Version: January 7, 2022

Economics 696F: Econometrics of Dynamic Industrial Organization Models
Comprehensive Assignment for Class

You may work in groups of 2-3 people for the programming. Everyone should turn in his or her own assignment. If you work in groups, please let me know who is in your group. You may turn in your assignment electronically or in hard-copy.

The assignment is based on Blundell, W., Gowrisankaran, G., and Langer A. (2020), "Escalation of Scrutiny: The Gains from Dynamic Enforcement of Environmental Regulations," *American Economic Review* 110: 2558-85 [BGL].

The dataset analysis_data.csv (which is zipped in Dropbox) is the final analysis data for this paper. The unit of observation is the plant/quarter. The dataset contains the following information in order:

- FRS number (a unique plant identifier)
- Quarter
- The variable state $\widetilde{\Omega}^1$, which includes NAICS, EPA region, and gravity
- Presence of an inspection
- Fine in millions of dollars
- A new notice of violation
- A new high priority violator (HPV) (you don't need to use this)
- Lagged investment (one quarter ago)
- Double lagged investment (two quarters ago)
- Investment
- NAICS recoded (coded 1-7)
- NAICS code (original)
- EPA region
- Gravity of violation measure (coded from 1-5)
- Indicator for compliance status
- Indicator for violator status
- Indicator for HPV status
- Ordered violator status (0, 1, or 2 based on the above variables)
- Number of HPVs (not used)
- Depreciated accumulated violations (DAV, described in the paper on p. 18)
- Other variables that follow that you don't need to use

The dataset data_for_bellman_computations.csv (which is also zipped in Dropbox) provides the conditional choice probabilities (CCPs) for the regulator actions and accompanying transitions. Each row in this file reflects one state. The dataset contains the following variables in order:

- The state:
    - The variable state $\widetilde{\Omega}^1$, which includes NAICS, EPA region, and gravity
    - NAICS recoded
    - EPA region
    - Gravity of violation measure
    - Lagged investment

    - Double lagged investment
    - Lagged ordered violator status
    - Lagged DAV as an integer grid in increments of 0.5, so 0 represents 0, 1 represents 0.5, 2 represents 1, etc.
    - Lagged DAV in increments of 0.5
- 80 discretized potential outcomes from this state. For each outcome:
    - Indicator for an inspection
    - Indicator for a new notice of violation
    - Fine in millions of dollars
    - Probability of the above three variables occurring at given state
    - Probability of transition to compliance, at given state and conditional on the above values of inspection, violation, and fine.
    - Probability of transition to regulator violator, at given state and conditional on the above values of inspection, violation, and fine.
    - Probability of transition to HPV status, at given state and conditional on the above values of inspection, violation, and fine.

The assignment has multiple parts, each of which will be handed out separately and is due separately.

Part 1: Reduced form analysis

Before completing Part 1 of the assignment, it will be helpful to read Section 3 of BGL in detail. This section explains the construction of the variables in the data. This section will help you understand the difference between being a violator and having a violation; the difference between being a regular violator and a HPV; and the fact that the past two lags of investment affect the state transition but the current investment does not.

1) Reduced-form regressions
    a. Regress investment on EPA region, NAICS code, and gravity indicators, HPV status, and DAV with a linear probability model. Report your results. What can you infer from them about plants' investment decisions?
    b. Papers in the empirical literature on environmental compliance have regressed compliance on the strictness of environmental regulation, as measured by inspections, violations, and fines. Collapse the data to obtain the mean levels of compliance, inspections, violations, and fines by EPA region / NAICS code / gravity / quarter level and run a regression of compliance on these variables with the collapsed data. Report your findings.
    c. What can you infer from part b about the impact of the strictness of environmental regulations on compliance? How might endogeneity affect your conclusions here?

2) Evidence of dynamic enforcement
    a. Regress inspections on lagged regular violator (lag_violator_not_HPV), lagged HPV status (lag_HPV_status), DAV (not lagged, because it already reflects lagged values), and their interactions. What do you find here about the presence of dynamic enforcement?
    b. Regress inspections on the same variables as in part a and indicators for EPA region / NAICS code / gravity. What are the differences between the results here and in part a—which has fewer controls—and how do we interpret them?
    c. Regress fines on the same variables as in part b and new inspections and violations. What do you find here about the presence of dynamic enforcement?

3) Estimation of a simple two-period structural model
    a. For plants not in compliance (based on their lagged reported status), regress fines on (1) lagged investment, (2) lagged HPV status, (3) lagged regular violator / HPV status interacted with DAV, (4) interactions of lagged investment with the variables in (2) and (3), and (5) EPA region / NAICS code / gravity indicators. Report your results.
    b. Using this regression, for every plant not in compliance, create measures of its expected fines in the next period given investment and given no investment. Report conditional means for these two variables, by violator / HPV status, for EPA region 1 and industry code 1. What do you find?

c. For every plant not in compliance, using a probit specification, regress investment on the difference between expected fines in the next period given investment and given no investment. Report your results.
d. Write down a simple two-period model where the plant invests in a period in order to lower fines in the subsequent period. Let the cost of investment be a constant plus a normally distributed residual. Interpret your results from part c) in the context of this model. What do your results imply about the cost of investment, under this model? Compare your investment costs to BGL and explain why they might be different.

Part 2: Quasi maximum likelihood estimation

Before completing Part 2 of the assignment, it will be helpful to read BGL, Section 4 and the sub-parts of Appendix A3 entitled "Plant Dynamic Optimization" and "Computing the Bellman Equation" in detail. Section 4 details the model including details on the CCPs that are used. Note that you don't have to estimate these CCPs as the file data_for_bellman_computations.csv provides simulations from these CCPs. The appendix subsections provide more details on the Bellman equation and the computation necessary for this part of the assignment.

For Part 2 of the assignment, please limit your data to one industrial sector, Mining and Extraction, as in the file that is distributed with this assignment. Note that this sector takes on a NAICS value of 21 or a naics_recode value of 1. Please use a discount factor of $\beta = 0.95^{1/4}$ as in the paper.

1) Computation of the plant's dynamic optimization decision for fixed parameters

a. Consider the value function at the point right after the regulator has moved, $\tilde{V}(\tilde{\Omega})$. What are the states $\tilde{\Omega}$ that enter here and how many such states are there? Use the grid of DAV from the data_for_bellman_computations.csv file. (The function Vtilde is defined in BGL, equation A2.)
b. Now consider the value function at the beginning of the period, $V(\Omega)$. What are the states Ω that enter here and how many states are there? (The function Vtilde is defined in BGL, equation A1.)
c. Compute the plant's dynamic optimization decision for the parameters $\hat{\theta} \equiv (\theta^X = 2, \theta^I = -0.5, \theta^V = -0.5, \theta^F = -5.0, \theta^H = -0.1)$. Report the investment probability for all states with DAV = 2 (for which DAVgrid takes on a value of 4). To check your Bellman code, the file value_part2_problem1_thetaBGL.csv contains $V(\Omega)$ and the file valuetilde_part2_problem1_thetaBGL.csv contains $\tilde{V}(\tilde{\Omega})$, both for the BGL QML parameter values, which are $\hat{\theta}^{BGL} = (2.872, -0.049, -0.077, -5.980, -0.065)$. (The reported value function does not include Euler's constant γ each period.)

Here is the outline of code that will call the Bellman function. You can base your program on this code.

```
Function Problem1
    Load in data_for_bellman_computations.csv
    Keep if orig_naics==21
    Declare variables:
      A) Vtilde (matrix of the value of being in every state Ω̃)
      B) NewV (updated matrix of the value of being in every state Ω)
      C) OldV (initial matrix of the value of being in every state Ω)
      D) Investprob (probability of investment at every state Ω̃)
      E) Coeff (θ coefficient vector)
    Initialize Coeff to the values noted above
    (NewV, Vtilde, Investprob) = Bellman(Coeff)
    Print out NewV, Vtilde, Investprob
End function
```

Here is the outline of the Bellman function itself.

```
function Bellman(Coeff)
    Declare variables OldV, NewV, Vtilde, Investment, norm
    Initialize OldV = 0, for every state
    Loop
      A) Loop through states Ω̃. For each Ω̃:
            1) Solve OldV(Ω) next period if investment this period for bins on
               both sides of the actual DAV. Find the actual value with linear
               interpolation.[1]
            2) Solve OldV(Ω) next period if no investment this period for bins
               on both sides of the actual DAV. Find the actual value with
               linear interpolation.
            3) Use logit inclusive value formula to solve for Vtilde
            4) Save Investprob(Ω) as the probability of an investment at this
               state
      B) Loop through states Ω. For each Ω:
            1) Loop through 240 regulatory actions and transitions: 80 cases
               of inspections, violations, and fines X three cases of
               regulatory status transitions to states Ω̃ (compliance, regular
               violator, HPV). For each of these cases:
                     a) Compute the resulting static utility
                     b) Compute Vtilde for Ω̃
                     c) Calculate the probability of this case
            2) Calculate the weighted sum over the 240 cases for this state to
               get NewV(Ω)
      C) Calculate "norm" as the sup norm difference between OldV and NewV
         across states
   Until norm < 1e-6
   Return: NewV, Vtilde, Investprob
End function
```

2) Nested fixed point quasi-maximum likelihood estimation

a. Report the quasi-likelihood for the parameter vector $\hat{\theta}$ given above. To check your code, the file likelihood_part2_problem2.csv contains the investment probabilities for and quasi-likelihood for each observation, at $\hat{\theta}^{BGL}$. (Note that an observation might have DAV between two values of DAV that are calculated. E.g., if $DAV = 3.7$, this implies a weight of 0.4 on 4 and a weight of 0.6 on 3.5. The file reports the likelihood values and weight on both calculated values.)
b. Maximize the likelihood with a non-linear search. For starting values, use the BGL parameters $\hat{\theta}^{BGL}$. Report your results. How do the results differ from BGL?
c. Calculate the standard errors for your parameters using the standard outer product approximation method, which is $\widehat{Var}(\theta^{ML}) = \left[\sum_j \frac{d \log L_j(\theta^{ML})}{\theta} \frac{d \log L_j(\theta^{ML})'}{\theta}\right]^{-1}$. Here,

[1] For example, if DAV next period would be 1.8, then since our grid points are are at intervals of 0.5, we would take OldV(Ω) for DAV of 1.5 and 2 and then calculate the weighted average, where the weights are .4 and .6, respectively.

please let *j* index a plant/quarter observation (unlike BGL, which accounts for correlations in residuals within a plant across time). Calculate the derivatives using a numerical approximation where you add 1e-6 to each element of θ.

Here is an additional outline of code to do this on which you can model your program:

```
Function Problem2
    Load in data_for_bellman_computations.csv
    Keep if orig_naics==21
    Load in analysis_data.csv
    Keep if orig_naics==21 and ordered_violator>0
    Call non-linear equation solver with function LogLike
    Report Final values

    /* Calculate standard errors */
    Loop over the 5 parameters
    For each parameter, calculate ∂LogL_j/∂θ_i with a numerical approximation
    Use the above formula to approximate the variance
    Report the square root of the diagonal elements as the standard errors
End function


function LogLike(Coeff)
    Declare variables NewV, Vtilde, Investprob, Loglike (a vector of log
    likelihoods for each observation)
    (Newv, Vtilde, Investprob) = Bellman(Coeff)

    Set Loglike=0
    Loop through observations i in data
      A) Find the two interpolated DAV states for the observation that are
         calculated in the Bellman equation
      B) Using "Investprob" solve for the probability of the observed action
         by interpolating across the two states
      C) Set LogLike[i] as the log likelihood value for this observation
    Return the sum of Loglike across observations
End function
```

3) Interpretation of results

a. What is the implicit cost of investment to a plant in dollars? What about the implicit cost of HPV status, inspections, and violations? Explain.
b. What if fines carried a stigma, so that \$1 of fines was actually equivalent to \$$\alpha$ of real cost to the plant? How would this change your implicit costs above? Explain.

Part 3: GMM estimation of random coefficients model[2]

Before completing Part 3 of the assignment, it will be helpful to read BGL, Section 4.4 "Empirical Implementation with Random Coefficients" and the sub-parts of Appendix A3 entitled "Choice of Fixed Grid Values for GMM Estimation" and "Weighting Matrix and Estimation of GMM Parameters η_j" in detail.

For Part 3 of the assignment, please again limit your data to one industrial sector, Mining and Extraction, as in the file that is distributed with this assignment. Note that this sector takes on a NAICS value of 21 or a naics_recode value of 1.

1) Calculation of transition matrix

For $\hat{\theta}^{BGL}$, start with your estimated investment probabilities for each state $\widetilde{\Omega}$. Then, for every fixed part of the state ($\widetilde{\Omega}_1$), calculate the transition matrix from $\widetilde{\Omega}_2$ (the variable part of the state) to the next period's value of $\widetilde{\Omega}_2$. To do this, for every state, you need to use:

- The probability of investment given $\hat{\theta}^{BGL}$.
- Given investment, the probability of each of the two beginning-of-period states Ω_2 in the following period. There are two values because DAV will be lie between two DAV state bins and you will need to interpolate between these bins (as in footnote 1 in Part 2).

This gives four potential states at the start of the next period. For each of these four states, you need to sum over the 240 possible violations, inspections, fines, and transitions cases. The 240 cases here lead a possibility of violation or no violation and compliance, regular violator, or HPV status for $\widetilde{\Omega}_2$. For each current state $\widetilde{\Omega}$, there are 17 possible $\widetilde{\Omega}$ states next period: compliance (1 state),[3] and every interaction of regular violator/HPV, DAV, lagged investment, and violation (2^4 states). You need to calculate the probability of transitions to each of these 17 values.

Report the state transition probabilities for the state with Gravity=1, Region = 1, DAVgrid = 4, Violation=1, Ordered Violator = 2, and Lag Investment = 1. The file transitionprob_part3_problem1_DAVgrid2_thetaBGL.csv has the state transition probabilities for the state Gravity=1, Region = 1, DAVgrid = 2, Violation=1, Ordered Violator = 2, and Lag Investment = 1.

2) Computation of the steady state distribution

For $\hat{\theta}^{BGL}$, calculate the steady state distribution π, which is the $1 \times N$ probability of being in each state, $1, \ldots, N$ in the steady state. By definition, this distribution satisfies the following formula:

$$\pi = \pi P,$$

[2] We thank Chase Eck for his help in writing and debugging the code for Part 3 of the assignment.
[3] In compliance, none of the other state variables matter.

where P is the right transition matrix, implying that P_{ij} indicates the probability of transition from row i to column j.

To solve for π:

- Define

$$A = \begin{bmatrix} P^T - I_N \\ 1 \dots 1 \end{bmatrix},$$

where P^T indicates the transpose of P and I_N is the $N \times N$ identity matrix, and

$$b = (0, \dots, 0, 1),$$

which is N zeros followed by a 1.

- Using the identity $A'A\pi = A'b$, solve for π as:

$$\pi = (A'A)^{-1}A'b.$$

Since each $\widetilde{\Omega}_1$ can only transition to 161 different states it is computationally faster to calculate π for each $\widetilde{\Omega}_1$ separately.

Report the state probabilities for the all states with Gravity=1, Region = 1, DAVgrid = 4. The file steadystate_part3_problem2_DAVgrid2_thetaBGL.csv has the state transition probabilities for all states with Gravity=1, Region = 1, DAVgrid = 2.

3) Computation of the model moment inputs $m_k(\theta_j)$ across parameter grid values

The file parameter_grid_assignment.csv reports parameter grid values for 500 θ values (including $\hat{\theta}^{BGL}$, which is the first element). For each of these, calculate $m_k(\theta_j)$. Store these values as a matrix where each row is the moment value and each column is the θ value. There are two types of moments:

- The first set of moments is the probability of the steady state distribution that you reported in problem 2 for all states in $\widetilde{\Omega}_1$.
- The second set of moments is the probability of the steady state distribution times the probability of investment. This is for all states in $\widetilde{\Omega}_1$ except for the states that are in compliance.

Report the mean and standard deviation for each of the first 10 parameter values across all moments. The file modelmoments_part3_problem3_param2.csv contains the moments for second parameter vector in the file.

4) Computation of the data moment inputs m_k^d

Use the panel data in analysis_data.csv to calculate the data moments, m_k^d, for each set of moments.

- For the first set of moments calculate the empirical distribution of states in the data conditional on being at a given state . That is, for each $\widetilde{\Omega}_1$, calculate the frequency of each $\widetilde{\Omega}_2$ in the data across time and firms. See equation (A5) in the paper.
- For the second set of moments calculate the empirical distribution of states and investments in the data. See equation (A7) in the paper.

Report the mean and standard deviation of the moments. The file datamoments_part3_problem4_DAVgrid2.csv contains the data moments for DAVgrid = 2, Violation=1, Gravity=1, Region = 1, Ordered Violator = 2, and Lag Investment = 1.

5) Calculate parameter estimates

- The program file solveforweights.m contains the Matlab code to calculate the parameter estimates, which are the weights η_j.
- The program requires two input files:
    - moment_data_class_assignment_use.csv: you need to create this file. It should have 501 columns and 14,445 rows. The first column should indicate the data moments that you created in problem 4. The next 500 columns should indicate the model moments that you created in problem 3 for each of the 500 parameter values, in turn. Each of the 14,445 rows corresponds to one moment.
    - moment_variance_class_assignment_use.csv: we have created this file and include it. This indicates the variance/covariance matrix of the moments.
- The program creates an output file, weightparams_class_assignment.csv, that reports the parameter numbers with positive weights (numbered from 1-500) and the weight of those parameters.

Create the first input file and then run the solveforweights.m program. Generate a table with the six θ parameter values with the highest weights, and the weights, η, of these parameters, as in BGL.