# Additional Topics on Linear Regression

Ping Yu

School of Economics and Finance
The University of Hong Kong

# Review of Our Assumptions

- **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): $E[y|\mathbf{x}] = \mathbf{x}'\beta$.
- **Assumption OLS.2′**: $y = \mathbf{x}'\beta + u$ 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 $E\left[\|\mathbf{x}\|^4\right] < \infty$.

| | $y = \mathbf{x}'\beta + u$ | | Implied Properties |
|---|---|---|---|
| $E[\mathbf{x}u] = \mathbf{0}$ | linear projection | $\Longrightarrow$ | consistency |
| | $\cup$ | | |
| $E[u\|\mathbf{x}] = 0$ | linear regression | $\Longrightarrow$ | unbiasedness |
| | $\cup$ | | |
| $E[u\|\mathbf{x}] = 0$<br>$E[u^2\|\mathbf{x}] = \sigma^2$ | homoskedastic<br>linear regression | $\Longrightarrow$ | Gauss-Markov Theorem |

Table 1: Relationship between Different Models

# Our Plan of This Chapter

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

# Tests for Functional Form Misspecification

# Ramsey's REgression Specification Error Test (RESET)

- Misspecification of $E[y|\mathbf{x}]$ may be due to omitted variables or misspecified functional forms. We only examine the second source of misspecification here.
- A Straightforward Test: add nonlinear functions of the regressors to the regression, and test their significance using a Wald test.
  - fit $y_i = \mathbf{x}_i'\widetilde{\beta} + \mathbf{z}_i'\widetilde{\gamma} + \widetilde{u}_i$ by OLS, and form a Wald statistic for $\gamma = \mathbf{0}$, where $\mathbf{z}_i = h(\mathbf{x}_i)$ denote functions of $\mathbf{x}_i$ which are not linear functions of $\mathbf{x}_i$ (perhaps squares of non-binary regressors).
- RESET: The null model is

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

Let

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

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

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

by OLS, and form the Wald statistic $W_n$ for $\gamma = \mathbf{0}$.

# continue...

- 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 is particularly powerful in detecting the single-index model,

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

  where $G(\cdot)$ is a smooth "link" function.
- Why? Note that (1) may be written as

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

  which has essentially approximated $G(\cdot)$ by an $m$th order polynomial.
- Other tests: Hausman test, Conditional Moment (CM) test, Zheng's score test.

# Nonlinear Least Squares

# Nonlinear Least Squares

- If the specification test rejects the linear specification, we may consider to use a nonlinear setup for $E[y|\mathbf{x}]$.
- Suppose $E[y_i|\mathbf{x}_i=\mathbf{x}]=m(\mathbf{x}|\theta)$. Nonlinear regression means that $m(\mathbf{x}|\theta)$ is a nonlinear function of $\theta$ (rather than $\mathbf{x}$).
- The functional form of $m(\mathbf{x}|\theta)$ 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.
- $m(\mathbf{x}|\theta)=\exp(\mathbf{x}'\theta)$: Exponential Link Regression
  - The exponential link function is strictly positive, so this choice can be useful when it is desired to constrain the mean to be strictly positive.
- $m(x|\theta)=\theta_1+\theta_2 x^{\theta_3}$, $x>0$: 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.

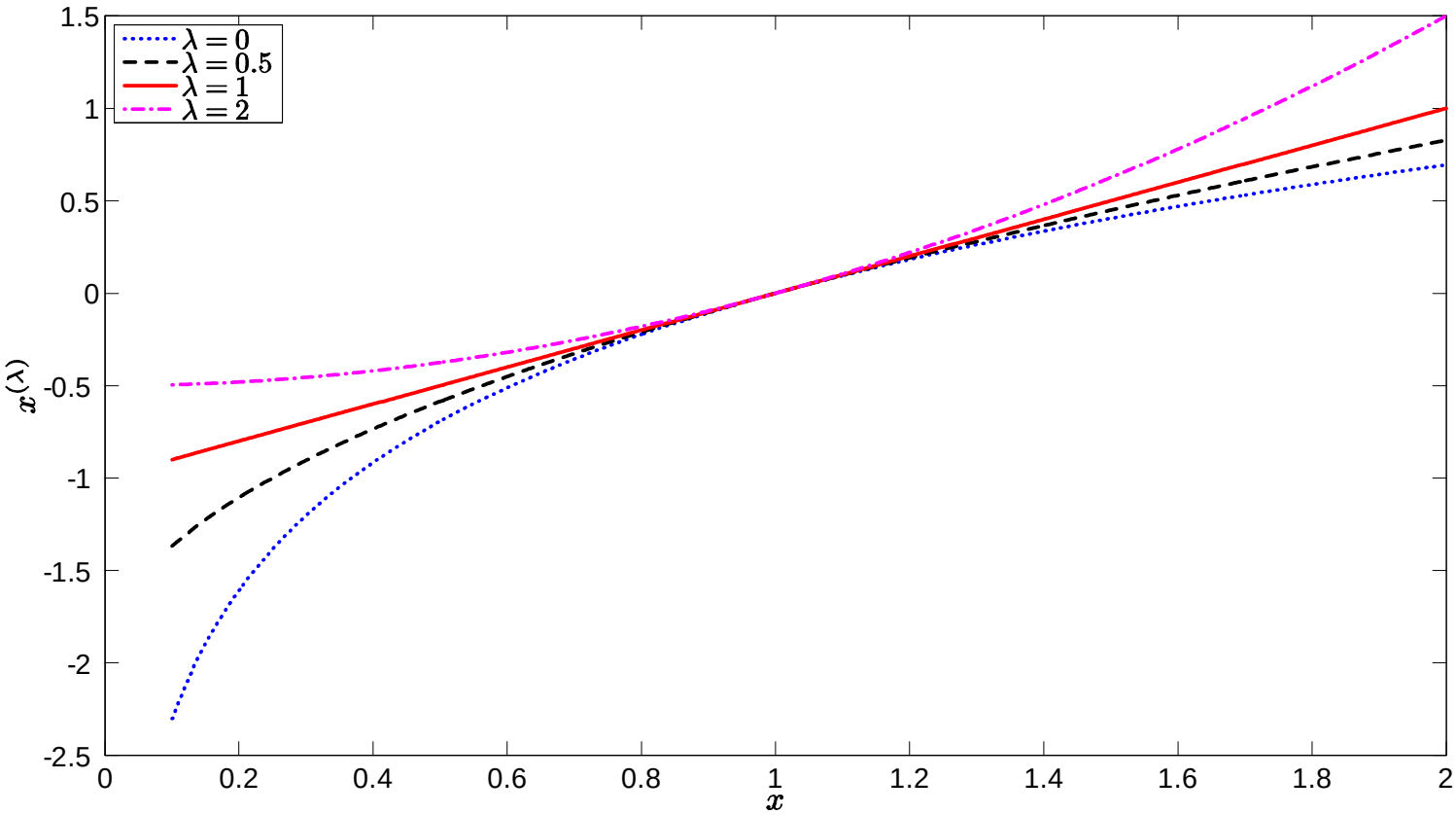


Figure: Box-Cox Transformation for Different $\lambda$ Values: all pass $(1, 0)$

# continue...

- $m(x|\theta) = \theta_1 + \theta_2 \exp(\theta_3 x)$: Exponentially Transformed Regressors
- $m(\mathbf{x}|\theta) = G(\mathbf{x}'\theta)$, $G$ known
  - When $G(\cdot) = (1+\exp(-\cdot))^{-1}$, the regression with $m(\mathbf{x}|\theta) = G(\mathbf{x}'\theta)$ 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**.
- $m(\mathbf{x}|\theta) = \theta_1'\mathbf{x}_1 + \theta_2'\mathbf{x}_1 G\left(\frac{x_2-\theta_3}{\theta_4}\right)$: Smooth Transition
- $m(x|\theta) = \theta_1 + \theta_2 x + \theta_3 (x-\theta_4) 1(x > \theta_4)$: Continuous Threshold Regression
- $m(\mathbf{x}|\theta) = (\theta_1'\mathbf{x}_1) 1(x_2 \le \theta_3) + (\theta_2'\mathbf{x}_1) 1(x_2 > \theta_3)$: Threshold Regression
  - When $\theta_4 = 0$, the smooth transition model (STM) reduces to the threshold regression (TR) model.
  - When $\theta_4 = \infty$, the STM reduces to linear regression.
  - For CTR and TR, $m(\mathbf{x}|\theta)$ is not smooth in $\theta$.

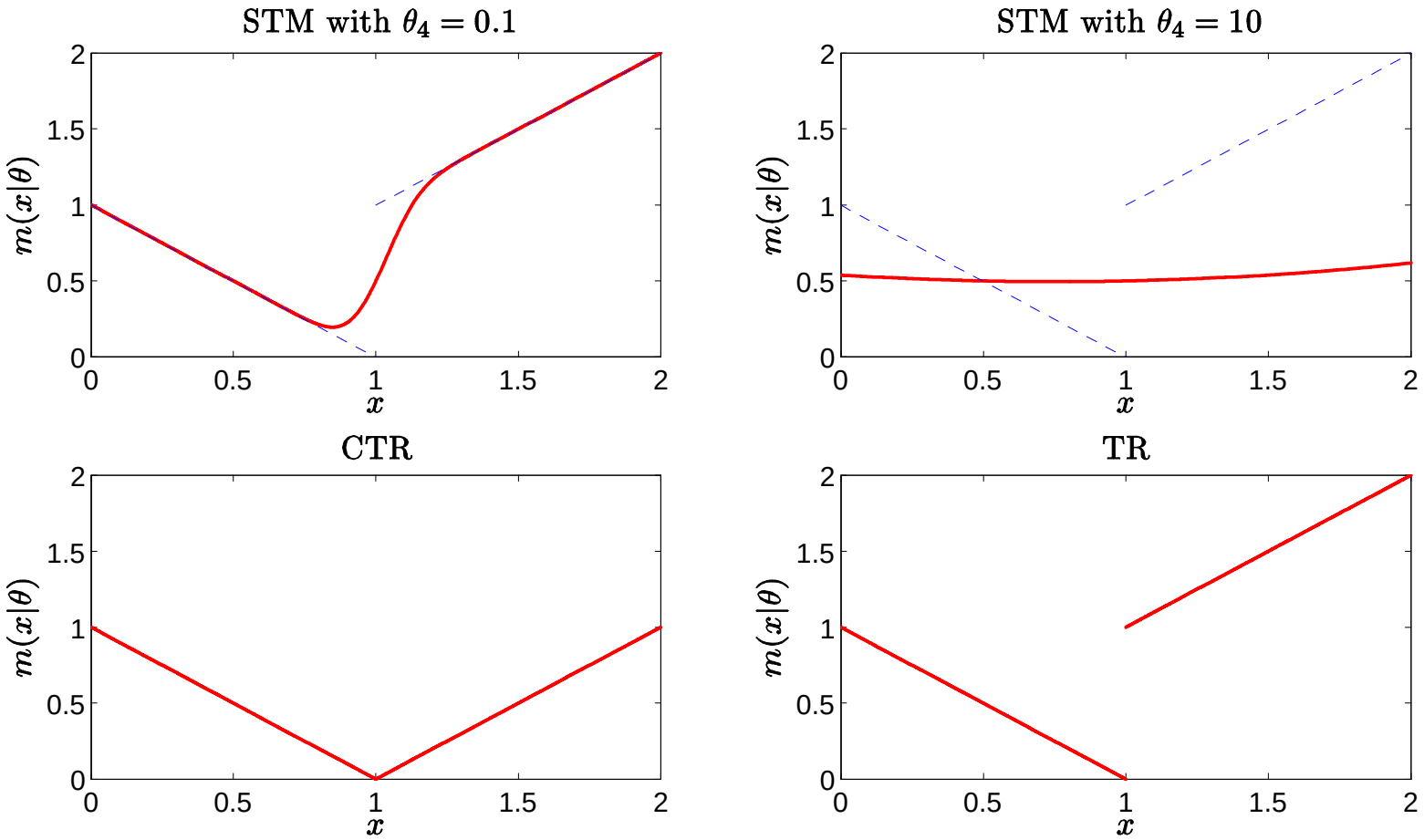


Figure: Difference Between STM, CTR and TR

# The NLLS Estimator

- The nonlinear least squares (NLLS) estimator is

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

- The FOCs for minimization are $\mathbf{0} = \sum_{i=1}^{n} \mathbf{m}_\theta(\mathbf{x}_i|\widehat{\theta})\widehat{u}_i$ with $\mathbf{m}_\theta(\mathbf{x}|\theta) = \frac{\partial}{\partial\theta} m(\mathbf{x}|\theta)$.

**Theorem**

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

$$\sqrt{n}\left(\widehat{\theta} - \theta_0\right) \xrightarrow{d} N(\mathbf{0}, \mathbf{V}),$$

*where* $\mathbf{V} = E\left[\mathbf{m}_{\theta i}\mathbf{m}_{\theta i}'\right]^{-1} E\left[\mathbf{m}_{\theta i}\mathbf{m}_{\theta i}' u_i^2\right] E\left[\mathbf{m}_{\theta i}\mathbf{m}_{\theta i}'\right]^{-1}$ *with* $\mathbf{m}_{\theta i} = \mathbf{m}_\theta(\mathbf{x}_i|\theta_0)$.

- $\mathbf{V}$ can be estimated by

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

  where $\widehat{\mathbf{m}}_{\theta i} = \mathbf{m}_\theta(\mathbf{x}_i|\widehat{\theta})$, and $\widehat{u}_i = y_i - m(\mathbf{x}_i|\widehat{\theta})$.

# Omitted and Irrelevant Variables

# Omitted Variables

- The true model is a **long regression**

$$y_i = \mathbf{x}_{1i}'\beta_1 + \mathbf{x}_{2i}'\beta_2 + u_i, E[\mathbf{x}_i u_i] = \mathbf{0}, \tag{2}$$

but we estimate a **short regression**

$$y_i = \mathbf{x}_{1i}'\gamma_1 + v_i, E[\mathbf{x}_{1i} v_i] = \mathbf{0}. \tag{3}$$

- $\beta_1 \neq \gamma_1$. Why?

$$\begin{aligned}
\gamma_1 &= E\left[\mathbf{x}_{1i}\mathbf{x}_{1i}'\right]^{-1} E[\mathbf{x}_{1i} y_i] \\
&= E\left[\mathbf{x}_{1i}\mathbf{x}_{1i}'\right]^{-1} E\left[\mathbf{x}_{1i}\left(\mathbf{x}_{1i}'\beta_1 + \mathbf{x}_{2i}'\beta_2 + u_i\right)\right] \\
&= \beta_1 + E\left[\mathbf{x}_{1i}\mathbf{x}_{1i}'\right]^{-1} E[\mathbf{x}_{1i}\mathbf{x}_{2i}']\beta_2 \\
&= \beta_1 + \Gamma\beta_2,
\end{aligned}$$

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

# Direct and Indirect Effects of $\mathbf{x}_1$ on $y$

- $\gamma_1 \neq \beta_1$ unless $\Gamma = \mathbf{0}$ or $\beta_2 = \mathbf{0}$. The former means that the regression of $\mathbf{x}_{2i}$ on $\mathbf{x}_{1i}$ yields a set of zero coefficients (they are uncorrelated), and the latter means that the coefficient on $\mathbf{x}_{2i}$ in (2) is zero.
- The difference $\Gamma\beta_2$ is known as **omitted variable bias**.
- $\gamma_1$ includes both the direct effect of $\mathbf{x}_1$ on $y$ ($\beta_1$) and the indirect effect ($\Gamma\beta_2$) through $\mathbf{x}_2$. [Figure here]
- To avoid omitted variable bias the standard advice is to include potentially relevant variables in the estimated model.
- But 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.

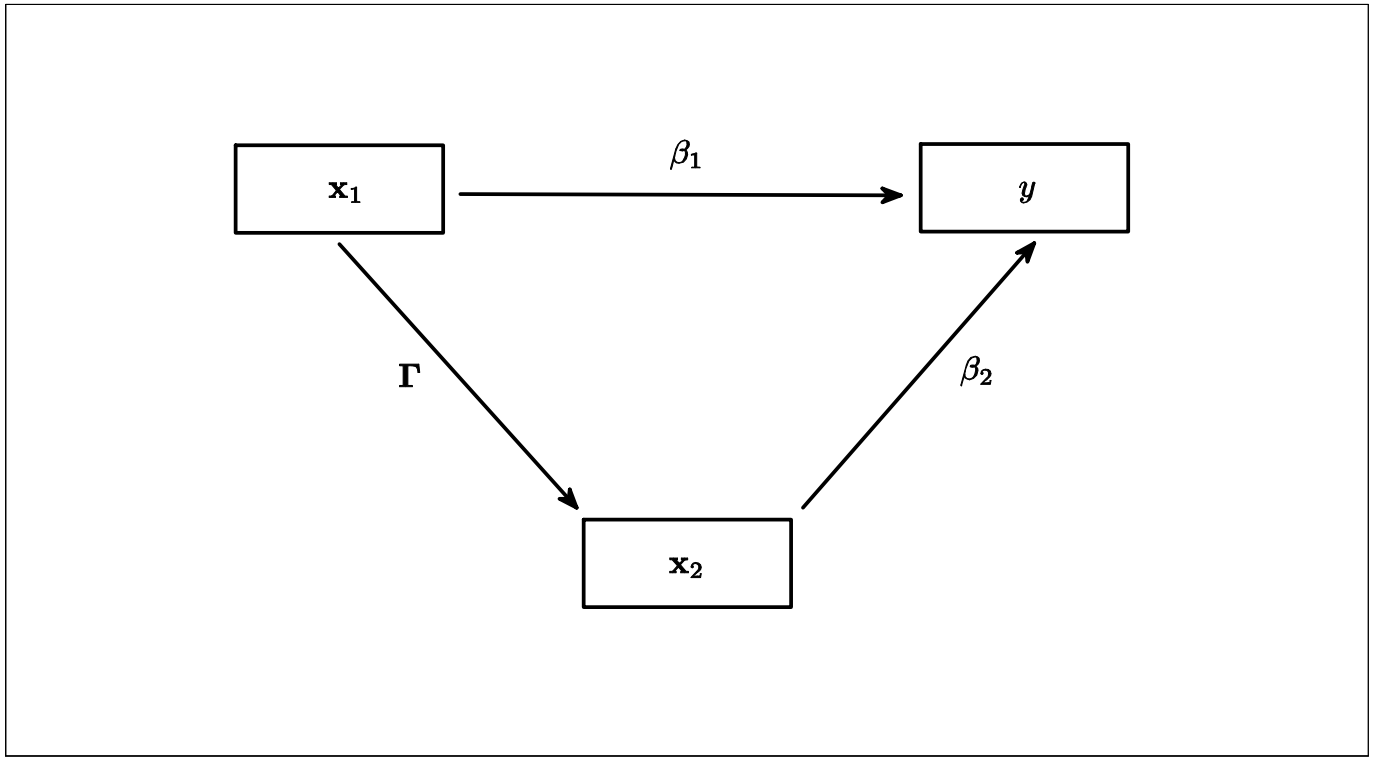


Figure: Direct and Indirect Effects of $\mathbf{x}_1$ on $y$

# Irrelevant Variables

- When $\beta_2 = \mathbf{0}$ and $\beta_1$ is the parameter of interest, $\mathbf{x}_{2i}$ is "irrelevant".
- In this case, the estimator of $\beta_1$ from the short regression, $\overline{\beta}_1 = (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{y}$, is consistent from the analysis above.
- Efficiency comparison under homoskedasticity:

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

and

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

- If $\mathbf{Q}_{12} = E\left[\mathbf{x}_{1i}\mathbf{x}_{2i}'\right] = \mathbf{0}$ (so the variables are orthogonal) then 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 $\overline{\beta}_1$ has a lower asymptotic variance matrix than $\widehat{\beta}_1$ if the irrelevant variables are correlated with the relevant variables.

# Intuition

- The irrelevant variable does not provide information for $y_i$, but introduces multicollinearity to the system, so decreases the denominator of $AVar\left(\widehat{\beta}_1\right)$ from $\mathbf{Q}_{11}$ to $\mathbf{Q}_{11.2}$ without decreasing its numerator $\sigma^2$.
- 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.).
- Then we can show under homoskedasticity,

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

while

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

where $E[x_i] = \mu$, and $Var(x_i) = \sigma_x^2$.

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

# When $\beta_2 \neq \mathbf{0}$ and $\mathbf{Q}_{12} = \mathbf{0}$

- If $\mathbf{x}_2$ is relevant ($\beta_2 \neq \mathbf{0}$) and uncorrelated with $\mathbf{x}_1$, then it decreases the numerator without affecting the denominator of $AVar\left(\widehat{\beta}_1\right)$, 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.

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

Table 2: Consistency and Efficiency with Omitted and Irrelevant Variables

# The Heteroskedastic Case

- From the last chapter, we know that when the model is heteroskedastic, it is possible that $\widehat{\beta}_1$ is more efficient than $\overline{\beta}_1$ $(=\widehat{\beta}_{1R})$ even if $\mathbf{x}_2$ is irrelevant, or adding irrelevant variables can actually decrease the estimation variance.
- This result 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.

# Model Selection

# Introduction

- 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?
- 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.
- 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.

# Setup

- 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

$$\mathbf{y} = \mathbf{X}_1\beta_1 + \mathbf{X}_2\beta_2 + \mathbf{u},$$

  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

$$\begin{aligned} \mathscr{M}_1 &: \mathbf{y} = \mathbf{X}_1\beta_1 + \mathbf{u}, & E[\mathbf{u}|\mathbf{X}_1, \mathbf{X}_2] &= \mathbf{0}, \\ \mathscr{M}_2 &: \mathbf{y} = \mathbf{X}_1\beta_1 + \mathbf{X}_2\beta_2 + \mathbf{u}, & E[\mathbf{u}|\mathbf{X}_1, \mathbf{X}_2] &= \mathbf{0}. \end{aligned}$$

- Note that $\mathscr{M}_1 \subset \mathscr{M}_2$. To be concrete, we say that $\mathscr{M}_2$ is true if $\beta_2 \neq \mathbf{0}$.
- Notations: OLS residual vectors $\widehat{\mathbf{u}}_1$ and $\widehat{\mathbf{u}}_2$, estimated variances $\widehat{\sigma}_1^2$ and $\widehat{\sigma}_2^2$, etc. for Model 1 and 2.
- For simplicity, use the homoskedasticity assumption $E[u_i^2|\mathbf{x}_{1i}, \mathbf{x}_{2i}] = \sigma^2$.

# Consistent Model Selection Procedure

- A model selection procedure is a data-dependent rule which selects one of the two models. We can write this as $\widehat{\mathscr{M}}$.
- A model selection procedure is **consistent** if

$$P\left(\widehat{\mathscr{M}} = \mathscr{M}_1 | \mathscr{M}_1\right) \quad \to \quad 1,$$
$$P\left(\widehat{\mathscr{M}} = \mathscr{M}_2 | \mathscr{M}_2\right) \quad \to \quad 1.$$

- Selection Based on the Wald $W_n$: For some significance level $\alpha$, let $c_\alpha$ satisfy $P\left(\chi^2_{k_2} > c_\alpha\right) = \alpha$. Then select $\mathscr{M}_1$ if $W_n \leq c_\alpha$, else select $\mathscr{M}_2$.
- **Inconsistency**: If $\alpha$ is held fixed, then $P(\widehat{\mathscr{M}} = \mathscr{M}_1 | \mathscr{M}_1) \to 1 - \alpha < 1$ although $P(\widehat{\mathscr{M}} = \mathscr{M}_2 | \mathscr{M}_2) = 1$.

# Information Criterion

- **Intuition**: 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.
- Most popular information criteria (IC) include the Akaike Information Criterion (**AIC**) and the Bayes Information Criterion (**BIC**). The former is inconsistent while the latter is consistent.
- The AIC under normality for model $m$ is

$$AIC_m = \log\left(\widehat{\sigma}_m^2\right) + 2\frac{k_m}{n}, \tag{4}$$

where $\widehat{\sigma}_m^2$ is the variance estimate for model $m$ and is roughly $-2l_n$ (neglecting the constant term) with $l_n$ being the average log-likelihood function, and $k_m$ is the number of coefficients in the model. The rule is to select $\mathscr{M}_1$ if $AIC_1 < AIC_2$, else select $\mathscr{M}_2$.

# Inconsistency of the AIC

- **Inconsistency**: The AIC tends to overfit. Why? Under $\mathscr{M}_1$,

$$LR = n\left(\log\left(\widehat{\sigma}_1^2\right) - \log\left(\widehat{\sigma}_2^2\right)\right) \xrightarrow{d} \chi^2_{k_2}, \tag{5}$$

so

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

# Consistency of the BIC

- The BIC is based on
$$BIC_m = \log\left(\widehat{\sigma}_m^2\right) + \log n \frac{k_m}{n}. \tag{6}$$
- Since $\log(n) > 2$ (if $n > 8$), the BIC places a larger penalty than the AIC on the number of estimated parameters and is more parsimonious.
- **Consistency**: Because (5) holds under $\mathscr{M}_1$,
$$\frac{LR}{\log(n)} \xrightarrow{p} 0,$$
so
$$\begin{aligned} P\left(\widehat{\mathscr{M}} = \mathscr{M}_1 | \mathscr{M}_1\right) &= P\left(BIC_1 < BIC_2 | \mathscr{M}_1\right) = P\left(LR < \log(n) k_2 | \mathscr{M}_1\right) \\ &= P\left(\frac{LR}{\log(n)} < k_2 | \mathscr{M}_1\right) \to P\left(0 < k_2\right) = 1. \end{aligned}$$
Also under $\mathscr{M}_2$, one can show that
$$\frac{LR}{\log(n)} \xrightarrow{p} \infty,$$
thus
$$P\left(\widehat{\mathscr{M}} = \mathscr{M}_2 | \mathscr{M}_2\right) = P\left(\frac{LR_n}{\log(n)} > k_2 \middle| \mathscr{M}_2\right) \to 1.$$

# Difference between the AIC and BIC

- Essentially, to consistently select $\mathscr{M}_1$, 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 $\mathscr{M}_2$, the penalty must be $o(n)$ so that $LR$ divided by the penalty converges in probability to infinity under $\mathscr{M}_2$.
- 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 the AIC tends to overfit, this need not be a defect of the AIC. The AIC is derived as an estimate of the Kullback-Leibler information distance $KLIC(\mathscr{M}) = E\left[\log f(\mathbf{y}|\mathbf{X}) - \log f(\mathbf{y}|\mathbf{X},\mathscr{M})\right]$ between the true density and the model density. In other words, the AIC attempts to select a good approximating model for inference. In contrast, the BIC and other IC attempt to estimate the "true" model.

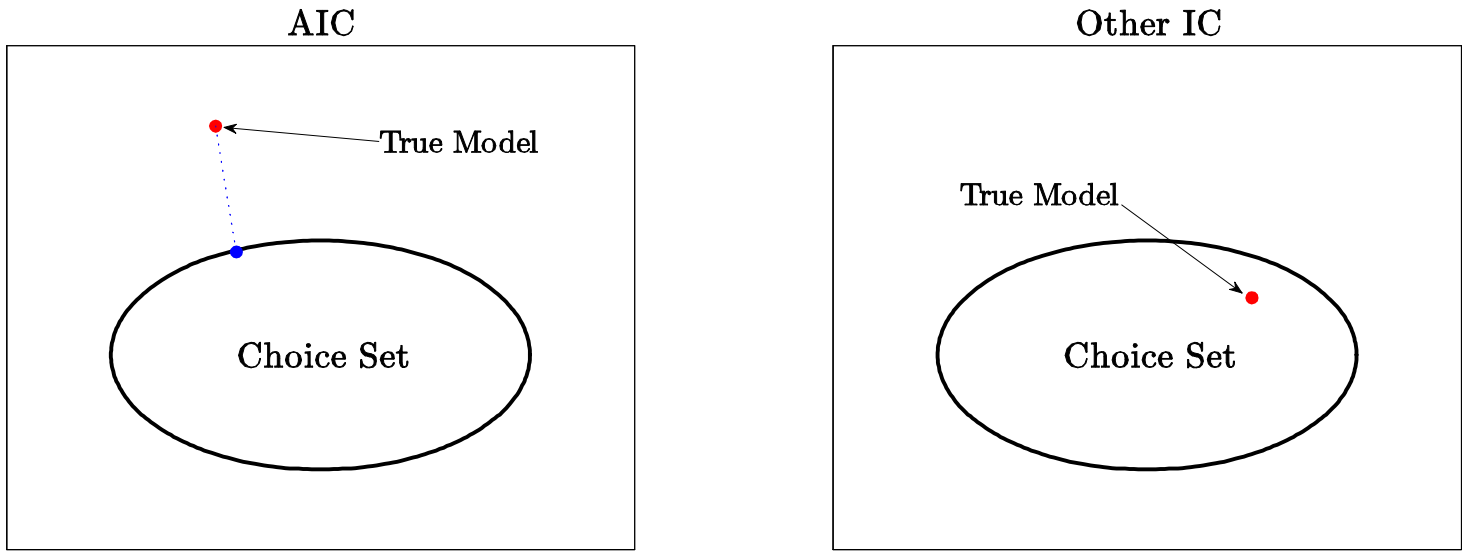


Figure: Difference Between AIC and Other IC

# Ordered and Unordered 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).
- Ordered regressors:

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

  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. Similarly for the BIC.
- Unordered regressors: a model consists of any possible subset of the regressors $\{x_{1i}, \cdots, x_{Ki}\}$, but there are $2^K$ such models, which can be a very large number; e.g., $2^{10} = 1024$, and $2^{20} = 1,048,576$. So this case seems computationally prohibitive.

# Recent Developments

- There is no clear answer as to which IC, if any, should be preferred.
- 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.
- 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.

- Shrinkage: continuously shrink the coefficients rather than discretely select the variables, e.g., LASSO, SCAD and MCP
- Model Averaging

# Generalized Least Squares

# The Weighted Least Squares (WLS) Estimator

- Recall that when the only information is $E[\mathbf{x}_i u_i] = \mathbf{0}$, the LSE is semi-parametrically efficient.
- When extra information $E[u_i|\mathbf{x}_i] = 0$ is available, the LSE is generally inefficient, while the WLS estimator is efficient.
- The WLS estimator:

$$\overline{\beta} = \left(\mathbf{X}'\mathbf{D}^{-1}\mathbf{X}\right)^{-1}\left(\mathbf{X}'\mathbf{D}^{-1}\mathbf{y}\right), \tag{7}$$

  where $\mathbf{D} = \text{diag}\left\{\sigma_1^2, \cdots, \sigma_n^2\right\}$ and $\sigma_i^2 = \sigma^2(\mathbf{x}_i) = E[u_i^2|\mathbf{x}_i]$.

- **Intuition**:

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

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

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

  is homoskedastic (why?). Under homoskedasticity, the Gauss-Markov theorem implies the efficiency of the LSE which is the WLS estimator in the original model.

# The Feasible GLS Estimator

- A feasible GLS (FGLS) estimator replaces the unknown $\mathbf{D}$ with an estimate $\widehat{\mathbf{D}} = \text{diag}\left\{\widehat{\sigma}_1^2, \cdots, \widehat{\sigma}_n^2\right\}$.
- Model the conditional variance using the parametric form

$$\sigma_i^2 = \alpha_0 + \mathbf{z}_{1i}'\alpha_1 = \alpha'\mathbf{z}_i,$$

  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

$$E\left[\eta_i|\mathbf{x}_i\right] = \alpha_0 + \mathbf{z}_{1i}'\alpha_1$$

  and we have the regression equation

$$\begin{aligned} \eta_i &= \alpha_0 + \mathbf{z}_{1i}'\alpha_1 + \xi_i, \\ E\left[\xi_i|\mathbf{x}_i\right] &= 0. \end{aligned} \tag{8}$$

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

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

# Skedastic Regression

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

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

and

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

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

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

$$\begin{aligned}\phi_i &= \widehat{\eta}_i - \eta_i = \widehat{u}_i^2 - u_i^2 = -2u_i\mathbf{x}_i'(\widehat{\beta}-\beta) + (\widehat{\beta}-\beta)'\mathbf{x}_i\mathbf{x}_i'(\widehat{\beta}-\beta),\\ \frac{1}{\sqrt{n}}\sum_{i=1}^n \mathbf{z}_i\phi_i &= -\frac{2}{n}\sum_{i=1}^n \mathbf{z}_i u_i\mathbf{x}_i'\sqrt{n}(\widehat{\beta}-\beta) + \frac{1}{n}\sum_{i=1}^n \mathbf{z}_i(\widehat{\beta}-\beta)'\mathbf{x}_i\mathbf{x}_i'\sqrt{n}(\widehat{\beta}-\beta) \xrightarrow{p} \mathbf{0}.\end{aligned}$$

- Let

$$\widetilde{\alpha} = \left(\mathbf{Z}'\mathbf{Z}\right)^{-1}\mathbf{Z}'\widehat{\eta} \tag{9}$$

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

$$\sqrt{n}\left(\widetilde{\alpha}-\alpha\right) = \sqrt{n}\left(\overline{\alpha}-\alpha\right) + \left(n^{-1}\mathbf{Z}'\mathbf{Z}\right)^{-1} n^{-1/2}\mathbf{Z}'\phi \xrightarrow{d} N\left(\mathbf{0}, \mathbf{V}_\alpha\right). \tag{10}$$

Thus the fact that $\eta_i$ is replaced with $\widehat{\eta}_i$ is asymptotically irrelevant.

- We call (9) the **skedastic regression**, as it is estimating $\sigma^2(\cdot)$.

# Practical Consideration

- Estimate $\sigma_i^2 = \mathbf{z}_i'\alpha$ by
$$\widetilde{\sigma}_i^2 = \widetilde{\alpha}'\mathbf{z}_i. \tag{11}$$
- Suppose that $\widetilde{\sigma}_i^2 > 0$ for all $i$. Then set
$$\widetilde{\mathbf{D}} = \text{diag}\left\{\widetilde{\sigma}_1^2, \cdots, \widetilde{\sigma}_n^2\right\}$$
and the FGLS estimator
$$\widetilde{\beta} = \left(\mathbf{X}'\widetilde{\mathbf{D}}^{-1}\mathbf{X}\right)^{-1}\left(\mathbf{X}'\widetilde{\mathbf{D}}^{-1}\mathbf{y}\right).$$
- We can iterate between $\widetilde{\mathbf{D}}$ and $\widetilde{\beta}$ until convergence.
- If $\widetilde{\sigma}_i^2 < 0$, or $\widetilde{\sigma}_i^2 \approx 0$ for some $i$, use a trimming rule
$$\overline{\sigma}_i^2 = \max\left\{\widetilde{\sigma}_i^2, \underline{\sigma}^2\right\}$$
for some $\underline{\sigma}^2 > 0$.

# Asymptotics for the FGLS Estimator

**Theorem**

*If the skedastic regression is correctly specified,*

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

*and thus*

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

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

- The natural estimator of the asymptotic variance of $\widetilde{\beta}$ 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 $n\to\infty$.

# Asymptotics for the Quasi-FGLS Estimator

- If $\alpha'\mathbf{z}_i$ is only an approximation to the true conditional variance $\sigma_i^2$, we have shown in the last chapter that

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

- $\mathbf{V}$ takes a sandwich form similar to the covariance matrix of the OLS estimator. Unless $\sigma_i^2=\alpha'\mathbf{z}_i$, $\widetilde{\mathbf{V}}^0$ is inconsistent for $\mathbf{V}$.
- An appropriate solution is to use a White-type estimator in place of $\widetilde{\mathbf{V}}^0$,

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

  where $\widehat{\mathbf{D}}=\text{diag}\left\{\widehat{u}_1^2,\cdots,\widehat{u}_n^2\right\}$.
- This estimator is robust to misspecification of the conditional variance.

# Choice Between the FGLS and OLS

- FGLS is asymptotically superior to OLS, but we do not exclusively estimate regression models by FGLS.
- 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 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.
- 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.

# Testing for Heteroskedasticity

# General Idea

- If heteroskedasticity is present, more efficient estimation is possible, so we need test for heteroskedasticity.
- 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

$$H_0 : \alpha_1 = \mathbf{0}$$

  in the skedastic regression (8). So We may test this hypothesis by the Wald test.
- We usually 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 for $\widetilde{\alpha}$ simplifies to $\mathbf{V}_\alpha = E\left[\mathbf{z}_i\mathbf{z}_i'\right]^{-1} E\left[\xi_i^2\right]$.

**Theorem**

*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 $\mathbf{z}_i$.

# The Breusch-Pagan Test

- Breusch-Pagan (1979) assume $u_i$ follows $N(0, \alpha' \mathbf{z}_i)$ and use the Lagrange Multiplier (LM) test to check whether $\alpha_1 = \mathbf{0}$:

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

  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.

# The White Test

- White (1980) observes that when $H_0$ holds, $\widehat{\Omega} = n^{-1}\sum_{i=1}^n \mathbf{x}_i\mathbf{x}_i'\widehat{u}_i^2$ 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.
- 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 auxiliary assumption that $u_i$ is independent of $\mathbf{x}_i$,

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

under $H_0$, 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 $\sqrt{n}\mathbf{D}_n$.

# Simplification

- Given that $u_i$ is independent of $\mathbf{x}_i$, $\widehat{\mathbf{B}}_n$ can be replaced by

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

- From Exercise 7, Koenker and White's test statistics can be expressed in the form of $nR^2$ in some regression.
- 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.
- $nR^2$ from $\widehat{u}_i^2$ on $1, \widehat{y}_i, \widehat{y}_i^2$ has a limiting $\chi_2^2$ distribution under $H_0$.

# Regression Intervals and Forecast Intervals

# Regression Intervals

- All previous sections consider the interval validity. We now consider the external validity, i.e., prediction.
- Suppose we want to estimate $m(\mathbf{x}) = E[y_i|\mathbf{x}_i = \mathbf{x}] = \mathbf{x}'\beta$ at a particular point $\mathbf{x}$ (which may or may not be the same as some $\mathbf{x}_i$) and note that this is a (linear) function of $\beta$.
- Letting $r(\beta) = \mathbf{x}'\beta$ and $\theta = r(\beta)$, we see that $\widehat{m}(\mathbf{x}) = \widehat{\theta} = \mathbf{x}'\widehat{\beta}$ 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

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

  Viewed as a function of $\mathbf{x}$, the width of the confidence set is dependent on $\mathbf{x}$.
- Figure 5: the confidence bands take a hyperbolic shape, which means that the regression line is less precisely estimated for very large and very small values of $x$.
- For a given value of $\mathbf{x}_i = \mathbf{x}$, we may also want to forecast (guess) $y_i$ out-of-sample.[1] A reasonable forecast is still $\widehat{m}(\mathbf{x}) = \mathbf{x}'\widehat{\beta}$ since $m(\mathbf{x})$ is the mean-square-minimizing forecast of $y_i$.

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

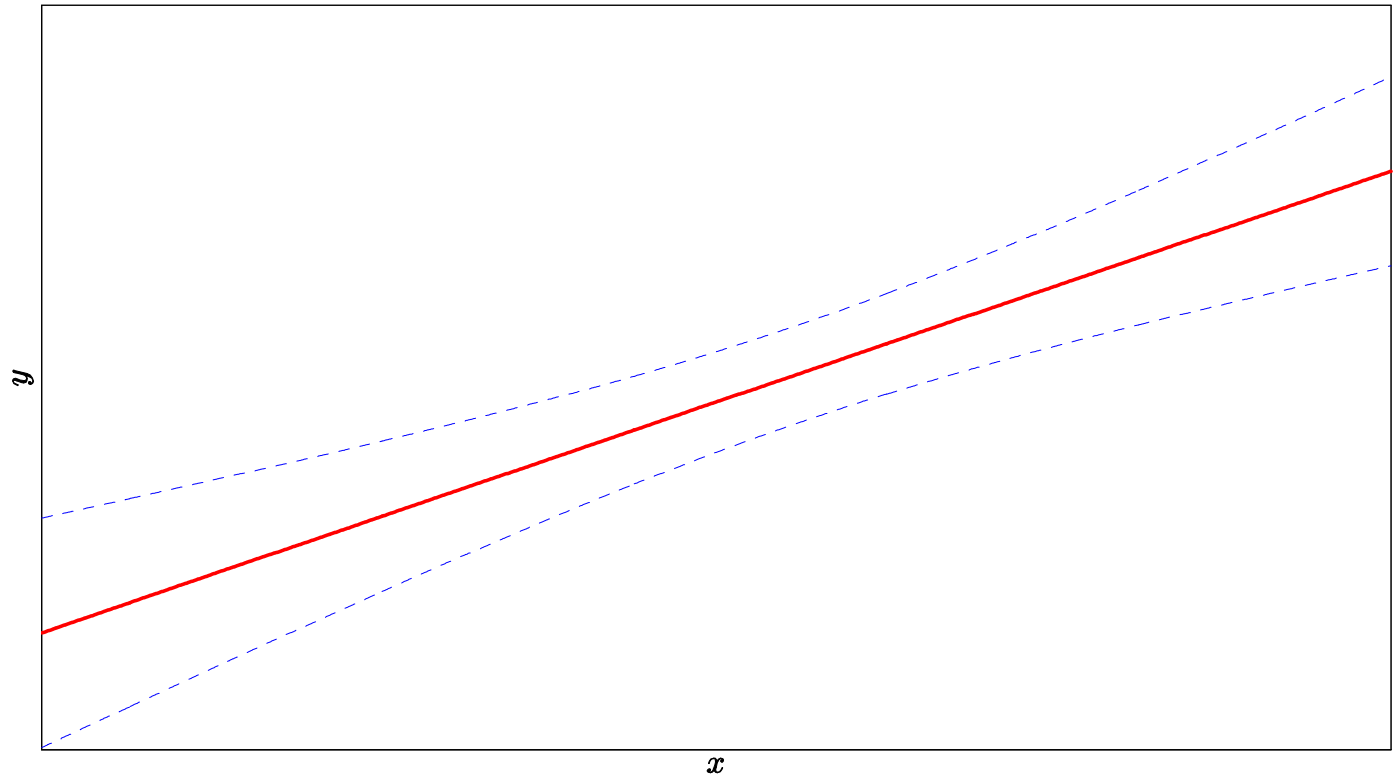


Figure: Typical Regression Intervals

# Forecast Intervals

- The forecast error is $\widehat{u}_i = y_i - \widehat{m}(\mathbf{x}) = u_i - \mathbf{x}'(\widehat{\beta} - \beta)$. As the out-of-sample error $u_i$ is independent of the in-sample estimate $\widehat{\beta}$, this has variance
$$E\left[\widehat{u}_i^2\right] = E\left[u_i^2 | \mathbf{x}_i = \mathbf{x}\right] + \mathbf{x}' E\left[(\widehat{\beta} - \beta)(\widehat{\beta} - \beta)'\right]\mathbf{x} = \sigma^2(\mathbf{x}) + n^{-1}\mathbf{x}'\mathbf{V}\mathbf{x}.$$
- 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 $\widehat{s}(\mathbf{x}) = \sqrt{\widehat{\sigma}^2 + n^{-1}\mathbf{x}'\widehat{\mathbf{V}}\mathbf{x}}$.
- This is different from the standard error for the conditional mean. If we have an estimate of the conditional variance function, e.g., $\widetilde{\sigma}^2(\mathbf{x}) = \widetilde{\alpha}'\mathbf{z}$, then the forecast standard error is $\widehat{s}(\mathbf{x}) = \sqrt{\widetilde{\sigma}^2(\mathbf{x}) + n^{-1}\mathbf{x}'\widehat{\mathbf{V}}\mathbf{x}}$.
- A natural asymptotic 95% forecast interval for $y_i$ is $\left[\mathbf{x}'\widehat{\beta} \pm 2\widehat{s}(\mathbf{x})\right]$, but its validity is based on the asymptotic normality of the studentized ratio, i.e., the asymptotic normality of $\frac{u_i - \mathbf{x}'(\widehat{\beta} - \beta)}{\widehat{s}(\mathbf{x})}$. But no such asymptotic approximation can be made unless $u_i \sim N(0, \sigma^2)$ 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 hard. Usually, people focus on the simple approximate interval $\left[\mathbf{x}'\widehat{\beta} \pm 2\widehat{s}(\mathbf{x})\right]$.