Method of moments¶

1. The method of moments principle

The population moment conditions will play a crucial role in the discussion so it is worth going back to the primitives to understand the mechanics of GMM.

The raw uncentered moments are easy to compute and they reveal important aspects of a distribution. For example, the first four moments tell us about the population mean, variance, skewness and kurtosis. Using them we can immediately place restrictions according to our theory on the location, scale or shape of the distribution without specifying a full model or distribution.

Once we have some information on the population, the question remains how to use the sample to estimate the parameters of interest. In general, sample statistics each have a counterpart in the population, for example, the correspondence between the sample mean and the population expected value. The natural next step in the analysis is to use this analogy to justify using the sample moments as bases of estimators of the population parameters. This was the original idea in *Karl Pearson*'s work [1893], [1894], [1895] in the late 19th century.

The Pearson family of distributions is a very flexible mathematical representation that has several important and frequently used distributions among its members depending on the parameterization you choose. Pearson's problem was to select an appropriate member of the family for a given dataset.

Example 1 – Simple method of moments estimator

To show a very simple example, assume that the population distribution has unknown mean μ and variance equal to one. In this case, the population moment condition states that $E[x_i] = \mu$. If $\{x_i : i = 1,2,\ldots,n\}$ is an independent and identically distributed sample from the distribution described formerly, then the sample average $\bar{x} = \frac{1}{n}\sum_{i=i}^{n} x_i$ is the sample analogue to the population mean $E[x_i]$. By utilizing this analogy principle, the method of moments (MM) estimator for $E[x_i] = \mu$ is simply given by $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \hat{\mu}_n$.

Basically we had to work out the first moment, then to replace it with the sample analogue and to solve the equation for the unknown parameter. What remains to be established is whether this approach is the best, or even a good way to use the sample data to infer the characteristics of the population.1 Our intuition suggests that the better the approximation is for the population quantity by the sample quantity, the better the estimates will be. To make a step further, it is time to introduce some more general definitions.

Definition 1 – Method of moments estimator

Suppose that we have an observed sample $\{x_i : i = 1,2,\ldots,n\}$ from which we want to estimate an unknown parameter vector $\theta \in \mathrm{R}^p$ with true value θ_0. Let $f(x_i,\theta)$ be a continuous and continuously differentiable $\mathrm{R}^p \to \mathrm{R}^q$ function of θ, and let $E[f(x_i,\theta)]$ exist and be finite for all i and θ. Then the population moment conditions are that $E[f(x_i,\theta_0)] = 0$. The corresponding sample moments are given by

$$f_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} f(x_i,\theta).$$

The method of moments estimator of θ_0 based on the population moments $E[f(x_i,\theta)]$ is the solution to the system of equations $f_n(\theta) = 0$.

Note that if $q = p$, then for an unknown parameter vector θ the population moment conditions $E[f(x_i,\theta)] = 0$ represent a set of p equations for p unknowns. Solving these moment equations would give the value of θ which satisfies the population moment conditions and this would be the true value θ_0. Our intuition suggests that if the sample moments provide good estimates of the population moments, we might expect that the estimator $\hat{\theta}$ that solves the sample moment conditions $f_n(\hat{\theta}) = 0$ would provide a good estimate of the true value θ_0 that solves the population moment conditions $E[f(x_i,\theta_0)] = 0$.

Now we present some common models in terms of the MM terminology.

Example 2 – Ordinary least squares (OLS)

Consider the linear regression model

$$y_i = x_i'\beta_0 + u_i,$$

GMM ESTIMATOR¶

rameters.

Example 5 – Motivation for GMM

Consider again Example 1. Notice that our estimation was based solely on the first raw moment of the distribution. Now suppose that we believe to know that the sample at hand is a result of n independent draws from a Poisson distribution with parameter λ. Thus the new (additional) population moment condition based on the second raw moment is $E[x_i^2] - \lambda^2 - \lambda = 0$. The MM estimator of λ should satisfy the system of equations based on the sample moments

$$\begin{bmatrix} \frac{1}{n}\sum_{i=1}^{n} x_i - \hat{\lambda} \\ \frac{1}{n}\sum_{i=1}^{n}\left(x_i^2\right) - \hat{\lambda}^2 - \hat{\lambda} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}.$$

Now we have two moment conditions and one unknown parameter which means that we do not have a general solution for $\hat{\lambda}$.

PDF GMM¶

https://www.stata.com/meeting/germany10/germany10_drukker.pdf

Method of Moments (MM)

- We estimate the mean of a distribution by the sample, the variance by the sample variance, etc
- We want to estimate $\mu = E[y]$
  - The population moment condition is $E[y] - \mu = 0$
  - The sample moment condition is

$$(1/N)\sum_{i=1}^{N} y_i - \mu = 0$$

  - Our estimator is obtained by solving the sample moment condition for the parameter
- Estimators that solve sample moment conditions to produce estimates are called method-of-moments (MM) estimators
  - This method dates back to Pearson (1895)

Ordinary least squares (OLS) is an MM estimator

- We know that OLS estimates the parameters of the condtional expectation of $y_i = \mathbf{x}_i\boldsymbol{\beta} + \epsilon_i$ under the assumption that $E[\epsilon|\mathbf{x}] = 0$
- Standard probability theory implies that

$$E[\epsilon|\mathbf{x}] = 0 \Rightarrow E[\mathbf{x}\epsilon] = \mathbf{0}$$

So the population moment conditions for OLS are

$$E[\mathbf{x}(y - \mathbf{x}\boldsymbol{\beta})] = \mathbf{0}$$

- The corresponding sample moment condtions are

$$(1/N)\sum_{i=1}^{N} \mathbf{x}_i(y_i - \mathbf{x}_i\boldsymbol{\beta}) = \mathbf{0}$$

Solving for $\boldsymbol{\beta}$ yields

$$\widehat{\boldsymbol{\beta}}_{OLS} = \left(\sum_{i=1}^{N} \mathbf{x}_i'\mathbf{x}_i\right)^{-1} \sum_{i=1}^{N} \mathbf{x}_i'y_i$$

8 / 29

Generalized method-of-moments (GMM)

- The MM only works when the number of moment conditions equals the number of parameters to estimate
  - If there are more moment conditions than parameters, the system of equations is algebraically over identified and cannot be solved
  - Generalized method-of-moments (GMM) estimators choose the estimates that minimize a quadratic form of the moment conditions
    - GMM gets as close to solving the over-identified system as possible
    - GMM reduces to MM when the number of parameters equals the number of moment condtions

https://online.stat.psu.edu/stat415/lesson/1/1.4

BLANK¶

https://bookdown.org/probability/inference2/method-of-moments.html

Recall from probability theory hat the **moments** of a distribution are given by:

$$\mu^k = E(X^k)$$

Where μ^k is just our notation for the k^{th} moment. So, the first moment, or μ, is just $E(X$ and the second moment, or μ^2, is $E(X^2)$. Recall that we could make use of MGFs (mom functions) to summarize these moments; don't worry, we won't really deal with MGFs mu

Instead, we are all about inference here; when we see the k^{th} moment of a distribution, w "how could I estimate that?" For example, μ^3 is just the average value of an individual ob just like μ is the average value of an individual observation. How would we then estimate

In inference, we're going to use something called **sample moments**. The k^{th} sample mo as:

$$\hat{\mu}_k = \frac{1}{n}\sum_{i=1}^{n} X_i^k$$

Where $\hat{\mu}_k$ is the k^{th} sample moment (remember, we put a hat on things when we mean t estimating something else. Here, $\hat{\mu}_k$ is estimating the same thing without a hat: μ_k, or the

Where $\hat{\mu}_k$ is the k^{th} sample moment (remember, we put a hat on things when we mean they are estimating something else. Here, $\hat{\mu}_k$ is estimating the same thing without a hat: μ_k, or the k^{th} mo

Let's break this down. This is basically saying that if we want μ_k, or $E(X^k)$ (they are the same th take a sample of n people, raise each of their values to the k, add them up and divide by the nur individuals in the sample (n). If you wanted to estimate the fourth moment for the weight of colleg you would take a sample of some college males, raise each of their weights to the power of 4 an by the number of people you sampled.

Does this make sense? Well, consider the case where $k = 1$, or μ. This is, of course, just the me distribution. The sample moment for this first moment is given by:

$$\frac{1}{n}\sum_{i=1}^{n} X_i$$

Which we got by just plugging in $k = 1$ to the above formula for sample moments. Hey-o, that's t sample mean, or what we've long-established is the natural estimator for the true mean! You can this builds, then, as we get to higher and higher moments.

Anyways, the takeaway here is that **we use sample moments to estimate the actual mom**
distribution. That's great, and we would be finished if we were asking you to estimate mom
distribution. However, you're rarely asked to estimate actual moments; instead, as you've se
generally asked to estimate *parameters* of a distribution.

So, we know that we can estimate moments with sample moments, and we know that we wa
estimate parameters. How can we use these two facts to get what we want; a solid estimate
parameters? Well, if we can write the parameters of a distribution in terms of that distributio
and then simply estimate those moments in terms of the sample moments, then we have cre
estimator for the parameter in terms of the sample moment.

Whoa…that's a little crazy, and probably too much of a mouthful right now. Let's try and lear
solid example of the most famous statistical distribution: the Normal.

Example normal dist¶

We already know, from what we learned earlier, that we have natural estimates for the mo
Normal distribution. What we're going to do, then, is try and write the *moments* of a Norma
parameters (the mean and variance). We only need to write out the first two moments, $E($
since we have to parameters (in general, if you have k parameters that you want to estima
write out k moments).

Let's go ahead and do that. How do we write $E(X)$ in terms of μ and σ^2? Well, you know
$E(X)$ is just μ, since they are both the mean for a Normal distribution. What about writing
terms of μ and σ? Well, this takes a little bit more cleverness. Recall that $Var(X) = E(X$
Re-writing this yields $Var(X) + E(X)^2 = E(X^2)$. We know that, for a Normal distributio
$Var(X) = \sigma^2$, and $E(X)^2 = \mu^2$. So, we can write $E(X^2) = \mu^2 + \sigma^2$, and this yields th
equations:

$$\mu_1 = \mu$$

$$\mu_2 = \sigma^2 + \mu^2$$

Where μ_1 is notation for the first moment, μ_2 notation for the second moment, etc.

Well now, we've written our moments in terms of the parameters that we're trying to es
that we have good estimators (the sample moments) for our moments μ_1 and μ_2, so le
this system of equations for the parameters *in terms of* the moments. Well, we now tha
can plug in μ_1 for μ in the second equation and then solve for σ^2. We get that:

$$\mu_2 = \sigma^2 + \mu^2 \rightarrow \sigma^2 = \mu_2 - \mu_1^2$$

So now our two equations for the parameters in terms of the moments are:

$$\mu = \mu_1$$

$$\sigma^2 = \mu_2 - \mu_1^2$$

That is, the first parameter, the mean μ, is equal to the first moment of the distribution,
parameter, the variance σ^2, is equal to the second moment of the distribution minus th
the distribution squared.

Why did we go through all of that work? Well, recall the ultimate goal of all of this: to es
parameters of a distribution. Recall also that we know how to estimate the moments of
the sample moments! That is, a good estimate for μ_k is $\frac{1}{n}\sum_{i=1}^{n} X_i^k$. So, now that we k
parameters in terms of the moments, estimating the parameters is the same as estima
We can plug in our estimates for the moments and get good estimates for the paramet

So, the sample moment for μ_1, by formula, is just $\frac{1}{n}\sum_{i=1}^{n} X_i$, and the sample momen
formula, $\frac{1}{n}\sum_{i=1}^{n} X_i^2$. Plugging these in for μ_1 and μ_2 yields:

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

$$\hat{\sigma^2} = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \left(\frac{1}{n}\sum_{i=1}^{n} X_i\right)^2$$

Where $\hat{\mu}$ and $\hat{\sigma^2}$ are just estimates for the mean and variance, respectively (remember
things to indicate that it's an estimator). We can test this in R by generating data from,
distribution and using the above MoM estimators to see if we can get close to the origi
(mean of 5 and variance of 9).

In [1]:

```
# replicate
set.seed(0)
n <- 20
mu <- 5
sigma <- 3

# generate
samples <- replicate(rnorm(n, mu, sigma), n = 100)
```

In [2]:

```
# calculate estimates
sample_means <- apply(samples, 2, mean)
sample_var <- apply(samples, 2, function(x) {
  return((1 / n) * sum(x ^ 2) - sum(x / n) ^ 2)
})
```

In [3]:

```
# check if estimates are close:
mean(sample_means); mean(sample_var)
```

4.93907597626289
8.95836332806096

Regression - Method of Moments

More generally, one can write the moment conditions as a vector of functions g the observed data, including all variables (y_i, X_i) and instruments $(\mathbf{Z}_i)$ in the r while β is the vector of parameters of length k. The model is identified if the so $Eg(X_i, \beta) = 0$ and $Eg(X_i, \hat{\beta}) = 0$ imply that $\beta = \hat{\beta}$. This requires that we ha restrictions for k parameters.

For the OLS regression, one can use the moment condition $E(\mathbf{X}_i U_i) = 0$ or E to solve for the usual OLS estimator.

The idea can be carried over to other more complicated regression models. For where $g(X_i, \beta)$ is linear in β i.e. $g(X_i, \beta) = \mathbf{Z}_i(y_i - \mathbf{X}_i'\beta)$ or $E(\mathbf{Z}_i U_i) = 0$, ar perfectly identified $(l = k)$, solving the moment condition yields the formula fo

$$
\begin{aligned}
0 &= \sum_{i=1}^{n} \mathbf{Z}_i(y_i - \mathbf{X}_i'\hat{\beta}^{IV}) \\
\hat{\beta}^{IV} &= \left(\sum_{i=1}^{n} \mathbf{Z}_i \mathbf{X}_i'\right)^{-1} \sum_{i=1}^{n} \mathbf{Z}_i y_i \\
&= (\mathbf{Z}'\mathbf{X})^{-1}\mathbf{Z}'\mathbf{y}
\end{aligned}
$$

Hence an IV regression could be thought of as substituting 'problematic' OLS m better moment conditions with the addition of instruments.

Extension - Generalised Method of Moments (GMM)

While it is not possible to identify β if there are too few restrictions, one could still ide
are $l > k$ restrictions (overidentified), as seen in the poisson example.[1] One might the
is the best way to combine these restrictions. The GMM approach, introduced by Han
finds an estimate of β that brings the sample moments as close to zero as possible. N
moment conditions for all the restrictions are still equal to zero, but the sample approx
drawn from a finite sample, may not be equal to zero. In other words, the GMM estima
as the value of β that minimizes the weighted distance of $\frac{1}{n}\sum_{i=1}^{n} g(X_i, \beta)$:

$$\begin{aligned}\hat{\beta}^{GMM} &= \arg\min_{\beta \in B} \left\| \frac{1}{n}\sum_{i=1}^{n} g(X_i, \beta) \right\|_W^2 \\ &= \arg\min_{\beta \in B} \left(\frac{1}{n}\sum_{i=1}^{n} g(X_i, \beta) \right)' \mathbf{W} \left(\frac{1}{n}\sum_{i=1}^{n} g(X_i, \beta) \right)\end{aligned}$$

where $\mathbf{W}$ is the $l \times l$ matrix of weights which is used to select the ideal linear combi
instruments. In the case of the regression model where $g(X_i, \beta)$ is linear in β but is o
the general GMM formula can be found by minimising the above condition and is give

$$\hat{\beta}^{GMM} = \left((\mathbf{X}'\mathbf{Z})\mathbf{W}(\mathbf{Z}'\mathbf{X}) \right)^{-1} (\mathbf{X}'\mathbf{Z})\mathbf{W}(\mathbf{Z}'\mathbf{y})$$

Note that when $\mathbf{W} = (\mathbf{Z}'\mathbf{Z})^{-1}$, $\hat{\beta}^{GMM} = \hat{\beta}^{IV}$.[2] Please google efficient GMM, for mo
on the optimal choice of the weighting matrix.

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]: