Problem Set 2 Solutions: Estimation frameworks

Thomas Mikaelsen*
Econometrics I

February 2, 2024

Problem 1

Extraordinary Least Squares. Consider the class of linear models: $Y = X\beta+\varepsilon$ where Y is an $n\times 1$ vector of outcomes, X is an $n\times k$ matrix of characteristics, β is a $k\times 1$ parameter vector, and ε is an $n\times 1$ vector of innovations or errors. Throughout this problem, A' denotes the transpose of some matrix A.

1. *Least squares.* Derive the $\hat{\beta}$ that minimizes the sum of squared errors: $\varepsilon'\varepsilon$.

2. *MLE.* Assume each error ε_i is mean zero iid normally distributed: $\varepsilon_i \sim N(0, \sigma^2)$. Find the maximum likelihood estimate of β.

3. *Bayesian.* Let $k = 1$. Assume each error ε_i is mean zero iid normally distributed: $\varepsilon_i \sim N(0, \sigma^2)$, where we assume that σ^2 is known (for simplicity). Your prior is that $\beta \sim N(\theta, \tau^2)$. Find the mean of the posterior distribution of β.

4. *Bayesian.* Assume each error ε_i is mean zero iid normally distributed: $\varepsilon_i \sim N(0, \sigma^2)$, where we assume that σ^2 is known (for simplicity). You don't want your prior to unduly influence your results, so you assume a uniform prior $\beta \sim U[-a, a]$ and take the limit as $a \to \infty$. This means

*Email: thomas.mikaelsen@ne.su.se, Office: A664

that your prior considers any value between $-a$ and a to be equally plausible, and you are taking the limit as this includes all possible values. Find the limit of the mode of the posterior distribution of β.

5. *GMM.* Assume that the errors are uncorrelated with the regressors (equivalently, all regressors are exogenous): $E(X'\varepsilon) = 0$. Derive the GMM estimate of β.

1.1: We have

$$\begin{aligned}\epsilon'\epsilon &= (Y - X\beta)'(Y - X\beta) = Y'Y - Y'X\beta - (X\beta)'Y + (X\beta)'X\beta \\ &= Y'Y - Y'X\beta - \beta'X'Y + \beta'X'X\beta \\ &= Y'Y - 2\beta'X'Y + \beta'X'X\beta\end{aligned}$$

To minimize we would like to impose first order conditions. For that to work, we need the expression to be convex in β. Since the expression is infinitely differentiable (it's a polynomium in β), we can use the second-order test. So let's take the derivatives. Bruce Hansen Theorem A.6 (p. 994) provides the properties of Matrix Calculus that we need (properties 1 and 3, specifically)

$$\begin{aligned}\frac{\partial \epsilon'\epsilon}{\partial \beta} &= -2X'Y + (X'X + (X'X)')\beta = -2X'Y + 2X'X\beta \\ \frac{\partial^2 \epsilon'\epsilon}{\partial \beta^2} &= 2X'X\end{aligned}$$

The expression $2X'X = A$ is a symmetric matrix and any such matrix induces a quadratic form $z \mapsto z'Az \in \mathbb{R}$. Similarly to the unidimensional case, the first order condition identifies a unique minimum if and only if the second derivative is strictly positive, which in the case of matrices means that A is positive definite ($z'Az > 0$ for all $z \in \mathbb{R}^k$). A quadratic form induced by a symmetric matrix $A = X'X$ is positive definite, in turn, if and only the columns of X are linearly independent. So, if the columns of X are linearly independent, we can impose first-order conditions and isolate β. We get

$$\begin{aligned}-2X'Y + 2X'X\beta &= 0 \\ \Rightarrow \quad \beta &= (X'X)^{-1}X'Y\end{aligned}$$

Notice that linear independence of the columns of X also implies that $(X'X)$ is invertible, ensuring that the expression above is well-defined.

1.2: The assumptions imply that, conditional on X, Y follows a Normal distribution

$$\begin{aligned} & Y \mid X = x \sim N(x'\beta, \sigma^2) \\ \Rightarrow \quad & f_{Y|X=x}(y) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp(-\frac{1}{2\sigma}(y - x'\beta)^2) \end{aligned}$$

Where the mean is equal to $x'\beta$ because for any normal $N \sim N(\mu, \sigma^2)$, we have $a + bN \sim N(a + b\mu, b^2\sigma^2)$ for constants a, b. So for a set of observations $(y, x) = \{(y_i, x_i)\}, i \in \{1, ..., n\}\}$ we can form the likelihood function

$$\begin{aligned} L(y, x, \beta) &= \prod_{i=1}^{n} f_{Y|X=x_i}(y_i) \\ \Rightarrow \quad l(y, x, \beta) &= \sum_{i=1}^{n} \log f_{Y|X=x_i}(y_i) \\ &= -n \log(\sqrt{2\pi\sigma^2}) - \frac{1}{2\sigma} \sum_{i=1}^{n} (y_i - x_i'\beta)^2 \\ &= \text{constant} - \frac{1}{2\sigma} \epsilon'\epsilon \end{aligned}$$

Notice that l only depends on β through $\epsilon'\epsilon$. So maximizing l with respect to β is equivalent to minimizing $\epsilon'\epsilon$ since it enters l with a minus. But we already did that in exercise (1.1) and showed that the β which minimizes $\epsilon'\epsilon$ is $\beta = (X'X)^{-1}X'Y$ and so we are done.

1.3: Recall the definition of the posterior distribution from slide 20 lecture 3

$$\begin{aligned} f(\beta \mid X_n, y_n) &= \frac{f(y_n \mid X_n, \beta) \cdot f(\beta)}{f(X_n, y_n)} \\ \Leftrightarrow \quad posterior &= \frac{likelihood \cdot prior}{normalization} \end{aligned}$$

We want to find the mean of the posterior distribution, so let's find its pdf by identifying the terms on the right hand side. We are told that $\epsilon \sim N(0, \sigma^2)$

which means that the distribution of Y conditional on X and β is $N(X\beta, \sigma^2)$ and thus

$$f(y \mid x, \beta) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp(-\frac{1}{2\sigma^2}(y - x'\beta)^2) \propto \exp(-\frac{1}{2\sigma^2}(y - x'\beta)^2)$$

$$\Rightarrow \quad L(y \mid x, \beta) \propto \prod_{i=1}^{n} \exp(-\frac{1}{2\sigma^2}(y_i - x_i'\beta)^2)$$

$$\Rightarrow \quad l(y \mid x, \beta) \propto -\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - x_i'\beta)^2$$

where the two last expressions are the likelihood and log-likelihood functions. We use the "proportional to" sign $\propto$ to avoid having to write terms that are constant with respect to β and the data. Next we are told the prior on β is $\beta \sim N(\theta, \tau^2 I_{k\times k})$, however for this first run-through I will assume we're

$$\begin{aligned} f(\beta) &= \frac{1}{\sqrt{(2\pi)^k|\tau^2 I|}}\exp(-\frac{1}{2}(\beta - \theta)'(\tau^2 I_{k\times k})^{-1}(\beta - \theta)) \\ &= \frac{1}{\sqrt{(2\pi)^k\tau^{2k}}}\exp(-\frac{1}{2}(\beta - \theta)'(\tau^{-2} I_{k\times k})(\beta - \theta)) \\ &= \frac{1}{\sqrt{(2\pi)^k\tau^{2k}}}\exp(-\frac{1}{2\tau^2}(\beta - \theta)'(\beta - \theta)) \\ &\propto \exp(-\frac{1}{2\tau^2}(\beta - \theta)'(\beta - \theta)) \end{aligned}$$

Where we used that the determinant of a diagonal matrix is the product of the diagonal entries, and we used that the inverse of a diagonal matrix is obtained by replacing the elements on the diagonal with their reciprocals. We find the implied posterior pdf

$$\begin{aligned} f(\beta \mid X, y) &\propto \prod_{i=1}^{n} \exp(-\frac{1}{2\sigma^2}(y_i - x_i'\beta)^2) \cdot \exp(-\frac{1}{2\tau^2}(\beta - \theta)'(\beta - \theta)) \\ &= \exp(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - x_i'\beta)^2 - \frac{1}{2\tau^2}(\beta - \theta)'(\beta - \theta)) \end{aligned}$$

For the following steps, let's assume that x_i is scalar. Then we get

$$f(\beta \mid X, y) \propto \exp(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - x_i\beta)^2 - \frac{1}{2\tau^2}(\beta - \theta)^2)$$

which is easier to work with. The strategy now is to realize that (i) the product of Gaussians is Gaussian, so $f(\beta \mid X, y)$ is the pdf of *some* Gaussian distribution – the question is (ii) *which* Gaussian distributions. To figure out (ii), we cleverly re-write it as something on the form

$$\exp(-\frac{1}{2\tau_n^2}(\beta - \theta_n)^2)$$

Then by appealing to (i) we can conclude immediately that the mean of the posterior distribution is θ_n. To achieve this re-writing, we will do something called completing the square. First, let's expand the brackets

$$f(\beta \mid X, y) \propto \exp(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i^2 + (x_i\beta)^2 - 2y_i x_i \beta) - \frac{1}{2\tau^2}(\beta^2 + \theta^2 - 2\beta\theta))$$

Let's collect terms based on β^2, β and everything else

$$f(\beta \mid X, y) \propto \exp(-\frac{\beta^2}{2}(\frac{\sum_i x_i^2}{\sigma^2} + \frac{1}{\tau^2}) + \beta(\frac{\sum_i x_i y_i}{\sigma^2} + \frac{\theta}{\tau^2}) - (\frac{\sum_i y_i^2}{2\sigma^2} + \frac{\theta^2}{2\tau^2})) \tag{1}$$

We now define τ_n^2 and θ_n to be whatever allows us to write equation (1) on the form

$$\begin{aligned}(1) &=: \exp(-\frac{1}{2\tau_n^2}(\beta^2 - 2\beta\theta_n + \theta_n^2)) \qquad (2)\\ &= \exp(-\frac{1}{2\tau_n^2}(\beta - \theta_n)^2)\end{aligned}$$

So let's find out what (2) implies about θ_n, τ_n^2 by matching the coefficients between (1) and (2). Matching the coefficients on β^2, we have

$$\begin{aligned}-\frac{\beta^2}{2\tau_n^2} &= -\frac{\beta^2}{2}(\frac{\sum_i x_i^2}{\sigma^2} + \frac{1}{\tau^2})\\ \Rightarrow \quad \frac{1}{\tau_n^2} &= (\frac{\sum_i x_i^2}{\sigma^2} + \frac{1}{\tau^2})\\ \Rightarrow \quad \tau_n^2 &= \frac{\sigma^2\tau^2}{\tau^2\sum_i x_i^2 + \sigma^2}\end{aligned}$$

Matching the β coefficients, we have

$$\begin{aligned}
\frac{\beta\theta_n}{\tau_n^2} &= \beta(\frac{\sum_i x_i y_i}{\sigma^2} + \frac{\theta}{\tau^2}) \\
\Rightarrow \quad \theta_n &= \tau_n^2(\frac{\sum_i x_i y_i}{\sigma^2} + \frac{\theta}{\tau^2}) \\
&= \frac{\sigma^2\tau^2}{\tau^2\sum_i x_i^2 + \sigma^2}(\frac{\sum_i x_i y_i}{\sigma^2} + \frac{\theta}{\tau^2}) \\
&= \frac{\tau^2\sum_i x_i y_i + \sigma^2\theta}{\tau^2\sum_i x_i^2 + \sigma^2}
\end{aligned}$$

which is the mean of the posterior distribution, as was argued previously, which is what we wanted. We have thus found the mean solely as a function of parameters we know, i.e. τ^2, σ^2 and θ as well as things we observe, i.e. x_i, y_i.

1.4: Now the prior has changed and $\beta \sim U[-a, a]$ implies $f(\beta) = \mathbb{1}_{\{\beta\in[-a,a]\}} \cdot \frac{1}{2a}$ but otherwise, the remaining ingredients to the posterior distribution are unchanged. The posterior pdf therefore is

$$\begin{aligned}
f(\beta \mid X, y) &= \frac{f(X, y \mid \beta, \sigma^2)f(\beta)}{\int_{-\infty}^{\infty} f(X, y \mid \beta^*, \sigma^2)f(\beta^*)d\beta^*} \\
&= \begin{cases} 0, & \text{if } \beta \neq [-a, a] \\ \frac{f(X,y|\beta,\sigma^2)\frac{1}{2a}}{\int_{-a}^{a} f(x,y|\beta^*,\sigma^2)\frac{1}{2a}d\beta^*}, & \text{otherwise} \end{cases}
\end{aligned}$$

We're asked to find the mode of the posterior, that is, the most common value or the maximum, in expectation, of the posterior distribution. If β is outside $[-a, a]$, the posterior is 0, so we won't find the maximum there – since the posterior is non-negative on at least some of $[-a, a]$. So we can restrict our attention to maximizing the second term with respect to β. For that, notice that the denominator is independent of β, because β is integrated out, as is $\frac{1}{2a}$ and so the only thing that depends on β is $f(X, y \mid \beta, \sigma^2)$. But this is just the likelihood function from question 1.b. Hence, the mode of the posterior is the MLE we found in (b), namely $\hat{\beta} = (X'X)^{-1}X'Y$.

1.5: We have

$$\begin{aligned} & E[X'\epsilon] = E[X'(Y - X\beta)] = E[X'Y] - E[X'X\beta] = 0 \\ \Rightarrow \quad & \beta E[X'X] = E[X'Y] \\ \Rightarrow \quad & \beta_{GMM} = (E[X'X])^{-1}E[X'Y] \end{aligned}$$

if $E[X'X]$ is invertible, or equivalently, the columns are linearly independent. The sample, or plug-in, estimator of β_{GMM} is

$$\begin{aligned} \hat{\beta}_{GMM} &= (\frac{1}{n}\sum_{i=1}^{n} X_i X_i')^{-1} \sum_{i=1}^{n} X_i Y_i \\ &= (\sum_{i=1}^{n} X_i X_i') \sum_{i=1}^{n} X_i Y_i \\ &= (X'X)^{-1}X'Y \\ &= \hat{\beta}_{OLS} \end{aligned}$$

So we have derived the OLS estimator using least squares, maximum likelihood (twice) and GMM.

Problem 2.a

Actually using MLE. You are interested in generating a teacher-level estimate of the propensity to inflate grades.[1] All students are given an identical test. Based on the test design, the number of points a student receives is known to be distributed according to a Poisson distribution with parameter λ. Assume λ is known because of special test-design-magic. If a student gets a score of 20 or more, s/he will pass the exam and graduate. Grading is done centrally (i.e., not by the students' teacher), but upon seeing the grades, the teacher can appeal to have the exam re-reviewed by another grader. This re-review process is known to be generous: In expectation, it increases the score received on the exam, although it varies by how much. Assume that the number of additional points the student receives, conditional on a re-review, is drawn from $\{0, 1\}$ with equal probabilities of each (i.e., a discrete

[1] This question was inspired by Diamond and Persson (2016), "The long-term consequences of teacher discretion in grading of high-stakes tests", *NBER Working Paper 22207*.

version of a uniform distribution). Assume that *i*) no teacher would bother appealing for a re-review of an exam with fewer than 19 points (since the student would never pass) or more than 20 points (since the student has already passed), and *ii*) a teacher realizes that if she appeals every exam with 19 points, s/he will get in trouble for gaming the system, and so s/he randomizes and only sends p share of those exams out for a re-review. You are interested in estimating p for each teacher, but you only observe the final grades (post re-review) and you don't observe how many or which exams were sent out for re-review.

1. Write the likelihood function for the problem.
2. Write the score.
3. Determine the maximum likelihood estimate $\hat{p}$.
4. Derive the Fisher Information Matrix for $\hat{p}$ (note it's a 1×1 "matrix" because there's only one parameter).
5. At what value of λ is the Fisher Information Matrix maximized? Why? Interpret what this means.

2.a(i): Let $f(y) = \frac{\lambda e^{-\lambda}}{y!}$ be the Poisson pmf, then the post-review pmf $f(p \mid y)$ is the following

$$f(y \mid p) = \begin{cases} f(y), & \text{if } y < 19 \\ (1-p)f(19) + \frac{p}{2}f(19), & \text{if } y = 19 \\ f(20) + \frac{p}{2}f(19), & \text{if } y = 20 \\ f(y), & \text{if } y > 20 \end{cases}$$

$$= \begin{cases} f(y), & \text{if } y < 19 \\ f(19) - \frac{p}{2}f(19), & \text{if } y = 19 \\ f(20) + \frac{p}{2}f(19), & \text{if } y = 20 \\ f(y), & \text{if } y > 20 \end{cases}$$

Why? The first equality is because: If the student gets $y < 19$, there's no reason to re-view. Same for $y > 20$. Now, how can $y = 19$ obtain? Either if the student gets 19, with probability $f(19)$, and the teacher doesn't

review, with probability $1-p$, or if the student gets 19, with probability $f(19)$, the teacher reviews, with probability p, but draws 0 extra points from the review, with probability 1/2. How can $y = 20$ obtain? If the student gets 20, with probability $f(20)$, or if the student gets 19, with probability $f(19)$, the teacher reviews, with probability p, and the review draws is 1, with probability 1/2. The second equality is just a re-writing to illustrate that there's excess mass at $y = 20$ and all of that excess comes from $y = 19$ scores that are being reviewed and get a review score of 1. Now we can calculate the likelihood functions

$$L(p \mid y) \propto \prod_{i=1}^{n} [f(19) - \frac{p}{2} f(19)]^{n_{19}} [f(20) + \frac{p}{2} f(19)]^{n_{20}}$$

$$\Rightarrow \quad l(p \mid y) = k + n_{19} \log(f(19) - \frac{p}{2} f(19)) + n_{20} \log(f(20) + \frac{p}{2} f(19))$$

2.a(ii): The score is the derivative of the log-likelihood with respect to the parameter p

$$\begin{aligned} s(p \mid y) = \frac{dl(p \mid y)}{dp} &= -\frac{n_{19} f(19)}{2f(19)(1 - \frac{p}{2})} + \frac{n_{20} f(19)}{2(f(20) + \frac{p}{2} f(19))} \\ &= -\frac{n_{19}}{2-p} + \frac{n_{20} f(19)}{2(f(20) + \frac{p}{2} f(19))} \end{aligned}$$

as we wanted.

2.a(iii): The MLE is found by setting the score equal to 0 and solving for p. I will skip the tedious algebra and go straight to the solution

$$0 = -\frac{n_{19}}{2-p} + \frac{n_{20} f(19)}{2(f(20) + \frac{p}{2} f(19))}$$

$$\Rightarrow \quad \hat{p} = 2[n_{20} - n_{19} \frac{f(20)}{f(19)}](n_{19} + n_{20})^{-1}$$

The interpretation of this number is the following: $f(20)/f(19)$ gives you the share of 20s relative to 19s you should see, absent manipulation, and so $n_{19} f(20)/f(19)$ tells you how many 20 point exams you should see, given that you saw n_{19} 19 point exams. Then $n_{20} - n_{19} \frac{f(20)}{f(19)}$ tells you the excess number of exams with 20 points you actually see. You multiply that by 2 because

only half of the randomly selected exams sent for regrading actually changes the score, so p should be double that, given the number of 20 point exams you observe. Because p is a probability, you divide it with the number of 19 and 20 point exams you actually observe. In one mouthful, $\hat{p}$ is two times the extra number of 20 point exams you observe, relative to how many you should have observed absent manipulation, divided by the total number of 19 and 20 point exams you observe. Even shorter: it's two times the excess share of observed 20 point exams, relative to how many there would have been absent manipulation.

2.a(iv): Let's calculate the Fisher Matrix by taking the negative expectation of the second order derivative of the log-likelihood. First notice that

$$\begin{aligned} f(20) &= \frac{\lambda^{20}e^{-\lambda}}{20!} = \frac{\lambda}{20}\frac{\lambda^{19}e^{-\lambda}}{19!} = \frac{\lambda}{20}f(19) \\ \Rightarrow \quad s(p \mid y) &= -\frac{n_{19}}{2-p} + \frac{10n_{20}}{10p+\lambda} \\ \Rightarrow \quad \frac{ds(p \mid y)}{dp} &= -\frac{n_{19}}{(2-p)^2} - \frac{100n_{20}}{(10p+\lambda)^2} \\ \Rightarrow \quad H_n &= -E[\frac{ds(p \mid y)}{dp}] \\ &= \frac{1}{(2-p)^2}E[n_{19}] + \frac{100}{(10p+\lambda)^2}E[n_{20}] \end{aligned}$$

which is strictly positive because all terms are positive. To calculate the two expectations, there's an intuitive way and a formal way. Both should give the same result. The intuitive way is to say: $E[n_{19}]$ is the number of 19 point exams we expect to observe. Since we know the probability of observing a 19 point exam, $f(19) - \frac{p}{2}f(19) =: p_{19}$, the expected number is the total number of exams we observe, n, times the probability of observing 19, hence

$$\begin{aligned} E[n_{19}] &= (f(19) - \frac{p}{2}f(19))n \\ E[n_{20}] &= (f(20) + \frac{p}{2}f(19))n \end{aligned}$$

The formal way is to note that n_{19} is the random variable that asks the question "If I draw n trials with probability p_{19} of success (observing a 19 point exam) how many successes do I draw?". This is a binomial distribution,

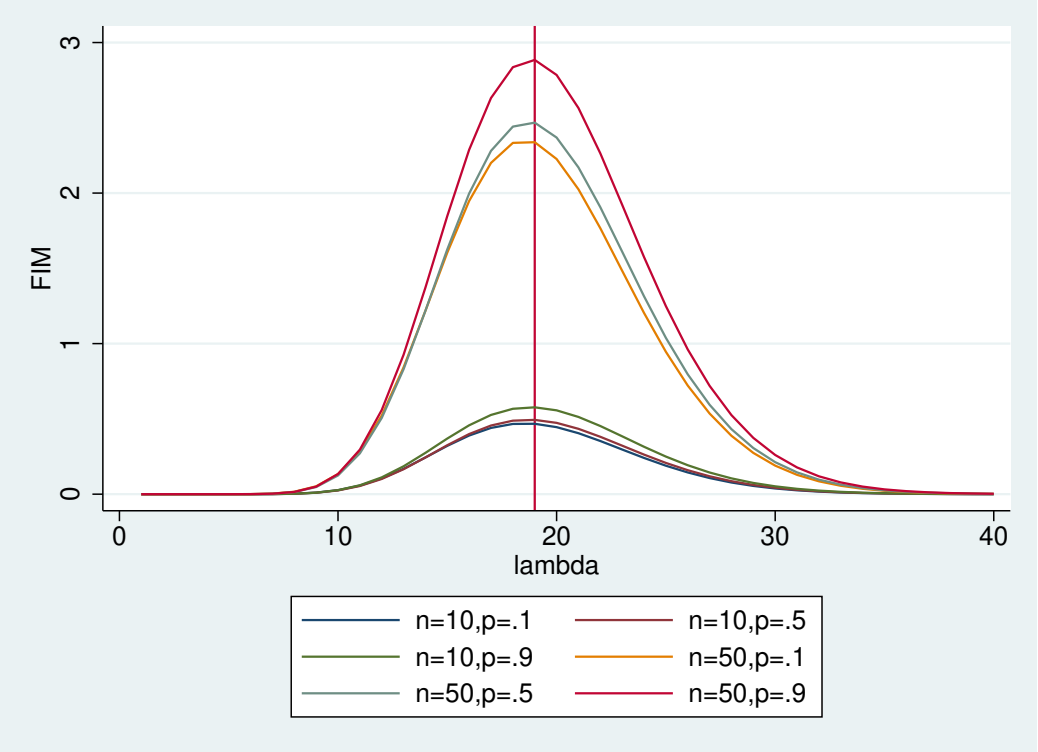
so $n_{19} \sim Bi(n, p_{19})$, and such a distribution has mean np_{19} – we did this calculation in the last TA session.

2.a(v): Plugging in the values for the expectation and using $f(20) = f(19)\lambda/20$ we get

$$H_n = \frac{n(1-\frac{p}{2})}{(2-p)^2}\frac{\lambda^{19}e^{-\lambda}}{19!} + \frac{100n}{(10p+\lambda)^2}(\frac{\lambda}{20} + \frac{p}{2})\frac{\lambda^{19}e^{-\lambda}}{19!}$$

It's not clear to me how to maximize this with respect to λ since we have λs in the denominator. However, we can give an intuitive explanation for why λ should be exactly 19: The FIM is a measure of how much information there is about the parameter of interest, p, in an observation. The only post-review scores that carry information about p are 19 and 20, and since λ is the mode of the distribution (the most often occurring observation), one could be tempted to think that λ should be set to 19.5 to maximize the number of 19 and 20 scores. However, that is not correct, because λ governs the underlying, pre-review score distribution; and from that point of view, only pre-review scores of 19 can put the review process in motion, thereby revealing information about p. A pre-review score of 20 will never be reviewed, so we would prefer that to have been a 19 instead – even though, from the post-review point of view, we cannot distinguish whether a post-review score of 20 came from an upgraded 19 or a pure 20 score – because the 19 has a chance of staying a 19 and that carries information about p as well. If we simulate FIM, we see that λ should indeed be exactly 19 which is the red line.

Figure 1: FIM simulation



Problem 2.b

You are interested in the parameters governing Calvo pricing. Calvo pricing asserts that firms cannot necessarily adjust their prices in response to changes in costs. Instead, in each period, with probability ϕ, a firm is randomly selected by the "Calvo fairy" to be allowed to change its prices. Note that by "random selection," I mean that being selected by the fairy is independent of other firm characteristics: It is a constant probability for all firms, regardless of prices or cost shocks. Assume that i) there is no aggregate inflation, only idiosyncratic firm-level cost shocks in which each firm, each period, draws a cost change from $N(0, \sigma^2)$, and ii) when selected, a firm sets its price equal to its marginal cost.[2] You observe only prices, not costs.

[2]Note that these two assumptions rule out a lot of the important economic issues around inflation expectations that macroeconomists are interested in with these models.

1. You are interested in estimating ϕ and σ^2. Why are those parameters interesting, from an economic perspective?

2. Assume that at time t, the Calvo fairy was feeling generous, and every firm was allowed to set price equal to marginal cost. At $t+1$, the fairy's behavior returned to normal and only a random ϕ share of firms were allowed to. You observe price changes at $t+1$. Write the likelihood function for the problem.

3. Assume that the Calvo fairy wasn't generous at time t, and that it has always been only a randomly selected ϕ share of firms who could change prices. Write the likelihood function for the problem.

4. Return to the likelihood function from ii (the one with the generous fairy, not iii with the mean one). Write the score.

5. Determine the maximum likelihood estimates $\hat{\phi}$ and $\hat{\sigma}^2$.

6. Derive the Fisher Information Matrix and check that it is positive semi-definite.

7. In our model, all of the costs are passed through to prices. Assume instead that only a share $\gamma \in [0, 1]$ is passed on (with γ being the same for all firms). Can you estimate ϕ, σ^2, and γ? Why or why not? In your answer, use a word that rhymes with mimentification.

2.b(i): σ^2 governs the variation in cost shocks that firms face. If firms face different costs, they have different optimal levels of employment, say. In a frictionless world, that wouldn't matter because workers just flow between firms, but if there are adjustment costs to switching employer, then this doesn't happen perfectly which leads to misallocation and that misallocation gets worse the more σ^2 increases. ϕ governs how cost shocks are passed through to consumers: pass-through is increasing in ϕ. This is important because if firms cannot adjust to cost shocks, that affects their equilibrium profits and so firms may have to exit, affecting the number of firms in the economy, which affects competition and therefore the potential gains from markets.

2.b(ii): Let $\Delta p_t = p_t - p_{t-1}$ and $\Delta c_t = c_t - c_{t-1}$. Then the assumption is that $\Delta c_t \sim N(0, \sigma^2)$ with pdf f_N. Consider how to think about the likelihood function: you observe a set of N price changes Δp_{t+1}. What's the likelihood of observing a given price change? If the price change is 0, it can only have come about if the firm wasn't allowed to adjust in $t+1$ because the probability of being allowed to adjust but drawing a cost shock of 0 is 0 by continuity of the normal distribution. What's the likelihood of observing a price change different from 0 in $t+1$? Well, the firm must have been allowed to adjust by the fairy, otherwise the price change would be 0, and the price change must equal the cost shock draw – because in t the firm was allowed to adjust so there is nothing "left over" to adjust from the previous period, hence all the change in price must be coming from the previous cost shock. So the likelihood of Δp_{t+1} equals the likelihood of Δc_{t+1} times ϕ in that case, and so we get the relevant pdf

$$f(\phi \mid \Delta p_{t+1}) = \begin{cases} 1-\phi, & \text{if } \Delta p_{t+1} = 0 \\ \phi f_N(\Delta c_{t+1} \mid \mu = 0, \sigma^2), & \text{otherwise} \end{cases}$$

$$\Rightarrow \quad L = (1-\phi)^{n_0} \phi^{n_1} \prod_{i \in N_1} f_N(\Delta c_{t+1} \mid \mu = 0, \sigma^2)$$

where N_1 is the set of firms that get to adjust, which contains n_1 firms, and N_0 is the set of firms that don't get to adjust, which contains n_0 firms.

2.b(iii): The probability of observing $\Delta p_t = 0$ is still the probability of not being allowed to adjust, $1 - \phi$. However, now, some firms might not have been allowed to adjust in $t-1$ so if $\Delta p_t \neq 0$ it will not necessarily equal Δc_t. Rather, it will equal the sum of all previous cost shocks since the last time the firm was allowed to adjust. Let $\tau \geq 0$ be the number of periods since a firm was last allowed to adjust, then we can think of a price change as the indicated sum of cost changes

$$\Delta p_t = \mathbb{1}_{\{\tau=0\}} \Delta c_t + \mathbb{1}_{\{\tau=1\}} (\Delta c_t + \Delta c_{t-1}) + ... + \mathbb{1}_{\{\tau=t\}} \sum_{j=1}^{t} \Delta c_j$$

We now need two ingredients to calculate the likelihood. First, for a given τ, what is the distribution, and hence the *pdf*, associated with this sum of cost shocks? For a given τ, the above equation is just a sum of τ iid normals $N(0, \sigma^2)$ which has distribution $N(0, \tau\sigma^2)$ and associated pdf $f_N(0, \tau\sigma^2)$.

Secondly, we need to calculate the probability that a given τ obtains. So what is the sequence of events that must have happened for a given τ to obtain? Well, $\tau+1$ periods ago the firm was allowed to adjust, which happened with probability ϕ. Then for the next τ periods, the firms was not allowed to adjust, which happened with probability $(1-\phi)^\tau$ and then in the current period, they were allowed to adjust again, which happened with probability ϕ. For a given Δp_t, we must take into account that this price change could have arisen from any τ between 0 and t, so we have to sum over each of these $\tau+1$ probabilities. Thus we have

$$f(\phi \mid \Delta p_t) = \begin{cases} 1-\phi, & \text{if } \tau = 1 \\ \phi \sum_{\tau=0}^{t} (1-\phi)^\tau \phi f_N(0, \tau\sigma^2) \end{cases}$$

$$\Rightarrow \quad L = (1-\phi)^{n_0} \phi^{2n_1} \prod_{i \in N_1} \sum_{\tau_i=0}^{t} (1-\phi)^{\tau_i} f_N(0, \tau_i \sigma^2)$$

Note that τ_i now depends on each firm i since each can have different τs. Notice also that the pmf of a Negative Binomial with $r=1$ is $(1-\phi)^\tau\phi$, so if you have phrased it in terms of a Negative Binomial that is fine as well.

2.b(iv): The log-likelihood is given by

$$l = n_0 \log(1-\phi) + n_1 \log(\phi) - n_1 \log(\sqrt{2\pi\sigma^2}) - \frac{1}{2\sigma^2} \sum_{i \in N_1} \Delta p_{i,t+1}^2$$

and the score is the derivative of the log-likelihood with respect to the parameters ϕ and σ^2

$$\frac{\partial l}{\partial \phi} = -\frac{n_0}{1-\phi} + \frac{n_1}{\phi}$$

$$\frac{\partial l}{\partial \sigma^2}[k - \frac{n_1}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\sum_{i \in N_1} \Delta p_{i,t+1}^2] = -\frac{n_1}{2\sigma^2} + \frac{1}{2(\sigma^2)^2} \sum_{i \in N_1} \Delta p_{i,t+1}^2$$

2.b(v): Set the score equal to 0 and solve for the parameters

$$\begin{aligned} 0 &= \frac{n_0}{1-\phi} = \frac{n_1}{\phi} \\ \Rightarrow \quad \phi n_0 &= (1-\phi) n_1 \\ \Rightarrow \quad \hat{\phi} &= \frac{n_1}{n_0 + n_1} \\ 0 &= -\frac{n_1}{2\sigma^2} + \frac{1}{2(\sigma^2)^2} \sum_{i \in N_1} \Delta p_{i,t+1}^2 \\ \Rightarrow \quad \hat{\sigma^2} &= \frac{1}{n_1} \sum_{i \in N_1} \Delta p_{i,t+1}^2 \end{aligned}$$

that is, the MLE estimate of the probability of a firm being allowed to adjust is the share of firms that adjust, which seems intuitive (to me, at least): The best estimate for the probability that a firm change change its price is the share of firms that change their prices. We do the same to get the variance estimate and see that it equals the sample variance of the price changes. This also makes sense: if we want to know about the variation in cost shocks, we only get information about that from the firms who adjust. For those firms, the best we can do is take the sample variance.

2.b(vi): The FIM is the negative of the second derivative of the log-likelihood, i.e. the Jacobian matrix of second order derivatives and cross-partials (which are 0)

$$\begin{aligned} -\frac{\partial^2 l}{\partial \phi^2} &= \frac{n_0}{(1-\phi)^2} + \frac{n_1}{\phi^2} \\ -\frac{\partial^2 l}{\partial (\sigma^2)^2} &= \frac{1}{(\sigma^2)^3} \sum_{i \in N_1} \Delta p_{i,t+1}^2 - \frac{n_1}{2(\sigma^2)^2} = \frac{n_1}{(\sigma^2)^3} \hat{\sigma^2} - \frac{n_1}{2(\sigma^2)^2} \end{aligned}$$

In the second derivative, if we plug in $\hat{\sigma^2} = \sigma^2$, we get

$$\frac{n_1}{(\sigma^2)^3} \hat{\sigma^2} - \frac{n_1}{2(\sigma^2)^2} = \frac{n_1}{\sigma^2}(1 - \frac{1}{2}) = \frac{n_1}{2\sigma^2}$$

Taking expectations, we use that ϕ and σ^2 are constants and so only n_0, n_1

have distributions

$$-E[\frac{\partial^2 l}{\partial \phi^2}] = (1-\phi)nE[\frac{1}{(1-\phi)^2}] + \phi nE[\frac{1}{\phi^2}] = \frac{n}{1-\phi} + \frac{n}{\phi}$$
$$-E[\frac{\partial^2 l}{\partial (\sigma^2)^2}] = \frac{n\phi}{2\sigma^2}$$

So we have a diagonal matrix, since the cross-partials are 0, and you might have learned that the eigenvalues of a diagonal matrix are the diagonal entries themselves. A square matrix is positive definite if all its eigenvalues are positive, so it is positive definite if and only if

$$\frac{n}{1-\phi} + \frac{n}{\phi} > 0 \quad \text{and}$$
$$\frac{n\phi}{2\sigma^2} > 0$$

Which is clearly true in both cases since all elements are positive.

2.b(vii): No, we can't estimate the three parameters because they are not identified. In particular $\Delta p_t = \gamma \Delta c_t$ along with $\Delta c_t \sim N(0, \sigma^2)$ implies that $\gamma \Delta p_t \sim N(0, \gamma^2 \sigma^2)$. If you go through the calculations again above, you will see that they are exactly the same, except where there before was σ^2 there now is $\gamma^2 \sigma^2$ and, in particular, nowhere is there any place where γ^2 and σ^2 appear separately. So the data doesn't allow us to distinguish the two parameters: for every set of data, there are infinite combinations of σ and γ that explains the data equally well.

Problem 3

Misspecified MLE. Consider a random sample $X_1, X_2, ..., X_n$ such that each X_i is iid distributed according to an exponential distribution: $X_i \sim exp(\lambda) \Leftrightarrow f(x) = 1 - e^{-\lambda x}$ if and only if $x > 0$ (and zero otherwise). Note that this implies that $E(x) = 1/\lambda$ and $Var(x) = 1/\lambda^2$. An analyst incorrectly assumes that the data comes from an iid normal distribution with mean μ and variance σ^2.

1. Derive the true MLE for λ assuming that you know the data are exponentially distributed. Show that this estimator is biased but consistent.

2. Let $\hat{\mu}$ denote the analyst's MLE estimator for the mean. Show that this is a consistent estimator for $E(x)$ despite the misspecification.

3. Let $\hat{\sigma}^2$ denote the analyst's MLE estimator for the variance. Show that this is a consistent estimator for $Var(x)$ despite the misspecification.

4. A property of the normal distribution is that the probability of an observation falling below the mean is .5: $F(\mu) = 1/2$. Propose an estimator for the probability that an observation falls below the mean, and show that it is consistent. Discuss how the analyst might use the estimates from your estimator to asess her assumptions about the distribution from which the data is drawn.

5. Simulations. For each $n \in \{50, 100, 250, 1000\}$, simulate n observations from the exponential distribution with $\lambda = 1$. Let $F_{exp}(x|\lambda)$ denote the CDF of the exponential distribution and $F_{norm}(x|\mu, \sigma^2)$ denote the CDF of the normal distribution. Consider four estimators – each a choice of $\hat{F}(\tilde{x})$ – for $Pr(x < \tilde{x})$ for fixed $\tilde{x}$: $F_{exp}(\tilde{x}|\hat{\lambda}_{MLE} = 1/\bar{x})$, $F_{exp}(\tilde{x}|\hat{\lambda}_2 = \frac{1}{2\bar{x}} + \frac{1}{2sd(x)})$,[3] $F_{norm}(\tilde{x}|\hat{\mu}_{MLE}, \hat{\sigma}^2_{MLE})$, and the share of the sample observed below $\tilde{x}$ (where $\hat{\mu}_{MLE}, \hat{\sigma}^2_{MLE}$ are the MLE estimators for the mean and variance, respectively, of a normal distribution). For each n and each $\tilde{x} \in \{1/2, 1, 2, 3\}$, conduct 500 iterations under each set of conditions, and compare these three estimators by plotting the empirical average bias (the sample mean of $\hat{F}(\tilde{x}) - F_{exp}(\tilde{x}|\lambda = 1)\big)$, the empirical variance (across the 500 iterations), and the MSE ($bias^2 + var$), all separately by n. Interpret these results. What do you learn?

---

[3]The mean of the exponential distribution is $1/\lambda$ and the variance is $1/\lambda^2$. Thus, unlike the normal distribution (where the centered second moment is independent of the first moment) with the exponential distribution, the second moment is a function of the first moment (other distributions are like this; the Poisson is an example). Thus, empirically, both the first and second (centered) moment are informative about the same parameter, unlike in the case of the normal distribution. So one estimator of λ is $1/\bar{x}$ and another is $\sqrt{1/Var(x)} = 1/sd(x)$. The estimator $\hat{\lambda}_2$ above is the simple mean of these two potential estimators.

3.1: We have

$$\begin{aligned} & f(x \mid \lambda) = \lambda e^{-\lambda x} \\ \Rightarrow \quad & L(\lambda \mid x) = \prod_{i=1}^{n} \lambda e^{-\lambda x_i} \\ \Rightarrow \quad & l(\lambda \mid x) = \sum_{i=1}^{n} \log \lambda e^{-\lambda x_i} = n \log \lambda - \lambda \sum_{i=1}^{n} x_i \end{aligned}$$

The logarithm is strictly concave and strictly increasing, $\lambda e^{-\lambda x}$ is strictly monotonic, hence $\log \lambda e^{-\lambda x}$ is strictly concave and so $l(\lambda \mid x)$ is a finite sum of strictly concave functions and is thus itself strictly concave. Hence we can use first-order conditions to find the MLE

$$\begin{aligned} 0 &= \frac{n}{\lambda} - \sum_{i=1}^{n} x_i \\ \Rightarrow \quad \hat{\lambda}_{MLE} &= \frac{1}{\frac{1}{n}\sum_i x_i} \end{aligned}$$

Let's show consistency first. By WLLN $\frac{1}{n}\sum_i x_i \xrightarrow{p} E[x_i] = \frac{1}{\lambda}$ and since the mapping $z \mapsto z^{-1}$ is continuous on $z > 0$, the continuous mapping theorem implies that $\hat{\lambda}_{MLE} = \frac{1}{\frac{1}{n}\sum_i x_i} \xrightarrow{p} \frac{1}{E[x]} = \lambda$, as we wanted.

To show bias, we want to show that the expectation of the estimator is not equal to the estimand. The mapping $z \mapsto z^{-1}$ is strictly convex on $z > 0$, so Jensen's Inequality (B.27, Hansen p. 1002) implies

$$\begin{aligned} E[\hat{\lambda}_{MLE}] &= E[\frac{1}{\frac{1}{n}\sum_i x_i}] \\ \text{[Jensen's Inequality]} \quad &> \frac{1}{E[\frac{1}{n}\sum_i x_i]} \\ &= \frac{1}{\frac{n}{n}\frac{1}{\lambda}} \\ &= \lambda \end{aligned}$$

as we wanted.

3.2: The analyst's likelihood function is

$$L(\mu, \sigma^2 \mid x) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp(-\frac{1}{2\sigma^2}(x_i - \mu)^2)$$

$$\Rightarrow \quad l(\mu, \sigma^2 \mid x) = -n \log \sqrt{2\pi\sigma^2} - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2$$

which is smooth (infinitely differentiable, or simply C^∞) and strictly concave, so we impose the first-order condition with respect to μ and get

$$0 = -\frac{1}{\sigma^2} \sum_{i=1}^{n} (x_i - \hat{\mu}_{MLE})$$

$$\Rightarrow \quad \hat{\mu}_{MLE} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

$$[WLLN] \quad \xrightarrow{p} E(x)$$

and so it is a consistent estimator of $E(x)$.

3.3: Let's now take the derivative of the log-likelihood with respect to σ^2 and set to 0

$$\frac{\partial}{\partial \sigma^2} l(\mu, \sigma^2 \mid x) = -\frac{n}{2\sigma^2} + \frac{1}{2(\sigma^2)^2} \sum_{i=1}^{n} (x_i - \mu)^2 = 0$$

$$\Rightarrow \quad \hat{\sigma^2}_{MLE} = \frac{1}{n} \sum_{i=1}^{n} (x_i - \mu)^2$$

which is the sample variance of x_i. By WLLN and the continuous mapping theorem for plim, we have

$$\frac{1}{n} \sum_{i=1}^{n} (x_i - \mu)^2 \xrightarrow{p} E[(x - \mu)^2] =: Var(x)$$

and so it is a consistent estimator of $Var(x)$, despite the misspecification, as we wanted.

3.4: Assume the value of the mean μ is known. Then consider the following estimator

$$\hat{P}(X \leq \mu) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{\{x_i \leq \mu\}}$$

Let $f_X(x)$ be the pdf, then by WLLN

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n}\mathbb{1}_{\{x_i\leq\mu\}} &\xrightarrow{p} E[\mathbb{1}_{\{X\leq\mu\}}]\\ &:= \int_{\mathbb{R}}\mathbb{1}_{\{X\leq\mu\}}f_X(x)dx\\ &= \int_{\{X\leq\mu\}}f_X(x)dx\\ &=: P(X\leq\mu)\end{aligned}$$

and so the estimator is consistent. If X is normal, then $P(X \leq \mu) = 1/2$. But if X is exponentially distributed, $X \sim exp(\lambda)$, then the mean is $1/\lambda$ and

$$P(X \leq \frac{1}{\lambda}) = 1 - e^{-\frac{\lambda}{\lambda}} = 1 - e^{-1} \approx 0.63$$

which is 26% bigger than the normally distributed probability. Hence, if the estimator is far from 1/2 and we have a decent iid sample with a decent sample size, this could cast doubt on our normality assumption because we know that WLLN would make the estimator converge to 1/2. Under misspecification, the estimator wouldn't converge to 1/2 so failure to converge might be a sign of misspecification.

3.5: There are seven points to be made

1. Even though we showed that the misspecified MLE was a consistent estimator of the mean and variance, if we want to use it to predict other quantities, such as the probability of falling below a certain threshold, the misspecification becomes important as we see in Figure 2 where the purple line is away from 0. It even seems inconsistent (bias doesn't vanish with n)

2. The exponential-based MLE (correctly specified) has a bit of bias in the start but it vanishes with n

3. The empirical fractions are unbiased

4. Turning to figure 3, one might have thought that $\hat{\lambda}_2$ were more efficient than $\hat{\lambda}_{MLE}$, since the former uses more information than the latter – by using the mean and standard deviation. However, iid + mild regularity

conditions, which hold for the exponential distribution, means that MLE achieves the Cramer-Rao lower bound and so is the most efficient estimate.

5. Perhaps surprisingly, the efficiency loss from the empirical fraction estimator is biggest toward the center of the distribution – compare Figure 3(a) and 3(b) to 3(c) and 3(d). This might seem surprising given that most of the data is toward the center. However, note that this estimator equals a sum of indicator functions $1_{\{x_i<\tilde{x}\}}$ and, as such, is a sum of Bernoulli trials. In the first TA session we showed that the sum of n Bernoulli trials has variance $np(1-p)$ and so the sample average of n Bernoulli trials has variance $p(1-p)/n$. This expression is maximized when $p = 0.5$ i.e. toward the center of the distribution

6. Most importantly, the empirical fraction is much more noisy (has higher variance) than the two correctly specified models. This is a fundamental principle of econometrics: estimators that rely on stronger assumptions are generally more powerful than those who rely on fewer. The reason to not always impose these assumptions is the risk of misspecification (Normal MLE) which introduces bias (figure 2) and is actually more noise in some of the cases as well – the worst of both worlds. This also becomes evident in Figure 4 where, indeed, the correctly specified MLE performs best with respect to MSE but the misspecified MLE performs the worst. This is because we're squaring the bias when calculating the MSE so any bias has lots of bite.

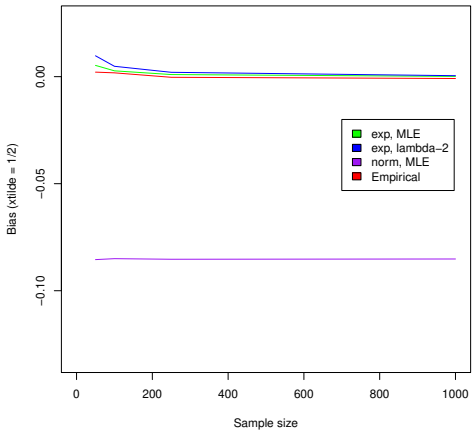
7. The gains from an MLE disappear asymptotically as the three MSEs (the two correctly specified and the empirical fraction) all converge to 0, but the risks (higher MSE if misspecified) do not disappear.

Problem 4

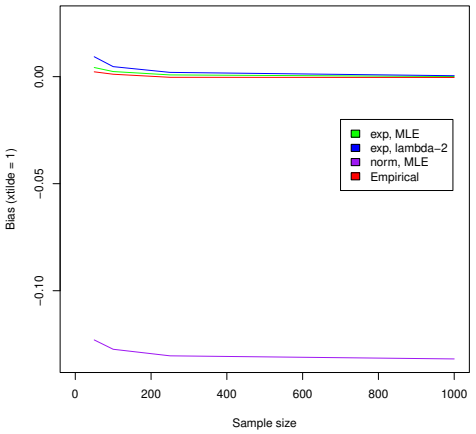
Actually using GMM. Hansen and Singleton (1982) analyze an asset pricing model in order to show that GMM can be used to estimate rational expectations models that aren't linear (unlike other methods).[4] Here, we analyze a

[4]Hansen, Lars Peter, and Kenneth J. Singleton. "Generalized instrumental variables estimation of nonlinear rational expectations models." *Econometrica* (1982): 1269-1286.

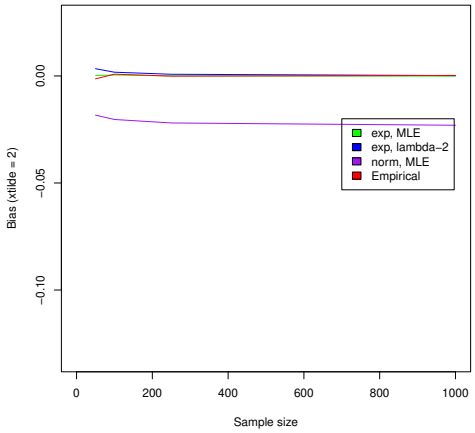
Figure 2: Bias



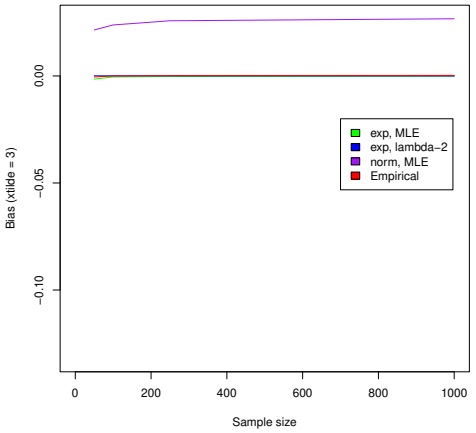
(a) $\tilde{x} = 1/2 : \ F(\tilde{x}) = .393$



(b) $\tilde{x} = 1 : \ F(\tilde{x}) = .632$

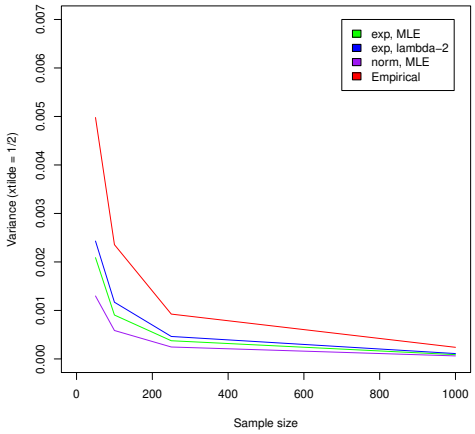


(c) $\tilde{x} = 2 : \ F(\tilde{x}) = .865$

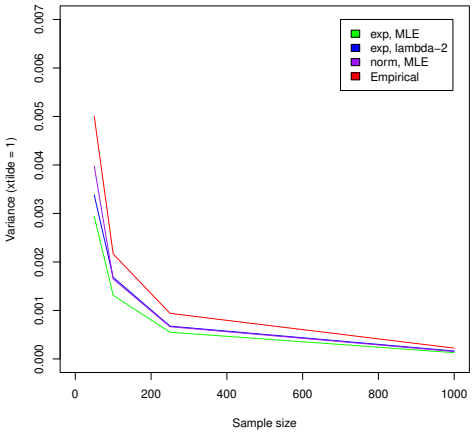


(d) $\tilde{x} = 3 : \ F(\tilde{x}) = .950$

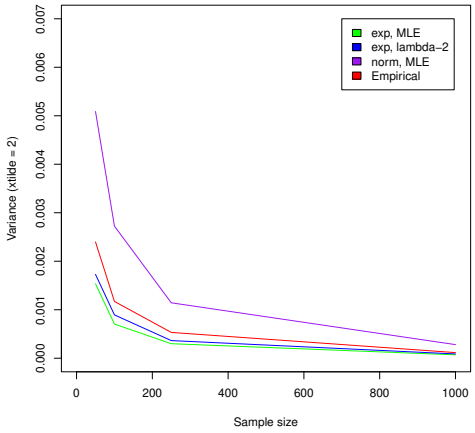
Figure 3: Variance



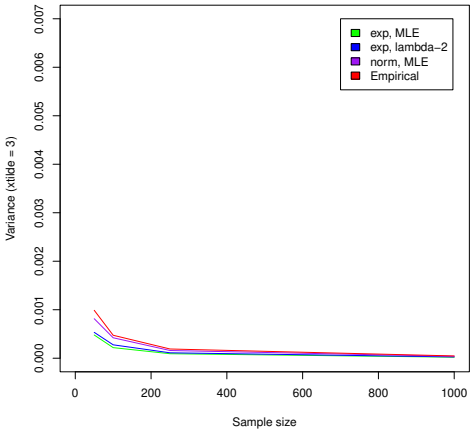
(a) $\tilde{x} = 1/2 : \ F(\tilde{x}) = .393$



(b) $\tilde{x} = 1 : \ F(\tilde{x}) = .632$

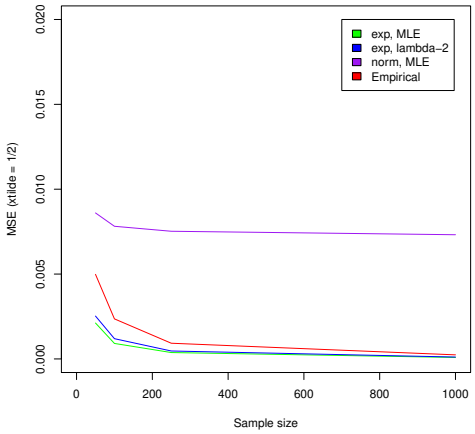


(c) $\tilde{x} = 2 : \ F(\tilde{x}) = .865$

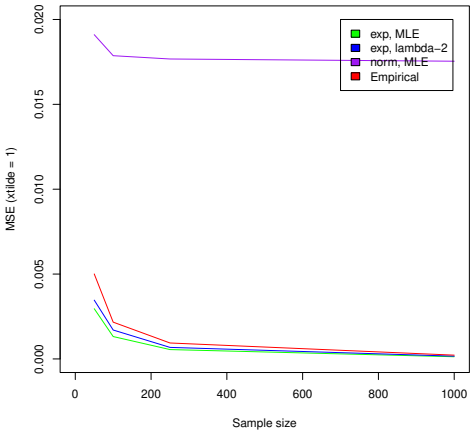


(d) $\tilde{x} = 3 : \ F(\tilde{x}) = .950$

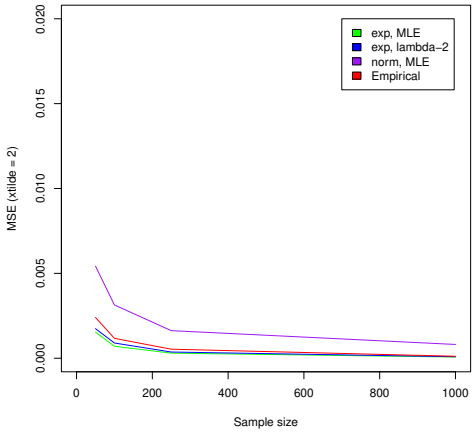
Figure 4: MSE



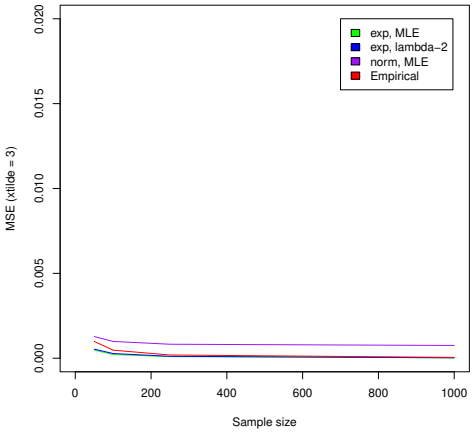
(a) $\tilde{x} = 1/2 : \ F(\tilde{x}) = .393$



(b) $\tilde{x} = 1 : \ F(\tilde{x}) = .632$



(c) $\tilde{x} = 2 : \ F(\tilde{x}) = .865$



(d) $\tilde{x} = 3 : \ F(\tilde{x}) = .950$

simplified version of their model. Suppose a representative agent maximizes expected discounted lifetime utility subject to a budget constraint. Assume a constant relative risk aversion utility function $U(c) = (c^\gamma - 1)/\gamma$. The agent's problem is to solve the following problem:

$$\max_{c_t, I_t} \; E_t\left[\sum_{\tau=0}^{\infty} \beta^\tau U(c_{t+\tau})\right]$$
$$s.t. \;\; c_t + I_t \leq r_t I_{t-1} + w_t \;\forall t$$

where $E_t(\cdot)$ denotes expectations given information available at time t, c_t is consumption, I_t is investment, r_t is the return on the single asset, and w_t is the wage. We assume all of these are observable to both the agent and the econometrician, for each t. However, we assume that the return to assets is drawn iid from some distribution. We assume that households are price takers, and so the return at time $t+1$ is independent of all quantities at time t (including c_t, w_t, r_t). We are interested in estimating β and γ.

1. Take the first order condition with respect to c_t. This generates the Euler equation, which can be written as saying that the marginal utility of consumption today is equal to the discounted expected marginal utility of consumption tomorrow (given expectations about the returns to investment today). Write this Euler equation.

2. Moment conditions are written as $E\big(g(\theta, x)\big) = 0$, where θ is a parameter vector. Rewrite this Euler equation as a moment condition. You have one moment condition and two parameters. You cannot solve this model; it is under-identified.

3. As noted above, we assume r_{t+1} is independent of c_t, w_t, r_t. Thus, it is uncorrelated with marginal utility in period t. Write this as a moment condition. Now you have two moment conditions and two parameters. You can solve this model; it is just-identified.

4. Let $g_b(\theta, x)$ be the moment condition you solved for in (b) above. Let $g_c(\theta, x)$ be the moment condition you solved for in (c) above. The gradient can be written as:

$$\begin{pmatrix} \frac{\partial g_b(\theta,x)}{\partial \beta} & \frac{\partial g_b(\theta,x)}{\partial \gamma} \\ \frac{\partial g_c(\theta,x)}{\partial \beta} & \frac{\partial g_c(\theta,x)}{\partial \gamma} \end{pmatrix}$$

   Solve for the gradient.

5. Download data from this dropbox link: `https://www.dropbox.com/scl/fi/7sh1te900ken36sydge23/gmmdata.csv?rlkey=k0iqjqcvp16x1tz4jyj54ut76&dl=0`. Use the gmm package in R to solve for β, γ (see `https://cran.r-project.org/web/packages/gmm/vignettes/gmm_with_R.pdf` for helpful documentation).

4.1: Utility is increasing in c, this implies local non-satiation whereby the budget constraint binds. So we can substitute for c_t in the objective function and get

$$\max_{I_t} E_t\left[\sum_{\tau=0}^{\infty} \beta^\tau U(r_{t+\tau} I_{t+\tau-1} + w_{t+\tau} - I_{t+\tau})\right]$$

Imposing the first order condition with respect to I_t gives us

$$\begin{aligned}
0 &= -E_t[\beta^0 U'(r_{t+0} I_{t+0-1} + w_{t+0} - I_{t+0})] + E_t[\beta^1 r_{t+1} U'(r_{t+1} I_{t+1-1} + w_{t+1} - I_{t+1})] \\
&= -U'(c_t) + \beta E_t[r_{t+1} U'(c_{t+1})] \\
\Rightarrow \quad U'(c_t) &= \beta E_t[r_{t+1} U'(c_{t+1})] && \text{(Euler)} \\
[CRRA] \Rightarrow \quad c_t^{\gamma-1} &= \beta E_t[r_{t+1} c_{t+1}^{\gamma-1}] && \text{(Euler-CRRA)}
\end{aligned}$$

as we wanted.

4.2: The moment condition follows from (Euler-CRRA)

$$0 = E[\beta r_{t+1} c_{t+1}^{\gamma-1} - c_t^{\gamma-1}] = E[g(\beta, \gamma; r_{t+1}, c_{t+1}, c_t)] =: E[g(\theta; x)]$$

with $\theta = (\beta, \gamma)$ and $x = (r_{t+1}, c_{t+1}, c_t)$.

4.3: When Mitch says marginal utility and returns are "uncorrelated" he means it in the correct sense, that is, the covariance of marginal utility and returns are 0, i.e., $Cov(U(c_t), r_{t+1}) = 0$. In applied economics, we often say uncorrelated means $E[XY] = 0$ but that is only true when (at least) one of X or Y is mean 0, since then $Cov(X, Y) = E[XY]$. However, neither returns r_{t+1} nor marginal utility of consumption $c_t^{\gamma-1}$ are mean 0 in our setting, but we still want to use the $E[XY] = 0$ condition, so what we do is to center returns around their mean, i.e. we consider $r_{t+1} - E(r_{t+1}) = r_{t+1} - \bar{r}$ where

we use that the returns process is stationary so the expectation is a constant. Then we can write the moment condition as

$$0 = E[c_t^{\gamma-1}(r_{t+1} - \bar{r})]$$

4.4: Translating the previous objects into Mitch's notation, we have

$$\begin{aligned}
g_b(\theta, x) &= \beta r_{t+1} c_{t+1}^{\gamma-1} - c_t^{\gamma-1} \\
g_c(\theta, x) &= c_t^{\gamma-1}(r_{t+1} - \bar{r}) \\
\Rightarrow \quad \frac{\partial}{\partial\beta} g_b &= \frac{\partial}{\partial\beta} \beta r_{t+1} c_{t+1}^{\gamma-1} - c_t^{\gamma-1} = r_{t+1} c_{t+1}^{\gamma-1} \\
\frac{\partial}{\partial\gamma} g_b &= \beta r_{t+1} c_t^{\gamma-1} \ln(c_{t+1}) - c_t^{\gamma-1} \ln(c_t) \\
\frac{\partial}{\partial\beta} g_c &= 0 \\
\frac{\partial}{\partial\gamma} g_c &= (r_{t+1} - \bar{r}) c_t^{\gamma-1} \ln(c_t)
\end{aligned}$$

4.5: I get the following results, with the caveat that the covariance matrix of the coefficients is singular

Figure 5: GMM estimates of β, θ

```
Method:  iterative

Kernel:  Quadratic Spectral

Coefficients:
       Estimate  Std. Error  t value  Pr(>|t|)
beta   0.95936       Inf     0.00000  1.00000
gamma  1.54547       Inf     0.00000  1.00000

J-Test: degrees of freedom is 0
                 J-test                  P-value
Test E(g)=0:    3.15446226291531e-11  *******
```