# Jointly Modeling the Adoption, Consumption, and Exclusive Use of Clean Cooking Fuels in Rural India

# Supporting Information

Carlos Gould
Columbia University

Xiaoxue Hou
Johns Hopkins SAIS

Jennifer Richmond
University of Maryland

Anjali Sharma
University of Maryland

Johannes Urpelainen*
Johns Hopkins SAIS

December 30, 2019

## Contents



*Corresponding author. Address: Rome Building, 4th Floor. 1619 Massachusetts Avenue, NW. Washington, DC 20036, USA. Tel: +1-734-757-0161. Email: JohannesU@jhu.edu.

# A1 Supplementary Note

## A1.1 Data

The Access to Clean Cooking Energy and Electricity – Survey of States (ACCESS) is the largest energy access survey conducted in India. A total of 8,568 households in 2015 and 9,072 households in 2018 were sampled to evaluate energy and electricity patterns over time. The increase in sample size in the second wave is attributed to an addition of 504 households in three added districts in the state of Odisha to balance the sample distribution among states. The full listing, therefore, contains 17,640 sampled households. However, for the purposes of our analysis, only 8,563 households were able to be interviewed in 2015, and only a total of 9,008 observations are able to be used for 2018 because of missing data for specific, key variables needed to perform our estimations.

The survey was designed to take 40-45 minutes to administer per household. The median time to complete the survey in the first wave was 37 minutes. The questionnaire included the following modules: socioeconomic information, a household's current source(s) of electricity (if any), satisfaction with the electricity service, cooking fuel access, household energy policy preferences, and willingness to pay for LPG or electricity.

Sampling was done using a three-stage probability-proportional-to-size (PPS) survey design. Among six energy-deprived states (Bihar, Jharkhand, Madhya Pradesh, Odisha, Uttar Pradesh, and West Bengal), a total of 51 districts, 714 rural villages, and 17,640 households were sampled for each of the three stages. Due to practical constraints, only one district was sampled from each large administrative division within a state. The number of districts sampled per state is therefore proportional to the number of administrative divisions. This is true for each state except for Odisha in the second wave, which was oversampled to include three additional districts since there are fewer administrative divisions (three) compared to the other states. Using the 2011 Indian Census, villages were divided into small and large villages in each district based on population size to be able to distinguish any differences due to the size of the community. A total of 12 households were then sampled from each of the 714 rural villages among 51 districts.

From 2015 to 2018, a total 86% of the same households were interviewed in the second wave.

| State | Responses | Retention |
|---|---|---|
| Bihar | 1,512 | 82% |
| Jharkhand | 840 | 84% |
| Madhya Pradesh | 1,680 | 77% |
| Odisha | 1,008 | 82% |
| Uttar Pradesh | 3,024 | 92% |
| West Bengal | 1,008 | 88% |
| **Total** | 9,072 | 86% |

Table A1: Retention rates for the 2018 wave are shown here for household responses in each state. Note that only 504 households were sampled and interviewed in Odisha in 2015, and sampling for Odisha doubled to 1,008 for the 2018 wave; therefore, retention is based only on the 504 households from the first wave.

Table A1 summarizes retention rates by study state. Households were identified by enumerators using a village listing. Enumerators requested heads of households to be interviewed, and if the head of the household was unavailable another willing adult was interviewed. If no adult in the household was available, or if the household was no longer willing to participate, enumerators replaced the household by interviewing the fifth household to the right of the originally sampled household. Enumerators were recruited and trained using role-playing exercises.

Questionnaires were designed to be easily understood and were tested and piloted in different settings to ensure that the instrument was effective. Regular quality checks were also done during data collection to ensure that data was coded correctly. Whereas paper questionnaires were used in 2015, digital questionnaires were used on tablets using the software program SurveyCTO for the 2018 wave. The digital survey program delivered regular, automated quality control reports, and it also made tracking the time and location of each survey more efficient. Finally, each respondent was briefed on the intent and purpose of the study as well as the nature of the questions before agreeing to participate. Written or oral consent was required from each respondent, depending on the respondent's writing ability.

## A2 Methods

First, the generalized ordered logit model (**gologit model**) is denoted as the following equation, as elsewhere[1]:

$$P(y_{it} > j) = \frac{exp(\alpha_j + X_{it}\beta_j + \Omega_{it}\beta n_j)}{1 + [exp(\alpha_j + X_{it}\beta_j + \Omega_{it}\beta n_j)]}, j = 2, 3..., M-1 \tag{1}$$

where $P(y_{it} > j)$ is the probability that the household's $y$ outcome category is greater than $j$ in time period $t$, in which $j$ is the collapsed category of $j = 1$, $j = 1, 2$, or $j = 1, 2, 3$. The $\alpha$ intercept term captures fixed effects, and differs across values of $j$ to allow for different regression lines. $M$ is the number of categories, which is four in this case. $X_{it}\beta_j$ is the main explanatory variable, logged monthly expenditure for household $i$ in wave $t$ with an estimated coefficient that may vary across values of $j$. The term $\Omega_{it}\beta n_j$ includes a set of control variables, including number of household members, education, caste, female head of household, religion, and age. The estimated coefficients for the control variables may remain constant across values of $j$ or may vary across values of $j$ depending on whether each variable meets the parallel lines assumption in the stepwise application of Wald tests in the gologit2 package in Stata. Standard errors are clustered at the village level. Fixed effects are included for the state and the survey wave (2015 and 2018).

Second, our **double-hurdle model** is denoted by the following equation[2]:

$$\begin{aligned} log(L) = \sum_{y_i=0} \Big[ log\Big\{1 - \Phi\Big(z_i\gamma, \frac{x_i\beta}{\sigma}, \rho\Big)\Big\}\Big] + \sum_{y_i>0} \Big( log\Big[\Phi\Big\{\frac{z_i\gamma + \frac{\rho}{\sigma}(y_i - x_i\beta)}{\sqrt{1-\rho^2}}\Big\}\Big] \\ - log[\sigma] + log\Big\{\phi\Big(\frac{y_i - x_i\beta}{\sigma}\Big)\Big\}\Big) \end{aligned} \tag{2}$$

where $y_i$ is either equal to 0, which would exclude the observation to the selection stage, or $y_i > 0$, which would allow the observation to impact the second stage. The first stage excludes those households without access to LPG while elevating households with LPG access to the second stage to estimate levels of consumption. This is stated as the following restriction for the double hurdle model:

$$y_i = \begin{cases} x_i\beta + \epsilon_i & \text{if } \min(x_i\beta + \epsilon_i, z_i\gamma + u_i) > 0 \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

If households select out of the model in stage one, the model for these households is essentially the same as the log likelihood function of a tobit model. For households that enter into the second stage, additional parameters are added to the tobit model to estimate the continuous outcome of LPG consumption.

## A3 Suppementary Figures

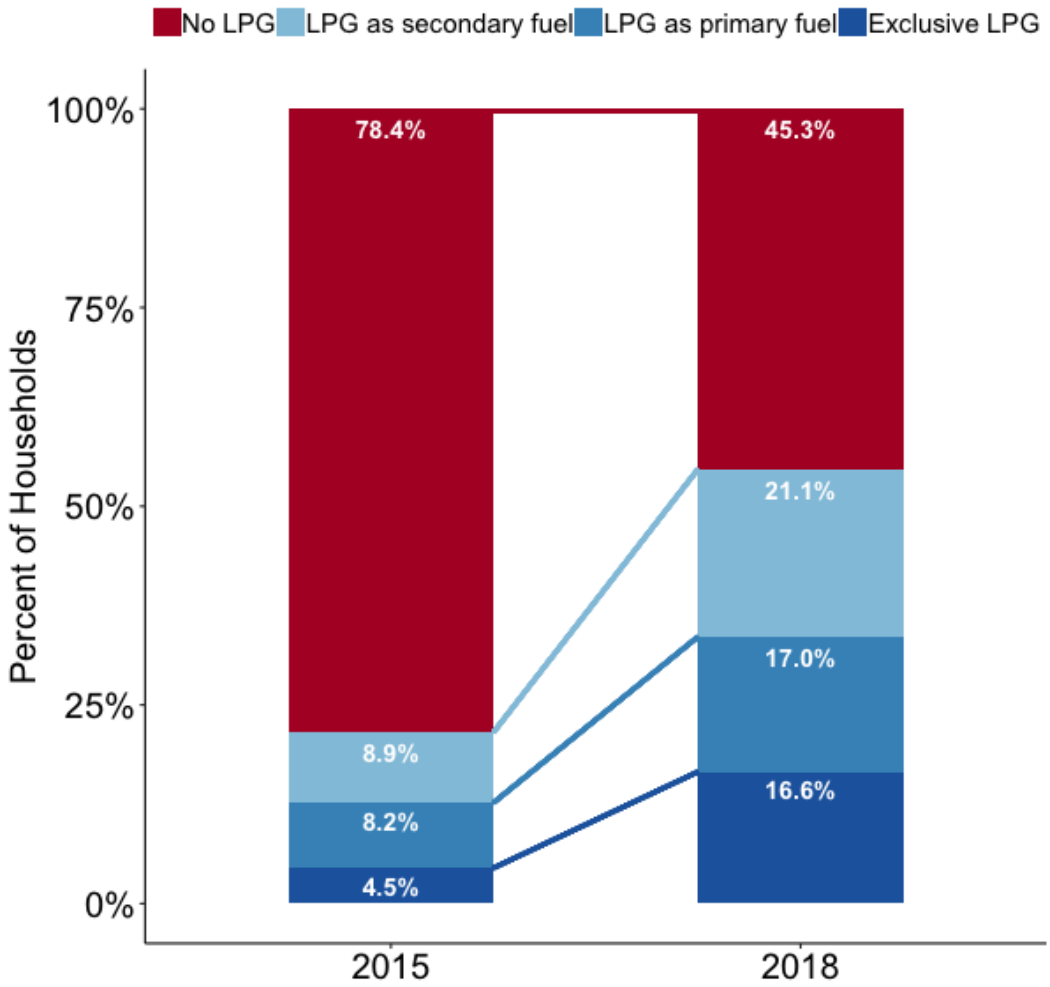


Figure A1: Shifts in cooking fuel stacking patterns from ACCESS I (2015) to ACCESS II (2018).

## A4 Supplementary Tables

| | Mean | SD | Min | Max | Observations |
|---|---|---|---|---|---|
| **Has LPG** (=1) | 0.216 | 0.412 | 0 | 1 | 8,563 |
| **Exclusive LPG Use** (=1) | 0.045 | 0.207 | 0 | 1 | 8,563 |
| **Fuel Stacking with LPG as Primary** (=1) | 0.083 | 0.275 | 0 | 1 | 8,563 |
| **Fuel Stacking with Solid Fuel as Primary** (=1) | 0.089 | 0.284 | 0 | 1 | 8,563 |
| **No LPG** (=1) | 0.784 | 0.412 | 0 | 1 | 8,563 |

Table A2: Summary Statistics of Dependent Variables (2015)

| | Mean | SD | Min | Max | Observations |
|---|---|---|---|---|---|
| **Has LPG** (=1) | 0.547 | 0.498 | 0 | 1 | 9,072 |
| **Exclusive LPG Use** (=1) | 0.166 | 0.372 | 0 | 1 | 9,072 |
| **Fuel Stacking with LPG as Primary** (=1) | 0.170 | 0.376 | 0 | 1 | 9,072 |
| **Fuel Stacking with Solid Fuel as Primary** (=1) | 0.211 | 0.408 | 0 | 1 | 9,072 |
| **No LPG** (=1) | 0.453 | 0.498 | 0 | 1 | 9,072 |

Table A3: Summary Statistics of Dependent Variables (2018)

| | Mean | SD | Min | Max | Observations |
|---|---|---|---|---|---|
| **Monthly Expenditure** (INR) | 5,300 | 3,900 | 500 | 60,000 | 8,563 |
| **Monthly Expenditure** (logarithmized) | 8.390 | 0.598 | 6 | 10 | 8,563 |
| **Household Size** | 6.740 | 3.530 | 1 | 50 | 8,563 |
| **Caste:** | | | | | 8,563 |
| Scheduled Caste | 0.183 | 0.387 | 0 | 1 | |
| Scheduled Tribe | 0.100 | 0.301 | 0 | 1 | |
| Other Backward Class | 0.477 | 0.499 | 0 | 1 | |
| General Caste | 0.240 | 0.427 | 0 | 1 | |
| **Household Head Education:** | | | | | 8,563 |
| No Formal Schooling | 0.320 | 0.466 | 0 | 1 | |
| Up To 5th Standard | 0.309 | 0.462 | 0 | 1 | |
| More Than 5th Standard | 0.371 | 0.483 | 0 | 1 | |
| **Religion:** | | | | | 8,563 |
| Hindu | 0.872 | 0.335 | 0 | 1 | 8,563 |
| Other | 0.128 | 0.335 | 0 | 1 | 8,563 |
| **Decision Maker Age** (Years) | 42.300 | 14.200 | 20 | 100 | 8,563 |
| **Decision-Maker:** | | | | | 8,563 |
| Man Household Head | 0.780 | 0.414 | 0 | 1 | |
| Woman Household Head | 0.057 | 0.231 | 0 | 1 | |
| Both Gender | 0.163 | 0.370 | 0 | 1 | |

Table A4: Summary Statistics of Independent Variables (2015)

| | Mean | SD | Min | Max | NA | Observations |
|---|---|---|---|---|---|---|
| **Monthly Expenditure** (INR) | 6,247 | 4,362 | 0 | 80,000 | 64 | 9,072 |
| **Monthly Expenditure** (logarithmized) | 8.548 | 0.640 | 0 | 10 | 64 | 9,072 |
| **Household Size** | 6.167 | 3.210 | 1 | 40 | 0 | 9,072 |
| **Caste:** | | | | | | 9,072 |
| Scheduled Caste | 0.194 | 0.396 | 0 | 1 | 0 | |
| Scheduled Tribe | 0.109 | 0.311 | 0 | 1 | 0 | |
| Other Backward Class | 0.467 | 0.499 | 0 | 1 | 0 | |
| General Caste | 0.230 | 0.421 | 0 | 1 | 0 | |
| **Household Head Education:** | | | | | | 9,072 |
| No Formal Schooling | 0.389 | 0.488 | 0 | 1 | 0 | |
| Up To 5th Standard | 0.309 | 0.462 | 0 | 1 | 0 | |
| More Than 5th Standard | 0.302 | 0.459 | 0 | 1 | 0 | |
| **Religion:** | | | | | | 9,072 |
| Hindu | 0.880 | 0.324 | 0 | 1 | 0 | |
| Other | 0.120 | 0.324 | 0 | 1 | 0 | |
| **Decision Maker Age** (Years) | 43.270 | 14.850 | 20 | 100 | 0 | 9,072 |
| **Decision-Maker:** | | | | | | 9,072 |
| Man Household Head | 0.668 | 0.471 | 0 | 1 | 0 | |
| Woman Household Head | 0.070 | 0.255 | 0 | 1 | 0 | |
| Both Gender | 0.262 | 0.440 | 0 | 1 | 0 | |

Table A5: Summary Statistics of Independent Variables (2018)

| | (1) | | | (2) | |
|---|---|---|---|---|---|
| | Two-Stage Model | | | Generalized Ordered Logit Model | |
| | Consumption | Selection | No LPG | No LPG & Stacking (LPG Secondary) | No LPG & Stacking (LPG Secondary) & Stacking (LPG Primary) |
| log (Monthly Expenditure) | 0.146*** | 0.430*** | 0.521*** | 0.543*** | 0.491*** |
| | (0.0152) | (0.0234) | (0.0254) | (0.0223) | (0.0182) |
| Household Size | 0.0132*** | -0.0152*** | 1.080*** | 1.042*** | 1.025*** |
| | (0.00253) | (0.00374) | (0.0120) | (0.00791) | (0.00634) |
| Education: Up to 5th Standard | 0.120*** | 0.275*** | 0.670*** | 0.596*** | 0.625*** |
| | (0.0223) | (0.0279) | (0.0433) | (0.0299) | (0.0285) |
| Education: More than 5th Standard | 0.231*** | 0.700*** | 0.321*** | 0.321*** | 0.321*** |
| | (0.0211) | (0.0319) | (0.0159) | (0.0159) | (0.0159) |
| Caste: OBC | -0.0968*** | -0.235*** | 1.488*** | 1.488*** | 1.488*** |
| | (0.0212) | (0.0370) | (0.0855) | (0.0855) | (0.0855) |
| Caste: ST | -0.218*** | -0.509*** | 2.192*** | 2.192*** | 2.192*** |
| | (0.0531) | (0.0682) | (0.256) | (0.256) | (0.256) |
| Caste: SC | -0.133*** | -0.362*** | 1.773*** | 1.773*** | 1.773*** |
| | (0.0268) | (0.0452) | (0.125) | (0.125) | (0.125) |
| Gender: Woman | 0.0374 | 0.192*** | 0.734*** | 0.734*** | 0.734*** |
| | (0.0298) | (0.0480) | (0.0536) | (0.0536) | (0.0536) |
| Gender: Both | 0.0298 | 0.118*** | 0.865* | 0.766*** | 0.804*** |
| | (0.0187) | (0.0292) | (0.0600) | (0.0398) | (0.0376) |
| Religion: Hindu | -0.0620* | 0.0371 | 1.067 | 1.099 | 0.950 |
| | (0.0313) | (0.0527) | (0.114) | (0.106) | (0.0836) |
| Decision Maker Age | 0.00196*** | 0.00553*** | 0.991*** | 0.991*** | 0.991*** |
| | (0.000546) | (0.000813) | (0.00125) | (0.00125) | (0.00125) |

Standard errors in parentheses, clustered by village

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table A6: Double hurdle model (1) and generalized ordered logit model (2). This table shows the regression results from our two main models. Coefficients in model (1) show the impact of variables on the two stages of LPG selection and consumption. Coefficients in model (2) show the impact of variables on three combined outcomes of LPG adoption. The coefficients for model (2) have been converted to odds ratio. We use relaxed parallel assumptions on household size, education of more than 5th standard, and caste of ST in generalized ordered logit model.

| | (1) | | | (2) | |
|---|---|---|---|---|---|
| | Two-Stage Model | | | Generalized Ordered Logit Model | |
| | Consumption | Selection | No LPG | No LPG & Stacking (LPG Secondary) | No LPG & Stacking (LPG Secondary) & Stacking (LPG Primary) |
| pmuy | -0.351*** | | | | |
| | (0.0248) | | | | |
| log (Monthly Expenditure) | 0.0850*** | 0.427*** | 0.577*** | 0.585*** | 0.511*** |
| | (0.0145) | (0.0314) | (0.0300) | (0.0251) | (0.0199) |
| Household Size | 0.0173*** | -0.0107* | 1.079*** | 1.043*** | 1.022** |
| | (0.00259) | (0.00487) | (0.0139) | (0.00869) | (0.00670) |
| Education: Up to 5th Standard | 0.0782*** | 0.293*** | 0.640*** | 0.640*** | 0.640*** |
| | (0.0225) | (0.0434) | (0.0304) | (0.0304) | (0.0304) |
| Education: More than 5th Standard | 0.126*** | 0.804*** | 0.329*** | 0.329*** | 0.329*** |
| | (0.0216) | (0.0481) | (0.0176) | (0.0176) | (0.0176) |
| Caste: OBC | -0.0878*** | -0.296*** | 1.522*** | 1.522*** | 1.522*** |
| | (0.0216) | (0.0476) | (0.0899) | (0.0899) | (0.0899) |
| Caste: ST | -0.163*** | -0.488*** | 2.068*** | 2.068*** | 2.068*** |
| | (0.0464) | (0.0819) | (0.229) | (0.229) | (0.229) |
| Caste: SC | -0.0434 | -0.573*** | 1.879*** | 1.879*** | 1.879*** |
| | (0.0267) | (0.0654) | (0.144) | (0.144) | (0.144) |
| Gender: Woman | 0.00969 | 0.212** | 0.819** | 0.819** | 0.819** |
| | (0.0300) | (0.0709) | (0.0600) | (0.0600) | (0.0600) |
| Gender: Both | 0.00734 | 0.127** | 0.920 | 0.787*** | 0.811*** |
| | (0.0182) | (0.0471) | (0.0680) | (0.0445) | (0.0414) |
| Religion: Hindu | -0.0756* | 0.184** | 1.008 | 1.020 | 0.847 |
| | (0.0306) | (0.0614) | (0.111) | (0.0982) | (0.0725) |
| Decision Maker Age | 0.000908 | 0.00652*** | 0.991*** | 0.991*** | 0.991*** |
| | (0.000516) | (0.00113) | (0.00134) | (0.00134) | (0.00134) |
| LPG Connection Years | 0.0183*** | | | | |
| | (0.00194) | | | | |
| Village Average Forest | 0.00396* | 0.0174*** | 0.974*** | 0.974*** | 0.974*** |
| | (0.00165) | (0.00435) | (0.00603) | (0.00603) | (0.00603) |
| log (Village Population) | 0.00615 | 0.129*** | 0.864*** | 0.864*** | 0.864*** |
| | (0.0133) | (0.0271) | (0.0313) | (0.0313) | (0.0313) |
| Village LPG Distance | -0.00380 | -0.0139** | 1.005 | 1.023** | 1.025*** |
| | (0.00242) | (0.00455) | (0.0102) | (0.00791) | (0.00677) |
| log (Village Town Distance) | -0.0268* | -0.0971*** | 1.138** | 1.138** | 1.138** |
| | (0.0136) | (0.0294) | (0.0471) | (0.0471) | (0.0471) |

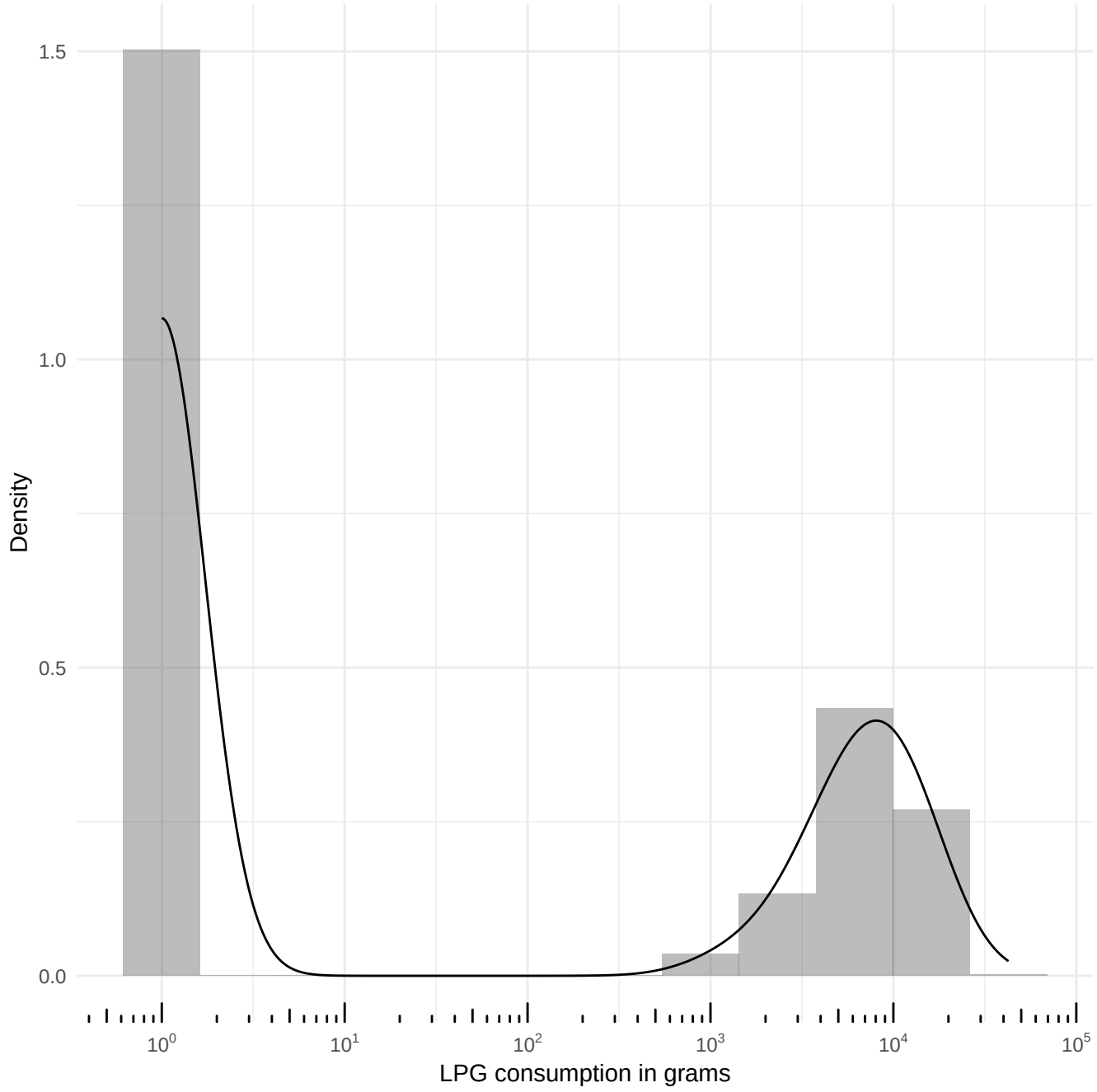


Figure A2: Distribution of household LPG consumption in the study sample. We use logarithmic scale axes and transform the data by plus one accordingly. This figure shows a zero-inflated distribution and a near-normal distribution of our dependent variable. It gives intuitions on using two-stage regression to model household consumption of LPG. The first part is a binary logit model to predict the selection of LPG and the second part is a truncated Poisson model to predict the consumption of LPG.

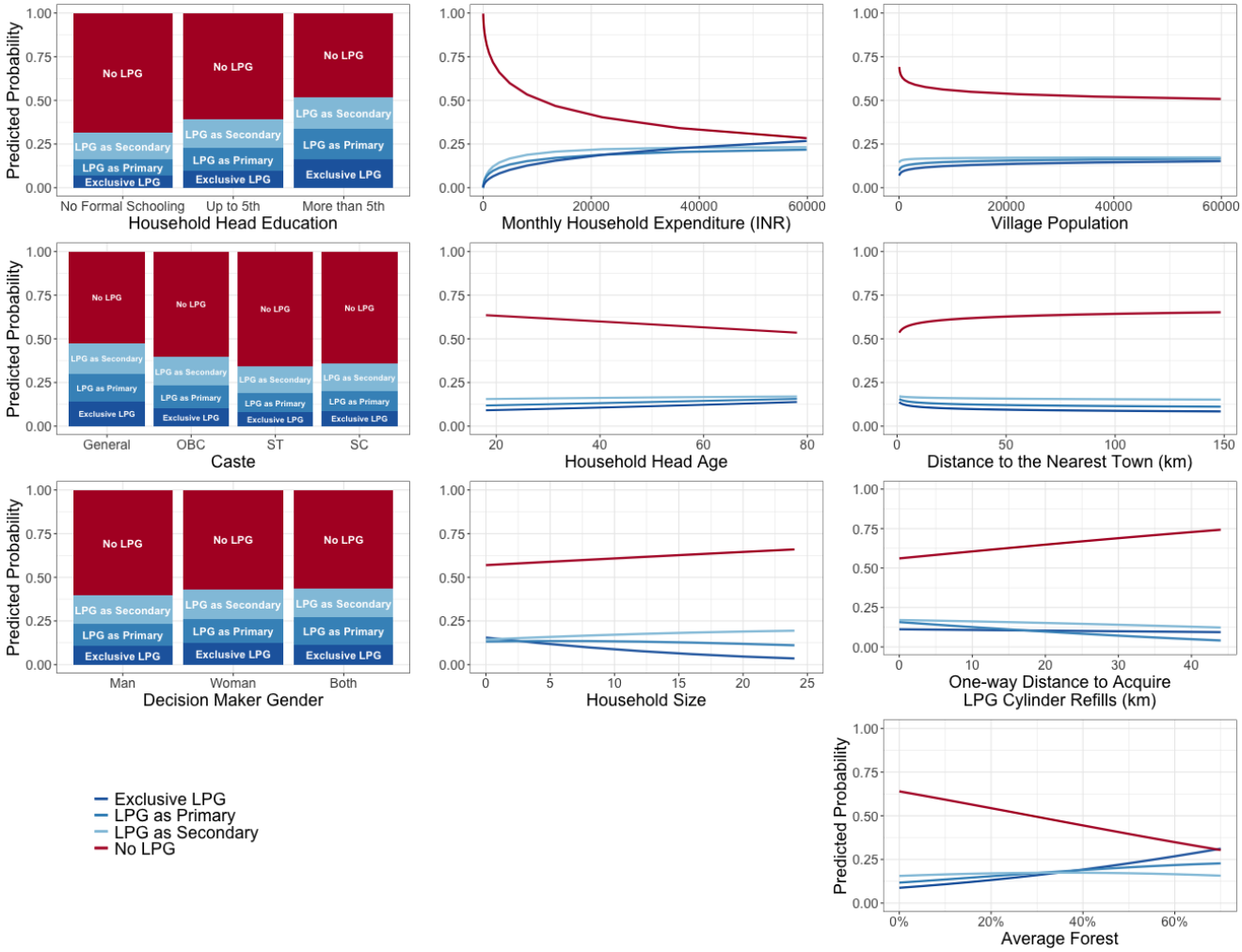


Figure A3: **Predicted probability of fuel stacking category from generalized ordered logit model with village-level covariates.** Figure shows the comparison of predicted probability (0-1) of four levels of LPG adoption: exclusive LPG use, LPG as primary, LPG as secondary, and no LPG use. The sum of the probabilities is equal to 1.

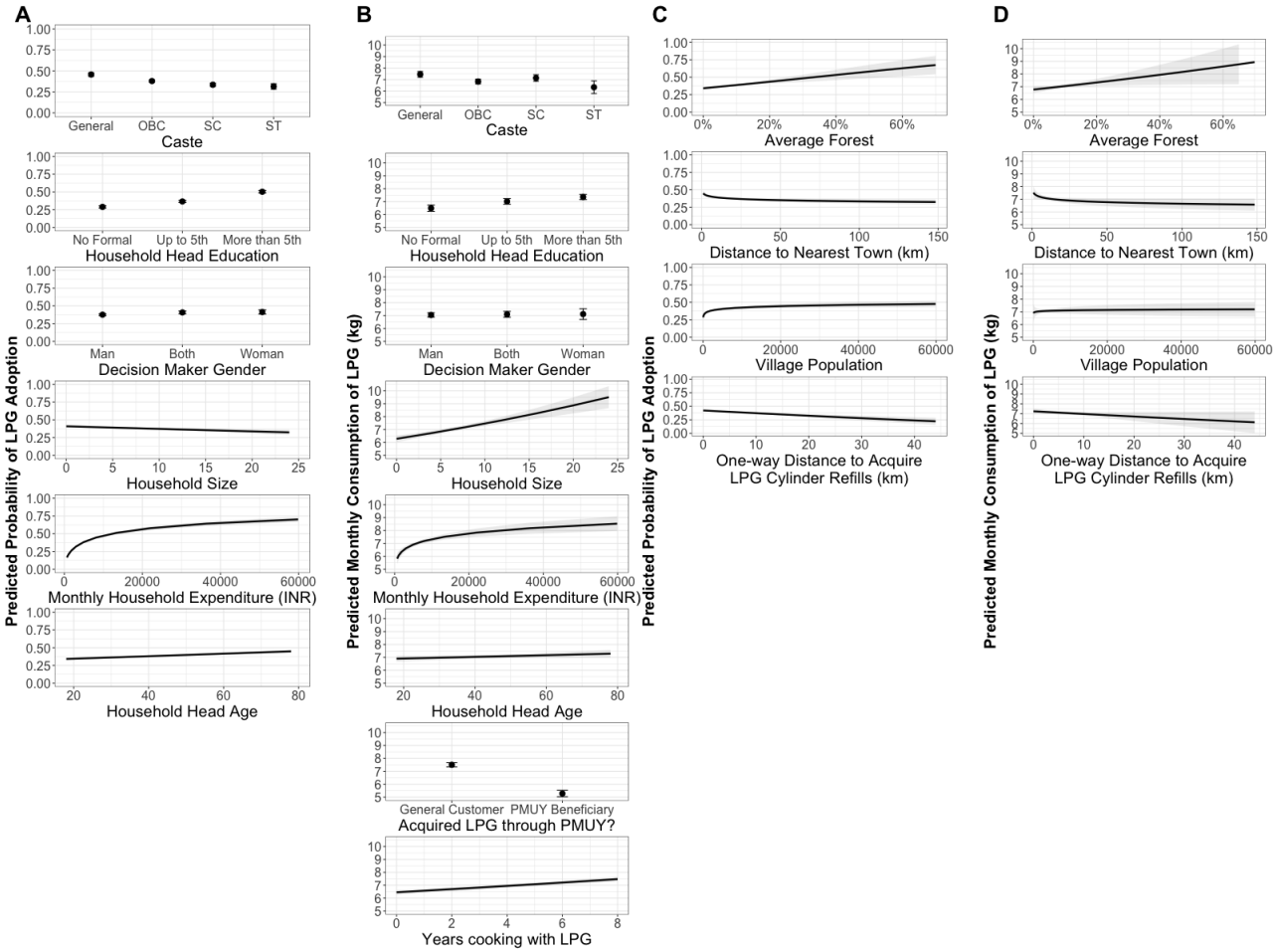


Figure A4: **Average adjusted predictions of LPG selection and consumption from two-stage double-hurdle model with village-level covariates. A.** The left panel shows the average-adjusted predicted probability of LPG adoption between 0 and 1 in the model's first stage with 95% confidence intervals. **B.** The right panel shows the average-adjusted prediction of monthly consumption – on condition of LPG adoption – in kilograms in the model's second stage. Panels **C.** and **D.** repeat the same for the village-level covariates. Standard errors in the both model stages are clustered by village.

## Supporting Information: References

[1] Williams, R. Understanding and interpreting generalized ordered logit models. *The Journal of Mathematical Sociology* **40** (2016).

[2] García, B. Implementation of a double-hurdle model. *The Stata Journal* **13**, 776–794 (2013).