# Chapter 6. Additional Topics on Linear Regression*

This chapter covers additional topics on linear regression. Related materials can be found in Chapters 23 and 28 of Hansen (2022).

We first collect assumptions in the previous chapters:

**Assumption OLS.0** (random sampling): $(y_i, \mathbf{x}_i)$, $i = 1, \cdots, n$, 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 $E\left[\|\mathbf{x}\|^4\right] < \infty$.

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 | $\Longrightarrow$ | consistency |
| | $\cup$ | | |
| $E[u\|\mathbf{x}] = 0$ | linear regression | $\Longrightarrow$ | unbiasedness |
| | $\cup$ | | |
| $E[u\|\mathbf{x}] = 0$ and $E[u^2\|\mathbf{x}] = \sigma^2$ | homoskedastic linear regression | $\Longrightarrow$ | Gauss-Markov Theorem |
| | $\cup$ | | |
| $u$ is independent of $\mathbf{x}$ | normal regression is a special case | $\Longrightarrow$ | UMVUE |

Table 1: Relationship Between Different Models

This chapter will examine the validity of these assumptions and cures when they fail. We first in Section 1 examine OLS.0. Although this assumption is more likely to fail in time series, we

*Email: pingyu@hku.hk

study a microeconometric scenario where the data manifest a group structure and violate OLS.0. We then examine two other key assumptions - OLS.2 and OLS.3′.[1] Assumption OLS.2 has many implications. For example, it implies (i) $\boldsymbol{\beta}$ is fixed; (ii) the conditional mean of $y$ given $\mathbf{x}$ is linear in $\mathbf{x}$; (iii) all relevant regressors are included in $\mathbf{x}$ and are fixed. Section 2 extends the first implication by allowing $\boldsymbol{\beta}$ to be random. This generates the so-called random coefficient model as mentioned in Chapter 1. Section 3 and 4 examine the second implication: Section 3 tests whether $E[y|\mathbf{x}]$ is indeed $\mathbf{x}'\boldsymbol{\beta}$ and Section 4 provides more flexible specifications of $E[y|\mathbf{x}]$. Section 5 and 6 examine the third implication: Section 5 checks the benefits and costs of including irrelevant variables or omitting relevant variables, and Section 6 provides some model selection procedures. Section 7 and 8 examine Assumption OLS.3′: Section 7 shows that there are more efficient estimators of $\boldsymbol{\beta}$ when this assumption fails and Section 8 tests whether this assumption fails. Finally, Section 9 examines the external validity of the model.

# 1 The Clustering Problem

In econometric practices, it is often that the data have an obvioius grouped structure. For example, the test scores of children observed within classes or schools or years of schooling are correlated because these children are subject to some of the same environmental and family background influences; similar grouped structure appears in neighborhood, family, siblings etc. For another example, researchers often merge aggregate data on characteristics of industries, occupations, or geographical localities with micro data obtained from a cross-sectional or panel survey; the grouped structure implied by these aggregate data is obvious.

We assume individuals are independent across groups but may be dependent within a group. The correlation within a group is called the **clustering problem**, or the **Moulton problem**, after Moulton (1986),[2] who made it famous. See Chapter 8 of Angrist and Pischke (2009) for an introduction and Wooldridge (2003), Cameron and Miller (2011, 2015) and MacKinnon (2012, 2016) for summaries on this problem; other early important references include Pakes (1983), Arellano (1987), Moulton (1987, 1990), Moulton and Randolph (1989), Angrist (1991), and Pepper (2002); important recent developments include Bertrand et al. (2004), Donald and Lang (2007), C.B. Hansen (2007), Stock and Watson (2008) and B.E. Hansen and Lee (2019).

Suppose we have $G$ groups and each group includes $n_g$ individuals, i.e., $n = \sum_{g=1}^{G} n_g$. Define parallel notations as in the standard case,

$$\mathbf{y} = \begin{pmatrix} \mathbf{y}_1 \\ \vdots \\ \mathbf{y}_G \end{pmatrix}, \mathbf{X} = \begin{pmatrix} \mathbf{X}_1 \\ \vdots \\ \mathbf{X}_G \end{pmatrix}, \mathbf{u} = \begin{pmatrix} \mathbf{u}_1 \\ \vdots \\ \mathbf{u}_G \end{pmatrix},$$

[1] We examine only OLS.0, OLS.2 and OLS.3′ in this chapter. OLS.1 is discussed in Chapter 3. OLS.4 is a parametric assumption and OLS.5 is a regularity assumption, so they are not seriously examined in the literature.

[2] Brent R. Moulton was an American economist at the Bureau of Economic Analysis. He earned his Ph.D. from the University of Chicago in 1985.

where

$$\mathbf{y}_g = \begin{pmatrix} y_{g1} \\ \vdots \\ y_{gn_g} \end{pmatrix}, \mathbf{X}_g = \begin{pmatrix} \mathbf{x}'_{g1} \\ \vdots \\ \mathbf{x}'_{gn_g} \end{pmatrix}, \mathbf{u}_g = \begin{pmatrix} u_{g1} \\ \vdots \\ u_{gn_g} \end{pmatrix},$$

$g = 1, \cdots, G$. Consider the **error components** linear regression model,

$$y_{gi} = \mathbf{x}'_{gi}\boldsymbol{\beta} + u_{gi}, E\left[u_{gi}|\mathbf{X}_g\right] = 0, g = 1, \cdots, G, i = 1, \cdots, n_g,$$

where $u_{gi} = \alpha_g + \varepsilon_{gi}$, $\alpha_g$ is the common latent factor in group $g$, and $\varepsilon_{gi}$ is the i.i.d. idiosyncratic error for each individual. Assume $Var\left(\alpha_g|\mathbf{X}_g\right) = \sigma^2_\alpha$ and $Var\left(\varepsilon_{gi}|\mathbf{X}_g\right) = \sigma^2_\varepsilon$. Then

$$E\left[\mathbf{u}_g\mathbf{u}'_g|\mathbf{X}\right] = \boldsymbol{\Psi}_g = \sigma^2_u \begin{pmatrix} 1 & \rho_u & \cdots & \rho_u \\ \rho_u & 1 & & \vdots \\ \vdots & & \ddots & \rho_u \\ \rho_u & \cdots & \rho_u & 1 \end{pmatrix} = \sigma^2_u\left[(1-\rho_u)\mathbf{I}_g + \rho_u\mathbf{1}_g\mathbf{1}'_g\right],$$

and

$$E\left[\mathbf{u}\mathbf{u}'\right] = \boldsymbol{\Psi} = \text{diag}\left\{\boldsymbol{\Psi}_1, \cdots, \boldsymbol{\Psi}_G\right\},$$

where $\sigma^2_u = \sigma^2_\alpha + \sigma^2_\varepsilon$, $\rho_u = corr\left(u_{gi}, u_{g_j}\right)_{i\neq j} = \frac{\sigma^2_\alpha}{\sigma^2_u} = \frac{\sigma^2_\alpha}{\sigma^2_\alpha+\sigma^2_\varepsilon}$ is the intragroup correlation, $\mathbf{I}_g$ is the $n_g \times n_g$ identity matrix, and $\mathbf{1}_g$ is a column vector of $n_g$ ones. That is, the errors are equicorrelated within groups. Now, the OLS estimator is

$$\widehat{\boldsymbol{\beta}} = \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{y} = \left(\sum_{g=1}^{G}\mathbf{X}'_g\mathbf{X}_g\right)^{-1}\left(\sum_{g=1}^{G}\mathbf{X}'_g\mathbf{y}_g\right);$$

from Chapter 3,

$$\begin{aligned} Var\left(\widehat{\boldsymbol{\beta}}|\mathbf{X}\right) &= \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\boldsymbol{\Psi}\mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1} = \left(\sum_{g=1}^{G}\mathbf{X}'_g\mathbf{X}_g\right)^{-1}\left(\sum_{g=1}^{G}\mathbf{X}'_g\boldsymbol{\Psi}_g\mathbf{X}_g\right)\left(\sum_{g=1}^{G}\mathbf{X}'_g\mathbf{X}_g\right)^{-1} \\ &= \sigma^2_u\left(\sum_{g=1}^{G}\sum_{i=1}^{n_g}\mathbf{x}_{gi}\mathbf{x}'_{gi}\right)^{-1}\left(\sum_{g=1}^{G}\mathbf{X}'_g\left[(1-\rho_u)\mathbf{I}_g + \rho_u\mathbf{1}_g\mathbf{1}'_g\right]\mathbf{X}_g\right)\left(\sum_{g=1}^{G}\sum_{i=1}^{n_g}\mathbf{x}_{gi}\mathbf{x}'_{gi}\right)^{-1} \\ &= \sigma^2_u\left(\sum_{g=1}^{G}\sum_{i=1}^{n_g}\mathbf{x}_{gi}\mathbf{x}'_{gi}\right)^{-1}\left[(1-\rho_u)\sum_{g=1}^{G}\sum_{i=1}^{n_g}\mathbf{x}_{gi}\mathbf{x}'_{gi} + \rho_u\sum_{g=1}^{G}\left(\sum_{i=1}^{n_g}\mathbf{x}_{gi}\right)\left(\sum_{i=1}^{n_g}\mathbf{x}_{gi}\right)'\right]\cdot\left(\sum_{g=1}^{G}\sum_{i=1}^{n_g}\mathbf{x}_{gi}\mathbf{x}'_{gi}\right)^{-1}. \end{aligned}$$

If $\rho_u = 0$, then $\boldsymbol{\Psi} = \sigma^2_u\mathbf{I}_n$ and $Var\left(\widehat{\boldsymbol{\beta}}|\mathbf{X}\right)$ degenerates to the standard homoskedastic covariance matrix, $Var_H\left(\widehat{\boldsymbol{\beta}}|\mathbf{X}\right) = \sigma^2_u\left(\mathbf{X}'\mathbf{X}\right)^{-1}$. The difference between $Var\left(\widehat{\boldsymbol{\beta}}|\mathbf{X}\right)$ and $Var_H\left(\widehat{\boldsymbol{\beta}}|\mathbf{X}\right)$ is

$$\begin{aligned} Var\left(\widehat{\boldsymbol{\beta}}|\mathbf{X}\right) - Var_H\left(\widehat{\boldsymbol{\beta}}|\mathbf{X}\right) &= \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\boldsymbol{\Psi}\mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1} - \sigma^2_u\left(\mathbf{X}'\mathbf{X}\right)^{-1} \\ &= \rho_u\sigma^2_u\left(\mathbf{X}'\mathbf{X}\right)^{-1}\left(\mathbf{N} - \mathbf{I}_n\right), \end{aligned}$$

where

$$\mathbf{N} = \mathbf{X}'\mathbf{D}\mathbf{D}'\mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1},$$

and $\mathbf{D}$ is an $n \times G$ matrix consisting of 0-1 indicator variables for membership (when the data is **balanced**, i.e., $n_g = T$, $\mathbf{D} = \mathbf{I}_G \otimes \mathbf{1}_T$).

To compare with $Var_H\left(\widehat{\boldsymbol{\beta}}|\mathbf{X}\right)$, we make further assumptions for simplification, following Kloek (1981), which provides a good approximation even in the general case. First, assume $\mathbf{x}_{gi} = \mathbf{x}_g$ for all $i = 1, \cdots, n_g$, i.e., all regressors are group or aggregate (market or policy) variables, then $\mathbf{X}_g = \mathbf{1}_g\mathbf{x}_g'$, $\mathbf{X}_g'\mathbf{X}_g = \sum_{i=1}^{n_g}\mathbf{x}_{gi}\mathbf{x}_{gi}' = n_g\mathbf{x}_g\mathbf{x}_g'$, and $\mathbf{X}_g'\mathbf{1}_g = \sum_{i=1}^{n_g}\mathbf{x}_{gi} = n_g\mathbf{x}_g$. Note that

$$\begin{aligned}
\mathbf{X}_g'\boldsymbol{\Psi}_g\mathbf{X}_g &= \mathbf{X}_g'\boldsymbol{\Psi}_g\mathbf{X}_g = \mathbf{x}_g\mathbf{1}_g'\boldsymbol{\Psi}_g\mathbf{1}_g\mathbf{x}_g' \\
&= \sigma_u^2\mathbf{x}_g\mathbf{1}_g'\begin{pmatrix} 1+(n_g-1)\rho_u \\ \vdots \\ 1+(n_g-1)\rho_u \end{pmatrix}\mathbf{x}_g' \\
&= \sigma_u^2 n_g\left[1+(n_g-1)\rho_u\right]\mathbf{x}_g\mathbf{x}_g' \equiv \sigma_u^2 n_g\tau_g\mathbf{x}_g\mathbf{x}_g',
\end{aligned}$$

so

$$Var\left(\widehat{\boldsymbol{\beta}}|\mathbf{X}\right) = \sigma_u^2\left(\sum_{g=1}^{G} n_g\mathbf{x}_g\mathbf{x}_g'\right)^{-1}\sum_{g=1}^{G} n_g\tau_g\mathbf{x}_g\mathbf{x}_g'\left(\sum_{g=1}^{G} n_g\mathbf{x}_g\mathbf{x}_g'\right)^{-1}.$$

Second, assume further that the data is balanced so that $\tau_g = \tau = 1+(T-1)\rho_u$, then

$$\begin{aligned}
Var\left(\widehat{\boldsymbol{\beta}}|\mathbf{X}\right) &= \sigma_u^2\tau\left(\sum_{g=1}^{G} T\mathbf{x}_g\mathbf{x}_g'\right)^{-1}\sum_{g=1}^{G} T\mathbf{x}_g\mathbf{x}_g'\left(\sum_{g=1}^{G} T\mathbf{x}_g\mathbf{x}_g'\right)^{-1} \\
&= \sigma_u^2\tau\left(\sum_{g=1}^{G} T\mathbf{x}_g\mathbf{x}_g'\right)^{-1} = \tau\left[\sigma_u^2\left(\mathbf{X}'\mathbf{X}\right)^{-1}\right],
\end{aligned}$$

i.e., $Var\left(\widehat{\boldsymbol{\beta}}|\mathbf{X}\right)$ is $\tau(>1)$ times the usual covariance matrix. $\tau$ is increasing in $T$ and $\rho_u$ is illustrated in an application of Pepper (2002). The under-estimation of standard errors can result in spurious findings of statistical significance, especially for the aggregate variable of interest as shown in the following exercise.

**Exercise 1** *Show that when* $\mathbf{x}_{gi} = (1, x_{gi})'$, *i.e.,* $\dim(\mathbf{x}_{gi}) = 2$ *and* $y_{gi} = \beta_0 + x_{gi}\beta_1 + u_{gi}$,

$$\frac{Var\left(\widehat{\beta}_1|\mathbf{X}\right)}{Var_H\left(\widehat{\beta}_1|\mathbf{X}\right)} = 1 + \left[\frac{Var(n_g)}{\overline{n}} + \overline{n} - 1\right]\rho_x\rho_u,$$

*where* $Var(n_g) = \frac{1}{G}\sum_{g=1}^{G}(n_g - \overline{n})^2$ *with* $\overline{n} = \frac{1}{G}\sum_{g=1}^{G} n_g$, *and* $\rho_x = \frac{\sum_g\sum_{i\neq j}(x_{gi}-\overline{x})(x_{gj}-\overline{x})}{Var(x_{gi})\sum_g n_g(n_g-1)}$ *with* $\overline{x} = \frac{1}{n}\sum_g\sum_i x_{gi}$ *and* $Var(x_{gi}) = \frac{1}{n}\sum_g\sum_i(x_{gi}-\overline{x})^2$ *is the intragroup correlation of* $x_{gi}$.

To estimate $Var\left(\widehat{\boldsymbol{\beta}}|\mathbf{X}\right)$, we need to estimate $\sigma_u^2$ and $\rho_u$. First, natural estimators of $\sigma_u^2$ and $\sigma_\alpha^2$ are

$$\widehat{\sigma}_u^2 = \frac{1}{n}\sum_{g=1}^{G}\sum_{i=1}^{n_g}\widehat{u}_{gi}^2 \text{ and } \widehat{\sigma}_\alpha^2 = \frac{1}{G}\sum_{g=1}^{G}\frac{\sum_{i\neq j}\widehat{u}_{gi}\widehat{u}_{gj}}{n_g\left(n_g-1\right)}.$$

So we can estimate $Var\left(\widehat{\boldsymbol{\beta}}|\mathbf{X}\right)$ by

$$\widehat{Var}\left(\widehat{\boldsymbol{\beta}}|\mathbf{X}\right) = \left(\sum_{g=1}^{G}\mathbf{X}_g'\mathbf{X}_g\right)^{-1}\left(\sum_{g=1}^{G}\mathbf{X}_g'\widehat{\boldsymbol{\Psi}}_g\mathbf{X}_g\right)\left(\sum_{g=1}^{G}\mathbf{X}_g'\mathbf{X}_g\right)^{-1}$$

where $\widehat{\boldsymbol{\Psi}}_g = \widehat{\sigma}_u^2\left[(1-\widehat{\rho}_u)\mathbf{I}_g + \widehat{\rho}_u\mathbf{1}_g\mathbf{1}_g'\right]$ with $\widehat{\rho}_u = \widehat{\sigma}_\alpha^2/\widehat{\sigma}_u^2$. Alternatively, as shown in Pepper (2002),

$$\sqrt{G}\left(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}\right) \xrightarrow{d} N\left(\mathbf{0}, E\left[\mathbf{X}_g'\mathbf{X}_g\right]^{-1}E\left[\mathbf{X}_g'\mathbf{u}_g\mathbf{u}_g'\mathbf{X}_g\right]E\left[\mathbf{X}_g'\mathbf{X}_g\right]^{-1}\right),$$

so we can estimate $Var\left(\widehat{\boldsymbol{\beta}}|\mathbf{X}\right)$ by

$$\widehat{Var}\left(\widehat{\boldsymbol{\beta}}|\mathbf{X}\right) = \left(\sum_{g=1}^{G}\mathbf{X}_g'\mathbf{X}_g\right)^{-1}\left(\sum_{g=1}^{G}\mathbf{X}_g'\widehat{\mathbf{u}}_g\widehat{\mathbf{u}}_g'\mathbf{X}_g\right)\left(\sum_{g=1}^{G}\mathbf{X}_g'\mathbf{X}_g\right)^{-1} = \left(\mathbf{X}'\mathbf{X}\right)^{-1}\left(\sum_{g=1}^{G}\mathbf{X}_g'\widehat{\mathbf{u}}_g\widehat{\mathbf{u}}_g'\mathbf{X}_g\right)\left(\mathbf{X}'\mathbf{X}\right)^{-1},$$

which allows for general within-group covariance and heteroskedasticity and was suggested by Liang and Zeger (1986). These estimators of $Var\left(\widehat{\boldsymbol{\beta}}|\mathbf{X}\right)$ are valid when $G\to\infty$.

We can extend the degrees-of-freedom correction of Bell and McCaffrey (2002) discussed in LN5 to the clustering case. Here, the counterpart of HC2 is

$$\widehat{\mathbf{V}}_{\text{LZ2}} = \left(\mathbf{X}'\mathbf{X}\right)^{-1}\left(\sum_{g=1}^{G}\mathbf{X}_g'\left(\mathbf{I}_{n_g}-\mathbf{P}_{gg}\right)^{-1/2}\widehat{\mathbf{u}}_g\widehat{\mathbf{u}}_g'\left(\mathbf{I}_{n_g}-\mathbf{P}_{gg}\right)^{-1/2}\mathbf{X}_g\right)\left(\mathbf{X}'\mathbf{X}\right)^{-1},$$

$\mathbf{G}$ is an $n\times G$ matrix with $g$th column equal to the $n$-vector

$$\mathbf{G}_g = \left(\mathbf{I}_n-\mathbf{P}\right)_g\left(\mathbf{I}_{n_g}-\mathbf{P}_{gg}\right)^{-1/2}\mathbf{X}_g\left(\mathbf{X}'\mathbf{X}\right)^{-1}e_{k,j},$$

and

$$K_{\text{BM}} = \frac{\left(\sum_{i=1}^{G}\lambda_i\right)^2}{\sum_{i=1}^{G}\lambda_i^2}$$

where $\left(\mathbf{I}_n-\mathbf{P}\right)_g$ consists of the $n_g$ columns of the $n\times n$ matrix $\left(\mathbf{I}_n-\mathbf{P}\right)$ corresponding to cluster $g$, $\mathbf{P}_{gg} = \mathbf{X}_g\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}_g'$, and $\lambda_i$ are the eigenvalues of $\mathbf{G}'\mathbf{G}$. If $n_g=1$ for all $g$, then the adjustment is the same as that in Chapter 5. Recall that under heteroskedasticity, $\lambda_i$ should be the eigenvalues of $\mathbf{G}'\boldsymbol{\Omega}\mathbf{G}$. Imbens and Kolesár (2016, IK hereafter) suggest to estimate $\boldsymbol{\Omega}$ using the error components model above. Their $\widehat{\sigma}_u^2$ is the same as above, but their $\widehat{\sigma}_\alpha^2 = \frac{\sum_{g=1}^{G}\sum_{i\neq j}\widehat{u}_{gi}\widehat{u}_{gj}}{\sum_{g=1}^{G}n_g(n_g-1)} = \sum_{g=1}^{G}\omega_g\frac{\sum_{i\neq j}\widehat{u}_{gi}\widehat{u}_{gj}}{n_g(n_g-1)}$

with $\omega_g = \frac{n_g(n_g-1)}{\sum_{g=1}^G n_g(n_g-1)}$; obviously, if all $n_g$'s are equal, their $\widehat{\sigma}_\alpha^2$ is the same as above. The simulations in Imbens and Kolesár (2016) show that the degrees-of-freedom corrections of BM and IK work well even in moderately sized samples.

**Exercise 2** *Show that* $\widehat{\sigma}_\alpha^2 = \frac{1}{n}\sum_{g=1}^G n_g \overline{\widehat{u}}_g^2$ *is inconsistent to* $\sigma_\alpha^2$ *when* $n_g = T$ *fixed and* $G \to \infty$.

(**) We can use the feasible GLS to improve the efficiency of OLS. Specifically, define

$$\widehat{\boldsymbol{\beta}}_{GLS} = \left(\sum_{g=1}^G \mathbf{X}_g' \widehat{\boldsymbol{\Psi}}_g^{-1} \mathbf{X}_g\right)^{-1} \left(\sum_{g=1}^G \mathbf{X}_g' \widehat{\boldsymbol{\Psi}}_g^{-1} \mathbf{y}_g\right),$$

whose variance can be estimated by $\left(\sum_{g=1}^G \mathbf{X}_g' \widehat{\boldsymbol{\Psi}}_g^{-1} \mathbf{X}_g\right)^{-1}$. To check whether there is indeed a grouped structure, we can use Breusch and Pagan (1980)'s LM test to test $H_0 : \rho_u = 0$ vs. $\rho_u > 0$, where we assume $\mathbf{u} \sim N(\mathbf{0}, \boldsymbol{\Psi})$. Honda (1985) and King and Evans (1986) find the one-sided LM statistic is more powerful, where the one-sided LM statistic is defined as

$$LM = \frac{\widehat{\mathbf{u}}'\mathbf{D}\mathbf{D}'\widehat{\mathbf{u}} - \widehat{\mathbf{u}}'\widehat{\mathbf{u}}}{\frac{\widehat{\mathbf{u}}'\widehat{\mathbf{u}}}{n}\sqrt{2\left(\sum n_g^2 - n\right)}} = \frac{\sum_g \left(n_g \overline{\widehat{u}}_g\right)^2 - \sum_g \sum_i \widehat{u}_{gi}^2}{\widehat{\sigma}_u^2 \sqrt{2\left(\sum n_g^2 - n\right)}} = n\frac{\frac{\sum_g \left(\sum_i \widehat{u}_{gi}\right)^2}{\sum_g \sum_i \widehat{u}_{gi}^2} - 1}{\sqrt{2\left(\sum n_g^2 - n\right)}} \equiv n\frac{s-1}{b},$$

which follows a standard normal asymptotic distribution as $n \to \infty$ and $G \to \infty$ under the null. In finite samples, $N(0,1)$ does not approximate $LM$ well, so Moulton and Randolph (1989) suggest to use the standardized LM statistic,

$$SLM = \frac{s - \mu_1}{\sqrt{\mu_2}},$$

where $\mu_1 = E[s|\mathbf{X}]$ and $\mu_2 = Var(s|\mathbf{X})$. The SLM statistic has a zero mean and unit variance and is asymptotically $N(0,1)$. The difference between $LM$ and $SLM$ depends on the size of $k$ and $\mathbf{D}'\mathbf{P}_\mathbf{X}\mathbf{D}$ (i.e., whether $\mathbf{X}$ can explain group membership). An alternative test is the $F$ test, which tries to test whether the columns of $\mathbf{D}$ jointly contribute to explaining the variance of $\mathbf{y}$. Since some columns of $\mathbf{X}$ may not have variation within groups, $\text{rank}([\mathbf{X}, \mathbf{D}])$ may be less than $k + G$, where $\mathbf{W} = [\mathbf{X}, \mathbf{D}]$. For example, suppose $\mathbf{X} = [\mathbf{X}_1, \mathbf{X}_2]$, where $\mathbf{X}_2$ contains the $k_2$ variables that do not change within groups (so the constant is also included in $\mathbf{X}_2$); then $\text{rank}(\mathbf{W}) = k_1 + G$, where $k_1 = k - k_2$. Define the $F$ statistic as

$$F = \frac{\mathbf{y}'\mathbf{M}_\mathbf{X}\mathbf{D}\left(\mathbf{D}'\mathbf{M}_\mathbf{X}\mathbf{D}\right)^- \mathbf{D}'\mathbf{M}_\mathbf{X}\mathbf{y}/(\text{rank}(\mathbf{W}) - k)}{\mathbf{y}'\mathbf{M}_\mathbf{W}\mathbf{y}/(n - \text{rank}(\mathbf{W}))},$$

which follows $F_{\text{rank}(\mathbf{W})-k, n-\text{rank}(\mathbf{W})}$ under the null, where $\mathbf{M}_\cdot$ is the annihilator, and $(\cdot)^-$ is a generalized inverse operator. Although the one-sided LM test is locally most powerful, the $F$ test may have better power when the deviation from the null is large.

All the above inferences are valid only when $G \to \infty$. When $G$ is small, as in the difference-

in-differences models, we need new inference results. Donald and Lang (2007) apply the two-step procedure of Amemiya (1978), as discussed in the next section, to estimate $\boldsymbol{\beta}_2$ (the coefficients of $\mathbf{X}_2$), and show that the usual inference methods based on the OLS estimator and $\widehat{Var}\left(\widehat{\boldsymbol{\beta}}|\mathbf{X}\right)$ (or $\widetilde{Var}\left(\widehat{\boldsymbol{\beta}}|\mathbf{X}\right)$) are quite misleading no matter $n_g$ is large or small. They also show that the $t$-statistic in their two-step procedure would follow a $t$ distribution in many scenarios where $n_g$ is either large or small. (**)

## 2 Random Coefficient Models

According to Swamy[3] and Tavlas (2001), **random coefficient models** (RCMs) grew out of Zellner (1969)[4] although already appeared in Rubin (1950), Klein (1953)[5] and Hildreth and Houck (1968). The following RCMs are called the first-generation RCMs in Swamy and Tavlas (2001). An exposition of this kind of RCMs at text-book level can be found in Judge et al. (1985, Chs 11, 13, and 19). See, also, Swamy (1971), Chow (1984), Nicholls and Pagan (1985) and Hsiao and Pesaran (2008).

Think about the return-to-schooling example. Suppose $y = \log(wage)$, $\mathbf{x} = (1, educ)'$, and $\boldsymbol{\beta} = (\beta_1, \beta_2)'$. Obviously, the return to education, i.e., $\beta_2$, may vary among individuals, so it is more convenient to write

$$y_i = \beta_1 + \beta_{2i} \cdot educ + u_i,$$

where $\beta_{2i}$ is the individual-specific return to education.[6] Generally, we write

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

where $\beta_{1i} = \beta_1 + u_i$ if there is an intercept. In other words, the usual linear regression models allow only the intercept to be random, while the RCMs further allow all slopes to be random.[7] In Chapter 1, we express $\boldsymbol{\beta}_i = \boldsymbol{\beta}(u_i)$. Since the $\boldsymbol{\beta}(\cdot)$ function is not restricted, these two expressions are equivalent.

Although when $\boldsymbol{\beta}_i$ is treated as fixed, the following exercise shows that the LSE converges to $\overline{\boldsymbol{\beta}}$ (the average of $\boldsymbol{\beta}_i$'s), it is more convenient to treat $\boldsymbol{\beta}_i$'s as random draws from a common

[3] P.A.V.B. Swamy (1934-) is an Indian-born statistician at the Federal Reserve System. He earned his Ph.D. from the University of Wisconsin–Madison under the supervision of Arthur Goldberger in 1968. Arnold Zellner is in his dissertation committee, which can explain why his thesis title was "Statistical Inference in a Random Coefficient Regression Model." His Ph.D. thesis led to a book, Swamy (1971), which quickly became a classic on RCMs.

[4] Arnold Zellner (1927-2010) was an American economist and statistician at the University of Chicago, specializing in the fields of Bayesian probability and econometrics. He earned his Ph.D. in economics from the University of California, Berkeley, in 1957 under supervision of George Kuznets whose older brother Simon Kuznets won the 1971 Nobel Prize.

[5] Lawrence Robert Klein (1920-2013) was an American econometrician at the University of Pennsylvania. He was awarded the Nobel Prize in 1980 "for the creation of econometric models and their application to the analysis of economic fluctuations and economic policies."

[6] $\beta_{2i}$ is the causal effect of education on log wage for individual $i$, while as shown below, $E[\beta_{2i}]$ is the corresponding average treatment effect.

[7] The clustering problem in the last section can be treated as a RCM where the intercept remains, and the coefficients on the group indicators are random with mean zero and independent of $u_i$.

distribution. We make the following two key assumptions:

**(i)** $\boldsymbol{\beta}_i$'s are i.i.d. with mean $\boldsymbol{\beta}$ and variance $\boldsymbol{\Sigma}$;

**(ii)** $\mathbf{x}_i$ is independent of $\boldsymbol{\beta}_i$.

Under these two assumptions we can show that the RCM can be expressed as a heteroskedastic linear regression model.

**Exercise 3 (*)** *Suppose $\boldsymbol{\beta}_i$ is fixed. Show that the LSE converges in probability to $\overline{\boldsymbol{\beta}} \equiv \lim_{n\to\infty} n^{-1}\sum_{i=1}^{n}\boldsymbol{\beta}_i$.*

Write $\boldsymbol{\beta}_i = \boldsymbol{\beta} + \boldsymbol{\varepsilon}_i$, where $E[\boldsymbol{\varepsilon}_i] = \mathbf{0}$ and $Var(\boldsymbol{\varepsilon}_i) = E[\boldsymbol{\varepsilon}_i\boldsymbol{\varepsilon}_i'] = \boldsymbol{\Sigma}$ from Assumption (i). Then

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

where

$$\begin{aligned} E[u_i|\mathbf{x}_i] &= E\left[\mathbf{x}_i'\boldsymbol{\varepsilon}_i|\mathbf{x}_i\right] = \mathbf{x}_i'E[\boldsymbol{\varepsilon}_i|\mathbf{x}_i] = \mathbf{x}_i'E[\boldsymbol{\varepsilon}_i] = 0, \\ Var(u_i|\mathbf{x}_i) &= E\left[u_i^2|\mathbf{x}_i\right] = E\left[\mathbf{x}_i'\boldsymbol{\varepsilon}_i\boldsymbol{\varepsilon}_i'\mathbf{x}_i|\mathbf{x}_i\right] = \mathbf{x}_i'E\left[\boldsymbol{\varepsilon}_i\boldsymbol{\varepsilon}_i'|\mathbf{x}_i\right]\mathbf{x}_i = \mathbf{x}_i'E\left[\boldsymbol{\varepsilon}_i\boldsymbol{\varepsilon}_i'\right]\mathbf{x}_i = \mathbf{x}_i'\boldsymbol{\Sigma}\mathbf{x}_i, \end{aligned}$$

and the second to last equality in both equations follow from Assumption (ii). In other words, RCMs are natural cases of heteroskedastic linear regression - the CEF is linear in $\mathbf{x}$ with the coefficients $\boldsymbol{\beta}$ equal to the mean of the random coefficient $\boldsymbol{\beta}_i$, and the conditional variance is quadratic in $\mathbf{x}$. As a result, the asymptotics developed in Chapter 5 for the heteroskedastic linear regression model can be applied.

**Exercise 4** *Suppose the components of $\boldsymbol{\beta}_i$, $\beta_{1i}, \cdots, \beta_{ki}$, in the RCM are independent. Find the asymptotic variance of LSE.*

(**) In some cases, we may restrict some components of $\boldsymbol{\beta}_i$ to be nonrandom, which would restrict the corresponding rows and columns of $\boldsymbol{\Sigma}$ as zero. As to random components, we can assume they are related to some observable covariates. For example, Amemiya (1978) considered the following setup,

$$\begin{aligned} \mathbf{y} &= \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{D}\boldsymbol{\gamma} + \boldsymbol{\varepsilon}, \\ \boldsymbol{\gamma} &= \mathbf{X}_2\boldsymbol{\beta}_2 + \boldsymbol{\alpha}, \end{aligned}$$

which implies

$$\mathbf{y} = \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{D}\mathbf{X}_2\boldsymbol{\beta}_2 + \mathbf{D}\boldsymbol{\alpha} + \boldsymbol{\varepsilon},$$

where $\boldsymbol{\alpha}$ and $\boldsymbol{\varepsilon}$ are uncorrelated,

$$Var(\boldsymbol{\varepsilon}) = \boldsymbol{\Lambda}, Var(\boldsymbol{\alpha}) = \boldsymbol{\Omega}, Var(\mathbf{D}\boldsymbol{\alpha} + \boldsymbol{\varepsilon}) = \mathbf{D}\boldsymbol{\Omega}\mathbf{D}' + \boldsymbol{\Lambda} \equiv \boldsymbol{\Sigma},$$

and $\mathbf{\Lambda}$, $\mathbf{\Omega}$, $\mathbf{X}_2$, and $[\mathbf{X}_1, \mathbf{D}]$ are all full-rank. First, assume $\mathbf{\Lambda}$ and $\mathbf{\Omega}$ are known. Then the BLUE of $\boldsymbol{\beta}_2$ is the GLS estimator,

$$\widehat{\boldsymbol{\beta}}_2 = \left[\mathbf{X}_2'\mathbf{D}'\mathbf{\Sigma}^{*-1}\mathbf{D}\mathbf{X}_2\right]^{-1}\mathbf{X}_2'\mathbf{D}'\mathbf{\Sigma}^{*-1}\mathbf{y},$$

where $\mathbf{\Sigma}^{*-1} = \mathbf{\Sigma}^{-1} - \mathbf{\Sigma}^{-1}\mathbf{X}_1\left(\mathbf{X}_1'\mathbf{\Sigma}^{-1}\mathbf{X}_1\right)^{-1}\mathbf{X}_1'\mathbf{\Sigma}^{-1}$. $\widehat{\boldsymbol{\beta}}_2$ is obtained by regressing $\mathbf{\Sigma}^{-1/2}\mathbf{y}$ on $\mathbf{\Sigma}^{-1/2}\mathbf{X}_1$ and $\mathbf{\Sigma}^{-1/2}\mathbf{D}\mathbf{X}_2$ and applying the FWL theorem. Another estimation method is to use a two-step procedure: estimate $\boldsymbol{\gamma}$ first and then regress the estimated $\boldsymbol{\gamma}$ on $\mathbf{X}_2$ to estimate $\boldsymbol{\beta}_2$. Specifically, in step one, define

$$\widehat{\boldsymbol{\gamma}} = \left(\mathbf{D}'\mathbf{\Lambda}^{*-1}\mathbf{D}\right)^{-1}\mathbf{D}'\mathbf{\Lambda}^{*-1}\mathbf{y},$$

which is the BLUE of $\boldsymbol{\gamma}$, where $\mathbf{\Lambda}^{*-1} = \mathbf{\Lambda}^{-1} - \mathbf{\Lambda}^{-1}\mathbf{X}_1\left(\mathbf{X}_1'\mathbf{\Lambda}^{-1}\mathbf{X}_1\right)^{-1}\mathbf{X}_1'\mathbf{\Lambda}^{-1}$. In step two, since

$$\widehat{\boldsymbol{\gamma}} = \boldsymbol{\gamma} + \widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma} = \mathbf{X}_2\boldsymbol{\beta}_2 + \boldsymbol{\alpha} + \left(\mathbf{D}'\mathbf{\Lambda}^{*-1}\mathbf{D}\right)^{-1}\mathbf{D}'\mathbf{\Lambda}^{*-1}\boldsymbol{\varepsilon},$$

the BLUE of $\boldsymbol{\beta}_2$ in this regression is

$$\widetilde{\boldsymbol{\beta}}_2 = \left[\mathbf{X}_2'\Phi^{*-1}\mathbf{X}_2\right]^{-1}\mathbf{X}_2'\Phi^{*-1}\widehat{\boldsymbol{\gamma}},$$

where

$$\Phi^{*-1} = \mathbf{D}'\mathbf{\Lambda}^{*-1}\mathbf{D}\left(\mathbf{D}'\mathbf{\Lambda}^{*-1}\mathbf{\Sigma}\mathbf{\Lambda}^{*-\mathbf{1}}\mathbf{D}\right)^{-1}\mathbf{D}'\mathbf{\Lambda}^{*-1}\mathbf{D}.$$

Amemiya showed that $\widehat{\boldsymbol{\beta}}_2 = \widetilde{\boldsymbol{\beta}}_2$. $\widetilde{\boldsymbol{\beta}}_2$ may have computational advantage over $\widehat{\boldsymbol{\beta}}_2$ when $\mathbf{\Lambda}^{-1}$ is easier to compute than $\mathbf{\Sigma}^{-1}$.

If feasible GLS is used, then provided the covariance matrices $\mathbf{\Lambda}$ and $\mathbf{\Omega}$ are estimated in the same fashion, numerical equivalence continue to hold. More commonly, the two approaches lend themselves to different methods for obtaining consistent estimates of the covariance matrices. If so, the equivalence is asymptotic rather than numeric.

In Donald and Lang (2007), $\mathbf{X}_2$ is $G \times k_2$, $\mathbf{\Lambda} = \sigma_\varepsilon^2\mathbf{I}_n$, $\mathbf{\Omega} = \sigma_\alpha^2\mathbf{I}_G$ and $\mathbf{\Sigma} = \mathbf{\Psi}$. So in step one, the BLUE of $(\boldsymbol{\beta}_1, \boldsymbol{\gamma})$ is the LSE $\left(\widehat{\boldsymbol{\beta}}_1, \widehat{\boldsymbol{\gamma}}\right)$. In step two,

$$\overline{\mathbf{y}}_g - \overline{\mathbf{X}}_{1g}'\widehat{\boldsymbol{\beta}}_1 \equiv \widehat{\beta}_{2g} = \mathbf{x}_{2g}'\boldsymbol{\beta}_2 + \boldsymbol{\alpha} + \overline{\boldsymbol{\varepsilon}}_g - \overline{\mathbf{X}}_{1g}'\left(\widehat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1\right), g = 1, \cdots, G,$$

where $\overline{\mathbf{y}}_g$ and $\overline{\mathbf{X}}_{1g}$ are the within-group means of $y_{gi}$ and $\mathbf{x}_{1gi}$. They use this form of regression in step two because when $\mathbf{X}_1$ is absent, $\widehat{\gamma}_g = \overline{\mathbf{y}}_g$. Obviously, new observations within group do not provide additional information for estimating $\boldsymbol{\beta}_2$ beyond affecting $\overline{\mathbf{y}}_g$, so $\boldsymbol{\beta}_2$ cannot be consistently estimated if $G$ is small even if $n_g \to \infty$. They provide a few scenarios where the error term $\boldsymbol{\alpha} + \overline{\boldsymbol{\varepsilon}}_g - \overline{\mathbf{X}}_{1g}'\left(\widehat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1\right)$ is homoskedastic so that the LSE is the BLUE; if the error term is further assumed to be normal, then the $t$-statistic would follow $t\left(G - k_2\right)$. (**)

**Exercise 5 (*)** *Show that* $\widehat{\boldsymbol{\beta}}_2 = \widetilde{\boldsymbol{\beta}}_2$.

(*) All the above models assume $\boldsymbol{\beta}_i$ and $\mathbf{x}_i$ are independent in some sense. Garen (1984) discusses a case in the estimation of return to schooling where $\boldsymbol{\beta}_i$ and $\mathbf{x}_i$ are correlated. Later on, Wooldridge (1997, 2003, 2008), Heckman and Vytlacil (1998) and Masten and Torgovitsky (2016) provide more discussions on identification and estimation in the so-called **correlated random coefficient model**. See also Heckman et al. (2010) and Heckman and Robb (1985, pp. 195-197) for testing the hypothesis that $\boldsymbol{\beta}_i$ and $\mathbf{x}_i$ are independent.

# 3 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 $y_i = \mathbf{x}_i'\widehat{\boldsymbol{\beta}} + \widehat{u}_i$ 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 $y_i = \mathbf{x}_i'\widetilde{\boldsymbol{\beta}} + \mathbf{z}_i'\widetilde{\boldsymbol{\gamma}} + \widetilde{u}_i$ 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).[8] The null model is

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

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

$$\mathbf{z}_i = \left(\widehat{y}_i^2, \cdots, \widehat{y}_i^m\right)'$$

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

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

by OLS, and form the Wald statistic $W_n$ for $\boldsymbol{\gamma} = \mathbf{0}$. $\widehat{y}_i$'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 $H_0$, $W_n \xrightarrow{d} \chi^2_{m-1}$. Thus the null is rejected at the $\alpha$ level if $W_n$ exceeds the upper $\alpha$ tail critical value of the $\chi^2_{m-1}$ 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

[8] James Bernard Ramsey (1937- ) is a Canadian econometrician. He is a professor of Economics at New York University. He got his Ph.D. in Economics from the University of Wisconsin-Madison in 1968. His most influential paper, Ramsey (1969), is based on his dissertation under the supervision of Arnold Zellner.

(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

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

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,

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

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

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

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 6** *(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 $Avar\left(\sqrt{n}\left(\widehat{\boldsymbol{\beta}}_{OLS} - \widehat{\boldsymbol{\beta}}_{WLS}\right)\right) = \mathbf{V}$ in (2). (ii) If $E[u_i^2|\mathbf{x}_i]$ is not a constant, what is the expression for $Avar\left(\sqrt{n}\left(\widehat{\boldsymbol{\beta}}_{OLS} - \widehat{\boldsymbol{\beta}}_{WLS}\right)\right)$? If $w_i = \sigma_i^{-2}$, what will the expression for $Avar\left(\sqrt{n}\left(\widehat{\boldsymbol{\beta}}_{OLS} - \widehat{\boldsymbol{\beta}}_{WLS}\right)\right)$ 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 $\chi^2$ distribution for a sequence of local alternative hypotheses, with a noncentrality parameter depending on $\text{plim}(\widehat{\boldsymbol{\beta}}_{OLS} - \widehat{\boldsymbol{\beta}}_{WLS})$. To generate power, the weights in $\widehat{\boldsymbol{\beta}}_{WLS}$ are important. White (1981) suggests to impose weights on the area where the linear approximation is bad. Specifically, the weights are predicted $\widehat{u}_i^2$ using all non-redundant elements of $\mathbf{x}_i$, its squares, and all cross-products.[9]

All these tests mentioned above are special cases of the conditional moment (CM) test of Newey (1985a) and Tauchen (1985).[10] Specifically, under $H_0$, $E[u|\mathbf{x}] = 0$, so any function of $\mathbf{x}$

[9] So 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.

[10] Note that these authors consider the CM test in the context of likelihood models.

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

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

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 $E[y|\mathbf{x}] = \mathbf{x}'\boldsymbol{\beta}$. $f(\mathbf{x})$ is added in to offset the denominator in $E[u|\mathbf{x}] = \int u\frac{f(u,\mathbf{x})}{f(\mathbf{x})}du$.[11] 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 Porter and Yu (2015) for more discussions. (**)

# 4 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, e.g., the mean of the Poisson distribution.

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

A generalized version of the power transformation is the famous Box-Cox (1964) transformation, where the regressor $x^{\theta_3}$ is generalized as $x^{(\theta_3)}$ with

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

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

[11] (*) 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.

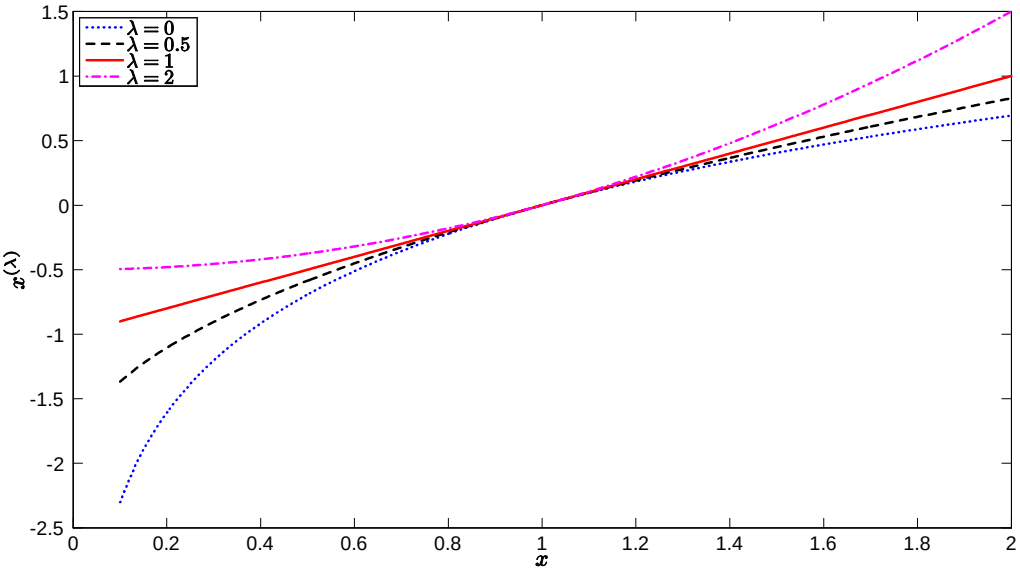


Figure 1: Box-Cox Transformation for Different $\lambda$ Values

- $m(x|\boldsymbol{\theta}) = \theta_1 + \theta_2 \exp(\theta_3 x)$: Exponentially Transformed Regressors
- $m(\mathbf{x}|\boldsymbol{\theta}) = G(\mathbf{x}'\boldsymbol{\theta})$, $G$ known[12]

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
- $m(\mathbf{x}|\boldsymbol{\theta}) = \left(\boldsymbol{\theta}_1'\mathbf{x}_1\right) 1(x_2 \leq \theta_3) + \left(\boldsymbol{\theta}_2'\mathbf{x}_1\right) 1(x_2 > \theta_3)$: 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 ($\mathbf{x}_1 = (1, x)'$, $x_2 = x$, $\boldsymbol{\theta}_1 = (1, -1)'$, $\boldsymbol{\theta}_2 = (-2, 2)'$, $\theta_3 = 1, \theta_4 = 0.1$ 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 $x \in [0, 2]$.

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

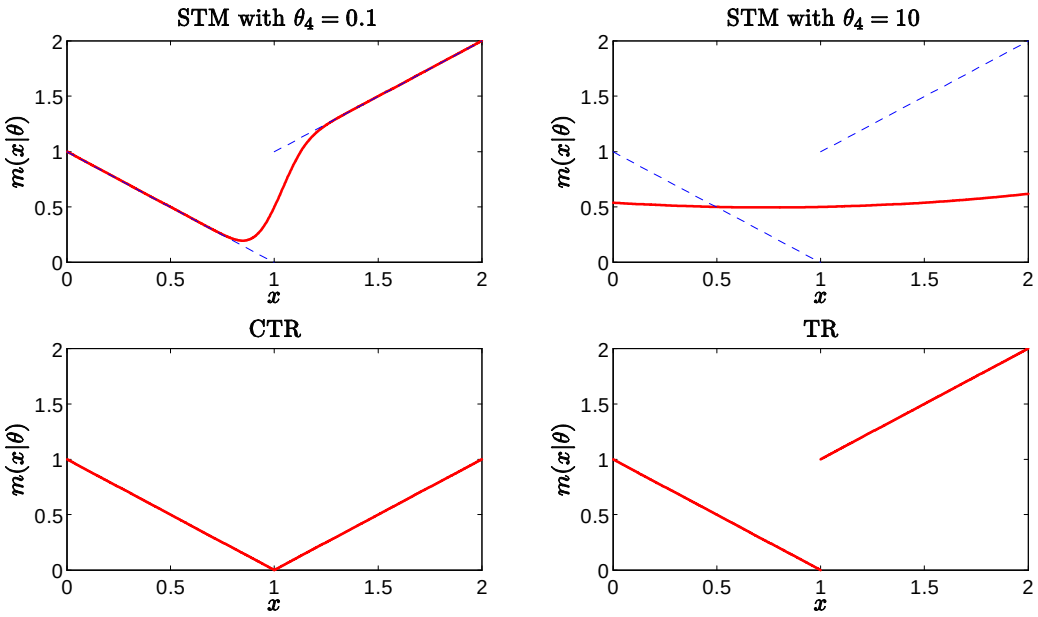


Figure 2: Difference Between STM, CTR and TR

In the first five examples, $m(\mathbf{x}|\theta)$ is (generically) differentiable with respect to the parameters $\boldsymbol{\theta}$. In the last two examples, $m$ is not differentiable with respect to $\theta_4$ and $\theta_3$ which alters some of the analysis. When it exists, let

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

The least squares estimator $\widehat{\boldsymbol{\theta}}$ 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 $\widehat{u}_i = y_i - m(\mathbf{x}_i|\widehat{\boldsymbol{\theta}})$. One motivation for the choice of NLLS as the estimation method is that the parameter is the solution to the population problem $\min_{\boldsymbol{\theta}} E\left[(y_i - m(\mathbf{x}_i|\boldsymbol{\theta}))^2\right]$.

Since sum-of-squared-errors function $S_n(\boldsymbol{\theta})$ is not quadratic, $\widehat{\boldsymbol{\theta}}$ 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

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

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

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

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

**Exercise 7** *Prove the above theorem. (Hint: Note that* $\widehat{\boldsymbol{\theta}}$ *is the MoM estimator associated with the moment conditions* $E\left[\mathbf{m}_{\boldsymbol{\theta}i}u_i\right] = \mathbf{0}$*.)*

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

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

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 the unknown parameter $\gamma$. Examples include $x_i(\gamma) = x_i^{\gamma}$, $x_i(\gamma) = \exp(\gamma x_i)$, $x_i(\boldsymbol{\gamma}) = x_i G\left(\frac{x_i-\gamma_1}{\gamma_2}\right)$ and $x_i(\gamma) = x_i 1(g(x_i) > \gamma)$. The model is linear when $\boldsymbol{\beta}_2 = 0$, 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 $\gamma$ have dropped out. This means that under $H_0$, $\gamma$ is not identified. Such tests are labeled as tests with nuisance parameter ($\gamma$) 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 $\theta_3$. 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. (**)

# 5 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

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

where $\boldsymbol{\Gamma} = E\left[\mathbf{x}_{1i}\mathbf{x}_{1i}'\right]^{-1-1} E[\mathbf{x}_{1i}\mathbf{x}_{2i}']$ 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 course of an empirical investigation.

When $\boldsymbol{\beta}_2 = \mathbf{0}$ and $\boldsymbol{\beta}_1$ is the parameter of interest, $\mathbf{x}_{2i}$ is "irrelevant". In this case, the estimator

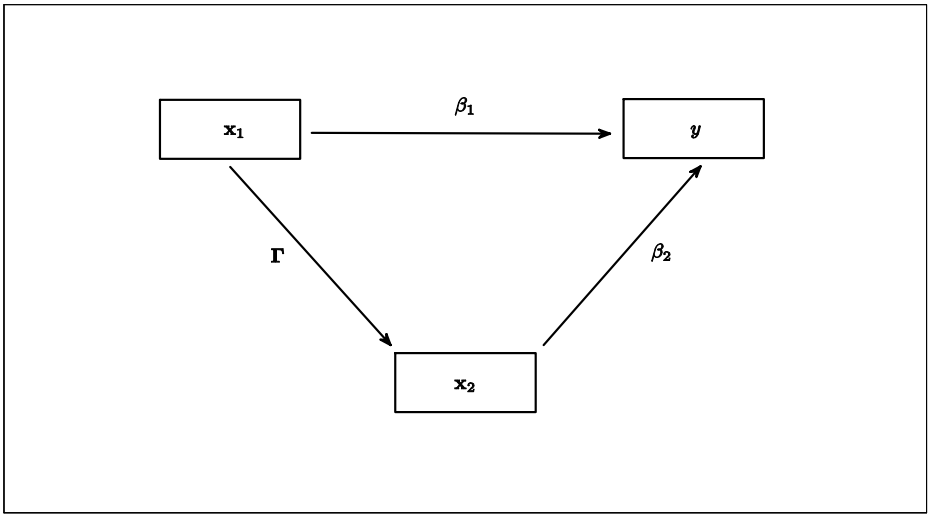


Figure 3: Direct and Indirect Effects of $\mathbf{x}_1$ on $y$

of $\boldsymbol{\beta}_1$ from the short regression, $\overline{\boldsymbol{\beta}}_1 = (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{y}$, is consistent from the analysis above. So we compare its efficiency relative to the estimator from the long regression, $\widehat{\boldsymbol{\beta}}_1$. 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,

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

and

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

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 $\overline{\boldsymbol{\beta}}_1$ has a lower asymptotic variance matrix than $\widehat{\boldsymbol{\beta}}_1$. 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 $AVar\left(\widehat{\boldsymbol{\beta}}_1\right)$ from $\mathbf{Q}_{11}$ to $\mathbf{Q}_{11.2}$ without decreasing its numerator $\sigma^2$. 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 $\beta_1$ 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,

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

while

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

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 ($\boldsymbol{\beta}_2 \neq \mathbf{0}$) and uncorrelated with $\mathbf{x}_1$, then it decreases the numerator without affecting the denominator of $AVar\left(\widehat{\boldsymbol{\beta}}_1\right)$, 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.

| | $\boldsymbol{\beta}_2 = \mathbf{0}$ | $\boldsymbol{\beta}_2 \neq \mathbf{0}$ |
|---|---|---|
| $\mathbf{Q}_{12} = \mathbf{0}$ | $\overline{\boldsymbol{\beta}}_1$ consistent<br>same efficiency | $\overline{\boldsymbol{\beta}}_1$ consistent<br>long more efficient |
| $\mathbf{Q}_{12} \neq \mathbf{0}$ | $\overline{\boldsymbol{\beta}}_1$ consistent<br>short more efficient | depends on $\mathbf{Q}_{12} \cdot \boldsymbol{\beta}_2$<br>undetermined |

Table 2: Consistency and Efficiency with Omitted and Irrelevant Variables

**Exercise 8** *If* $\boldsymbol{\beta}_2 \neq \mathbf{0}$, $\mathbf{Q}_{12} \neq \mathbf{0}$, *could* $\mathbf{Q}_{12}\boldsymbol{\beta}_2$ *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 $\widehat{\boldsymbol{\beta}}_1$ is more efficient than $\overline{\boldsymbol{\beta}}_1$ ($= \widehat{\boldsymbol{\beta}}_{1R}$) 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.

# 6 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 and Linhart and Zucchini (1986) and Burham and Anderson (2002) for general references on model selection.

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 $(x_1, \cdots, x_K)$ enters the regression function $E[y_i|x_{1i} = x_1, \cdots, x_{Ki} = x_K]$?" is well posed.

## 6.1 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 $\widehat{\mathbf{u}}_1$ and $\widehat{\mathbf{u}}_2$, estimated variances $\widehat{\sigma}_1^2$ and $\widehat{\sigma}_2^2$, 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

$$P\left(\widehat{\mathcal{M}} = \mathcal{M}_1|\mathcal{M}_1\right) \to 1 \text{ and } P\left(\widehat{\mathcal{M}} = \mathcal{M}_2|\mathcal{M}_2\right) \to 1.$$

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 $\alpha$, let $c_\alpha$ satisfy $P\left(\chi^2_{k_2} > c_\alpha\right) = \alpha$. 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 $\alpha$ in hypothesis testing (controlling Type I error) is not relevant for model selection. That is, if $\alpha$ 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 $\alpha$ is held fixed, then this model selection procedure is inconsistent, as $P(\widehat{\mathcal{M}} = \mathcal{M}_1|\mathcal{M}_1) \to 1 - \alpha < 1$ although $P(\widehat{\mathcal{M}} = \mathcal{M}_2|\mathcal{M}_2) \to 1$.

### 6.1.1 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).[13] From Appendix A, the AIC under normality for model $m$ is

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

where $\widehat{\sigma}_m^2$ is the variance estimate for model $m$ and is roughly $-2\ell_n$ (neglecting the constant term) with $\ell_n$ 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,

$$LR = n\left(\log\left(\widehat{\sigma}_1^2\right) - \log\left(\widehat{\sigma}_2^2\right)\right) \xrightarrow{d} \chi^2_{k_2}, \tag{6}$$

then

$$\begin{aligned} P\left(\widehat{\mathcal{M}} = \mathcal{M}_1|\mathcal{M}_1\right) &= P\left(AIC_1 < AIC_2|\mathcal{M}_1\right) \\ &= P\left(\log\left(\widehat{\sigma}_1^2\right) + 2\frac{k_1}{n} < \log\left(\widehat{\sigma}_2^2\right) + 2\frac{k_1+k_2}{n}\middle|\mathcal{M}_1\right) \\ &= P\left(LR < 2k_2|\mathcal{M}_1\right) \to P\left(\chi^2_{k_2} < 2k_2\right) < 1. \end{aligned} \tag{7}$$

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 $KLIC(\mathcal{M}) = E\left[\log f(\mathbf{y}|\mathbf{X}) - \log f(\mathbf{y}|\mathbf{X}, \mathcal{M})\right]$ between the true density and the model density.[14] 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).[15] The BIC is based on approximating

[13] Hirotugu Akaike (1927-2009) was a Japanese statistician at the Institute of Statistical Mathematics, Tokyo.

[14] This is also the term "information criterion" originated.

[15] Another Baysian measure of model complexity, especially when the number of parameters is not clearly defined, is the deviance information criterion (DIC) by Spiegelhalter et al. (2002).

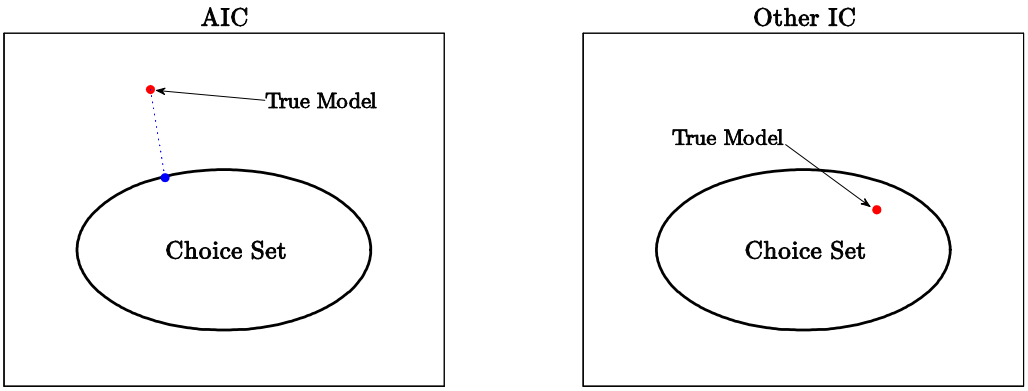


Figure 4: Difference Between AIC and Other IC

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

$$BIC_m = \log\left(\widehat{\sigma}_m^2\right) + \log n \frac{k_m}{n}. \tag{8}$$

Since $\log(n) > 2$ (if $n \geq 8$), 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

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

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

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

thus

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

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. Although BIC leads to a consistent model selector if the true data generating model belongs to the finite-parameter family under investigation, as shown by Haughton (1989) for expoential families, BIC selected models tend to underfit if this assumption does not hold.

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 $\widehat{\sigma}^2$ for each model, and then selects the model with the lowest AIC (5). Similarly, the BIC selects based on (8). 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 9** *In the ordered regressors case, compare the difference between the AIC, the BIC and the LR test in the conclusion of model selection.*

### 6.1.2 Limitation and Extension of Information Criteria (*)

Given their simplicity, penalized likelihoood 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. Such shrinkage estimtors are especially useful when the number of regressors, $k$, is large relative to (or even larger than) the number of available observations, $n$, where the standard information criteria such asAIC and BIC are not applicable; see Hastie et al. (2009) and Bühlmann and van de Geer (2012) for comprehensive reviews of literature. 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 (2007);[16] see Claeskens and Hjort (2008) for a summary of literature.

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

### 6.2 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 $\widehat{\sigma}_m^2$ in (5) or (8) 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

[16] Rigorously speaking, Hansen's model averaging is to approximate an infinite-dimensional model rather than a parametric model as Hjort and Claeskens. In this sense, Hansen's work is close to Li (1987) where the best model rather than an average is selected.

case often take one of two approaches. **Artificial nesting**, proposed in Dividson 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. See Gourieroux and Monfort (1994, 1995 Ch. 22) and Pesaran and Weeks (2001) for a summary of literature.

We will not discuss these nonnested tests in details, but only 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 $u_i \sim N(0, \sigma^2)$.[17]

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.

# 7 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

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

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

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

$$\overline{\boldsymbol{\beta}} = \left(\mathbf{X}'\mathbf{D}^{-1}\mathbf{X}\right)^{-1}\left(\mathbf{X}'\mathbf{D}^{-1}\mathbf{y}\right), \tag{9}$$

where $\mathbf{D} = \text{diag}\left\{\sigma_1^2, \cdots, \sigma_n^2\right\}$ 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

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

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 10** *Consider the WLS estimator (9) with* $\mathbf{D} = \text{diag}\left\{x_{j1}^2, \cdots, x_{jn}^2\right\}$, *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 (9) is infeasible since the matrix $\mathbf{D}$ is unknown. A feasible GLS (FGLS) estimator replaces the unknown $\mathbf{D}$ with an estimate $\widehat{\mathbf{D}} = \text{diag}\left\{\widehat{\sigma}_1^2, \cdots, \widehat{\sigma}_n^2\right\}$. 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

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

and we have the regression equation

$$\begin{aligned} \eta_i &= \alpha_0 + \mathbf{z}_{1i}'\boldsymbol{\alpha}_1 + \xi_i, \\ E\left[\xi_i|\mathbf{x}_i\right] &= 0. \end{aligned} \tag{10}$$

This regression error $\xi_i$ is generally heteroskedastic and has the conditional variance

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

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

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

and

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

where

$$\mathbf{V}_{\boldsymbol{\alpha}} = \left(E\left[\mathbf{z}_i\mathbf{z}_i'\right]\right)^{-1}\left(E\left[\mathbf{z}_i\mathbf{z}_i'\xi_i^2\right]\right)\left(E\left[\mathbf{z}_i\mathbf{z}_i'\right]\right)^{-1}. \tag{11}$$

**Exercise 11** *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* $\left(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\gamma}}\right)$ *for* $(\boldsymbol{\beta}, \boldsymbol{\gamma})$.

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

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

And then

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

Let

$$\widetilde{\boldsymbol{\alpha}} = \left(\mathbf{Z}'\mathbf{Z}\right)^{-1}\mathbf{Z}'\widehat{\boldsymbol{\eta}} \tag{12}$$

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

$$\sqrt{n}\left(\widetilde{\boldsymbol{\alpha}} - \boldsymbol{\alpha}\right) = \sqrt{n}\left(\overline{\boldsymbol{\alpha}} - \boldsymbol{\alpha}\right) + \left(n^{-1}\mathbf{Z}'\mathbf{Z}\right)^{-1} n^{-1/2}\mathbf{Z}'\boldsymbol{\phi} \xrightarrow{d} N\left(\mathbf{0}, \mathbf{V}_{\boldsymbol{\alpha}}\right). \tag{13}$$

Thus the fact that $\eta_i$ is replaced with $\widehat{\eta}_i$ is asymptotically irrelevant.[18] We call (12) 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

$$\widetilde{\sigma}_i^2 = \widetilde{\boldsymbol{\alpha}}'\mathbf{z}_i. \tag{14}$$

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

$$\widetilde{\mathbf{D}} = \operatorname{diag}\left\{\widetilde{\sigma}_1^2, \cdots, \widetilde{\sigma}_n^2\right\}$$

[18] Such a result appears as early as in Amemiya (1977).

and

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

This is the feasible GLS, or FGLS, estimator of $\boldsymbol{\beta}$. In practice, we can iterate between $\widetilde{\mathbf{D}}$ and $\widetilde{\boldsymbol{\beta}}$ 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 $\overline{\boldsymbol{\beta}}$.

One typical problem with implementation of FGLS estimation is that in a linear regression specification, there is no guarantee that $\widetilde{\sigma}_i^2 > 0$ for all $i$. If $\widetilde{\sigma}_i^2 < 0$ for some $i$, then the FGLS estimator is not well defined. Furthermore, if $\widetilde{\sigma}_i^2 \approx 0$ 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:

$$\overline{\sigma}_i^2 = \max\left\{\widetilde{\sigma}_i^2, \underline{\sigma}^2\right\}$$

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

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

*and thus*

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

*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 $\widetilde{\boldsymbol{\beta}}$ is

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

which is consistent for $\mathbf{V}$ as $n \to \infty$. This estimator $\widetilde{\mathbf{V}}^0$ is appropriate when the skedastic regression (10) is correctly specified.

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

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

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 $\sigma_i^2 = \boldsymbol{\alpha}'\mathbf{z}_i$, $\widetilde{\mathbf{V}}^0$ is inconsistent for $\mathbf{V}$. An appropriate solution is to use a White-type estimator in place of $\widetilde{\mathbf{V}}^0$. This may be written as

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

where $\widehat{\mathbf{D}} = \operatorname{diag}\left\{\widehat{u}_1^2, \cdots, \widehat{u}_n^2\right\}$. 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 perfoms 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.

# 8 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 resources, 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 (10). We may therefore test this hypothesis by the estimation (12) and constructing a Wald statistic.

This hypothesis does not imply that $\xi_i$ 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 $\xi_i$ is independent of $\mathbf{x}_i$ and the asymptotic variance (11) for $\widetilde{\boldsymbol{\alpha}}$ simplifies to

$$\mathbf{V}_{\boldsymbol{\alpha}} = E\left[\mathbf{z}_i\mathbf{z}_i'\right]^{-1} E\left[\xi_i^2\right]. \tag{15}$$

Hence the standard test of $H_0$ is a classic $F$ (or Wald) test for exclusion of all regressors from the skedastic regression (12). The asymptotic distribution (13) and the asymptotic variance (15) 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 $\chi_q^2$.*

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 $N(0, h(\boldsymbol{\alpha}'\mathbf{z}_i))$ for a general $h(\cdot) > 0$, and use the LM test to check whether $\boldsymbol{\alpha}_1 = \mathbf{0}$.[19] Because $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are "informationally" independent and $H_0$ involves only $\boldsymbol{\alpha}$, the LM test statistic can be much simplified. It turns out that

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

where $f_i = \widehat{u}_i^2 - \widehat{\sigma}^2$. This LM test statistic is similar to the Wald test statistic; a key difference is that $\widehat{E}\left[\xi_i^2\right]$ is replaced by $2\widehat{\sigma}^4$ which is a consistent estimator of $E\left[\xi_i^2\right]$ 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 $2\widehat{\sigma}^4$ to achieve the correct size. We denote the resulting LM test statistic as $LM_K$.

White (1980c) observes that when $H_0$ holds, $\widehat{\boldsymbol{\Omega}} = n^{-1}\sum_{i=1}^n \mathbf{x}_i\mathbf{x}_i'\widehat{u}_i^2$ and $\widehat{\sigma}^2\widehat{\mathbf{Q}} = \left(n^{-1}\sum_{i=1}^n \widehat{u}_i^2\right)\left(n^{-1}\sum_{i=1}^n \mathbf{x}_i\mathbf{x}_i'\right)$ 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 $\mathbf{D}_n = n^{-1}\sum_{i=1}^n \mathbf{z}_{1i}\left(\widehat{u}_i^2 - \widehat{\sigma}^2\right) \approx \mathbf{0}$. Under the

[19] The 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.

auxiliary assumption that $u_i$ is independent of $\mathbf{x}_i$, we can show that under $H_0$,

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

where

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

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

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

**Exercise 12** *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*

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

*If* $x_i = 1$ *for* $i = 1, \cdots, n_1$ *and* $x_i = 0$ *for* $i = n_1 + 1, \cdots, n$, *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 13** *(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 causes 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 $\chi_2^2$ 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)

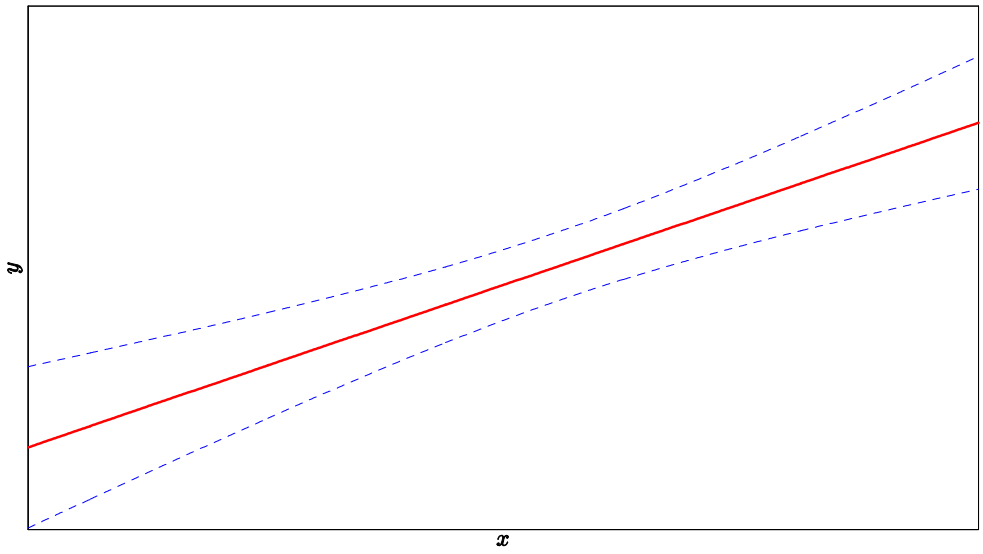


Figure 5: Typical Regression Intervals

where the testing idea follows from Zheng (1996). (**)

# 9 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.

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 $\beta$. 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 14** *In the wage equation,* $\log wage = \beta_1 + \beta_2 \cdot educ + \beta_3 \cdot exper + \beta_3 \cdot exper^2 + u$, *how*

*to construct regression intervals for experience when education is fixed at its mean?*

**Exercise 15** *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* $\exp\left\{\widehat{\log y}\right\}$ *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.[20] 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

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

Assuming $E\left[u_i^2|\mathbf{x}_i\right] = \sigma^2$, 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 (14), 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 16** *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* $\sigma^2$*. You wish to forecast an out-of-sample value of* $y_{n+1}$ *given that* $\mathbf{x}_{n+1} = \mathbf{x}$*.*

[20] $\mathbf{x}$ cannot be the same as any $\mathbf{x}_i$ observed, why?

*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.*

**Exercise 17 (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*

$$\log TC_i = \beta_1 + \beta_2 \log Q_i + \beta_3 \log PL_i + \beta_4 \log PK_i + \beta_5 \log PF_i + u_i. \tag{16}$$

**(a)** *Following Nerlove, add the variable* $(logQ_i)^2$ *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 (16) 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* $\beta_7$ *is selected so that several values (in this example, at least 10 to 15) of* $logQ_i$ *are both below and above* $\beta_7$*. Examine the data and pick an appropriate range for* $\beta_7$*.*

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

**(d)** *Calculate standard errors for all the parameters* $(\beta_1, \cdots, \beta_7)$*.*

**Exercise 18 (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.*

## 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). The AIC is used to select the model whose estimated density is closest to the true density. It is designed for parametric models estimated by maximum likelihood.

Recall from Chapter 4 that the **Kullback-Leibler information distance** (or the **relative entropy**[21]) 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.[22] The AIC is just an estimate of $D$ when $q$ is the true density while $p$ is an approximation of $q$. The rough idea is that we estimate $D(q,p)$ by $D(q,\widehat{p})$, where $\widehat{p}$ is an estimator of $p$. Since $D(q,\widehat{p})$ is random, we take expectation of $D(q,\widehat{p})$ to get

$$\begin{aligned} E\left[D(q,\widehat{p})\right] &= \int q(x) \log q(x) dx - E\left[q(x) \int \log \widehat{p}(x) dx\right] \\ &= C - E\left[\log \widehat{p}(\widetilde{x})\right], \end{aligned}$$

where $\widetilde{x}$ is an independent copy of $y$. In other words, $E\left[D(q,\widehat{p})\right]$ is the expected log-likelihood fit using the estimated model $\widehat{p}$ of an out-of-sample realization $\widetilde{x}$; thus, $E\left[D(q,\widehat{p})\right]$ can be interpreted as an expected predictive log likelihood.

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

[22] $D$ 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)).

Suppose we want to pick up some $p_\theta$ 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 $(d/d\theta) \int \log \left( q(x) / p_{\bar{\theta}}(x) \right) q(x)\, dx = 0$. By the second order Taylor approximation around $\bar{\theta}$,

$$\begin{aligned} D(q, p_\theta) &\approx D(q, p_{\bar{\theta}}) + \frac{1}{2} \left(\theta - \bar{\theta}\right)' H(p_{\bar{\theta}}) \left(\theta - \bar{\theta}\right) \\ &= \int [\log q(x)]\, q(x)\, dx - \int [\log p_{\bar{\theta}}(x)]\, q(x)\, dx + \frac{1}{2} \left(\theta - \bar{\theta}\right)' H(p_{\bar{\theta}}) \left(\theta - \bar{\theta}\right) \end{aligned} \tag{17}$$

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

We consider the minimization of the RHS in (17) with respect to $\theta$. The first term of the RHS in (17) is independent of $\theta$ and $\bar{\theta}$, so can be ignored. The second term is independent of $\theta$, 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 estimated by

$$-\frac{1}{n} \sum_{i=1}^{n} \log p_{\bar{\theta}}(X_i). \tag{18}$$

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

$$-\frac{1}{n} \sum_{i=1}^{n} \log p_{\bar{\theta}}(X_i) \approx -\frac{1}{n} \sum_{i=1}^{n} \log p_{\tilde{\theta}}(X_i) - \frac{1}{2} \left(\tilde{\theta} - \bar{\theta}\right)' \frac{1}{n} \sum_{i=1}^{n} \left. \frac{\partial^2 \log p_\theta(X_i)}{\partial\theta\partial\theta'} \right|_{\theta=\tilde{\theta}} \left(\tilde{\theta} - \bar{\theta}\right) \tag{19}$$

since $\dfrac{1}{n} \sum_{i=1}^{n} (\partial/\partial\theta) \log p_\theta(X_i)|_{\theta=\tilde{\theta}} = 0$.[23] Only the third term $\dfrac{1}{2} \left(\theta - \bar{\theta}\right)' H(p_{\bar{\theta}}) \left(\theta - \bar{\theta}\right)$ depends on $\theta$, 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

$$-\frac{1}{2} \left(\tilde{\theta} - \bar{\theta}\right)' \frac{1}{n} \sum_{i=1}^{n} \left. \frac{\partial^2 \log p_\theta(X_i)}{\partial\theta\partial\theta'} \right|_{\theta=\tilde{\theta}} \left(\tilde{\theta} - \bar{\theta}\right) \tag{20}$$

[23] A naive estimator of $-\int \log [p_{\bar{\theta}}(x)]\, q(x)\, dx$ 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 $\theta$). So, (19) can be interpreted as a bias corrected estimator.

In summary, we have

$$\begin{aligned}\min_{\theta} D\left(q, p_{\theta}\right) &= D\left(q, p_{\bar{\theta}}\right) \\ &\approx -\frac{1}{n}\sum_{i=1}^{n}\log p_{\tilde{\theta}}\left(X_i\right) - \left(\tilde{\theta}-\bar{\theta}\right)' \frac{1}{n}\sum_{i=1}^{n}\left.\frac{\partial^2 \log p_{\theta}\left(X_i\right)}{\partial\theta\partial\theta'}\right|_{\theta=\tilde{\theta}}\left(\tilde{\theta}-\bar{\theta}\right) + C, \quad (21)\end{aligned}$$

where $C$ is some constant. Compared to the straightforward estimator $-\frac{1}{n}\sum_{i=1}^{n}\log p_{\tilde{\theta}}\left(X_i\right)$ of $-\int\left[\log p_{\bar{\theta}}\left(x\right)\right]q\left(x\right)dx$, the second term of (21) is a penalty term. It comes from two resources - one half is from in-sample over-fitting in (19) and another half is from parameter estimation in (20). The RHS is random (and involves the unknown $\bar{\theta}$), so we take expectation to further simplify the formula.

Taking the expectation of the both sides of (21), we have

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

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

$$\sqrt{n}\left[\tilde{\theta}-\bar{\theta}\right] \xrightarrow{d} N\left(0, H\left(p_{\bar{\theta}}\right)^{-1} J\left(p_{\bar{\theta}}\right) H\left(p_{\bar{\theta}}\right)^{-1}\right) \quad (22)$$

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

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

we have

$$-E\left[\left(\tilde{\theta}-\bar{\theta}\right)' \frac{1}{n}\sum_{i=1}^{n}\left.\frac{\partial^2 \log p_{\theta}\left(X_i\right)}{\partial\theta\partial\theta'}\right|_{\theta=\tilde{\theta}}\left(\tilde{\theta}-\bar{\theta}\right)\right] \approx E\left[\left(\tilde{\theta}-\bar{\theta}\right)' H\left(p_{\bar{\theta}}\right)\left(\tilde{\theta}-\bar{\theta}\right)\right]. \quad (23)$$

It holds

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

From this and (22), we have

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

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

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

Therefore,

$$-E\left[\left(\tilde{\theta}-\bar{\theta}\right)' \frac{1}{n}\sum_{i=1}^{n} \left.\frac{\partial^2 \log p_\theta\left(X_i\right)}{\partial\theta\partial\theta'}\right|_{\theta=\tilde{\theta}} \left(\tilde{\theta}-\bar{\theta}\right)\right] \approx \frac{k}{n} \tag{24}$$

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

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

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

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

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

From the proof above, there are some errors in approximating $D\left(q, p_{\bar{\theta}}\right)$ by the AIC. First, we use a quadratic approximation to the log likelihood. Second, we use a normal approximation to the distribution of $\tilde{\theta}$. Third, we have assumed that the true model is in the choice set to apply the information equality although the target of AIC is to choose the best approximation model rather than select the true model.

## 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,

$$B_{12} = \frac{p\left(\mathcal{M}_1|y\right)/p(\mathcal{M}_2|y)}{p(\mathcal{M}_1)/p(\mathcal{M}_2)} = \frac{p\left(y|\mathcal{M}_1\right)p(\mathcal{M}_1)/p(y)}{p\left(y|\mathcal{M}_2\right)p(\mathcal{M}_2)/p(y)} \Big/ \frac{p(\mathcal{M}_1)}{p(\mathcal{M}_2)} = \frac{p(y|\mathcal{M}_1)}{p\left(y|\mathcal{M}_2\right)} = \frac{\int p(y|\theta_1,\mathcal{M}_1)\cdot p(\theta_1|\mathcal{M}_1)d\theta_1}{\int p(y|\theta_2,\mathcal{M}_2)\cdot p(\theta_2|\mathcal{M}_2)d\theta_2}. \tag{25}$$

Here, $p(\mathcal{M}_i)$ is the prior probability of model $i$, $p\left(\mathcal{M}_i|y\right)$ is the posterior probability of model $i$,

$p(y|\mathcal{M}_i)$ is the **marginal likelihood** of $y$ in $\mathcal{M}_i$,[24] $i = 1, 2$, and $y$ is the data set. Intuitively, if $B_{12} > 1$, i.e., $\mathcal{M}_1$ can explain the data better than $\mathcal{M}_2$, then we prefer model 1. If $p(\mathcal{M}_1) = p(\mathcal{M}_2)$, then $B_{12}$ is actually the posterior odds, so we are selecting the model with the highest approximate probability of being the true model.

**Laplace expansion** is used to approximate some integrals like the numerator and denominator of equation (25) by normal density functions. The procedure is as follows,

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

where $k_i$ is the dimension of $\theta_i$, $\widehat{\theta}_i$ is the minima of $h(\theta_i)$,[25] $H$ is the Hessian matrix of $h$, and $L_i(\widehat{\theta}_i)$ 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 equailty is from the definition of the density function of mutivariate nomral. So we could approximate equation (25) as

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

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

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

where $\lambda_n$ is the standard likelihood ratio for the comparison of $\mathcal{M}_1$ with $\mathcal{M}_2$, $\lambda_n \equiv L_{1,n}(\widehat{\theta}_{1,n})/L_{2,n}(\widehat{\theta}_{2,n})$, and $R(\widehat{\theta}_{1,n}, \widehat{\theta}_{2,n})$ 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 $R(\widehat{\theta}_{1,n}, \widehat{\theta}_{2,n})$ is negligible compared with both other terms (i.e., $R(\widehat{\theta}_{1,n}, \widehat{\theta}_{2,n}) = O(1)$). Obviously, when $BIC > 0$, we should select model 1.

[24] "marginal" here means that $\theta_i$ is integrated out.

[25] Note 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 $\widehat{\theta}_i$ as the MLE.

Applying this criterion to the normal regression model, we have

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

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