariance matrix **S** that is the same for both categories.

Two remarkable facts about the LDM:

Fact 1. The LDM specifies the conditional distribution of **x** given the value of *y*. Using Bayes' theorem, it can be re-expressed to give the conditional probability of of *y*, given **x**. This yields a logistic regression model:

$$\log[P(y=1|\mathbf{x})/Pr(y=0|\mathbf{x})] = a + \mathbf{b'x},$$

where *a* and **b** are functions of the two mean vectors and the covariance matrix.

Fact 2. Maximum likelihood estimates of *a* and **b** can be obtained (indirectly) by OLS regression of *y* on **x** (Haggstrom,1983).

- To estimate **b**, the OLS slope coefficients must each be multiplied by $K = N / RSS$ where *N* is the sample size and *RSS* is the residual sum of squares. *K* is typically substantially larger than 1.
- The intercept *a* is obtained as follows. Let *m* be the sample mean of *y* and let *c* be the intercept in the OLS regression. Then,

$$a = \log[m/(1-m)] + K(c-.5) + .5[1/m - 1/(1-m)]$$

A better way to get predicted probabilities

The LDM method:

1. Estimate the LPM by OLS.
2. Transform the parameters as described in Fact 2.
3. Generate predicted probabilities using the logistic equation in Fact 1.

This produces predicted values guaranteed to lie in the (0,1) interval!

Three tools to make this easy:

- Stata ado file: **net install predict_ldm, from(https://www3.nd.edu/~rwilliam/stata)**
- SAS macro: https://statisticalhorizons.com/resources/macros
- R function: https://statisticalhorizons.com/better-predicted-probabilities

Is the LDM method any good?

A major reason for concern:

- LDM assumes multivariate normality of the predictors.
- Few applications will meet that assumption, or even come close.

Several studies have suggested that the LDM is pretty robust to violations of MVN

- Press and Wilson 1978, Wilensky and Rossiter 1978, Chatla and Shmueli 2017
- None of those investigations was very systematic.
- Focused on coefficients and test statistics, not on predicted probabilities.

I've applied the method to 15 data sets to test it out.

- I first fit the linear model and applied the LDM method to get predicted probabilities.
- Then I fit a logistic model using the standard ML method.
- I compared predicted probabilities from LDM and standard logistic regression in several ways.

Standard logit should be the gold standard. LDM can't do any better than conventional logit because both rely on the same underlying model for *y*, but LDM makes additional assumptions about the predictor variables.

Example 1. Women's Labor Force Participation

753 married women (Mroz 1987)

Dependent variable: **inlf** = 1 if the woman is currently in the labor force (478 women), otherwise 0.

Predictor variables: number of children under the age of six, age, education (in years), and years of labor force experience, as well as the square of experience.

```
use "https://statisticalhorizons.com/wp-content/uploads/MROZ.dta",clear
reg inlf kidslt6 age educ exper expersq
predict_ldm
```

This new command generates the usual predicted values (with the default name **yhat**) and predicted values based on LDM (with the default name **yhat_ldm**)

```
logit inlf kidslt6 age educ exper expersq
predict yhat_logit
summarize yhat yhat_ldm yhat_logit
corr yhat yhat_ldm yhat_logit

scatter yhat yhat_logit, xtitle(Logistic Predicted Probabilities) ///
  ytitle(Fitted Values from Linear Model)
scatter yhat_ldm yhat_logit, xtitle(Logistic Predicted Probabilities) ///
  ytitle(LDM Predicted Probabilities)
```

Example 1 Results

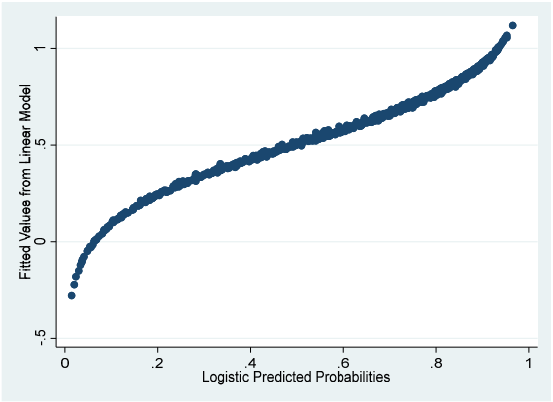
14 cases < 0

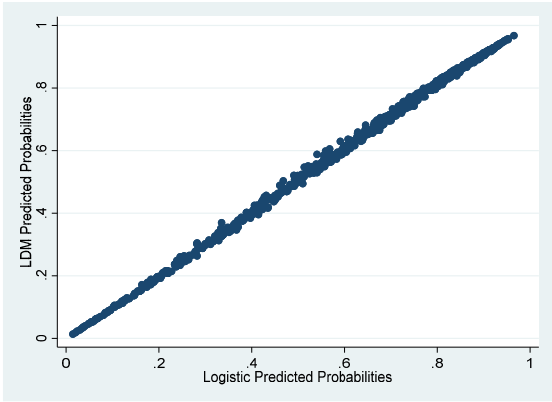
12 cases >1

```
    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
        yhat |        753    .5683931    .2517531  -.2782369   1.118993
    yhat_ldm |        753    .5745898    .2605278   .0136687   .9676127
  yhat_logit |        753    .5683931    .2548012   .0145444   .9651493

             |     yhat yhat_ldm yhat_logit
-------------+---------------------------
        yhat |   1.0000
    yhat_ldm |   0.9870   1.0000
  yhat_logit |   0.9880   0.9994   1.0000
```


Example 1 Scatterplots





12 of the 15 data sets produced similar results.

Example 2. Mortality For Lung Cancer Patients

1,029 patients. Death is the DV, and 764 of the patients died. All predictors are categorical:

- surgery (1 or 0)
- a randomized treatment (1 or 0)
- hospital (Mayo Clinic, Johns Hopkins, or Sloan Kettering)
- cancer stage at diagnosis (1, 2, or 3).

Both of the 3-category variables are represented by two dummy variables. The fitted models included main effects of these variables but no interactions.

```
use "https://statisticalhorizons.com/wp-content/uploads/lung.dta",clear
gen dead=surv!=0
reg dead operated treat i.instit i.stage
predict ldm
logit dead operated treat i.instit i.stage
predict yhat_logit
sum yhat yhat_ldm yhat_logit
corr yhat yhat_ldm yhat_logit
scatter yhat yhat_logit, xtitle(Logistic Predicted Probabilities) ///
  ytitle(Fitted Values from Linear Model)
scatter yhat_ldm yhat_logit, xtitle(Logistic Predicted Probabilities) ///
  ytitle(LDM Predicted Probabilities)
```

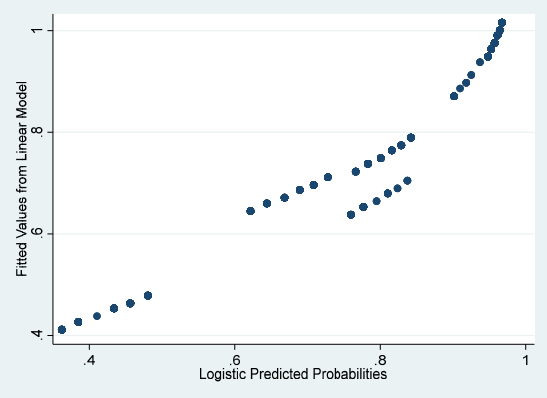

Example 2 Results

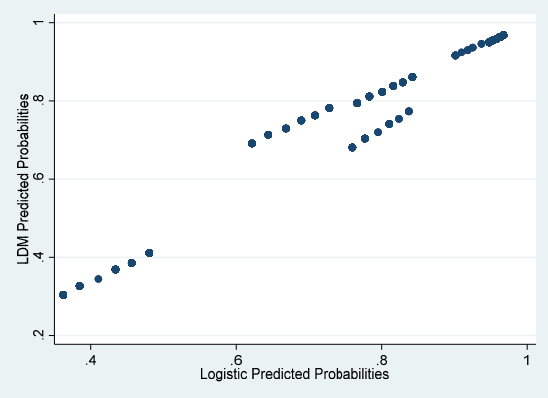
```
    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
        yhat |      1,029    .7424684    .2206033   .4116536    1.01567
    yhat_ldm |      1,029    .7299414     .251858    .303894   .9679288
  yhat_logit |      1,029    .7424684    .2247691   .3621077   .9674457
```

99 cases had this value.

```
             |     yhat yhat_ldm yhat_logit
-------------+---------------------------
        yhat |   1.0000
    yhat_ldm |   0.9754   1.0000
  yhat_logit |   0.9815   0.9908   1.0000
```

Example 2 Scatterplots





Example 3. Diabetes in NHANES data

Here's one example that didn't work so well.

10,337 respondents. DV=1 if they had diabetes, else 0. Predictor variables are sex, age, and race, with an interaction between age and race.

```
webuse nhanes2f, clear
reg diabetes black female age c.age#c.black
predict_ldm

logit diabetes black female age c.age#c.black
predict yhat_logit

sum yhat yhat_ldm yhat_logit
corr yhat yhat_ldm yhat_logit
scatter yhat yhat_logit, xtitle(Logistic Predicted Probabilities) ///
  ytitle(Fitted Values from Linear Model)
scatter yhat_ldm yhat_logit, xtitle(Logistic Predicted Probabilities) ///
  ytitle(LDM Predicted Probabilities)
```

This interaction is highly significant.

This interaction is not significant.

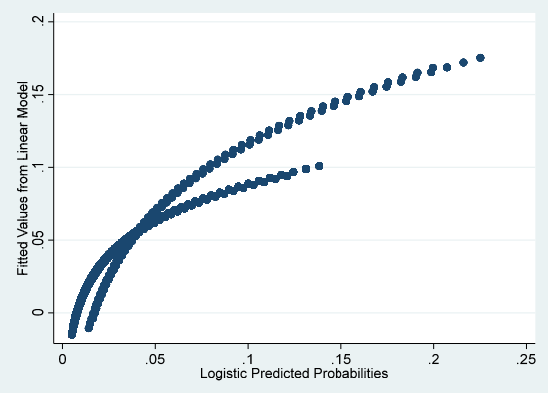
Example 3. Results

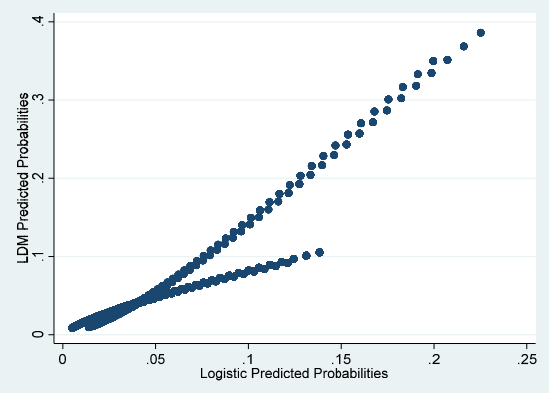
1,352 cases < 0

Variable	Obs	Mean	Std. Dev.	Min	Max
yhat	10,337	.0482781	.0394012	-.0149728	.1753265
yhat_ldm	10,337	.0489435	.0490787	.0085806	.3861544
yhat_logit	10,337	.0482765	.0414374	.0050582	.2251559

	yhat	yhat_ldm	yhat_logit
yhat	1.0000		
yhat_ldm	0.8700	1.0000	
yhat_logit	0.9509	0.8968	1.0000

Example 3. Scatterplots





If you remove the interaction from both models, LDM does much better than LPM.

Where to go from here?

Simulations to study impact of departures from MVN.

- Skewed distributions
- Categorical predictors
- Models with interactions
- Large numbers of predictors
- Non-linear effects

Get confidence intervals for predicted probabilities (maybe with delta method).

Extend methods to unordered categorical variables with more than two categories.

- Haggstrom already developed the theory: For a *k*-category variable, estimate *k*-1 OLS regressions and do appropriate transformations of the coefficients.

Write Stata command to do MVN imputation when some variables categorical.

- Will probably use EMB algorithm of Honaker and King rather than MCMC
- Compare with "chained equations" method

References

Allison, P.D. (2005). Multiple imputation of categorical variables under the multivariate normal model. Presented at the Annual Meeting of SUGI (SAS User's Group International), Philadelphia, PA, April 2005. https://statisticalhorizons.com/resources/unpublished-papers

Chatla, S. B., & Shmueli, G. (2017). An extensive examination of regression models with a binary outcome variable. *Journal of the Association for Information Systems*, *18*(4), 1.

Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. *Annals of eugenics*, *7*(2), 179-188.

Greene, W.H. (2017) *Econometric Analysis*. Pearson.

Haggstrom, G. W. (1983). Logistic regression and discriminant analysis by ordinary least squares. *Journal of Business & Economic Statistics*, 1(3), 229-238.

Hellevik, Ottar (2009): Linear versus logistic regression when the dependent variable is a dichotomy. *Quality & Quantity* 43.1 59-74.

Long, J. S. (1997) *Regression models for categorical and limited dependent variables*. Sage Thousand Oaks.

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.

Press, S. J., & Wilson, S. (1978). Choosing between logistic regression and discriminant analysis. *Journal of the American Statistical Association*, *73*(364), 699-705.

Westin, R. B. (1974). Predictions from binary choice models. *Journal of Econometrics*, *2*(1), 1-16.

Wilensky, G. R., & Rossiter, L. F. (1978). OLS and logit estimation in a physician location study. In *Proceedings of the Social Statistics Section, American Statistical Association* (pp. 260-265).

Wooldridge, J. M. (2010). *Econometric analysis of cross section and panel data*. MIT press