POLS6382 Quantitative Method III
Maximum Likelihood Estimation
Lab 7: Censored and Trancated Data

Ling Zhu and Emily Lee
Department of Political Science
University of Houston

2025/10/22

1. Learning Objectives

Objectives

- Learn how to estimate a Tobit regression model.
- Compare Tobit and OLS results with censored data.
- Learn how to estimate a Heckman selection model.
- Assessing bias due to sample selection.

```
> rm(list=ls())
> setwd("/Users/lingzhu/Dropbox/UH Teaching/POLS6382_2025/2025 Labs/Lab 7")
> my_packages <- c("foreign", "ggplot2","dplyr","ggpubr","VGAM","MASS",
+                  "GGally","censReg", "sampleSelection", "tidyr","mvtnorm")
> invisible(lapply(my_packages, require, character.only = TRUE))
```

2. Compare Tobit and OLS with Censored Data

In this section, we will use a simulated data et to compare the Tobit and OLS model. First, we simulate a dataset that contains a dependent variable y, which is left-censored. All the censored cases are assigned with value "0". The explanatory variable is x, generated by randomly drawing numbers from a normal distribution. We also set the latent variable y.star as a linear function of x, with a slope coefficient of 5.

```
> N = 10
> f = rep(c("s1","s2","s3","s4","s5","s6","s7","s8"),N)
> fcoeff = rep(c(-1,-2,-3,-4,-3,-5,-10,-5),N)
> set.seed(100)
> x = rnorm(20*N)+1
> beta = 5
> epsilon = rnorm(20*N,sd = sqrt(1/5))
> y.star = x*beta+fcoeff+epsilon ## latent response
> y = y.star
> y[y<0] <- 0 ## left censored response
> simdata<-data.frame(cbind(x,y))
```

Next, we fit an OLS model with this simulated dataset. We see that the OLS model produces a slope coefficient of 2.93, which is substantially smaller than the true parameter value that we set (β=5). The biased coefficient is caused by ignoring the fact that values of y are censored at 0.

The proper model specification is Tobit. We use the `vglm()` function to estimate a Tobit model.[1]. The first coefficient labeled as `(intercept):1" is the intercept term. The second coefficient labeled as`(intercept):2" an ancillary statistic. If we exponentiate this value, we get a number that is analogous to the square root of the residual variance in OLS regression. $exp(0.718) \approx 2.05$. We also see that the Tobit model produces a mean slope of 4.856 for variable x. This is quite close to the true parameter value (5).

```
> fitols<-lm(y~x)
> summary(fitols)


Call:
lm(formula = y ~ x)

Residuals:
    Min      1Q  Median      3Q     Max
-5.3648 -1.3024 -0.2074  1.1920  5.4142

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -0.4943     0.2032  -2.432   0.0159 *
x             2.9351     0.1497  19.610   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.928 on 198 degrees of freedom
Multiple R-squared:  0.6601,    Adjusted R-squared:  0.6584
F-statistic: 384.5 on 1 and 198 DF,  p-value: < 2.2e-16
> fittobit<-vglm(formula=y~x, family=tobit(Lower=0))
> summary(fittobit)


Call:
vglm(formula = y ~ x, family = tobit(Lower = 0))

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept):1 -3.83444    0.37275  -10.29   <2e-16 ***
(Intercept):2  0.71856    0.06651   10.80   <2e-16 ***
x              4.85623    0.23724   20.47   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Names of linear predictors: mu, loglink(sd)

Log-likelihood: -282.8849 on 397 degrees of freedom

Number of Fisher scoring iterations: 5

No Hauck-Donner effect found in any of the estimates
> coef(fitols)

(Intercept)           x
```

[1]Various `R` functions can be used to estimate a Tobit model, e.g. `censReg()` from package `censReg` and `tobit()` from package `AER`.

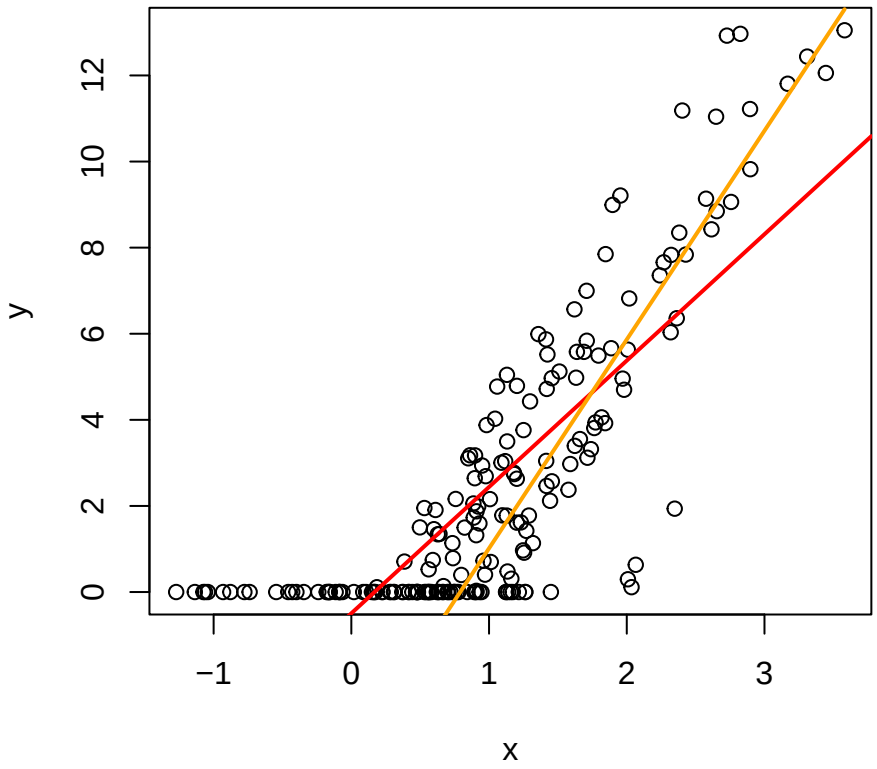
```
 -0.4942938   2.9351363
```

```
> coef(fittobit) # More satisfying estimates
```

```
(Intercept):1 (Intercept):2             x
   -3.8344367     0.7185647     4.8562283
```

We can also compare the two models by plotting the observations and the two regression lines. Figure 1 shows the Tobit model (orange line) and the OLS model (red line) produces very different slope coefficients. Ignoring that y is censored, the OLS model underestimates the slope coefficient of x.

```
> # Compare two regression lines
> plot(x,y)
> abline(lm(y~x),col="red",lwd=2,lty=1)
> curve(-3.83444  + 4.85 *x, col="orange", lwd="2", add=TRUE)
```



3. Tobit Model for Left- and Right-Censored Data

We use a dataset named `tobit.csv`. This dataset considers the situation in which we have a measure of academic aptitude (scaled 200-800), which we want to model using reading and math test scores, as well as the type of program the student is enrolled in (academic, general, or vocational). The problem here is that students who answered all questions on the academic aptitude test correctly receive a score of 800, even though it is likely that these students are not "truly" equal in aptitude. The same is true of students who answer all of the questions incorrectly. All such students would score 200, although they may not all be of equal aptitude.

The dataset contains 200 observations. The academic aptitude variable is `apt`, the reading and math test scores are `read` and `math` respectively. The variable `prog` is the type of program the student is in, it is a categorical (nominal) variable that takes on three values:academic (`prog` = 1), general (`prog` = 2), and vocational (`prog` = 3). The variable `id` indexes different students.

Let us start from looking at the dependent variable `apt` descriptively. Note that in this dataset, the lowest

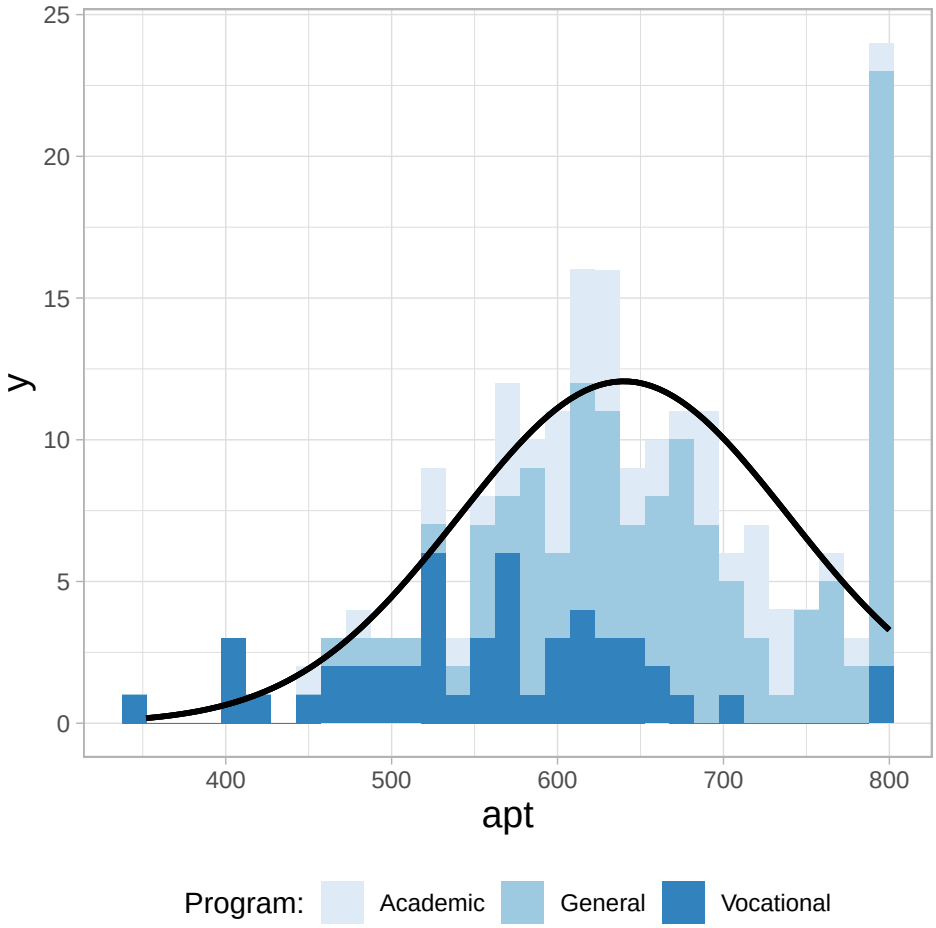
value of apt is 352. That is, no students received a score of 200 (the lowest score possible), meaning that even though censoring at the bottom was possible, it does not occur in the data set.

The following figure is a histogram plot, showing the frequency of different scores. It clearly shows that the censoring in the values of apt. There are far more cases with scores of 750 to 800 than one would expect looking at the rest of the distribution.

```
> mydata<-read.csv("tobit.csv")
> attach(mydata)
> summary(mydata$apt)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  352.0   575.5   633.0   640.0   705.2   800.0
```

```
> f <- function(x, var, bw = 15) {
+   dnorm(x, mean = mean(var), sd(var)) * length(var)  * bw
+ }
>
> ggplot(mydata, aes(x = apt, fill=prog))+
+   stat_bin(binwidth=15) +
+   stat_function(fun = f, size = 1,args = list(var = mydata$apt))+
+   scale_fill_brewer(palette = "Blues",
+                         name = "Program:",
+                         labels = c("Academic",
+                                    "General",
+                                    "Vocational"))+
+   theme_light()+
+   theme(legend.position = "bottom",
+         axis.title = element_text(size = 14))
```

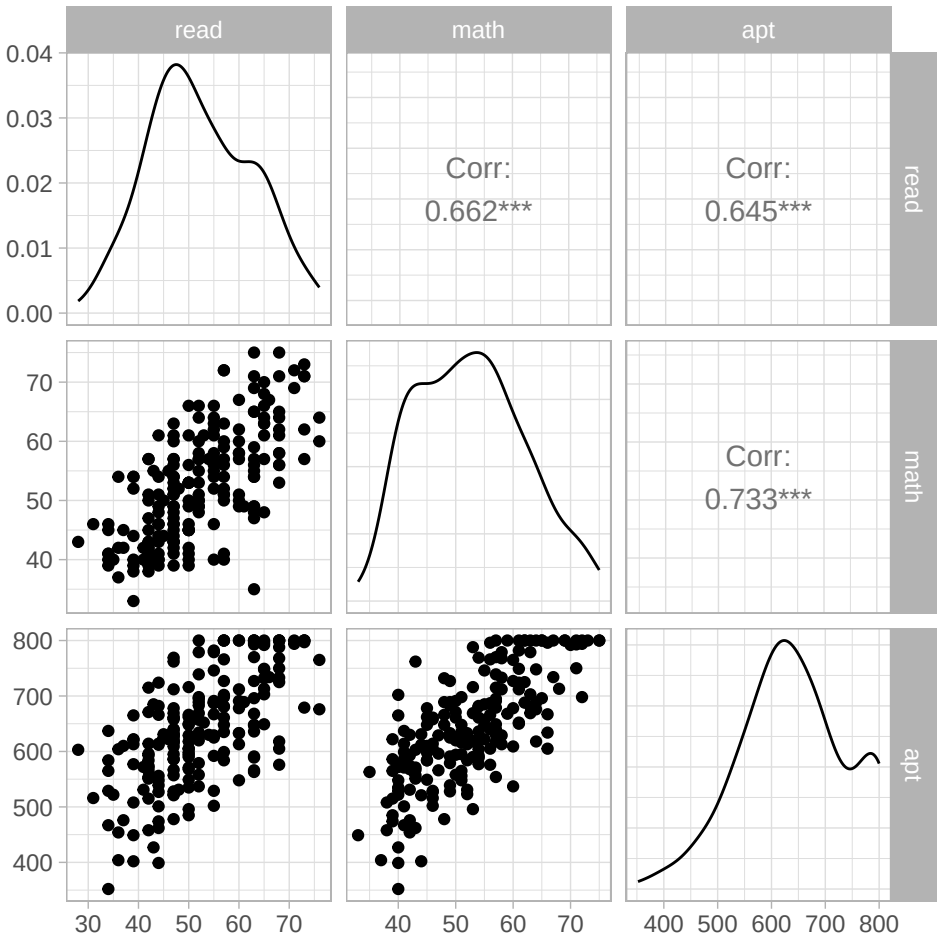


We can also describe variables by using a correlation matrix figure, showing the pair-wise correlations between reading score, math score and the apt scale.

```
> cor(mydata[, c("read", "math", "apt")])
```

```
          read      math       apt
read 1.0000000 0.6622801 0.6451215
math 0.6622801 1.0000000 0.7332702
apt  0.6451215 0.7332702 1.0000000
```

```
> ggpairs(mydata[, c("read", "math", "apt")])+
+   theme_light()
```



Because the data are top-censored, we can we run a Tobit model, using the `vglm` function from the VGAM package.

```
> tobitmodel<-vglm(formula=apt~read+math+as.factor(prog),
+                  family=tobit(Upper=800))
> summary(tobitmodel)
```

```
Call:
vglm(formula = apt ~ read + math + as.factor(prog), family = tobit(Upper = 800))

Coefficients:
                           Estimate Std. Error z value Pr(>|z|)
(Intercept):1             209.55956   32.54590   6.439 1.20e-10 ***
(Intercept):2               4.18476    0.05235  79.944  < 2e-16 ***
read                        2.69796    0.61928   4.357 1.32e-05 ***
math                        5.91460    0.70539   8.385  < 2e-16 ***
as.factor(prog)general    -12.71458   12.40857  -1.025 0.305523
as.factor(prog)vocational -46.14327   13.70667  -3.366 0.000761 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Names of linear predictors: mu, loglink(sd)

Log-likelihood: -1041.063 on 394 degrees of freedom
```

```
Number of Fisher scoring iterations: 5

No Hauck-Donner effect found in any of the estimates
```

Based on this model, we can interpret the results substantively as the following:

- For a one unit increase in `read`, there is a 2.6981 point increase in the predicted value of `apt`.
- A one unit increase in `math` is associated with a 5.9146 unit increase in the predicted value of `apt`.
- The terms for `prog` have a slightly different interpretation. The predicted value of `apt` is 46.1419 points lower for students in a vocational program than for students in an academic program.
- We do not observe statistically different predicted `apt` scores between students in a general program and those in an academic program.

We can test the significance of program type overall by fitting a model without the variable " program", and using the likelihood ratio test to compare two models. The LRT with two degrees of freedom is associated with a p-value of 0.0032, indicating that the overall effect of `prog` is statistically significant.

```
> tobitmodel2<-vglm(apt ~ read + math, tobit(Upper = 800),
+                   data = mydata)
> (p <- pchisq(2 * (logLik(tobitmodel) - logLik(tobitmodel2)),
+              df = 2, lower.tail = FALSE))
```

```
[1] 0.003155176
```

4. Heckman Sample Selection Models: Simulated Data

In this section, we use a simulated dataset to show how we can estimate a correctly specified the Heckman selection model with exclusion restriction. We simulate the data by the following steps.

- Using `mvtnorm`, we create bivariate normal disturbances with correlation -0.7. This is the correlation parameter ρ between our selection and outcome equation.
- We generate a uniformly distributed explanatory variable for the selection equation, `xs`, the selection outcome `ys` by Probit data generating process,
- The explanatory variable for the outcome equation `xo` is also drawn from a uniform distribution.
- All our true intercepts are equal to 0 and our true slopes are equal to 1, both in this and the following examples.
- The latent outcome variable is `yoX`, and the observable outcome is `yo`. Note that the vectors of explanatory variables for the selection (xs) and outcome equation (xo) are independent and hence the exclusion restriction is fulfilled.

```
> set.seed(0)
> eps <- rmvnorm(500, c(0, 0),
+      matrix(c(1, -0.7, -0.7, 1), 2, 2))
> # selection
> xs <- runif(500)
> ys <- xs + eps[, 1] > 0
> # outcome
> xo <- runif(500)
> yoX <- xo + eps[, 2]
> yo <- yoX * (ys > 0)
```

Next, we run a Heckman selection model using function `selection` from `sampleSelection` package. The first model component is the selection equation, and the second component is the outcome equation. We see that the estimates are reasonably precise.

```
> summary(selection(ys ~ xs, yo ~ xs))
```

```
--------------------------------------------
Tobit 2 model (sample selection model)
Maximum Likelihood estimation
Newton-Raphson maximisation, 14 iterations
Return code 8: successive function values within relative tolerance limit (reltol)
Log-Likelihood: -721.7615
500 observations (172 censored and 328 observed)
6 free parameters (df = 494)
Probit selection equation:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -0.2004     0.1115  -1.798   0.0728 .
xs            1.2955     0.2087   6.206 1.15e-09 ***
Outcome equation:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.1468     0.4658   0.315    0.753
xs            0.2171     0.3805   0.571    0.569
   Error terms:
      Estimate Std. Error t value Pr(>|t|)
sigma  0.89989    0.09899   9.091   <2e-16 ***
rho   -0.35525    0.53106  -0.669    0.504
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
--------------------------------------------
```

Now we repeat the same exercise, but without the exclusion restriction, generating `yo` using `xs` instead of `xo`. The estimates are still unbiased but standard errors are substantially larger in this case. The exclusion restriction–independent information about the selection process–has a certain identifying power that we now have lost. We are solely relying on the functional form identification.

```
> yoX <- xs + eps[, 2]
> yo <- yoX * (ys > 0)
> summary(selection(ys ~ xs, yo ~ xs))
```

```
--------------------------------------------
Tobit 2 model (sample selection model)
Maximum Likelihood estimation
Newton-Raphson maximisation, 14 iterations
Return code 8: successive function values within relative tolerance limit (reltol)
Log-Likelihood: -712.8298
500 observations (172 censored and 328 observed)
6 free parameters (df = 494)
Probit selection equation:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -0.1984     0.1114  -1.781   0.0756 .
xs            1.2907     0.2085   6.191 1.25e-09 ***
Outcome equation:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -0.5499     0.5644  -0.974  0.33038
xs            1.3987     0.4482   3.120  0.00191 **
   Error terms:
```

```
      Estimate Std. Error t value Pr(>|t|)
sigma  0.85091    0.05352  15.899   <2e-16 ***
rho   -0.13226    0.72684  -0.182    0.856
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
--------------------------------------------
```

5. Heckman Selection Model with Observational Data

The dataset used in this example is included in `sampleSelection`, named `Mroz87`. This dataset was used by Mroz (1987) for analyzing female labor supply. In this example, labor force participation (described by dummy `lfp`) is modeled by a quadratic polynomial in age (`age`), family income (`faminc`, in 1975 dollars), presence of children (`kids`), and education in years (`educ`). The wage equation includes a quadratic polynomial in experience (`exper`), education in years (`educ`), and residence in a big city (`city`). First, we estimate the model by the Heckman two-step method.

```
> data("Mroz87")
> Mroz87$kids <- (Mroz87$kids5 + Mroz87$kids618 > 0)
> selectmod1 <- selection(lfp ~ age + I(age^2) + faminc + kids + educ,
+   +   wage ~ exper + I(exper^2) + educ + city, data = Mroz87,
+     method = "2step")
> summary(selectmod1)
```

```
--------------------------------------------
Tobit 2 model (sample selection model)
2-step Heckman / heckit estimation
753 observations (325 censored and 428 observed)
14 free parameters (df = 740)
Probit selection equation:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -4.157e+00  1.402e+00  -2.965 0.003127 **
age          1.854e-01  6.597e-02   2.810 0.005078 **
I(age^2)    -2.426e-03  7.735e-04  -3.136 0.001780 **
faminc       4.580e-06  4.206e-06   1.089 0.276544
kidsTRUE    -4.490e-01  1.309e-01  -3.430 0.000638 ***
educ         9.818e-02  2.298e-02   4.272 2.19e-05 ***
Outcome equation:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.9712003  2.0593505  -0.472    0.637
exper        0.0210610  0.0624646   0.337    0.736
I(exper^2)   0.0001371  0.0018782   0.073    0.942
educ         0.4170174  0.1002497   4.160 3.56e-05 ***
city         0.4438379  0.3158984   1.405    0.160
Multiple R-Squared:0.1264,  Adjusted R-Squared:0.116
   Error terms:
             Estimate Std. Error t value Pr(>|t|)
invMillsRatio   -1.098      1.266  -0.867    0.386
sigma            3.200         NA      NA       NA
rho             -0.343         NA      NA       NA
--------------------------------------------
```

In this exercise, we are interested in examining the determinants of female workers' wages. To do so, we have to consider the selection process that determines female labor participation. The above model shows that education affects both labor participation and wages. Notice that the above model also produces "NAs" for sigma and ρ, meaning we have a nonpositive and definite variance-covariance matrix. This could be because

we add quadratic polynomial terms in both equations. When this issue occurs, we can consider the ML estimation.

```
> selectmod2 <- selection(lfp ~ age + I(age^2) + faminc + kids + educ,
+           +wage ~ exper + I(exper^2) + educ + city, data = Mroz87,
+           maxMethod = "BHHH", iterlim = 500)
> summary(selectmod2)

--------------------------------------------
Tobit 2 model (sample selection model)
Maximum Likelihood estimation
BHHH maximisation, 62 iterations
Return code 8: successive function values within relative tolerance limit (reltol)
Log-Likelihood: -1581.259
753 observations (325 censored and 428 observed)
13 free parameters (df = 740)
Probit selection equation:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -4.120e+00  1.410e+00  -2.921  0.00359 **
age          1.840e-01  6.584e-02   2.795  0.00532 **
I(age^2)    -2.409e-03  7.735e-04  -3.115  0.00191 **
faminc       5.676e-06  3.890e-06   1.459  0.14493
kidsTRUE    -4.507e-01  1.367e-01  -3.298  0.00102 **
educ         9.533e-02  2.400e-02   3.973  7.8e-05 ***
Outcome equation:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.9537242  1.6745690  -1.167    0.244
exper        0.0284295  0.0753989   0.377    0.706
I(exper^2)  -0.0001151  0.0023339  -0.049    0.961
educ         0.4562471  0.0959626   4.754 2.39e-06 ***
city         0.4451424  0.4255420   1.046    0.296
   Error terms:
      Estimate Std. Error t value Pr(>|t|)
sigma  3.10350    0.08368  37.088   <2e-16 ***
rho   -0.13328    0.22296  -0.598     0.55
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
--------------------------------------------
```