

Week 4: Estimating Dynamic Models

Advanced Econometrics 4EK608

Vysoká škola ekonomická v Praze

Outline

- 1 Finite and infinite distributed lag models
- 2 Polynomial distributed lag
- 3 Geometric distributed lag (Koyck)
- 4 PAM, AEH, Rational expectations
- 5 FWL theorem - repetition from BSc courses

Czech terminology

Modely konečných a nekonečných rozložených zpoždění,
řád modelu, okamžitý a dlouhodobý multiplikátor,
rozdělení zpoždění, dynamicky úplné modely,
polynomicky (geometricky, racionálně) rozdělené zpoždění,
Koyckova transformace, model částečného přizpůsobení,
cílová (optimální) úroveň vysvětlované proměnné,
koeficient přizpůsobení, rychlost přizpůsobení,
adaptivní a racionální očekávání,
hypotéza efektivních trhů.

Finite and infinite distributed lag models

- Static models

$$y_t = \beta_0 + \beta_1 x_t + u_t, \quad t = 1, 2, \dots, T$$

- Finite distributed lag (FDL) Models

$$y_t = \alpha_0 + \delta_0 x_t + \delta_1 x_{t-1} + \delta_2 x_{t-2} + u_t$$

- Order of the FDL model,
impact multiplier vs. long-run multiplier,
temporary vs. permanent change in x ,
lag distribution

Finite and infinite distributed lag models

Dynamically complete models

- Model is dynamically complete if we have a “sufficient” number of lags of regressors, so that no more additional lags would help with explanation of variance in the dependent variable.
- In dynamically incomplete models, we usually detect autocorrelation in the error term of the LRM.

Finite and infinite distributed lag models

Infinite distributed lag (IDL) models

- Lagged regressors extend back to infinity
- We cannot estimate IDL models without the use of simplifying restrictions on parameters, i.e. restrictions on lag distribution
- IDL models are useful under the assumption of lagged coefficients converging to zero as lag increases
- Order of the IDL model (∞), impact multiplier vs. long-run multiplier, temporary vs. permanent change in x , ... all analogous to FDL models

FDL: Polynomial distributed lag (Almon)

Used in Finite distributed lag models

... example below also extends to higher order polynomials

$$y_t = \alpha + \beta_0 x_t + \beta_1 x_{t-1} + \cdots + \beta_m x_{t-m} + u_t \quad (1)$$

Simplifying assumption:

$$\beta_i = k_0 + k_1 i + k_2 i^2 \quad (2)$$

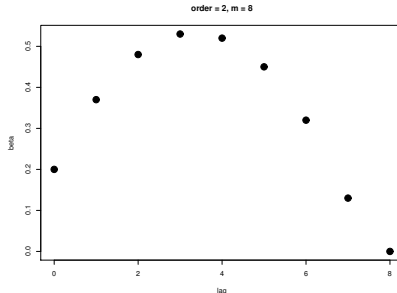
$$\beta_0 = k_0$$

$$\beta_1 = k_0 + k_1 + k_2$$

$$\beta_2 = k_0 + k_1 \cdot 2 + k_2 \cdot 4$$

...

$$\beta_m = k_0 + k_1 m + k_2 m^2$$



Polynomial distributed lag

- Almon-type transformation of (1) for $m = 8$ and $k = 2$:

$$y_t = \alpha + k_0 x_t + (k_0 + k_1 + k_2)x_{t-1} + (k_0 + 2k_1 + 4k_2)x_{t-2} + \dots + (k_0 + 8k_1 + 64k_2)x_{t-8} + u_t \quad (3)$$

$$y_t = \alpha + k_0(x_t + x_{t-1} + \dots + x_{t-8}) + k_1(x_{t-1} + 2x_{t-2} + \dots + 8x_{t-8}) + k_2(x_{t-1} + 4x_{t-2} + \dots + 64x_{t-8}) =$$

$$y_t = \alpha + k_0 \underbrace{\sum_{i=0}^8 x_{t-i}}_{W_{0t}} + k_1 \underbrace{\sum_{i=1}^8 i x_{t-i}}_{W_{1t}} + k_2 \underbrace{\sum_{i=1}^8 i^2 x_{t-i}}_{W_{2t}} + u_t \quad (4)$$

$$Y_t = \alpha + k_0 W_{0t} + k_1 W_{1t} + k_2 W_{2t} + u_t \quad (5)$$

We estimate (5), then calculate β_j as in (2)

...note the reduction in estimated parameters (10 vs 4).

Polynomial distributed lag

- Method developed by Shirley Almon in the 60ies.
- Equation (5) can be generalized: for m lags, sums go to m , for higher order polynomials, we add more W -terms.
- Advantages of this approach:
 - Saves degrees of freedom
 - Removes the problem of multicollinearity
 - Does not affect the assumptions for u , because errors do not change during transformation
- In EViews, transformation is slightly modified.
- In R, routines are available.

Geometric distributed lag (Koyck)

IDL linear regression model: $y_t = f(x_t, x_{t-1}, x_{t-2}, \dots)$:

$$y_t = \alpha + \delta_0 x_t + \delta_1 x_{t-1} + \delta_2 x_{t-2} + \delta_3 x_{t-3} + \dots + u_t$$

Assumptions for the geometric δ_t weights:

- ❶ $\delta_t = \delta_{t-1}\rho, \quad 0 < \rho < 1,$
- ❷ $\delta_j = \gamma\rho^j, \quad \text{where } \gamma \equiv \delta_0, \quad j = 0, 1, 2, \dots$

Instantaneous propensity (multiplier): δ_0

Long-run propensity (multiplier):

$$\delta_0 + \delta_1 + \dots = \gamma(1 + \rho + \rho^2 + \rho^3 + \dots) = \frac{\gamma}{1-\rho}$$

Geometric distributed lag (Koyck)

Koyck transformation of the IDL model:

$$y_t = \alpha + \delta_0 x_t + \delta_1 x_{t-1} + \delta_2 x_{t-2} + \cdots + u_t$$

$$y_t = \alpha + \gamma(\rho^0)x_t + \gamma\rho x_{t-1} + \gamma\rho^2 x_{t-2} + \cdots + u_t \quad (6)$$

$$y_{t-1} = \alpha + \gamma x_{t-1} + \gamma\rho x_{t-2} + \gamma\rho^2 x_{t-3} + \cdots + u_{t-1} \quad | \times \rho \quad (7)$$

$$\rho y_{t-1} = \alpha\rho + \gamma\rho x_{t-1} + \gamma\rho^2 x_{t-2} + \cdots + \rho u_{t-1} \quad (8)$$

Now, we subtract (8) from (6):

$$y_t - \rho y_{t-1} = \underbrace{\alpha(1 - \rho)}_{\alpha_0} + \gamma x_t + \underbrace{u_t - \rho u_{t-1}}_{v_t} \quad (9)$$

$$y_t = \alpha_0 + \gamma x_t + \rho y_{t-1} + v_t \quad (10)$$

Geometric distributed lag (Koyck)

IDL model: $y_t = \alpha + \delta_0 x_t + \delta_1 x_{t-1} + \delta_2 x_{t-2} + \cdots + u_t$

Koyck transf.: $y_t = \alpha_0 + \gamma x_t + \rho y_{t-1} + v_t$

Using the Koyck transformation, we can calculate parameters of the IDL model from the estimated model after Koyck transformation:

$$\hat{\delta}_0 = \hat{\gamma}$$

$$\hat{\delta}_j = \hat{\gamma} \hat{\rho}^j ; j = 1, 2, 3, \dots$$

$$\hat{\alpha} = \frac{\hat{\alpha}_0}{1 - \hat{\rho}}$$

Problems of the Koyck transformation:

- in (10), regressor y_{t-1} is not exogenous ($v_t = u_t - \rho u_{t-1}$)
... IVR will be discussed in Week 8
- $v_t = u_t - \rho u_{t-1}$ is not i.i.d.

Rational distributed lag

The geometric distributed lag is a special case of rational distributed lag (RDL) model:

$$\begin{aligned} y_t &= \alpha_0 + \gamma x_t + \rho y_{t-1} + v_t \quad (\text{geometric distributed lag}) \\ y_t &= \alpha_0 + \gamma_0 x_t + \rho y_{t-1} + \gamma_1 x_{t-1} + v_t \quad (\text{RDL}) \end{aligned} \quad (11)$$

This can be shown by successive substitution ($0 < \rho < 1$):

$$\begin{aligned} y_t &= \alpha + \gamma_0(x_t + \rho x_{t-1} + \rho^2 x_{t-2} + \dots) \\ &\quad + \gamma_1(x_{t-1} + \rho x_{t-2} + \rho^2 x_{t-3} + \dots) + u_t \end{aligned} \quad (12)$$

$$\begin{aligned} y_t &= \alpha + \gamma_0 x_t + (\rho \gamma_0 + \gamma_1) x_{t-1} + \rho(\rho \gamma_0 + \gamma_1) x_{t-2} + \\ &\quad + \rho^2(\rho \gamma_0 + \gamma_1) x_{t-3} + \dots + u_t \end{aligned} \quad (13)$$

After estimating (11), we can calculate lag distributions for (13)

Rational distributed lag

RDL specification:

$$y_t = \alpha_0 + \gamma_0 x_t + \rho y_{t-1} + \gamma_1 x_{t-1} + v_t$$

can be used to calculate δ_h in the IDL model:

$$y_t = \alpha + \delta_0 x_t + \delta_1 x_{t-1} + \delta_2 x_{t-2} + \cdots + u_t$$

With RDLs, impact propensity $\gamma_0 \equiv \delta_0$ can differ in sign from lagged coefficients.

$\delta_h = \rho^{h-1}(\rho\gamma_0 + \gamma_1)$ corresponds to the x_{t-h} variable for $h \geq 1$.

... Note: for $\rho > 0$, δ_h doesn't change sign with growing $h \geq 1$.

Long-run propensity: $LRP = \frac{\gamma_0 + \gamma_1}{1 - \rho}$,

where $|\rho| < 1 \Rightarrow$ the sign of LRP follows the sign of $(\gamma_0 + \gamma_1)$.

Also, $u_t = v_t + \rho v_{t-1} + \rho^2 v_{t-2} + \cdots \quad MA(\infty)$

Partial adjustment model

Partial adjustment model (PAM)

is based on two main assumptions:

- 1 LRM describes behavior of y_t^* , which is the unobserved, expected/equilibrium/target/optimum value of y_t :

$$y_t^* = \alpha + \beta x_t + u_t \quad (14)$$

- 2 Between two time periods, y_t follows the process:

$$y_t - y_{t-1} = \theta(y_t^* - y_{t-1}), \quad 0 < \theta < 1 \quad (15)$$

Hence, the actual Δy_t is only a fraction of the “desirable” change from y_{t-1} to the optimum value of y_t^* .

...in the special case of $\theta = 1$, Δy_t leads to optimum.

Note: (15) can be re-written as: $y_t = \theta y_t^* + (1 - \theta)y_{t-1}$

Partial adjustment model

Parameter estimation of PAM:

$$y_t^* = \alpha + \beta x_t + u_t \quad (14)$$

$$y_t = \theta y_t^* + (1 - \theta)y_{t-1}, \quad 0 < \theta < 1 \quad (15)$$

- ❶ Substitute for y_t^* in (15) from (14):

$$\begin{aligned} y_t &= \alpha\theta + \beta\theta x_t + (1 - \theta)y_{t-1} + \theta u_t \\ y_t &= \beta'_0 + \beta'_1 x_t + \beta'_2 y_{t-1} + \theta u_t \end{aligned} \quad (16)$$

- ❷ Estimate (16) and then calculate sought parameters of the PAM in (14) and (15):

$$\hat{\theta} = 1 - \hat{\beta}'_2$$

$$\hat{\alpha} = \hat{\beta}'_0 / \hat{\theta}$$

$$\hat{\beta} = \hat{\beta}'_1 / \hat{\theta}$$

Note that θu_t can be independent of y_{t-1} and i.i.d.

Adaptive expectations hypothesis

Adaptive expectations hypothesis (AEH)
model is based on two main assumptions:

- 1 LRM describes behavior of y_t , as a function of x_t^* : the unobserved, expected/equilibrium/target/optimum value of x_t (permanent income, potential output, etc.):

$$y_t = \alpha + \beta x_t^* + u_t. \quad (17)$$

- 2 The unobserved x_t^* process is defined as:

$$\begin{aligned} x_t^* - x_{t-1}^* &= \phi(x_t - x_{t-1}^*), \quad 0 < \phi < 1 \\ &\Downarrow \\ x_t^* &= \phi x_t + (1 - \phi)x_{t-1}^*. \end{aligned} \quad (18)$$

with $\phi = 0$ for static expectations
and $\phi = 1$ for immediate adjustment.

Note: alternative 2^{nd} hypothesis: $x_t^* = \phi x_{t-1} + (1 - \phi)x_{t-1}^*$.

Adaptive expectations hypothesis

Parameter estimation of AEH model:

$$y_t = \alpha + \beta x_t^* + u_t, \quad (17)$$

$$x_t^* = \phi x_t + (1 - \phi)x_{t-1}^* \quad (18)$$

Successive substitution for x_t^* from (18) to (17): IDL process

$$y_t = \alpha + \beta\phi x_t + \beta\phi(1 - \phi)x_{t-1} + \beta\phi(1 - \phi)^2 x_{t-2} + \cdots + u_t \quad (19)$$

After applying Koyck transformation, we get

$$\begin{aligned} y_t &= \alpha\phi + \beta\phi x_t + (1 - \phi)y_{t-1} + v_t \\ y_t &= \beta'_0 + \beta'_1 x_t + \beta'_2 y_{t-1} + v_t \end{aligned} \quad (20)$$

Estimate (20), then calculate parameters in (17) and (18).

$$\hat{\phi} = 1 - \hat{\beta}'_2$$

$$\hat{\alpha} = \hat{\beta}'_0 / \hat{\phi}$$

$$\hat{\beta} = \hat{\beta}'_1 / \hat{\phi}$$

Note: Both problems with Koyck-transformed model estimation apply.

Koyck, PAM, AEH: regression of y_t on x_t and y_{t-1}

The same underlying regression model (statistical form) is used:

- The Koyck transformation:

$$y_t = \alpha_0 + \gamma x_t + \rho y_{t-1} + v_t$$

- The Partial adjustment model (PAM):

$$y_t = \alpha\theta + \beta\theta x_t + (1 - \theta)y_{t-1} + \theta u_t$$

- The Model with adaptive expectations (AEM):

$$y_t = \alpha\phi + \beta\phi x_t + (1 - \phi)y_{t-1} + v_t$$

We can make three different interpretations from one estimated equation.

Of course, not all interpretations are always relevant, we must choose according to application and test assumptions.

Koyck, PAM, AEH: regression of y_t on x_t and y_{t-1}

Example: We have an estimated model for $c_t \leftarrow f(x_t)$:
private consumption (c_t) as a function of disposable GDP (x_t)

$$\hat{c}_t = 1,038 + 0.404x_t + 0.501c_{t-1}$$

$$\textbf{Koyck: } \hat{c}_t = \hat{\alpha}_0 + \hat{\gamma}x_t + \hat{\rho}c_{t-1}$$

Koyck: IDL, geometric decay in δ parameters assumed:

- $\hat{\rho} = 0.501$
- $\hat{\alpha}_0 = 1,038 = \hat{\alpha}(1 - \hat{\rho}) = \hat{\alpha}(1 - 0.501) \Rightarrow \hat{\alpha} = \frac{1,038}{0.499} = 2,080$
- $\hat{\delta}_j = \hat{\gamma}\hat{\rho}^j = 0.404 \times 0.501^j$
- $LRP = \frac{\hat{\gamma}}{1-\hat{\rho}} = \frac{0.404}{0.499} \doteq 0.81$
- IDL: $\hat{c}_t = 2,080 + \underbrace{0.404}_{\hat{\gamma}\hat{\rho}^0}x_t + \underbrace{0.202}_{\hat{\gamma}\hat{\rho}}x_{t-1} + \underbrace{0.101}_{\hat{\gamma}\hat{\rho}^2}x_{t-2} + \dots$

Koyck, PAM, AEH: regression of y_t on x_t and y_{t-1} **Example** continued:

$$\hat{c}_t = 1,038 + 0.404x_t + 0.501c_{t-1}$$

$$\text{PAM: } \hat{c}_t = \hat{\alpha}\hat{\theta} + \hat{\beta}\hat{\theta}x_t + (1 - \hat{\theta})c_{t-1}$$

- $(1 - \hat{\theta}) = 0.501 \Rightarrow \hat{\theta} = 0.499$
- $\hat{\alpha}\hat{\theta} = 1,038 \Rightarrow \hat{\alpha} = \frac{1,038}{0.499} = 2,080$
- $\hat{\beta}\hat{\theta} = 0.404 \Rightarrow \hat{\beta} = \frac{0.404}{0.499} \doteq 0.81$
- PAM: $\hat{c}_t^* = 2,080 + 0.81x_t$
 $c_t - c_{t-1} = 0.499 \cdot (c_t^* - c_{t-1})$
- If c_t has a prominent inertia and Δc_t significantly follows changes in habits, we might use the PAM approach.

Koyck, PAM, AEH: regression of y_t on x_t and y_{t-1} **Example** continued:

$$\hat{c}_t = 1,038 + 0.404x_t + 0.501c_{t-1}$$

$$\mathbf{AEH:} \quad \hat{c}_t = \hat{\alpha}\hat{\phi} + \hat{\beta}\hat{\phi}x_t + (1 - \hat{\phi})c_{t-1}$$

- $(1 - \hat{\phi}) = 0.501 \Rightarrow \hat{\phi} = 0.499$
- $\hat{\alpha}\hat{\phi} = 1,038 \Rightarrow \hat{\alpha} = \frac{1,038}{0.499} = 2,080$
- $\hat{\beta}\hat{\phi} = 0.404 \Rightarrow \hat{\beta} = \frac{0.404}{0.499} \doteq 0.81$
- AEH: $\hat{c}_t = 2,080 + 0.81x_t^*$

$$x_t^* = 0.499x_t + 0.501x_{t-1}^*$$
- If c_t is formed as a function of expected (e.g. permanent) GDP, we might prefer AEH.

Rational expectations

- Rational expectations

$$\mathbf{E}_{t-1}(x_t) = a_0 + a_1 x_{t-1} + b_1 z_{1,t-1} + b_2 z_{2,t-2} + \dots$$

$\mathbf{E}_{t-1}(x_t)$: expected value of x_t at time $t - 1$

$z_{k,t-j}$: exogenous variables with impact on $\mathbf{E}_{t-1}(x_t)$

We put $x_t^* = \mathbf{E}_{t-1}(x_t)$ into (17)

We assume that agents:

- know all relevant information
- know how to use this information

Agents can make prediction errors (v_t), so:

$$x_t = x_t^* + v_t$$

Rational vs. adaptive expectations

Under rational expectation:

- Expected value of prediction errors must be zero.
If they were systematically different from zero, rational agents would immediately adjust their forecasting methods accordingly.
- Prediction error must be uncorrelated with any information available when the prediction is made.
If not, this would imply that the forecaster has not made use of all available information.
- These properties can be used for testing the rational expectations hypothesis in different applications.

Some economic application that use expectations

- Philips curve
- Efficient market hypothesis (EMH)
- Consumption function - Permanent income hypothesis

FWL theorem - repetition from BSc courses

Frisch and Waugh in the 30-ies: detrending

Lovell in the 60-ies: deseasonalizing

- Example: Spurious regression with trend-stationary series:
e.g. Regression of the US GDP on salmon production in Norway

Solution (identical $\hat{\beta}_2$ estimates)

$$\text{M1: } gdp_t = \beta_0 + \beta_1 year_t + \beta_2 salmon_t + u_t$$

$$\text{M2: } gdp.detrended_t = \beta_2 salmon.detrended_t + u_t$$

$$\text{M2: } gdp.detrended_t = \beta_0 + \beta_1 year_t + \beta_2 salmon_t + u_t$$

$$\text{M4: } gdp_t = \beta_2 salmon.detrended_t + u_t$$

- Note that β_0 , β_1 and u_t may differ among equations.

FWL theorem

- Partitioned regression

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u} = \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2 + \mathbf{u} \quad (21)$$

- Projection matrices and residual makers

$$\mathbf{P} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}', \quad \mathbf{M} = \mathbf{I} - \mathbf{P}$$

$$\mathbf{P}_1 = \mathbf{X}_1(\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1', \quad \mathbf{M}_1 = \mathbf{I} - \mathbf{P}_1$$

$$\mathbf{P}_2 = \mathbf{X}_2(\mathbf{X}_2'\mathbf{X}_2)^{-1}\mathbf{X}_2', \quad \mathbf{M}_2 = \mathbf{I} - \mathbf{P}_2$$

- Some properties

$$\mathbf{P}\mathbf{P}_1 = \mathbf{P}_1, \quad \mathbf{P}\mathbf{X}_1 = \mathbf{X}_1 \quad (\mathbf{X}_1 \text{ is in the span } (\mathbf{X}))$$

$$\mathbf{P}\mathbf{P}_1 = (\mathbf{P}_1\mathbf{P})' = \mathbf{P}_1\mathbf{P} = \mathbf{P}_1$$

$$\mathbf{M}\mathbf{M}_1 = (\mathbf{M}_1\mathbf{M})' = \mathbf{M}_1\mathbf{M} = \mathbf{M}$$

$$\mathbf{M}_1\mathbf{X}_1 = \mathbf{0}, \quad \mathbf{X}_2'\mathbf{M} = \mathbf{0}$$

FWL theorem

$$\text{M1: } gdp_t = \beta_0 + \beta_1 year_t + \beta_2 salmon_t + u_t$$

$$\mathbf{y} = \mathbf{X}_1\beta_1 + \mathbf{X}_2\beta_2 + \mathbf{u} = \mathbf{X}\beta + \mathbf{u}$$

$$\text{M2: } gdp.detrended_t = \beta_2 salmon.detrended_t + u_t$$

$$\mathbf{M}_1\mathbf{y} = \mathbf{M}_1\mathbf{X}_2\beta_2 + \mathbf{u}$$

$$\text{M3: } gdp.detrended_t = \beta_0 + \beta_1 year_t + \beta_2 salmon_t + u_t$$

$$\mathbf{M}_1\mathbf{y} = \mathbf{X}_1\beta_1 + \mathbf{X}_2\beta_2 + \mathbf{u}$$

$$\text{M4: } gdp_t = \beta_2 salmon.detrended_t + u_t$$

$$\mathbf{y} = \mathbf{M}_1\mathbf{X}_2\beta_2 + \mathbf{u}$$

FWL theorem

FWL theorem & applications

- ① Estimates of β_2 for a partitioned model (21) in all M1 - M4 models are the same
- ② Residuals from M1, M2, M3 are the same

Lemma: Based on properties of projection matrices and residual makers, we can show that $\hat{\beta}_2$ estimators in all four (M1 to M4) specifications can be expressed as:

$$\hat{\beta}_2 = (\mathbf{X}_2' \mathbf{M}_1 \mathbf{X}_2)^{-1} \mathbf{X}_2' \mathbf{M}_1 \mathbf{y}$$

Application: deseasonalizing, centered coefficient of determination, interchangeability of reference categories (e.g. quarterly dummies), simplification of many proofs in econometrics

FWL theorem (detrrending)

$$\mathbf{y} = \mathbf{X}\hat{\boldsymbol{\beta}} + \hat{\mathbf{u}} = \mathbf{P}\mathbf{y} + \mathbf{M}\mathbf{y} = \mathbf{X}_1\hat{\boldsymbol{\beta}}_1 + \mathbf{X}_2\hat{\boldsymbol{\beta}}_2 + \mathbf{M}\mathbf{y}$$

Skeleton of the proof (for M1 equation):

$$\begin{aligned} \mathbf{y} &= \mathbf{X}_1\hat{\boldsymbol{\beta}}_1 + \mathbf{X}_2\hat{\boldsymbol{\beta}}_2 + \mathbf{M}\mathbf{y} \quad | \quad \mathbf{X}_2'\mathbf{M}_1 \times \\ \mathbf{X}_2'\mathbf{M}_1\mathbf{y} &= \mathbf{X}_2' \underbrace{\underbrace{\mathbf{M}_1\mathbf{X}_1}_{0}}_0 \hat{\boldsymbol{\beta}}_1 + \mathbf{X}_2'\mathbf{M}_1\mathbf{X}_2\hat{\boldsymbol{\beta}}_2 + \mathbf{X}_2' \underbrace{\underbrace{\mathbf{M}_1\mathbf{M}}_M}_{0} \mathbf{y} \\ \mathbf{X}_2'\mathbf{M}_1\mathbf{y} &= \mathbf{X}_2'\mathbf{M}_1\mathbf{X}_2\hat{\boldsymbol{\beta}}_2 \Rightarrow \hat{\boldsymbol{\beta}}_2 = (\mathbf{X}_2'\mathbf{M}_1\mathbf{X}_2)^{-1} \mathbf{X}_2'\mathbf{M}_1\mathbf{y} \Rightarrow \text{QED} \end{aligned}$$