ARE 212
Problem Set 3a, Second Half
Scott Kjorlien, Thor Larson, Sarah O'Brien, Bora Ozaltun, Garrison Schlauch

1. Exercises (GMM)

Part One

After describing the parameter space and set of moment conditions, we have a set of parameters $\beta \in B \subset \mathbb{R}^k$ that we are trying to estimate. We are given a set of ℓ moment conditions that are zero in expectation. The GMM estimation routine for an observed dataset $\{W_i\}_{i=1}^n$[1] solves the following objective:

$$\hat{b}_{GMM} = arg\min_{b \in B} E_n[g(W_i, b)]^T \; W \; E_n[g(W_i, b)],$$

where we are using the following conventions

- $g(W_i, b)$ is the moment function evaluated at b for the i^{th} observation. It is a $(\ell \times 1)$ vector.
- $E_n[g(W_i, b)] = \frac{1}{n}\sum_{i=1}^n g(W_i, b)$ is the empirical expectation. It is a $(\ell \times 1)$ vector.
- W is an $(\ell \times \ell)$ weighting matrix that is symmetric and positive semi-definite.

W is some weighting matrix of moment conditions that we usually do not know. The selection of W however impacts the efficiency properties of our estimator so we care about how we pick this matrix. The optimal weighting matrix from an efficiency point of view is the inverse of the moments' variance:

$$\Omega^{-1} = (Var(g(W_i, \beta)))^{-1} = (\mathbb{E}[g(W_i, \beta)g(W_i, \beta)]' - \mathbb{E}[g(W_i, \beta)]\mathbb{E}[g(W_i, \beta)]')^{-1}$$

Since we do not know the true value of β, the optimal weighting matrix is not directly implementable. However, one can estimate the optimal weighting matrix through a two-step GMM process:

- Step One: Estimate GMM using an arbitrary weighting matrix (for example $\hat{W}_0 = I_\ell$) to obtain consistent estimates for β: $\hat{b}_0$.
- Step Two: Use parameters obtained in step one to estimate $\hat{\Omega}^{-1}$ and re-estimate GMM using the estimated efficient weighting matrix

Finally, we can construct the asymptotic variance of our parameters using the equation $avar(\beta) = (G'\Omega^{-1}G)/N$ and replacing the terms with their sample analogs. Note that $G = \mathbb{E}[\frac{\partial g(W_i,\beta)}{\partial \beta}]$, an $(\ell \times k)$ Jacobian of our moments. This asymptotic variance of our parameters is derived using the optimal weighting matrix discussed above.

1. W_i here can be (X_i, y_i, Z_i).

Part Two

For this section, the question asks us to analyze different properties of the models that are listed. Particularly, we are asked to provide: (1) causal diagrams, (2) construct optimally weighted GMM estimators for the unknown parameters, and (3) give an estimator for the covariance matrix of our estimates. We provide causal diagrams in the jupyter notebook associated with this question. The notebook can be found in `PS3/Q1/Question 1.ipynb`. In order to construct optimally weighted GMM estimators, the main differentiator between the different models below is determining the moments, $g(W_i, b)$, for each model. Once these moments are determined, the process of constructing the optimally weighted GMM estimator is repetetive and the details can be found in the previous part. Finally, getting the covariance matrix is also a repetitive process that was detailed in the previous part.

(a) This setup describes properties for a single variable y. The moment conditions for this problem can be written as:

$$g(y_i, (\mu, \sigma^2)) = \begin{bmatrix} y_i - \mu \\ (y_i - \mu)^2 - \sigma^2 \\ (y_i - \mu)^3 \end{bmatrix}$$

The two step estimation procedure described in the previous section for the optimal weighting matrix can be calculated using the moments. An estimate of the variance can be constructed after deriving the Jacobian:

$$G(y_i, (\mu, \sigma^2)) = \begin{bmatrix} -1 & 0 \\ -2(y_i - \mu) & -2\sigma \\ -3(y_i - \mu)^2 & 0 \end{bmatrix}$$

where μ, σ are the parameters estimated from the GMM using a optimal weighting matrix.

(b) The setup is analogous to the linear model. For ease of notation, consider the parameter $\gamma = [\alpha \ \beta]'$ and $w_i = [1 \ x_i']'$. We have three moments for this setup with the moment condition:

$$g((y_i, w_i), \gamma) = \begin{bmatrix} w_i(y_i - w_i'\gamma) \end{bmatrix}$$

and Jacobian:

$$G((y_i, w_i), \gamma) = -w_i w_i'$$

(c) Assuming the goal is to estimate (α, β, σ) jointly, we add an additional moment to part (b):

$$g((y_i, w_i), (\gamma, \sigma)) = \begin{bmatrix} w_i(y_i - w_i'\gamma) \\ (y_i - w_i'\gamma)^2 - \sigma^2 \end{bmatrix}$$

where as in the previous section: $\gamma = [\alpha \ \beta]'$ and $w_i = [1 \ x_i']'$.

The Jacobian is:

$$G((y_i, w_i), (\gamma, \sigma)) = \begin{bmatrix} -w_i w_i' & 0 \\ 2(y_i - w_i'\gamma)w_i' & -2\sigma \end{bmatrix}$$

(d) This problem introduces heteroskedasticity into the model through $\mathbb{E}[u^2|X] = e^{X\sigma}$.[2] Our moments for this problem are similar to part (c), but with the change to the last moment:

$$g((y_i, w_i), (\gamma, \sigma)) = \begin{bmatrix} w_i(y_i - w_i'\gamma) \\ (y_i - w_i'\gamma)^2 - e^{x_i\sigma} \end{bmatrix}$$

where as in the previous section: $\gamma = [\alpha\ \beta]'$ and $w_i = [1\ x_i]'$.

The Jacobian is:

$$G((y_i, w_i), (\gamma, \sigma)) = \begin{bmatrix} -w_i w_i' & 0 \\ 2(y_i - w_i'\gamma)w_i' & -x_i e^{x_i\sigma} \end{bmatrix}$$

(e) In this model, we have a new set of r.v. that are analogous to instruments: Z. The problem setup means that we have the dimension of Z equal to the dimension of X (since we have $\mathbb{E}[Z'X] = Q$). *Critically, we assume that Q has full column rank within some neighborhood of the true γ.* The moment conditions that the model gives us are:

$$g((y_i, w_i), \gamma) = \begin{bmatrix} z_i(y_i - w_i'\gamma) \\ y_i - w_i'\gamma \end{bmatrix}$$

where as in the previous section: $\gamma = [\alpha\ \beta]'$ and $w_i = [1\ x_i]'$.

The Jacobian is:

$$G((y_i, w_i), \gamma) = \begin{bmatrix} -z_i w_i' \\ -w_i' \end{bmatrix} = -\begin{bmatrix} z_i \\ 1 \end{bmatrix} w_i'$$

(f) We move to a new model for this part, particularly, a more general linear model[3]. *Additionally, we assume that Q(b) has full column rank within some neighborhood of the true β and the f is continuously differntiable.*.We can now write our moments as:

$$g((y_i, w_i), \beta) = \begin{bmatrix} z_i(y_i - f(x_i'\beta)) \\ y_i - f(x_i'\beta) \end{bmatrix}$$

The Jacobian is:

$$G((y_i, w_i), \beta) = -\begin{bmatrix} z_i \\ 1 \end{bmatrix} x_i' \cdot f'(x_i'\beta)$$

Lets discuss the last restriction. For simplicity, lets denote $\frac{\partial f(x_i'\beta)}{\partial \beta'} = \nabla_\beta f(x_i\beta)$. Notice that $\nabla_\beta f(x_i'\beta) = x_i f'(x_i'\beta)$. Our final moment is an analogous restriction to the final moment in part (e). Particularly, they are both restrictions on the first derivative of our first moment: $\frac{\partial z_i(y_i - f(x_i'\beta))}{\partial \beta'} = -(z_i'\nabla_\beta f(x_i'\beta)) = -Q(\beta)$.

2. Note that we are assuming for this problem that X is uni-variate even though it has a transpose in the problem definition. If not, the heteroskedasticity doesn't make sense since u is univariate and so is σ (if not the dimensions don't work).

3. Relationship between β and X is still linear.

(g) We now generalize the model to the non-linear model where X and β can interact non-linearly. We retain the assumption that f is continously differentiable. Our moments are:

$$g((y_i, w_i), \beta) = \begin{bmatrix} z_i(y_i - f(x_i, \beta)) \\ y_i - f(x_i, \beta) \end{bmatrix}$$

The Jacobian is:

$$G((y_i, w_i), \beta) = - \begin{bmatrix} z_i \\ 1 \end{bmatrix} f_\beta(x_i, \beta)$$

(h) Our final model has a transformation on y. Lets write out the moments:

$$g((y_i, w_i), (\alpha, \gamma)) = \begin{bmatrix} z_i(y_i^\gamma - \alpha) \\ y_i^\gamma - \alpha \end{bmatrix}$$

The Jacobian is:

$$g((y_i, w_i), (\alpha, \gamma)) = \begin{bmatrix} z_i y_i^\gamma \log(y_i) & -z_i \\ y_i^\gamma \log(y_i) & -1 \end{bmatrix} = \begin{bmatrix} z_i \\ 1 \end{bmatrix} [y_i^\gamma \log(y_i) \ \ -1]$$

Part Three

For each model detailed in the previous part, we've constructed a dgp process. The code for the dgp process can be found in `PS3/Q1/dgp.py` and examples of the dgp process being called can be found in the notebook `PS3/Q1/Question 1.ipynb`. Additionally, for each model, we estimate the model using two step GMM once in the same notebook to find the estimation routine working well. Below are some details of the dgp:

(a) $y \sim N(\mu, \sigma^2)$.

(b) $u \sim N(0, 1)$, $X \sim N(0, I_k)$.

(c) $u \sim N(0, \sigma^2)$, $X \sim N(0, I_k)$.

(d) $u|X \sim N(0, expX\sigma)$, $X \sim N(0, I)$.

(e) $Cov(X, Z, u)$ is such that $Cov(Z, u) = Cov(u, Z) = 0$, $X, Z, u \sim N(0, Cov(X, Z, u))$.

(f) $Cov(X, Z, u)$ is such that $Cov(Z, u) = Cov(u, Z) = 0$, $X, Z, u \sim N(0, Cov(X, Z, u))$. $f(a) = e^a$.

(g) $Cov(X, Z, u)$ is such that $Cov(Z, u) = Cov(u, Z) = 0$, $X, Z, u \sim N(0, Cov(X, Z, u))$. $f(a, b) = a^2 * b^2$.

(h) For this part, we created a dgp for a specific α, γ. Specifically, $\alpha = 1$ and $\gamma = 3$. Let $v \sim Uniform[-1, 1]$, $u = v^\gamma$, $z = v^{\gamma-1}$, $y^\gamma = \alpha + u$.

Part Four

For this section, we decided to write a new GMM estimation routine based on the code provided but using a different package: `PyTorch`. Our reasons for writing this code in a new framework were: (1) good exercise, (2) PyTorch provides automatic differentiation, making analytical Jacobians and Hessians trivial to calculate (helpful especially for non-linear models), and (3) computational efficiency that could be provided if a GPU were available.

The results can be found at the end of the noteboook `PS3/Q1/Question 1.ipynb`. We can see that as n increases our parameter estimates are behaving as we would expect them to for consistency to be true as well as the limiting distribution.

2. Breusch-Pagan Extended

Part One

Assumption (a.i) is untestable since we can never observe the true error term u.

Part Two

Homoskedasticity assumes that the variance of the error is a constant, or that $Var[u|x] = \sigma^2$. The residuals from an OLS regression are constructed to be mean zero, so if error terms are homoskedastic the variance should equal the average square of the residuals, and would be constant for any value of x. On the other hand, heteroskedasicity allows for the variance of disturbances to vary with x, $Var[u|x] = \sigma^2 f(x)$. The Breusch-Pagan test essentially tests the null hypothesis that the squared residuals of the OLS are constant, which would be evidence for constant homoskedastic errors. The alternative, that the independent variables have a significant relationship with the squared residuals, is evidence that the variance of the error does indeed depend on x.

Part Three

Note: we are interpreting this question as follows. In class, we discussed using GMM to construct an overidentification test for testing whether the moment restrictions hold (discussed below). Thus, we contructed a GMM version of the Breusch-Pagan test that adds an additional moment condition such that we have more moment conditions than parameters we are trying to estimate.

The GMM population moment conditions a.i and a.ii are

$$\begin{aligned} \mathbb{E}[u] &= 0 \\ \mathbb{E}[ux] &= 0 \\ \mathbb{E}[u^2 - \sigma^2] &= 0 \end{aligned}$$

Moreover, $\mathbb{E}[u^2 - \sigma^2|x] = 0$ implies $\mathbb{E}[h(x)(u^2 - \sigma^2)] = 0$ for any function h, in particular, $h(x) = x$:

$$\mathbb{E}[x'(u^2 - \sigma^2)] = 0$$

Using the analogy principle, we can construct sample moment analogs:

$$g_N(x, \alpha, \beta, \sigma) = \frac{1}{n} \sum_{i=1}^{n} \begin{bmatrix} y_i - \alpha - x_i'\beta \\ x_i(y_i - \alpha - x_i'\beta) \\ (y_i - \alpha - x_i'\beta)^2 - \sigma^2 \\ x_i'[(y_i - \alpha - x_i'\beta)^2 - \sigma^2] \end{bmatrix}$$

We can then estimate the model as described in Exercise 1 Part 1 using two-step GMM to obtain consistent and efficient estimators of the parameters α, β, and σ^2.

Finally, we can use a J-test to test the null hypothesis that our four sample moment conditions equal zero. A significant J-test statistic would be evidence for heteroskedasticity, where the J statistic is

$$J = g_N(x, \alpha, \beta, \sigma)' \hat{\Omega}^{-1} g_N(x, \alpha, \beta, \sigma)$$

Under the null of homoskedasticity with the above moment restrictions, $J \sim \chi^2_1$.

Part Four

Breusch-Pagan test

Merits:

- Conceptually simpler
- Less computationally intensive

Drawbacks:

- Sensitive to outliers
- Only tests for heteroskedasticity as a linear function of x
- Assumes i.i.d. normally distributed error terms

GMM alternative

Merits:

- More flexible in that it can test for heteroskedasticity as a non-linear function of x
- Can account for serially correlated errors with the weighting matrix

Drawbacks:

- Conceptually more complicated
- More computationally intensive
- Requires more assumptions, namely
  - The empirical moments obey the law of large numbers
  - The derivatives of the empirical moments converge in probability
  - The empirical moments obey the central limit theorem
  - The parameters are identified
  - The weighting matrix converges in probability to a finite symmetric positive definite matrix

Part Five

The Breusch-Pagan test will fail to reject the null hypothesis of homoskedasticity here. This is because the expectation of $\sigma^2 \sin(2x)$ over the interval $[0, 2\pi]$ is 0.

Our GMM test from part 3 would perform better in this case since it's a more general test. In particular, $\mathbb{E}[x \sin(2x)] \neq 0$ over this interval.

Part Six

- We can use the same regression we use to estimate the residuals (equation 1 in the problem).
- In the auxiliary regression, we could regress the squared residuals on a non-linear function of x, such as a higher-order polynomial in x. This is similar to the White test, although we could consider additional functions of x beyond a quadratic. Graphing the squared residuals against x could help us determine what functional specifications we should consider.
- We could then use an F-test or LM test to test the null hypothesis that the coefficients on all of the terms involving transformations of x equal zero.
- This method is still limited in that it requires specifying the functional form of the potential heteroskedasticity.

Part Seven

If our ideas about estimating $f(x)$ are correct, we could add more moment conditions in constructing our over-identified GMM estimator, which would improve its efficiency. However, our GMM estimator would still be less efficient than the optimal GLS estimator since GLS uses the true $f(x)$ to construct its weighting matrix.

3. Tests of Normality

Part One

Using the analogy principal, we can construct estimators for the first k moments of the distribution x by replacing the expectation with the sample mean. The stacked sample moments are collected as follows (if k odd). Under the null hypothesis of normality, the following moments hold:

$$g_N(x, \mu, \sigma) = \frac{1}{n} \sum_{i=1}^{n} \begin{bmatrix} x_i - \mu \\ (x_i - \mu)^2 - \hat{\sigma}^2 \\ (x_i - \mu)^3 \\ (x_i - \mu)^4 - 3\sigma^4 \\ \dots \\ (x_i - \mu)^k \end{bmatrix}$$

Part Two

Under the null, we have that $\mathbb{E}g_i(\mu, \sigma) = 0$, so we can estimate the sample moment covariance as:

$$\hat{\Omega} = \frac{1}{n} \sum_{i=1}^{n} g_i(x_i, \mu, \sigma) g_i(x_i, \mu, \sigma)'$$

Where g_i are the individual moment conditions.

Part Three

A GMM-based test of normality essentially tests the null that our objective function is equal to zero, where the objective function is given by

$$J = g_N(x, \mu, \sigma)' \hat{\Omega}^{-1} g_N(x, \mu, \sigma)$$

Under the null of normality, $J \sim \chi^2_{k-2}$, where k is the number of moment restrictions. Thus, we can compare our estimate of J to the critical value for a chi-squared distribution with $k-2$ degrees of freedom to test the null that the moment-restrictions hold, which would give evidence for normally distributed data.

Part Four

This has been done on the jupyter notebook for question 3. I generate two samples, one from a normal distribution and one from the uniform distribution. I reject the null of normality with K = 5 for the uniformly distributed dgp and fail to reject for the normal dgp with a test size of 0.05.

Part Five

The optimal choice of K should lead to a test that does not reject the null of normality when the data is normally distributed but does reject the null for $x_1, ..., x_N$ drawn from a non-normal distribution. As shown in the plots in the jupyter notebook, a value of K that is too small (less than 4 in this example) does not distinguish between a normal and uniform distribution with the same mean. With too few moment restrictions there are many distributions which may satisfy them. On the other hand, as k grows there is an increasing chance that sample variation may lead even normally distributed data to fail higher moment conditions. The optimal K must balance these two concerns. In the simulations done in the jupyter notebook we find that k between 4 and 7 balance these two concerns when comparing data drawn from a Uniform[-1,1] and Normal(0,1) data generating process. Another metric to base the optimal choice of K is the actual parameter estimates. As the number of moment conditions become significantly larger than the number of parameters, the parameter estimates begin to diverge from their true values as depicted in the graph in the jupyter notebook.

Part Six

The GMM estimates of (μ, σ) are nearly identical to the MLE estimates when K = 2 (the just identified case). The moment conditions used in GMM are derived from the likelihood function in MLE, so it is unsurprising that the estimates essentially coincide. As K increases and we are overidentified, the estimates begin to diverge as the higher moment conditions are no longer part of the maximization problem for MLE. We also note that as n increases the parameter estimates are growing closer and closer together as they are both consistent for the true parameters.

4. Logit

1

First, show that in this model, $E(Y - \sigma(X\beta)|X) = 0$

$$\begin{aligned} E(Y - \sigma(X\beta)|X) &= E(Y|X) - E(\sigma(X\beta)|X) \\ &= P(Y = 1|X) - \sigma(X\beta) \\ &= P(Y = 1|X) - P(Y = 1|X) \\ &= 0 \end{aligned}$$

Our population moment condition is $E[X_i(y_i - \sigma(\beta' X_i))] = 0$. Define $g_i(\beta) = X_i(y_i - \sigma(\beta' X_i))$. Now define the sample analogue:

$$\bar{g}_N(\beta) = \frac{1}{N} \sum_{i=1}^{N} X_i(y_i - \sigma(\beta' X_i))$$

This is a just-identified setting where $l = k$.

2

A single Bernoulli with parameter p and $k \in \{0, 1\}$ has a pmf of $p^k(1-p)^{1-k}$. Here, y is an $N-$vector of Bernoullis, $p = P(Y = 1|X) = \sigma(\beta' X)$, so the distribution for a single y_i we have

$$f(y_i|X, \beta) = \sigma(\beta' X_i)^{y_i}(1 - \sigma(\beta' X_i))^{1-y_i}$$

and by independence of the N observations, we can derive our likelihood function by taking the product of the pmf of all y_i.

$$\mathcal{L}(\beta|y, X) = \prod_{i=1}^{N} \sigma(\beta' X_i)^{y_i}(1 - \sigma(\beta' X_i)^{1-y_i}$$

3

Set up the problem as:

$$\max_{b \in \mathbb{R}^k} \sum_{i=1}^{N} y_i \ln \sigma(b' X_i) + (1 - y_i) \ln (1 - \sigma(b' X_i))$$

Note that $\sigma'(z) = \sigma(z)(1 - \sigma(z))$, and with $z_i = b' X_i$, $\frac{dz_i}{db_j} = x_{ij}$.

Taking the FOC, the jth first order condition (for $j \in 1 \dots k$)

$$\sum_{i=1}^{N} y_i \frac{1}{\sigma(z_i)} \frac{d\sigma}{db_j} - (1 - y_i) \left(\frac{1}{1 - \sigma(z_i)} \right) \frac{d\sigma}{db_j} = 0$$

$$\sum_{i=1}^{N} y_i \frac{\sigma(z_i)(1 - \sigma(z_i)}{\sigma(z_i)} \frac{dz_i}{db_j} - (1 - y_i) \left(\frac{\sigma(z_i)(1 - \sigma(z_i)}{1 - \sigma(z_i)} \right) \frac{dz_i}{db_j} = 0$$

$$\sum_{i=1}^{N} x_{ij} \left[y_i(1 - \sigma(z_i)) - (1 - y_i)\sigma(z_i) \right] = 0$$

$$\sum_{i=1}^{N} x_{ij} \left[y_i - y_i\sigma(z_i) - \sigma(z_i) + \sigma(z_i)y_i) \right] = 0$$

$$\sum_{i=1}^{N} x_{ij} \left[y_i - \sigma(b'X_i) \right] = 0$$

If we scale by $\frac{1}{N}$, We can represent all j of our sample moment conditions as:

$$\bar{g}_N(\beta) = \frac{1}{N} \sum_{i=1}^{N} X_i(y_i - \sigma(\beta' X_i))$$

which is identical to the GMM estimator from the first part of this question. Since both are just-identified, the efficiency of the estimator does not depend on a weighting matrix W, so these estimators are equally efficient.