Notes on Regression Asymptotics 15 August 2007

DA Freedman

We're in the OLS model $Y = X\beta + \epsilon$, the ϵ_i being IID, with mean 0 and finite variance σ^2. Take the $n \times p$ matrix X as fixed; or assume the errors are independent of X and condition on X. We impose the following regularity conditions: $n \to \infty$, p is fixed, $X'X/n \to V$ positive definite $p \times p$, and the largest element of X is $o(\sqrt{n})$.

Theorem 1. Under the foregoing regularity conditions, $n^{1/2}(\hat{\beta} - \beta)$ is asymptotically normal, with covariance matrix V^{-1}.

Theorem 2. Under the foregoing regularity conditions, when the null hypothesis restricts p_0 components of β to vanish, the asymptotic distribution of F is $\chi^2_{p_0}/p_0$.

Argument for Theorem 1. Let X_i be the ith row of X. Fix c, a $p \times 1$ vector. Now

$$c'X'\epsilon = \sum_{i=1}^{n} T_i \quad \text{with} \quad T_i = (X_i c)\epsilon_i.$$

The T_i are independent with mean 0. And $X_i c = o(\sqrt{n})$. Furthermore, var $(c'X'\epsilon) = \sigma^2 c'X'Xc$ is of order n. Now we can appeal to a central limit theorem for independent non-identically distributed components, each being small relative to the total (e.g., Lindeberg's theorem, Feller Vol. II 1971 p. 518). Finally, $\hat{\beta} - \beta = (X'X)^{-1}X'\epsilon$.

All would seem to go through if the ϵ_i are independent, mean 0, constant variance σ^2, not identically distributed, although some uniform integrability is needed; triangular arrays are probably ok too. If e.g. there is an a priori bound on $E(|\epsilon_i|^3)$, we can presumably get a Berry-Esseen type of error bound on the difference between scaled $\hat{\beta}$ and the approximating normal distribution. Probably $p = o(\sqrt{n})$ is ok too.

Argument for Theorem 2. The error variance is a consistent estimator for σ^2, so the denominator of F goes to σ^2. In a little more detail, let $H = X(X'X)^{-1}X'$ be the hat matrix. The residuals are $e = (I - H)Y = (I - H)\epsilon$. The denominator of the F-statistic is $\|e\|^2/(n - p)$. Now $E(\|H\epsilon\|^2) = \sigma^2 p = o(n)$. Thus, $E(\|e - \epsilon\|^2) = o(n)$. That's all we need for convergence in distribution.

For the numerator, let X_u be the $p - p_0$ columns of X whose coefficients are unconstrained by the null hypothesis (u for unconstrained). Let $\hat{\beta}_u$ be the OLS estimator for those coefficients, i.e., in the small model with the p_0 constraints imposed. We have to get our hands on $\|X\hat{\beta}\|^2$ and $\|X_u\hat{\beta}_u\|^2$, and then the difference. Let X_c be the p_0 columns of X whose coefficients are constrained to 0 by the null hypothesis (c for constrained). Let $\hat{\beta}_c$ be the OLS estimator for those coefficients, i.e., in the full model with no constraints.

1) F depends only on Y and the column spaces of X_u and X: indeed, $X\hat{\beta}$ is the projection of Y onto X, whilst $X_u\hat{\beta}_u$ is the projection of Y onto X_u. AWLOG that X_u consists of the first $p - p_0$ columns of X; the null hypothesis constrains the last p_0 entries of β to be 0.

2) Let $W = X'X$.

3) In the leading special case, X has orthogonal columns with squared length n, so $W = nI_{p\times p}$; the elements of X are uniformly $o(\sqrt{n})$. The numerator of F is $n\|\hat{\beta}_c\|^2/p_0$ and Theorem 1 applies. Pause to verify the numerator of F. First, $X_c\hat{\beta}_c \perp X_u\hat{\beta}_u$. So $\|X\hat{\beta}\|^2 = \|X_c\hat{\beta}_c\|^2 + \|X_u\hat{\beta}_u\|^2$ and the numerator of F is $\|X_c\hat{\beta}_c\|^2/p_0$. (See, e.g., section 4.8 in Freedman 2005.) But $\|X_c\hat{\beta}_c\|^2 = \hat{\beta}_c'X_c'X_c\hat{\beta}_c = n\|\hat{\beta}_c\|^2$. Under the null, $E(\hat{\beta}_c) = 0_{p_0\times 1}$, and $\text{cov}(\hat{\beta}_c) = I_{p_0\times p_0}/n$. That is where the $\chi^2_{p_0}$ comes from.

4) Reduce the general case to the special case by doing Gram-Schmidt on X; normalize the output columns to have squared length n. If A is $p \times p$ non-singular, the column space of XA coincides with the column space of X; for Gram-Schmidt, A is upper triangular. Call the output matrix $\tilde{X}$. By construction, $\tilde{X}'\tilde{X} = nI_{p\times p}$. The column space of X coincides with the column space of $\tilde{X}$. Likewise, the linear space $\mathcal{L}$ spanned by the first $p - p_0$ columns of X coincides with the linear space spanned by the first $p - p_0$ columns of $\tilde{X}$. The null hypothesis says that $E(Y) \in \mathcal{L}$.

5) In order to use Theorem 1, we need to check that the maximum element of $\tilde{X}$ is $o(\sqrt{n})$. This can be done by induction on p. The case $p = 1$ is obvious. Let's go from $p - 1$ to p. Recall that $W = X'X$, so $W = nV + o(n)$. Let W_0 denote the top left $(p-1)\times(p-1)$ corner of W, and let $W_1 = (W_{p,1}, \ldots, W_{p,p-1})'$, so W_1 is $(p-1)\times 1$. Define V_0 and V_1 in a similar way. Let X^p be column p in X and $X_{(p-1)}$ the first $p-1$ columns. The projection of X^p onto $X_{(p-1)}$ is $X_{(p-1)}W_0^{-1}W_1$, whose elements are $o(\sqrt{n})$—because $W_0^{-1}W_1 \to V_0^{-1}V_1$ and the elements of $X_{(p-1)}$ are $o(\sqrt{n})$. A similar conclusion must therefore apply to $X^p - X_{(p-1)}W_0^{-1}W_1$.

6) We must also check that $X^p - X_{(p-1)}W_0^{-1}W_1$ has length of order $\sqrt{n}$; otherwise, renormalizing length could make trouble. The squared length of the projection is $W_1'W_0^{-1}W_1$. The squared length of the original vector is W_{pp}. The difference is $n(V_{pp} - V_1'V_0^{-1}V_1) + o(n)$ and $\Delta = V_{pp} - V_1'V_0^{-1}V_1 > 0$ because V is positive definite. In more detail, V can be realized as the inner products of pairs of a set of p linearly independent vectors of dimension $p \times 1$. The difference Δ is the squared length of the pth vector net of the first $p-1$ vectors. (A weird argument, but I don't see a direct calculation; more below.)

A more elegant set of conditions might be—

Let $W = X'X$. Let s be the smallest eigenvalue of W, and B the biggest. We require $s \to \infty$, $B = O(s)$, and the largest element of X is $o(\sqrt{s})$. Argument seems to be the same, not checked though. Presumably, normalize Gram-Schmidt so squared length is s. We should get that $W^{-1/2}(\hat{\beta} - \beta)$ tends in law to $N(0_{p\times 1}, I_{p\times 1})$. Check also that W/s is precompact in the set of positive definite matrices (see below). Confirm that

$$s = \min_x x'Wx, \quad B = \max_x x'Wx, \quad s = \min_x \|Wx\|, \quad B = \max_x \|Wx\|,$$

the min and max being taken over x with ℓ_2-norm equal to 1. In particular, the eigenvalues of W_0 are between s and B. (In fact, although irrelevant here, the eigenvalues of the two matrices are interlaced.) Also, B is the L_2 norm of W, so any row (or column) of W has ℓ_2-norm at most B. Especially, W_1 has ℓ_2-norm which is $O(s)$, so $\|W_0^{-1}W_1\| = O(1)$.

Precompactness of W/s

If $0 < \alpha < \beta < \infty$, the set of $p \times p$ symmetric matrices with $\alpha \le x'Wx \le \beta$ for all x having $\|x\| = 1$ is a closed bounded set.

The argument for V_{pp}

We can realize V above as $Z'Z$, where $VR = RD$ with R orthogonal and D diagonal, and e.g. $Z = \sqrt{D}R'$. The difference $V_{pp} - V_1' V_0^{-1} V_1$ is the squared length of the pth column of Z, net of the projection into the first $p - 1$ columns. This length has to be positive: Z is nonsingular because R is nonsingular.

References

Anderson TW (1971). *The Statistical Analysis of Time Series*. New York, Wiley. §2.6.

Anderson TW and Taylor JB (1979). Strong consistency of least squares estimates in dynamic models. *Annals of Statistics* 7: 484–89.

Drygas H (1971). Consistency of the least squares and Gauss-Markov estimators in regression models. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete* 17: 309–326.

Feller W (1971). *An Introduction to Probability Theory and its Applications*. Vol. II, 2nd ed., Wiley, New York.

Freedman DA (1981). Bootstrapping regression models. *Annals of Statistics* 6: 1218–28.

Freedman DA (2005). *Statistical Models: Theory and Practice*. Cambridge University Press.

Lai TL, Robbins H, Wei CZ (1979). Strong consistency of least squares estimates in multiple regression. *Journal of Multivariate Analysis* 9: 343–361.