The Effects of Cement Floors on Maternal Wellbeing: A Causal Analysis using Instrumental Variable Estimation

Rose Porta

2023-12-14

Background and Introduction

As outlined in the paper “Housing, Health, and Happiness”, previous research has shown that housing quality is significantly associated with health and wellbeing of families, as shelter is a foundational survival need. Despite this, there has not been much previous research specifically analysing the causal effects of housing improvements on wellbeing. There have been many government initiatives across different countries with the aim to improve housing quality for families living in poverty. In this analysis, we focus on one such government program: *Piso Firme*. *Piso Firme* had the primary goal of replacing dirt floors with cement floors in Mexico starting in the year 2000. The program was first initiated in the state of Coahuila, and we focus on this state for the analysis. The program was offered to all households with dirt floors within Coahuila, but was not offered to households in the surrounding states (yet). Thus, we employ similar families in the neighboring State of Durango as a control group. Based on this, we use whether or not the program was offered as an instrumental variable to assess the causal effect of the implementation of cement floors on the mental wellbeing of mothers. The data was sourced from the aforementioned paper, and this analysis aims to replicate a piece of the analysis presented in the paper (Titiunik *et al.*, 2007). The paper analysed the causal effect of share of cement floors on both health outcomes of children and satisfaction/wellbeing outcomes of mothers. This analysis focuses only on the outcome variables associated with mental wellbeing of mothers. Specifically, ‘wellbeing’ is measured using two different response variables: Perceived Stress Scale (PSS) and Center for Epidemiologic Studies Depression Scale (CES-D Scale).

The mechanism for causation hypothesized in the paper is that dirt floors are thought to be associated with increased risk of dangerous parasites and other health concerns for children, so we would expect replacement of dirt floors with cement floors to improve health for children. We would expect wellbeing of mothers to improve also both because the mothers would be happier seeing their children healthy and because the presence of solid floors would improve the cleanliness and overall quality of the home environment for everyone in the family (Titiunik *et al.*, 2007).

The methodology of this analysis differs from the methodology in the paper in a couple of ways. Firstly, the paper combines a regression discontinuity approach with an instrumental variable (IV) approach, while this analysis focuses only on the IV approach. Secondly, this analysis extends the statistical analysis from the paper to be more robust by employing both two-stage least squares and an inverse weighed probability (IPW) estimator, as the IPW estimator is less commonly used but better satisfies the statistical assumptions needed for the analysis than two-stage least squares. Thirdly, this analysis classifies the treatment as a binary indicator of 100 percent cement floors versus less than 100 percent cement floors. The original paper defined the treatment as the proportion of cement floors.

Thus, the causal question of interest for this analysis is, what is the causal effect of receiving completely cement floors on the mental wellbeing of mothers as measured by perceived stress and perceived depression, among mothers in families who received complete installment of cement floors due to the program?

Data

The data for this analysis was collected through a cross-sectional household survey conducted by the Mexican National Institute of Public Health during the spring of 2005. The survey was distributed to 3,000 households (evenly split between treatment and control groups), and the response rates were high, with a final sample size of 2,755 (Titiunik *et al.*, 2007, p. 9).

Our response variables of interest are Perceived Stress Scale (PSS) and Depression Scale (CES-D). The PSS measure is on a scale of 0-40, with a higher score indicating higher perceived stress. The questions incorporated into the score aim to measure the degree to which the mothers perceived their lives to be unpredictable, uncontrollable and overloaded during the most recent month. The CES-D measure is on a scale of 0-60, with a higher score indicating higher perceived depression. The questions incorporated into the depression score aim to measure the degree to which the mothers are experiencing symptoms of depression (Titiunik *et al.*, 2007, p. 11).

Our treatment variable indicating installment of cement floors is defined in the data as “share of cement floors” as a proportion. However, in this analysis, we have collapsed the treatment into a binary indicator variable where the value 0 means the household less than 100 percent cement floors post-treatment and the value 1 means the household has 100 percent cement floors post-treatment. This choice was made because when looking at summary statistics and exploratory analysis of the "share of cement floors" variable, we see that the distribution is very skewed toward higher proportions, with more than 50 percent of the values equal to 1 (100 percent). Thus, it seemed more useful for the analysis to convert the treatment to a binary variable.

Our instrumental variable indicating whether or not the household was offered the program is a binary variable where the value 0 means the household was not offered the program and the value 1 means the household was offered the program.

The data also includes several covariates which we have included in the analysis. There are 19 covariates total which describe the demographic characteristics of the household such as the total number of members and the proportion of household members within several age-gender strata as well as other characteristics which may impact household well being such as water and electricity access. The data originally included 2,783 observations, but after removing rows with missing values for any of the included variables, the number of rows reduced to 2,758. The full list of included covariates as well as plots visualizing the distributions of the main variables of interest are in the Appendix.

Causal Methodology and Assumptions For IV Analysis

For this section, we will use the following notation: Z represent the the IV (0 = program not offered, 1 = program offered); A represents the binary treatment (0 = less than 100 percent cement floors, 1 = complete cement floors); Y represents the response variable (maternal stress or depression score); W represents a vector of covariates.

Definition of Potential Outcomes

Before presenting the identification of the causal effect, it is necessary to define the potential outcomes for the IV Model. In the IV model, there are two sets of potential outcomes: potential outcomes for the treatment A and potential outcomes for the response Y. The potential outcomes for treatment A have the form $A_i^{Z=z}$ where $z \in \{0, 1\}$ and i represents a particular participant. These potential outcomes represent the treatment status under theoretical assignment of Z = z. For each participant, we would only observe one potential outcome in practice, but in identifying a causal effect, it is necessary to define both "counter-factual" outcomes in order to estimate what may happen on average if we were to assign all participants to a particular value of Z. In this study, we can interpret the potential outcomes of A as the value of complete cement floors (yes/no) which would have been observed for any given participant if we had assigned that participant to not be offered the program (Z = 0) or be offered the program (Z = 1).

The potential outcomes for response Y have the form $Y_i^{Z=z,A=a}$ where $z \in \{0, 1\}$, $a \in \{0, 1\}$, and i represents a particular participant. We interpret the potential outcomes for Y as the outcome value which would have been observed for any particular participant if we could set both the IV and the treatment status to a particular value. Since we are considering manipulation of both Z and A, there will be four potential outcomes for response Y. In the context of this study, we can interpret the potential outcomes of Y as the value of depression score or stress score which would have been observed for any given participant if we had assigned that participant to not be offered the program (Z = 0) or be offered the program (Z = 1), and then to receive complete cement floors (A = 0) or not (A = 1). Again, we will only observe one potential outcome for Y for each participant, but we define all counterfactual outcomes in order to estimate the causal effect.

Causal Assumptions

In order to identify the causal effect using the IV model, we assume the following causal assumptions.

A1: There is only one version of treatment, which is well defined. In an IV model, this means that all combinations of the IV status and the treatment status are well defined. In this context, this means that all four of the following cases must be clearly defined:

1. Z = 0, A = 0; case where an individual is not offered the program and they do not receive complete cement floors.
2. Z = 0, A = 1; case where an individual is not offered the program and they do receive complete cement floors
3. Z = 1, A = 0; case where an individual is offered the program and they do not receive complete cement floors.
4. Z = 1, A = 1; case where an individual is offered the program and they do receive complete cement floors.

A2: There is no interference, i.e. the treatment status of one participant (receive complete cement floors) has no impact on the outcomes of any other participants.

A3: There is consistency. This means that we have accurately recorded the values of the IV (offered the program) and the treatment (receive complete cement floors) for each participant in the data, and the observed values of both the treatment and the response (depression or stress score) correspond to the potential outcomes that we would expect assuming there are no data entry errors. For example, consistency says that if $Z_i = 0$, then $A_i = A_i^{Z=0}$. An example for the response is, if $Z_i = 0$ and $A_i = 1$, $Y_i = Y_i^{Z=0,A=1}$. For both the treatment and response, the corresponding conditions must hold for each combination of observed values for Z and A.

A4: There is positivity, i.e. $0 < P(Z_i = 1|W_i) < 1, \forall i$. This means that conditional on observed confounders, all participants have a non-zero probability of being exposed to the IV, which is being offered the program. In this context, this would mean that all participants (within each subset of categories of W) have a non-zero probability of being offered the program.

A5: There is no unobserved confounding between the IV and the treatment or between the IV and the outcome. In this context, this would mean that there are no unobserved common causes of being offered the program and whether or not the family receives complete cement floors, and likewise there are no unobserved common causes of being offered the program and perceived stress/perceived depression score. It is okay to have unobserved confounding between whether or not the family received complete cement floors and perceived stress/perceived depression score.

A6: IV relevance. being offered the program must be a direct cause of receiving complete cement floors, i.e. $P(A^{Z=1}) \neq P(A^{Z=0})$.

A7: There is no direct effect of the IV on the outcome. In this context, this means that there is no direct effect of being offered the program on perceived stress/perceived depression score. We can also frame this as the idea that being offered the program only has an effect on perceived stress/perceived depression through whether or not a family receives complete cement floors.

A8: Either monotonicity (no defiers) or homogeneity. Both could be reasonable, but for this analysis I will assume monotonicity, which would mean that we are assuming there is nobody who would do the opposite of what they were assigned based on their IV status. i.e., there are no individuals who would receive complete cement floors if not offered the program and would not receive complete cement floors if offered the program.

Plausibility of Assumptions

A1 is satisfied because all IV-treatment combinations are well defined.

There could be concerns with A2 (interference) because the families being offered the treatment all live in the same state, so it is possible that they could impact each other in various ways. For example, suppose one family receives the cement floors and their next-door neighbors do not. Then, suppose that the presence of cement floors does lower the stress of the mother for the family who recieved the treatment, and because that mother is less stressed, she is more friendly to the family next door, thus causing the stress of the neighboring mother to also decrease even though she did not directly get cement floors. Thus, we should be skeptical about A2.

A3 is reasonable, as we are assuming all data is recorded correctly.

A4 is reasonable as long as we assume that all strata of the covariates include at least one family in the control state (Durango) and one in the treatment state (Coahuila).

A5 could definitely have some concerns. There were likely several reasons why the Mexican government decided to implement the treatment in Coahuila first, and it is plausible that some of these reasons are unobserved and may also impact maternal wellbeing (the outcome). For example, maybe supplies to make the floors are more readily available in Coahuila. If this is the case, maybe more of the public buildings in Coahuila would have already had cement floors prior to the program, such as the children's schools. In this case, maybe the families in Coahuila would have better wellbeing metrics partly due to the cement floors in schools in addition to the effect from the cement floors being installed in homes. This same possible unobserved confounder of availability of cement could also be confounding the relationship between the IV and the treatment. If there is more access to cement in Coahuila, the families there may be more likely to install cement floors regardless of whether or not they were offered the program. Thus, we should be skeptical of A5.

A6 is reasonable given that unless the offer in the program is drastically unappealing, we would assume that being offered the program would increase a family's probability of receiving cement floors on average. We will further verify this assumption based on the data in the following sections.

A7 is reasonable because there is no obvious way that simply being offered the program could directly affect the wellbeing of families.

There could be concerns with A8 (monotonicity), although it is fairly reasonable. If we assume that people not offered the program do not have access to cement floors at all, then this assumption would be fully verified, as the design would mimic one-sided non-compliance (i.e., those offered the program could refuse, but those not offered the program could not receive cement floors by any other means). However, this is likely not the case. We would need to know more about the context in order to know how plausible it would be that people may be installing cement floors who are not offered the program. Although monotonicity has some concerns, we choose to assume monotonicity over homogeneity because assuming homogeneity may have even more serious concerns. One factor which could violate homogeneity is share of cement floor prior to the intervention of the program, which we are not accounting for. It is very plausible that families who started with 0 percent cement floors and received 100 percent cement floors through the treatment may benefit much more on average than families who started with 80 percent cement floors and increased to 100 percent after the treatment. In this case we do not have pre-post data, so there is no way to know how much influence prior share of cement floors may have.

Although we have identified a few concerns with the assumptions, we will proceed with the analysis assuming A1-A8, and under these assumptions, we can identify the Local Average Treatment Effect (LATE), meaning the average treatment effect among compliers (people who would receive cement floors if offered the program and would not receive cement floors if not offered the program). We estimate the LATE using two different statistical methods: two-stage least squares and an Inverse Probability Weighted (IPW) Estimator. We present the two-stage least squares because this is the most common technique in practice, and it is more intuitive to understand than IPW. However, we also present the IPW estimate because it is more statistically sound, as described further in the methods section below.

Statistical Methods

Under A1-A8 described above, we identify the theoretical LATE as follows

$$E[Y^{A=1} - Y^{A=1} | A^{Z=1} > A^{Z=0}] = \frac{E[E(Y|Z=1, W) - E(Y|Z=0, W)]}{E[E(A|Z=1, W) - E(A|Z=0, W)]}$$

Where Y represents the response, A represents the treatment, Z represents the IV, and W represents a vector of covariates. In plain language, the above formula says that the average causal effect of the treatment on the outcome among compliers is identified to be the ratio of the average effect of the IV on the outcome to the average effect of the IV on treatment, conditional on covariates W.

We use two approaches to estimating this causal effect as described below.

Two-Stage Least Squares

The two-stage least squares approach proceeds as follows.

First, we estimate the effect of Z on A conditional on covariates using a main effects linear regression model of treatment A on IV Z and covariates W.

$$E(A|Z, W) = \beta_0 + \beta_Z Z + \beta_W^T W$$

When we fit the model, we get fitted values for A defined as follows.

$$\hat{A} = \hat{\beta_0} + \hat{\beta_Z} Z + \hat{\beta_W^T} W$$

Next, we estimate the causal effect of A on Y given covariates using a second main effects linear regression of Y on the fitted values $\hat{A}$ from the first regression model and covariates W.

$$E(Y|\hat{A}, W) = \beta_0 + \beta_A \hat{A} + \beta_W^T W$$

And we fit the model as follows.

$$\hat{Y} = \hat{\beta_0} + \hat{\beta_A} \hat{A} + \hat{\beta_W^T} W$$

Based on this second fitted model defined above, the coefficient estimate for β_A, $\hat{\beta_A}$, represents our estimate of the LATE of the treatment A on the outcome Y.

This two-stage least squares approach is popular both because it is intuitive to understand and because it allows us to easily compute confidence intervals and p-values for the estimate based on the estimate of the standard error of β_A produced by the linear regression output. Further, the first-stage model provides a means for estimating the causal effect of the IV on the treatment, which can be used to assess the IV relevance assumption. In that first stage model, the coefficient estimate $\hat{\beta_Z}$ represents our estimate of the effect of IV on treatment conditional on covariates, and this can also be interpreted to be an estimate of the proportion of compliers.

Despite these advantages of the two-stage least squares approach, it is important that we also consider the key assumptions of linear regression which should hold when applying this approach. One central assumption of linear regression is that the errors should be approximately normally distributed. This assumption generally holds when the response variable is continuous (or at least numeric with a wide range of possible values) and our sample size is large enough, but it cannot hold when the response variable is binary or categorical. In our case, both of our outcome variables are numeric and roughly normally distributed, but the treatment A is binary. Thus, the first stage model which uses A as the response variable does not fit the assumptions of linear regression. For this reason, we will implement a second approach to estimation which is more statistically sound in the following section.

We note that in the case of an unadjusted analysis with no covariates, two-stage least squares is a correct specification regardless of whether the treatment and outcome are continuous or binary, as it serves simply as a "computational trick" in this scenario and does not actually require linear regression assumptions.

Inverse Probability Weighted Estimator (IPW)

The IPW approach estimates the causal effect of A on Y using the following formula:

$$\hat{\theta_{IPW}} = \frac{\frac{1}{n}\sum_{i=1}^{n} Y_i[\frac{Z_i}{\hat{\gamma}(W_i)} - \frac{1-Z_i}{1-\hat{\gamma}(W_i)}]}{\frac{1}{n}\sum_{i=1}^{n} A_i[\frac{Z_i}{\hat{\gamma}(W_i)} - \frac{1-Z_i}{1-\hat{\gamma}(W_i)}]}$$

Where $\gamma(W_i) = P(Z = 1|W)$ and $\hat{\gamma}(W_i)$ is estimated using a main effects logistic regression of IV Z on covariates W.

$$logit(P(Z = 1|W)) = \beta_0 + \beta_W^T W$$

Since the IV Z is binary (offered program/not offered program), logistic regression is a correct specification for modeling this propensity score.

This approach does not directly produce a standard error estimate for the causal effect of A on Y, but we can compute confidence intervals for the effect via a bootstrap approach.

Results

Effect of Offering the Program on Cement Floors Installment

Based on fitting a linear regression model of the treatment, complete cement floors, on the IV, offered program, and covariates (first stage of two-stage least squares), we estimate the causal effect of the IV on the treatment conditional on covariates to be 0.25 (95 percent CI: (0.22, 0.28)). We can also interpret this estimate as the proportion of compliers. Since 0 is not included in the confidence interval, we have evidence that there exists a significant causal effect of offering the program on receiving complete cement floors conditional on covariates, and thus the IV relevance assumption is reasonable.

It is notable that based on this regression of cement floors on program offered or not and covariates, we find many of the covariates to be significant predictors of cement floors conditional on the IV. Specifically, we find p-values for coefficient estimates <0.05 for number of household members, spouse age, number of years of schooling for head of household, water connection, water connection in house, animals allowed inside, use of garbage collection services, and number of times respondent washed hands the day before. We find p-values <0.1 for the coefficient estimates for all variables relating to gender-age distributions of household members as well as the age of the household head.

Despite our finding that many of the covariates are significantly associated with cement floors, the estimate for the the effect of the IV on treatment (proportion of compliers) not adjusting for covariates is very similar to the estimate found in the adjusted model. The unadjusted estimate is 0.24 with 95 percent CI (0.20, 0.27). This indicates that although these covariates are correlated with cement floors, they are likely not strongly correlated with the IV (program offered). Further, this indicates that the researchers who collected the data were roughly successful in drawing a sample from the control region which is similar to the sample from the treatment region with respect to these observed covariates.

However, it is important to keep in mind that the treatment (complete cement floors) is a binary variable, which violates the assumptions of linear regression for the model of the treatment on the IV and covariates. Thus, we may not fully trust the findings based on this model. Unfortunately, this is the only model that we have for assessing the relationships between the covariates and the treatment, as the IPW approach only provides an estimate of the local average treatment effect without any of the additional information that a regression model offers.

Unadjusted Analysis

We focus on estimation of the causal effect adjusting for covariates, but first we present the results of an unadjusted analysis (not adjusting for any covariates) for reference for comparison. We used a two-stage least squares technique for the unadjusted analysis, but as mentioned above, this technique is only a computational trick and does not rely on linear regression assumptions for the unadjusted case.

Using an unadjusted IV analysis, we find statistically significant evidence of an average causal effect of completely cement floors on both maternal depression score and maternal perceived stress score among participants who received cement floors due to the *Piso Firme* program (compliers). We estimate that the average effect of the the program on maternal depression score was -9.22 (95% CI: (-12.0, -6.4); p = 0). That is, the average depression score of mothers who received 100 percent cement floors was 9.22 units lower than it would have been had they not received the cement floors. We estimate that the average effect of the the program on maternal stress score was -7.26 (95% CI: (-9.35, -5.18); p = 0). That is, the average perceived stress score of mothers who received 100 percent cement floors was 7.26 units lower than it would have been had they not received the cement floors.

Adjusted Analysis: Two-Stage Least Squares

Using an adjusted IV analysis via two-stage least squares estimation, we find statistically significant evidence of an average causal effect of completely cement floors on both maternal depression score and maternal perceived stress score among participants who received cement floors due to the *Piso Firme* program (compliers). We estimate that the average effect of the the program on maternal depression score was -9.52 (95% CI: (-12.17, -6.87); p = 0). That is, the average depression score of mothers who received 100 percent cement floors was 9.52 units lower than it would have been had they not received the cement floors, adjusting for covariates. We estimate that the average effect of the the program on maternal stress score was -7.16 (95% CI: (-9.18, -5.13); p = 0). That is, the average perceived stress score of mothers who received 100 percent cement floors was 7.16 units lower than it would have been had they not received the cement floors, adjusting for covariates.

Further, we find that several of the covariates relating to household environment/availability of resources seem to be significant predictors of maternal depression and stress, while demographic characteristics tend not to show up as significant predictors. Specifically, the only covariates with significant p-values (<0.05) based on the linear regression model output for stress score were garbage collection (uses garbage collection services yes/no), animals allowed inside the house (yes/no), and water connection inside the house (yes/no). All three of these same covariates were also significant in the model for depression score, and there was one additional significant covariate for depression score, which was water connection (access to water but not inside the house, yes/no).

Adjusted Analysis: IPW Estimator

Using an adjusted IV analysis via IPW estimation, we find statistically significant evidence of an average causal effect of completely cement floors on both maternal depression score and maternal perceived stress score among participants who received cement floors due to the *Piso Firme* program (compliers). We estimate that the average effect of the the program on maternal depression score was -9.28 (95% CI: (-12.21, -6.46)).

That is, the average depression score of mothers who received 100 percent cement floors was 9.28 units lower than it would have been had they not received the cement floors, adjusting for covariates. We estimate that the average effect of the the program on maternal stress score was -6.93 (95% CI: (-9.14, -4.75)). That is, the average perceived stress score of mothers who received 100 percent cement floors was 6.93 units lower than it would have been had they not received the cement floors, adjusting for covariates.

Unfortunately, the IPW approach does not provide a means for assessing the relationships between individual covariates with the treatment and the response (respectively) like the two-stage least squares approach does.

The results of all three estimation approaches are summarized in Table 1 and Table 2 below.

Table 1: Estimates of Causal Effects for Maternal Depression Score

Method	Estimate	CI	p_value
Unadjusted	-9.22	(-12.0, -6.4)	0
Adjusted Two-Stage Least Squares	-9.52	(-12.17, -6.87)	0
Adjusted IPW	-9.28	(-12.21, -6.46)	–

Table 2: Estimates of Causal Effects for Maternal Stress Score

Method	Estimate	CI	p_value
Unadjusted	-7.26	(-9.35, -5.18)	0
Adjusted Two-Stage Least Squares	-7.16	(-9.18, -5.13)	0
Adjusted IPW	-6.93	(-9.14, -4.75)	–

Conclusion

Overall, we have evidence that there is a statistically significant average causal effect of complete cement floors on both perceived depression and perceived stress among mothers of families who received complete cement floors due to the *Piso Firme* program. Specifically, we find that complete cement floors reduces perceived depression score by about 9 points on average and reduces perceived stress score by about 7 points, which are quite large effect sizes considering that the scales are 0-60 and 0-40 respectively. Further, the confidence intervals for all three methods were nowhere near including zero, and for the approaches where obtaining p-values was possible, the p-values were all extremely small (not near the borderline of significance).

Interestingly, we find that all three statistical approaches to estimating the causal effect (unadjusted, adjusted two-stage least squares, and adjusted IPW) all yield very similar results, and the overall conclusion is the same for all methods. Based on the two stage least-squares approach, we found that only four of the 19 covariates were significantly associated with maternal stress and depression scores, so this may partially explain why the result of the adjusted analysis is not much different from the adjusted analysis. The fact that the IPW analysis yielded very similar estimates to the two-stage least squares approach suggests that the two stage least-squares approach is still reasonably valid for this particular data despite the mispecification of the model with the binary treatment as the response variable.

Based on the agreement of all three estimation methods and the large estimated effect sizes, we have very strong evidence of an average causal effect among compliers. However, it is important to highlight the limitations and cautions of our analysis. As described in the assumptions section, we have some serious concerns with a few of the causal assumptions, namely interference and unobserved confounding of the IV/treatment and IV/response relationships. We can be reassured by knowing that our results are consistent with the results found in the reference paper, and we actually found effect sizes slightly smaller than those found in the paper (estimates reported in paper are about -11 for depression and -9 for stress) (Titiunik *et al.*, 2007, p. 36 (table IX)).

A further step which could help to verify the robustness of our findings would be to conduct a sensitivity analysis to assess how much unobserved confounding between the IV/treatment and IV/response relationships there would need to be to explain away the estimated effect found here. If a new study were to be conducted and pre/post data could be collected, another approach could be a difference in difference causal analysis which could help to account for initial conditions such as pre-treatment share of cement floors.

Beyond the concerns with violations of assumptions, it is important to highlight that using the IV model under the monotonicity assumption, we are only able to estimate the local average treatment effect, meaning the treatment effect among compliers. Thus, we cannot be sure that this causal effect would apply to everyone in the population. There could be systematic differences between compliers and non-compliers which are impossible to identify. The two-stage least squares approach allows us to estimate the proportion of compliers, but it does not give us any way of knowing or estimating the characteristics of compliers as compared to non-compliers. For example, the reference paper mentioned that there was a small monetary fee associated with participation in the program. Although we would assume that the fee is affordable for most people given that the program is aimed at improving quality of life for people living in poverty, it could be the case that the cost of the program is still unaffordable for the poorest families. If this were the case, it may be true that compliers are significantly wealthier than non-compliers on average, and thus our conclusions may only apply to families of the same wealth status as compliers. We would need more context on the specifics of the program implementation in order to assess how much of a concern systematic differences between compliers and non-compliers may be, yet it is important to keep this potential limitation in mind.

Overall, this analysis finds significant effects of cement floors on maternal stress and depression among participants who received cement floors due to the *Piso Firme* program, which agrees with the findings in the reference paper. Yet, there are still limitations of the analysis and further research which could be done in order to further verify these findings.

Appendix

Exploratory Data Analysis

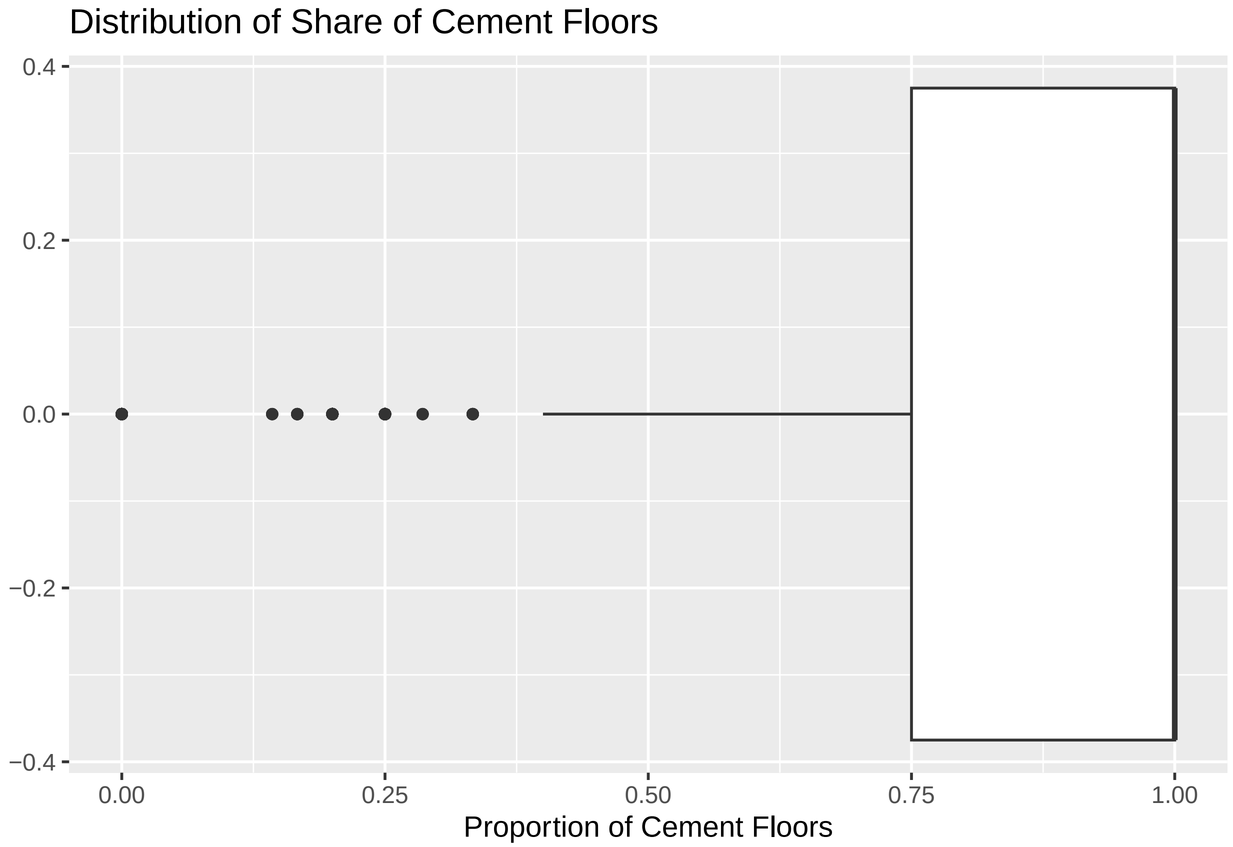
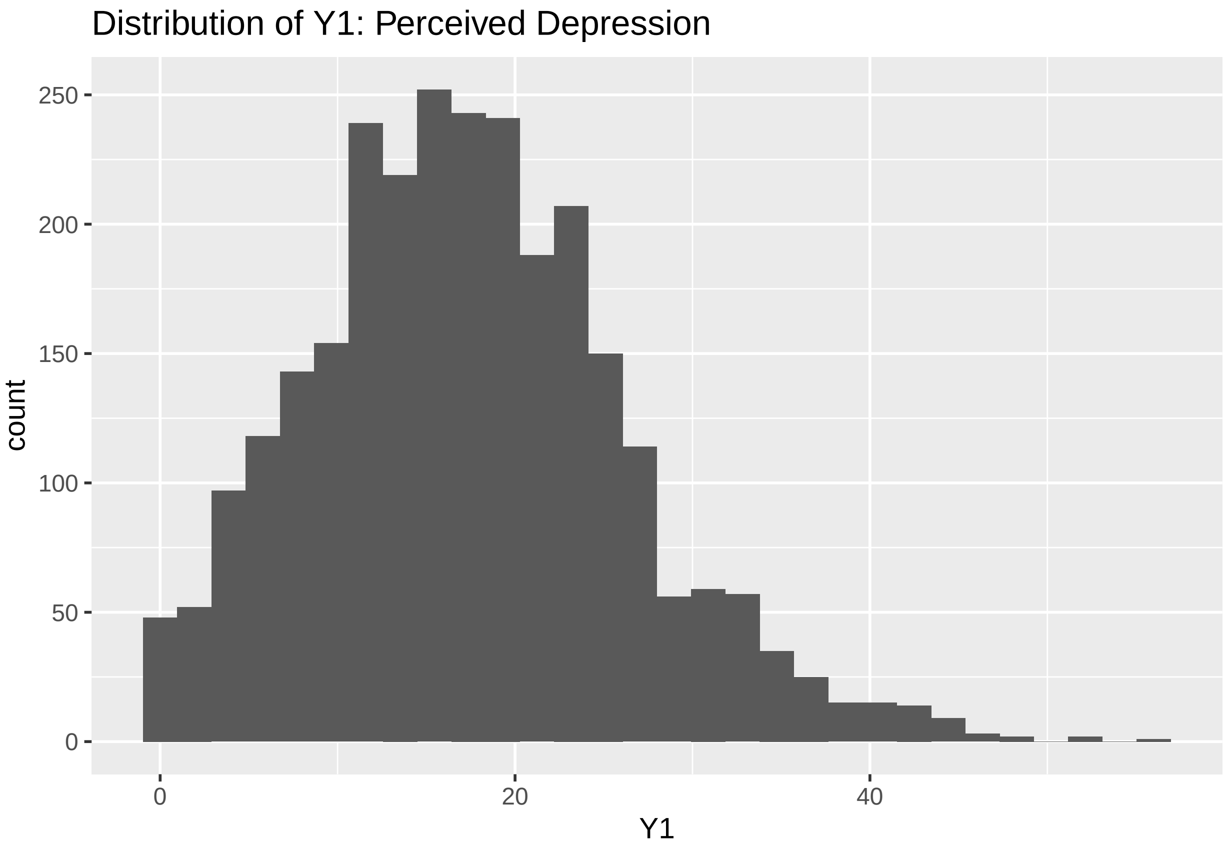


Table 3: Distribution of Treatment: Complete Cement Floors

A_bin	count	percent
0	924	33.5
1	1834	66.5

Table 4: Distribution of IV: Offered Program

Z	count	percent
0	1381	50.1
1	1377	49.9



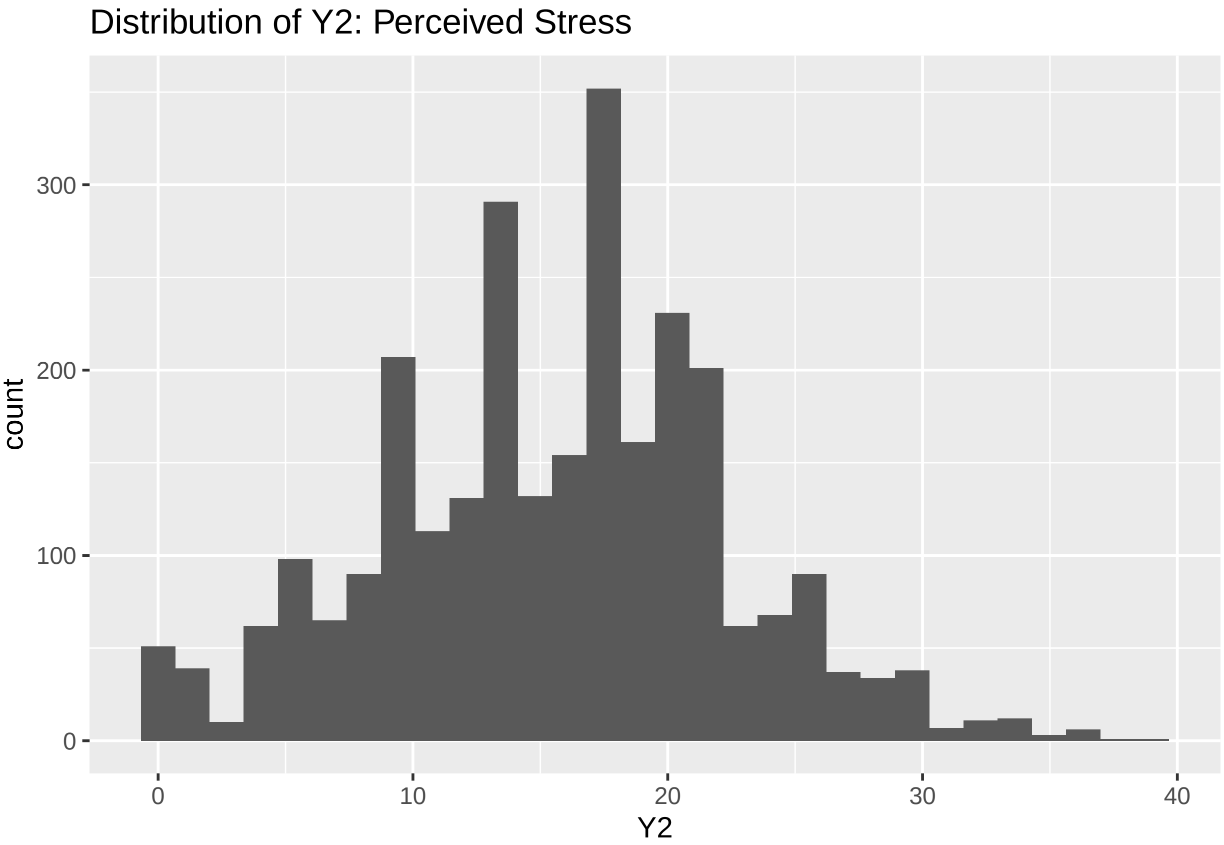
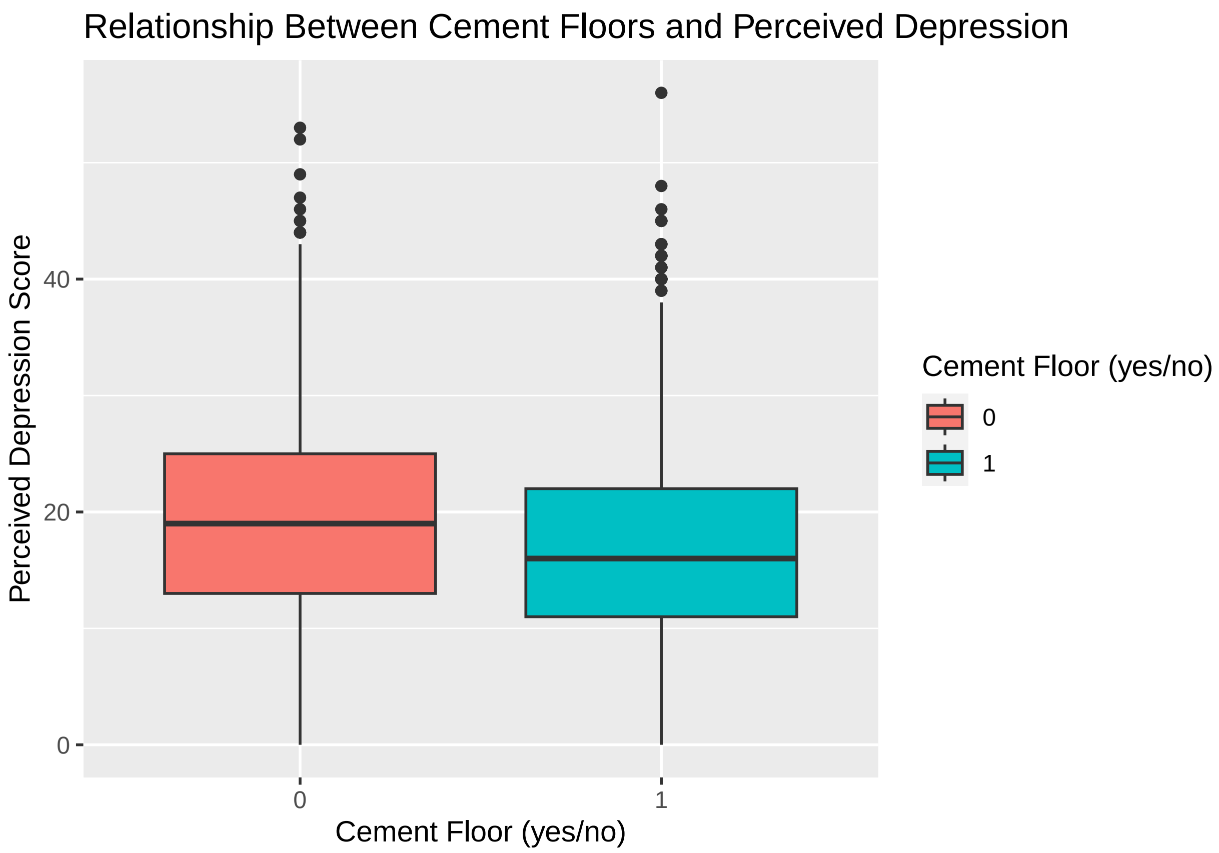
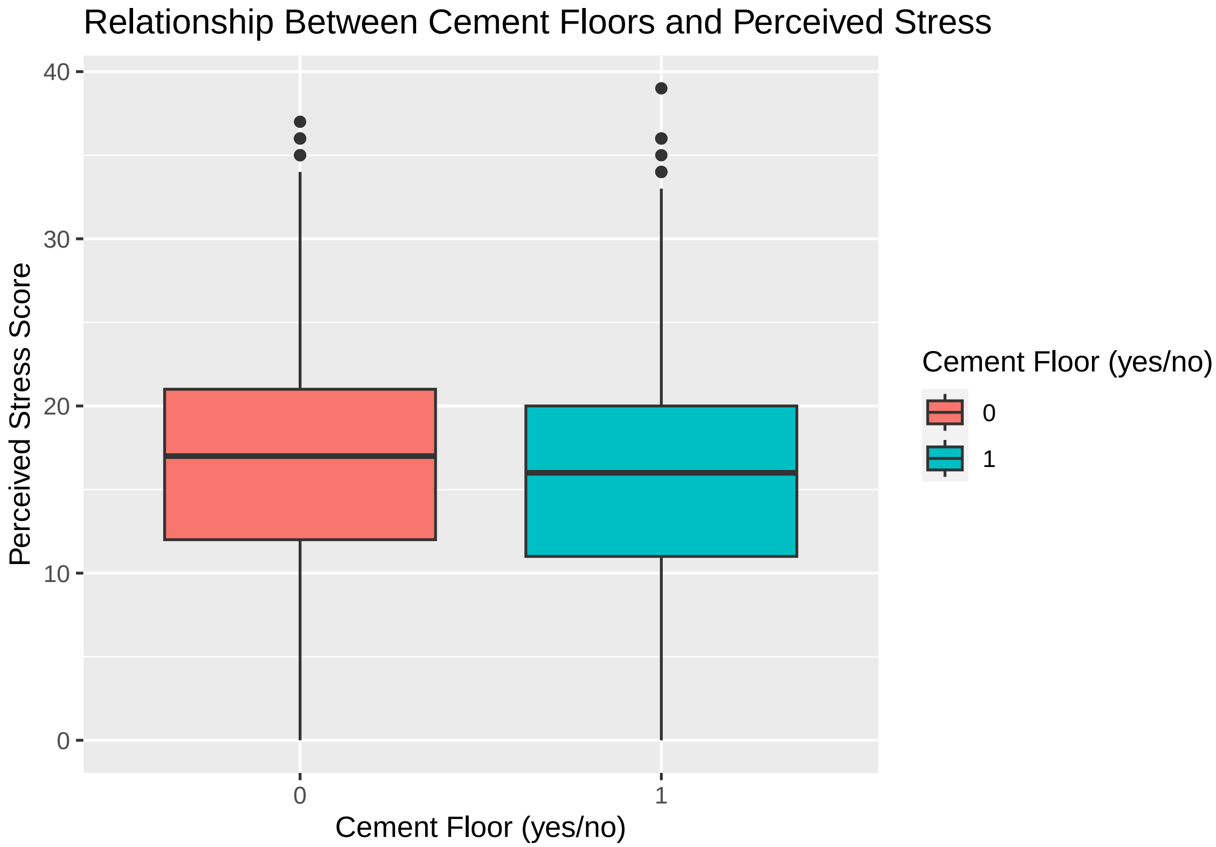


Table 5: Joint Distribution of A and Z

Z	A_bin	count
0	0	627
0	1	754
1	0	297
1	1	1080


Relationship Between Cement Floors and Perceived Depression
Perceived Depression Score
0
20
40
0
1
Cement Floor (yes/no)
Cement Floor (yes/no)
0
1



Included Covariates

Demographics:

- Number of household members (S_HHpeople)
- Head of household's age (S_headage)
- Spouse's age (S_spouseage)
- Head of household's years of schooling (S_headeduc)
- Proportion of Males 0-5yrs in household (S_dem1)
- Proportion of Males 6-17yrs in household (S_dem2)
- Proportion of Males 18-49yrs in household (S_dem3)
- Proportion of Males 50+yrs in household (S_dem4)
- Proportion of Females 0-5yrs in household (S_dem5)
- Proportion of Females 6-17yrs in household (S_dem6)
- Proportion of Females 18-49yrs in household (S_dem7)
- Proportion of Females 50+yrs in household (S_dem8)

Environment/Access to Resources:

- Water connection (=1) (S_waterland)
- Water connection inside the house (=1) (S_waterhouse)
- Electricity (=1) (S_electricity)
- Household has animals on land (=1) (S_hasanimals)
- Animals allowed to enter the house (=1) (S_animalsinside)
- Uses garbage collection service (=1) (S_garbage)
- Number of times respondent washed hands the day before (S_washhands)

Note: Spouse's years of schooling (S_spouseeduc) was included in the dataset, but I excluded it in the final analysis because it had about 350 missing values and was not a significant predictor of either of the response variables in the two-stage least squares models.

R Code

```r
library(tidyverse)
```

Data Exploration/Cleaning

```r
# read in data
df_household <- readRDS("cleaned_data.rds")
dim(df_household)
```

```r
### Outcome
# Maternal Depression Scale (CES-D Scale)
Y1 <- df_household$S_cesds
# Maternal Perceived Stress Scale (PSS)
Y2 <- df_household$S_pss

### endogeneous variable
# the share of cement floors
A <- df_household$S_shcementfloor
```

```r
### IV
# a dummy variable that indicates whether or not the household was offered the program treatment
Z <- df_household$dpisofirme
```

```r
# distribution of share of cement floors
ggplot(mapping = aes(x = A)) +
  geom_boxplot() +
  labs(title = "Distribution of Share of Cement Floors",
       x = "Proportion of Cement Floors")
```

```r
# convert treatment to binary
df_bin <- df_household |>
  select(S_cesds, S_pss, S_shcementfloor, dpisofirme) |>
  mutate(A_bin = ifelse(S_shcementfloor == 1, 1, 0))
```

```r
# check for missing values for main variables
df_missing_y1 <- df_bin |>
  filter(is.na(S_cesds))

df_missing_y2 <- df_bin |>
  filter(is.na(S_pss))

df_missing_A <- df_bin |>
  filter(is.na(S_shcementfloor))

df_missing_Z <- df_bin |>
  filter(is.na(dpisofirme))
```

```r
# filter out missing values
df_bin2 <- df_bin |>
  filter(!is.na(S_cesds)) |>
  filter(!is.na(S_pss))
```

```
# select variables for adjusted analysis
data_adj <- df_household |>
  select(Y1 = S_cesds, Y2 = S_pss, A = S_shcementfloor, Z = dpisofirme,
         S_HHpeople, S_headage, S_spouseage, S_headeduc,
         S_dem1, S_dem2, S_dem3, S_dem4, S_dem5, S_dem6, S_dem7, S_dem8,
         S_waterland, S_waterhouse, S_electricity, S_hasanimals,
         S_animalsinside, S_garbage, S_washhands) |>
  mutate(A_bin = ifelse(A == 1, 1, 0)) |>
  filter(!is.na(Y1)) |>
  filter(!is.na(Y2))
```

```
# check for missing values in covariates
sum(is.na(data_adj))
```

```
# check specifically which covariates include missing values and how many
missing_summmary <- data_adj |>
  summarise_all(funs(sum(is.na(.))))
```

```
# remove missing values for covariates
which(!complete.cases(data_adj))
na_df <- which(!complete.cases(data_adj))
data_adj <- data_adj[-na_df,]
```

```
# table of distribution of binary treatment
df_summaryA <- data_adj |>
  group_by(A_bin) %>%
  summarise(count = n(), percent = round(count/nrow(.)*100, 1))


knitr::kable(df_summaryA)
```

```
# table of distribution of IV
df_summaryZ <- data_adj |>
  group_by(Z) %>%
  summarise(count = n(), percent = round(count/nrow(.)*100, 1))
```

```r
knitr::kable(df_summaryZ)
```

```r
# histogram of distribution of response 1: depression
ggplot(data_adj, aes(x = Y1)) +
  geom_histogram() +
  labs(title = "Distribution of Y1: Perceived Depression")
```

```r
# histogram of distribution of response 2: stress
ggplot(data_adj, aes(x = Y2)) +
  geom_histogram() +
  labs(title = "Distribution of Y2: Perceived Stress")
```

```r
# table of joint distribution between IV and binary treatment
df_summaryAZ <- data_adj |>
  group_by(Z, A_bin) |>
  summarise(count = n())

df_summaryAZ
```

```r
# Side-by-side boxplots of depression score by treatment status
ggplot(data = data_adj, aes(y = Y1, x = as.factor(A_bin), fill = as.factor(A_bin))) +
  geom_boxplot() +
  labs(title = "Relationship Between Cement Floors and Perceived Depression",
       x = "Cement Floor (yes/no)",
       fill = "Cement Floor (yes/no)",
       y = "Perceived Depression Score")
```

```r
# Side-by-side boxplots of stress score by treatment status
ggplot(data = data_adj, aes(y = Y2, x = as.factor(A_bin), fill = as.factor(A_bin))) +
  geom_boxplot() +
  labs(title = "Relationship Between Cement Floors and Perceived Stress",
       x = "Cement Floor (yes/no)",
       fill = "Cement Floor (yes/no)",
       y = "Perceived Stress Score")
```

Unadjusted Analysis

```
# define variables
Y1 <- df_bin2$S_cesds

Y2 <- df_bin2$S_pss

A <- df_bin2$A_bin

Z <- df_bin2$dpisofirme
```

```
# estimate and CI for proportion of Compliers
summary(lm(A ~ Z))
confint(lm(A ~ Z))
```

Y1: Depression Scale

```
# two equivalent methods for unadjusted analysis:

## method 1: ratio of difference in means
(mean(Y1[Z == 1]) - mean(Y1[Z == 0])) / (mean(A[Z == 1]) - mean(A[Z == 0]))

## method 2: two-stage least squares
pred <- lm(A ~ Z)$fitted.values
summary(lm(Y1 ~ pred))

# 95 percent CI
confint(lm(Y1~pred))
```

Y2: Stress Scale

```
# two equivalent methods for unadjusted analysis:

## method 1: ratio of difference in means
(mean(Y2[Z == 1]) - mean(Y2[Z == 0])) / (mean(A[Z == 1]) - mean(A[Z == 0]))

## method 2: two-stage least squares
pred <- lm(A ~ Z)$fitted.values
summary(lm(Y2 ~ pred))

# 95 percent CI
confint(lm(Y2~pred))
```

Adjusted Analysis: Two-Stage Least Squares

```
# redefine variables
W <- data_adj |>
  select(S_HHpeople, S_headage, S_spouseage, S_headeduc,
         S_dem1, S_dem2, S_dem3, S_dem4, S_dem5, S_dem6, S_dem7, S_dem8,
         S_waterland, S_waterhouse, S_electricity, S_hasanimals,
         S_animalsinside, S_garbage, S_washhands
         )

Y1 <- data_adj$Y1

Y2 <- data_adj$Y2

A <- data_adj$A_bin

Z <- data_adj$Z
```

```
# estimate for proportion of compliers
summary(lm(A ~ Z + ., data=W))
```

```r
confint(lm(A ~ Z + ., data=W))
```

Y1: Depression Scale

```r
# Two-stage least squares with binary instrument adjusting for other covariates:
pred <- lm(A ~ Z + ., data=W)$fitted.values
coef(lm(Y1 ~ pred + ., data = W))['pred']
confint(lm(Y1 ~ pred + .,data = W))['pred',]

summary(lm(Y1 ~ pred + ., data = W))
```

Y2: Stress Scale

```r
# Two-stage least squares with binary instrument adjusting for other covariates:
pred <- lm(A ~ Z + ., data=W)$fitted.values
coef(lm(Y2 ~ pred + ., data = W))['pred']
confint(lm(Y2 ~ pred + .,data = W))['pred',]

summary(lm(Y2 ~ pred + ., data = W))
```

Adjusted Analysis: IPW Estimator

Y1: Depression Scale

```r
# Adjusted IV analysis using propensity score approach
Z.prop <- glm(Z ~ ., family = binomial, data=W)
Z.prop.score <- Z.prop$fitted.values
IPW.est <- (mean(Y1 * Z / Z.prop.score) - mean(Y1 * (1-Z) / (1-Z.prop.score)))/
  (mean(A * Z / Z.prop.score) - mean(A * (1-Z) / (1-Z.prop.score)) )

n <- nrow(data_adj)
boot.ests <- sapply(1:500, function(j) {
  set.seed(j)
  if(j %% 50 == 0) print(j)
  boot.inds <- sample(1:n, n, replace=TRUE)
  Z.prop <- glm(Z[boot.inds] ~ ., family = binomial, data=W[boot.inds,])
  Z.prop.score <- Z.prop$fitted.values
  (mean(Y1[boot.inds] * Z[boot.inds] / Z.prop.score)
    - mean(Y1[boot.inds] * (1-Z[boot.inds]) / (1-Z.prop.score)))/
    (mean(A[boot.inds] * Z[boot.inds] /
            Z.prop.score) - mean(A[boot.inds] * (1-Z[boot.inds]) / (1-Z.prop.score)))
})

IPW.est
quantile(boot.ests, c(.025, .975))
```

Y2: Stress Scale

```
# Adjusted IV analysis using propensity score approach
Z.prop <- glm(Z ~ ., family = binomial, data=W)
Z.prop.score <- Z.prop$fitted.values
IPW.est <- (mean(Y2 * Z / Z.prop.score) - mean(Y2 * (1-Z) / (1-Z.prop.score)))/
  (mean(A * Z / Z.prop.score) - mean(A * (1-Z) / (1-Z.prop.score)) )

n <- nrow(data_adj)
boot.ests <- sapply(1:500, function(j) {
  set.seed(j)
  if(j %% 50 == 0) print(j)
  boot.inds <- sample(1:n, n, replace=TRUE)
  Z.prop <- glm(Z[boot.inds] ~ ., family = binomial, data=W[boot.inds,])
  Z.prop.score <- Z.prop$fitted.values
  (mean(Y2[boot.inds] * Z[boot.inds] / Z.prop.score)
    - mean(Y2[boot.inds] * (1-Z[boot.inds]) / (1-Z.prop.score)))/
    (mean(A[boot.inds] * Z[boot.inds] /
            Z.prop.score) - mean(A[boot.inds] * (1-Z[boot.inds]) / (1-Z.prop.score)))
})
```

```
IPW.est
quantile(boot.ests, c(.025, .975))
```

Summary Tables

```
# table 1: depression
summary_estimates_Y1 <- tibble(
  Method = c("Unadjusted", "Adjusted Two-Stage Least Squares", "Adjusted IPW"),
  Estimate = c(-9.22, -9.52, -9.28),
  CI = c("(-12.0, -6.4)", "(-12.17, -6.87)", "(-12.21, -6.46)"),
  p_value = c("0", "0", "--")
)

# table 2: stress
summary_estimates_Y2 <- tibble(
  Method = c("Unadjusted", "Adjusted Two-Stage Least Squares", "Adjusted IPW"),
  Estimate = c(-7.26, -7.16, -6.93),
  CI = c("(-9.35, -5.18)", "(-9.18, -5.13)", "(-9.14, -4.75)"),
  p_value = c(0, 0, "--" )
)
```

References

Titiunik, R. *et al.* (2007) ‘Housing, Health, And Happiness’, *Research Working papers*, 1, pp. 1–34. Available at: https://doi.org/10.1596/1813-9450-4214.