test

2023-11-29

Load data

```
# Clear all
# rm(list = ls())

# load and require packages
pacman::p_load(foreign, tidyverse, here, randomForest, boot, MASS, ivmte, aod)

# Load data
data <- read.csv(here("data/andres_cleaned.csv"))
data$X <- NULL

set.seed(1234) # For reproducibility
```

Question 1a - Relate structural equation and probit regression

Suppose we specify the potential outcome equations as $Y_i(1) = X_i'\theta_1 + \epsilon_{1i}$ and $Y_i(0) = X_i'\theta_0 + \epsilon_{0i}$, and the structural selection equation as $D_i = 1\{u(X_i, Z_i) \geq V_i\}$. where $u(X_i, Z_i) - V_i$ is the latent utility derived from additional children.

The structural selection equation describes the switch from $D = 0$ to $D = 1$ in terms of utility, i.e., whether, as a function of your observables and instrument, the family has a higher utility than some level of having more than three children. The propensity score quantifies this probability, capturing the likelihood, given your observables and instrument value, of having $D = 1$.

Question 1b - Assumptions for MTE

Clarify a set of assumptions that the marginal treatment effect $M(x, u)$, $u \in (0, 1)$ can be identified by the derivative of the OLS regression equation WRT the propensity score:

$$MTE(x, u) = \frac{\partial}{\partial p}\bigg|_{p=u} \left(x'\beta_0 + px'\beta_1 + \kappa_1 p + \kappa_2 p^2 + \kappa_3 p^3\right)$$

The assumptions are that (1) the instrument is independent $(U_1, U_0, V) \perp Z|X$, (2) that $\mu_D(Z, X)$ has a nondegenerate distribution given X, (3) that scalar V is continuously distributed, (4) $E(|Y(1)|)$ and $E(|Y(0)|)$ are finite, and (5) $0 < P(D = 1|X) < 1$.

Question 1c - Estimate propensity score

Estimate the propensity score as instructed above, and assess whether the coefficient of the instrumental variable in the probit regression is significantly different from zero or not.

Yes, we see that the z-value for the instrument is extremely large at 37, and very significant. So we have confidence that the instrument is predicting treatment.

```
# Fit probit
probit_model <- glm(d ~ blackm + hispm + othracem + agem1 + agefstm + boy1st + z,
                    data = data,
                    family = binomial(link = "probit"))

# Predict propensity scores
data$prop_score <- predict(probit_model, type = "response")

summary(probit_model)
```

```
##
## Call:
## glm(formula = d ~ blackm + hispm + othracem + agem1 + agefstm +
##     boy1st + z, family = binomial(link = "probit"), data = data)
##
## Coefficients:
##               Estimate Std. Error  z value Pr(>|z|)
## (Intercept) -0.3878595  0.0261706  -14.820  < 2e-16 ***
## blackmTRUE   0.1712398  0.0115636   14.808  < 2e-16 ***
## hispmTRUE    0.4462240  0.0159873   27.911  < 2e-16 ***
## othracem     0.1631770  0.0151467   10.773  < 2e-16 ***
## agem1        0.0822504  0.0008432   97.540  < 2e-16 ***
## agefstm     -0.1220241  0.0010082 -121.031  < 2e-16 ***
## boy1stTRUE  -0.0308048  0.0051877   -5.938 2.88e-09 ***
## zTRUE        0.1911553  0.0051907   36.827  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 338352  on 254651  degrees of freedom
## Residual deviance: 317827  on 254644  degrees of freedom
## AIC: 317843
##
## Number of Fisher Scoring iterations: 4
```