Paper 10, IIB: Solutions supervision 5

Eoghan O'Neill*

22 February 2020

Sections with a ∗ are for your information only. You are not expected to be able to do or know this.

(1) The example in this question is in section 10.2.3. of Wooldridge (2010) Example 10.1

1. a. To estimate δ, consider a simple DiD estimator. Let N_T and N_C be the set of treated and non-treated individuals, respectively. Let $\bar{y}_{j,t} = \frac{1}{|N_j|}\sum_{i\in N_j} y_{it}$ for $j = \{C, T\}$ and $t = 1, 2$. The DiD estimator for δ can then be obtained by

$$\hat{\delta} = \begin{cases} (\bar{y}_{T,2} - \bar{y}_{T,1}) - (\bar{y}_{C,2} - \bar{y}_{C,1}) \\ (\bar{y}_{T,2} - \bar{y}_{C,2}) - (\bar{y}_{T,1} - \bar{y}_{C,1}). \end{cases}$$

Note that both estimators are equivalent (see lecture notes). Furthermore, we have ignored the unobserved $\mathbf{x}_i$ and the presence of the unobserved heterogeneity. Therefore, this estimator will, in general, be inconsistent and biased with biased standard errors.

b. For the remainder we shall ignore the time fixed effect. It really depends on how you want to model it. If we treat it as a constant, we include a set of $T-1$ dummies (why $T-1$?) to estimate a different intercept for each time period common to all individuals. However, if we treat it as a proper time fixed effect, i.e. it is random and possibly $E[\theta_t|\mathbf{x}_i] \neq 0$ then it will be treated the same as c_i. The FE estimator in that case relies on the following within transformation

$$y_{it} - \bar{y}_t - \bar{y}_i + \bar{\bar{y}} = (\mathbf{x}_{it} - \bar{\mathbf{x}}_t - \bar{\mathbf{x}}_i + \bar{\bar{\mathbf{x}}})'\boldsymbol{\gamma} + u_{it} - \bar{u}_t - \bar{u}_i + \bar{\bar{u}},$$

where $\bar{z}_t = N^{-1}\sum_{i=1}^{N} z_{it}$, $\bar{z}_i = T^{-1}\sum_{t=1}^{T} z_{it}$ and $\bar{\bar{z}} = (NT)^{-1}\sum_{t=1}^{T}\sum_{i=1}^{N} z_{it}$ for $z_{it} = \{y_{it}, \mathbf{x}_{it}, u_{it}\}$. This transformation eliminates the individual and time fixed effect.

To answer the question, in the model

$$y_{it} = \mathbf{x}_{it}'\boldsymbol{\beta} + c_i + \varepsilon_{it}, \tag{1}$$

*Thank you to Alexis De Boeck for creating the original version of this document, and to Melvyn Weeks for providing the abridged solutions on which these are based. Please let me know of any mistakes and typos at epo21@cam.ac.uk.

show that the first difference and fixed effects estimator are numerically identical for $T = 2$. The first difference transformation yields

$$\Delta y_i = \Delta \mathbf{x}_i' \boldsymbol{\beta} + \Delta \varepsilon_i,$$

Since $T = 2$ the time dimension is eliminated. Therefore, OLS will give

$$\hat{\boldsymbol{\beta}}^{FD} = \Big(\sum_{i=1}^{N} \Delta \mathbf{x}_i \Delta \mathbf{x}_i' \Big)^{-1} \Big(\sum_{i=1}^{N} \Delta \mathbf{x}_i \Delta y_i \Big).$$

Next, consider the usual within transformation such that

$$y_{it} - \bar{y}_i = (\mathbf{x}_{it} - \bar{\mathbf{x}}_i)' \boldsymbol{\beta} + u_{it} - \bar{u}_i.$$

Rewrite this in matrix notation

$$\begin{bmatrix} \mathbf{y}_1 - \dot{\mathbf{y}}_1 \\ \vdots \\ \mathbf{y}_N - \dot{\mathbf{y}}_N \end{bmatrix} = \begin{bmatrix} (\mathbf{x}_1 - \dot{\mathbf{x}}_1)' \\ \vdots \\ (\mathbf{x}_N - \dot{\mathbf{x}}_N)' \end{bmatrix} \boldsymbol{\beta} + \mathbf{u} - \dot{\mathbf{u}},$$

where $(\mathbf{y}_i - \dot{\mathbf{y}}_i)$ are 2×1 vectors and $(\mathbf{x}_i - \dot{\mathbf{x}}_i)'$ are $2 \times K$ matrices, with $\dot{\mathbf{z}}_i = \mathbf{1}\bar{z}_i$ for $\bar{z}_i = \{\bar{y}_i, \bar{\mathbf{x}}_i'\}$ and $\mathbf{1}$ a 2×1 vector of ones. The fixed effects estimator for $\boldsymbol{\beta}$ is

$$\hat{\boldsymbol{\beta}}^{FE} = \Big(\sum_{i=1}^{N} (\mathbf{x}_i - \dot{\mathbf{x}}_i)(\mathbf{x}_i - \dot{\mathbf{x}}_i)' \Big)^{-1} \Big(\sum_{i=1}^{N} (\mathbf{x}_i - \dot{\mathbf{x}}_i)(\mathbf{y}_i - \dot{\mathbf{y}}_i) \Big).$$

Then,

$$\mathbf{y}_i - \dot{\mathbf{y}}_i = \begin{bmatrix} y_{i,1} - \frac{y_{i,1}+y_{i,2}}{2} \\ y_{i,2} - \frac{y_{i,1}+y_{i,2}}{2} \end{bmatrix} = \frac{1}{2} \begin{bmatrix} -\Delta y_i \\ \Delta y_i \end{bmatrix},$$

and

$$(\mathbf{x}_i - \dot{\mathbf{x}}_i)' = \begin{bmatrix} \mathbf{x}_{i,1}' - \frac{\mathbf{x}_{i,1}'+\mathbf{x}_{i,2}'}{2} \\ \mathbf{x}_{i,2}' - \frac{\mathbf{x}_{i,1}'+\mathbf{x}_{i,2}'}{2} \end{bmatrix} = \frac{1}{2} \begin{bmatrix} -\Delta \mathbf{x}_i' \\ \Delta \mathbf{x}_i' \end{bmatrix},$$

for $i = 1, \ldots, N$. Plugging this into $\hat{\boldsymbol{\beta}}^{FE}$,

$$\begin{aligned} \hat{\boldsymbol{\beta}}^{FE} &= \left(\frac{1}{4} \sum_{i=1}^{N} \begin{bmatrix} -\Delta \mathbf{x}_i & \Delta \mathbf{x}_i \end{bmatrix} \begin{bmatrix} -\Delta \mathbf{x}_i' \\ \Delta \mathbf{x}_i' \end{bmatrix} \right)^{-1} \left(\frac{1}{4} \sum_{i=1}^{N} \begin{bmatrix} -\Delta \mathbf{x}_i & \Delta \mathbf{x}_i \end{bmatrix} \begin{bmatrix} -\Delta y_i \\ \Delta y_i \end{bmatrix} \right) \\ &= \Big(\sum_{i=1}^{N} \Delta \mathbf{x}_i \Delta \mathbf{x}_i' \Big)^{-1} \Big(\sum_{i=1}^{N} \Delta \mathbf{x}_i \Delta y_i \Big). \end{aligned}$$

Hence, $\hat{\boldsymbol{\beta}}^{FD} = \hat{\boldsymbol{\beta}}^{FE}$ when $T = 2$.

2. a. We need the following assumptions

Assumption (A1) (1) is the true model; **(A2)** $E[\tilde{\mathbf{X}}_i'\tilde{\mathbf{X}}_i]$ is positive definite (or at least non-singular);[1] **(A3)** $E[u_{it}|\mathbf{x}_i, c_i] = 0$ for all t, i (strict exogeneity); **(A4)** We have a random sample from the cross section. $\mathbf{X}_i$ and $\mathbf{u}_i$ are iid.

Proposition 1 (∗)**.** *Under A1-4, the FE estimator is consistent, i.e.* $\hat{\boldsymbol{\beta}} \xrightarrow{p} \boldsymbol{\beta}$.

The FE estimator is $\hat{\boldsymbol{\beta}} = (\tilde{\mathbf{X}}'\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}'\tilde{\mathbf{y}}$, where $\tilde{}$ denotes that the data has been time demeaned. Rewrite this using summation notation

$$\hat{\boldsymbol{\beta}} = \Big(\sum_{i=1}^{N} \tilde{\mathbf{X}}_i'\tilde{\mathbf{X}}_i\Big)^{-1}\Big(\sum_{i=1}^{N} \tilde{\mathbf{X}}_i'\tilde{\mathbf{y}}_i\Big),$$

under A1 this quantity is

$$\hat{\boldsymbol{\beta}} = \boldsymbol{\beta} + \Big(\sum_{i=1}^{N} \tilde{\mathbf{X}}_i'\tilde{\mathbf{X}}_i\Big)^{-1}\Big(\sum_{i=1}^{N} \tilde{\mathbf{X}}_i'\tilde{\mathbf{u}}_i\Big),$$

or

$$\hat{\boldsymbol{\beta}} = \boldsymbol{\beta} + \Big(\frac{1}{N}\sum_{i=1}^{N} \tilde{\mathbf{X}}_i'\tilde{\mathbf{X}}_i\Big)^{-1}\Big(\frac{1}{N}\sum_{i=1}^{N} \tilde{\mathbf{X}}_i'\tilde{\mathbf{u}}_i\Big), \tag{2}$$

By the Weak Law of Large Numbers under A2 and A4

$$\frac{1}{N}\sum_{i=1}^{N} \tilde{\mathbf{X}}_i'\tilde{\mathbf{X}}_i \xrightarrow{p} E[\tilde{\mathbf{X}}_i'\tilde{\mathbf{X}}_i] \Rightarrow \Big(\frac{1}{N}\sum_{i=1}^{N} \tilde{\mathbf{X}}_i'\tilde{\mathbf{X}}_i\Big)^{-1} \xrightarrow{p} (E[\tilde{\mathbf{X}}_i'\tilde{\mathbf{X}}_i])^{-1}, \tag{3}$$

where the implication follows by appealing to the continuous mapping theorem under A2 and the continuity of matrix inversion. Similarly by the WLLN and the law of iterated expectations under A3 and A4 we have

$$\Big(\frac{1}{N}\sum_{i=1}^{N} \tilde{\mathbf{X}}_i'\tilde{\mathbf{u}}_i\Big) \xrightarrow{p} E[\tilde{\mathbf{X}}_i'\tilde{\mathbf{u}}_i] = \mathbf{0}. \tag{4}$$

Finally, combining (2), (3) and (4) and applying the continuous mapping theorem yields

$$\hat{\boldsymbol{\beta}} \xrightarrow{p} \boldsymbol{\beta}.$$

b. The strict exogeneity assumption is given in A3. It states that the idiosyncratic error is mean independent of the covariates and the fixed effect contemporaneously and at all leads and lags. The strict exogeneity assumption is really about mean independence, which in turn *implies* contemporaneous uncorrelatedness between the regressors and the residual. This follows directly from the law of iterated expectations.

[1] This implies the rank condition: $rk(\tilde{\mathbf{X}}_i'\tilde{\mathbf{X}}_i) = K$.

Another way to state this assumption is

$$E[y_{it}|\mathbf{x}_i, c_i] = c_i + \mathbf{x}_{it}'\boldsymbol{\beta}.$$

The interpretation of this is that once $\mathbf{x}_{is}$ has been controlled for it has no partial effect on y_{it}, for $s \neq t$.

To answer the remainder of the question you can make a nice story of why you think strict exogeneity may or may not hold. This is really about you convincing the reader (i.e. the examiner) of your opinion.

c. Stationarity is of no concern if we are interested in the consistency of the estimator. Recall that we proved consistency under a large N, small T asymptotic framework. Of course, the time series properties will matter for obtaining correct standard errors. How to deal with this when T is small is a whole different matter.

3. The object, $\omega = \frac{\sigma_c^2}{\sigma_c^2+\sigma_u^2}$, is the intraclass correlation coefficient. It is derived from $\varepsilon_{it} = c_i + u_{it}$ such that for $s \neq t$

$$\omega = \frac{\text{cov}(\varepsilon_{is}, \varepsilon_{it})}{\sqrt{\text{var}(\varepsilon_{is})\,\text{var}(\varepsilon_{it})}} = \frac{\sigma_c^2}{\sigma_c^2 + \sigma_u^2}.$$

This can be obtained under either a random or fixed effects specification.

It is a measure for the between variance as a proportion of the total variance of the composite error term. In the population, if $\sigma_c^2 = 0$, the unobserved heterogeneity is a point mass, there is no between variance and $\omega = 0$. Would use POLS. If $\omega \to 1$, POLS or RE are inconsistent under a FE specification. Therefore, use the FE estimator. For $\omega \in (0, 1)$, POLS is consistent, but inefficient. Use the RE estimator.

However, we do not know the true ω, but only a value estimated from the data. Therefore, $\hat{\omega}$ is a random variable, perhaps with a mean and variance. Its value here is 'low', but not zero. Hence, may want to use RE over POLS. Yet, if the sample variance of $\hat{\omega}$ is particularly high, we may prefer POLS or even FE in an extreme case. Also RE, requires one more parameter to be estimated (which one?), thus we may prefer POLS on the grounds of that in very small samples.

Of course, the caveat is that there is no requirement for $\sigma_c^2 \to \infty$ for the true model to be the FE model. Therefore, even with small ω, the FE model could be correct. This statistic can obviously not discriminate between the competing models in that case. In then end, you will have to make a choice on which estimator to use and justify it accordingly to the best of your abilities.

(2) a. We make use of the variance operator to show this

$$\begin{aligned}\boldsymbol{\Sigma} := \operatorname{var}(\boldsymbol{w}_i) &= \operatorname{var}(\alpha_i \mathbf{1}_T + \boldsymbol{\varepsilon}_i) \\ &= \operatorname{var}(\alpha_i \mathbf{1}_T) + \operatorname{var}(\boldsymbol{\varepsilon}_i) \quad \text{(indepence of } \alpha_i, \boldsymbol{\varepsilon}_i) \\ &= \mathbf{1}_T \operatorname{var}(\alpha_i) \mathbf{1}_T' + \operatorname{var}(\boldsymbol{\varepsilon}_i) \\ &= \sigma_\alpha^2 \mathbf{1}_T \mathbf{1}_T' + \sigma_\varepsilon^2 \boldsymbol{I}_T \\ &= \begin{bmatrix} \sigma_\alpha^2 + \sigma_\varepsilon^2 & \sigma_\alpha^2 & \dots & \sigma_\alpha^2 \\ \sigma_\alpha^2 & \sigma_\alpha^2 + \sigma_\varepsilon^2 & \dots & \sigma_\alpha^2 \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_\alpha^2 & \sigma_\alpha^2 & \dots & \sigma_\alpha^2 + \sigma_\varepsilon^2 \end{bmatrix}.\end{aligned}$$

b. In (2), we have $E[\boldsymbol{w}_i|\mathbf{x}_i] = \mathbf{0}$, but $\boldsymbol{\Sigma} \neq \sigma^2 \boldsymbol{I}_T$. This is a violation of the Gauss-Markov assumptions. The equicorrelation does not vanish as $|s-t| \to \infty$. The RE estimator takes care of this by incorporating the variance-covariance structure of the errors. Hence, it is a particular case of a GLS estimator. Note that in the RE model we are dealing with serial correlation. The errors are homoskedastic across individuals. The estimator is not feasible without knowledge of the parameters σ_α^2 and σ_ε^2.

$*$ Rewrite the variance matrix of $\boldsymbol{w}_i$ as

$$\boldsymbol{\Sigma} = \sigma_\varepsilon^2 \left(\boldsymbol{I}_T + \frac{\sigma_\alpha^2}{\sigma_\varepsilon^2} \mathbf{1}_T \mathbf{1}_T' \right).$$

Consider the following lemma

Lemma 1 (Abadir and Magnus, 2005, p. 87)**.** *For a non-singular matrix* $\boldsymbol{A}$ *and two conformable vectors* $\boldsymbol{a}$ *and* $\boldsymbol{b}$*, if* $\boldsymbol{b}'\boldsymbol{A}^{-1}\boldsymbol{a} \neq -1$ *then*

$$(\boldsymbol{A} + \boldsymbol{a}\boldsymbol{b}')^{-1} = \boldsymbol{A}^{-1} - \frac{1}{1 + \boldsymbol{b}'\boldsymbol{A}^{-1}\boldsymbol{a}} \boldsymbol{A}^{-1}\boldsymbol{a}\boldsymbol{b}'\boldsymbol{A}^{-1}.$$

Using lemma 2 you can show that

$$\boldsymbol{\Sigma}^{-1} = \frac{1}{\sigma_\varepsilon^2} \left(\boldsymbol{D} + \frac{\psi}{T} \mathbf{1}_T \mathbf{1}_T' \right),$$

where $\boldsymbol{D}$ is the demeaning matrix and $\psi = \frac{\sigma_\varepsilon^2}{\sigma_\varepsilon^2 + T\sigma_\alpha^2}$, see appendix C for more details. By verification,

$$\boldsymbol{\Sigma}^{-1/2} = \frac{1}{\sigma_\varepsilon} \left(\boldsymbol{I}_T - \frac{\theta}{T} \mathbf{1}_T \mathbf{1}_T' \right),$$

with $\theta = 1 - \psi^{1/2} = 1 - \frac{\sigma_\varepsilon}{\sqrt{\sigma_\varepsilon^2 + T\sigma_\alpha^2}}$.

c. $*$ The RE estimator is a shrinkage estimator in the sense that depending on ψ it will treat all individuals equally (no heterogeneity) or will treat everyone as different

(heterogeneity). Since, we are estimating the mean we are really considering a linear combination between the individual and global means. From the answer above you can see that the shrinkage sits very much in the variance of $\boldsymbol{w}_i$ as this determines how the data will be transformed to obtain the RE estimator.

Of course, we do not need to constrain ourselves to means. In a Bayesian perspective, this would imply that we consider setting the prior as the product of N identical priors, the product of N different priors or a mixture of the two.

For those of you who want to know more about shrinkage estimators a good place to start is to look up the James-Stein estimator.

d. This question relates to what was said above. Consider the transformed model

$$\boldsymbol{\Sigma}^{-1/2}\mathbf{y}_i = \boldsymbol{\Sigma}^{-1/2}\mathbf{X}_i\boldsymbol{\beta} + \boldsymbol{\Sigma}^{-1/2}\boldsymbol{w}_i.$$

The different cases are

Case 1: $\theta \to 1$. This happens when $\psi \to 0$. Hence, $\sigma_\alpha^2 \to \infty$ (or $T \to \infty$). Unobserved heterogeneity becoming more important. The RE estimator becomes the FE estimator as $\boldsymbol{\Sigma}^{-1/2} \to \boldsymbol{D}$, which performs the within transformation.

Case 2: $\theta \in (0, 1)$. We perform a quasi-demeaning of the data. This is the RE estimator.

Case 3: $\theta = 0$. For fixed T, $\sigma_\alpha^2 = 0$. There is no individual heterogeneity. The RE estimator becomes POLS as $\boldsymbol{\Sigma}^{-1/2} \to \boldsymbol{I}_T$.

(3) By construction, the disturbance and y_{it} are correlated. This is a classical endogenous regressor problem: consider $(y_{it}, x_{it+1}, x_{it})'$, then for all t

$$\begin{aligned} E[v_{it}|y_{it}, x_{it+1}, x_{it}, \eta_i] &= E[\varepsilon_{it}|y_{it}] \quad \text{(strict exogeneity of } x_{it}, \eta_i) \\ &= E[v_{it}|\alpha y_{it-1} + \beta_0 x_{it} + \beta_1 x_{it-1} + \eta_i + v_{it}] \\ &\neq 0. \end{aligned}$$

The FE estimator is inconsistent for fixed T. Nickell (1981) showed that as $T \to \infty$

$$E[\hat{\alpha}] = \alpha + O(T^{-1}).$$

Thus as $T \to \infty$ we may achieve asymptotic unbiasedness of $\hat{\alpha}$. This is not the same as consistency (why?).

Definition 1 ($*$)**.** Define the notation: O (big Oh) and o (small Oh) as follows for a non-stochastic sequence $\{X_n\}_{n \in \mathcal{N}_0}$

$X_n = O(n^\lambda)$, if for some real $M > 0$, there is an N such that for all $n \geq N$, $|n^{-\lambda} X_n| < M$,

and

$$X_n = o(n^\lambda), \text{ if } n^{-\lambda}X_n \to 0 \text{ as } n \to \infty.$$

Remark. The definition for O may not seem immediately obvious, but it comes down to saying that the sequence X_n eventually is bounded as $n \to \infty$ if we divide through by n^λ. Extensions to O_p (big Oh in probability) and o_p (small Oh in probability) readily follow from this definition.

In our setting T is fixed, but we can use internal instruments to deal with the endogeneity. Eliminate the fixed effect by first differencing. However, there is serial correlation in v_{it} of unspecified form. This, in general, rules out using lagged dependent variables as instruments. For example, consider using y_{it-2}

$$\begin{aligned} E[y_{it-2}\Delta v_{it}] &= E[(\alpha y_{it-3} + \beta_0 x_{it-2} + \beta_1 x_{it-3} + \eta_i + v_{it-2})\Delta v_{it}] \\ &= \alpha E[y_{it-3}\Delta v_{it}] + E[v_{it-2}\Delta v_{it}] \\ &\neq 0, \end{aligned}$$

unless $\alpha E[y_{it-3}\Delta v_{it}] = -E[v_{it-2}\Delta v_{it}]$.

Yet, the added regressor, x, is included in the model. Thus, if this is the true model it explains variation in y. Combined with the strict exogeneity assumption, we can specify a set of moment conditions for $t \geq 3$ and $s \geq 1$

$$E[x_{is}\Delta v_{it}] = 0.$$

The three parameters of interest can then be estimated by GMM. How many moment conditions do you need? And how many could you potentially have as a function of T?

See appendix A for a short note on general GMM estimation with examples and some questions.

(4) a. This is a standard probit model

$$\Pr\{Y = 1|\boldsymbol{Z} = \mathbf{z}\} = E[Y|\boldsymbol{Z} = \mathbf{z}] = \Phi(\mathbf{z}'\boldsymbol{\delta}).$$

The partial effect with respect to z_2 is

$$\frac{\partial E[Y|\boldsymbol{Z} = \mathbf{z}]}{\partial z_2} = (\gamma_1 + 2\gamma_2 z_2)\phi(\mathbf{z}_1'\boldsymbol{\delta}_1 + \gamma_1 z_2 + \gamma_2 z_2^2),$$

where ϕ is the standard normal density. In this nonlinear model, the partial (or marginal) effect is heterogeneous across individuals depending on the covariates. Given the specified parametric model, we can estimate this partial derivative by plugging in the ML estimates. Note that you have now estimated the partial effect. Suppose that there is only one covariate, z, then we could plot this function in (z, y)-space.

Plotting quickly becomes infeasible as the dimensions of the covariate increase. Therefore, we can consider measures that summarise the partial effect. Frequently used ones are the average partial effect and partial effect at the average. Just as the estimated partial effect is a random variable (really a random process), so are the latter two. Of course, we do not need to constrain ourselves to averages, we could look at whichever quantile of the covariates we are interested in.

$*$ I have included a small simulation exercise to show how these quantities can be estimated, see appendix B.

b. Analogously, the partial effect with respect to z_2 is

$$\frac{\partial E[Y|\boldsymbol{Z}=\mathbf{z}, D=d]}{\partial z_2} = (\gamma_1 + \gamma_3 d)\phi(\mathbf{z}_1'\boldsymbol{\delta}_1 + \gamma_1 z_2 + \gamma_2 d + \gamma_3 z_2 d).$$

Since D is a binary random variable we obtain the partial effect with respect to D as follows,

$$\Pr\{Y=1|\boldsymbol{Z}=\mathbf{z}, D=1\} - \Pr\{Y=1|\boldsymbol{Z}=\mathbf{z}, D=0\} = \Phi(\mathbf{z}_1'\boldsymbol{\delta}_1 + \gamma_2 + (\gamma_1+\gamma_3)z_2) - \Phi(\mathbf{z}_1'\boldsymbol{\delta}_1 + \gamma_1 z_2).$$

These two quantities can again be estimated by plugging in the ML estimates. We can also look at the partial effect at the average, the average partial effect or any other interesting quantity of $\mathbf{z}$.

A Generalised Method of Moments

Basic idea is straightforward. Suppose that we can write down our model through a collection of moment conditions

$$E[g(\mathbf{X}, \boldsymbol{\theta}_0)] = \mathbf{0}, \tag{5}$$

where $g(\cdot)$ is a K-dimensional vector valued function and $\boldsymbol{\theta}_0$ is the true, but unknown K-dimensional parameter vector which sets the moment conditions exactly to zero in the population. If there are multiple $\boldsymbol{\theta}_0 \in \Theta$ which satisfy (5) then the model in not identified and if for all $\boldsymbol{\theta} \in \Theta$ we have $E[g(\mathbf{X}, \boldsymbol{\theta})] \neq \mathbf{0}$ then the model is misspecified. I won't discuss these two cases here.

This is a just-identified case, thus we can solve (5) exactly. However, since we don't specify the distribution of $\mathbf{X}$ (otherwise we would use maximum likelihood instead) we don't know these expectations. We can replace them (i.e. estimate them) by their sample analogue. Then we estimate $\boldsymbol{\theta}_0$ by solving

$$\frac{1}{n}\sum_{i=1}^{n} g(\mathbf{X}_i, \hat{\boldsymbol{\theta}}) = \mathbf{0}. \tag{6}$$

This is known as the Method of Moments (MoM) estimator.

Example A.1. Let $(X_i)_{i=1}^n$ be an iid sample from a uniform distribution on $[0, \theta]$. The expectation of this uniform distribution is $E[X] = \theta/2$. A valid moment condition is

$$E[X - \theta/2] = 0,$$

with sample analogue

$$\frac{1}{n}\sum_{i=1}^{n} X_i - \hat{\theta}/2 = 0.$$

From this, it follows that the MoM estimator is

$$\hat{\theta} = 2\frac{1}{n}\sum_{i=1}^{n} X_i.$$

Note that this is not the same as the maximum likelihood estimator. The ML estimator of θ in this example is $\hat{\theta}^{ML} = \max_{1 \leq i \leq n} X_i$.

Example A.2 (OLS through MoM)**.** Consider the linear model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

with $E[\boldsymbol{\varepsilon}] = \mathbf{0}$ and $\mathbf{X}$ fixed. A valid moment condition in this model is

$$E[\mathbf{x}_i \varepsilon_i] = \mathbf{0},$$

or rewritten

$$E[\mathbf{x}_i(y_i - \mathbf{x}_i'\boldsymbol{\beta})] = \mathbf{0}.$$

The sample analogue is

$$\frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_i(y_i - \mathbf{x}_i'\hat{\boldsymbol{\beta}}) = \mathbf{0}.$$

Observe that $\mathbf{x}_i\varepsilon_i$ is a K-dimensional vector and there are K unknown parameters such that we have the same number of moment conditions as unknown parameters. Proceeding by MoM, after some rewriting and assuming that $\frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_i\mathbf{x}_i'$ is invertible we obtain the familiar expression

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}.$$

Generalised Method of Moments (GMM) builds on this by allowing for over-identified models. The set-up from (5) and (6) carries over, but $g(\cdot)$ is now an M-dimensional vector valued function with $M > K$. Thus, we have more moment conditions than parameters. We want to solve the sample analogue to this problem, but as you all know there is no (unique) solution to such an over-identified system of equations. What GMM tries to do instead is to set (6) as close as possible to zero. Therefore, it solves the following minimisation problem

$$\hat{\boldsymbol{\theta}}_{(1)}^{GMM} = \arg\min_{\boldsymbol{\theta}\in\Theta} \left\| \frac{1}{n}\sum_{i=1}^{n} g(\mathbf{X}_i, \boldsymbol{\theta}) \right\|^2 = \arg\min_{\boldsymbol{\theta}\in\Theta} \Big(\frac{1}{n}\sum_{i=1}^{n} g(\mathbf{X}_i, \boldsymbol{\theta})\Big)' \Big(\frac{1}{n}\sum_{i=1}^{n} g(\mathbf{X}_i, \boldsymbol{\theta})\Big), \tag{7}$$

with $\|\cdot\|$ the Euclidean norm.

Often, GMM is stated more generally as minimising the weighted norm of the sample moment conditions

$$\hat{\boldsymbol{\theta}}_{(2)}^{GMM} = \arg\min_{\boldsymbol{\theta}\in\Theta} \left\| \frac{1}{n}\sum_{i=1}^{n} g(\mathbf{X}_i, \boldsymbol{\theta}) \right\|_{\boldsymbol{W}}^2 = \arg\min_{\boldsymbol{\theta}\in\Theta} \Big(\frac{1}{n}\sum_{i=1}^{n} g(\mathbf{X}_i, \boldsymbol{\theta})\Big)' \boldsymbol{W} \Big(\frac{1}{n}\sum_{i=1}^{n} g(\mathbf{X}_i, \boldsymbol{\theta})\Big), \tag{8}$$

where $\boldsymbol{W}$ is a positive definite matrix. The choice of $\boldsymbol{W}$ is an important issue and can be shown to be related to the asymptotic efficiency of $\hat{\boldsymbol{\theta}}_{(2)}^{GMM}$. Trivially, if we choose $\boldsymbol{W} = \boldsymbol{I}$ then $\hat{\boldsymbol{\theta}}_{(1)}^{GMM} \equiv \hat{\boldsymbol{\theta}}_{(2)}^{GMM}$. The choice of $\boldsymbol{W}$ is irrelevant for consistency. Let us now see how GMM works through some examples.

Example A.3 (OLS through GMM)**.** Consider again the standard linear regression model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

with $E[\boldsymbol{\varepsilon}] = \mathbf{0}$ and $\mathbf{X}$ fixed. Instead of using $E[\mathbf{x}_i\varepsilon_i] = \mathbf{0}$, let us use $E[\boldsymbol{\varepsilon}] = \mathbf{0}$ or $E[\mathbf{y} - \mathbf{X}\boldsymbol{\beta}] = \mathbf{0}$. This specifies n moment conditions, whereas we only have K parameters with $n > K$. Using (7) we can estimate $\boldsymbol{\beta}$ by solving

$$\hat{\boldsymbol{\beta}}^{GMM} = \arg\min_{\boldsymbol{\beta}\in\mathcal{R}^K} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}).$$

Either you notice that this is exactly the same as minimising the sum of squared residuals or after some tedious matrix calculus we can obtain

$$\hat{\boldsymbol{\beta}}^{GMM} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y},$$

from which we see that $\hat{\boldsymbol{\beta}}^{GMM} = \hat{\boldsymbol{\beta}}^{MoM} = \hat{\boldsymbol{\beta}}^{OLS}$. Hence, the OLS estimator can also be motivated through a MoM or GMM framework.

Example A.4 (Instrumental Variables)**.** Consider again the standard linear regression model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

with $E[\boldsymbol{\varepsilon}|\mathbf{X}] \neq \mathbf{0}$, but with M instruments available, captured in $\boldsymbol{Z}$, such that $E[\boldsymbol{Z}'\boldsymbol{\varepsilon}] = \mathbf{0}$ and $E[\boldsymbol{Z}'\mathbf{X}] \neq \mathbf{0}$. Similarly to the examples above, there are M valid moment conditions

$$E[\boldsymbol{Z}'\boldsymbol{\varepsilon}] = \mathbf{0}.$$

Again, setting up the estimator to solve (7)

$$\hat{\boldsymbol{\beta}}^{GMM} = \arg\min_{\boldsymbol{\beta}\in\mathcal{R}^K} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'\boldsymbol{Z}\boldsymbol{Z}'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}).$$

The first-order conditions are

$$-2\mathbf{X}'\boldsymbol{Z}\boldsymbol{Z}'(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}) = \mathbf{0}.$$

After rearranging and making suitable invertibility assumptions we get

$$\hat{\boldsymbol{\beta}}^{GMM} = (\mathbf{X}'\boldsymbol{Z}\boldsymbol{Z}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{Z}\boldsymbol{Z}'\mathbf{y}.$$

Please do note that this is not the efficient IV estimator.

Question 1. A second-year friend of yours received an iid sample on a single variable $(Z_i)_{i=1}^n$ and was asked for a supervision to estimate the mean of Z. Friends of your friend thought this was too easy (maybe rightly so?) and copied the data into two new samples and added and subtracted some noise such that

$$X_i = Z_i + \varepsilon_i$$
$$Y_i = Z_i - \nu_i,$$

with $E[\varepsilon_i] = E[\nu_i] = 0$, $E[\varepsilon_i] = E[\nu_i] = \lambda^2$ and $E[\varepsilon_i \nu_i] = \lambda$. The parameter λ is known.

Your friend, being a student, left the supervision work until the last moment and unfortunately just now the Faculty's servers are down such that the original data cannot be retrieved before the deadline. Using these two series obtain a GMM estimator for $E[Z] = \theta$ when:

(a) $\boldsymbol{W} = \boldsymbol{I}_2$;

(b) $\boldsymbol{W} = E[g(X_i, Y_i; \theta) g(X_i, Y_i; \theta)]^{-1}$;

(c) your friend's friends cannot remember λ.

Question 2. Consider the following model

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 w_i + \varepsilon_i,$$

where X_i is endogenous and $(W_i, Z_i)'$ are valid included and excluded instruments respectively. Data simulated form this model can be found in the file paper10_gmm_simulation_data.csv.

(a) Write down a set of moment conditions valid in this model;

(b) Using the data estimate $(\beta_0, \beta_1, \beta_2)'$ through GMM using STATA, or any other software package of your choosing. What do you think is the true vector $(\beta_0, \beta_1, \beta_2)'$ used to generate the data;[2]

(c) What happens when you add more and more moment conditions?

[2]This may be helpful: http://www.stata.com/manuals13/rgmm.pdf.

B Probit simulation

This appendix is only meant to help improve your understanding of the material; it is not part of the syllabus.

Let us simulate a random sample $(Y_i, Z_i)_{i=1}^N$ from the following probit model

$$Y_i = \mathbf{1}(\alpha + \beta Z_i + \varepsilon_i > 0) \quad \varepsilon_i | Z_i \sim \mathcal{N}(0,1),\ i = 1, \ldots, N,$$

where we choose $N = 100$, $\alpha = 0.25$ and $\beta = -1.25$. Furthermore, we choose $Z_i \sim \mathcal{U}(-2, 2)$ for all $i = 1, \ldots, N$. The density of Z is $f_Z(z) = 1/4$ for all $z \in [-2, 2]$. Below is the script which implements this in R, but since I provide the data it should not be too hard to replicate the results in a package you are more familiar with, e.g. STATA.

Table 1: Probit estimation results.

	y
z	-1.214^{***}
	(0.217)
Constant	0.478^{***}
	(0.168)
Observations	100
Log Likelihood	-37.163

Note: standard errors in parenthesis. $^{*}p < 0.1;^{**}p < 0.05;^{***}p < 0.01$.

Table 1 presents the estimation results using the simulated data. Figure 1 plots the simulated sample as well as the estimated probability of $Y = 1$ conditional on $Z = z$, i.e. this is an estimate as a function of z. We can then use the ML estimates to estimate the partial effect in this simple univariate probit model. Figure 2 plots the estimated partial effect, $\hat{\beta}\phi(\hat{\alpha} + \hat{\beta}z)$, and the true partial effect, $\beta\phi(\alpha + \beta z)$, as a function of z. This shows the heterogeneity of the partial effect quite nicely.[3]

If we are interested in the average partial effect we could estimate this by

$$\widehat{APE} = \frac{1}{N}\sum_{i=1}^{N} \hat{\beta}\phi(\hat{\alpha} + \hat{\beta}z_i) \approx -0.241.$$

Since we know the true model, we can compute the population average partial effect exactly

$$APE = E_Z\left[\frac{\partial E[Y = 1|Z = z]}{\partial z}\right] = \int_{-2}^{2} -1.25\phi(.25 - 1.25z) \times 0.25 \, \mathrm{d}z = -0.246.$$

[3]Recall that in a linear probability model the partial effect will be a constant at the value of the slope parameter for all z.

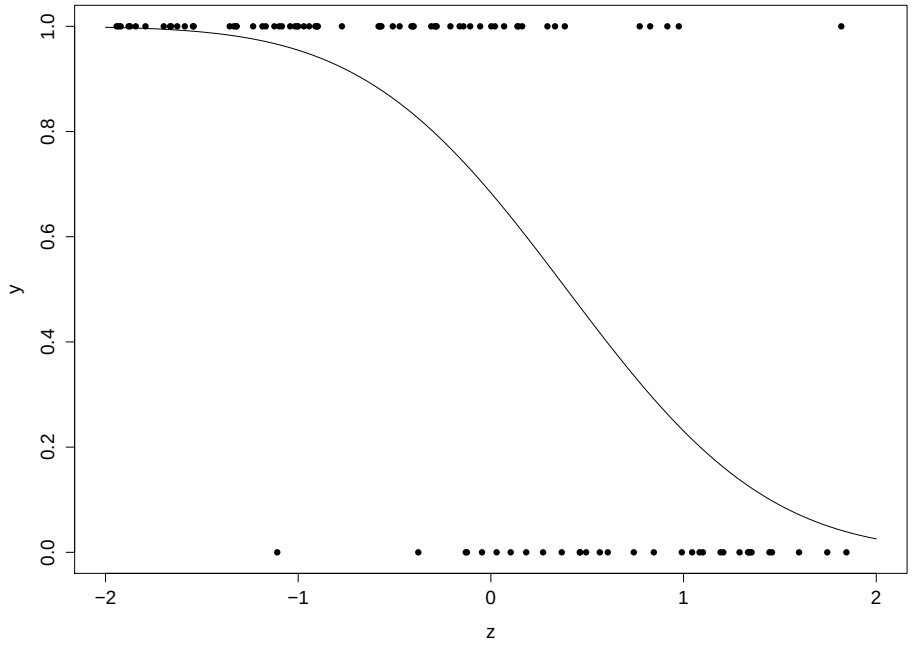

Figure 1: Plot of the estimated model and the sample points.

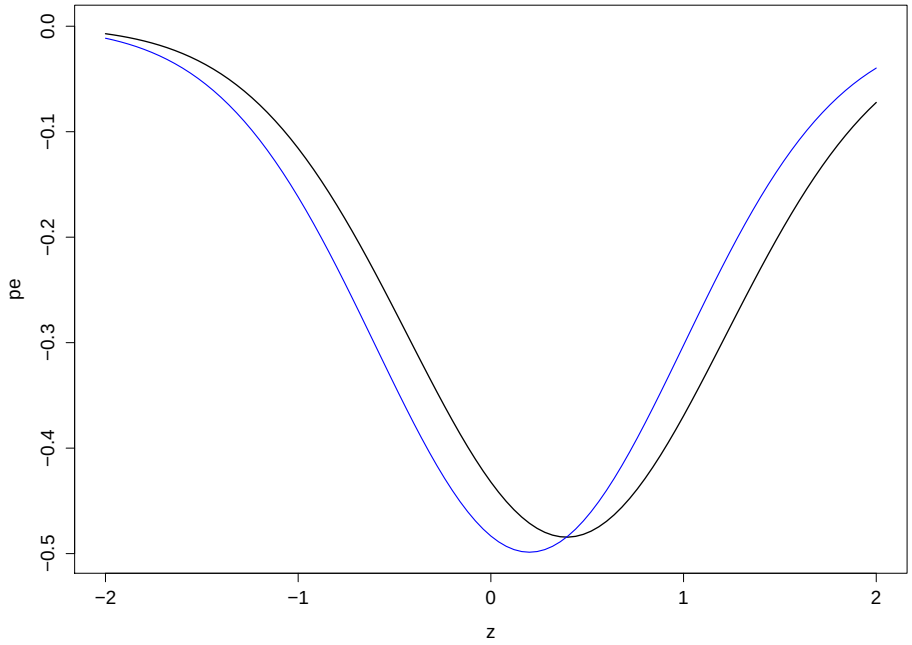

Figure 2: Plot of the true partial effect (blue) and the estimated partial effect (black).

```r
##' Set seed to be able to reproduce the results later.
set.seed(10)

##' Define simulation parameters.
N <- 100
alpha <- .25
beta <- -1.25

##' Simulate variables: covariate, error and response.
z <- runif(N, -2, 2)
e <- rnorm(N)
y <- ifelse(alpha + beta*z + e > 0, 1, 0)

##' Estimate the simulated model using probit.
reg <- glm(y ~ z, family = binomial(link = "probit"))
summary(reg)

##' Creates LaTeX table of estimation results
# library(stargazer)
# stargazer(reg, omit.stat = c("aic"), no.space = TRUE, title="Probit estimation results.", align = TRUE)

##' Evaluate the estimated function on an equally spaced grid
##' between -2 and 2, and plot the estimated function together
##' with the sample points.
grid <- seq(-2, 2, length.out = 101)
y.pred <- predict(reg, list(z = grid), type = "response")

dat <- data.frame(z = sort(grid), y = y.pred[order(grid)])
plot(z, y, pch = 16, xlim = c(-2, 2))
lines(dat)

##' Estimate the partial effect and plot this together with the true partial effect
##' implied by the model.
pe <- reg$coefficients["z"] * dnorm(reg$coefficients[1] + grid * reg$coefficients["z"])
pe.dat <- data.frame(z = sort(grid), pe = pe[order(grid)])
plot(pe.dat, type = "l", lwd = 1.5, ylim = c(optimize(pe.true, c(-2, 2))$objective, 0), xlim = c(-2, 2))
curve(pe.true, -2, 2, add = TRUE, col = "blue", lwd = 1.25)

##' Compute the true average partial effect and the estimated average partial effect.
ape.true <- integrate(mean.pe.true, -2, 2)
ape <- mean(pe)

##' Below are helper functions.
pe.true <- function(z){
  ##' Function for the true partial effect
  ##' evaluated at a point z.
  beta * dnorm(alpha + beta * z)
}

mean.pe.true <- function(z){
  ##' Function that if we integrate over yields that average partial effect.
  ##' Factor '0.25' is the density of Z, f_Z(z).
 0.25 * pe.true(z)
}
```

C Random Effects: Inverse of Covariance Matrix

Lemma 2 (Abadir and Magnus, 2005, p. 87)**.** *For a non-singular matrix* $\boldsymbol{A}$ *and two conformable vectors* $\boldsymbol{a}$ *and* $\boldsymbol{b}$*, if* $\boldsymbol{b}'\boldsymbol{A}^{-1}\boldsymbol{a} \neq -1$ *then*

$$(\boldsymbol{A} + \boldsymbol{a}\boldsymbol{b}')^{-1} = \boldsymbol{A}^{-1} - \frac{1}{1 + \boldsymbol{b}'\boldsymbol{A}^{-1}\boldsymbol{a}}\boldsymbol{A}^{-1}\boldsymbol{a}\boldsymbol{b}'\boldsymbol{A}^{-1}.$$

Apply this lemma, letting $\boldsymbol{A} = \boldsymbol{I}$, $\boldsymbol{a} = \frac{\sigma_\alpha^2}{\sigma_\varepsilon^2}\boldsymbol{i}_T$, $\boldsymbol{b} = \boldsymbol{i}_T$

$$(\boldsymbol{I} + \frac{\sigma_\alpha^2}{\sigma_\varepsilon^2}\boldsymbol{i}_T\boldsymbol{i}_T')^{-1} = \boldsymbol{I} - \frac{1}{1 + \boldsymbol{i}_T'\boldsymbol{I}\frac{\sigma_\alpha^2}{\sigma_\varepsilon^2}\boldsymbol{I}\boldsymbol{i}_T\boldsymbol{i}_T'}\frac{\sigma_\alpha^2}{\sigma_\varepsilon^2}\boldsymbol{I}\boldsymbol{i}_T\boldsymbol{i}_T'\boldsymbol{I} = \boldsymbol{I} - \frac{\frac{\sigma_\alpha^2}{\sigma_\varepsilon^2}}{1 + (\frac{\sigma_\alpha^2}{\sigma_\varepsilon^2})T}\boldsymbol{i}_T\boldsymbol{i}_T'$$

$$= \boldsymbol{I} - \frac{\frac{\sigma_\alpha^2}{\sigma_\varepsilon^2}}{\frac{\sigma_\varepsilon^2 + \sigma_\alpha^2 T}{\sigma_\varepsilon^2}}\boldsymbol{i}_T\boldsymbol{i}_T' = \boldsymbol{I} - \frac{\sigma_\alpha^2}{\sigma_\varepsilon^2 + \sigma_\alpha^2 T}\boldsymbol{i}_T\boldsymbol{i}_T' = \boldsymbol{I} - \frac{1}{T}\boldsymbol{i}_T\boldsymbol{i}_T' + (\frac{1}{T} - \frac{\sigma_\alpha^2}{\sigma_\varepsilon^2 + \sigma_\alpha^2 T})\boldsymbol{i}_T\boldsymbol{i}_T'$$

$$= \boldsymbol{D} + \frac{1}{T}(\frac{\sigma_\varepsilon^2 + \sigma_\alpha^2 T}{\sigma_\varepsilon^2 + \sigma_\alpha^2 T} - \frac{\sigma_\alpha^2 T}{\sigma_\varepsilon^2 + \sigma_\alpha^2 T})\boldsymbol{i}_T\boldsymbol{i}_T' = \boldsymbol{D} + \frac{1}{T}(\frac{\sigma_\varepsilon^2}{\sigma_\varepsilon^2 + \sigma_\alpha^2 T})\boldsymbol{i}_T\boldsymbol{i}_T' = \boldsymbol{D} + \frac{\psi}{T}\boldsymbol{i}_T\boldsymbol{i}_T'$$

Alternatively, begin with

$$\Sigma = \sigma_\varepsilon^2(\boldsymbol{I} + \frac{\sigma_\alpha^2}{\sigma_\varepsilon^2}\boldsymbol{i}_T\boldsymbol{i}_T')$$

Then note that

$$\boldsymbol{I} + \frac{\sigma_\alpha^2}{\sigma_\varepsilon^2}\boldsymbol{i}_T\boldsymbol{i}_T' = \boldsymbol{I} - \frac{1}{T}\boldsymbol{i}_T\boldsymbol{i}_T' + (\frac{1}{T} + \frac{\sigma_\alpha^2}{\sigma_\varepsilon^2})\boldsymbol{i}_T\boldsymbol{i}_T' = \boldsymbol{D} + \frac{\sigma_\varepsilon^2 + T\sigma_\alpha^2}{T\sigma_\varepsilon^2}\boldsymbol{i}_T\boldsymbol{i}_T' = \boldsymbol{D} + \frac{\sigma_\varepsilon^2 + T\sigma_\alpha^2}{T\sigma_\varepsilon^2}\frac{T}{T}\boldsymbol{i}_T\boldsymbol{i}_T'$$

$$= \boldsymbol{D} + \frac{\sigma_\varepsilon^2 + T\sigma_\alpha^2}{\sigma_\varepsilon^2}\frac{1}{T}\boldsymbol{i}_T\boldsymbol{i}_T'$$

Note that $\boldsymbol{I} - \boldsymbol{D} = \boldsymbol{I} - (\boldsymbol{I} - \frac{1}{T}\boldsymbol{i}_T\boldsymbol{i}_T') = \frac{1}{T}\boldsymbol{i}_T\boldsymbol{i}_T'$

$$\boldsymbol{I} + \frac{\sigma_\alpha^2}{\sigma_\varepsilon^2}\boldsymbol{i}_T\boldsymbol{i}_T' = \boldsymbol{D} + \frac{\sigma_\varepsilon^2 + T\sigma_\alpha^2}{\sigma_\varepsilon^2}(\boldsymbol{I} - \boldsymbol{D}) = \boldsymbol{D} + \frac{1}{\psi}(\boldsymbol{I} - \boldsymbol{D})$$

Note that $\boldsymbol{Q}'\boldsymbol{Q} = \boldsymbol{Q}$. Therefore

$$\Sigma^{-1} = \frac{1}{\sigma_\varepsilon^2}(\boldsymbol{D} + \psi(\boldsymbol{I} - \boldsymbol{D}))$$

$$\Sigma^{-1/2} = \frac{1}{\sigma_\varepsilon}(\boldsymbol{D} + \psi^{1/2}(\boldsymbol{I} - \boldsymbol{D})) = \frac{1}{\sigma_\varepsilon}(\boldsymbol{I} - \frac{1}{T}\boldsymbol{i}_T\boldsymbol{i}_T' + \frac{\psi^{1/2}}{T}\boldsymbol{i}_T\boldsymbol{i}_T') = \frac{1}{\sigma_\varepsilon}(\boldsymbol{I} - \frac{(1 - \psi^{1/2})}{T}\boldsymbol{i}_T\boldsymbol{i}_T')$$

$$= \frac{1}{\sigma_\varepsilon}(\boldsymbol{I} - \frac{\theta}{T}\boldsymbol{i}_T\boldsymbol{i}_T')$$