Examining the Influence of Regressor Distributions and Exclusion Restrictions on Heckman Selection Model Performance

Emmanuel Honny
Department of Agricultural Economics, Oklahoma State University
emmanuel.honny@okstate.com

Chanjin Chung
Department of Agricultural Economics, Oklahoma State University
chanjin.chung@okstate.edu

Selected Paper prepared for presentation at the 2025 AAEA & WAEA Joint Annual Meeting in Denver, CO; July 27-29, 2025

Introduction

Accurate estimation of economic models underpins reliable inferences, driving effective policy and business decisions. Sample selection bias, which occurs when the sample used in the analysis is not representative of the population due to a systematic exclusion of certain observations, threatens accuracy when data derives from a non-random sample, leading to biased and inconsistent estimates (Winship & Mare, 1992). When researchers fail to address this bias, they risk drawing conclusions that do not reflect the broader population (Sartori, 2003; Puhani, 2000). One widely used method for correcting sample selection bias is the Heckman model (Heckman, 1976). The Heckman model is a two-step procedure to alleviate selection bias by treating the selection problem as an omitted variable problem. To gain consistent estimates from the Heckman model, it is imperative to meet the assumptions under which the model effectively corrects for selection bias.

Collinearity between independent variables and the Inverse Mills Ratio (IMR) in the second stage of the Heckman model can severely compromise its performance by inflating standard errors, and destabilizing parameter estimates (Leung & Yu, 1996). This issue becomes particularly problematic when exclusion restrictions are absent as it increases the correlation between the IMR and the regressors. Leung and Yu demonstrated how collinearity is worsened with narrow regressor distributions, such as Uniform (0,3), compared to broader ones like Uniform (0,10) and how these undermine the Heckman model's reliability. The authors used condition numbers to measure the correlation between the independent variables and the IMR and suggested the Heckman model can be productively estimated if the condition number is less than 20. Their suggested threshold of 20 may be more specific to the experimental designs and the conditions they examined and may not universally apply across different datasets.

Subsequent research, such as Puhani (2000), Bushway et al. (2007), and Thapa, Morrison, and Parton (2021), has adopted Leung and Yu's threshold to diagnose collinearity problems in the Heckman model, reflecting the influence of their work. The reliance on this specific threshold without considering alternative experimental conditions could overlook whether the threshold remains valid across different scenarios or datasets. Building on their work, this study investigates how alternative distributions, such as Normal and Gamma distributions, influence the correlation between regressors and the IMR, aiming to provide a deeper understanding of the model's limitations and enhance its application in diverse empirical contexts. Specifically, this research seeks to answer: How do different distributions of the independent variable and varying strengths of exclusion restrictions affect the correlation between the independent variable and the IMR, and consequently, the performance of the Heckman model?

This study aims to (1) assess how different distributions of the independent variable influence the correlation between the independent variable and the Inverse Mills Ratio (IMR) in the outcome equation of the Heckman model using metrics like correlation coefficients and condition numbers, and how this correlation affects model performance using bias and mean squared error (MSE) metrics, (2) to investigate whether the condition number threshold of 20, proposed by Leung and Yu (1996), consistently holds across different distributions of the independent variable, (3) to assess how different levels of strengths of exclusion restrictions affect the correlation between the independent variable and the IMR, (4) to explore how to identify the strength of exclusion restrictions.

The effectiveness of the Heckman model does not only rely on the correlation between the independent variable and the IMR in the outcome equation of the Heckman model but also on certain assumptions that need to be met to identify the effect correlation in independent

variable and the IMR on the performance of the Heckman model. The primary identifying assumption of the Heckman estimator is that the error terms in both the selection equation and the outcome equation follow a joint normal distribution (Sartori, 2003). As demonstrated by Arabmazar and Schmidt (1982) and Robinson (1982), the estimator becomes inconsistent when the errors deviate from this normality assumption, leading to biased estimates (Lai and Tsay, 2018). According to Certo et al. (2016), researchers testing for potential sample selection bias should check the significance of the independent variables of interest in the first stage of the Heckman model. If the independent variables of interest are not statistically significant in the initial selection equation, it suggests that sample selection bias may not be present. In such cases, alternative estimators, such as ordinary least squares (OLS), can be used. If the independent variable is statistically significant in the first stage, a Heckman model can be utilized to assess the significance of lambda (λ) in the model's second stage, where lambda represents the Inverse Mills Ratio (IMR). The IMR is included as a regressor in the second stage of the Heckman model to correct for sample selection bias (Tucker, 2011). A significant lambda indicates the presence of sample selection bias. In terms of hypothesis tests, the null hypothesis (H_0) is $\lambda = 0$ meaning there is no sample selection bias, while the alternative hypothesis (H_1) is $\lambda \neq 0$ meaning sample selection is present. If the estimated coefficient for λ is statistically significant we reject the null hypothesis and conclude that sample selection bias is present, however, if λ is not statistically significant, we fail to reject the null hypothesis and conclude that sample selection bias is not present.

A key component of the Heckman model's implementation is the assumption of valid exclusion restrictions, which are variables that affect the selection process (the first stage) but not the outcome of interest (the second stage). The Heckman model's effectiveness depends on the

proper selection and validity of these exclusion restrictions (Gomes et al., 2020; Sundaram-Stukel, 2021). The Heckman model tends to produce inflated standard errors when exclusion restrictions are absent, meaning all variables included in the first stage are identical to the covariates used in the second stage (Bushway et al., 2007). Stolzenberg and Relles (1990) suggest that when identical covariates are used in both stages of the Heckman model, the Two-Part Model (TPM) which separately models the probability of selection and the level of the outcome assuming independence between the two stages, may be a preferable approach, despite its estimates being inherently biased in the presence of selection. Other research has explored alternative approaches to address identical covariates in both stages of the Heckman model, such as maximum likelihood (Sartori, 2003) and semiparametric extensions (Honoré & Hu, 2022). The Heckman model is more preferred when exclusion restriction is present (Leung & Yu, 1996). The strength and validity of this exclusion restriction in the Heckman model are vital as weak exclusion restrictions are unlikely to yield significant lambda, even when sample selection bias is present (Certo et al., 2016).

Conceptual Framework

Leung and Yu (1996) conducted a series of Monte Carlo simulations to evaluate how collinearity between the regressors and the Inverse Mills Ratio (IMR) affects the performance of the Heckman sample selection model. They specifically varied the regressor's distribution from a narrow uniform interval ($U(0,3)$) to a wider one ($U(0,10)$) to see how this change influenced collinearity and model's performance. They showed that when regressors were drawn from a narrow uniform distribution ($U(0,3)$), the regressors and the IMR becomes highly correlated. By widening the regressor range to $U(0,10)$, they observed that collinearity was reduced, resulting

in better performance of the Heckman model. They measured collinearity using the condition number of the matrix formed by the regressors and the IMR, proposing a threshold of 20 as an indicator of collinearity problems.

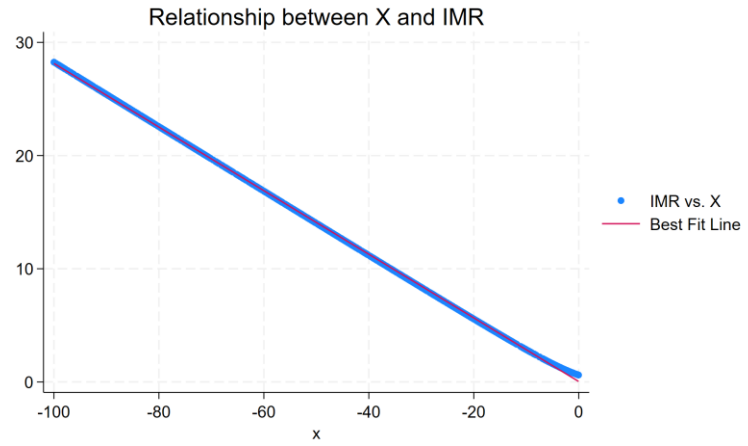
However, Leung and Yu's work primarily focused on uniform regressor distributions and a specific range of sample sizes, leaving open the question of whether their recommended condition number threshold of 20 holds more broadly across different regressor distributions (e.g., normal, gamma) and varying sample sizes. This paper extends their work by exploring broader regressor distributions and sample sizes to evaluate the robustness of their proposed condition number threshold and its applicability across diverse scenarios. Leung and Yu's design approach focused only on positive-valued regressors. Extending the distribution to include distributions with both positive and negative values, such as from $N(0,3)$ and $G(1,1)$, may result in different IMR–regressor correlation patterns. Having negative value included could influence correlation as the IMR behaves different both the negative and positive region. This is due to how the IMR is being calculated, that is,

$$\lambda(\boldsymbol{w_i'\gamma}) = \frac{\phi\,(\boldsymbol{w_i'\gamma})}{\Phi(\mathbf{w_i'\gamma})} \tag{3}$$

where $\phi\,(\boldsymbol{w_i'\gamma})$ is the probability density function of the standard normal distribution and $\Phi(\mathbf{w_i'\gamma})$ is the cumulative distribution function of the standard normal distribution (Heckman, 1979). $\boldsymbol{w_i'\gamma}$ is the estimated latent index obtained from the first stage probit model. For example, a regressor having all negative values ($U(-100,0)$) results in $\boldsymbol{w_i'\gamma}$ being negative. When $\boldsymbol{w_i'\gamma}$ is large and negative, $\Phi(\mathbf{w_i'\gamma})$ becomes very small, making the IMR very large. As $\boldsymbol{w_i'\gamma}$ increases toward 0, that is the regressor value increases toward 0, $\Phi(\mathbf{w_i'\gamma})$ increases, and the IMR declines. The IMR flattens near zero with higher positive values of the regressor. Therefore,

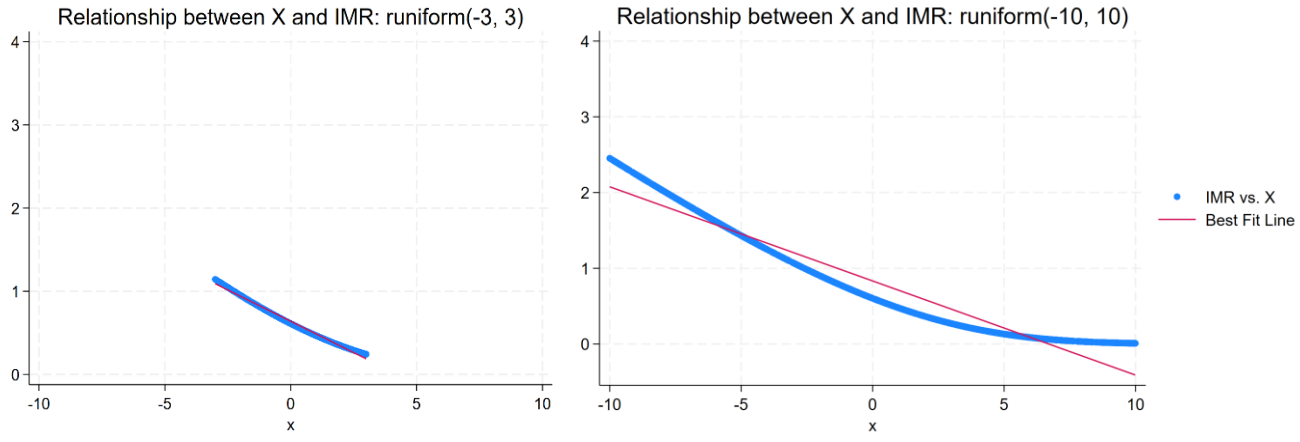
increasing the range of regressor with negative may not necessarily decrease the correlation between the regressor and IMR (figure 1).

Figure 1. Inverse Mills Ratio Behavior Across Negative Values of the Regressor (U(0,-100))



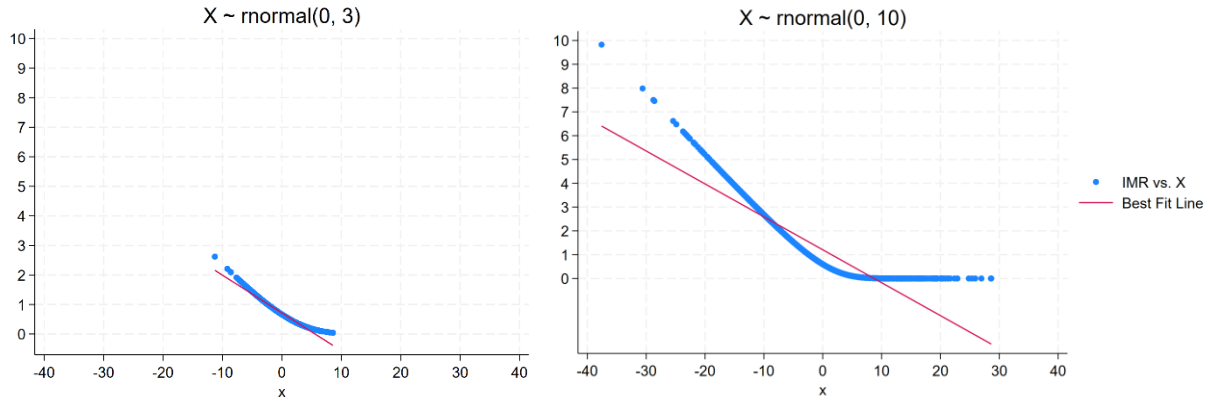
Also, having the regressor drawn from both positive and negative values may follow the rule of increasing the range of the variable will reduce the correlation between the regressor and IMR. For example, a regressor that is generated from $U(-3,3)$ will have a higher correlation between the regressor and IMR than a regressor that is generated from $U(-10,10)$ since the latter will have the IMR flattens near zero for the more positive values is have (figure 2).

Figure 2. Inverse Mills Ratio Behavior Across Negative and Positive Values of the Regressor (U(-3,3) and U(-10,10))



Alternative regressor distributions may affect the correlation between the IMR and the regressor differently. Normal distributions are symmetric and concentrate most values around the mean. In this case, simply increasing the range may not meaningfully reduce the correlation between the regressor and IMR, since the bulk of values still lie in the nonlinear middle portion of the IMR curve However, increasing the variance of a normal distribution causes values to spread more widely into the tails, placing more observations into regions where the IMR declines rapidly toward zero (in the left tail) or flattens around zero (in the right tail). This spread weakens the correlation between the regressor and IMR (figure 3).

Figure 3. Inverse Mills Ratio Behavior Across Negative and Positive Values of the Regressor ((0,3) and N(0,10))



Data

The Monte Carlo simulation is conducted to evaluate the performance of the Heckman model under varying conditions, using metrics such as bias and mean squared error (MSE) to assess its effectiveness. First, datasets with known parameters that follow the structure of the Heckman model are generated. The simulation aims to compare the performance of the Heckman model against the true parameters and evaluate its performance in the presence of collinearity of the independent variable and IMR in the outcome equation.

Data was generated with different distributions of the independent variable, varying degrees of censoring, and exclusion restriction strength. First, a full population of 1,000 observations (N) was generated and nine different percentages of censoring are considered, which are 10% up to 90%. An effect of the dependent and independent variable was modeled using the linear regression model specified as

(1) $$\ln(y_i) = \alpha + \beta x_i + \varepsilon_i$$

where y_i is the dependent variable for observation i ($i = 1,\ldots,n$), x_i is the independent variable of interest, α and β are the intercept and slope parameters to be estimated respectively and ε_i is the error term which is independently and identically distributed and expressed as $\varepsilon_i \sim N(0, \sigma_\varepsilon^2)$. In line with the model presented by Certo et al. (2016), a small effect of the independent variable on the dependent variable was considered for this study. The coefficient, β, was set at 1 for all simulations. The independent variable of interest, x, was generated under six distinct scenarios: two following a normal distribution with a mean of 0 and variances of 3 and 10, and two following a gamma distribution with parameters (1, 1) and (5, 1). In comparison, Leung and Yu (1996), independent variable of interest, x, was also drawn from a uniform distribution with two different ranges: (0, 3), and (0, 10). Therefore, this study extends the work of Leung and Yu (1996) who focused on uniform distribution by drawing the independent variable from a normal and gamma distribution to evaluate how it affects collinearity between the independent variable of interest and the IMR.

After the relationship between the dependent and independent variable has been established, sample selection is then introduced through the following equation:

(2) $$d_i = \delta + \lambda_1 z_{1i} + \lambda_2 z_{2i} + \lambda_3 z_{3i} + \lambda_4 x_i + u_i$$

where d_i is a latent variable that determines whether an observation i is included in the selected sample, δ is the intercept, z_{1i}, z_{2i} and z_{3i} are exclusion restriction variables, x_i is the independent variable of interest, λ_1, λ_2, λ_3 and λ_4 are coefficients to be estimated and μ_i is the error term is independently and identically distributed and expressed as $u_i \sim N(0, \sigma_u^2)$. In the first stage, the dependent variable d_i represents a continuous latent variable. When d_i exceeds a specified threshold, the selection indicator D is set to 1 (indicating that the observation is included in the selected sample). Otherwise, D is set to 0 (indicating that the observation is excluded from the sample). The value intercept δ determines the percentage of observations censored. Three exclusion restrictions were included to model small, moderate, and large effects. The exclusion restrictions z_{1i}, z_{2i} and z_{3i} are all assumed to follow a normal distribution with a mean of 0 and variances of 1. The coefficients of the exclusion restrictions which are, λ_1, λ_2 and λ_3 are set to a number to model the strength of the exclusion restriction using McFadden pseudo-R^2 values as a guide. According to Cohen's (1992) pseudo R^2 values of 0.02, 0.10, and 0.20 indicate small, medium and large effect respectively. These pseudo R^2 value were used as guides to in generating the coefficient of the exclusion restrictions.

For each scenario that incorporates an exclusion restriction, only one of the exclusion restriction variables is included in the Heckman model. Although data will be generated using three exclusion restrictions as shown in equation 2, in the analysis, we assume access to only one exclusion restriction, simulating a realistic research scenario where data availability is limited.

Correlation between the error terms of the selection and outcome equation (ρ) is considered to be 0.5 which indicates the presence of selection bias since $\rho \neq 0$ (Campbell & Nagel, 2016). For each experiment, we generate a dataset of size N and conduct 1,000 Monte

Carlo simulations. The results are summarized by reporting the average values of the statistics obtained across these simulations.

Procedure

This research will first replicate the work of Leung and Yu, that is, replicating their experimental design using uniform distribution of the regressor and sample size before alternative distribution of the regressors as well as the sample size is the changed for comparison. As Belsley (1991) cautions, condition indexes like 28 and 35 are "essentially the same," given the nonlinear progression of the condition index scale (1, 3, 10, 30, 100, 300, 1,000, …). This means that small numerical differences between close to a considered high condition index do not necessarily reflect meaningful differences in collinearity severity. As a result, strictly applying a threshold like 20 can be misleading, as it may label a condition index of 21 as problematic while overlooking a value like 19, even though both fall within the same general range of collinearity intensity. As further emphasized by Callaghan and Chen (2008) who applied a threshold of 30 but they considered a condition index of 29 as noteworthy, despite it falling just below the threshold of 30.

Unlike Certo et al., (2016) who used limited scenarios to examine the relationship between varying exclusion restriction strengths and the correlation between IMR and regressors, this study also includes various regressor distributions, censoring levels, and exclusion restriction strengths, to establish practical benchmarks for the correlation between regressors and the IMR that signify strong, moderate, or weak exclusion restrictions.

To assess how different regressor distributions and varying exclusion restriction strengths influence the correlation between the regressor and the IMR, we employ the condition number diagnostic approach, as outlined by Belsley et al. (1980). The condition number measures the sensitivity of the regression estimates to linear dependencies among the predictors. The condition number (κ) is computed using the singular value decomposition (SVD) of the matrix X (matrix of predictors). In our analysis, matrix X includes the independent variable and the IMR from the outcome equation of the Heckman model. The decomposition expresses the matrix X as

$$X = U\Sigma V' \tag{4}$$

where U is an orthogonal matrix, Σ is a diagonal matrix containing singular values and V' is the transpose of an orthogonal matrix. The condition number is then calculated as

$$\kappa = \frac{s_{max}}{s_{min}} \tag{5}$$

where s_{max} is the maximum singular value of X and s_{min} is the minimum singular value of X. Belsley et al. (1980) suggest that a condition number and condition index above 30 indicates serious multicollinearity. The condition number is computed after estimating the Heckman model.

To measure the impact of the correlation between the regressor and the IMR on the performance of the model, two key metrics were used: bias and MSE. These metrics will help determine how accurately and precisely the Heckman model estimates the true parameters under different conditions. Bias is calculated as

$$Bias(\hat{\theta}) = E[\hat{\theta}] - \theta \tag{6}$$

where $E[\hat{\theta}]$ is the expected value of the estimator $\hat{\theta}$ and θ is the true parameter. This measures how consistently the estimator aligns with the true parameter value across simulations. A nonzero bias indicates systematic deviation of the estimate from the true parameter. The MSE sums the bias and variance of the estimator $\hat{\theta}$ capturing overall estimation accuracy. The MSE is calculated as

$$MSE(\hat{\theta}) = Var(\hat{\theta}) + Bias(\hat{\theta})^2 \tag{7}$$

where $Var(\hat{\theta})$ is the variance of the estimator and $Bias(\hat{\theta})$ is the bias calculated in equation 6.

To examine whether the Stock and Yogo (2005) framework can assess the strength of exclusion restrictions in the Heckman model, the F-statistic will be computed for the first stage of the Heckman model and compared to the Stock-Yogo weak ID F-test critical values. When the F-statistic is greater or smaller than all the critical values, the exclusion restriction is deemed strong or weak, respectively. These results will then be cross-checked with the actual strength of the exclusion restrictions used that are generated from the Monte Carlo simulations. This ensures that the Stock and Yogo framework can reliably diagnose the strength of exclusion restrictions within the Heckman model.

Results

This study first we replicate the simulation setup of Leung and Yu (1996), focusing on their experimental design involving a uniform regressor distribution and a fixed sample size 0f 1,000. Leung and Yu's original simulations used regressors drawn from a uniform distribution $U(0,3)$ to illustrate the impact of severe collinearity between the regressor and the IMR and later expanded to $U(0,10)$ to demonstrate improvement when the regressor range is widened. This study

replication follows Leung and Yu's specification closely, though differences in random seed generation naturally yield numerical discrepancies, however, the patterns remain consistent with their results. From tables 1 and 2, it can be observed that collinearity is exacerbated by a narrow regressor range and high censoring rates. Also, high collinearity between the regressor and the IMR, as indicated by high condition numbers and strong negative correlation, undermines the performance of the Heckman estimator.

Table 1. Effect of Censoring Rate on Collinearity and Estimation Accuracy under U(0,3) Regressor Distribution

Proportion of Censored Observation (%)	Corr(x, IMR)	Condition Number	Bias	MSE
10	-0.9205	13.66	-0.00505	0.00755
20	-0.9559	19.54	-0.00762	0.01514
30	-0.974	26.99	0.00009	0.03048
40	-0.9844	36.97	0.01396	0.05514
50	-0.9907	50.82	-0.00386	0.11789
60	-0.9945	70.54	-0.03180	0.32573
70	-0.997	101.46	-0.03931	0.85527
80	-0.9985	156.2	0.01848	3.08957
90	-0.9994	272.4	-0.07536	19.22484

Table 2. Effect of Censoring Rate on Collinearity and Estimation Accuracy under U(0,10) Regressor Distribution

Proportion of Censored Observation (%)	Corr(x, IMR)	Condition Number	Bias	MSE
10	-0.7005	6.087	-0.00015	0.00021
20	-0.7845	7.341	0.00047	0.00033
30	-0.8477	8.993	-0.00159	0.00051
40	-0.898	11.37	-0.00247	0.00096
50	-0.9375	15.15	-0.00532	0.00171
60	-0.9671	21.97	-0.00202	0.00418
70	-0.9858	35.45	-0.00181	0.01630
80	-0.9951	64.29	0.02115	0.09357
90	-0.9988	139.4	0.25235	2.16613

Figure 4. Relationship Between Condition Number and MSE for Different Regressor Distributions

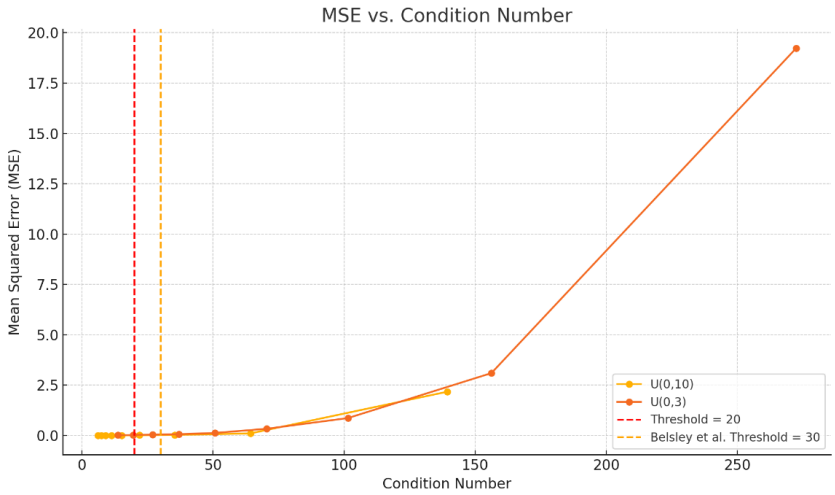


Figure 3 shows below condition number 20 MSE remains very low and stable for both distributions and continues to stay low between 20 and 30. Beyond the condition number of 30, MSE rises sharply, especially beyond 50 to 100, where the detrimental effects of collinearity on estimator precision become substantial. This indicates Leung and Yu's condition number of 20 is more conservative and an early-warning threshold, especially if the goal is to flag incipient multicollinearity before it escalates. The difference in MSE under both regressor distributions demonstrates that this threshold does not universally correspond to degradation in estimator performance. Under the $U(0,3)$ regressor distribution (Table 1), a condition number of 19.54, which is just below the threshold, corresponds to MSE of 0.01514. In contrast, under the $U(0,10)$ distribution (Table 2), a condition number of 21.97, slightly above the threshold, is associated with a considerably smaller MSE of 0.00418. This result reveals that an estimator with a condition number above the threshold of 20 (21.97 under $U(0,10)$) performs significantly better in terms of estimation accuracy than one just below the threshold (19.54 under $U(0,3)$).

When the independent variable is generated from a normal distribution, the results show overall lower condition numbers and improved model performance (in terms of MSE) compared

to the uniform distribution scenarios. At a high censoring rate of 90%, the condition numbers for the $N(0,3)$ and $N(0,10)$ cases are 17.91 and 10.24, respectively, both falling below Leung and Yu's threshold of 20 and Belsley's threshold of 30 (Tables 3 and 4). Across all censoring rates under the normal distribution, the model consistently yields MSE values below 0.0300, outperforming uniform distribution scenarios. This supports Leung and Yu's observation that increasing the variability of the independent variable reduces its correlation with the IMR. The results suggest that collinearity between the regressor and the IMR is less of a concern when the independent variable is normally distributed. However, despite a lower condition number of 17.91 under $N(0,3)$ with 90% censoring, the model's MSE is 0.02186, worse than the MSE of 0.00418 observed under the $U(0,10)$ scenario with a higher condition number of 21.97. This indicates that a model with a slightly higher condition number may perform better than one with a lower condition number, challenging the strict application of condition number thresholds which may be misleading.

Table 3. Effect of Censoring Rate on Collinearity and Estimation Accuracy under N(0,3) Regressor Distribution

Proportion of Censored Observation (%)	Corr(x, IMR)	Condition Number	Bias	MSE
10	-0.7317	5.14	0.01003	0.00123
20	-0.7247	5.07	0.00006	0.00128
30	-0.7256	5.09	-0.00398	0.00122
40	-0.7222	5.04	0.00731	0.00080
50	-0.7274	5.09	0.00000	0.00115
60	-0.7614	6.64	0.00047	0.00194
70	-0.7922	8.70	0.00178	0.00330
80	-0.8257	11.88	-0.00761	0.00545
90	-0.8617	17.91	0.00718	0.02186

Table 4. Effect of Censoring Rate on Collinearity and Estimation Accuracy under N(0,10) Regressor Distribution

Proportion of Censored Observation (%)	Corr(x, IMR)	Condition Number	Bias	MSE
10	-0.4002	3.47	-0.00001	0.00006
20	-0.4014	3.48	0.00119	0.00008
30	-0.4027	3.49	0.00002	0.00005
40	-0.4026	3.50	-0.00094	0.00007
50	-0.4033	3.49	-0.00206	0.00008
60	-0.4259	4.39	-0.00056	0.00010
70	-0.4502	5.66	-0.00149	0.00011
80	-0.4780	7.51	-0.00130	0.00025
90	-0.5191	10.24	0.00434	0.00073

Independent variables generated from the gamma distribution also produced lower condition numbers and better model performance (in terms of MSE) compared to those from the uniform distribution. However, normal distributions outperformed gamma distributions, exhibiting even lower condition numbers and MSE values. As the variability of the gamma-distributed regressor increases from the more skewed $G(1,1)$ to the more dispersed $G(5,1)$, condition numbers tend to rise, particularly at higher censoring rates. Despite this increase, model performance under $G(5,1)$ remains superior. For example, at a 90% censoring rate, the condition number under $G(1,1)$ was 19.45 with an MSE of 0.04990, whereas $G(5,1)$ yielded a higher condition number of 24.22 but a much lower MSE of 0.00920. This suggests that higher condition numbers do not always indicate poorer model performance, particularly when the regressor distribution is more dispersed.

Table 5. Effect of Censoring Rate on Collinearity and Estimation Accuracy under G(1,1) Regressor Distribution

Proportion of Censored Observation (%)	Corr(x, IMR)	Condition Number	Bias	MSE
10	-0.8511	9.44	0.00230	0.00474
20	-0.8515	9.46	-0.01060	0.00377
30	-0.8665	10.29	0.00124	0.00591
40	-0.8808	11.22	-0.00302	0.00584
50	-0.8848	11.62	0.00247	0.00557
60	-0.9049	13.44	-0.00383	0.00906
70	-0.9116	14.56	-0.01614	0.01521
80	-0.9158	15.93	-0.00341	0.01740
90	-0.9250	19.45	-0.00298	0.04990

Table 6. Effect of Censoring Rate on Collinearity and Estimation Accuracy under G(5,1) Regressor Distribution

Proportion of Censored Observation (%)	Corr(x, IMR)	Condition Number	Bias	MSE
10	-0.6047	7.28	0.00148	0.00043
20	-0.6649	8.66	0.00029	0.00054
30	-0.6959	9.74	-0.00060	0.00061
40	-0.7226	11.04	-0.00162	0.00073
50	-0.7459	12.50	0.00305	0.00118
60	-0.7655	14.00	-0.00561	0.00161
70	-0.7819	16.12	-0.00696	0.00266
80	-0.7921	18.50	0.01066	0.00544
90	-0.8248	24.22	0.01228	0.00920

When no exclusion restriction is present, relying solely on condition numbers to assess whether collinearity affects model performance can be misleading. In such cases, introducing a valid exclusion restriction can help reduce collinearity between the independent variable and the IMR, thereby enhancing the model's performance. However, the question remains, how strong must the exclusion restriction be to effectively mitigate collinearity across different regressor distributions? Figures 5 through 10 illustrate how varying levels of exclusion restriction strength

influence the correlation between the independent variable and the IMR across different distributional scenarios. The inclusion of a strong exclusion restriction consistently lowers the correlation below -0.85 across all distributions and censoring rates. Even a medium-strength exclusion restriction brings correlations below -0.90, suggesting that moderate exclusion strength may be sufficient to address collinearity concerns in most cases.

Figure 5. Impact of Exclusion Restriction Strengths on Correlation Between Regressor and IMR (U(0,3))

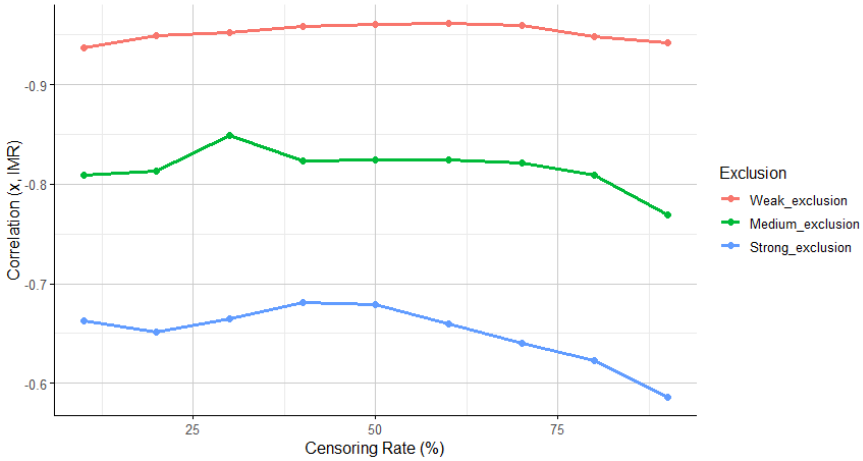


Figure 6. Impact of Exclusion Restriction Strengths on Correlation Between Regressor and IMR (U(0,10))

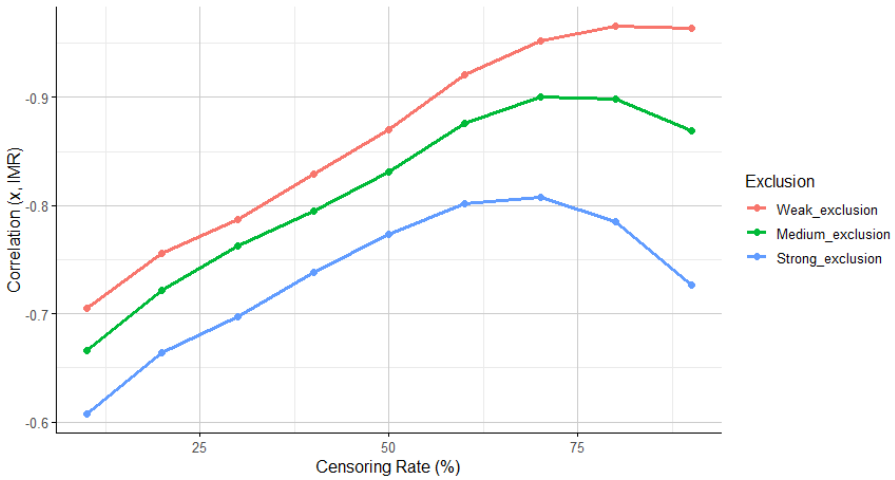


Figure 7. Impact of Exclusion Restriction Strengths on Correlation Between Regressor and IMR (N(0,3))

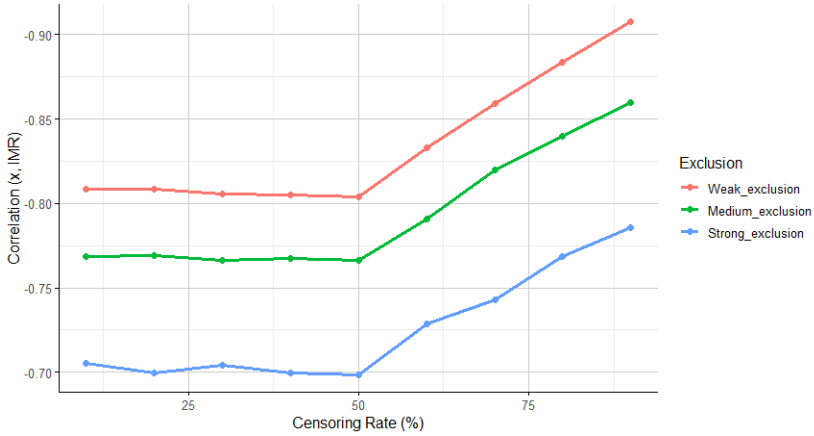


Figure 8. Impact of Exclusion Restriction Strengths on Correlation Between Regressor and IMR (N(0,10))

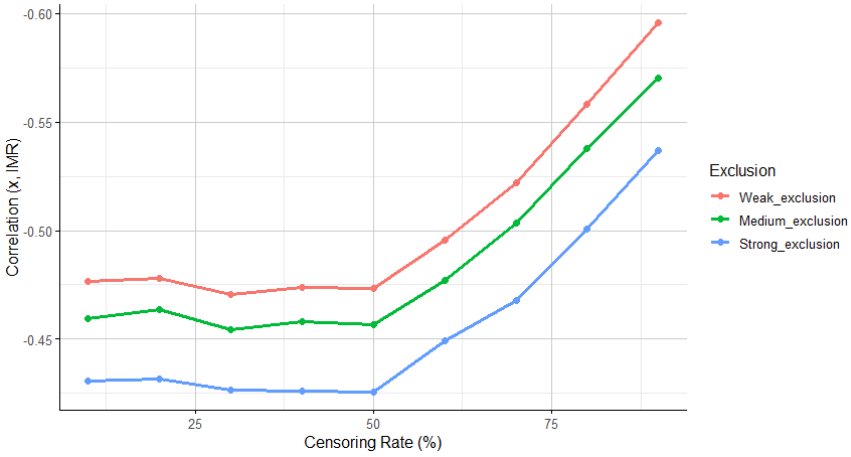


Figure 9. Impact of Exclusion Restriction Strengths on Correlation Between Regressor and IMR (G(1,1))

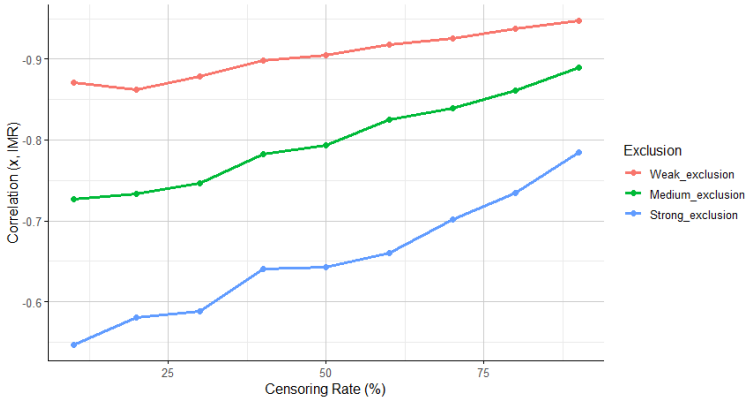
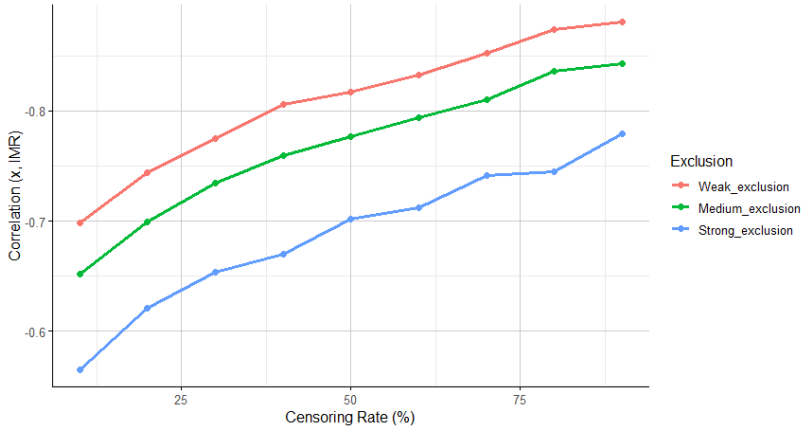


Figure 10. Impact of Exclusion Restriction Strengths on Correlation Between Regressor and IMR (G(5,1))



Conclusion

This study investigated how the performance of the Heckman selection model is affected by the correlation between the regressor and the Inverse Mills Ratio (IMR), focusing on the role of regressor distributions and the strength of exclusion restrictions. Through Monte Carlo simulations, this paper examined model performance across different censoring rates, regressor types (Uniform, Normal, Gamma), and three levels of exclusion restriction strength (weak,

medium, strong). Collinearity was measured using correlation coefficients, and condition numbers and model performance was assessed using bias and mean squared error (MSE).

The findings of this research reaffirm the work of Leung and Yu (1996) that collinearity between the regressor and IMR can severely degrade the accuracy of the Heckman estimator, particularly under narrow uniform distributions such as $U(0,3)$. Across all scenarios, high censoring rates produced extremely large condition numbers and MSE values. It was also observed that even when the narrow and broader variance distributions yield similar condition numbers, the broader variance scenarios tend to result in lower MSE values. For instance, both $N(0,3)$ and $N(0,10)$ may produce condition numbers below 20, yet the latter consistently provides more accurate estimates. This suggests that condition numbers alone may not fully capture the impact of distributional shape and spread on estimator performance.

Additionally, while Leung and Yu proposed a condition number threshold of 20 as a diagnostic benchmark, our results show that this threshold may be too rigid. In several cases, models with condition numbers just above 20 demonstrated better performance (lower MSE) than those just below the threshold. This indicates that relying solely on condition numbers to assess whether collinearity affects model performance can be misleading.

Rather than relying solely on condition number thresholds, especially in scenarios where no exclusion restriction is present and it becomes unclear whether the model is performing reliably, researchers should prioritize including a valid exclusion restriction. A strong exclusion restriction consistently reduces correlation below -0.85 across all distributional settings, while a medium-strength restriction generally brings correlations below -0.90 levels that correspond with significant improvements in MSE. These findings suggest that moderate to strong exclusion restrictions are typically sufficient to mitigate collinearity-related estimation issues.

References

Arabmazar, A., and P. Schmidt. 1982. "An Investigation of the Robustness of the Tobit Estimator to Non-Normality." *Econometrica* 50(4):1055–1063. doi:10.2307/1912776.

Bushway, S., B. D. Johnson, and L. A. Slocum. 2007. "Is the Magic Still There? The Use of the Heckman Two-Step Correction for Selection Bias in Criminology." *Journal of Quantitative Criminology* 23(2):151–178. doi:10.1007/s10940-007-9024-4.

Callaghan, K., and J. Chen. 2008. “Revisiting the Collinear Data Problem: An Assessment of Estimator ‘Ill-Conditioning’ in Linear Regression.” Practical Assessment, Research & Evaluation 13(5). Available online at https://pareonline.net/getvn.asp?v=13&n=5.

Certo, S. T., J. R. Busenbark, H.-S. Woo, and M. Semadeni. 2016. "Sample Selection Bias and Heckman Models in Strategic Management Research." *Strategic Management Journal* 37(13):2639–2657. doi:10.1002/smj.2475.

Cohen, J. 1992. “A Power Primer.” Psychological Bulletin 112(1):155–159.

Gomes, M., M. G. Kenward, R. Grieve, and J. Carpenter. 2020. "Estimating Treatment Effects under Untestable Assumptions with Non-Ignorable Missing Data." *Statistics in Medicine* 39(10):1992–2006. doi:10.1002/sim.8526.

Heckman, J. J. 1976. "The Common Structure of Statistical Models of Truncation, Sample Selection, and Limited Dependent Variables and a Simple Estimator for Such Models." *Annals of Economic and Social Measurement* 5(4):475–492.

Honoré, B. E., and L. Hu. 2022. "Sample Selection Models without Exclusion Restrictions: Parameter Heterogeneity and Partial Identification." Federal Reserve Bank of Chicago Working Paper WP 2022-33. doi:10.21033/wp-2022-33.

Lai, H. P., and W. J. Tsay. 2018. "Maximum Simulated Likelihood Estimation of the Panel Sample Selection Model." *Econometric Reviews* 37(7):744–759.

Leung, S. F., and S. Yu. 1996. "On the Choice Between Sample Selection and Two-Part Models." *Journal of Econometrics* 72(1):197–229. doi:10.1016/0304-4076(94)01720-4.

Puhani, P. A. 2000. "The Heckman Correction for Sample Selection and Its Critique." *Journal of Economic Surveys* 14(1):53–68. doi:10.1111/1467-6419.00104.

Robinson, P. M. 1982. "On the Asymptotic Properties of Estimators of Models Containing Limited Dependent Variables." *Econometrica* 50(1):27–41. doi:10.2307/1912527.

Sartori, A. E. 2003. "An Estimator for Some Binary-Outcome Selection Models Without Exclusion Restrictions." *Political Analysis* 11(2):111–138. doi:10.1093/pan/mpg001.

Stock, J. H., and M. Yogo. 2005. “Testing for Weak Instruments in Linear IV Regression.” In Identification and Inference for Econometric Models: A Festschrift in Honor of Thomas Rothenberg, edited by D. W. K. Andrews and J. H. Stock, 80–108. New York: Cambridge University Press.

Stolzenberg, R. M., and D. A. Relles. 1990. "Theory Testing in a World of Constrained Research Design: The Significance of Heckman’s Censored Sampling Bias Correction for Nonexperimental Research." *Sociological Methods & Research* 18(4):395–415.

Sundaram-Stukel, R. n.d. "Heckman-Selection or Two-Part Models for Alcohol Studies? Depends." University of Wisconsin-Madison, Department of Economics. Available online at https://arxiv.org/abs/2112.10542.

Thapa, S., Morrison, M., and Parton, K. A. 2021. "Willingness to Pay for Domestic Biogas Plants and Distributing Carbon Revenues to Influence Their Purchase: A Case Study in Nepal." Energy Policy 158:112521. https://doi.org/10.1016/j.enpol.2021.112521.

Tucker, J. W. 2011. "Selection Bias and Econometric Remedies in Accounting and Finance Research." *Journal of Accounting Literature* 29:31. Available online at https://ssrn.com/abstract=1756911.

Winship, C., and R. D. Mare. 1992. "Models for Sample Selection Bias." *Annual Review of Sociology* 18(1):327–350. doi:10.1146/annurev.so.18.080192.001551.