B Extensions

B.1 Extension to Conditional Similarity

Assumption 1 assumes that the distributions of characteristics are identical between experimental and observational data. This assumption is plausible if the experimental unit is randomly selected. This assumption can be relaxed to assume that the distributions are identical after adjusting for observed covariate Z_i:

Assumption 3 (Conditional Similarity in Unobservables)

$$G_i \perp U_i | Z_i$$

This conditional independence condition is common in the related literature (e.g., Athey et al. (2020), Kallus et al. (2018)) to ensure effects learned in one dataset can be transferred to another dataset.

Compared to Assumption 1, this relaxed assumption allows for the possibility of stratified sampling in which the probability of selecting into the experimental dataset depends on observed covariates. For example, firms may choose to reduce the experimental traffic during certain business hours. If firms believe that experimenting with new customers is relatively riskier, firms may also consider reducing the amount of experimental traffic for relatively new customers. In this case, Assumption 3 holds if Z includes the variables firms use for stratification.

My method can be extended to account for this relaxed assumption by re-weighting or re-sampling the observational data such that the distribution of Z for these two datasets match.

B.2 Nonlinear Model Efficiency Gain: A GMM Perspective

Section 6 takes a constraint optimization perspective to discuss why the efficiency gain may be increased in nonlinear models. This section takes a GMM perspective to discuss why efficiency gain can be improved under the nonlinear model. For illustration, consider the case when Z is a continuous covariate observed by researcher, X is a continuous variable, and the function h is smooth in the continuous X and Z, such that the partial derivatives of the function h are well defined.[17]

Consider the neighborhood of any (X_0, Z_0). Let $b_2 \equiv \frac{\partial h(X_0, Z_0; \theta)}{\partial Z}$, and $\beta_1 \equiv \frac{\partial h(X_0, Z_0; \theta)}{\partial X}$. Then around the neighborhood of (X_0, Z_0):

[17] Additional parametric assumptions may be needed for binary Z_i and X_i.

$$
\begin{aligned}
h(X_i, Z_i; \theta) &\approx h(X_0, Z_0; \theta) + \frac{\partial h(X_0, Z_0; \theta)}{\partial X}(X_i - X_0) + \frac{\partial h(X_0, Z_0; \theta)}{\partial Z}(Z_i - Z_0) \\
&= h(X_0, Z_0; \theta) + \beta_1(X_i - X_0) + b_2(Z_i - Z_0) \\
&= h(X_0, Z_0; \theta) - \beta_1 X_0 - b_2 Z_0 + \beta_1 X_i + b_2 Z_i
\end{aligned}
$$

Let $\alpha_0 = h(X_0, Z_0; \theta) - \beta_1 \cdot X_0 - b_2 \cdot Z_0$ be the local intercept such that

$$
h(X_i, Z_i; \theta) \approx \alpha_0 + \beta_1 X_i + b_2 Z_i
$$

Then we can derive local moment conditions from experimental data:

$$
\begin{aligned}
E[(Y_i - \alpha_0 - \beta_1 X_i - b_2 Z_i)1 | G_i = E] &\approx 0 \\
E[(Y_i - \alpha_0 - \beta_1 X_i - b_2 Z_i)X_i | G_i = E] &\approx 0 \\
E[(Y_i - \alpha_0 - \beta_1 X_i - b_2 Z_i)Z_i | G_i = E] &\approx 0
\end{aligned}
$$

as well as additional moment conditions from the observational data

$$
\begin{aligned}
E[(Y_i - \alpha_0 - \beta_1 X_i - b_2 Z_i)1 | G_i = O] &\approx 0 \\
E[(Y_i - \alpha_0 - \beta_1 X_i - b_2 Z_i)Z_i | G_i = O] &\approx 0
\end{aligned}
$$

Because these local moment conditions are similar to the moment conditions in the linear case, the accuracy of β_1 can also be improved based on Theorem 1, where the determinants of efficiency gain is the local variance and first-stage relevance of Z_i, as well as the proportion of observational data. If the accuracy of β_1 is improved, then by definition the accuracy for $\frac{\partial h(X,Z;\theta)}{\partial X}$ will also be improved, which means that we can better estimate the marginal impact of X.

B.3 Extension to Binary Probit

In this section, I illustrate how the observational data can be incorporated into the case of binary probit with endogenous and continuous regressors following a textbook example in Wooldridge (2010, Section 15.7.2), where:

$$
Y_i = \begin{cases} 1 & Y_i^* > 0 \\ 0 & \text{otherwise} \end{cases} \tag{19}
$$

where $Y_i^* = \beta_1 X_i + \beta_2 Z_i + U_i$. I hold the first-stage equation to be the same as in Equation 2:

$$
X_i = \begin{cases} \gamma Z_i + V_i & \text{if } G_i = O \\ randomized & \text{if } G_i = E \end{cases}.
$$

The standard probit assumes that (U_i, V_i) follows a standard normal distribution that is independent of the instrumental variable Z_i. To make the case more general, I allow Z_i to be correlated with U_i:[18]

$$(U_i, V_i|Z_i) \sim N\left(\begin{bmatrix}\rho_{zu} \times Z_i \\ 0\end{bmatrix}, \begin{bmatrix}1 & \rho_{uv}\sigma_v \\ \rho_{uv}\sigma_v & \sigma_v^2\end{bmatrix}\right)$$

Z_i is therefore not an instrumental variable because it either directly affects Y_i^*, or correlated with unobservables U_i.

As in my linear case, not all parameters can be separately identified in this setup and I need to aggregate some parameters for estimation convenience. I define b_2 as

$$b_2 \equiv \beta_2 + \rho_{zu}$$

which captures both the direct effect of Z on Y^* and an indirect correlation with the unobservables U. Let Θ denote the set of unknown parameters:

$$\Theta = \{\beta_1, b_2, \rho_{uv}, \gamma, \sigma_v\}$$

A maximum likelihood estimator that only uses the experimental data chooses (β_1, b_2) to maximizes:

$$\ln \mathcal{L}_E(\Theta) = \ln \prod_{i:G_i=E} P(Y_i|X_i, Z_i, \beta_1, b_2)$$

The log-likelihood of the observational data can be written as

$$\begin{aligned}\ln \mathcal{L}_O(\Theta) &= \ln \prod_{i:G_i=O} P(Y_i, X_i|Z_i, \Theta) \\ &= \ln \prod_{i:G_i=O} P(Y_i|Z_i, X_i, \Theta)P(X_i|Z_i, \Theta)\end{aligned}$$

The data can be combined by optimizing the sum of these two log-likelihoods:

$$\max_{\Theta} \ln \mathcal{L}_E(\Theta) + \ln \mathcal{L}_O(\Theta)$$

To understand why incorporating the observational data can help improve efficiency in this probit case, consider when the size of the observational data is infinite. (γ, σ_v) are identified using the first-stage equation, where

$$P(X_i|Z_i, G_i = O, \Theta) = \phi(\frac{X_i - \gamma Z_i}{\sigma_v})$$

[18] I maintain the assumption that $Cov(Z_i, V_i|G_i = O) = 0$ since γ is a linear projection parameter, Since I assume X_i is determined by (Z_i, V_i) and I do not consider simultaneous equation, I do not allow the distribution to depend on X_i

Following Wooldridge, under normality, the second stage can be written as:

$$P(Y_i = 1|X_i, Z_i, G_i = O, \Theta) = \Phi\left[\frac{\beta_1 X_i + b_2 Z_i + \frac{\rho_{uv}}{\sigma_v}(X_i - \gamma Z_i)}{\sqrt{1-\rho_{uv}^2}}\right]$$
$$= \Phi\left[\frac{1}{\sqrt{1-\rho_{uv}^2}}(\beta_1 + \frac{\rho_{uv}}{\sigma_v})X_i + \frac{1}{\sqrt{1-\rho_{uv}^2}}(b_2 - \frac{\rho_{uv}}{\sigma_v}\gamma)Z_i\right]$$

A probit of Y_i on (X_i, Z_i) on observational data can then consistently identify two constraints:

$$C_1 = \frac{1}{\sqrt{1-\rho_{uv}^2}}(\beta_1 + \frac{\rho_{uv}}{\sigma_v})$$
$$C_2 = \frac{1}{\sqrt{1-\rho_{uv}^2}}(b_2 - \frac{\rho_{uv}}{\sigma_v}\gamma)$$

Since (γ, σ_v) are already identified in the first-stage, $(b_2, \beta_1, \rho_{uv})$ are the three unknowns left. If $b_2 = 0$, the problem is reduced to an endogenous binary probit with Z_i being an instrumental variable; then (β_1, ρ_{uv}) are just identified under C_1 and C_2. However, because b_2 is unknown, the observational data alone are not sufficient for identification.

Although the observational data are not sufficient for identification, it can help improve efficiency of the experimental data by imposing additional constraints onto the maximum likelihood problem:

$$\max_{\beta_1, b_2, \rho_{uv}} \ln \mathcal{L}_E(\beta_1, b_2)$$
$$\text{subject to } C_1 = \frac{1}{\sqrt{1-\rho_{uv}^2}}(\beta_1 + \frac{\rho_{uv}}{\sigma_v})$$
$$C_2 = \frac{1}{\sqrt{1-\rho_{uv}^2}}(b_2 - \frac{\rho_{uv}}{\sigma_v}\gamma)$$

Intuitively, when the two constraints C_1 and C_2 are precisely measured, any deviations from these constraints must be attributed to the inaccuracy of (β_1, b_2). Incorporating the observational data can therefore help improving the efficiency.

B.4 Extension to Misspecified Observational Model

My method can be generalized to settings when the experimental data are still sufficient for identification, but the observational model is misspecified. Recall that Equation 1 assumes a model with additive separability that rules out any interactions. Equation 9 assumes a more flexible nonlinear model with additive unobservables. Assumption 3 assumes that distributions of unobservables are similar between observational and experimental units after conditioning on covariates. To alleviate the concerns that these assumptions may be violated in practice, I discuss strategies that help researchers deter-

mine: 1) whether they should incorporate the observational data and 2) how to tune the hyperparameter when the model may be misspecified. The method is still valuable if the goal is to minimize mean-squared error (MSE), which does not only include the bias but also the variance.

One strategy is to use cross validation, a popular procedure for tuning hyperparameters in the regularized regression. The key is to cross validate on out-of-sample experimental data. Intuitively, a better causal estimator should more accurately predict Y given a randomized X. Researchers can therefore test the usefulness of a combined estimator $\widehat{\beta}_1^{combine}$ relative to the experiment-only estimator $\widehat{\beta}_1^{E}$ using out-of-sample experimental data.

Another strategy is to average multiple GMM estimators given the recent development in the literature. The key is to recognize that two estimators can be generated when both observational and experimental estimates are available:[19] 1) a valid experiment-only estimate $\hat{\beta}_1^{E}$ that does not require the additional assumptions, and 2) a more efficient estimator $\widehat{\beta}_1^{combine}$ that is potentially biased due to model misspecification. This setting fits into the classical Hausman specification test (Hausman (1978)), where researchers can decide to use the more efficient combined estimator if it is not significantly different from the less efficient one that only uses the experimental data. This setting also fits into the more recent framework developed by Cheng et al. (2019) that averages two GMM estimators, where one is always consistent, and the other one could be inconsistent due to misspecifications of moment conditions. The framework guarantees that given certain weights, the mean-squared-error can be improved. Following their framework, one can combine the two estimators:

$$\widehat{\beta}_1^{combine,robust} = w\widehat{\beta}_1^{combine} + (1-w)\hat{\beta}_1^{E}$$

where the optimal weight w is determined by the relative variance and bias differences.

$$\begin{aligned} w^* &= \frac{\mathbb{V}(\hat{\beta}_1^E) - \mathbb{V}(\widehat{\beta}_1^{combine})}{(E[\hat{\beta}_1^E] - E[\widehat{\beta}_1^{combine}])^2 + \mathbb{V}(\hat{\beta}_1^E) - \mathbb{V}(\widehat{\beta}_1^{combine})} \\ &= \frac{ImprovedVariance}{Bias^2 + ImprovedVariance} \end{aligned}$$

Therefore, when the combined estimator is consistent, the optimal weight is approximately 1, suggesting that one can safely use the combined estimate. When the combined estimator has large bias relative to the variance improvement, w approaches 0, suggesting that one should avoid using the combined estimate. When the bias is small, researchers can still use the combined method to improve the overall MSE.

[19] I thank the anonymous associate editor for this valuable insight.

B.5 Extension to Other Empirical Settings

This section discusses additional examples under which my method could be useful. A common reason for firms to analyze data in the past is to improve decisions and profit in the future. Firms are only interested in conducting such analyses if past customers are not too different from future customers, otherwise insights gained from the past are not useful for the future. For this type of firms, if the observational data include all customers that arrive in the past, and the experimental data include all customers that arrive in the near future, it is plausible to assume that these two populations are similar. To improve future decisions, firms need to determine the size of the experiment as well as what to do after they obtain the experimental results.

My method can improve the profit by lowering the size of the experiment needed to achieve certain statistical accuracy. According to Feit and Berman (2019), the profit depends on the size of the experimental sample because larger experiments imply that more customers receive suboptimal treatments under randomization, leading to lower profit. Because my method can reduce the variance by up to 50%, only half of the experimental sample is required to attain the same statistical precision, thus improving the profit.

B.5.1 Credit Loan

In general, it is difficult to find a dataset where both experimental and observational data are available in other contexts. To best demonstrate the efficiency gain of my method in another context, I simulated a semi-realistic setting based on a direct mail field experiment (Bertrand et al. (2010)) implemented by a consumer lender that randomizes creative content and interest rates. One parameter of interest in the experiment is how the offered interest rate affects the customer loan amount, an important managerial input that firms can use to optimize future offering of interest rates. Bertrand et al. (2010) assume a linear causal model to answer this question:[20]

$$\text{LoanAmount}_i = \beta_1 \text{InterestRate}_i + \beta_2 \text{ConsumerCharacteristics}_i + \epsilon_i \tag{20}$$

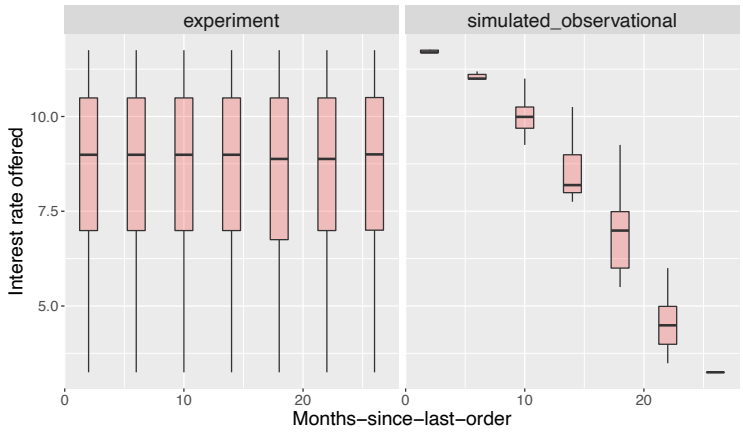
where consumer characteristics include variables such as credit risk as well as months-since-last-loan, β_2 is the vector of corresponding coefficient, and β_1 is the parameter of interest. The experiment is sufficient for identifying the impacts of interest rate on loan because the interest rate is randomly drawn from a distribution after stratifying on credit risk.[21] However, in practice the interest rate is directly affected by many consumer characteristics. For example, the firm may give a lower interest rate to churned customers who have not made a loan for a few months. If researchers also have access to

[20] The OLS specification in Table (3) Column 4 of the paper.

[21] For illustrative purposes, I focus on the customers in the high-risk stratum, whose interest rate is randomly drawn from the same distribution because the experiment is stratified. My method can be easily extended to other strata by combining experimental and observational data within those strata.

the large amount of business-as-usual direct mail data before the experiment, then my method can incorporate this observational data to improve the efficiency of the experiment estimate. Because Bertrand et al. (2010) does not include such observational data, additional assumptions are needed to to simulate such a semi-realistic setting. I first follow Equation 20 to estimate the model using all the experimental sample and take the estimated parameters and unobservables as ground truth. I then assume that the interest rate is determined by a weighted average between the months-since-last-loan and unobservables, where months-since-last-loan receive 95% of the weight.[22] Figure 5 shows that the interest rate is random in the experimental data, but is negatively correlated with the months-since-last-order in the simulated observational data.

Figure 5: Distribution of interest rate vs months since last order



Similar to the Expedia application, I examine the effectiveness of my method when units are randomly assigned into the experimental and observational groups. I focus on a case when the observational group is large with $20,000$ units and the experimental group is small with $1,000$ units. I repeatedly draw $M = 10,000$ such samples and perform estimation on each sample. Table 9 reports the bias, variance, and MSE of different estimators across 10,000 replications. To characterize the efficiency gain, I compare the MSE of these methods with a benchmark that uses only a random sample of experimental data for estimation. The OLS using observational data has large bias due to endogeneity. The incorrect IV approach also has large bias, because the variable Z_i violates the exclusion restriction and is not a valid IV. In comparison, the combined GMM method has a efficiency gain of 45.7%.

For the two unbiased estimators, Table 10 reports how often the increased efficiency changes the statistical significance at the 95% level. The combined GMM approach is more likely to detect that the effect of interest rate on loan is negative, and also more likely to detect that this effect is statistically significant.

[22]In practice, many other features are used for determining the interest rate offer, such as a continuous credit score. All these variables can be incorporated as well

Table 9: Summary for different estimators when units are randomly assigned into experimental and observational groups

Estimator	Bias^2	Variance	MSE	Relative MSE	Efficiency Gain
Experiment Only	0.007	16.796	16.803	1.000	0.000
Combined GMM	0.002	9.127	9.129	0.543	0.457
OLS (Obs)	12107.454	56.715	12164.169	723.928	-722.928
IV (Obs)	105.028	0.524	105.552	6.282	-5.282

Table 10: Summary of statistical significance ($p < 0.05$) over $10,000$ samples

Method	Negative (Correct Sign)	Significantly Negative
Experiment Only	80.38%	13.99%
Combined GMM	87.61%	21.18%

B.5.2 Incentives

Rideshare platforms such as Uber and Lyft are often interested in how workers respond to wages and incentives. Causally measuring the elasticity of labor supply is crucial for satisfying consumer demand. The experimental approach is to randomly change the wages. However, because randomizing wages is controversial and may lead to public backlash, firms are sometimes only willing to conduct a small-scale experiment to minimize the risk. For example, Chen et al. (2019) mentioned that Uber has conducted some randomized wage experiments on a limited basis in several cities. One such experiment is conducted in Orange County, California that covers approximately $3,000$ Uber drivers during April 2016. In the experiment, a random set of drivers receive an email indicating that the driver would receive a 10 percent increase in wages for 3 weeks. Chen et al. (2019) use this small-scale experiment to show that the results obtained using a larger observational dataset that covers $200,000$ drivers are valid.[23]

In this setting, the experimental dataset is small, containing one market for three experimental weeks, and the observational dataset is large, containing all other non-experimental markets and weeks. In the experimental data, the wage is set randomly; in the observational data, the wage is affected by many factors observed by firms, such as local consumer demand and worker characteristics. Because workers' willingness to work may be correlated with these factors, observational methods that incorrectly use these factors as IV may generate biased results. Because my method can account for this bias, researchers can derive an additional bias-corrected estimate that is uncorrelated with the experiment-only estimate to improve the efficiency.

[23] An implicit assumption behind this exercise is that observational and experimental population are similar, satisfying my Assumption 1