Abe Burton, Nicholas Oxendon, Carter Jensen, Nate Tollet

ECON 588

3/11/22

Dr. Frandsen

Binary Prediction in Econometrics

Introduction and Theory:

Binary prediction is a common econometric tool that can be done in several ways. It is very useful in a variety of scenarios where an outcome's probability can be turned into a prediction of an outcome. Several methods have been developed to deal with this type of question, each with its own strengths, weaknesses, and relevant assumptions. Our purpose is to build the most common of these tools and test their performance with data that challenges their assumptions.

One possible way to handle a binary dependent variable is through a linear probability model. This is done by essentially running an OLS regression,

$$\Pr\left(Y = 1 \mid X = x_i\right) = x'\beta$$

where Y is interpreted as the probability of our outcome being 1 and our β as the marginal effect of a 1 unit increase in x_i.

Advantages of the linear probability model include its computational efficiency and well-understood implications. Since it is just a normal regression, we can calculate β by simply applying our existing OLS formula,

$$\hat{b} = (X'X)^{-1}XY$$

This estimator shares all understood properties of OLS, including unbiasedness, efficiency, and asymptotic normality. However, there are clear drawbacks to using an OLS regression to predict a binary outcome. One obvious issue is the presence of nonsensical predictions. By construction,

OLS assumes an unconstrained and continuous outcome variable. So, for extreme values of x_i, fitted values will be above 1 and below 0. It also will assume a constant partial effect of any explanatory variable, even though it is likely that the effect of a change in x_i may vary depending on its magnitude. Lastly, having a binary outcome variable will necessarily cause our estimator to be heteroskedastic, since the variance of a binary variable is

$\Pr(Y = 1 \mid X_i)(1 - Pr(Y_i = 1|X_i)$, and cannot be some σ^2 for all x_i. Intuitively, since all observations have outcomes of either 1 or 0, but our linear model gives us fitted values somewhere between 1 and 0, (when they are sensical predictions) the size of our residuals will depend on our value of x_i. This issue can be solved like other cases of heteroskedasticity by using robust standard errors.

An effective way to address the issues of Linear Probability ability estimation is by using MLE estimation. MLE is a type of M-estimator, where we have some likelihood function representing the joint distribution of the observed data and our parameters. For computational reasons, we often use the log-likelihood function,

$$\ln L(\beta|X) = \sum_{i=1}^{n} \ln g(x_i\beta)$$

Where $g(x_i\beta)$ is the underlying density function of our data. In this case, our MLE estimator is given by,

$$\hat{b} = \arg max_b \sum_{i=1}^{n} \ln g(x_i\beta)$$

Similarly to other M-estimators, our MLE estimator is consistent and asymptotically normal. To perform MLE estimation, we need to make an assumption about the underlying distribution of our data. We will look at two options, Probit and Logit estimation.

Probit

Probit estimation assumes that our distribution $g(x\beta) = \Phi(x\beta)$ (standard normal CDF). Because the values given by the CDF function will always be bounded between 0 and 1, our first issue with our linear probability model, nonsensical predictions, is avoided when using Probit. It also allows for nonlinear trends in our data, allowing for a potentially better fit than the traditional LPM.

Logit

Logit estimation works similar to Probit, but instead of the standard normal CDF, our underlying distribution is assumed to be,

$$g(x\beta) = \frac{e^{x\beta}}{1+e^{x\beta}}$$

It can be seen that, like Probit, our logit estimator is bounded between 1 and 0. It also allows for nonlinear trends in our data. There is no consensus on which is the "better" estimator. One possible advantage of Logit is it is more computationally efficient than Probit. It also will generally have fatter tails, so it will approach 0 and 1 slightly slower, adding robustness to our model. Ultimately, whichever estimator is closer to the true underlying distribution should give a more accurate estimate.

Literature Review:

The linear probability model has been the subject of some criticism among researchers concerning its usefulness in calculating binary outcomes in RCTs. A similar question exists for

its usefulness in quasi-experimental data, which in reality is the majority of what is available to estimate causality. John Deke from the HHS explains in a brief why the linear probability model is appropriate for both of these situations. The linear probability model, of course, is the use of binary outcomes in the place of continuous ones in OLS. The estimate can be interpreted as a probability. Deke notes that while logistic regression (logit) and the linear probability model both yield the same average treatment effect, there are certain advantages to the generality of the linear probability model, and he illustrates this with a sample regression of predicted probabilities. The estimates given by the linear probability model can be directly interpreted as the mean marginal effect, while those of the logistic model cannot. This gives the advantage of being more intuitively interpretable, giving estimates that can be easily shared and understood by a more broad audience. The drawback to this is that it misses the true non-linear relationship of a binary outcome and a continuous covariate. There then remains the possibility that the linear probability model can regurgitate nonsense predictions of less than 0 percent or more than 100 percent.

The empirical evidence as given by Deke suggests that perhaps we should not be overly concerned with the linear probability model giving nonsense estimates. Using Monte Carlo simulations, he shows that the linear probability model performs as well or better in most scenarios than logistic regression. He summarizes these in 4 key findings- first, if treatment perfectly predicts outcome, logistic regression will fail to appropriately estimate the impact, while the linear probability can. Second, he found that the linear probability model faced issues of bias in far fewer cases than logistic regression. Third, and a disadvantage for the linear probability model, is that logistic regression is typically more precise. Finally, and perhaps a reason to disregard the disadvantage presented in the third finding, is that the standard error

given by the linear probability model is far more often correct than the one given by logistic regression, whose standard error is often too small. Linear probability models seem to sacrifice some specificity for interpretability.

Logistic regression, or logit, could be considered a sister model to the linear probability model. Both are interpreted similarly, but logit provides a better approximation of the true nonlinear form that a probabilistic function necessarily takes. Unlike the linear probability model, we will not see predicted probabilities below 0 percent or above 100 percent. And, perhaps an advantage over probit regression, logit tails are fatter, approaching the limits of 0 and 100 more slowly and providing room for the possibility of more extreme values. Stone and Rasp (1991) use accounting choice studies in gathering evidence on the utility of logit regression. Their key findings show that even in sample sizes as small as 50 observations, logit may still be preferable to OLS, although, with such a small sample, the results are likely to be biased. Thus, even in a small sample, logit may outperform an OLS estimate on a binary outcome variable.

This confirms some preconceptions while also setting some bounds on its usefulness. In conjunction with Noreen (1988), they conclude that the assumption that researchers have often made that the sample size is sufficient for logit is often not true. After running 10,000 replications of a model with one predictor, they conclude that at least 50 observations may be needed for logit to outperform OLS, and 100 observations when there is skewness in the predictor. This jumps to 200 observations in a model with skewness and multiple predictors. In cases where the sample size does not exceed these bounds, OLS test statistics might be better calibrated, but we may still be willing to sacrifice some accuracy for flexibility. Even in this case where logit seems like it may be outperformed, logit is likely to lead to lower misclassification rates, less meaningless predictions, and more powerful tests of parameter estimates. Researchers,

therefore, find that logit, while biased in small samples, is a better classifier and provides more useful estimates than OLS. We see fewer nonsense predictions than in the linear probability model, and it is computationally easier than probit. Its accuracy and test power increase with sample size. Logit may be the estimation method of choice for research with dichotomous outcomes.

Although logit and probit are nearly identical in their estimates, Noreen (1988) notes that probit has been the method of choice for researchers dealing with dichotomous outcomes, despite its computational inefficiency. Noreen's results on probit regression are nearly identical to logit regression in comparison with OLS. Using 1000 Monte Carlo trials, he found that OLS seemingly performed better on small sample sizes (<50) than probit. He concludes that in these scenarios, the evidence does not support the use of probit over OLS. Like logit, probit increases in accuracy and test power as sample size increases, but it seems that logit tends to be more useful than probit in small sample sizes. Why then is probit so often utilized, when it is computationally more difficult than logit and less powerful than OLS? There are a few reasons noted in the literature. First, the marginal effects in probit are more intuitive than logit or OLS because it is based on a normal distribution. The linear probability model may be more intuitive, but we may still prefer probit for its nonlinearity and accuracy. Second, because probit approaches the bounds more slowly than logit, it may reduce the effect of outliers on our estimates. If we are concerned that extreme cases in the tails of the distribution may skew our inference, we may prefer probit to logit. Probit, although widely used by researchers in the accounting field as noted by Noreen, may only be more useful than logit or the linear probability model in the specified cases above. Otherwise, we may prefer logit when nonlinearity is more important, or the linear probability model if we seek intuitiveness.

Research Design

To further test the properties and performance of the linear probability, logit, and probit models, we wrote linear probability, logit, and probit models and estimated models for data generated using a modified version of Julian Winkel and Tobias Krebs' Python package Opossum.

For the linear probability model, we used the classic closed-form solution

$$\hat{b} = (X'X)^{-1}XY$$

to identify the coefficients that minimize the sum of squared error, which coefficients in the case of ordinary least squares also minimize the negative log-likelihood of the least squares model based upon the strong assumption that the error term in the model, ε, is normally distributed with mean zero and constant variance σ^2.

For the logit model, we use maximum likelihood estimation to solve for the unique parameters that minimize the negative log-likelihood of the logistic model, which posits that event likelihood is linked to the logistic function. For the probit model, we adopt a similar approach. Using maximum likelihood estimation, we solve for the unique parameters that minimize the negative log-likelihood of the joint likelihoods of each observation in the entire sample characterized by the standard normal cumulative distribution function.

Specifically, to successfully minimize the log-likelihood functions of the logit and probit models, we use the L-BFGS-B gradient descent algorithm provided by the popular Python package Scipy to numerically estimate the coefficients, $\widehat{b}$.

The covariate data used in our model is sampled randomly from a multivariate normal distribution with mean zero and variance Σ, where Σ is a positive definite matrix constructed from a uniform distribution multiplied element-wise with an overlay matrix consisting of random values in the set {-1, 1} to allow for some covariates with negative correlation.

The outcome variable in our data is linearly related to the covariates in our model. True coefficients are drawn from a beta distribution and then linearly combined with the covariates to generate a probability estimate for y. The probability estimates for y are clipped between .1 and .9, and then y is converted to a binary outcome via draws from a Bernoulli distribution. To create additional noise in our model, we apply both logistic and normal error terms to the linear combination of covariates and true coefficients before drawing outcome variable values from the Bernoulli distribution.

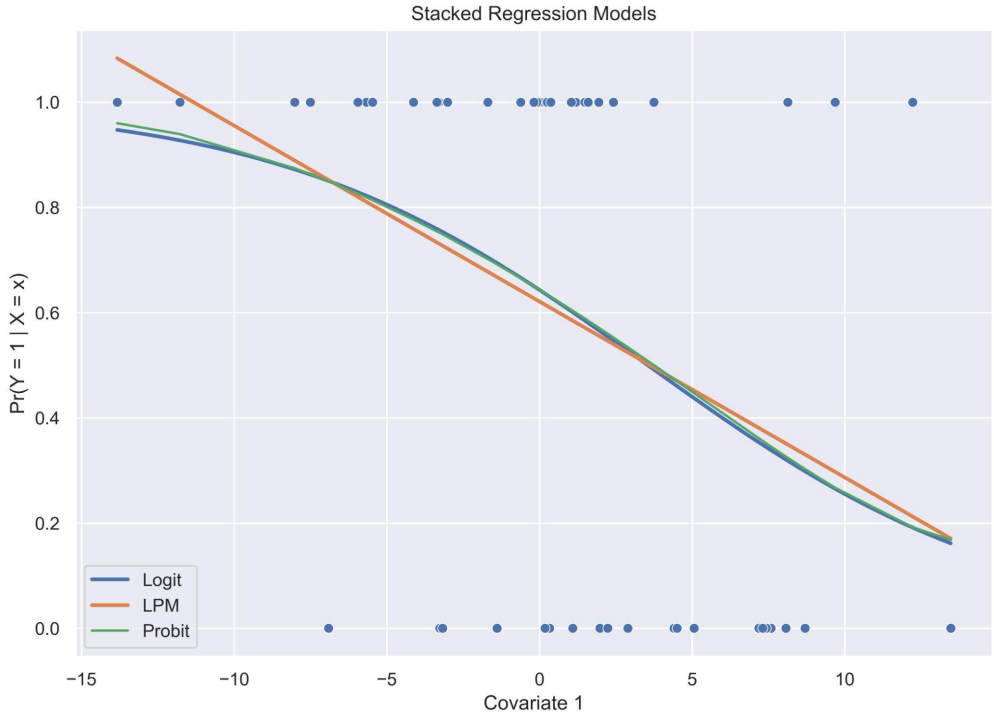
Results:

Our models for logit, probit, and LPM performed well on the data that was generated. We tested the validity of the models we built by comparing their results to the results of their respective models in the SKLEARN package in Python. Our coefficients and mean squared error matched the results of those models which are from a commonly used package. The general performance of our models was encouraging, and the results of testing their assumptions on the data were for the most part consistent with what we expected.

	Logit Model Performance	Probit Model Performance	LPM Model Performance
Data with normal error	**R-squared:** 0.53 **β:**[-.08,-.18,-.17]	**R-squared:** .51 **β:** [.48, -.03, -.03]	**R-squared:** 0.50 **β:** [.48, -.03, -.03]
Data with logistic error	**R-squared:**.525 **β:** [.62,-.059,.019]	**R-squared:** .502 **β:** [.64, -.01,.004]	**R-squared** -.523 **β:** [.64,-.01,.004]

Consistent with the theory, logistic regression performed the best of the three when the data had a logistic error term. Probit and LPM gave similar results to each other and both were worse than logit. Probit did better on the data with the normal error term than it did with logit, which also is what we would have expected in the beginning. However, logit still outperformed both other models on the data with normal error where we would have thought probit would do better. This could be due to the phenomena identified by Noreen that logit does better with smaller sample sizes than probit. For the most part, these

results conform to our expectations of how the models should perform based on the theory they were based on.



Conclusion:

By following an MLE and OLS framework we were able to replicate results from models that are commonly used by the Python community. Our results supported the theory that the distribution of the error term is an important factor in the model performance of logit and probit. Even though logit did better in all cases, each model performed its individual best when matched to its theoretical best error. Further research could be done to test the circumstances that lead to logit's high performance in this simulation. LPM is useful for its computational ease and interpretability but is not the best model because of its potential for nonsensical output and its slight disadvantage in predictive power in some cases. Our results are for the most part an encouraging affirmation of the theory that inspired this simulation research from the beginning.

Works Cited

Deke, John. “Using the Linear Probability Model to Estimate Impacts on Binary Outcomes in Randomized Controlled Trials.” *Mathematica*, https://mathematica.org/publications/using-the-linear-probability-model-to-estimate-impacts-on-binary-outcomes-in-randomized.

Noreen, Eric. “An Empirical Comparison of Probit and OLS Regression Hypothesis Tests.” *Journal of Accounting Research*, vol. 26, no. 1, 1988, p. 119., https://doi.org/10.2307/2491116.

Rosett, Richard N., and Forrest D. Nelson. “Estimation of the Two-Limit Probit Regression Model.” *Econometrica*, vol. 43, no. 1, 1975, p. 141., https://doi.org/10.2307/1913419.

Stone, Mary, and John Rasp. *Tradeoffs in the Choice between Logit and OLS for Accounting Choice Studies*. American Accounting Association, Jan. 1991, https://www.jstor.org/stable/247712.