Problem Set 2

Advanced Microeconometrics

Instituto Tecnológico Autónomo de México

Carlos Lezama

The Data

The following results are inspired by Rau, Sanchez, and Urzua (2019). The file `data.csv` contains real data for Chilean individuals that track their schooling and labor market decisions and outcomes from their early teens through their late twenties.

The goal is to estimate a generalized Roy model for the decision to attend a private high school in lieu of a public high school, and labor market outcomes, i.e., wages. The model includes an unobserved factor that approximates a combination of individuals' scholastic abilities.

The following is a description of the variables in the data:

- `test_X` is students' performance on standardized test `X`. The unit measure is standard deviations.
- `privateHS` is a dummy indicating having attended a private high school.
- `wage` is the natural log of wage.
- `male`, `momschoolingX`, `dadschoolingX`, `broken_homeX`, `incomehhX`, `north`, `center` are demographic variables.
- `share_private` and `avg_price` are instruments for the decision of attending a private high school, and denote the local share of private high schools and the average fees local private high schools charge.

The Model

Formally, the model includes potential outcomes as follows,

$$Y_1 = X\beta_1 + \theta\alpha_1 + U_1,$$
$$Y_0 = X\beta_0 + \theta\alpha_0 + U_0,$$

where Y_1 is the potential outcome (i.e., `wages`) in the counterfactual of attending a private high school, and Y_0 is similarly defined for the counterfactual of attending a public high school. All relevant observable demographics are included in X, while θ is the unobserved one-dimensional factor (i.e., ability) determining labour market outcomes. The unobserved factor is normally distributed with mean zero and standard deviation σ_θ. The terms U_1 and U_0 are idiosyncratic error terms that are normally distributed with mean zero and standard deviations σ_1 and σ_0, respectively.

Individual decide wheter or not to attend a private high school based on a latent variable I:

$$I = Z\gamma + \theta\alpha_I + V,$$

where Z include observable demographics and instruments, and V is an idiosyncratic error term with mean zero and unit variance. Note that the unobservable factor is also present in this part of the model. We can thus define a binary variable D indicating treatment status,

$$D = 1[I \geq 0].$$

The model includes a measurement system, that help with the identification of the distribution of θ. Specifically,

$$T_k = W\omega_k + \theta\alpha_{T_k} + \varepsilon_k, \quad \forall k = 1, 2, 3, 4,$$

where T_k is the test score k, W include demographics determining test scores, and ε_k normally distributed error term with mean zero and standard deviation σ_{ε_k}.

Finally, we assume that the error terms in the model are all independent from each other conditioning on the observables and the unobserved factor, i.e., $U_1 \perp\!\!\!\perp U_0 \perp\!\!\!\perp V \perp\!\!\!\perp \varepsilon \mid X, Z, W, \theta$, and θ is independent of all observables.

For the empirical implementation of the model, in X include `male`, `north`, and `center`. In Z include all variables in X plus `share_private` and `avg_price`. In W include `male`, `momschoolingX`, `dadschoolingX`, `broken_homeX`, `incomehhX`, `north`, and `center`. The outcome variable is `wage`. D is `privateHS`. And, the measurement system is comprised by the four test scores `test_X`.

1

As stated in Heckman, et al. (2003), we may have

$$T_k = \theta\alpha_{T_k} + \varepsilon_k, \quad \forall k = 1, 2, 3, 4,$$

such that $\mathrm{Cov}(T_i, T_j) = \alpha_{T_i}\alpha_{T_j}\sigma_\theta^2$ for any $i \neq j$. This way, we can obtain the values for the rest of the loadings in the measuring system such that

$$\frac{T_k}{\alpha_{T_k}} = \theta + \varepsilon_{T_k}^*, \quad \forall k = 1, 2, 3, 4,$$

where $\varepsilon_{T_k}^* = \varepsilon_{T_k}/\alpha_{T_k}$. Thus, we can compute the densities of every ε_{T_k}, and θ. The assumption that one of the loadings in the measuring system is equal to one is necessary in order to compute the other ones; otherwise, we may have n equations and $n+1$ unknowns. Namely, we can approximate the distribution of θ — assuming mean zero — as a mixture of normal distributions as shown below:

$$\theta = p\mathcal{N}(\mu_1, \sigma_1^2) + (1-p)\mathcal{N}(\mu_2, \sigma_2^2).$$

2

We may expect that our fitted model coefficients are unbiased, but there is a endogeneity problem in our case since our error terms contain part of the "unobserved abilities". Therefore, our assumptions for our OLS to estimate the **best linear predictor** do not hold.

```
Call:
lm(formula = wage ~ privateHS + male + north + center, data = data.frame(data))

Residuals:
    Min      1Q  Median      3Q     Max 
-6.6353 -0.4254  0.2473  0.7468  2.4423 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  7.75885    0.05362 144.689  < 2e-16 ***
privateHS    0.12157    0.04527   2.685  0.00729 ** 
male         0.29620    0.04332   6.837 9.86e-12 ***
north        0.48118    0.10199   4.718 2.50e-06 ***
center       0.25462    0.05103   4.990 6.41e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.151 on 2834 degrees of freedom
Multiple R-squared:  0.03264,	Adjusted R-squared:  0.03128 
F-statistic: 23.91 on 4 and 2834 DF,  p-value: < 2.2e-16
```

3

By 2SLS, with $D \sim Z$ on the first stage, and $Y \sim D$ on the second stage; our coefficient and standard error related to `privateHS*` increased. So we may have to investigate further for more significant results.

```
Call:
lm(formula = wage ~ privateHS_est, data = data_2sls)

Residuals:
    Min      1Q  Median      3Q     Max 
-6.8361 -0.3975  0.2367  0.7473  2.4764 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)    8.05902    0.05639 142.911   <2e-16 ***
privateHS_est  0.15519    0.06121   2.536   0.0113 *  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.168 on 2837 degrees of freedom
Multiple R-squared:  0.002261,	Adjusted R-squared:  0.001909 
F-statistic: 6.429 on 1 and 2837 DF,  p-value: 0.01128
```

with

```
privateHS_est <- round(lm(d ~ z)$fitted.values)
```

4

New coefficient of D on Y: `0.11297`.

```
Call:
lm(formula = wage ~ privateHS + male + north + center + test_lect + 
    test_mate + test_soc + test_nat, data = data.frame(data))

Residuals:
    Min      1Q  Median      3Q     Max 
-6.5270 -0.4147  0.2607  0.7434  2.3179 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  7.75909    0.05374 144.381  < 2e-16 ***
privateHS    0.11297    0.04524   2.497   0.0126 *  
male         0.29951    0.04474   6.694 2.60e-11 ***
north        0.46606    0.10179   4.579 4.88e-06 ***
center       0.25163    0.05094   4.940 8.27e-07 ***
test_lect    0.02457    0.03557   0.691   0.4898    
test_mate    0.08600    0.03364   2.557   0.0106 *  
test_soc    -0.02924    0.03234  -0.904   0.3660    
test_nat     0.02231    0.03335   0.669   0.5036    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.148 on 2830 degrees of freedom
Multiple R-squared:  0.03915,	Adjusted R-squared:  0.03643 
F-statistic: 14.41 on 8 and 2830 DF,  p-value: < 2.2e-16
```

5

ATE

> Average of the treatment gains over the entire student population.

$$ATE = \iint E\left[Y_1 - Y_0 \mid X = x, \theta = \hat{\theta}\right] dF_{X,\theta}(x, \hat{\theta})$$

TT

> Average of the treatment gains over the subset of students that actually choose to be treated.

$$TT = \iint E\left[Y_1 - Y_0 \mid X = x, \theta = \hat{\theta}, D = 1\right] dF_{X,\theta|D=1}(x, \hat{\theta})$$

MTE

> Average of the treatment gains over the subset of students who would be indifferent betweent choosing to be treated or not.

$$TT = \iint E\left[Y_1 - Y_0 \mid X = x, \theta = \hat{\theta}, V = Z\gamma - \theta\alpha_I\right] dF_{X,\theta|V=Z\gamma-\theta\alpha_I}(x, \hat{\theta})$$

6

Likelihood function of the model

$$\begin{aligned}\mathcal{L} = \prod_i \int_{\mathbb{R}} \Bigg[& \prod_{j=1}^{4} \left\langle \phi_{0,\sigma^2_{T_j}}\left(T_j - W\omega_j - \theta\alpha_{T_j}\right)\right\rangle \\ & \times \left(1 - \Phi\left(Z\gamma + \theta\alpha_I\right)\right)^{1-D_i} \\ & \times \phi_{0,\sigma_0^2}\left(Y_0 - X\beta_0 - \theta\alpha_0\right) \\ & \times \phi_{0,\sigma_1^2}\left(Y_1 - X\beta_1 - \theta\alpha_1\right) \\ & \times \Phi(Z\gamma + \theta\alpha_I)^{D_i} \\ & \times \phi_{0,\sigma_\theta^2}\left(\theta\right)\Bigg] d\theta\end{aligned}$$

7

For simplicity, `MATLAB` was used to compute the optimal values for $\mathcal{L}$. Some functions are described below:

- Ordinary least squares

```
% OLS.m
function [beta, sigma] = OLS(X, Y)
    beta = inv(X' * X) * X' * Y;
    er = Y - X*beta;
    sigma = sqrt((er' * er) / (size(Y, 1) - size(X, 2)));
end
```

- $\log(\mathcal{L})$

```
% roy_likelihood.m
function output = roy_loglikelihood(initial_guess)

    global D T W X Y Z

    [...]

    ithL = @(theta) (normpdf(Y - X * beta_0 - theta * alpha_0, 0, exp(sigma_0)) ...
        .* normpdf(Y - X * beta_1 - theta * alpha_1, 0, exp(sigma_1)) ...
        .* normpdf(T(:, 1) - W * beta_T1 - theta, 0, exp(sigma_T1)) ...
        .* normpdf(T(:, 2) - W * beta_T2 - theta * alpha_T2, 0, exp(sigma_T2)) ...
        .* normpdf(T(:, 3) - W * beta_T3 - theta * alpha_T3, 0, exp(sigma_T3)) ...
        .* normpdf(T(:, 4) - W * beta_T4 - theta * alpha_T4, 0, exp(sigma_T4)) ...
        .* (1 - normcdf(Z * beta_D + theta * alpha_I, 0, 1)) .^ (1 - D) ...
        .* normcdf(Z * beta_D + theta * alpha_I, 0, 1) .^ D ...
        .* normpdf(theta, 0, exp(sigma_theta)));

    q = integral(ithL, -Inf, Inf, 'ArrayValued', true);

    output = -sum(log(q));

end
```

Aforementioned functions helped to compute optimal parameters with a runtime of almost 45 minutes using the following master code:

```
tic
options = optimoptions(@fminunc, 'Algorithm', 'quasi-newton', 'Display', 'iter',...
    'GradObj', 'off', 'HessUpdate', 'bfgs', 'UseParallel', false,...
    'TolFun', 1e-6, 'TolX', 1e-6, 'MaxIter', 1e6, 'MaxFunEvals', 1e6);
[estimates, estimatesF, exitflag, output, grad, hessian] = fminunc('roy_loglikelihood', initial_guess,
options);

runtime = toc;

se = sqrt(diag(inv(hessian)));
```

The table below describes the output results.

Parameter	Estimated Value	Standard Error
β_0	7.81921708	0.04870465
β_0	0.30213284	0.04328226
β_0	0.47635371	0.10201014
β_0	0.2734988	0.05056296
β_1	7.8192177	0.04870461
β_1	0.30213257	0.04328222
β_1	0.47635472	0.10201044

β_1	0.27349878	0.05056286
γ	-0.4353483	0.08313967
γ	0.1362415	0.04872624
γ	-0.1255597	0.11287811
γ	0.27133711	0.06139946
γ	0.84976264	0.12406831
γ	-0.1442248	0.27753458
ω_1	-0.1439185	0.05197017
ω_1	-0.2302358	0.033553
ω_1	0.23011268	0.04289702
ω_1	0.28724934	0.06432697
ω_1	-0.1596855	0.09643413
ω_1	0.1715684	0.04324387
ω_1	0.31945617	0.06239367
ω_1	-0.0757068	0.08807978
ω_1	0.00521931	0.04304264
ω_1	0.28378209	0.11002384
ω_1	0.12959975	0.04906602
ω_1	0.20919163	0.05453735
ω_1	-0.0065869	0.10816752
ω_1	0.02681309	0.07944891
ω_1	0.04333963	0.03976408
ω_2	-0.265488	0.05135096
ω_2	0.08274793	0.03314235
ω_2	0.18261303	0.04240422
ω_2	0.30313879	0.06358904
ω_2	-0.2159232	0.09532613
ω_2	0.15561507	0.04275075

ω_2	0.23616125	0.06166408
ω_2	-0.0339236	0.08707646
ω_2	0.06489823	0.04254685
ω_2	0.31801493	0.10876592
ω_2	0.09823527	0.04849112
ω_2	0.2084306	0.0539072
ω_2	-0.0080477	0.10693163
ω_2	0.08531422	0.07847687
ω_2	0.00511872	0.03927803
ω_3	-0.3435333	0.05253868
ω_3	0.15621576	0.0339065
ω_3	0.22157547	0.04338929
ω_3	0.38560581	0.06506685
ω_3	-0.0768919	0.09754074
ω_3	0.13860485	0.04374495
ω_3	0.2558973	0.0630931
ω_3	-0.123407	0.08910147
ω_3	0.02093546	0.04353481
ω_3	0.28451287	0.1112932
ω_3	0.18254442	0.04961437
ω_3	0.24637216	0.05515842
ω_3	-0.0221353	0.10941769
ω_3	0.0040397	0.08028637
ω_3	0.01983037	0.04018385
ω_4	-0.230814	0.05358091
ω_4	0.01256138	0.03458659
ω_4	0.18288027	0.04423705
ω_4	0.36445828	0.06633709

ω_4	-0.16968	0.0994465
ω_4	0.14931767	0.04459695
ω_4	0.2966095	0.06433515
ω_4	-0.0446379	0.09083632
ω_4	-0.0183219	0.04438648
ω_4	0.2552503	0.1134633
ω_4	0.10896538	0.050592
ω_4	0.16435029	0.05623886
ω_4	0.0080692	0.11155019
ω_4	0.00102197	0.08189654
ω_4	-0.0113865	0.04098944
σ_0	0.13653706	0.01329632
σ_1	0.13653707	0.01329631
σ_{T_1}	-0.6704545	0.02038455
σ_{T_2}	-0.5504599	0.01710144
σ_{T_3}	-0.501899	0.01669583
σ_{T_4}	-0.5622838	0.01834335
α_0	0.14479556	0.03259601
α_1	0.14479572	0.03259603
α_I	0.03696093	0.03619097
α_{T_2}	0.91102389	0.02197503
α_{T_3}	0.91323262	0.0228353
α_{T_4}	0.98708182	0.0233811
σ_θ	-0.3167767	0.02020435

8

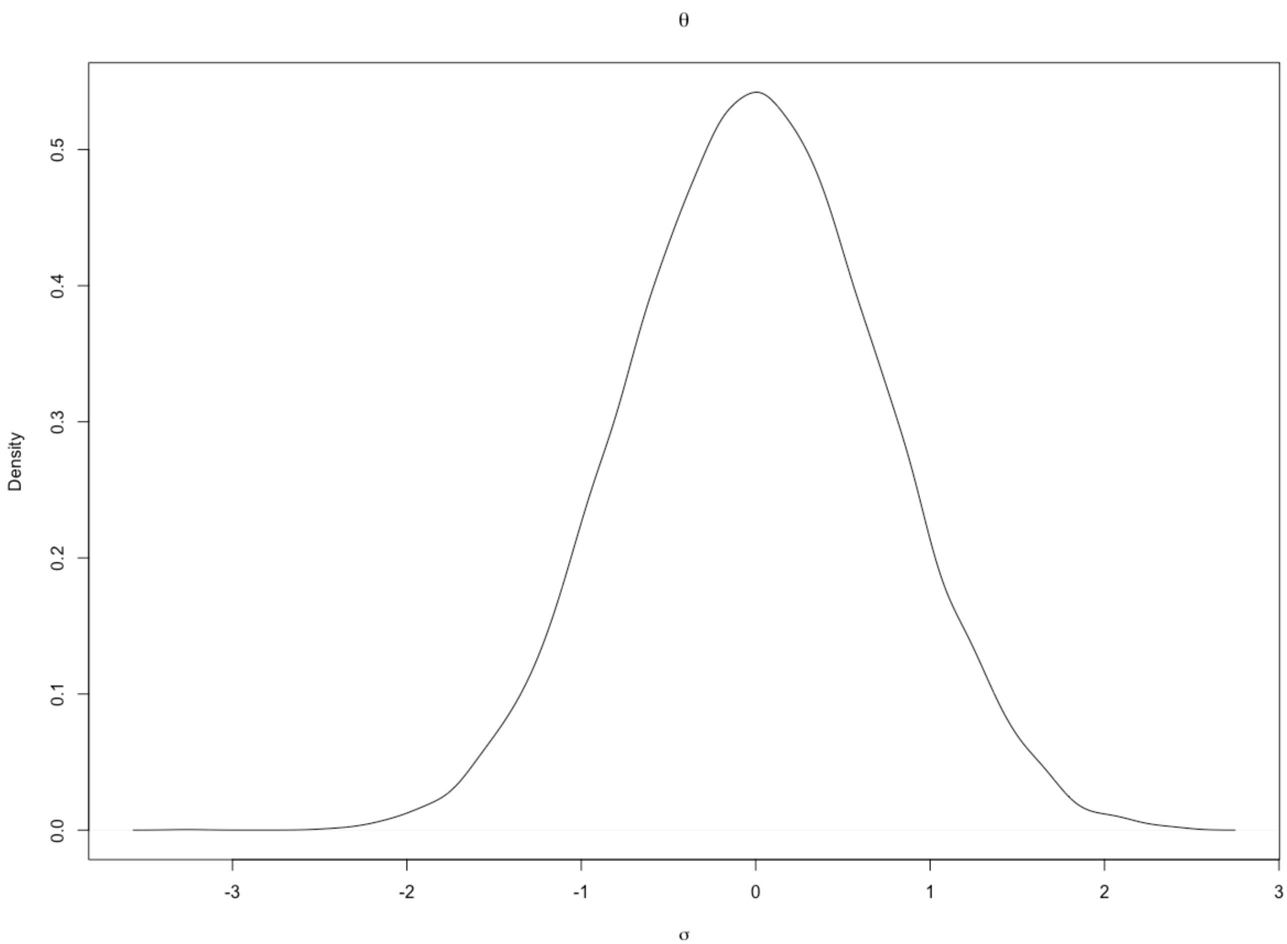
Actual Data Averages

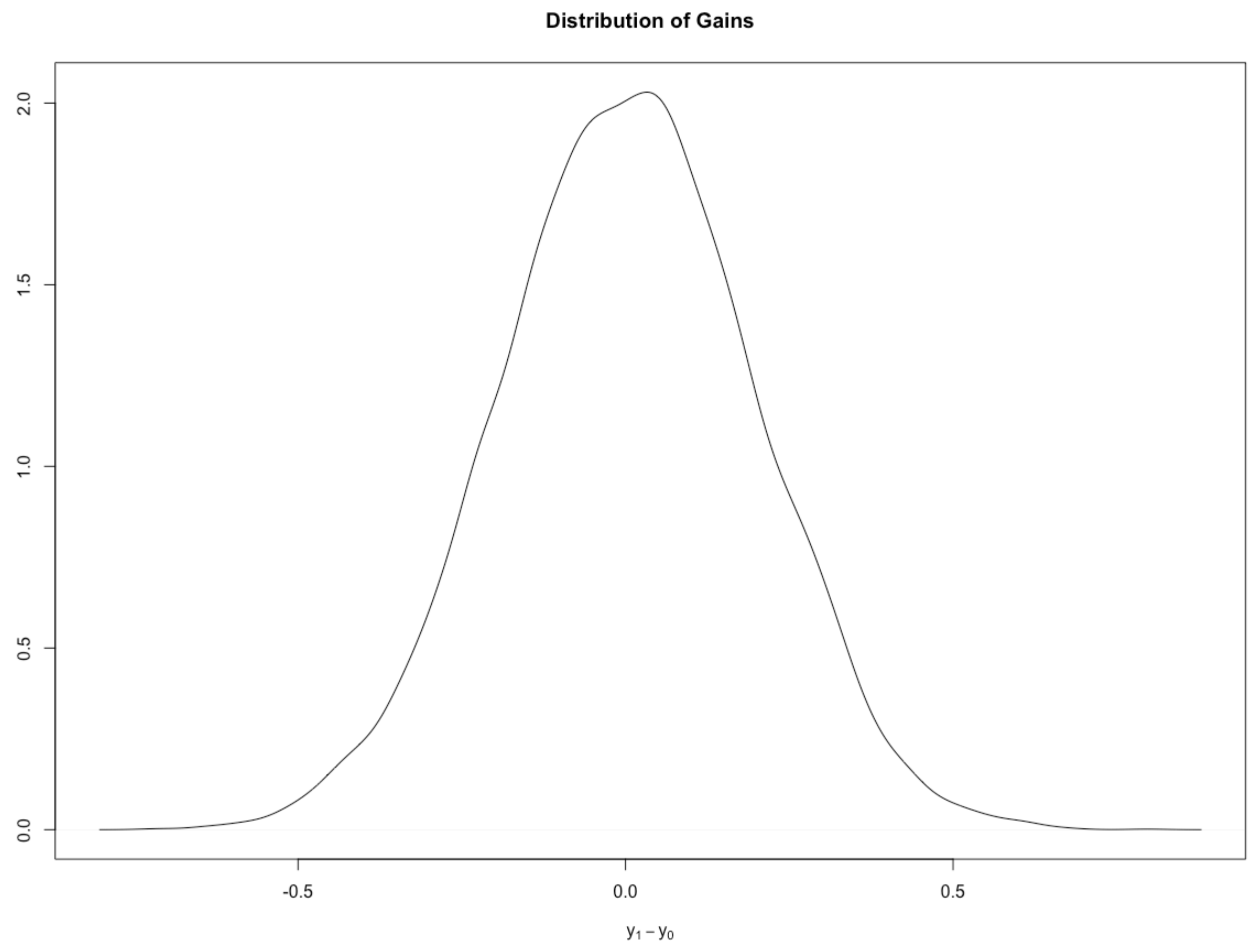
```
           test_lect            test_mate             test_soc
 "Mean   : 0.07369  " "Mean   : 0.07013  " "Mean   : 0.07624  "
           test_nat             privateHS                 male
 "Mean   : 0.03437  "   "Mean   :0.6277  "   "Mean   :0.5114  "
       momschooling2        momschooling3    momschooling_miss
   "Mean   :0.2769  "   "Mean   :0.1067  "   "Mean   :0.1384  "
       dadschooling2        dadschooling3    dadschooling_miss
   "Mean   :0.2895  "   "Mean   :0.1233  "    "Mean   :0.156  "
        broken_home1     broken_home_miss            incomehh2
   "Mean   :0.2166  "   "Mean   :0.1976  "   "Mean   :0.3244  "
           incomehh3         incomehh_miss                north
   "Mean   :0.2691  "   "Mean   :0.2138  "  "Mean   :0.05495  "
              center
   "Mean   :0.6978  "
```

Simulated Data Averages

```
           test_lect             test_mate             test_soc
 "Mean   : 0.07379  " "Mean   : 0.070643  "  "Mean   : 0.07558  "
           test_nat              privateHS                 male
 "Mean   : 0.03494  "    "Mean   :0.6274  "    "Mean   :0.5119  "
       momschooling2         momschooling3    momschooling_miss
   "Mean   :0.2768  "    "Mean   :0.1066  "    "Mean   :0.1385  "
       dadschooling2         dadschooling3    dadschooling_miss
     "Mean   :0.29  "     "Mean   :0.123  "    "Mean   :0.1559  "
        broken_home1      broken_home_miss            incomehh2
   "Mean   :0.2163  "    "Mean   :0.1974  "    "Mean   :0.3242  "
           incomehh3          incomehh_miss                north
   "Mean   :0.2695  "    "Mean   :0.2136  "   "Mean   :0.05503  "
              center
   "Mean   :0.6978  "
```

9




Distribution of Gains
0.0
0.5
1.0
1.5
2.0
-0.5
0.0
0.5
y1 – y0

11

```
ATE <- sim_df %>%
    pull(beta) %>%
    mean()

TT <- sim_df %>%
    filter(decision == 1) %>%
    pull(beta) %>%
    mean()

> ATE
[1] 0.003077
> TT
[1] 0.00288
```