

Chapter 6. Additional Topics on Linear Regression*

This chapter covers additional topics on linear regression. Related materials can be found in Chapter 5 of Hansen (2007).

We first collect assumptions in the previous chapters:

Assumption OLS.0 (random sampling): $(y_i, \mathbf{x}_i), i = 1, \dots, n$, are i.i.d.

Assumption OLS.1 (full rank): $\text{rank}(\mathbf{X}) = k$.

Assumption OLS.1': $\text{rank}(E[\mathbf{x}\mathbf{x}']) = k$.

Assumption OLS.2 (first moment): $E[y|\mathbf{x}] = \mathbf{x}'\boldsymbol{\beta}$.

Assumption OLS.2': $y = \mathbf{x}'\boldsymbol{\beta} + u$ with $E[\mathbf{x}u] = \mathbf{0}$.

Assumption OLS.3 (second moment): $E[u^2] < \infty$.

Assumption OLS.3' (homoskedasticity): $E[u^2|\mathbf{x}] = \sigma^2$.

Assumption OLS.4 (normality): $u|\mathbf{x} \sim N(0, \sigma^2)$.

Assumption OLS.5: $E[u^4] < \infty$ and $E[\|\mathbf{x}\|^4] < \infty$.

Different assumptions imply different properties of the LSE as summarized in Table 1.

	$y = \mathbf{x}'\boldsymbol{\beta} + u$		Implied Properties
$E[\mathbf{x}u] = \mathbf{0}$	linear projection	\implies	consistency
	\cup		
$E[u \mathbf{x}] = 0$	linear regression	\implies	unbiasedness
	\cup		
$E[u \mathbf{x}] = 0$ and $E[u^2 \mathbf{x}] = \sigma^2$	homoskedastic linear regression	\implies	Gauss-Markov Theorem
	\cup		
u is independent of \mathbf{x}	normal regression is a special case	\implies	UMVUE

Table 1: Relationship between Different Models

*Email: pingyu@hku.hk

This chapter will examine the validity of these assumptions and cures when they fail. Two key assumptions we will examine are OLS.2 and OLS.3'. Assumption OLS.2 has many implications. For example, it implies (i) the conditional mean of y given \mathbf{x} is linear in \mathbf{x} , and (ii) all relevant regressors are included in \mathbf{x} and are fixed. Section 1 and 2 examine the first implication: Section 1 tests whether $E[y|\mathbf{x}]$ is indeed $\mathbf{x}'\boldsymbol{\beta}$ and Section 2 provides more flexible specifications of $E[y|\mathbf{x}]$. Section 3 and 4 examine the second implication: Section 3 checks the benefits and costs of including irrelevant variables or omitting relevant variables, and Section 4 provides some model selection procedures based on information criteria. Section 5 and 6 examine Assumption OLS.3': Section 5 shows that there are more efficient estimators of $\boldsymbol{\beta}$ when this assumption fails and Section 6 checks whether this assumption fails. Finally, Section 7 examines the external validity of the model.

1 Tests for Functional Form Misspecification

Misspecification of $E[y|\mathbf{x}]$ may be due to omitted variables or misspecified functional forms. In this section, we only examine the second source of misspecification and provide a general test of the adequacy of the specification of $E[y|\mathbf{x}]$.

One simple test for neglected nonlinearity is to add nonlinear functions of the regressors to the regression, and test their significance using a Wald test. Thus, if the model $y_i = \mathbf{x}_i'\hat{\boldsymbol{\beta}} + \hat{u}_i$ has been fit by OLS, let $\mathbf{z}_i = h(\mathbf{x}_i)$ denote functions of \mathbf{x}_i which are not linear functions of \mathbf{x}_i (perhaps squares of non-binary regressors) and then fit $y_i = \mathbf{x}_i'\tilde{\boldsymbol{\beta}} + \mathbf{z}_i'\tilde{\boldsymbol{\gamma}} + \tilde{u}_i$ by OLS, and form a Wald statistic for $\boldsymbol{\gamma} = \mathbf{0}$.

Another popular approach is the REgression Specification Error Test (RESET) proposed by Ramsey (1969, 1970). The null model is

$$y_i = \mathbf{x}_i'\boldsymbol{\beta} + u_i,$$

which is estimated by OLS, yielding predicted values $\hat{y}_i = \mathbf{x}_i'\hat{\boldsymbol{\beta}}$. Now let

$$\mathbf{z}_i = \begin{pmatrix} \hat{y}_i^2 \\ \vdots \\ \hat{y}_i^m \end{pmatrix}$$

be an $(m - 1)$ -vector of powers of \hat{y}_i . Then run the auxiliary regression

$$y_i = \mathbf{x}_i'\tilde{\boldsymbol{\beta}} + \mathbf{z}_i'\tilde{\boldsymbol{\gamma}} + \tilde{u}_i \tag{1}$$

by OLS, and form the Wald statistic W_n for $\boldsymbol{\gamma} = \mathbf{0}$. \hat{y}_i 's are *generated regressors* in the term of Pagan (1984) (see also Murphy and Topel, 1985). However, for *testing* purposes, using estimates from earlier stages causes no complications (because under H_0 , the coefficients associated with generated regressors are zero). So under H_0 , $W_n \xrightarrow{d} \chi_{m-1}^2$. Thus the null is rejected at the α level if W_n exceeds the upper α tail critical value of the χ_{m-1}^2 distribution. To implement the test, m

must be selected in advance. Typically, small values such as $m = 2, 3$, or 4 seem to work best.

The RESET test appears to work well as a test of functional form against a wide range of smooth alternatives. It is particularly powerful in detecting the single-index model of Ichimura (1993),

$$y_i = G(\mathbf{x}'\boldsymbol{\beta}) + u_i,$$

where $G(\cdot)$ is a smooth "link" function. To see why this is the case, note that (1) may be written as

$$y_i = \mathbf{x}'_i \tilde{\boldsymbol{\beta}} + \left(\mathbf{x}'_i \hat{\boldsymbol{\beta}}\right)^2 \tilde{\gamma}_1 + \cdots + \left(\mathbf{x}'_i \hat{\boldsymbol{\beta}}\right)^m \tilde{\gamma}_{m-1} + \tilde{u}_i,$$

which has essentially approximated $G(\cdot)$ by an m th order polynomial.

(**) Another specification test is White's (1980a, 1981) version of Hausman's (1978) test. The idea of this test is that if the model is correctly specified, then both the LSE and the WLS estimator are consistent and their difference should be close to zero. Specifically, under the null,

$$n \left(\hat{\boldsymbol{\beta}}_{OLS} - \hat{\boldsymbol{\beta}}_{WLS} \right)' \hat{\mathbf{V}}^{-1} \left(\hat{\boldsymbol{\beta}}_{OLS} - \hat{\boldsymbol{\beta}}_{WLS} \right) \xrightarrow{d} \chi_k^2,$$

where $\hat{\mathbf{V}}$ is a consistent estimator of \mathbf{V} , the asymptotic variance of $\sqrt{n} \left(\hat{\boldsymbol{\beta}}_{OLS} - \hat{\boldsymbol{\beta}}_{WLS} \right)$ under the null. \mathbf{V} generally takes a complicated form, but under the auxiliary assumption of homoskedasticity, it takes a neat form,

$$\begin{aligned} \mathbf{V} &= n \left[AVar(\hat{\boldsymbol{\beta}}_{WLS}) - AVar(\hat{\boldsymbol{\beta}}_{OLS}) \right] \\ &= E \left[w_i \mathbf{x}_i \mathbf{x}'_i \right]^{-1} E \left[w_i^2 \mathbf{x}_i \mathbf{x}'_i u_i^2 \right] E \left[w_i \mathbf{x}_i \mathbf{x}'_i \right]^{-1} - \sigma^2 E \left[\mathbf{x}_i \mathbf{x}'_i \right]^{-1}. \end{aligned} \quad (2)$$

This simplification is due to the fact that although both the LSE and the WLS estimator are consistent under the null, the LSE is efficient under homoskedasticity.

Exercise 1 (i) Suppose in model $y_i = \mathbf{x}'_i \boldsymbol{\beta} + u_i$, $E[u_i | \mathbf{x}_i] = 0$ and $E[u_i^2 | \mathbf{x}_i] = \sigma^2$. Show that $Avar \left(\sqrt{n} \left(\hat{\boldsymbol{\beta}}_{OLS} - \hat{\boldsymbol{\beta}}_{WLS} \right) \right) = \mathbf{V}$ in (2). (ii) If $E[u_i^2 | \mathbf{x}_i]$ is not a constant, what is the expression for $Avar \left(\sqrt{n} \left(\hat{\boldsymbol{\beta}}_{OLS} - \hat{\boldsymbol{\beta}}_{WLS} \right) \right)$? If $w_i = \sigma_i^{-2}$, what will the expression for $Avar \left(\sqrt{n} \left(\hat{\boldsymbol{\beta}}_{OLS} - \hat{\boldsymbol{\beta}}_{WLS} \right) \right)$ change to?

\mathbf{V} can be estimated by its sample analog. But such an estimator is generally consistent only under the null (and the auxiliary assumption of homoskedasticity). A covariance matrix estimator that is consistent regardless of misspecification is given in White (1980b). Theorem 2.2 of Hausman (1978) shows that the test statistic has a noncentral χ^2 distribution for a sequence of local alternative hypotheses, with a noncentrality parameter depending on $\text{plim}(\hat{\boldsymbol{\beta}}_{OLS} - \hat{\boldsymbol{\beta}}_{WLS})$. To generate power, the weights in $\hat{\boldsymbol{\beta}}_{WLS}$ are important. White (1981) suggests to impose weights on the area where the linear approximation is bad. Specifically, the weights are predicted \hat{u}_i^2 using all non-redundant elements of \mathbf{x}_i , its squares, and all cross-products.¹

¹So we should include $2(k-1) + C_{k-1}^2$ nonconstant regressors on the right-hand side. But if \mathbf{x}_i includes polynomial or dummy terms, we should adjust the number of regressors accordingly.

All these tests mentioned above are special cases of the conditional moment (CM) test of Newey (1985a) and Tauchen (1985).² Specifically, under H_0 , $E[u|\mathbf{x}] = 0$, so any function of \mathbf{x} is orthogonal to u . Different tests just pick different functions of \mathbf{x} to form orthogonal moment conditions. But all such tests can only detect some form of deviation from the null. To detect any possible deviation from the null, we essentially need infinite moment conditions, which seems formidable. Nevertheless, Bierens (1982, 1990) extends the CM test to achieve this goal; see also Bierens and Ploberger (1997) for the integrated conditional moment (ICM) test. Another test that can detect any deviation from the null is proposed by Zheng (1996). The idea of his test is that

$$E[uE[u|\mathbf{x}]f(\mathbf{x})] = E[E[u|\mathbf{x}]^2 f(\mathbf{x})] = \int \left[\int u f(u, \mathbf{x}) du \right]^2 d\mathbf{x} \geq 0,$$

where $f(\mathbf{x})$ is the density of \mathbf{x} . Equality can be achieved only if $E[u|\mathbf{x}] = 0$, so this test would have power to detect any deviation from $E[y|\mathbf{x}] = \mathbf{x}'\beta$. $f(\mathbf{x})$ is added in to offset the denominator in $E[u|\mathbf{x}] = \int u \frac{f(u, \mathbf{x})}{f(\mathbf{x})} du$.³ This test is constructed under the null (e.g., u_i is estimated in the null model), so is similar to the score test in spirit; see Yu (2014) for more discussions. (**)

2 Nonlinear Least Squares

If the specification test rejects the linear specification in the least squares estimation, we may consider to use a nonlinear setup for the regression function $E[y|\mathbf{x}]$. Specifically, suppose $E[y_i|\mathbf{x}_i = \mathbf{x}] = m(\mathbf{x}|\boldsymbol{\theta})$. For a comprehensive treatment of nonlinear regression, see Seber and Wild (2003). Nonlinear regression means that $m(\mathbf{x}|\boldsymbol{\theta})$ is a nonlinear function of $\boldsymbol{\theta}$ (rather than \mathbf{x}). The functional form of $m(\mathbf{x}|\boldsymbol{\theta})$ can be suggested by an economic model, or as in the LSE, it can be treated as a nonlinear approximation to a general conditional mean function (see White (1981)). Examples of nonlinear regression functions include the following.

- $m(\mathbf{x}|\boldsymbol{\theta}) = \exp(\mathbf{x}'\boldsymbol{\theta})$: Exponential Link Regression

The exponential link function is strictly positive, so this choice can be useful when it is desired to constrain the mean to be strictly positive.

- $m(x|\boldsymbol{\theta}) = \theta_1 + \theta_2 x^{\theta_3}$, $x > 0$: Power Transformed Regressors

A generalized version of the power transformation is the famous Box-Cox (1964) transformation, where the regressor x^{θ_3} is generalized as $x^{(\lambda)}$ with

$$x^{(\lambda)} = \begin{cases} \frac{x^\lambda - 1}{\lambda}, & \text{if } \lambda > 0, \\ \log x, & \text{if } \lambda = 0. \end{cases}$$

²Note that these authors consider the CM test in the context of likelihood models.

³(*) This is essentially to avoid random denominators in the nonparametric kernel estimation of $E[u|\mathbf{x}]$ which will not be covered in this course.

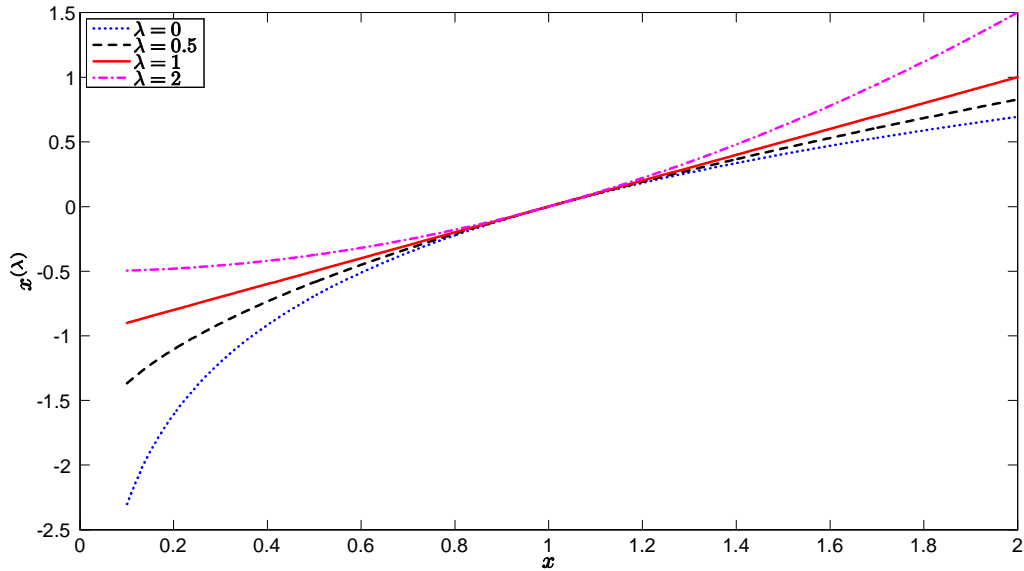


Figure 1: Box-Cox Transformation for Different λ Values

The function $x^{(\lambda)}$ nests linearity ($\lambda = 1$) and logarithmic ($\lambda = 0$) transformations continuously. Figure 1 shows the Box-Cox transformations for different λ values. All transformations pass the point $(1, 0)$.

- $m(x|\boldsymbol{\theta}) = \theta_1 + \theta_2 \exp(\theta_3 x)$: Exponentially Transformed Regressors
- $m(\mathbf{x}|\boldsymbol{\theta}) = G(\mathbf{x}'\boldsymbol{\theta})$, G known⁴

When $G(\cdot) = (1 + \exp(-\cdot))^{-1}$, the regression with $m(\mathbf{x}|\boldsymbol{\theta}) = G(\mathbf{x}'\boldsymbol{\theta})$ is called the **logistic link regression**; when $G(\cdot) = \Phi(\cdot)$ with $\Phi(\cdot)$ being the cdf of standard normal, it is called the **probit link regression**.

- $m(\mathbf{x}|\boldsymbol{\theta}) = \boldsymbol{\theta}'_1 \mathbf{x}_1 + \boldsymbol{\theta}'_2 \mathbf{x}_1 G\left(\frac{x_2 - \theta_3}{\theta_4}\right)$: Smooth Transition
- $m(x|\boldsymbol{\theta}) = \theta_1 + \theta_2 x + \theta_3 (x - \theta_4) 1(x > \theta_4)$: Continuous Threshold Regression
- $m(\mathbf{x}|\boldsymbol{\theta}) = (\boldsymbol{\theta}'_1 \mathbf{x}_1) 1(x_2 \leq \theta_3) + (\boldsymbol{\theta}'_2 \mathbf{x}_1) 1(x_2 > \theta_3)$: Threshold Regression

For surveys of the smooth transition model (STM), see Teräsvirta (1998), Teräsvirta et al. (2010) and van Dijk et al. (2002); for the continuous threshold regression (CTR) model, see Chan and Tsay (1998); for surveys of the threshold regression (TR) model, see Hansen (2011) and Tong (1990). When $\theta_4 = 0$, the STM reduces to the TR model, and when $\theta_4 = \infty$, the STM reduces to linear regression. Figure 2 shows the difference between the STM ($\mathbf{x}_1 = (1, x)'$, $x_2 = x$,

⁴This is different from the single index model mentioned in the Introduction where $G(\cdot)$ is unknown.

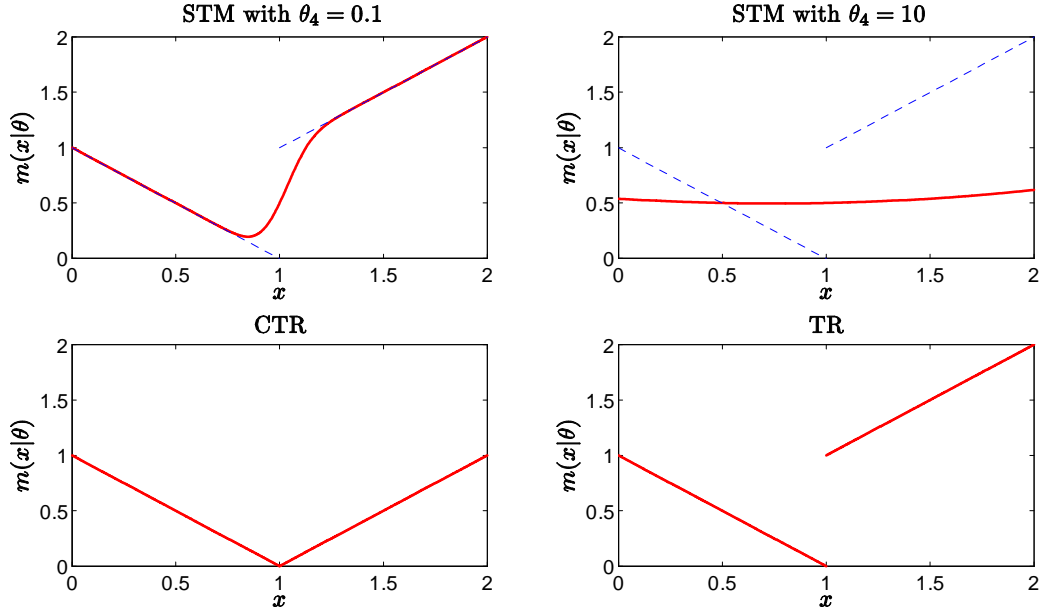


Figure 2: Difference Between STM, CTR and TR

$\boldsymbol{\theta}_1 = (1, -1)'$, $\boldsymbol{\theta}_2 = (-2, 2)'$, $\theta_3 = 1$, $\theta_4 = 0.1$ in the top left panel and $\theta_4 = 10$ in the top right panel, and $G(\cdot) = \Phi(\cdot)$, the CTR model ($(\theta_1, \theta_2) = (1, -1)$ and $(\theta_3, \theta_4) = (2, 1)$) and the TR model ($\boldsymbol{\theta}_1 = (1, -1)'$, $\boldsymbol{\theta}_2 = (0, 1)'$ and $\theta_3 = 1$) for $x \in [0, 2]$.

In the first five examples, $m(\mathbf{x}|\boldsymbol{\theta})$ is (generically) differentiable with respect to the parameters $\boldsymbol{\theta}$. In the last two examples, m is not differentiable with respect to θ_4 and θ_3 which alters some of the analysis. When it exists, let

$$\mathbf{m}_\theta(\mathbf{x}|\boldsymbol{\theta}) = \frac{\partial}{\partial \boldsymbol{\theta}} m(\mathbf{x}|\boldsymbol{\theta}).$$

The least squares estimator $\hat{\boldsymbol{\theta}}$ minimizes the sum of squared errors

$$S_n(\boldsymbol{\theta}) = \sum_{i=1}^n (y_i - m(\mathbf{x}_i|\boldsymbol{\theta}))^2.$$

When the regression function is nonlinear, we call this the **nonlinear least squares** (NLLS) estimator. The NLLS residuals are $\hat{u}_i = y_i - m(\mathbf{x}_i|\hat{\boldsymbol{\theta}})$. One motivation for the choice of NLLS as the estimation method is that the parameter is the solution to the population problem $\min_{\boldsymbol{\theta}} E [(y_i - m(\mathbf{x}_i|\boldsymbol{\theta}))^2]$.

Since sum-of-squared-errors function $S_n(\boldsymbol{\theta})$ is not quadratic, $\hat{\boldsymbol{\theta}}$ must be found by numerical methods (as in the ML estimation). When $m(\mathbf{x}|\boldsymbol{\theta})$ is differentiable, then the FOCs for minimization are

$$\mathbf{0} = \sum_{i=1}^n \mathbf{m}_\theta(\mathbf{x}_i|\hat{\boldsymbol{\theta}}) \hat{u}_i.$$

Theorem 1 *If the model is identified and $m(\mathbf{x}|\boldsymbol{\theta})$ is differentiable with respect to $\boldsymbol{\theta}$,*

$$\sqrt{n}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \xrightarrow{d} N(\mathbf{0}, \mathbf{V}),$$

where $\mathbf{V} = E[\mathbf{m}_{\boldsymbol{\theta}_i} \mathbf{m}'_{\boldsymbol{\theta}_i}]^{-1} E[\mathbf{m}_{\boldsymbol{\theta}_i} \mathbf{m}'_{\boldsymbol{\theta}_i} u_i^2] E[\mathbf{m}_{\boldsymbol{\theta}_i} \mathbf{m}'_{\boldsymbol{\theta}_i}]^{-1}$ with $\mathbf{m}_{\boldsymbol{\theta}_i} = \mathbf{m}_{\boldsymbol{\theta}}(\mathbf{x}_i|\boldsymbol{\theta}_0)$.

Exercise 2 *Prove the above theorem. (Hint: Note that $\widehat{\boldsymbol{\theta}}$ is the MoM estimator associated with the moment conditions $E[\mathbf{m}_{\boldsymbol{\theta}_i} u_i] = \mathbf{0}$.)*

Based on this theorem, an estimate of the asymptotic variance \mathbf{V} is

$$\widehat{\mathbf{V}} = \left(\frac{1}{n} \sum_{i=1}^n \widehat{\mathbf{m}}_{\boldsymbol{\theta}_i} \widehat{\mathbf{m}}'_{\boldsymbol{\theta}_i} \right)^{-1} \left(\frac{1}{n} \sum_{i=1}^n \widehat{\mathbf{m}}_{\boldsymbol{\theta}_i} \widehat{\mathbf{m}}'_{\boldsymbol{\theta}_i} \widehat{u}_i^2 \right) \left(\frac{1}{n} \sum_{i=1}^n \widehat{\mathbf{m}}_{\boldsymbol{\theta}_i} \widehat{\mathbf{m}}'_{\boldsymbol{\theta}_i} \right)^{-1},$$

where $\widehat{\mathbf{m}}_{\boldsymbol{\theta}_i} = \mathbf{m}_{\boldsymbol{\theta}}(\mathbf{x}_i|\widehat{\boldsymbol{\theta}})$.

(**) Identification is often tricky in nonlinear regression models. Suppose that

$$m(\mathbf{x}_i|\boldsymbol{\theta}) = \boldsymbol{\beta}'_1 \mathbf{z}_i + \boldsymbol{\beta}'_2 \mathbf{x}_i(\gamma),$$

where $\mathbf{x}_i(\gamma)$ is a function of \mathbf{x}_i with an unknown parameter γ . Examples include $x_i(\gamma) = x_i^\gamma$, $x_i(\gamma) = \exp(\gamma x_i)$, $x_i(\gamma) = x_i G\left(\frac{x_i - \gamma_1}{\gamma_2}\right)$ and $x_i(\gamma) = x_i 1(g(x_i) > \gamma)$. The model is linear when $\boldsymbol{\beta}_2 = \mathbf{0}$, and this is often a useful hypothesis (sub-model) to consider. Thus we want to test

$$H_0 : \boldsymbol{\beta}_2 = \mathbf{0}.$$

However, under H_0 , the model is

$$y_i = \boldsymbol{\beta}'_1 \mathbf{z}_i + u_i$$

and both $\boldsymbol{\beta}_2$ and γ have dropped out. This means that under H_0 , γ is not identified. Such tests are labeled as tests with nuisance parameter (γ) unidentified under the null. This renders the distribution theory presented in the last chapter invalid. Thus when the truth is that $\boldsymbol{\beta}_2 = \mathbf{0}$, the parameter estimates are not asymptotically normally distributed. Furthermore, tests of H_0 do not have asymptotic normal or chi-square distributions. This kind of tests was first considered by Davies (1977, 1987). More discussions on the asymptotic theory of such tests can be found in Andrews (1993), Andrews and Ploberger (1994) and Hansen (1996) among others. In particular, Hansen (1996) shows how to use simulation (similar to the bootstrap) to construct the asymptotic critical values (or p -values) in a given application.

The asymptotic theory for $\widehat{\boldsymbol{\theta}}$ may also be complicated when $m(\mathbf{x}|\boldsymbol{\theta})$ is not smooth in $\boldsymbol{\theta}$. For example, in threshold regression, $m(\mathbf{x}|\boldsymbol{\theta})$ is discontinuous in θ_3 . The asymptotic distribution of $\widehat{\theta}_3$ depends on the magnitude of the threshold effect $\boldsymbol{\theta}_2 - \boldsymbol{\theta}_1$: when $\boldsymbol{\theta}_2 - \boldsymbol{\theta}_1$ shrinks to zero, the asymptotic distribution is related to a two-sided Brownian motion, while when $\boldsymbol{\theta}_2 - \boldsymbol{\theta}_1$ is fixed, the asymptotic distribution is related to a compound Poisson process; see Chan (1993), Hansen (2000) and Yu (2012) for further discussions. (**)

3 Omitted and Irrelevant Variables

Let the regressors be partitioned as

$$\mathbf{x}_i = \begin{pmatrix} \mathbf{x}_{1i} \\ \mathbf{x}_{2i} \end{pmatrix}.$$

Suppose we are interested in the coefficient on \mathbf{x}_{1i} alone in the regression of y_i on the full set \mathbf{x}_i . We can write the model as

$$\begin{aligned} y_i &= \mathbf{x}'_{1i}\boldsymbol{\beta}_1 + \mathbf{x}'_{2i}\boldsymbol{\beta}_2 + u_i, \\ E[\mathbf{x}_i u_i] &= \mathbf{0}, \end{aligned} \tag{3}$$

where the parameter of interest is $\boldsymbol{\beta}_1$.

Now suppose that instead of estimating equation (3) by least-squares, we regress y_i on \mathbf{x}_{1i} only. This is estimation of the equation

$$\begin{aligned} y_i &= \mathbf{x}'_{1i}\boldsymbol{\gamma}_1 + v_i, \\ E[\mathbf{x}_{1i} v_i] &= \mathbf{0}. \end{aligned} \tag{4}$$

Notice that we have written the coefficient on \mathbf{x}_{1i} as $\boldsymbol{\gamma}_1$ rather than $\boldsymbol{\beta}_1$ and the error as v_i rather than u_i . This is because the model being estimated is different from (3). Goldberger (1991) calls (3) the **long regression** and (4) the **short regression** to emphasize the distinction.

Typically, $\boldsymbol{\beta}_1 \neq \boldsymbol{\gamma}_1$, except in special cases. To see this, we calculate

$$\begin{aligned} \boldsymbol{\gamma}_1 &= E[\mathbf{x}_{1i}\mathbf{x}'_{1i}]^{-1} E[\mathbf{x}_{1i}y_i] \\ &= E[\mathbf{x}_{1i}\mathbf{x}'_{1i}]^{-1} E[\mathbf{x}_{1i}(\mathbf{x}'_{1i}\boldsymbol{\beta}_1 + \mathbf{x}'_{2i}\boldsymbol{\beta}_2 + u_i)] \\ &= \boldsymbol{\beta}_1 + E[\mathbf{x}_{1i}\mathbf{x}'_{1i}]^{-1} E[\mathbf{x}_{1i}\mathbf{x}'_{2i}]\boldsymbol{\beta}_2 \\ &= \boldsymbol{\beta}_1 + \boldsymbol{\Gamma}\boldsymbol{\beta}_2, \end{aligned}$$

where $\boldsymbol{\Gamma} = E[\mathbf{x}_{1i}\mathbf{x}'_{1i}]^{-1} E[\mathbf{x}_{1i}\mathbf{x}'_{2i}]$ is the coefficient from a regression of \mathbf{x}_{2i} on \mathbf{x}_{1i} .

Observe that $\boldsymbol{\gamma}_1 \neq \boldsymbol{\beta}_1$ unless $\boldsymbol{\Gamma} = \mathbf{0}$ or $\boldsymbol{\beta}_2 = \mathbf{0}$. Thus the short and long regressions have the same coefficient on \mathbf{x}_{1i} only under one of two conditions. First, the regression of \mathbf{x}_{2i} on \mathbf{x}_{1i} yields a set of zero coefficients (they are uncorrelated), or second, the coefficient on \mathbf{x}_{2i} in (3) is zero. In general, least squares estimation of (4) is an estimate of $\boldsymbol{\gamma}_1 = \boldsymbol{\beta}_1 + \boldsymbol{\Gamma}\boldsymbol{\beta}_2$ rather than $\boldsymbol{\beta}_1$. The difference $\boldsymbol{\Gamma}\boldsymbol{\beta}_2$ is known as **omitted variable bias**. It is the consequence of omitting a relevant correlated variable. Intuitively, $\boldsymbol{\gamma}_1$ includes both the direct effect of \mathbf{x}_1 on y ($\boldsymbol{\beta}_1$) and the indirect effect ($\boldsymbol{\Gamma}\boldsymbol{\beta}_2$) through \mathbf{x}_2 . Figure 3 illustrates these two effects.

To avoid omitted variable bias the standard advice is to include potentially relevant variables in the estimated model. By construction, the general model will be free of the omitted variables problem. Typically there are limits, as many desired variables are not available in a given dataset. In this case, the possibility of omitted variable bias should be acknowledged and discussed in the

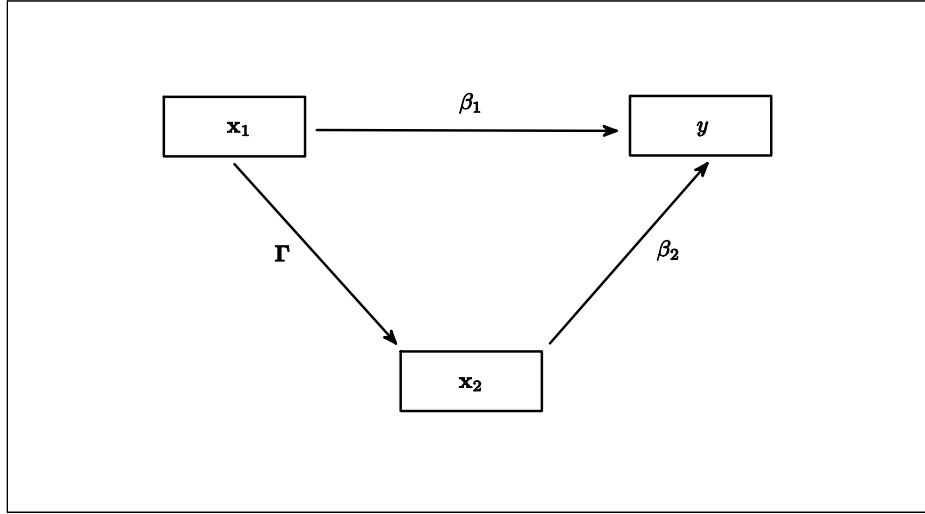


Figure 3: Direct and Indirect Effects of \mathbf{x}_1 on y

course of an empirical investigation.

When $\beta_2 = \mathbf{0}$ and β_1 is the parameter of interest, \mathbf{x}_{2i} is "irrelevant". In this case, the estimator of β_1 from the short regression, $\bar{\beta}_1 = (\mathbf{X}'_1 \mathbf{X}_1)^{-1} \mathbf{X}'_1 \mathbf{y}$, is consistent from the analysis above. So we compare its efficiency relative to the estimator from the long regression, $\hat{\beta}_1$. The comparison between the two estimators is straightforward when the error is conditionally homoskedastic $E[u_i^2 | \mathbf{x}_i] = \sigma^2$. In this case,

$$n \cdot AVar(\bar{\beta}_1) = E[\mathbf{x}_{1i} \mathbf{x}'_{1i}]^{-1} \sigma^2 \equiv \mathbf{Q}_{11}^{-1} \sigma^2,$$

and

$$n \cdot AVar(\hat{\beta}_1) = \mathbf{Q}_{11.2}^{-1} \sigma^2 \equiv (\mathbf{Q}_{11} - \mathbf{Q}_{12} \mathbf{Q}_{22}^{-1} \mathbf{Q}_{21})^{-1} \sigma^2$$

as discussed in Section 1.2 of the last chapter. If $\mathbf{Q}_{12} = E[\mathbf{x}_{1i} \mathbf{x}'_{2i}] = \mathbf{0}$ (so the variables are orthogonal) then these two variance matrices equal, and the two estimators have equal asymptotic efficiency. Otherwise, since $\mathbf{Q}_{12} \mathbf{Q}_{22}^{-1} \mathbf{Q}_{21} > 0$, $\mathbf{Q}_{11} > \mathbf{Q}_{11.2}$ and consequently

$$\mathbf{Q}_{11}^{-1} \sigma^2 < \mathbf{Q}_{11.2}^{-1} \sigma^2.$$

This means that $\bar{\beta}_1$ has a lower asymptotic variance matrix than $\hat{\beta}_1$. We conclude that inclusion of irrelevant variables reduces estimation efficiency if these variables are correlated with the relevant variables. Intuitively, the irrelevant variable does not provide information for y_i , but introduces multicollinearity to the system, so decreases the denominator of $AVar(\hat{\beta}_1)$ from \mathbf{Q}_{11} to $\mathbf{Q}_{11.2}$ without decreasing its numerator σ^2 . For example, take the model $y_i = \beta_0 + \beta_1 x_i + u_i$ and suppose

that $\beta_0 = 0$. Let $\widehat{\beta}_1$ be the estimate of β_1 from the unconstrained model, and $\overline{\beta}_1$ be the estimate under the constraint $\beta_0 = 0$ (the least squares estimate with the intercept omitted.). Let $E[x_i] = \mu$, and $Var(x_i) = \sigma_x^2$. Then we can show under homoskedasticity,

$$n \cdot AVar(\overline{\beta}_1) = \frac{\sigma^2}{\sigma_x^2 + \mu^2},$$

while

$$n \cdot AVar(\widehat{\beta}_1) = \frac{\sigma^2}{\sigma_x^2}.$$

When $\mu = E[1 \cdot x] \neq 0$, we see that $\overline{\beta}_1$ has a lower asymptotic variance.

On the contrary, if \mathbf{x}_2 is relevant ($\beta_2 \neq \mathbf{0}$) and uncorrelated with \mathbf{x}_1 , then it decreases the numerator without affecting the denominator of $AVar(\widehat{\beta}_1)$, so should be included in the regression. For example, including individual characteristics in a regression of beer consumption on beer prices leads to more precise estimates of the price elasticity because individual characteristics are believed to be uncorrelated with beer prices but affect beer consumption. The analysis above is summarized in Table 2.

	$\beta_2 = \mathbf{0}$	$\beta_2 \neq \mathbf{0}$
$\mathbf{Q}_{12} = \mathbf{0}$	$\overline{\beta}_1$ consistent same efficiency	$\overline{\beta}_1$ consistent long more efficient
$\mathbf{Q}_{12} \neq \mathbf{0}$	$\overline{\beta}_1$ consistent short more efficient	depends on $\mathbf{Q}_{12} \cdot \beta_2$ undetermined

Table 2: Consistency and Efficiency with Omitted and Irrelevant Variables

Exercise 3 If $\beta_2 \neq \mathbf{0}$, $\mathbf{Q}_{12} \neq \mathbf{0}$, could $\mathbf{Q}_{12}\beta_2$ be zero?

We have concentrated on the homoskedastic linear regression model. From Exercise 8 of the last chapter, we know that when the model is heteroskedastic, it is possible that $\widehat{\beta}_1$ is more efficient than $\overline{\beta}_1$ ($= \widehat{\beta}_{1R}$) even if \mathbf{x}_2 is irrelevant, or adding irrelevant variables can actually decrease the estimation variance. This result is strongly counter-intuitive. It seems to contradict our initial motivation for pursuing restricted estimation (or short regression) - to improve estimation efficiency. It turns out that a more refined answer is appropriate. Constrained estimation is desirable, but not the RLS estimation. While least squares is asymptotically efficient for estimation of the unconstrained projection model, it is not an efficient estimator of the constrained projection model; the efficient minimum distance estimator is the choice.

4 Model Selection

In the last section, we discussed the costs and benefits of inclusion/exclusion of variables. How does a researcher go about selecting an econometric specification, when economic theory does not provide complete guidance? This is the question of model selection. Model selection is an important

topic in linear regression analysis. In practice, a large number of variables usually are introduced at the initial stage of modeling to attenuate possible modeling biases. On the other hand, to enhance predictability and to select significant variables, econometricians usually use stepwise deletion and subset selection. See Miller (2002) for a comprehensive treatment of subset selection in regression.

It is important that the model selection question be well-posed. For example, the question: "What is the right model for y ?" is not well-posed, because it does not make clear the conditioning set. In contrast, the question, "Which subset of (x_1, \dots, x_K) enters the regression function $E[y_i | x_{1i} = x_1, \dots, x_{Ki} = x_K]$?" is well posed.

4.1 Selection Among Nested Models

In many cases the problem of model selection can be reduced to the comparison of two nested models, as the larger problem can be written as a sequence of such comparisons. We thus consider the question of the inclusion of \mathbf{X}_2 in the linear regression

$$\mathbf{y} = \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2 + \mathbf{u},$$

where \mathbf{X}_1 is $n \times k_1$ and \mathbf{X}_2 is $n \times k_2$. This is equivalent to the comparison of the two models

$$\begin{aligned} \mathcal{M}_1 : \mathbf{y} &= \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{u}, & E[\mathbf{u} | \mathbf{X}_1, \mathbf{X}_2] &= \mathbf{0}, \\ \mathcal{M}_2 : \mathbf{y} &= \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2 + \mathbf{u}, & E[\mathbf{u} | \mathbf{X}_1, \mathbf{X}_2] &= \mathbf{0}. \end{aligned}$$

Note that $\mathcal{M}_1 \subset \mathcal{M}_2$. To be concrete, we say that \mathcal{M}_2 is true if $\boldsymbol{\beta}_2 \neq \mathbf{0}$. To fix notation, models 1 and 2 are estimated by OLS, with residual vectors $\hat{\mathbf{u}}_1$ and $\hat{\mathbf{u}}_2$, estimated variances $\hat{\sigma}_1^2$ and $\hat{\sigma}_2^2$, etc., respectively. To simplify some of the statistical discussion, we will on occasion use the homoskedasticity assumption $E[u_i^2 | \mathbf{x}_{1i}, \mathbf{x}_{2i}] = \sigma^2$.

A model selection procedure is a data-dependent rule which selects one of the two models. We can write this as $\widehat{\mathcal{M}}$. There are many possible desirable properties for a model selection procedure. One useful property is consistency, i.e., it selects the true model with probability one if the sample is sufficiently large. Formally, a model selection procedure is **consistent** if

$$\begin{aligned} P\left(\widehat{\mathcal{M}} = \mathcal{M}_1 | \mathcal{M}_1\right) &\rightarrow 1, \\ P\left(\widehat{\mathcal{M}} = \mathcal{M}_2 | \mathcal{M}_2\right) &\rightarrow 1. \end{aligned}$$

However, this rule only makes sense when the true model is finite dimensional. If the truth is infinite dimensional, it is more appropriate to view model selection as determining the best finite sample approximation.

A common approach to model selection is to use a statistical test such as the Wald W_n . The model selection rule is as follows. For some significance level α , let c_α satisfy $P(\chi_{k_2}^2 > c_\alpha) = \alpha$. Then select \mathcal{M}_1 if $W_n \leq c_\alpha$, else select \mathcal{M}_2 . A major problem with this approach is that the significance level is indeterminate. The reasoning which helps guide the choice of α in hypothesis

testing (controlling Type I error) is not relevant for model selection. That is, if α is set to be a small number, then $P(\widehat{\mathcal{M}} = \mathcal{M}_1 | \mathcal{M}_1) \approx 1 - \alpha$ but $P(\widehat{\mathcal{M}} = \mathcal{M}_2 | \mathcal{M}_2)$ could vary dramatically, depending on the sample size, etc. Another problem is that if α is held fixed, then this model selection procedure is inconsistent, as $P(\widehat{\mathcal{M}} = \mathcal{M}_1 | \mathcal{M}_1) \rightarrow 1 - \alpha < 1$ although $P(\widehat{\mathcal{M}} = \mathcal{M}_2 | \mathcal{M}_2) \rightarrow 1$.

4.1.1 Information Criteria

Another common approach to model selection is to use an information criterion. The essential intuition is that there exists a tension between model fit, as measured by the maximized log-likelihood value, and the principle of parsimony that favors a simple model. The fit of the model can be improved by increasing model complexity, but parameters are only added if the resulting improvement in fit sufficiently compensates for loss of parsimony. One popular choice is the Akaike Information Criterion (AIC) proposed in Akaike (1973). From the Technical Appendix A, the AIC under normality for model m is

$$AIC_m = \log(\widehat{\sigma}_m^2) + 2\frac{k_m}{n}, \quad (5)$$

where $\widehat{\sigma}_m^2$ is the variance estimate for model m and is roughly $-2\ell_n$ (neglecting the constant term) with ℓ_n being defined in the Introduction, and k_m is the number of coefficients in the model. The rule is to select \mathcal{M}_1 if $AIC_1 < AIC_2$, else select \mathcal{M}_2 . AIC selection is inconsistent, as the rule tends to overfit as observed in Shibata (1976). Indeed, since under \mathcal{M}_1 , from Section 5 of Chapter 4,

$$LR = n(\log(\widehat{\sigma}_1^2) - \log(\widehat{\sigma}_2^2)) \xrightarrow{d} \chi_{k_2}^2, \quad (6)$$

then

$$\begin{aligned} P(\widehat{\mathcal{M}} = \mathcal{M}_1 | \mathcal{M}_1) &= P(AIC_1 < AIC_2 | \mathcal{M}_1) \\ &= P\left(\log(\widehat{\sigma}_1^2) + 2\frac{k_1}{n} < \log(\widehat{\sigma}_2^2) + 2\frac{k_1 + k_2}{n} \mid \mathcal{M}_1\right) \\ &= P(LR < 2k_2 | \mathcal{M}_1) \rightarrow P(\chi_{k_2}^2 < 2k_2) < 1. \end{aligned}$$

Although the AIC tends to overfit, this need not be a defect of the AIC. This is because the AIC is derived as an estimate of the Kullback-Leibler information distance $KLIC(\mathcal{M}) = E[\log f(\mathbf{y}|\mathbf{X}) - \log f(\mathbf{y}|\mathbf{X}, \mathcal{M})]$ between the true density and the model density.⁵ In other words, the AIC attempts to select a good approximating model for inference. In contrast, other information criteria as mentioned below attempt to estimate the "true" model. Figure 4 intuitively shows the difference between the AIC and other information criteria.

Many other criteria similar to the AIC have been proposed, the most popular is the Bayes Information Criterion (BIC) introduced by Schwarz (1978). The BIC is based on approximating

⁵This is also the term "information criterion" originated.

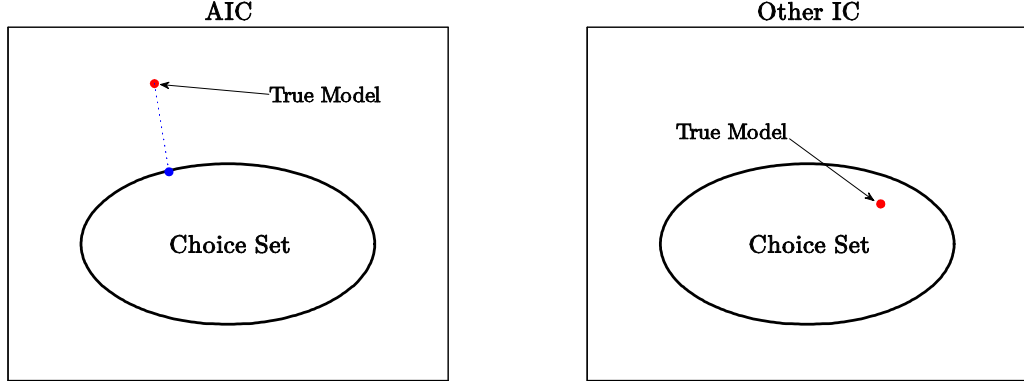


Figure 4: Difference Between AIC and Other IC

the **Bayes factor**; from the Technical Appendix B, it turns out to be

$$BIC_m = \log(\hat{\sigma}_m^2) + \log n \frac{k_m}{n}. \quad (7)$$

Since $\log(n) > 2$ (if $n > 8$), the BIC places a larger penalty than the AIC on the number of estimated parameters and is more parsimonious. Another criterion, which is often cited but seems to have seen little use in practice, is the HQIC by Hannan and Quinn (1979). This criterion replaces $\log n$ in BIC_m by $Q \log \log n$ for some $Q > 2$. This criterion imposes a penalty larger than the AIC and smaller than the BIC.

In contrast to the AIC, the BIC and HQIC model selection procedures are consistent. Take the BIC as an example since the argument for the HQIC is similar. Because (6) holds under \mathcal{M}_1 ,

$$\frac{LR}{\log(n)} \xrightarrow{p} 0,$$

so

$$\begin{aligned} P(\widehat{\mathcal{M}} = \mathcal{M}_1 | \mathcal{M}_1) &= P(BIC_1 < BIC_2 | \mathcal{M}_1) = P(LR < \log(n) k_2 | \mathcal{M}_1) \\ &= P\left(\frac{LR}{\log(n)} < k_2 | \mathcal{M}_1\right) \rightarrow P(0 < k_2) = 1. \end{aligned}$$

Also under \mathcal{M}_2 , one can show that

$$\frac{LR}{\log(n)} \xrightarrow{p} \infty,$$

thus

$$P(\widehat{\mathcal{M}} = \mathcal{M}_2 | \mathcal{M}_2) = P\left(\frac{LR_n}{\log(n)} > k_2 \middle| \mathcal{M}_2\right) \rightarrow 1.$$

Essentially, to consistently select \mathcal{M}_1 , we must let the significance level of the LR test approach zero to asymptotically avoid choosing a model that is too large, so the critical value (or the penalty)

must diverge to infinity. On the other hand, to consistently select \mathcal{M}_2 , the penalty must be $o(n)$ so that LR divided by the penalty converges in probability to infinity under \mathcal{M}_2 . Compared with the fixed penalty scheme such as the AIC, the consistent selection procedures sacrifice some power to exchange for an asymptotically zero type I error.

We have discussed model selection between two models. The methods extend readily to the issue of selection among multiple regressors. The general problem is the model

$$y_i = \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_K x_{Ki} + u_i, E[u_i | \mathbf{x}_i] = 0$$

and the question is which subset of the coefficients are non-zero (equivalently, which regressors enter the regression). There are two leading cases: ordered regressors and unordered regressors. In the ordered case, the models are

$$\begin{aligned} \mathcal{M}_1 & : \beta_1 \neq 0, \beta_2 = \beta_3 = \cdots = \beta_K = 0, \\ \mathcal{M}_2 & : \beta_1 \neq 0, \beta_2 \neq 0, \beta_3 = \cdots = \beta_K = 0, \\ & \vdots \\ \mathcal{M}_K & : \beta_1 \neq 0, \beta_2 \neq 0, \cdots, \beta_K \neq 0, \end{aligned}$$

which are nested. The AIC estimates the K models by OLS, stores the residual variance $\hat{\sigma}^2$ for each model, and then selects the model with the lowest AIC (5). Similarly, the BIC selects based on (7). In the unordered case, a model consists of any possible subset of the regressors $\{x_{1i}, \cdots, x_{Ki}\}$, and the AIC or BIC in principle can be implemented by estimating all possible subset models. However, there are 2^K such models, which can be a very large number. For example, $2^{10} = 1024$, and $2^{20} = 1,048,576$. So in the unordered case, a full-blown implementation of these information criteria would seem computationally prohibitive.

Exercise 4 *In the ordered regressors case, compare the difference between the AIC, the BIC and the LR test in the conclusion of model selection.*

4.1.2 Limitation and Extension of Information Criteria

Given their simplicity, penalized likelihood criteria are often used for selecting "the best model". However, there is no clear answer as to which criterion, if any, should be preferred. Considerable approximation is involved in deriving the formulas for the AIC and related measures and other criteria than the AIC and BIC might be more appropriate. From a decision-theoretic viewpoint, the choice of the model from a set of models should depend on the intended use of the model. For example, the purpose of the model may be to summarize the main features of a complex reality, or to predict some outcome, or to test some important hypothesis. In applied work it is quite rare to see an explicit statement of the intended use of an econometric model. Recently, Claeskens and Hjort (2003) propose the focused information criterion (FIC) that focuses on the parameter singled out for interest. This criterion seems close to the target-based principle mentioned above.

(**) Another serious drawback of the information criteria (and many other subset variable selection procedures) is their lack of stability as analyzed in Breiman (1996). That is, a small change of data may cause large changes in selected variables. To avoid such a problem, recent statistical literature on model selection proposes penalty-function-based criteria where penalty functions continuously shrink the coefficients rather than discretely select the variables. Different penalty functions induce different selection criteria; outstanding examples include the least absolute shrinkage and selection operator (LASSO) of Tibshirani (1996), L_q -penalty of Frank and Friedman (1993), the smoothly clipped absolute deviation (SCAD) penalty function of Fan and Li (2001), adaptive LASSO of Zou (2006) and the minimax concave penalty (MCP) of Zhang (2010) among others. Another strand of literature uses so-called model averaging to alleviate model misspecification on estimation. The idea of model averaging was first put forward in Hjort and Claeskens (2003) and applied in least squares estimation by Hansen (2007b).

Finally, because k and its estimate \hat{k} are integers, all the established asymptotic theory when k is known applies also when \hat{k} can consistently estimate k . However, see Leeb and Pötscher (2005) for cautions on this result. (**)

4.2 Tests against Nonnested Alternatives (*)

In the model selection procedures above, the model to be selected can be nonnested, i.e., the null is not a special case of the alternative, e.g., $y = \mathbf{x}'\boldsymbol{\beta} + u$ vs $y = \mathbf{z}'\boldsymbol{\gamma} + v$ with \mathbf{x} and \mathbf{z} not covering each other completely. Nevertheless, the dependent variable should be the same to define $\hat{\sigma}_m^2$ in (5) or (7) in the same scale for different m 's. In some cases, nonnested models have different dependent variables, e.g., $y = \mathbf{x}'\boldsymbol{\beta} + u$ vs $\log y = \mathbf{x}'\boldsymbol{\gamma} + v$. How to choose among such nonnested models is challenging.

Tests against nonnested alternatives date back at least to Cox (1961, 1962) and Atkinson (1969, 1970). Breusch and Pagan (1980) interpret Cox's LR test as a LM test. Cox's basic ideas were adapted to linear regression models by Pesaran (1974) and to nonlinear regression models by Pesaran and Deaton (1978). Vuong (1989) provides a very general distribution theory for the LR test statistic that covers both nested and nonnested models and more remarkably permits the DGP to be an unknown density that differs from both the densities under the null and alternative. However, all these methods are restricted to fully parametric models. Tests in the nonlikelihood case often take one of two approaches. **Artificial nesting**, proposed in Davidson and MacKinnon (1981, 1984), embeds the two nonnested models into a more general artificial model, which leads to so-called J tests and P tests and related tests. The **encompassing principle**, proposed by Mizon and Richard (1986), leads to a quite general framework for testing one model against a competing nonnested model. White (1994) links this approach with CM tests. We will present Cox's LR test and Breusch and Pagan's LM interpretation for a taste. See Gourieroux and Monfort (1994, 1995 Ch. 22) and Pesaran and Weeks (2001) for a summary of literature.

4.2.1 Cox's LR Test

Consider choosing between two parametric models. Let model $F_{\boldsymbol{\theta}}$ have density $f(y|\mathbf{x}, \boldsymbol{\theta})$ and model $G_{\boldsymbol{\gamma}}$ have density $g(y|\mathbf{x}, \boldsymbol{\gamma})$. A likelihood ratio test of the model $F_{\boldsymbol{\theta}}$ against $G_{\boldsymbol{\gamma}}$ is based on

$$LR = 2 \left(\mathcal{L}_f(\hat{\boldsymbol{\theta}}) - \mathcal{L}_g(\hat{\boldsymbol{\gamma}}) \right) = 2 \sum_{i=1}^n \ln \frac{f(y_i|\mathbf{x}_i, \hat{\boldsymbol{\theta}})}{g(y_i|\mathbf{x}_i, \hat{\boldsymbol{\gamma}})},$$

where \mathcal{L}_f and \mathcal{L}_g are log likelihood functions under $F_{\boldsymbol{\theta}}$ and $G_{\boldsymbol{\gamma}}$, and $\hat{\boldsymbol{\theta}}$ and $\hat{\boldsymbol{\gamma}}$ are the corresponding MLEs. If $G_{\boldsymbol{\gamma}}$ is nested in $F_{\boldsymbol{\theta}}$ then LR is chi-square distributed under the null that $F_{\boldsymbol{\theta}} = G_{\boldsymbol{\gamma}}$. However, this result no longer holds if the models are nonnested.

Cox proposed solving this problem in the special case that $F_{\boldsymbol{\theta}}$ is the true model but the models are not nested, by applying a CLT under the assumption that $F_{\boldsymbol{\theta}}$ is the true model. To understand Cox's statistic, we will use Breusch and Pagan's LM interpretation. The connection between the LM and Cox's statistic was mentioned in Atkinson (1970, pp332-335) and discussed in more details in Breusch and Pagan (1980). Let $L_0(\boldsymbol{\theta})$ and $L_1(\boldsymbol{\gamma})$ be the likelihoods under H_0 and H_1 . To make a choice between H_0 and H_1 , one procedure would be to form a joint likelihood of the form $L = \left(\int L_0^\lambda L_1^{1-\lambda} \right)^{-1} L_0^\lambda L_1^{1-\lambda}$ and test if $\lambda = 0$ or 1 , where the factor $\int L_0^\lambda L_1^{1-\lambda}$ makes this a proper pdf. The log likelihood is then $\mathcal{L} = \lambda \mathcal{L}_0 + (1 - \lambda) \mathcal{L}_1 - \log \int L_0^\lambda L_1^{1-\lambda}$, where $\mathcal{L}_j = \log L_j$, and the LM test would be based upon $\partial \mathcal{L} / \partial \lambda$. It is not hard to show that

$$\frac{\partial \mathcal{L}}{\partial \lambda} = \mathcal{L}_0 - \mathcal{L}_1 - E[\mathcal{L}_0 - \mathcal{L}_1] = \mathcal{L}_{01} - E[\mathcal{L}_{01}],$$

where $E[\cdot]$ is the expectation with respect to the parameters $\boldsymbol{\theta}$ and $\boldsymbol{\gamma}$, and $\mathcal{L}_{01} = \mathcal{L}_0 - \mathcal{L}_1$. The LM test for $\lambda = 1$ is based upon the score $\mathcal{L}_0 - \mathcal{L}_1$. To do so joint distribution of the scores under H_0 is required and it is instructive to observed that

$$\frac{\partial \mathcal{L}}{\partial \boldsymbol{\gamma}} = (1 - \lambda) \frac{\partial \mathcal{L}_1}{\partial \boldsymbol{\gamma}} - \left(\int L_0^\lambda L_1^{1-\lambda} \right)^{-1} \int L_0^\lambda (1 - \lambda) \frac{\partial L_1}{\partial \boldsymbol{\gamma}} = \mathbf{0}$$

when $\lambda = 1$. Therefore the relevant parts of the information matrix are (when evaluated under H_0)

$$\begin{aligned} J_{\lambda\lambda} &= -E_0 \left[\frac{\partial^2 \mathcal{L}}{\partial \lambda \partial \lambda'} \right] = E_0 \left[(\mathcal{L}_{01} - E[\mathcal{L}_{01}])^2 \right] = \text{Var}_0(\mathcal{L}_{01}) \equiv V_0(\mathcal{L}_{01}), \\ J_{\lambda\boldsymbol{\theta}} &= -E_0 \left[\frac{\partial^2 \mathcal{L}}{\partial \lambda \partial \boldsymbol{\theta}'} \right] = -E_0 \left[\frac{\partial \mathcal{L}_{01}}{\partial \boldsymbol{\theta}'} \right] = \boldsymbol{\eta}', \\ J_{\boldsymbol{\theta}\boldsymbol{\theta}} &= -E_0 \left[\frac{\partial^2 \mathcal{L}}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} \right] = -E_0 \left[\frac{\partial^2 \mathcal{L}_0}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} \right] = \mathbf{Q}, \end{aligned}$$

which follows by noting that $\partial \mathcal{L}_1 / \partial \boldsymbol{\theta} = 0$, $\lambda = 1$ and $E[\partial \mathcal{L}_0 / \partial \boldsymbol{\theta}] = 0$, where $E_0[\cdot]$ is the expectation with respect to the randomness in the data when they follow the distribution under H_0 . Thus the

LM statistic becomes

$$\left(\widehat{\mathcal{L}}_{01} - E_0 \left[\widehat{\mathcal{L}}_{01}\right]\right)' \left[V_0(\mathcal{L}_{01}) - \boldsymbol{\eta}'\mathbf{Q}^{-1}\boldsymbol{\eta}\right]^{-1} \left(\widehat{\mathcal{L}}_{01} - E_0 \left[\widehat{\mathcal{L}}_{01}\right]\right),$$

which is Cox's test in the type of notation in Pesaran and Deaton (1978, p681), where $\widehat{\mathcal{L}}_{01} = \mathcal{L}_0(\widehat{\boldsymbol{\theta}}) - \mathcal{L}_1(\widehat{\boldsymbol{\gamma}}) \equiv \widehat{\mathcal{L}}_0 - \widehat{\mathcal{L}}_1$ is exactly $\mathcal{L}_f(\widehat{\boldsymbol{\theta}}) - \mathcal{L}_g(\widehat{\boldsymbol{\gamma}})$ in the LR statistic of Cox. Cox (1961) showed that $\widehat{\mathcal{L}}_{01} - E_0 \left[\widehat{\mathcal{L}}_{01}\right]$ is asymptotically normally distributed with mean zero and variance $V_0(\mathcal{L}_{10}) - \boldsymbol{\eta}'\mathbf{Q}^{-1}\boldsymbol{\eta}$.

This approach is computationally awkward to implement if one cannot analytically obtain $E_0 \left[\widehat{\mathcal{L}}_{01}\right]$. Note that $\widehat{\mathcal{L}}_{01} = \widehat{\mathcal{L}}_0 - \widehat{\mathcal{L}}_1$ and $E_0 \left[\widehat{\mathcal{L}}_{01}\right] = \widehat{\mathcal{L}}_0 - E_0 \left[\widehat{\mathcal{L}}_1\right]$ so that $\widehat{\mathcal{L}}_{01} - E_0 \left[\widehat{\mathcal{L}}_{01}\right] = -\left(\widehat{\mathcal{L}}_1 - E_0 \left[\widehat{\mathcal{L}}_1\right]\right)$, i.e., the test is based on comparing the observed value of $\widehat{\mathcal{L}}_1$ with its expected value if H_0 were true. But \mathcal{L}_1 is a function of $\boldsymbol{\gamma}$ and $\boldsymbol{\gamma}$ does not appear under H_0 so that it is not entirely clear how to evaluate $\widehat{\mathcal{L}}_1$. In fact two methods have been adopted. Cox forms $\widehat{\mathcal{L}}_1 - E_0 \left[\widehat{\mathcal{L}}_1(\widehat{\boldsymbol{\gamma}}_0)\right]$ where $\widehat{\boldsymbol{\gamma}}_0$ is the plim of $\widehat{\boldsymbol{\gamma}}$ under H_0 , whereas Atkinson forms $\mathcal{L}_1(\widehat{\boldsymbol{\gamma}}_0) - E_0 \left[\widehat{\mathcal{L}}_1(\widehat{\boldsymbol{\gamma}}_0)\right]$. It would seem logical to adopt the latter from an LM viewpoint as the score is then unbiased, but de Pereira (1977) claims that the resulting test statistic is inconsistent, whereas Cox's is consistent.

Another problem of Cox's test is that if a similar test statistic is obtained with the roles of $F_{\boldsymbol{\theta}}$ and $G_{\boldsymbol{\gamma}}$ reversed it is possible to find both that model $F_{\boldsymbol{\theta}}$ is rejected in favor of $G_{\boldsymbol{\gamma}}$ and that model $G_{\boldsymbol{\gamma}}$ is rejected in favor of $F_{\boldsymbol{\theta}}$. The test is therefore not necessarily for model selection as it does not necessarily select one or the other; instead it is a model specification test that zero, one, or two of the models can pass.

The Cox statistic has been obtained analytically in some cases. For nonnested linear regression models $y = \mathbf{x}'\boldsymbol{\beta} + u$ and $y = \mathbf{z}'\boldsymbol{\gamma} + v$ with homoskedastic normally distributed errors, see Pesaran (1974). For nonnested transformation models $h(y) = \mathbf{x}'\boldsymbol{\beta} + u$ and $g(y) = \mathbf{z}'\boldsymbol{\gamma} + v$, where $h(y)$ and $g(y)$ are known transformations, see Pesaran and Pesaran (1993), who use a simulation-based approach. This permits, for example, discrimination between linear and log-linear parametric models, with $h(\cdot)$ the identity transformation and $g(\cdot)$ the log transformation. Pesaran and Pesaran (1993) apply the idea to choose between logit and probit models.

4.2.2 $\log(y)$ versus y as Dependent Variable

We briefly discuss choice between $\log(y)$ versus y as the dependent variable. There is a large literature on this subject, much of it quite misleading. The plain truth is that either regression is "okay", in the sense that both $E[y_i|\mathbf{x}_i]$ and $E[\log(y_i)|\mathbf{x}_i]$ are well-defined (so long as $y_i > 0$). It is perfectly valid to estimate either or both regressions. They are *different* regression functions, neither is more nor less valid than the other. To test one specification versus the other, or select one specification over the other, requires the imposition of additional structure, such as the assumptions that the conditional expectation is linear in \mathbf{x}_i , and $u_i \sim N(0, \sigma^2)$.⁶

⁶For example, under such assumptions, Amemiya (1980) shows that R^2 and \bar{R}^2 based on $\log(y)$ are larger than those based on y .

There still may be good reasons for preferring the $\log(y)$ regression over the y regression. First, it may be the case that $E[\log(y_i)|\mathbf{x}_i]$ is roughly linear in \mathbf{x}_i over the support of \mathbf{x}_i , while the regression $E[y_i|\mathbf{x}_i]$ is non-linear, and linear models are easier to report and interpret. Second, it may be the case that the errors in $u_i = \log(y_i) - E[\log(y_i)|\mathbf{x}_i]$ may be less heteroskedastic than the errors from the linear specification (although the reverse may be true!). Finally, and probably most importantly, if the distribution of y_i is highly skewed (as the wage example in the Introduction), the conditional mean $E[y_i|\mathbf{x}_i]$ may not be a useful measure of central tendency, and estimates will be undesirably influenced by extreme observations ("outliers"). In this case, the conditional mean-log $E[\log(y_i)|\mathbf{x}_i]$ may be a better measure of central tendency, and hence more interesting to estimate and report. In the classical return-to-schooling example, it is commonly believed that the wage rate follows the log-normal distribution which is highly skewed, while the log-wage follows the normal distribution which is symmetric.

Log transformation is often used for a percentage interpretation of the coefficients. This is because $\log(y+\Delta) - \log(y) \approx \Delta/y$ which is the percentage change of y . When the log transformation is also conducted to \mathbf{x} , the coefficients are elasticities of y with respect to \mathbf{x} . Finally, note that variables measured in units such as years or in percentage points should not be logged.

5 Generalized Least Squares

In the linear projection model, we know that the least squares estimator is semi-parametrically efficient for the projection coefficient. However, in the linear regression model

$$\begin{aligned} y_i &= \mathbf{x}'_i \boldsymbol{\beta} + u_i, \\ E[u_i|\mathbf{x}_i] &= 0, \end{aligned}$$

the least squares estimator may not be efficient. The theory of Chamberlain (1987) can be used to show that in this model the semiparametric efficiency bound is obtained by the weighted least squares (WLS) estimator

$$\bar{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{D}^{-1}\mathbf{X})^{-1} (\mathbf{X}'\mathbf{D}^{-1}\mathbf{y}), \quad (8)$$

where $\mathbf{D} = \text{diag}\{\sigma_1^2, \dots, \sigma_n^2\}$ and $\sigma_i^2 = \sigma^2(\mathbf{x}_i) = E[u_i^2|\mathbf{x}_i]$. We provide some intuition for this result. Note that

$$\bar{\boldsymbol{\beta}} = \arg \min_{\boldsymbol{\beta}} \sum_{i=1}^n (y_i - \mathbf{x}'_i \boldsymbol{\beta})^2 \frac{1}{\sigma_i^2} = \arg \min_{\boldsymbol{\beta}} \sum_{i=1}^n \left(\frac{y_i}{\sigma_i} - \frac{\mathbf{x}'_i}{\sigma_i} \boldsymbol{\beta} \right)^2,$$

where the objective function takes the form of weighted sum of squared residuals, and the model

$$\frac{y_i}{\sigma_i} = \frac{\mathbf{x}'_i}{\sigma_i} \boldsymbol{\beta} + \frac{u_i}{\sigma_i}$$

is homoskedastic (why?). Under homoskedasticity, the Gauss-Markov theorem implies the efficiency of the LSE which is the WLS estimator in the original model. An interesting aspect of this efficiency result is that only the second moment of u_i is relevant and higher moments are not. Of course, this is because we are comparing the asymptotic variance which involves only the second moment of u_i .

Exercise 5 Consider the WLS estimator (8) with $\mathbf{D} = \text{diag}\{x_{j1}^2, \dots, x_{jn}^2\}$, where x_{ji} is one of \mathbf{x}_i . (i) Is this estimator unbiased and consistent? (ii) Using your intuition, in which situations would you expect that this estimator would perform better than OLS?

The GLS estimator (8) is infeasible since the matrix \mathbf{D} is unknown. A feasible GLS (FGLS) estimator replaces the unknown \mathbf{D} with an estimate $\widehat{\mathbf{D}} = \text{diag}\{\widehat{\sigma}_1^2, \dots, \widehat{\sigma}_n^2\}$. We now discuss this estimation problem. As in Goldfeld and Quandt (1972, Ch.3), we model the conditional variance using the parametric form

$$\sigma_i^2 = \alpha_0 + \mathbf{z}'_{1i}\boldsymbol{\alpha}_1 = \boldsymbol{\alpha}'\mathbf{z}_i,$$

where \mathbf{z}_{1i} is some $q \times 1$ function of \mathbf{x}_i . Typically, \mathbf{z}_{1i} are squares (and perhaps levels) of some (or all) elements of \mathbf{x}_i . Often the functional form is kept simple for parsimony. Let $\eta_i = u_i^2$. Then

$$E[\eta_i|\mathbf{x}_i] = \alpha_0 + \mathbf{z}'_{1i}\boldsymbol{\alpha}_1$$

and we have the regression equation

$$\begin{aligned} \eta_i &= \alpha_0 + \mathbf{z}'_{1i}\boldsymbol{\alpha}_1 + \xi_i, \\ E[\xi_i|\mathbf{x}_i] &= 0. \end{aligned} \tag{9}$$

This regression error ξ_i is generally heteroskedastic and has the conditional variance

$$\text{Var}(\xi_i|\mathbf{x}_i) = \text{Var}(u_i^2|\mathbf{x}_i) = E[u_i^4|\mathbf{x}_i] - (E[u_i^2|\mathbf{x}_i])^2.$$

Suppose u_i (and thus η_i) were observed. Then we could estimate $\boldsymbol{\alpha}$ by OLS:

$$\bar{\boldsymbol{\alpha}} = (\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\boldsymbol{\eta}$$

and

$$\sqrt{n}(\bar{\boldsymbol{\alpha}} - \boldsymbol{\alpha}) \xrightarrow{d} N(\mathbf{0}, \mathbf{V}_\alpha),$$

where

$$\mathbf{V}_\alpha = (E[\mathbf{z}_i\mathbf{z}'_i])^{-1} (E[\mathbf{z}_i\mathbf{z}'_i\xi_i^2]) (E[\mathbf{z}_i\mathbf{z}'_i])^{-1}. \tag{10}$$

Exercise 6 Take the model

$$\begin{aligned} y_i &= \mathbf{x}'_i\boldsymbol{\beta} + u_i, E[\mathbf{x}_i u_i] = \mathbf{0}, \\ u_i^2 &= \mathbf{z}'_i\boldsymbol{\gamma} + \xi_i, E[\mathbf{z}_i \xi_i] = \mathbf{0}. \end{aligned}$$

Find the MoM estimators $(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}})$ for $(\boldsymbol{\beta}, \boldsymbol{\gamma})$.

While u_i is not observed, we have the OLS residual $\hat{u}_i = y_i - \mathbf{x}_i' \hat{\boldsymbol{\beta}} = u_i - \mathbf{x}_i'(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})$. Thus

$$\phi_i = \hat{\eta}_i - \eta_i = \hat{u}_i^2 - u_i^2 = -2u_i \mathbf{x}_i'(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) + (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})' \mathbf{x}_i \mathbf{x}_i' (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}).$$

And then

$$\frac{1}{\sqrt{n}} \sum_{i=1}^n \mathbf{z}_i \phi_i = -\frac{2}{n} \sum_{i=1}^n \mathbf{z}_i u_i \mathbf{x}_i' \sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) + \frac{1}{n} \sum_{i=1}^n \mathbf{z}_i (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})' \mathbf{x}_i \mathbf{x}_i' \sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \xrightarrow{p} \mathbf{0}.$$

Let

$$\tilde{\boldsymbol{\alpha}} = (\mathbf{Z}'\mathbf{Z})^{-1} \mathbf{Z}'\hat{\boldsymbol{\eta}} \quad (11)$$

be from OLS regression of $\hat{\eta}_i$ on \mathbf{z}_i . Then

$$\sqrt{n}(\tilde{\boldsymbol{\alpha}} - \boldsymbol{\alpha}) = \sqrt{n}(\bar{\boldsymbol{\alpha}} - \boldsymbol{\alpha}) + (n^{-1}\mathbf{Z}'\mathbf{Z})^{-1} n^{-1/2}\mathbf{Z}'\boldsymbol{\phi} \xrightarrow{d} N(\mathbf{0}, \mathbf{V}_{\boldsymbol{\alpha}}). \quad (12)$$

Thus the fact that η_i is replaced with $\hat{\eta}_i$ is asymptotically irrelevant.⁷ We call (11) the **skedastic regression**, as it is estimating the conditional variance of the regression of y_i on \mathbf{x}_i . We have shown that $\boldsymbol{\alpha}$ is consistently estimated by a simple procedure, and hence we can estimate $\sigma_i^2 = \mathbf{z}_i' \boldsymbol{\alpha}$ by

$$\tilde{\sigma}_i^2 = \tilde{\boldsymbol{\alpha}}' \mathbf{z}_i. \quad (13)$$

Suppose that $\tilde{\sigma}_i^2 > 0$ for all i . Then set

$$\tilde{\mathbf{D}} = \text{diag} \{ \tilde{\sigma}_1^2, \dots, \tilde{\sigma}_n^2 \}$$

and

$$\tilde{\boldsymbol{\beta}} = (\mathbf{X}'\tilde{\mathbf{D}}^{-1}\mathbf{X})^{-1} (\mathbf{X}'\tilde{\mathbf{D}}^{-1}\mathbf{y}).$$

This is the feasible GLS, or FGLS, estimator of $\boldsymbol{\beta}$. In practice, we can iterate between $\tilde{\mathbf{D}}$ and $\tilde{\boldsymbol{\beta}}$ until convergence. Specifically, we can start from a $\boldsymbol{\beta}$ estimate to get residuals which are used to estimate \mathbf{D} , and this \mathbf{D} estimate is plugged in the FGLS formula to get a new estimate of $\boldsymbol{\beta}$. Repeat this process until convergence. Since there is not a unique specification for the conditional variance the FGLS estimator is not unique, and will depend on the model (and estimation method) for the skedastic regression. Robinson (1987) shows that even if $\sigma^2(\cdot)$ is nonparametrically specified, $\boldsymbol{\beta}$ can be estimated as if $\sigma^2(\cdot)$ were known. In other words, the nonparametric estimation of $\sigma^2(\cdot)$ does not affect the efficiency of the FGLS estimator or the FGLS estimator can achieve the same efficiency as $\bar{\boldsymbol{\beta}}$.

One typical problem with implementation of FGLS estimation is that in a linear regression specification, there is no guarantee that $\tilde{\sigma}_i^2 > 0$ for all i . If $\tilde{\sigma}_i^2 < 0$ for some i , then the FGLS

⁷Such a result appears as early as in Amemiya (1977).

estimator is not well defined. Furthermore, if $\tilde{\sigma}_i^2 \approx 0$ for some i , then the FGLS estimator will force the regression equation to pass through the point (y_i, \mathbf{x}_i) , which is typically undesirable. This suggests that there is a need to bound the estimated variances away from zero. A trimming rule might make sense:

$$\bar{\sigma}_i^2 = \max \{ \tilde{\sigma}_i^2, \underline{\sigma}^2 \}$$

for some $\underline{\sigma}^2 > 0$. Of course, we can assume $\sigma_i^2 = h(\boldsymbol{\alpha}'\mathbf{z}_i) > 0$ at the beginning, e.g., $h(\cdot) = \exp\{\cdot\}$ or $h(\cdot) = |\cdot|^m$ with m a prespecified integer. If h is invertible such as $\exp\{\cdot\}$, we can regress $h^{-1}(\hat{u}_i^2)$ on \mathbf{z}_i to get $\tilde{\boldsymbol{\alpha}}$ and then estimate σ_i^2 by $h(\tilde{\boldsymbol{\alpha}}'\mathbf{z}_i)$. If h is not invertible such as $|\cdot|^m$, we must estimate $\boldsymbol{\alpha}$ using the nonlinear least squares and lose the elegance of the OLS estimation.

It is possible to show that if the skedastic regression is correctly specified, then FGLS is asymptotically equivalent to GLS, but the proof of this can be tricky. Below we just state the result without proof.

Theorem 2 *If the skedastic regression is correctly specified,*

$$\sqrt{n} \left(\bar{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}} \right) \xrightarrow{p} \mathbf{0},$$

and thus

$$\sqrt{n} \left(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta} \right) \xrightarrow{d} N(\mathbf{0}, \mathbf{V}),$$

where $\mathbf{V} = E \left[\sigma_i^{-2} \mathbf{x}_i \mathbf{x}_i' \right]^{-1}$.

Examining the asymptotic distribution in the above theorem, the natural estimator of the asymptotic variance of $\tilde{\boldsymbol{\beta}}$ is

$$\tilde{\mathbf{V}}^0 = \left(\frac{1}{n} \sum_{i=1}^n \tilde{\sigma}_i^{-2} \mathbf{x}_i \mathbf{x}_i' \right)^{-1} = \left(\frac{1}{n} \mathbf{X}' \tilde{\mathbf{D}}^{-1} \mathbf{X} \right)^{-1},$$

which is consistent for \mathbf{V} as $n \rightarrow \infty$. This estimator $\tilde{\mathbf{V}}^0$ is appropriate when the skedastic regression (9) is correctly specified.

It may be the case that $\boldsymbol{\alpha}'\mathbf{z}_i$ is only an approximation to the true conditional variance σ_i^2 . In this case we interpret $\boldsymbol{\alpha}'\mathbf{z}_i$ as a linear projection of u_i^2 on \mathbf{z}_i . $\tilde{\boldsymbol{\beta}}$ should perhaps be called a quasi-FGLS estimator of $\boldsymbol{\beta}$. Its asymptotic variance is not that given in the above theorem. Instead,

$$\mathbf{V} = \left(E \left[(\boldsymbol{\alpha}'\mathbf{z}_i)^{-1} \mathbf{x}_i \mathbf{x}_i' \right] \right)^{-1} E \left[(\boldsymbol{\alpha}'\mathbf{z}_i)^{-2} \sigma_i^2 \mathbf{x}_i \mathbf{x}_i' \right] \left(E \left[(\boldsymbol{\alpha}'\mathbf{z}_i)^{-1} \mathbf{x}_i \mathbf{x}_i' \right] \right)^{-1}$$

as shown in Section 1.3 of the last chapter. \mathbf{V} takes a sandwich form similar to the covariance matrix of the OLS estimator. Unless $\sigma_i^2 = \boldsymbol{\alpha}'\mathbf{z}_i$, $\tilde{\mathbf{V}}^0$ is inconsistent for \mathbf{V} . An appropriate solution

is to use a White-type estimator in place of $\tilde{\mathbf{V}}^0$. This may be written as

$$\begin{aligned}\tilde{\mathbf{V}} &= \left(\frac{1}{n} \sum_{i=1}^n \tilde{\sigma}_i^{-2} \mathbf{x}_i \mathbf{x}_i' \right)^{-1} \left(\frac{1}{n} \sum_{i=1}^n \tilde{\sigma}_i^{-4} \hat{u}_i^2 \mathbf{x}_i \mathbf{x}_i' \right) \left(\frac{1}{n} \sum_{i=1}^n \tilde{\sigma}_i^{-2} \mathbf{x}_i \mathbf{x}_i' \right)^{-1} \\ &= n \left(\mathbf{X}' \tilde{\mathbf{D}}^{-1} \mathbf{X} \right)^{-1} \left(\mathbf{X}' \tilde{\mathbf{D}}^{-1} \hat{\mathbf{D}} \tilde{\mathbf{D}}^{-1} \mathbf{X} \right) \left(\mathbf{X}' \tilde{\mathbf{D}}^{-1} \mathbf{X} \right)^{-1}\end{aligned}$$

where $\hat{\mathbf{D}} = \text{diag}\{\hat{u}_1^2, \dots, \hat{u}_n^2\}$. This is an estimator proposed by Cragg (1992), which is robust to misspecification of the conditional variance.

In the linear regression model, FGLS is asymptotically superior to OLS. Why then do we not exclusively estimate regression models by FGLS? This is a good question. There are three reasons. First, FGLS estimation depends on specification and estimation of the skedastic regression. Since the form of the skedastic regression is unknown, and it may be estimated with considerable error, the estimated conditional variances may contain more noise than information about the true conditional variances. In this case, FGLS performs worse than OLS in practice. Second, individual estimated conditional variances may be negative, and this requires trimming to solve. This introduces an element of arbitrariness which is unsettling to empirical researchers. Third, OLS is a more robust estimator of the parameter vector. It is consistent not only in the regression model ($E[u|\mathbf{x}] = 0$), but also under the assumptions of linear projection ($E[\mathbf{x}u] = \mathbf{0}$). The GLS and FGLS estimators, on the other hand, require the assumption of a correct conditional mean. If the equation of interest is a linear projection, and not a conditional mean, then the OLS and FGLS estimators will converge in probability to different limits, as they will be estimating two different projections. And the FGLS probability limit will depend on the particular function selected for the skedastic regression. The point is that the efficiency gains from FGLS are built on the stronger assumption of a correct conditional mean, and the cost is a reduction of robustness to misspecification.

6 Testing for Heteroskedasticity

If heteroskedasticity is present, more efficient estimation is possible, so we discuss testing for heteroskedasticity in this section. Heteroskedasticity may come from many sources, e.g., random coefficients, misspecification, stratified sampling, etc. So rejection of the null may be an indication of other deviations from our basic assumptions.

The hypothesis of homoskedasticity is that $E[u^2|\mathbf{x}] = \sigma^2$, or equivalently that

$$H_0 : \boldsymbol{\alpha}_1 = \mathbf{0}$$

in the regression (9). We may therefore test this hypothesis by the estimation (11) and constructing a Wald statistic.

This hypothesis does not imply that ξ_i is independent of \mathbf{x}_i . Typically, however, we impose the stronger hypothesis and test the hypothesis that u_i is independent of \mathbf{x}_i , in which case ξ_i is

independent of \mathbf{x}_i and the asymptotic variance (10) for $\tilde{\alpha}$ simplifies to

$$\mathbf{V}_{\alpha} = E [\mathbf{z}_i \mathbf{z}_i']^{-1} E [\xi_i^2]. \quad (14)$$

Hence the standard test of H_0 is a classic F (or Wald) test for exclusion of all regressors from the skedastic regression (11). The asymptotic distribution (12) and the asymptotic variance (14) under independence show that this test has an asymptotic chi-square distribution.

Theorem 3 *Under H_0 and u_i independent of \mathbf{x}_i , the Wald test of H_0 is asymptotically χ_q^2 .*

Most tests for heteroskedasticity take this basic form. The main difference between popular "tests" lies in which transformation of \mathbf{x}_i enters \mathbf{z}_i . Breusch-Pagan (1979) assume u_i follows $N(0, h(\alpha' \mathbf{z}_i))$ for a general $h(\cdot) > 0$, and use the LM test to check whether $\alpha_1 = \mathbf{0}$.⁸ Because α and β are "informationally" independent and H_0 involves only α , the LM test statistic can be much simplified. It turns out that

$$LM = \frac{1}{2\hat{\sigma}^4} \left(\sum_{i=1}^n \mathbf{z}_i f_i \right)' \left(\sum_{i=1}^n \mathbf{z}_i \mathbf{z}_i' \right)^{-1} \left(\sum_{i=1}^n \mathbf{z}_i f_i \right),$$

where $f_i = \hat{u}_i^2 - \hat{\sigma}^2$. This LM test statistic is similar to the Wald test statistic; a key difference is that $\hat{E} [\xi_i^2]$ is replaced by $2\hat{\sigma}^4$ which is a consistent estimator of $E [\xi_i^2]$ under H_0 where u_i follows $N(0, \sigma^2)$. Koenker (1981) shows that the asymptotic size and power of the Breusch-Pagan test is extremely sensitive to the kurtosis of the distribution of u_i , and suggests to renormalize LM by $n^{-1} \sum_{i=1}^n f_i^2$ rather than $2\hat{\sigma}^4$ to achieve the correct size. We denote the resulting LM test statistic as LM_K .

White (1980c) observes that when H_0 holds, $\hat{\Omega} = n^{-1} \sum_{i=1}^n \mathbf{x}_i \mathbf{x}_i' \hat{u}_i^2$ and $\hat{\sigma}^2 \hat{\mathbf{Q}} = (n^{-1} \sum_{i=1}^n \hat{u}_i^2) (n^{-1} \sum_{i=1}^n \mathbf{x}_i \mathbf{x}_i')$ should have the same probability limit, so the difference of them should converge to zero. In some sense, this is a Hausman-type test. Collecting non-redundant elements of $\mathbf{x}_i \mathbf{x}_i'$, denoted as $\mathbf{z}_i = (1, \mathbf{z}_{1i}')'$ as above, we are testing whether $\mathbf{D}_n = n^{-1} \sum_{i=1}^n \mathbf{z}_{1i} (\hat{u}_i^2 - \hat{\sigma}^2) \approx \mathbf{0}$. Under the auxiliary assumption that u_i is independent of \mathbf{x}_i , we can show that under H_0 ,

$$n \mathbf{D}_n' \hat{\mathbf{B}}_n^{-1} \mathbf{D}_n \xrightarrow{d} \chi_q^2,$$

where

$$\hat{\mathbf{B}}_n = \frac{1}{n} \sum_{i=1}^n (\hat{u}_i^2 - \hat{\sigma}^2)^2 (\mathbf{z}_{1i} - \bar{\mathbf{z}}_1) (\mathbf{z}_{1i} - \bar{\mathbf{z}}_1)'$$

is an estimator of the asymptotic variance matrix of $\sqrt{n} \mathbf{D}_n$. Given that u_i is independent of \mathbf{x}_i ,

⁸The general form of heteroskedasticity $h(\cdot)$ does not play any role in their arguments because under the null, any $h(\cdot)$ function reduces to a constant.

$\widehat{\mathbf{B}}_n$ can be replaced by

$$\widetilde{\mathbf{B}}_n = \frac{1}{n} \sum_{i=1}^n (\widehat{u}_i^2 - \widehat{\sigma}^2)^2 \frac{1}{n} \sum_{i=1}^n (\mathbf{z}_{1i} - \bar{\mathbf{z}}_1) (\mathbf{z}_{1i} - \bar{\mathbf{z}}_1)'$$

Exercise 7 Show that when $\mathbf{z}_i = (1, x_i)' \in \mathbb{R}^2$, $LM = (\widehat{\mathbf{u}}' \mathbf{D} \widehat{\mathbf{u}} / \widehat{\mathbf{u}}' \widehat{\mathbf{u}})^2$, where

$$\mathbf{D} = \text{diag} \left\{ n(x_i - \bar{x}) / \sqrt{2 \sum_{i=1}^n (x_i - \bar{x})^2}, i = 1, \dots, n \right\}.$$

If $x_i = 1$ for $i = 1, \dots, n_1$ and $x_i = 0$ for $i = n_1 + 1, \dots, n$, show that LM reduces to

$$\frac{1}{2} \frac{n}{n_1(n - n_1)} \left[\sum_{i=1}^{n_1} \left(\frac{\widehat{u}_i^2}{\widehat{\sigma}^2} \right)^2 - n_1 \right]^2.$$

Exercise 8 (i) Show that $\widehat{\sigma}^4 \cdot LM$ is one half of the explained sum of squares in the regression of \widehat{u}_i^2 upon \mathbf{z}_i . (ii) Show that $LM_K = nR_u^2 = nR^2$, where R_u^2 is the uncentered R^2 in the regression of f_i on \mathbf{z}_i , and R^2 is the centered R^2 in the regression of \widehat{u}_i^2 on \mathbf{z}_i . (iii) Show that $n\mathbf{D}'_n \widetilde{\mathbf{B}}_n^{-1} \mathbf{D}_n = nR^2$, where R^2 is the centered R^2 in the regression of \widehat{u}_i^2 on \mathbf{z}_i .

The Breusch-Pagan and White tests have degrees of freedom that depend on the number of regressors in $E[y|\mathbf{x}]$. Sometimes we want to conserve on degrees of freedom. A test that combines features of the Breusch-Pagan and White tests but has only two dfs takes $\mathbf{z}_{1i} = (\widehat{y}_i, \widehat{y}_i^2)'$, where \widehat{y}_i are the OLS fitted values. This reduced White test has some similarity to the RESET test. \widehat{y}_i are generated regressors, but as argued in the RESET test, this will cause no complications in the testing environment. So nR^2 from \widehat{u}_i^2 on $1, \widehat{y}_i, \widehat{y}_i^2$ has a limiting χ_2^2 distribution under H_0 .

(**) All the above tests are based on a key assumption under H_0 that u_i is independent of \mathbf{x}_i , especially, $E[u_i^4|\mathbf{x}_i]$ is constant. This assumption is usually called the **homokurtosis** (constant conditional fourth moment) assumption. When this assumption fails, Wooldridge (1990) proposes a **heterokurtosis-robust test** for heteroskedasticity, which is very similar to the heteroskedasticity-robust LM test.

There are some alternative tests of heteroskedasticity in the literature. Koenker and Bassett (1982) propose a robust test of heteroskedasticity based on quantile regression. For testing a more general deviation from homoskedasticity, i.e., $E[u_i^2|\mathbf{x}_i] = \sigma^2(\mathbf{x}_i)$ for a general $\sigma^2(\cdot)$, see Zheng (2009) where the testing idea follows from Zheng (1996). (**)

7 Regression Intervals and Forecast Intervals

All previous sections consider the internal validity. This section considers the external validity, i.e., prediction. We specifically concentrate on regression intervals and forecast intervals. First note that for prediction, misspecification is less important.

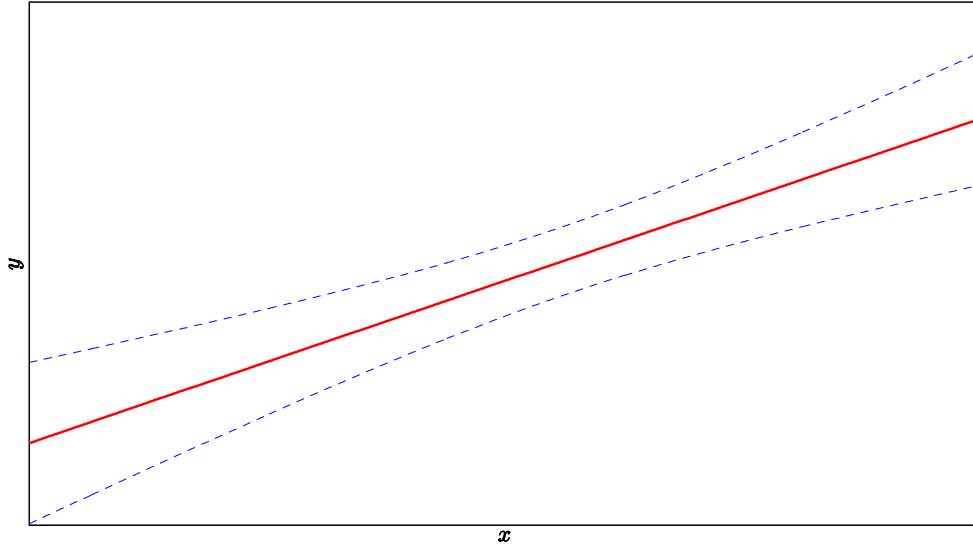


Figure 5: Typical Regression Intervals

In the linear regression model the conditional mean of y_i given $\mathbf{x}_i = \mathbf{x}$ is

$$m(\mathbf{x}) = E[y_i | \mathbf{x}_i = \mathbf{x}] = \mathbf{x}'\boldsymbol{\beta}.$$

In some cases, we want to estimate $m(\mathbf{x})$ at a particular point \mathbf{x} (which may or may not be the same as some \mathbf{x}_i). Notice that this is a (linear) function of $\boldsymbol{\beta}$. Letting $r(\boldsymbol{\beta}) = \mathbf{x}'\boldsymbol{\beta}$ and $\theta = r(\boldsymbol{\beta})$, we see that $\widehat{m}(\mathbf{x}) = \widehat{\theta} = \mathbf{x}'\widehat{\boldsymbol{\beta}}$ and $\mathbf{R} = \mathbf{x}$, so $s(\widehat{\theta}) = \sqrt{n^{-1}\mathbf{x}'\widehat{\mathbf{V}}\mathbf{x}}$. Thus an asymptotic 95% confidence interval for $m(\mathbf{x})$ is

$$\left[\mathbf{x}'\widehat{\boldsymbol{\beta}} \pm 2\sqrt{n^{-1}\mathbf{x}'\widehat{\mathbf{V}}\mathbf{x}} \right].$$

It is interesting to observe that if this is viewed as a function of \mathbf{x} , the width of the confidence set is dependent on \mathbf{x} . Typical regression intervals are shown in Figure 5, where the nonconstant covariate is only one-dimensional. Notice that the confidence bands take a hyperbolic shape. This means that the regression line is less precisely estimated for very large and very small values of x .

Exercise 9 In the wage equation, $\log \text{wage} = \beta_1 + \beta_2 \cdot \text{educ} + \beta_3 \cdot \text{exper} + \beta_3 \cdot \text{exper}^2 + u$, how to construct regression intervals for experience when education is fixed at its mean?

Exercise 10 In the linear regression $\log(y) = \mathbf{x}'\boldsymbol{\beta} + u$, denote $\widehat{\log y} = \mathbf{x}'\widehat{\boldsymbol{\beta}}$. If we want to predict $E[y_i | \mathbf{x}_i = \mathbf{x}]$, is $\exp\{\widehat{\log y}\}$ suitable? If not, does this predictor under- or over-estimate $E[y_i | \mathbf{x}_i = \mathbf{x}]$? Can you provide a more suitable predictor under the additional assumption that u is independent of \mathbf{x} ?

For a given value of $\mathbf{x}_i = \mathbf{x}$, we may want to forecast (guess) y_i out-of-sample.⁹ A reasonable rule is the conditional mean $m(\mathbf{x})$ as it is the mean-square-minimizing forecast. A point forecast is the estimated conditional mean $\widehat{m}(\mathbf{x}) = \mathbf{x}'\widehat{\boldsymbol{\beta}}$. We would also like a measure of uncertainty for the forecast.

The forecast error is $\widehat{u}_i = y_i - \widehat{m}(\mathbf{x}) = u_i - \mathbf{x}'(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})$. As the out-of-sample error u_i is independent of the in-sample estimate $\widehat{\boldsymbol{\beta}}$, this has variance

$$\begin{aligned} E[\widehat{u}_i^2] &= E[u_i^2 | \mathbf{x}_i = \mathbf{x}] + \mathbf{x}' E[(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})'] \mathbf{x} \\ &= \sigma^2(\mathbf{x}) + n^{-1} \mathbf{x}' \mathbf{V} \mathbf{x}. \end{aligned}$$

Assuming $E[u_i^2 | \mathbf{x}_i] = \sigma^2$, the natural estimate of this variance is $\widehat{\sigma}^2 + n^{-1} \mathbf{x}' \widehat{\mathbf{V}} \mathbf{x}$, so a standard error for the forecast is $\widehat{s}(\mathbf{x}) = \sqrt{\widehat{\sigma}^2 + n^{-1} \mathbf{x}' \widehat{\mathbf{V}} \mathbf{x}}$. Notice that this is different from the standard error for the conditional mean. If we have an estimate of the conditional variance function, e.g., $\widetilde{\sigma}^2(\mathbf{x}) = \widetilde{\boldsymbol{\alpha}}' \mathbf{z}$ from (13), then the forecast standard error is $\widehat{s}(\mathbf{x}) = \sqrt{\widetilde{\sigma}^2(\mathbf{x}) + n^{-1} \mathbf{x}' \widehat{\mathbf{V}} \mathbf{x}}$.

It would appear natural to conclude that an asymptotic 95% forecast interval for y_i is

$$\left[\mathbf{x}' \widehat{\boldsymbol{\beta}} \pm 2 \widehat{s}(\mathbf{x}) \right],$$

but this turns out to be incorrect. In general, the validity of an asymptotic confidence interval is based on the asymptotic normality of the studentized ratio. In the present case, this would require the asymptotic normality of the ratio

$$\frac{u_i - \mathbf{x}'(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})}{\widehat{s}(\mathbf{x})}.$$

But no such asymptotic approximation can be made. The only special exception is the case where u_i has the exact distribution $N(0, \sigma^2)$, which is generally invalid.

To get an accurate forecast interval, we need to estimate the conditional distribution of u_i given $\mathbf{x}_i = \mathbf{x}$, which is a much more difficult task. Given the difficulty, many applied forecasters focus on the simple approximate interval $\left[\mathbf{x}' \widehat{\boldsymbol{\beta}} \pm 2 \widehat{s}(\mathbf{x}) \right]$.

Exercise 11 In the homoskedastic regression model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}$ with $E[u_i | \mathbf{x}_i] = 0$ and $E[u_i^2 | \mathbf{x}_i] = \sigma^2$, suppose $\widehat{\boldsymbol{\beta}}$ is the OLS estimate with covariance matrix $\widehat{\mathbf{V}}$, based on a sample of size n . Let $\widehat{\sigma}^2$ be the estimate of σ^2 . You wish to forecast an out-of-sample value of y_{n+1} given that $\mathbf{x}_{n+1} = \mathbf{x}$. Thus the available information is the sample (\mathbf{y}, \mathbf{X}) , the estimates $(\widehat{\boldsymbol{\beta}}, \widehat{\mathbf{V}}, \widehat{\sigma}^2)$, the residuals $\widehat{\mathbf{u}}$, and the out-of-sample value of the regressors, \mathbf{x}_{n+1} .

(i) Find a point forecast of y_{n+1} .

(ii) Find an estimate of the variance of this forecast.

⁹ \mathbf{x} cannot be the same as any \mathbf{x}_i observed, why?

Exercise 12 (Empirical) Reconsider Nerlove's dataset in the last chapter, where you estimated a cost function on a cross-section of electric companies. The equation you estimated was

$$\log TC_i = \beta_1 + \beta_2 \log Q_i + \beta_3 \log PL_i + \beta_4 \log PK_i + \beta_5 \log PF_i + u_i. \quad (15)$$

- (a) Following Nerlove, add the variable $(\log Q_i)^2$ to the regression. Assess the merits of this new specification using (i) a hypothesis test; (ii) AIC criterion; (iii) BIC criterion. Do you agree with this modification?
- (b) Now try a non-linear specification. Consider model (15) plus the extra term $\beta_6 z_i$, where

$$z_i = \log Q_i (1 + \exp(-(\log Q_i - \beta_7)))^{-1}.$$

In addition, impose the restriction $\beta_3 + \beta_4 + \beta_5 = 1$. This is the smooth transition model. The model works best when β_7 is selected so that several values (in this example, at least 10 to 15) of $\log Q_i$ are both below and above β_7 . Examine the data and pick an appropriate range for β_7 .

- (c) Estimate the model by non-linear least squares. I recommend the concentration method: Pick 10 (or more or you like) values of β_7 in this range. For each value of β_7 , calculate z_i and estimate the model by OLS. Record the sum of squared errors, and find the value of β_7 for which the sum of squared errors is minimized.
- (d) Calculate standard errors for all the parameters $(\beta_1, \dots, \beta_7)$.

Exercise 13 (Empirical) The data file *cps78.dat* contains 550 observations on 20 variables taken from the May 1978 current population survey. Variables are listed in the file *cps78.pdf*. The goal of the exercise is to estimate a model for the log of earnings (variable *LNWAGE*) as a function of the conditioning variables.

- (a) Start by an OLS regression of *LNWAGE* on the other variables. Report coefficient estimates and standard errors.
- (b) Consider augmenting the model by squares and/or cross-products of the conditioning variables. Estimate your selected model and report the results.
- (c) Are there any variables which seem to be unimportant as a determinant of wages? You may re-estimate the model without these variables, if desired.
- (d) Test whether the error variance is different for men and women. Interpret.
- (e) Test whether the error variance is different for whites and nonwhites. Interpret.
- (f) Construct a model for the conditional variance. Estimate the model, test for general heteroskedasticity and report the results.

- (g) Using this model for the conditional variance, re-estimate the model from part (c) using FGLS. Report the results.
- (h) Do the OLS and FGLS estimates differ greatly? Note any interesting differences.
- (i) Compare the estimated standard errors. Note any interesting differences.

Technical Appendix A: Derivation of the AIC

We first derive AIC in a general model and then apply it to the normal regression model. An alternative simplified derivation can be found in Amemiya (1980).

Recall from Chapter 4 that the **Kullback-Leibler information distance** (or the **relative entropy**¹⁰) between probability densities q and p is defined as

$$D(q, p) \equiv \int \log \left(\frac{q(x)}{p(x)} \right) q(x) dx.$$

From Jensen's inequality, we know that (i) $D(q, p) \geq 0$ for all probability densities q and p ; (ii) $D(q, p) = 0$ iff $q = p$. So we can treat D as a criterion to measure a difference between two densities.¹¹ The AIC is just an estimate of D as shown below.

Consider q as the true density, and we want to pick up some p_θ that minimizes D out of the set of densities $\mathcal{P} = \{p_\theta : \theta \in \Theta\}$, where $\Theta \subset \mathbb{R}^k$ is a parameter space. Let

$$\bar{\theta} = \arg \min_{\theta \in \Theta} \int \log \left[\frac{q(x)}{p_\theta(x)} \right] q(x) dx,$$

and then, we have $(d/d\theta) \int \log (q(x)/p_\theta(x)) q(x) dx = 0$. By the second order Taylor approximation around $\bar{\theta}$,

$$\begin{aligned} D(q, p_\theta) &\approx D(q, p_{\bar{\theta}}) + \frac{1}{2} (\theta - \bar{\theta})' H(p_{\bar{\theta}}) (\theta - \bar{\theta}) \\ &= \int [\log q(x)] q(x) dx - \int [\log p_{\bar{\theta}}(x)] q(x) dx + \frac{1}{2} (\theta - \bar{\theta})' H(p_{\bar{\theta}}) (\theta - \bar{\theta}) \end{aligned} \quad (16)$$

where $H(p_\theta) \equiv - \int \left[\frac{\partial^2 \log p_\theta(x)}{\partial \theta \partial \theta'} \right] q(x) dx$.

We consider the minimization of the RHS in (16) with respect to θ . The first term of the RHS in (16) is independent of θ and $\bar{\theta}$, so can be ignored. The second term is independent of θ , but is related to the model parameter $\bar{\theta}$, so needs to be estimated. $-\int [\log p_{\bar{\theta}}(x)] q(x) dx$ can be

¹⁰Entropy is defined as $-E[\log p(X|\theta)]$, and is a measurement of the disorder of a system.

¹¹ D cannot serve as a metric in the space of densities in a strict sense, since it does not satisfy the symmetricity and triangle inequality. However, the plausibility of D as a criterion of difference between densities can be justified by Shannon's information theory, and the theory of information geometry (Efron (1978) and Amari (1985)).

estimated by

$$-\frac{1}{n} \sum_{i=1}^n \log p_{\tilde{\theta}}(X_i). \quad (17)$$

This is a consistent and unbiased estimator. But, $\bar{\theta}$ is still unknown; we estimate it by

$$\tilde{\theta} = \arg \min_{\theta \in \Theta} -\frac{1}{n} \sum_{i=1}^n \log p_{\theta}(X_i).$$

Expanding (17) around $\tilde{\theta}$, we get

$$-\frac{1}{n} \sum_{i=1}^n \log p_{\tilde{\theta}}(X_i) \approx -\frac{1}{n} \sum_{i=1}^n \log p_{\tilde{\theta}}(X_i) - \frac{1}{2} (\tilde{\theta} - \bar{\theta})' \frac{1}{n} \sum_{i=1}^n \frac{\partial^2 \log p_{\theta}(X_i)}{\partial \theta \partial \theta'} \Big|_{\theta=\tilde{\theta}} (\tilde{\theta} - \bar{\theta}) \quad (18)$$

since $\frac{1}{n} \sum_{i=1}^n (\partial/\partial \theta) \log p_{\theta}(X_i)|_{\theta=\tilde{\theta}} = 0$.¹² Only the third term $\frac{1}{2} (\theta - \bar{\theta})' H(p_{\bar{\theta}}) (\theta - \bar{\theta})$ depends on θ , which is minimized at $\bar{\theta}$. Since $\bar{\theta}$ is unknown, we estimate it by $\tilde{\theta}$. So the third term can be estimated by

$$-\frac{1}{2} (\tilde{\theta} - \bar{\theta})' \frac{1}{n} \sum_{i=1}^n \frac{\partial^2 \log p_{\theta}(X_i)}{\partial \theta \partial \theta'} \Big|_{\theta=\tilde{\theta}} (\tilde{\theta} - \bar{\theta})$$

In summary, we have

$$\begin{aligned} \min_{\theta} D(q, p_{\theta}) &= D(q, p_{\tilde{\theta}}) \\ &\approx -\frac{1}{n} \sum_{i=1}^n \log p_{\tilde{\theta}}(X_i) - (\tilde{\theta} - \bar{\theta})' \frac{1}{n} \sum_{i=1}^n \frac{\partial^2 \log p_{\theta}(X_i)}{\partial \theta \partial \theta'} \Big|_{\theta=\tilde{\theta}} (\tilde{\theta} - \bar{\theta}) + C, \end{aligned} \quad (19)$$

where C is some constant. We approximate $D(q, p_{\tilde{\theta}})$ further to simplify the formula.

Taking the expectation of the both sides of (19), we have

$$\begin{aligned} D(q, p_{\tilde{\theta}}) &\approx E[D(q, p_{\tilde{\theta}})] \\ &\approx -E \left[\frac{1}{n} \sum_{i=1}^n \log p_{\tilde{\theta}}(X_i) \right] - E \left[(\tilde{\theta} - \bar{\theta})' \frac{1}{n} \sum_{i=1}^n \frac{\partial^2 \log p_{\theta}(X_i)}{\partial \theta \partial \theta'} \Big|_{\theta=\tilde{\theta}} (\tilde{\theta} - \bar{\theta}) \right] + C. \end{aligned}$$

For large n , $n^{-1} \sum_{i=1}^n \log p_{\tilde{\theta}}(X_i)$ is close to its expectation. From White (1982), the asymptotic distribution of $\tilde{\theta}$ is

$$\sqrt{n} [\tilde{\theta} - \bar{\theta}] \xrightarrow{d} N \left(0, H(p_{\bar{\theta}})^{-1} J(p_{\bar{\theta}}) H(p_{\bar{\theta}})^{-1} \right) \quad (20)$$

¹²A naive estimator of $-\int \log [p_{\tilde{\theta}}(x)] q(x) dx$ would be $-n^{-1} \sum_{i=1}^n \log p_{\tilde{\theta}}(X_i)$. But this is a biased estimator (the bias comes from the nonlinearity of p in θ). So, (18) can be interpreted as a bias corrected estimator.

where $J(p_\theta) \equiv \int \left[\frac{\partial \log p_\theta(x)}{\partial \theta} \frac{\partial \log p_\theta(x)}{\partial \theta'} \right] q(x) dx$. Using the approximation of

$$-\frac{1}{n} \sum_{i=1}^n \frac{\partial^2 \log p_\theta(X_i)}{\partial \theta \partial \theta'} \Big|_{\theta=\bar{\theta}} \approx H(p_{\bar{\theta}}),$$

we have

$$-E \left[\left(\tilde{\theta} - \bar{\theta} \right)' \frac{1}{n} \sum_{i=1}^n \frac{\partial^2 \log p_\theta(X_i)}{\partial \theta \partial \theta'} \Big|_{\theta=\bar{\theta}} \left(\tilde{\theta} - \bar{\theta} \right) \right] \approx E \left[\left(\tilde{\theta} - \bar{\theta} \right)' H(p_{\bar{\theta}}) \left(\tilde{\theta} - \bar{\theta} \right) \right]. \quad (21)$$

It holds

$$\begin{aligned} \underbrace{\left(\tilde{\theta} - \bar{\theta} \right)' H(p_{\bar{\theta}}) \left(\tilde{\theta} - \bar{\theta} \right)}_{1 \times 1} &= \text{tr} \left\{ \left(\tilde{\theta} - \bar{\theta} \right)' H(p_{\bar{\theta}}) \left(\tilde{\theta} - \bar{\theta} \right) \right\} \\ &= \text{tr} \left\{ H(p_{\bar{\theta}}) \left(\tilde{\theta} - \bar{\theta} \right) \left(\tilde{\theta} - \bar{\theta} \right)' \right\}. \end{aligned}$$

From this and (20), we have

$$\begin{aligned} E \left[\left(\tilde{\theta} - \bar{\theta} \right)' H(p_{\bar{\theta}}) \left(\tilde{\theta} - \bar{\theta} \right) \right] &= \text{tr} \left\{ H(p_{\bar{\theta}}) E \left[\left(\tilde{\theta} - \bar{\theta} \right) \left(\tilde{\theta} - \bar{\theta} \right)' \right] \right\} \\ &\approx \frac{1}{n} \text{tr} \left\{ J(p_{\bar{\theta}}) H(p_{\bar{\theta}})^{-1} \right\} \end{aligned}$$

If $p_{\bar{\theta}}$ is close to the true model q , then we can say $H(p_{\bar{\theta}}) \approx J(p_{\bar{\theta}})$ by the usual information equality.

As a result,

$$\text{tr} \left\{ J(p_{\bar{\theta}}) H(p_{\bar{\theta}})^{-1} \right\} = \text{tr} \{ I_k \} = k (= \dim(\Theta)).$$

Therefore,

$$-E \left[\left(\tilde{\theta} - \bar{\theta} \right)' \frac{1}{n} \sum_{i=1}^n \frac{\partial^2 \log p_\theta(X_i)}{\partial \theta \partial \theta'} \Big|_{\theta=\bar{\theta}} \left(\tilde{\theta} - \bar{\theta} \right) \right] \approx \frac{k}{n} \quad (22)$$

In summary, we have the estimator of $D(q, p_{\bar{\theta}})$ as

$$-\frac{1}{n} \sum_{i=1}^n \log p_{\bar{\theta}}(X_i) + \frac{k}{n} + C,$$

and we minimize $D(q, p_{\bar{\theta}})$ with respect to k to select the correct model. In the normal regression model,

$$-\frac{1}{n} \sum_{i=1}^n \log [p_{\bar{\theta}}(X_i)] + \frac{k}{n} = \frac{1}{2} \log (2\pi \hat{\sigma}^2) + \frac{1}{2} + \frac{k}{n},$$

so minimize $D(q, p_{\bar{\theta}})$ is equivalent to minimize $\log(\hat{\sigma}^2) + \frac{2k}{n}$, which is the AIC in the main text.

Technical Appendix B: Derivation of the BIC

The derivation in this appendix follows from Robert (2001). The BIC is related to the asymptotic approximation to the Bayes factor (see Lavine and Schervish (1999) for an elementary exposition for Bayes factors). We will first review Bayesian factors and Laplace expansion, and then show that BIC is a natural corollary of this expansion.

The Bayes factor is the ratio of posterior odds to prior odds, that is,

$$B_{12} = \frac{p(\mathcal{M}_1|y)/p(\mathcal{M}_2|y)}{p(\mathcal{M}_1)/p(\mathcal{M}_2)} = \frac{p(y|\mathcal{M}_1)p(\mathcal{M}_1)/p(y)}{p(y|\mathcal{M}_2)p(\mathcal{M}_2)/p(y)} \bigg/ \frac{p(\mathcal{M}_1)}{p(\mathcal{M}_2)} = \frac{p(y|\mathcal{M}_1)}{p(y|\mathcal{M}_2)} = \frac{\int p(y|\theta_1, \mathcal{M}_1) \cdot p(\theta_1|\mathcal{M}_1)d\theta_1}{\int p(y|\theta_2, \mathcal{M}_2) \cdot p(\theta_2|\mathcal{M}_2)d\theta_2}. \quad (23)$$

Here, $p(\mathcal{M}_i)$ is the prior probability of model i , $p(\mathcal{M}_i|y)$ is the posterior probability of model i , $i = 1, 2$, and y is the data set. Intuitively, if $B_{12} > 1$, then we prefer model 1.

Laplace expansion is used to approximate some integrals like the numerator and denominator of equation (23) by normal density functions. The procedure is as follows,

$$\begin{aligned} & \int p(y|\theta_i, \mathcal{M}_i) \cdot p(\theta_i|\mathcal{M}_i)d\theta_i \\ &= \int \exp\{-nh(\theta_i)\} d\theta_i \\ &= \int \exp\left\{-nh(\hat{\theta}_i) - \frac{n}{2}(\theta_i - \hat{\theta}_i)' H(\hat{\theta}_i)(\theta_i - \hat{\theta}_i)\right\} d\theta_i + O\left(\frac{1}{n}\right) \\ &= \exp\{-nh(\hat{\theta}_i)\} \int \exp\left\{-\frac{1}{2}(\theta_i - \hat{\theta}_i)' \left[\frac{H^{-1}(\hat{\theta}_i)}{n}\right]^{-1} (\theta_i - \hat{\theta}_i)\right\} d\theta_i + O\left(\frac{1}{n}\right) \\ &= \exp\{-nh(\hat{\theta}_i)\} (2\pi)^{k_i/2} \left|\frac{H^{-1}(\hat{\theta}_i)}{n}\right|^{1/2} + O\left(\frac{1}{n}\right) \\ &= L_i(\hat{\theta}_i) \left(\frac{2\pi}{n}\right)^{k_i/2} [H(\hat{\theta}_i)]^{-1/2} + O\left(\frac{1}{n}\right), \end{aligned}$$

where k_i is the dimension of θ_i , $\hat{\theta}_i$ is the minima of $h(\theta_i)$,¹³ H is the Hessian matrix of h , and $L_i(\hat{\theta}_i)$ is the likelihood function of model i evaluated at its maxima. The second equality is from the second order Taylor expansion at the maxima of $-nh(\theta_i)$, or minima of $h(\theta_i)$, and the fourth equality is from the definition of the density function of multivariate normal. So we could approximate equation (23) as

$$B_{12} \approx \frac{L_{1,n}(\hat{\theta}_{1,n})}{L_{2,n}(\hat{\theta}_{2,n})} \left[\frac{|H_1(\hat{\theta}_{1,n})|}{|H_2(\hat{\theta}_{2,n})|} \right]^{-1/2} \left(\frac{n}{2\pi}\right)^{(k_2-k_1)/2}$$

¹³Note that the prior $p(\theta_i|\mathcal{M}_i)$ is asymptotically neglectable, so we can let $p(\theta_i|\mathcal{M}_i) = 1$ on the parameter space and $\hat{\theta}_i$ as the MLE.

The subscript n is to emphasize the fact that we based our inference on a size n sample. Therefore,

$$\log(B_{12}) \approx \log(\lambda_n) + \frac{k_2 - k_1}{2} \log(n) + R(\hat{\theta}_{1,n}, \hat{\theta}_{2,n})$$

where λ_n is the standard likelihood ratio for the comparison of \mathcal{M}_1 with \mathcal{M}_2 , $\lambda_n \equiv L_{1,n}(\hat{\theta}_{1,n})/L_{2,n}(\hat{\theta}_{2,n})$, and $R(\hat{\theta}_{1,n}, \hat{\theta}_{2,n})$ denotes the remainder term.

This approximation leads to Schwartz's criterion,

$$BIC = \log(\lambda_n) + \frac{k_2 - k_1}{2} \log(n)$$

when $\mathcal{M}_1 \subset \mathcal{M}_2$, if the remainder term $R(\hat{\theta}_{1,n}, \hat{\theta}_{2,n})$ is negligible compared with both other terms (i.e., $R(\hat{\theta}_{1,n}, \hat{\theta}_{2,n}) = O(1)$). Obviously, when $BIC > 0$, we should select model 1.

Applying this criterion to the normal regression model, we have

$$\begin{aligned} BIC &= \left(-\frac{n}{2} \log(2\pi\hat{\sigma}_1^2) - \frac{1}{2\hat{\sigma}_1^2} S_{n,1}(\hat{\beta}_1) \right) - \left(-\frac{n}{2} \log(2\pi\hat{\sigma}_2^2) - \frac{1}{2\hat{\sigma}_2^2} S_{n,2}(\hat{\beta}_2) \right) + \frac{(k_2 + 1) - (k_1 + 1)}{2} \log(n) \\ &= -\frac{n}{2} \left[\left(\log(\hat{\sigma}_1^2) + \frac{S_{n,1}(\hat{\beta}_1)/n}{\hat{\sigma}_1^2} + \log(n) \frac{k_1}{n} \right) - \left(\log(\hat{\sigma}_2^2) + \frac{S_{n,2}(\hat{\beta}_2)/n}{\hat{\sigma}_2^2} + \log(n) \frac{k_2}{n} \right) \right] \\ &= -\frac{n}{2} (BIC_1 - BIC_2), \end{aligned}$$

where $S_{n,i}$ is the sum of squared residuals in model i , $\hat{\sigma}_i^2 = S_{n,i}(\hat{\beta}_i)/n$, and $BIC_i \equiv \log(\hat{\sigma}_i^2) + \log(n) \frac{k_i}{n}$. We could see that $BIC > 0$ is equivalent to $\hat{\sigma}_1^2 + k_1 \frac{\log(n)}{n} < \hat{\sigma}_2^2 + k_2 \frac{\log(n)}{n}$, that is, Schwartz's criterion is equivalent to the BIC in the main text. Here, we should note that the dimension of parameter in model i is $(k_i + 1)$, where k_i is the dimension of β_i , $i = 1, 2$, because σ^2 is an extra parameter. Similarly, k in Appendix A should be the dimension of β plus 1. However, the AIC is equivalent to the criterion with k replaced by the dimension of β .