2024-06-04

Selection results (FE from baseline)

3 variables and 1 control.

Selection comparison (Known variance)

We set capacity constraint $\alpha = 0.22$ and FDR constraint $\gamma = 0.05$. We compare the selection of the following methods: 1. TPKWs and PMKWs 2. TPKWs and MLE 3. TPKWs and JS(linear)

Left tail selection result

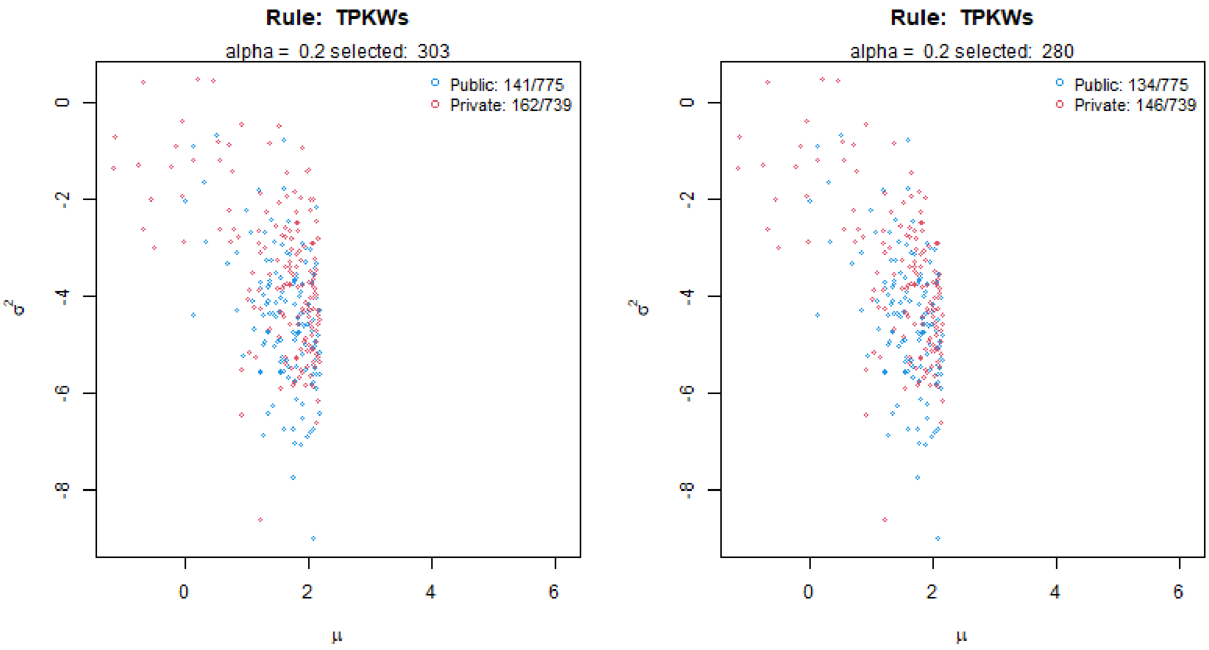
Right tail selection result

Selection comparison (Unknown variance)

Right tail selection

Selection classification

Consider all 4 categories:



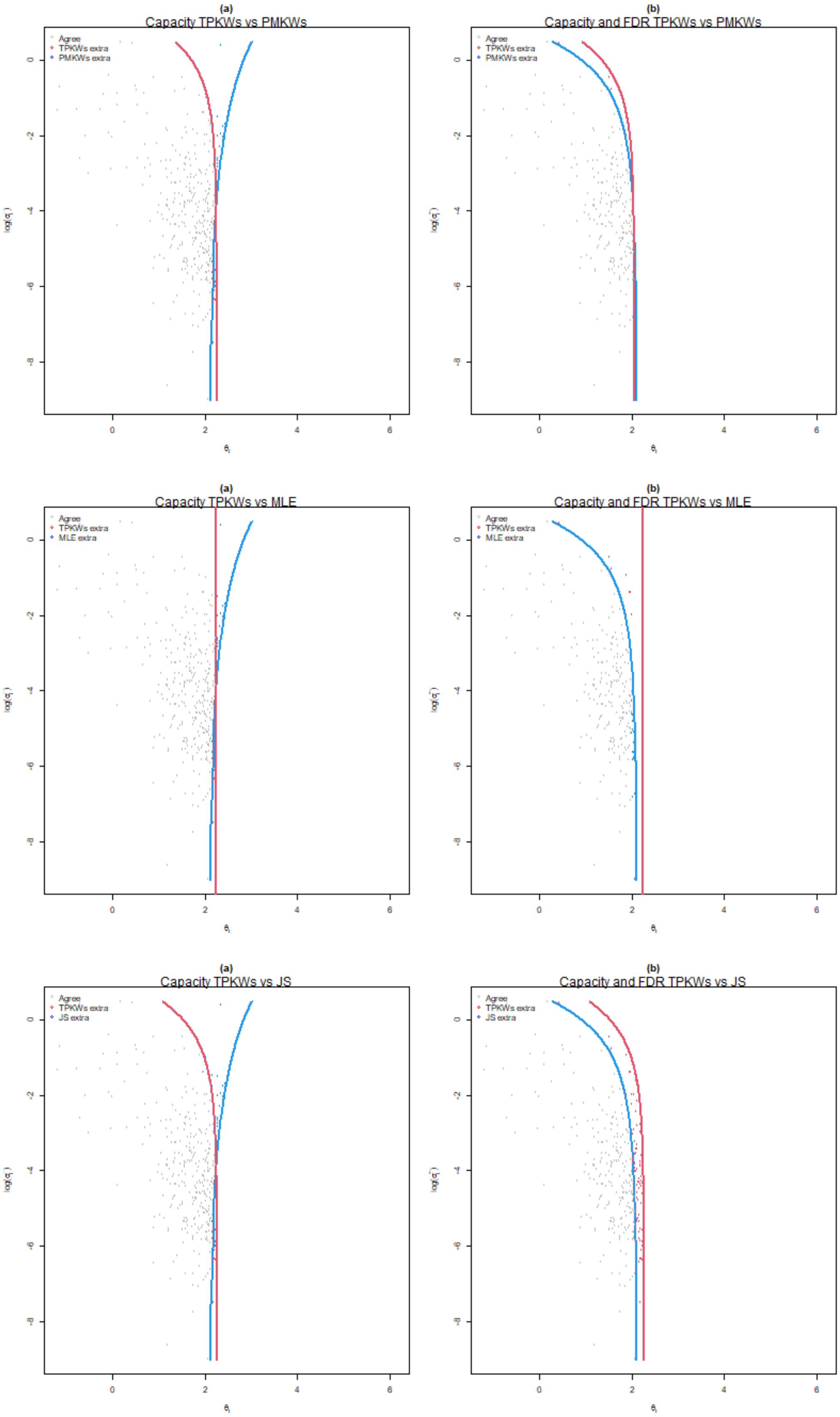


Figure 1: Left tail $\alpha = 0.22$, $\gamma = 0.05$

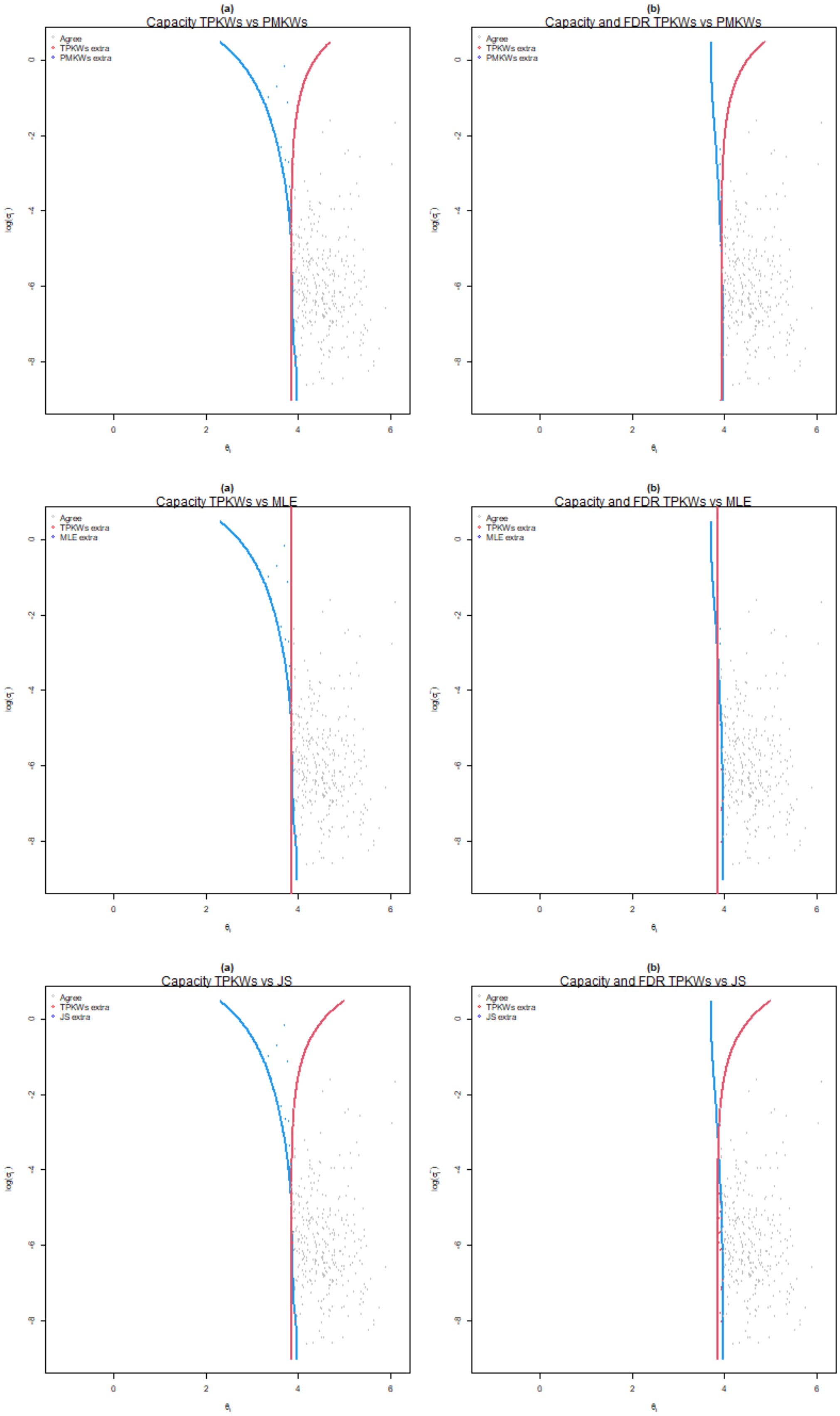


Figure 2: Right tail $\alpha = 0.22$, $\gamma = 0.05$

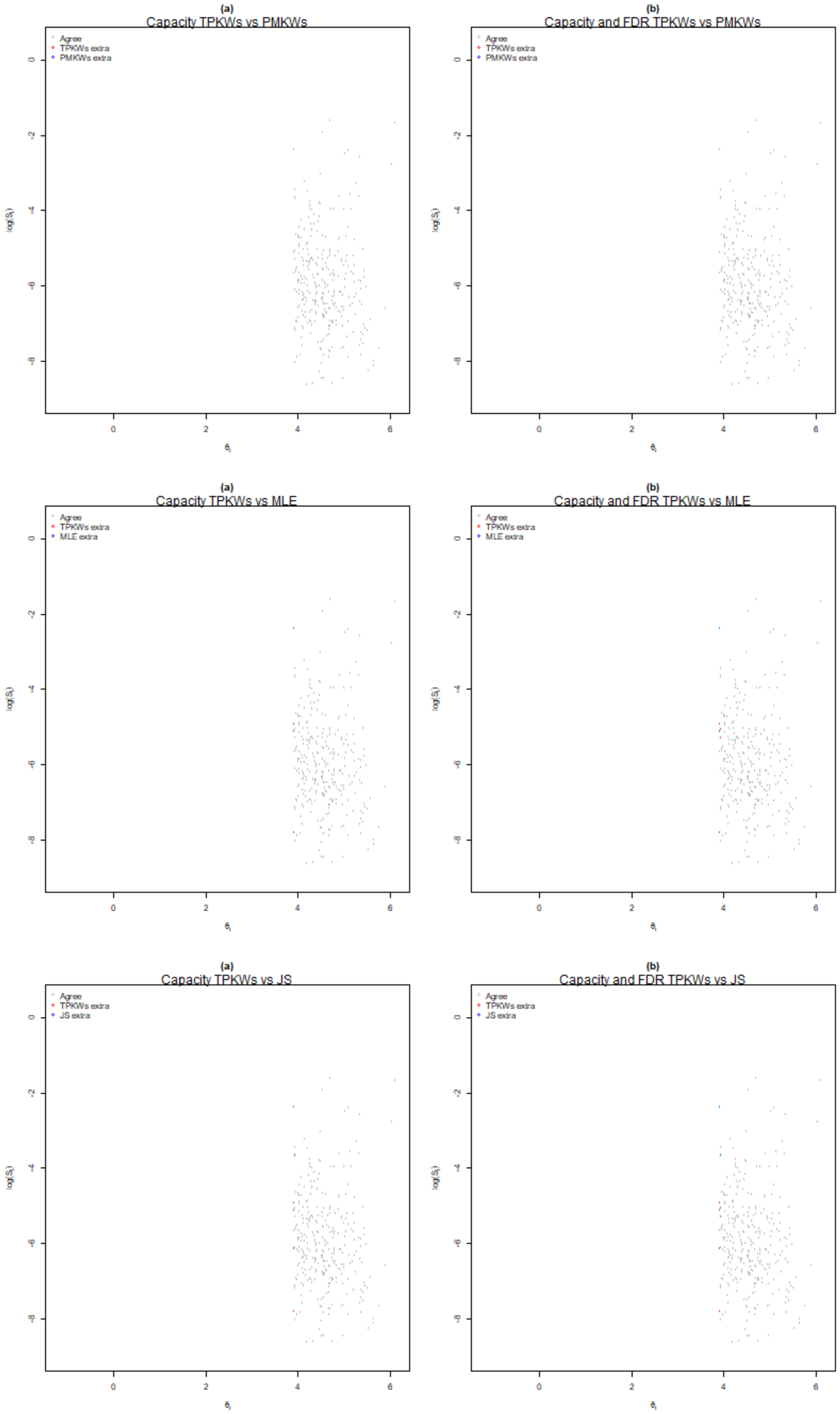
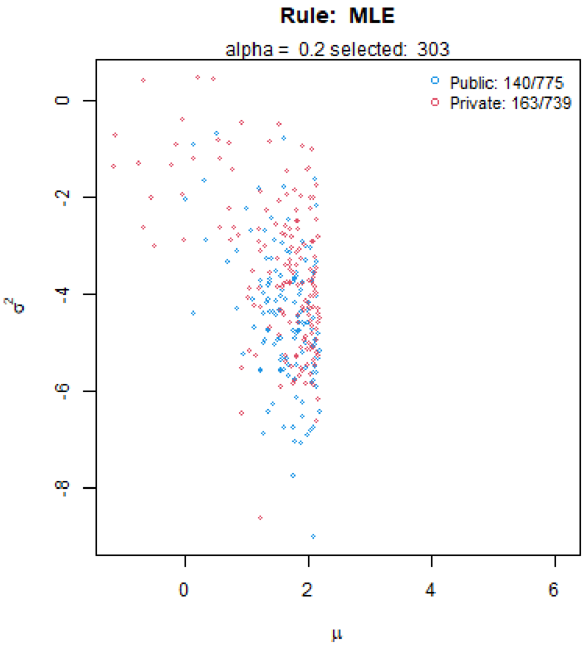
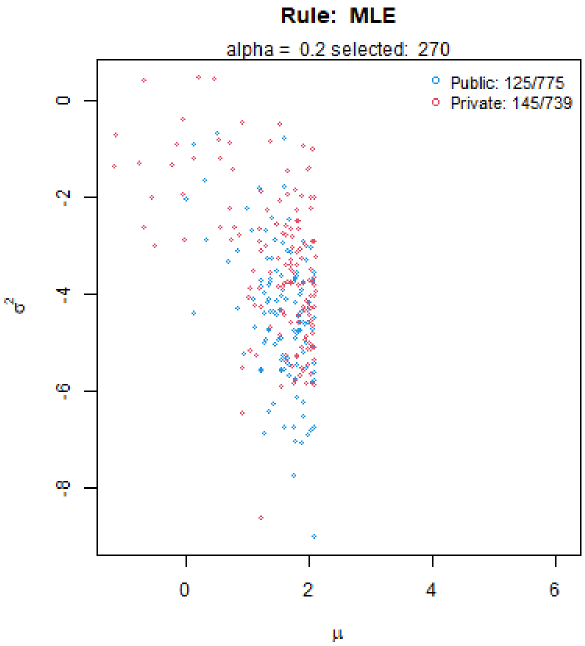


Figure 3: Left tail $\alpha = 0.22$, $\gamma = 0.1$



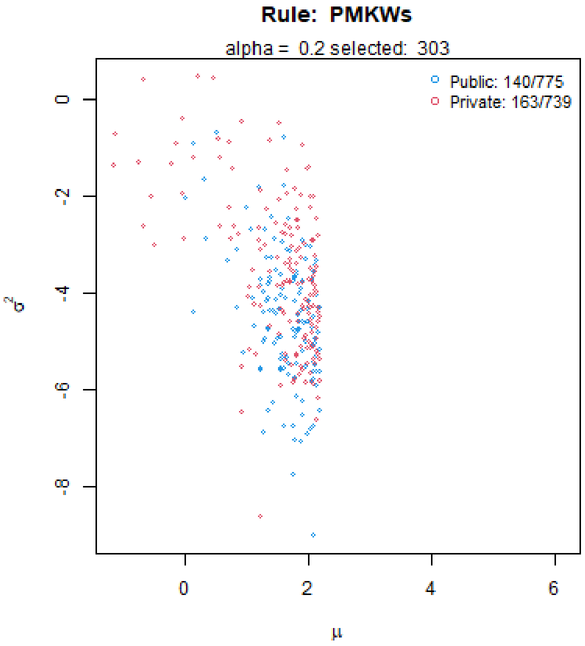
Rule: MLE

alpha = 0.2 selected: 303



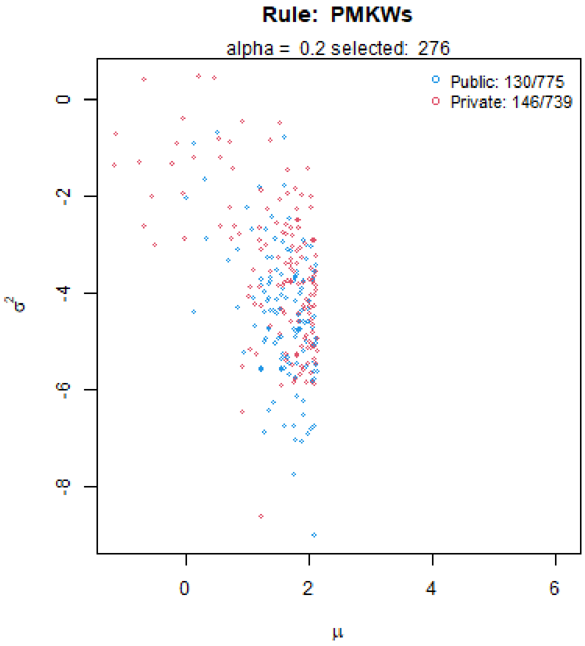
Rule: MLE

alpha = 0.2 selected: 270



Rule: PMKWs

alpha = 0.2 selected: 303



Rule: PMKWs

alpha = 0.2 selected: 276

Consider only 2 categories:

Consider a different capacity constraint $\alpha = 0.40$

Issues

1. Public is not quite inefficient? Cautious about the interpretation of the results.

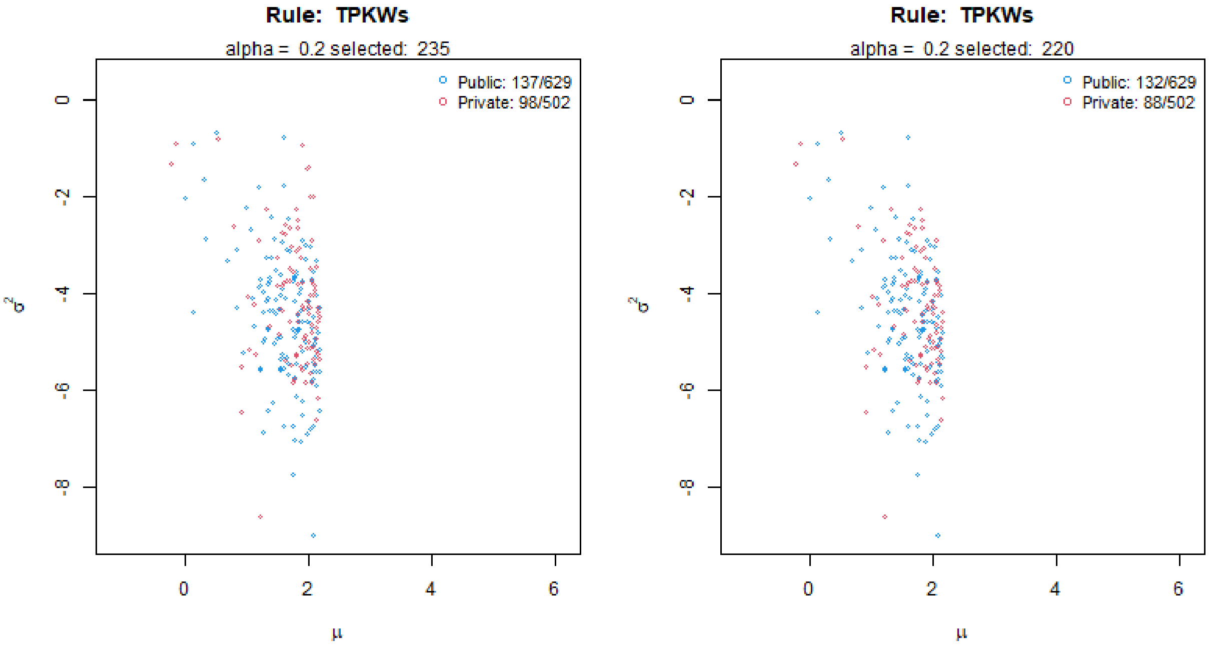


Figure 4: Left tail $\alpha = 0.20$, $\gamma = 0.05$

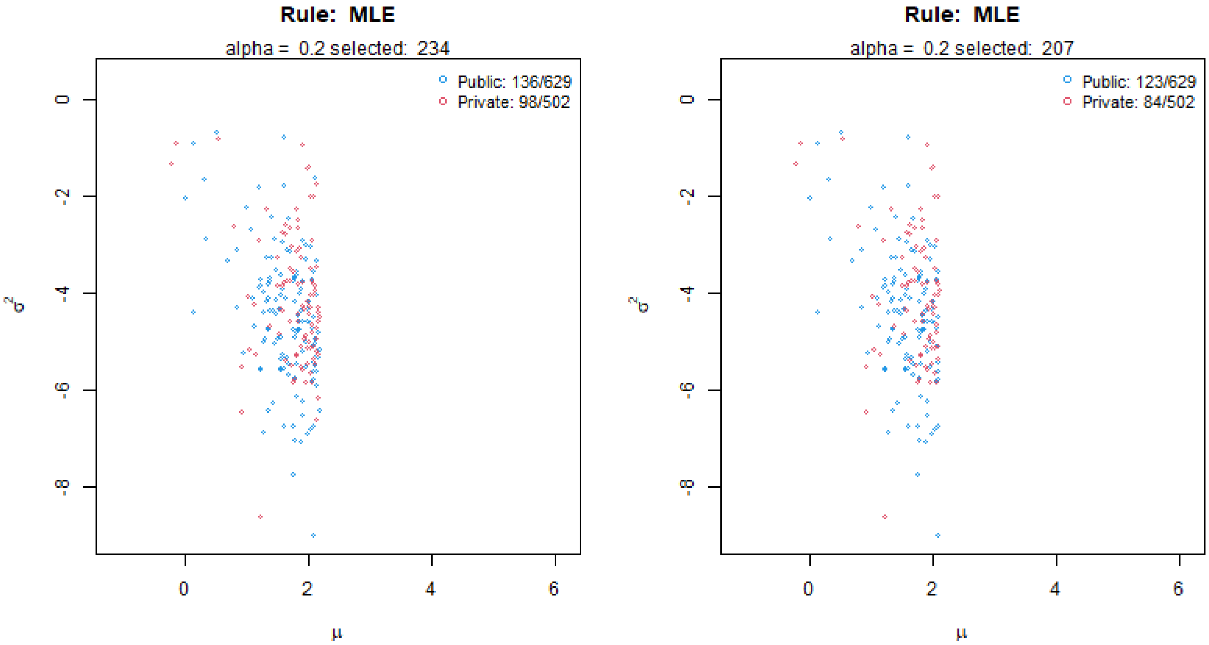


Figure 5: Left tail $\alpha = 0.20$, $\gamma = 0.05$

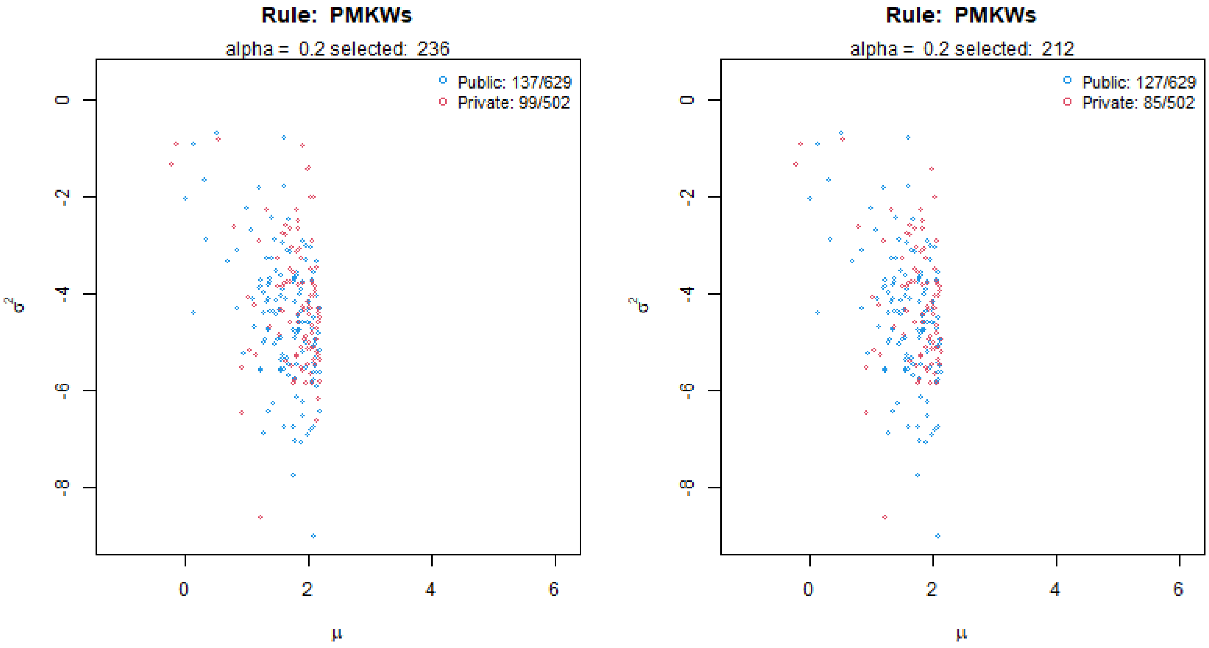


Figure 6: Left tail $\alpha = 0.20$, $\gamma = 0.05$

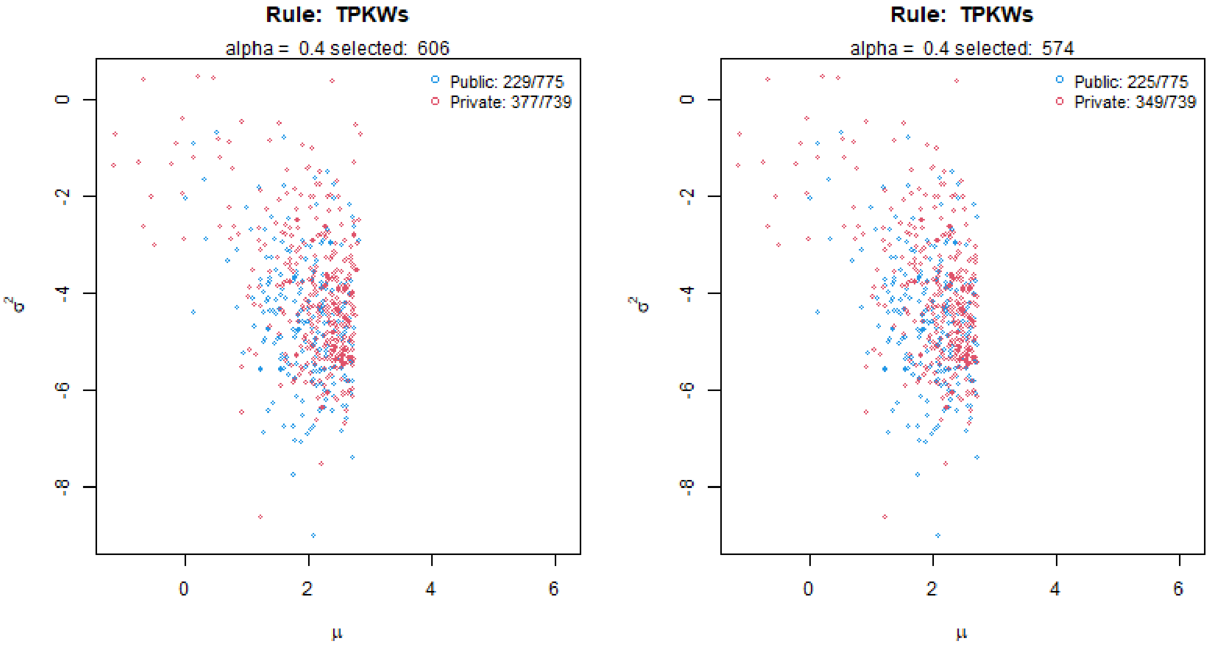


Figure 7: Left tail $\alpha = 0.40$, $\gamma = 0.05$

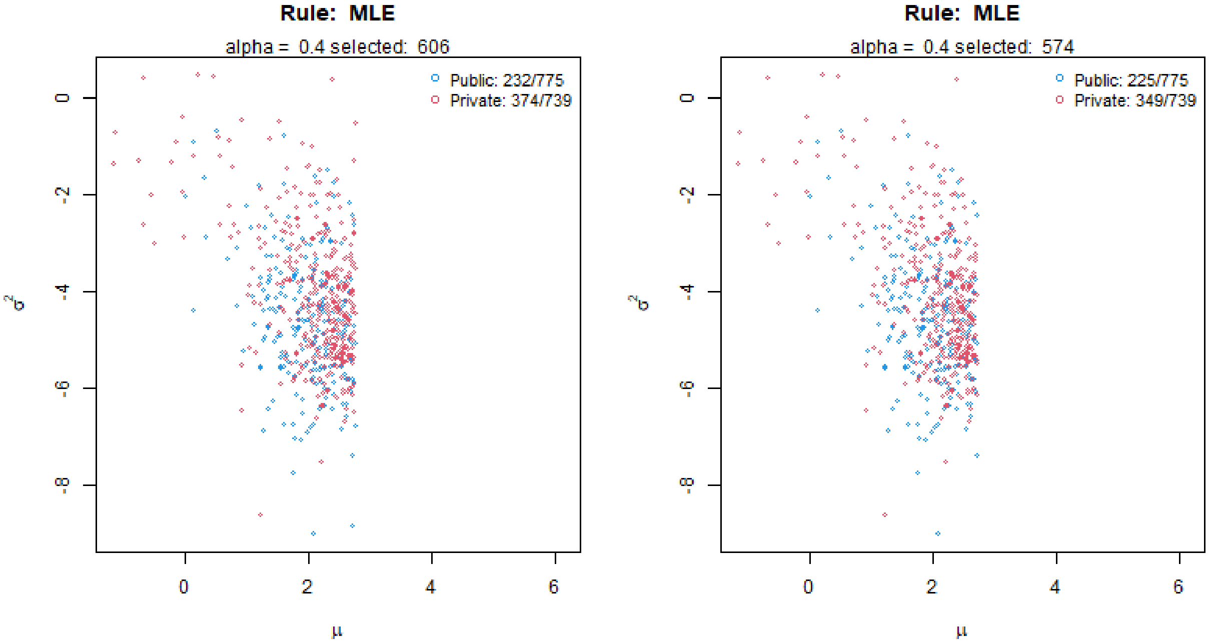


Figure 8: Left tail $\alpha = 0.40$, $\gamma = 0.05$

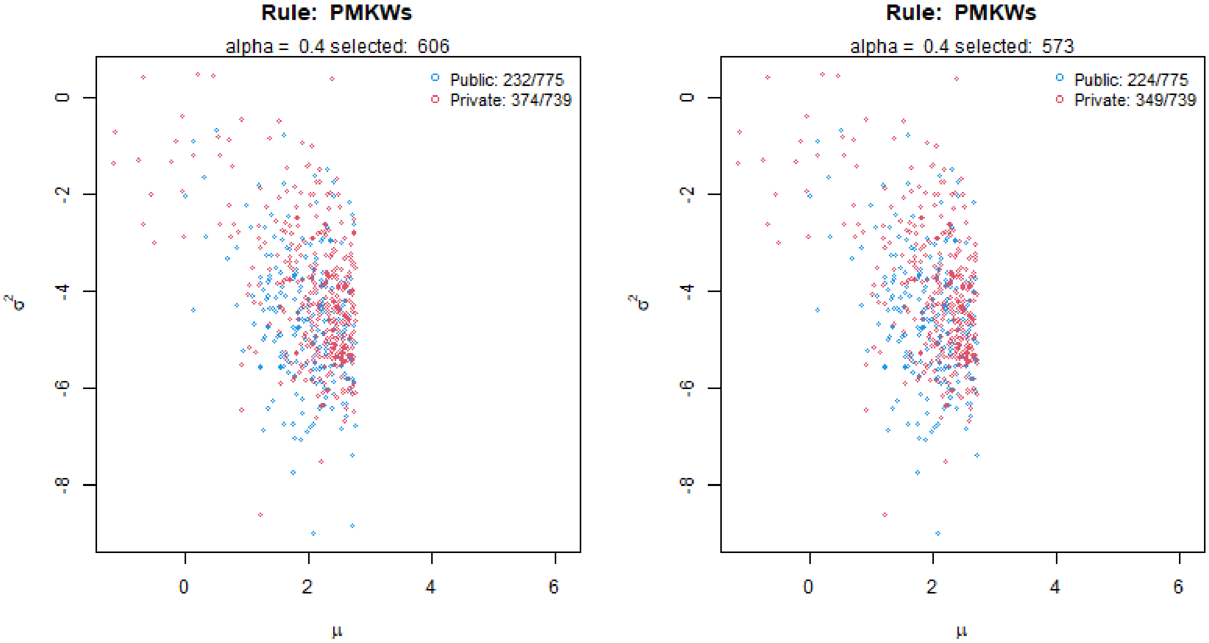


Figure 9: Left tail $\alpha = 0.40$, $\gamma = 0.05$

Regression results

Specification 1

$$\log(ETP\_INF) = \beta_0 + \beta_1 \log(SEJHC\_MCO) + \beta_2 \log(SEJHP\_MCO) + \beta_3 \log(SEANCES\_MED) + \beta_4 \text{CASEMIX} + u_i + \epsilon_{it}$$

Strict exogeneity

regression table

$$E[\epsilon_{it}|x_{i1}, \dots, x_{iT}, z_{i1}, \dots, z_{iT}] = 0$$

	Within-group	Within-group (GLS)	First difference	First difference (GLS)
log(SEJHC_MCO)	0.1283***	0.1088***	0.1136***	0.1063***
	(0.0238)	(0.0059)	(0.0207)	(0.0059)
log(SEJHP_MCO)	0.0307***	0.0212***	0.0226***	0.0206***
	(0.0083)	(0.0024)	(0.0063)	(0.0024)
log(SEANCES_MED)	0.0308***	0.0245***	0.0239***	0.0216***
	(0.0042)	(0.0023)	(0.0046)	(0.0021)
CASEMIX	0.0021**	0.0007	0.0007	0.0007
	(0.0007)	(0.0004)	(0.0006)	(0.0004)
R2	0.1145	0.9910	0.0692	0.9907
Num. obs.	9038	9038	7554	7554
***p < 0.001; **p < 0.01; *p < 0.05				

Endogeneity Current errors affect/correlate with current regressors and future regressors.

$$E[\epsilon_{it}|x_1, \dots, x_{it-1}] = 0$$

We use past values of $\log(SEJHC\_MCO)$ as instruments.

`Arellano-Bond (1991)`

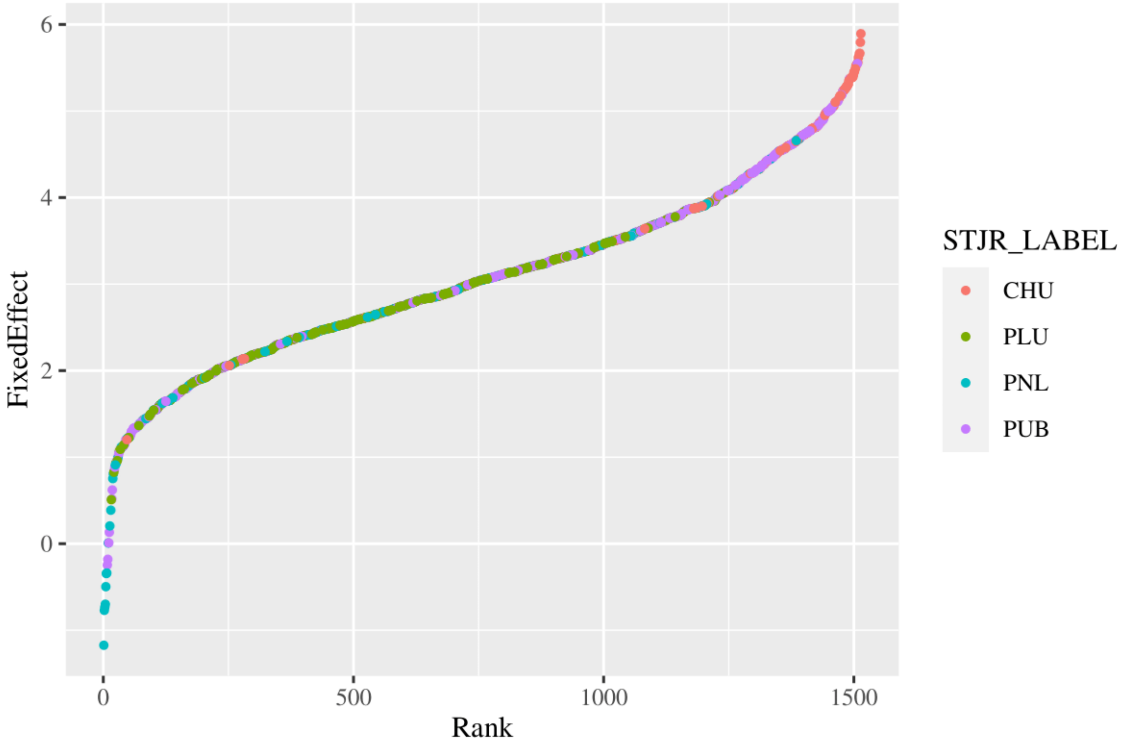
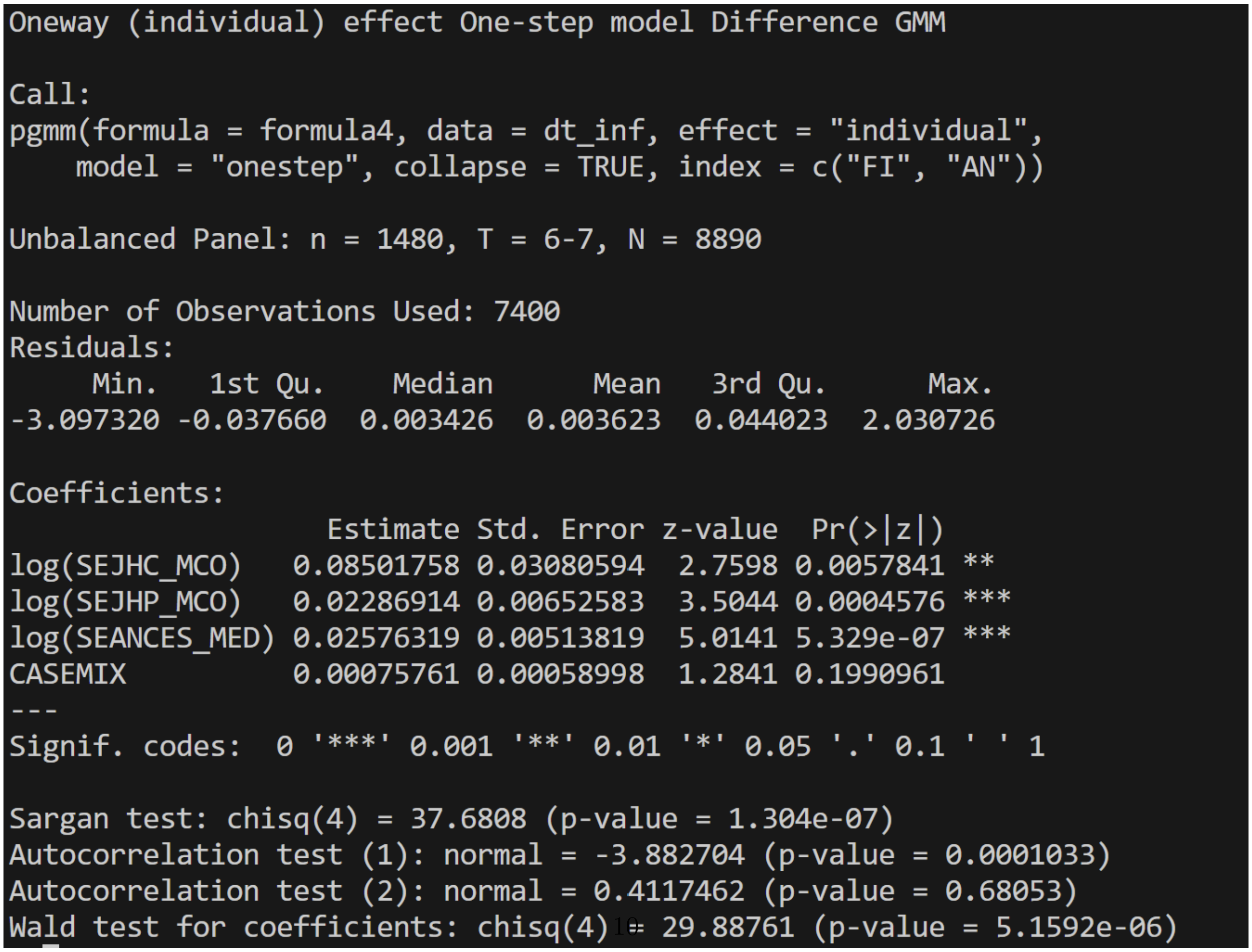
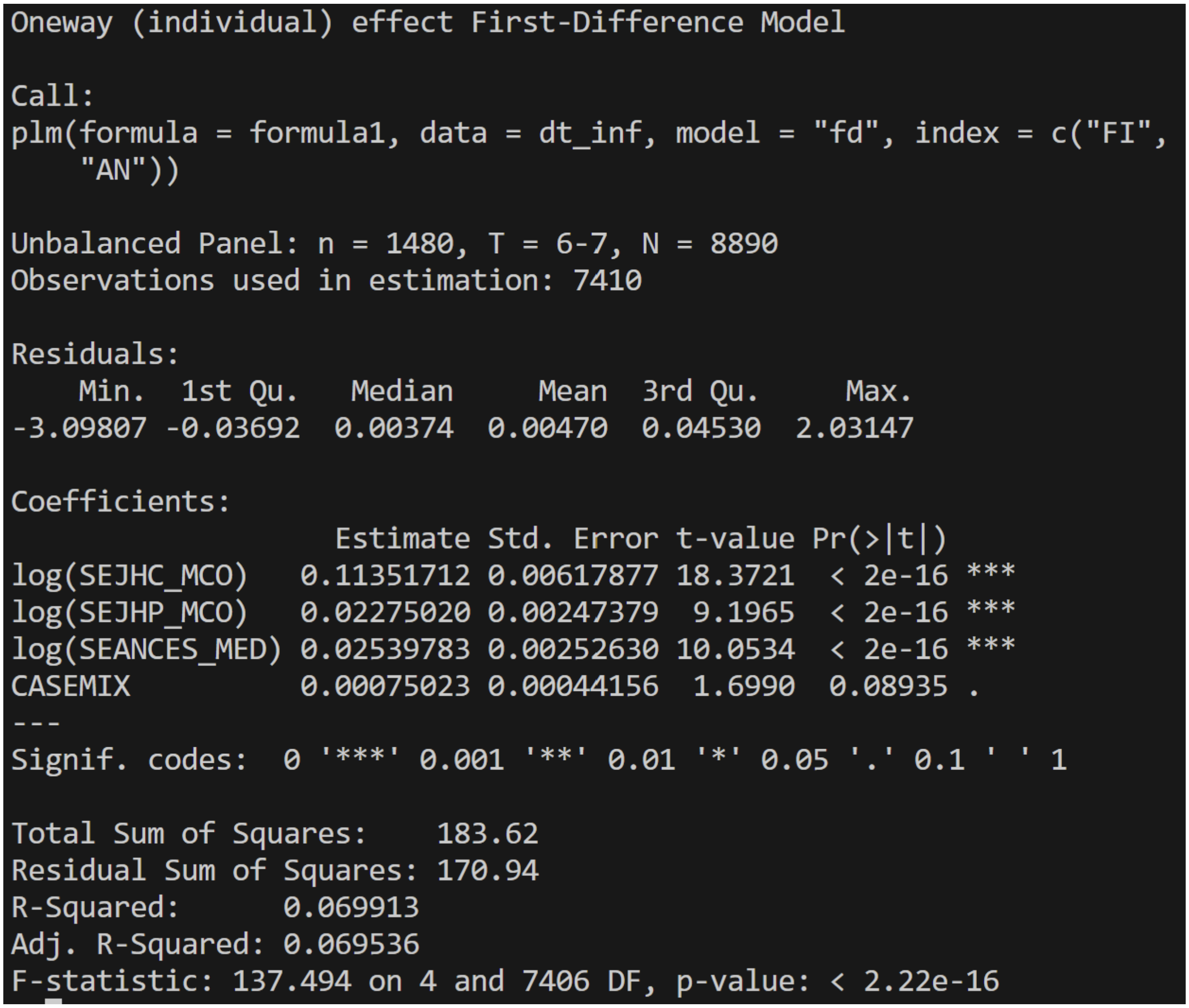


Figure 10: fixed effect



```
Oneway (individual) effect One-step model Difference GMM

Call:
pgmm(formula = formula4, data = dt_inf, effect = "individual",
    model = "onestep", collapse = TRUE, index = c("FI", "AN"))

Unbalanced Panel: n = 1480, T = 6-7, N = 8890

Number of Observations Used: 7400
Residuals:
     Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
-3.097320 -0.037660  0.003426  0.003623  0.044023  2.030726

Coefficients:
                   Estimate Std. Error z-value  Pr(>|z|)
log(SEJHC_MCO)   0.08501758 0.03080594  2.7598 0.0057841 **
log(SEJHP_MCO)   0.02286914 0.00652583  3.5044 0.0004576 ***
log(SEANCES_MED) 0.02576319 0.00513819  5.0141 5.329e-07 ***
CASEMIX          0.00075761 0.00058998  1.2841 0.1990961
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Sargan test: chisq(4) = 37.6808 (p-value = 1.304e-07)
Autocorrelation test (1): normal = -3.882704 (p-value = 0.0001033)
Autocorrelation test (2): normal = 0.4117462 (p-value = 0.68053)
Wald test for coefficients: chisq(4) = 29.88761 (p-value = 5.1592e-06)
```

```
First difference
```



```
Oneway (individual) effect First-Difference Model

Call:
plm(formula = formula1, data = dt_inf, model = "fd", index = c("FI",
    "AN"))

Unbalanced Panel: n = 1480, T = 6-7, N = 8890
Observations used in estimation: 7410

Residuals:
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
-3.09807 -0.03692  0.00374  0.00470  0.04530  2.03147

Coefficients:
                    Estimate Std. Error t-value Pr(>|t|)
log(SEJHC_MCO)    0.11351712 0.00617877 18.3721  < 2e-16 ***
log(SEJHP_MCO)    0.02275020 0.00247379  9.1965  < 2e-16 ***
log(SEANCES_MED) 0.02539783 0.00252630 10.0534  < 2e-16 ***
CASEMIX          0.00075023 0.00044156  1.6990  0.08935 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Total Sum of Squares:    183.62
Residual Sum of Squares: 170.94
R-Squared:      0.069913
Adj. R-Squared: 0.069536
F-statistic: 137.494 on 4 and 7406 DF, p-value: < 2.22e-16
```

Specification 2

	Within-group	Within-group (GLS)	First difference	First difference (GLS)
log(SEJHC_MCO)	0.1262***	0.1072***	0.1122***	0.1049***
	(0.0234)	(0.0059)	(0.0200)	(0.0059)
log(SEJHP_MCO)	0.0198**	0.0135***	0.0152**	0.0128***
	(0.0073)	(0.0027)	(0.0054)	(0.0027)
log(SEANCES_MED)	0.0294***	0.0239***	0.0232***	0.0215***
	(0.0041)	(0.0023)	(0.0045)	(0.0021)
log(PASSU)	0.0019	0.0024	-0.0003	0.0017
	(0.0043)	(0.0043)	(0.0035)	(0.0043)
log(VEN_TOT)	0.0106*	0.0075*	0.0039	0.0024
	(0.0041)	(0.0031)	(0.0031)	(0.0030)
log(SEJ_HTP_TOT)	0.0224**	0.0233***	0.0158*	0.0182***
	(0.0084)	(0.0053)	(0.0066)	(0.0054)
log(PLA_MCO)	0.0472**	0.0409***	0.0416**	0.0424***

	Within-group	Within-group (GLS)	First difference	First difference (GLS)
	(0.0182)	(0.0067)	(0.0161)	(0.0067)
CANCER	0.0018	0.0015*	0.0017	0.0015**
	(0.0012)	(0.0006)	(0.0009)	(0.0006)
CASEMIX	0.0020**	0.0008	0.0008	0.0007
	(0.0007)	(0.0004)	(0.0006)	(0.0004)
R2	0.1267	0.9911	0.0761	0.9908
Num. obs.	9038	9038	7554	7554
***p < 0.001; **p < 0.01; *p < 0.05				

Specification 3

Strict exogeneity

regression table

	Within-group	Within-group (GLS)	First difference	First difference (GLS)
log(SEJHC_MCO)	0.1256***	0.1071***	0.1124***	0.1049***
	(0.0231)	(0.0059)	(0.0196)	(0.0059)
log(SEJHP_MCO)	0.0201**	0.0142***	0.0159**	0.0135***
	(0.0073)	(0.0027)	(0.0055)	(0.0027)
log(SEANCES_MED)	0.0292***	0.0237***	0.0229***	0.0213***
	(0.0041)	(0.0023)	(0.0044)	(0.0021)
log(PASSU)	0.0021	0.0021	-0.0002	0.0015
	(0.0043)	(0.0043)	(0.0034)	(0.0043)
log(VEN_TOT)	0.0106**	0.0075*	0.0039	0.0024
	(0.0041)	(0.0031)	(0.0031)	(0.0029)
log(SEJ_HTP_TOT)	0.0220**	0.0231***	0.0158*	0.0182***
	(0.0084)	(0.0053)	(0.0065)	(0.0054)
log(PLA_MCO)	0.0673*	0.0616***	0.0656**	0.0637***
	(0.0264)	(0.0080)	(0.0234)	(0.0081)
CANCER	0.0018	0.0015**	0.0017*	0.0016**
	(0.0012)	(0.0006)	(0.0009)	(0.0006)
CASEMIX	0.0027**	0.0013**	0.0014	0.0013**
	(0.0009)	(0.0004)	(0.0007)	(0.0004)
log(PLA_MCO):CASEMIX	-0.0013	-0.0013***	-0.0015*	-0.0013***
	(0.0007)	(0.0003)	(0.0007)	(0.0003)
R2	0.1291	0.9911	0.0797	0.9908
Num. obs.	9038	9038	7554	7554
***p < 0.001; **p < 0.01; *p < 0.05				

Issues

1. Why do WG and FD give different results?
2. Arella-Bond: how many instruments to be used?
3. Even if some coefficients are insignificant, they are still included in the model because what we care about is the unobserved heterogeneity.

- Calculate the R squared.

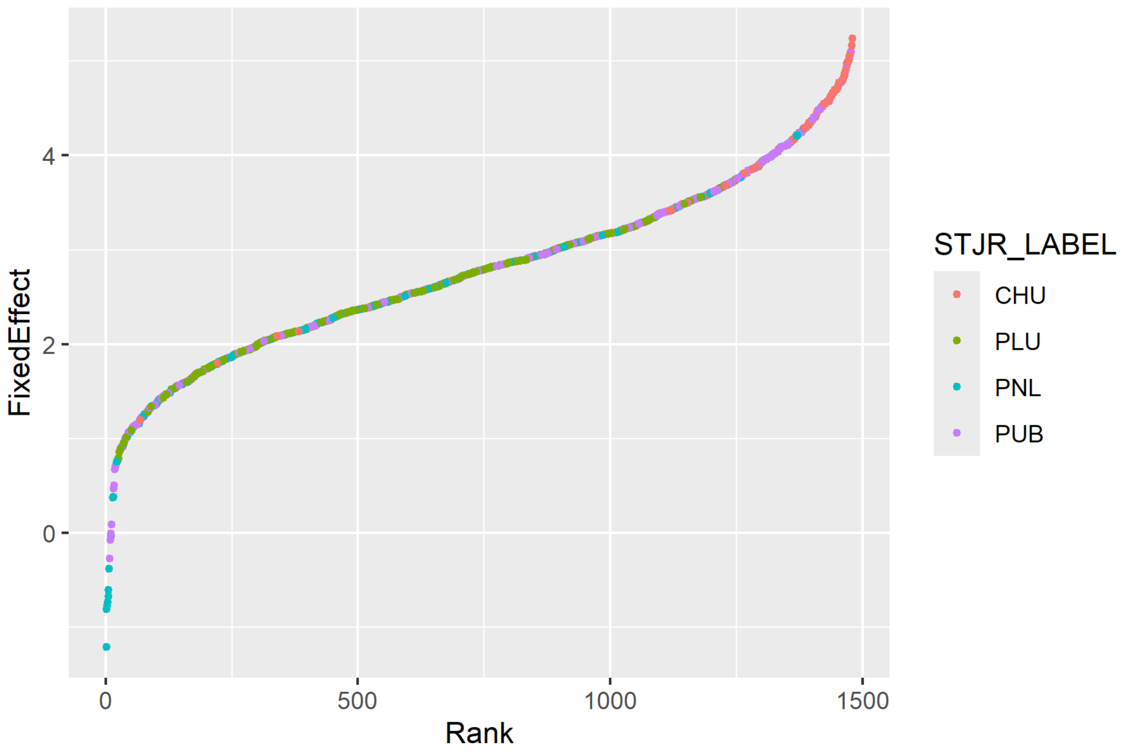


Figure 11: fixed effect

Other thoughts

Separate G

Idea 1:

Step 1: estimate G separately Following `Walter2022nber`, it is possible to estimate G while incorporating the covariates. The simplest case is when the covariate is just dummy variable, which means we estimate two different distributions for each group (standard vs charter, public vs private, etc.).

Following `KlineRoseWalter2022`, we can estimate two distributions for estimates ($\hat{\theta}$) with high (s_h) vs low precision (s_l).

Step 2: aggregate G Following `KlineRoseWalter2022`, a possible way to obtain the marginal distribution is to aggregate G_1 and G_2 by taking average. > The marginal density is compute as the average of the group-specific densities.

Maybe we can weighted by the size of each group in computing the marginal G

Step 3: Same as before Proceed as usual in defining tail probability etc.

Idea 2

If we do not aggregate, can we compute tail probability?
The tail is not obvious here since previously

$$\theta_\alpha = G(1 - \alpha)$$

We only have a capacity constraint over the total number of selection but not for each group. As shown in GuKoenker2023, different tail θ_α can lead to different ranking statistics $v_\alpha(y)$. Different ranking statistics will affect the calculation of **local False discovery rate**.

Yet if we only impose capacity constraint, the ranking statistics won't matter. ranking will be the same regardless of whether we use $v_{\alpha_1}(y)$ or $v_{\alpha_2}(y)$.

But FDR depends on both on **ranking statistics** $T(y)$ and **tail probability** $v_\alpha(y)$.

```
function (lambda, stat, v)
{
    mean((1 - v) * (stat > lambda))/mean(stat > lambda)
}
```

To deal with the issue of selecting α for each group, we can 1. Use a common α_0 to define the tail $v_{\alpha_0}(y)$. 3. Given the posterior tail probability $v_{\alpha_0}(y)$ for all i as the **ranking statistics**, perform the selection based on the ranking statistics $v_{\alpha_0}(y)$ such that capacity constraint is satisfied. 4. See how many i are selected in each group is selected and update the α_0 to individual group specific α_i. 5. For each group, find the threshold λ_{1i} that satisfy the LFdr constraint using the tail probability $v_{\alpha_i}(y)$.

Issues

Conventional and capacity dependent null rules do not give the same selection result.

1. How to aggregate the G? Just weighted average?
2. Or do we simply have 4 different regression for 4 types of hospitals, thus separate G–>aggregate G?

Presentation & Paper organization

1. Talk about the measure of labor efficiency. (literature on efficiency/producitivity).
2. endogeneity issue.
3. about the purpose: selection–>compound decision
4. the method of selection–>empirical bayes
5. the result of selection.
6. conclusion.
7. discussion on the improvement on model specification, the selection (FDR).

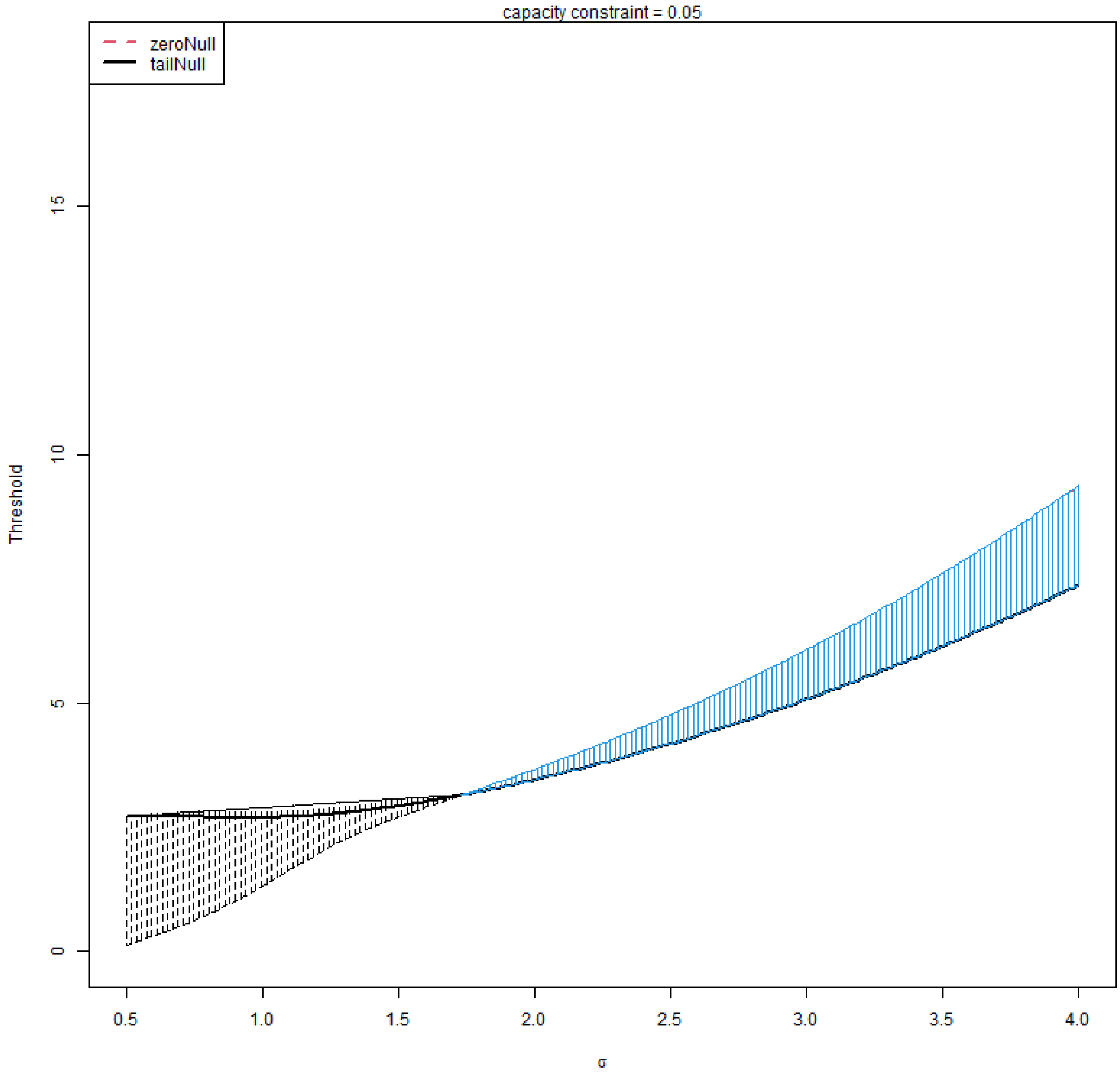


Figure 12: cap=0.05

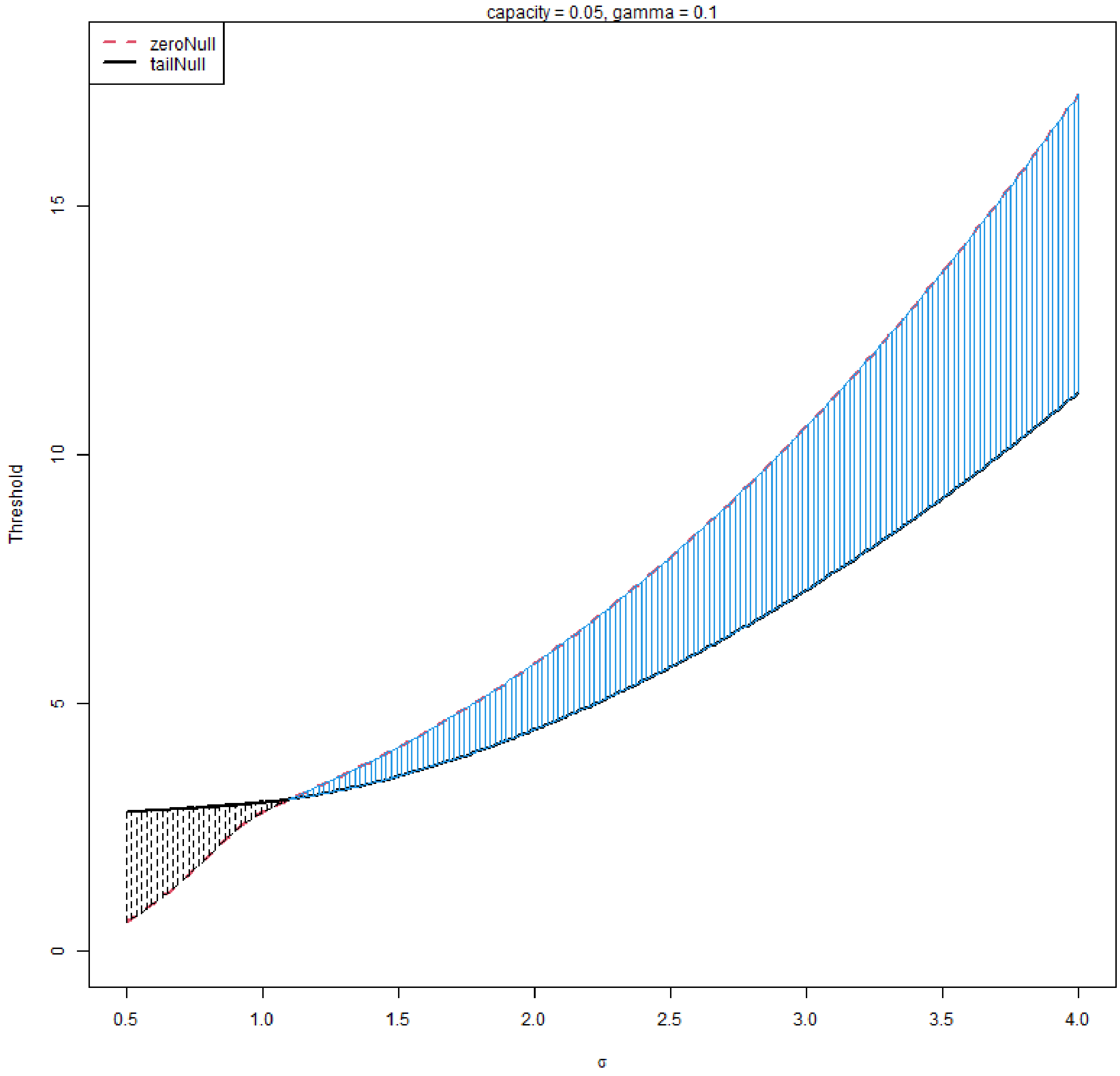


Figure 13: fdr=0.1 and cap=0.05