Econometrics

These slides follow very closely
William H. Green, Econometric Analysis, 6th Edition
All graphics and formula are taken out of this book.

Please visit our homepage for updates and new information:

$http://www.uibk.ac.at/statistics/personal/janettewalde/lehre/phdcourseeconometricsss2010.html$

Inhaltsverzeichnis

2. The Classical Multiple Linear Regression Model
2.1 Introduction
2.2 The Linear Regression Model
2.3 Assumptions of the Classical Linear Regression Model
  Linearity of the Classical Regression Model
  Full Rank
  Exogeneity of the Independent Variables
  Spherical Disturbances
  Data Generating Process for the Regressors
  Normality

2.2 The Linear Regression Model

$$\begin{aligned} y &= f(x_1, x_2, ..., x_k) + \varepsilon \\ &= x_1\beta_1 + x_2\beta_2 + ... + x_K\beta_K + \varepsilon \end{aligned}$$

y ... dependent variable (explained variable, regressand)
x_i ... independent variables (explanatory variables, regressors, covariates)
f ... deterministic relationship
ε ... noise (random term, disturbance)
β_i ... parameters (coefficients)

$$\begin{aligned} y_i &= x_{i1}\beta_1 + x_{i2}\beta_2 + ... + x_{iK}\beta_K + \varepsilon_i, \qquad i = 1, ..., n \\ y_i &= \mathbf{x}_i'\beta + \varepsilon_i \\ \mathbf{y} &= \mathbf{X}\beta + \boldsymbol{\varepsilon} \end{aligned}$$

What are the Assumptions of the Classical Linear Regression Model?

2.3.1 Linearity

Assumption 1: $\mathbf{y} = \mathbf{X}\beta + \boldsymbol{\varepsilon}$

Is this a restrictive assumption?

1. Loglinear model (e.g., knowledge production function)
2. Semilog model (growth rates)
3. Variable transformation (e.g., $\frac{1}{x}, x^2$)
4. $y = Ax^{\alpha}e^{\varepsilon}$ versus $y = Ax^{\alpha} + \varepsilon$

2.3.2 Full rank

Assumption 2: $\mathbf{X}$ ia an $n \times K$ matrix with rank K.

What if multicollinearity exists?

$$
\begin{aligned}
y &= \beta_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \varepsilon \\
\text{Assume } x_4 &= x_2 + x_3 \\
\text{Then set } \beta_2' &= \beta_2 + a \\
\beta_3' &= \beta_3 + a \\
\beta_4' &= \beta_4 - a \\
\text{Hence } y &= \beta_1 + \beta_2' x_2 + \beta_3' x_3 + \beta_4' x_4 + \varepsilon.
\end{aligned}
$$

Identification condition.

2.3.3 Regression

$E[\varepsilon_i|\mathbf{X}] = 0$

or

$$\text{Assumption 3: } E[\boldsymbol{\varepsilon}|\mathbf{X}] = E\begin{bmatrix} E[\varepsilon_1|\mathbf{X}] \\ E[\varepsilon_2|\mathbf{X}] \\ \vdots \\ E[\varepsilon_n|\mathbf{X}] \end{bmatrix} = \mathbf{0}$$

This assumption implies that

- $E[\varepsilon_i] = 0$
- $Cov[\varepsilon_i, \mathbf{X}] = 0$ for all i
- $E[\mathbf{y}|\mathbf{X}] = \mathbf{X}\beta$ The regression of $\mathbf{y}$ on $\mathbf{X}$ is the conditional mean.

2.3.4 Spherical Disturbances

$$\begin{aligned} Var[\varepsilon_i|\mathbf{X}] &= \sigma^2, \quad \text{for all } i = 1, ..., n, \\ Cov[\varepsilon_i, \varepsilon_j|\mathbf{X}] &= 0, \quad \text{for all } i \neq j \end{aligned}$$

Assumption 4: $E[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'|\mathbf{X}] = \sigma^2\mathbf{I}$

By using the variance decomposition formula, we find
$Var[\boldsymbol{\varepsilon}] = E[Var[\boldsymbol{\varepsilon}|\mathbf{X}]] + Var[E[\varepsilon|\mathbf{X}]] = \sigma^2 I$

Spherical disturbances are homoscedastic and uncorrelated.

2.3.5 Data Generating Process for the Regressors

Assumption 5: $\mathbf{X}$ may be fixed or random.

2.3.6 Normality

Assumption 6: $\varepsilon|\mathbf{X} \sim N[\mathbf{0}, \sigma^2\mathbf{I}]$.

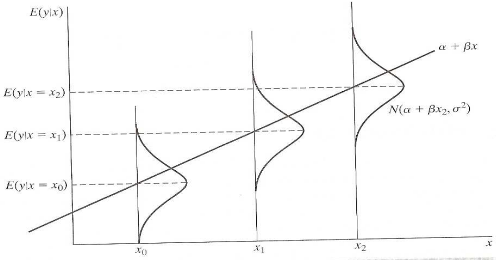


FIGURE 2.2 The Classical Regression Model.

2.3 Assumptions of the Classical Linear Regression Model

1. Linearity
2. Full rank
3. Exogeneity of the independent variables
4. Spherical Disturbance
5. Data Generating Process for the Regressors
6. Normal distribution

Contents

3. Least Squares
3.1 Introduction
3.2 Least Squares Regression
3.3 Partitioned Regression and Partial Regression
3.4 Partial Regression and Partial Correlation Coefficient
3.5 Goodness of Fit and the Analysis of Variance

3.2.1 The Least Squares Coefficient Vector I

Population Regression: $E[y_i|x_i] = x_i'\beta$ with $\varepsilon_i = y_i - x_i'\beta$
Estimate of $E[y_i|x_i]$: $\hat{y}_i = x_i'b$ with $e_i = y_i - x_i'b$

Hence, $y_i = x_i'\beta + \varepsilon_i = x_i'b + e_i$.

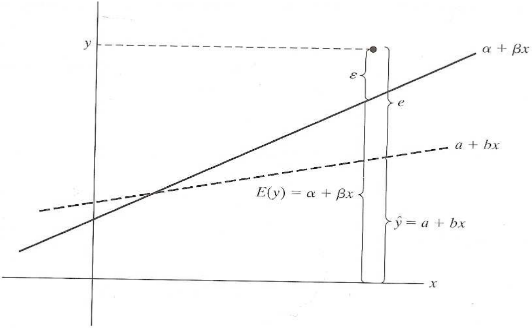


FIGURE 3.1 Population and Sample Regression.

3.2.1 Least Squares Coefficient Vector II

The least squares coefficient vector minimizes the sum of squared residuals:

$$\sum_{i=1}^{n} e_{i0}^2 = \sum_{i=1}^{n}(y_i - x_i' b_0)^2$$

Or in matrix notation:

$$\begin{aligned}
min_{b_0} S(b_0) &= e_0' e_0 = (y - Xb_0)'(y - Xb_0) \\
S(b_0) &= y'y - b_0' X'y - y'Xb_0 + b_0 X'Xb_0 \\
&= y'y - 2y'Xb_0 + b_0 X'Xb_0 \\
\frac{\partial S(b_0)}{\partial b_0} &= -2X'y + 2X'Xb_0 = 0
\end{aligned}$$

Let b be the solution, then b satisfies the **least squares normal equations**: $X'Xb = X'y$. If $(X'X)^{-1}$ exists, the solution is:

$$b = (X'X)^{-1}X'y$$

As $\frac{\partial^2 S(b_0)}{\partial b_0 \partial b_0'} = 2X'X$ is a positive definite matrix b minimizes the sum of squares.

3.2.3 Algebraic Aspects of the Least Squares Solution

Recall the normal equations:

$$X'Xb - X'y = 0 = -X'(y - xb) = -X'e = 0$$

Hence, for every column x_k of X, $x_k e = 0$. If the first column of X is a column of 1s, then there are three implications.

1. The least squares residuals sum to zero: $x_1' e = i'e = \sum_i e = 0$.
2. The regression hyperplane passes through the point of means of the data: $\bar{y} = \bar{x}'b$.
3. The mean of the fitted values from the regression equals the mean of the actual values ($\hat{y} = Xb$).

It is important to note that none of these results need hold if the regression does not contain a constant term.

3.2.4 Projection I

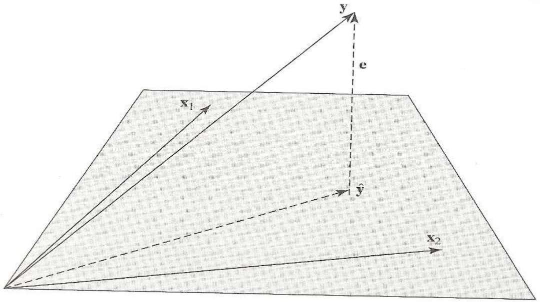
Residual maker:
$e = y - Xb = y - X(X'X)^{-}1X'y = \left(I - X(X'X)^{-1}X'\right)y = My$
M is symmetric ($M = M'$) and idempotent ($M = M^2$), and $MX = 0$.

Projection matrix:
$\hat{y} = y - e = (I - M)y = X(X'X)^{-1}X'y = Py$.
The matrix P which is also symmetric and idempotent, is a projection matrix, and $PX = X$.

As $PM = MP = 0$ it follows that M and P are orthogonal.

3.2.4 Projection II

FIGURE 3.2 Projection of **y** into the Column Space of **X**.



The least squares partitions the vector y into two orthogonal parts, $y = Py + My = projection + residual$.

Using the Pythagorean theorem we get:
$y'y = y'P'Py + y'M'My = \hat{y}'\hat{y} + e'e$.

3.3 Partitioned Regression and Partial Regression I

Suppose that the regression involves two sets of variables X_1 and X_2. Thus,

$$y = X\beta + \varepsilon = X_1\beta_1 + X_2\beta_2 + \varepsilon.$$

What are the algebraic solutions for b_1 and b_2?

The normal equations are

$$\begin{matrix}(1)\\(2)\end{matrix} \quad \begin{pmatrix} X_1'X_1 & X_1'X_2 \\ X_2'X_1 & X_2'X_2 \end{pmatrix} \begin{pmatrix} b_1 \\ b_2 \end{pmatrix} = \begin{pmatrix} X_1'y \\ X_2'y \end{pmatrix}$$

'Solve' (1) for b_1:

$b_1 = (X_1'X_1)^{-1}X_1'(y - X_2b_2)$

Orthogonal Partitioned Regression

> In the multiple linear least squares regression of y on two sets of variables X_1 and X_2, if the two sets of variables are orthogonal, then the separate coefficient vectors can be obtained by separate regressions of y on X_1 alone and y on X_2 alone.

3.3 Partitioned Regression and Partial Regression II

Plug b_1 into (2) and solve the normal equations for b_2:

$$b_2 = (X_2' M_1 X_2)^{-1} (X_2' M_1 y) = (X_2^{*\prime} X_2^{*\prime})^{-1} X_2^{*\prime} y^*,$$

where $X_2^* = M_1 X_2$ and $y^* = M_1 y$.

Frisch-Waugh-Lovell Theorem

> In the linear least squares regression of vector y on two sets of variables, X_1 and X_2, the subvector b_2 is the set of coefficients obtained when the residuals from a regression of y on X_1 alone are regressed on the set of residuals obtained when each column of X_2 is regressed on X_1.

3.3 Partitioned Regression and Partial Regression III

Individual Regression Coefficients

The coefficient on z in a multiple regression of y on $[X, z]$ is computed as $c = (z'Mz)^{-1}(z'My) = (z^{*'}z^{*})^{-1}z^{*'}y^{*}$ where z^{*} and y^{*} are the residual vectors from least squares regressions of z and y on X, $z^{*} = Mz$ and $y^{*} = My$ and M is the residual maker using X.

Application: Slopes of a regression with a constant term

Regression of any variable z on i

is $(i'i)^{-1}i'z = \bar{z}$, the fitted values are $i\bar{z}$, and the residuals are $z_i - \bar{z}$.

The slopes in a multiple regression that contains a constant term are obtained by transforming the data to deviations from their means and then regressing the variable y in deviations form on the explanatory variables, also in deviation form.

3.4 Partial Correlation Coefficients I

Partial correlation coefficient of y and z controlling for X

First regress y on X, in order to obtain $y* = MX$, and regress z on X, to get $z^* = Mz$, where M is the residual maker using X. Then the partial correlation coefficient r^*_{yz} is just the correlation between y^* and z^*

How does the sum of squares change when adding a variable?

$$
\begin{aligned}
e &= y - Xb \\
u &= y - Xd - zc \\
& \qquad\qquad d = (X'X)^{-1}X'(y - zc) = b - (X'X)^{-1}X'zc \\
u &= y - Xb - X(X'X)^{-1}X'zc - zc \\
&= e - Mzc = e - z^*c \\
u'u &= e'e + c^2 z^{*'} z^* - 2cz^{*'} e = e'e - c^2 z^{*'} z^* \\
& \qquad\qquad\qquad\qquad z^{*'} e = z^{*'} y^* = c(z^{*'} z^*) \qquad e = My
\end{aligned}
$$

3.4 Partial Correlation Coefficients II

Change in the sum of squares when a variable is added to a regression

If $e'e$ is the sum of squared residuals when y is regressed on X and $u'u$ is the sum of squared residuals when y is regressed on X and z, then

$$u'u = e'e - c^2({z^*}'z^*) \leq e'e, \tag{1}$$

where c is the coefficient on z in the long regression and $z^* = [I - X(X'X)^{-1}X']z$ is the vector of residuals when z is regressed on X.

3.5 Goodness of Fit and the Analysis of Variance

How well does the regression line fit the data?
Starting point is the total variation in y : $SST = \sum(y_i - \bar{y})^2$

$$\begin{aligned}
y_i &= \hat{y}_i + e_i = x_i'b + e_i \\
y_i - \bar{y} &= \hat{y}_i - \bar{y} + e_i = (x_i - \bar{x})'b + e_i \\
M^0 y &= M^0 Xb + e \qquad\qquad \text{where } M^0 \equiv I - \frac{1}{n} ii' \\
y'M^0 y &= b'X'M^0 Xb + e'e \qquad \text{using } e'M^0 Xb = e'Xb = 0 \\
SST &= SSR + SSE
\end{aligned}$$

Total sum of squares = regression sum of squares + error sum of squares
Coefficient of determination $= R^2 = \frac{SSR}{SST} = \frac{b'X'M^0Xb}{y'M^0y} = 1 - \frac{e'e}{y'M^0y}$

3.5.1 The Adjusted R-Squared and a Measure of Fit I

Use the previous result (1):

$$\begin{aligned} u'u &= e'e - c^2(z^{*\prime}z^*) = e'e - ((z^{*\prime}z^*)^{-1}(z^{*\prime}y^*))^2(z^{*\prime}z^*) \\ &= e'e - \frac{(z^{*\prime}y^*)^2}{z^{*\prime}z^*} = e'e(1 - r_{yz}^{*2}), \end{aligned}$$

where r_{yz}^* is the partial correlation between y and z, controlling for X.

Change in R^2 When a Variable is Added to a Regression

Let R_{Xz}^2 be the coefficient of determination in the regression of y on X and on an additional variable z, let R_X^2 be the same for the regression of y on X alone, and let r_{yz}^* be the partial correlation between y and z, controlling for X. Then

$$R_{Xz}^2 = R_X^2 + (1 - R_X^2)r_{yz}^{*2}.$$

3.5.1 The Adjusted R-Squared and a Measure of Fit II

The adjusted R^2, which incorporates a penalty, is computed as follows:

$$\bar{R}^2 = 1 - \frac{e'e/(n-K)}{y'M^0y/(n-1)} = 1 - \frac{n-1}{n-K}(1-R^2)$$

Whether $\bar{R}^2$ rises or falls depends on whether the contribution of the new variable to the fit of the regression more than offsets the correction for the loss of an additional degree of freedom.

Change in $\bar{R}^2$ When a New Variable is Added to a Regression

In a multiple regression, $\bar{R}^2$ will fall (rise) when the variable x is deleted from the regression if the square of the t ratio associated with this variable is greater (less) than 1.

3.5.2 R-Squared and the Constant Term in the Model

A second difficulty with R^2 concerns the constant term in the model. The proof that $0 \leq R^2 \leq 1$ requires X to contain a column of 1s. If not, then $M^0 e \neq e$ and $e' M^0 X \neq 0$, and the term $2e' M^0 X b$ in $y' M^0 y$ in the preceding expansion will not drop out. Consequently, when computing R^2 the result is unpredictable.

Some final remarks to R^2:

- What value of R^2 is high?
- Comparisons across different models (e.g. linear versus loglinear models) is not possible (variation in y is different from variation in $\ln y$).
- R^2 is a measure of *linear* association.

Contents

4. Statistical Properties of the Least Squares Estimator
4.1 Introduction
4.2 Motivating Least Squares
4.3 Unbiased Estimation
4.4 The Variance of the Least Squares Estimator
4.5 The Implications of Stochastic Regressors
4.6 Estimating the Variance of the Least Squares Estimator
4.7 The Normality Assumption and Basic Statistical Inference
4.8 Finite-Sample Properties of the Least Squares Estimator
4.9 Large Sample Properties of the Least Squares Estimator

4.3 Unbiased Estimation I

The least squares estimator is unbiased in every sample:

$$\begin{aligned} b &= (X'X)^{-1}X'y = (X'X)^{-1}X'(X\beta+\varepsilon) = \beta + (X'X)^{-1}X'\varepsilon \\ E[b|X] &= \beta + E[(X'X)^{-1}X'\varepsilon|X] \qquad \text{Assumption 3: } E[\varepsilon|X] = 0 \\ E[b|X] &= \beta \\ E[b] &= E_X\left[E[b|X]\right] = E_X[\beta] = \beta \end{aligned}$$

Example 4.1: The Sampling Distribution of a Least Squares Estimator

$y_i = 0.5 + 0.5x_i + \varepsilon_i, i = 1, ..., 10000$

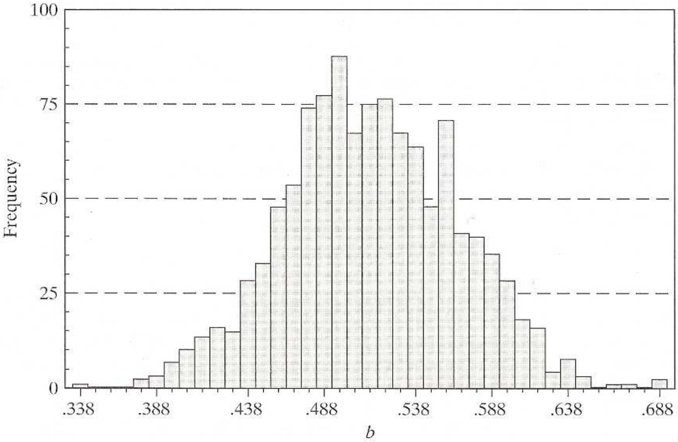
$x_i \sim N(0,1)$

$\varepsilon_i \sim N(0,1/4)$

$MC = 500$ samples, $n = 100$

4.3 Unbiased Estimation II

FIGURE 4.1 Histogram for Sampled Least Squares Regression Slopes.



4.4 The Variance of the Least Squares Estimator I

The covariance matrix of the least squares slope estimator is:

$$\begin{aligned}
b &= (X'X)^{-1}X'(X\beta + \varepsilon) = \beta + (X'X)^{-1}X'\varepsilon \\
Var[b|X] &= E[(b-\beta)(b-\beta)'|X] \\
&= E[(X'X)^{-1}X'\varepsilon\varepsilon'X(X'X)^{-1}|X] \\
&= (X'X)^{-1}X'E[\varepsilon\varepsilon'|X]X(X'X)^{-1} \\
&= (X'X)^{-1}X'(\sigma^2 I)X(X'X)^{-1} \\
&= \sigma^2(X'X)^{-1}
\end{aligned}$$

4.4 The Variance of the Least Squares Estimator II

Example 4.2 Sampling Variance in the Two-Variable Regression Model

Suppose that X contains only a constant term and a single regressor x. Thus, $Var[b|x] = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$.

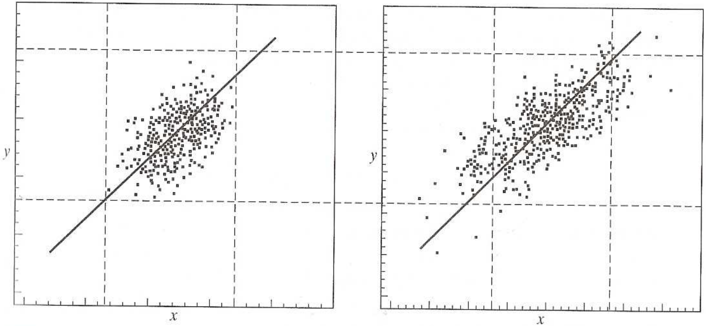


FIGURE 4.2 Effect of Increased Variation in x Given the Same Conditional and Overall Variation in y.

4.4 The Gauss-Markov Theorem I

Since we can write $b = \beta + (X'X)^{-1}X'\varepsilon = \beta + A\varepsilon$, b is a linear estimator. Regardless of the distribution of ε b is a linear, unbiased estimator of β.

Let $b_0 = Cy$ be another linear unbiased estimator of β, where C is a $K \times n$ matrix. If b_0 is unbiased, then $E[b_0|X] = E[Cy|X] = E[(CX\beta + C\varepsilon)|X] = \beta$, which implies that $CX = I$.

Now calculate the covariance matrix of b_0:

$$\begin{aligned} b_0 &= Cy = CX\beta + C\varepsilon = \beta + C\varepsilon \\ Var[b_0|X] &= E[(b_0 - \beta)(b_0 - \beta)'|X] \\ &= E[C\varepsilon\varepsilon' C'] \\ &= \sigma^2 CC' \end{aligned}$$

4.4 The Gauss-Markov Theorem II

Now let $D \equiv C - (X'X)^{-1}X'$ so that $Dy = b_0 - b$. Then

$$\begin{aligned} Var[b_0|X] &= \sigma^2 CC' = \sigma^2[(D + (X'X)^{-1}X')(D + (X'X)^{-1}X')'] \\ & \quad CX = I = DX + (X'X)^{-1}(X'X) = DX + I \rightarrow DX = 0 \\ Var[b_0|X] &= \sigma^2(X'X)^{-1} + \sigma^2 DD' \end{aligned}$$

As DD' is a nonnegative definite matrix $\Rightarrow$ Gauss-Markov Theorem

In the classical linear regression model with regressor matrix X, the least squares estimator b is the minimum variance linear unbiased estimator of β. For any vector of constants w, the minimum variance linear unbiased estimator of $w'\beta$ in the classical regression model is $w'b$, where b is the least squares estimator.

4.5 The Implications of Stochastic Regressors

We showed

$$Var[b|X] \leq Var[b_0|X]$$

for any $b_0 \neq b$ and for the specific X in the sample.
If this inequality holds for every particular X, then it must hold for $Var[b] = E_X[Var[b|X]]$.
If it holds for every particular X, then it must hold over the average values of X.

Gauss-Markov Theorem (Concluded)

In the classical linear regression model, the least squares estimator b is the minimum variance unbiased estimator of β whether X is stochastic or nonstochastic, so long as the other assumptions of the model continue to hold.

4.6 Estimating the Variance of the Least Squares Estimator

In order to test hypotheses about β or to form confidence intervals we require a sample estimate of the covariance matrix $Var[b|X] = \sigma^2(X'X)^{-1}$.

$$
\begin{aligned}
e &= My = M(X\beta + \varepsilon) = M\varepsilon \\
e'e &= \varepsilon' M\varepsilon \\
E[e'e|X] &= E[\varepsilon' M\varepsilon|X] = E[tr(\varepsilon' M\varepsilon)|X] = E[tr(M\varepsilon\varepsilon')|X] \\
&= tr(ME[\varepsilon\varepsilon'|X]) = tr(M\sigma^2 I) = \sigma^2 tr(M) \\
&= \sigma^2 tr(I_n - X(X'X)^{-1}X') = \sigma^2(tr(I_n) - tr((X'X)^{-1}X'X)) \\
&= \sigma^2(tr(I_n) - tr(I_K)) = \sigma^2(n-K)
\end{aligned}
$$

Hence, an unbiased estimator of σ^2 is $s^2 = \frac{e'e}{n-K}$. The estimator is unbiased unconditionally as well:
$E[s^2] = E_X[E[s^2|X]] = E_X[\sigma^2] = \sigma^2$.
Finally, we have Est.Var$[b|X] = s^2(X'X)^{-1}$.

4.7 The Normality Assumption and Basic Statistical Inference I

Using Assumption 6 and the fact that b is a linear function in ε, we obtain the exact finite sample distribution of b:

$$\begin{aligned} b|X &\sim N(\beta, \sigma^2 (X'X)^{-1}) \\ b_k|X &\sim N(\beta_k, \sigma^2 (X'X)^{-1}_{kk}) \end{aligned}$$

Consequently,

1. Testing a hypothesis about a coefficient
   $t_{k,df=n-K} = \frac{b_k - \beta_k}{\sqrt{s^2 (X'X)^{-1}_{kk}}}$
2. Confidence intervals for parameters
   $Prob(b_k - t_{\alpha/2} s_{b_k} \leq \beta_k \leq b_k + t_{\alpha/2} s_{b_k}) = 1 - \alpha$

4.7 The Normality Assumption and Basic Statistical Inference II

3. Confidence intervals for a linear combination of coefficients
   Let w denote a $K \times 1$ vector of known constants. Then $c \equiv w'b \sim N(w'\beta, w'[\sigma^2(X'X)^{-1}]w)$, which we estimate with $s_c^2 = w'[s^2(X'X)^{-1}]w$. Consequently, $Prob[c - t_{\alpha/2}s_c \leq w'\beta \leq c + t_{\alpha/2}s_c] = 1 - \alpha$.

4. Testing the significance of the regression
   $H_0 : \beta_2 = \beta_3 = ... = \beta_K = 0$
   $F[K-1, n-K] = \frac{R^2/(K-1)}{(1-R^2)/(n-K)}$

5. If the disturbances are normally distributed, then the distributions of the hypotheses tests and confidence intervals remain the same regardless whether the regressors are stochastic, nonstochastic or some mix of the two.

4.8.1 Multicollinearity I

For convenience, define the data matrix, X to contain a constant and $K-1$ other variables measured in deviations from their means. Let x_k denote the kth variable, and let $X_{(k)}$ denote all the other variables (including the constant term). Then, in the inverse matrix, $(X'X)^{-1}$, the kth diagonal element is

$$\begin{aligned}(x_k' M_{(k)} x_k)^{-1} &= [x_k' x_k - x_k' X_{(k)} (X_{(k)}' X_{(k)})^{-1} X_{(k)}' x_k]^{-1} \\ &= \left[x_k' x_k \left(1 - \frac{x_k' X_{(k)} (X_{(k)}' X_{(k)})^{-1} X_{(k)}' x_k}{x_k' x_k} \right) \right]^{-1} \\ &= \frac{1}{(1 - R_{k.}^2) S_{kk}},\end{aligned}$$

where $R_{k.}^2$ is the R^2 in the regression of x_k on all the other variables.

4.8.1 Multicollinearity II

The result

$$Var[b_k|X] = \frac{\sigma^2}{(1-R^2_{k.})\sum_{i=1}^{n}(x_{ik}-\bar{x}_k)^2}$$

shows the three ingredients of the precision of the kth least squares estimator:

1. Ceteris paribus, the greater the correlation of x_k with the other variables (multicollinearity) the higher the variance.
2. Ceteris paribus, the greater the variation in x the lower the variance.
3. Ceteris paribus, the greater the overall fit of the regression the lower the variance will be.

Since nonexperimental data will never be orthogonal ($R^2_{k.} = 0$), to some extent multicollinearity will always be present. When is multicollinearity a problem? Some computer packages report a variance inflation factor (VIF), $1/(1 - R^2_{k.})$, for each coefficient in a regression as a diagnostic statistic.

4.9.1 Consistency of the Least Squares Estimator of β I

Assumption 6 (Normal distribution) is dropped, but two additional assumptions:

- Modification of Assumption 5: $(x_i, \varepsilon_i), i = 1, ..., n$ is a sequence of *independent* observations
- Behavior of the data in large samples: $\text{plim}_{n\to\infty} \frac{X'X}{n} = Q$ is a positive definite matrix

The least squares estimator may be written:

$$\begin{aligned} b &= (X'X)^{-1}X'y = (X'X)^{-1}X'(X\beta + \varepsilon) \\ &= \beta + \left(\frac{X'X}{n}\right)^{-1}\left(\frac{X'\varepsilon}{n}\right) \\ \text{plim } b &= \beta + Q^{-1}\text{plim }\left(\frac{X'\varepsilon}{n}\right) \end{aligned}$$

4.9.1 Consistency of the Least Squares Estimator of β II

We need to establish plim $\left(\frac{X'\varepsilon}{n}\right)$. Hence, to compute the expectation value and the variance:

$$
\begin{aligned}
\frac{1}{n}X'\varepsilon &= \frac{1}{n}\sum_{i=1}^{n} x_i\varepsilon_i \\
E[x_i\varepsilon_i] &= E_X[E[x_i\varepsilon_i|x_i]] = E_X[x_i E[\varepsilon_i|x_i]] = 0 \quad \text{Assumption 3} \\
E\left[\frac{1}{n}X'\varepsilon\right] &= 0 \\
Var\left[\frac{1}{n}X'\varepsilon\right] &= E\left[Var\left[\frac{1}{n}X'\varepsilon|X\right]\right] + Var\left[E\left[\frac{1}{n}X'\varepsilon|X\right]\right] \\
Var\left[\frac{1}{n}X'\varepsilon|X\right] &= E\left[Var\left[\frac{1}{n}X'\varepsilon|X\right]\right] + 0 \\
&= E\left[\frac{1}{n}X'\varepsilon(\frac{1}{n}X'\varepsilon)'|X\right] = \frac{1}{n}X'E\left[\varepsilon\varepsilon'|X\right]X\frac{1}{n}
\end{aligned}
$$

4.9.1 Consistency of the Least Squares Estimator of β III

$$
\begin{aligned}
Var\left[\frac{1}{n}X'\varepsilon|X\right] &= \left(\frac{\sigma^2}{n}\right)\left(\frac{X'X}{n}\right) \\
Var\left[\frac{1}{n}X'\varepsilon\right] &= \left(\frac{\sigma^2}{n}\right)E\left(\frac{X'X}{n}\right) \\
lim_{n\to\infty}Var\left[\frac{1}{n}X'\varepsilon\right] &= 0\cdot Q = 0
\end{aligned}
$$

Since the mean of $\frac{1}{n}X'\varepsilon$ is identically to zero and its variance converges to zero, the expression converges in mean square to zero, so plim $\frac{1}{n}X'\varepsilon = 0$.
Hence, plim $b = \beta$. b is a consistent estimator of β.

Further Asymptotic Results I

Asymptotic Distribution of b With Independent Observations

Given our assumptions we obtain the asymptotic distribution of b as $b \sim^a N\left(\beta, \frac{\sigma^2}{n} Q^{-1}\right)$. In practice, it is necessary to estimate $(1/n)Q^{-1}$ with $(X'X)^{-1}$ and σ^2 with $e'e/(n-K)$.

Consistency of s^2 and the Estimator of Asy.Var[b]

Under fairly weak conditions, we get plim $s^2 = \sigma^2$ and the appropriate estimator of the asymptotic covariance matrix of b is $Est.Asy.Var[b] = s^2(X'X)^{-1}$.

Further Asymptotic Results II

Asymptotic Distribution of a Function of b

If $f(b)$ is a set of continuous and continuously differentiable functions of b such that $\Gamma = \partial f(\beta)/\partial \beta'$ and if the assumptions hold, then

$$f(b) \sim^{a} N\left[f(\beta), \Gamma\left(\frac{\sigma^2}{n} Q^{-1}\right)\Gamma'\right].$$

In practice the estimator of the asymptotic covariance matrix would be

$$Est.Asy.Var[f(b)] = C[s^2(X'X)^{-1}]C',$$

where $C = \frac{\partial f(b)}{\partial b'}$.

Further Asymptotic Results III

Asymptotic Efficiency

> An Estimator is asymptotically efficient if it is consistent, asymptotically normally distributed and has an asymptotic covariance matrix that is not larger than the asymptotic covariance matrix of any other consistent, asymptotically normally distributed estimator.

If the disturbances are normally distributed then the least squares estimator is also a maximum likelihood estimator (MLE). MLEs are asymptotically efficient among consistent and asymptotically normally distributed estimators. If some other distribution is specified for ε and it emerges that b is not the MLE then least squares may not be efficient.

Contents

5. Inference and Prediction
5.1 Introduction
5.2 Restrictions and Nested Models
5.3 Two Approaches to Testing Hypotheses
5.4 Nonnormal Disturbances and Large Sample Tests
5.5 Testing Nonlinear Restrictions
5.6 Prediction

5.2 Restrictions and Nested Models

$y = X\beta + \varepsilon$
Considering a set of linear restrictions:

$$
\begin{aligned}
r_{11}\beta_1 + r_{12}\beta_2 + \ldots + r_{1K}\beta_K &= q_1 \\
r_{21}\beta_1 + r_{22}\beta_2 + \ldots + r_{2K}\beta_K &= q_2 \\
&\vdots \\
r_{J1}\beta_1 + r_{J2}\beta_2 + \ldots + r_{JK}\beta_K &= q_J
\end{aligned}
$$

These can be combined in a single equation: $R\beta = q$

The restriction imposes J restrictions on K otherwise free parameters, i.e. $R\beta = R_1\beta_1 + R_2\beta_2 = q$, if the J columns of R_1 are independent, then $\beta_1 = R_1^{-1}(q - R_2\beta_2)$. Thus, only the $K - J$ elements of β_2 are free parameters in the restricted model.

5.3 Two Approaches to Testing Hypothesis 5.3.1 The F Statistic and the Least Sqares Discrepancy

$H_0 : R\beta - q = 0$ versus $H_1 : R\beta - q \neq 0$
$Rb - q = m$
$b \sim N \rightarrow m \sim N$
H_0 true

$$\begin{aligned} E[m|X] &= RE[b|X] - q = R\beta - q = 0 \\ Var[m|X] &= Var[Rb - q|X] = RVar[b|X]R' = \sigma^2 R(X'X)^{-1}R' \end{aligned}$$

Wald
$W = m' Var[m|X]^{-1} m = \frac{(Rb-q)'(R(X'X)^{-1}R')^{-1}(Rb-q)}{\sigma^2} \sim \chi^2(J)$
F-statistic
$F = \frac{W}{J}\frac{\sigma^2}{s^2} = \frac{(Rb-q)'(R(X'X)^{-1}R')^{-1}(Rb-q)}{Js^2} \sim F(J, n-K|X)$

5.3.2 The Restricted Least Squares Estimator

Minimize$_{b_0}(y - Xb_0)'(y - Xb_0)$ subject to $Rb_0 = q$

$$
\begin{aligned}
L(b_0, \lambda) &= (y - Xb_0)'(y - Xb_0) + 2\lambda'(Rb_0 - q) \\
\frac{\partial L}{\partial b_0} &= -2X'(y - Xb_0) + 2R'\lambda = 0 \\
\frac{\partial L}{\partial \lambda} &= Rb_0 - q = 0 \\
\begin{pmatrix} X'X & R' \\ R & 0 \end{pmatrix} \begin{pmatrix} b_* \\ \lambda_* \end{pmatrix} &= \begin{pmatrix} X'y \\ q \end{pmatrix} \\
b_* &= b - (X'X)^{-1}R'(R(X'X)^{-1}R')^{-1}(Rb - q)
\end{aligned}
$$

5.3.3 The Loss of Fit From Restricted Least Squares

Focus on the fit of the regression: $R^2 = 1 - e'e/y'M^0y$

Let $e_* = y - Xb_*$:

$$\begin{aligned} e_* &= y - Xb_* + Xb - Xb = y - Xb - X(b_* - b) = \\ &= e - X(b_* - b) \\ e_*'e_* &= e'e - (b_* - b)'X'e - e'X(b_* - b) + (b_* - b)'X'X(b_* - b) \\ &= e'e + (b_* - b)'X'X(b_* - b) \geq e'e \end{aligned}$$

$$\begin{aligned} e_*'e_* - e'e &= (b_* - b)'X'X(b_* - b) \\ &= (Rb - q)'(R(X'X)^{-1}R')^{-1}(Rb - q) \end{aligned}$$

$$F(J, n-K) = \frac{e_*'e_* - e'e/J}{e'e/(n-K)} \qquad \div \sum_i (y_i - \bar{y})^2$$

$$F(J, n-K) = \frac{(R^2 - R_*^2)/J}{(1 - R^2)/(n-K)}$$

5.4 Nonnormal Disturbances and Large Sample Tests I

$b \sim^a N(\beta, \frac{\sigma^2}{n} Q^{-1})$, where $Q = plim\left(\frac{X'X}{n}\right)$

$plim\ s^2 = \sigma^2$, where $s^2 = e'e/(n-K)$

$t_k = \frac{\sqrt{n}(b_k - \beta_k^0)}{\sqrt{s^2 (X'X/n)^{-1}_{kk}}} \sim^a N(0,1)$

$$\begin{aligned} F(J, n-K) &= \frac{(Rb-q)'(R(X'X)^{-1}R')^{-1}(Rb-q)}{Js^2} \\ F = &= \frac{(Rb-q)'(R\sigma^2(X'X)^{-1}R')^{-1}(Rb-q)}{J(s^2/\sigma^2)} \end{aligned}$$

5.4 Nonnormal Disturbances and Large Sample Tests II

$$
\begin{aligned}
W^* &= \frac{1}{J}(Rb-q)'(R(\sigma^2/n)Q^{-1}R')^{-1}(Rb-q) \\
&= \frac{1}{J}(Rb-q)'(Ass.Var[Rb-q])^{-1}(Rb-q)
\end{aligned}
$$

Limiting Distribution of the Wald Statistic

If $\sqrt{n}(b-\beta) \rightarrow^d N\left(0, \sigma^2 Q^{-1}\right)$ and if $H_0 : R\beta - q = 0$ is true, then

$$W = (Rb-q)'[Rs^2(X'X)^{-1}R']^{-1}(Rb-q) = JF \rightarrow^d \chi^2(J).$$

5.6 Prediction I

Suppose we wish to predict the value of y_0 associated with a regressor vector x_0.
$y_0 = x_0'\beta + \varepsilon_0$
$E[y_0|x_0] = \hat{y}_0 = x_0'b$ is the minimum variance linear unbiased estimator.

Prediction variance:
$e_0 = y_0 - \hat{y}_0 = (\beta - b)'x_0 + \varepsilon_0$
$Var[e_0|X, x_0] = \sigma^2 + Var[(\beta - b)'x_0|X, x_0] = \sigma^2 + x_0'\sigma^2(X'X)^{-1}x_0$

$Var[e_0] =$
$\sigma^2 \left[1 + \frac{1}{n} + \sum_{j=1}^{K-1}\sum_{k=1}^{K-1}(x_{j0} - \bar{x}_j)(x_{k0} - \bar{x}_k)(Z'M_0Z)^{jk}\right]$,
where Z is the $K-1$ columns of X not including the constant.

5.6 Prediction II

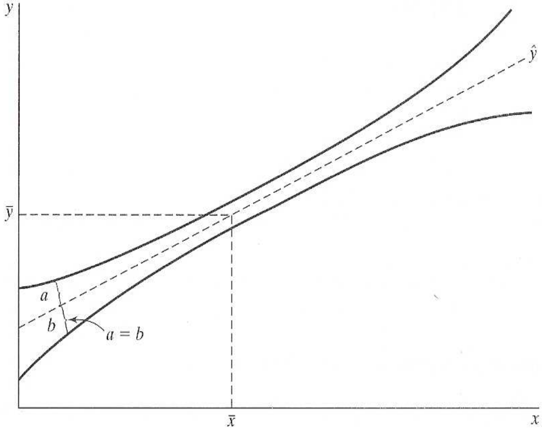


FIGURE 5.1 Prediction Intervals.

5.6 Prediction III

$RMSE = \sqrt{\frac{1}{n} \sum_i (y_i - \hat{y}_i)^2}$

$MAE = \frac{1}{n} \sum_i |y_i - \hat{y}_i|$

Theil U statistic:

$$U = \sqrt{\frac{(1/n) \sum_i (y_i - \hat{y}_i)^2}{(1/n) \sum_i y_i^2}}$$

$$U_\Delta = \sqrt{\frac{(1/n) \sum_i (\Delta y_i - \Delta \hat{y}_i)^2}{(1/n) \sum_i (\Delta y_i)^2}}$$

where $\Delta y_i = y_i - y_{i-1}$ and $\Delta \hat{y}_i = \hat{y}_i - \hat{y}_{i-1}$ or $\Delta y_i = \frac{y_i - y_{i-1}}{y_{i-1}}$ and $\Delta \hat{y}_i = \frac{\hat{y}_i - \hat{y}_{i-1}}{\hat{y}_{i-1}}$.

Contents

6. Functional Form and Structural Change
6.1 Introduction
6.2 Using Binary Variables
6.3 Nonlinearity in the Variables
6.4 Modeling and Testing for a Structural Break

6.2.1 Binary Variables in Regression

- Dummy variables are variables taking the value 1 or 0 to represent a particular observation either having or not having a particular property.
- Studying the treatment on some kind of response, e.g. effect of college on lifetime income, sex differences in labor supply behaviour, pre- versus postregime shifts in macroeconomic models
- A dummy covering a single observation, e.g. an anomalous year, effectively takes that observation out of the sample from the point of view of calculating the coefficient estimates. In minimizing the sum of squared residuals (SSR), can choose coefficients of the other variables that minimize the SSR for the other observations, then choose coefficient on dummy to give error 0 on that observation.

6.2.1 Binary Variables in Regression

A model involving a dummy variable or dummy variables essentially allows different intercept terms for the different categories represented by the variables.

E.g., an individual earnings equation might take the form:
$Earnings = \beta_0 + \beta_1 Education + \beta_2 Age + \beta_3 Age^2 + \beta_4 Gender + \varepsilon$,
where $Gender = 1$ for females and 0 for males.

We can break this down into two separate equations:
Males: $Earnings = \beta_0 + \beta_1 Education + \beta_2 Age + \beta_3 Age^2 + \varepsilon$
Females:
$Earnings = (\beta_0 + \beta_4) + \beta_1 Education + \beta_2 Age + \beta_3 Age^2 + \varepsilon$

6.2.2 Several Categories

When using dummies for categories, one must be careful to avoid dummy variable trap. E.g. if you have a dummy for each quarter, Q_1, Q_2, Q_3 and Q_4, which are 1 in their corresponding quarter and 0 otherwise, then you would have $Q_1 + Q_2 + Q_3 + Q_4 = 1$, which give perfect multicollinearity in a model with constant term.
Confer:

$$
\begin{aligned}
C_t &= \beta_1 + \beta_2 x_t + \delta_1 Q_1 + \delta_2 Q_2 + \delta_3 Q_3 + \delta_4 Q_4 + \varepsilon \\
X &= \begin{pmatrix} 1 & x_1 & 1 & 0 & 0 & 0 \\ 1 & x_2 & 0 & 1 & 0 & 0 \\ 1 & x_3 & 0 & 0 & 1 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_n & 0 & 0 & 0 & 1 \end{pmatrix}
\end{aligned}
$$

Clearly, the columns of this matrix are not linearly independent and consequently X has not full column rank.

6.2.4 Threshold Effects and Categorical Variables I

$income = \beta_1 + \beta_2 age + effect\ of\ education + \varepsilon$

The data on education might consist of the highest level of education attained, such as high school (HS), undergraduate (B), master's (M), or PhD (P).

Why not variable 'Education' with values 1,2,3,4?

6.2.4 Threshold Effects and Categorical Variables II

Using the following dummy variables:
$income = \beta_1 + \beta_2 age + \delta_B B + \delta_M M + \delta_P P + \varepsilon$

Coding 1:

	B	M	P
HS	0	0	0
B	1	0	0
M	0	1	0
PhD	0	0	1

The correspondence between the coefficients and income for a given age is:

High school: $E[income|age, HS] = \beta_1 + \beta_2 age$
Bachelor's: $E[income|age, B] = \beta_1 + \beta_2 age + \delta_B$
Master's: $E[income|age, M] = \beta_1 + \beta_2 age + \delta_M$
PhD: $E[income|age, P] = \beta_1 + \beta_2 age + \delta_P$

6.2.4 Threshold Effects and Categorical Variables III

Using the following dummy variables:

$income = \beta_1 + \beta_2 age + \delta_B B + \delta_M M + \delta_P P + \varepsilon$

Coding 2:

	B	M	P
HS	0	0	0
B	1	0	0
M	1	1	0
PhD	1	1	1

The correspondence between the coefficients and income for a given age is:

High school: $E[income|age, HS] = \beta_1 + \beta_2 age$

Bachelor's: $E[income|age, B] = \beta_1 + \beta_2 age + \delta_B$

Master's: $E[income|age, M] = \beta_1 + \beta_2 age + \delta_B + \delta_M$

PhD: $E[income|age, P] = \beta_1 + \beta_2 age + \delta_B + \delta_M + \delta_P$

6.2.5 Spline Regression I

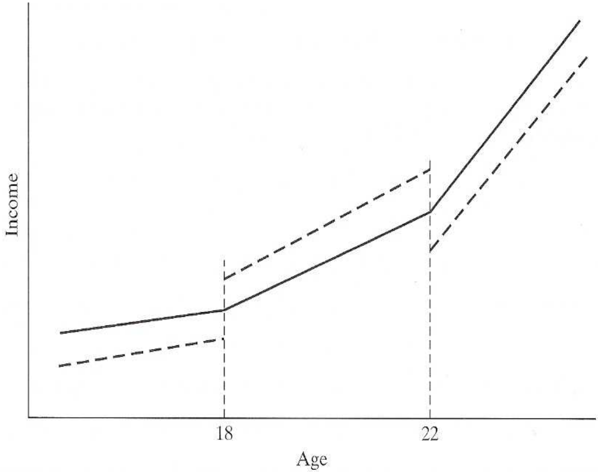


FIGURE 6.2 Spline Function.

6.2.5 Spline Regression II

The function we wish to estimate is

$$\begin{aligned} E[income|age] &= \alpha^0 + \beta^0 age \quad \text{if } age < 18 \\ &= \alpha^1 + \beta^1 age \quad \text{if } age \geq 18 \text{ and } age < 22 \\ &= \alpha^2 + \beta^2 age \quad \text{if } age \geq 22 \end{aligned}$$

The threshold values, 18 and 22, are called knots. Let

$$\begin{aligned} d_1 = 1 \quad \text{if } age \geq t_1^* = 18 \\ d_2 = 1 \quad \text{if } age \geq t_2^* = 22 \end{aligned}$$

$income = \beta_1 + \beta_2 age + \gamma_1 d_1 + \delta_1 d_1 age + \gamma_2 d_2 + \delta_2 d_2 age + \varepsilon$

This relationship is the dashed function in the figure before. The slopes in the three segments are β_2, $\beta_2 + \delta_1$, and $\beta_2 + \delta_1 + \delta_2$.

6.2.5 Spline Regression III

To make the functions piecewise continuous, we require that the segments join at the knots:

$$\begin{aligned} \beta_1 + \beta_2 t_1^* &= (\beta_1 + \gamma_1) + (\beta_2 + \delta_1) t_1^* \\ (\beta_1 + \gamma_1) + (\beta_2 + \delta_1) t_2^* &= (\beta_1 + \gamma_1 + \gamma_2) + (\beta_2 + \delta_1 + \delta_2) t_2^* \end{aligned}$$

These are linear restrictions on the coefficients:

$$\gamma_1 + \delta_1 t_1 * \rightarrow \gamma_1 = -\delta_1 t_1 *$$

$$\gamma_2 + \delta_2 t_2 * \rightarrow \gamma_2 = -\delta_2 t_2 *$$

Hence, $income = \beta_1 + \beta_2 age + \delta_1 d_1 (age - t_1^*) + \delta_2 d_2 (age - t_2^*) + \varepsilon$. Constrained least squares estimates are obtainable by multiple regression using a constant and the three variables.

6.3 Nonlinearity I

The linear regression model in a very general form:

$$\begin{aligned} g(y) &= \beta_1 f_1(z) + \beta_2 f_2(z) + ... + \beta_K f_K(z) + \varepsilon \\ &= \beta_1 x_1 + \beta_2 x_2 + ... + \beta_K x_K + \varepsilon \\ &= x'\beta + \varepsilon, \end{aligned}$$

where $z = z_1, z_2, ..., z_L$ are L independent variables, $f_1, f_2, ..., f_K$ are K linearly independent functions of z, and $g(y)$ is an observable function of y.

- Loglinear model $lny = ln\alpha + \sum_k \beta_k lnX_k + \varepsilon = \beta_1 + \sum_k \beta_k x_k + \varepsilon$. The coefficients are elasticities.
- Semilog model $lny = \beta_1 + \beta_2 + \varepsilon$. The coefficients are partial or semi-elasticities.

6.3 Nonlinearity I

- Age, age^2
- Interaction terms.
  E.g. $D = \beta_1 + \beta_2 S + \beta_3 W + \beta_4 SW + \varepsilon$.
  In this model $\frac{\partial E[D|S,W]}{\partial S} = \beta_2 + \beta_4 W$ and
  $Var\left(\frac{\partial \hat{E}[D|S,W]}{\partial S}\right) = Var[\hat{\beta}_2] + W^2 Var[\hat{\beta}_4] + 2WCov[\hat{\beta}_2, \hat{\beta}_4]$.
- $y_i = \alpha + \beta x_{i1} + \gamma x_{i2} + \beta\gamma x_{i3} + \varepsilon_i$

6.4 Modeling and Testing for a Structural Break I

Application of the F test (Chow test) to test structural breaks.

1. Different parameter vectors

$$
\begin{aligned}
\begin{pmatrix} y_1 \\ y_2 \end{pmatrix} &= \begin{pmatrix} X_1 & 0 \\ 0 & X_2 \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \end{pmatrix} \\
b = (X'X)^{-1}X'y &= \begin{pmatrix} X_1'X_1 & 0 \\ 0 & X_2'X_2 \end{pmatrix}^{-1} \begin{pmatrix} X_1'y_1 \\ X_2'y_2 \end{pmatrix} = \begin{pmatrix} b_1 \\ b_2 \end{pmatrix} \\
e'e &= e_1'e_1 + e_2'e_2
\end{aligned}
$$

$$
\begin{pmatrix} y_1 \\ y_2 \end{pmatrix} = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \beta + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \end{pmatrix} \rightarrow e_*'e_*
$$

$F(J, n_1 + n_2 - 2K) = \frac{(e_*'e_* - e'e)/J}{e'e/(n_1+n_2-2K)}$, where J, the number of restrictions, is the number of columns in X_2.

6.4 Modeling and Testing for a Structural Break II

2. Way out if insufficient observations are available in one period:

   1. Estimate the regression using the full data set $\rightarrow e_*' e_*$.
   2. Use the longer subperiod (n_1 observations) and compute the unrestricted sum of squares $e_1' e_1$.
   3. The F statistic is then computed, using $F(n_2, n_1 - K) = \frac{(e_*' e_* - e_1' e_1)/n_2}{e_1' e_1/(n_1 - K)}$

6.4 Modeling and Testing for a Structural Break III

3. Change in a subset of coefficients

$$X_U = \begin{pmatrix} i & 0 & W_{pre73} & 0 \\ 0 & i & 0 & W_{post73} \end{pmatrix}$$

$$X_R = \begin{pmatrix} i & 0 & W_{pre73} \\ 0 & i & W_{post73} \end{pmatrix}$$

or

$$X_U = \begin{pmatrix} i_{pre} & Z_{pre} & 0 & 0 & W_{pre} & 0 \\ 0 & 0 & i_{post} & Z_{post} & 0 & W_{post} \end{pmatrix}$$

$$X_R = \begin{pmatrix} i_{pre} & Z_{pre} & 0 & 0 & W_{pre} \\ 0 & 0 & i_{post} & Z_{post} & W_{post} \end{pmatrix}$$

6.4 Modeling and Testing for a Structural Break IV

4. Tests of structural break with unequal variances:
   If the sample size is reasonably large, we have a test whether or not the disturbance variances are the same.
   Wald statistic $W = (\hat{\theta}_1 - \hat{\theta}_2)'(\hat{V}_1 + \hat{V}_2)^{-1}(\hat{\theta}_1 - \hat{\theta}_2) \sim \chi^2(K)$

Contents

7. Specification Analysis and Model Selection
7.1 Introduction
7.2 Specification Analysis and Model Building
7.3 Choosing between Nonnested Models
7.4 Model Selection Criteria
7.5 Model Selection

7.2.1 Bias Caused by Omission of Relevant Variables I

Suppose that a correctly specified regression model would be

$$y = X_1\beta_1 + X_2\beta_2 + \varepsilon,$$

where the two parts of X have K_1 and K_2 columns.

If we regress y on X_1 without including X_2, then the estimator is

$$b_1 = (X_1'X_1)^{-1}X_1'y = \beta_1 + (X_1'X_1)^{-1}X_1'X_2\beta_2 + (X_1'X_1)^{-1}X_1'\varepsilon.$$

Unless $X_1'X_2 = 0$ or $\beta_2 = 0$, b_1 is biased.
Omitted variable formula:

$$E[b_1|X] = \beta_1 + P_{1.2}\beta_2, \quad P_{1.2} = (X_1'X_1)^{-1}X_1'X_2$$

Each column of the $K_1 \times K_2$ matrix $P_{1.2}$ is the column of slopes in the least squares regression of the corresponding column of X_2 on the columns of X_1.

7.2.1 Bias Caused by Omission of Relevant Variables II

Note: The direction of the bias can be derived if there is just a single variable included and one omitted variable. However, if more than one variable is included the sign of the bias cannot be determined.

7.2.3 Inclusion of Irrelevant Variables

If the regression model is correctly given by

$$y = X_1\beta_1 + \varepsilon$$

but we estimate $y = X_1\beta_1 + X_2\beta_2 + \varepsilon$. Then the inclusion of the irrelevant variables X_2 in the regression i equivalent to failing to impose $\beta_2 = 0$in estimation. But $y = X_1\beta_1 + X_2\beta_2 + \varepsilon$ is not incorrect, it simply fails to incorporate $\beta_2 = 0$.

7.2.3 Inclusion of Irrelevant Variables

Consequently, the estimator b is unbiased:

$$E[b|X] = \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix} = \begin{pmatrix} \beta_1 \\ 0 \end{pmatrix}$$

and also the estimator of s^2

$$E\left[\frac{e'e}{n-K_1-K_2}\middle| X\right] = \sigma^2.$$

The cost is the reduced precision of the estimates. The covariance matrix in the short regression (omitting X_2) is never larger than the covariance matrix for the estimator obtained in the presence of the superfluous variables.

7.3 Choosing Between Nonnested Models

Focus on comparing two competing linear models, e.g.:

$H_0 : y = X\beta + \varepsilon_0$ versus $H_1 : y = Z\gamma + \varepsilon_1$

We discuss two testing possibilities:

1. **Encompassing principle**: The principle directs attention to the question of whether a maintained model can explain the features of its competitors, that is, whether the maintained model encompasses the alternative.
2. **Comprehensive model**: The approach is based on forming a comprehensive model that contains both competitors as special cases.

7.3.2 An Encompassing Model I

The encompassing approach is one in which the ability of one model to explain features of another is tested. Model 0 encompasses Model 1 if the features of Model 1 can be explained by Model 0 but the reverse is not true.

Procedure: Artificial nesting of the two models
$\bar{X}$... set of variables in X that are not in Z
$\bar{Z}$... set of variables in Z that are not in X
W... set of variables that the models have in common

Then H_0 and H_1 can be combined in a 'supermodel':

$$y = \bar{X}\bar{\beta} + \bar{Z}\bar{\gamma} + W\delta + \varepsilon.$$

In principle, H_1 is rejected if it is found that $\bar{\gamma} = 0$ by a conventional F test, whereas H_0 is rejected if it is found that $\bar{\beta} = 0$.

7.3.2 An Encompassing Model II

Two obvious disadvantages:

- δ remains a mixture of parts of β and γ, and it is not established by the F test that either of these parts is zero. Hence, this test does not really distinguish between H_0 and H_1; it distinguishes between H_1 and a hybrid model.
- The compound model may have an extremely large number of regressors (time-series setting: problem of collinearity).
- Alternative approach with the augmented regression:
  $y = X\beta + \bar{Z}\gamma_1 + \varepsilon_1$.
  Testing the hypothesis that $\gamma_1 = 0$ by using an appropriate F test.

7.3.3 Comprehensive Approach - The J Test

$H_0 : y = X\beta + \varepsilon_0$ versus $H_1 : y = Z\gamma + \varepsilon_1$

$y = (1 - \lambda)X\beta + \lambda(Z\gamma) + \varepsilon$

A test of λ would be a test against H_1. The problem is that λ cannot be separately estimated in this model; it would amount to a redundant scaling of the regression coefficients.
Davidson and MacKinnon's J test consists of the following steps:

1. Regress y on Z and obtain $\hat{\gamma}$.
2. Regress y on X and $(Z\hat{\gamma})$.
3. $H_0 : \lambda = 0$ Asymptotically, the ratio $\frac{\hat{\lambda}}{se(\hat{\lambda})}$ (i.e., the usual t ratio) is distributed as standard normal.

7.4 Model Selection Criteria I

$$\bar{R}^2 = 1 - \frac{n-1}{n-K}(1 - R^2) = 1 - \frac{n-1}{n-K}\left(\frac{e'e}{\sum_{i=1}^{n}(y_i - \bar{y})^2}\right)$$

The adjusted R^2 penalizes the loss of degrees of freedom that occurs when a model is expanded. There is, however, some question about whether the penalty is sufficiently large to ensure that the criterion will necessarily lead the analyst to the correct model as the sample size increases.

Two alternative fit measures have been suggested: Akaike Information Criterion (AIC) and Schwarz or Bayesian Information Criterion (BIC):

7.4 Model Selection Criteria II

$$
\begin{aligned}
AIC(K) &= s_y^2(1-R^2)e^{2K/n} \\
BIC(K) &= s_y^2(1-R^2)n^{k/n}
\end{aligned}
$$

Logs are more convenient:

$$
\begin{aligned}
AIC(K) &= \ln(\frac{e'e}{n}) + \frac{2K}{n} \\
BIC(K) &= \ln(\frac{e'e}{n}) + \frac{K \ln n}{n}
\end{aligned}
$$

Both prediction criterion have their virtues, and neither has an obvious advantage over the other. The Schwarz criterion, with its heavier penalty for degrees of freedom lost, will lean towards a simpler model.

Contents

8. The Generalized Regression Model and Heteroscedasticity
8.1 Introduction
8.2 Least Squares Estimation
8.3 Efficient Estimation by Generalized Least Squares (GLS)
8.4 Heteroscedasticity
8.5 Testing for Heteroscedasticity
8.6 Weighted Least Squares When Ω is Known
8.7 Estimation When Ω Contains Unknown Parameters

8.1 Introduction I

The generalized linear regression model is

$$\begin{aligned} y &= X\beta + \varepsilon, \\ E[\varepsilon|X] &= 0, \\ E[\varepsilon\varepsilon'|X] &= \sigma^2\Omega = \Sigma, \end{aligned}$$

where Ω is a positive definite matrix.

In this section we will consider heteroscedastic disturbances that are still uncorrelated:

$$\sigma^2\Omega = \sigma^2 \begin{pmatrix} \omega_1 & 0 & \cdots & 0 \\ 0 & \omega_2 & \cdots & 0 \\ & & \vdots & \\ 0 & 0 & \cdots & \omega_n \end{pmatrix} = \begin{pmatrix} \sigma_1^2 & 0 & \cdots & 0 \\ 0 & \sigma_2^2 & \cdots & 0 \\ & & \vdots & \\ 0 & 0 & \cdots & \sigma_n^2 \end{pmatrix}$$

8.1 Introduction II

Autocorrelation is usually found in time-series data. Economic time series often display a 'memory' in that variation around the regression function is not independent from one period to the next. Time-series data are usually homoscedastic, so $\sigma^2\Omega$ might be

$$\sigma^2\Omega = \sigma^2 \begin{pmatrix} 1 & \rho_1 & \cdots & \rho_{n-1} \\ \rho_1 & 1 & \cdots & \rho_{n-2} \\ & & \vdots & \\ \rho_{n-1} & \rho_{n-2} & \cdots & 1 \end{pmatrix}$$

In most cases, consistent with the notion of a fading memory, the values decline as we move away from the diagonal.

Our earlier results for the classical model will have to be modified.

8.2 Least Squares Estimation

To summarize, the least squares estimator remains unbiased, consistent, and asymptotically normally distributed. It will however, no longer be efficient and the usual inference procedures are no longer appropriate.

The least squares estimator is unbiased in the generalized regression model and its variance is given by:

$$
\begin{aligned}
b &= (X'X)^{-1}X'y = \beta + (X'X)^{-1}X'\epsilon \\
E[b] &= E_X[E[b|X]] = \beta \qquad\qquad \text{as } E[\varepsilon|X] = 0 \\
Var[b|X] &= E[(b-\beta)(b-\beta)'|X] \\
&= E[(X'X)^{-1}X'\varepsilon\varepsilon' X(X'X)^{-1}|X] \\
&= (X'X)^{-1}X'(\sigma^2\Omega)X(X'X)^{-1}
\end{aligned}
$$

8.2.1 Finite-Sample Properties of Ordinary Least Squares

b is a linear function of ε. Therefore, if ε is normally distributed, then

$$b|X \sim N(\beta, \sigma^2(X'X)^{-1}X'\Omega X(X'X)^{-1})$$

8.2.2 Asymptotic Properties

$$\begin{aligned} Var[b|X] &= (X'X)^{-1}X'(\sigma^2\Omega)X(X'X)^{-1} \\ &= \frac{\sigma^2}{n}\left(\frac{1}{n}X'X\right)^{-1}\left(\frac{1}{n}X'\Omega X\right)\left(\frac{1}{n}X'X\right)^{-1} \end{aligned}$$

Consistency of OLS in the Generalized Regression Model

If $Q = \text{plim}\ (X'X/n)$ and $\text{plim}\ (X'\Omega X/n)$ are both finite positive definite matrices, then b is consistent for β. Under the assumed conditions, $\text{plim}\ b = \beta$.

8.2.2 Asymptotic Properties of Least Squares

Asymptotic Distribution of b in the Generalized Regression Model

If the regressors are sufficiently well behaved and the off-diagonal terms in Ω diminish sufficiently rapidly, then the least squares estimator is asymptotically normally distributed with mean β and covariance matrix

$$Asy.Var[b] = \frac{\sigma^2}{n} Q^{-1} \text{plim} \left(\frac{1}{n} X'\Omega X\right) Q^{-1}.$$

8.2.3 Robust Estimation of Asymptotic Covariance Matrices

If $\sigma^2\Omega$ were known, then the estimator of the asymptotic covariance matrix of b would be

$$Asy.Var[b] \equiv V_{OLS} = \frac{1}{n}\left(\frac{1}{n}X'X\right)^{-1}\left(\frac{1}{n}X'\sigma^2\Omega X\right)^{-1}\left(\frac{1}{n}X'X\right)^{-1}.$$

It might seem that to estimate $(1/n)X'\Sigma X$, an estimator of Σ, which contains $n(n+1)/2$ unknown parameters, is required. But fortunately, this is not quite right. What is required is an estimator of the $K(K+1)/2$ unknown elements in the matrix

$$Q_* = \text{plim } \frac{1}{n}X'\Sigma X = \text{plim } \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n}\sigma_{ij}x_i x_j'.$$

8.3 Efficient Estimation by Generalized Least Squares (GLS) I

Because Ω is a positive definite symmetric matrix, it can be factored into

$$\Omega = C\Lambda C',$$

where C are the characteristic vectors of Ω and the diagonal matrix Λ contains the characteristic roots of Ω (Eigenvalue decomposition, spectral decomposition).

Let $\Lambda^{1/2}$ be the diagonal matrix with ith diagonal element $\sqrt{\lambda_i}$.
Let $P' = C\Lambda^{-1/2}$, so $\Omega^{-1} = P'P$.

Premultiply the model by P:

$$\begin{aligned} Py &= PX\beta + P\varepsilon \\ y_* &= X_*\beta + \varepsilon_* \\ E[\varepsilon_*\varepsilon_*'|X_*] &= P\sigma^2\Omega P' = \sigma^2 I \end{aligned}$$

8.3 Efficient Estimation by GLS II

Because Ω is assumed to be known, y_* and X_* are observed data. In the classical model, ordinary least squares is efficient; hence,

$$\begin{aligned}\hat{\beta} &= (X_*'X_*)^{-1}X_*'y_* \\ &= (X'P'PX)^{-1}X'P'Py \\ &= (X'\Omega^{-1}X)^{-1}X'\Omega^{-1}y\end{aligned}$$

$\hat{\beta}$ is the efficient estimator of β. This estimator is the generalized least squares (GLS) estimator for β.

To summarize, all the results for the classical model, including the usual inference procedures, apply to the transformed model.

8.3 Efficient Estimation by Generalized Least Squares (GLS)

Properties of the Generalized Least Squares Estimator

If $E[\varepsilon_*|X_*] = 0$, then
$E[\hat{\beta}|X_*] = E[(X_*'X_*)^{-1}X_*'y_*|X_*] = \beta + E[(X_*'X_*)^{-1}X_*'\varepsilon_*|X_*] = \beta$.
The GLS estimator $\hat{\beta}$ is unbiased. The assumption is equivalent to $E[P\varepsilon|PX] = 0$, but because P is a matrix of known constants, we can keep the original assumption $E[\varepsilon|X] = 0$. The requirements that the regressors and disturbances be uncorrelated is unchanged. The GLS estimator is consistent if plim $(1/n)X_*'X_* = Q_*$, where Q_* is a positive definite matrix. Making the substitution, we see that this implies plim $[(1/n)X'\Omega^{-1}X]^{-1} = Q_*^{-1}$. The GLS estimator is asymptotically normally distributed, with mean β and sampling variance $Var[\hat{\beta}|X_*] = \sigma^2(X_*'X_*)^{-1} = \sigma^2(X'\Omega^{-1}X)^{-1}$.
The GLS estimator $\hat{\beta}$ is the minimum variance linear unbiased estimator in the generalized regression model. This statement follows by applying the Gauss-Markov theorem to the transformed model.

8.3.1 Problems regarding R^2

- R_*^2
  - May not be a constant in the transformed model, i.e. not bounded in the unit interval.
  - The transformed model is a computational device not the model of interest.
- $R_G^2 = 1 - \frac{(y - X\hat{\beta})'(y - X\hat{\beta})}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$
  - Not bounded in the unit interval.
  - No insurance that dropping a variable from the model will result in a decrease in R_G^2.
- $r_{y,\hat{y}}^2 = corr^2(y, \hat{y}) = corr^2(y, x'\beta)$. It is not a proportion of variance explained as in R^2; it is a measure of agreement of the model predictions.

8.3.2 Feasible Generalized Least Squares (FGLS)

With an unrestricted Ω, there are $n(n+1)/2$ additional parameters to estimate. This number is far too many to estimate with n observations. Obviously, some structure must be imposed on the model if we are to proceed: $\Omega(\theta)$.

Suppose then, that $\hat{\theta}$ is a consistent estimator of θ. Can we use $\hat{\Omega} = \Omega(\hat{\theta})$ to make GLS estimation feasible and in order to preserve the properties of the estimator for β?

Efficiency of the FGLS Estimator

An asymptotically efficient FGLS estimator does not require that we have an efficient estimator of θ; only a consistent one is required to achieve full efficiency for the FGLS estimator.

Except for the simplest cases, the finite-sample properties and exact distributions of FGLS estimators are unknown.

8.4 Heteroscedasticity

A heteroscedastic regression model:

$$
\begin{aligned}
Var[\varepsilon_i|X] &= \sigma_i^2, i = 1, ..., n \qquad Cov[\varepsilon_i, \varepsilon_j|X] = 0, i \neq j \\
E[\varepsilon\varepsilon'] &= \sigma^2\Omega = \sigma^2 \begin{pmatrix} \omega_1 & 0 & 0 & \cdots & 0 \\ 0 & \omega_2 & 0 & \cdots & 0 \\ & & & \vdots & \\ 0 & 0 & 0 & \cdots & \omega_n \end{pmatrix} \\
&= \begin{pmatrix} \sigma_1^2 & 0 & 0 & \cdots & 0 \\ 0 & \sigma_2^2 & 0 & \cdots & 0 \\ & & & \vdots & \\ 0 & 0 & 0 & \cdots & \sigma_n^2 \end{pmatrix}
\end{aligned}
$$

It will sometimes prove useful to write $\sigma_i^2 = \sigma^2\omega_i$. This form is an arbitrary scaling which allows us to use a normalization, $tr(\Omega) = \sum_{i=1}^{n} \omega_i = n$. This makes the classical regression a simple special case with $\omega_i = 1, i = 1, ..., n$.

8.4 When Does Heteroscedasticity Arise? I

If omitted variables are not correlated with the included variables (i.e., no omitted variable bias), but have a different order of magnitude for (groups of) observations.

- Cross-sectional data on units of different size, e.g. states, cities, firms. Omitted variables may be larger for more populous states or cities. For example, even for accounting for firm sizes, we expect to observe greater variation in the profits of large firms than in those of small ones.

- Cross-sectional data on units at different points in time. Omitted variables may be more important at some points in time. For example the income may vary greater with more years of work experience or volatility in the stock market.

8.4 When Does Heteroscedasticity Arise? II

- Cross sectional data on units that face different restrictions on their behavior. For instance, high income individuals have more discretion in their spending or profits of firms might depend on industry characteristics and consequently variation vary across firms of similar sizes.

8.4 How Can Heteroscedasticity Be Detected Graphically? I

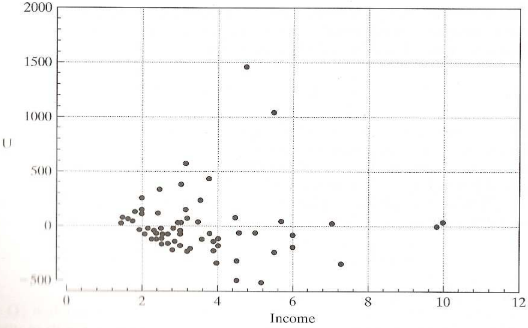


FIGURE 8.1 Plot of Residuals Against Income.

Figure 8.1: Regression of monthly expenditure on a constant, age, income and its square, dummy for home ownership.

8.4 How Can Heteroscedasticity Be Detected Graphically? II

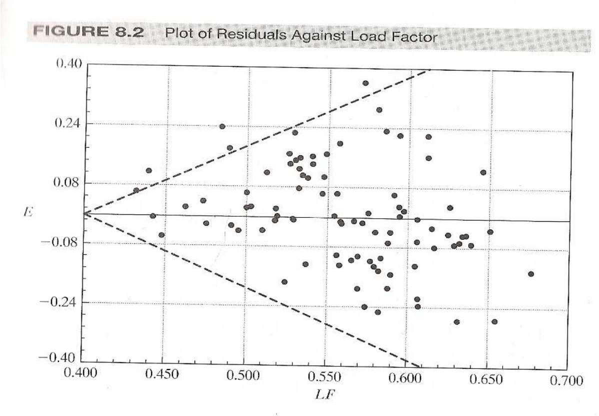


Figure 8.2: Cost function for the U.S airline industry.

$\ln C_i t = \beta_1 + \beta_2 Q_{it} + \beta_3(\ln(Q_{it}))^2 + \beta_4 \ln P_{it} + \varepsilon_{it}$, where C_{it} is total cost, Q_{it} is output, and P_{it} is the price of fuel. 6 firms observed for 15 years.

8.4 How Can Heteroscedasticity Be Detected Graphically? III

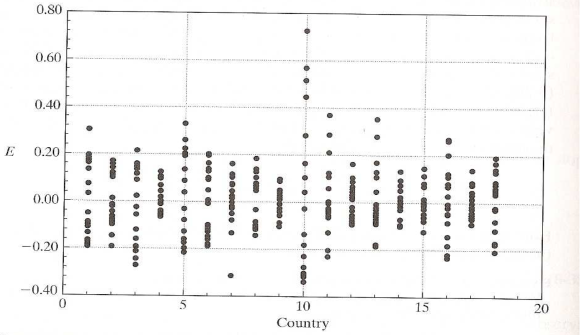


FIGURE 8.3 Plot of OLS Residuals by Country.

Figure 8.3: Gasoline consumption.

$\ln(Gasolineusage/capita)_{it} = \beta_1 + \beta_2 \ln(Percapitaincome_{it}) + \beta_3 \ln(Price_{it}) + \varepsilon_{it}$ plus country dummies.

8.4.1 Ordinary Least Squares Estimation

In the presence of heteroscedasticity, the least squares estimator b is still unbiased, consistent, and asymptotically normally distributed. The asymptotic covariance matrix is

$$\begin{aligned} Asy.Var[b] &= \frac{\sigma^2}{n}\left(\text{plim } \frac{1}{n}X'X\right)^{-1}\left(\text{plim } \frac{1}{n}X'\Omega X\right)\left(\text{plim } \frac{1}{n}X'X\right)^{-1} \\ &= \frac{\sigma^2}{n}Q^{-1}Q^{*}Q^{-1}. \end{aligned}$$

Estimation of the asymptotic covariance matrix would be based on

$$Var[b|X] = (X'X)^{-1}\left(\sigma^2 \textstyle\sum_{i=1}^{n} \omega_i x_i' x_i\right)(X'X)^{-1}.$$

Consequently, if $(X'X/n)$ and $(X'\Omega X)/n)$ converge to positive definite matrices, we obtain

$$b \sim^a N\left(\beta, \frac{\sigma^2}{n}Q^{-1}Q^{*}Q^{-1}\right)$$

8.4.2 Inefficiency of Least Squares I

b is inefficient relative to the GLS estimator. By how much will depend on the setting, but there is some generality to the pattern: The greater the dispersion in ω_i across observations, the greater the efficiency of GLS over OLS.

How erroneous are the conventional standard errors?

$$Est.Var[b|X] - Var[b|X]$$

$$\begin{aligned} &= s^2(X'X)^{-1} - \sigma^2(X'X)^{-1}(X'\Omega X)(X'X)^{-1} && \text{Large sample} \\ &\approx \sigma^2(X'X)^{-1}(X'X)(X'X)^{-1} - \sigma^2(X'X)^{-1}(X'\Omega X)(X'X)^{-1} \\ &= \sigma^2(X'X)^{-1}[(X'X) - (X'\Omega X)](X'X)^{-1} \end{aligned}$$

8.4.2 Inefficiency of Least Squares II

$$\Delta \;=\; \frac{X'X}{n} - \frac{X'\Omega X}{n} = \sum_{i=1}^{n} \frac{1}{n} x_i x_i' - \sum_{i=1}^{n} \frac{\omega_i}{n} x_i x_i' = \frac{1}{n} \sum_{i=1}^{n} (1 - \omega_i) x_i x_i'$$

The scaling $tr(\Omega) = n$ implies that $\sum_i (\omega_i / n) = 1$.

Whether the weighted average based on ω_i / n differs much from the one using $1/n$ depends on the weights. If the weights are related to the values in x_i, then the difference can be considerable. If the weights are uncorrelated with $x_i x_i'$ however, then the weighted average will tend to equal the unweighted average.

8.4.2 Inefficiency of Least Squares III

For example: Suppose that X contains a single column and that both x_i and ω_i are i.i.d. random variables. Then $x'x/n$ converges to $E[x_i^2]$, whereas $x'\Omega x/n$ converges to $Cov[\omega_i, x_i^2] + E[\omega_i]E[x_i^2]$. $E[\omega_i] = 1$, so if ω and x^2 are uncorrelated, then the sums have the same probability limit.

If the heteroscedasticity is not correlated with the variables in the model, then at least in large samples, the ordinary least squares computations, although not the optimal way to use the data, will not be misleading.

8.4.4 Estimating the Appropriate Covariance Matrix for Ordinary Least Squares I

We seek an estimator for

$$Q_* = \frac{1}{n}\sum_{i=1}^{n} \sigma_i^2 x_i x_i'.$$

White shows that under very general conditions, the estimator

$$S_0 = \frac{1}{n}\sum_{i=1}^{n} e_i^2 x_i x_i'$$

has

$$\text{plim } S_0 = \text{plim } Q_*.$$

8.4.4 Estimating the Appropriate Covariance Matrix for Ordinary Least Squares II

We obtain the White heteroscedastic consistent estimator

$$
\begin{aligned}
Est.Asy.Var[b] &= \frac{1}{n}\left(\frac{1}{n}X'X\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n} e_i^2 x_i x_i'\right)\left(\frac{1}{n}X'X\right)^{-1} \\
&= n(X'X)^{-1}S_0(X'X)^{-1}
\end{aligned}
$$

Without actually specifying the type of heteroscedasticity, we can still make appropriate inferences based on the results of least squares.

Two improvements by Davidson and MacKinnon regarding small/finite samples:

- Scaling up the standard errors by $\frac{n}{n-K}Est.Asy.Var[b]$.
- Using the squared residuals scaled by its true variance e_i^2/m_{ii} instead of e_i^2, where $m_{ii} = 1 - x_i'(X'X)^{-1}x_i$.

8.5 Testing for Heteroscedasticity I

Possible models:

$$\begin{aligned} \sigma_i^2 &= \alpha_1 + \alpha_2 z_{i2} + \alpha_3 z_{i3} + \ldots + \alpha_L z_{iL} && \text{Breusch Pagan} \\ & && \text{(Koenker variation)} \\ \sigma_i &= \alpha_1 + \alpha_2 z_{i2} + \alpha_3 z_{i3} + \ldots + \alpha_L z_{iL} && \text{Glesjer} \\ \ln(\sigma_i^2) &= \alpha_1 + \alpha_2 z_{i2} + \alpha_3 z_{i3} + \ldots + \alpha_L z_{iL} && \text{Harvey-Godfrey} \end{aligned}$$

$$z_{il} = \{x_{ij}, x_{ij}^2, x_{ij} \cdot x_{ik}\}$$

The Z's may be regressors or squares or products of regressors. If there is no idea what may cause heteroscedasticity then choose as Z's $x_2, \ldots x_K, x_2^2, \ldots, x_K^2, x_1 x_2, \ldots$, i.e. regressors, their squares and their crossproducts. The Breusch-Pagan test with this choice is the White's general test.

8.5 Testing for Heteroscedasticity II

Test procedure:

1. Obtain OLS residuals $e_i, i = 1, ..., n$.
2. Estimate the linear regression depending on the model of the test and compute R^2.
3. Compute the test statistic $LM = nR^2$ for the null hypothesis $H_o : \alpha_1 = 0, \alpha_2 = 0, ..., \alpha_L = 0$. If H_0 is true (homoscedastic errors) then LM has a χ^2 distribution with $L - 1$ degrees of freedom.

8.6 Weighted Least Squares When Ω is known I

Having tested for and found evidence of heteroscedasticity, the logical next step is to revise the estimation technique to account for it. The GLS estimator is obtained by regressing Py on PX.

$$Py = PX\beta + P\varepsilon \qquad Var[\varepsilon_i|X] = \sigma^2\omega_i$$

$$Py = \begin{pmatrix} y_1/\sqrt{\omega_1} \\ y_2/\sqrt{\omega_2} \\ \vdots \\ y_n/\sqrt{\omega_n} \end{pmatrix} \qquad PX = \begin{pmatrix} x_1'/\sqrt{\omega_1} \\ x_2'/\sqrt{\omega_2} \\ \vdots \\ x_n'/\sqrt{\omega_n} \end{pmatrix}$$

$$\begin{aligned} \hat{\beta} &= (X'\Omega^{-1}X)^{-1}X'\Omega^{-1}y \\ &= \left(\sum_{i=1}^{n} w_i x_i x_i'\right)^{-1}\left(\sum_{i=1}^{n} w_i x_i y_i\right) \qquad w_i = \frac{1}{\omega_i} \end{aligned}$$

Weighted Least Squares (WLS) Estimator

8.6 Weighted Least Squares When Ω is known II

Two important notes:

1. A common specification is that the variance is proportional to one of the regressors or its square: $\sigma_i^2 = \sigma^2 x_{ik}^2$, then the transformed regression model for GLS is

$$\frac{y}{x_k} = \beta_k + \beta_1 \left(\frac{x_1}{x_k}\right) + \beta_2 \left(\frac{x_2}{x_k}\right) + ... + \frac{\varepsilon}{x_k}$$

   - If the variance is proportional to any power of x_k other than two, no constant will be in the transformed model. Consequently, R^2 of no use.
   - The good fit of the weighted regression might be due to the presence of $1/x_k$ on both sides of the equality.

8.6 Weighted Least Squares When Ω is known III

2. The weighted least squares estimator is consistent regardless of the weights used, as long as the weights are uncorrelated with the disturbances.
   Using the wrong weights has two other less benign consequences
   - The improperly weighted least squares estimator is inefficient.
   - The standard errors will be incorrect.

The standard approach in the literature is to use OLS with the White estimator or some variant for the asymptotic covariance matrix.
(Using the wrong weights is better than using no weights at all?)

8.7 Estimation When Ω Contains Unknown Parameters I

The model is restricted by formulating $\sigma^2\Omega$ as a function of a few parameters, $\Omega(\alpha)$, e.g. $\sigma_i^2 = \sigma^2 x_i^\alpha$ or $\sigma_i^2 = \sigma^2(\mathbf{x}_i'\alpha)^2$.

FGLS based on a consistent estimator of $\Omega(\alpha)$ is asymptotically equivalent to full GLS. Two methods are typically used to find a consistent estimator for α: Two-step GLS and maximum likelihood.

For the heteroscedastic model, the GLS estimator is

$$\hat{\beta} = \left(\sum_{i=1}^n \left(\frac{1}{\sigma_i^2}\right) x_i x_i'\right)^{-1} \left(\sum_{i=1}^n \left(\frac{1}{\sigma_i^2}\right) x_i y_i\right).$$

The two-step estimators are computed by first obtaining estimates $\hat{\sigma}_i^2$, usually using some function of the ordinary least squares residuals (β_{OLS} although inefficient is still consistent.). Then plug these estimates in the above formula.

8.7 Estimation When Ω Contains Unknown Parameters II

For example, if $\sigma_i^2 = \mathbf{z}_i'\boldsymbol{\alpha}$, then a consistent estimator of $\boldsymbol{\alpha}$ will be the least squares slopes, $\mathbf{a}$, in the regression $e_i^2 = \mathbf{z}_i'\boldsymbol{\alpha} + \epsilon_i$. Consistency is all that is required for asymptotically efficient estimation of $\boldsymbol{\beta}$ using $\Omega(\hat{\boldsymbol{\alpha}})$.

Note: The two-step estimator may be iterated by recomputing the residuals after computing the FGLS estimates and then reentering the computation. Iteration in this context provides little additional benefit, if any.