Notes on estimation of the emissions model (1/6/2021)

Summary of model

Emissions in region i at time t equal

$$\begin{aligned} e_{it} &= \frac{1}{b}\left(b_{0i}+\beta x_{it}\right)+\frac{1}{b}\left(\rho\nu_{t-1}+\alpha_t+\mu_{it}\right) && (1)\\ \text{with } \nu_{it} &= \rho\nu_{t-1}+\alpha_t+\mu_{it} && (2)\end{aligned}$$

and

$$\alpha_t \sim iid\left(0,\sigma_\alpha^2\right), \mu_{i,t} \sim iid\left(0,\sigma_\mu^2\right), \mathbf{E}\left(\alpha_t\mu_{i,\tau}\right)=0\forall t,\tau. \qquad (3)$$

The aggregate shock is

$$\begin{aligned} \nu_t &\equiv \frac{\sum_i \nu_{it}}{n}=\rho\nu_{t-1}+\alpha_t+\theta_t \text{ with } \theta_t \equiv \frac{\sum_i \mu_{it}}{n} && (4)\\ \nu_t &= \rho\nu_{t-1}+\eta_t \text{ with } \eta_t\equiv\alpha_t+\theta_t && (5)\end{aligned}$$

The covariance matrix

Defining

$$s=|\tau-t|$$

and

$$\iota(i,j)=\begin{cases}1 \text{ for } i=j\\ 0 \text{ for } i\neq j\end{cases}$$

I can write (after some calculation) the covariance as

$$\mathbf{E}\nu_{i,t}\nu_{j,t+s}=\frac{\rho^s}{1-\rho^2}\sigma_\alpha^2+\left(\iota(i,j)+\frac{\rho^{s+2}}{n}\frac{1}{1-\rho^2}\right)\sigma_\mu^2. \qquad (6)$$

Dividing by b I have

$$\begin{gathered}\frac{\mathbf{E}\nu_{i,t}\nu_{j,t+s}}{b^2}=\sigma^2\left[\frac{\rho^s}{1-\rho^2}+\left(\iota(i,j)+\frac{\rho^{s+2}}{n}\frac{1}{1-\rho^2}\right)\lambda\right]\\ \text{with } \sigma^2\equiv\frac{\sigma_\alpha^2}{b^2} \text{ and } \lambda\equiv\frac{\sigma_\mu^2}{\sigma_\alpha^2}.\end{gathered} \qquad (7)$$

(I can estimate ρ, σ^2 and λ, but not the scaling factor b.)

For future use, note that

$$\begin{gathered}\frac{d\left(\frac{\rho^s}{1-\rho^2}+\left(\iota+\frac{\rho^{s+2}}{n}\frac{1}{1-\rho^2}\right)\lambda\right)}{d\rho}=\\ \rho^{s-1}\rho^2\frac{s-s\rho^2+2}{n(\rho-1)^2(\rho+1)^2}\lambda+\rho^{s-1}\frac{s+2\rho^2-s\rho^2}{(\rho-1)^2(\rho+1)^2}\end{gathered} \qquad (8)$$

and

$$\frac{d\left(\frac{\rho^s}{1-\rho^2}+\left(\iota+\frac{\rho^{s+2}}{n}\frac{1}{1-\rho^2}\right)\lambda\right)}{d\lambda}=\left(\iota\left(i,j\right)+\frac{\rho^{s+2}}{n}\frac{1}{1-\rho^2}\right) \tag{9}$$

Write the panel as a system of equations
Define

$$\underset{4x1}{\mathbf{e}_t}=\begin{pmatrix} e_{1t} \\ e_{2t} \\ e_{3t} \\ e_{4t} \end{pmatrix} \text{ and } \mathbf{v}_t=\begin{pmatrix} v_{1t} \\ v_{2t} \\ v_{3t} \\ v_{4t} \end{pmatrix} \tag{10}$$

I am stacking the system in such a way that, for example, the first four elements equal emissions of the four regions in the first period.

$$\underset{6x6}{\mathbf{x}_t}=\begin{bmatrix} 1 & 0 & 0 & 0 & t & t^2 \\ 0 & 1 & 0 & 0 & t & t^2 \\ 0 & 0 & 1 & 0 & t & t^2 \\ 0 & 0 & 0 & 1 & t & t^2 \\ 0 & 0 & 0 & 0 & t & t^2 \\ 0 & 0 & 0 & 0 & t & t^2 \end{bmatrix} \text{ and } \underset{4Tx1}{\mathbf{e}}=\begin{pmatrix} \mathbf{e}_1 \\ \mathbf{e}_2 \\ \mathbf{e}_3 \\ \vdots \\ \vdots \\ \mathbf{e}_T \end{pmatrix} \text{ and } \underset{4Tx6}{\mathbf{X}}=\begin{pmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \\ \mathbf{x}_3 \\ \vdots \\ \vdots \\ \mathbf{x}_T \end{pmatrix} \tag{11}$$

$$\text{and } \underset{6x1}{\beta}=\begin{pmatrix} \delta_1 \\ \delta_2 \\ \delta_3 \\ \delta_4 \\ \phi_1 \\ \phi_2 \end{pmatrix} \text{ and } \underset{4Tx1}{\mathbf{v}}=\begin{pmatrix} \mathbf{v}_1 \\ \mathbf{v}_2 \\ \mathbf{v}_3 \\ \vdots \\ \vdots \\ \mathbf{v}_T \end{pmatrix}, \text{ and } \mathbf{E}\left(\mathbf{v}\mathbf{v}'\right)=\sigma^2\mathbf{V} \tag{12}$$

(I defined $\sigma^2=\frac{\sigma_\alpha^2}{b^2}$ above.) With this notation we can write the stacked system as

$$\mathbf{e}=\mathbf{X}\beta+\mathbf{v} \text{ with } \mathbf{E}\left(\mathbf{v}\mathbf{v}'\right)=\sigma^2\mathbf{V} \tag{13}$$

The first four elements of β equal the region-specific constants (the δ's); the next two elements equal the coefficients of the linear and quadratic time trend.

I want a procedure for associating an arbitrary element of the vector $\mathbf{v}$ with a particular time period and region. (It seems to me that we need this information for programming....) Define the floor function $\lfloor y \rfloor$ as the largest integer no greater than y. For example, $\lfloor 7.2 \rfloor = 7$ and $\lfloor 7.0 \rfloor = 7$, and $\lfloor 0.25 \rfloor = 0$. Consider v_m, defined as the m'th element of $\mathbf{v}$. This element corresponds to time period $t = \left\lfloor \frac{m-1}{4} \right\rfloor + 1$ and region $j = m - 4(t-1)$. (Check!)

We can also invert this relation. For example, observation t, j corresponds to $m = (t-1)\,4 + j$. With this information we can translate the elements defined by equation 7 into elements of the matrix $\mathbf{V}$.

The (m, p) element, $\mathbf{V}_{m,p}$ is

$$\mathbf{E}v_m v_p = \mathbf{E}v_{m-4(t-1),t} v_{p-4(\tau-1),\tau} \text{ with } t = \left\lfloor \frac{m-1}{4} \right\rfloor + 1 \text{ and } \tau = \left\lfloor \frac{p-1}{4} \right\rfloor + 1 \tag{14}$$

(I see that this notation might be ambiguous. By $\mathbf{E}v_m v_p$ I mean the expectation of the product of the m'th and the p'th element of the $4T \times 1$ vector $\mathbf{v}$. By $\mathbf{E}v_{m-4(t-1),t} v_{p-4(\tau-1),\tau}$ I mean the expectation of the product of the errors associated with time period t and region $m - 4(t-1)$, and time period τ and region $p - 4(\tau - 1)$.)

Above I defined the absolute value of the difference between two time indices as s. Using the expressions for any two arbitrary elements of the vector of stacked errors, $\mathbf{v}$, m and p, and equation 14, I have

$$\tau - t = \left\lfloor \frac{p-1}{4} \right\rfloor + 1 - \left(\left\lfloor \frac{m-1}{4} \right\rfloor + 1 \right) = \left\lfloor \frac{p-1}{4} \right\rfloor - \left\lfloor \frac{m-1}{4} \right\rfloor . \tag{15}$$

Therefore,

$$s \equiv |\tau - t| = \left| \left\lfloor \frac{p-1}{4} \right\rfloor - \left\lfloor \frac{m-1}{4} \right\rfloor \right| \tag{16}$$

The estimation procedure

Define $\Gamma = \mathbf{V}^{-1}$. Greene calls the covariance matrix Ω. I call this matrix $\mathbf{V}$. Greene calls the vector of LHS variables y. I call this vector $\mathbf{e}$. Greene assumes that there are n observations. For my problem there are $4T$ observations. Green calls the vector of residuals $\boldsymbol{\epsilon}$. I refer to these residuals as $\tilde{\mathbf{v}} = \mathbf{e} - \mathbf{X}\tilde{\beta}$.

Greene pg 471 gives the estimation equations for β and σ as the solutions to the first order conditions (from minimizing the likelihood function, assuming normality) as

$$\begin{aligned}\mathbf{X}'\Gamma\left(\mathbf{e}-\mathbf{X}\boldsymbol{\beta}\right) &= 0 \qquad (17)\\ -\frac{4T}{2\sigma^2}+\frac{1}{2\sigma^4}\left(\mathbf{e}-\mathbf{x}\boldsymbol{\beta}\right)'\Gamma\left(\mathbf{e}-\mathbf{x}\boldsymbol{\beta}\right) &= 0. \qquad (18)\end{aligned}$$

The derivative of the likelihood function with respect to Γ returns the first order condition (eq 11-31, pg 471 in Greene).

$$\frac{1}{2\sigma^2}\left(\sigma^2\mathbf{V}-\tilde{\mathbf{v}}\tilde{\mathbf{v}}'\right)=0. \qquad (19)$$

Greene notes that the covariance matrix must be restricted in some way, i.e. it must be possible to write the covariance matrix as $\mathbf{V}=\mathbf{V}\left(\gamma\right)$, where γ is a vector of parameters. For my problem, $\gamma=(\rho,\lambda)$. (Greene refers to this parameter vector as θ, but I used that symbol above.) Greene outlines the Oberhofer-Kmenta algorithm for achieving consistent estimators: begin with a consistent estimator of γ. Using these, solve equations 17 and 18 to obtain estimates of β and σ^2. Using these, solve equation 19 to obtain an estimate of γ. Repeat until satisfactory convergence. He gives conditions (pg 472) for this process to converge to the MLE. The only condition that I do not know how to satisfy is the requirement that the starting guess for γ is consistent. I do not know how to guarantee this. A practical solution would be to begin with different initial values of the guess for γ and show that the convergence does not depend on the initial guess.

The remaining issue is how to solve 19 to obtain an estimate of γ, conditional on estimates of β and σ^2. There may be a clever way of doing this without using derivatives. However, because we have a fairly simple expression for $\mathbf{V}$ it might be easier to use derivatives. Define L as the likelihood function.

$$L=\text{"}terms\text{"}-\frac{1}{2\sigma^2}\tilde{\mathbf{v}}'\boldsymbol{\Gamma}\tilde{\mathbf{v}}+\frac{1}{2}\ln\left|\Gamma\right| \qquad (20)$$

where "terms" are independent of Γ. Use the rule for the derivative of the inverse of a matrix:

$$\frac{dA^{-1}}{d\rho}=-A^{-1}\frac{dA}{d\rho}A^{-1}. \qquad (21)$$

Consider the FOC for ρ. I have

$$\frac{\partial\tilde{\mathbf{v}}'\boldsymbol{\Gamma}\tilde{\mathbf{v}}}{\partial\rho}=\frac{\partial\tilde{\mathbf{v}}'\mathbf{V}^{-1}\tilde{\mathbf{v}}}{\partial\rho}=-\tilde{\mathbf{v}}'\mathbf{V}^{-1}\frac{d\mathbf{V}}{d\rho}\mathbf{V}^{-1}\tilde{\mathbf{v}} \qquad (22)$$

I also have the rule (see "cookbook" eqn 10)

$$
\begin{aligned}
\partial\left(\ln|\Gamma|\right) &= Tr\left(\Gamma^{-1}\partial\Gamma\right)\Rightarrow\frac{\partial\left(\ln|\Gamma|\right)}{\partial\rho}=Tr\left(\Gamma^{-1}\frac{\partial\Gamma}{\partial\rho}\right) &(23)\\
&= -Tr\left(\mathbf{V}\mathbf{V}^{-1}\frac{\partial\mathbf{V}}{\partial\rho}\mathbf{V}^{-1}\right)=-Tr\left(\frac{\partial\mathbf{V}}{\partial\rho}\mathbf{V}^{-1}\right) &(24)
\end{aligned}
$$

Using these two equations I have the FOC for ρ:

$$\frac{d\ln L}{d\rho}=\frac{1}{2\sigma^2}\tilde{\mathbf{v}}'\mathbf{V}^{-1}\frac{d\mathbf{V}}{d\rho}\mathbf{V}^{-1}\tilde{\mathbf{v}}-Tr\left(\frac{d\mathbf{V}}{d\rho}\mathbf{V}^{-1}\right)=0 \tag{25}$$

I also have the FOC for λ

$$\frac{d\ln L}{d\lambda}=\frac{1}{2\sigma^2}\tilde{\mathbf{v}}'\mathbf{V}^{-1}\frac{d\mathbf{V}}{d\lambda}\mathbf{V}^{-1}\tilde{\mathbf{v}}-Tr\left(\frac{d\mathbf{V}}{d\lambda}\mathbf{V}^{-1}\right)=0 \tag{26}$$

Conditional on estimates of σ^2 and β, equations 25 and 25 comprise two equations in two unkowns, ρ and λ.

We can estimate the parameters using the method of steepest gradient ascent (see https://en.wikipedia.org/wiki/Gradient_descent). Denote $a=(\rho,\lambda)$, the parameters we want to estimate (along with β and σ^2). Denote $F=\ln L$, our maximand, and denote $\nabla F(a)$, the gradient. The first element of $\nabla F(a)$ is given by the left side of equation 25 and the second element is given by the left side of equation 26. (I'm using these definitions so that things line up with the Wikipedia article.) Here is the proposed algorithm.

1. We need an initial guess to begin the method of gradient ascent. $\rho_0=0.9$ is a reasonable starting value for ρ. (The subscript 0 denotes the starting value.) To find a reasonable starting value for λ, use a grid, e.g. $\lambda\in\{0.5,1,1.5,2....20\}$. For each of these values calculate V and V^{-1}. Using these matrices calculate the estimate of β and σ^2 :

$$
\begin{aligned}
\hat{\beta} &= \left(\mathbf{X}'V^{-1}\mathbf{X}\right)^{-1}\mathbf{X}'V^{-1}\mathbf{e} &(27)\\
\hat{\sigma}^2 &= \frac{1}{n}\left(\mathbf{e}-\mathbf{X}'\beta\right)'V^{-1}\left(\mathbf{e}-\mathbf{X}'\beta\right) &(28)
\end{aligned}
$$

with $n=4T$. For each of these values of λ, calculate the log likelihood function

$$\ln L=-\frac{n}{2}\left[\ln\left(2\pi+\ln\sigma^2\right)\right]-\frac{1}{2\sigma^2}\tilde{\mathbf{v}}'V^{-1}\tilde{\mathbf{v}}+\frac{1}{2}\ln\left|V^{-1}\right| \tag{29}$$

with $\tilde{\mathbf{v}} = \mathbf{e} - \mathbf{X}\tilde{\beta}$. Choose the λ that maximizes the criteria; call this value λ_0. Graph $\ln L$ over λ to see if the function is well behaved. If it reaches a maximum at either 0.5 or 20, you need to change the initial grid to go beyond these points.

2. Using $a_0 = (\rho_0, \lambda_0)$, update the guess using $a_{n+1} = a_n + \gamma_n \nabla F(a_n)$, where $\gamma_n > 0$ is the step size at iteration n. The Wikipedia article gives the formula for γ_n, called the Barzilai-Borwein method. This looks simple to calculate.

3. Stop when $\| a_j - a_{j-1} \|$ is small. (I don't know what small means in this context; we have to experiment.)

I think that perhaps the greatest danger of a coding error arises in correctly assigning the entries of the V matrix and the derivative of this matrix. Therefore, I suggest writing the code where T (the number of periods) and n (the number of regions) are parameters. To check the code, set $T = 3$ and $n = 2$, so that V and the derivative matrices $\frac{d\mathbf{V}}{d\rho}$ and $\frac{d\mathbf{V}}{d\lambda}$ are all 6x6 matrices. You can construct this matrix by hand, and then confirm that your code returns the correct matrices.