Estimation

Devesh Raval

All econometrics problems are at their heart conditional prediction problems. We start with $\{Y, X\}$ where Y is a scalar random variable and X a 1 by k random vector. We then want to predict Y given that we know X. For example, Y can be wages and X schooling and experience, or Y birthweight and X smoking, or Y consumption and X (permanent) income. What does it mean to predict Y given X?

One object we could want is the entire conditional distribution, $P(Y|X)$. Or we could want certain moments of this distribution only:

$E(Y|X)$: the conditional expectation function

$M(Y|X)$: the conditional median function

$q_\alpha(Y|X)$: the alpha quantile function

I will now examine a number of different approaches to estimation: the best predictor under a loss function, maximum likelihood, and method of moments- and show how they can lead to OLS under certain assumptions. Ordinary Least Squares or OLS is the most widely used estimator in economics and so it is worth understanding why we should use it.

1 Best Predictors under a Loss Function

Let us define $\theta = \theta(X)$ to be the best predictor of Y given X. What does "best" predictor mean? We need some way to assess how bad we are doing. To do so, we introduce a loss function $L(Y - \theta)$, where we give the loss function some nice properties:

$$\begin{aligned} L(0) &= 0 \\ L(u) &= L(-u) \end{aligned}$$

and, for $0 < u < v$,

$$L(u) < L(v)$$

Thus, the loss function is zero when $\theta = Y$, symmetric, and increasing in the absolute value of its argument. We can then call the "best predictor" the value of $\theta(x)$ which minimizes loss at every value of x.

Thus,

$$\theta(x) \quad = \quad \arg\min_c E(L(Y-c)|X)$$

arg min means the value c that solves the minimization problem. In this case only Y is random so the expectation is over the conditional distribution of $Y|X$.

We will in general generate a different best predictor depending on the loss function. Here are a couple of examples:

1.1 Squared Loss

$$L(Y-\theta) \quad = \quad (Y-\theta)^2$$

Under squared loss predictors that are far away from the data get larger and larger marginal penalties. The best predictor under squared loss is then the conditional expectation function:

$$\theta(X) \quad = \quad E(Y|X)$$

Proof. Let $\mu(x) = E(Y|X)$. Then we can write the problem as:

$$\begin{aligned} E((Y-\theta)^2|X) &= E(((Y-\mu(X))+(\mu(X)-\theta))^2|X) \\ &= E((Y-\mu(X))^2+(\mu(X)-\theta)^2+2(Y-\mu(X))(\mu(X)-\theta)|X) \\ &= Var(Y|X)+(\mu(X)-\theta)^2+2(E(Y|X)-\mu(X))(\mu(X)-\theta) \\ &= Var(Y|X)+(\mu(X)-\theta)^2 \end{aligned}$$

We set θ to make $(\mu(X)-\theta)^2 = 0$ which means that $\theta = \mu(X) = E(Y|X)$.

Thus, the conditional expectation function minimizes expected loss if we have a quadratic loss function. If we define the prediction error as $\epsilon = Y - E(Y|X)$ it is clear that $E(\epsilon|X) = 0$.

Note that if the CEF minimizes expected loss for every value of X, it also minimizes averaged expected loss (averaging over X)- so it solves the problem

$$\theta(x) \quad = \quad \arg\min_c E(L(Y-c))$$

as well, where c is allowed to vary by X. □

1.2 Absolute Loss

Under absolute loss the loss function is $L(Y-\theta) = |Y-\theta|$. The best predictor then becomes $\theta(X) = M(Y|X)$ or the conditional median function. The mean and median thus come from different loss functions- they do not always move together! The median has some advantages over the mean- for example it is less sensitive to outliers. We will not cover much more about median or quantile (median is the 50% quantile, one can think of other quantiles) in this class, but some references are:

Angrist and Pischke, Mostly Harmless Econometrics

Roger Koenker, Quantile Estimation (or 2001 JEP article)

1.3 Best Predictor under Shape Restrictions and Squared Loss

So far, we have seen why economists and statisticians are interested in the conditional expectation function- the CEF solves the best predictor problem under squared loss.

1. Best Constant Predictor

   Under this, $\theta(x) = c$. Then our expected loss function becomes:

$$\begin{aligned} E((Y-\theta)^2|X) &= E((Y-c)^2|X) \\ &= E((Y-c)^2) \end{aligned}$$

   Since $\theta(x)$ is constant it does not depend on X. The first order conditions are:

$$\begin{aligned} -2E(Y-c) &= 0 \\ c &= E(Y) \end{aligned}$$

   The best constant predictor of Y is just the expectation of Y, and does not depend on X!

2. Best Linear Predictor

Let $\{Y, X\} \sim P$. The Best Linear Predictor of Y given X is found by minimizing $E(U^2) = E((Y-Xb)^2)$.

FOC:

$$\begin{aligned} -2E(X'(Y-Xb) &= 0 \\ E[X'X]b &= E(X'Y) \end{aligned}$$

These are k by 1 first order conditions, since X is k by 1. If $X'X$ is nonsingular (there is no perfect linear relationship among the Xs,

$$b = E(X'X)^{-1}E(X'Y)$$

The second order conditions are positive indicating a unique minimum:

$$2E(X'X)$$

The best linear predictor is the OLS estimator in the population.

Note: R^2: the common measure of goodness of fit- is really measuring how much better the BLP is from the BCP.

What are the differences between the BLP and CEF? How does this relate to OLS?

1. If the CEF is actually linear, the BLP is the CEF and so OLS gives us the CEF. This is only true in rare cases, however. One such rare case is if X and Y are bivariate normal. This case can be very useful (for ex when proving the Heckman selection model) so we show it below.

   Example 1. Let X and Y be distributed bivariate normal. The conditional distribution of Y given X and X given Y are both linear. You will do the proof for this in your homework, using completing the square.

2. If the CEF is not actually linear, the BLP will serve as the best linear approximation to the CEF. Thus OLS can be seen as giving us a linear approximation to a true nonlinear CEF. This may be fine if the true CEF is not too nonlinear so the linear approx is close to reality. I will prove both 1 and 2 below:

   Proof. Let us minimize the distance between the CEF and a linear function: □

$$\beta = \arg\min_b E(E(Y|X) - Xb)^2)$$

But we can think of our earlier problem for the BLP,

$$\begin{aligned}
\arg\min E(U^2) &= \arg\min_b E((Y - Xb)^2) \\
&= \arg\min_b E(((Y - E(Y|X)) - (E(Y|X) - Xb))^2) \\
&= \arg\min_b E(E(Y|X) - Xb)^2)
\end{aligned}$$

Thus the BLP is the best linear approximation for the CEF.

3. The BLP FOC state that $E(X'U) = 0$. We can always form this moment condition and get the BLP.- U is just defined as Y minus the BLP.

   In a sense, we can say:

$$\begin{aligned} BLP &= E^*[Y|X] = bX \\ Y &= bX + U \\ U &= Y - bX \\ E(XU) &= 0 \end{aligned}$$

   where the last holds by the definition of the BLP. We have not really shown anything. However, the BLP does not imply that $E(U|X) = 0$. If we say $E(U|X) = 0$, we have made an assumption that the CEF is linear.

4. In general the variance of the prediction error of the CEF is lower than that of the BLP. From the previous derivation,

$$\begin{aligned} V(U) &= E(U^2) - E(U)^2 \\ &= E(U^2) \\ &= E((Y - bX)^2|X) = Var(Y|X) + (\mu(X) - bX)^2 \\ &< Var(Y|X) \\ &= Var(\epsilon|X) \end{aligned}$$

   These will be equal when the CEF=BLP.

5. The CEF only depends on the conditional distribution of $Y|X$. The BLP depends on both the conditional distribution of $Y|X$ and the marginal distribution of X. We can see this in the definitions of both- the CEF is solved for for every value of X (so by definition does not depend on marginal distribution of X). The BLP minimization problem averages over all X as well- so depends on the marginal distribution of X. Here is a simple example where this matters.

   Example 2. 3 point distribution:

Y	1	0	1
X	-1	0	1

   The true relationship is quadratic- draw the graph.

   The CEF is $E(Y|X = 1) = E(Y|X = -1) = 1$ and $E(Y|X = 0) = 0$. If each point is sampled with 1/3 probability, $Cov(X, Y) = 0$ (calculate $E(XY)$ using LIE logic conditioning on X) and so BLP will show no relationship between X and Y. (b/c linear estimator $\frac{Cov(X,Y)}{Var(X)}$). If the marginal distribution of X moves closer to 1 from -1. the BLP is upwards sloping and so we find a positive relationship- thus the BLP sign is determined by the density of X!!

   For this reason, may miss something if estimate a line when true relationship is not linear.

6. Finally, the CEF is only defined on the support of X. To extrapolate outside of the support of X, we must impose some sort of functional form-such as linearity with the BLP. With sampling variability, the difference between inside and outside the support of X is less clear though.

1.4 Analogy Principle

Before, everything we have done has been at the level of the population. But in practice, we don't observe the population. Instead we only observe a sample from the population.

Suppose we observe a random sample of size n, on $\{Y_i, X_i\}$ from P where

$$\{Y, X\} \quad \sim \quad P$$

where P is unknown. How should we proceed?

We need to construct an estimator, $T_n = h(Y, X)$. Here T_n is a random variable that depends on the distribution of the data. t_N is a single observation of the random variable T_N- a realization of the estimator given the data observed.

T_N is a random variable because with different random samples, we would obtain different estimates.

In general, T_N depends on the underlying population distribution P, the sampling process, and the sample size N.

What does a sample look like?

Y, n by 1,

$$Y = \begin{matrix} Y_1 \\ Y_2 \\ Y_3 \\ ... \\ Y_n \end{matrix}$$

$$X = \begin{matrix} X_{11} & X_{21} & ... & X_{k1} \\ X_{12} & & & \\ ... & & & \\ & & & \\ X_{1n} & & & X_{kn} \end{matrix}$$

Here Y are all i.i.d. from random sampling, X n by k are all independent, with each row (a different observation) identically distributed with next row.

How might we choose T_n?

Analogy Principle: To estimate a population parameter, use the corresponding feature in the sample.

Example 3. $\theta = E(Y)$

$T_N = \frac{1}{N} \sum_i Y_i = \bar{Y}$

The sample moment replaces the population moment.

Example 4. $\beta = E(X'X)^{-1}E(X'Y)$

This is the population version of OLS. There are three different sample analogies, in this case all leading to the same estimator:

1.
$$\begin{aligned} T_N &= E_N(X'X)^{-1}E_N(X'Y) \\ &= [\frac{1}{N}\sum_i X_i'X_i]^{-1}[\frac{1}{N}\sum_i X_i'Y_i] \\ &= (X'X)^{-1}(X'Y) \end{aligned}$$

   This is just two sample averages!

2. We use the least squares minimization problem from the population:
$$\begin{aligned} \beta &= \arg\min_c E((Y - Xc)^2) \\ T_N &= \arg\min_c E_N((Y_i - X_ic)^2) \\ T_N &= \arg\min_c \frac{1}{N}\sum_i (Y_i - X_ic)^2 \end{aligned}$$

   and then find the argument that minimizes this function in the sample- it will be the OLS estimator.

3. We use the first order conditions for β, $E(X'U) = 0$, $U = Y - X\beta$
$$\begin{aligned} \frac{1}{N}\sum_i X_ie_i &= 0 \\ e_i &= Y_i - X_iT_N \end{aligned}$$

With three different analogies, we will get the same T_N in the linear model (not always!).

Example 5. Nonparametric estimator in discrete case:
$$\begin{aligned} \theta(x_0) &= E(Y|X = x_0) \\ T_N &= \frac{1}{N(x_0)}\sum_i Y_i 1(X_i = x_0) \\ N(x_0) &= \sum_i 1(X_i = x_0) \end{aligned}$$

This is the nonparametric approximation to the CEF- take the average Y value at $X = x_0$.

5 min digression into nonparametrics- Nonparametrics is just doing cell averages to estimate statistical quantities like the CEF- without assuming a functional form. The advantage of this is we dont want our results driven by functional form if our functional form is wrong!! In the discrete case, its just a cell average. Given a particular value of X, what is the average value of Y? Thats what the estimator above does.

What happens if X is continuous? Then at any particular value of X, there is no data! Instead, we take an average in a band around X, as follows:

$$T_N = \frac{\frac{1}{N}\sum_i Y_i 1(|X_i - x_0| < \delta)}{\frac{1}{N}\sum_i 1(|X_i - x_0| < \delta)}$$

This is called the Uniform Kernel Estimator.

δ is the bandwidth parameter- determines how are far away data points can get away from the value we are interested in. Bias-variance tradeoff- as increase δ, decrease variance (as more observations are included) but increase bias (more data points away from x_0 which we are interested in).

Draw a quadratic example to show this.

Also curse of dimensionality- harder to estimate in multiple Xs b/c have to average within a small cell- may have very few observations. Ex: average log wages conditional on education easy. Hard if condition on education, experience, race, sex, union status, etc. etc.- there may be no one in cell- if very few people in cell your estimate will have a large variance. To do good nonparametrics you may need lots of data!

Another approach- polynomials - throw in squares, cubes, etc. to try to capture curvature. Also depends on econ application- avg. cost curves should be quadratic (if fixed cost + MC that eventually increases, or fixed cost + constant MC) so know to put in constant term.

1.5 Comparison of Estimators

How do we assess estimators?

Finite Sample-

Unbiasedness- Is the expectation of the estimator equal to its true value?

Lets look at the expectation of the sample mean.

$$\begin{aligned} E(\bar{Y}) &= E(\frac{1}{N}\sum_i Y_i) = \frac{1}{N}\sum_i E(Y_i) \\ &= \frac{1}{N}\sum_i \mu_Y = \mu_Y \end{aligned}$$

So the sample mean is unbiased. But so are an infinite number of other estimators. For example, $T'_N = Y_1$. $E(T'_N) = E(Y_1) = \mu_Y$.

To separate out unbiased estimators, we can look at the variance of the estimator as well.

$$\begin{aligned} V(\bar{Y}) &= V(\frac{1}{N}\sum_i Y_i) = \frac{1}{N^2}\sum_i V(Y_i) \\ &= \frac{1}{N^2} N V(Y_i) = \frac{V(Y_i)}{N} \end{aligned}$$

By contrast,$V(T'_N) = V(Y_1)$ which is always bigger.

Theorem 6. *Sample Mean Thereom: In a random sample of sample size N, from any population, the sample mean has an expectation equal to the population mean, and variance is equal to the population mean divided by N.*

The sample average is the minimum variance linear unbiased estimator of the population mean- we will prove this later as a special case of the Gauss Markov Thereom with X only a constant term.

However, it is possible for a biased estimator to have lower variance. To compare bias and variance differences, we can use the mean square error criterion:

Mean Squared Error:

$$\begin{aligned} E((T_N - \theta)^2) &= V(T_N) + (\theta - E(T_N))^2 \\ &= variance + bias^2 \end{aligned}$$

If restricted to unbiased estimators, the MSE is only determined by the variance. If no restrictions, can trade off bias and variance.

Large Samples-

We often consider large sample or asymptotic results as it is too difficult to derive the distribution of T_N in finite samples. In large samples, consistency is the analog of unbiasedness (does the estimator converge to its true value) and the limiting distribution of the estimator, and particular the variance of the limiting distribution, the analog of variance in the finite sample case.

2 Method of Moments

Remember- moment can be any expectation- nothing special!

In general, the method of moments proceeds by specifying a set of population moments that include functions of the data and parameters to zero, as below:

$$E(g(y, x, \beta)) = 0$$

a J by 1 set of moment restrictions.

Almost every estimator can be written as some kind of method of moments estimator.

We have already seen one example for the BLP:

$$E(X'(Y - X\beta)) = 0$$

The first question is the identification question. In the population, is there a unique solution to this equation? If there are multiple solutions, the model is not identified.

For example, in the BLP case X must have full rank for identification- so can invert $E(X'X)$ and solve for β.

If the model is identified, we just use the analogy principle to form the sample moments:

$$\begin{aligned} E_N(g(y, x, b_N) &= 0 \\ \frac{1}{N}\sum_i g(y_i, x_i, b_N) &= 0 \end{aligned}$$

If $J = k$- so we have as many equations as parameters- just solve out the system. If $J > k$- so have more equations than parameters- may be no unique solution in sample (though must exist in population). Will cover this case if get to GMM at end. (Basic idea- change analogy to minimize squared moments- zero at true parameter in population and positive elsewhere).

What do you do if $J > k$? One solution- take k of the moment conditions and only use those- then we are in previous case. Problem- not using some part of the statistical model- why throw out restrictions?

Another case- weight the moment conditions in such a way so only use k. Example: Lets say we have two moment conditions for one parameter. That is we have:

$$\begin{aligned} E(g_1(y, x, \beta_1)) &= 0 \\ E(g_2(y, x, \beta_1)) &= 0 \end{aligned}$$

We could use either one of these moment conditions in isolation, or instead use a weighted average of them:

$$\frac{1}{2}E(g_1(y, x, \beta_1)) + \frac{1}{2}E(g_2(y, x, \beta_1)) = 0$$

Making this more general, we can have:

$$aE(g(y_i, x_i, \beta) = 0$$

where a is k by J, E() is J by 1 (in the population). Thus a transforms the over identified case to the just identified case. In the example above, a was:

$$a = 1/2 \quad 1/2$$

or equal weighting of the moments.

If we use one moment condition in isolation, its equivalent to the following weighting matrix:

$$a = 1 \quad 0$$

In the just identified case, there is no need for a weighting matrix as we have just enough moments. You may have heard of Hansen's GMM- Generalized Method of Moments. GMM is really valuable in the overidentified case where we need to decide on the weighting matrix to use. Hansen derives the optimal weighting matrix to use- which depends on the asymptotic variance of the moment conditions. The intuition here is to weight the high variance moments less and the low variance moments more. He also shows that one can test using the overidentification restrictions. Basically, imagine we only use one moment condition. In the population, the second moment condition should be zero under the null. So under the null that the model is correct, the other moment condition should be close to zero in the sample. If its far away from zero, our model is not very good!! One can test the model using this logic. We cant use the moment we used for estimation to test as it was set to zero in the estimation procedure.

With a weighting matrix its more complicated- but we are basically testing the model with the part of the moment conditions that we havent used for estimation.

Fun example: Ed Prescott, Progress vs. Regress, macro models fail the overidentification tests.

I will now show that the GMM estimator is consistent and asymptotically normal. Using the Mean Value Theorem, we have that:

$$\frac{1}{N}\sum_i g(y_i, x_i, b_N) \approx \frac{1}{N}\sum_i g(y_i, x_i, \beta_0) + \frac{1}{N}\sum_i \frac{dg(y_i, x_i, b_N)}{d\beta}(b_N - \beta_0)$$

By the sample estimation,

$$\frac{1}{N}\sum_i g(y_i, x_i, b_N) = 0$$

As the sample size gets large, $\frac{1}{N}\sum_i g(y_i, x_i, \beta_0) \to_p E(g(y_i, x_i, \beta_0)) = 0$ by the LLN. But then $\frac{1}{N}\sum_i \frac{dg(y_i, x_i, b_N)}{d\beta}(b_N - \beta_0) \to_p 0$. This will mean that in general $(b_N - \beta_0) \to_p 0$.

Not the greatest proof!

By the CLT,

$$\frac{1}{\sqrt{N}}\sum_i g(y_i, x_i, \beta_0) \quad \to_d \quad N(0, V)$$

Multiplying by $\sqrt{N}$ and combining results, we have that:
Let

$$(\frac{1}{N}\sum_i \frac{dg(y_i, x_i, \beta_0)}{d\beta})^{-1} = d_N^{-1}$$

By the LLN and Slutsky, $d_N^{-1} \to_p d^{-1}$

$$\begin{aligned}\sqrt{N}(b_N - \beta_0) &\approx (-\frac{1}{N}\sum_i \frac{dg(y_i, x_i, b_N)}{d\beta})^{-1}\frac{1}{\sqrt{N}}\sum_i g(y_i, x_i, \beta_0) \\ &\to_d N(0, d^{-1}Vd^{-1})\end{aligned}$$

If we have a selection matrix a, this becomes:

$$\begin{aligned}\sqrt{N}(b_N - \beta_0) &\approx a_N(\frac{1}{N}\sum_i \frac{dg(y_i, x_i, b_N)}{d\beta})^{-1}a_N\frac{1}{\sqrt{N}}\sum_i g(y_i, x_i, \beta_0) \\ &\to_d N(0, (ad)^{-1}a'Va(ad)^{-1})\end{aligned}$$

Hansen then shows that under optimal GMM, $a = d'V^{-1}$. The asymp variance becomes $(dV^{-1}d)^{-1}$.

Example 7. Bivariate Regression Model

$$\{\alpha, \beta\} = \arg\min_{a,b} E((Y - (a + bX))^2)$$

Here we are regressing Y on a constant and scalar X. We can then derive the first order conditions:

$$\begin{aligned}-2E(Y - (a + bX)) &= 0 \\ -2E(X(Y - (a + bX)) &= 0\end{aligned}$$

We can define U to be the prediction error so $U = (Y - (a + bX)$. We then have two moments: $E(U) = 0$ and $E(XU) = 0$. Since there are two parameters to estimate, α and β, we have the just identified case. Also notice- if we include a constant, the average prediction error is zero- this is why you should always include a constant in the regression!

Using these two moments:

$$
\begin{aligned}
E(U) &= 0 \\
\alpha &= E(Y) - \beta E(X) \\
E(XU) &= 0 \\
Cov(X,U) &= E(XU) - E(X)E(U) \\
&= 0 \\
Cov(X,U) &= Cov(X, Y - (\alpha + \beta X)) \\
&= Cov(X,Y) - \beta Var(X) \\
\beta &= \frac{Cov(X,Y)}{Var(X)} \\
\alpha &= E(Y) - \frac{Cov(X,Y)}{Var(X)} E(X)
\end{aligned}
$$

Thus, if X and Y positively covary, the slope of β is positive. Notice $Var(X) \neq 0$- this is the rank condition- in this case that there is no perfect linear relationship between X and a constant. If there was that means that X doesnt vary- so its variance is zero. In the multivariate case we would use the moment $E(X'X)$ has to be invertible.

Example 8. Gamma Distribution

The gamma distribution has density

$$\frac{x^{\alpha-1}e^{-x/\beta}}{\beta^{\alpha}\Gamma(\alpha)}$$

for $x > 0$. We can start with the population moment conditions:

$$
\begin{aligned}
E(X) &= \alpha\beta \\
E(X^2) &= \beta^2\alpha(\alpha+1)
\end{aligned}
$$

Since there are two equations, two unknowns, there should be a unique solution to these equations. Using the analogy principle, we then use the sample equivalents:

$$
\begin{aligned}
\frac{1}{N}\sum_i x_i &= \alpha_N\beta_N \\
\frac{1}{N}\sum_i x_i^2 &= \beta_N^2\alpha_N(\alpha_N+1)
\end{aligned}
$$

and solve them for α and β. And we are done!

But so far we have only used information from the mean and variance of the distribution. To improve our estimates, we could also use information from

the skewness and kurtosis of the distribution (which are the third and fourth normalized moment resp., skewness is a measure of whether one tail of the distribution is bigger than the other, while kurtosis is how "peaked" the distribution is- how much of the variance is coming from the tails vs. the center.) Then we get two more moments:

$$\begin{aligned} E((\frac{x-\mu}{\sigma})^3) &= \frac{2}{\sqrt{\alpha}} \\ E((\frac{x-\mu}{\sigma})^4) &= \frac{6}{\alpha} \end{aligned}$$

We can then form these moments in the sample:

$$\begin{aligned} \frac{1}{N}\sum_i (\frac{x_i-\bar{x}}{\bar{\sigma}})^3 &= \frac{2}{\sqrt{\alpha}} \\ \frac{1}{N}\sum_i (\frac{x_i-\bar{x}}{\bar{\sigma}})^4 &= \frac{6}{\alpha} \end{aligned}$$

Notice we could take any combination of these moments, except just the last two, to estimate α and β. If we decided to weight the moments, we would have a 2 by 4 weighting matrix:

$$W = \begin{matrix} w_{11} & w_{12} & w_{13} & w_{14} \\ w_{21} & w_{22} & w_{23} & w_{24} \end{matrix}$$

For example, the first moment could just be the mean and the second moment include $E(X^2)$, the skewness, and kurtosis.

This approach is often used to estimate complicated economic models- particular parameters in the model are set equal to moments from the data. For example, in a neoclassical production function $Y = Ak^\alpha l^{1-\alpha}$ for the economy, α would be set to the capital share of GDP. Other reasons are to avoid specifying everything going on in the model- if one is not sure about the distributions of those parts of the model. For ex, dont know how stock prices move so dont want to model their movements econometrically- use other parts of the model to identify parameter. Calibration!

3 Maximum Likelihood

We can think of many different stages of knowledge about the conditional prediction problem, $P(Y|X)$:

1. We know nothing about the distribution. In this case we can do nonparametrics to estimate some functionals of this distribution, such as the conditional expectation, or estimate the BLP to find a linear approx to the CEF.

2. We know some shape restrictions on the distribution. For example, we know that the CEF is linear- in which case we do OLS to get the CEF. Or we know that certain moments are equal to zero in the population- in which case we might try method of moments.

3. We know the entire conditional distribution up to some set of unknown parameters θ- then we do maximum likelihood.

Maximum likelihood begins by defining the likelihood function, the probability that the data was generated by a given model, where the model is specified by a density f known up to parameters θ:

$$l = f(y;\theta)$$

For example, if y was distributed normally with mean μ, variance σ^2 f would be the normal density with θ including both parameters. We now consider what happens if we change θ of the likelihood function- intuitively the likelihood function should be higher at the true value of the parameters θ_0 for any given y generated from the model with θ_0. I will prove this now.

We often use the log likelihood rather than the likelihood as it is more convenient (sum instead of product of density functions)- in the sample these both produce the same estimator as the log is a monotone transformation.

3.1 Identification

Let the population y be generated from the density $f(y,\theta_0)$. Then:

$$\begin{aligned}\theta_0 &= \arg\max_c \int \log f(y,c) f(y,\theta_0) dy \\ &= \arg\max_c E_{\theta_0}[\log f(y,c)]\end{aligned}$$

Here the expectation is taken over the density wrt θ_0 because the data comes from that density.

Proof. Let us compare any parameter vector c to the true parameter vector θ_0. If θ_0 does maximize the function then $E_{\theta_0}[\log f(y,c)] < E_{\theta_0}[\log f(y,\theta_0)]$.

$$\begin{aligned}E_{\theta_0}[\log f(y,c)] - E_{\theta_0}[\log f(y,\theta_0)] &= E_{\theta_0}[\log \frac{f(y,c)}{f(y,\theta_0)}] \\ E_{\theta_0}[\log \frac{f(y,c)}{f(y,\theta_0)}] &\leq \log E_{\theta_0}[\frac{f(y,c)}{f(y,\theta_0)}] \\ E_{\theta_0}[\frac{f(y,c)}{f(y,\theta_0)}] &= \int [\frac{f(y,c)}{f(y,\theta_0)}] f(y,\theta_0) dy \\ &= \int f(y,c) dy = 1\end{aligned}$$

where the inequality is Jensen's inequality:

If $g(.) : \Re \to \Re$ is concave, then for any random variable x for which $E|x| < \infty$ and $E|g(x)| < \infty$,

$$g(E(x)) \quad \geq \quad E(g(x))$$

and the last equality because the density function must integrate to 1.

We then have that : □

$$\begin{aligned} E_{\theta_0}[\log \frac{f(y,c)}{f(y,\theta_0)}] &\leq \log 1 = 0 \\ E_{\theta_0}[\log f(y,c)] &\leq E_{\theta_0}[\log f(y,\theta_0)] \end{aligned}$$

Thus, θ_0 is at least one solution to this problem- if it is the only one we have identification. What happens if there are other solutions??

We can also compute the first order conditions to this problem: they form the Zero Expected Score Rule (ZES Rule):

$$E_{\theta_0}[\frac{d \log f(y;\theta_0)}{d\theta}] \quad = \quad 0$$

The zero expected score rule is thus a set of moment conditions, one for each parameter- a special application of Method of Moments!!

3.2 Estimation

Let's assume we have an iid sample from the population. How do we estimate using maximum likelihood?

We can use the analogy principle in two ways here- with the maximization problem or with the ZES Rule:

1. Maximization Problem

$$\begin{aligned} \theta_0 &= \arg\max_c E_{\theta_0}[\log f(y,c)] \\ \theta_N &= \arg\max_c \frac{1}{N}\sum_i \log f(y_i,c) \end{aligned}$$

(Note: this problem solves the same problem as maximizing $\Pi_i f(y_i, c)$)

2. ZES Rule

$$\begin{aligned} E_{\theta_0}[\frac{d\log f(y;\theta_0)}{d\theta}] &= 0 \\ \frac{1}{N}\sum_i Z_i &= 0 \\ Z_i &= \frac{d\log f(y_i,\theta_0)}{d\theta} \end{aligned}$$

Conditional distributions- often in economics we are interested in the conditional distribution of $Y|X$ instead of the joint of X, Y- we do not want to model how X is sampled. As long as the parameters being estimated are only part of the conditional likelihood, we can just maximize the conditional problem:

$$\begin{aligned} \theta_0 &= \arg\max_c E(\log f(y,c|X)) \\ &= \arg\max_c E(\log f(y,c,X)) \end{aligned}$$

If we maximize the likelihood conditional on X for all values of X- will maximize on average too.

Another way to see it- can filter joint distribution into conditional and marginal:

$$\begin{aligned} f(X,Y,\theta) &= f(Y|X,\theta)f(X) \\ \ln f(X,Y,\theta) &= \ln f(Y|X,\theta) + \ln f(X) \end{aligned}$$

as long as the marginal distribution of X does not depend on the parameters estimated, can throw it away as will not affect estimation (for ex, if take foc marginal disappears).

Let use define the information matrix:

$$W = -\frac{dz}{d\theta'} = \frac{d'^2 \ln f(y,\theta)}{d\theta d\theta'}$$

Information Rule: $E(W) = V(Z) = E(ZZ')$. The last holds since $E(Z) = 0$.

Proof. We know that $E(Z) = 0$. The derivative of $E(Z)$ will the also be equal to zero:

$$\begin{aligned} \frac{dE(Z)}{d\theta'} &= 0 \\ &= \int \frac{d[Zf(y,\theta)]}{d\theta'} dy \\ &= \int (\frac{d[Z]}{d\theta'} f(y,\theta) + Z\frac{d[f(y,\theta)]}{d\theta'}) dy \end{aligned}$$

$$
\begin{aligned}
&= \int (\frac{d[Z]}{d\theta'} f(y,\theta) + Z\frac{d[f(y,\theta)]}{d\theta'})dy \\
&= \int -Wf(y,\theta)dy + \int Z\frac{d[\ln f(y,\theta)]}{d\theta'} * f(y,\theta)dy \\
&= \int -Wf(y,\theta)dy + \int Z * Z' * f(y,\theta)dy \\
E(W) &= V(Z)
\end{aligned}
$$

□

I can now show that the MLE estimator is consistent and asymptotically normal (being a little loose):

Use the "proof" from the Method of Moments section to conclude that MLE is consistent and asymptotically normal- because the expectation of the score vector equal to zero is a moment condition. Its asympotic variance will be:

$$
\begin{aligned}
d^{-1}Vd^{-1} &= E(\frac{d[Z]}{d\theta'})^{-1}E(ZZ')E(\frac{d[Z]}{d\theta'})^{-1} \\
&= E(W)^{-1}V(Z)E(W)^{-1} \\
&= E(W)^{-1} = V(Z)^{-1}
\end{aligned}
$$

by the Information Matrix Equality. Intuitively, where the score vector has a lot of variance, or the likelihood a lot of curvature- we have a lot of information to figure out the parameter so should have a low variance. If the likelihood is flat- how can we tell what the parameter is? Draw a picture with flat and curved likelihood.

Cramer Rao lower bound:

Theorem 9. *Information (Cramer-Rao) Lower Bound*

Let θ_0 solve a conditional likelihood problem. $\theta_0 = \arg\max_c E(\log f(y|x,\theta))$. Let t_N be any consistent estimator of θ_0. Then the limiting distribution can not have a smaller variance than I^{-1}.

$$
I = E(\frac{d\log f()}{d\theta}\frac{d\log f()}{d\theta}')
$$

But we have just shown that MLE attains this bound!!

Two consistent estimators of this variance, by the analogy principle, are:

$$
(\frac{1}{N}\sum_i Z_i Z_i')^{-1}
$$

$$
-(\frac{1}{N}\sum_i \frac{dZ_i}{d\theta'})^{-1}
$$

3.3 Why maximum likelihood?

The maximum likelihood estimator turns out to have some nice properties, including:

1. Consistency
2. Asymptotic Normality
3. Asymptotic Efficiency- attains the Cramer-Rao lower bound, so no consistent estimator has a lower asymptotic variance
4. Asymptotic Variance is equal to the inverse of the information matrix:

$$E[\frac{d\log f(y_i,\theta_0)}{d\theta}\frac{d\log f(y_i,\theta_0)}{d\theta}']^{-1}$$

   which makes estimation easy- have already calculated score and dont have to calculate Hessian matrix.
5. However- is generally biased in finite samples
6. Usually should do MLE- dont do it if dont want to model parts of the problem b/c unsure of the model- then go to method of moments using moments that believe
7. We get the entire conditional distribution P(Y|X) (or joint P(Y,X))!

Example 10. Gamma Distribution
Earlier we showed that the gamma distribution has density

$$\frac{x^{\alpha-1}e^{-x/\beta}}{\beta^\alpha\Gamma(\alpha)}$$

Given an iid sample from this distribution, we can form the log likelihood function and maximize it:

$$\begin{aligned} L &= \prod_i \frac{x_i^{\alpha-1}e^{-x_i/\beta}}{\beta^\alpha\Gamma(\alpha)} \\ \log L &= \sum_i \log(\frac{x_i^{\alpha-1}e^{-x_i/\beta}}{\beta^\alpha\Gamma(\alpha)}) \\ &= -N\alpha\log\beta - \log\Gamma(\alpha) + \sum_i(\alpha-1)\log x_i - x_i/\beta \end{aligned}$$

The maximum likelihood estimator is:

$$\{\alpha_N,\beta_N\} = \max -N\alpha\log\beta - \log\Gamma(\alpha) + \sum_i(\alpha-1)\log x_i - x_i/\beta_{\alpha,\beta}$$

We can also form the score vector but will probably be ugly!

Example 11. Classical Normal Regression Model

Let us assume that Y is distributed by $Y = X\beta + \epsilon$, $\epsilon|X \sim N(0, \sigma^2 I)$

Then we can write the likelihood of the OLS model as:

$$\begin{aligned} L(Y|X;\beta,\sigma^2) = L(\beta,\sigma^2) &= \prod_i (2\pi\sigma^2)^{-1/2} \exp\left(-(Y_i - X_i\beta)'(Y_i - X_i\beta)/2\sigma^2\right) \\ &= (2\pi\sigma^2)^{-n/2} \exp\left(-(Y - X\beta)'(Y - X\beta)/2\sigma^2\right) \end{aligned}$$

Now taking the log likelihood:

$$\log L = K - \frac{n}{2}\log\sigma^2 - (Y - X\beta)'(Y - X\beta)/2\sigma^2$$

The maximum likelihood estimator is:

$$\{\beta, \sigma^2\} = \arg\max \log L(\beta, \sigma^2)$$

Taking first order conditions of the log likelihood, we have:

$$\begin{aligned} \frac{d\log L}{d\beta} &= \frac{-2X'(Y - X\beta)}{-2\sigma^2} = 0 \\ X'Y - X'X\beta &= 0 \\ \beta &= (X'X)^{-1}X'Y \end{aligned}$$

$$\begin{aligned} \frac{d\log L}{d\sigma^2} &= -\frac{n}{2}\frac{1}{\sigma^2} + \frac{(Y - X\beta)'(Y - X\beta)}{2(\sigma^2)^2} = 0 \\ \sigma^2 &= \frac{(Y - X\beta)'(Y - X\beta)}{n} = \frac{\hat{\epsilon}'\hat{\epsilon}}{n} \end{aligned}$$

The estimate for β is the OLS estimator- thus the OLS estimator is the MLE if errors are normally distributed with a variance matrix $\sigma^2 I$. The MLE of σ^2 is biased- the standard OLS estimator is different as we will show later!