IV. Estimation Frameworks in Econometrics (GR: Ch. 14-18)

1. Parametric, semiparametric, and nonparametric specifications

2. Maximum likelihood

3. M-estimators and partially adaptive methods of estimation

4. Generalized method of moments

5. Extremum estimation

6. Kernel estimation

7. Empirical likelihood

8. Homework exercises

IV. Estimation Frameworks in Econometrics (GR: Ch. 14-18)

1. Parametric, semiparametric, nonparametric specifications

The literature on econometric modeling and methodologies refer to parametric, semi-parametric, and non-parametric methods. The following table (modified from the table in Mittlehammer, Judge, and Miller) summarizes attributes of these methods.

	Parametric	**Partially adaptive (QMLE)**	**Semiparametric**	**Nonparametric**
Model: Y	$Y_t = h(X_t, \beta) + \varepsilon_t$	$Y_t = h(X_t, \beta) + \varepsilon_t$	$Y_t = h(X_t, \beta) + \varepsilon_t$	$Y_t = h(X_t) + \varepsilon_t$
Systematic component	$h(X_t, \beta) = X_t\beta$ X_t **known and fixe**	$h(X_t, \beta) = X_t\beta$ X_t **known and fixe**	$h(X_t, \beta) = X_t\beta$ X_t **known and fixed**	* $h(X_t)$ **is unspecified** X_t **known and fixed**
Error term	ε_t's **are iid** $N\left[0,\sigma^2\right]$ ε_t's	ε_t's **are iid** $\sim f(\varepsilon,\theta)$	ε_t's are independent with zero mean and variance σ^2	ε_t's are independent with zero mean and variance σ^2
Parameters	$\beta \in R^k, \sigma^2 > 0$	$\beta \in R^k$, θ (*dist. parameters*) *capture skew & kurt*	$\beta \in \mathrm{R}^k, \sigma^2 > 0$	$\sigma^2 > 0$

*** If the function is not linear, then it is of a known functional form with possible unknown parameters in the parametric, QMLE, and semiparametric formulations; however, no functional form is specified in the nonparametric formulation.**

The differences between the approaches depends upon the detail in the model specification. Parametric methods correspond to a complete specification of the functional form of the systematic component as well as a rather restrictive error distribution. The other extreme is a nonparametric specification in which neither the functional form of the systematic component nor the error distribution are specified. In the parametric and semiparametric specifications, the systematic component need not be linear as shown above, but only needs to be a specified functional form with possibly unknown parameters. The partially adaptive formulation is similar to the parametric specification, except that the error distribution may allow for flexibility in the skewness and kurtosis of the error distribution.

The parametric specification is often estimated using methods of maximum likelihood . Kernel methods can be used to estimate the functional form in nonparametric models. Kernel methods can also be used to approximate the unknown error distribution in semiparametric models. An intermediate step between the parametric and semiparametric specifications is that of partially adaptive or quasi maximum likelihood specifications in which the assumption of normally distributed errors is replaced with the assumption of of a more general flexible error distribution.

2. Maximum likelihood estimation

a. Theoretical development and main results

Assume that the the underlying model is given by

$$y_i = h(X_i, \beta) + \varepsilon_i \qquad , i = 1, 2, \ldots, n ,$$

where the disturbances are independently and identically distributed with the probability density function $f(\varepsilon, \theta)$ and θ denotes a vector of unknown distributional parameters. We will review and contrast several alternative estimators. Recall that such a comparison is often based upon their relative bias and variance and whether they are consistent.[1] We now turn to the development of maximum likelihood estimation and then we briefly summarize the notion of extremum estimators.

The likelihood function associated with the random sample $\{y_t | t = 1, 2, \ldots, n\}$ is given by

$$L(\psi;Y) = \prod_{i=1}^{n} f(\varepsilon_i = y_i - h(X_i, \beta); \theta)$$

with corresponding log likelihood function

$$\begin{aligned} \ell(\psi : Y) &= \ln \mathrm{L}(\psi : \mathrm{Y}) = \sum_{i=1}^{n} \ln \mathrm{f}(y_i - h(X_i, \beta), \theta) \\ &= \sum_{i=1}^{n} \ell(\varepsilon_i ; \Psi) \end{aligned}$$

[1] Recall that $\hat{\theta}$ is an *unbiase*d estimator $E(\hat{\theta}) = \theta$.

Consistency requires that $p\lim_{n\to\infty} \hat{\theta} = \theta \text{ or } \lim_{n\to\infty} \Pr\left(\left|\hat{\theta} - \theta\right| \geq \varepsilon\right) = 0 \text{ for any } \varepsilon > 0$

If the variance and bias of $\hat{\theta}$ approach zero, th $\hat{\theta}$ will be consistent; however, there are cases in which the variance/bias may not be defined and the estimator is still consistent. An unbiased estimator is said to be efficient if its variance is less than the variance of any other unbiased estimator. In the case of a vector of estimators, one would investigate whether the difference of the variance covariance matrices os positive definite. In the case of biased estimators, comparisons can be made using the mean squared error which is equal to sum of the variance and square of the bias, i.e. $MSE(\hat{\theta}) = VAR(\hat{\theta}) + BIAS^2(\hat{\theta})$.

where $\ell(\varepsilon_t;\psi)$, $\ell(\varepsilon_i;\psi) = \ln\ f(\varepsilon_i = y_i - h(X_i,\beta);\theta)$
(multivariate pdf) as a product of independent pdf's assumes that the observations are independent. If there is a correlation between the observations, then the likelihood function will be a multivariate pdf such as a multivariate normal, multivariate t, or some other multivariate pdf.

The maximum likelihood estimators (MLE) of the parameters ψ are implicitly defined by the equation

$$\frac{d\ell(Y;\psi)}{d\psi} = \begin{pmatrix} \frac{\partial\ell}{\partial\beta} \\ \vdots \\ \frac{\partial\ell}{\partial\theta} \end{pmatrix} = \begin{pmatrix} 0 \\ \vdots \\ 0 \end{pmatrix} = 0.$$

These estimators will optimize the log likelihood function $\ell(y;\theta)$. It is possible that multiple solutions to will exist; however, a unique maximum will exist if $\ell(Y;\theta)$ is concave in θ. The concavity can be investigated by determining whether the matrix

$$\frac{d^2\ell(Y;\psi)}{d\psi^2} = \begin{pmatrix} \ell_{11} & \cdots & \ell_{1k} \\ \vdots & & \\ \ell_{k1} & \cdots & \ell_{kk} \end{pmatrix}$$

is negative definite where $\ell_{ij} = \dfrac{\partial^2\ell(Y;\psi)}{\partial\psi_i\partial\psi_j}$.

Under rather general regularity conditions the MLE of

$\psi(\textit{the distributional and } \mathrm{model}\,\textit{parameters}),\ \hat{\psi}$

distribution. The regularity conditions will be given and the main results cited. The literature contains a number of alternative formulations of "regularity conditions." The following conditions are fairly typical.

Regularity Conditions:

Let Ψ denote the set of permissible values ψ

(R.1) For almost all y_t

$$\frac{d\ell(y_i;\psi)}{d\psi}, \frac{d^2\ell(y_i;\psi)}{d\psi^2}, \frac{d^3\ell(y_i;\psi)}{d\psi^3}$$

exist for all $\psi \in \Psi$

(R.2) There exist integrable functions $F_1(y)$, $F_2(y)$ and $F_3(y)$ such that

$$\frac{d\ell(y_i;\psi)}{d\psi} < F_1(y_i)$$

$$\frac{d^2\ell(y_i;\psi)}{d\psi^2} < F_2(y_i)$$

$$\frac{d^3\ell(y_i;\theta)}{d\psi^3} < F_3(y_i)$$

for all $\psi \in \Psi$

and

$$E(F_3(y)) = \int_{-\infty}^{\infty} F_3(y)f(y;\theta)dy < M$$

where M is independent of θ.

$$\text{(R.3)} \quad E\left[\frac{\partial\ell(y_i;\psi)}{\partial\psi}\right]^2 = \int_{-\infty}^{\infty}\left[\frac{\partial \ln f(y;\theta)}{\partial\psi}\right]^2 f(y;\theta)dy < +\infty.$$

These assumptions are important in demonstrating the asymptotic normality of MLE of ψ . R.1 insures the existence of necessary Taylor Series expansions while R.2 permits differentiation under an integral and (R.3) implies that the random variable(s) $\left\{\frac{\partial \ln f(y_t;\theta)}{\partial \psi}\right\}$

will have a finite variance.

Theorem. Under the regularity conditions (R.1), (R.2) and (R.3) the MLE of ψ will be consistent and asymptotically normal.

$$\boxed{\psi_{MLE} \overset{a}{\sim} N[\theta,\ \Sigma_{\psi_{MLE}}]}$$

where

$$\Sigma_{\psi_{MLE}} = -\left(\frac{Ed^2\ell}{d\psi d\psi'}\right)^{-1}$$

$$= \left(E\left(\frac{d\ell}{d\psi}\right)\left(\frac{d\ell}{d\psi}\right)\right)^{-1}.$$

This distributional result can be alternatively written

$$\boxed{\sqrt{n}\left(\psi_{MLE} - \psi\right) \overset{a}{\sim} N[0, n\Sigma_{\psi_{MLE}}].}$$

2. A simple example

Recall from the statistics section what an asymptotic distribution is. Example from previous homework. MGF for sample mean (exponential elements) $(1-\beta t/n)^{-n}$

	Exact	Asymptotic
	GA(z; β/n, p =n)	N(β, β^2 /n)
Mean	β	β
Variance	β^2 /n	β^2 /n
Skewness	$2\beta^3/n^2$	0
Kurtosis	$6\ \beta^4/n^3 + 3\ \beta^4/n^2$	$3\ \beta^4/n^2$

3. M-estimators and partially adaptive methods of estimation

A large body of statistics and econometric literature, summarized in Koenker (1982), discusses robust regression estimation. Huber (1981), Hampel (1986) and Rey (1983) overview many topics related to robust statistics. Much of the material in this literature can be structured in terms of influence functions and M-estimators.

In order to consider robust regression, first define the simple regression model as

$$y_i = X_i\beta + \varepsilon_i \qquad , \ i = 1, 2, \ldots, n ,$$

where the y_i and X_i denote the i-th observation on the dependent variable and a 1 x K vector of independent variables. The β represents a K x 1 parameter vector of regressor coefficients. The ε_i are independently, identically distributed random variables and are also independent of X_i.

a. M-Estimators.

As commonly described in the literature, the M-estimation technique selects an estimator of β which minimizes a function of the residuals $\rho(\varepsilon_i)$, i.e.,

$\min_\beta \Sigma\, \rho(Y_i - X_i\beta = \varepsilon_i.)$

The solution to this minimization problem yields an influence function which follows the Hampel (1974) definition. In the notation of the above linear model, the influence function is given by:

$$\psi(\varepsilon_i) = \rho`(\varepsilon_i)$$

The ψ function measures the "influence" that a residual will have in the estimation process.

M-estimators include

Estimator	ρ ( )	ψ ( )
OLS	$\rho(\varepsilon) = \varepsilon^2$	$\psi(\varepsilon) = 2\varepsilon$
LAD	$\rho(\varepsilon) = \lvert\varepsilon\rvert$	$\psi(\varepsilon) = \text{sign}(\varepsilon)$
L_p	$\rho(\varepsilon) = \lvert\varepsilon\rvert^p$	$\psi(\varepsilon) = p\lvert\varepsilon\rvert^{p-1}\,\text{sign}(\varepsilon)$

sign(ε)=1 if (ε>0) and -1 if (ε<0)

Consider the shapes of these influence functions. Note that large disturbances or outliers will have a large impact on the OLS estimators; whereas, LAD is not as sensitive to outliers. The L_p estimators obviously include the OLS and LAD estimators as special cases, corresponding to p=2 and p=1. OLS, LAD, and L_p correspond to MLE with a normal, Laplace, and GED, respectively. Maximum likelihood estimators will be obtained if the ρ-function is the negative of the natural logarithm of the density of the residuals. These estimators are also referred to as quasi-maximum likelihood estimators (QMLE). They may offer some advantages over arbitrarily selected OLS, LAD, L_p or other estimators. However, the optimality properties of partially adaptive estimators is conditional on the degree of agreement between the assumed and actual pdf for the errors.

STATA allows for various forms of M-estimation. The general form of the ***r***obust ***reg***ression command is

```
rreg y x's, options
```

A common variation is the command

```
qreg y x's
```

which yields the LAD, LAE, or MAD estimator obtained by selecting the regression estimates to minimize the sum of absolute values of the residuals.

Options: tune(#) use # as the biweight tuning constant; default is tune(7) meaning 7 times the median absolute deviation from the median residual (MAD). Lower tuning constants downweight outliers rapidly but may lead to unstable estimates (below 6 is not recommended). Higher tuning constants produce milder downweighting.

See rreg postestimation for additional capabilities of estimation commands.

The section of the course, "Nonlinear models", will outline how STATA can be used to obtain additional M-estimators.

b. Partially adaptive estimators.

Partially adaptive estimators allow some additional flexibility to distributional characteristics of the errors if the estimators are defined by

$$\min_{\beta,\varphi} \Sigma\, \rho(\varepsilon_i = y_i - X_i\beta; \theta \quad)$$

$$= -\min_{\beta,\varphi} \Sigma \ln f(\varepsilon_t; \theta \quad)$$

where θ denotes the distributional parameters for the error distribuiton. Thus the estimation procedure may be able to accommodate diverse distributional assumptions. Optimizing the objective function over the regression and distributional parameters allows the influence function to *adapt* to the distributional characteristics. For example, selecting the pdf to be the GED not only includes the OLS, LAD, and L_p estimators, but endogenizes the selection of the parameter *p* if the objective function is optimized over *p.*

Partially adaptive estimation of regression models requires the use of nonlinear optimization algorithms which will be covered in the section on nonlinear models.

4. Generalized Method of Moments

a. Basic Theory

Recall that the **method of moments** (MOM) estimators of the parameters in the data generating process

DGP: f(Y, θ)

are obtained by solving

$$g\left(\hat{\theta}\right) = \hat{m} - m\left(\hat{\theta}\right) \qquad =0$$

where $\hat{m} = \left(\hat{m}_1, \hat{m}_1, \ldots \hat{m}_m\right)'$ & $m(\theta) = \left(m_1(\theta), m_2(\theta), \ldots, m_m(\theta)\right)'$

with $\hat{m}_h = \frac{1}{n}\sum_{i=1}^{n} y_i^h$, (sample moments), and $m_h(\theta) = E\left(Y^h\right)$, (theoretical moments)
h = 1,2, .. , m, the number of moments (m) used in estimation (number of equations) is the same as the number of parameters (p). If a solution exits, then not only will $g\left(\hat{\theta}\right) = \hat{m} - m\left(\hat{\theta}\right) = 0$

but $$g\left(\hat{\theta}\right)'g\left(\hat{\theta}\right) = \left(\hat{m} - m\left(\hat{\theta}\right)\right)'\left(\hat{m} - m\left(\hat{\theta}\right)\right) = 0$$ a

$$g\left(\hat{\theta}\right)'Wg\left(\hat{\theta}\right) = \left(\hat{m} - m\left(\hat{\theta}\right)\right)'W\left(\hat{m} - m\left(\hat{\theta}\right)\right) = 0$$

where W is any positive definite weighting matrix.

Generalized method of moments estimation arises when there are more moment conditions (m) imposed than parameters (p) in the model and there is not a solution to

$$g\left(\hat{\theta}\right) = \hat{m} - m\left(\hat{\theta}\right) \qquad =0.$$

How can you take account of more sample moments than parameters? Generalized method of moments estimators (GMM) estimators are defined by

$$\hat{\theta} = \arg\min_{\theta} g(\theta)'g(\theta) = \left(\hat{m} - m(\theta)\right)'\left(\hat{m} - m(\theta)\right)$$

or more generally by

$$\hat{\theta} = \arg\min_{\theta} g(\theta)'Wg(\theta)$$

where W denotes a positive definite weighting matrix. Both of these quadratic forms will be nonnegative and their value can be used to perform a test of the consistency of the data with the hypothesized model in overidentified cases where there are more moment conditions than parameters (m>p). The GMM estimator of θ in the second equation depends upon the weighting matrix W. W is often selected to minimize the asymptotic variance of the GMM estimator. The W matrix which does this is given by $W = Var\left(g\left(\hat{\theta}\right)\right)^{-1}$.

In this case, $g\left(\hat{\theta}\right)'Var^{-1}\left(g\left(\hat{\theta}\right)\right)g\left(\hat{\theta}\right)$ has an asymptotic $\chi^2(m-p)$ distribution.

We will now consider a simple example and also demonstrate that GMM includes OLS, IV, and MLE as special cases.

b. Examples.

(1) Exponential

The pdf for the exponential is given by $f(y;\beta) = \frac{e^{-y/\beta}}{\beta}$ for y>0 and has me

and variance equal to β and β^2 , respectively.

Estimators:

MLE: $\hat{\beta} = \bar{Y}$

MOM: $\hat{\beta} = \bar{Y}$

What if the sample variance is not equal to the square of the sample mean?

Alternatively, what if the sample variance is not equal to the square of the sample mean and you want the estimator to take account of information about the sample mean and sample variance? GMM is one approach of doing this (two sample moments and one parameter)

GMM: minimize over beta sum of squares or weighted sum of squares

$$H(\beta)=\left(\bar{Y}-\beta, s^2-\beta^2\right)W\left(\bar{Y}-\beta, s^2-\beta^2\right)'$$

$$= \left(\bar{\mathrm{Y}}\text{-}\beta \quad \mathrm{s}^2-\beta^2\right)\begin{bmatrix} w_1^2 & w_{12} \\ w_{21} & w_2^2 \end{bmatrix}\begin{pmatrix} \bar{\mathrm{Y}}\text{-}\beta \\ \mathrm{s}^2-\beta^2 \end{pmatrix}$$

(2) OLS

Let $g(\beta)=X'(Y-X\beta)=X'\varepsilon$, then

$$Var\left(g(\beta)\right)=Var(X'\varepsilon)=\sigma^2\left(X'X\right)=W^{-1}$$

The corresponding minimization problem is to minimize

$$\mathrm{H}(\theta)=\left(\mathrm{g}(\theta)\right)'\mathrm{W}\left(\mathrm{g}(\theta)\right)$$

$$=(Y-X\beta)'X\left(X'X\right)^{-1}X'\left(Y-X\beta\right)/\sigma^2$$

The necessary conditions for a minimum of H( ) are

$$\frac{dH(\beta)}{d\beta}=\frac{-2}{\sigma^2}X'X\left(X'X\right)^{-1}X'\left(Y-X\hat{\beta}\right)=\frac{-2}{\sigma^2}X'\left(Y-X\beta\right)=0$$

which yields the OLS estimator

$$\hat{\beta}=\left(X'X\right)^{-1}X'Y$$

which is based on the estimated covariances of the errors and the x’s being zero.

(3) Instrumental variables

Let $g\left(\beta\right)=Z'\left(Y-X\beta\right)=Z'\varepsilon$

Z_{nXm} denotes a matrix including n observations on m instrumental variables which we would like to be uncorrelated with the estimated error terms. It can be shown that

$$Var\left(g\left(\beta\right)\right)=Var(Z'\varepsilon)=\sigma^{2}\left(Z'Z\right)$$

The corresponding minimization problem is to minimize

$$\mathrm{H}\left(\theta\right)=\left(\mathrm{g}\left(\theta\right)\right)'\mathrm{D}\left(\mathrm{g}\left(\theta\right)\right)$$

$$=(Y-X\beta)'Z\left(Z'Z\right)^{-1}Z'(Y-X\beta)/\sigma^{2}$$

The necessary conditions for a minimum are

$$\frac{dH\left(\beta\right)}{d\beta}=\frac{-2}{\sigma^{2}}X'Z\left(Z'Z\right)^{-1}Z'\left(Y-X\hat{\beta}\right)=0$$

which yields the IV estimator

$$\tilde{\beta}=\left(X'Z\left(Z'Z\right)^{-1}Z'X\right)^{-1}X'Z\left(Z'Z\right)^{-1}Z'Y$$

which is based on the estimated covariances between the instrumental variables and the errors being zero.

(4) Maximum likelihood estimation

The assumed log-likelihood function is given by

$$\ell(\beta) = \Sigma_t \ln f\left(\varepsilon_t = Y - X_t\beta\right)$$

Let $g(\beta) = \dfrac{d\ell}{d\beta} = \Sigma_t \left(\dfrac{d \ln f(\ )}{d\varepsilon_t} \right) X_t^{'}$

with corresponding variance estimated by

$$Var\left(\frac{d\ell}{d\beta}\right) = \Sigma_t \left(X_t X_t^{'}\right)\left(\frac{d \ln f(\ )}{d\varepsilon_t}\right)^2 / n$$

The objective function is given by:

$$H(\theta) = \left(g(\theta)\right)'\left(Var\left(\frac{d\ell}{d\theta}\right)\right)^{-1}\left(g(\theta)\right)$$

The variance matrix involves the unknown parameters; thus, this can be a very nonlinear optimization problem.

c. Comments

(1) GMM encompasses most of the common estimation procedures such as MLE, OLS, IV, and 2SLS.

(2) In the exactly identified case (m=p) we can often get exact answers

(3) In the overidentified case, GMM requires the specification of a weighting matrix case. The estimators will depend upon the weighting matrix.

(4) The optimal weighting matrix minimizes the asymptotic variance of the estimator

(5) Tests of the overidentifying assumptions: The optimized objective function corresponding to the optimal weighting matrix is asymptotically distributed as a chi square with degrees of freedom m-p.

(6) Chamberlain (1987, Journal of Econometrics, 305-334) showed that the class of GMM estimators achieve the semiparametric efficiency bound given the set of moment restrictions.

(7) In the case of overidentification (number of moment restrictions exceed the number of parameters), in some cases GMM estimators exhibit substantial bias with related test statistics being problematical.

(8) Woolridge (2001) survey article, "Applications of Generalized Method of Moments Estimation," explores some different areas in which GMM can be productively applied, including cross-sectional applications, testing functional forms, for censored or truncated regression models, various time series problems, and with panel data sets. However, Woolridge notes that the "majority of empirical researchers in economics probably have never used a GMM procedure."

(9) The method of empirical likelihood appears to have less bias and yields test statistics similar in form to LR, W, and LM statistics. Imbens (2002, JBES, 493-506). Imbens' paper includes a nice discussion of the method of empirical likelihood.

5. Extremum estimators (GR 6th: 14.5, 7th: 12.5)

a. Basic Theory and results

The book by Huber (Robust Statistics) discusses properties of more general extremum estimators which include the case of "MLE estimators." Consider the extremum estimator $\left(\hat{\theta}\right)$ defined by

$$\hat{\theta} = \arg\max_{\theta\in\Theta} H(\theta : data) \qquad ,$$

the estimator maximizes a function of the parameters, conditional on the data. Under fairly general regularity conditions the asymptotic distribution of the extremum estimator, $\hat{\theta}$, is given by $\sqrt{n}\left(\hat{\theta}-\theta\right) \sim N[0,C]$

where $C = A^{-1}\, B\, A^{-1}$, $A = lim_{n\to\infty}\ E\left[\frac{1}{n}\frac{d^2H(\ )}{d\theta d\theta'}\right]$, and

$$B = lim_{n\to\infty}\ \left\{Var\left(\frac{1}{\sqrt{n}}\frac{dh(\ )}{d\theta}\right) = E\left[\frac{1}{n}\frac{dh(\ )}{d\theta}\frac{dh(\ )}{d\theta'}\right]\right\}$$

The variance-covariance matrix *C* is often referred to as the ***sandwich estimator***.

Proof: Consider the Taylor series expansion of $\frac{dH(\ )}{d\theta}$, the derivative of the objective function,

$$\frac{dH(\ )}{d\theta}|_{\hat{\theta}} = 0 = \frac{dH(\ )}{d\theta}|_{\theta_0} + \frac{d^2H(\ )}{d\theta d\theta'}|_{\theta_*}\left(\hat{\theta}-\theta_0\right)$$

where θ_* lies betwe $\hat{\theta}$ and θ_0 .

Since the extremum estimator is defined by $\frac{dH(\;)}{d\theta}|_{\hat{\theta}}=0$,

we can express $\left(\hat{\theta}-\theta_0\right)$ as

$$\sqrt{n}\left(\hat{\theta}-\theta_0\right)=-\left(\frac{1}{n}\frac{d^2H(\;)}{d\theta d\theta'}|_{\theta_*}\right)^{-1}\left(\frac{1}{\sqrt{n}}\frac{dH(\;)}{d\theta}|_{\theta_0}\right)$$

where the regularity conditions are sufficient to insure that the second term on the right hand side converges to

$$N\left[0,\ B\ \right]$$

and the desired result follows from an application of the *Useful Theorem*.

b. Special cases:

(1) OLS

Let $H(\beta)=-SSE(\beta)=-(Y'Y-2\beta'X'Y+\beta'X'X\beta)$

$$\frac{dH}{d\beta}=2X'\varepsilon$$,

$$Var\left(\frac{dH}{d\beta}\right)=4\sigma^2(X'X)$$ =B

$$\frac{d^2H}{d\beta d\beta'}=-2(X'X)$$ =A

$$Var\left(\hat{\beta}\right)=A^{-1}BA^{-1}=\left(-2\left(X'X\right)\right)^{-1}4\sigma^{2}\left(X'X\right)\left(-2\left(X'X\right)\ \right)^{-1}$$

$$=\sigma^{2}\left(X'X\right)^{-1}$$

(2) MLE

Let $H\left(\hat{\theta}\right)=\ell\left(\hat{\theta}\right)$.

If the model is correctly specified, B=-A, and the sandwich estimator gives the same results as MLE, the Cramer-Rao Matrix.

(3) QMLE

Quasi Maximum likelihood estimation ("MLE" with the wrong pdf) can be included in this framework. In the context of a QMLE framework a specification test (Hal White) can be performed by testing A(θ) + B(θ) = 0, using $\mathrm{A}\left(\hat{\theta}\right)+\mathrm{B}\left(\hat{\theta}\right)$.

The expectations are taken with respect to the true distribution of Y. In the case of correct model specification A(θ) = -B(θ), see Theil (pp. 384-6) and Appendix B of the Statistics secion. In this case, C(θ) simplifies to the the result for MLE. In this case B(θ) provides an alternative estimator of the variance covariance matrix which is based on the first derivatives of ℓ(Y;θ). This estimator is often useful in empirical work.

(5) GMM

Extremum estimators also include GMM as special cases by defining

$$H(\theta) = -g(\theta)'Wg(\theta) \qquad (\theta)$$

6. Kernel or adaptive estimation

a. Semiparametric applications

Kernel estimators can be used to obtain semiparametric or nonparametric estimators of the relationship between variables. We first consider their use as semiparametric estimators of a known functional form for a regression relationship with an unknown error distribution. We will then consider the use of kernels to estimate an unknown relationship between variables with an unknown error distribution. These methods may require large samples for the estimators to reflect desirable large sample properties.

As discussed earlier, kernel estimators make use of an approximating pdf

$$f(\varepsilon) = \left(\frac{1}{sn}\right) \sum_{t=1}^{n} K\left(\frac{\varepsilon - e_t}{s}\right)$$

where K and e_t denote the a specified kernel pdf and least squares residuals, $e_t = Y_t - X_t\hat{\beta}$, respectively. K ( ) can be any pdf, but is often selected to be the standard normal. *s* is a smoothing parameter. Some variations of this procedure use of a trimming parameter. Silverman suggests selecting the smoothing parameter to be to $1.06\hat{\sigma}n^{-2}$. The approximating pdf, f( ), can be thought of as a smoothed histogram.

"MLE" is then performed using the kernel in the likelihood function,

$$\ell(\beta:Y,s,K)=\ln \mathrm{L}(\beta\mathrm{:Y,s,K})=\sum_{i=1}^{n}\ln \mathrm{f}(y_i-X_i\beta,K,s)$$

where the estimators are conditional on the smoothing parameter (s) and the selected kernel (K). These estimators are also conditional on the OLS estimated parameters to obtain the estimated disturbances. An alternative approach is to iterate the procedure. Start with the OLS residuals and obtain estimates of β . Given the new estimate β , calculate the corresponding residuals, then re-optimize over β . This procedure until convergence is achieved. Kernel estimators are quite robust and, at least in this form, apply to the classical linear regression model with independently (no autocorrelation) and identically (homoskedasticity) error terms. Variations can be used for multiple regressors and to cover departures from the assumptions independence and identically distributed errors a. The interested reader is encouraged to review the paper by Hsieh and Manski [1987]. Pagan and Ullah [1999] provide an excellent introduction to many related techniques and application. An alternative adaptive estimator is discussed in Newey [1988].

b. Nonparametric application

(Reference: Wikipedia, the free encyclopedia and Greene (14.4.2))

Kernel regression methods are nonparametric procedures to estimate the conditional expectation of a random variable. These conditional expectations can involve nonlinear relations between random variables and can be written as

$$E(Y|X)=m(X) \quad .$$

The Nadaraya-Watson estimator of $m(X)$ is defined by

$$\hat{m}(x) = \frac{\sum_{i=1}^{n} K_h\left(\frac{x - X_i}{h}\right) Y_i}{\sum_{i=1}^{n} K_h\left(\frac{x - X_i}{h}\right)}$$

where $K_h(\ )$ denotes a kernel with window width *h*. While alternate regression estimators, such as the Priestly-Chao or Gasser-Muller kernel estimators have been considered, the Nadaraya-Watson estimator is one of the most commonly used.. The Nadaraya-Watson kernel estimator can be obtained by evaluating

$$E(Y|X) = m(X)] = \int y f(y|x) dy = \int y \frac{f(x,y)}{f(x)} dy$$

where kernel estimates are used for $f(x,y)\ and\ f(x)$

$$\hat{f}(x,y) = \frac{\sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right) K\left(\frac{y - y_i}{h}\right)}{nh^2}$$

$$\hat{f}(x) = \frac{\sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right)}{nh}$$

The **STATA command** to implement kernel regression estimation is

kernreg yvar xvar [if exp] [in range], bwidth(#) kercode(#) npoint(#)
[gen(mhvar gridvar) nograph graph_options]

Description: kernreg calculates the Nadaraya-Watson nonparametric regression. By default, kernreg draws the graph of the estimated conditional mean over the grid points used for calculation connected by a line without any symbol.

Options:

bwidth(#) specifies the smoothing parameter (bandwidth or halfwidth) of the kernel density estimation for xvar. This parameter defines the width of the weight function window around each grid point.

kercode(#) specifies the weight function (kernel) to calculate the required univariate densities according to the following numerical codes:

1 = Uniform
2 = Triangle
3 = Epanechnikov
4 = Quartic (Biweight)
5 = Triweight
6 = Gaussian
7 = Cosinus

npoint(#) specifies the number of equally spaced points (which define a grid) in the range of xvar used for the regression estimation.

gen(mhvar gridvar) creates two variables containing the estimated regression (conditional mean) values and the corresponding grid points, respectively.

nograph suppresses the graph.

graph_options are any of the options allowed with graph, twoway.

Remarks: bwidth, kercode, and nppoint are not optional. If the user does not provide them, the program halts and displays an error message on screen.

This program uses kernel density estimators modified from Salgado-Ugarte, et al. (1993) and based on the equations provided by Haerdle (1991) and Scott (1992). The smoothness of the resulting estimate can be regulated by changing the bandwidth: wide intervals produce smooth results; narrow intervals give noisier estimates.

Except for the Gaussian kernel, all the functions are supported on [-1,1].

While using the gen option, if the number of cases is less than npoint then the program sets the number of the observations = npoint to obtain the full set of estimations.

This procedure can be regarded as a descriptive smoother of scatterplots as well as a nonparametric regression estimator (Nadaraya-Watson).

Examples:

```
kernreg wait dura, bwidth(0.65) kercode(4) npoint(100)
kernreg accel time, b(2.4) k(4) np(100) gen(m2p4 g2p4) nog
```

A recent reference: **The July 2010 issue of the *Journal of Econometrics* is entitled "Nonlinear and nonparametric methods in econometrics and includes a variety of papers dealing with applications of nonparametric methods in econometrics.

c. An application with parametric and nonparametric components

A researcher may have a particular interest in modeling the relationship between a dependent variable (y) and several independent variables (X's and z), but has a particular interest in exploring the functional relationship between the dependent variable and one of the independent variables(z). This situation might be modeled as

$$y_t = X_t\beta + h(z_t) + \varepsilon_t$$

where the form of h(z) is unspecified. Kennedy (2008), Robinson (1988), and Anglin and Gencay (1996) suggest methods which are similar to estimating the relationship between y and X using a traditional method such as least squares. The unknown form for h(z) is estimated by using Kernel methods to estimate the relationship between the residuals from the first stage and the variable(s) z.

7. Empirical likelihood estimation

While GMM estimators have desirable large sample properties, their small sample properties can be improved upon by considering different estimation procedures. In particular, some applied studies have shown that GMM may be strongly biased in certain applications. Variations of GMM which update the weighting matrix, Continues updating estimators (CUE), appear to reduce the finite sample bias. Another family of estimation procedures, generalized empirical likelihood (GEL), also appears to reduce estimator bias. EL, GMM, CUE, and an estimator called exponential tilting (ET), share a common structure. We will use the notation used by Whitney and Smith (2004) to explore these interrelationships. .

Let z_i (i=i,...,n) be independent and identically distributed observations on a data vector. Further let β denote a *px1* parameter vector a $g(z,\beta)$ be an *mx1* vector of functions where the number of functions is greater than or equal to the number of parameters, $m \geq p$. Further assume that $E_z\left[g(z,\beta_0)\right]=0$ for the β_0 parameter vector

The GMM estimator is defined by

$$\hat{\beta}_{GMM} = \arg\min_{\beta \in B} \hat{g}(\beta)'\Omega(\tilde{\beta})\hat{g}(\beta)$$

where $\tilde{\beta}$ denotes a preliminary estimate β to calculate the weighting matrix. A variation of the GMM estimator is the continuous updating estimator (CUE) which updates the weighting matrix in the objective function and is defined by

$$\hat{\beta}_{CUE} = \arg\min_{\beta \in B} \hat{g}(\beta)'\Omega(\beta)\hat{g}(\beta)$$

To define empirical likelihood (EL) and exponential tilting (ET) estimators, consider a concave function $\rho(\nu)$ on its domain, an open interval containing zero. Let

$\hat{\Lambda}_n(\beta)=\{\lambda:\lambda' g_i(\beta)\in\nu\}\lambda$. The generalized empirical likelihood (GE

defined by the solution to

$$\hat{\beta}_{GEL}=\arg\min_{\beta\in B}\sup_{\lambda\in\hat{\Lambda}}\sum_{i=1}^{n}\rho(\lambda' g_i(\beta))$$.

The exponential tilting estimator corresponds to a special case of the GEL estimator where $\rho(\nu)=-e^{\nu}$, Kitamura and Stutzer (1997) and Smith (1997). The empirical likelihood estimator is also a special case of GEL where $\rho(\nu)=\ln(1-\nu)$, Qin and Lawless (1994). $\rho(\nu)$ is quadrati $\hat{\beta}_{GEL}=\hat{\beta}_{CUE}$

An interesting interpretation of the GEL estimators is available by considering

. $\hat{\pi}_i=\dfrac{\rho_1(\hat{\lambda}'\hat{g}_i)}{\sum_{i=1}^{n}\rho_1(\hat{\lambda}'\hat{g}_i)}$ $\rho_1(\ )=\dfrac{\partial\rho(\nu)}{\partial\nu}$ 'here .

The $\hat{\pi}_i$ sum to one and satisfy $\sum_{i=1}^{n}\hat{\pi}_i\hat{g}_i=0$ $\hat{\pi}_i$ Thus, the ca empirical probabilities for the observations.

Minimum discrepancy (MD) estimators are defined as the solution to the optimization problem

$$\bar{\beta}=\arg\ \min_{\beta\in B,\pi_i' s}\sum_{i=1}^{n}h(\pi_i)$$ and

subject to $\sum_{i=1}^{n}\pi_i g_i(\beta)=0$ $\sum_{i=1}^{n}\pi_i=1$ and .

Selecting $h(\pi)$ $=-\ln(\pi)$ $=-\pi\ln(\pi)$ or yields EL and ET, respective

selecting $h(\pi)=\dfrac{n^2\pi^2-1}{2n}$ yields CUE.

Thus the EL estimator can be seen to solve the following optimization problem:

$$\bar{\beta} = \arg \ \max_{\beta \in B, \pi_i' s} \sum_{i=1}^{n} \ln\left(\pi_i\right)$$

subject to $\sum_{i=1}^{n} \pi_i g_i\left(\beta\right) = 0$ $\sum_{i=1}^{n} \pi_i = 1$ and .

In a recent paper Chausse' presents an R program which can be used to obtain GMM and GEL estimators.

Guggenberger (2008) uses Monte Carlo methods to compare the finite sample properties of the GEL, CUE, and various k-class estimators based on sample median, mean, meas squared errors and coverage probabilities and length of confidence intervals. The results suggest that GEL and LIML behave similarly, having thick tails.

References:

Chausse', P., "Computing Generalized Method of Moments and Generalized Empirical Likelihood with R,' Rmetrics Workshop, 2009. (Listed in 588 resources)

Guggenberger, P., "Finite Sample Evidence Suggesting a Heavy Tail Problem of the Generalized Empirical Likelihood Estimator," *Econometric Reviews*, 26 (2008), 526-541.

Kitamura, Y. And M. Stutzer, "An Information-Theoretic Alternative to Generalized Method of Moments Estimation," *Econometrica*, 65 (1997), 861-874.

Newey, W. K. And R. J. Smith, "Higher Order Properties of GMM and Generalized Empirical Likelihood Estimators," *Econometrica, 72 (2004), 219-255.*

Smith, R. J. , "Alternative Semi-parametric Likelihood Approaches to Generalized Method of Moments Estimation," *Economic Journal*, 107(1997), 503-519

Qin, J. And J. Lawless, "Empirical Likelihood and General Estimating Equations," *Annals of Statistics*, 22(1994), 300-325.

8. Homework problems

1. Consider the model $Y = X\beta + \varepsilon$ $Y, X, \beta,$ and ε denote *nx1*, *nxk*, *k*

matrices. Additionally, let *Z* denote a matrix of *nxm* observations on *m* instrumental variables.

a. Derive the IV/GMM estimator which maximizes $H(\beta)$ where

$$\begin{aligned} H(\beta) &= -((Z'\varepsilon)'W(Z'\varepsilon)) = -((Y - X\beta)'ZWZ'(Y - X\beta)) \\ &= -(Y - X\beta)'ZWZ'(Y - X\beta) \end{aligned}$$

The estimator will depend upon the weighting matrix, W, which is assumed to be positive

definite. The optimal weighting matrix is given by $W = (Var(Z'\varepsilon))^{-1} = (Z'Var(\varepsilon)Z)^{-1}$

For parts (b) and (c) the following results will be helpful.

Note: $\dfrac{dH}{d\beta} = 2X'ZWZ'(Y - X\beta) = 2X'ZWZ'\varepsilon$

$$\frac{dH}{d\beta}\frac{dH}{d\beta'} = 4X'ZWZ'\varepsilon\varepsilon'ZWZ'X$$

$$\frac{d^2H}{d\beta d\beta'} = -X'ZWZ'X$$

b. If $Var(\varepsilon) = \sigma^2 I$, demonstrate that the optimal weighting matrix is give by

$W = (Z'Z)/\sigma^2$ and evaluate the corresponding IV estimator and the Sandwich estimator o

variance.

c. . If $Var(\varepsilon) = \Sigma$, demonstrate that the optimal weighting matri: $W = \left(Z'\Sigma Z\right)^{-1}$

and evaluate the corresponding IV estimator and the sandwich estimator of its variance.

d. Derive the sandwich estimator of the variance of the OLS estimator when $Var(\varepsilon) = \Sigma$.

Hint: Let $H(\beta) = -(Y - X\beta)'(Y - X\beta)$. Evaluate and take expected values of

$$\frac{dH(\beta)}{d\beta}, \frac{dH(\beta)}{d\beta}\frac{dH(\beta)}{d\beta'}, \text{ and } \frac{d^2H(\beta)}{d\beta\, d\beta'}.$$

This estimator should look familiar from Econ 388 and is what the reg y x's and newey y x's, lags(#) estimate when heteroskedasticity and/or autocorrelation of the errors are present.

e. Consider how you might estimate the sandwich estimator of the OLS estimator derived in 1(d).

(1) Heteroskedasticity

(2) Heteroskedasticity and autocorrelation

2. How can you obtain a sandwich estimator for MLE/QMLE estimators of β in the linear regression model with flexible error distributions?

3. Consider a model for panel data,

$$y_1 = X_1\beta + \varepsilon_1$$
$$.$$
$$.$$
$$y_n = X_n\beta + \varepsilon_n$$

where y_i, X_i, β, and ε_i , respectively denote *Tx1, Txk, kx1, and tx.* βıatrices. Note that

the same for each panel. This system of equations can be written as

$$Y = X\beta + \varepsilon$$

where $Y = \begin{pmatrix} y_1 \\ . \\ . \\ y_n \end{pmatrix}$, $\mathrm{X} = \begin{pmatrix} X_1 \\ . \\ . \\ X_n \end{pmatrix}$, $\beta = \begin{pmatrix} \beta_1 \\ . \\ . \\ \beta_k \end{pmatrix}$, $\varepsilon = \begin{pmatrix} \varepsilon_1 \\ . \\ . \\ \varepsilon_n \end{pmatrix}$

where *Y, X*, and ε will be *Tnx1* matrices and

$$Var(\varepsilon) = \begin{bmatrix} Var(\varepsilon_1) & \dots & 0 \\ . & & . \\ . & . & . \\ 0 & 0 & Var(\varepsilon_n) \end{bmatrix} = \begin{bmatrix} \Omega_1 & \dots & 0 \\ . & & . \\ . & . & . \\ 0 & 0 & \Omega_n \end{bmatrix} = \Omega$$

is *Tn x Tn.*

Demonstrate that the OLS estimator and corresponding sandwich estimator can be written as

$$\hat{\beta} = \left(\sum_{i=1}^{n} X_i' X_i\right)^{-1} \left(\sum_{i=1}^{n} X_i' Y_i\right)$$

$$Sandwich\ estimator = \left(\sum_{i=1}^{n} X_i' X_i\right)^{-1} \left(\sum_{i=1}^{n} X_i' \Omega_i X_i\right) \left(\sum_{i=1}^{n} X_i' X_i\right)^{-1}$$

This estimator is related to what are referred to as clustered standard errors,

see C. B. Hansen, “Asymptotic Properties of a Robust Variance Matrix Estimator for Panel Data when T is Large,” *Journal of Econometrics*, 141 (2007), 597-620.

Homework key:

1. Extremum, IV, and GMM estimation

a. Derive the IV/GMM estimator

(points)

(1) Min $H = (Y - X\beta)'ZWZ'(Y - X\beta)$

$$\frac{dh}{d\beta} = -2X'ZWZ'(Y - X\beta) = -2X'ZWZ'\epsilon = 0$$

$$\rightarrow \hat{\beta}_{IV} = (X'ZWZ'X)^{-1}X'ZWZ'Y$$

b. Special case of $Var(\varepsilon) = \sigma^2 I$

Substitute $(\sigma^2(Z'Z))^{-1} = W$ into the equation for the IV estimator

$$\rightarrow \hat{\beta}_{IV} = (X'Z(Z'Z)^{-1}Z'X)^{-1}X'Z(Z'Z)^{-1}Z'Y$$

Sandwich estimator:

$$\frac{d^2H}{d\beta d\beta'} = 2X'ZWZ'X = \frac{2}{\sigma^2}X'Z(Z'Z)^{-1}Z'X$$

$$\left(\frac{dH}{d\beta}\right)\left(\frac{dH}{d\beta'}\right) = 4X'ZWZ'\epsilon\epsilon'ZWZ'X = \frac{4}{\sigma^4}X'Z(Z'Z)^{-1}Z'\epsilon\epsilon'Z(Z'Z)^{-1}Z'X$$

$$\rightarrow E\left[\left(\frac{dH}{d\beta}\right)\left(\frac{dH}{d\beta'}\right)\right] = \frac{4}{\sigma^4}X'Z(Z'Z)^{-1}Z'X$$

Sandwich:

$$\left(\frac{d^2H}{d\beta d\beta'}\right)^{-1} E\left[\left(\frac{dH}{d\beta}\right)\left(\frac{dH}{d\beta'}\right)\right]\left(\frac{d^2H}{d\beta d\beta'}\right)^{-1}$$
$$= \frac{\sigma^4}{4}(X'Z(Z'Z)^{-1}Z'X)^{-1}\left(\frac{4}{\sigma^2}\right)(X'Z(Z'Z)^{-1}Z'X)(X'Z(Z'Z)^{-1}Z'X)^{-1}$$
$$= \sigma^2(X'Z(Z'Z)^{-1}Z'X)^{-1}.$$

1.c. Special case of $Var(\varepsilon) = \Sigma$

Substitute $(Z'\Sigma Z)^{-1}$ for W

(1) $$\hat{\beta}_{IV} = (X'ZWZ'X)^{-1}X'ZWZ'Y$$

(2) $$\hat{\beta}_{IV} == (X'Z(Z'\Sigma Z)^{-1}Z'X)^{-1}X'Z(Z'\Sigma Z)^{-1}Z'Y$$

(3) Sandwich estimator

$$\frac{d^2H}{d\beta d\beta'} = 2X'ZWZ'X = 2X'Z(Z'\Sigma Z)^{-1}Z'X$$

$$\left(\frac{dH}{d\beta}\right)\left(\frac{dH}{d\beta'}\right) = 4X'ZWZ'\epsilon\epsilon'ZWZ'X = 4X'Z(Z'\Sigma Z)^{-1}Z'\epsilon\epsilon'Z(Z'\Sigma Z)^{-1}Z'X$$

$$\to E\left[\left(\frac{dH}{d\beta}\right)\left(\frac{dH}{d\beta'}\right)\right] = 4X'Z(Z'\Sigma Z)^{-1}Z'\Sigma Z(Z'\Sigma Z)^{-1}Z'X = 4X'Z(Z'\Sigma Z)^{-1}X$$

So sandwich is

$$\left(\frac{d^2H}{d\beta d\beta'}\right)^{-1} E\left[\left(\frac{dH}{d\beta}\right)\left(\frac{dH}{d\beta'}\right)\right]\left(\frac{d^2H}{d\beta d\beta'}\right)^{-1} = (X'Z(Z'\Sigma Z)^{-1}Z'X)^{-1}.$$

1. d. Sandwich estimator for the OLS estimator

$$Max_{\beta}H(\beta)\{-(Y-X\beta)'(Y-X\beta)\}$$

$$\frac{dH}{d\beta} = (-2)(-X')(Y-X\beta) = 2X'(Y-X\beta) = 2X'\epsilon.$$

$$\frac{d^2H}{d\beta d\beta'} = 2(X'X).$$

(1)

$$E\left[\frac{dH}{d\beta}\frac{dH}{d\beta'}\right] = E(2X'\epsilon)(2\epsilon'X') = 4X'E(\epsilon\epsilon')X = 4X'\Sigma X.$$

(2)

Sandwich estimator:

$$\left(-E\frac{d^2H}{d\beta d\beta'}\right)^{-1}\left(E\frac{dH}{d\beta}\frac{dH}{d\beta'}\right)\left(-E\frac{d^2H}{d\beta d\beta'}\right)^{-1} = \frac{1}{2}(X'X)^{-1}(4X'\Sigma X)\left(\frac{1}{2}\right)(X'X)^{-1}$$
$$= (X'X)^{-1}X'\Sigma X(X'X)^{-1}.$$

1.e. (1) Heteroskedasticity

Estimate

$$E(\epsilon\epsilon') = \begin{bmatrix} E(\epsilon_1^2) & E(\epsilon_1\epsilon_2) & \cdots & E(\epsilon_1\epsilon_n) \\ E(\epsilon_2\epsilon_1) & E(\epsilon_2^2) & \cdots & E(\epsilon_2\epsilon_n) \\ \vdots & \vdots & \ddots & \vdots \\ E(\epsilon_n\epsilon_1) & E(\epsilon_n\epsilon_2) & \cdots & E(\epsilon_n^2) \end{bmatrix} = \begin{bmatrix} \sigma_1^2 & 0 & \cdots & 0 \\ 0 & \sigma_2^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_n^2 \end{bmatrix}$$

Estimate

$$\Sigma = \begin{bmatrix} \sigma_1^2 & 0 & \cdots & 0 \\ 0 & \sigma_2^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_n^2 \end{bmatrix}$$

using $\begin{bmatrix} e_1^2 & 0 & \cdots & 0 \\ 0 & e_2^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & e_n^2 \end{bmatrix}$

1.e. (2) Autocorrelation

Part (b): $E(\epsilon\epsilon') = \begin{bmatrix} E(\epsilon_1^2) & E(\epsilon_1\epsilon_2) & \cdots & E(\epsilon_1\epsilon_n) \\ E(\epsilon_2\epsilon_1) & E(\epsilon_2^2) & \cdots & E(\epsilon_2\epsilon_n) \\ \vdots & \vdots & \ddots & \vdots \\ E(\epsilon_n\epsilon_1) & E(\epsilon_n\epsilon_2) & \cdots & E(\epsilon_n^2) \end{bmatrix} = \begin{bmatrix} \sigma_1^2 & \sigma_1\sigma_2 & \cdots & \sigma_1\sigma_n \\ \sigma_2\sigma_1 & \sigma_2^2 & \cdots & \sigma_2\sigma_n \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_n\sigma_1 & \sigma_n\sigma_2 & \cdots & \sigma_n^2 \end{bmatrix}$

A naive approach might be to use e_j^2 to estima$E(\epsilon_j^2)$,
Use e_ie_j to estim$E(\epsilon_i\epsilon_j)$.

The Newey-West provides a variation on this method.

3. Sandwich estimator for the MLE

$Max\{\sum_{i=1}^{n} \ln(y_i - X_i\beta;\theta\}$

(1)

$$\left(-E\frac{d^2H}{d\psi d\psi'}\right)^{-1}\left(E\frac{dH}{d\psi}\frac{dH}{d\psi'}\right)\left(-E\frac{d^2H}{d\psi d\psi'}\right)^{-1}.$$

(2)

$$E\frac{d^2H}{d\psi d\psi'} = \frac{1}{n}\sum\frac{d^2H}{d\psi d\psi'}$$

$$E\frac{dH}{d\psi}\frac{dH}{d\psi'} = \frac{1}{n}\sum\frac{dH}{d\psi}\frac{dH}{d\psi'}.$$

4. Panel data estimation

Coefficient estimator

$$\begin{aligned}\hat{\beta} = (X'X)^{-1}X'Y &= \left\{[X_1' \quad \cdots \quad X_n'] \begin{bmatrix} X_1 \\ \vdots \\ X_n \end{bmatrix}\right\}^{-1} \left\{[X_1' \quad \cdots \quad X_n'] \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}\right\} \\ &= \left(\sum_{i=1}^{n} X_i'X_i\right)^{-1} \sum_{i=1}^{n} X_i'Y_i\end{aligned}$$

Sandwich estimator

$$\begin{aligned}(X'X)^{-1}X'\Omega X(X'X)^{-1} &= \left(\sum_{i=1}^{n} X_i'X_i\right)^{-1} [X_1' \quad \cdots \quad X_n'] \begin{bmatrix} \Omega_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \Omega_n \end{bmatrix} \begin{bmatrix} X_1 \\ \vdots \\ X_n \end{bmatrix} \left(\sum_{i=1}^{n} X_i'X_i\right)^{-1} \\ &= \left(\sum_{i=1}^{n} X_i'X_i\right)^{-1} \left(\sum_{i=1}^{n} X_i'\Omega_i X_i\right) \left(\sum_{i=1}^{n} X_i'X_i\right)^{-1}.\end{aligned}$$

Selected References

Amemiya, Takeshi (1985), Advanced Econometrics, Cambridge: Harvard Press (especially Chapter 4, Asymptotiac Properties of Extremum Estimators).

Anglin, P. M., and G. Ramazan (1996), “Semiparametric Estimation of a Hedonic Price Function,” Journal of Applied Econometrics 11, no. 6, 633-648.

Anscombe, F. J. (1973), "Graphs in Statistical Analysis," The American Statistician, 27, 17-21.

Boyer, B. H., J. B. McDonald, and W. K. Newey, (2003) “A Comparison of Patially Adaptive and Reweighted Least Squares Estimation, “ Econometric Reviews, 115-134.

Hampel, F. R. (1974), "The Influence Curve and Its Role in Robust Estimation," Journal of the American Statistical Association, 69, 383-393.

Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J., and Stahel, W. A. (1986), Robust Statistics: The Approach Based on Influence Functions, New York: John Wiley and Sons.

Hsieh, D. A. and C. F. Manski, "Monte Carlo Evidence on Adaptive Maximum Likelihood Estimation of a Regression," Annals of Statistics, 15(1987), 541-551.

Huber, P. J. (1981), Robust Statistics, New York: Wiley.

Kennedy, P. (2008), A Guide to Econometrics, 6th edition, Malden, MA: Blackwell Publishing.

Koenker, R. (1982), "Robust Methods in Econometrics," Econometric Reviews, 1, 213-255.

Martin, R. D. and T. T. Simin (2003), “Outlier-Resistant Estimates of Beta,” Financial Analysts Journal, 56-69.

McDonald, J. B. and Newey, W. K. (1988), "Partially Adaptive Estimation of Regression Models via the Generalized T Distribution," Econometric Theory, 4(1988), 428-457.

McDonald, J. B. an S. B. White, "A Comparison of Some Robust, Adaptive and Partially Adaptive Estimators of Regression Models," Econometric Reviews, 12(1993), 103-124.

Newey, W. K., "Adaptive Estimation of Regression Models Via Moment Restrictions," Journal of Econometrics, 38(1988), 301-339.

Pagan, A. And A. Ullah, Nonparametric Econometrics, Cambridge: Cambridge University Press, 1999.

Rey, W. J. J. (1983), Introduction to Robust and Quasi-Robust Statistical Methods, New York: Springer-Verlag.

Robinson, P. M. (1998), “Root-N consistent Semiparametric Regression,” Econometrica 56, no. 4, 931-954.

Rousseeuw, P. J. (1984), “Least Median Squares Regression,” JASA, 871-880.

Sharpe, W. F. (1964), "Capital Asset Prices: A Theory of Equilibrium Under Conditions of Risk," Journal of Finance, 19, 425-442.

Silverman, B. W. 1986. Density estimation for statistics and data analysis. Chapman & Hall.

APPENDIX: DATA FOR CAPM EXAMPLES

		VALUE-		
		WEIGHTED		MARTIN
	RISKFREE	MARKET	AMERICAN	MARIETTA
DATE	RATE	RATE	CAN CO	CORP
JAN 1982	0.0080	-0.022042	-0.05164	-0.12847
FEB 1982	0.0092	-0.049210	-0.16078	-0.06773
MAR 1982	0.0098	-0.008324	0.03738	-0.04769
APR 1982	0.0113	0.041886	0.01712	0.06393
MAY 1982	0.0106	-0.029107	0.00000	-0.03433
JUN 1982	0.0096	-0.019897	0.05455	-0.07627
JUL 1982	0.0105	-0.021088	-0.01379	-0.06373
AUG 1982	0.0076	0.125243	0.12108	0.69550
SEP 1982	0.0051	0.012631	-0.02800	-0.07187
OCT 1982	0.0059	0.115704	0.05267	0.09091
NOV 1982	0.0063	0.047138	-0.01600	0.00926
DEC 1982	0.0067	0.016163	0.00407	0.08208
JAN 1983	0.0069	0.037046	0.02348	-0.03429
FEB 1983	0.0062	0.028331	0.04049	-0.08284
MAR 1983	0.0063	0.033159	0.05058	0.23819
APR 1983	0.0071	0.072562	0.15481	0.11579
MAY 1983	0.0069	0.003940	0.17974	0.04443
JUN 1983	0.0067	0.039217	-0.03047	0.10251
JUL 1983	0.0074	0.030019	-0.03771	0.02479
AUG 1983	0.0076	0.012462	-0.01813	-0.06484
SEP 1983	0.0076	0.018097	0.05846	0.08261
OCT 1983	0.0076	0.018098	0.01105	-0.05120
NOV 1983	0.0070	0.025632	0.14035	-0.05498
DEC 1983	0.0073	0.008210	-0.03846	-0.03051
JAN 1984	0.0076	0.008873	0.12747	0.02448
FEB 1984	0.0071	0.036881	-0.06954	-0.07276
MAR 1984	0.0073	0.016679	-0.04381	-0.00743
APR 1984	0.0081	0.005258	-0.01132	0.01873
MAY 1984	0.0078	0.051286	-0.06094	-0.03426
JUN 1984	0.0075	0.023251	0.10030	0.00385
JUL 1984	0.0082	0.015648	-0.06220	0.09579
AUG 1984	0.0083	0.111437	0.09884	0.11077
SEP 1984	0.0086	0.002052	0.03175	-0.04127
OCT 1984	0.0100	0.003322	0.00462	0.20530
NOV 1984	0.0073	0.009374	0.01554	-0.06407
DEC 1984	0.0064	0.025158	0.03061	0.05325
JAN 1985	0.0065	0.079807	0.02178	0.13764
FEB 1985	0.0058	0.016262	0.00737	0.05185
MAR 1985	0.0062	0.000849	0.03415	-0.02657
APR 1985	0.0072	0.002710	0.01604	-0.00243
MAY 1985	0.0066	0.058722	0.08000	0.13382
JUN 1985	0.0055	0.017185	0.03704	-0.00216
JUL 1985	0.0062	0.003678	-0.02143	0.02273
AUG 1985	0.0055	0.004681	0.02174	-0.00952
SEP 1985	0.0060	0.036768	-0.09149	-0.14194
OCT 1985	0.0065	0.044130	0.06276	0.00000

APPENDIX:R CAPM EXAMPLES, CONT'D.

		VALUE-		
		WEIGHTED		MARTIN
	RISKFREE	MARKET	AMERICAN	MARIETTA
DATE	RATE	RATE	CAN CO	CORP
NOV 1985	0.0061	0.068925	0.14286	0.04511
DEC 1985	0.0065	0.045591	-0.06250	0.02878
JAN 1986	0.0056	0.005765	0.11000	-0.06338
FEB 1986	0.0053	0.074143	0.16888	0.13910
MAR 1986	0.0060	0.054636	-0.01299	0.15182
APR 1986	0.0052	-0.012209	-0.06941	0.01153
MAY 1986	0.0049	0.050867	0.04821	0.07407
JUN 1986	0.0052	0.015193	0.03578	-0.01867
JUL 1986	0.0052	-0.054215	0.06875	-0.05163
AUG 1986	0.0046	0.072633	0.09627	0.08596
SEP 1986	0.0045	-0.079405	-0.07507	-0.08443
OCT 1986	0.0046	0.052730	0.08239	-0.08406
NOV 1986	0.0039	0.017544	-0.03138	0.10759
DEC 1986	0.0049	-0.027270	-0.00884	-0.11143