AMSI VACATIONRESEARCH SCHOLARSHIPS 2020–21

Get a Thirst for Research this Summer

Robustness of Limited Dependent Variable Models to Misspecification

Luke Thomas

Supervised by A/Prof Leandro Magnusson and Professor Inge Koch

University of Western Australia

Vacation Research Scholarships are funded jointly by the Department of Education, Skills and Employment and the Australian Mathematical Sciences Institute.

Abstract

This project explores the validity of using linear regression to model Limited Dependent Variable models, in particular binary choice data. It is very common in Economics to use only the Linear Probability Model, and it is also common to do so without investigating the required assumptions on the residuals. However, if the likelihood function does not represent the data's generative process, then the model may not be correctly specified. We use Monte Carlo simulations to estimate the distribution of coefficients and marginal effects, for different Binary Regression models, with proper and misspecified residuals. Surprisingly, for our Monte Carlo models the Linear Probability Model coefficients and marginal effects are reasonably robust, whereas the logit and probit models — although accurate when correctly specified — are not robust to misspecification.

Contents

1 Introduction

In Economics, we are frequently interested in considering whether a policy works or not, or understanding the root causes of social phenomena. This is often done in econometrics by identifying causal channels or variables, and then testing the size, significance, and direction of regression coefficients. When the outcome is binary, the most common regression model used is the "Linear Probability Model". However, supposing the residuals of the generative process are not as specified by the model, this has the potential to yield incorrect results. In this research project we are interested in applying Monte Carlo simulation to test the robustness of estimators from Limited Dependent Variable (LDV) models — particularly Binary Choice models — to misspecification of the residuals. This project is interdisciplinary, combining the Economics and statistics literature.

There is significant research done into LDV which demonstrates that "the consequences of violation of the normality assumption in LDV situations can be quite severe" as the MLE are inconsistent when they are specified differently to the generative process (Jarque, Bera and Lee, 1984). There are a number of different kinds of LDV models, each designed for differently censored or truncated variables. In this project we are concerned with "binary choice data".

In particular for binary choice data, there are trade-offs between using Linear Probability Models (LPM) through a simple ordinary least squares (OLS) procedure, or using some nonlinear index function such as the probit or logit model. LPM is arguably simpler and easier to interperet, whereas logit and probit are more sophisticated but overcome some of the LPM difficulties, such as predicting results outside of the support. Contentious debate surrounds the trade-offs of these approach, and their relative robustness. A popular econometrics book concludes that "while a nonlinear model may fit the [Conditional Expectation Function] for LDVs more closely than a linear model, when it comes to marginal effects this probably matters little. This optimistic conclusion is not a theorem, but ... it seems to be fairly robustly true." (Angrist & Pischke, 2008, p. 80).

In this report we will first outline some of the motivation behind this project, and background on the techniques used. Next, we will explore the two experiments: the first, a simulated cross-sectional model, and the second, a simulated fixed effects panel model. I find in each case that the results are somewhat robust, with the LPM generally performing better under misspecification, and all models performing worse in the fixed effects panel case.

1.1 Statement of Authorship

A/Prof Leandro Magnusson, UWA, and Professor Inge Koch, UWA, formulated the project idea and supervised the research. I, Luke Thomas performed the simulations, and interpreted the results with support from Magnusson and Koch. I would like to thank AMSI for their generous financial support.

2 Background

In this section, I lay out the motivation for the project, the background econometrics, and outline the models used.

2.1 Motivation

Economists and social scientists are often concerned with big causal questions. Their work often involves evaluating and testing government policy, and using econometric techniques to isolate the impact of different variables on one another.

These questions often involve assessing how a factor affects an outcome. While there are a range of econometric techniques used to try and isolate causal channels of effect, the final model specification is often similar to regressions seen in applied statistics.

For example, consider the following important economics questions:

- **Does U.S. Food Aid cause civil conflict in developing countries?** (AER, 2014)
- **Do workplace smoking bans reduce smoking?** (AER, 1999)
- **Is medical care use sensitive to cost?** (RAND Health Insurance Experiment)

For each of these cases, researchers were trying to isolate the impact of some variable — respectively, US food aid, workplace smoking bans, and medical costs — on relevant outcomes. Also, for each of these cases, the outcomes considered are binary variables. For example: whether or not a country is in conflict, whether or not someone smokes, and whether or not someone uses medical care.

As a result, it is important that we understand the models used to assess these questions. In this Vacation Research Project, we look at simulating the models commonly used, in order to assess how robust they are to misspecification. That is, what if the models are specified incorrectly: would the results still hold?

2.2 Model Specification

There are three common types of binary regression models which we consider in this research. These are the linear probability model, the probit model, and the logit model.

2.2.1 Linear Probability Model

Arguably the most common approach used in econometrics is the linear probability model (LPM). A linear probability model is "any regression where the dependent variable is zero-one" (Angrist Pishcke, p. 36).

The model is of the form, with y being a binomial variable:

$$E[y|\boldsymbol{X}] = P(y = 1|\boldsymbol{x}) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots$$

The coefficients here represent the change in probability of success based on a unit change of the covariate. For binomial data, the following hold:

$$E[y|\boldsymbol{X}] = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots,$$

and

$$Var[y|\boldsymbol{X}] = \boldsymbol{X}\beta(1 - \boldsymbol{X}\beta).$$

This shows that the OLS regression produces " unbiased estimators" of the coefficients (Wooldridge, 2002). The variance of y depends on $\boldsymbol{x}$ however, which implies heteroskedasticity (by design, there is non-constant variance).

Of course, the major flaw with the linear probability model is that it has limited predictive power, and for some covariate values will produce estimates that are outside of the probabilistic bounds of $[0, 1]$.

2.2.2 Probit Model

The probit and logit models are a nonlinear types of regression used in econometrics. They are termed "Index Models" by modern econometrics texts, as they are reliant on a secondary nonlinear transformation.

That is, they can be thought of as:

$$P(y = 1|\boldsymbol{x}) = F(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots).$$

Where the function $F(\cdot)$ may take different forms. It must however be a Cumulative Distribution Function, which is limited at 0 and 1. The motivation behind this approach relies on what is called a **Latent Variable Model**. A Latent Variable model in econometrics is different to the Factor Analysis concept of the same name in statistics, and instead refers to an underlying data generating process which is not observed.

We denote the unobserved latent variable as y^*, and the observed variable as y. To outline this concept, we use the notation in Wooldridge, 2002.

A latent variable model in econometrics is slightly different to its Factor Analysis meaning in statistics. Instead, Latent Variable Models are constructed to represent the underlying data generating process behind binary observations.

The latent variable is a variable which determines the result of the observed variable. For example, consider the case of the following possible observed binary data: whether or not a family chooses to have children. This decision might be a result of a number of variables: income, age, country, and many others. We consider that these variables might be related to an unobserved variable: "the utility (or internal cost-benefit analysis) of having children". If the utility of the decision is positive, an individual has children, and if it is negative, they do not.

This process of an unobserved latent variable (such as utility) being transformed into an observed variable (the decision or outcome), is the motivation behind the nonlinear models we consider.

The latent variable is assumed to have some distribution based on observable indepedent variables ($\boldsymbol{x}$) and noise (e):

$$y^* = \boldsymbol{x}\beta + e,$$

where the observed variable depends on whether that variable is above or below 0:

$$y = 1[y^* > 0].$$

It can be shown that the probability of the observed variable, then, is linked to the predicted value of the latent variable by the distribution of the error term.

$$P(y = 1|\boldsymbol{x}) = P(y^* > 0|\boldsymbol{x}) = P(e > -\boldsymbol{x}\beta|\boldsymbol{x}) = 1 - F(-\boldsymbol{x}\beta) = F(\boldsymbol{x}\beta).$$

For the probit model, the CDF and residual distribution is that of the standard normal:

$$F(z) \equiv \Phi(z) \equiv \int_{-\infty}^{z} \frac{1}{2\pi} e^{\frac{-t^2}{2}} dt.$$

2.2.3 Logit Model

The logit model, also known as "Logistic Regression", is commonly seen in statistics and machine learning as a regression of covariates on the log-odds of an event. The model is often motivated differently in econometrics, as an index function, but it is inherently the same model. That is, it has the same likelihood function and estimators.

The logit model is an index model as above, but instead of the residual being distributed according to the standard normal, it follows the "logistic distribution". That is,

$$F(z) = \Lambda(z) \equiv \frac{e^z}{1 + e^z}.$$

With a little massaging, it can be seen that the above approach leads to the same result as a log-odds regression with the same statistics (Wooldridge, 2002).

2.3 Marginal Effects

The nature of Ordinary Least Squares (OLS) is that its coefficient estimates provide a simple interpretation of how a unit change in the independent variables affects the dependent variable. For other regression techniques including Generalised Linear Models (GLM) such as logit and probit, which involve a transformation on the linear predictor, the impact of unit changes in independent variables is less immediately obvious.

In these cases, it is useful to consider the Marginal Effects. Stated simply, for a regression of the form:

$$\hat{y} = \beta_0 + \beta_1 X_1 + \beta_2 X_2.$$

The marginal effect with respect to X_1 is $\frac{\partial \hat{y}}{\partial X_1} = \beta_1$.

There are a number of common ways to cite the marginal effects of nonlinear regressions. The three common ways are the Marginal Effects at Means (MEMs), Marginal Effects at Representative Values (MERs) and Average Marginal Effects (AMEs). The Marginal Effects at Means are found by taking the marginal effects of each covariate at the mean of each covariate. The Marginal Effects at Representative Values is similar, but instead of letting the covariates be means, they are set at specific important values. Finally, the Average Marginal Effects are the most commonly used, and find the marginal effects at every value of X, and then take the mean across these.

In this simulation, we calculate Average Marginal Effects using Thomas Leeper's `margins` package in R (Leeper, 2017). This package, similar to the STATA statistical software often used in econometrics, calculates the partial derivates numerically:

$$f'(x) = \lim_{h \to 0} \frac{f(x+h) - f(x)}{h}.$$

3 Case 1: Cross-Sectional Model

Now, we explore our first experiment. Here we are assessing the robustness of the coefficients and marginal effects under a simple cross-sectional model.

3.1 Data Generating Process (DGP)

In order to isolate for the robustness of the regression techniques, we construct a unique Data Generating Process which can be replicated. Binomial regression concerns itself with estimating the probability of a binary choice (labelled 1 a *success* or 0 a *failure*), given some exogenous independent variables. As described, the underlying process motivating the probit/logit model of such data, is the estimation of a "latent variable model".

We simulate this in the following way:

$$y_i^* = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \upsilon_i \tag{1}$$

$$y_i = 1_{\{y_i^* \geq 0\}} \tag{2}$$

For each y_i^*, υ_i represents the residuals or noise, and is our primary object of interest. For these simulations we set $n = 100$, and describe three independent variables and a constant. Note that we refer to independent variables as analagous to the explanatory variables or covariates — they are not necessarily uncorrelated. Equation (2) shows an indicator function representing censoring: if and only if the value of y^* is positive will the observed value will be a success.

As a result, for the particular DGP we investigate, we describe the above latent variable model in the linear form:

$$\overbrace{y^*}^{(100\times1)} = \overbrace{\boldsymbol{X'}}^{(100\times4)} \overbrace{\beta}^{(4\times1)} + \overbrace{\upsilon}^{(100\times1)} \tag{3}$$

Where the true parameter (β) of the latent variable model are:

$$\beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix} = \begin{bmatrix} -0.8 \\ 0.2 \\ 0.3 \\ -0.5 \end{bmatrix}$$

Below in Figure 1 you can see an example histogram of the latent variable and its corresponding censored observed variable.

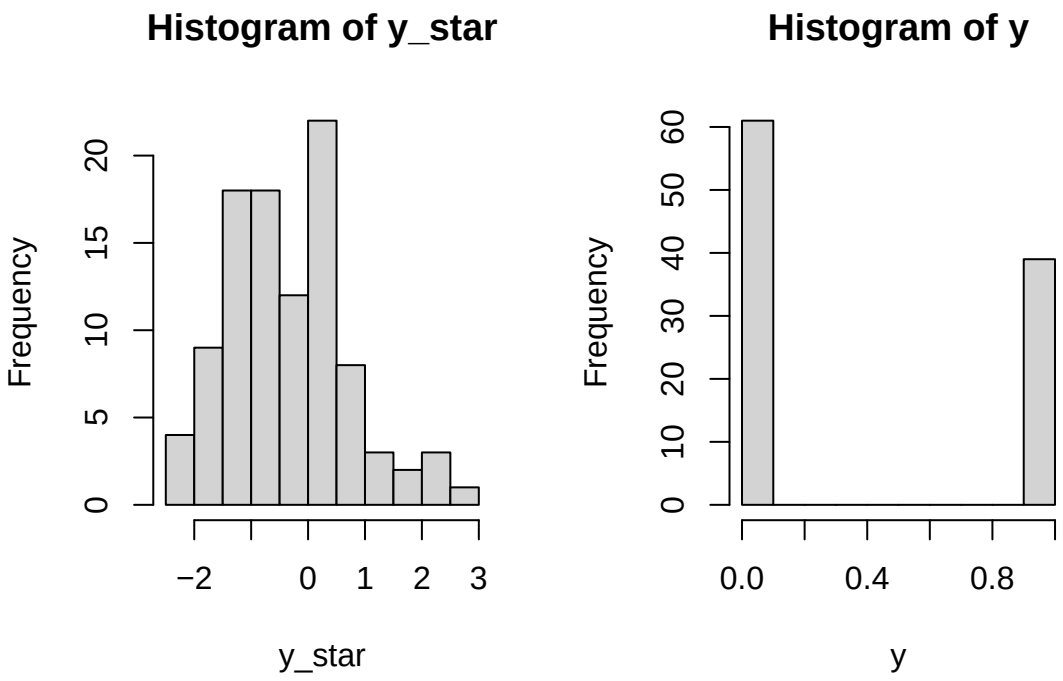

Figure 1: Histogram of Simulated y_i^* and y_i

In order to emulate the variety of independent variables that may be found in the real world, we simulate three differently distributed independent variables as follows:

$$X_{0i} = 1, \qquad \text{A vector of 1's}$$
$$X_{1i} \sim N(1, 1)$$
$$X_{2i} \sim Poisson(\lambda = 1)$$
$$X_{3i} \sim Unif(-1, 1)$$

It should be noted that **the parameters and independent variable distributions are arbitrary**. They are designed purely to provide a reproducible simulation of some data. It should also be noted that only one set of independent variables is ever drawn from the above distributions. All further resimulations involve *resimulation of the residual* and not of the whole model.

3.2 Methodology

Our methodology involves two major steps: first, we simulate the data under the correct specification of the model (using a standard normal distribution on the residual), and then we simulate it using residuals which are autocorrelated. In each case we find the coefficients and the marginal effects, and then assess their deviation from the true values.

For the first step, we model the residuals as independently and identically distributed (iid) standard normal, which would mean that the probit model is properly specified.

Using the fixed independent variables, we simulate the residuals $N = 10,000$ times to find the distribution of estimators given the correct specification of the model as outlined.

Next, we are interested in the coefficients and marginal effects under misspecification. A key candidate for misspecification, termed endogeneity in economics, is error autocorrelation. This is where the residuals may be correlated with one another. In economics, this may be observed when a specified model is fitted with omitted or simultaneously determined variables. For example, there may be two variables which are both correlated with a single hidden variable, and hence the residuals may be autocorrelated.

To model misspecification, we choose to simulate residuals with an AR(1) structure (that is, autoregressive with one lag), such that $\upsilon_i = \rho\upsilon_{i-1} + \nu_i$.

We can demonstrate this through the following covariance matrix:

$$\Omega = \underbrace{SS'}_{M\times M} = \begin{bmatrix} 1 & \rho & \rho^2 & \dots & \rho^M \\ \rho & 1 & \rho & \dots & \rho^{M-1} \\ \rho^2 & \rho & 1 & \dots & \vdots \\ \vdots & \vdots & \vdots & \ddots & \rho \\ \rho^M & \rho^{M-1} & \dots & \rho & 1 \end{bmatrix}$$

where $S'S$ is the Cholesky decomposition. It can be shown that $S\upsilon_i$, premultiplying residuals by the lower triangular S, will embed this autocorrelation structure in the covariance matrix.

It can be shown that $S'\upsilon$ is the new residuals of interest.

$$\begin{aligned} Var(S'\upsilon) &= S'Var(\upsilon)S && (4)\\ &= S'IS' && (5)\\ &= S'S = \Omega && (6) \end{aligned}$$

Here ρ represents the degree of correlation (the proportion of one residual explained by another), and υ_i are the correctly specified residuals (standard normal).

Fortunately, we can use this covariance matrix to directly simulate random normal variables using the `mvrnorm()` function from the `MASS` package in R.

We resimulate the noise using this structure, and then use this to construct the latent variable model. The latent variable model is then censored to become binary choice data. We use this new data to test the robustness of the estimators calculated.

3.3 Results

Firstly, for the correctly specified simulations we can find the mean coefficients in Table 1.

Table 1: Mean Coefficients of Regression output

Estimator	LPM	probit	logit
$\hat{\beta}_0$	0.22	-0.84	-1.39
$\hat{\beta}_1$	0.07	0.21	0.35
$\hat{\beta}_2$	0.11	0.32	0.53
$\hat{\beta}_3$	-0.18	-0.53	-0.88

The results from the nonlinear logit and probit in Table 1 clearly do not tell us much about the magnitude of the change in probability caused by a unit change in each covariate. Between the LPM, logit, and probit we can see the direction of the change in probability from a unit change in covariates is the same. However, except for the linear probability model, the relative magnitude of that change is not clear. To find the more comparable average marginal effects, see Table 2.

It can be seen that mean marginal effects across models are much more similar, with the probit model slightly overstating the impact of a unit change in covariate. As expected, the coefficients of the LPM are equivalent to the marginal effects.

Table 2: Mean Marginal Effects of Regression output

Estimator	LPM	probit	logit
$dydx\_X1$	0.07	0.08	0.08
$dydx\_X2$	0.11	0.11	0.11
$dydx\_X3$	-0.18	-0.19	-0.20

We take density plot of these simulated coefficients under the correct specifiation, and now look at comparing them to their misspecified counterparts. These resulting distributions can be seen in Figure 2 and Figure 3 for coefficients and marginal effects respectively. Each of these only shows the coefficients and marginal effects on the first covariate.

In the density plots shown in Figure 2 you can respectively find the distributions of the LPM, probit, and logit under misspecifcation and different degrees of autocorrelation for the coefficients of X_1. This exercise can be repeated for each covariate, but for simplicity we chose to focus on X_1. We consider this first correctly

specified, and then with ρ increased to 0.5. A non-parametric kernel estimation technique is used to model the density function of each distribution.

For each case we can see that the correctly specified model, without autocorrelation, tends to produce estimators with a lower variance than the misspecified model. This suggests that the correctly specified coefficient may converge more quickly to the mean. While only a small difference, the higher the autocorrelation, the larger the difference between the correct and misspecified model estimates' distributions.

The probit model, properly specified, correctly estimates the mean of its latent variable model coefficient around 0.2. However, it is potentially concerning that the marginal effects of the probit model do not appear to converge with the LPM. This could be an interesting question for further research.

For both the probit and logit model, the distribution of the marginal effects appears strongly influenced by endogeneity. It not only appears to shift the central tendency of the distribution, but also induce peculiar behaviour around the correctly specified means. While this behaviour seems more strong for the logit model, which is misspecified from the beginning, this could just be due to random variation in sampling. The distribution of marginal effects at the tail (95th percentile) can be seen in Appendix III — while we had limited time to investigate effects at the tails, it is an interesting area for future research.

Overall, for misspecification in the cross-sectional model, each of the LPM, probit and logit seem remarkably robust.

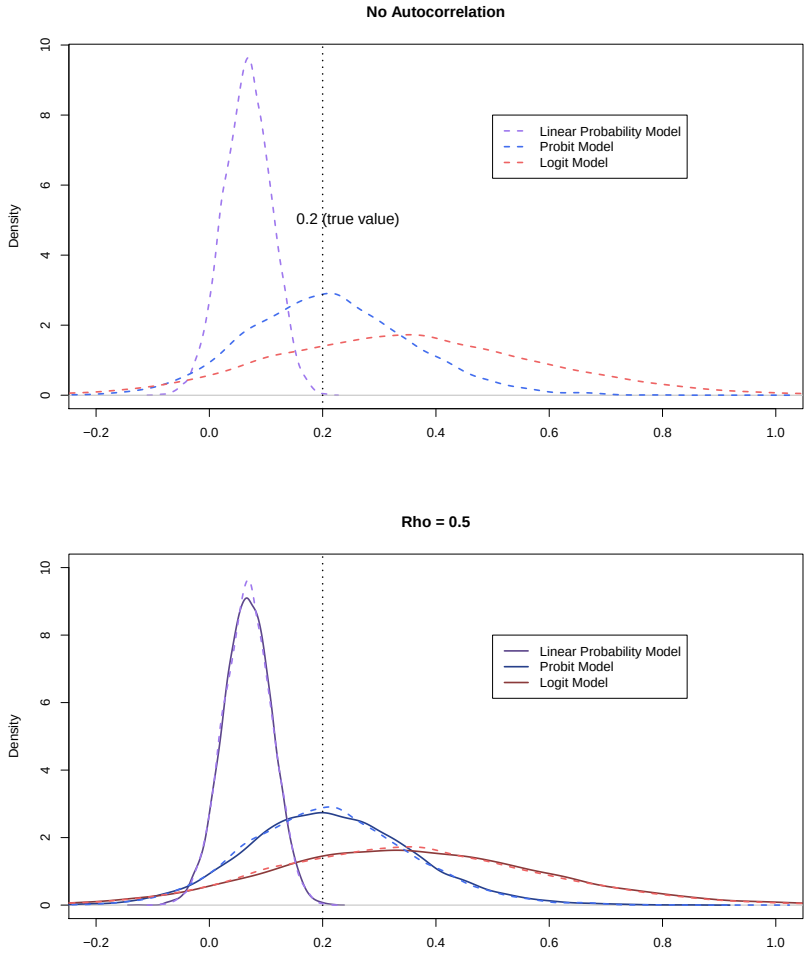

Figure 2: Simulated Distribution of X_1 Coefficient Estimates under Different Autocorrelation Structures

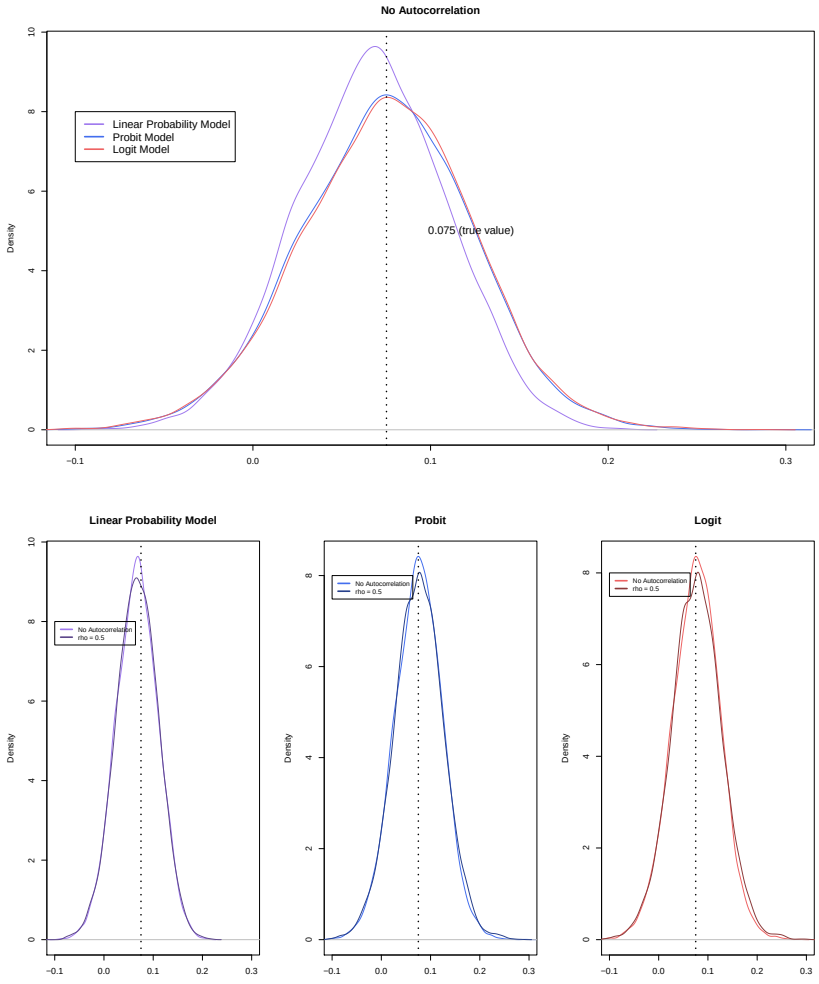

Figure 3: Simulated Distribution of X_1 Marginal Effects Estimates under Different Autocorrelation Structures

4 Case 2: Fixed Effects Panel Model

In econometrics, statistical techniques have been designed to investigate panel data. Panel data are data where particular observations are followed over time — a given unit, such as a country, an individual, or a business, will have regular variables recorded. The data can be thought to have a two-unit index: unit and time.

Like other kinds of data, the variable of interest in a panel dataset may be binary choice. In this kind of data, linear probability models and Index models such as the logit can be motivated as above.

4.1 Data Generating Process

As above, we simulate a panel dataset with a predictable Data Generating Process.

We reduce some of the complexity in the above model to form more predictable behaviour. We construct our Latent Variable Model:

$$y^*_{i,t} = \beta_0 + \beta_1 X_{1,it} + \phi_i + \upsilon_{it} \tag{7}$$

$$y_{it} = 1_{y^*_{i,t}>0}. \tag{8}$$

$y^*_{i,t}$ is our unobserved Latent Variable Model, and y_{it} is the observed variable. X_1 is our only explanatory variable, and is distributed:

$$X_{it} \sim N(1,1)$$

The true values of β used in our simulation are arbitrarily set as:

$$\beta = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} = \begin{bmatrix} -0.8 \\ 0.2 \end{bmatrix}$$

Finally, ϕ_i is the individual-specific fixed effect. That is, an adjustment of the intercept depending on the observation. We derive this from a random uniform distribution on $[-1, 1]$ — again, set arbitrarily to generate data.

We ignore the probit model in these simulations and focus on correctly specifying the logit model. As a result, the residual of the latent variable model, υ is specified correctly with a *standard logistic distribution*.

4.2 Methodology

We follow a similar methodology as before: first simulating the binary data under the correct specification (in this case, of the logit model), and then under residual autocorrelation.

First, we assume that we have a balanced panel dataset with 1,000 observations. These observations comprise 100 individuals across 10 time periods. The distributions of the explanatory variables over time is assumed to remain the same.

Next, we pre-define $\boldsymbol{\phi}$, the vector of fixed effects across all individuals. To do this, we sample from a uniform distribution on $[-1, 1]$ 99 times, assuming that $\phi_1 = 0$. When the dataset is resimulated, we assume that these ϕ_i remain constant.

Finally, for each individual i we sample $X_{1,it}$ 10 times, indexed by $t = 1, \ldots, 10$, then simulate the residuals from a logistic distribution and produce the y^* and y as in equation (8). We gather all of the observations into one dataset.

We then repeat this process 1,000 times to produce 1,000 datasets. Using the 1,000 simulated datasets, we apply the linear probability model and logit model as earlier specified to derive coefficients.[1]

We now introduce a misspecification of the residuals in our simulations. That is, the residuals no longer follow the logistic distribution as is assumed by the logit model. We consider two degrees of autocorrelation, where $\rho = 0.5$ and 0.9.

[1] We use the `plm` package and `cquad` package respectively to model the linear panel models and the conditional logit models. To understand the fixed effects more clearly, it can be useful to think of these packages as modelling a series of dummy variables representing each individual. These variables "switch on" for each of the individuals, to provide a fixed adjustment of the intercept.

We introduce this misspecification by pre-multiplying the residuals within individuals by the lower triangular Cholesky decomposition of the previously seen AR(1) matrix. To summarise:

- Correct Specification: $\upsilon_1 \sim Logis(0,1)$[2]
- Misspecifiaction: $\upsilon_2 \equiv S\upsilon_1$

Where:

$$Var(S\upsilon_1) = \underbrace{SVar(\upsilon_1)S'}_{M\times M} \tag{9}$$

$$= S(\frac{\pi^2}{3}I)S' \tag{10}$$

$$= \frac{\pi^2}{3}SS' \tag{11}$$

$$= \frac{\pi^2}{3}\begin{bmatrix} 1 & \rho & \rho^2 & \dots & \rho^M \\ \rho & 1 & \rho & \dots & \rho^{M-1} \\ \rho^2 & \rho & 1 & \dots & \vdots \\ \vdots & \vdots & \vdots & \ddots & \rho \\ \rho^M & \rho^{M-1} & \dots & \rho & 1 \end{bmatrix} \tag{12}$$

For each of these misspecifications we resimulate as in the previous methodology, but using the new residuals.

4.3 Results

The means of the coefficients under each resimulation can be seen in Table 4. It can be seen then in Table 4 that in each the estimators move away from the unbiased or true result as misspecification increases. This implies that the coefficients are not robust to this misspecification.

Table 3: Mean and Mean Standard Deviation of $\hat{\beta}_1$ Regression output

Estimator	LPM	logit
Mean	0.04	0.20
Mean Standard Deviation	0.02	0.07

Turning to Figures 4 and 5, we can see the distributions of these estimators under each condition. Interestingly, the coefficients of the LPM model seem more robust to autocorrelation than the logit model. This may be because of the natural lack of sensitivity in the LPM (as it calculates marginal effects), but it would be

[2]The Standard Logistic Distribution

Table 4: Mean of Misspecified $\hat{\beta}_1$ Regression output

Estimator	LPM	logit
True Mean	N/A	0.2
Mean ($\rho = 0$)	0.04	0.20
Mean ($\rho = 0.2$)	0.04	0.21
Mean ($\rho = 0.5$)	0.04	0.22
Mean ($\rho = 0.9$)	0.04	0.38

prudent to also find the marginal effects of the logit model and investigate how they change as a result of the coefficients changing.

The marginal effects, calculated as described in Appendix I, are presented in Appendix II. This plot shows that a similar trend is observed: the linear probability model seems significantly more robust.

In the logit model the mode reliably increases with autocorrelation, whereas the mode in the LPM appears to adjust less predictably. The logit model coefficients are distributed with a larger standard deviation as the autocorrelation increases, although this trend seems to be almost inverted for the LPM.

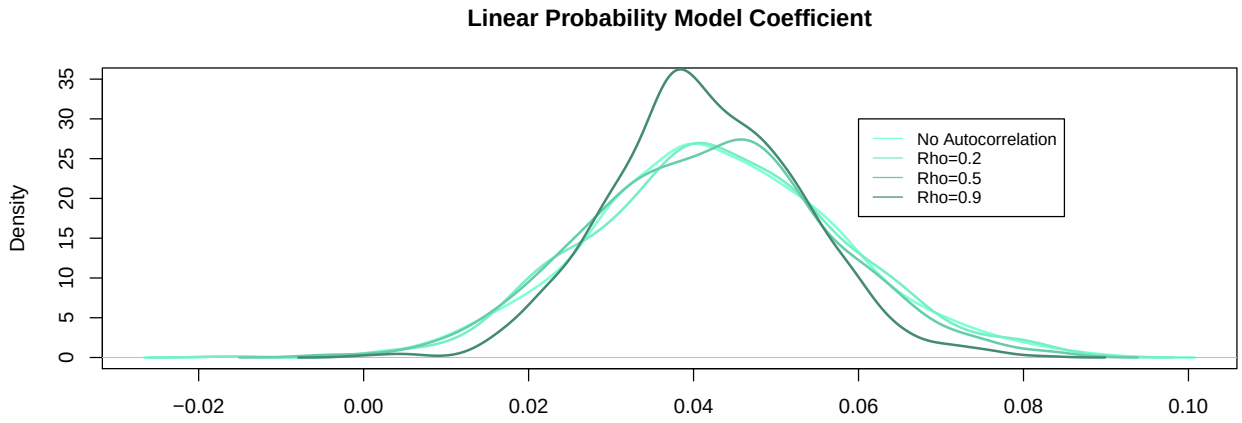

Figure 4: Simulated Distribution of X_1 Estimates (LPM)

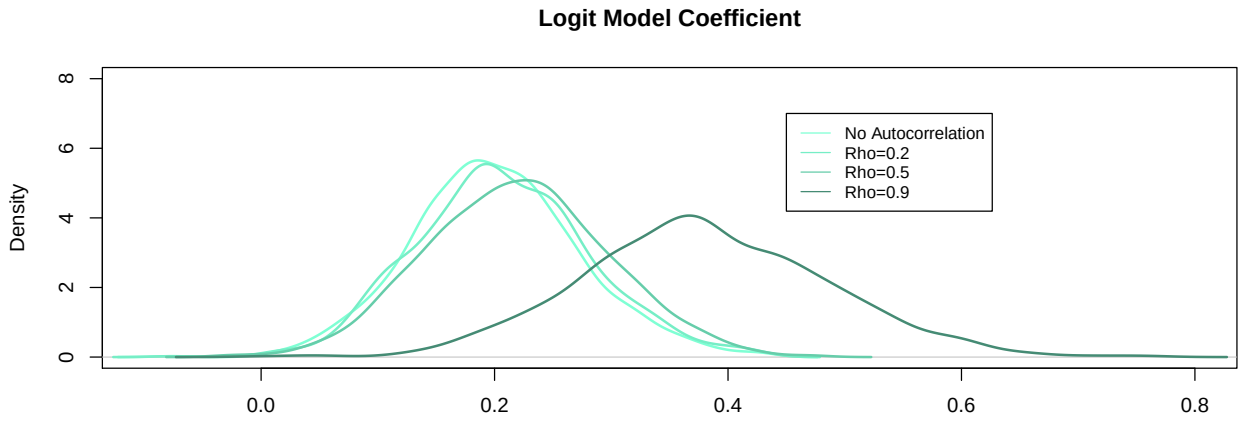

Figure 5: Simulated Distribution of X_1 Estimates (logit)

In summary, the results appear as expected with more autocorrelation responding with a larger deviation from the true coefficients. The logit model, correctly specified, returns a close approximation to the truth, but when misspecified also appears the most sensitive.

Broadly speaking, the Fixed Effects Panel Model is markedly less robust to misspecification than the simple cross sectional model.

5 Conclusion

In this project we were interested in how robust the coefficients are under model misspecification. This is an important question, as it details the consequence of poorly fitted regressions when considering major social science questions.

We use Monte Carlo simulations to estimate the distribution of coefficients and marginal effects, for different Binary Regression models, with proper and misspecified residuals.

We find that the linear probability model tends to be more robust to specification, but is not as accurate as a properly specified logit or probit model is. These effects are more pronounced for the fixed effects panel model than the simple cross-sectional model.

These results are specific to the Monte Carlo simulations we have run, and it motivates interesting future research into the behaviour of these coefficients and marginal effects under misspecification. It would be valuable to consider these effects through a theoretical lense in future research.

Appendix I: Calculating Marginal Effects

For the second case, it is not possible to use the `margins` package in R. Instead, take the following approach to calculating the marginal effects:

Say

$$y^*_{i,t} = \beta_0 + \beta_1 X_{1,it} + \phi_i + \upsilon_{it} \tag{13}$$

$$y_{it} = 1_{y^*_{i,t}>0} \tag{14}$$

Therefore:

$$\begin{aligned} P(y_{it} = 1|\boldsymbol{X}) &= P(y^* > 0|\boldsymbol{X}) && (15)\\ &= P(\beta_0 + \beta_1 X_{1,it} + \phi_i + \upsilon_{it} > 0|\boldsymbol{X}) && (16)\\ &= P(\upsilon_{it} > -(\beta_0 + \beta_1 X_{1,it} + \phi_i)|\boldsymbol{X}) && (17)\\ &= 1 - \Lambda(-(\beta_0 + \beta_1 X_{1,it} + \phi_i)) && (18)\\ &= \Lambda(\beta_0 + \beta_1 X_{1,it} + \phi_i) && (19) \end{aligned}$$

Note that Λ refers to the CDF oof the logistic distribution, as described earlier. Therefore the marginal effects of X_1 are:

$$\begin{aligned} \frac{P(y_{it} = 1|\boldsymbol{X})}{\partial X_{it}} &= \frac{\partial}{\partial X_{it}} \Lambda(\beta_0 + \beta_1 X_{1,it} + \phi_i) && (20)\\ &= \lambda(\beta_0 + \beta_1 X_{1,it} + \phi_i) \cdot \beta_1 && (21) \end{aligned}$$

To find an estimate of the marginal effects, we take β_0, ϕ_i to be the true values from the simulation. We take $\beta_1 = \hat{\beta}_1$. To find the Average Marginal Effects, we calculate this at each observation of X and then take the mean:

$$AME = \frac{\sum_{it} \lambda(\beta_0 + \hat{\beta}_1 X_{1,it} + \phi_i) \cdot \hat{\beta}_1}{N} \tag{22}$$

Appendix II: Marginal Effects (Case 2)

Here you can see the marginal effects, calculated as in Appendix I, for case 2. It can be seen that where the LPM stays fairly robust, the Logit model performs much worse under misspecification and appears to be biased upwards.

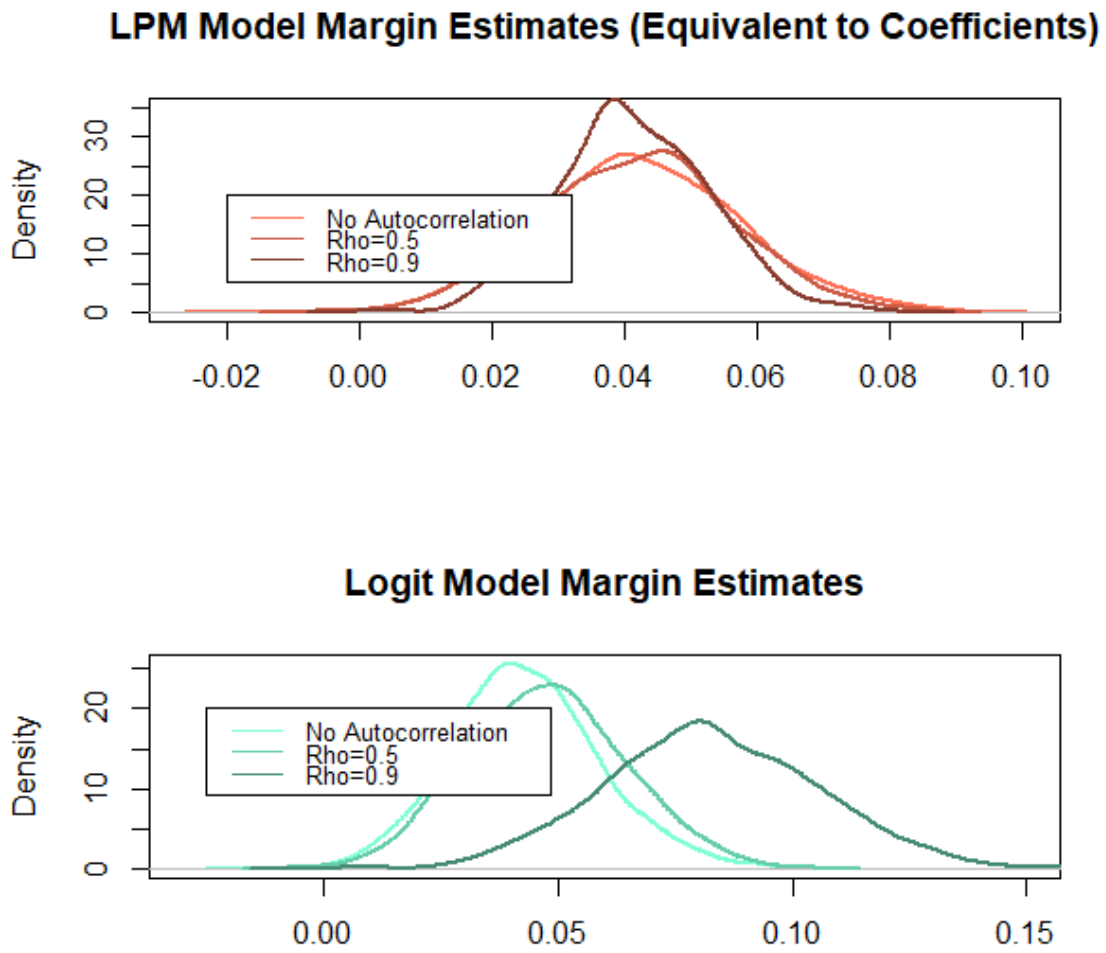

Figure 6: Simulated Distribution of X_1 Marginal Effects

Appendix III: Marginal Effects at the Tails (Case 1)

These plots show the distribution of marginal effects, where we take the marginal effect of a change in the covariate when X_1 is at the 95th percentile instead of at the mean. We can see even at the tail the results are fairly robust in case 1. We did look into tail phenomena in depth, but this is an interesting area for future research.

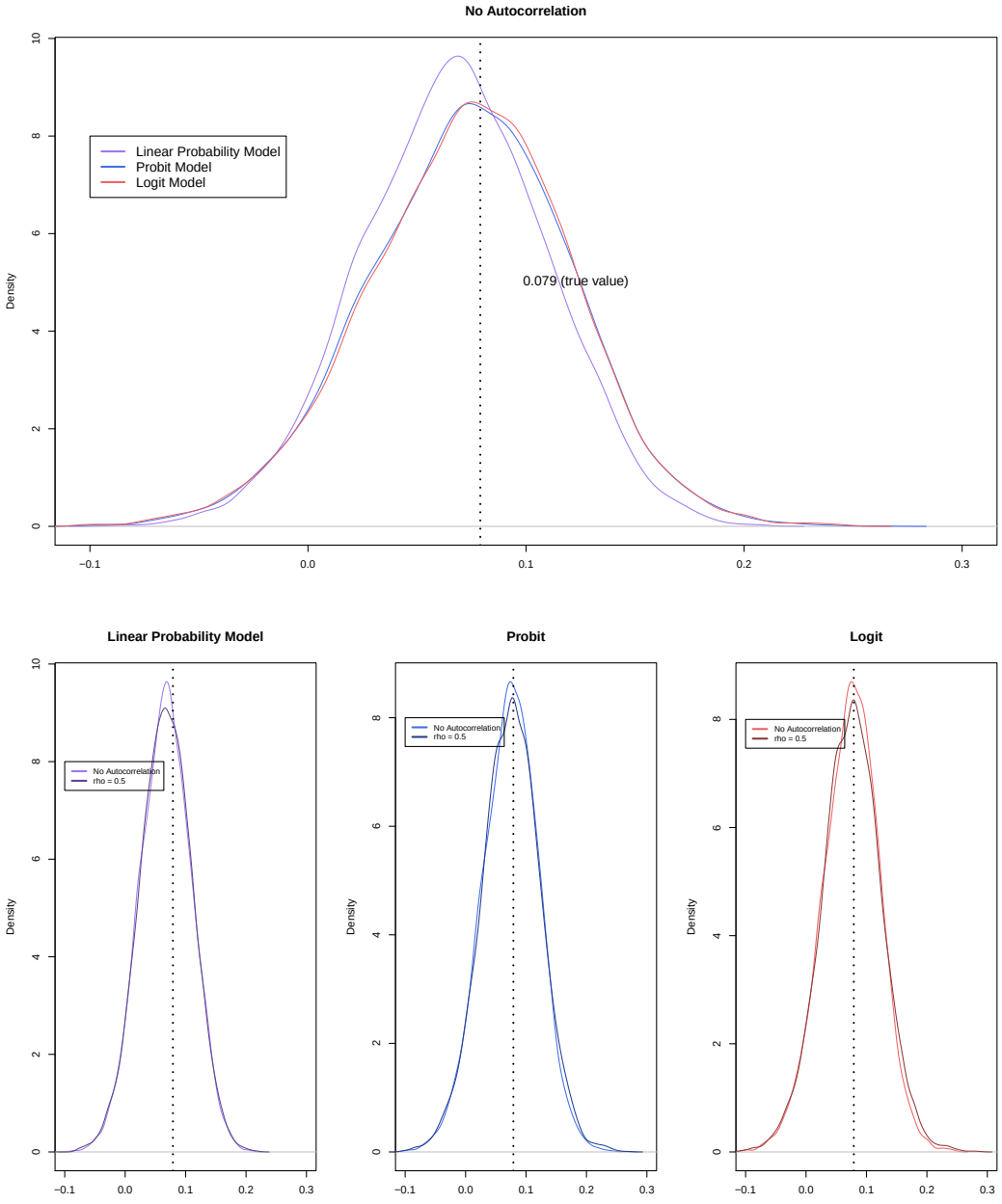

Figure 7: Distribution of X_1 Marginal Effects, at the 95th Percentile Value of X_1

References

Angrist, J., & Pischke, J. (2009). Mostly harmless econometrics: an empiricist's companion . Princeton University Press.

Aron-Dine, A., Einav, L. & Finkelstein, A. (2013). The RAND Health Insurance Experiment, Three Decades Later. The Journal of Economic Perspectives, 27(1), 197–222. https://doi.org/10.1257/jep.27.1.197

Bera, A., Jarque, C., & Lee, L. (1984). Testing the Normality Assumption in Limited Dependent Variable Models. *International Economic Review*, 25(3), 563-578. doi:10.2307/2526219

Davidson, R., & MacKinnon, J. (2004) Econometric Theory and Methods. New York: Oxford University Press.

Evans, W., Farrelly, M. & Montgomery, E. (1999). Do Workplace Smoking Bans Reduce Smoking? The American Economic Review, 89(4), 728–747. Available at: https://doi.org/10.1257/aer.89.4.728

Leeper, TJ. (2017). Interpreting regression results using average marginal effects with R's margins, Available at the comprehensive R Archive Network (CRAN)

Nunn, N. & Qian, N. (2014). US Food Aid and Civil Conflict. The American Economic Review, 104(6), 1630–1666. https://doi.org/10.1257/aer.104.6.1630

Wooldridge, J. (2010). Econometric analysis of cross section and panel data (2nd ed.). MIT Press.