LSTAT 2120
Linear Models
2020/21

Christian M. Hafner
LSBA, Université catholique de Louvain
christian.hafner@uclouvain.be

Contact

Prof. Dr. Christian M. Hafner

Université catholique de Louvain

Institut de statistique, biostatistique, et sciences actuarielles

Voie du Roman Pays, 20

B-1348 Louvain-la-Neuve, Belgium

Email: `christian.hafner@uclouvain.be`

 @CMHafner

LinkedIn cmhafner

Assistant: Stefka Asenova

Email: `stefka.asenova@uclouvain.be`

Web site

`http://moodleucl.uclouvain.be/course/view.php?id=5683`

Support:

- Slides on the web site
- Additional literature will be indicated throughout the lectures

Evaluation

1. An oral exam (60 %)
2. A project, where the material of the class is applied to real data. This project will be distributed during the semester. A written report has to be delivered before the exam session. (40 %).

Agenda

Lectures on Wednesday, 10h45, BARB91 (if not indicated otherwise)
The three dates for the exercise sessions are:

1. Thursday, October 22, 16h15
2. Thursday, November 5, 8h30
3. Friday, November 20, 16h15

The room will be c.045 in the CV09 building.

Prerequisites

- Basic course in mathematics (equivalent to LSTAT2011)
- Basic course in probability (equivalent to LSTAT2012)
- Basic course in statistics (equivalent to LSTAT2013)
- Linear algebra (equivalent to LINGE1121)

e-learning

You are invited to follow the following course (in french) sur edx:
`Introduction à l'économétrie`

It will teach basic statistics and econometrics, with many applications to case studies.

Table of contents

Recommended literature

I will use mainly the following sources:

KNNW: Kutner, Nachtsheim, Neter and Wasserman, Applied Linear Statistical Models, McGraw-Hill, Fifth edition (2005)

DM: Davidson and McKinnon, Econometric Theory and Methods, Oxford University Press (2004)

MOOC `Introduction à l'économétrie` , MOOC on edx

For the individual chapters I recommend the following texts:

1. The model, specification and interpretation: DM 1.3, MOOC Ch.2
2. Estimation and geometry: DM 2.3, 2.4
3. Statistical properties of OLS: DM 3.2-3.6, MOOC Ch.3
4. Maximum likelihood estimator: DM 10.2
5. Inference and hypothesis tests: KNNW 7.1-7.4, MOOC Ch.4
6. Multicollinearity: KNNW 7.6, 11.2
7. Discrete variables: KNNW 8.3, 8.5, MOOC Ch.5
8. Variable selection: KNNW 9.1-9.5
9. Heteroskedasticity, autocorrelation: DM 7.2-7.8, MOOC Ch. 7 and 8
10. Diagnostics (outliers, influential observations): KNNW 10.2-10.4, DM 2.6
11. Panel data: DM 7.10

Part I

The model, specification and interpretation

Objectives of a regression model

- Explain and predict a response variable, Y, by one (simple regression) or several (multiple regression) explanatory variables X.
- Estimate the marginal effects of one variable, controlling for the effects of the other variables.
- Test relevant relationships, significance of parameters, etc.

Simple regression

The regression estimates the average value of the response variable (Y) as a function of the explanatory variable (X), $\mathbb{E}[Y \mid X]$.
This conditional expectation is a function of x:

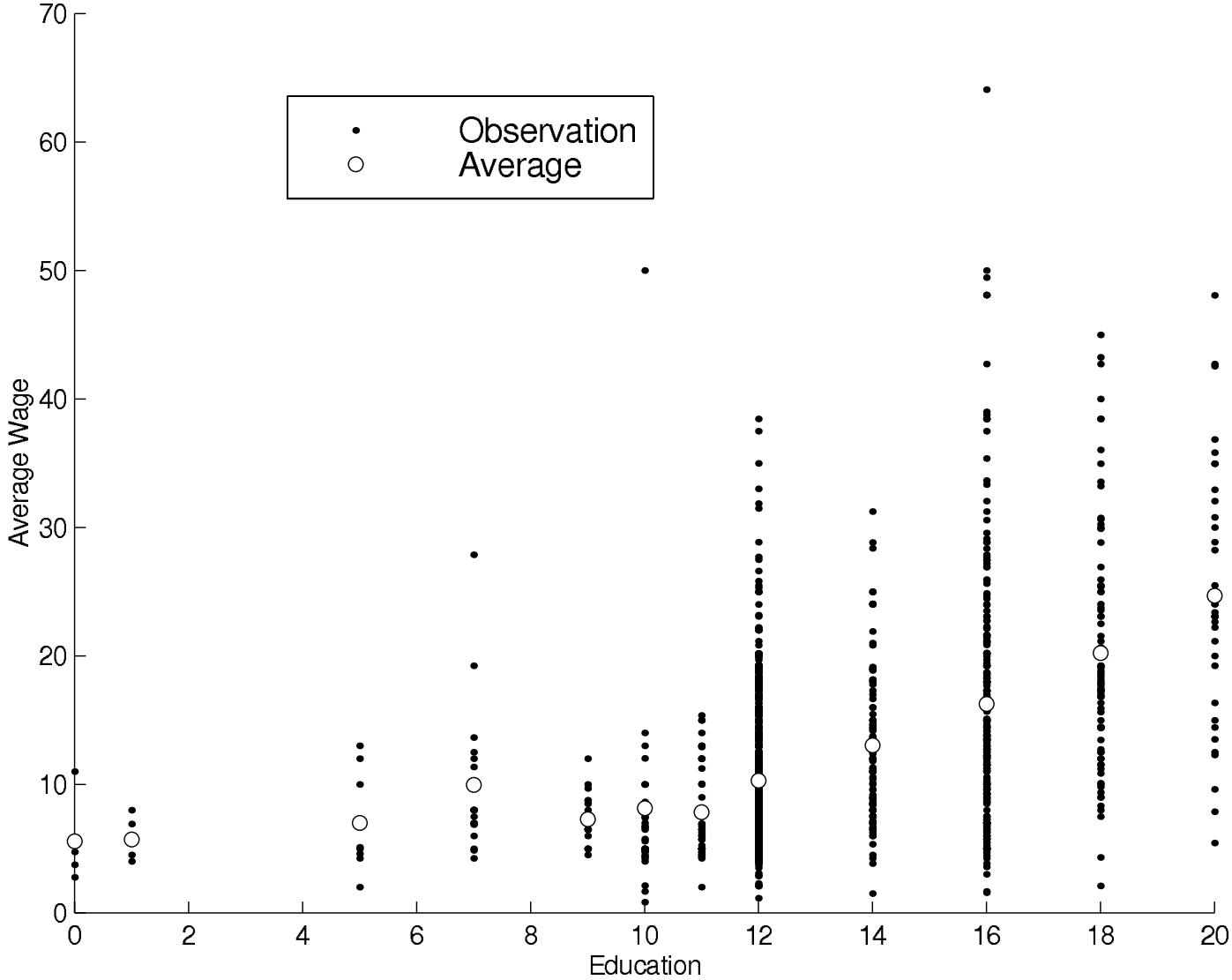
$$\mathbb{E}[Y \mid X = x] = f(x)$$

This *regression function* can be linear:

$$\mathbb{E}[Y \mid X = x] = \beta_1 + \beta_2 x$$

Example

Relation between the salary and the education level



OLS estimators for individuals with at least 11 years of higher education

Variable	Coefficient	St. Err.	t-stat	P-value
C	-9.56468	1.19172	-8.02597	[.000]
ED	1.64218	.086710	18.9388	[.000]

OLS estimators for individuals with less than 11 years of higher education

Variable	Coefficient	St. Err.	t-stat	P-value
C	6.27886	1.71133	3.66898	[.000]
ED	.212891	.212669	1.00104	[.320]

Multiple regression: examples

Example

- Y_i: Sales of a new product in region i
- X_{i1}: Population size in region i
- X_{i2}: Average salary in region i

Model:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \varepsilon_i, i = 1 \ldots, n,$$

Example

- Y_i: Grade of student i at an exam
- X_{i1}: Number of preparation hours for the exam
- X_{i2}: Binary variable: $X_{2i} = 1$ if student i is "talented", 0 otherwise.

Model:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \varepsilon_i, i = 1 \ldots, n,$$

Parameter of interest: β_1; Control variable: X_2.

Log-linear models

Example

(Cobb-Douglas production function)

- Y_i: Production (output)
- X_{i1}: entities of human resources in the production process
- X_{i2}: entities of capital

Model:

$$Y_i = A X_{i1}^{\alpha_1} X_{2i}^{\alpha_2} \eta_i,$$

where η_i is an error term.
Transformation:

$$\log Y_i = \log A + \alpha_1 \log X_{i1} + \alpha_2 \log X_{i2} + \log \eta_i$$

After transformation, the model becomes linear. The variables can appear in non-linear form.

The general multiple regression model

Consider the multiple regression model

$$Y_i = \beta_0 + \beta_1 X_{i1} + \ldots + \beta_{p-1} X_{i,p-1} + \varepsilon_i, \, i = 1 \ldots, n,$$

or equivalently in matrix form

$$\begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix} = \begin{pmatrix} 1 & X_{11} & \ldots & X_{1,p-1} \\ \vdots & \vdots & & \\ 1 & X_{n1} & \ldots & X_{n,p-1} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \vdots \\ \beta_{p-1} \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

or compactly,

$$Y = X\beta + \varepsilon$$

with the vectors $(n \times 1)$ Y and ε, the matrix $(n \times p)$ X and the parameter vector $(p \times 1)$, β. The number p is the number of parameters, including the intercept. For $p = 2$, the model reduces to a simple regression.

The parameters β and the errors ε are not observed.
$\Rightarrow$ we have to find estimators of β in order to interpret the link between X and Y.

Marginal effects

Consider the function

$$\mu : \mathbb{R}^p \to \mathbb{R} : X \mapsto \mathbb{E}(Y|X)$$

The marginal effect of a variable X_i is defined by

$$\frac{\partial \mu}{\partial x_i}$$

Elasticity

$$\begin{aligned}\nu &:= \text{relative variation of} \mathbb{E}(Y|X) \\ &= \frac{\partial \mathbb{E}(Y|X)}{\partial X} \frac{X}{\mathbb{E}(Y|X)}\end{aligned}$$

The elasticity measures a proportional change of Y as a function of a proportional change of X.

Example: linear model

If

$$\mathbb{E}(Y|X) = \beta_1 + \beta_2 X$$

then the marginal effect with respect to X is

$$\beta_2$$

and the elasticity is

$$\beta_2 X / \mathbb{E}(Y|X)$$

Example: log-linear model

If

$$\log(\mathbb{E}(Y|X)) = \beta_1 + \beta_2 \log(X)$$

then the marginal effect with respect to X is

$$\beta_2 \mathbb{E}(Y|X)/X$$

and the elasticity is

$$\beta_2$$

If

$$\log(\mathbb{E}(Y|X)) = \beta_1 + \beta_2 X$$

then the marginal effect with respect to X is

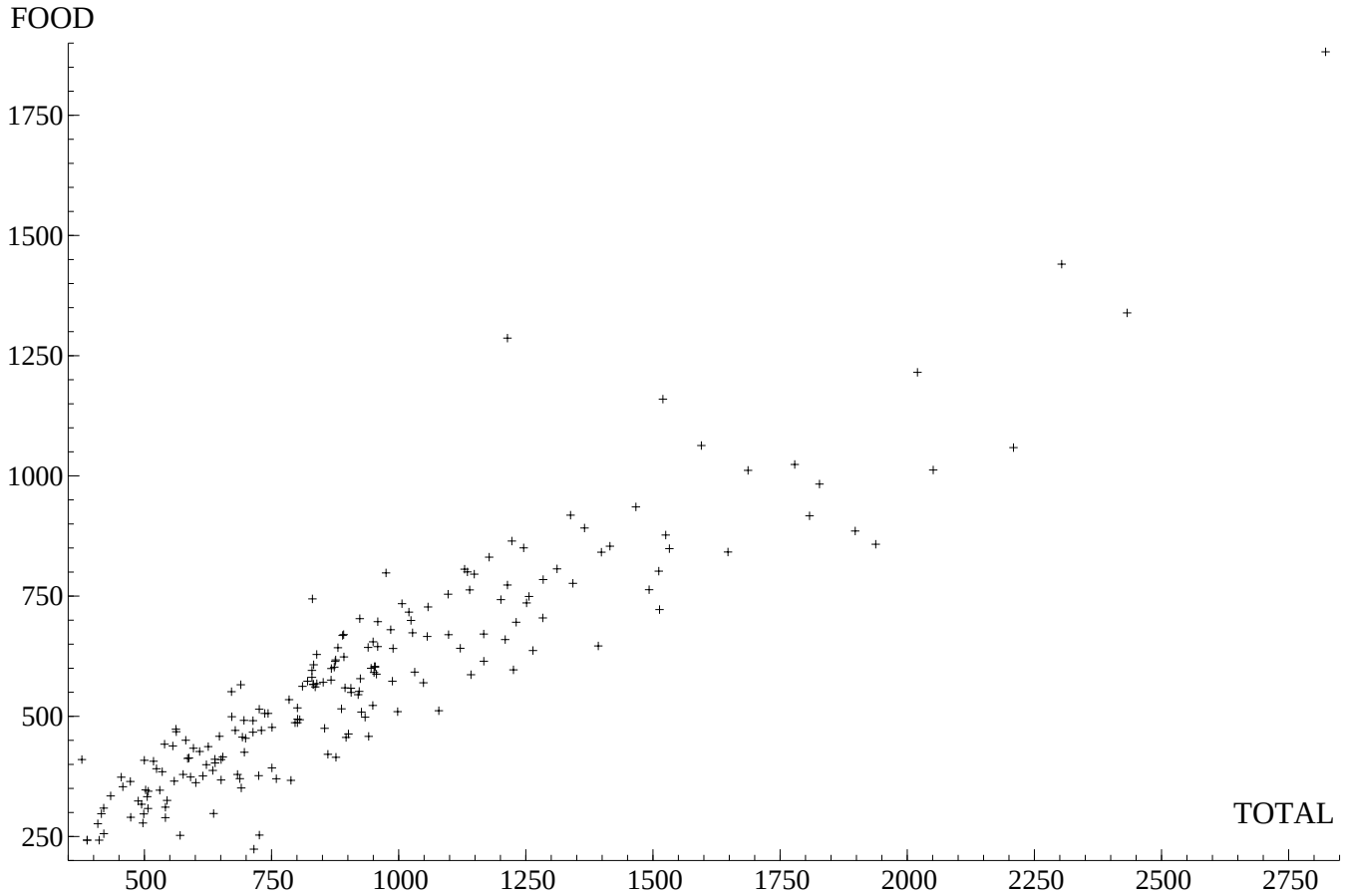
$$\beta_2 \mathbb{E}(Y|X)$$

and the elasticity is

$$\beta_2 X$$

Example

Relation between food expenditures (FOOD) of 198 Belgian households and their revenus, measured by total expenditures (TOTAL). Data of Edouard Ducpetiaux, 1855.



The law of Engel (1857)

Relation between expenditures for food Y and revenues X:
$\mathbb{E}(Y|X) = f(X)$.

Engel (1857) postulates: *The marginal propensity to consume food df/dX, is between 0 and 1. Moreover, the average propensity to consume, f/X, decreases when the revenues increase.*

If f is linear, these hypotheses can be expressed as constraints on the parameters.

If the model is linear,

$$f(X) = \beta_1 + \beta_2 X$$

and the Engel law implies that $\beta_1 > 0$ et $0 < \beta_2 < 1$.
These hypotheses can be tested using empirical observations and statistical methods.

OLS estimation output for the linear model:

	Coefficient	Std.Error	t-prob	Part.R^2
Constant	84.1009	16.60	0.0003	0.0656
TOTAL	0.533252	0.01629	0.0000	0.6630

sigma	94.0812	RSS	1734848.21
R^2	0.845303	F(1,196) =	1071 [0.000]**
log-likelihood	-1179.69	DW	1.6
no. of observations	198	no. of parameters	2
mean(FOOD)	581.422	var(FOOD)	56638.9

The condition $d(f/X)/dX < 0$ is equivalent to the condition that the **elasticity** of f with respect to X is smaller than 1.
For example, in the log-linear model

$$\log f(X) = \beta_1 + \beta_2 \log(X)$$

the Engel law implies that $0 < \beta_2 < 1$.
Again, we will be able to test this hypothesis.

OLS estimation output for the log-linear model:

	Coefficient	Std.Error	t-prob	Part.R^2
Constant	0.432916	0.1883	0.0221	0.0264
LTOT	0.867333	0.02784	0.0000	0.8349

sigma	0.157366	RSS	4.85377182
R^2	0.832022	F(1,196) =	970.8 [0.000]**
log-likelihood	86.1928	DW	1.54
no. of observations	198	no. of parameters	2
mean(LFOOD)	6.29116	var(LFOOD)	0.145936

Part II

Estimation and geometry

Estimation of β

Objective: minimize the distance of the observations from the regression function, taking as distance the sum of squared residuals. We define the loss function

$$\begin{aligned} S(\beta) &= \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_{i1} - \ldots - \beta_{p-1} X_{i,p-1})^2 \\ &= (Y - X\beta)'(Y - X\beta) \end{aligned}$$

The least squares estimator is then implicitly defined by the equation $dS(\beta)/d\beta = 0$.

Convention for the derivative of a scalar-valued function with respect to a vector of parameters:

$$\frac{dS(\beta)}{d\beta} = \begin{pmatrix} \frac{\partial S(\beta)}{\partial \beta_0} \\ \vdots \\ \frac{\partial S(\beta)}{\partial \beta_{p-1}} \end{pmatrix}$$

Calculating the partial derivatives gives the following result:

$$\frac{\partial S(\beta)}{\partial \beta_j} = -2 \sum_{i=1}^{n} X_{ij}(Y_i - \beta_0 - \beta_1 X_{i1} - \ldots - \beta_{p-1} X_{i,p-1})$$

In matrix form we obtain

$$\frac{dS(\beta)}{d\beta} = -2X'Y + 2X'X\beta.$$

This derivative, evaluated at the least squares estimator, is equal to zero:

$$\left.\frac{dS(\beta)}{d\beta}\right|_{\beta=\hat{\beta}} = 0$$

We thus obtain the system of equations:

$$X'X\hat{\beta} = X'Y$$

which are often called the normal equations.
If X has full column rank ($rk(X) = p$), then $rk(X'X) = p$ and $X'X$ is invertible. From the normal equations, we get the unique solution for the least squares estimator, also called Ordinary Least Squares (OLS) estimator:
$\hat{\beta}$:

$$\hat{\beta} = (X'X)^{-1}X'Y \tag{1}$$

Components of this expression: $(X'X)^{-1}$ is a $p \times p$ matrix, and $X'Y$ a $p \times 1$ vector. The product gives a $p \times 1$ vector, the required dimension for the estimator of β.
The formula requires the inversion of a $p \times p$ matrix. If this matrix is ill-conditioned (smallest eigenvalue very close to zero), then this can be numerically unstable. We will discuss remedies in the chapter on multicollinearity.

We have derived the OLS estimator $\hat{\beta} = (X'X)^{-1}X'Y$ of the parameter β in the model

$$Y = X\beta + \varepsilon.$$

The estimated model will be written as

$$Y = X\hat{\beta} + e$$

where e is the $(n \times 1)$ vector of residuals.
We will call $\hat{Y} = X\hat{\beta}$ the fitted values of the response variable.

Simple regression

For the special case of a simple regression (i.e., $p = 2$) we obtain

$$X'X = \begin{pmatrix} n & \sum_{i=1}^n X_i \\ \sum_{i=1}^n X_i & \sum_{i=1}^n X_i^2 \end{pmatrix}$$

and

$$X'Y = \begin{pmatrix} \sum_{i=1}^n Y_i \\ \sum_{i=1}^n X_i Y_i \end{pmatrix}$$

As it can be checked, after inversion of the (2×2) matrix $X'X$, we obtain the well known expressions

$$\begin{aligned} \hat{\beta}_1 &= \frac{\sum_i (Y_i - \bar{Y})(X_i - \bar{X})}{\sum_i (X_i - \bar{X})^2} \\ \hat{\beta}_0 &= \bar{Y} - \hat{\beta}_1 \bar{X} \end{aligned}$$

Link between correlation and simple regression

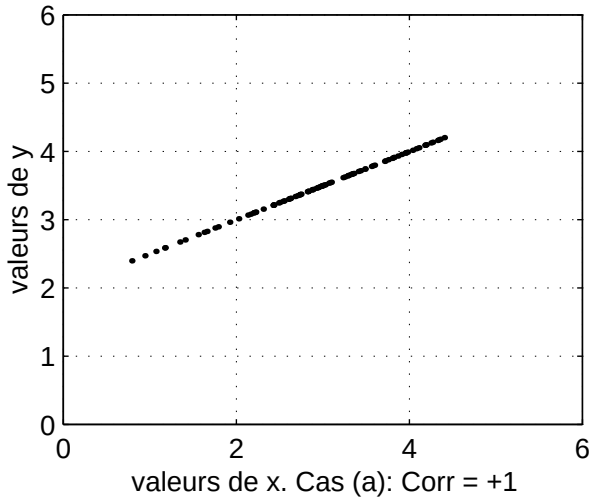
Recall that the empirical correlation coefficient between X and Y is given by

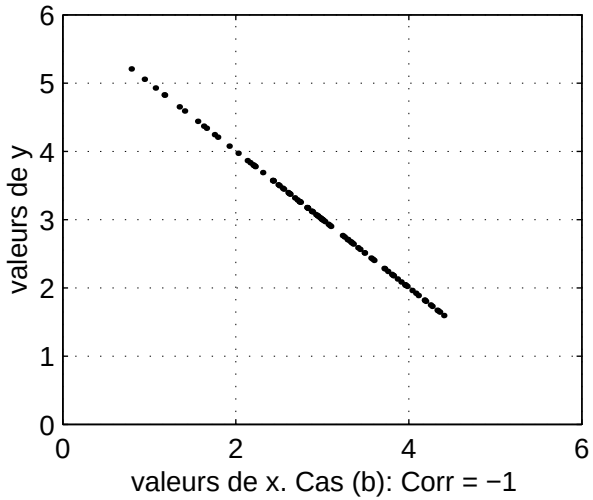
$$r = \frac{\sum_i (Y_i - \bar{Y})(X_i - \bar{X})}{\sqrt{\sum_i (X_i - \bar{X})^2}\sqrt{\sum_i (Y_i - \bar{Y})^2}}$$

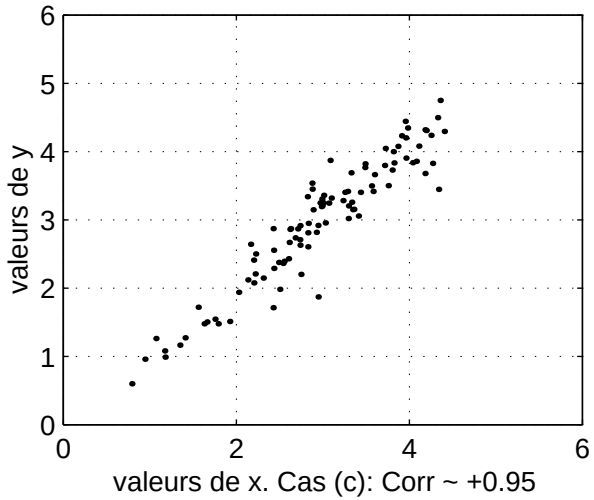
and r estimates the population correlation coefficient

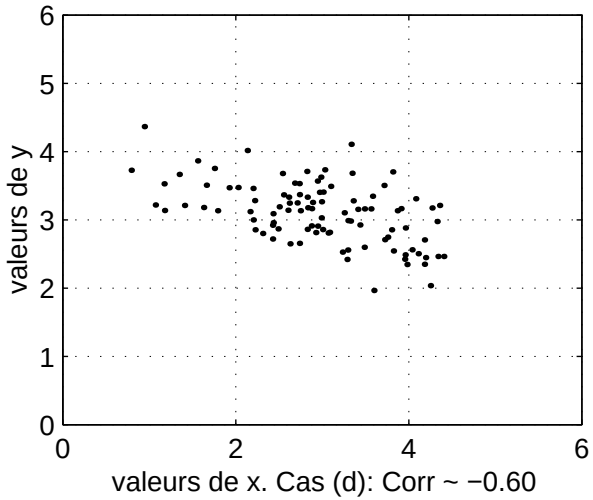
$$\rho = \frac{\mathrm{Cov}(X, Y)}{\sqrt{\mathrm{Var}(X)}\sqrt{\mathrm{Var}(Y)}}$$

Correlation is a standardized measure for the linear dependence between X et Y. Independence implies $\rho = 0$, but the reverse is not true in general.


valeurs de y
valeurs de x. Cas (a): Corr = +1


valeurs de y
valeurs de x. Cas (b): Corr = −1


valeurs de y
valeurs de x. Cas (c): Corr ~ +0.95


valeurs de y
valeurs de x. Cas (d): Corr ~ −0.60

Link between $\hat{\beta}_1$ and r:

$$r = \hat{\beta}_1 \frac{\sqrt{\sum_i (X_i - \bar{X})^2}}{\sqrt{\sum_i (Y_i - \bar{Y})^2}}$$

Therefore, r and $\hat{\beta}_1$ always have the same sign.

Geometric interpretation

Preliminary definitions:

Definition

An $(m \times m)$ matrix A is idempotent if $A^2 = A$.

Examples: the identity matrix I_m, or the matrix

$$A = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}$$

Note that, for any $(m \times m)$ matrix A with $rk(A) = m$, $A(A'A)^{-1}A'$ is idempotent.

Properties of an idempotent matrix

Let $A(m \times m)$ be an idempotent matrix. Then,

1. A' is idempotent
2. $I_m - A$ is idempotent
3. All eigenvalues of A are either 0 or 1.
4. $\text{rk}(A) = \text{tr}(A)$

Projection matrix

Definition (Projection matrix)

An $(m \times m)$ matrix A is a projection matrix if it is symmetric and idempotent.

Trivial example: the identity matrix.
Idempotence ensures that a projected vector remains unchanged after iterated projections.
Symmetry ensures that the projection is orthogonal.

Column space of a matrix

The column space of an $(n \times p)$ matrix X is the sub-space of $\mathbb{R}^n$ which consists of all linear combinations of the columns of X:

$$\mathcal{C}(X) = \{z \in \mathbb{R}^n | z = Xb, b \in \mathbb{R}^p\}$$

The dimension of this sub-space is equal to the rank of X, that is:

$$\dim(\mathcal{C}(X)) = p$$

We will later often call the rank of a projection matrix its degrees of freedom.

The "hat" matrix H

The $(n \times n)$ matrix $H = X(X'X)^{-1}X'$ is a projection matrix.
The operation Hy projects the vector y onto the space $\mathcal{C}(X)$.
It is an orthogonal projection:

$$(y - Hy)'X = y'(I_n - H)X = y'(X - X) = 0$$

The matrix M

The matrix $M = I_n - X(X'X)^{-1}X'$ is also a projection matrix. It projects a vector onto the orthogonal complement of $\mathcal{C}(X)$, denoted as $\mathcal{C}^{\perp}(X)$. The dimension of this space is:

$$\mathbf{dim}(\mathcal{C}^{\perp}(X)) = \mathrm{rk}(M) = \mathrm{tr}(M) = n - p$$

Orthogonal decomposition of $\mathbb{R}^n$

The projection matrices H and M define complementary projections because $H + M = I_n$. That is, $Hy + My$ reconstructs the original vector, y.
By idempotence, $H + M = I_n$ implies that $HM = 0$.
The orthogonality of the spaces $\mathcal{C}(X)$ and $\mathcal{C}^{\perp}(X)$ is due to the symmetry: Let $z \in \mathcal{C}(X)$ and $w \in \mathcal{C}^{\perp}(X)$. We have $z = Hz$ and $w = Mw$. Thus, the scalar product of the two vectors z and w is:

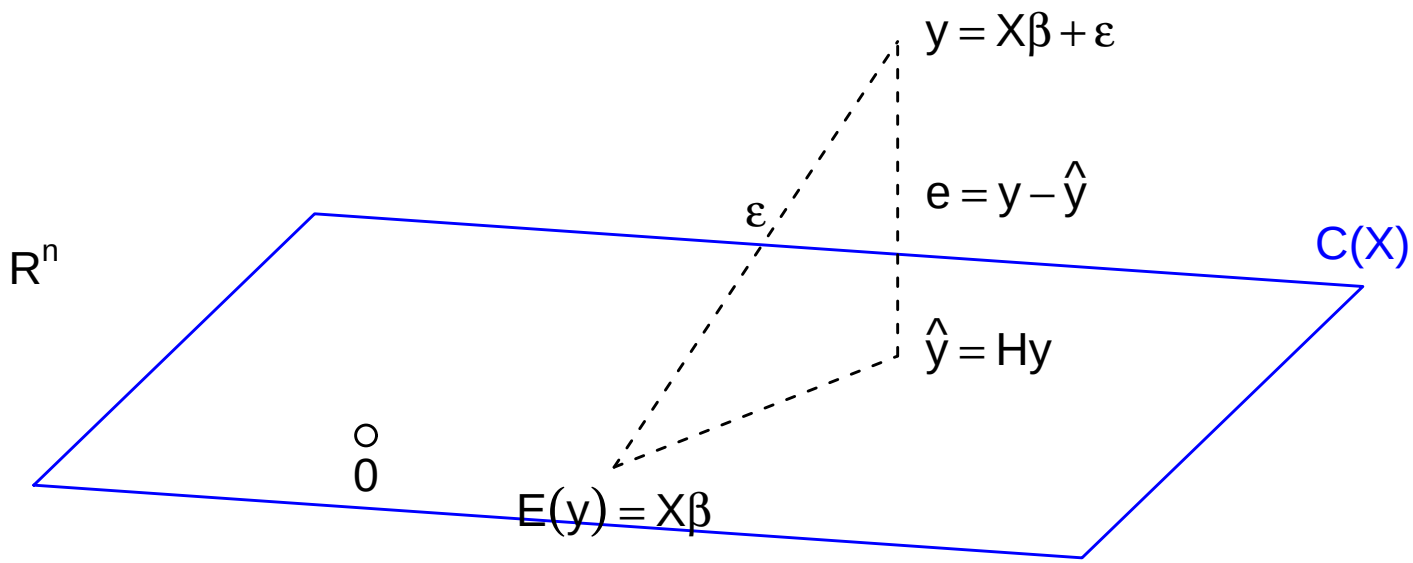
$$z'w = z'H'Mw = z'HMw = 0$$

For example, we can interpret the fitted regression as

$$\begin{aligned} Y &= X\hat{\beta} + e \\ &= HY + MY \end{aligned}$$

- The fitted values are obtained by projecting Y onto the space $\mathcal{C}(X)$
- The residuals are obtained by projecting Y onto the space $\mathcal{C}^{\perp}(X)$

Orthogonal projection of Y onto $\mathcal{C}(X)$



Residuals

Note that

$$e = MY = M(X\beta + \varepsilon) = M\varepsilon$$

and, consequently,

$$e'X = \varepsilon' MX = 0$$

The vector of residuals is orthogonal with respect to all explanatory variables.
In particular, if there is a constant intercept in the model, then

$$e'\iota = 0$$

where $\iota = (1, \ldots, 1)'$ $(n \times 1)$, which implies that the residuals have a mean zero.

Partitioned regression

Suppose we have two groups of regressors, X_1 and X_2 of dimension $n \times p_1$ and $n \times p_2$, $p_1 + p_2 = p$. We can write the regression $Y = X\beta + \varepsilon$ as

$$Y = X_1\beta_1 + X_2\beta_2 + \varepsilon \tag{2}$$

There are two approaches to estimate β_2:

1. Estimate model (2) by OLS
2. Calculate the residuals of a regression of Y on X_1, and regress them by OLS on the residuals of a regression of X_2 on X_1.

The second approach is the OLS estimator of β_2 in the regression

$$M_1 Y = M_1 X_2 \beta_2 + u$$

with $M_1 := I_n - H_1$ and $H_1 := X_1(X_1'X_1)^{-1}X_1'$.
Note that

$$H_1 H = H H_1 = H_1,$$

because $HX_1 = X_1$, as all columns of X_1 are contained in $\mathcal{C}(X)$.
It follows that

$$M_1 M = (I_n - H_1)(I_n - H) = I_n - H_1 - H + H_1 H = I_n - H = M$$

Frisch-Waugh-Lovell Theorem

Theorem (Frisch-Waugh-Lovell)

The two estimation approaches of β_2 are equivalent. Moreover, the residuals of both regressions are identical.

Proof

The estimator of the second approach is

$$(X_2' M_1 X_2)^{-1} X_2' M_1 y$$

The estimators $\hat{\beta}_1$ and $\hat{\beta}_2$ of the first approach are such that

$$Y = X_1 \hat{\beta}_1 + X_2 \hat{\beta}_2 + MY \tag{3}$$

Multiplying (3) by $X_2' M_1$ yields $X_2' M_1 Y = X_2' M_1 X_2 \hat{\beta}_2$ because $M_1 X_1 = 0$ and $X_2' M_1 M = X_2' M = 0$. Therefore, we obtain

$$\hat{\beta}_2 = (X_2' M_1 X_2)^{-1} X_2' M_1 y,$$

which shows the first part of the theorem.

For the second part, multiply (3) by M_1:

$$M_1 y = M_1 X_2 \hat{\beta}_2 + MY$$

The response variable of this regression is the same as that of the second approach.
Because $\hat{\beta}_2$ is the estimator of this regression, the first term on the right hand side has to be equal to the fitted values of the second approach.
Consequently, the residuals MY have to be the same for the two approaches, which shows the second part of the theorem.

Fundamental decomposition

Let $\bar{Y} = n^{-1} \sum_{i=1}^{n} Y_i$ be the empirical mean of Y, and $\iota = (1, \ldots, 1)'$ an $(n \times 1)$ vector. Note that $\bar{Y} = \iota' Y / n$.

We define the following quantities:

$$
\begin{aligned}
SST &= \sum_{i=1}^{n} (Y_i - \bar{Y})^2 = (Y - \bar{Y}\iota)'(Y - \bar{Y}\iota) \\
SSE &= \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 = (Y - \hat{Y})'(Y - \hat{Y}) \\
SSR &= \sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2 = (\hat{Y} - \bar{Y}\iota)'(\hat{Y} - \bar{Y}\iota)
\end{aligned}
$$

For the total variation, we can write

$$\begin{aligned} SST &= (Y - \bar{Y}\iota)'(Y - \bar{Y}\iota) \\ &= (Y - \iota\frac{\iota' Y}{n})'(Y - \iota\frac{\iota' Y}{n}) \\ &= \{(I_n - \frac{\iota\iota'}{n})Y\}'\{(I_n - \frac{\iota\iota'}{n})Y\} \\ &= Y'(I_n - \frac{\iota\iota'}{n})(I_n - \frac{\iota\iota'}{n})Y \\ &= Y'(I_n - \frac{\iota\iota'}{n})Y \end{aligned}$$

where $(I_n - \frac{\iota\iota'}{n})$ is a projection matrix of rank $n-1$. That means, SST has $n-1$ degrees of freedom.

$$
\begin{aligned}
SSE &= (Y - \hat{Y})'(Y - \hat{Y}) \\
&= (Y - HY)'(Y - HY) \\
&= \{(I_n - H)Y\}'\{(I_n - H)Y\} \\
&= \{MY\}'\{MY\} \\
&= Y'MMY \\
&= Y'MY
\end{aligned}
$$

where $M = I_n - X(X'X)^{-1}X$ is a projection matrix of rank $n - p$. That means, SSE has $n - p$ degrees of freedom.

$$
\begin{aligned}
SSR &= (\hat{Y} - \bar{Y}\iota)'(\hat{Y} - \bar{Y}\iota) \\
&= (HY - \iota\frac{\iota' Y}{n})'(HY - \iota\frac{\iota' Y}{n}) \\
&= \{(H - \frac{\iota\iota'}{n})Y\}'\{(H - \frac{\iota\iota'}{n})Y\} \\
&= Y'(H - \frac{\iota\iota'}{n})(H - \frac{\iota\iota'}{n})Y \\
&= Y'(H - \frac{\iota\iota'}{n})Y
\end{aligned}
$$

where $(H - \frac{\iota\iota'}{n})$ is a projection matrix of rank $p - 1$. That means, SSR has $p - 1$ degrees of freedom.
Then we have the following fundamental decomposition.

The sum of squares and associated degrees of freedom can be decomposed in the following way:

Theorem

$$\begin{aligned} SST &= SSR + SSE \\ n-1 &= (p-1) + (n-p) \end{aligned}$$

Coefficient of determination

A measure for the goodness-of-fit of a linear model is given by the coefficient of determination, R^2. It is defined by

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$$

and measures the proportion of the total variation, SST, that is explained by the variables $X_1, \ldots, X_{p-1}$. As $SST = SSR + SSE$, it is clear that $0 \leq R^2 \leq 1$.

Geometric interpretation:
The R^2 is equal to the square cosine of the angle between the vectors Y and $\hat{Y}$.

To take the number of variables into account, the adjusted R^2 is defined as

$$R_a^2 = 1 - \frac{SSE}{SST}\,\frac{n-1}{n-p}$$

The correction factor $\frac{n-1}{n-p}$ increases by adding variables to the regression (p increases), and therefore the R_a^2 can decrease by adding variables.

For the special case of a simple regression, $p = 2$, one can show (simple exercise) that the R^2 takes a simple form:

$$R^2 = r^2 = \frac{\{\sum_i (X_i - \bar{X})(Y_i - \bar{Y})\}^2}{\sum_i (X_i - \bar{X})^2 \sum_i (Y_i - \bar{Y})^2}$$

which is the squared correlation between X and Y (r^2). In other words, the quality of fit of a simple regression is directly related to the correlation between the response and the explanatory variable.

Part III

Statistical properties of OLS

Notation

We will write $\mathbf{X}$ for the matrix $(n \times p)$ of explanatory variables in the model

$$Y = \mathbf{X}\beta + \varepsilon.$$

On the other hand, we will denote X_i the vector $p \times 1$, which contains the explanatory variables of the i-th observation, and by X the vector $p \times 1$ of explanatory variables of a generic observation (fixed or random).

The classical assumptions

We suppose that

1. $rk(\mathbf{X}) = p$
2. $\mathbf{X}$ is a fixed matrix (deterministic)
3. $\varepsilon_1, \ldots, \varepsilon_n$ are independent
4. $\mathbb{E}[\varepsilon] = 0$
5. $\text{Var}(\varepsilon) = \sigma^2 I_n$

Assumption 5 requires that the variances are constant (homoskedasticity) and all co-variances are zero.

Some results of linear algebra

Let A and B be $(n \times n)$ matrices.

Lemma

1. $(AB)' = B'A'$
2. $(AB)^{-1} = B^{-1}A^{-1}$, *if A and B are invertible*
3. $tr(AB) = tr(BA)$

Some useful results

Let Y be an $(n \times 1)$ random vector with expectation μ and variance-covariance Σ, A an $(n \times n)$ fixed matrix, and b an $(n \times 1)$ fixed vector.

Lemma

1. $\mathbb{E}[AY + b] = A\mu + b$
2. $Var(AY + b) = A\Sigma A'$
3. $\mathbb{E}[Y'AY] = tr\{A\Sigma\} + \mu' A\mu$

Properties of $\hat{\beta}$

The properties of the OLS estimator can be summarized by the following theorem.

Theorem

1. $E[\hat{\beta}] = \beta$
2. $Var(\hat{\beta}) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$
3. *If* $\varepsilon \sim N(0, \sigma^2 I_n)$, *then* $\hat{\beta} \sim N(\beta, \sigma^2(\mathbf{X}'\mathbf{X})^{-1})$

Definition (Mean square error)

The mean square error (MSE) of an estimator $\hat{\theta}$ is defined by

$$MSE(\hat{\theta}) = \mathbb{E}[(\hat{\theta} - \theta)^2]$$

Lemma

$$MSE(\hat{\theta}) = Bias(\hat{\theta})^2 + Var(\hat{\theta})$$

Definition (Efficiency, one-dimensional)

Let $\theta^{(1)}$ and $\theta^{(2)}$ be two estimators of a parameter $\theta \in \mathbb{R}$. We say that $\theta^{(1)}$ is more efficient than $\theta^{(2)}$ if

$$\mathbb{E}[(\theta^{(1)} - \theta)^2] \leq \mathbb{E}[(\theta^{(2)} - \theta)^2]$$

for all $\theta \in \mathbb{R}$. This definition implies that, if $\mathbb{E}[\theta^{(1)}] = \mathbb{E}[\theta^{(2)}] = \theta$, then $\theta^{(1)}$ is more efficient than $\theta^{(2)}$ if $Var(\theta^{(1)}) \leq Var(\theta^{(2)})$ for all $\theta \in \mathbb{R}$.

Definition (Efficiency, multi-dimensional)

Let $\theta^{(1)}$ and $\theta^{(2)}$ two estimators of a parameter $\theta \in \mathbb{R}^k, k > 1$. We say that $\theta^{(1)}$ is more efficient than $\theta^{(2)}$ if

$$\mathbb{E}[(a'\theta^{(1)} - a'\theta)^2] \leq \mathbb{E}[(a'\theta^{(2)} - a'\theta)^2]$$

for all $\theta \in \mathbb{R}^k$ and all $a \in \mathbb{R}^k$. This implies that, if $\mathbb{E}[\theta^{(1)}] = \mathbb{E}[\theta^{(2)}] = \theta$, then $\theta^{(1)}$ is more efficient than $\theta^{(2)}$ if

$$a' Var(\theta^{(1)}) a \leq a' Var(\theta^{(2)}) a$$

for all $\theta \in \mathbb{R}^k$ and all $a \in \mathbb{R}^k$.

Definition (Linear estimator in Y)

Let $\hat{\theta}$ be an estimator of $\theta \in \mathbb{R}^k$, $k \geq 1$. We say that the estimator is linear in Y if there exists a matrix A of dimension $k \times n$ such that $\hat{\theta} = AY$.

We note that by equation (1), the OLS estimator is linear in Y.

Theorem (Gauss-Markov)

Let $\tilde{\beta}$ be a linear unbiased estimator of β and let $\hat{\beta}$ be the OLS estimator (1). Then,

$$a' Var(\hat{\beta}) a \leq a' Var(\tilde{\beta}) a$$

for all $\theta, a \in \mathbb{R}^p$.

The OLS estimator is the best linear unbiased estimator (BLUE), meaning that it has the smallest variance of all linear unbiased estimators.

Proof

As $\tilde{\beta}$ is linear, it can be written as $\tilde{\beta} = AY$ for some fixed $(p \times n)$ matrix A. The unbiasedness condition of $\tilde{\beta}$ requires that

$$\mathbb{E}[\tilde{\beta}] = \mathbb{E}[AY] = A\mathbf{X}\beta = \beta$$

for all β, which implies that $A\mathbf{X} = I_p$. Calculate the variance-covariance matrix of $\tilde{\beta}$:

$$
\begin{aligned}
\mathrm{Var}(\tilde{\beta}) &= \mathbb{E}[(\tilde{\beta}-\beta)(\tilde{\beta}-\beta)'] \\
&= \mathbb{E}[(AY - A\mathbf{X}\beta)(AY - A\mathbf{X}\beta)'] \\
&= \mathbb{E}[A(Y - \mathbf{X}\beta)(Y - \mathbf{X}\beta)'A'] \\
&= A\mathbb{E}[\varepsilon\varepsilon']A' \\
&= \sigma^2 AA' \\
&= \sigma^2\left\{AA' - (\mathbf{X}'\mathbf{X})^{-1}\right\} + \sigma^2(\mathbf{X}'\mathbf{X})^{-1} \\
&= \sigma^2\left\{(A - (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')(A - (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')'\right\} + \mathrm{Var}(\hat{\beta})
\end{aligned}
$$

The first term on the right hand side is a positive semi-definite matrix, which shows that the difference between the variance-covariance matrix of $\tilde{\beta}$ and that of $\hat{\beta}$ is positive semi-definite. □

Difference between least squares and maximum likelihood?

Theorem

If $\varepsilon \sim N(0, \sigma^2 I_n)$, then the OLS estimator of β is the same as the maximum likelihood estimator.

Under normality of the error terms, there is hence no difference between the two estimators. We study maximum likelihood estimators a bit more in detail later.

Convergence in probability

The sequence Z_n converges to the value θ in probability, if

$$\lim_{n\to\infty} P(|Z_n - \theta| < \varepsilon) = 1, \quad \forall \varepsilon > 0$$

We will then write that

$$Z_n \xrightarrow{p} \theta, \quad \text{ou} \quad \text{plim}(Z_n) = \theta$$

A sufficient condition is that $\mathbb{E}[Z_n] = \theta$ and $\lim_{n\to\infty} \text{Var}(Z_n) = 0$. One can use the Chebycheff inequality to show this result.

The law of large numbers (LLN)

- Let $Y_1, Y_2, \ldots, Y_n$ i.i.d. random variables with mean μ and variance $\sigma^2 < \infty$.
- We then can show that

$$\overline{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i \xrightarrow{p} \mu$$

Indeed, the sufficient condition for convergence in probability holds:

$$\begin{aligned} \mathbb{E}[\overline{Y}] &= \mu \\ \lim_{n\to\infty} \mathrm{Var}(\overline{Y}) &= \lim_{n\to\infty} \frac{\sigma^2}{n} = 0, \end{aligned}$$

Assumptions concerning the explanatory variables

When $n \to \infty$, we have to add an assumption about $(n \times p)$ $\mathbf{X}$, still assumed to be fixed, but whose dimension grows to infinity:
We assume that

$$\frac{1}{n}\sum_{i=1}^{n} X_i X_i' \to \mathbf{\Sigma}$$

where Σ is a $(p \times p)$ positive definite matrix, and X_i is the p-vector of the i-th observation.

Consistency of the OLS estimator

Consider the decomposition:

$$\hat{\beta} - \beta = \left(\frac{1}{n} \sum_{i=1}^{n} X_i X_i' \right)^{-1} \left(\frac{1}{n} \sum_{i=1}^{n} X_i \epsilon_i \right)$$

The first term converges to $\mathbf{\Sigma}^{-1}$. The second term, which is random, has mean zero:

$$\mathbb{E}[\frac{1}{n} \sum_{i=1}^{n} X_i \epsilon_i] = \frac{1}{n} \sum_{i=1}^{n} X_i \mathbb{E}\epsilon_i = 0$$

and variance:

$$\begin{aligned} \text{Var}[\frac{1}{n}\sum_{i=1}^{n} X_i\epsilon_i] &= \frac{1}{n^2}\sum_{i=1}^{n} X_i X_i' \text{Var}(\epsilon_i) \\ &= \frac{\sigma^2}{n}\frac{1}{n}\sum_{i=1}^{n} X_i X_i' \to 0 \end{aligned}$$

Thus, the sufficient condition holds and the OLS estimator converges in probability toward the population parameter, β. It is therefore consistent.

Convergence in distribution

Let V be a random variable with cumulative distribution function F. The sequence Z_n tends to V in distribution, if

$$\lim_{n\to\infty} P(Z_n < u) = F(u), \quad \forall u \in \mathbb{R}$$

We will then write

$$Z_n \xrightarrow{\mathcal{L}} V.$$

Central Limit Theorem (CLT)

- Let $Y_1, Y_2, \ldots, Y_n$ be i.i.d. random variables with mean μ and variance σ^2.
- The standardised random variable $\overline{Y}$

$$Z_n = \left(\frac{\overline{Y} - \mu}{\sigma / \sqrt{n}} \right) = \sqrt{n} \left(\frac{\overline{Y} - \mu}{\sigma} \right) \xrightarrow{\mathcal{L}} N(0, 1)$$

- Remark: This result does not depend on the distribution of Y_i!

Asymptotic distribution of $\hat{\beta}$

Based on the decomposition

$$\sqrt{n}(\hat{\beta} - \beta) = \left(\frac{1}{n} \sum_{i=1}^{n} X_i X_i' \right)^{-1} \sqrt{n} \left(\frac{1}{n} \sum_{i=1}^{n} X_i \epsilon_i \right)$$

Again, the first term converges to $\mathbf{\Sigma}^{-1}$.
The second term is a mean, standardized by the rate at which its standard error converges to zero: $\sqrt{n}$. We can therefore use the CLT to obtain

$$\frac{\sqrt{n}}{n} \sum_{i=1}^{n} X_i \epsilon_i \xrightarrow{\mathcal{L}} \mathcal{N}\left(0, \sigma^2 \mathbf{\Sigma}\right)$$

This result does not depend on the distribution of ε!

Modification of the assumptions about X

If X, the vector of explanatory variables, is random, we change our assumptions in the following way:
The classical hypotheses 1 and 2 are replaced by

1. The rank of $\mathbf{X}$ is equal to p with probability 1 (i.e., almost surely).
2. The $(p \times 1)$ random vector X is i.i.d., independent of ε, and $\mathbb{E}[XX'] = \mathbf{\Sigma} < \infty$ is a positive definite matrix.

The independence between X and ε can be weakened by assuming only $\mathbb{E}[\varepsilon|X] = 0$, sometimes called strict exogeneity.

Consequences of the modification

- The unconditional moments $\mathbb{E}[\hat{\beta}]$ and $\text{Var}[\hat{\beta}]$ are replaced by conditional moments, $\mathbb{E}[\hat{\beta}|X]$ and $\text{Var}[\hat{\beta}|X]$.
- For the asymptotic theory, we have to slightly change the arguments, which we will do in the following.

A. Consistency

To show the convergence of the term

$$\left(\frac{1}{n}\sum_{i=1}^{n} X_i X_i'\right)^{-1}$$

we can apply the following lemma.

Lemma (Continuous mapping theorem)

Let g be a continuous function (potentially multivariate) which does not depend on the sample size n. If $\hat{\theta}_n$ is a consistent estimator of θ, then $g(\hat{\theta}_n)$ is a consistent estimator of $g(\theta)$.

In our case, the matrix-valued function $g(\cdot)$ is the inverse function.

Consider the decomposition:

$$\hat{\beta} - \beta = \left(\frac{1}{n} \sum_{i=1}^{n} X_i X_i' \right)^{-1} \left(\frac{1}{n} \sum_{i=1}^{n} X_i \epsilon_i \right)$$

By the continuous mapping theorem and the LLN,

$$\left(\frac{1}{n} \sum_{i=1}^{n} X_i X_i' \right)^{-1} \xrightarrow{p} \left\{ \mathbb{E}(XX') \right\}^{-1}$$

and by the LLN,

$$\frac{1}{n} \sum_{i=1}^{n} X_i \epsilon_i \xrightarrow{p} \mathbb{E}(X\epsilon) = 0$$

Consistency

We have shown the following proposition.

Proposition

In the multiple linear regression model, if the matrix $\mathbb{E}(XX')$ is non-singular, then the OLS estimator is consistent.

Asymptotic distribution

It is based on the decomposition

$$\sqrt{n}\left(\hat{\beta}-\beta\right)=\left(\frac{1}{n}\sum_{i=1}^{n}X_iX_i'\right)^{-1}\left(\frac{\sqrt{n}}{n}\sum_{i=1}^{n}X_i\epsilon_i\right)\ .$$

and, again, on the convergence,

$$\left(\frac{1}{n}\sum_{i=1}^{n}X_iX_i'\right)^{-1}\xrightarrow{p}\left\{\mathbb{E}(XX')\right\}^{-1}\qquad\text{(if }\mathbb{E}(XX')\text{ is invertible)}$$

For the second factor we use the CLT,

$$\frac{\sqrt{n}}{n}\sum_{i=1}^{n} X_i\epsilon_i \xrightarrow{\mathcal{L}} \mathcal{N}\left(0, \sigma^2\mathbf{\Sigma}\right)$$

Recall:

Lemma (Slutsky's lemma)

If the vector $\underline{Z}$ converges in distribution towards a multivariate normal distribution $\mathcal{N}(\underline{0}, \mathbf{S})$ and if the random matrix $\mathbf{A}_n$ is such that $\mathbf{A}_n \xrightarrow{p} \mathbf{A}$, then

$$\mathbf{A}_n \underline{Z} \xrightarrow{\mathcal{L}} \mathcal{N}\left(\underline{0}, \mathbf{A}\mathbf{S}\mathbf{A}'\right) \ .$$

We therefore have the following proposition:

Proposition

In the multiple linear regression model, if the matrix $\mathbf{\Sigma} := \mathbb{E}(XX')$ is non-singular, then the OLS estimator is such that

$$\sqrt{n}\left(\hat{\beta} - \beta\right) \xrightarrow{\mathcal{L}} \mathcal{N}\left(\underline{0}, \sigma^2 \mathbf{\Sigma}^{-1}\right) \ .$$

Part IV

Maximum Likelihood

Introduction

We have so far worked with Least Squares estimation of linear models.
An alternative, general, method of estimation is maximum likelihood.

- estimation of a fully specified parametric model
- can be applied in more general contexts
- has generally excellent asymptotic properties
- requires distributional assumptions about the error term

Joint distribution

For an n-vector of random variables y, let $f(y, \theta)$ be its joint distribution, characterized by a parameter vector θ.
If $f(y, \theta)$ is evaluated at the observations $y_1, \ldots, y_n$, then $f(y, \theta)$ is called the likelihood function of the model for a given data set.
The parameter vector θ that maximized the likelihood function $f(y, \theta)$ for a particular data set is called the maximum likelihood estimator.

Independence

Under the assumption of independence, the joint density is just equal to the product of the marginal densities, i.e.

$$f(y, \theta) = \prod_{i=1}^{n} f_i(y_i, \theta)$$

It is often much easier to maximize not the likelihood function itself, but its logarithm:

$$l(y, \theta) := \log f(y, \theta) = \sum_{i=1}^{n} l_i(y_i, \theta),$$

where $l_i(y_i, \theta)$ is the contribution of the i-th observation to the log-likelihood function, equal to $\log f_i(y_i, \theta)$.

Normal linear regression model

In the linear regression model, we add the normality assumption of the error term:

$$Y = X\beta + \varepsilon, \quad \varepsilon \sim N(0, \sigma^2 I_n)$$

The distribution of y_i conditional on X_i is $N(X_i\beta, \sigma^2)$:

$$f_i(y_i, \beta, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(y_i - X_i\beta)^2}{2\sigma^2}\right).$$

The contribution of the i-th observation to the log-likelihood function is given by

$$l_i(y_i, \beta, \sigma) = -\frac{1}{2}\log 2\pi - \frac{1}{2}\log\sigma^2 - \frac{1}{2\sigma^2}(y_i - X_i\beta)^2$$

Log-likelihood function

Due to independence of the error terms, the log-likelihood function is just the sum of the individual contributions:

$$l(y, \beta, \sigma) = -\frac{n}{2} \log 2\pi - \frac{n}{2} \log \sigma^2 - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - X_i\beta)^2$$

Concentrated log-likelihood function

Concentrate $l(y, \beta, \sigma)$ wrt σ. This means to do the following steps

1. Differentiate $l(y, \beta, \sigma)$ wrt σ
2. Solve the resulting first-order condition for σ
3. Substitute the result back into $l(y, \beta, \sigma)$.

The next step will be to maximize the concentrated log-likelihood wrt β.

We obtain the first order condition

$$\frac{\partial l(y, \beta, \sigma)}{\partial \sigma} = -\frac{n}{\sigma} + \frac{1}{\sigma^3} \sum_{i=1}^{n} (y_i - X_i\beta)^2 = 0$$

and solving this gives

$$\hat{\sigma}^2(\beta) = \frac{1}{n} \sum_{i=1}^{n} (y_i - X_i\beta)^2$$

Substituting back into $l(y, \beta, \sigma)$ yields the concentrated log-likelihood function

$$l^c(y, \beta) = -\frac{n}{2} \log 2\pi - \frac{n}{2} \log \left(\frac{1}{n} \sum_{i=1}^{n} (y_i - X_i\beta)^2 \right) - \frac{n}{2}$$

The maximum likelihood estimators (MLE)

Now maximize $l^c(y, \beta)$ wrt β. The first and third terms do not depend on β. The second term is a negative constant times the log of the sum of squared residuals. Maximizing this term is equivalent to minimizing the sum of squared residuals.
Thus, the MLE of β is the same as OLS in the Gaussian linear regression model.
The MLE of σ is

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} (y_i - X_i \hat{\beta})^2$$

which is different from S^2 which was an unbiased estimator. Hence, the MLE of σ^2 is biased!
However, the bias tends to zero as $n \to \infty$.

Caveat

We have seen that in the Gaussian linear regression model, the MLE of β is the same as OLS, but the estimator of σ^2 is different.

Keep in mind that the equality of MLE and OLS for β only holds in the Gaussian case. If another distribution is assumed for the error term, the MLE of β will be different!

Minimum variance unbiased (MVU)

- An estimator is MVU if it has the smallest variance among all unbiased estimators
- Note: This is a stronger concept than BLUE of the Gauss-Markov theorem. Why?
- One can show that in the Gaussian linear regression model, the MLE of β (which is the same as OLS in the Gaussian case) is MVU.
- One can also show that in the Gaussian linear regression model, S^2 is MVU for σ^2.

Asymptotic properties of MLE

In general, under weak regularity conditions, the MLE is

- consistent
- asymptotically normally distributed
- asymptotically efficient in the sense that its asymptotic variance is smallest in the class of consistent and asymptotically normally distributed estimators

Part V

Inference and hypothesis tests

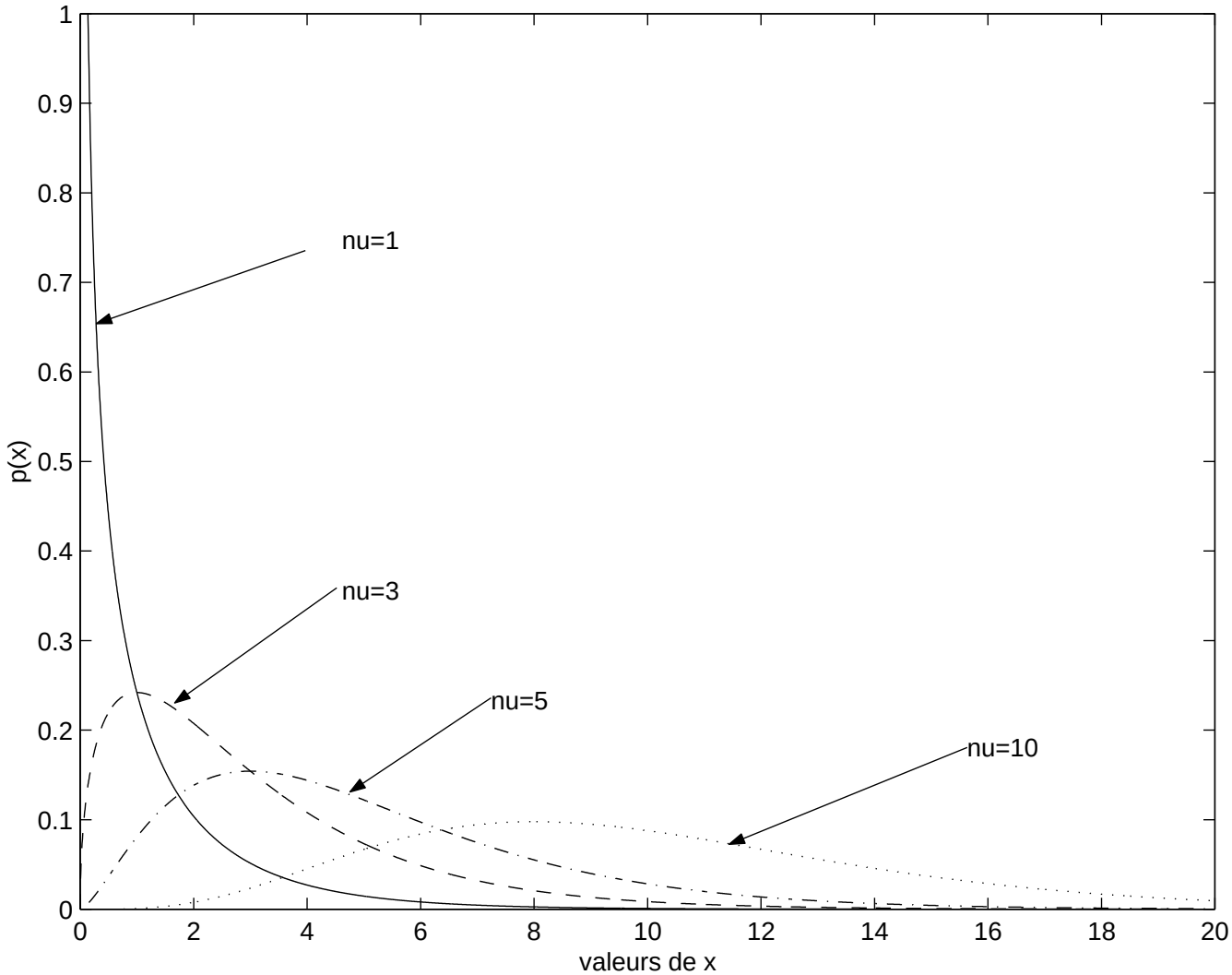
Definition

Let Z be a random $(n \times 1)$ vector with $Z \sim N(0, I_n)$, and A a symmetric idempotent $(n \times n)$ matrix of rank $p \leq n$. Then,

$$Z'AZ \sim \chi^2_p$$

We call p the degrees of freedom of the quadratic form $Z'AZ$.

Special case: if $A = I_n$, then $Z'Z = \sum_{i=1}^{n} Z_i^2 \sim \chi^2_n$.


nu=1
nu=3
nu=5
nu=10
p(x)
valeurs de x
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
0
2
4
6
8
10
12
14
16
18
20

Estimation of σ^2

An estimator of σ^2 is given by

$$\begin{aligned} S^2 &= \frac{1}{n-p}\sum_{i=1}^{n}(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_{i1} - \ldots - \hat{\beta}_{p-1} X_{i,p-1})^2 \\ &= \frac{1}{n-p}(Y - X\hat{\beta})'(Y - X\hat{\beta}) \\ &= \frac{1}{n-p}(Y - \hat{Y})'(Y - \hat{Y}) \\ &= \frac{1}{n-p}e'e \\ &= \frac{1}{n-p}SSE \end{aligned}$$

with $SSE = (Y - \hat{Y})'(Y - \hat{Y})$ the *sum of squared residuals*.

The properties of S^2 are summarized in the following:

Theorem

1. $\mathbb{E}[S^2] = \sigma^2$
2. *If* $\varepsilon \sim N(0, \sigma^2 I_n)$, *then*
   1. $\hat{\beta}$ *and* S^2 *are independent*
   2. $(n-p)S^2/\sigma^2 \sim \chi^2_{n-p}$

Proof

The first property follows by the fact that $SSE = \varepsilon' M \varepsilon$. Indeed,

$$\begin{aligned}
\mathbb{E}[SSE] &= \mathbb{E}[\varepsilon' M \varepsilon] \\
&= \mathbb{E}[\text{tr}(\varepsilon' M \varepsilon)] \\
&= \mathbb{E}[\text{tr}(M \varepsilon \varepsilon')] \\
&= \text{tr}[M E(\varepsilon \varepsilon')] \\
&= \sigma^2 \text{tr}(M) \\
&= \sigma^2 (n - p)
\end{aligned}$$

by the properties of the projection matrix M.

The property 2.a follows by the fact that $\hat{\beta}$ and S^2 depend on orthogonal random vectors, $H\varepsilon$ et $M\varepsilon$:

$$\begin{aligned}\hat{\beta} &= \beta + (X'X)^{-1}X'\varepsilon \\ &= \beta + (X'X)^{-1}X'(H\varepsilon + M\varepsilon) \\ &= \beta + (X'X)^{-1}X'H\varepsilon\end{aligned}$$

and $S^2 = \varepsilon' M\varepsilon/(n-p)$.
The two vectors $H\varepsilon$ and $M\varepsilon$ are orthogonal because $\varepsilon' HM\varepsilon = 0$.
Normality of ε implies independence of $\hat{\beta}$ and S^2.

Finally, the property 2.b follows directly by the definition of the χ^2 density. Indeed,

$$\begin{aligned}(n-p)S^2/\sigma^2 &= \frac{e'e}{\sigma^2} \\ &= \frac{\varepsilon' M \varepsilon}{\sigma^2} \\ &= Z' M Z\end{aligned}$$

where $Z := \varepsilon/\sigma \sim N(0, I_n)$. The degrees of freedom are equivalent to the dimension of the vector space $\mathcal{C}^{\perp}(X)$ into which the matrix M projects. As we have shown, this dimension is equal to $\text{tr}(M) = n - p$.

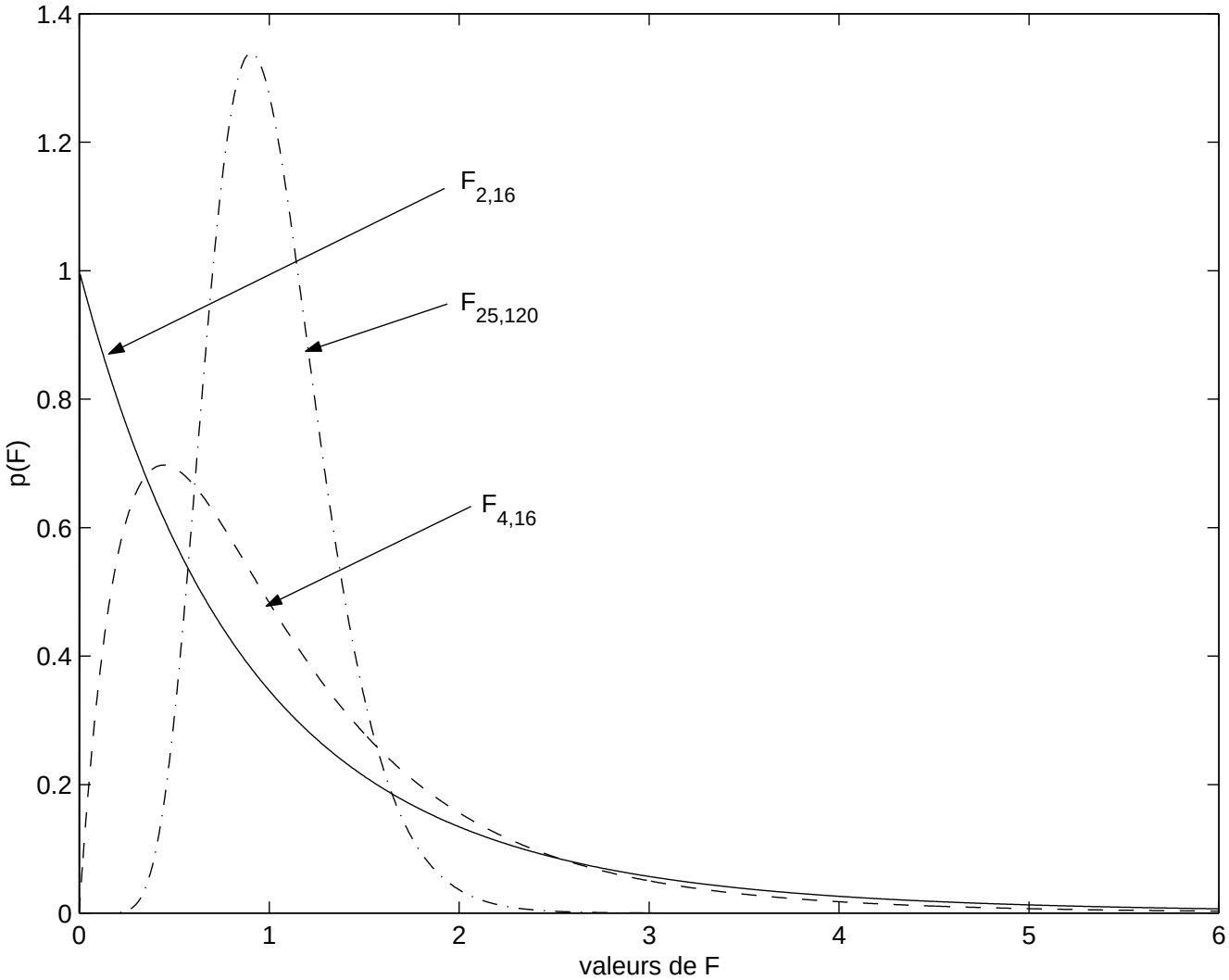
Definition (The F distribution)

Let there be two independent random variables, $X_1 \sim \chi^2_{n_1}$ *and* $X_2 \sim \chi^2_{n_2}$. *Then,*

$$F := \frac{X_1/n_1}{X_2/n_2} \sim F_{n_1,n_2}$$

Properties:

$$\begin{aligned} \mathbb{E}[F] &= n_2/(n_2-2), \quad n_2 > 2 \\ n_1 F &\xrightarrow{\mathcal{L}} \chi^2_{n_1}, \quad n_2 \to \infty \end{aligned}$$


1.4
1.2
1
0.8
0.6
0.4
0.2
0
p(F)
0
1
2
3
4
5
6
valeurs de F
F2,16
F25,120
F4,16

The F-Test

Suppose that $\varepsilon \sim N(0, \sigma^2 I_n)$. We are interested in testing the following linear system of hypotheses for β:

$$H_0 : A\beta = c$$

where A is a known $q \times p$ matrix of rank $q \leq p$, and c is a known $q \times 1$ vector. A special case is, for example,

$$A = \begin{pmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & 0 & \cdots & 1 \end{pmatrix}, \quad c = 0$$

which is equivalent to $H_0 : \beta_1 = \cdots = \beta_{p-1} = 0$.

Using the method of Lagrange multipliers, we show that the least squares estimator under the constraint $A\beta = c$ is given by

$$\hat{\beta}_{H_0} = \hat{\beta} + (X'X)^{-1}A'[A(X'X)^{-1}A']^{-1}(c - A\hat{\beta})$$

The Lagrange function of the minimization problem is given by

$$L(\beta, \lambda) = \frac{1}{2} SSE(\beta) + \lambda'(A\beta - c)$$

with first order conditions

$$\begin{aligned} X'(Y - X\hat{\beta}_{H_0}) + A'\lambda &= 0 \quad (4) \\ A\hat{\beta}_{H_0} - c &= 0 \quad (5) \end{aligned}$$

Pre-multiplying (4) by $(X'X)^{-1}$ and re-arranging gives

$$\hat{\beta}_{H_0} = \hat{\beta} + (X'X)^{-1} A'\lambda \quad (6)$$

Pre-multiplying this equation by A, and using (5), we obtain

$$\begin{aligned} \lambda &= -[A(X'X)^{-1}A']^{-1}A(\hat{\beta} - \hat{\beta}_{H_0}) \\ &= -[A(X'X)^{-1}A']^{-1}(A\hat{\beta} - c) \end{aligned}$$

Finally, from (6),

$$\hat{\beta}_{H_0} = \hat{\beta} + (X'X)^{-1}A'[A(X'X)^{-1}A']^{-1}(c - A\hat{\beta}) \tag{7}$$

Check that $A\hat{\beta}_{H_0} = c$.

From the expression (7) found for $\hat{\beta}_{H_0}$, it follows that the residuals of the constrained model, e_0, are obtained as

$$e_0 := Y - X\hat{\beta}_{H_0} = e + X(X'X)^{-1}A'[A(X'X)^{-1}A']^{-1}(A\hat{\beta} - c) \qquad (8)$$

Let $SSE_0 = e_0'e_0$ be the sum of squared residuals of the constrained model. By (8), the difference between the sum of squared residuals of the constrained and un-constrained model is given by[1]

$$SSE_0 - SSE = (A\hat{\beta} - c)'[A(X'X)^{-1}A']^{-1}(A\hat{\beta} - c) \qquad (9)$$

[1] simple exercise, by noting that $X'e = 0$.

Theorem

Suppose that $\varepsilon \sim N(0, \sigma^2 I_n)$. *Under* H_0,

$$\frac{(SSE_0 - SSE)/q}{SSE/(n-p)} = \frac{(A\hat{\beta} - c)'[A(X'X)^{-1}A']^{-1}(A\hat{\beta} - c)}{qS^2} \sim F(q, n-p)$$

Proof

We have to show three things: (i) the distribution of the numerator, (ii) the distribution of the denominator, and (iii) the independence of both.

(i) The numerator is given by (9). Since, under H_0, $A\hat{\beta} - c \sim N(0, G)$, with $G = \sigma^2 A(X'X)^{-1}A'$ a covariance matrix of rank q, we have $Z := G^{-1/2}(A\hat{\beta} - c) \sim N(0, I_q)$. By the definition of the chi-square distribution, $Z'Z$ follows a chi-square distribution with q degrees of freedom.

(ii) The distribution of the denominator is already known: The SSE is equal to $\varepsilon' M \varepsilon$, and $\varepsilon' M \varepsilon / \sigma^2$ follows a chi-square distribution with $n - p$ degrees of freedom. Note that σ^2 cancels against σ^2 in the numerator.

(iii) The independence of numerator and denominator follows immediately by the fact that the numerator only depends on $\hat{\beta}$, while the denominator depends only on S^2, which under normality are independent as shown before.

This proves the stated result. □

Motivation of the F-test

- Under H_0, SSE and SSE_0 should be close, and the F statistic takes small values.
- If H_0 is false, then F will tend to take large values. We reject H_0, as usual, if the F statistic is beyond a critical value at a given level of the test.

Special case:

$$H_0 : \beta_1 = \ldots = \beta_{p-1} = 0$$

Under H_0, the model becomes $Y_i = \beta_0 + \varepsilon_i, i = 1 \ldots, n$ and $\hat{\beta}_{0,H_0}$ is the value of v that minimizes $\sum_i (Y_i - v)^2$, thus $\hat{\beta}_{0,H_0} = \bar{Y}$ and $\hat{\beta}_{i,H_0} = 0, i = 1, \ldots, p-1$.
The SSE under H_0 becomes

$$\begin{aligned} SSE_0 &= (Y - X\hat{\beta}_{H_0})'(Y - X\hat{\beta}_{H_0}) \\ &= (Y - \bar{Y}\iota)'(Y - \bar{Y}\iota) \\ &= \sum_{i=1}^{n} (Y_i - \bar{Y})^2 = SST \end{aligned}$$

The statistic simplifies to:

$$\frac{(SST - SSE)/(p-1)}{SSE/(n-p)} = \frac{SSR/(p-1)}{SSE/(n-p)} = \frac{MSR}{MSE} \sim_{H_0} F(p-1, n-p)$$

Inference for the parameters β

Recall that if $\varepsilon \sim N(0, \sigma^2 I_n)$, then $\hat{\beta} \sim N(\beta, \sigma^2 (X'X)^{-1})$. This implies that, for all $a \in \mathbb{R}^p$, $a \neq 0$,

$$a'\hat{\beta} \sim N(a'\beta, \sigma^2 a'(X'X)^{-1}a),$$

so that each non-trivial linear combination of $\hat{\beta}$ also follows a normal distribution. By standardizing this linear combination, one obtains

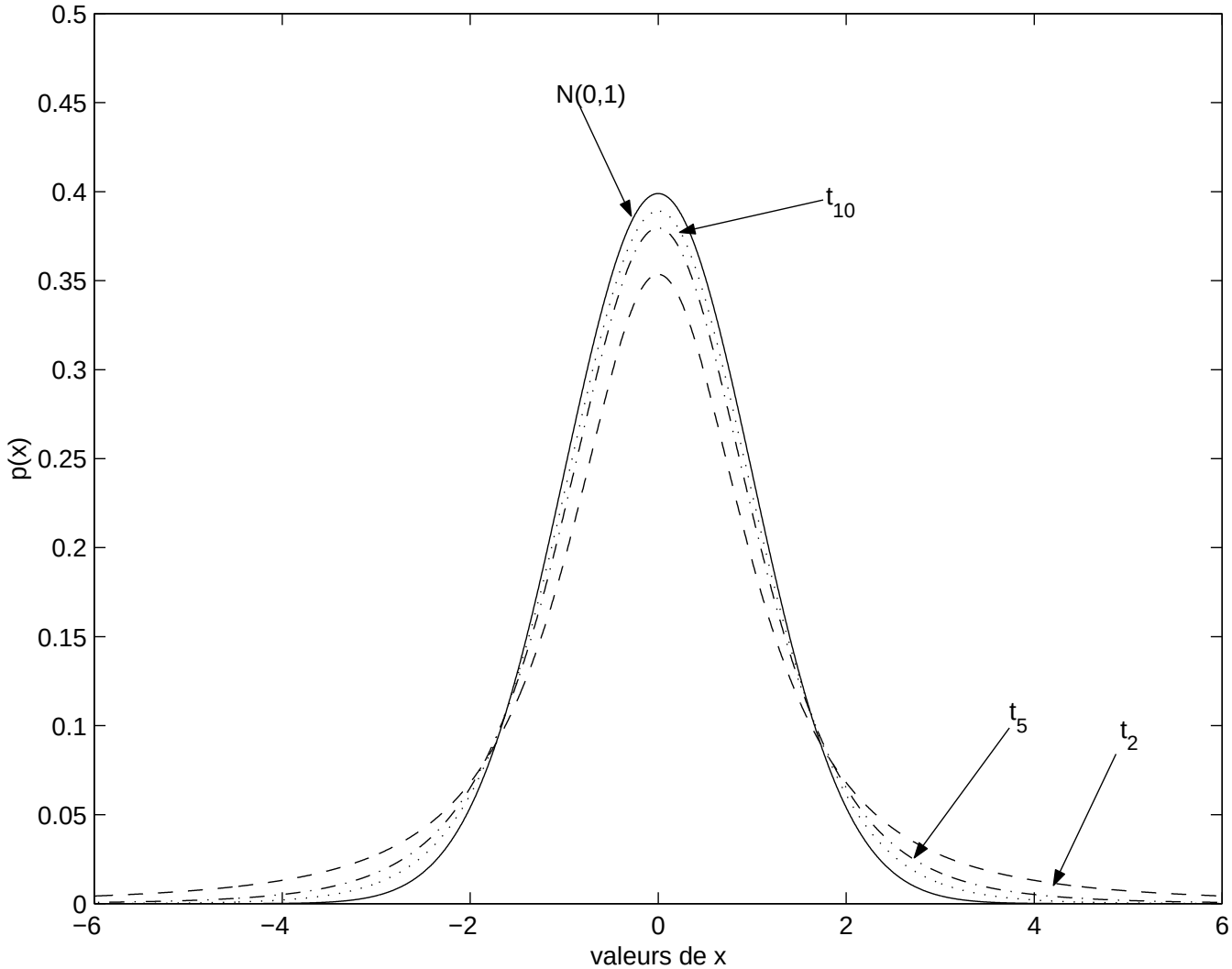
$$U = \frac{a'\hat{\beta} - a'\beta}{\sigma\{a'(X'X)^{-1}a\}^{1/2}} \sim N(0, 1)$$

Definition (The student-t distribution)

Let $Z_0, Z_1, \ldots Z_n$ be independent standard normal r.v., then

$$T = \frac{Z_0}{\sqrt{\sum_{i=1}^{n} Z_i^2 / n}} \sim t_n$$

which means that T follows a student-t distribution with n degrees of freedom.


N(0,1)
t10
t5
t2
p(x)
valeurs de x
0.5
0.45
0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
−6
−4
−2
0
2
4
6

Because, under normality of the error terms,

$$V = \frac{(n-p)S^2}{\sigma^2} \sim \chi^2_{n-p}$$

and $\hat{\beta}$ and S^2 are independent, we have

$$\frac{U}{\sqrt{V/(n-p)}} = \frac{a'\hat{\beta} - a'\beta}{S\{a'(X'X)^{-1}a\}^{1/2}} \sim t_{n-p}.$$

An $(1-\alpha)100\%$ confidence interval for $a'\beta$ is given by

$$a'\hat{\beta} \pm t_{n-p,1-\alpha/2} S\{a'(X'X)^{-1}a\}^{1/2}$$

We can also test the hypothesis

$$H_0 : a'\beta = c \quad \text{versus}$$
$$H_A : a'\beta \neq c$$

This test is in fact equivalent to an F-test with $q = 1$. The equivalence is due to the fact that

$$T_\nu^2 \sim F_{1,\nu}$$

We are now interested in constructing a confidence region for the parameters $\beta_0, \ldots, \beta_{p-1}$ simultaneously.
First, we have $\hat{\beta} - \beta \sim N(0, \sigma^2(X'X)^{-1})$. This implies that

$$\frac{(\hat{\beta} - \beta)'X'X(\hat{\beta} - \beta)}{\sigma^2} \sim \chi^2_p.$$

With $\frac{(n-p)S^2}{\sigma^2} \sim \chi^2_{n-p}$ and independence between $\hat{\beta}$ and S^2, we have

$$\frac{(\hat{\beta} - \beta)'X'X(\hat{\beta} - \beta)}{pS^2} \sim F(p, n-p).$$

A confidence region of level $1 - \alpha$ is given by

$$\left\{ \beta \mid \frac{(\hat{\beta} - \beta)'X'X(\hat{\beta} - \beta)}{pS^2} \leq F_{1-\alpha}(p, n-p) \right\}$$

Bonferroni method (conservative)

A classical confidence interval of level $1-\alpha/p$ for β_j is given by

$$I_j = [\hat{\beta}_j \pm t_{n-p,1-\alpha/2p} s(\hat{\beta}_j)]$$

By the Bonferroni inequality, we have

$$\begin{aligned} P(\beta_0 \in I_0, \ldots, \beta_{p-1} \in I_{p-1}) &\geq 1 - P(\beta_0 \notin I_0) - \ldots - P(\beta_{p-1} \notin I_{p-1}) \\ &= 1 - p\frac{\alpha}{p} = 1 - \alpha \end{aligned}$$

Therefore, a $(1-\alpha)100\%$ confidence region for $\beta_0, \ldots, \beta_{p-1}$ is given by $I_0 \times \ldots \times I_{p-1}$.

Inference for $\mathbb{E}[Y]$

Let us again suppose that $\varepsilon \sim N(0, \sigma^2 I_n)$. We will first derive a confidence interval for $\mathbb{E}[Y]$ at a point x.

Confidence interval for $\mathbb{E}[Y]$

Consider a fixed point of the explanatory variables,

$$x_h = \begin{pmatrix} 1 \\ x_{h1} \\ \vdots \\ x_{h,p-1} \end{pmatrix}$$

with $\mathbb{E}[Y_h] = x_h'\beta$ and $\hat{Y}_h = x_h'\hat{\beta}$.

As

$$\frac{\hat{Y}_h - \mathbb{E}[Y_h]}{S\{x_h'(X'X)^{-1}x_h\}^{1/2}} \sim t_{n-p}$$

a $(1-\alpha)100\%$ confidence interval for $\mathbb{E}[Y_h]$ is given by

$$\hat{Y}_h \pm t_{n-p,1-\alpha/2} S\{x_h'(X'X)^{-1}x_h\}^{1/2}$$

Confidence region for the regression surface

We now search for a confidence region for $\mathbb{E}[Y] = x'\beta$ for all possible x simultaneously.

Theorem (Scheffé, 1953)

For all $l \in \mathbb{R}^p$,

$$P(|l'\hat{\beta} - l'\beta| \leq \{pF(p, n-p; 1-\alpha)\}^{1/2} S(l'(X'X)^{-1}l)^{1/2}) = 1 - \alpha$$

A $(1-\alpha)100\%$ confidence region for the regression surface is given by (*Working-Hotelling confidence region)*

$$\hat{Y}_h \pm WS(x_h'(X'X)^{-1}x_h)^{1/2}$$

where $W^2 = pF(p, n-p; 1-\alpha)$

Confidence region for several $\mathbb{E}[Y_h]$

If one wants to obtain simultaneous confidence intervals for $\mathbb{E}[Y_{h_1}], \ldots, \mathbb{E}[Y_{h_q}]$, there are in principle two possibilities:

- the confidence region of Scheffé
- the confidence region of Bonferroni:

$$\hat{Y}_h \pm BS(x'_{h_k}(X'X)^{-1}x_{h_k})^{1/2},\ k = 1, \ldots, q$$

where $B = t_{n-p;1-\alpha/2q}$.

Both methods are conservative. For large q, the method of Scheffé will be preferred because this region will be smaller than the Bonferroni region, and hence more informative.

Prediction

Consider the prediction of a new variable Y for some given x_h.
Here we want to predict Y and associate a prediction interval.
As before, we fix a x_h:

$$x_h = \begin{pmatrix} 1 \\ x_{h1} \\ \vdots \\ x_{h,p-1} \end{pmatrix}$$

The observation to predict is

$$Y_h = x_h'\beta + \varepsilon_h$$

where $\varepsilon_h \sim N(0, \sigma^2)$ is independent of $\varepsilon_1, \ldots, \varepsilon_n$.

Let $\hat{Y}_h = x_h'\hat{\beta}$ be the predictor of Y_h. As $\mathbb{E}[\hat{Y}_h - Y_h] = 0$ and

$$\begin{aligned} \mathrm{Var}(\hat{Y}_h - Y_h) &= \mathrm{Var}(\hat{Y}_h) + \mathrm{Var}(Y_h) \\ &= \sigma^2 x_h'(X'X)^{-1}x_h + \sigma^2 \\ &= \sigma^2(1 + x_h'(X'X)^{-1}x_h) \end{aligned}$$

we obtain by standard arguments

$$\frac{\hat{Y}_h - Y_h}{S(1 + x_h'(X'X)^{-1}x_h)^{1/2}} \sim t_{n-p}$$

Thus, a prediction interval at level $(1-\alpha)100\%$ is given by

$$\hat{Y}_h \pm t_{n-p;1-\alpha/2} S(1 + x_h'(X'X)^{-1}x_h)^{1/2}$$

Extra sum of squares

Let $X_1, \ldots, X_q$ be the explanatory variables of our model, and let $\tilde{X}_1, \ldots, \tilde{X}_r$ be other variables to be considered to be added to our model. We will define the *extra sum of squares* as

$$\begin{aligned} SSR(\tilde{X}_1, \ldots, \tilde{X}_r \mid X_1, \ldots, X_q) &= SSE(X_1, \ldots, X_q) \\ &- SSE(X_1, \ldots, X_q, \tilde{X}_1, \ldots, \tilde{X}_r) \end{aligned}$$

It measures the reduction of the SSE when the variables $\tilde{X}_1, \ldots, \tilde{X}_r$ are added to a model that already contains $X_1, \ldots, X_q$.

Example: the case of two variables X_1 et X_2

$$\begin{aligned} SSR(X_2|X_1) &= SSE(X_1) - SSE(X_1, X_2) \\ &= SST - SSR(X_1) - SST + SSR(X_1, X_2) \\ &= SSR(X_1, X_2) - SSR(X_1) \end{aligned}$$

and

$$\begin{aligned} SSR(X_1, X_2, X_3) &= SST - SSE(X_1, X_2, X_3) \\ &= SST - SSE(X_1) + SSE(X_1) - SSE(X_1, X_2) \\ &+ SSE(X_1, X_2) - SSE(X_1, X_2, X_3) \\ &= SSR(X_1) + SSR(X_2|X_1) + SSR(X_3|X_1, X_2) \end{aligned}$$

Tests about the parameters $\beta_0, \ldots, \beta_{p-1}$

Consider the model

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$$

If we want to test the hypothesis $H_0 : \beta_2 = 0$ versus $H_a : \beta_2 \neq 0$, we have two equivalent possibilities:

1. Either calculate the student-t statistic $\hat{\beta}_2 / \sqrt{\hat{\text{Var}}(\hat{\beta}_2)}$ which under H_0 follows a student-t distribution with $n - 3$ degrees of freedom.
2. Or calculate the F-statistic,

$$F = \frac{SSE(X_1) - SSE(X_1, X_2)}{SSE(X_1, X_2)/(n-3)} = \frac{SSR(X_2|X_1)}{SSE(X_1, X_2)/(n-3)} \sim F_{1,n-3}$$

If we are now interested in testing the hypothesis $H_0 : \beta_1 = \beta_2 = 0$ versus $H_a : \beta_1 \neq 0$ or $\beta_2 \neq 0$ (or both), we have the Fisher statistic

$$F = \frac{SSR/2}{SSE/(n-p)} = \frac{MSR}{MSE} \sim F(2, n-3)$$

with by definition, $MSR = SSR/2$ and $MSE = SSE/(n-p)$.

In general, for the model

$$Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_{p-1} X_{p-1} + \varepsilon$$

we have three possibilities for hypothesis tests:

1. "global" hypothesis, $H_0 : \beta_1 = \ldots = \beta_{p-1} = 0$
2. "individual" hypothesis, $H_0 : \beta_k = 0$
3. "partial" hypothesis, $H_0 : \beta_q = \beta_{q+1} = \ldots = \beta_{p-1} = 0$

Test of the "global" hypothesis, $H_0 : \beta_1 = \ldots = \beta_{p-1} = 0$
statistic:

$$F = \frac{SSR(X_1, \ldots, X_{p-1})/p-1}{SSE(X_1, \ldots, X_{p-1})/(n-p)} = \frac{MSR}{MSE} \sim F(p-1, n-p)$$

under H_0

Test of the "individual" hypothesis, $H_0 : \beta_k = 0$
statistic:

$$\begin{aligned} F &= \frac{SSR(X_k|X_1 \ldots, X_{k-1}, X_{k+1}, \ldots, X_{p-1})/1}{SSE(X_1, \ldots, X_{p-1})/(n-p)} \\ &= \frac{MSR(X_k|X_1 \ldots, X_{k-1}, X_{k+1}, \ldots, X_{p-1})}{MSE(X_1, \ldots, X_{p-1})} \sim F(1, n-p) \end{aligned}$$

under H_0

Test of the "partial" hypothesis, $H_0 : \beta_q = \beta_{q+1} = \ldots = \beta_{p-1} = 0$
statistic:

$$\begin{aligned} F &= \frac{SSR(X_q, \ldots, X_{p-1} | X_1 \ldots, X_{q-1})/(p-q)}{SSE(X_1, \ldots, X_{p-1})/(n-p)} \\ &= \frac{MSR(X_q, \ldots, X_{p-1} | X_1 \ldots, X_{q-1})}{MSE(X_1, \ldots, X_{p-1})} \sim F(p-q, n-p) \end{aligned}$$

under H_0.

Example (Neter et al., pp. 260)

http://www.ats.ucla.edu/stat/sas/examples/alsm/alsmsasch7.htm

Explain the body fat (Y, difficult to measure) by:

1. X_1: triceps skinfold thickness
2. X_2: thigh circumference
3. X_3: midarm circumference

Data (20 healthy women, 25-34 years old):

Sujet i	X_{i1}	X_{i2}	X_{i3}	Y_i
1	19.5	43.1	29.1	11.9
2	24.7	49.8	28.2	22.8
3	30.7	51.9	37.0	18.7
⋮				
19	22.7	48.2	27.1	14.8
20	25.2	51.0	27.5	21.1

We will try four models (M1 to M4):

1. M1: $Y = \beta_0 + \beta_1 X_1 + \varepsilon$
2. M2: $Y = \beta_0 + \beta_1 X_2 + \varepsilon$
3. M3: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$
4. M4: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon$

```
proc reg data = ch7tab01;
  model y = x1;
  model y = x2;
  model y = x1 x2;
  model y = x1-x3;
run;
quit;
```

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	352.26980	352.26980	44.30	<.0001
Error	18	143.11970	7.95109		
Corrected Total	19	495.38950			

Root MSE	2.81977	R-Square	0.7111
Dependent Mean	20.19500	Adj R-Sq	0.6950
Coeff Var	13.96271		

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	-1.49610	3.31923	-0.45	0.6576
X1	Triceps	1	0.85719	0.12878	6.66	<.0001

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	381.96582	381.96582	60.62	<.0001
Error	18	113.42368	6.30132		
Corrected Total	19	495.38950			

Root MSE	2.51024	R-Square	0.7710
Dependent Mean	20.19500	Adj R-Sq	0.7583
Coeff Var	12.43002		

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	-23.63449	5.65741	-4.18	0.0006
X2	Thigh cir.	1	0.85655	0.11002	7.79	<.0001

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	385.43871	192.71935	29.80	<.0001
Error	17	109.95079	6.46769		
Corrected Total	19	495.38950			

Root MSE	2.54317	R-Square	0.7781
Dependent Mean	20.19500	Adj R-Sq	0.7519
Coeff Var	12.59305		

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	-19.17425	8.36064	-2.29	0.0348
X1	Triceps	1	0.22235	0.30344	0.73	0.4737
X2	Thigh cir.	1	0.65942	0.29119	2.26	0.0369

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	396.98461	132.32820	21.52	<.0001
Error	16	98.40489	6.15031		
Corrected Total	19	495.38950			

Root MSE	2.47998	R-Square	0.8014
Dependent Mean	20.19500	Adj R-Sq	0.7641
Coeff Var	12.28017		

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	117.08469	99.78240	1.17	0.2578
X1	Triceps	1	4.33409	3.01551	1.44	0.1699
X2	Thigh cir.	1	-2.85685	2.58202	-1.11	0.2849
X3	Midarm cir.	1	-2.18606	1.59550	-1.37	0.1896

We obtain the following residual sum of squares:

$$\begin{aligned} SSE(X_1) &= 143.12 \\ SSE(X_2) &= 113.42 \\ SSE(X_1, X_2) &= 109.95 \\ SSE(X_1, X_2, X_3) &= 98.41 \end{aligned}$$

and the reduction of the sum of squares, measured by the extra sum of squares:

$$\begin{aligned} SSR(X_2|X_1) &= SSE(X_1) - SSE(X_1, X_2) = 33.17 \\ SSR(X_1|X_2) &= SSE(X_2) - SSE(X_1, X_2) = 3.47 \\ SSR(X_3|X_1, X_2) &= SSE(X_1, X_2) - SSE(X_1, X_2, X_3) = 11.54 \end{aligned}$$

Test of the "individual" hypothesis, $H_0 : \beta_2 = 0$ in the model M3: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$:
Two possibilities:

1. Student test:

$$T = \frac{\hat{\beta}_2}{\sqrt{\hat{\text{Var}}(\hat{\beta}_2)}} \sim_{H_0} t_{n-3} = t_{17}$$

$T_{obs} = 2.26 > t_{17;0.975} = 2.11 \Rightarrow RH_0$

2. F-test:

$$\frac{SSE(X_1) - SSE(X_1, X_2)}{SSE(X_1, X_2)/17} = \frac{SSR(X_2|X_1)}{SSE(X_1, X_2)/17} \sim_{H_0} F_{1,17}$$

$F_{obs} = 5.13 = T^2_{obs} > F_{1,17;0.95} = 4.45 = 2.11^2 \Rightarrow RH_0$

Test of the “global" hypothesis, $H_0 : \beta_1 = \beta_2 = 0$:
F-test:

$$F = \frac{MSR}{MSE} = \frac{SSR/2}{SSE/17} \sim_{H_0} F_{2,17}$$

$F_{obs} = 29.79 > F_{2,17;0.95} = 3.68 \Rightarrow RH_0$

Partial coefficient of determination

Recall that

$$R^2 = \frac{SSR}{SST} = \frac{SST - SSE}{SST},\ 0 \leq R^2 \leq 1$$

- It measures the proportion of the reduction of the residual variation from an "empty" model (SST) to a "full" model (SSE).
- The idea of a partial R^2 is to replace the "empty" model by a model having a certain number of explanatory variables, and the "full" model by a model that adds one variable to the first one.

Example: two variables are in the model, X_1 et X_2, and one considers adding the variable X_3.
We replace SST in the expression for R^2 by $SSE(X_1, X_2)$, and obtain the partial coefficient of determination of Y explained by X_3, conditional on X_1 and X_2 being in the model,

$$r^2_{y3\cdot 12} = \frac{SSE(X_1, X_2) - SSE(X_1, X_2, X_3)}{SSE(X_1, X_2)} = \frac{SSR(X_3|X_1, X_2)}{SSE(X_1, X_2)}$$

- The partial R^2 is a measure of reduction of $SSE(X_1, X_2)$ associated with the variable X_3.
- $0 \leq r^2_{y3\cdot 12} \leq 1$.
- If $r^2_{y3\cdot 12}$ is large (close to 1), it is useful to add X_3 in the model.
- $r^2_{y3\cdot 12}$ can be shown to be equal to the squared partial correlation of Y and X_3, having both adjusted for their linear relationships with X_1 and X_2.
- A generalization to several variables is possible.

So in our example *body fat*:

$$
\begin{aligned}
r^2_{y2\cdot1} &= \frac{SSR(X_2|X_1)}{SSE(X_1)} = \frac{33.17}{143.12} = 0.232 \\
r^2_{y3\cdot12} &= \frac{SSR(X_3|X_1,X_2)}{SSE(X_1,X_2)} = \frac{11.54}{109.95} = 0.105 \\
r^2_{y1\cdot2} &= \frac{SSR(X_1|X_2)}{SSE(X_2)} = \frac{3.47}{113.42} = 0.031
\end{aligned}
$$

```
proc reg data = ch7tab01;
  model y = x1 x2 / pcorr2;
  model y = x1-x3 / pcorr2;
run;
quit;
```

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Squared Partial Corr Type II
Intercept	Intercept	1	-19.17425	8.36064	-2.29	0.0348	.
X1	Triceps	1	0.22235	0.30344	0.73	0.4737	0.03062
X2	Thigh cir.	1	0.65942	0.29119	2.26	0.0369	0.23176

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Squared Partial Corr Type II
Intercept	Intercept	1	117.08469	99.78240	1.17	0.2578	.
X1	Triceps	1	4.33409	3.01551	1.44	0.1699	0.11435
X2	Thigh cir.	1	-2.85685	2.58202	-1.11	0.2849	0.07108
X3	Midarm cir.	1	-2.18606	1.59550	-1.37	0.1896	0.10501

Part VI

Multicollinearity

Multicollinearity

Analysis of the important effect of correlations between explanatory variables

Basic model:

$$Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_{p-1} X_{p-1} + \varepsilon$$

Consequences of multicollinearity:

- coefficients are extremely sensitive to adding or discarding variables
- variance of estimators becomes very large
- the *extra sum of squares*, strongly depends on other variables in the model

We will consider the following situations:

- uncorrelated variables
- perfect multicollinearity (e.g. a correlation equal to one).
- approximate multicollinearity

Uncorrelated variables

Theorem

If all pairs (X_j, X_k), $1 \leq j < k \leq p-1$ *are uncorrelated, then:*

1. $\hat{\beta}_k$, $k = 1, \ldots, p-1$, *does not depend on the number of variables in the model*
2. $SSR(X_1, \ldots, X_{p-1}) = SSR(X_1) + \ldots + SSR(X_{p-1})$
3. $SSR(X_k|X_j) = SSR(X_k)$, $(1 \leq j \neq k \leq p-1)$

Perfectly correlated variables

An example:

i	X_{i1}	X_{i2}	Y_i
1	2	6	23
2	8	9	83
3	6	8	63
4	10	10	103

Here, $X_2 = 5 + 0.5X_1 \Rightarrow$ the correlation between X_1 and X_2 is equal to 1.
$\Rightarrow$ there is an infinite number of equivalent regression functions, e.g.

$$\begin{aligned} \hat{Y} &= -87 + X_1 + 18X_2 \\ \hat{Y} &= -7 + 9X_1 + 2X_2 \end{aligned}$$

$\Rightarrow$ the coefficients are not identifiable.

Approximate multicollinearity

In general, a high degree of correlation can have the following consequences:

1. By adding or discarding a variable that is highly correlated with other variables, the coefficients of the regression can completely change.
2. The standard errors of the estimated coefficients are very large.
3. If X_1 and X_2 are highly correlated, then $SSR(X_1|X_2) \neq SSR(X_1)$, because if X_2 is already in the model, the marginal effect of X_1 to reduce the SSE is relatively small.

Diagnostics

1. See whether the coefficients are highly sensitive to the specification of the model.
2. See whether the signs of the coefficients are counter-intuitive.
3. Calculate the correlation matrix of explanatory variables. Attention: high pairwise correlations are not necessary to indicate multicollinearity!
4. See whether the confidence intervals for β_k become very wide and un-informative.

Variance inflation factor

$$\text{Var}(\hat{\beta}_k) = \frac{\sigma^2}{\sum_{i=1}^{n}(X_{ik} - \bar{X}_k)^2} \frac{1}{1 - R_k^2}, \quad k = 1, \ldots, p-1$$

where R_k^2 is the coefficient of determination of a regression of X_k on $X_1, \ldots, X_{k-1}, X_{k+1}, \ldots, X_{p-1}$.

Definition

The variance inflation factor (VIF) for $\hat{\beta}_k$ is defined by

$$VIF_k = \frac{1}{1 - R_k^2}, \quad k = 1 \ldots, p-1$$

Decision rule: there is a mulicollinearity problem if either the maximum VIF is very large, e.g. higher than 10, or if the average VIF is considerably larger than 1.

Example *Body Fat*
Three explanatory variables

Correlation matrix:

$$\begin{pmatrix} 1 & 0.92 & 0.46 \\ 0.92 & 1 & 0.08 \\ 0.46 & 0.08 & 1 \end{pmatrix}$$

Estimation results of different models:

Variable	estimated coefficient	standard error
X_1	0.8572	0.1288
X_2	0.8565	0.1100
X_1	0.2224	0.3034
X_2	0.6594	0.2912
X_1	4.334	3.016
X_2	-2.857	2.582
X_3	-2.186	1.596

Coefficient of determination between the three variables:

k	R_k^2	VIF_k
1	0.9986	708.84
2	0.9982	564.34
3	0.9904	104.61

The coefficients R_k^2 are very close to one and, consequently, the VIF are very large $\Rightarrow$ multicollinearity problem.
What remedial measures?

Ridge regression

Main idea: render the matrix $X'X$ well conditioned by adding a term κI_p, $\kappa > 0$.

Ridge regression is based on the minimization of SSE with a penalty term:

$$S(\beta, \kappa) \quad = \quad (Y - X\beta)'(Y - X\beta) + \kappa ||\beta||_2^2$$

where $||\beta||_2 = \sqrt{\beta'\beta}$

Minimization of $S(\beta, \kappa)$ with respect to β gives the result (easy exercise):

$$\hat{\beta}_R = (X'X + \kappa I_p)^{-1} X'Y$$

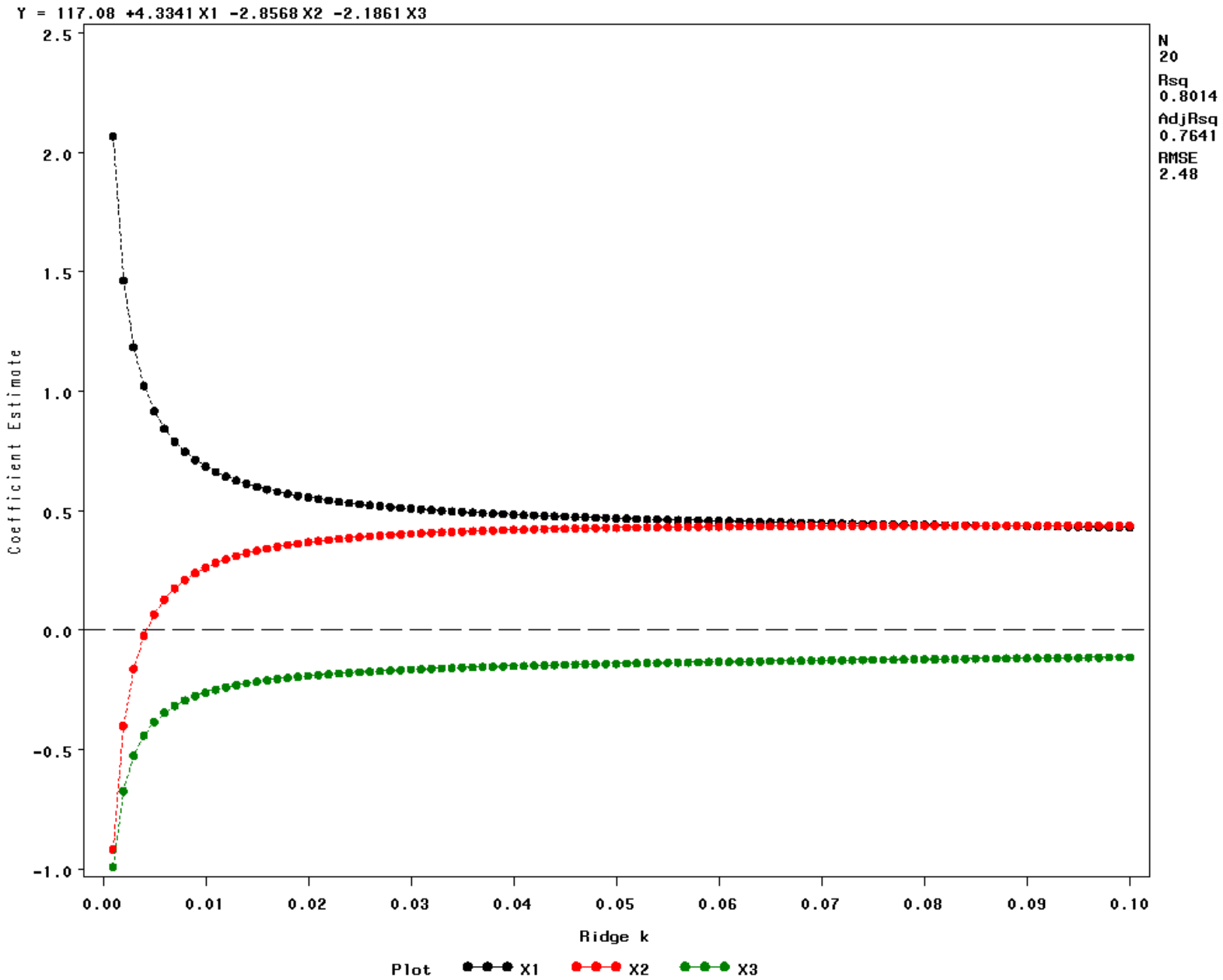
If $\kappa > 0$, then $\hat{\beta}_R$ is biased but might have a smaller MSE than $\hat{\beta}$.

In practice, how do we choose κ?

- There is no generally optimal solution.
- Often used: the *ridge trace*, a graph that shows the estimates of β as a function of κ. This is for a regression with *standardized* Y and X variables.

Generalized Ridge regression: one adds a diagonal matrix, $\text{diag}(\kappa_1, \ldots, \kappa_p)$, with $\kappa_j > 0$ for all j.

```
symbol1 v=dot h=.8;
proc reg data = ch7tab01 noprint outest = temp outstb noprint;
  model y = x1-x3/ ridge = (0.001 to 0.1 by .001) outvif  ;
  plot / ridgeplot vref=0;
run;
quit;
```


Y = 117.08 +4.3341 X1 -2.8568 X2 -2.1861 X3
N
20
Rsq
0.8014
AdjRsq
0.7641
RMSE
2.48
Coefficient Estimate
2.5
2.0
1.5
1.0
0.5
0.0
-0.5
-1.0
0.00
0.01
0.02
0.03
0.04
0.05
0.06
0.07
0.08
0.09
0.10
Ridge k
Plot
X1
X2
X3

Part VII

Discrete variables

Dummy variables

Definition: A dummy variable is a special explanatory variable that only takes two values, 0 or 1.

Example

Regression model for the log salary

$$\mathbb{E}(\log S | Ed, D) = \beta_1 + \beta_2 Ed + \beta_3 D$$

where D takes the value 1 if the individual is a man, and 0 otherwise.

Example

Phillips curve: unemployment vs inflation

$$U_t = \beta_1 + \beta_2 \pi_t + \varepsilon_t$$

where U_t is the unemployment rate in year t, and π_t is the inflation rate in year t.
One has data for the period 1975 to 2019. Has there been a structural break caused by the financial crisis in 2007 to 2009? We define

$$\begin{aligned} D_t &= 0 \qquad \text{for } t = 1975 \text{ to } 2006 \\ D_t &= 1 \qquad \text{for } t \geq 2007 \end{aligned}$$

$$U_t = \beta_1 + \beta_2 \pi_t + \beta_3 D_t + \varepsilon_t$$

Stability test (Chow test)

Objective: Test structural changes of the model

Consider the dummy variable D which specifies in which sub-sample a given individual lies (man or woman, before or after the crisis, etc.)

Consider

$$\mathbb{E}(Y|X) = X\beta + DX\gamma$$

We test

$$H_0 : \gamma = 0_p \ .$$

$\Rightarrow$ special case of an F-test (here called *Chow test*.)

The Chow test statistic

Under H_0, there are p constraints

The unconstrained model contains $2p$ parameters

As a consequence, the Chow test statistic can be written as

$$F = \frac{\left(\widehat{SSE}_0 - \widehat{SSE}_1\right)/p}{\widehat{SSE}_1/(n-2p)} \sim F^p_{n-2p}$$

Analysis of variance (one factor)

One discrete variable (factor ou criterion) with r levels to explain one quantitative variable Y.
We can write the model by using dummy variables D_{kj}:

$$Y_k = \sum_{j=1}^{r} \beta_j D_{kj} + \varepsilon_k, \quad k = 1 \ldots, n$$

where $D_{kj} = 1$ if the k-th observation belongs to level j of the factor, and 0 otherwise.
Equivalent version with reference level, for example the first one:

$$Y_k = \beta_1 + \sum_{j=2}^{r} \beta_j D_{kj} + \varepsilon_k, \quad k = 1 \ldots, n$$

Equivalent representation ("ANOVA"):

$$Y_{ij} = \mu_i + \varepsilon_{ij}, \quad i = 1, \ldots, r;\ j = 1, \ldots, n_i$$

We are interested, for example, in the hypothesis of equality of the means:

$$\begin{aligned} H_0 &: \mu_1 = \ldots = \mu_r \\ H_1 &: \text{at least one mean is different} \end{aligned}$$

This is a special case of the test for linear restrictions.

Alternative specification of the model

The preceding model is equivalent to

$$Y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$$

where $\mu_i = \mu + \alpha_i$ and $\varepsilon_{ij} \sim N(0, \sigma^2)$.

- Without constraint, this model is not identified.
- One possible constraint:

$$\sum_{i=1}^{r} \alpha_i = 0$$

- If the design is balanced, this constraint implies that μ corresponds to the total mean.

Relation between regression and ANOVA

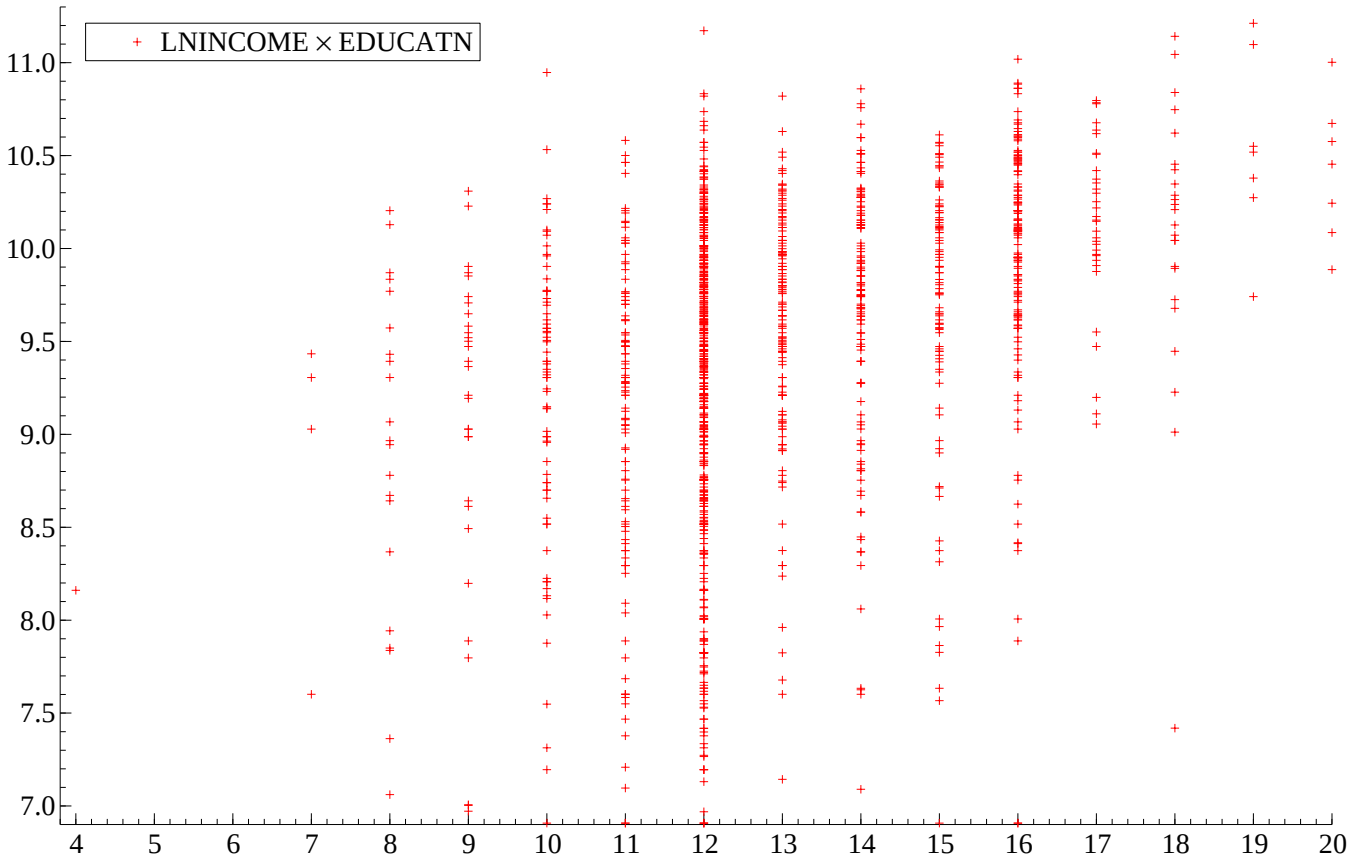
The ANOVA model can be written in the form of a multiple regression:

$$Y = X\beta + \varepsilon$$

Example: $r = 3$, balanced design with $n = 2$, and the definitions:

$$Y = \begin{pmatrix} Y_{11} \\ Y_{12} \\ Y_{21} \\ Y_{22} \\ Y_{31} \\ Y_{32} \end{pmatrix}, \quad X = \begin{pmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \\ 1 & -1 & -1 \\ 1 & -1 & -1 \end{pmatrix}, \quad \varepsilon = \begin{pmatrix} \varepsilon_{11} \\ \varepsilon_{12} \\ \varepsilon_{21} \\ \varepsilon_{22} \\ \varepsilon_{31} \\ \varepsilon_{32} \end{pmatrix}, \quad \beta = \begin{pmatrix} \mu \\ \alpha_1 \\ \alpha_2 \end{pmatrix}$$

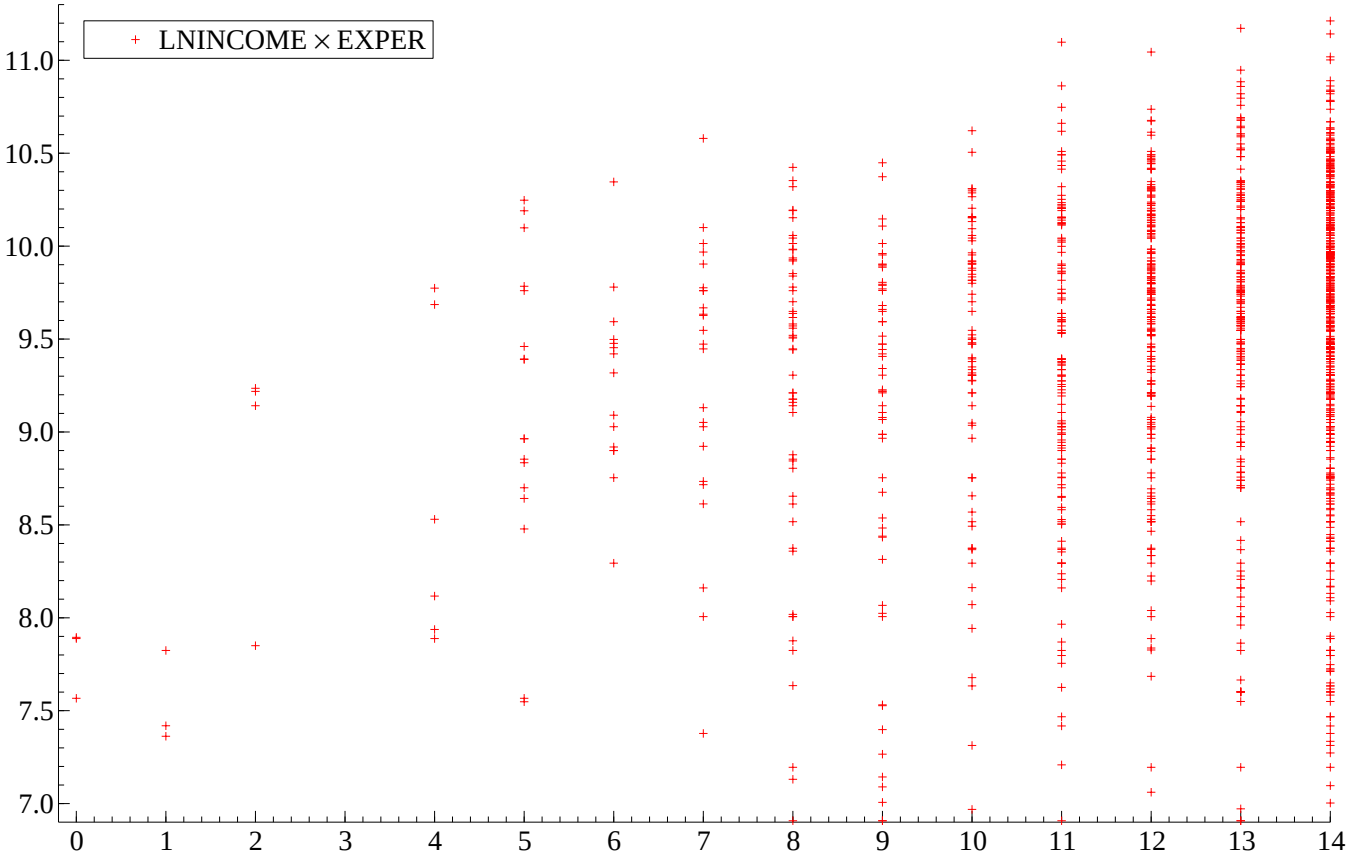
Case study: Education and salary



The research question

- What is the gain in salary by an additional year of education?
- May depend on the professional experience of the employee
- Therefore, we wish to control for the effect of experience.

Experience and salary



The regression

$$\log S = \beta_0 + \beta_1 E + \beta_2 X + \varepsilon$$

où

- S: Salary
- E: Number of years of education
- X: Number of years of professional experience

Results

```
           Coefficient  Std.Error  t-value  t-prob Part.R^2
Constant     6.93366     0.1547      44.8  0.0000   0.5965
EDUCATN     0.137031    0.009368     14.6  0.0000   0.1360
EXPER       0.0623686   0.008508     7.33  0.0000   0.0380

sigma       0.750061  SSE               764.562667
R^2       0.172837  F(2,1359) =       142 [0.000]**
Adj.R^2   0.17162   log-likelihood    -1539.38
```

We predict that an additional year of education yields, in mean, a salary increase of 13.7%.

Refined analysis taking gender into account

Are the salaries of men and women different on average?

Define the dummy variable

$$W_i = \begin{cases} 1, & \text{if i is a woman} \\ 0, & \text{otherwise} \end{cases}$$

More general model:

$$\log S = \beta_0 + \beta_1 E + \beta_2 X + \beta_3 W + \varepsilon$$

Results

	Coefficient	Std.Error	t-value	t-prob	Part.R^2
Constant	6.98627	0.1536	45.5	0.0000	0.6036
EDUCATN	0.146139	0.009454	15.5	0.0000	0.1496
EXPER	0.0566627	0.008505	6.66	0.0000	0.0317
FEMALE	-0.210517	0.04133	-5.09	0.0000	0.0187

sigma	0.743272	SSE	750.232104
R^2	0.188341	F(3,1358) =	105 [0.000]**
Adj.R^2	0.186548	log-likelihood	-1526.5

1. The effect of education increases from 13.7% to 14.6%.
2. We reject the hypothesis that gender has no effect for the salary.
3. Women have, on average, 21% lower salaries than men, for the same level of education and experience.

Dummy variables trap

Why did we not introduce an additional dummy variable for men,

$$M_i = \begin{cases} 1, & \text{if i is a man} \\ 0, & \text{otherwise} \end{cases}$$

and specify the model

$$\log S = \beta_0 + \beta_1 E + \beta_2 X + \beta_3 W + \beta_4 M + \varepsilon$$

What is the problem here?

Dummy variables trap (contd)

To see the problem, write the model separately for men and women:

$$\begin{aligned} \log S_i &= \beta_0 + \beta_3 + \beta_1 E_i + \beta_2 X_i + \varepsilon_i, \quad \text{i is a woman} \\ \log S_i &= \beta_0 + \beta_4 + \beta_5 E_i + \beta_6 X_i + \varepsilon_i, \quad \text{i is a man} \end{aligned}$$

We have two intercepts, but three parameters $(\beta_0, \beta_3, \beta_4) \Rightarrow$ Identification problem!

Dummy variables trap (contd)

This identification problem corresponds to multicollinearity:

$$\forall i, W_i + M_i = 1$$

There is a linear combination of W and M that, for all i, is identical to another variable (here the constant).

Dummy variables trap (contd)

We could have discarded the constant and leave W and M in the model:

$$\log S = \beta_1 E + \beta_2 X + \beta_3 W + \beta_4 M + \varepsilon$$

```
            Coefficient  Std.Error  t-value  t-prob Part.R^2
EDUCATN       0.146139   0.009454     15.5   0.0000   0.1496
EXPER         0.0566627  0.008505     6.66   0.0000   0.0317
FEMALE        6.77576    0.1564       43.3   0.0000   0.5802
MALE          6.98627    0.1536       45.5   0.0000   0.6036

sigma         0.743272   SSE                750.232104
R^2           0.188341   F(3,1358) =     105 [0.000]**
Adj.R^2       0.186548   log-likelihood       -1526.5
```

Different marginal effects

Do experience and education have the same effects for men and women?
To answer this question, consider the models

$$\log S = \beta_0 + \beta_1 E + \beta_2 X + \beta_3 W + \beta_4 E \cdot W + \beta_5 X \cdot W + \varepsilon$$

If i is a man ($W = 0$), this reduces to

$$\log S = \beta_0 + \beta_1 E + \beta_2 X + \varepsilon$$

and if i is a woman ($W = 1$), we obtain

$$\log S = \beta_0 + \beta_3 + (\beta_1 + \beta_4) E + (\beta_2 + \beta_5) X + \varepsilon$$

Estimation of the unconstrained model

	Coeff	s.e.	t-value	t-prob	Part.R^2
Constant	7.05818	0.2315	30.5	0.0000	0.4067
EDUCATN	0.136130	0.01279	10.6	0.0000	0.0771
EXPER	0.0609992	0.01306	4.67	0.0000	0.0158
FEMALE	-0.395794	0.3133	-1.26	0.2067	0.0012
EDUCATN*FEMALE	0.0233039	0.01911	1.22	0.2228	0.0011
EXPER*FEMALE	-0.00973617	0.01730	-0.563	0.5737	0.0002

sigma	0.743355	SSE	749.293618
R^2	0.189356	F(5,1356) =	63.35 [0.000]**
Adj.R^2	0.186367	log-likelihood	-1525.64
no. of observations	1362	no. of parameters	6
mean(LNINCOME)	9.45855	se(LNINCOME)	0.824104

Chow test

In order to test whether the regressions for log salary are the same for men and women, let

$$H_0 : \beta_3 = \beta_4 = \beta_5 = 0$$

Test statistic:

$$F = \frac{\left(\widehat{S}SE_0 - \widehat{S}SE_1\right)/J}{\widehat{S}SE_1/(n-p)} \sim F^J_{n-p} \quad \text{under } H_0$$

With $SSE_0 = 764,56$, $SSE_1 = 749.29$, $J = 3$, $n - p = 1356$, we obtain $F = 9.21$ and a p-value very close to zero. We reject H_0.

Test of marginal restrictions

In order to test whether the marginal effects of education and experience are the same for men and women, let

$$H_0 : \beta_4 = \beta_5 = 0$$

With $SSE_0 = 750,23$, $SSE_1 = 749.29$, $J = 2$, $n - p = 1356$, we obtain

$$F = 0.84919$$

and a p-value of 0.4280. We do not reject H_0.

Summary of the case study

- The salaries of women are on average 21% lower than those of men, for the same level of education and professional experience.
- This difference does not depend on the level of education or experience.
- One additional year of education increases the expected salary by 14%, and this effect is the same for men and women.

Part VIII

Variable selection

Selection of explanatory variables

Trois types:

- Type I: Choice of "best" model among all possible models according to a determined criterion
- Type II: Sequential approach to include variables starting from a simple model, or sequentially exclude variables from a complex one.
- Type III: Variable selection by penalization (LASSO)

Procedures of type I

Mallows criterion

The complete model:

$$Y = X\beta + \varepsilon$$

with $X = (X_1, \ldots, X_P)$ and $\beta = (\beta_0, \beta_1, \ldots, \beta_{P-1})'$ and the estimator

$$\hat{\beta} = (X'X)^{-1}X'Y$$

If we reduced the number of parameters from P to p, $p < P$, how much would we lose in terms of mean square error?

Reduced model:

$$Y = X_{(p)}\beta_{(p)} + \varepsilon$$

with $X_{(p)} = (X_1, \ldots, X_p)$ et $\beta_{(p)} = (\beta_0, \beta_1, \ldots, \beta_{p-1})'$ and the estimator:

$$\tilde{\beta}_{(p)} = (X'_{(p)}X_{(p)})^{-1}X'_{(p)}Y$$

What are the properties of the reduced model?

Let

$$\begin{aligned} \tilde{Y}_{i(p)} &= \tilde{\beta}_0 + \tilde{\beta}_1 x_{i1} + \ldots + \tilde{\beta}_{p-1} x_{i,p-1} \\ &= x'_{i(p)} \tilde{\beta}_{(p)} \end{aligned}$$

Then

$$\begin{aligned} \mathbb{E}[\tilde{Y}_{i(p)}] &= x'_{i(p)} \mathbb{E}[\tilde{\beta}_{(p)}] \\ &\neq \mathbb{E}[\hat{Y}_i] = x'_i \mathbb{E}[\hat{\beta}] = x'_i \beta \end{aligned}$$

Thus, the prediction of Y using the reduced model is biased. But this is not the only criterion.

The mean square error

$$\begin{aligned}\mathbb{E}[(\tilde{Y}_{i(p)} - x_i'\beta)^2] &= \mathbb{E}[(\tilde{Y}_{i(p)} - \mathbb{E}[\tilde{Y}_{i(p)}])^2] + \mathbb{E}[(\mathbb{E}[\tilde{Y}_{i(p)}] - x_i'\beta)^2] \\ &= \mathrm{Var}(\tilde{Y}_{i(p)}) + Bias(\tilde{Y}_{i(p)})^2\end{aligned}$$

The Mallows criterion is a standardized sum of these mean square errors

Let us define

$$\Gamma_p = \frac{1}{\sigma^2}\left\{\sum_{i=1}^{n} \text{Var}(\tilde{Y}_{i(p)}) + \sum_{i=1}^{n} Biais(\tilde{Y}_{i(p)})^2\right\}$$

We can show that

$$\sum_{i=1}^{n} \text{Var}(\tilde{Y}_{i(p)}) = \sigma^2 p$$

and

$$\sum_{i=1}^{n} Biais(\tilde{Y}_{i(p)})^2 = \mathbb{E}[SSE_p] - \sigma^2(n-p)$$

Estimating σ^2 by MSE, we can estimate Γ_p by

$$\begin{aligned} C_p &= p + \frac{SSE_p - (n-p)MSE}{MSE} \\ &= \frac{SSE_p}{MSE} - (n-2p) \end{aligned}$$

For two models with the same number of parameters p, we prefer that with smaller C_p.

Link with the coefficient of determination:

$$C_p = (n - P)\frac{1 - R_p^2}{1 - R_P^2} + 2p - n$$

The first term is minimized by $p = P$ (because R_p^2 is maximized for the full model), but the second term imposes as penalty.

The adjusted coefficient of determination also imposes a penalty for the number of parameters:

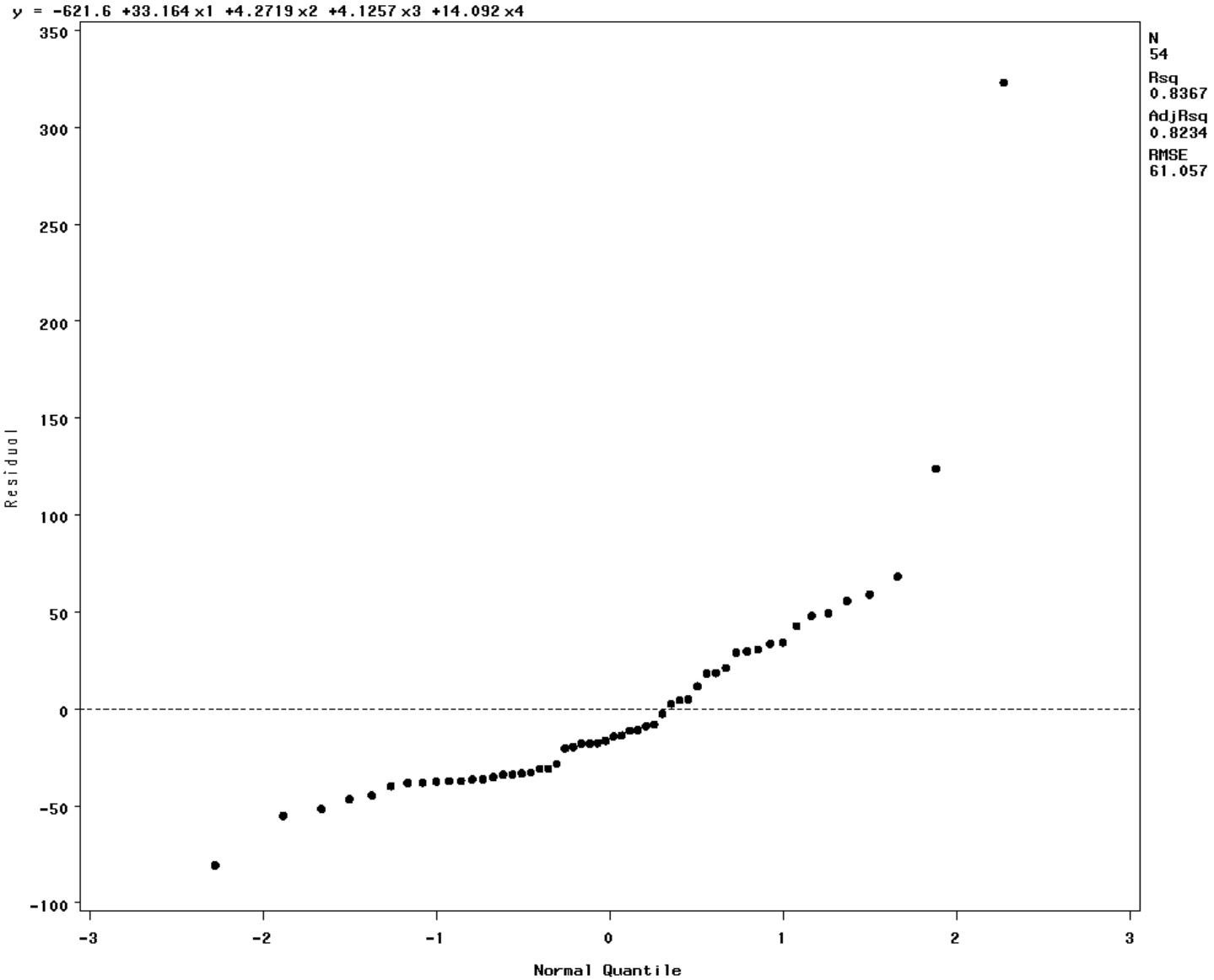
$$R^2_{a(p)} = 1 - \frac{n-1}{n-p}\frac{SSE_p}{SST}$$

Exemple (Neter et al., pp. 334):
Data about 54 patients undergoing a liver operation.

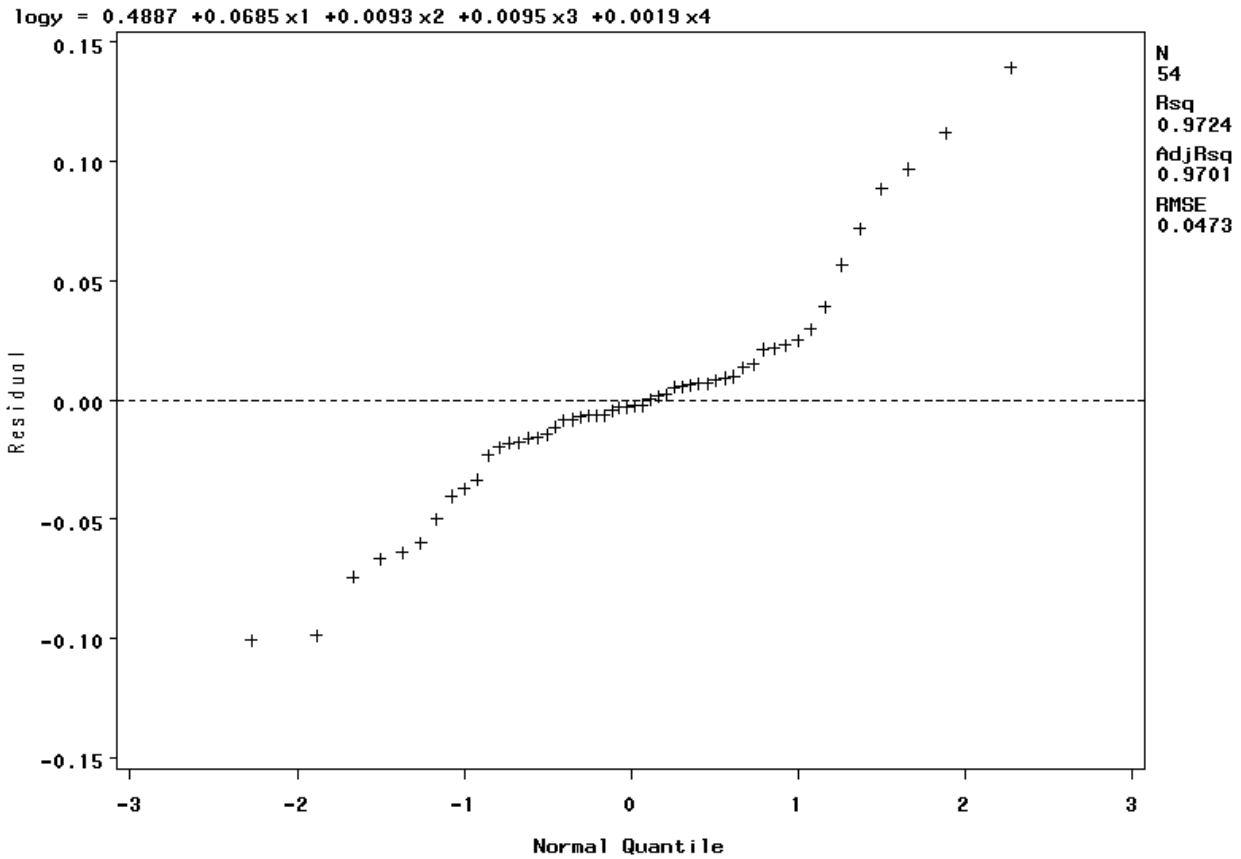
- Y: survival time
- X_1: blood clotting score
- X_2: prognostic index, which includes the age of the patient
- X_3: enzyme function test score
- X_4: liver function test score

Before selecting a model, it has been verified if the model $Y = \beta_0 + \sum_{i=1}^{4} \beta_i X_i + \varepsilon$, $\varepsilon \sim N(0, \sigma^2)$ was adequat.

QQ-plot of the residuals of the model $Y = \beta_0 + \sum_{i=1}^{4} \beta_i X_i + \varepsilon$



QQ-plot of the residuals of the model $\log Y = \beta_0 + \sum_{i=1}^{4} \beta_i X_i + \varepsilon$



```
proc reg data=ch8tab01 ;
  model logy = x1-x4/selection=rsquare adjrsq cp press mse sse;
run;
quit;
```

Number in Model	R-Square	Adjusted R-Square	C(p)	MSE	SSE	Variables in Model
1	0.5274	0.5183	787.9471	0.03611	1.87765	x4
1	0.4424	0.4316	938.6707	0.04260	2.21539	x3
1	0.3515	0.3391	1099.691	0.04954	2.57620	x2
1	0.1200	0.1031	1510.148	0.06723	3.49594	x1
2	0.8129	0.8056	283.6276	0.01457	0.74310	x2 x3
2	0.6865	0.6742	507.8069	0.02442	1.24544	x3 x4
2	0.6496	0.6358	573.2766	0.02730	1.39214	x2 x4
2	0.6458	0.6319	580.0075	0.02759	1.40722	x1 x3
2	0.5278	0.5093	789.1422	0.03678	1.87585	x1 x4
2	0.4381	0.4160	948.2417	0.04377	2.23235	x1 x2
3	0.9723	0.9707	3.0390	0.00220	0.10989	x1 x2 x3
3	0.8829	0.8758	161.6520	0.00931	0.46530	x2 x3 x4
3	0.7192	0.7023	451.8957	0.02231	1.11567	x1 x3 x4
3	0.6500	0.6290	574.5468	0.02781	1.39051	x1 x2 x4
4	0.9724	0.9701	5.0000	0.00224	0.10980	x1 x2 x3 x4

```
goptions reset=all;
symbol1 v=star c=blue h = .8;
proc reg data = ch8tab01  outest = temp covout;
  model logy = x1-x4/ selection= rsquare adjrsq cp mse noprint;
  plot cp.*np. / cmallows = blue;
  plot mse.*np.;
  plot rsq.*np.;
run;
quit;
```

Figure of R_p^2 versus p

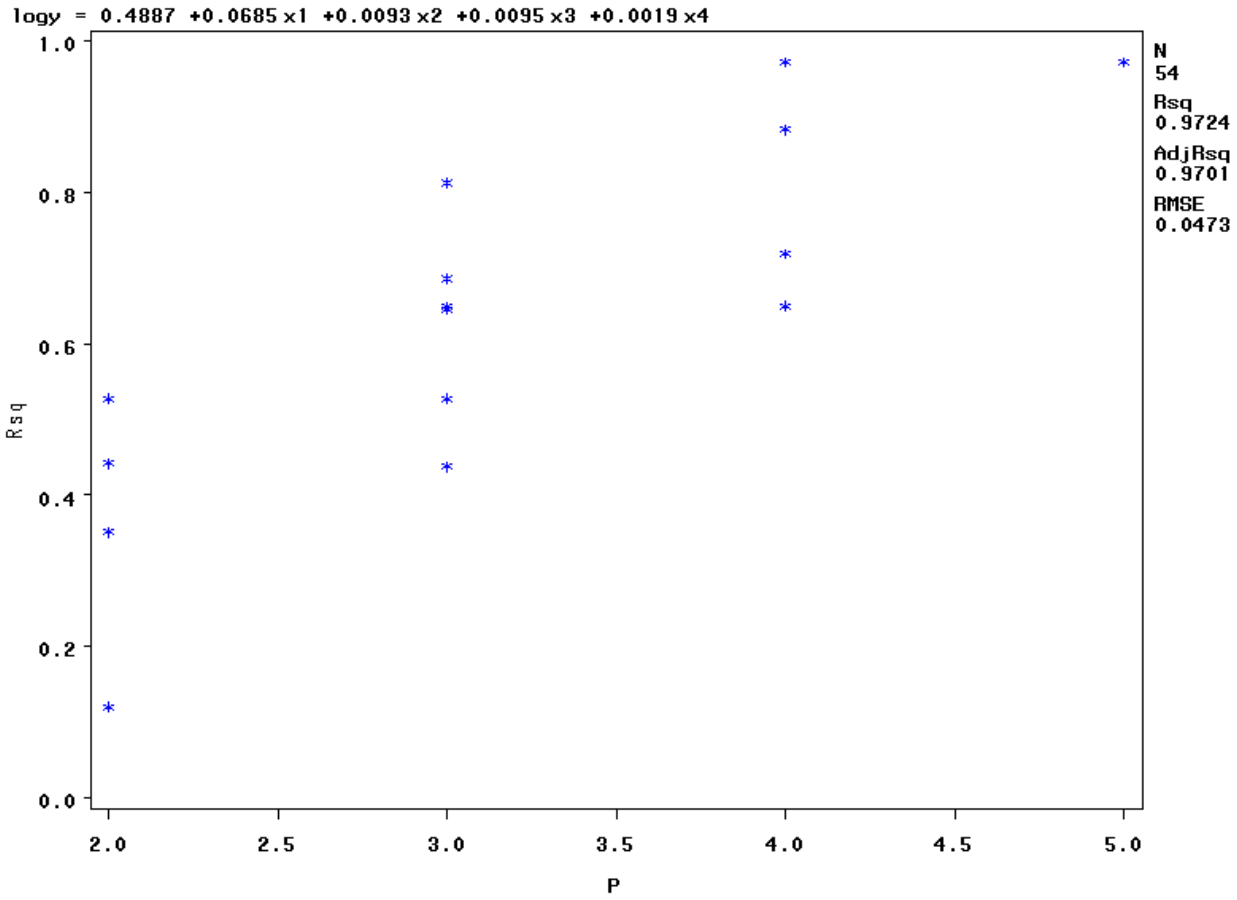


Figure of MSE_p versus p

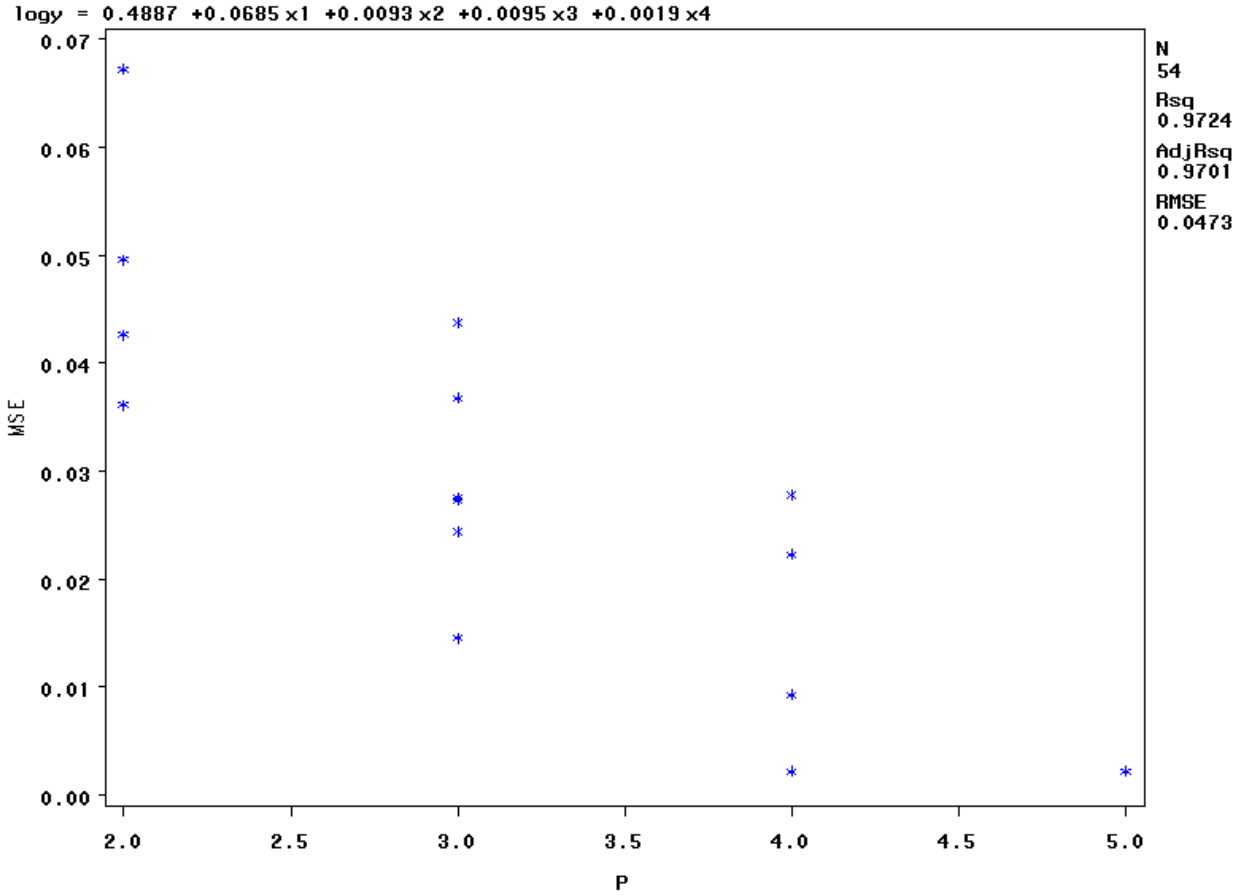
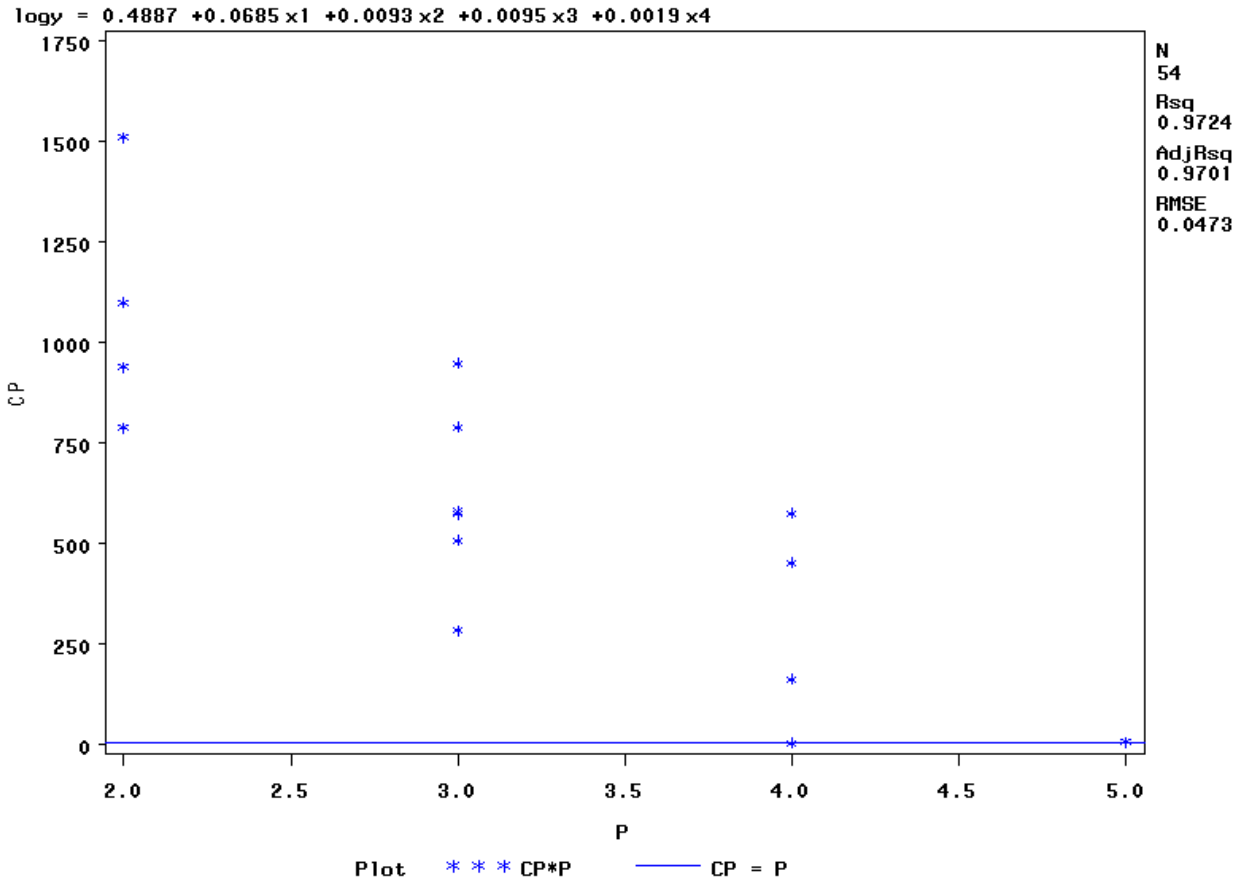


Figure of C_p versus p



Selection procedures of type II

Useful when the number of variables is large.

Forward stepwise regression

The steps:

1. Let $r^2_{yk} = \frac{SSR(X_k)}{SST}, \quad k = 1, \ldots, p-1$ and

$$F_k = \frac{SSR(X_k)}{SSE(X_k)/(n-2)} = \frac{MSR(X_k)}{MSE(X_k)}$$

Designate by X_{s_1} the variable for which r^2_{yk} is maximum (which also maximizes F_k)

Rule: include X_{s_1} if p-value which corresponds to F_{s_1} is smaller than SLE (*significance level to enter*)

2. Suppose that X_{s_1} is selected
Let $r^2_{yk.s_1} = \frac{SSR(X_k|X_{s_1})}{SSE(X_{s_1})}$, and

$$F_{k|s_1} = \frac{SSR(X_k|X_{s_1})}{MSE(X_{s_1}, X_k)}, \quad k = 1, \ldots, p-1, k \neq s_1$$

Designate by X_{s_2} the variable for which $r^2_{yk.s_1}$ is maximum (which also maximizes $F_{k|s_1}$)
Rule: include X_{s_2} if p-value which corresponds to $F_{s_2|s_1}$ is smaller than SLE

3. Suppose that X_{s_1} and X_{s_2} are selected
Let

$$F_{s_1|s_2} = \frac{MSR(X_{s_1}|X_{s_2})}{MSE(X_{s_1}, X_{s_2})}$$

Rule: Discard X_{s_1} if p-value which corresponds to $F_{s_1|s_2}$ is larger than SLS (*significance level to stop*)
For example, choose SLE=SLS=0.05.

Forward selection

A simplified version of "forward stepwise regression", where one does not test whether a variable that has been included previously can be discarded at a later stage.

Backward elimination

This is essentially the converse of "forward selection":
One starts with a "complete" model and then suppresses sequentially variables that minimize

$$F_k^* = \frac{MSR(X_k|\text{other variables})}{MSE(X_k|\text{other variables})}$$

and for which the p-value is larger than SLS.

Backward stepwise regression

This is the converse of "forward stepwise regression":
A variable that has been suppressed before can be reincluded at a later stage.

```
Exemple: Surgery unit data

proc reg data = ch8tab01;
  model logy = x1-x4/ selection = stepwise slentry= .01 slstay= .05;
run;
quit;


Stepwise Selection: Step 1

Variable x4 Entered: R-Square = 0.5274 and C(p) = 787.9471

                            Analysis of Variance

                                    Sum of           Mean
Source                    DF       Squares         Square    F Value    Pr > F

Model                      1       2.09508        2.09508      58.02    <.0001
Error                     52       1.87765        0.03611
Corrected Total           53       3.97273

             Parameter     Standard
Variable      Estimate        Error   Type II SS  F Value  Pr > F

Intercept      1.69639      0.07174     20.18819   559.10  <.0001
x4             0.18575      0.02439      2.09508    58.02  <.0001

Bounds on condition number: 1, 1
------------------------------------------------------------------------------------------
```

```
Stepwise Selection: Step 2

Variable x3 Entered: R-Square = 0.6865 and C(p) = 507.8069

                          Analysis of Variance

                                    Sum of           Mean
Source                    DF       Squares         Square    F Value    Pr > F

Model                      2       2.72729        1.36364      55.84    <.0001
Error                     51       1.24544        0.02442
Corrected Total           53       3.97273

             Parameter     Standard
Variable      Estimate        Error   Type II SS  F Value  Pr > F

Intercept      1.38881      0.08447      6.60107   270.31  <.0001
x3             0.00565      0.00111      0.63221    25.89  <.0001
x4             0.13902      0.02206      0.96995    39.72  <.0001

The REG Procedure
Model: MODEL1
Dependent Variable: logy

Stepwise Selection: Step 2

Bounds on condition number: 1.2098, 4.8392
------------------------------------------------------------------------------------------
```

```
Stepwise Selection: Step 3

Variable x2 Entered: R-Square = 0.8829 and C(p) = 161.6520

                            Analysis of Variance

                                    Sum of           Mean
Source                    DF       Squares         Square    F Value    Pr > F

Model                      3       3.50743        1.16914     125.63    <.0001
Error                     50       0.46530        0.00931
Corrected Total           53       3.97273

             Parameter     Standard
Variable      Estimate        Error   Type II SS  F Value  Pr > F

Intercept      0.94229      0.07140      1.62097   174.18  <.0001
x2             0.00790   0.00086269      0.78014    83.83  <.0001
x3             0.00700   0.00070135      0.92684    99.60  <.0001
x4             0.08185      0.01498      0.27780    29.85  <.0001

Bounds on condition number: 1.4642, 11.822
------------------------------------------------------------------------------------------
```

```
Stepwise Selection: Step 4

Variable x1 Entered: R-Square = 0.9724 and C(p) = 5.0000

                            Analysis of Variance

                                    Sum of           Mean
Source                    DF       Squares         Square    F Value    Pr > F

Model                      4       3.86293        0.96573     430.98    <.0001
Error                     49       0.10980        0.00224
Corrected Total           53       3.97273

The REG Procedure
Model: MODEL1
Dependent Variable: logy

Stepwise Selection: Step 4

             Parameter     Standard
Variable      Estimate        Error   Type II SS  F Value  Pr > F

Intercept      0.48874      0.05024      0.21206    94.64  <.0001
x1             0.06853      0.00544      0.35550   158.65  <.0001
x2             0.00925   0.00043679      1.00587   448.90  <.0001
x3             0.00947   0.00039630      1.28071   571.55  <.0001
x4             0.00192      0.00971   0.00008741     0.04  0.8442

Bounds on condition number: 2.5553, 29.286
------------------------------------------------------------------------------------------
```

```
Stepwise Selection: Step 5

Variable x4 Removed: R-Square = 0.9723 and C(p) = 3.0390

                          Analysis of Variance

                                   Sum of           Mean
Source                     DF       Squares         Square    F Value    Pr > F

Model                       3       3.86284        1.28761     585.89    <.0001
Error                      50       0.10989        0.00220
Corrected Total            53       3.97273

             Parameter     Standard
Variable      Estimate        Error   Type II SS  F Value  Pr > F

Intercept      0.48362      0.04263      0.28280   128.68  <.0001
x1             0.06923      0.00408      0.63322   288.13  <.0001
x2             0.00929   0.00038255      1.29734   590.31  <.0001
x3             0.00952   0.00030644      2.12247   965.76  <.0001

Bounds on condition number: 1.0308, 9.1864
------------------------------------------------------------------------------------------

All variables left in the model are significant at the 0.0500 level.

No other variable met the 0.0100 significance level for entry into the model.
```

The REG Procedure
Model: MODEL1
Dependent Variable: logy

Summary of Stepwise Selection

Step	Variable Entered	Variable Removed	Label	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	x4		liver function	1	0.5274	0.5274	787.947	58.02	<.0001
2	x3		enzyme	2	0.1591	0.6865	507.807	25.89	<.0001
3	x2		prognostic	3	0.1964	0.8829	161.652	83.83	<.0001
4	x1		blood-clotting	4	0.0895	0.9724	5.0000	158.65	<.0001
5		x4	liver function	3	0.0000	0.9723	3.0390	0.04	0.8442

Type III: Variable selection by penalization

A third selection type is possible when the number of variables is very large.
We will only look at one method: The Least absolute shrinkage and selection operator, or LASSO, introduced by Tibshirani (1996).

LASSO

Suppose that all variables are centered. If that is not the case, we can first center them and obtain an estimator of the intercept in a second stage.
The LASSO estimator is the solution of the problem

$$\min_{\beta} \sum_{i=1}^{n} (y_i - \sum_{j=1}^{p} \beta_j x_{ij})^2, \quad \text{tel que } \sum_{j=1}^{p} |\beta_j| \leq t$$

where $t \geq 0$ is a coefficient to be determined.
The LASSO estimator minimizes the sum of squared residuals (as OLS), but imposes a constraint on the L_1 norm of β.

Lagrange multiplier form

An equivalent form is given by the optimization problem with Lagrange multiplier:

$$\min_{\beta} \sum_{i=1}^{n} (y_i - \sum_{j=1}^{p} \beta_j x_{ij})^2 + \lambda \sum_{j=1}^{p} |\beta_j|$$

where λ is the Lagrange multiplier.
Solution: Very efficient algorithms from linear programming, such as *least angle regression* (LARS) of Efron et al. (2004).

LASSO characteristics

- LASSO is non-linear in the observations.
- Analytic solution is not available.
- Numerical calculation very fast and efficient for several values λ simultaneously.
- $\lambda = 0$ corresponds the OLS estimator.
- The higher λ, the more the coefficients are shrunk to zero (*shrinkage*).
- Shrinkage degree is often measured by $s = |\beta_{LASSO}|/|\beta_{OLS}|$, $s \in [0, 1]$.

Example: Prostate cancer

- analyzed by Tibshirani (1996)
- 97 patients
- Variables: volume of tumor, age of patient, benign prostatic hyperplasia (bph), capsular penetration (cp), prostate-specific antigen (psa), Gleason score

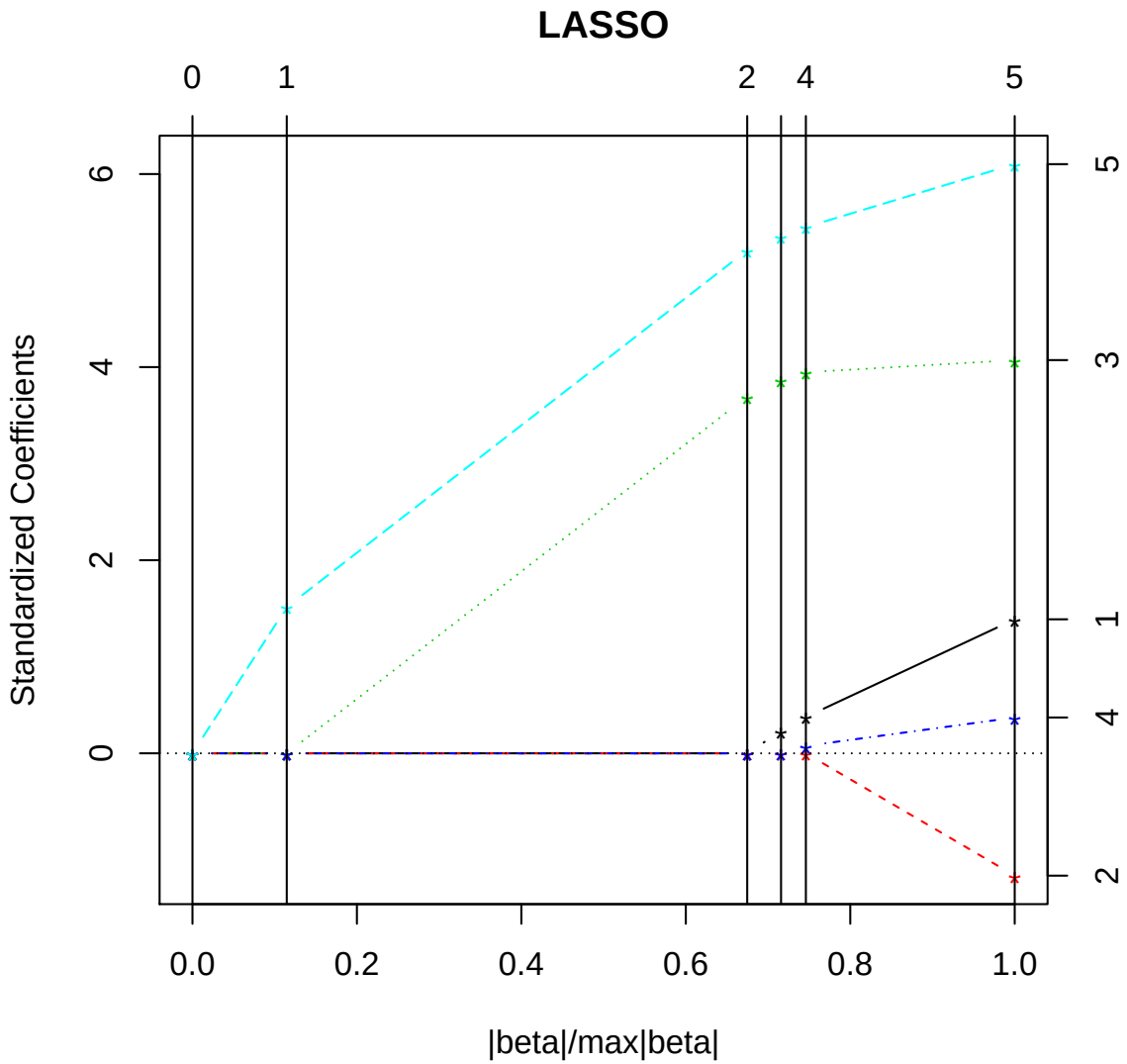
Estimation par OLS

```
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.49371    0.94261  -1.585   0.1165
age          0.01902    0.01063   1.789   0.0769 .
lbph        -0.08918    0.05376  -1.659   0.1006
lcp          0.29727    0.06762   4.396 2.98e-05 ***
gleason      0.05240    0.11965   0.438   0.6625
lpsa         0.53955    0.07648   7.054 3.30e-10 ***
---
Residual standard error: 0.7015 on 91 degrees of freedom
Multiple R-squared:  0.6642,    Adjusted R-squared:  0.6457
F-statistic:    36 on 5 and 91 DF,  p-value: < 2.2e-16
```

Estimation by LASSO

```
Sequence of LASSO moves:
     lpsa lcp age gleason lbph
Var     5   3   1       4    2
Step    1   2   3       4    5

             age         lbph        lcp    gleason      lpsa   s
[1,] 0.000000000  0.000000000 0.06519506 0.00000000 0.2128290   0.25
[2,] 0.000000000  0.000000000 0.18564339 0.00000000 0.3587292   0.5
[3,] 0.005369985 -0.001402051 0.28821232 0.01136331 0.4827810   0.75
[4,] 0.019023772 -0.089182565 0.29727207 0.05239529 0.5395488   1
```


LASSO
0
1
2
4
5
Standardized Coefficients
6
4
2
0
5
3
1
4
2
0.0
0.2
0.4
0.6
0.8
1.0
|beta|/max|beta|

Part IX

Violation of the classical hypotheses

Homoskedasticity

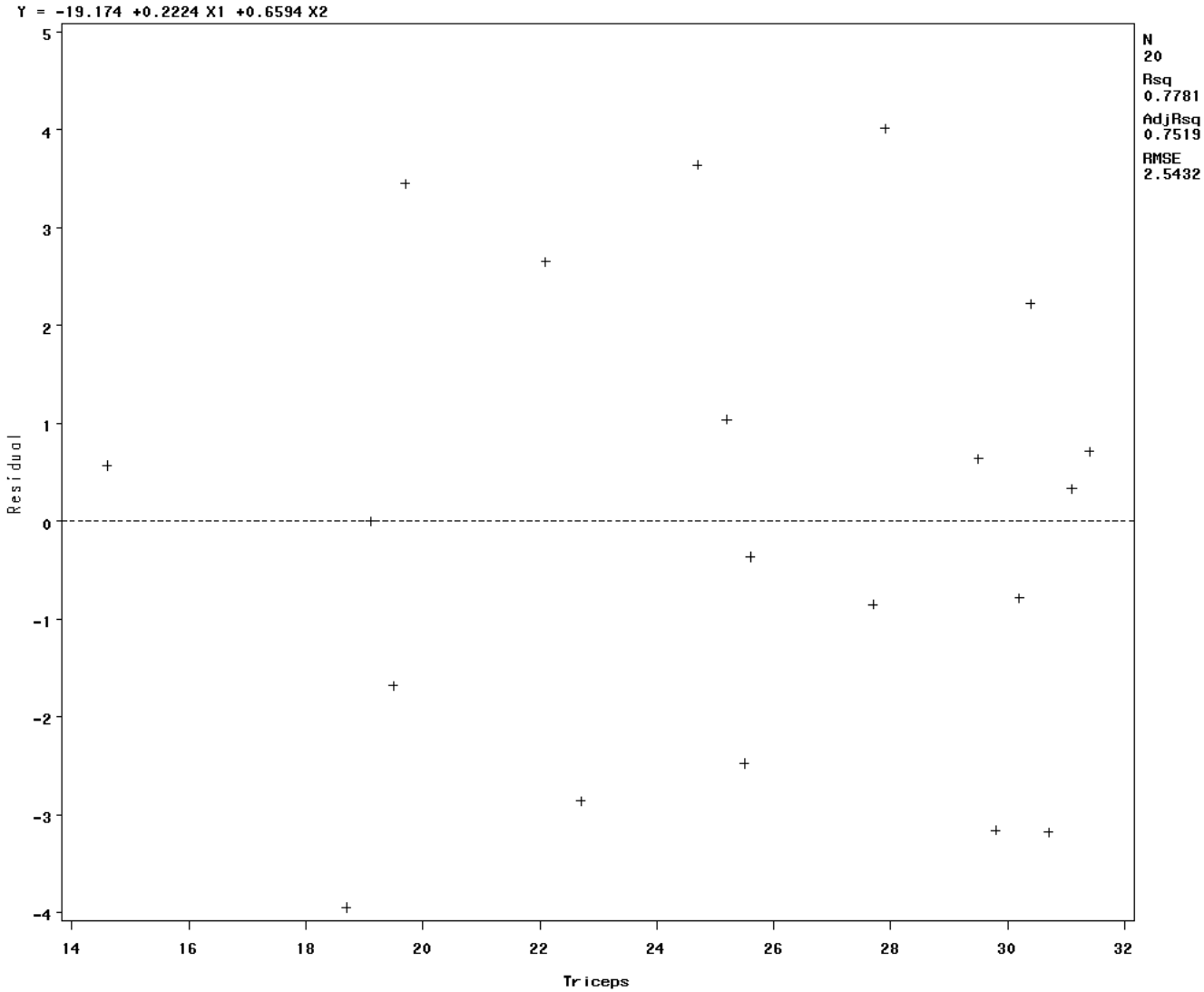
Diagnostics:

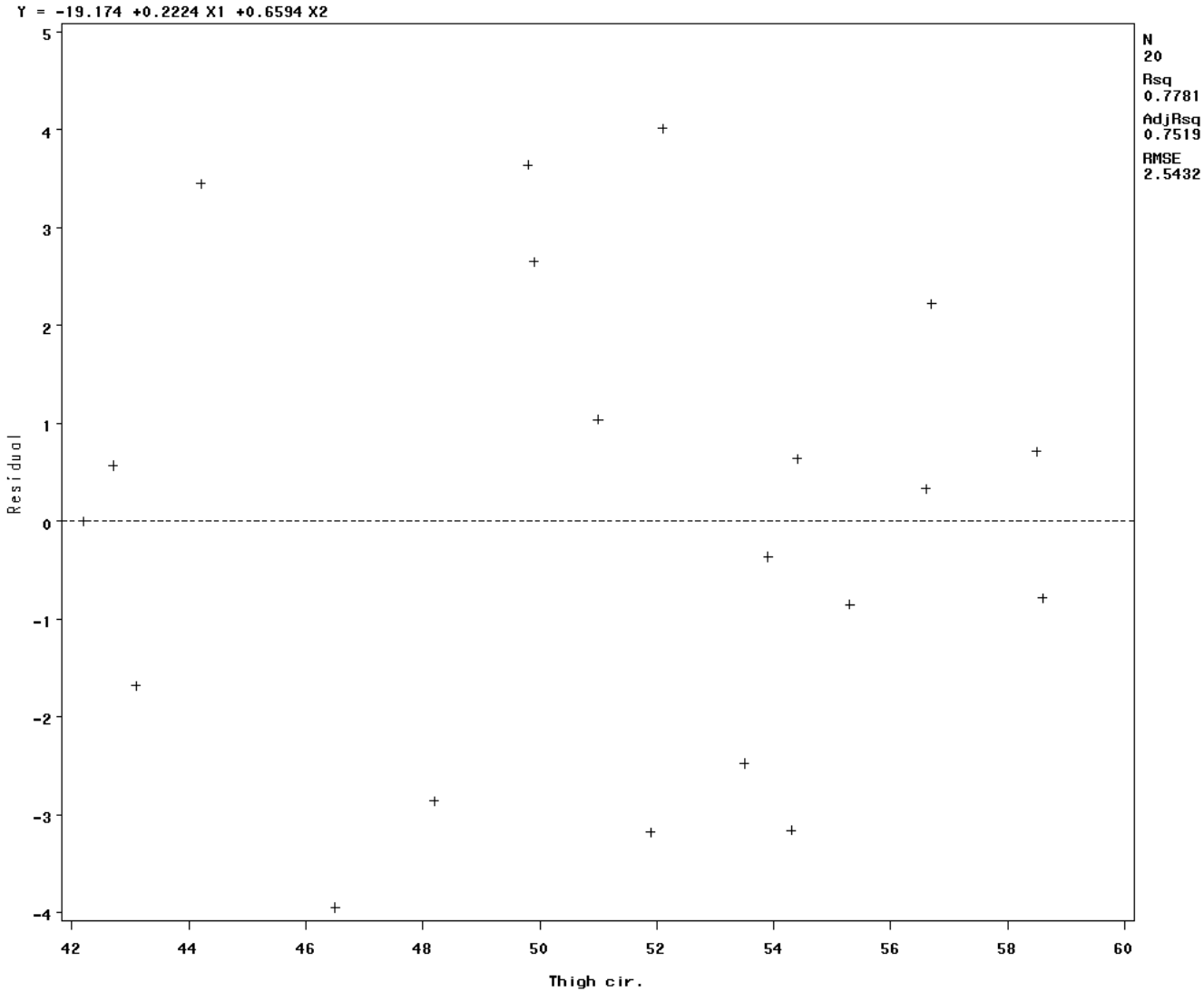
- Figure of e_i against X_{ij}, $j = 1 \ldots, p-1$
- Figure of e_i against $\hat{Y}_i$

Remedial measures: Box-Cox transformation, or the method of weighted least squares

Example: Body fat

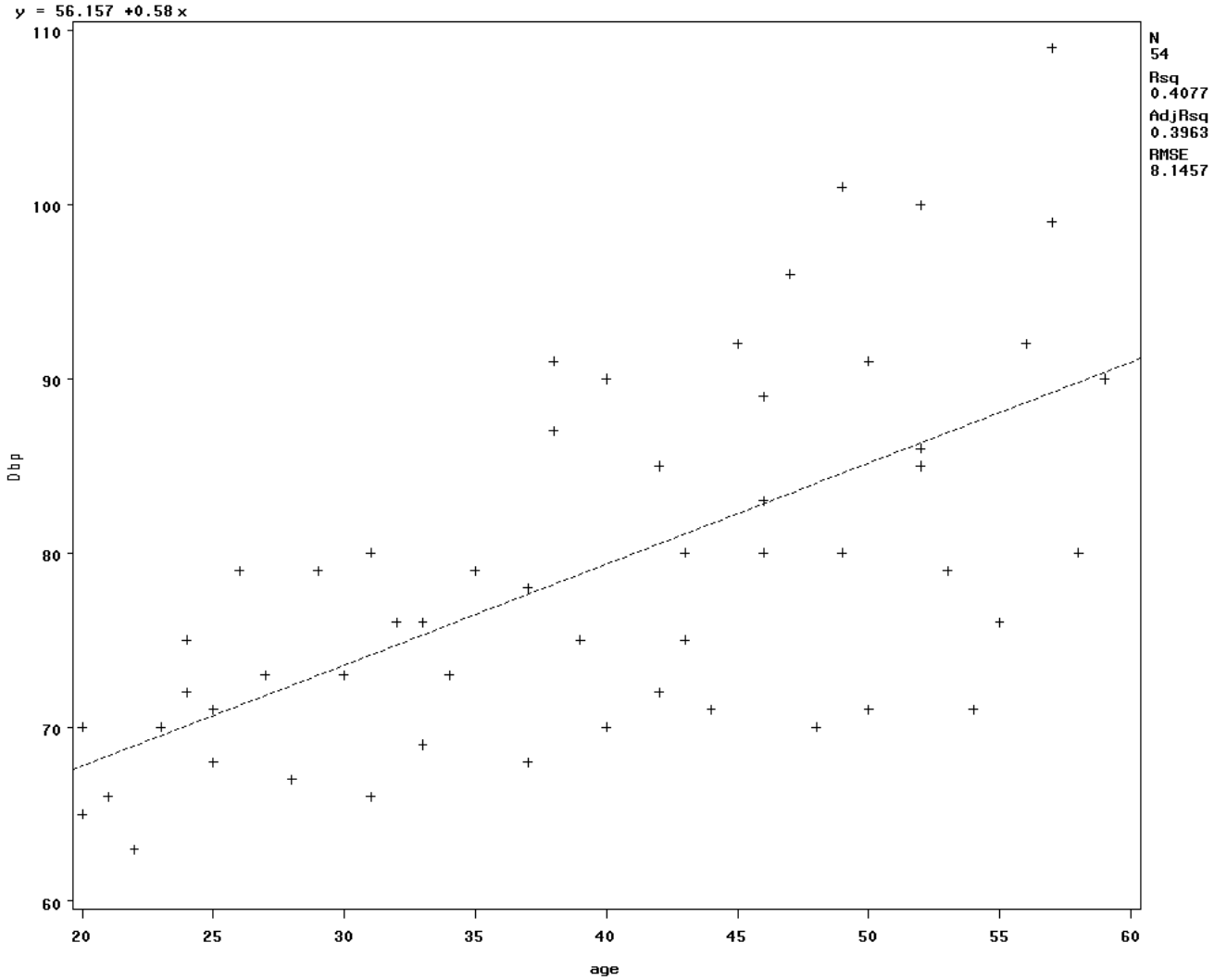
```
proc reg data = ch7tab01;
  model y = x1 x2;
  plot r.*x1 r.*x2;
run;
quit;
```

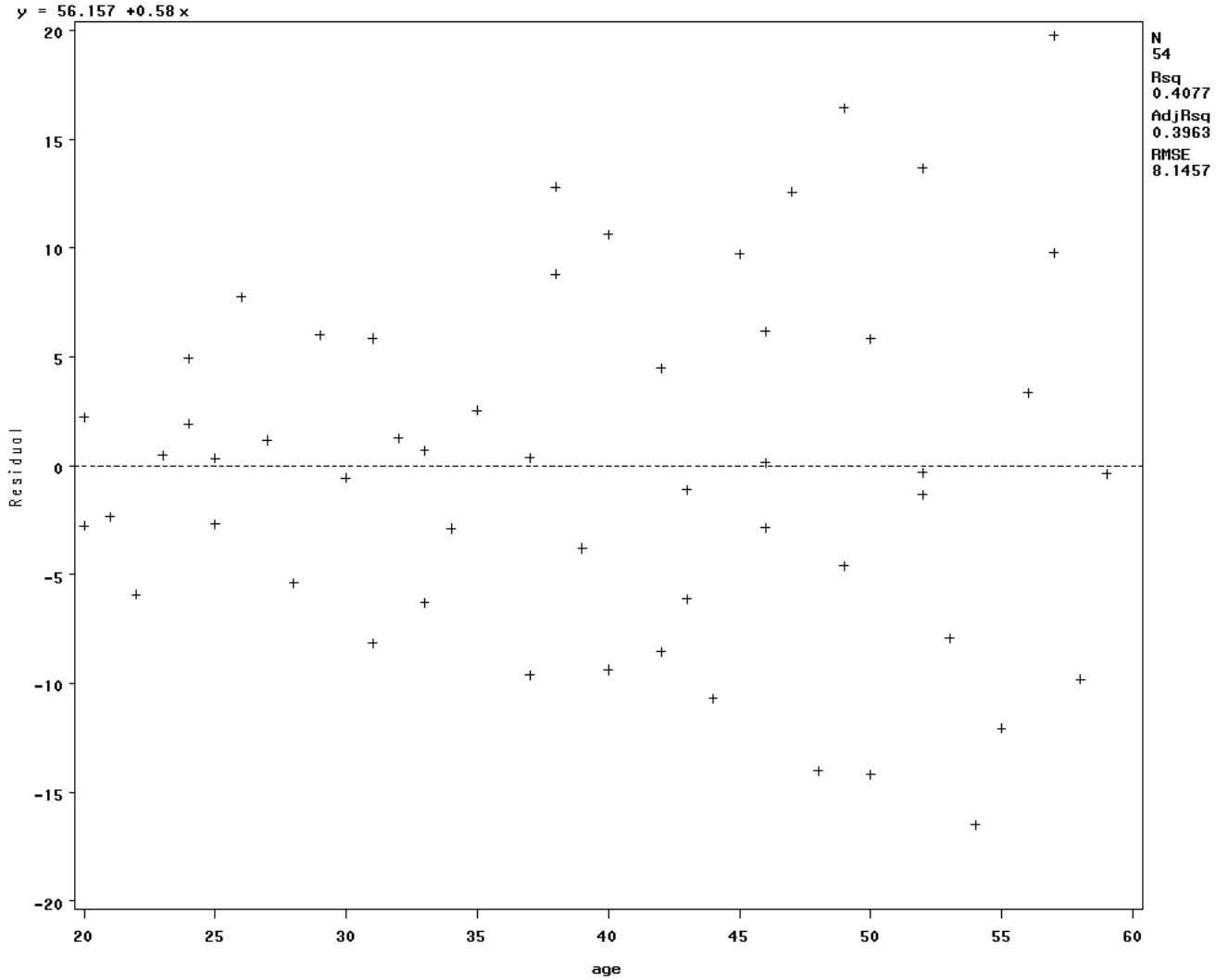

Y = -19.174 +0.2224 X1 +0.6594 X2
N
20
Rsq
0.7781
AdjRsq
0.7519
RMSE
2.5432
Residual
5
4
3
2
1
0
-1
-2
-3
-4
14
16
18
20
22
24
26
28
30
32
Triceps


Y = -19.174 +0.2224 X1 +0.6594 X2
N
20
Rsq
0.7781
AdjRsq
0.7519
RMSE
2.5432
Residual
5
4
3
2
1
0
-1
-2
-3
-4
42
44
46
48
50
52
54
56
58
60
Thigh cir.

Other example (Neter et al., blood pressure):

- Y: blood pressure (diastolic blood pressure)
- X: age
- $n = 54$


y = 56.157 +0.58 x
N
54
Rsq
0.4077
AdjRsq
0.3963
RMSE
8.1457
Dbp
age
110
100
90
80
70
60
20
25
30
35
40
45
50
55
60


y = 56.157 +0.58 x
N
54
Rsq
0.4077
AdjRsq
0.3963
RMSE
8.1457
Residuol
20
15
10
5
0
-5
-10
-15
-20
20
25
30
35
40
45
50
55
60
age

Testing homoskedasticity

A. Test of Goldfeld-Quandt (1965)

Split the sample in two groups: A and B

The idea is to test

$$H_0 : \sigma_A^2 = \sigma_B^2 \ .$$

How to compare $\hat{\sigma}_A^2$ and $\hat{\sigma}_B^2$? We can use that

$$(n_j - p)\frac{\hat{\sigma}_j^2}{\sigma_j^2} \sim \chi^2_{n_j - p}$$

where n_j is the sample size of the sub-sample $j = A$ or B

Since $\hat{\sigma}_A^2 \perp\!\!\!\perp \hat{\sigma}_B^2$, we have

$$\frac{\hat{\sigma}_A^2/\sigma_A^2}{\hat{\sigma}_B^2/\sigma_B^2} \sim F_{n_B-p}^{n_A-p} .$$

Hence, a natural test statistic is given by

$$T = \hat{\sigma}_A^2/\hat{\sigma}_B^2 \sim F_{n_B-p}^{n_A-p} \qquad \text{under} H_0 .$$

Inconvenient: the test depends on an a priori choice of A and B

B. Test of White (1980)

Idea: compare the matrices $X'X$ and $X'W^{-1}X$, where $W := \text{diag}(\sigma^2/\sigma_i^2)$.

Procedure:

1. Calculate $\hat{e}_i^2$ by OLS
2. Run the regression $\hat{e}_i^2$ with respect to all explanatory variables, their squares and cross-products.
3. Calculate the R^2 of this auxiliary regression
4. We can show that $T = nR^2 \sim \chi_J^2$ under H_0, where J is the number of variables in the auxiliary regression (not counting the intercept).

Example

If $Y = \beta_1 + \beta_2 X_1 + \beta_3 X_2 + \epsilon$, then the auxiliary regression is

$$\hat{e}_i^2 | 1, X_1, X_2, X_1^2, X_2^2, X_1 X_2$$

so here $J = 5$

Method of weighted least squares

Consider the heteroskedastic model

$$Y = X\beta + \varepsilon$$

where $rg(X) = p$, $\mathbb{E}[\varepsilon] = 0$, $\text{Var}(\varepsilon) = \text{diag}(\sigma_1^2, \ldots, \sigma_n^2)$ and $\sigma_i^2 = \sigma^2/w_i$
Let $W = \text{diag}(w_1, \ldots, w_n)$. Then we can rewrite the model as

$$Z = B\beta + \eta$$

where $Z = W^{1/2}Y$, $B = W^{1/2}X$ and $\eta = W^{1/2}\varepsilon$.

This new model satisfies $rg(B) = p$, $\mathbb{E}[\eta] = 0$ and $\text{Var}(\eta) = \sigma^2 I_n \Rightarrow$ Classical homoskedastic model

$$\begin{aligned} \hat{\beta} &= (B'B)^{-1}B'Z = (X'WX)^{-1}X'WY \\ \text{Var}(\hat{\beta}) &= \sigma^2(B'B)^{-1} = \sigma^2(X'WX)^{-1} \end{aligned}$$

This estimator is BLUE according to Gauss-Markov for a heteroskedastic model.
Problem: find the weights w_i

Estimation of the weights w_i

Two methods:

- sometimes σ_i^2 depends on a certain explanatory variable, for example

$$\begin{aligned} \sigma_i^2 &= \sigma^2 X_i \rightarrow w_i = 1/X_i \\ \sigma_i^2 &= \sigma^2 X_i^2 \rightarrow w_i = 1/X_i^2 \\ \sigma_i^2 &= \sigma^2 \sqrt{X_i} \rightarrow w_i = 1/\sqrt{X_i} \end{aligned}$$

- group the observations into s categories according to the value of one of the explanatory variables and estimate the variance $\sigma_j^2, j = 1 \ldots, s$. For each observation of the category j we have $w_i = 1/\sigma_j^2$.

OLS with robust inference

If we maintain estimation by OLS (which is unbiased), then its variance will change:

$$\text{Var}(\hat{\beta}_{OLS}) = \sigma^2 (X'X)^{-1} X'W^{-1}X(X'X)^{-1} \neq \sigma^2 (X'X)^{-1}$$

We have to estimate this matrix because W is unknown. Note first that

$$\sigma^2 X'W^{-1}X = \sum_{i=1}^{n} \sigma_i^2 X_i X_i'$$

The estimator of White (1980) replaces σ_i^2 by e_i^2, to obtain

$$V = \text{Var}(\hat{\beta}_{OLS}) = (X'X)^{-1} \left(\sum_{i=1}^{n} e_i^2 X_i X_i' \right) (X'X)^{-1}$$

which is given by most statistical software packages. This estimator is consistent and can be used for inference under heteroskedasticity.

Independence of the error terms

Another classical hypothesis was that the errors are independent, which implies that

$$\text{Cov}(\varepsilon_i, \varepsilon_j) = 0$$

for all $i \neq j$. Often, when variables are observed at different moments in time, one observes a problem of autocorrelation, i.e.

$$\text{Cov}(\varepsilon_t, \varepsilon_{t-k}) \neq 0$$

for some $k \neq 0$, which is a violation of the classical hypothesis:

$$\Psi := \text{Var}(\varepsilon_t)$$

is no longer a diagonal matrix.

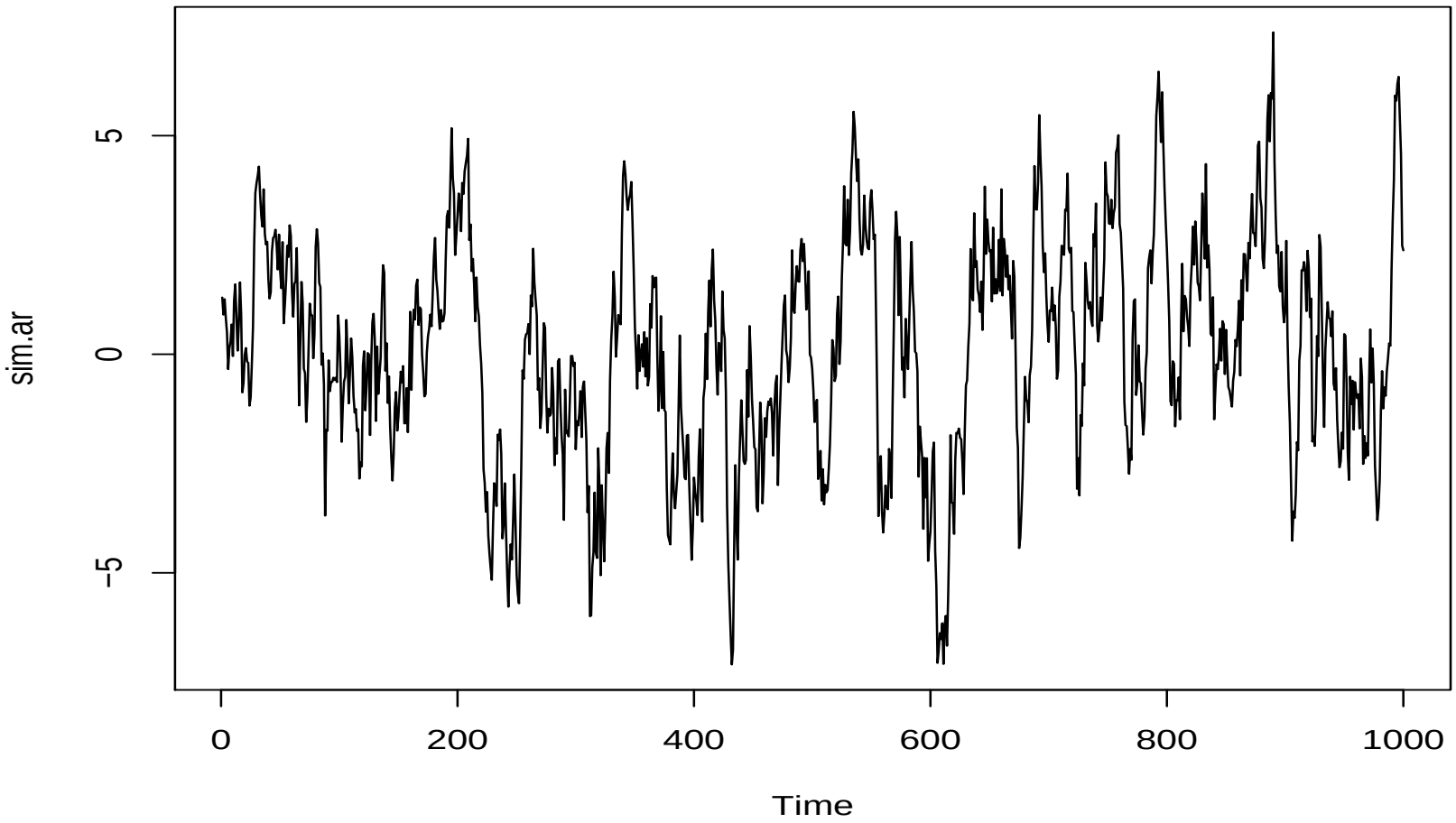
Example: Autoregressive process of order 1, AR(1)

$$\varepsilon_t = \rho\varepsilon_{t-1} + u_t, \quad t = 1, \ldots, T$$

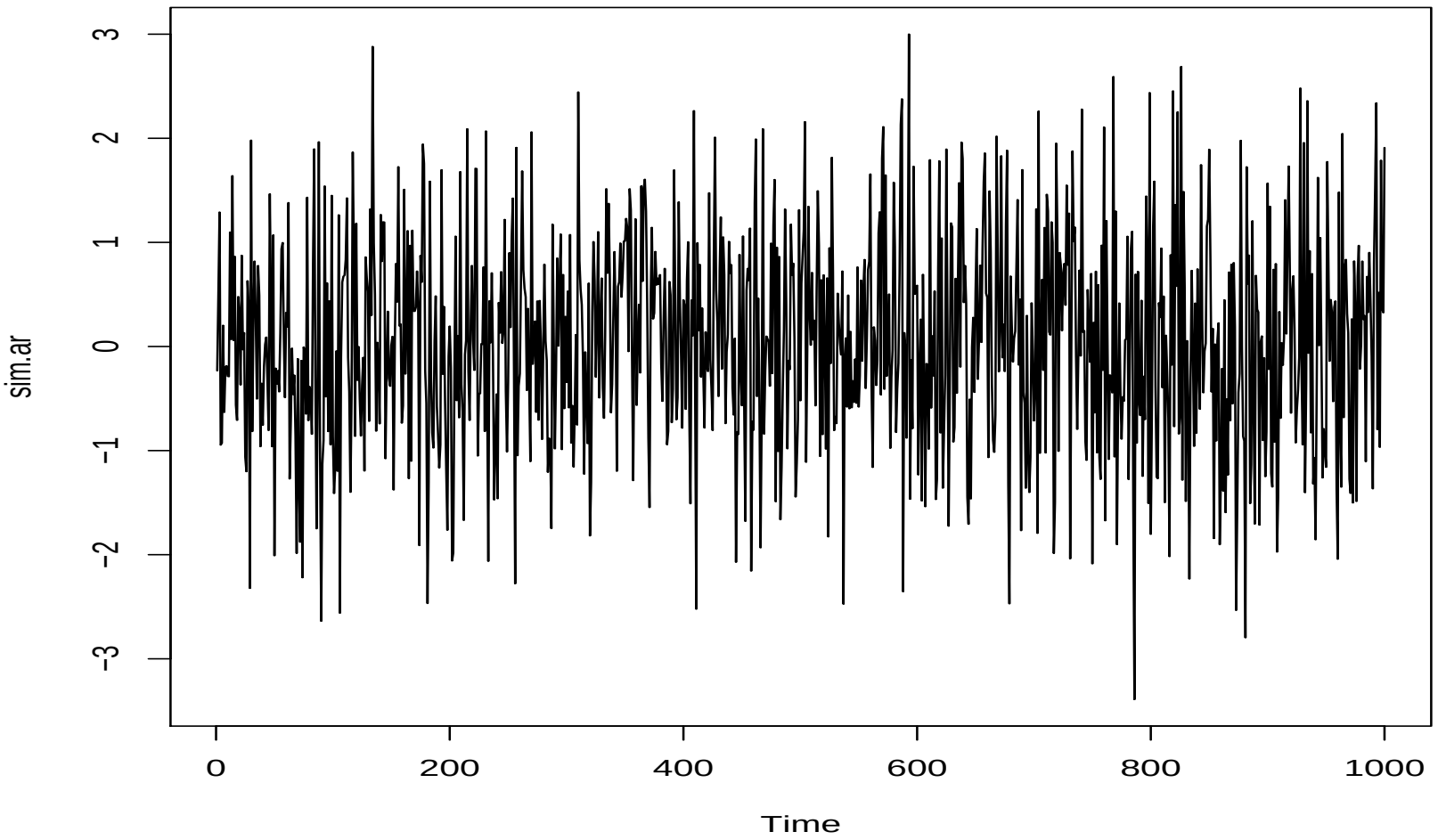
with $\rho \in (-1, 1)$, and u_t "white noise", i.e.,

$$\begin{aligned} \mathbb{E}[u_t] &= 0 \\ \mathrm{Var}(u_t) &= \sigma^2 \\ \mathrm{Cov}(u_t, u_{t-k}) = 0, &\quad \forall k \neq 0, \forall t \end{aligned}$$

AR(1) with $\rho = 0.9$



Gaussian white noise



Properties of an AR(1) process (to be shown)

$$\begin{aligned} \mathbb{E}[\varepsilon_t] &= 0 \\ \text{Var}(\varepsilon_t) &= \frac{\sigma^2}{1-\rho^2} \\ \text{Corr}(\varepsilon_t, \varepsilon_{t-k}|X) &= \rho^k, \quad k \in \mathbb{Z} \end{aligned}$$

The variance-covariance matrix of ε, for this example, can be written as

$$\Psi = \frac{\sigma^2}{1-\rho^2} \begin{pmatrix} 1 & \rho & \rho^2 & \dots & \rho^{T-1} \\ \rho & 1 & \rho & \dots & \rho^{T-2} \\ \rho^2 & \rho & 1 & \dots & \rho^{T-3} \\ \vdots & & & \ddots & \vdots \\ \rho^{T-1} & \rho^{T-2} & \dots & & 1. \end{pmatrix}$$

Example of autocorrelation

Poverty rate in the U.S. as a function of the unemployment rate. Yearly data from 1980 to 2003.

	Coefficient	Std.Error	t-value	t-prob	Part.R^2
Constant	9.79205	0.6112	16.0	0.0000	0.9211
UNEMPLOY	0.586614	0.09473	6.19	0.0000	0.6355

sigma	0.676259	RSS	10.0611649
R^2	0.63546	F(1,22) =	38.35 [0.000]**
Adj.R^2	0.61889	log-likelihood	-23.6221
no. of observations	24	no. of parameters	2

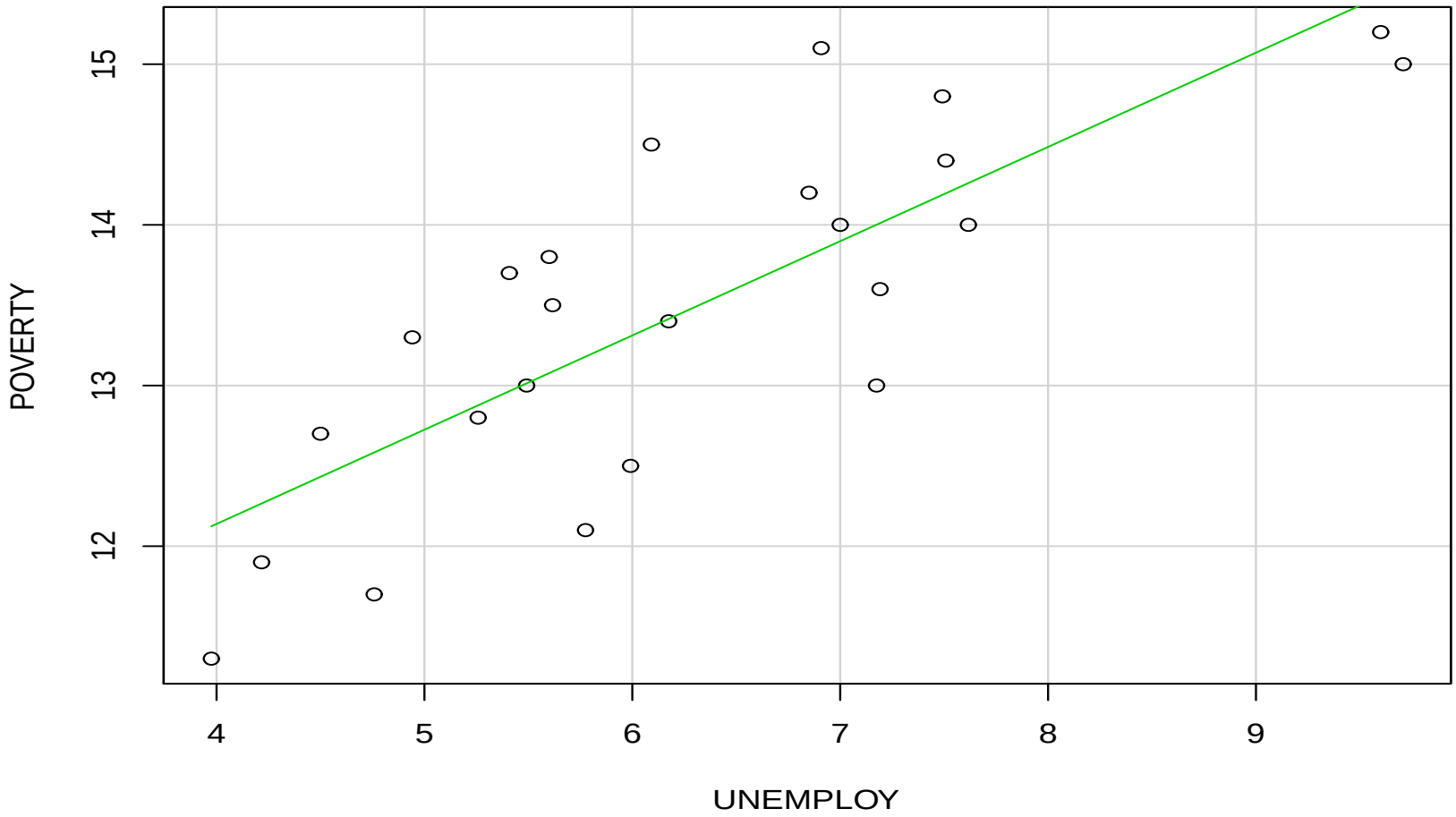


Figure: Poverty rate versus unemployment rate, U.S., 1980 to 2003. The straight line is the OLS regression fit.

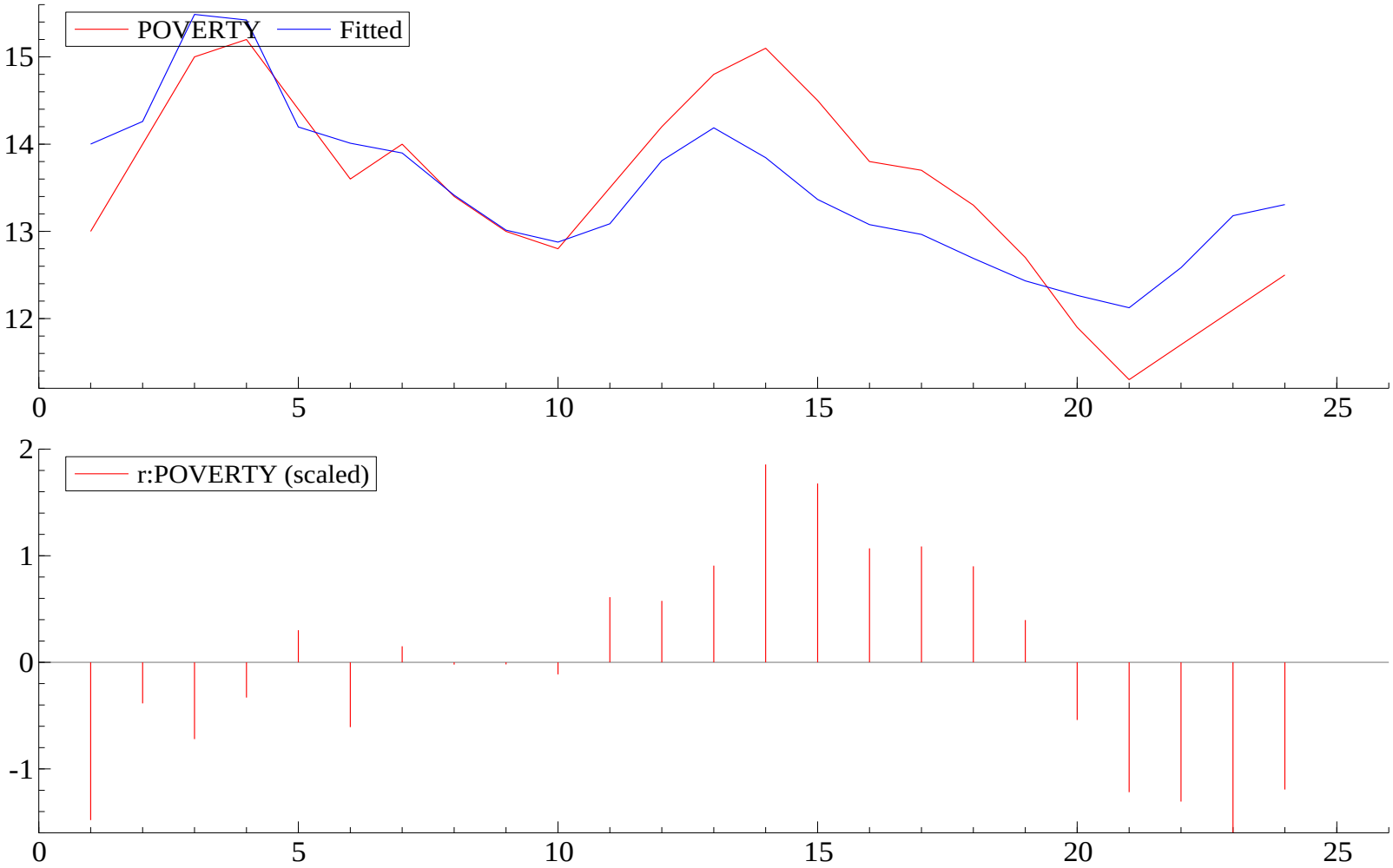


Figure: Above: time series of observed poverty rates (in red) and the fitted rates by OLS (in blue). Below: Residuals of the OLS regression.

Consequences of autocorrelation for the OLS estimator?

The OLS estimator

1. remains without bias
2. is no longer efficient
3. in general, has a variance different from $\sigma^2(X'X)^{-1}$.

Autocorrelation test of Breusch (1978) and Godfrey (1978)

Consider the model

$$Y_t = \alpha + \beta X_t + \varepsilon_t$$

and the hypothesis to be tested,

$$H_0 : Corr(\varepsilon_t, \varepsilon_{t-k}) = 0, \quad k = 1, \ldots, p$$

The steps of the Breusch-Godfrey test are the following:

1. Estimation of the model by OLS, which gives the residuals e_t
2. Estimate the regression

$$e_t = \alpha + \beta X_t + \rho_1 e_{t-1} + \ldots + \rho_p e_{t-p} + u_t$$

   and obtain the R^2 of this regression.
3. Under H_0, the statistic nR^2 follows an asymptotic χ^2 distribution with p degrees of freedom:

$$nR^2 \to \chi^2_p$$

How to treat autocorrelation?

A. Autocorrelation is known

In this case, we can construct a BLUE estimator.
For example, in the AR(1) case with known coefficient ρ, we can transform

$$Y_t - \rho Y_{t-1} = \alpha(1-\rho) + \beta(X_t - \rho X_{t-1}) + \varepsilon_t - \rho\varepsilon_{t-1}$$

and estimate the model by OLS.
Note that the error $u_t = \varepsilon_t - \rho\varepsilon_{t-1}$ is white noise.

B. Unknown autocorrelation

En supposant une forme d'autocorrélation, par exemple AR(1), on peut estimer le coefficient et construire un estimateur FGLS.

By supposing a form of autocorrelation, for example AR(1), we can estimate the coefficient and construct a feasible generalized least squares estimator (FGLS).

The Cochrane-Orcutt procedure:

1. Estimate the model by OLS and obtain $\hat{e}_t$.
2. Run the regression $\hat{e}_t = \rho \hat{e}_{t-1} + u_t$ and obtain the estimator

$$\hat{\rho} = \frac{\sum_{t=2}^{T} \hat{e}_t \hat{e}_{t-1}}{\sum_{t=2}^{T} \hat{e}_{t-1}^2}$$

3. Estimate by OLS

$$Y_t - \hat{\rho} Y_{t-1} = \alpha(1 - \hat{\rho}) + \beta(X_t - \hat{\rho} X_{t-1}) + \varepsilon_t - \hat{\rho} \varepsilon_{t-1}$$

C. Model specification

A residual autocorrelation can be an indicator of

1. Omitted variables, which are autocorrelated
2. Nonlinearities

Moreover, a supposed form of autocorrelation can be erroneous.

Normality of the error terms

Diagnostics: Figure which compares the empirical quantiles of the residuals with those of a normal distribution.
If $\varepsilon_i \sim N(0, \sigma^2)$, then $Z_i = \varepsilon_i/\sigma \sim N(0,1)$ and

$$\Phi(Z_1), \ldots, \Phi(Z_n) \sim U(0,1)$$

Let $\Phi(Z_{1:n}), \ldots, \Phi(Z_{n:n})$ be the order statistics.

$$\begin{aligned} \Phi(Z_{i:n}) &\approx \frac{i}{n} \\ \Rightarrow Z_{i:n} &\approx \Phi^{-1}(\frac{i}{n}) \\ \Rightarrow \frac{e_{i:n}}{\sqrt{MSE}} &\approx \Phi^{-1}(\frac{i}{n}) \\ \Rightarrow e_{i:n} &\approx \sqrt{MSE}\Phi^{-1}(\frac{i}{n}) \end{aligned}$$

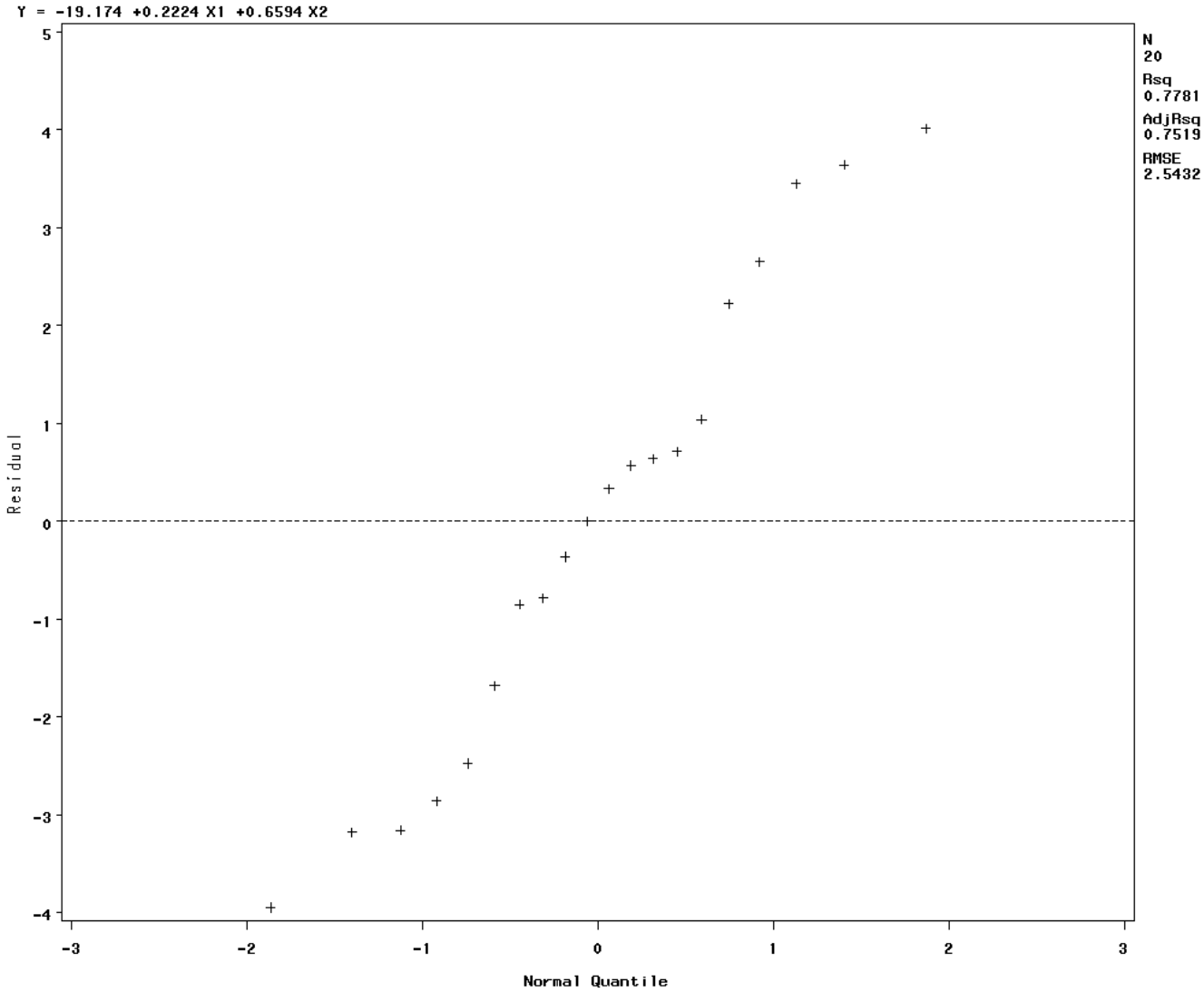
A better approximation is given by

$$e_{i:n} \approx \sqrt{MSE}\Phi^{-1}(\frac{i-0.375}{n+0.25})$$

The QQ plot (*Normal probability plot*) compares $e_{i:n}$ with $\Phi^{-1}(\frac{i-0.375}{n+0.25})$.
If the points are more or less on a line, then the normality hypothesis is acceptable.
Other normality tests, for example the test of Jarque-Bera (1980).

Exemple: Body fat

```
proc reg data = ch7tab01;
  model y = x1 x2;
  plot r.*nqq.;
run;
quit;
```


Y = -19.174 +0.2224 X1 +0.6594 X2
N
20
Rsq
0.7781
AdjRsq
0.7519
RMSE
2.5432
Residual
Normal Quantile
5
4
3
2
1
0
-1
-2
-3
-4
-3
-2
-1
0
1
2
3

The normality test of Jarque-Bera (1980)

Idea of the test: compare the coefficients of *skewness* S and *kurtosis* κ of the residuals with the theoretical values for a normal distribution:

$$H_0 : S = 0 \quad \text{and} \quad \kappa = 3$$

The test statistic of the Jarque-Bera test is given by

$$JB = \frac{n}{6}\left[S^2 + \frac{(\kappa - 3)^2}{4}\right] .$$

$$JB \sim \chi^2_2 \qquad \text{sous } H_0 .$$

Part X

Diagnostics

Diagnostics and remedial measures

Question:
Is the chosen model adequat?
In particular:

- Is the regression function linear?
- Are there outliers?
- Are there influential observations?

Analysis of the residuals $e_i = Y_i - \hat{Y}_i$ is important for the diagnostics.

Linearity of the regression function

Diagnostic: Figure of e_i against X_{ij}, $j = 1, \ldots, p-1$
For example, if the scatterplot is quadratic, then add a quadratic term in X_{ij} to the model.
In general, there are two types of remedial measures:

1. Modify the regression function, examples:
   - add quadratic terms: $\mathbb{E}[Y] = \beta_0 + \beta_1 X + \beta_2 X^2$
   - exponential form: $\mathbb{E}[Y] = \beta_0 \beta_1^X$
2. Transformation of the response variable (example: Box-Cox, which includes the logarithm)

Outliers

To identify outlying observations, we will first classify:

1. outliers with respect to X
2. outliers with respect to Y
3. outliers with respect to X and Y

Identify outliers with respect to X

Projection matrix $H = X(X'X)^{-1}X'$ with *leverage*

$$h_{ii} = (1, X_{i1} \dots X_{i,p-1})(X'X)^{-1}(1, X_{i1} \dots X_{i,p-1})'$$

Properties:

1. $0 \leq h_{ii} \leq 1$
2. $\sum_{i=1}^{n} h_{ii} = p$

Criterion: Mahalanobis distance

$$MD_i = (X_i - \bar{X})' C^{-1} (X_i - \bar{X})$$

with $X_i = (X_{i1}, \ldots, X_{i,p-1})'$, $\bar{X} = (\bar{X}_1, \ldots, \bar{X}_{p-1})'$ and the variance-covariance matrix of X,

$$C = \frac{1}{n} X_c' X_c, \quad X_c = \begin{pmatrix} X_{11} - \bar{X}_1 & \ldots & X_{1,p-1} - \bar{X}_{p-1} \\ \vdots & & \vdots \\ X_{n1} - \bar{X}_1 & \ldots & X_{n,p-1} - \bar{X}_{p-1} \end{pmatrix}$$

MD_i measures the distance between the point X_i and the mean by taking into account the "structure" of the data.

Link between the Mahalanobis distance and the leverage:

$$MD_i \approx nh_{ii} - 1$$

$\Rightarrow$ it suffices to study the leverages $h_{ii}, i = 1, \ldots, n$.

If h_{ii} is large (and hence X_i is an outlier), Y_i has a strong influence for $\hat{Y}_i$:

- The fitted values are

$$\hat{Y}_i = h_{i1} Y_1 + \ldots + h_{ii} Y_i + \ldots + h_{in} Y_n$$

- The residuals are $e = Y - \hat{Y} = (I_n - H)Y$ with variance-covariance

$$\begin{aligned} \text{Var}(e) &= (I_n - H)\text{Var}(Y)(I_n - H)' \\ &= (I_n - H)\sigma^2 I_n (I_n - H)' \\ &= \sigma^2 (I_n - H) \end{aligned}$$

and hence $\text{Var}(e_i) = \sigma^2(1 - h_{ii})$
For the observations with large leverages, the variances of the corresponding residuals are small and $\hat{Y}_i$ is close to Y_i.

Which threshold should we take to decide if a leverage is "large"?
Note that

$$\frac{1}{n}\sum_{i=1}^{n} h_{ii} = \frac{p}{n}$$

We say that an observation X_i is an outlier if

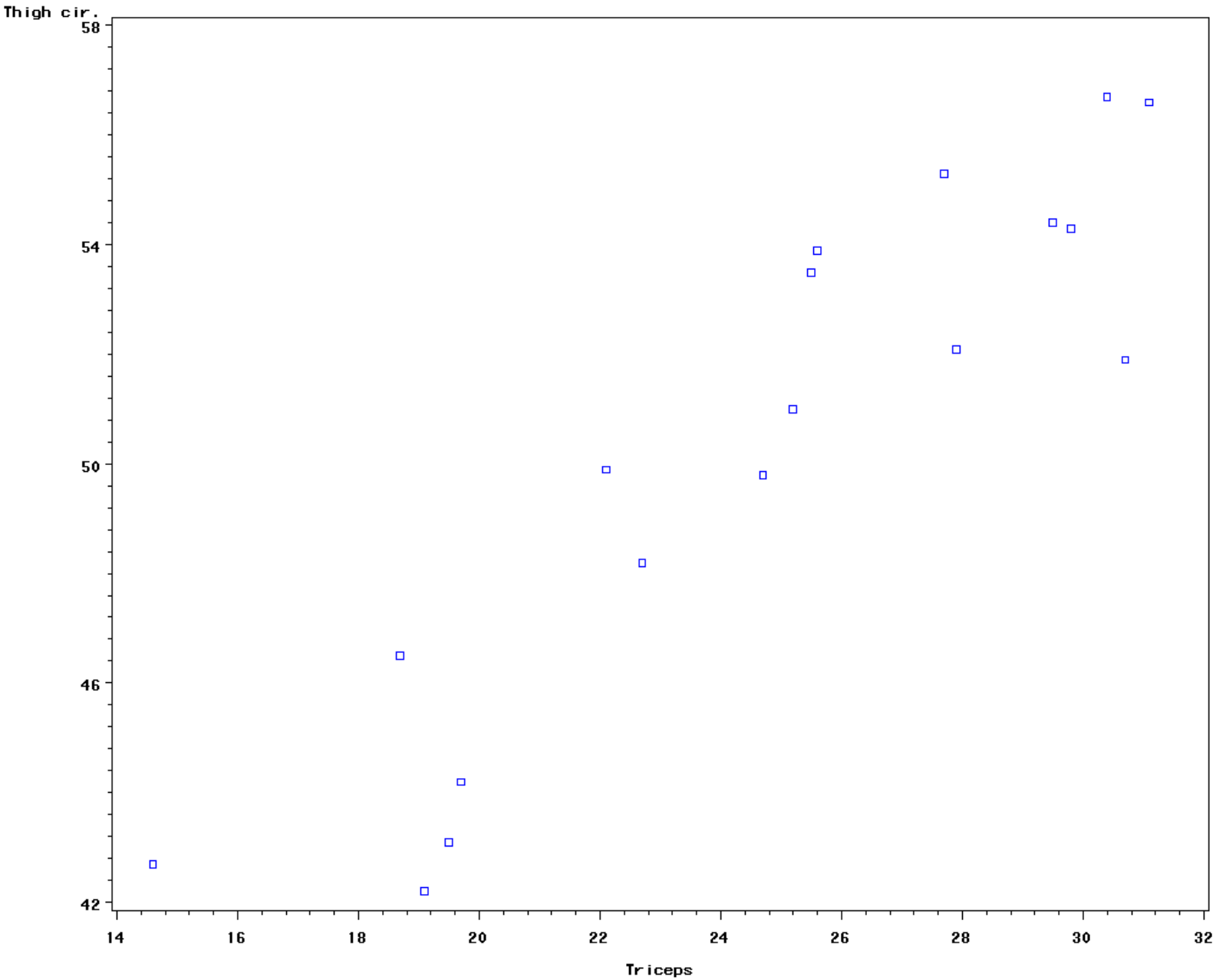
$$h_{ii} > \frac{2p}{n},$$

which will be our decision rule to detect outliers with respect to X.

Exemple: Body fat

```
proc reg data = ch7tab01 noprint ;
  model y = x1 x2  ;
  output out=temp r=residual h=hat rstudent=rstudent;
run;
proc print data = temp;
 var residual hat rstudent;
run;
```

Obs	residual	hat	rstudent
1	-1.68271	0.20101	-0.72999
2	3.64293	0.05889	1.53425
3	-3.17597	0.37193	-1.65433
4	-3.15847	0.11094	-1.34848
5	-0.00029	0.24801	-0.00013
6	-0.36082	0.12862	-0.14755
7	0.71620	0.15552	0.29813
8	4.01473	0.09629	1.76009
9	2.65511	0.11464	1.11765
10	-2.47481	0.11024	-1.03373
11	0.33581	0.12034	0.13666
12	2.22551	0.10927	0.92318
13	-3.94686	0.17838	-1.82590
14	3.44746	0.14801	1.52476
15	0.57059	0.33321	0.26715
16	0.64230	0.09528	0.25813
17	-0.85095	0.10559	-0.34451
18	-0.78292	0.19679	-0.33441
19	-2.85729	0.06695	-1.17617
20	1.04045	0.05009	0.40936


Thigh cir.
58
54
50
46
42
14
16
18
20
22
24
26
28
30
32
Triceps

Identify outliers with respect to Y

Idea: study the residuals $e_i = Y_i - \hat{Y}_i$

standardized residuals:

$$e_i^* = \frac{e_i}{\sqrt{MSE(1 - h_{ii})}}$$

If e_i^* is "large", then the corresponding values Y_i are outliers.

Second method: calculate the regression by suppressing the i-th observation from the data.
Let $\hat{Y}_{i(i)}$ be the fitted value of Y_i based on the observations $1, \ldots, i-1, i+1, \ldots, n$.
"deleted residual", $d_i = Y_i - \hat{Y}_{i(i)}$
Fortunately (will be shown later):

$$d_i = \frac{e_i}{1 - h_{ii}}$$

This means that we can obtain the deleted residuals from the complete regression.

Sherman-Morrison formula

Let A be a matrix, and a, b column vectors. Then,

$$(A - ab')^{-1} = A^{-1} + \frac{A^{-1}ab'A^{-1}}{1 - b'A^{-1}a}$$

We will use this formula with $A = (X'X)$ and $a = b = x_i$

The estimator of β without the i-th observation:

$$\hat{\beta}_{(i)} = (X'_{(i)}X_{(i)})^{-1}X'_{(i)}Y_{(i)}$$

By Sherman-Morrison,

$$(X'_{(i)}X_{(i)})^{-1} = (X'X)^{-1} + \frac{(X'X)^{-1}x_i x_i'(X'X)^{-1}}{1 - h_{ii}}$$

$$
\begin{aligned}
\hat{\beta} - \hat{\beta}_{(i)} &= (X'X)^{-1} \left\{ X'Y - X'_{(i)} Y_{(i)} - \frac{x_i x_i' (X'X)^{-1} X'_{(i)} Y_{(i)}}{1 - h_{ii}} \right\} \\
&= (X'X)^{-1} \left\{ x_i y_i - \frac{x_i x_i' (X'X)^{-1}}{1 - h_{ii}} X'_{(i)} Y_{(i)} \right\} \\
&= (X'X)^{-1} \frac{x_i y_i (1 - h_{ii}) - x_i x_i' (X'X)^{-1} X'_{(i)} Y_{(i)}}{1 - h_{ii}} \\
&= (X'X)^{-1} \frac{x_i}{1 - h_{ii}} \left\{ y_i - x_i' (X'X)^{-1} x_i y_i - x_i' (X'X)^{-1} X'_{(i)} Y_{(i)} \right\} \\
&= (X'X)^{-1} \frac{x_i}{1 - h_{ii}} \left\{ y_i - x_i' (X'X)^{-1} [x_i y_i + X'_{(i)} Y_{(i)}] \right\} \\
&= (X'X)^{-1} \frac{x_i}{1 - h_{ii}} \left\{ y_i - x_i' (X'X)^{-1} X'Y \right\} \\
&= (X'X)^{-1} \frac{x_i}{1 - h_{ii}} e_i
\end{aligned}
$$

Difference in fitted values

$$\begin{aligned}
\hat{Y}_i - \hat{Y}_{i(i)} &= x_i'(\hat{\beta} - \hat{\beta}_{(i)}) \\
&= x_i'(X'X)^{-1}\frac{x_i}{1-h_{ii}}e_i \\
&= \frac{h_{ii}}{1-h_{ii}}e_i
\end{aligned}$$

This implied for the deleted residuals:

$$\begin{aligned}
d_i &= e_i + \hat{Y}_i - \hat{Y}_{i(i)} \\
&= e_i + \frac{h_{ii}}{1-h_{ii}}e_i \\
&= \frac{e_i}{1-h_{ii}}
\end{aligned}$$

Standardized deleted residuals

$$d_i^* = \frac{d_i}{\sqrt{MSE_{(i)}/(1-h_{ii})}}$$

where $MSE_{(i)}$ is the MSE without the i-th observation, i.e.,

$$\begin{aligned} MSE_{(i)} &= \frac{1}{n-p-1}\sum_{j\neq i}(Y_j - \hat{Y}_{j(i)})^2 \\ &= \frac{1}{n-p-1}SSE_{(i)} \end{aligned}$$

$$
\begin{aligned}
SSE_{(i)} &= \sum_{j \neq i} (y_j - x_j' \hat{\beta}_{(i)})^2 \\
&= \sum_{j \neq i} \left(y_j - x_j' \hat{\beta} + \frac{x_j'(X'X)^{-1} x_i e_i}{1 - h_{ii}} \right)^2 \\
&= \sum_{j \neq i} \left(e_j + \frac{h_{ji} e_i}{1 - h_{ii}} \right)^2 \\
&= \sum_{j=1}^{n} \left(e_j + \frac{h_{ji} e_i}{1 - h_{ii}} \right)^2 - \frac{e_i^2}{(1 - h_{ii})^2}
\end{aligned}
$$

$$
\begin{aligned}
&= SSE + \frac{2e_i}{1-h_{ii}} \underbrace{\sum_{j=1}^{n} e_j h_{ji}}_{=0} + \frac{e_i^2}{(1-h_{ii})^2} \sum_j h_{ji}^2 - \frac{e_i^2}{(1-h_{ii})^2} \\
&= SSE + \frac{e_i^2}{(1-h_{ii})^2} h_{ii} - \frac{e_i^2}{(1-h_{ii})^2} \\
&= SSE - \frac{e_i^2}{1-h_{ii}}
\end{aligned}
$$

Consequently,

$$
MSE_{(i)} = \frac{1}{n-p-1} \left\{ SSE - \frac{e_i^2}{1-h_{ii}} \right\}
$$

The standardized deleted residuals

$$d_i^* = e_i \left[\frac{n-p-1}{SSE(1-h_{ii}) - e_i^2} \right]^{1/2},$$

are thus easy to obtain.
We have the property

$$d_i^* \sim t_{n-p-1}.$$

Decision rule: Y_i is an outlier if

$$|d_i^*| > t_{n-p-1;1-\alpha/2}$$

Influence for the fitted values: DFFITS

The i-th observation is influential for its fitted value if $\hat{Y}_i - \hat{Y}_{i(i)}$ is "large". To standardize the difference we have to calculate the variance of $\hat{Y}_i$:

$$\begin{aligned}
\text{Var}(\hat{Y}_i) &= \text{Var}(\sum_{j=1}^{n} h_{ij} Y_j) \\
&= \sum_{j=1}^{n} h_{ij}^2 \text{Var}(Y_j) \\
&= \sigma^2 \sum_{j=1}^{n} h_{ij}^2 \\
&= \sigma^2 h_{ii}
\end{aligned}$$

Definition:

$$DFFITS_i = \frac{\hat{Y}_i - \hat{Y}_{i(i)}}{\sqrt{MSE_{(i)} h_{ii}}}$$

which will be our measure for the influence of the i-th observation for its fitted value.
This measure can be decomposed:

$$DFFITS_i = d_i^* \sqrt{\frac{h_{ii}}{1 - h_{ii}}}$$

The first factor, d_i^*, is large if Y_i is an outlier and the second is large if X_i is an outlier.

Decision rule:

- For $n < 30$, the observation is influential if $|DFFITS_i| > 1$
- For $n \geq 30$, the observation is influential if $|DFFITS_i| > 2\sqrt{p/n}$.

Influence for the regression coefficients: DFBETAS

Idea: study the difference between the estimated coefficients,

$$\hat{\beta} - \hat{\beta}_{(i)} = (X'X)^{-1} \frac{x_i}{1 - h_{ii}} e_i$$

The standardized difference for the k-th element (with $c_k = (X'X)^{-1}_{kk}$):

$$DFBETAS_{k,i} = \frac{\hat{\beta}_k - \hat{\beta}_{k(i)}}{\sqrt{MSE_{(i)} c_k}}$$

Deision rule:

- For $n < 30$, the i-th observation is influential if $|DFBETAS_{k,i}| > 1$
- For $n \geq 30$ the i-th observation i is influential if $|DFBETAS_{k,i}| > 2/\sqrt{n}$

Cook's distance

Recall that a confidence region for the regression coefficients is determined by

$$\frac{(\hat{\beta}-\beta)'X'X(\hat{\beta}-\beta)}{pMSE} \sim F_{p,n-p}$$

Define Cook's distance as

$$D_i = \frac{(\hat{\beta}-\hat{\beta}_{(i)})'X'X(\hat{\beta}-\hat{\beta}_{(i)})}{pMSE}$$

measures the distance between $\hat{\beta}$ and $\hat{\beta}_{(i)}$.

Moreover, one can show that

$$\begin{aligned} D_i &= \frac{e_i^2}{pMSE}\frac{h_{ii}}{(1-h_{ii})^2} \\ &= \frac{(\hat{Y}-\hat{Y}_{(i)})'(\hat{Y}-\hat{Y}_{(i)})}{pMSE} \end{aligned}$$

measures the distance between $\hat{Y}$ and $\hat{Y}_{(i)}$.
Thus, D_i measures the influence of the i-th observation on the coefficients and on the fitted values
Decision rule:

$$D_i > F_{p,n-p;1-\alpha}$$

Remedial measures for influential observations

- In general, it is not recommended to exclude outlying or influential observations from the analysis.
- Select another model, for example by adding explanatory variables, adding quadratic or interaction terms, or by transforming the variables.
- Select a "robust" estimation method, such as minimizing the sum of absolute residuals:

$$\sum_{i=1}^{n} |Y_i - \beta_0 - \beta_1 X_{i1} - \ldots - \beta_{p-1} X_{i,p-1}|$$

which is more complex than OLS and does not yield closed form solutions for the coefficients.

Part XI

Panel data

Introduction

- Panel data are time series of cross-sections.
- Examples: panel of households, firms, or countries
- By pooling time series, we can obtain more efficient parameter estimates, test more sophisticated behavioral models with less restrictive assumptions
- ability to control for individual heterogeneity
- identify and estimate dynamic effects, e.g. the change of the proportion of unemployed from one period to the next.
- The discussion follows Baltagi (2008), Econometric Analysis of Panel Data, Wiley.

Error components model

The model is a usual linear regression, but with double-subscripts:

$$Y_{it} = \alpha + X_{it}'\beta + u_{it}$$

where $i = 1, \ldots, N$ is the cross-section index, $t = 1, \ldots, T$ is the time index, and X_{it} is a K-dimensional vector of explanatory variables for Y_{it}. The error components model specifies

$$u_{it} = \mu_i + \nu_{it}$$

where μ_i are cross-section specific components, and ν_{it} an error term.

Examples of individual effects

- individual ability in explaining earnings
- managerial skills in the production function
- country specific effects

The observations can be stacked into vectors, to obtain

$$Y = \alpha \iota_{NT} + X\beta + u = Z\delta + u$$

where Y is $(NT \times 1)$, $\iota_{NT} = (1, \ldots, 1)'$, X is $(NT \times K)$, $Z = [\iota_{NT}, X]$, and $\delta' = (\alpha, \beta')$.
The error components are, in vector form,

$$u = Z_\mu \mu + \nu$$

where $Z_\mu = I_N \otimes \iota_T$.

Time averages

Note that $Z_\mu Z_\mu' = I_N \otimes J_T$, where J_T is a $(T \times T)$ matrix of ones.
Let $P := Z_\mu (Z_\mu' Z_\mu)^{-1} Z_\mu$ be the projection matrix on Z_μ. This reduces to

$$P = I_N \otimes \bar{J}_T$$

where $\bar{J}_T = J_T / T$.
The matrix P averages the observations over time for each individual.
For example, Pu has typical element $\bar{u}_{i\cdot} = \sum_{t=1}^{T} u_{it}/T$, repeated T times for each individual.

Deviations from time averages

Define the orthogonal projection $Q = I_{NT} - P$. This projection obtains the deviations from individual means.
For example, Qu has typical element $u_{it} - \bar{u}_{i\cdot}$.
P and Q are symmetric and idempotent, and orthogonal complements, i.e. $PQ = 0$ and $P + Q = I_{NT}$.

The Fixed Effects Model

Consider the μ_i as fixed parameters.

$$Y_{it} = \alpha + X_{it}'\beta + \sum_{i=1}^{N} \mu_i D_i + \nu_{it}$$

where ν_{it} is an error term with variance σ_ν^2, and D_i is a dummy variable that takes value 1 for individual i, and 0 otherwise.
Need an identification restriction, e.g. $\sum_{i=1}^{N} \mu_i = 0$.

Estimation

Joint estimation of the parameters α, β, μ by OLS is possible, but problematic if N is large: Inversion of $(N + K + 1 \times N + K + 1)$ dimensional matrices.
It is preferable to apply the FWL theorem: first obtain deviations from the time averages, and then apply OLS to obtain estimates of α and β. This is called the least squares dummy variables estimator, or Within estimator.

Within estimator

Suppose $K = 1$ for simplicity. Then

$$Y_{it} = \alpha + \beta X_{it} + \mu_i + \nu_{it}$$

Averaging over time gives

$$\bar{Y}_{i\cdot} = \alpha + \beta \bar{X}_{i\cdot} + \mu_i + \bar{\nu}_{i\cdot}$$

Denote $\tilde{Y}_{it} := Y_{it} - \bar{Y}_{i\cdot}$ and $\tilde{X}_{it} := X_{it} - \bar{X}_{i\cdot}$. Then

$$\tilde{Y}_{it} = \beta \tilde{X}_{it} + \tilde{\nu}_{it}$$

Then, estimate β by OLS.

In matrix notation,

$$Y = \alpha \iota_{NT} + X\beta + Z_\mu \mu + \nu = Z\delta + Z_\mu \mu + \nu$$

where Z is $(NT \times (K+1))$ and Z_μ is $(NT \times N)$.
Reduce the dimension by premultiplying by Q:

$$QY = QX\beta + Q\nu$$

because $QZ_\mu = 0$. The Q matrix wipes out individual effects. Note that $Q\nu \sim (0, \sigma_\nu^2 Q)$.
Then apply OLS to obtain the Within estimator:

$$\hat{\beta}_W = (X'QX)^{-1} X'QY$$

Test for individual effects: One can test for individual effects, $H_0 : \mu_1 = \ldots = \mu_{N_1} = 0$ using a standard ANOVA-type test.

The Between estimator

The Between estimator runs the regression of averages across time, e.g.

$$\bar{Y}_{i\cdot} = \alpha + \bar{X}_{i\cdot}\beta + \bar{u}_{i\cdot}, \quad i = 1, \ldots, N$$

or in vector notation

$$PY = \alpha P\iota_{NT} + PX\beta + Pu$$

We can get rid of α by

$$(P - \bar{J}_{NT})Y = (P - \bar{J}_{NT})X\beta + (P - \bar{J}_{NT})u$$

because $(P - \bar{J}_{NT})\iota_{NT} = 0$. This gives

$$\hat{\beta}_B = (X'(P - \bar{J}_{NT})X)^{-1}X'(P - \bar{J}_{NT})Y$$

The random effects model

We now assume that μ_i are random variables:

$$\mu_i \sim iid(0, \sigma_\mu^2), \quad \nu_{it} \sim iid(0, \sigma_\nu^2)$$

and the μ_i are independent of ν_{it}.
By construction,

$$\text{Cov}(u_{it}, u_{js}) = \left\{ \begin{array}{ll} \sigma_\mu^2 + \sigma_\nu^2, & i = j, t = s \\ \sigma_\mu^2, & i = j, t \neq s \\ 0 & \text{otherwise} \end{array} \right\}$$

In vector form, since $u = Z_\mu \mu + \nu$, it has var-cov matrix

$$\begin{aligned} \Omega := \mathbb{E}(uu') &= Z_\mu \mathbb{E}(\mu\mu') Z_\mu' + \mathbb{E}(\nu\nu') \\ &= \sigma_\mu^2 (I_N \otimes J_T) + \sigma_\nu^2 (I_N \otimes I_T) \end{aligned}$$

Alternative expression

Ω is huge: $NT \times NT$. The previous expression is not convenient to compute Ω^{-1}.
Reformulation: replace J_T by $T\bar{J}_T$, and define $E_T := I_T - \bar{J}_T$. Then,

$$\begin{aligned}
\Omega &= T\sigma_\mu^2(I_N \otimes \bar{J}_T) + \sigma_\nu^2(I_N \otimes E_T) + \sigma_\nu^2(I_N \otimes \bar{J}_T) \\
&= (T\sigma_\mu^2 + \sigma_\nu^2)(I_N \otimes \bar{J}_T) + \sigma_\nu^2(I_N \otimes E_T) \\
&= \sigma_1^2 P + \sigma_\nu^2 Q
\end{aligned}$$

where $\sigma_1^2 := T\sigma_\mu^2 + \sigma_\nu^2$
Easy to check: $\Omega^{-1} = P/\sigma_1^2 + Q/\sigma_\nu^2$, and $\Omega^{-1/2} = P/\sigma_1 + Q/\sigma_\nu$

GLS estimator

Consider the system

$$\begin{pmatrix} QY \\ (P - \bar{J}_{NT})Y \end{pmatrix} = \begin{pmatrix} QX \\ (P - \bar{J}_{NT})X \end{pmatrix} \beta + \begin{pmatrix} Qu \\ (P - \bar{J}_{NT})u \end{pmatrix}$$

where the error term has covariance matrix

$$\begin{pmatrix} \sigma_\nu^2 Q & 0 \\ 0 & \sigma_1^2 (P - \bar{J}_{NT}) \end{pmatrix}$$

Easy to check: Estimation of this system by OLS is equivalent to estimating $Y^* = X^*\beta + u^*$ by OLS, and estimation by GLS is equivalent to estimating $Y^* = X^*\beta + u^*$ by GLS, where Y^* and X^* are the de-meaned Y and X, i.e. $Y^* = (I_{NT} - \bar{J}_{NT})Y$ and $X^* = (I_{NT} - \bar{J}_{NT})X$, and $u^* = (I_{NT} - \bar{J}_{NT})u$ has var-cov matrix $\Omega^* = \sigma_1^2(P - \bar{J}_{NT}) + \sigma_\nu^2 Q$.

Thus,

$$\begin{aligned}\hat{\beta}_{GLS} &= \left[\frac{X'QX}{\sigma_\nu^2} + \frac{X'(P-\bar{J}_{NT})X}{\sigma_1^2}\right]^{-1}\left[\frac{X'QY}{\sigma_\nu^2} + \frac{X'(P-\bar{J}_{NT})Y}{\sigma_1^2}\right] \\ &= [W_{XX} + \phi^2 B_{XX}]^{-1}[W_{XY} + \phi^2 B_{XY}]\end{aligned}$$

where $W_{XX} := X'QX$, $W_{XY} := X'QY$, $B_{XX} = X'(P-\bar{J}_{NT})X$, $B_{XY} = X'(P-\bar{J}_{NT})Y$, and $\phi^2 = \sigma_\nu^2/\sigma_1^2$.

GLS as mixture of Within and Between

We can write

$$\hat{\beta}_{GLS} = G\hat{\beta}_W + (I_K - G)\hat{\beta}_B$$

where

$$G = (W_{XX} + \phi^2 B_{XX})^{-1} W_{XX}$$

The GLS estimator of the random effects model is a matrix-weighted average of the Within and Between estimators.

Special cases

We have three extreme cases:

1. If $\sigma_\mu^2 = 0$, then $\phi^2 = 1$ and $\hat{\beta}_{GLS} = \hat{\beta}_{OLS}$
2. If $T \to \infty$, then $\phi^2 \to 0$ and $\hat{\beta}_{GLS}$ tends to $\hat{\beta}_W$
3. If $\phi^2 \to \infty$, then $\hat{\beta}_{GLS}$ tends to $\hat{\beta}_B$

Note also that

- The Within estimator ignores the between variation, and the Between estimator ignores the within variation.
- The OLS estimator gives equal weight to the between and within estimators.

Efficiency

We have

$$\begin{aligned} \mathrm{Var}(\hat{\beta}_{GLS}) &= \sigma_\nu^2 (W_{XX} + \phi^2 B_{XX})^{-1} \\ \mathrm{Var}(\hat{\beta}_{W}) &= \sigma_\nu^2 W_{XX}^{-1} \end{aligned}$$

so that

$$\mathrm{Var}(\hat{\beta}_{W}) - \mathrm{Var}(\hat{\beta}_{GLS})$$

is positive definite: GLS is more efficient than OLS.
However, if $T \to \infty$, for N fixed, then $\phi \to 0$ and the variances are equal.

Feasible GLS

GLS requires estimation of σ_1^2 and σ_μ^2. Several estimators have been proposed.
One possible estimator, proposed by Amemiya (1971), is

$$\begin{aligned} \hat{\sigma}_1^2 &= \frac{\hat{u}' P \hat{u}}{\text{tr}(P)} \\ \hat{\sigma}_\nu^2 &= \frac{\hat{u}' Q \hat{u}}{\text{tr}(Q)} \end{aligned}$$

where $\hat{u}$ are the residuals from the Within regression.

Specification test of Hausman

In the error components model, a critical assumption is $\mathbb{E}(u_{it}|X_{it}) = 0$ (strict exogeneity). If this does not hold, because e.g. μ_i is correlated with X_{it}, then the fixed effects estimator $\hat{\beta}_W$ is still consistent, but not the GLS estimator.
That is,

- Under strict exogeneity, both fixed and random effects estimators are consistent, and the random effects estimator is efficient.
- If strict exogeneity is violated, only the fixed effects estimator is consistent.

Idea of the Hausman test is to compute $\hat{q} := \hat{\beta}_{GLS} - \hat{\beta}_W$. If $\hat{q}$ is 'large', then strict exogeneity is rejected.

Recall that

$$\hat{\beta}_{GLS} - \beta = (X'\Omega^{-1}X)^{-1}X'\Omega^{-1}u$$

and

$$\hat{\beta}_{W} - \beta = (X'QX)^{-1}X'Qu$$

It is easy to see that $\mathbb{E}(\hat{q}) = 0$, and

$$\begin{aligned}
\mathrm{Cov}(\hat{\beta}_{GLS}, \hat{q}) &= \mathrm{Var}(\hat{\beta}_{GLS}) - \mathrm{Cov}(\hat{\beta}_{GLS}, \hat{\beta}_{W}) \\
&= (X'\Omega^{-1}X)^{-1} - (X'\Omega^{-1}X)^{-1}X'\Omega^{-1}\underbrace{\mathbb{E}(uu')}_{\Omega}QX(X'QX)^{-1} \\
&= (X'\Omega^{-1}X)^{-1} - (X'\Omega^{-1}X)^{-1}X'QX(X'QX)^{-1} \\
&= 0
\end{aligned}$$

Therefore,

$$\text{Var}(\hat{\beta}_W) = \text{Var}(\hat{\beta}_{GLS} - \hat{q}) = \text{Var}(\hat{\beta}_{GLS}) + \text{Var}(\hat{q})$$

and

$$\text{Var}(\hat{q}) = \text{Var}(\hat{\beta}_W) - \text{Var}(\hat{\beta}_{GLS}) = \sigma_\nu^2 (X'QX)^{-1} - (X'\Omega^{-1}X)^{-1}$$

Then, the Hausman test statistic is given by

$$m = \hat{q}' \text{Var}(\hat{q})^{-1} \hat{q}.$$

Under H_0, m has an asymptotic χ^2 with K degrees of freedom, where K is the number of parameters in $\hat{\beta}$.
In practice, replace Ω by a feasible estimator (FGLS).

Empirical Example

Baltagi and Griffin (1983) consider the following gasoline demand equation:

$$\log \frac{Gas}{Car} = \beta_0 + \beta_1 \log \frac{Y}{N} + \beta_2 \log \frac{P_{MG}}{P_{GDP}} + \beta_3 \log \frac{Car}{N} + u$$

where

- Gas/Car: gasoline consumption per car
- Y/N: real income per capita
- P_{MG}/P_{GDP}: real gasoline price
- Car/N: stock or cars per capita

Data are for 18 OECD countries, annual data from 1960-78.

Results

	FE (Within)	Between	GLS
Y/N	0.66	0.96	0.60
P_{MG}/P_{GDP}	-0.32	-0.96	-0.36
Car/N	-0.64	-0.79	-0.62
Constant	2.40	2.54	1.99

The Hausman test rejects the null ($m = 306.1$), so that the FGLS estimator is inconsistent. Before accepting the FE Within estimator, other tests should be performed (e.g. that of Chamberlain, 1982).

Extensions

- Versions of the Hausman test that are robust to heteroskedasticity and/or autocorrelation (Arellano, 1993)
- Other specification tests (Chow, Breusch-Pagan, etc).
- Dynamic panel models: include a lagged endogenous variable as regressor.

Part XII

Summary and conclusions

Summary of learning outcomes

You should now be able to

- interpret parameters of linear and log-linear models
- know what an orthogonal projection is, and what it implies for the residuals of the estimated model
- know what the FWL-theorem says, and how it can help to estimate partitioned regressions
- understand the geometry of the R^2
- show the link between the four distributions: normal, student, chi-square, and Fisher
- apply the concepts of consistency, unbiasedness, and efficiency to estimators of the model
- say where normality of the errors is needed, and where it is not needed.

- know the assumptions of the Gauss-Markov theorem, what it says, and what it does not say
- know the difference between the law of large numbers and the CLT, and how they are used in the theory of least squares
- show the construction of the test of a general set of linear restrictions on the regression coefficients
- show how to obtain individual and simultaneous confidence intervals, and explain the difference between them.
- explain the problem of multicollinearity, and how to find remedies
- explain how to include discrete variables in the model, how to interpret them, and how to test for their significance

- present alternative strategies to select variables as regressors
- explain the impact of heteroskedasticity on the properties of OLS estimators, and what strategies exist to solve the problem
- explain the impact of autocorrelation on the properties of OLS estimators, and what strategies exist to solve the problem
- understand the difference and the relationship between outlying and influential observations, and how to detect them
- explain the difference between a fixed and random effects model, and the properties of the Within and GLS estimators
- compare the properties of OLS and MLE in the linear regression model