V.C. 14.381 Class Notes[1]

1. Fundamentals of Regression

1.1. **Regression and Conditional Expectation Function.** Suppose y_t is a real response variable, and w_t is a d-vector of covariates. We are interested in the conditional mean (expectation) of y_t given w_t:

$$g(w_t) := E[y_t|w_t].$$

It is also customary to define a regression equation:

$$y_t = g(w_t) + \varepsilon_t, \quad E[\varepsilon_t|w_t] = 0,$$

where y_t is thought of as a dependent variable, w_t as independent variables, and ε_t as a disturbance.

The regression function can also be defined as the solution to the best conditional prediction problem under square loss: for each w, we have

$$g(w) = \arg\min_{\tilde{g}\in\mathbb{R}} E[(y_t - \tilde{g})^2|w].$$

[1]The notes are very rough and are provided for your convenience only. Please e-mail me if you notice any mistakes (vchern@mit.edu).

Therefore, the conditional mean function also solves the unconditional prediction problem:

$$g(\cdot) = \arg \min_{\tilde{g}(\cdot) \in \mathcal{G}} E[(y_t - \tilde{g}(w_t))^2],$$

where the argmin is taken over $\mathcal{G}$, the class of all measurable functions of w. This formulation does not easily translate to either estimation or computation.

Thus in this course we will learn to

- first approximate $g(w_t)$ by $x_t'\beta$, for $\beta \in \mathbb{R}^K$ and x_t formed as transformations of the original regressor,

  $$x_t = f(w_t),$$

  where the choice of transformations f is based on the approximation theory,
- then, estimate $x_t'\beta$ reasonably well using data, and make small-sample and large sample inferences on $x_t'\beta$ as well as related quantities.

Example 1: In Engel (1857), y_t is a household's food expenditure, and w_t is the household's income. A good approximation appears to be a power series such as $g(w_t) = \beta_0 + \beta_1 w_t + \beta_2 w_t^2$. Engle actually used a Haar series (an approximation based on many dummy terms). Of interest is the effect of income on food consumption

$$\frac{\partial g(w)}{\partial w}.$$

See the figure distributed in class.

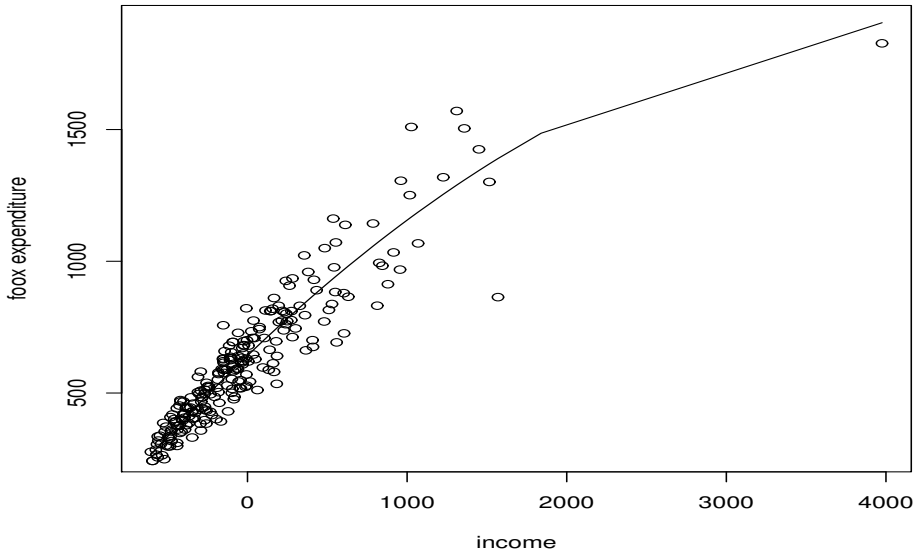

FIGURE 1.

Example 2: Suppose y_t^* is the birthweight, and w_t is smoking or quality of medical care. Clearly $E[y_t^*|w_t]$ is an interesting object. Suppose we are interested in the impact of w_t on very low birthweights; one way to do this is to define

$$y_t := 1\{y_t^* < c\},$$

where c is some critical level of birthweight. Then we can study

$$g(w_t) = E[y_t|w_t] = P[y_t^* < c|w_t].$$

This regression function measures the dependence of the probability of occurrence of extreme birthweight on covariates.[2]

Suppose we are able to recover $E[y_t|w_t]$. What do we make of it?

Schools of thought:

1. Descriptive. Uncover interesting stylized facts. Smoking "reduces" mean birthweight (but does not reduce occurrence of extreme birthweight).

2. Treatment Effects. An ideal setup for inferring causal effect is thought to be a perfectly controlled randomized trial. Look for natural experiments when the latter is not available.

3. Structural Effects. Estimate a parameter of an economic (causal, structural) model. Use economics to justify why $E[y_t|x_t]$ might be able to identify economic parameters. See Varian's chapter.

1.2. **OLS in Population and Finite Sample.**

1.2.1. *BLP Property.* In population LS β is defined as the minimizer (argmin) of

$$Q(b) = E[y_t - x_t'b]^2$$

[2]Another way is to look at the quantiles of y_t^* as a function of w_t, which is what quantile regression accomplishes.

and thus $x_t'\beta$ is the best linear predictor of y_t in population under square loss. In finite samples LS $\hat{\beta}$ is defined as the minimizer of

$$Q_n(\beta) = E_n[y_t - x_t'\beta]^2,$$

where E_n is the empirical expectation (a shorthand for $\frac{1}{n}\sum_{t=1}^n$). Thus $x_t'\hat{\beta}$ is the best linear predictor of y_t in the sample under square loss.

We can also state an explicit solution for β. Note that β solves the first-order condition

$$E[x_t(y_t - x_t'\beta)] = 0, \text{ i.e. } \beta = E[x_t x_t']^{-1}E[x_t y_t],$$

provided $E[x_t x_t']$ has full rank. LS in sample replaces population moments by empirical moments:

$$E_n[x_t(y_t - x_t'\hat{\beta})] = 0, \text{ i.e. } \hat{\beta} = E_n[x_t x_t']^{-1}E_n[x_t y_t],$$

provided $E_n[x_t x_t']$ has full rank.

1.2.2. *OLS as the Best Linear Approximation.* Observe that

$$Q(b) = E[y_t - x_t'b]^2 = E[E[y_t|w_t] - x_t'b \ + \varepsilon_t]^2 = \ E[E[y_t|w_t] - x_t'b]^2 \ + E[\varepsilon_t]^2,$$

where $\varepsilon_t = y_t - E[y_t|w_t]$. Therefore β solves

$$\min_b E(E[y_t|w_t] - x_t'b)^2,$$

and $x_t'\beta$ is the best linear approximation to the conditional mean function $E[y_t|w_t]$. This provides a link to approximation theory.

1.2.3. *Building Functional Forms.* The link to approximation theory is useful because approximation theory can be used to build good functional forms for our regressions. Here we focus the discussion on the approximation schemes that are most useful in econometric applications.[3]

1. Spline Approximation: Suppose we have one regressor w. Then the linear spline (spline of order 1) with a finite number of equally spaced knots $k_1, k_2, ..., k_r$ takes the form:

$$x_t = f(w_t) = (1, w_t, \ (w_t - k_1)_+, ..., (w_t - k_r)_+)',$$

where $(u)_+$ denotes $u \times 1\,(u > 0)$. The cubic spline takes the form:

$$x_t \ = f(w_t) = (1, (w_t, w_t^2, w_t^3), \ (w_t - k_1)_+^3, ..., (w_t - k_r)_+^3)'.$$

When specifying splines we may control K – the dimension of x_t. The function $w \mapsto f(w)'b$ constructed using splines is twice differentiable in w for any b.

2. Power Approximation: Suppose we have one regressor w, transformed to have support in $[0, 1]$. Then the r-th degree polynomial series is given by:

$$x_t = f(w_t) = (1, w_t, \ ..., w_t^r)'.$$

[3]W. Newey's paper provides a good treatment from an estimation prospective. See K. Judd's book for a good introduction to approximations methods.

Chebyshev polynomials are often used instead of the simple polynomials. Suppose w_t is transformed to have values ranging between $[-1, 1]$, the Chebyshev polynomials can be constructed as

$$x_t = f(w_t) = (cos(j \cdot cos^{-1}(w_t)), j = 0, ..., r)$$

(They are called polynomials because $f(w_t) = (1, w_t, 2w^2 - 1, 4w_t^3 - 3w_t, ...,\ )$, and thus are indeed polynomials in w_t.)[4]

3. Wavelet Approximation. Suppose we have one regressor w, transformed to have support in $[0, 1]$. Then the r-th degree wavelet series is given by:

$$x_t = f(w_t) = (e^{j(i2\pi w_t)}, j = 0, ..., r),$$

or one can use sines and cosines bases separately.

The case with multiple regressors can be addressed similarly: suppose the basic regressors are $w_{1t}, ...w_{dt}$. Then we can create d series–one for each basic regressor– then create all interactions of the d series, called tensor products, and collect them into the regressor vector x_t. If each series for a basic regressor has J terms, then the final regressor has dimension $K \approx J^d$, which explodes exponentially in the dimension d (a manifestation of the curse of dimensionality). For a formal definition of the tensor products see, e.g., Newey.

[4]See K. Judd's book for further detail; also see http://mathworld.wolfram.com/

Theorem 1.1. *Suppose w_t has a bounded support on a cube in $\mathbb{R}^d$ and has a positive, bounded density. If $g(w)$ is s-times continuously differentiable with bounded derivatives (by a constant M), then using K-term series $x = f(w)$ of the kind described above, the approximation error is controlled as:*

$$\min_b [E[g(w_t) - x_t'b]^2]^{1/2} \leq \ const_M \cdot K^{-\gamma/d},$$

where $\gamma = s$ for power series and wavelets, and for splines $\gamma = \min(s, r)$, where r is the order of the spline.

Thus, as $K \to \infty$, the approximation error converges to zero. The theorem also says that it is easier to approximate a smooth function, and it is harder to approximate a function of many basic regressors (another manifestation of the curse of the dimensionality). It should be noted that the statement that the approximation error goes to zero as $K \to \infty$ is true even *without* smoothness assumptions; in fact, it suffices for $g(w)$ to be measurable and square integrable.

The approximation of functions by least squares using splines and Chebyshev series has good properties, not only in minimizing mean squared approximation error, but also in terms of the maximum distance to the approximand (the latter property is called co-minimality).

1.3. **Examples on Approximation.** Example 1.(Synthetic) Suppose function $g(w) = w + 2sin(w)$, and that w_t is uniformly distributed on integers $\{1, ..., 20\}$. Then OLS in population solves the approximation problem:

$$\beta = \arg\min_b E[g(w_t) - x_t'b]^2$$

for $x_t = f(w_t)$. Let us try different functional forms for f. In this exercise, we form $f(w)$ as (a) linear spline (Figure 2,left) and (b) Chebyshev series (Figure 2,right), such that the dimension of $f(w)$ is either 3 or 8.

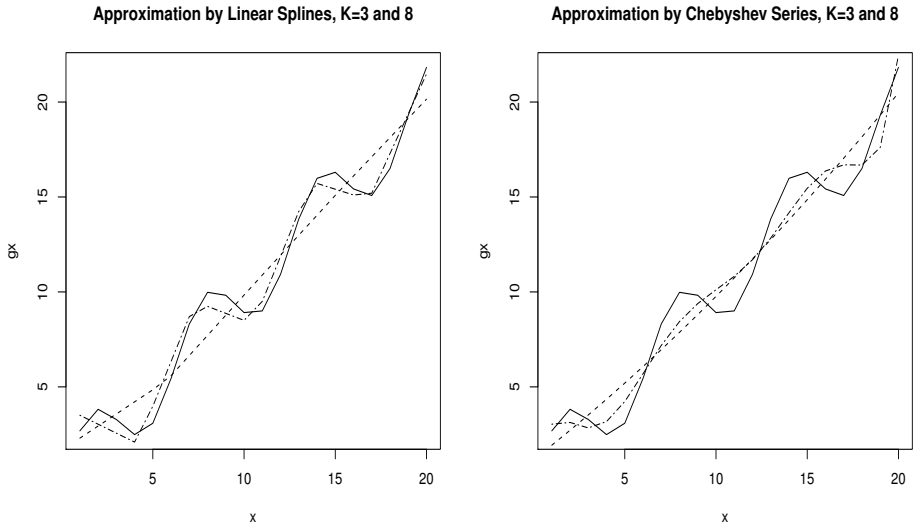

FIGURE 2.

	spline $K=3$	spline $K=8$	Chebyshev $K=3$	Chebyshev $K=8$
RMSAE	1.37	0.65	1.39	1.09
MAE	2.27	0.95	2.19	1.81

Then we compare the function $g(w)$ to the linear approximation $f(w)'\beta$ graphically. In Figure 2 we see that the parsimonious model with $K = 3$ accurately approximates the global shape ("big changes") in the conditional expectation function, but does not accurately approximate the local shape ("small changes"). Using a more flexible form with $K = 8$ parameters leads to a better approximation of the local shape. We also see the splines do much better in this example than Chebyshev polynomials.

We can also look at the formal measures of approximation error such as the root mean square approximation error (RMSAE):

$$[E[g(w_t) - f(w_t)'\beta]^2]^{1/2},$$

and the maximum approximation error (MAE):

$$\max_w |g(w) - f(w)'\beta|.$$

These measures are computed in the following table:

Example 2.(Real) Here $g(w)$ is the mean of log wage (y) conditional on education

$$w \in \{8, 9, 10, 11, 12, 13, 14, 16, 17, 18, 19, 20\}.$$

The function $g(w)$ is computed using population data – the 1990 U.S. Census data for men of prime age [5]. We would like to know how well this function is approximated by OLS when common approximation methods are used to form the regressors. For simplicity we assume that w_t is uniformly distributed (otherwise we can weight by the frequency). In the population, OLS solves the approximation problem: $\min E[g(w_t) - x_t'b]^2$ for $x_t = f(w_t)$, where we form $f(w)$ as (a) linear spline (Figure 3,left) and (b) Chebyshev series (Figure 3,right), such that dimension of $f(w)$ is either $K = 3$ or $K = 8$.

Then we compare the function $g(w)$ to the linear approximation $f(w)'\beta$ graphically. We also record RMSAE as well as the maximum error MAE. The approximation errors are given in the following table:

	spline $K = 3$	spline $K = 8$	Chebyshev $K = 3$	Chebyshev $K = 8$
RMSAE	0.12	0.08	0.12	0.05
MAE	0.29	0.17	0.30	0.12

[5] See Angrist, Chernozhukov, Fenrandez-Val, 2006, Econometrica, for more details

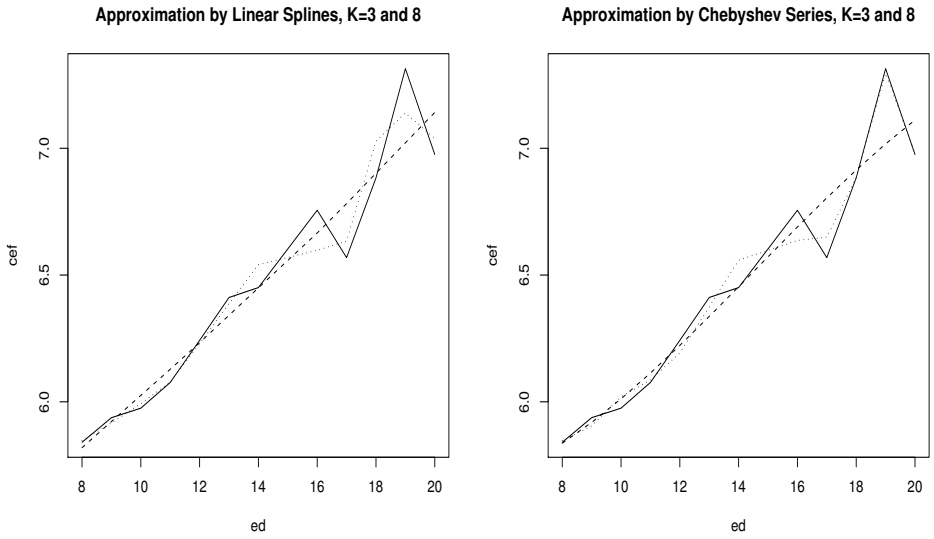

FIGURE 3.

REFERENCES:

1. Newey, Whitney K. Convergence rates and asymptotic normality for series estimators. J. Econometrics 79 (1997), no. 1, 147–168. (The definition of regression splines follows this reference).

2. Judd, Kenneth L. Numerical methods in economics. MIT Press, Cambridge, MA, 1998. (Chapter Approximation Methods. This is quite an advanced reference, but it is useful to have for many econometric and non-econometric applications.)

3. Hal R. Varian. Microeconomic Analysis (Chapter Econometrics. This is a great read on Structural Econometrics, and has a lot of interesting ideas.).

These materials have been posted on the MIT Server.

2. Regression Calculus

2.1. **Matrix Calculations.** The following calculations are useful for finite sample inference. Let $Y = (y_1, ..., y_n)'$ and X be the $n \times K$ matrix with rows x_t', $t = 1, ..., n$. Using this notation we can write

$$\hat{\beta} = \arg\min_{b \in \mathbb{R}^k} (Y - Xb)'(Y - Xb),$$

If $\text{rank}(X) = K$, the Hessian for the above program, $2X'X$, is positive definite; this verifies strict convexity and implies that the solution is unique. The solution $\hat{\beta}$ is determined by the first order conditions, called normal equations:

$$X'(Y - X\beta) = 0. \tag{1}$$

Solving these equations gives

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

The fitted or predicted values are given by the vector

$$\widehat{Y} := X\hat{\beta} = X(X'X)^{-1}X'Y = P_X Y,$$

Also define the residuals as

$$\hat{e} := \ Y - X'\hat{\beta} = (I - P_X)Y = M_X Y.$$

Geometric interpretation: Let $L := \text{span}(X) := \{Xb : b \in \mathbb{R}^k\}$ be the linear space spanned by the columns of X. Matrix P_X is called the projection matrix

because it projects Y onto L. Matrix M_X is a projection matrix that projects Y onto the subspace that is orthogonal to L.

Indeed, LS solves the problem

$$\min_{Y^* \in L} (Y - Y^*)'(Y - Y^*)$$

The solution $Y^* = \widehat{Y}$ is the orthogonal projection of Y onto L, that is

$$\text{(i) } \hat{Y} \in L \quad \text{(ii) } S'\widehat{e} = 0, \ \forall S \in L, \ \widehat{e} := (Y - \widehat{Y}) = M_X Y.$$

To visualize this, take a simple example with $n = 2$, one-dimensional regressor, and no intercept, so that $Y = (y_1, y_2)' \in \mathbb{R}^2$, $X = (x_1, x_2)' \in \mathbb{R}^2$, and $\beta \in \mathbb{R}$. (See Figure drawn in Class).

Some properties to note:

1. If regression has intercept, i.e. a column of X is $\mathbf{1} = (1, .., 1)'$, then $\bar{Y} = \bar{X}'\hat{\beta}$. The regression line passes through the means of data. Equivalently, since $\mathbf{1} \in L$, $\mathbf{1}' \ \hat{e} = 0$ or $\bar{\hat{e}} = 0$.

2. Projection ("hat") matrix $P_X \ = X(X'X)^{-1}X'$ is symmetric ($P_X' = P_X$) and idempotent ($P_X P_X = P_X$). P_X is an orthogonal projection operator mapping vectors in $\mathbb{R}^n$ to L. In particular, $P_X X = X$, $P_X \hat{e} = \mathbf{0}$, $P_X Y = X \ \hat{\beta}$.

3. Projection matrix $M_X = I - P_X$ is also symmetric and idempotent. M_X maps vectors in $\mathbb{R}^n$ to the linear space that is orthogonal to L. In particular, $M_X Y = \widehat{e}$, $M_X X = 0$, and $M_X \hat{e} = \hat{e}$. Note also $M_X P_X = 0$.

2.2. **Partitioned or Partial Regression.** Let X_1 and X_2 partition X as

$$X = [X_1, X_2]$$

and think of a regression model

$$Y = X_1\beta_1 + X_2\beta_2 + \varepsilon.$$

Further, let $P_{X_2} = X_2(X_2'X_2)^{-1}X_2'$ and $M_{X_2} = I - P_{X_2}$. Let $\hat{V}_1 = M_{X_2}X_1$ and $\hat{U} = M_{X_2}Y$, that is, $\hat{V}_1$ is the residual matrix from regressing the columns of X_1 on X_2, and $\hat{U}$ is the residual vector from regressing Y on X_2.

Theorem 2.1. *The following estimators are equivalent: 1. the component $\hat{\beta}_1$ of vector estimate $(\hat{\beta}_1', \hat{\beta}_2')'$ obtained from regressing Y on X_1 and X_2, 2. $\tilde{\beta}_1$ obtained from regressing Y on $\hat{V}_1$, 3. $\bar{\beta}_1$ obtained from regressing $\hat{U}$ on $\hat{V}_1$.*

Proof. Recall the following Fact shown above in equation (1):

$$\hat{\gamma} \text{ is OLS of } Y \text{ on } Z \text{ iff } Z'(Y - Z\hat{\gamma}) = 0. \tag{2}$$

Write

$$Y = X_1\hat{\beta}_1 + X_2\hat{\beta}_2 + \hat{e} = \hat{V}_1\hat{\beta}_1 + \underbrace{(X_1 - \hat{V}_1)\hat{\beta}_1 + X_2\hat{\beta}_2 + \hat{e}}_{\hat{\eta}}.$$

By the fact (2) above $\hat{\beta}_1 = \bar{\beta}_1$ if and only if $\hat{V}_1'\hat{\eta} = 0$. The latter follows because $\hat{V}_1'\hat{e} = 0$ by $\hat{V}_1 = M_{X_2}X_1 \in \text{span}(X)$, $X_2'\hat{V}_1 = X_2'M_{X_2}X_1 = 0$, and $(X_1 - \hat{V}_1)'\hat{V}_1 = (P_{X_2}X_1)'M_{X_2}X_1 = 0$.

To show $\hat{\beta}_1 = \bar{\beta}_1$, write

$$M_{X_2} Y = M_{X_2}(X_1 \hat{\beta}_1 + X_2 \hat{\beta}_2 + \hat{e}),$$

which can be equivalently stated as (noting that $M_{X_2} X_2 = 0$)

$$\widehat{U} = \widehat{V}_1 \hat{\beta}_1 + M_{X_2} \hat{e}.$$

Since $\widehat{V}_1' M_{X_2} \hat{e} = (M_{X_2} X_1)'(M_{X_2} \hat{e}) = X_1' M_{X_2} \hat{e} = \hat{V}_1' \hat{e} = 0$, it follows that $\hat{\beta}_1 = \bar{\beta}_1$.

Applications of Partial Regression:

1. Interpretation: $\hat{\beta}_1$ is a *partial* regression coefficient.

2. De-meaning: If $X_2 = \mathbf{1}$, then $M_{X_2} = I - \mathbf{1}(\mathbf{1}'\mathbf{1})^{-1}\mathbf{1}' = I - \mathbf{1}\mathbf{1}'/n$, and $M_{X_2}Y = Y - \mathbf{1}\bar{Y}$, $M_{X_2}X_1 = X_1 - \mathbf{1}\bar{X}_1'$. Therefore regressing Y on constant $X_2 = \mathbf{1}$ and a set, X_1, of other regressors produces the same slope coefficients as (1) regressing deviation of Y from its mean on deviation of X_1 from its mean or (2) regressing Y on deviations of X_1 from its mean.

3. Separate Regressions: If X_1 and X_2 are orthogonal, $X_1'X_2 = 0$, then $\hat{\beta}_1$ obtained from regressing Y on X_1 and X_2 is equivalent to $\tilde{\beta}_1$ obtained from regressing Y on X_1. To see this, write $Y = X_1\hat{\beta}_1 + (X_2\hat{\beta}_2 + \hat{e})$ and note $X_1'(X_2\hat{\beta}_2 + \hat{e}) = 0$ by $X_1'X_2 = 0$ and $X_1'\hat{e} = 0$. By the fact (2), it follows that $\hat{\beta}_2 = \tilde{\beta}_2$.

4. Omitted Variable Bias: If X_1 and X_2 are not orthogonal, $X_1'X_2 \neq 0$, then $\hat{\beta}_1$ obtained from regressing Y on X_1 and X_2 is not equivalent to $\tilde{\beta}_1$ obtained from regressing Y on X_1. However, we have that

$$\tilde{\beta}_1 = (X_1'X_1)^{-1}X_1'(X_1\hat{\beta}_1 + X_2\hat{\beta}_2 + \hat{e}) = \hat{\beta}_1 + (X_1'X_1)^{-1}X_1'(X_2\hat{\beta}_2).$$

That is, the coefficient in "short" regression equals the coefficient in the "long" regression plus a coefficient obtained from the regression of "omitted" term $X_2\hat{\beta}_2$ on

the included regressor X_1. It should be clear that this relation carries over to the population in a straightforward manner.

Example: Suppose Y are earnings, X_1 is education, and X_2 is unobserved ability. Compare the "long" coefficient $\hat{\beta}_1$ to the "short" coefficient $\tilde{\beta}_1$.

2.3. **Projections, R^2, and ANOVA.** A useful application is the derivation of the R^2 measure that shows how much of variation of Y is explained by variation in X. In the regression through the origin, we have the following analysis of variance decomposition (ANOVA)

$$Y'Y = \widehat{Y}'\widehat{Y} + \widehat{e}'\widehat{e}.$$

Then $R^2 := \widehat{Y}'\widehat{Y}/Y'Y \ = 1 - \widehat{e}'\widehat{e}/Y'Y$, and $0 \leq R^2 \leq 1$ by construction.

When the regression contains an intercept, it makes sense to de-mean the above values. Then the formula becomes

$$(Y - \bar{Y})'(Y - \bar{Y}) = (\hat{Y} - \bar{Y})'(\hat{Y} - \bar{Y}) + \widehat{e}'\widehat{e},$$

and

$$R^2 := (\widehat{Y} - \bar{Y})'(\widehat{Y} - \bar{Y})/((Y - \bar{Y})'(Y - \bar{Y})) = 1 - \widehat{e}'\widehat{e}/(Y - \bar{Y})'(Y - \bar{Y}),$$

where $0 \leq R^2 \leq 1$, and residuals have zero mean by construction, as shown below.

3. Estimation and Basic Inference in Finite Samples

3.1. **Estimation and Inference in the Gauss-Markov Model.** The GM model is a collection of probability laws $\{P_\theta, \theta \in \Theta\}$ for the data (Y, X) with the following properties:

GM1	$Y = X\beta + \varepsilon,$	$\beta \in \mathbb{R}^k$	linearity
GM2	$\text{rank}(X) = k$		identification
GM3	$E_\theta[\varepsilon \mid X] = 0$	$\forall\, \theta \in \Theta$	orthogonality, correct specification, exogeneity
GM4	$E_\theta[\varepsilon\varepsilon' \mid X] = \sigma^2 I_{n\times n}$	$\forall\, \theta \in \Theta$	sphericality

Here the parameter θ, that describes the probability model P_θ, consists of

$$(\beta, \sigma^2, F_{\varepsilon|X}, F_X),$$

where β is the regression parameter vector, σ^2 is variance of disturbances, $F_{\varepsilon|X}$ is the conditional distribution function of errors ϵ given X, and F_X is the distribution function of X. Both $F_{\varepsilon|X}$ and F_X are left unspecified, that is, nonparametric.

The model rules out many realistic features of actual data, but is a useful starting point for the analysis. Moreover, later in this section we will restrict the distribution of errors to follow a normal distribution.

GM1 & GM3: These assumptions should be taken together, but GM3 will be discussed further below. The model can be written as a linear function of the parameters and the error term, i.e.$y_i = \beta' x_i + \varepsilon_i$, where GM3 imposes that $E[\varepsilon_i | x_i] = 0$, and in fact much more, as discussed below.

This means that we have correctly specified the conditional mean function of Y by a functional form that is linear in parameters. At this point we can recall how we built functional forms using approximation theory. There we had constructed x_t as transformations of some basic regressors, $f(w_t)$. Thus, the assumptions GM1 and GM3 can be interpreted as stating $E[y_t|w_t] = E[y_t|x_t] = x_t'\beta$, which is an assumption that we work with a perfect functional form and that the approximation error is numerically negligible. Many economic functional forms will match well with the stated assumptions.[6]

[6]For example, a non-linear model such as the Cobb-Douglas production function $y_i = AK_i^{\alpha} L_i^{1-\alpha} e^{\varepsilon_i}$, can easily be transformed into a linear model by taking logs:

$$\ln y_i = \ln A + \alpha \ln L_i + (1-\alpha) \ln K_i + \varepsilon_i$$

This also has a nice link to polynomial approximations we developed earlier. In fact, putting additional terms $(\ln L)^2$ and $(\ln K^2)$ in the above equation gives us a translog functional form and is also a second degree polynomial approximation. Clearly, there is an interesting opportunity to explore the connections between approximation theory and economic modeling.

GM2: Identification.[7] The assumption means that explanatory variables are linearly independent. The following example highlights the idea behind this requirement. Suppose you want to estimate the following wage equation:

$$\log(wage_i) = \beta_1 + \beta_2 edu_i + \beta_3 tenure_i + \beta_4 exper_i + \varepsilon_i$$

where edu_i is education in years, $tenure_i$ is years on the current job, and $exper_i$ is experience in the labor force (i.e., total number of years at all jobs held, including the current). But what if no one in the sample ever changes jobs so $tenure_i = exper_i$ for all i. Substituting this equality back into the regression equation, we see that

$$\log(wage_i) = \beta_1 + \beta_2 edu_i + (\beta_3 + \beta_4) exper_i + \varepsilon_i.$$

We therefore can estimate the linear combination $\beta_3 + \beta_4$, but not β_3 and β_4 separately. This phenomenon is known as the *partial identification*. It is more common in econometrics than people think.

GM3: Orthogonality or Strict Exogeneity. The expected value of the disturbance term does not depend on the explanatory variables:

$$E[\varepsilon|X] = 0.$$

Notice that this means not only that $E[\varepsilon_i|x_i] = 0$ but also that $E[\varepsilon_i|x_j] = 0$ for all j. That is, the expected value of the disturbance for observation i not only does not

[7] This example and some of the examples below were provided by Raymond Guiterras.

depend on the explanatory variables for that observation, but also does not depend on the explanatory variables for any other observation. The latter may be an unrealistic condition in a time series setting, but we will relax this condition in the next section. Also, as we have noticed earlier, the assumption may also be unrealistic since it assumes perfect approximation of the conditional mean function.

There is another reason why this assumption should be looked with a caution. Recall that one of the main purposes of the econometric analysis is to uncover causal or structural effects. If the regression is to have a causal interpretation, the disturbances of a true causal equation:

$$y_i = x_i'\gamma + u_i$$

must satisfy the orthogonality restrictions such as $E[u_i|x_i] = 0$. If it does, then the causal effect function $x_i'\gamma$ coincides with regression function $x_i'\beta$.

The following standard example helps us clarify the idea. Our thinking about the relationship between income and education and the true model is that

$$y_i = \gamma_1 x_i + u_i, \quad u_i = \gamma_2 A_i + \varepsilon_i$$

where x_i is education, u_i is a disturbance that is composed of an ability effect $\beta_2 A_i$ and another disturbance ε_i which is independent of x_i, A_i, and ε. Suppose that both education and ability are de-meaned so that x_i measures deviation from average education, and A_i is a deviation from average ability. In general education is related

to ability, so

$$E[u_i|x_i] = \gamma_2 E[A_i|x_i] \neq 0.$$

Therefore orthogonality fails and γ_1 can not be estimated by regression of y_i on x_i. Note however, if we could observe ability A_i and ran the long regression of y_i on education x_i and ability A_i, then the regression coefficients β_1 and β_2 would recover the coefficients γ_1 and γ_2 of the causal function.

GM4: **Sphericality.** This assumption embeds two major requirements. The first is *homoscedasticity*: $E[\varepsilon_i^2|X] = \sigma^2, \quad \forall\, i$. This means that the conditional variance of each disturbance term is the same for all observations. This is often a highly unrealistic assumption.

The second is *nonautocorrelation*: $E[\varepsilon_i \varepsilon_j|X] = 0 \quad \forall\, i \neq j$. This means that the disturbances to any two different observations are uncorrelated. In time series data, disturbances often exhibit autocorrelation.

Moreover the assumption tacitly rules out many binary response models (and other types of discrete response). For example, suppose $y_i \in \{0, 1\}$ than

$$y_i = E[y_i|x_i] + \varepsilon_i = Pr[y_i = 1|x_i] + \varepsilon_i,$$

where ε_i has variance $P[y_i = 1|x_i](1 - P[y_i = 1|x_i])$ which does depend on x_i, except for the uninteresting case where $P[y_i = 1|x_i]$ does not depend on x_i.

3.2. **Properties of OLS in Gauss-Markov Model.** We are interested in various functionals of β, for example,

- β_j, a j-th component of β that may be of interest,
- $(x_1 - x_0)'\beta$, a partial difference of the conditional mean that results from a change in regressor values,
- $\frac{\partial x(w)}{\partial w_k}'\beta$, a partial derivative of the conditional mean with the elementary regressor w_k.

These functionals are of the form

$$c'\beta \quad \text{for} \quad c \in \mathbb{R}^K.$$

Therefore it makes sense to define efficiency of $\hat{\beta}$ in terms of the ability to estimate such functionals as precisely as possible.

Under the assumptions stated, it follows that

$$E_\theta[\hat{\beta}|X] = E_\theta[(X'X)^{-1}X'(X\beta + \varepsilon) \mid X] = I\beta + 0 = \beta \quad \forall \theta \in \Theta.$$

This property is called mean-unbiasedness. It implies in particular that the estimates of linear functionals are also unbiased $E_\theta[c'\hat{\beta}|X] = c'\beta$.

Next, we would like to compare efficiency of OLS with other estimators of the regression coefficient β. We take a candidate competitor estimator $\tilde{\beta}$ to be linear and unbiased. Linearity means that $\tilde{\beta} = a + AY$, where a and A are measurable function of X, and $E_\theta[\tilde{\beta}|X] = \beta$ for all θ in Θ. Note that unbiasedness requirement imposes

that $a + AX\beta = \beta$, for all $\beta \in \mathbb{R}^k$, that is,

$$AX = I \text{ and } a = 0.$$

Theorem 3.1. *Gauss-Markov Theorem. In the GM model, conditional on X, $\hat{\beta}$ is the minimum variance linear unbiased estimator (MVLUE) of β, meaning that any other unbiased linear estimator $\tilde{\beta}$ satisfies the relation:*

$$Var_\theta[c'\tilde{\beta} \mid x] \geq Var_\theta[c'\hat{\beta} \mid X], \quad \forall c \in \mathbb{R}^K, \quad \forall \theta \in \Theta.$$

The above property is equivalent to $c'\mathrm{Var}_\theta[\tilde{\beta} \mid X]c - c'\mathrm{Var}_\theta[\hat{\beta} \mid X]c \geq 0 \quad \forall c \in \mathbb{R}^K, \quad \forall \theta \in \Theta$, which is the same as saying

$$\mathrm{Var}_\theta[\tilde{\beta} \mid X] - \mathrm{Var}_\theta[\hat{\beta} \mid X] \text{ is } \textit{positive definite} \quad \forall \theta \in \Theta.$$

Example. Suppose y_t represents earnings, x_t is schooling. The mean effect of a change in schooling is $E[y_t \mid x_t = x^*] - E[y_t \mid x_t = x] = (x^* - x)'\beta$. By GM Theorem, $(x^* - x)'\hat{\beta}$ is MVLUE of $(x^* - x)'\beta$.

Example. One competitor of OLS is the weighted least squares estimator (WLS) with weights $W = \mathrm{diag}(w(x_1), ..., w(x_n))$. WLS solves $\min_\beta E_n[(y_t - x_t'\beta)^2 w(x_t)]$, or, equivalently $\min_\beta (Y - X\beta)'W(Y - X\beta)$. The solution is $\hat{\beta}_{WLS} = (X'WX)^{-1}X'WY$,

and $\hat{\beta}_{WLS}$ is linear and unbiased (show this). Under GM1-GM4 it is less efficient than OLS, unless it coincides with OLS.

3.3. **Proof of GM Theorem.** Here we drop indexing by θ: Unbiasedness was verified above. Take also any other unbiased estimator $\tilde{\beta} = AY$. By unbiasedness, $AX = I$. Observe that $\text{var}[c'\tilde{\beta} \mid X] = c'AA'c\sigma^2$ and $\text{var}[c'\hat{\beta} \mid X] = c'(X'X)^{-1}c \cdot \sigma^2$. It suffices to show the equality

$$\text{var}[c'\tilde{\beta} \mid X] - \text{var}[c'\hat{\beta} \mid X] = \text{var}[c'\tilde{\beta} - c'\hat{\beta} \mid X],$$

since the right hand side is non-negative. Write

$$\begin{aligned}
\text{var}[c'\tilde{\beta} - c'\hat{\beta} \mid X] &= \text{var}[\underbrace{c'(A - (X'X)^{-1}X')}_{M}(X\beta + \varepsilon) \mid X] \\
&= \text{var}[M\varepsilon \mid X] = E[M\varepsilon\varepsilon'M' \mid X] = MM'\sigma^2 \quad \text{(by A4)} \\
&= c'[AA' - (X'X)^{-1}]c \cdot \sigma^2 \quad (\Longleftarrow AX = I) \\
&= c'AA'c\sigma^2 - c'(X'X)^{-1}c \cdot \sigma^2 \square
\end{aligned}$$

Remark: This proof illustrates a general principle that is used in many places, such as portfolio theory and Hausman-type tests. Note that variance of the difference has very simple structure that does not involve covariance:

$$\text{var}[c'\tilde{\beta} - c'\hat{\beta} \mid X] = \text{var}[c'\tilde{\beta} \mid X] - \text{var}[c'\hat{\beta} \mid X]$$

This is because

$$\text{cov}[c'\tilde{\beta}, c'\hat{\beta} \mid X] = \text{var}[c'\hat{\beta} \mid X].$$

This means that an inefficient estimate $c'\tilde{\beta}$ equals $c'\hat{\beta}$, an efficient estimate, plus additional estimation noise that is uncorrelated with the efficient estimate.

3.4. **OLS Competitors and Alternatives. Part I.** Let us consider the following examples.

Example [*Expert Estimates vs OLS*] As an application, suppose $\beta = (\beta_1, ...\beta_k)'$, where β_1 measures elastiticity of demand for a good. In view of the foregoing definitions and GM theorem, analyze and compare two estimators: the fixed estimate $\beta_1^* = 1$, provided by an industrial organization expert, and $\hat{\beta}_1$, obtained as the ordinary least squares estimate.

- When would you prefer one over the other?
- Is GM theorem relevant for this decision?

The estimates may be better than OLS in terms of the mean squared error:

$$E_\theta[(\beta^* - \beta)^2] < E_\theta[(\hat{\beta} - \beta)^2]$$

for some or many values of $\theta \in \Theta$, which also translates into smaller estimation error of linear functionals. The crucial aspect to the decision is what we think Θ is. See

notes taken in class for further discussion. Take a look at Section 3.9 as well.

$X_n \Downarrow X$ and $X_n \Rightarrow X$

Example [*Shrinkage Estimates vs OLS*] Shrinkage estimators experienced a revival in learning theory, which is a modern way of saying regression analysis.

An example of a shrinkage estimator is the one that solves the following problem:

$$\min_b \left[(Y - Xb)'(Y - Xb)/2 \ + \lambda(b - \beta^*)X'X(b - \beta^*)\right]$$

The first term is called fidelity as it rewards goodness of fit for the given data, while the second term is called the shrinkage term as it penalizes deviations of Xb from the values $X\beta^*$ that we think are reasonable *a priori* (theory, estimation results from other data-sets etc.) The normal equations for the above estimator are given by:

$$X'(Y - X\tilde{\beta}) + \lambda X'X(\tilde{\beta} - \beta^*) = 0.$$

Solving for $\tilde{\beta}$ gives

$$\tilde{\beta} = (X'X(1 + \lambda))^{-1} \ (X'Y + \lambda X'X\beta^*) = \frac{\hat{\beta}}{1 + \lambda} + \frac{\lambda}{1 + \lambda}\beta^*$$

Note that setting $\lambda = 0$ recovers OLS $\hat{\beta}$, and setting $\lambda \approx \infty$ recovers the expert estimate β^*.

The choice of λ is often left to practitioners. For estimation purposes λ can be chosen to minimize the mean square error. This can be achieved by a device called *cross-validation*.

3.5. **Finite-Sample Inference in GM under Normality.** For inference purposes, we'll need to estimate the variance of OLS. We can construct an unbiased estimate of $\sigma^2(X'X)^{-1}$ as $s^2(X'X)^{-1}$ where

$$s^2 = \widehat{e}'\widehat{e}/(n-K).$$

Unbiasedness $E_\theta[s^2|X] = \sigma^2$ follows from

$$\begin{aligned} & E_\theta[\widehat{e}'\widehat{e}|X] = E_\theta[\varepsilon' M_X \varepsilon|X] = E_\theta[tr(M_X \varepsilon\varepsilon')|X] \\ = \quad & tr(M_X E[\varepsilon\varepsilon'|X]) \ = tr(M_X \sigma^2 I) \\ = \quad & \sigma^2 tr(I - P_X) = \sigma^2(tr(I_n) \ - tr(P_X)) = \sigma^2(tr(I_n) \ - tr(I_K)) = \sigma^2(n-K), \end{aligned}$$

where we used $tr(AB) = tr(BA)$, linearity of trace and expectation operators, and that $tr(P_X) = tr((X'X)^{-1}X'X) = tr(I_K)$.

We also have to add an additional major assumption:

$$\text{GM5.} \quad \varepsilon|X \sim N(0, \sigma^2 \ I) \qquad \text{Normality}$$

This makes our model for the conditional distribution of Y given X a parametric one and reduces the parameter vector of the conditional distribution to $\theta = (\beta, \sigma^2)$.

How does one justify the normality assumption? We discussed some heuristics in class.

Theorem 3.2. *Under GM1-GM5 the following are true:*

1. $\hat{\beta}|X \sim N(\beta, \sigma^2(X'X)^{-1})$, $Z_j := (\hat{\beta}_j - \beta_j)/\sqrt{\sigma^2(X'X)^{-1}_{jj}} \sim N(0,1)$.
2. $(n-K)s^2/\sigma^2 \sim \chi^2(n-K)$.
3. s^2 *and* $\hat{\beta}$ *are independent.*
4. $t_j := (\hat{\beta}_j - \beta_j)/se(\hat{\beta}_j) \sim t(n-K) \approx N(0,1)$ *for* $se(\hat{\beta}_j) = \sqrt{s^2(X'X)^{-1}_{jj}}$*; approximation* $\approx$ *is accurate when* $(n-K) \geq 30$.

3.6. **Proof of Theorem 3.2.** We will use the following facts:

(a) a linear function of a normal is normal, i.e. if $Z \sim N(0, \Omega)$, then $AZ \sim N(0, A\Omega A')$,

(b) if two normal vectors are uncorrelated, then they are independent,

(c) if $Z \sim N(0, I)$, and Q is symmetric, idempotent, then $Z'QZ \sim \chi^2(rank(Q))$,

(d) if a standard normal variable $N(0,1)$ and a chi-square variable $\chi^2(J)$ are independent, then $t(J) = N(0,1)/\sqrt{\chi^2(J)/J}$ is said to be a Student's t-variable with J degrees of freedom.

Proving properties (a)-(c) is left as a homework exercise.

Now let us prove each of the claims:

(1) $\hat{e} = M_X\varepsilon$ and $\hat{\beta} - \beta = (X'X)^{-1}X'\varepsilon$ are jointly normal with mean zero, since a linear function of a normal vector is normal.

(3) $\hat{e}$ and $\hat{\beta}$ are uncorrelated because their covariance equals $\sigma^2(X'X)^{-1}X'M_X = 0$, and therefore they are independent by joint normality. s^2 is a function of $\hat{e}$, so it is independent of $\hat{\beta}$.

(2)

$$(n-K)s^2/\sigma^2 = (\varepsilon/\sigma)' M_X (\varepsilon/\sigma) \sim \chi^2(rank(M_X)) = \chi^2(n-K).$$

(4) By properties 1-3, we have

$$t_j = Z_j \Big/ \sqrt{(s^2/\sigma^2)} \sim N(0,1) \Big/ \sqrt{\chi^2(n-K)/(n-K)} \sim t(n-K).$$

□

Property 4 enables us to do hypothesis testing and construct confidence intervals. We have that the event

$$t_j \in [t_{\alpha/2}, t_{1-\alpha/2}] \text{ has probability } 1-\alpha,$$

where t_α denotes the α-quantile of a $t(n-K)$ variable. Therefore, a confidence region that contains β_j with probability $1-\alpha$ is given by

$$I_{1-\alpha} = [\hat{\beta}_j \pm t_{1-\alpha/2} se(\hat{\beta}_j)].$$

This follows from event $\beta_j \in I_{1-\alpha}$ being equivalent to the event $t_j \in [t_{\alpha/2}, t_{1-\alpha/2}]$.

Also in order to test

$$H_o : \beta_j = 0 \quad \text{vs.} \quad H_a : \beta_j > 0$$

we check if $t_j = (\hat{\beta}_j - 0)/se(\hat{\beta}_j) \geq t_{1-\alpha}$, a critical value. It is conventional to select critical value $t_{1-\alpha}$ such that the probability of falsely rejecting the null when the null is right is equal to some small number α, with the canonical choice of α equal to .01,

.05 or .1. Under the null, t_j follows a $t(n-K)$ distribution, so the number $t_{1-\alpha}$ is available from the standard tables.

In order to test

$$H_o : \beta_j = 0 \quad \text{vs.} \quad H_a : \beta_j \neq 0$$

we check if $|t_j| \geq t_{1-\alpha/2}$. The critical value $t_{1-\alpha/2}$ is chosen such that the probability of false rejection equals α.

Instead of merely reporting "do not reject" or "reject" decisions it is also common to report the p-value – the probability of seeing a statistic that is larger or equal to t_j under the null:

$$P_j = \ 1 - Pr[t(n-K) \leq t]\Big|_{t=t_j}$$

for one sided alternatives, and $|t_j|$

$$P_j = \ 1 - Pr[-t \leq t(n-K) \leq t]\Big|_{t=|t_j|}$$

for two-sided alternatives. The probability $Pr[t(n-K) \leq t]$ is the distribution function of $t(n-K)$, which has been tabulated by Student.

P-values can be used to test hypotheses in a way that is equivalent to using t-statistics. Indeed, we can reject a hypothesis if $P_j \leq \alpha$.

Example. Temin's Roman Wheat Prices. Temin estimates a distance discount model:

$$price_i = \beta_1 \ + \beta_2 \cdot distance_i \ + \epsilon_i, \quad i = 1, ..., 6$$

where $price_i$ is the price of wheat in Roman provinces, and $distance_i$ is a distance from the province i to Rome. The estimated model is

$$price_i = \underset{(.49)}{-1.09} - \underset{(.0003)}{.0012} \cdot distance_i + \hat{\epsilon}_i, \quad R^2 = .79,$$

with standard errors shown in parentheses. The t-statistic for testing $\beta_2 = 0$ vs $\beta_2 < 0$ is $t_2 = -3.9$. The p-value for the one sided test is $P_2 = P[t(4) < -3.9] = 0.008$. A 90% confidence region for β_2 is $[-0.0018, -0.0005]$; it is calculated as $[\hat{\beta}_2 \pm t_{.95}(4) \cdot se(\hat{\beta}_2)] = [.0012 \pm 2.13 \cdot .0003]$.

Theorem 3.3. *Under GM1-GM5, $\hat{\beta}$ is the maximum likelihood estimator and is also the minimum variance unbiased estimator of β.*

Proof. This is done by verifying that the variance of $\hat{\beta}$ achieves the Cramer-Rao lower bound for the variance of unbiased estimators. Then, the density of y_i at $y_i = y$ conditional on x_i is given by

$$f(y|x_i, \beta, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\{-\frac{1}{2\sigma^2}(y - x_i'\beta)^2\}.$$

Therefore the likelihood function is

$$L(b, \sigma^2) = \prod_{i=1}^{n} f(y_i|x_i, b, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\{-\frac{1}{2\sigma^2}(Y - Xb)'(Y - Xb)\}.$$

It is easy to see that OLS $\hat{\beta}$ maximizes the likelihood function over β (check).

The (conditional) Cramer-Rao lower bound on the variance of unbiased estimators of $\theta = (\beta, \sigma^2)$ equals (verify)

$$\left[-E\left[\frac{\partial^2 \ln L}{\partial\theta\partial\theta'}\middle| X\right]\right]^{-1} = \begin{bmatrix} \sigma^2 (X'X)^{-1} & 0 \\ 0 & 2\sigma^4/n \end{bmatrix}.$$

It follows that least squares achieves the lower bound. ∎

3.7. **OLS Competitors and Alternatives. Part II.** When the errors are not normal, the performance of OLS relative to other location estimators can deteriorate dramatically.

Example. Koenker and Bassett (1978). See Handout distributed in class.

3.8. **Omitted Heuristics: Where does normality of ε come from?** Poincare: "Everyone believes in the *Gaussian* law of errors, the experimentalists because they think it is a mathematical theorem, the mathematicians because they think it is an experimental fact."

Gauss (1809) worked backwards to construct a distribution of errors for which the least squares is the maximum likelihood estimate. Hence normal distribution is sometimes called Gaussian.

Central limit theorem "justification": In econometrics, Haavelmo, in his "Probability Approach to Econometrics", *Econometrica* 1944, was a prominent proponent of this justification. Under the CLT justification, each error ε_i is thought of as a sum

of a large number of small and independent elementary errors v_{ij}, and therefore will be approximately Gaussian due to the central limit theorem considerations.

If elementary errors $v_{ij}, j = 1, 2, ...$ are i.i.d. mean-zero and $E[v_{ij}^2] < \infty$, then for *large* N

$$\varepsilon_i \ = \sqrt{N} \left[\frac{\sum_{j=1}^N v_{ij}}{N} \right] \approx_d N(0, Ev_{ij}^2),$$

as follows from the CLT.

However, if elementary errors v_{ij} are i.i.d. symmetric and $E[v_{ij}^2] = \infty$, then *for large* N (with additional technical restrictions on the tail behavior of v_{ij})

$$\varepsilon_i = N^{1-\frac{1}{\alpha}} \left[\frac{\sum_{j=1}^N v_{ij}}{N} \right] \approx_d Stable$$

where α is the largest finite moment: $\alpha = \sup\{p : E|v_{ij}|^p < \infty\}$. This follows from the CLT proved by Khinchine and Levy. The Stable distributions are also called sum-stable and *Pareto-Levy* distributions.

Densities of symmetric stable distributions have thick tails which behave approximately like power functions $x \mapsto \text{ const} \cdot |x|^{-\alpha}$ in the tails, with $\alpha < 2$.

Another interesting side observation: If $\alpha > 1$, the sample mean $\sum_{j=1}^N v_{ij}/N$ is a converging statistic, if $\alpha < 1$ the sample mean $\sum_{j=1}^N v_{ij}/N$ is a diverging statistic, which has interesting applications to diversification and non-diversification. (see R. Ibragimov's papers for the latter).

References: Embrechts et al. *Modelling Extremal Events*

3.9. **Testing and Estimating under General Linear Restrictions.** We now consider testing a linear equality restriction of the form

$$H_0 : R\beta = r, \quad \text{rank}R = p.$$

where R is a $p \times K$ matrix, and r is a p-vector. The assumption that R has full row rank simply means that there are no redundant restrictions – i.e., there are no restrictions that can be written as linear combinations of other restrictions. The alternative is

$$H_0 : R\beta \ \neq r.$$

This formulation allows us to test a variety of hypotheses. For example,

$R = [0, 1, 0, .., 0] \quad r = 0 \quad$ generates the restriction $\beta_2 = 0$

$R = [1, 1, 0, ..., 0] \quad r = 1 \quad$ generates the restriction $\beta_1 + \beta_2 = 1$

$R = [\mathbf{0}_{p\times(K-p)} \ I_{p\times p}] \quad r = (0, 0, ...)' \quad$ generates the restriction $\quad \beta_{K-p+1} = 0, ..., \beta_K = 0.$

To test H_0, we check whether the Wald statistic exceeds a critical value:

$$W := \ (R\hat{\beta} - r)'[\widehat{Var}(R\hat{\beta})]^{-1}(R\hat{\beta} - r) > c_\alpha,$$

where the critical value c_α is chosen such that probability of false rejection when the null is true is equal to α. Under GM1-5, we can take

$$\widehat{Var}(R\hat{\beta}) = s^2 R(X'X)^{-1}R'. \tag{3}$$

We have that $W_0 = (R\hat{\beta} - r)'[\sigma^2 R(X'X)^{-1}R']^{-1}(R\hat{\beta} - r) = N(0, I_p)^2 \sim \chi^2(p)$ and it is independent of s^2 that satisfies $s^2/\sigma^2 \sim \chi^2(n-K)/(n-K)$. We therefore have that

$$W/p \ = (W_0/p)/(s^2/\sigma^2) \sim \frac{\chi^2(p)/p}{\chi(n-K)/(n-K)} \sim F(p, n-K),$$

so c_α can be taken as the α-quantile of $F(p, n-K)$ times p. The statistic W/p is called the F-statistic.

Another way to test the hypothesis would be the distance function test (quasi-likelihood ratio test) which is based on the difference of the criterion function evaluated at the unrestricted estimate and the restricted estimate:

$$Q_n(\hat{\beta}_R) - Q_n(\hat{\beta}),$$

where in our case $Q_n(b) = (Y - X'b)'(Y - X'b)/n$, $\hat{\beta} = \arg\min_{b \in \mathbb{R}^K} Q_n(b)$ and

$$\hat{\beta}_R = \arg \min_{b \in \mathbb{R}^K : Rb = r} Q_n(b)$$

It turns out that the following equivalence holds for the construction given in (3):

$$DF = n[Q_n(\hat{\beta}_R) - Q_n(\hat{\beta})]/s^2 = W,$$

so using the distance function test is equivalent to using Wald test. This equivalence does not hold more generally, outside of the GM model.

Another major testing principle is the LM test principle, which is based on the value of the Lagrange Multiplier for the constrained optimization problem described

above. The Lagrangean for the problem is

$$\mathcal{L} = \ nQ_n(b)/2 + \lambda'(Rb - r).$$

The conditions that characterize the optimum are

$$X'(Y - Xb) + R'\lambda = 0, \ Rb - r = 0$$

Solving these equations, we get

$$\hat{\beta}_R = (X'X)^{-1}(X'Y + \ R'\hat{\lambda}) = \hat{\beta} + (X'X)^{-1}R'\hat{\lambda}$$

Putting $b = \hat{\beta}_R$ into constraint $Rb - r = 0$ we get $R(\hat{\beta} + (X'X)^{-1}R'\hat{\lambda}) - r = 0$ or

$$\hat{\lambda} = \ -[R(X'X)^{-1}R']^{-1}(R\hat{\beta} - r)$$

In economics, we call the multiplier the shadow price of a constraint. In our testing problem, if the price of the restriction is too high, we reject the hypothesis. The test statistic takes the form:

$$LM = \hat{\lambda}'[\widehat{Var}(\hat{\lambda})]^{-1}\hat{\lambda},$$

In our case we can take

$$\widehat{Var}(\hat{\lambda}) = [R(X'X)^{-1}R']^{-1}\widehat{Var}(R\hat{\beta}) \ [R'(X'X)^{-1}R]^{-1}$$

for $\widehat{Var}(R\hat{\beta}) = s^2R(X'X)^{-1}R'$. This construction also gives us the equivalence for our particular case

$$LM = W.$$

Note that this equivalence need not hold for non-linear estimators (though generally we have asymptotic equivalence for these statistics).

Above we have also derived the restricted estimate:

$$\hat{\beta}_R = \hat{\beta} - (X'X)^{-1}R'[R(X'X)^{-1}R']^{-1}(R\hat{\beta} - r).$$

When the null hypothesis is correct, we have that

$$E[\hat{\beta}_R|X] = \beta$$

and

$$Var[\hat{\beta}_R|X] = F\ Var(\hat{\beta}|X)F', \quad F = (I - (X'X)^{-1}R'[R(X'X)^{-1}R']^{-1}R).$$

It is not difficult to show that

$$Var[\hat{\beta}_R|X] \leq Var[\hat{\beta}|X]$$

in the matrix sense, and therefore ROLS $\hat{\beta}_R$ is unbiased and more efficient than OLS. This inequality can be verified directly; another way to show this result is given below. Does this contradict the GM theorem? No. Why?

It turns out that, in GM model with the *restricted* parameter $\{\beta \in \mathbb{R}^K : R\beta = r\}$, the ROLS is minimum variance unbiased linear estimator. Similarly, in GM normal model with the *restricted* parameter $\{\beta \in \mathbb{R}^K : R\beta = r\}$, the ROLS is also the minimum variance unbiased estimator. Note that previously we worked with the unrestricted parameter space $\{\beta \in \mathbb{R}^K\}$.

The constrained regression problem can be turned into a regression without constraints. This point can be made by first considering an example with the restriction $\beta_1 + \beta_2 = 1$. Write

$$\begin{aligned} Y &= X_1\beta_1 + X_2\beta_2 + \epsilon \\ &= X_1(\beta_1 + \beta_2 - 1) + X_2\beta_2 - X_1(\beta_2 - 1) + \epsilon \\ &= (X_2 - X_1)\beta_2 - X_1 + \epsilon \end{aligned}$$

or

$$Y - X_1 = (X_2 - X_1)\beta_2 + \epsilon,$$

It is easy to check that the new model satisfies the Gauss-Markov assumptions with the parameter space consisting of

$$\theta = (\beta_2, \sigma^2, F_{\epsilon|X}, \ F_X),$$

where $\beta_2 \in \mathbb{R}$ is unrestricted. Therefore $\hat{\beta}_{R2}$ obtained by applying LS to the last display is the efficient linear estimator (efficient estimator under normality). The same is true of $\hat{\beta}_{R1} = 1 - \hat{\beta}_{R2}$ because it is a linear functional of $\hat{\beta}_{R2}$. The ROLS $\hat{\beta}_R$ is therefore more efficient than the unconstrained linear least squares estimator $\hat{\beta}_2$.

The idea can be readily generalized. Without loss of generality, we can rearrange the order of regressors so that

$$R = [R_1 \quad R_2],$$

where R_1 is a $p \times p$ matrix of full rank p. Imposing $H_0 : R\beta = r$ is equivalent to $R_1\beta_1 + R_2\beta_2 = r$ or $\beta_1 = R_1^{-1}(r - R_2\beta_2)$, so that

$$Y = X\beta_1 + X_2\beta_2 + \epsilon \stackrel{H_0}{=} X_1(R_1^{-1}(r - R_2\beta_2)) + X_2\beta_2 + \epsilon,$$

that is

$$Y - X_1R_1^{-1}r = (X_2 - X_1R_1^{-1}R_2)\beta_2 + \epsilon.$$

This again gives us a model with a new dependent variable and a new regressor, which falls into the previous GM framework. The estimate $\hat{\beta}_{2R}$ as well as the estimate $\hat{\beta}_{1R} = R_1^{-1}(r - R_2\hat{\beta}_{2R})$ are efficient in this framework.

3.10. **Finite Sample Inference Without Normality.** The basic idea of the finite-sample Monte-Carlo method (MC) can be illustrated with the following example.

Example 1. Suppose $Y = X\beta + \epsilon$, $E[\epsilon|X] = 0$, $E[\epsilon\epsilon'|X] = \sigma^2 I$, $\epsilon = (\epsilon_1, ...\epsilon_n)' = \sigma U$, where

$$U = (U_1, ..., U_n)|X \text{ are i.i.d. with law } F_U \tag{4}$$

where F_U is known. For instance, taking $F_U = t(3)$ will better match the features of many financial return datasets than $F_U = N(0, 1)$ will.

Consider testing $H_0 : \beta_j = \beta_j^0$ vs. $H_A : \beta_j > \beta_j^0$. Under H_0

$$t_j = \frac{\hat{\beta}_j - \beta_j^0}{\sqrt{s^2(X'X)^{-1}_{jj}}} \stackrel{\text{by } H_0}{=} \frac{((X'X)^{-1}X'\epsilon)_j}{\sqrt{\frac{\epsilon' M_X \epsilon}{n-K}(X'X)^{-1}_{jj}}} = \frac{((X'X)^{-1}X'U)_j}{\sqrt{\frac{U' M_X U}{n-K}(X'X)^{-1}_{jj}}} \tag{5}$$

Note that this statistic nicely cancels the unknown parameter σ^2.

The p-value for this test can be computed via Monte-Carlo. Simulate many draws of the t-statistics under H_0:

$$\{t^*_{j,d}, d = 1, ..., B\}, \tag{6}$$

where d enumerates the draws and B is the total number of draws, which needs to be large. To generate each draw $t^*_{j,d}$, generate a draw of U according to (4) and plug it in the right hand side of (5). Then the p-value can be estimated as

$$P_j = \frac{1}{B}\sum_{d=1}^{B} 1\{t^*_{j,d} \geq t_j\}, \tag{7}$$

where t_j is the empirical value of the t-statistic. The p-value for testing $H_0 : \beta_j = \beta^0_j$ vs. $H_A : \beta_j \neq \beta^0_j$ can be estimated as

$$P_j = \frac{1}{B}\sum_{d=1}^{B} 1\{|t^*_{j,d}| \geq |t_j|\}. \tag{8}$$

Critical values for confidence regions and tests based on the t-statistic can be obtained by taking appropriate quantiles of the sample (6).

Example 2. Next generalize the previous example by allowing F_U to depend on an unknown nuisance parameter γ, whose true value, γ_0, is known to belong to region Γ. Denote the dependence as $F_U(\gamma)$.

For instance, suppose $F_U(\gamma)$ is the t-distribution with the "degrees of freedom" parameter $\gamma \in \Gamma = [3, 30]$, which allows us to nest distributions that have a wide

range of tail behavior, from very heavy tails to light tails. The normal case is also approximately nested by setting $\gamma = 30$.

Then, obtain a p-value for each $\gamma \in \Gamma$ and denote it as $P_j(\gamma)$. Then use

$$\sup_{\gamma \in \Gamma} P_j(\gamma)$$

for purposes of testing. Since $\gamma_0 \in \Gamma$, this is a valid upper bound on the true P-value $P_j(\gamma_0)$. Likewise, one can obtain critical values for each $\gamma \in \Gamma$ and use the *least favorable* critical value. The resulting confidence regions could be quite conservative if Γ is large; however, see the last paragraph.

The question that comes up naturally is: why not use an estimate $\hat{\gamma}$ of the true parameter γ_0 and obtain $P_j(\hat{\gamma})$ and critical values using MC where we set $\gamma = \hat{\gamma}$? This method is known as the parametric *bootstrap*. Bootstrap is simply a MC method for obtaining p-values and critical values using the estimated data generating process. Bootstrap provides asymptotically valid inference, but bootstrap does not necessarily provide valid finite sample inference. However, bootstrap often provides a more accurate inference in finite samples than the asymptotic approach does.

The finite-sample approach above also works with Γ that can be data-dependent. Let us denote the data dependent set of nuisance parameters as $\hat{\Gamma}$. If the set $\hat{\Gamma}$ contains γ_0 with probability $1 - \beta_n$, where $\beta_n \to 0$, we can adjust the estimate of the p-value

to be

$$\sup_{\gamma \in \hat{\Gamma}} P_j(\gamma) + \beta_n.$$

In large samples, we can expect that

$$\sup_{\gamma \in \hat{\Gamma}} P_j(\gamma) + \beta_n \approx P_j(\gamma_0),$$

provided $\hat{\Gamma}$ converges to γ_0 and $\beta_n \to 0$. Thus, the finite-sample method can be efficient in large samples, but also retain validity in finite samples. The asymptotic method or bootstrap cannot (necessarily) do the latter. This sets the methods apart. However, as someone mentioned to me, the finite-sample method can be thought of as a kind of "fancy bootstrap".

Example 3. (HW) Consider Temin's (2005) paper that models the effect of distance from Rome on wheat prices in the Roman Empire. There are only 6 observations. Calculate the p-values for testing the null that the effect is zero versus the alternative that the effect is negative. Consider first the case with normal disturbances (no need to do simulations for this case), then analyze the second case where disturbances follow a t-distribution with 8 and 16 "degrees of freedom".

3.11. **Appendix: Some Formal Decision Theory under Squared Loss.**

Amemiya (1985) sets up the following formalisms to discuss efficiency of estimators.

1. Let $\hat{\beta}$ and β^* be scalar estimators of a scalar parameter β. $\hat{\beta} \succcurlyeq$ (as good as) β^* if $E_\beta(\hat{\beta} - \beta)^2 \leq E_\beta(\beta^* - \beta)^2$, $\forall \beta \in \mathcal{B}$

Definition of "better" is tied down to quadratic loss.

2. $\hat{\beta}$ is **better** (more efficient, $\succ$) than β^* if $\hat{\beta} \succcurlyeq \beta^*$ and $E_\beta(\hat{\beta} - \beta)^2 < E_\beta(\beta^* - \beta)^2$, for some $\beta \in \mathcal{B}$

3. Let $\hat{\beta}$ and β^* be **vector** estimates of vector parameter β. $\hat{\beta} \succcurlyeq \beta^*$ if for all $c \in \mathbb{R}^k$, $c'\hat{\beta} \succcurlyeq c'\beta^*$ (for estimating $c'\beta$).

4. $\hat{\beta} \succ \beta$ if $c'\hat{\beta} \succ c'\beta^*$ for some $c \in \mathbb{R}^k$ and $c'\hat{\beta} \succcurlyeq c'\beta^*$ for all $c \in \mathbb{R}^k$.

It should be obvious that Definition 3 is equivalent to Definition 5.

5. $\hat{\beta} \succcurlyeq \beta^*$ if for $A_\beta \equiv E_\beta(\hat{\beta} - \beta)(\hat{\beta} - \beta)'$ and $B_\beta \equiv E_\beta(\beta^* - \beta)(\beta^* - \beta)'$, $A_\beta - B_\beta$ is semi-negative definite $\forall\ \beta \in \mathcal{B}$, or $A_\beta \leq B_\beta$ in the matrix sense.

6. $\hat{\beta}$ is a best in a class of estimators if there is no better estimator in this class.

4. Estimation and Basic Inference in Large Samples

A good reference for this part is Newey's lecture note that is posted on-line. Below I only highlight the main issues that we have discussed in class.

4.1. **The Basic Set-up and Implications.** In large sample we can conduct valid inference under much more general conditions than the previous GM model permitted.

One sufficient set of conditions we can work with is the following:

L1 $y_t = x_t'\beta + e_t, \quad t = 1, ..., n$

L2 $Ee_t x_t = 0, \quad t = 1, ..., n$

L3 $X'X/n \to_p Q$ finite and full rank

L4 $X'e/\sqrt{n} \to_d N(0, \Omega)$, where Ω is finite and non-degenerate.

Discussion:

1) L1 and L2 imply that β is the parameter that describes the best linear approximation of the conditional expectation function $E[y_t|x_t]$. Under L1-L4, OLS $\hat{\beta}$ turns out to be a consistent and asymptotically (meaning approximately) normally distributed estimator of β.

2) We can decompose

$$e_t = a_t + \varepsilon_t,$$

where $\varepsilon_t = y_t - E[y_t|x_t]$ and

$$a_t = E[y_t|x_t] - x_t'\beta.$$

The The error is therefore the sum of the "usual" disturbance ε_t, defined as a deviation of y_t from the conditional mean, and the approximation (specification) error a_t, defined as the error resulting from using linear functional form in place of the true function $E[y_t|x_t]$. In the previous GM framework, we assumed away the approximation error. In the present framework, we can "afford" it. We can explicitly acknowledge that we merely estimate an approximation to the true conditional mean function, and we can explicitly account for the fact that approximation error is a non-trivial source of heteroscedasticity (why?) that impacts the variance of our estimator.

3) L3 is merely an analog of the previous identification condition. It also requires that t product of regressors $\{x_t x_t', t = 1, ...n\}$ satisfy a LLN, thereby imposing some stability on them. This condition can be relaxed to include trending regressors (see e.g. Newey's handout or Amemiya's Advanced Econometrics).

4) L4 requires the sequence $\{x_t e_t, t = 1, ..., n\}$ to satisfy a CLT. This condition is considerably more general than the previous assumption of normality of errors that we made. In large samples, it will lead us to estimation and inference results that are similar to the results we obtained under normality.

Proposition 4.1. *Under L1-L4,* $\hat{\beta} \to_p \beta$.

We have that $\hat{\beta}$ approaches β as the sample size increases. Obviously, for consistency, we can replace L4 by a less stringent requirement $X'e/n \to_p 0$.

Proposition 4.2. *Under L1-L4,* $\sqrt{n}(\hat{\beta} - \beta) \to_d N(0, V)$, $V = Q^{-1}\Omega Q^{-1}$.

The results suggest that in large samples $\hat{\beta}$ is approximately normally distributed with mean β and variance V/n:

$$\hat{\beta} \approx_d N\left(\beta, V/n\right).$$

Proposition 4.3. *Under L1-L4, suppose there is* $\hat{V} \to V$, *then*

$$t_j := \left(\hat{\beta}_j - \beta_j\right)/s.e.\left(\hat{\beta}_j\right) := \left(\hat{\beta}_j - \beta_j\right)/\sqrt{\hat{V}_{jj}/n} \to_d N\left(0,1\right),$$

and if $R\beta = r$ *for* R *having full row rank* p

$$W = \left(R\hat{\beta} - r\right)' \left[R\left(\hat{V}/n\right)R'\right]^{-1}\left(R\hat{\beta} - r\right) \to_d \chi^2\left(p\right).$$

In large samples the appropriately constructed t-statistic and W-statistic are approximately distributed as the standard normal variable and a chi-square variable with p degrees of freedom; that is $t \approx_d N(0,1)$ and $W \approx_d \chi^2(p)$.

Basic use of these results is exactly the same as in the finite-sample case, except now all the statements are approximate. For example, under the null hypothesis, a t-statistic satisfies $t_j \to_d N(0,1)$, and that implies

$$\lim_{n\to\infty} Pr[t_j < c] = Pr[N(0,1) < c] = \Phi(c),$$

for every c, since Φ is continuous. In finite, large samples, we merely have

$$Pr[t_j < c] \approx \Phi(c),$$

where quality of approximation may be good or poor, depending on a variety of circumstances.

Remark: Asymptotic results for restricted least squares, which is a linear transformation of the unrestricted OLS, and other test statistics (e.g. LM) readily follow from the results presented above.

The main tools we will use to prove the results are the Continuous Mapping Theorem and the Slutsky Theorem. The underlying metric spaces in these results are finite-dimensional Euclidian spaces.

Lemma 4.1 (CMT)**.** *Let X be a random element, and $x \mapsto g(x)$ be continuous at each $x \in D_0$, where $X \in D_0$ with probability one. Suppose that $X_n \to_d X$, then $g(X_n) \to_d g(X)$; if $X_n \to_p X$, then $g(X_n) \to_p g(X)$.*

The proof of this lemma follows from an application of an a.s. representation theorem and then invoking the continuity hypothesis. The following lemma is a corollary of the continuous mapping theorem.

Lemma 4.2 (Slutsky Lemma)**.** *Suppose that matrix $A_n \to_p A$ and vector $a_n \to_p a$, where matrix A and vector a are constant. If $X_n \to_d X$, then $A_n X_n + a_n \to_d AX + a$.*

Proof of Proposition 1: Conditions L4 and L3 imply respectively that

$$X'e/n \xrightarrow{p} 0,\ X'X/n \xrightarrow{p} Q. \tag{2}$$

Then, by nonsingularity of Q, the fact that the inverse of a matrix is a continuous function of the elements of the matrix at any nonsingular matrix, and the Slutsky Lemma it follows that

$$\hat{\beta} = \beta + (X'X/n)^{-1}X'e/n \xrightarrow{p} \beta + Q^{-1} \cdot 0 = \beta.$$

Proof of Proposition 2: Conditions L4 and L3 imply respectively that

$$X'e/\sqrt{n} \xrightarrow{d} N(0,\Omega),\ X'X/n \xrightarrow{p} Q. \tag{2}$$

By the Slutsky Lemma it follows that

$$\sqrt{n}(\hat{\beta} - \beta) = (X'X/n)^{-1}X'e/\sqrt{n} \xrightarrow{d} Q^{-1}N(0,\Omega) = N(0, Q^{-1}\Omega Q^{-1}).$$

Proof of Proposition 3: By $\hat{V} \xrightarrow{p} V$, $V_{jj} > 0$, and the CMT, $\left(V_{jj}/\hat{V}_{jj}\right)^{1/2} \xrightarrow{p} 1$. It follows by the Slutsky Theorem that

$$\left(\hat{\beta}_j - \beta_j\right) / \left[V_{jj}\right]^{1/2} = (V_{jj}/\hat{V}_{jj})^{1/2}\sqrt{n}(\hat{\beta}_j - \beta_j)/\sqrt{V_{jj}} \xrightarrow{d} 1 \cdot N(0,1) = N(0,1).$$

Let $\Sigma = RVR'$. Matrix Σ is nonsingular by R having rank p and nonsingularity of V, so by the CMT, $\hat{\Sigma}^{-1/2} \xrightarrow{p} \Sigma^{-1/2}$. Also, by the Slutsky Lemma $Z_n = \widehat{\Sigma}^{-1/2}R\sqrt{n}\left(\hat{\beta} - \beta\right) \xrightarrow{d} Z = \Sigma^{-1/2}N(0,\Sigma) =_d N(0,I)$. Then by the CMT, $W = Z_n'Z_n \to_d Z'Z =_d \chi^2(p)$.

4.2. **Independent Samples.** Here we consider two models:

IID Model: Suppose that *(a) L1 and L2 hold, (b) vectors* (y_t, x_t) *are independent and identically distributed across* t*, (c)*

$$\Omega = Var[x_t e_t] = E[e_t^2 x_t x_t']$$

is finite and non-degenerate (full rank) and that

$$Q = E[x_t x_t']$$

is finite and is of full rank.

It should be emphasized that this condition does not restrict in any away the relationship between y_t and x_t; it only requires that the joint distribution function of (y_t, x_t) does not depend on t and that there is no dependence of data points across t.

This model allows for two sources of heteroscedasticity in the error $e_t = \varepsilon_t + a_t$ – one is the heteroscedasticity of ε_t and another is the heteroscedasticity of the approximation error $a_t = E[y|x_t] - x_t'\beta$. By heteroscedasticity we mean that $E[e_t^2|x_t]$ depends on x_t.

Example: Recall the Engel curve example discussed in section 1.1, where ε_t was clearly heteroscedastic. Therefore, e_t should be heteroscedastic as well. Many regression problems in economics have heteroscedastic disturbances.

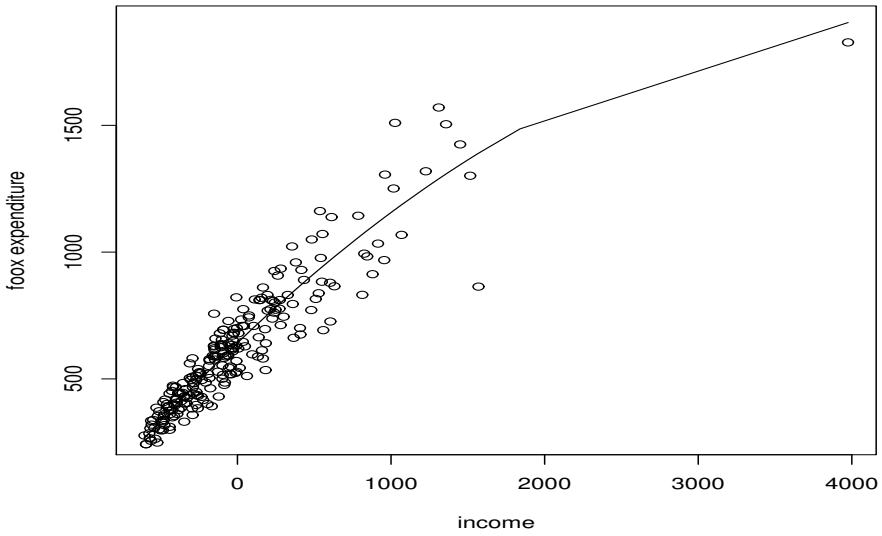

FIGURE 4.

Example: We saw in the wage census data that $a_t \neq 0$ for basic functional forms. Therefore e_t should be heteroscedastic due to $a_t \neq 0$.

Theorem 4.1. *The conditions of the iid model imply that L1-L4 hold with Ω and Q defined above. Therefore the conclusions of Propositions 1-3 hold as well.*

Proof: This follows from the Khinchine LLN and the multivariate Lindeberg-Levy CLT.

A consistent estimator (called heteroscedastiticy robust estimator) of variance V is given by:

$$\hat{V} = \hat{Q}^{-1}\hat{\Omega}\hat{Q}^{-1}, \quad \hat{\Omega} = E_n[e_t^2 x_t x_t'], \quad \hat{Q} = X'X/n.$$

Theorem 4.2. *The conditions of the iid model and additional assumptions (e.g. bounded fourth moments of x_t) imply that $\hat{\Omega} \to_p \Omega$ and $\hat{Q} \to_p Q$, so that $\hat{V} \to_p V$.*

Proof: Consistency of $\hat{Q} \to_p Q$ follows by the Khinchine LLN, and consistency of $\hat{\Omega} \to_p \Omega$ can be shown as follows. Consider the scalar case to simplify the notation. We have that

$$\hat{e}_t^2 = e_t^2 - 2(\hat{\beta} - \beta)' x_t e_t + (\hat{\beta} - \beta)^2 x_t^2.$$

Multiply both sides by x_t^2 and average over t to get

$$E_n[\hat{e}_t^2 x_t^2] = E_n[e_t^2 x_t^2] - 2(\hat{\beta} - \beta) E_n[e_t x_t^3] + (\hat{\beta} - \beta)^2 E_n[x_t^4].$$

Then $E_n[e_t^2 x_t^2] \to_p E[e_t^2 x_t^2]$, $E_n[x_t^4] \to_p E[x_t^4]$, and $E_n[e_t x_t^3] \to_p E[e_t x_t^3]$ by the Khinchine LLN, since $E[x_t^4]$ is finite by assumption, and $E[e_t x_t^3]$ is finite by

$$|E[e_t x_t x_t^2]|^2 \leq E[e_t^2 x_t^2]\ E[x_t^4],$$

which follows from the Cauchy-Schwartz inequality and the assumed finiteness of $E[e_t^2 x_t^2]$ and $E[x_t^4]$. Using the consistency of $\hat{\beta}$, the facts mentioned, and CMT, we get the consistency. □.

Homoscedastic IID Model: *In addition to conditions (a)-(c) of the iid model, suppose that* $E[e_t|x_t] = 0$ *and* $E[e_t^2|x_t] = \sigma^2$*, then* $Var[e_t^2|x_t] = \sigma^2$*, so that we have the simplification*

$$\Omega = \Omega_0 := \sigma^2 E[x_t x_t'].$$

In this model, there is no approximation error, i.e. $e_t = \varepsilon_t$, and there is no heteroscedasticity. This model is a Gauss-Markov model without imposing normality.

Theorem 4.3. *In the homoscedastic iid model, L1-L4 hold with* $\Omega_0 = \sigma^2 E[x_t x_t']$ *and* $Q = E[x_t x_t']$*, so that* $V = V_0 := \sigma^2 Q^{-1}$.

Proof. This is a direct corollary of the previous result.

For the homoscedastic iid model, a consistent estimator (non-robust estimator) of variance V is given by:

$$\hat{V}_0 = s^2 (X'X/n)^{-1}$$

Theorem 4.4. *The conditions of the homoscedastic iid model imply that* $\hat{V}_0 \to_p V_0$.

Proof. The matrix Q^{-1} is consistently estimated by $(X'X/n)^{-1}$ by $X'X/n \xrightarrow{p} Q$, holding by LLN, and by CMT. We have that $s^2 = \hat{\varepsilon}'\hat{\varepsilon}/(n-K)$, where $\hat{\varepsilon} = y - X\hat{\beta}$.

It follows from $X'\varepsilon/n \xrightarrow{p} 0$ and $X'X/n \xrightarrow{p} Q$ that

$$\begin{array}{lccccccc} s^2 = & \left[\frac{n}{n-K}\right] & \left[\frac{\varepsilon'\varepsilon}{n} + \right. & 2\left(\beta-\hat{\beta}\right)' & \left(\frac{X'\varepsilon}{n}\right) + & \left(\beta-\hat{\beta}\right)' & \left(\frac{X'X}{n}\right) & \left.\left(\beta-\hat{\beta}\right)\right] \underset{(CMT)}{\to_p} \sigma^2, \\ & \downarrow & \downarrow p & \downarrow p & \downarrow p & \downarrow p & \downarrow p & \\ & 1 & \sigma^2 & 0 & 0 & 0 & Q & \\ & & (\text{LLN}) & (\text{OLS}) & (\text{LLN}) & (\text{OLS}) & (\text{LLN}) & \end{array}$$

where (OLS) refers to consistency of the OLS estimator. Thus, by the CMT, $s^2 \left(X'X/n\right)^{-1} \xrightarrow{p} \sigma^2 Q^{-1}$.□

Comment: Under heteroscedasticity $V \neq V_0$, and V may be larger or smaller than V_0. Convince yourself of this. In practice, V is often larger than V_0.

4.3. **Dependent Samples.** In many macro-economic and financial settings, time series are serially dependent (dependent across t). Think of some examples.

There are many ways to model the dependence. You will see some parametric models in 14.382 in connection to GLS. Here we describe basic non-parametric models.

In what follows it will be convenient to think of the data $z_t = (y_t, x_t')', t = 1, ..., n$ as a subset of some infinite stream $\{z_t\} = \{z_t, t = \pm 1, \pm 2, ...\}$..

$\{z_t\}$ is said to be *stationary*, if for each $k \geq 0$, distribution of $(z_t, ..., z_{t+k})$ equals the distribution of $(z_1, ..., z_{1+k})$, i.e. does not depend on t. As a consequence, we have e.g. mean and covariance stationarity: $E[z_t] = E[z_1]$ for each t and $E[z_t z_{t+k}'] = E[z_1 z_{1+k}']$ for each t.

(a) Mixing. Mixing is a fairly mathematical way of thinking about temporal dependence.

Let$\{h_t\}$ be a stationary process. Define for $k \geq 1$

$$\alpha_k = \sup_{A,B} \Big| \ P(A \cap B) - P(A)P(B) \Big|$$

where sup is taken over events $A \in \sigma(h_0, h_{-1}, h_{-2}, ...,)$ and $B \in \sigma(h_k, h_{k+1}, ...)$. A simple way to think of these sigma-fields is as information generated by variables enclosed in brackets. $\{h_t\}$ is said to be *strongly mixing* or *alpha mixing* if $\alpha_k \to 0$ as $k \to \infty$. If the data are i.i.d. $\alpha_k = 0$ for each k.

The mixing condition states that dependence between blocks of data separated by k units of time dissipates as k increases. Many parametric models have been shown to be mixing under some regularity conditions.

Lemma 4.3 (An Ergodic LLN)**.** *If $\{h_t\}$ is a stationary strongly mixing with a finite mean $E[h_t]$, then $E_n[h_t] \to_p E[h_t]$.*

Remark. There is a more general version of this LLN, called Birkhoff ergodic theorem.[8]

[8]See e.g. http://mathworld.wolfram.com/BirkhoffsErgodicTheorem.html

Lemma 4.4 (Gordin's CLT)**.** *If $\{h_t\}$ is a stationary strongly mixing process, with $E[h_t] = 0$, $\sum_{k=1}^{\infty} \alpha_k^{\delta/(2+\delta)} < \infty$, $E\|h_t\|^{2+\delta} < \infty$, then*

$$\sum_{t=1}^{n} h_t/\sqrt{n} \ \to_d N(0, \Omega),$$

where

$$\Omega \ = \lim_n Var(\sum_{t=1}^{n} h_t/\sqrt{n}) = \lim_n \left(\Omega_0 + \sum_{k=1}^{n-1} \frac{n-k}{n} (\Omega_k + \Omega_k') \right) = \Omega_0 + \sum_{k=1}^{\infty} (\Omega_k + \Omega_k') < \infty,$$

where $\Omega_0 = Eh_1 h_1'$ and $\Omega_k = Eh_1 h_{1+k}'$.

The restriction on the rate of mixing and the moment condition imply the covariances sum up to a finite quantity Ω; see remark below. If this happens, the series is said to be weakly-dependent.

It is helpful to think of a sufficient condition that implies covariance summability: as $k \to \infty$ it suffices to have

$$\Omega_k / k^{-c} \to 0, \text{ for } c > 1.$$

Covariances should decay faster than $1/k$. If this does not happen, then the series is said to have long memory. High frequency data in finance often is thought of as having long memory, because the covariances decrease very slowly. The asymptotic theory under long memory is significantly different from the asymptotic theory presented here. [Reference: H. Koul.]

Remark. In Gordin's theorem, covariance summability follows from Ibragimov's mixing inequality for stationary series (stated here for the scalars):

$$|\Omega_k| = |Eh_t h_{t+k}| \leq \alpha_k^{1-\gamma} [E[h_t]^p]^{1/p} [E[h_t]^q]^{1/q}, \quad \frac{1}{p} + \frac{1}{q} = \gamma \in (0,1).$$

Setting $p = 2 + \delta$, we see that the covariance summability $\sum_{k=-\infty}^{\infty} |\Omega_k| < \infty$ follows from the restriction made on the mixing coefficients.

Theorem 4.5. *Suppose that the series $\{(y_t, x_t)\}$ is stationary and strongly mixing and that L1 and L2 hold. Suppose that $\{\tilde{h}_t\} = \{x_t x_t'\}$ has finite mean. Then L3 holds with $Q = E[x_t x_t']$. Suppose that $\{h_t\} = \{x_t e_t\}$ satisfies Gordin's conditions. Then L4 holds with Ω of the form stated above.*

Proof: The result follows from the previous two theorems, and from the definition of mixing.

The formula above suggests the following estimator for Ω:

$$\hat{\Omega} = \hat{\Omega}_0 + \sum_{k=1}^{L-1} \frac{L-k}{L} \left(\hat{\Omega}_k + \Omega_k' \right),$$

where $\hat{\Omega}_0 = E_n[h_t h_t'] = \frac{1}{n} \sum_{t=1}^{n} h_t h_t'$ and $\hat{\Omega}_k = E_n[h_t h_{t+k}'] = \frac{1}{n-k} \sum_{t=1}^{n-k} h_t h_{t+k}'$. Under certain technical conditions and conditions on the truncation lag, such as $L/n \to 0$ and $L \to \infty$, the estimator has been shown to be consistent.

The estimator $\hat{V} = \hat{Q}^{-1}\hat{\Omega}\hat{Q}^{-1}$ with $\hat{\Omega}$ of the form stated above is often called a HAC estimator ("heteroscedasticity and autocorrelation consistent" estimator). Under some regularity conditions, it is indeed consistent. For the conditions and the proof, see Newey and West.

(b) Martingale difference sequences (MDS). Data can be temporally dependent but covariances Ω_k can still be zero ("all-pass" series), simplifying Ω above to Ω_0. MDSs are one example where this happens. MDSs are important in connection with the rational expectations (cf. Hansen and Singleton) as well as the efficient market hypothesis (cf. Fama). A detailed textbook reference is H. White, Asymptotic Theory for Econometricians.

Let h_t be an element of z_t and $I_t = \sigma(z_t, z_{t-1}, ...)$. The process $\{h_t\}$ is a martingale with respect to filtration I_{t-1} if $E[h_t|I_{t-1}] = h_{t-1}$. The process $\{h_t\}$ is martingale difference sequence with respect to I_{t-1} if $E[h_t|I_{t-1}] = 0$.

Example. In Hall's model of a representative consumer with quadratic utility and rational expectations, we have that

$$E[y_t|I_{t-1}] = 0,$$

where $y_t = c_t - c_{t-1}$ and I_t is the information available at period t. That is, a rational, optimizing consumer sets his consumption today such that no change in the mean of subsequent consumption is anticipated.

Lemma 4.5 (Billingsley's Martingale CLT)**.** *Let $\{h_t\}$ be a martingale difference sequence that is stationary and strongly mixing with $\Omega = E[h_t h_t']$ finite. Then $\sqrt{n}E_n[h_t] \to_d N(0, \Omega)$.*

The theorem makes some intuitive sense, since h_t's are identically distributed and also are uncorrelated. Indeed, $E[h_t h_{t-k}'] = E[E[h_t h_{t-k}'|I_{t-k}]] = E[E[h_t|I_{t-k}]h_{t-k}'] = 0$ for $k \geq 1$, since $E[h_t|I_{t-k}] = E[E[h_t|I_{t-1}]|I_{t-k}] = 0$. There is a nice generalization of this theorem due to McLeish.

Theorem 4.6. *Suppose that series $\{(y_t, x_t)\}$ is stationary and strongly mixing. Further suppose that (a) $e_t = y_t - x_t'\beta$ is a martingale difference sequence with respect to the filtration $I_{t-1} = \sigma((e_{t-1}, x_t)', (e_{t-2}, x_{t-1})', ...)$, that is $E[e_t|I_{t-1}] = 0$, (b) $\Omega = E[e_t^2 x_t x_t']$ is finite, and (c) $Q = E[x_t x_t']$ is finite and of full rank. Then L1-L4 hold with Ω and Q defined above. Therefore, the conclusions of Propositions 1-3 also hold.*

Proof. We have that $E[e_t|x_t] = E[E[e_t|I_{t-1}]|x_t] = 0$, which implies L1 and L2. We have that $E_n[x_t x_t'] \to_p Q = E[x_t x_t']$ by the Ergodic LLN, which verifies L3. We have that $\sqrt{n}E_n[e_t x_t] \to_d N(0, \Omega)$ with $\Omega = E[e_t^2 x_t x_t']$ by the martingale CLT, which verifies L4. ∎

Remark 4.1. *We can use the same estimator for Ω and Q as in the i.i.d. case,*

$$\widehat{\Omega} = E_n[\hat{e}_t^2 x_t x_t'], \quad \hat{Q} = E_n[x_t x_t'].$$

Consistency of $\widehat{\Omega}$ follows as before under the assumption that $E[\|x_t\|^4] < \infty$. The proof is the same as before except that we should use the Ergodic LLN instead of the usual LLN for iid data. Consistency of $\hat{Q}$ also follows by the ergodic LLN.

Example. In Hall's example,

$$E[y_t - x_t'\beta_0 | x_t] = 0,$$

for $\beta_0 = 0$ and x_t representing any variables entering the information set at time $t-1$. Under the Hall's hypothesis,

$$\sqrt{n}(\hat{\beta} - \beta_0) \ \to_d N(0,\ V), \quad V = Q^{-1}\Omega Q^{-1}, \quad \beta_0 = 0$$

where $\Omega = E[y_t^2 x_t x_t']$ and $Q = E x_t x_t'$. Then one can test the Hall's hypothesis by using the Wald statistic:

$$W = \ \sqrt{n}(\hat{\beta} - \beta_0)' \widehat{V}^{-1} \sqrt{n}(\hat{\beta} - \beta_0),$$

where $\widehat{V} = \widehat{Q}^{-1}\widehat{\Omega}\widehat{Q}^{-1}$ for $\widehat{\Omega}$ and $\widehat{Q}$ defined above.