Jerdan Daniels & Chandler Zachary
"Factors Affecting Women's Healthcare Access in Malawi"
MATH 5387 -- Final Project
Due 27 July 2018



## Introduction

The scope of our analysis is the factors that affect whether and the extent to which women access healthcare services in developing nations. We are interested in this question because compelling evidence has been found that, when women's outcomes improve, other outcomes, such as children's health and human capital development, also improve. Our data are taken from from the Demographic and Health Surveys ("DHS"), which collect microdata in low- and middle-income countries, totalling 24 African and 4 Asian countries. The data are maintained by the Integrated Public Use Microdata Series ("IPUMS") at the Minnesota Population Center in the University of Minnesota. Our initial specification contains 10 predictors, and we end with four in our final model.

## Exploratory Data Analysis

Based on the availability of data on the variables of interest, we narrow our scope to Malawi in 2016 and Zambia in 2013. We choose Malawi because it is the more recent year. One frustration we experienced with this data is the overall lack of continuous variables in our scope. Most variables are categorical. There are no missing values; however, there are 23,021 "not in universe" values, indicating respondents who were not asked a particular question. Since we have no way of knowing why those questions are not in the universe for the sample, we drop all of them. We are left with a sample size of 1,541.

Our outcome of interest is the variable INJHCWORKER. This variable measures the number of injections a woman received from a doctor, nurse, pharmacist, dentist, or any other professional healthcare worker in the six - 12 months prior to taking the survey. We use this variable as our outcome of interest because it satisfies the criterion of being continuous and, of all variables in the data, it provides the best indicator of a woman's ability to access healthcare services and the extent to which a woman accesses those services when necessary. In our sample, the mean number of injections received is three, and the median is two. The range is 0 - 60, and the standard deviation is 3.12. Table 1 presents definitions for all variables used the the analysis. One potential drawback of this outcome variable is that we have no indication why these women are receiving injections. It may be that women who do not feel the need for a specific injection (e.g. - they do not exhibit symptoms) will not obtain an injection.

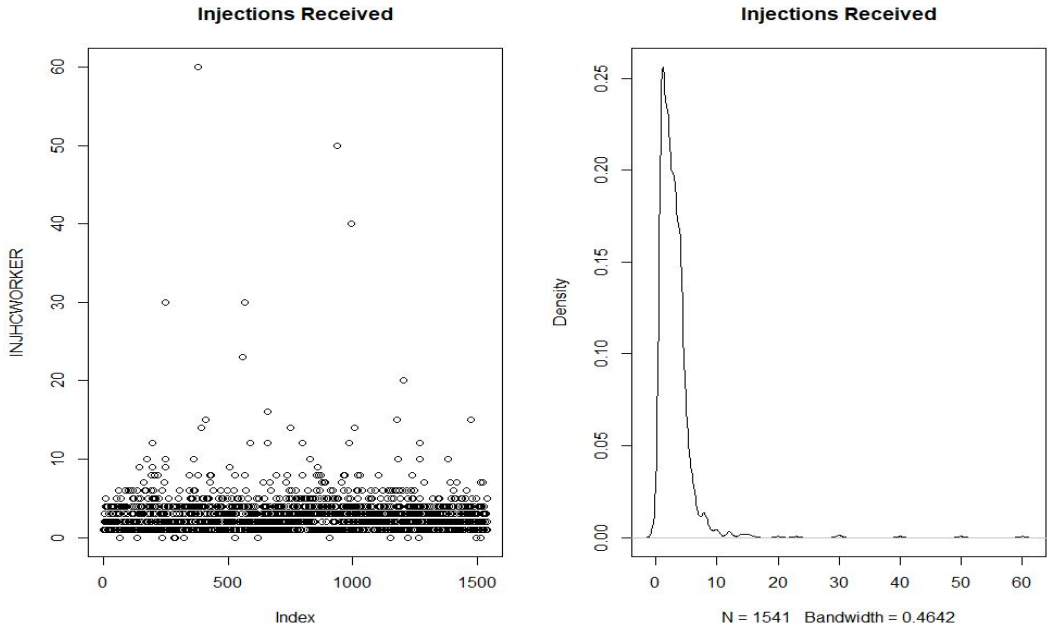
Table 1: Variable names and definitions

| Variable Name | Definition |
|---|---|
| INJHCWORKER (outcome variable) | Number of injections received in the 6-12 months prior to survey |
| AGE | Respondent's age |
| FEMOWNLAND | Categorical variable that describes the respondent's ownership of land |

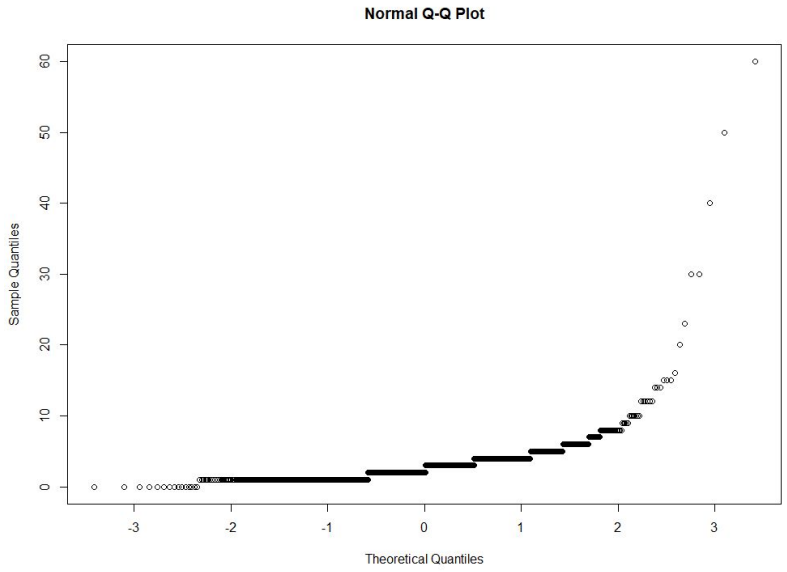
| | |
|---|---|
| EDACHIEVER | Categorical variable that describes the respondent's education level |
| DECFEMEARN | Categorical variable that describes who has the final say on how the respondent's earnings are spent |
| BHCPERMIT | Categorical variable describing whether getting permission to visit a healthcare facility is a problem for the respondent |
| BHCMONEY | Categorical variable describing whether paying for a visit to a healthcare facility is a problem for the respondent |
| BHCDISTANCE | Categorical variable describing whether the distance to a healthcare facility is a problem for the respondent |
| BHCALONE | Categorical variable describing whether going alone to a healthcare facility is a problem for the respondent |
| BHCNOFEMDR | Categorical variable describing whether lack of a female healthcare provider is a problem for the respondent |
| BHCNOPROV | Categorical variable describing whether lack of a healthcare provider is a problem for the respondent |
| INSCOVERYN | Binary variable indicating whether the respondent has health insurance |

We categorized the predictor variables into two segments: empowerment and barrier. The empowerment variables represent socio-economic, cultural, and financial attributes that indicate whether women can act on their own decisions. The barrier variables represent conditions that might restrict access to healthcare services, such as physical, cultural, and financial barriers. The variable age is also included as a continuous predictor. The mean age for the women in the sample is 30 years old, and the median is 29. The range is 15 - 49 years old with a standard deviation of 9.26 years.

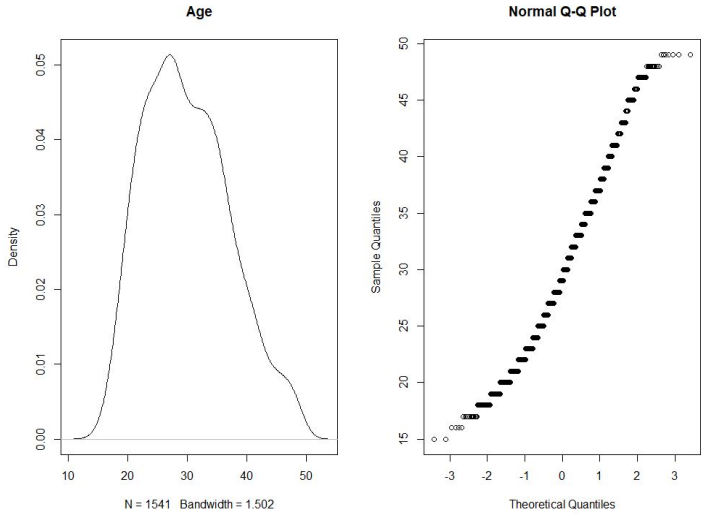
We begin the exploratory analysis with an investigation of the outcome variable, INJHCWORKER. Based on the graphs below, we observe that values are clustered between zero and 10. The univariate plot on the left prompts the question of which values may be considered outliers. These women may be receiving injections for a host of different reasons, such as HIV, so we can not exclude them without knowing more about about them.



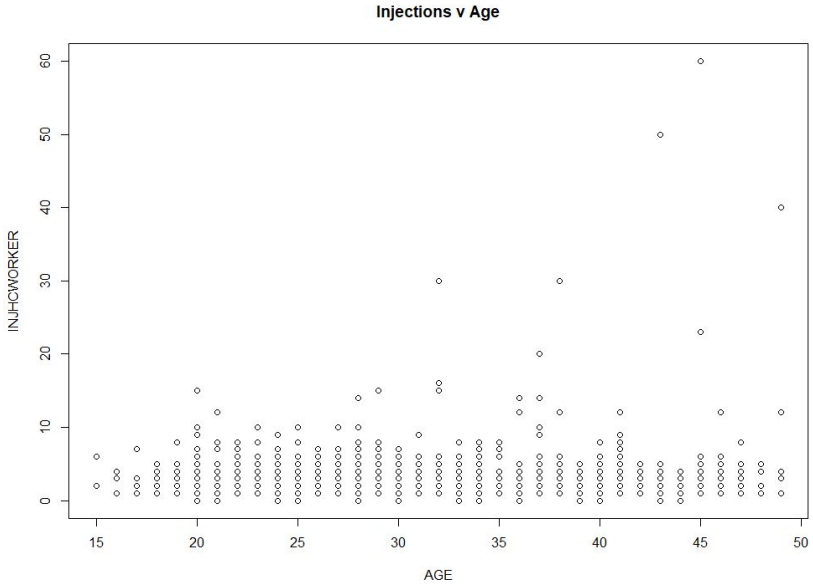

The graph on the right is the distribution of injections which suggests a chi-squared distribution. The QQ-plot below provides further evidence that we can not assume INJHCWORKER is normally distributed. Using the Shapiro-Wilk test on the null hypothesis that INJHCWORKER is normally distributed against the alternative that it is not, the p-value is approximately zero. We conclude that there is enough evidence to reject the null; therefore, we can not assume that the number of injections received by respondents is normally distributed. Transformations of this variable are limited because of the high concentration of zero values. A square root transformation was attempted, but the distribution was still not normal.



We next investigate the variable AGE. Based on the density graph and QQ plot below, we are confident in assuming that AGE is normally distributed.




Based on the scatterplot below that compares injections against age, we observe that the outlying injection values correspond with higher ages.




Based on the correlation matrix, which is not reproduced here due to space concerns, we note that the empowerment variables are correlated with the outcome, while the barrier variables are not. Therefore, we do not expect the barrier variable to be significant in the regression.

**Regression Analysis & Model Selection**

First, we examine the overall significance of the regression that includes all predictors. Based on the exploratory and graphical analysis, we expect only AGE, FEMOWNLAND, and EDACHIEVER to be significant predictors of INJHCWORKER. The specification is:

$$\begin{aligned} INJHCWORKER_i = {} & \beta_0 + \beta_1 AGE_i + \beta_2 FEMOWNLAND_i + \beta_3 EDACHIEVER_i \\ & + \beta_4 DECFEMEARN_i + \beta_5 BHCPERMIT_i + \beta_6 BHCMONEY_i \\ & + \beta_7 BHCDISTANCE_i + \beta_8 BHCALONE_i + \beta_9 BHCNOFEMDR_i \\ & + \beta_{10} BHCNOPROV_i + \beta_{11} INSCOVERYN_i + \varepsilon_i \end{aligned}$$

Regression results are reproduced below:

Coefficients:

| | Estimate | Std. Error | t value | Pr(>\|t\|) |
|---|---|---|---|---|
| (Intercept) | 1.628725 | 0.524178 | 3.107 | 0.00192 ** |
| AGE | 0.036152 | 0.011186 | 3.232 | 0.00126 ** |
| FEMOWNLANDOwns alone only | 0.011306 | 0.199147 | 0.057 | 0.95473 |
| FEMOWNLANDOwns jointly | 0.135833 | 0.206259 | 0.659 | 0.51028 |
| FEMOWNLANDOwns jointly only | 1.086046 | 0.459445 | 2.364 | 0.01821 * |
| EDACHIEVERIncomplete primary | 0.194736 | 0.317856 | 0.613 | 0.54020 |
| EDACHIEVERComplete primary | 0.093338 | 0.389440 | 0.240 | 0.81062 |
| EDACHIEVERIncomplete secondary | -0.026170 | 0.356478 | -0.073 | 0.94149 |
| EDACHIEVERComplete secondary | -0.033545 | 0.389724 | -0.086 | 0.93142 |
| EDACHIEVERHigher | -0.318628 | 0.418827 | -0.761 | 0.44692 |
| DECFEMEARNWoman and husband\\partner | 0.178372 | 0.195730 | 0.911 | 0.36227 |
| DECFEMEARNWoman and someone else | 0.239705 | 0.232855 | 1.029 | 0.30345 |
| DECFEMEARNHusband\\partner | 5.328518 | 3.121606 | 1.707 | 0.08803 . |
| BHCPERMITNo problem at all | -0.086701 | 0.297255 | -0.292 | 0.77058 |
| BHCMONEYNo problem at all | 0.016753 | 0.192765 | 0.087 | 0.93075 |
| BHCDISTANCENo problem at all | -0.007527 | 0.189751 | -0.040 | 0.96836 |
| BHCALONENo problem at all | 0.015114 | 0.233860 | 0.065 | 0.94848 |
| BHCNOFEMDRNo problem at all | 0.390391 | 0.267841 | 1.458 | 0.14517 |
| BHCNOPROVNo problem at all | -0.180433 | 0.175865 | -1.026 | 0.30507 |
| INSCOVERYNYes | 0.584010 | 0.430954 | 1.355 | 0.17557 |

---

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.105 on 1521 degrees of freedom
Multiple R-squared: 0.01959, Adjusted R-squared: 0.007345
F-statistic: 1.6 on 19 and 1521 DF, p-value: 0.04854

Rather than offer an exhaustive interpretation of all coefficients, we focus only on those that are significant at or above the 10% level. The intercept can be interpreted as for a woman who: does not own land, has no education, has the final say on spending her earnings, has no insurance, and for whom getting permission, lacking money, distance to facility, going alone, no female

healthcare provider, and no healthcare provider are not big problems, the average number of injections received from a healthcare worker in the 6-12 months preceding the survey was 1.63, for a woman at zero years of age. Evaluated at the mean age, the number of injections is: $1.628725 + 30 \times 0.036152 = 2.71$.

Age is another significant predictor. An increase of one year in age is associated with about 0.036 more injections. Owning land jointly and having one's husband or partner make the final decision on how one's earnings are spent are also significant predictors. Holding other factors constant, owning land jointly is associated with $1.628725 + 30 \times 0.036152 + 1.086046 = 3.80$ injections. Having one's husband or partner make the final decision over spending one's earnings is associated with $1.628725 + 30 \times 0.036152 + 5.328518 = 8.04$ injections, both calculated at the mean age. Crucially, no barrier variables are significant, which we anticipated.

Next, we use a permutation F-test for overall significance of the regression. The permutation test is used since we can not assume our outcome is normally distributed. The null hypothesis is that all coefficients equal zero, and the alternative is that at least one coefficient is different from zero. The observed p-value is 0.04854, and the permutation p-value is 0.049. Thus, we reject the null at the 5% significance level and conclude that at least one predictor is different from zero. Please see R code for the computation of the F-statistic and p-value.

We reproduce the results of the Variance Inflation Factor test in Table 2. The values are low enough for us to conclude that collinearity is not a problem.

Table 2: Variance Inflation Factor Test Results

| | GVIF | Df | GVIF^(1/(2*Df)) |
|---|---|---|---|
| AGE | 1.048969 | 1 | 1.024192 |
| FEMOWNLAND | 1.172172 | 3 | 1.026830 |
| EDACHIEVER | 1.474082 | 5 | 1.039566 |
| DECFEMEARN | 1.105567 | 3 | 1.016867 |
| BHCPERMIT | 1.188690 | 1 | 1.090270 |
| BHCMONEY | 1.384940 | 1 | 1.176835 |
| BHCDISTANCE | 1.404376 | 1 | 1.185064 |
| BHCALONE | 1.370156 | 1 | 1.170536 |
| BHCNOFEMDR | 1.267065 | 1 | 1.125640 |
| BHCNOPROV | 1.225232 | 1 | 1.106902 |
| INSCOVERYN | 1.163703 | 1 | 1.078751 |

We reproduce the results of the Akaike Information Criterion ("AIC") and the Bayesian Information Criterion ("BIC") to narrow our choices of covariates after excluding barrier variables.

```
Step:  AIC=3493.36
INJHCWORKER ~ AGE + FEMOWNLAND
```

```
Step:  AIC=3505.35
INJHCWORKER ~ AGE
```

The AIC recommends the predictors AGE and FEMOWNLAND, while the BIC recommends only AGE.

We provide an F-test for the model using age and the empowerment variables, the restricted model, against the model that uses all predictors, the unrestricted model. The null hypothesis is that the restricted model is sufficient to explain variation in the outcome variable, and the alternative is that the unrestricted model is necessary to explain variation in the outcome variable. The F-statistic is 0.6506, and the p-value is 0.7141, suggesting that we fail to reject the null and conclude there is enough evidence to support the restricted model.

The graphical and regression evidence support choosing a model that includes only age and the empowerment variables. The supplementary AIC and BIC suggest that we restrict the model even further by excluding DECFEMEARN and EDACHIEVER. However, we believe that, based on the interpretation of the intercept in the unrestricted model, we would ignore valuable information by excluding these covariates. The empowerment variables suggest that there is a socio-economic status effect on women receiving injections, while the barrier variables suggest that women seem to experience few physical and financial barriers to receiving injections. Therefore, we choose the model that uses age and the empowerment variables to predict number of injections received from a healthcare worker.

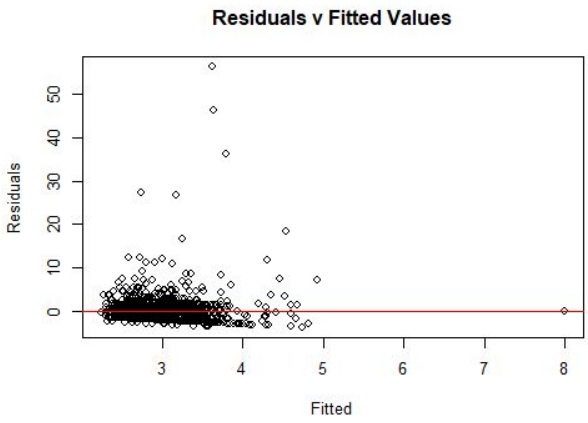
Table 3 below presents comparative measures of the two models. We acknowledge that the RMSE for the unrestricted model is greater than that for the restricted model, but the difference is less than 1%. We prefer the Adjusted $R^2$ as a measure of explanatory power because it does not penalize us for removing the barrier variables.

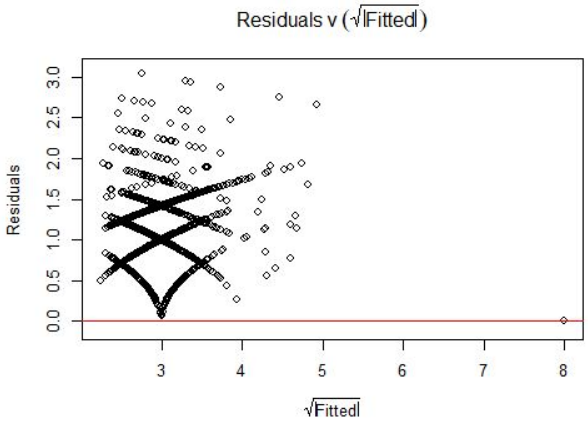
Table 3: Comparing the Unrestricted and Restricted Models

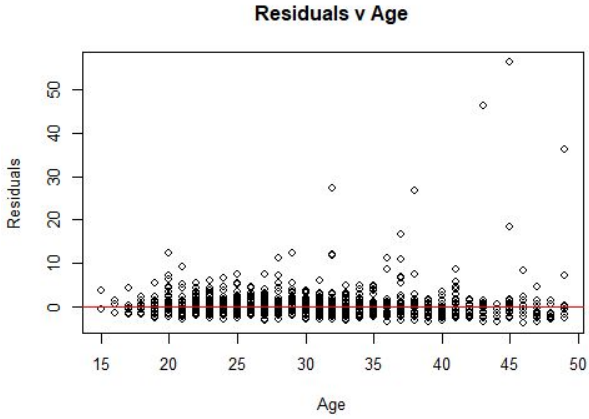
| **Measure** | **Unrestricted** | **Restricted** |
|---|---|---|
| RMSE | 3.0852 | 3.0898 |
| Adjusted $R^2$ | 0.0073 | 0.0089 |

**Diagnostics**

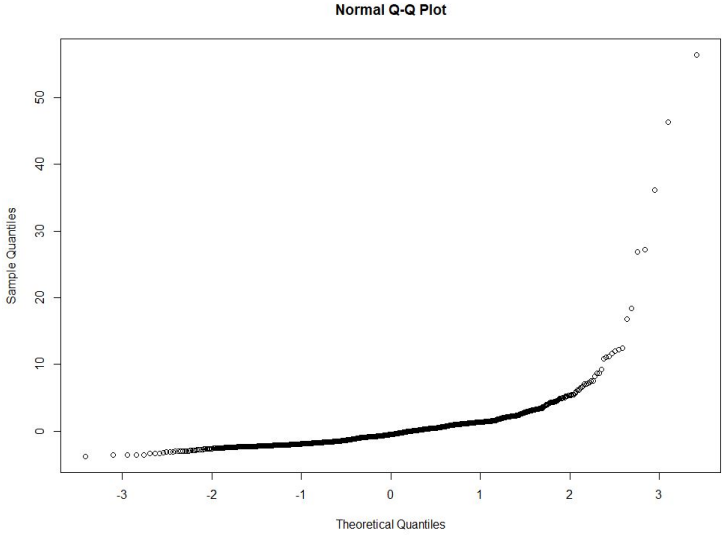
The first assumption we evaluate is constant error variance. Covariates other than age are omitted because they are not numeric. We observe clear patterns in the relationship between residuals and fitted values, suggesting that our residuals are heteroskedastic. Observe also that the relationship between residuals and AGE is strikingly similar to the relationship between INJHCWORKER and AGE. We believe this is evidence that AGE is correlated with the error term.



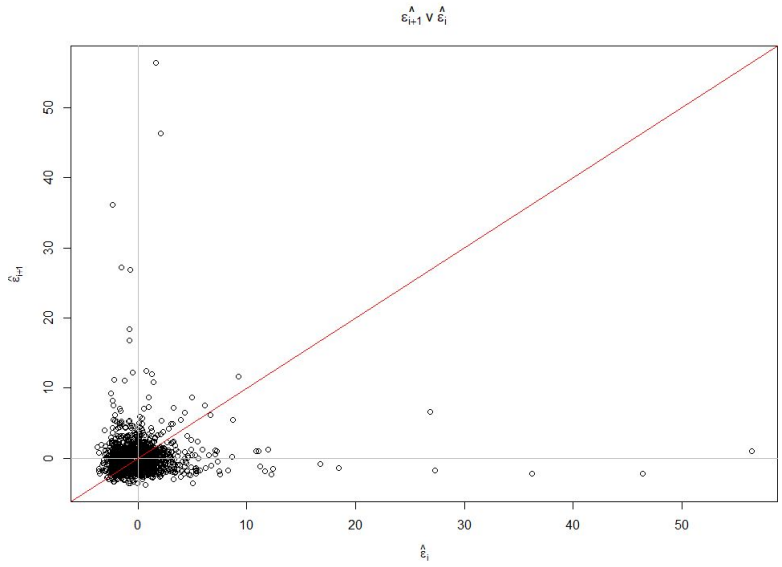




Next, we check for normally distributed errors. Based on the graph below, we do not believe our residuals satisfy this assumption.



We check for serial correlation. The plot below suggests that errors are correlated in our model.




Given the presence of heteroskedasticity and correlated errors, one possible modification would be to use a generalized least squares model. We could also employ robust standard errors, or White standard errors, to increase the threshold for statistical significance.

We provide additional diagnostics in the Added-Variable plot ("AVP") on the following page. The added-variable plots suggest that there are no non-linear relationships between the predictors and the outcome of interest. They do not introduce any new information about the other marginal effects. We do not reproduce Component-Plus-Residual plots because the categorical variables make interpretation difficult.

**Conclusion**

Our final model includes only four predictors after excluding variables that were insignificant and including the predictors that we believe contribute the most information in explaining variation in the outcome. However, it is unclear exactly what the roles those variables play. In future studies, the model could be improved by using a better measure of healthcare access in the outcome and by including variables that control for factors such as why women seek healthcare treatment, such as HIV or STIs. Longitudinal data could be useful, too, since women may only access healthcare when they need it, rather than seeking treatment for specific symptoms or for regular health checks. We have no way of knowing how our model treats such individuals. More technically, a generalized least squares model could be used because there is clear evidence of correlation and heteroskedasticity in the residuals.

**Added-Variable Plots**