Universidad de San Andrés
DEPARTAMENTO DE ECONOMÍA

YOUNG RESEARCHERS WORKING PAPER SERIES

EVALUATING EFFICIENCY GAINS IN THE LINEAR PROBABILITY MODEL

Tomás Pacheco (Universidad de San Andrés)

Primera versión: septiembre 2025
Esta versión: septiembre 2025

Young Researchers Working Paper Nº 18
Departamento de Economía Universidad de San Andrés

Vito Dumas 284, B1644BID, Victoria,
San Fernando, Buenos Aires, Argentina
Teléfono 7078-0400
Email: economia@udesa.edu.ar

Evaluating efficiency gains in the Linear Probability Model

Tomás Pacheco*

This version: September, 2025

Abstract

This paper evaluates the efficiency gains of the Adaptive Least Squares (ALS) estimator proposed by Romano and Wolf (2017) in the context of Linear Probability Models (LPM), where heteroskedasticity is inherent to the model. Using empirical applications and Monte Carlo simulations, we compare ALS to OLS and Probit estimators under three strategies for handling predicted probabilities outside the $(0, 1)$ interval: bounding, sigmoid transformation, and trimming. The results show that efficiency gains from ALS are not systematic and depend on the correction method, with the bounding approach yielding the most substantial improvements.

Keywords: efficiency; linear probability model; weighted least squares
JEL Classification: C01, C12, C50

*Department of Economics, Universidad de San Andres, Vito Dumas 284, B1644BID, Victoria, Buenos Aires, Argentina. Email: tpacheco@udesa.edu.ar

1 Introduction

In recent years, economic research has leaned towards empirical rather than theoretical work (Hamermesh, 2013; Angrist et al., 2017). In applied economic research, ordinary least squares (OLS) remains the default estimator for linear models, as it leads to consistent and unbiased estimates under mild regularity conditions. However, in real-world data, the assumption of homoskedasticity is frequently violated. Although heteroskedasticity does not bias OLS coefficient estimates, it leads to incorrect variance estimates and undermines the validity of hypothesis tests and confidence intervals.

The two main solutions to the heteroskedasticity problem in linear regression are the use of Weighted Least Squares (WLS, henceforth)—a particular case of the Generalized Least Squares (GLS) estimator that consists of estimating a transformed model using OLS—or the use of the well-known White's robust estimator (White, 1980). The fact that WLS requires modeling the variance of the error term, combined with the property that White's estimator provides a consistent estimate of the covariance matrix under heteroskedasticity of unknown form, has led researchers to favor the latter approach. Petersen (2008), for example, found that in finance panel data sets only 3% of studies use GLS to estimate random effects models. This suggests that approaches requiring explicit modeling of the error variance are widely avoided.

Romano and Wolf (2017) challenge the conventional wisdom by showing that even with a misspecified variance, WLS can achieve substantial efficiency improvements over OLS. They propose the Adaptive Least Squares estimator (ALS), which consists in performing a heteroskedasticity test on the linear model and, if the null hypothesis is rejected, estimating the model via WLS with a generic skedastic function and also using robust standard errors. In their paper, they show that this method works both theoretically and through simulations, and both suggest that modest modeling of heteroskedasticity need not be perfect to yield gains in precision.

The contribution of Romano and Wolf (2017) to the literature is to show that the Adaptive Least Squares (ALS) estimator can lead to efficiency gains. However, the more relevant question is how large those gains can be, and how they depend on the specification of the skedastic function. To address this gap, this paper leverages the Linear Probability Model (LPM), where heteroskedasticity arises not from the data, but from the structure of the model itself. This setting provides a natural laboratory

in which to evaluate the performance of WLS: the true variance function is known up to parameter estimates, yet researchers routinely ignore it in practice.

The LPM remains widely used for binary outcomes, despite the availability of nonlinear alternatives such as Probit, Tobit, or Logit, which some researchers argue are superior. There is ongoing debate about whether the LPM or nonlinear models are more appropriate in such contexts. For example, Angrist and Pischke (2009) argue that while nonlinear models may fit the conditional expectation function (CEF) more closely, marginal effects—often the primary interest—are similarly estimated in linear models. Moreover, computing average marginal effects in nonlinear settings requires additional assumptions and is more complex, especially when extended to panel data or instrumental variable settings. These complications make the LPM an attractive and accessible alternative.

In contrast, Giles (2012a,b) criticizes the LPM, offering three main arguments. First, based on Horrace and Oaxaca (2006), he argues that the LPM does not recover structural parameters of the underlying nonlinear model. Second, he claims LPM estimates are neither unbiased nor consistent. Third, he highlights that in the presence of misclassification error in the dependent variable, LPM parameters are unidentified. Steffen Pischke responds to these criticisms in Angrist and Pischke (2012), arguing that researchers are usually interested in marginal effects rather than structural parameters, and that linear models approximate these well. More fundamentally, he argues that one cannot assert the LPM is biased or inconsistent without knowing the true model form—a claim that is rarely defensible. Thus, he contends that the choice between a linear and nonlinear model is no less arbitrary than the choice between different nonlinear specifications.

This debate shows the relevance of the LPM in the actual state of the art. And, regarding to this, the research question addressed in this paper, is particularly relevant in light of publication bias. The LPM is frequently used to estimate treatment effects in randomized controlled trials (RCTs), where the outcome variable is binary. Many such studies remain unpublished because they fail to reject the null hypothesis of no effect. However, these results are only credible if the estimator used is efficient—that is, if the associated hypothesis tests have sufficient statistical power. If the estimator is inefficient, the failure to reject the null may reflect a lack of power rather than the absence of a true effect.

In this paper, I empirically and quantitatively assess Romano and Wolf's efficiency-gain claim within the LPM framework. I replicate multiple regressions using well-known datasets and a published paper to test whether ALS leads to efficiency gains. Finally, I test these gains using a Monte Carlo simulation.

The paper is organized as follows: Section 2 presents the ALS estimator and the inherent heteroskedasticity problem in the LPM; Section 3 presents the empirical applications; Section 4 presents the Monte Carlo simulation; Section 5 concludes.

2 ALS estimator in the LPM

Consider the following linear model:

$$y_i = x_i'\beta + u_i, \tag{1}$$

where all classical assumptions are met except for homoskedasticity. In this case, I will assume that the variance of the error term is not constant, i.e., $var(u_i) = f(u_i)$, where $f(\cdot)$ is an uknown function. Now, in this setup, the classical estimator of the variance of the OLS estimator, $S^2 = e'e/(n-K)$, where e is a $n \times 1$ vector of OLS residuals, is neither consistent nor unbiased. In this context, two possible solutions may arise. The first is the use of the heteroskedasticity-consistent covariance matrix estimator proposed by White (1980):

$$\hat{var}^{HC}(\hat{\beta}) = (X'X)^{-1}\left(\sum_{i=1}^{n}\hat{u}_i^2 x_i x_i'\right)(X'X)^{-1}, \tag{2}$$

where $\hat{u}_i = y_i - x_i'\beta$ is the i-th OLS residual. In this paper, the key result (Theorem 1) is that:

$$\left|\frac{1}{n}\sum_{i=1}^{n}\hat{u}_i^2 x_i x_i' - \frac{1}{n}\sum_{i=1}^{n}E(u_i^2 x_i x_i')\right| \xrightarrow{p} 0 \tag{3}$$

This is the original estimator proposed by White (1980), which has since been modified in several ways to improve its finite-sample properties (MacKinnon and White, 1985) [1].

[1]For an overview of the developments following White's seminal work, see MacKinnon (2012), which provides a clear and comprehensive discussion of both theoretical refinements and practical considerations.

With the homoskedasticity assumption not held, the second solution to the problem consists of using Weighted Least Squares (Aitken, 1936). To illustrate this idea, I will assume that the variance of the error term is as follows: $var(u_i) = x_i^2\sigma^2$. Again, the S^2 estimator is not valid in this context. To overcome this issue, I divide (1) by x_i:

$$\frac{y_i}{x_i} = \frac{x_i'\beta}{x_i} + \frac{u_i}{x_i}, \tag{4}$$

Note that in this model, the variance of the error term is:

$$var(u_i^*) = var\left(\frac{u_i}{x_i}\right) = \frac{1}{x_i^2}var(u_i) = \frac{\sigma^2 x_i^2}{x_i^2} = \sigma^2, \tag{5}$$

which means that the model is homoskedastic. Therefore, in this case, the OLS estimator is the most efficient, given that all of the assumptions required by the Gauss-Markov Theorem are met. As might be noticed, the key for this method to work is knowing the exact form (up to a constant) of the skedastic function. If this function is known by the researcher, then using WLS will be the best estimator among the class of linear unbiased estimators (Aitken, 1936). The issue in practice is that this function is hardly ever known by the researcher, which has led to its infrequent use.

Romano and Wolf (2017) propose to "resurrect" Weighted Least Squares, which is rarely used in practice due to the need to assume a specific functional form for heteroskedasticity. They argue and demonstrate that combining WLS with robust standard errors produces a consistent estimator, even when the heteroskedasticity function is misspecified. Moreover, this approach can yield substantial efficiency gains. Specifically, the authors introduce the Adaptive Least Squares (ALS) estimator, which involves testing for conditional homoskedasticity using the same conditional variance function that will be used to weight the data, applying traditional heteroskedasticity tests, such as Breusch and Pagan (1979)[2]. If the test fails to reject the null hypothesis, the preferred estimator is OLS with robust standard errors; otherwise, WLS with robust standard errors is used. The central idea is to employ WLS only when there is sufficient evidence of heteroskedasticity in the tested form.

In the previous example, I assumed that the model was heteroskedastic. However,

[2]The authors do not recommend using a more general test for conditional heteroskedasticity, such as White (1980), except when the parametric specification of the skedastic function matches the one the researcher intends to test.

this may not be the case with real-world data, which is why it is important to test for heteroskedasticity. Nevertheless, there is one case in which heteroskedasticity does not need to be tested, as it is inherent to the model. This is the case of the Linear Probability Model, which is given by:

$$p_i = \Pr(y_i = 1|x_i) = x_i'\beta + u_i, \tag{6}$$

Since y_i is a Bernoulli random variable, the variance of the error term is given by:

$$var(u_i) = p_i(1 - p_i) = x_i'\beta(1 - x_i'\beta), \tag{7}$$

As the variance depends on the values that the observations take, we can conclude that the error term is heteroskedastic. Therefore, to estimate (6) using WLS, I should first estimate it by OLS and compute $\hat{p}_i = x_i'\hat{\beta}$, and then weight the observations using the functional form of the heteroskedasticity:

$$\frac{y_i}{\sqrt{\hat{p}_i(1-\hat{p}_i)}} = \frac{x_i'\beta}{\sqrt{\hat{p}_i(1-\hat{p}_i)}} + \frac{u_i}{\sqrt{\hat{p}_i(1-\hat{p}_i)}}, \tag{8}$$

By doing this, the model will meet the homoskedasticity assumption. The issue that this strategy might face is that in the first step, the OLS estimate might not yield valid probabilities; that is, it might produce negative values or numbers greater than one. Thus, the application of the WLS estimator to the LPM is limited to the predicted probabilities that fall within the $(0, 1)$ interval (Wooldridge, 2010, 2015).

In this paper, I explore Romano and Wolf's claim regarding the efficiency of the ALS estimator in contexts where heteroskedasticity is present, using both real and simulated data. Ideally, I would like to use data from unpublished studies affected by publication bias—that is, studies whose results were not statistically significant and thus remained unpublished—and test whether these results become significant when applying a more efficient estimator. However, obtaining such data poses two challenges: first, unpublished research data is generally difficult to access online; and second, to our knowledge, there are very few papers that use linear probability models with robust standard errors and also provide their data and replication files publicly. For this reason, I use four datasets originally featured in published papers and also used by Wooldridge (2015, 2010) for empirical applications. Additionally, I present an application using a paper published in the American Economic Journal: Applied Economics,

a journal that makes all codes and data available for replication. Finally, I present a Monte Carlo simulation to illustrate how this estimator performs in the context of the Linear Probability Model.

The next section presents the empirical applications. As mentioned above, a common issue with the LPM is that it can produce predicted probabilities outside the $(0, 1)$ interval. To address this, I propose three alternative solutions: (i) replacing probabilities greater than 1 with 0.999 and those less than 0 with 0.001; (ii) applying a sigmoid transformation to the predicted probabilities; and (iii) iteratively dropping observations with predicted probabilities outside the $(0, 1)$ range.

3 Empirical applications

In this section, I use four different datasets featured in Wooldridge (2015) to assess whether WLS leads to efficiency gains in this context. I estimate each model using OLS with robust standard errors (from now on, OLS refers to OLS with robust standard errors) and WLS with robust standard errors (hereafter, ALS).

To measure and compare the efficiency of each method, I present the estimated coefficients along with the corresponding t-statistics and the ratio between the lengths of the 95% confidence intervals from the ALS and OLS estimations, following Romano and Wolf (2017). It is important to note that when both estimations use the same number of observations, this ratio is equivalent to the ratio of the standard errors. A ratio below 1 indicates that the ALS confidence interval is narrower than that of OLS, while a ratio above 1 implies the opposite.

The first dataset is from Mroz (1987), who uses data from the 1975 Michigan Panel Study of Income Dynamics (PSID) to test various assumptions in empirical models of female labor supply. In this case, I use the data to estimate a model explaining the determinants of labor force participation in 1975. The estimated model is:

$$\begin{aligned} inlf_i = \beta_0 + \beta_1 \text{nwifeinc}_i + \beta_2 \text{educ}_i + \beta_3 \text{exper}_i + \beta_4 \text{expersq}_i + \beta_5 \text{age}_i + \beta_6 \text{kidslt6}_i \\ + \beta_7 \text{kidsge6}_i + u_i, \end{aligned} \tag{9}$$

Variable definitions and summary statistics are provided in Table A1 in the Appendix.

The second dataset is a subset of the data used in Grogger (1991), who uses criminal and labor market information to estimate an economic model of crime. I use this data to model the probability of an individual being arrested in 1986 based on several covariates, including prior convictions, sentence length, total time spent in prison since age 18, employment experience, and physical characteristics. The estimated model is:

$$\begin{aligned} arr86_i = \beta_0 + \beta_1 \text{pcnv}_i + \beta_2 \text{avgsen}_i + \beta_3 \text{tottime}_i + \beta_4 \text{ptime86}_i + \beta_5 \text{qemp86}_i + \beta_6 \text{black}_i \\ + \beta_7 \text{hispan}_i + u_i, \end{aligned} \tag{10}$$

Variable definitions and summary statistics are presented in Table A2 in the Appendix.

The third dataset comes from Hunter and Walker (1996), who test the Cultural Affinity Hypothesis in the mortgage lending market. I use the data to estimate a credit model—that is, the probability of a loan being approved—based on whether the applicant is white, housing expenses as a share of income, other financial obligations, loan amount, and other relevant characteristics. The estimated model is:

$$\begin{aligned} approve_i = \beta_0 + \beta_1 \text{white}_i + \beta_2 \text{hrat}_i + \beta_3 \text{obrat}_i + \beta_4 \text{loanprc}_i + \beta_5 \text{unem}_i + \beta_6 \text{male}_i \\ + \beta_7 \text{married}_i + \beta_8 \text{dep}_i + \beta_9 \text{sch}_i + \beta_{10} \text{cosign}_i + \beta_{11} \text{chist}_i + u_i, \end{aligned} \tag{11}$$

Variable definitions and summary statistics are reported in Table A3 in the Appendix.

The fourth dataset, originally used in Abadie (2003), relates eligibility for a pension program to individual characteristics such as income and age. The estimated model is:

$$e401k_i = \beta_0 + \beta_1 \text{inc}_i + \beta_2 \text{age}_i + \beta_3 \text{agesq}_i + u_i, \tag{12}$$

Variable definitions and summary statistics are presented in Table A4 in the Appendix.

Finally, I replicate an empirical application from Augsburg et al. (2015), who conduct a randomized controlled trial (RCT) to evaluate the impact of microcredit in Bosnia and Herzegovina. The authors assess the effect of microcredit on several outcomes, including inventory ownership, self-employment income, business ownership, business activity in services or agriculture, and whether a business was started or closed in the past 14 months. I replicate the results from Table 3 of the paper, excluding four outcomes that are continuous variables and not compatible with the methods applied

in this paper.

The following subsections estimate each of the presented models using both OLS and ALS, in order to compare the results and assess whether efficiency gains are achieved. Each subsection also reports the results for the different approaches used to ensure that the probabilities predicted by the OLS model fall within the open unit interval $(0, 1)$.

3.1 Approach 1: Bounding Inconsistent Probabilities

The first strategy I implement is the one proposed by Wooldridge (2015), which consists of replacing predicted probabilities greater than 1 with 0.999 and negative probabilities with 0.001. Thus, the procedure is as follows:

1. Estimate the model by OLS.
2. Compute $\hat{p}_i = x_i'\hat{\beta}$.
3. Replace $\hat{p}_i$ with 0.999 if $\hat{p}_i > 1$ and with 0.001 if $\hat{p}_i < 0$. These modified probabilities are denoted as $\hat{p}_i^*$.
4. Compute $\hat{v}^*(x_i) = \hat{p}_i^* \, (1 - \hat{p}_i^*)$.
5. Transform the model and estimate using OLS with robust standard errors.

3.1.1 Dataset 1: labor supply

For the labor supply dataset, the results are presented in Table 1. The first important result worth mentioning is that all ratios of the lengths of the confidence intervals (or standard error ratios) are less than one, indicating that the ALS confidence intervals are narrower than those of OLS. Moreover, some coefficients exhibit considerable efficiency gains, such as years of experience (`expersq`), age (`age`), and the number of children under age 6 (`kidslt6`). This application, particularly in the last column, provides direct evidence of the efficiency gains achieved by the ALS estimator over OLS.

However, it is important to note that several estimated coefficients change noticeably. For example, the first coefficient estimated by OLS is -0.0034, while for ALS it is -0.0001. Despite this, variables for which the estimated coefficients remain similar—such as the number of children under age 6 (`kidslt6`)—still exhibit considerable efficiency gains.

Some changes in coefficients are not large in absolute terms but represent substantial percentage differences. In this case, fewer than 5% of the observations are modified, suggesting that the estimates are sensitive to even small adjustments. It is also noteworthy that one variable becomes statistically insignificant after applying the model transformation.

Table 1: OLS and ALS estimates for the labor supply dataset after bounding inconsistent probabilities

	OLS Coef	OLS t-stat	ALS Coef	ALS t-stat	Ratio: length of CI
nwifeinc	-0.0034**	-2.2330	-0.0001	-0.0592	0.7250
educ	0.0380***	5.2292	0.0240***	4.8723	0.6788
exper	0.0395***	6.7973	0.0459***	9.7354	0.8110
expersq	-0.0006***	-3.1384	-0.0008***	-5.6068	0.7622
age	-0.0161***	-6.7073	-0.0138***	-8.5443	0.6751
kidslt6	-0.2618***	-8.2374	-0.2169***	-11.8641	0.5752
kidsge6	0.0130	0.9615	0.0086	1.2090	0.5252
_cons	0.5855***	3.8455	0.5929***	5.7569	0.6764

Note: The White test statistic for heteroskedasticity is 121 (34 degrees of freedom). The R-squared from the OLS estimation is 0.264, while that from ALS is 0.957. The estimation was performed with 753 observations. A total of 17 observations with $p \geq 1$ (2.26%) and 16 with negative probabilities (2.12%) were replaced. No variable became significant, and 1 variable became non-significant.

3.1.2 Dataset 2: crime

For the crime dataset, the results are presented in Table 2. Again, it can be seen that the ratios of the confidence interval lengths are less than 1, indicating that the ALS intervals are narrower than those from OLS.

In this case, unlike the previous example, I observe few changes in the estimated coefficients. A possible explanation for this might be that only 0.29% of the observations were replaced. Despite the small number of modifications, the efficiency gains were large enough that one variable—total time in prison since age 18 (`tottime`)—became significant at the 5% level when it was previously insignificant. This result clearly illustrates the efficiency gains from using ALS.

Table 2: OLS and ALS estimates for the crime dataset after bounding inconsistent probabilities

	OLS Coef	OLS t-stat	ALS Coef	ALS t-stat	Ratio: length of CI
pcnv	-0.1521***	-7.9466	-0.1450***	-8.5104	0.8907
avgsen	0.0046	0.7774	0.0045	1.2002	0.6275
tottime	-0.0026	-0.6040	-0.0029**	-2.0847	0.3332
ptime86	-0.0237***	-8.0470	-0.0222***	-13.6319	0.5538
qemp86	-0.0385***	-7.0584	-0.0396***	-7.4805	0.9712
black	0.1698***	6.5806	0.1633***	6.5547	0.9655
hispan	0.0962***	4.5164	0.0863***	4.4249	0.9159
_cons	0.3804***	19.3242	0.3824***	20.0261	0.9701

Note: The White test statistic for heteroskedasticity is 327.3 (32 degrees of freedom). The R-squared from the OLS estimation is 0.068, while that from ALS is 0.298. The estimation was performed with 2,725 observations. A total of 0 observations with $p \geq 1$ (0.00%) and 8 with negative probabilities (0.29%) were replaced. One variable became significant, and no variable became non-significant.

3.1.3 Dataset 3: loan applications

The results for the loan applications dataset are shown in Table 3. In this case, the results differ from the previous two applications. First, it is important to note that around 11% of the observations were replaced, which may have contributed to changes in the estimated coefficients.

In terms of efficiency gains, the results are mixed. Half of the coefficients have an interval ratio less than 1, while the other half have a ratio greater than 1, indicating that the ALS confidence intervals are wider than those of OLS in some cases. Only one variable—more than 12 years of education (`sch`)—became statistically significant at the 10% level, while three variables lost statistical significance.

Still, this application is informative in showing that ALS does not always lead to efficiency gains in the Linear Probability Model. In some cases, the estimates may differ, and the confidence intervals may become wider.

Table 3: OLS and ALS estimates for the loan applications dataset after bounding inconsistent probabilities

	OLS Coef	OLS t-stat	ALS Coef	ALS t-stat	Ratio: length of CI
white	0.1288***	4.9796	0.1484***	5.6520	1.0152
hrat	0.0018	1.2495	0.0015	0.6340	1.6027
obrat	-0.0054***	-4.0810	-0.0010	-0.5909	1.3128
loanprc	-0.1473***	-3.8932	-0.0364	-0.9170	1.0481
unem	-0.0073**	-1.9662	-0.0038**	-2.1150	0.4888
male	-0.0041	-0.2147	0.0246	0.6694	1.9068
married	0.0458***	2.6584	-0.0206	-1.0193	1.1714
dep	-0.0068	-0.9889	0.0001	0.0153	0.9426
sch	0.0018	0.1022	-0.0217*	-1.9076	0.6625
cosign	0.0098	0.2469	0.0118	1.0975	0.2722
chist	0.1330***	5.4031	0.1404***	5.7650	0.9892
pubrec	-0.2419***	-5.6535	-0.2775***	-6.5065	0.9965
mortlat1	-0.0573	-0.8645	0.0556	1.1066	0.7589
mortlat2	-0.1137	-1.2488	-0.0963	-1.2801	0.8264
vr	-0.0314**	-2.1705	-0.0626*	-1.8387	2.3503
_cons	0.9367***	15.7729	0.7481***	11.9154	1.0571

Note: The White test statistic for heteroskedasticity is 289.75 (122 degrees of freedom). The R-squared from the OLS estimation is 0.166, while that from ALS is 0.981. The estimation was performed with 1,971 observations. A total of 213 observations with $p \geq 1$ (10.81%) and 0 with negative probabilities (0.00%) were replaced. One variable became significant, and three variables became non-significant.

3.1.4 Dataset 4: pensions

The results for the pensions dataset are presented in Table 4. This dataset differs from the others in terms of sample size. While the previous datasets had fewer than 3,000 observations, the pensions dataset includes 9,275 observations.

Despite the larger sample and the fact that only 0.25% of the observations were modified to fall within the open unit interval, the results show no evidence of efficiency gains in this case: all confidence interval ratios are greater than one—and in some cases, substantially so. Thus, in this application, using ALS does not lead to any efficiency improvement over OLS.

Table 4: OLS and ALS estimates for the pensions dataset after bounding inconsistent probabilities

	OLS Coef	OLS t-stat	ALS Coef	ALS t-stat	Ratio: length of CI
inc	0.0052***	23.7783	0.0019***	2.9857	2.9302
age	0.0298***	7.7459	0.0792***	2.8981	7.0968
agesq	-0.0003***	-7.7783	-0.0009***	-2.9266	7.1779
_cons	-0.4222***	-5.3582	-1.2899**	-2.3420	6.9894

Note: The White test statistic for heteroskedasticity is 480.08 (8 degrees of freedom). The R-squared from the OLS estimation is 0.078, while that from ALS is 0.538. The estimation was performed with 9,275 observations. A total of 26 observations with $p \geq 1$ (0.28%) and 0 with negative probabilities (0.00%) were replaced. No variable became significant, and no variable became non-significant.

3.1.5 Dataset 5: microcredit in Bosnia and Herzegovina

The results for the microcredit paper are presented in Table 5. In this case, the interpretation of the table differs slightly from the previous applications, as each row corresponds to a different dependent variable. In all cases, the regressor of interest is the treatment variable. The estimation includes additional control variables, which are not displayed in the table as they are not central to the analysis.

In terms of efficiency, all confidence interval ratios are less than one, indicating that there are efficiency gains from using ALS instead of OLS. Although these gains did not cause any treatment effects to become statistically significant, the results still demonstrate that ALS provides more precise estimates in this context.

Table 5: OLS and ALS estimates for the microcredit dataset after bounding inconsistent probabilities

	OLS		ALS		Ratio:	White test		R^2		Obs.	Replaced	
	Coef	t-stat	Coef	t-stat	CI len.	Stat	df	OLS	ALS	Total	$\hat{p}_i \geq 1$	$\hat{p}_i < 0$
ownership	0.0513**	2.5573	0.0368**	2.3232	0.7884	152.04	(140)	0.025	0.128	994	0	17
self-emp income	0.0602**	2.0505	0.0567*	1.9467	0.9929	131.04	(140)	0.038	0.725	994	0	0
business own	0.0584*	1.8755	0.0554*	1.7860	0.9962	191.11	(140)	0.073	0.582	994	0	0
business in serv	0.0312	1.2659	-0.0252	-0.5675	1.8049	184.87	(140)	0.050	0.321	994	0	8
business in agric	0.0350	1.2621	0.0220	1.5570	0.5098	168.71	(140)	0.060	0.277	994	0	9
started bus 14m	0.0210	0.9630	0.0144	0.7188	0.9175	142.40	(140)	0.012	0.139	994	0	1
closed bus 14m	-0.0168	-0.6354	-0.0218	-0.8286	0.9930	130.41	(140)	0.027	0.222	994	0	0

These five applications suggest that using WLS instead of OLS can lead to significant efficiency gains—but not always. For instance, the labor supply and crime datasets show that the gains can be large enough for some coefficients that were not statistically significant under OLS to become significant when estimated with ALS. In contrast, other datasets, such as the loan applications and the microcredit paper, suggest that while some coefficients benefit from increased efficiency, others may lose precision relative to the OLS estimates. Finally, the pension dataset illustrates a case where no efficiency gains are achieved at all.

Together, these results suggest that while efficiency gains are possible, they are not guaranteed. It is important to note that these conclusions are based on a specific treatment of inconsistent predicted probabilities, namely, bounding them to fall within the $(0, 1)$ interval. In the next section, I explore what happens when, instead of bounding, the predicted probabilities are transformed to lie within the valid range using the sigmoid function.

3.2 Approach 2: Sigmoid Transformation

The second alternative to address the issue that predicted probabilities should lie within the interval $(0, 1)$ consists of applying a transformation that maps them into this range. Let $\hat{p}_i$ be a probability predicted by the linear model estimated via OLS. The sigmoid transformation is defined as:

$$\hat{p}_i^{\sigma}(\hat{p}_i) = \frac{1}{1 + e^{-\hat{p}_i}} \tag{13}$$

It can be easily shown that this function has horizontal asymptotes at 0 and 1 as $\hat{p}_i \to -\infty$ and $\hat{p}_i \to +\infty$, respectively, which guarantees that the transformed values are valid probabilities. Figure A1 in the Appendix illustrates this transformation.

Thus, the steps to apply the ALS estimator are as follows:

1. Estimate the model by OLS.
2. Compute $\hat{p}_i = x_i'\hat{\beta}$.
3. Transform $\hat{p}_i$ to obtain $\hat{p}_i^{\sigma}$.
4. Compute $\hat{v}^{\sigma}(x_i) = \hat{p}_i^{\sigma}\,(1 - \hat{p}_i^{\sigma})$.
5. Transform the model and estimate using OLS with robust standard errors.

The next subsections show the results of this application.

3.2.1 Dataset 1: labor supply

The results are presented in Table 6. In this case, the results suggest that there are minimal efficiency gains, as 7 out of the 8 coefficients have a ratio below one. No variable changes its significance level, and the estimated coefficients remain the same.

Table 6: OLS and ALS estimates for the labor supply dataset with sigmoid-transformed inconsistent probabilities

	OLS Coef	OLS t-stat	ALS Coef	ALS t-stat	Ratio: length of CI
nwifeinc	-0.0034**	-2.2330	-0.0035**	-2.3018	0.9854
educ	0.0380***	5.2292	0.0383***	5.3990	0.9770
exper	0.0395***	6.7973	0.0394***	6.8176	0.9952
expersq	-0.0006***	-3.1384	-0.0006***	-3.1604	0.9935
age	-0.0161***	-6.7073	-0.0162***	-6.8517	0.9877
kidslt6	-0.2618***	-8.2374	-0.2640***	-8.2526	1.0064
kidsge6	0.0130	0.9615	0.0116	0.8743	0.9840
_cons	0.5855***	3.8455	0.5922***	3.9591	0.9824

Note: The White test statistic for heteroskedasticity is 121 (34 degrees of freedom). The R-squared from the OLS estimation is 0.264, while that from ALS is 0.696. The estimation was performed with 753 observations. No variable became significant, and no variable became non-significant.

3.2.2 Dataset 2: crime

The results are presented in Table 7. The same pattern is observed as in the previous dataset: all ratios are essentially equal to 1, indicating that using ALS instead of OLS with robust standard errors does not lead to any efficiency gains. The significance levels remain unchanged for all variables included in the model. It is also worth to note that the coefficients estimated by both estimators are very similar—unlike the case when probabilities outside the $(0, 1)$ range are imputed as 0.999 or 0.001.

Table 7: OLS and ALS estimates for the crime dataset with sigmoid-transformed inconsistent probabilities

	OLS Coef	OLS t-stat	ALS Coef	ALS t-stat	Ratio: length of CI
pcnv	-0.1521***	-7.9466	-0.1512***	-7.8737	1.0036
avgsen	0.0046	0.7774	0.0045	0.7477	1.0096
tottime	-0.0026	-0.6040	-0.0025	-0.5769	1.0117
ptime86	-0.0237***	-8.0470	-0.0236***	-7.9515	1.0087
qemp86	-0.0385***	-7.0584	-0.0383***	-7.0124	1.0028
black	0.1698***	6.5806	0.1705***	6.5948	1.0019
hispan	0.0962***	4.5164	0.0963***	4.5101	1.0029
_cons	0.3804***	19.3242	0.3796***	19.2567	1.0013

Note: The White test statistic for heteroskedasticity is 327.3 (32 degrees of freedom). The R-squared from the OLS estimation is 0.068, while that from ALS is 0.328. The estimation was performed with 2,725 observations. No variable became significant, and no variable became non-significant.

3.2.3 Dataset 3: loan applications

In the loan applications data, whose results are presented in Table 8, we find slightly different results compared to the previous cases. All estimated coefficients have a confidence interval length ratio below 1, indicating that there are efficiency gains from using the ALS estimator.

However, although the ratios are below 1, they remain close to one, suggesting that the efficiency gains are modest. In this case, no variable becomes statistically significant.

Table 8: OLS and ALS estimates for the loan applications dataset with sigmoid-transformed inconsistent probabilities

	OLS Coef	OLS t-stat	ALS Coef	ALS t-stat	Ratio: length of CI
white	0.1288***	4.9796	0.1276***	4.9903	0.9885
hrat	0.0018	1.2495	0.0018	1.2502	0.9717
obrat	-0.0054***	-4.0810	-0.0052***	-4.0444	0.9669
loanprc	-0.1473***	-3.8932	-0.1419***	-3.9083	0.9594
unem	-0.0073**	-1.9662	-0.0073**	-2.0345	0.9707
male	-0.0041	-0.2147	-0.0051	-0.2749	0.9691
married	0.0458***	2.6584	0.0449***	2.6862	0.9703
dep	-0.0068	-0.9889	-0.0066	-0.9834	0.9696
sch	0.0018	0.1022	0.0010	0.0574	0.9669
cosign	0.0098	0.2469	0.0110	0.2906	0.9596
chist	0.1330***	5.4031	0.1306***	5.3747	0.9867
pubrec	-0.2419***	-5.6535	-0.2422***	-5.6753	0.9973
mortlat1	-0.0573	-0.8645	-0.0856	-0.9072	0.0752
mortlat2	-0.1137	-1.2488	-0.1121	-1.2556	0.9805
vr	-0.0314**	-2.1705	-0.0315**	-2.2371	0.9730
_cons	0.9367***	15.7729	0.9314***	16.1559	0.9708

Note: The White test statistic for heteroskedasticity is 289.75 (122 degrees of freedom). The R-squared from the OLS estimation is 0.166, while that from ALS is 0.901. The estimations were performed with 1,975 observations. No variable became significant, and no variable became non-significant.

3.2.4 Dataset 4: pensions

The results of the pensions dataset are presented in Table 9. This case is very similar to the second one, in the sense that all confidence interval ratios are virtually equal to 1. This indicates that the length of the confidence intervals is the same for both OLS and ALS, meaning there are neither efficiency gains nor losses from using one estimator over the other.

Table 9: OLS and ALS estimates for the pensions dataset with sigmoid-transformed inconsistent probabilities

	OLS Coef	OLS t-stat	ALS Coef	ALS t-stat	Ratio: length of CI
inc	0.0052***	23.7783	0.0050***	22.6256	1.0150
age	0.0298***	7.7459	0.0302***	7.7807	1.0070
agesq	-0.0003***	-7.7783	-0.0003***	-7.8077	1.0074
_cons	-0.4222***	-5.3582	-0.4228***	-5.3337	1.0060

Note: The White test statistic for heteroskedasticity is 480.08 (8 degrees of freedom). The R-squared from the OLS estimation is 0.078, while that from ALS is 0.443. The estimation was performed with 9,275 observations. No variable became significant and no variable became non-significant.

3.2.5 Dataset 5: microcredit in Bosnia and Herzegovina

The results of this application is presented in Table 10. The results do not differ much from the previous cases, as all the ratios are very close to one, indicating that there are no difference between the OLS and ALS estimators.

Table 10: OLS and ALS estimates for the microcredit dataset with sigmoid-transformed inconsistent probabilities

	OLS		ALS		Ratio:	White test		R^2		ALS
	Coef	t-stat	Coef	t-stat	CI len.	Stat	df	OLS	ALS	Obs.
ownership	0.0513**	2.5573	0.0513**	2.5544	1.0010	152.04	(140)	0.025	0.142	994
self-emp income	0.0602**	2.0505	0.0596**	2.0377	0.9968	131.04	(140)	0.038	0.715	994
business own	0.0584*	1.8755	0.0580*	1.8640	1.0002	191.11	(140)	0.073	0.579	994
business in serv	0.0312	1.2659	0.0313	1.2702	1.0027	184.87	(140)	0.050	0.230	994
business in agric	0.0350	1.2621	0.0359	1.2913	1.0032	168.71	(140)	0.060	0.304	994
started bus 14m	0.0210	0.9630	0.0210	0.9638	1.0005	142.40	(140)	0.012	0.146	994
closed bus 14m	-0.0168	-0.6354	-0.0169	-0.6375	1.0015	130.41	(140)	0.027	0.239	994

Using the sigmoid transformation, the results show that only minimal efficiency gains can be achieved by using ALS instead of OLS. Across all five applications, the ratios of the lengths of the confidence intervals are equal to or very close to one, indicating that any efficiency gains are small, if present at all. In the previous section, where I bounded predicted probabilities, both the labor supply and crime datasets showed substantial efficiency gains for most coefficients. In contrast, those gains are now considerably reduced or disappear entirely under the sigmoid transformation.

Regarding the loan applications and microcredit datasets, the confidence interval ratios for coefficients where ALS was previously more efficient than OLS are now close to one. Similarly, for coefficients where OLS was previously more efficient, the ratios also shrink to one. This same pattern is observed in the pensions dataset: ratios that were as high as 2 or 7 in the bounding approach are now nearly 1 after applying the sigmoid transformation.

To sum up, I do not observe any confidence interval ratio significantly greater than one, suggesting that ALS with the sigmoid transformation is rarely less efficient than OLS. In the worst-case scenario, both estimators yield confidence intervals of similar length. This result contrasts with the previous section, where bounding led to substantially larger efficiency gains, although not in all cases. In the next section, I apply the trimming estimator to address predicted probabilities that fall outside the $(0, 1)$ interval.

3.3 Approach 3: Trimming Estimator

The article by Horrace and Oaxaca (2006) shows that estimating the LPM using OLS can lead to inconsistent estimates[3]. Toward the end of the article, the authors propose an estimator—the *trimming estimator*—which consists of removing from the dataset those observations that yield predicted probabilities outside the valid range. This was merely a comment suggested for future research. A work in progress article by Chen, Martin, and Wooldridge (2025) shows that iteratively removing the observations that produce predicted probabilities outside the interval $(0, 1)$ is equivalent to estimating the *ramp function*. They show that this is a consistent and asymptotically normal estimator. For this reason, in this alternative we apply the *trimming estimator* and then correct the inference using WLS.

[3]Assuming the true model follows a "ramp function".

The estimation steps are as follows:

1. Estimate the model by OLS.
2. Compute $\hat{p}_i = x_i'\hat{\beta}$
   - 2.1 If any $\hat{p}_i \notin (0, 1)$, remove those observations.
   - 2.2 Re-estimate the model by OLS.
   - 2.3 Recompute $\hat{p}_i = x_i'\hat{\beta}$.
   - 2.4 Repeat steps 2.1–2.3 until all predicted probabilities fall within the open unit interval.
3. Compute: $\hat{v}(x_i) = \hat{p}_i(1 - \hat{p}_i)$.
4. Transform the model and estimate it using OLS with robust standard errors.

3.3.1 Dataset 1: labor supply

The results for the labor supply dataset are presented in Table 11. Out of the 8 coefficients, 5 have a ratio below one. I find that one variable—the number of children between ages 6 and 18 (`kidsage6`)—becomes statistically significant at the 10% level. Additionally, family income (`inlf`) becomes statistically insignificant at any conventional level, as also occurred after bounding inconsistent probabilities in Subsection 3.1.

In this case, 45 observations were excluded from the WLS estimation, meaning that 94% of the original sample was retained.

3.3.2 Dataset 2: crime

Table 12 shows the results for the crime dataset. In this case, all confidence interval ratios are below 1, suggesting efficiency gains from using ALS. For some coefficients, the gains are substantial. For example, the t-statistic for total time in prison since age 18 more than doubles in magnitude. Only 8 observations out of the full sample of 2,725 were dropped. Overall, this case demonstrates notable efficiency gains.

Table 11: OLS and ALS estimates for the labor supply dataset using the trimming estimator

	OLS Coef	OLS t-stat	ALS Coef	ALS t-stat	Ratio: length of CI
nwifeinc	-0.0034**	-2.2330	-0.0032	-1.4982	1.3954
educ	0.0380***	5.2292	0.0433***	6.9753	0.8540
exper	0.0395***	6.7973	0.0413***	6.9527	1.0214
expersq	-0.0006***	-3.1384	-0.0007***	-3.1215	1.1233
age	-0.0161***	-6.7073	-0.0170***	-8.4515	0.8385
kidslt6	-0.2618***	-8.2374	-0.2856***	-9.3938	0.9568
kidsge6	0.0130	0.9615	0.0156*	1.7728	0.6499
_cons	0.5855***	3.8455	0.5571***	4.8254	0.7584

Note: The White test statistic for heteroskedasticity is 121 (34 degrees of freedom). The R-squared from the OLS estimation is 0.264, while that from ALS is 0.825. The OLS estimation was performed with 753 observations and the ALS estimation with 708 (45 observations were dropped). One variable became significant, and one variable became non-significant.

Table 12: OLS and ALS estimates for the crime dataset using the trimming estimator

	OLS Coef	OLS t-stat	ALS Coef	ALS t-stat	Ratio: length of CI
pcnv	-0.1521***	-7.9466	-0.1582***	-8.6066	0.9608
avgsen	0.0046	0.7774	0.0046	1.2150	0.6417
tottime	-0.0026	-0.6040	-0.0028	-1.5983	0.4084
ptime86	-0.0237***	-8.0470	-0.0248***	-16.2369	0.5187
qemp86	-0.0385***	-7.0584	-0.0392***	-7.3915	0.9728
black	0.1698***	6.5806	0.1638***	6.5936	0.9632
hispan	0.0962***	4.5164	0.0884***	4.5442	0.9139
_cons	0.3804***	19.3242	0.3870***	20.1210	0.9771

Note: The White test statistic for heteroskedasticity is 327.3 (32 degrees of freedom). The R-squared from the OLS estimation is 0.068, while that from ALS is 0.300. The OLS estimation was performed with 2,725 observations and the ALS estimation with 2,717 (8 observations were dropped). No variable became significant, and no variable became non-significant.

3.3.3 Dataset 3: loan applications

The results are shown in Table 13. In this case, the results are not unidirectional: out of the 16 coefficients, 10 have a ratio greater than one, while the remaining ones have a ratio below one. There are no changes in the significance of the coefficients. It is important to note that 592 observations were excluded in this case, which represents 30% of the original sample.

Table 13: OLS and ALS estimates for the loan applications dataset using the trimming estimator

	OLS Coef	OLS t-stat	ALS Coef	ALS t-stat	Ratio: length of CI
white	0.1288***	4.9796	0.1296***	4.1928	1.1954
hrat	0.0018	1.2495	0.0001	0.0425	2.3020
obrat	-0.0054***	-4.0810	-0.0059***	-3.0361	1.4604
loanprc	-0.1473***	-3.8932	-0.1806**	-2.4581	1.9429
unem	-0.0073**	-1.9662	-0.0084**	-2.2578	1.0042
male	-0.0041	-0.2147	-0.0071	-0.5515	0.6679
married	0.0458***	2.6584	0.0398**	2.3663	0.9754
dep	-0.0068	-0.9889	-0.0128	-1.3886	1.3349
sch	0.0018	0.1022	0.0200	1.4451	0.8055
cosign	0.0098	0.2469	-0.0997	-0.7378	3.4134
chist	0.1330***	5.4031	0.2034***	4.3494	1.9003
pubrec	-0.2419***	-5.6535	-0.2271***	-5.1325	1.0341
mortlat1	-0.0573	-0.8645	-0.0522	-1.0414	0.7575
mortlat2	-0.1137	-1.2488	-0.1050	-1.4481	0.7962
vr	-0.0314**	-2.1705	-0.0366**	-2.2244	1.1357
_cons	0.9367***	15.7729	0.9836***	8.5608	1.9352

Note: The White test statistic for heteroskedasticity is 289.75 (122 degrees of freedom). The R-squared from the OLS estimation is 0.166, while that from ALS is 0.981. The OLS estimation was performed with 1,971 observations and the ALS estimation with 1,397 (592 observations were dropped). No variable became significant, and no variable became non-significant.

3.3.4 Dataset 4: pensions

The results are presented in Table 14. Very few observations were excluded, and no clear efficiency gains are observed. For the first coefficient, the t-statistic is reduced

by more than half, while for the other coefficients, the t-statistics remain essentially unchanged. This suggests that there are no efficiency gains in this application.

Table 14: OLS and ALS estimates for the pensions dataset using the trimming estimator

	OLS Coef	OLS t-stat	ALS Coef	ALS t-stat	Ratio: length of CI
inc	0.0052***	23.7783	0.0042***	8.1784	2.3651
age	0.0298***	7.7459	0.0286***	7.4670	0.9934
agesq	-0.0003***	-7.7783	-0.0003***	-7.6024	0.9888
_cons	-0.4222***	-5.3582	-0.3665***	-4.5033	1.0327

Note: The White test statistic for heteroskedasticity is 480.08 (8 degrees of freedom). The R-squared from the OLS estimation is 0.078, while that from ALS is 0.426. The OLS estimation was performed with 9,275 observations and the ALS estimation with 9,242 (33 observations were dropped). No variable became significant, and no variable became non-significant.

3.3.5 Dataset 5: microcredit in Bosnia and Herzegovina

The results for the microcredit dataset are presented in Table 15. In all but one case, the confidence interval ratio is below one, although the values are close to 1. This essentially indicates that, for the majority of outcomes, small efficiency gains are observed.

Table 15: OLS and ALS estimates for the microcredit dataset using the trimming estimator

	OLS		ALS		Ratio:	White test		R^2		ALS
	Coef	t-stat	Coef	t-stat	CI len.	Stat	df	OLS	ALS	Obs.
ownership	0.0513**	2.5573	0.0430**	2.0326	1.0535	152.04	(140)	0.025	0.145	977
self-emp income	0.0602**	2.0505	0.0567*	1.9467	0.9929	131.04	(140)	0.038	0.725	994
business own	0.0584*	1.8755	0.0554*	1.7860	0.9962	191.11	(140)	0.073	0.582	994
business in serv	0.0312	1.2659	0.0268	1.1723	0.9278	184.87	(140)	0.050	0.198	980
business in agric	0.0350	1.2621	0.0162	0.6958	0.8374	168.71	(140)	0.060	0.277	984
started bus 14m	0.0210	0.9630	0.0159	0.7852	0.9289	142.40	(140)	0.012	0.139	993
closed bus 14m	-0.0168	-0.6354	-0.0218	-0.8286	0.9930	130.41	(140)	0.027	0.222	994

This section of the paper presents, using five different datasets, multiple applications of the ALS estimator and its comparison with the OLS estimator, employing three different strategies to address a key issue of OLS when estimating a Linear Probability Model (LPM): the predicted probabilities may fall outside the $(0, 1)$ interval, which makes it impossible to apply the WLS estimator. The three strategies considered are: (i) bounding inconsistent probabilities by imputing 0.999 for predicted probabilities above 1 and 0.001 for those below 0; (ii) applying the sigmoid transformation to all predicted probabilities, which guarantees they fall within the open unit interval; and (iii) using a trimming estimator, which consists of iteratively dropping observations with predicted probabilities that fall outside the valid range.

The results from this section show that there are no uniform conclusions—that is, ALS does not always yield efficiency gains over OLS. In some datasets, I find that, regardless of the correction strategy used, ALS consistently leads to greater efficiency. In others, the opposite occurs: no efficiency gains are observed, even after applying any of the proposed strategies. It is worth noting that the first four datasets may suffer from structural issues, such as the potential endogeneity of some regressors, which could influence the results. In contrast, the dataset from Augsburg et al. (2015) provides a setting where the treatment variable is exogenous due to random assignment, ensuring that the OLS estimator is unbiased and consistent. In this particular application, the results suggest that efficiency gains are possible—but not guaranteed—and that such gains are relatively modest. Moreover, these gains were not sufficient to render any of the treatment effects statistically significant.

Although the empirical applications using real-world data yielded several interesting findings, it is important to acknowledge that in none of these settings (nor in real life) is the data-generating process known to the researcher. To complement this analysis, the next section presents the results of applying the ALS estimator to a model where the data-generating process is known. I implement this using a Monte-Carlo simulation.

4 Monte-Carlo simulation

In this section, I present a Monte-Carlo simulation that allows us to evaluate the properties of using ALS in the linear probability model. In this case, I am also interested in comparing its performance with the probit model. For the simulation, consider the

following latent model:

$$y_i^* = x_i\beta + u_i \tag{14}$$

$$y_i = 1[y_i^* > 1] \tag{15}$$

with $u_i \sim \mathcal{N}(0,1)$ and $x_i \sim \mathcal{N}(0,1)$. Since this is, by construction, a probit model, I compare the performance of ALS to that of probit. To do so, I compare the OLS and ALS estimates to the marginal effect at the mean from the probit model.

The steps for conducting the Monte Carlo simulation are as follows:

1. Generate a dataset with n observations.
2. Define the variables x_i and u_i.
3. Specify the latent model.
4. Estimate equation (14) using OLS with robust standard errors.
5. Estimate equation (14) using ALS (applying any of the three approaches to overcome with probabilities falling outside the $(0,1)$ interval).
6. Estimate equation (14) using a probit model and compute the marginal effect at the mean. Standard errors are calculated using the delta method.
7. Store the estimated coefficients and the lengths of the confidence intervals.

Table 16 shows the results of the simulation in which predicted probabilities falling outside the open unit interval were replaced with 0.999 and 0.001. Each version of the simulation was run with 1,000 repetitions, varying both the sample size and the true value of β. Simulations were conducted with 25, 50, 100, and 500 observations, and with the following values of the true parameter: $\beta \in \{0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8\}$. The motivation for varying the values of β is that smaller values are associated with a lower likelihood of predicted probabilities falling outside the $(0,1)$ interval.

Each column in the table presents averages. The first three columns report the average estimated coefficients from OLS, ALS, and Probit (marginal effect at the mean), respectively. The next two columns present the average ratio between the length of the confidence interval from ALS and that from OLS, and between ALS and Probit. The last column shows the average number of observations with predicted values outside the $(0,1)$ interval.

Table 16: Monte Carlo results — Predicted probabilities bounded to $(0, 1)$

		Coef OLS	Coef WLS	Coef Probit	Ratio WLS/OLS	Ratio WLS/Prob	Obs out of range
$\beta = 0.1$							
	$n = 25$	0.0422	0.0409	0.0468	0.9183	0.9297	0.109
	$n = 50$	0.0418	0.0411	0.0441	0.9659	0.9557	0.028
	$n = 100$	0.0402	0.0399	0.0414	0.9864	0.9700	0.008
	$n = 500$	0.0396	0.0396	0.0400	0.9991	0.9837	0.000
$\beta = 0.2$							
	$n = 25$	0.0811	0.0746	0.0921	0.8861	0.8741	0.191
	$n = 50$	0.0799	0.0766	0.0854	0.9282	0.8942	0.078
	$n = 100$	0.0790	0.0773	0.0826	0.9629	0.9166	0.042
	$n = 500$	0.0777	0.0776	0.0796	0.9912	0.9357	0.004
$\beta = 0.3$							
	$n = 25$	0.1178	0.1074	0.1364	0.8402	0.7947	0.318
	$n = 50$	0.1161	0.1103	0.1269	0.8644	0.7945	0.197
	$n = 100$	0.1155	0.1105	0.1236	0.8988	0.8120	0.179
	$n = 500$	0.1141	0.1119	0.1196	0.9444	0.8360	0.077
$\beta = 0.4$							
	$n = 25$	0.1517	0.1352	0.1821	0.7799	0.6985	0.524
	$n = 50$	0.1498	0.1373	0.1693	0.7822	0.6789	0.479
	$n = 100$	0.1486	0.1354	0.1640	0.7967	0.6715	0.557
	$n = 500$	0.1476	0.1353	0.1595	0.8368	0.6831	0.738
$\beta = 0.5$							
	$n = 25$	0.1822	0.1588	0.2302	0.7180	0.6074	0.783
	$n = 50$	0.1789	0.1608	0.2105	0.6836	0.5523	0.863
	$n = 100$	0.1795	0.1565	0.2054	0.6777	0.5243	1.243
	$n = 500$	0.1775	0.1493	0.1989	0.8422	0.6200	3.174
$\beta = 0.6$							
	$n = 25$	0.2088	0.1773	0.2731	0.6542	0.5147	1.091
	$n = 50$	0.2060	0.1780	0.2520	0.5942	0.4412	1.428
	$n = 100$	0.2065	0.1720	0.2468	0.5673	0.3958	2.285
	$n = 500$	0.2046	0.1647	0.2392	0.9146	0.5993	8.006
$\beta = 0.7$							
	$n = 25$	0.2323	0.1937	0.3100	0.5918	0.4303	1.423
	$n = 50$	0.2298	0.1911	0.2933	0.5186	0.3454	2.117
	$n = 100$	0.2309	0.1875	0.2886	0.4874	0.2983	3.633
	$n = 500$	0.2284	0.1811	0.2794	0.9291	0.5400	14.906
$\beta = 0.8$							
	$n = 25$	0.2542	0.2069	0.3603	0.5307	0.3503	1.782
	$n = 50$	0.2520	0.2053	0.3387	0.4661	0.2736	2.866
	$n = 100$	0.2519	0.2018	0.3305	0.4870	0.2612	5.210
	$n = 500$	0.2488	0.1959	0.3192	0.8549	0.4403	22.586

Note: Each simulation was conducted with 1,000 repetitions.

At first glance, the results in Table 16 show considerable efficiency gains from using ALS over OLS for all values of β and sample sizes. For all true values of β, efficiency gains diminish as the sample size increases. This phenomenon can be explained by the fact that increasing the sample size naturally reduces the variance of the estimator.

The most interesting result is that the efficiency gain becomes larger as the true value of β increases. This is intuitive: when the parameter is large, there is more room for efficiency gains. In this case, the number of observations with predicted probabilities outside the $(0, 1)$ range increases with the value of β. This justifies exploring multiple values of the parameter, as it allows us to examine how the estimator performs when a growing proportion of predictions are inconsistent with the definition of a probability.

Table 17 presents the simulation results using the sigmoid transformation to address the issue of predicted probabilities outside the $(0, 1)$ interval. The results are very similar to those found in the empirical applications. This transformation causes ALS estimates to closely resemble the OLS estimates, both in terms of point estimates and confidence interval lengths. This suggests that ALS can lead to efficiency gains, but they tend to be modest. As in the previous table, we observe that as β increases, the scope for efficiency gains also increases, and these gains diminish with larger sample sizes—again, due to the inverse relationship between sample size and estimator variance.

Finally, Table 18 shows the results of the Monte Carlo simulation using the trimming estimator. In this case, the results are less intuitive, as no clear pattern emerges. First, we observe that as β increases, the number of observations deleted by the trimming procedure also increases. Second, there do not appear to be systematic efficiency gains from using this method. Only when the true parameter is 0.10 do we observe efficiency gains for all sample sizes—i.e., the confidence interval length ratio is below one. It appears that as β increases, ALS becomes less efficient than OLS. One possible explanation is that larger values of β lead the trimming estimator to drop more observations, which increases the variance of the ALS estimator, given the inverse relationship between sample size and variance.

In this section, I presented the results of the Monte Carlo simulation, which aims to assess whether there are efficiency gains from using ALS in the context of the Linear Probability Model. Once again, the results are not unidirectional: in two of the three strategies used to correct for inconsistent predicted probabilities, there is evidence that the approach proposed by Romano and Wolf (2017) may lead to efficiency gains.

When I imputed constant values (0.999 and 0.001) to predicted probabilities falling outside the $(0, 1)$ interval, the results showed substantial efficiency gains, which increased with the magnitude of the true parameter of interest. In the case where I applied the sigmoid transformation, there were also efficiency gains, though they were relatively modest in size. Lastly, when using the trimming estimator (Chen, Martin, and Wooldridge, 2025), the previous findings did not hold. In this case, no systematic efficiency gains were observed.

Table 17: Monte Carlo Results — predicted probabilities transformed via sigmoid function

		Coef OLS	Coef WLS	Coef Probit	Ratio WLS/OLS	Ratio WLS/Prob	Obs out of range
$\beta = 0.1$							
	$n = 25$	0.0422	0.0419	0.0468	0.9971	0.9856	0.109
	$n = 50$	0.0418	0.0417	0.0441	0.9995	0.9801	0.028
	$n = 100$	0.0402	0.0402	0.0414	0.9997	0.9828	0.008
	$n = 500$	0.0396	0.0396	0.0400	1.0000	0.9846	0.000
$\beta = 0.2$							
	$n = 25$	0.0811	0.0809	0.0921	0.9952	0.9471	0.191
	$n = 50$	0.0799	0.0797	0.0854	0.9987	0.9443	0.078
	$n = 100$	0.0790	0.0789	0.0826	0.9993	0.9437	0.042
	$n = 500$	0.0777	0.0777	0.0796	0.9999	0.9432	0.004
$\beta = 0.3$							
	$n = 25$	0.1178	0.1175	0.1364	0.9936	0.8918	0.318
	$n = 50$	0.1161	0.1159	0.1269	0.9970	0.8837	0.197
	$n = 100$	0.1155	0.1154	0.1236	0.9977	0.8815	0.179
	$n = 500$	0.1141	0.1140	0.1196	0.9994	0.8802	0.077
$\beta = 0.4$							
	$n = 25$	0.1517	0.1512	0.1821	0.9915	0.8235	0.524
	$n = 50$	0.1498	0.1494	0.1693	0.9948	0.8116	0.479
	$n = 100$	0.1486	0.1484	0.1640	0.9957	0.8072	0.557
	$n = 500$	0.1476	0.1474	0.1595	0.9979	0.8034	0.738
$\beta = 0.5$							
	$n = 25$	0.1822	0.1812	0.2302	0.9898	0.7632	0.783
	$n = 50$	0.1789	0.1783	0.2105	0.9919	0.7343	0.863
	$n = 100$	0.1795	0.1790	0.2054	0.9940	0.7264	1.243
	$n = 500$	0.1775	0.1771	0.1989	0.9957	0.7227	3.174
$\beta = 0.6$							
	$n = 25$	0.2088	0.2074	0.2731	0.9875	0.6921	1.091
	$n = 50$	0.2060	0.2052	0.2520	0.9887	0.6581	1.428
	$n = 100$	0.2065	0.2057	0.2468	0.9917	0.6455	2.285
	$n = 500$	0.2046	0.2039	0.2392	0.9933	0.6403	8.006
$\beta = 0.7$							
	$n = 25$	0.2323	0.2307	0.3150	0.9853	0.6259	1.423
	$n = 50$	0.2298	0.2285	0.2933	0.9865	0.5848	2.117
	$n = 100$	0.2309	0.2297	0.2886	0.9893	0.5698	3.633
	$n = 500$	0.2284	0.2272	0.2794	0.9911	0.5642	14.906
$\beta = 0.8$							
	$n = 25$	0.2542	0.2522	0.3603	0.9833	0.5583	1.782
	$n = 50$	0.2520	0.2502	0.3387	0.9845	0.5156	2.866
	$n = 100$	0.2519	0.2502	0.3305	0.9875	0.5025	5.210
	$n = 500$	0.2488	0.2472	0.3192	0.9898	0.4983	22.586

Note: Each simulation was conducted with 1,000 repetitions.

Table 18: Monte Carlo Results — Predicted probabilities trimmed outside $(0, 1)$ interval

		Coef OLS	Coef WLS	Coef Probit	Ratio WLS/OLS	Ratio WLS/Prob	Deleted obs
$\beta = 0.1$							
	$n = 25$	0.0422	0.0438	0.0468	0.9742	0.9642	0.182
	$n = 50$	0.0418	0.0421	0.0441	0.9759	0.9618	0.033
	$n = 100$	0.0402	0.0400	0.0414	0.9921	0.9768	0.008
	$n = 500$	0.0396	0.0396	0.0400	0.9991	0.9837	0.000
$\beta = 0.2$							
	$n = 25$	0.0811	0.0884	0.0921	1.0042	0.9559	0.334
	$n = 50$	0.0799	0.0786	0.0854	0.9645	0.9183	0.090
	$n = 100$	0.0790	0.0781	0.0826	0.9808	0.9297	0.053
	$n = 500$	0.0777	0.0776	0.0796	0.9936	0.9377	0.004
$\beta = 0.3$							
	$n = 25$	0.1178	0.1319	0.1364	1.0267	0.9025	0.592
	$n = 50$	0.1161	0.1159	0.1269	0.9367	0.8410	0.247
	$n = 100$	0.1155	0.1125	0.1236	0.9722	0.8639	0.227
	$n = 500$	0.1141	0.1126	0.1196	0.9821	0.8666	0.079
$\beta = 0.4$							
	$n = 25$	0.1517	0.1703	0.1821	1.0620	0.8475	0.913
	$n = 50$	0.1498	0.1511	0.1693	0.9327	0.7665	0.669
	$n = 100$	0.1486	0.1454	0.1640	0.9358	0.7653	0.730
	$n = 500$	0.1476	0.1434	0.1595	0.9722	0.7871	0.804
$\beta = 0.5$							
	$n = 25$	0.1822	0.2112	0.2302	1.2102	0.8575	1.396
	$n = 50$	0.1789	0.1814	0.2105	0.9574	0.7008	1.261
	$n = 100$	0.1795	0.1781	0.2054	0.9330	0.6828	1.726
	$n = 500$	0.1775	0.1697	0.1989	1.0171	0.7417	3.676
$\beta = 0.6$							
	$n = 25$	0.2088	0.2493	0.2731	1.3189	0.8164	1.955
	$n = 50$	0.2060	0.2131	0.2520	0.9898	0.6429	2.230
	$n = 100$	0.2065	0.2089	0.2468	0.9594	0.6218	3.401
	$n = 500$	0.2046	0.1942	0.2392	1.1548	0.7457	10.388
$\beta = 0.7$							
	$n = 25$	0.2323	0.2847	0.3150	1.4673	0.7979	2.598
	$n = 50$	0.2298	0.2437	0.2933	1.0433	0.5951	3.392
	$n = 100$	0.2309	0.2393	0.2886	1.0073	0.5709	5.652
	$n = 500$	0.2284	0.2254	0.2794	1.2440	0.7083	21.343
$\beta = 0.8$							
	$n = 25$	0.2542	0.3207	0.3603	1.5455	0.7406	3.230
	$n = 50$	0.2520	0.2819	0.3387	1.1472	0.5597	4.980
	$n = 100$	0.2519	0.2714	0.3305	1.1391	0.5608	8.735
	$n = 500$	0.2488	0.2499	0.3192	1.5124	0.7603	34.974

Note: Each simulation was conducted with 1,000 repetitions.

5 Final remarks

In this paper, I evaluated the extent to which the Adaptive Least Squares (ALS) estimator, proposed by Romano and Wolf (2017), leads to efficiency gains in the estimation of the Linear Probability Model (LPM). This setting is particularly interesting because heteroskedasticity is inherent to the model—it arises from the functional form itself, rather than from the structure of the data. As demonstrated in the paper, the LPM is always heteroskedastic by construction.

Leveraging this feature, I assessed whether the ALS estimator—which involves using Weighted Least Squares (WLS) combined with robust standard errors—improves efficiency relative to the standard approach of using Ordinary Least Squares (OLS) with robust standard errors (when clustering is not required). I evaluated this question through two strategies: a set of empirical applications using real-world data, and a Monte Carlo simulation study. Since WLS requires all predicted probabilities to lie within the open unit interval $(0, 1)$, I implemented three different approaches to enforce this condition: (i) replacing negative predicted probabilities with 0.001 and those greater than one with 0.999; (ii) applying the sigmoid transformation to all predicted probabilities; and (iii) using a trimming estimator, which sequentially removes observations that yield probabilities outside the valid range.

The results from the empirical applications indicate that efficiency gains from ALS are not consistent. In some cases, ALS clearly improves precision, while in others, it does not. I found that applying the sigmoid transformation results in ALS and OLS estimates that are very similar, with efficiency gains that are real but minimal.

The Monte Carlo simulation revealed similar patterns: efficiency gains are not systematic and depend on how inconsistent probabilities are treated. When using the imputation method (0.001 and 0.999), the results showed meaningful efficiency gains, which increased with the magnitude of the true parameter β. This is intuitive, as larger parameter values produce a greater proportion of out-of-bounds probabilities, providing more room for potential gains. Another consistent finding is that efficiency gains diminish as sample size increases, since variance naturally decreases with more observations. When applying the sigmoid transformation, the results mirrored the empirical findings: ALS and OLS produced nearly identical estimates, and the gains were modest. Lastly, with the trimming estimator, no consistent pattern of efficiency gains

emerged. However, I did observe that as β increases, more observations are dropped from the sample, which increases the variance of the ALS estimator—again highlighting the relationship between sample size and variance.

To sum up, this paper tested the efficiency claims of the ALS estimator in the LPM framework using multiple correction strategies and both real and simulated data. The findings show that while ALS does not offer systematic efficiency gains, in certain contexts it can yield substantial improvements in precision.

References

Abadie, A. 2003. "Semiparametric instrumental variable estimation of treatment response models." *Journal of Econometrics* 113 (2):231–263.

Aitken, A. C. 1936. "IV.—On Least Squares and Linear Combination of Observations." *Proceedings of the Royal Society of Edinburgh* 55:42–48.

Angrist, Joshua, Pierre Azoulay, Glenn Ellison, Ryan Hill, and Susan Feng Lu. 2017. "Economic Research Evolves: Fields and Styles." *American Economic Review* 107 (5):293–97.

Angrist, Joshua D. and Jörn-Steffen Pischke. 2009. "Making Regression Make Sense." In *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton University Press.

Angrist, Joshua D. and Jørn-Steffen Pischke. 2012. "Probit better than LPM?" https://www.mostlyharmlesseconometrics.com/2012/07/probit-better-than-lpm/.

Augsburg, Britta, Ralph De Haas, Heike Harmgart, and Costas Meghir. 2015. "The Impacts of Microcredit: Evidence from Bosnia and Herzegovina." *American Economic Journal: Applied Economics* 7 (1):183–203.

Breusch, T. S. and A. R. Pagan. 1979. "A Simple Test for Heteroscedasticity and Random Coefficient Variation." *Econometrica* 47 (5):1287–1294.

Chen, Kaicheng, Robert S Martin, and Jeffrey M Wooldridge. 2025. "Another Look at the Linear Probability Model and Nonlinear Index Models." *BLS Working Papers* .

Giles, David E. 2012a. "Another Gripe About the Linear Probability Model." https://davegiles.blogspot.com/2012/06/another-gripe-about-linear-probability.html.

———. 2012b. "Yet Another Reason for Avoiding the Linear Probability Model." https://davegiles.blogspot.com/2012/06/yet-another-reason-for-avoiding-linear.html.

Grogger, Jeffrey. 1991. "Certainty vs. severity of punishment." *Economic inquiry* 29 (2):297–309.

Hamermesh, Daniel S. 2013. "Six Decades of Top Economics Publishing: Who and How?" *Journal of Economic Literature* 51 (1):162–72.

Horrace, William C and Ronald L Oaxaca. 2006. "Results on the bias and inconsistency of ordinary least squares for the linear probability model." *Economics letters* 90 (3):321–327.

Hunter, W.C. and M.B. Walker. 1996. "The cultural affinity hypothesis and mortgage lending decisions." *Journal of Real Estate Finance and Economics* 13:57–70.

MacKinnon, James G. 2012. "Thirty years of heteroskedasticity-robust inference." In *Recent advances and future directions in causality, prediction, and specification analysis: Essays in honor of Halbert L. White Jr.* Springer, 437–461.

MacKinnon, James G. and Halbert White. 1985. "Some heteroskedasticity-consistent covariance matrix estimators with improved finite sample properties." *Journal of Econometrics* 29 (3):305–325.

Mroz, Thomas. 1987. "The Sensitivity of an Empirical Model of Married Women's Hours of Work to Economic and Statistical Assumptions." *Econometrica* 55 (4):765–99.

Petersen, Mitchell A. 2008. "Estimating Standard Errors in Finance Panel Data Sets: Comparing Approaches." *The Review of Financial Studies* 22 (1):435–480.

Romano, Joseph P and Michael Wolf. 2017. "Resurrecting weighted least squares." *Journal of Econometrics* 197 (1):1–19.

White, Halbert. 1980. "A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity." *Econometrica: Journal of the Econometric Society* :817–838.

Wooldridge, Jeffrey M. 2010. *Econometric Analysis of Cross Section and Panel Data*. Cambridge, MA: The MIT Press, 2nd ed.

———. 2015. *Introductory Econometrics: A Modern Approach*. Mason, OH: South-Western College Pub, 6th ed.

Appendix

Table A1: Variable descriptions and summary statistics from the dataset used in Mroz (1987)

Variable	Description	Obs	Mean	SD
inlf	=1 if in the labor force in 1975	753	0.57	0.50
nwifeinc	(Family income - earnings)/1000	753	20.13	11.63
educ	Years of schooling	753	12.29	2.28
exper	Current labor market experience	753	10.63	8.07
expersq	Experience squared	753	178.04	249.63
age	Age of the woman (in years)	753	42.54	8.07
kidslt6	Number of children under age 6	753	0.24	0.52
kidsge6	Number of children between 6–18	753	1.35	1.32

Table A2: Variable descriptions and summary statistics from the dataset used in Grogger (1991)

Variable	Description	Count	Mean	SD
arr86	=1 if arrested in 1986	2725	0.28	0.45
pcnv	Proportion of prior convictions	2725	0.36	0.40
avgsen	Average sentence length (months)	2725	0.63	3.51
tottime	Total time in prison since age 18 (mo.)	2725	0.84	4.61
ptime86	Months spent in prison during 1986	2725	0.39	1.95
qemp86	Number of quarters employed, 1986	2725	2.31	1.61
black	=1 if Black	2725	0.16	0.37
hispan	=1 if Hispanic	2725	0.22	0.41

Table A3: Variable descriptions and summary statistics from the dataset used in Hunter and Walker (1996)

Variable	Description	Count	Mean	SD
approve	=1 if the loan was approved	1989	0.88	0.33
white	=1 if the applicant is white	1989	0.85	0.36
hrat	Housing expenses as % of total income	1989	24.79	7.12
obrat	Other obligations as % of total income	1989	32.39	8.26
loanprc	Loan amount divided by property price	1989	0.77	0.19
unem	Industry-specific unemployment rate	1989	3.88	2.16
male	=1 if the applicant is male	1974	0.81	0.39
married	=1 if the applicant is married	1986	0.66	0.47
dep	Number of dependents	1986	0.77	1.10
sch	=1 if education $>$ 12 years	1989	0.77	0.42
cosign	=1 if there is a co-signer	1989	0.03	0.17
chist	=0 if any accounts are delinquent $\geq$ 60 days	1989	0.84	0.37
pubrec	=1 if bankruptcy has been declared	1989	0.07	0.25
mortlat1	One or two late mortgage payments	1989	0.02	0.14
mortlat2	>2 late mortgage payments	1989	0.01	0.10
vr	=1 if vacancy rate in tract $>$ MSA median	1989	0.41	0.49

Table A4: Variable descriptions and summary statistics from the dataset used in Abadie (2003)

Variable	Description	Count	Mean	SD
e401k	=1 if eligible for the 401(k) pension program	9275	0.39	0.49
inc	Income	9275	39.25	24.09
age	Age	9275	41.08	10.30
agesq	Age squared	9275	1793.65	895.65

Figure A1: Sigmoid transformation