# Instrumental Variables

Hartel, W.

# Universiteit Leiden
# Opleiding Wiskunde

Instrumental Variables


Name: Wout Hartel


Date: 08/07/2016

Supervisor: Prof.dr. A.W. Van der Vaart

BACHELOR THESIS


Mathematical Institute (MI)
Leiden University
Niels Bohrweg 1
2333 CA Leiden
The Netherlands

# Contents

# 1 Introduction

A major complication in microeconomics is the possibility of biased parameter estimation due to endogenous variables. One of the solutions to avoid the inconsistency of parameter estimation is the use of instrumental variables (IV). This type of variable provides a way for consistent parameter estimation.

The aim of this thesis is to understand the use of instrumental variables and to prove the consistency of the parameter found by ordinary least squares and two stage least squares.

In this thesis we will focus on the use of instrumental variables for linear models using the least squares method to estimate parameters of the best line of fit. First we will discuss these methods and compare them. thereafter we will explain the different causes of endogeneity. In Section 3, we will have a closer look at the proof of the consistency of the parameter found by ordinary least squares method and two stage least squares method. At the end we will analyze an endogeneity test named Durbin-Wu-Hausman test and a simulation of the explained methods.

# 2 Instrumental variable estimation

In this section we introduce instrumental variable estimation in three subsections: definitions, the ordinary least squares method, the two stage least squares method and finally the concept of endogeneity.

## 2.1 Definition

Instrumental variables estimation is a tool to estimate linear equation's parameters when the ordinary least squares estimator is biased. This concept will be explicitly explained in the next subsections.

The instrument must satisfy the following assumptions: (1) it should be associated with the treatment, (2) it should only affect the outcome trough the treatment (exclusion restriction), and (3) it should not share a common cause with the outcome (independence assumption). The following definition [3] is specific for a linear regression with one variable:

**Definition 2.1.** A variable $Z$ is called instrumental variable for the regressor $X$ in $Y = \alpha + \beta X + \epsilon$ where $E(\epsilon) = 0$ if $Z$ is uncorrelated with the error term $\epsilon$, and $Z$ is correlated with the regressor $X$.

The use of this instrumental variable is when $X$ is correlated with $\epsilon$, so $\text{Cov}(X, \epsilon) \neq 0$. Let $Z$ be a instrumental variable then $\text{Cov}(\epsilon, Z) = 0$ and $\text{Cov}(X, Z) \neq 0$. We will describe this effect later in part 2.4

In the next subsection we explain how the ordinary least squares method works.

## 2.2 Ordinary least squares Method

Instrumental variables estimation uses the ordinary least squares (OLS) method. This method is used to determine the line of best fit for a model. Linear regression is the way to find a line that fits best with a set of data points.

We assume that we have N independent pairs of measurements $\{Y_i, X_i\}$ following the model:

$$Y_i = \alpha + \beta X_i + \epsilon_i \tag{1}$$

Where $\epsilon_1, \epsilon_2, ..., \epsilon_N$ are measurement errors. We wish to find estimators $\hat{\alpha}$ and $\hat{\beta}$ for the parameters $\alpha$ and $\beta$ using the data $\{Y_i, X_i\}$.

The ordinary least squares method estimates these parameters as the values which minimize the sum of the squares of the differences between the model and the data points [7]

$$E = \sum_i (Y_i - \hat{Y}_i)^2 = \sum_i (Y_i - (\alpha + \hat{\beta} X_i))^2 \tag{2}$$

When its partial derivatives reaches zero, then the equation attains its minimum:

$$\frac{\partial E}{\partial \hat{\alpha}} = 2N\hat{\alpha} + 2\hat{\beta} \sum_i X_i - \sum_i Y_i = 0 \tag{3}$$

and

$$\frac{\partial E}{\partial \hat{\beta}} = 2\hat{\beta} \sum_i X_i^2 + 2\hat{\alpha} \sum_i X_i - 2 \sum_i Y_i X_i = 0 \tag{4}$$

Solving these equations gives the least squares estimates of $\alpha$ and $\beta$:

$$\hat{\alpha} = \bar{Y} - \hat{\beta} \bar{X} \tag{5}$$

$$\hat{\beta} = \frac{\sum_i (Y_i - \bar{Y})(X_i - \bar{X})}{\sum_i (X_i - \bar{X})^2} \tag{6}$$

With $\bar{X}$, $\bar{Y}$ the averages of $X_1, X_2, ..., X_N$ and $Y_1, Y_2, ..., Y_N$.

As shown by the following theorem the ordinary least squares gives unbiased estimators when $X_i$ is not correlated with the error term $\epsilon_i$.

**Theorem 2.1.** *Suppose $\{Y_i, X_i\}$ for $i = 1, 2, .., N$ are independent, identically distributed and $X_i$ from a distribution with a positive variance with $Y_i = \alpha + \beta X_i + \epsilon_i$ for $\epsilon_i \sim N(0, \sigma^2)$ and $E(\epsilon_i | X_i) = 0$ for all i. Then $\hat{\beta}_{OLS} \to \beta$ in probability as $N \to \infty$ and $P(\sqrt{N}(\hat{\beta}_{OLS} - \beta) \leq x) \to \Phi(x/\sigma_\beta)$ for all x.*

A crucial assumption of the theorem is that $E(\epsilon_i | X_i) = 0$, or the $\epsilon_i$ is exogenous.If this not the case, we use an instrumental variable with the two stage least squares.

## 2.3 Two stage least squares

The two stage least squares method (2SLS) is used for making a linear regression of a data set. The instrumental variable will be used to estimate the parameter with an endogenous variable (definition in 2.4). In this subsection we explain the method with one variable.

### 2.3.1 Method

Take the same equation as (1):

$$Y_i = \alpha + \beta X_i + \epsilon_i \tag{7}$$

We use the two stage least squares if $\text{Cov}(X, \epsilon) \neq 0$ because of the biased estimator of the OLS method. This method is composed of two stages [8]:

First stage:

Let $Z$ be a instrumental variable, so $\text{Cov}(X, Z) \neq 0$ and $\text{Cov}(\epsilon, Z) = 0$.

Perform ordinary least squares of $X$ on $Z$, i.e. determine $\hat{\gamma_1}$ and $\hat{\gamma_2}$ of the equation $X_i = \gamma_1 + \gamma_2 Z_i + \nu_i$, where $\nu_i$ is the measurement error term, to minimize

$$\sum_i (X_i - \hat{\gamma_1} - \hat{\gamma_2} Z_i)^2 \tag{8}$$

Define:

$$\hat{X_i} = \hat{\gamma_1} + \hat{\gamma_1} Z_i \tag{9}$$

Second stage:

Perform ordinary least squares of $Y_i$ on $\hat{X_i}$ i.e find the $\hat{\alpha}$ and $\hat{\beta}$ to minimize

$$\sum_i (Y_i - \hat{\alpha} - \hat{\beta} \hat{X})^2 \tag{10}$$

We find $\hat{\alpha}_{2SLS}$ and $\hat{\beta}_{2SLS}$.

The consistency of the estimator found with the 2SLS method is stated in the following theorem:

**Theorem 2.2.** *Suppose $\{Y_i, X_i, Z_i\}$ for $i = 1, 2, .., N$ are independent, identically distributed and $X_i$ from a distribution with a positive variance with $Y_i = \alpha + \beta X_i + \epsilon_i$ for $\epsilon_i \sim N(0, \sigma^2)$, $E(\epsilon_i | Z_i) = 0$ and $\text{Cov}(X_i, Z_i) \neq 0$ for all i. Then $\hat{\beta}_{2SLS} \to \beta$ in probability as $N \to \infty$ and $P(\sqrt{N}(\hat{\beta}_{OLS} - \beta) \leq x) \to \Phi(x/\sigma_\beta)$ for all $x$, for some $\sigma_\beta > 0$.*

### 2.3.2 Example

For a better understanding of the two stage least square method we will work out an example [6].

We investigate the score of a student for a course at the university. The score depends on many variables, so to simplify this model we only use one variable: the class attendance ($CA$), for $N$ students.

$$\textbf{Score}_i = \alpha + \beta CA_i + \epsilon_i \tag{11}$$

There exist variables that have influence on the score and the class attendance, like the interest of the student in the course. When a student is interested in the course, he is more likely to attend classes than when he is not interested. So we can assume that the class attendance is correlated with the interest. The factor of interest is processed in the error term $\epsilon$. This means that there exists a correlation between $\epsilon$ and the class attendance $CA$: $\text{Cov}(CA_i, \epsilon_i) \neq 0$. If we had used the OLS method we would had a biased estimator. That is why we are using the 2SLS method.

We have to find a instrumental variable: take the distance (*dist*) between the university and the student's home. This distance is correlated with the class attendance. The further away the student lives from the university, the less likely the student is to attend to the class. But the distance has no correlation with the interest of the student in the course. Therefore, distance is an instrumental variable.

First stage:

Perform ordinary least squares of $CA_i$ on $dist_i$

$$\hat{CA}_i = \hat{\gamma}_0 + \hat{\gamma}_1 dist_i \tag{12}$$

Second stage:

Perform ordinary least squares of $\textbf{Score}_i$ on $\hat{CA}_i$

$$\textbf{Score}_i = \hat{\alpha} + \hat{\beta}\hat{CA}_i \tag{13}$$

So in this case $\hat{\beta}_{2SLS} \to \beta$ in probability as $N \to \infty$ as stated in theorem 2.2

## 2.4 Endogeneity

The use of instrumental variables with the two stage least squares method is due to the correlation of the variable with the error term $\epsilon$. This is called endogeneity. In mathematical terms: $E(\epsilon|X) \neq 0$ or $\text{Cov}(X, \epsilon) \neq 0$.

This effect can be due to the following complications [5]:

1) Omitted variables Bias

This mean that there is a linear dependency between the error and the "independent" variable. That makes the expectation $E(\epsilon|X) \neq 0$.

If we take the example 2.3.2:

$$\textbf{Score}_i = \alpha + \beta CA_i + \epsilon_i \tag{14}$$

We saw that the interest ($int$) is integrated in the error term $\epsilon$. This is an example of an omitted variable bias.

One possible solution for this problem is to add the interest as a variable in the equation:

$$\textbf{Score}_i = \alpha + \beta CA_i + \delta int_i + \epsilon_i \tag{15}$$

Perform afterwards the OLS with an extra variable. This method can be difficult to perform due to the lack of information on the extra variables which we add in the equation. In this example, it is very difficult to measure the interest of a student. That's why we use instrumental variable estimation.

2) Measurement error

This error can lead of a non real correlation between the error $\epsilon_i$ and the independent variable $X_i$.

In the example 2.3.2, there could be a measurement error, when we ask $N$ students how many times they attend the course. A student could overestimate this number.

3) Simultaneous causality

The primary aim of a linear regression is to know how $X$ causes $Y$, but in some cases $Y$ causes $X$, this means there is simultaneous causality. This implies that $\text{Cov}(X, \epsilon) \neq 0$.

In the example 2.3.2, let the score of the course be calculate with $2/3$ with the final exam and $1/3$ with the grades of the homework that they have to hand in every week. If the student gets bad homework grades, it can have an influence on the class attendance. So the score the student gets for the course is causal with the class attendance: simultaneous causality.

These three complications make that $E(\epsilon_i|X_i) \neq 0$ or $\text{Cov}(X_i, \epsilon_i) \neq 0$. When you make a linear regression with the least squares method with an endogenous variable, you get biased/inconsistent parameters. Therefore, these complications are avoided when you use instrumental variable estimation.

# 3 Estimators properties

In the previous section we stated two theorems about the consistency and the distribution of the difference between the estimators and the parameter $\beta$. In this section we prove these theorems (2.1 and 2.2). First the consistency of $\hat{\beta}_{OLS}$ and $\hat{\beta}_{2SLS}$, thereafter the proof of the distribution of $\sqrt{N}(\hat{\beta}_{2SLS} - \beta)$.

## 3.1 Consistency estimator from OLS

In this subsection we prove the first part of the theorem 2.1. The first part of the theorem, stated in the previous section, was as follow:

**Theorem 3.1.** *Suppose $\{Y_i, X_i\}$ for $i = 1, 2, .., N$ are independent, identically distributed and $X_i$ from a distribution with a positive variance with $Y_i = \alpha + \beta X_i + \epsilon_i$ for $\epsilon_i \sim N(0, \sigma^2)$ and $E(\epsilon_i | X_i) = 0$ for all $i$. Then $\hat{\beta}_{OLS} \to \beta$ in probability as $N \to \infty$.*

To prove this theorem we will need the following lemmas:

**Lemma 3.2.** *If $\{Y_i, X_i\}$ for $i = 1, 2, .., N$ are independent and identically distributed with $Y_i = \alpha + \beta X_i + \epsilon_i$ for $\epsilon_i \sim N(0, \sigma^2)$ and $E(\epsilon_i | X_i) = 0$ for all $i$ then $E(\hat{\beta}_{OLS} | X_1, X_2, \ldots, X_N) = \beta$.*

*Proof.* We saw that $\hat{\beta}_{OLS} = \frac{\sum_i (Y_i - \bar{Y})(X_i - \bar{X})}{\sum_i (X_i - \bar{X})^2} = \frac{N \sum_i X_i Y_i - (\sum_i X_i)(\sum_i Y_i)}{N \sum_i X_i^2 - (\sum_i X_i)^2}$.

We know that $Y_i = \alpha + \beta X_i + \epsilon_i$ so $E(Y_i | X_1, X_2, \ldots, X_N) = \alpha + \beta X_i$.

This means that:

$$
\begin{aligned}
& E(\hat{\beta}_{OLS} | X_1, X_2, \ldots, X_N) \\
= & \frac{N \sum_i X_i E(Y_i | X_1, X_2, \ldots, X_N) - (\sum_i X_i)(\sum_i E(Y_i | X_1, X_2, \ldots, X_N))}{N \sum_i X_i^2 - (\sum_i X_i)^2} \\
= & \frac{N \sum_i X_i (\alpha + \beta X_i) - (\sum_i X_i)(\sum_i (\alpha + \beta X_i))}{N \sum_i X_i^2 - (\sum_i X_i)^2} \\
= & \frac{N\alpha \sum_i X_i + \beta \sum_i X_i^2 - N\alpha \sum_i X_i + \beta (\sum_i X_i)^2}{N \sum_i X_i^2 - (\sum_i X_i)^2} \\
= & \frac{\beta (N \sum_i X_i^2 - (\sum_i X_i)^2)}{N \sum_i X_i^2 - (\sum_i X_i)^2} \\
= & \beta
\end{aligned}
$$

□

**Lemma 3.3.** *If $\{Y_i, X_i\}$ for $i = 1, 2, .., N$ are independent and identically distributed with $Y_i = \alpha + \beta X_i + \epsilon_i$ for $\epsilon_i \sim N(0, \sigma^2)$, $E(\epsilon_i|X_i) = 0$ and $\text{Var}(\epsilon_i|X_i) = \sigma^2$ for all $i$, then $\text{Var}(\hat{\beta}_{OLS}|X_1, X_2, \ldots, X_N) = \frac{\sigma^2}{\sum_i (X_i - \bar{X})^2}$.*

*Proof.* It holds that $\hat{\beta}_{OLS} = \frac{\sum_i (X_i - \bar{X}) Y_i}{\sum_i (X_i - \bar{X})^2} = \frac{\sum_i (X_i - \bar{X})(\alpha + \beta X_i)}{\sum_i (X_i - \bar{X})^2} + \frac{\sum_i (X_i - \bar{X}) \epsilon_i}{\sum_i (X_i - \bar{X})^2}$

Then:

$$\begin{aligned}\text{Var}(\hat{\beta}_{OLS}|X_1, X_2, \ldots, X_N) &= \frac{\sum_i (X_i - \bar{X})^2 \text{Var}(\epsilon_i|X_1, X_2, \ldots, X_N)}{(\sum_i (X_i - \bar{X})^2)^2} \\ &= \frac{\sigma^2}{\sum_i (X_i - \bar{X})^2}\end{aligned}$$

□

**Proposition 3.4.** *If $X_i$ for $i = 1, 2, .., N$ are independent and identically distributed from a distribution with positive variance, then $\lim_{N\to\infty} \text{Var}(\hat{\beta}_{OLS}|X_1, X_2, \ldots, X_N) \overset{\text{a.s}}{=} 0$*

*Proof.*

$$\begin{aligned}\frac{1}{N} \sum_i (X_i - \bar{X})^2 &= \frac{1}{N} \sum_i (X_i^2 - 2 X_i \bar{X} + \bar{X}) \\ &= \frac{\sum_i X_i^2}{N} - (\bar{X})^2\end{aligned}$$

Following the law of strong numbers we know that

$$\lim_{N\to\infty} \frac{\sum_i X_i^2}{N} \overset{\text{a.s}}{=} E(X_1^2) \tag{16}$$

$$\lim_{N\to\infty} \bar{X} \overset{\text{a.s}}{=} E(X_1) \tag{17}$$

so $\lim_{N\to\infty} \frac{1}{N} \sum_i (X_i - \bar{X})^2 = E(X_1^2) - E(X_1)^2 = \text{Var}(X_1)$ =constant.

So if $\lim_{N\to\infty} \frac{1}{N} \sum_i (X_i - \bar{X})^2$ =constant then $\lim_{N\to\infty} \sum_i (X_i - \bar{X})^2 = \infty$ and $\lim_{N\to\infty} \text{Var}(\hat{\beta}_{OLS}) = \lim_{N\to\infty} \frac{\sigma^2}{\sum_i (X_i - \bar{X})^2} = 0$ □

Now we have enough knowledge to do the proof of the first part of the theorem 2.1:

$$\hat{\beta}_{OLS} \to \beta \text{ in probability as } N \to \infty$$

*Proof.* The above statement is the same as proving that $P(|\beta_{OLS} - \beta| \geq \epsilon) \to 0$ when $N \to \infty$, for any $\epsilon > 0$ .

By Chebyshev's inequality :

$$P(|\hat{\beta}_{OLS} - E(\hat{\beta}_{OLS}|X_1, X_2, \ldots, X_N)| \geq \epsilon | X_1, X_2, \ldots, X_N) \leq \frac{\text{Var}(\hat{\beta}_{OLS}|X_1, X_2, \ldots, X_N)}{\epsilon^2}.$$

By lemmas 3.2, 3.3 and proposition 3.4 we state that $P(|\hat{\beta}_{OLS} - \beta| \geq \epsilon | X_1, X_2, \ldots, X_N) \to 0$ a.s, when $N \to \infty$ with $\epsilon > 0$. This implies that $P(|\hat{\beta}_{OLS} - \beta| \geq \epsilon) \to 0$, by the law of iterated expectation.

□

## 3.2 Consistency estimator from 2SLS

In this subsection we prove the first part of the theorem 2.2:

**Theorem 3.5.** *Suppose $\{Y_i, X_i, Z_i\}$ for $i = 1, 2, .., N$ are independent and identically distributed with $Y_i = \alpha + \beta X_i + \epsilon_i$ for $\epsilon_i \sim N(0, \sigma^2)$, $E(\epsilon_i|Z_i) = 0$ and $E(\epsilon_i|X_i) \neq 0$ for all $i$. Then $\hat{\beta}_{2SLS} \to \beta$ in probability as $N \to \infty$.*

First we prove the following equation:

$$\hat{\beta}_{2SLS} = \frac{\sum_i (\hat{X}_i - \bar{\hat{X}}) Y_i}{\sum_i (\hat{X}_i - \bar{\hat{X}})^2} = \frac{\sum_i (Z_i - \bar{Z}) Y_i}{\sum_i (Z_i - \bar{Z}) \hat{X}_i}. \tag{18}$$

Applying (5) and (6) with $(X_i, Z_i)$ substituted for $(Y_i, X_i)$ gives:

$$\hat{X}_i = \hat{\gamma}_0 + \hat{\gamma}_1 Z_i \tag{19}$$
$$\bar{\hat{X}} = \hat{\gamma}_0 + \hat{\gamma}_1 \bar{Z} \tag{20}$$

Using (6) for $\gamma_1$

$$\hat{\gamma}_1 = \frac{\sum_i (Z_i - \bar{Z})(X_i - \bar{X})}{\sum_i (Z_i - \bar{Z})^2} = \frac{\sum_i (Z_i - \bar{Z}) X_i}{\sum_i (Z_i - \bar{Z})^2} \tag{21}$$

Using (19) and (20) gives:

$$\hat{X}_i - \bar{\hat{X}} = \hat{\gamma}_1 (Z_i - \bar{Z}) \tag{22}$$

By (6) applied to $(Y_i, \hat{X}_i)$ instead of $(Y_i, X_i)$ we find:

$$
\begin{aligned}
\hat{\beta}_{2SLS} &= \frac{\sum_i (\hat{X}_i - \bar{\hat{X}}) Y_i}{\sum_i (\hat{X}_i - \bar{\hat{X}})^2} \\
&= \frac{\sum_i \hat{\gamma}_1 (Z_i - \bar{Z}) Y_i}{\sum_i \hat{\gamma}_1^2 (Z_i - \bar{Z})^2} \\
&= \frac{1}{\hat{\gamma}_1} \frac{\sum_i (Z_i - \bar{Z}) Y_i}{\sum_i (Z_i - \bar{Z})^2} \\
&= \frac{\sum_i (Z_i - \bar{Z})^2}{\sum_i (Z_i - \bar{Z}) X_i} \frac{\sum_i (Z_i - \bar{Z}) Y_i}{\sum_i (Z_i - \bar{Z})^2} \\
&= \frac{\sum_i (Z_i - \bar{Z}) Y_i}{\sum_i (Z_i - \bar{Z}) X_i}.
\end{aligned} \tag{23}
$$

We can $\hat{\beta}_{2SLS}$ write as:

$$
\begin{aligned}
\hat{\beta}_{2SLS} &= \frac{\frac{1}{N} \sum_i (Z_i - \bar{Z}) Y_i}{\frac{1}{N} \sum_i (Z_i - \bar{Z}) X_i} \\
&= \frac{\frac{1}{N} \sum_i (Z_i - \bar{Z}) (\alpha + \beta X_i)}{\frac{1}{N} \sum_i (Z_i - \bar{Z}) X_i} + \frac{\frac{1}{N} \sum_i (Z_i - \bar{Z}) \epsilon_i}{\frac{1}{N} \sum_i (Z_i - \bar{Z}) X_i} \\
&= \beta + \frac{\frac{1}{N} \sum_i (Z_i - \bar{Z}) \epsilon_i}{\frac{1}{N} \sum_i (Z_i - \bar{Z}) X_i}
\end{aligned} \tag{24}
$$

**Proposition 3.6.** *If $\{Y_i, X_i, Z_i\}$ for $i = 1, 2, .., N$ are independent and identically distributed then $lim_{N \to \infty} \frac{\frac{1}{N} \sum_i (Z_i - \bar{Z}) \epsilon_i}{\frac{1}{N} \sum_i (Z_i - \bar{Z}) \hat{X}_i} = 0$ in probability.*

First we state two lemmas that we are using in the proof of the previous proposition 3.6.

**Lemma 3.7.** *If $lim_{N \to \infty} P_N = C$ and $lim_{N \to \infty} Q_N = D$ in probability with $C, D$ constant and $D \neq 0$, then $lim_{N \to \infty} \frac{P_N}{Q_N} = \frac{C}{D}$ in probability.*

**Lemma 3.8.** *If $lim_{N \to \infty} P_N = C$ almost surely (a.s) then $lim_{N \to \infty} P_N = C$ in probability .*

The proofs of these two lemmas are not interesting for this thesis, therefore they are not explained here. You can find them in the references ([7], [4]).

Now we prove proposition 3.6:

*Proof.* Let $T_N = \frac{1}{N} \sum_i (Z_i - \bar{Z}) \epsilon_i$ and $U_N = \frac{1}{N} \sum_i (Z_i - \bar{Z}) X_i$

$$E(T_N|Z_1, Z_2, \ldots, Z_N) = \frac{1}{N}\sum_i (Z_i - \bar{Z})E(\epsilon_i|Z_1, Z_2, \ldots, Z_N) = 0 \quad (25)$$

$$\mathrm{Var}(T_N|Z_1, Z_2, \ldots, Z_N) = \frac{1}{N^2}\sum_i (Z_i - \bar{Z})^2\sigma^2 \quad (26)$$

With $\sigma^2 = \mathrm{Var}(\epsilon_i|Z_1, Z_2, \ldots, Z_N)$

By the law of large numbers $lim_{N\to\infty}\frac{1}{N}\sum_i (Z_i - \bar{Z})^2 \overset{\text{a.s}}{=} \mathrm{Var}(Z)$ .

This means that

$$lim_{N\to\infty}\mathrm{Var}(T_N|Z_1, Z_2, \ldots, Z_N) = lim_{N\to\infty}\frac{1}{N}\sigma^2(\frac{1}{N}\sum_i (Z_i - \bar{Z})^2) \overset{\text{a.s}}{=} 0 \quad (27)$$

Chebyshev inequality states:

$$P(|T_N - E(T_N|Z_1, Z_2, \ldots, Z_N)| \geq t|Z_1, Z_2, \ldots, Z_N) \leq \frac{\mathrm{Var}(T_n|Z_1, Z_2, \ldots, Z_N)}{t^2}$$

when $N \to \infty$ with $t > 0$.

This means that $P(|T_N - 0| \geq t) = 0$ when $N \to \infty$ with $t > 0$

So $lim_{N\to\infty}T_N = 0$ in probability.

We still have to prove that $U_N$ is a constant when $N$ goes to infinity.

$$\begin{aligned} U_N &= \frac{1}{N}\sum_i (Z_i - \bar{Z})X_i \\ &= \frac{1}{N}\sum_i Z_i X_i - \bar{Z}\frac{1}{N}\sum_i X_i \\ &= \frac{1}{N}\sum_i Z_i X_i - \bar{Z}\bar{X}_i \end{aligned} \quad (28)$$

By the law of large numbers

$$\begin{aligned} \lim_{N\to\infty} U_N &= \lim_{N\to\infty}\frac{1}{N}\sum_i Z_i X_i - \bar{Z}\bar{X}_i \\ &\overset{\text{a.s}}{=} E(Z_1X_1) - E(Z_1)E(X_1) \\ &= \mathrm{Cov}(Z_1, X_1) \end{aligned} \quad (29)$$

We choose $Z_i$ with a non-zero covariance with $X_i$, so $U_N$ is a non-zero constant.

Using Lemmas 3.8 and 3.7, we know that $\lim_{N\to\infty} U_N = \text{Cov}(Z_1, X_1)$ in probability. We can conclude that in probability:

$$\lim_{N\to\infty} \frac{T_N}{U_N} = \lim_{N\to\infty} \frac{\frac{1}{N}\sum_i (Z_i - \bar{Z})\epsilon_i}{\frac{1}{N}\sum_i (Z_i - \bar{Z})X_i} = 0 \tag{30}$$

□

Using proposition 3.6 and equation (24) we proved that $\hat{\beta}_{2SLS} \to \beta$ in probability as $N \to \infty$

We proved the first part of the Theorem 2.2

## 3.3 Distribution of $\sqrt{N}(\hat{\beta}_{2SLS} - \beta)$

To complete the proofs of theorems 2.1 and 2.2, we have to prove the second part. We will only do the proof of the theorem 2.2, because it is the most interesting for this thesis. The proof for theorem 2.1 follows the same steps as the following proof.

The second part of the theorem states:

Suppose $\{Y_i, X_i, Z_i\}$ for $i = 1, 2, .., N$ are independent, identically distributed and $X_i$ from a distribution with a positive variance with $Y_i = \alpha + \beta X_i + \epsilon_i$ for $\epsilon_i \sim N(0, \sigma^2)$, $E(\epsilon_i|Z_i) = 0$ and $E(\epsilon_i|X_i) \neq 0$ for all $i$. Then $P(\sqrt{N}(\hat{\beta}_{OLS} - \beta) \leq x) \to \Phi(x/\sigma_\beta)$ for all $x$, for some $\sigma_\beta > 0$.

Before proving this theorem we will state the Lindeberg-Feller Central limit theorem and a lemma that we will use [1]

**Theorem 3.9.** *(***Lindeberg-Feller Central limit theorem***) For every $N$, let $X_{N1}, X_{N2}, \ldots, X_{NN}$ be i.i.d with $E(X_{Ni}) = 0$ and*

$$\frac{1}{N}\sum_i E(X_{Ni}^2) \quad \to \quad \tau^2 \tag{31}$$

$$\frac{1}{N}\sum_i E(X_{Ni}^2 \mathbb{1}_{\{|X_{Ni}|>\mu\sqrt{N}\}}) \quad \to \quad 0 \tag{32}$$

*for all $\mu > 0$, when $N \to \infty$,*

*Then*

$$\frac{1}{\sqrt{N}}\sum_i X_{Ni} \sim \mathbb{N}(0, \tau^2)$$

*when $N \to \infty$.*

**Lemma 3.10.** *Let $Z_1, Z_2, \ldots, Z_N$ be i.i.d and $E(Z_1^2) < \infty$, then*

$$\frac{1}{\sqrt{N}} \max_{1\le i\le N} |Z_i| \to 0$$

*Almost surely when $N \to \infty$.*

*Proof.* For fixed $M$ define $Y_i$ to be 0 if $Z_i^2 \le M$ and to be $Z_i^2$ otherwise. Then $E(Y_i) = E(Z_i^2 1_{Z_i^2>M})$, which can be made smaller than any given $\epsilon > 0$ by choosing a sufficiently large $M$, since $E(Z_i^2) < \infty$.

Now $\max_{1\le i\le N} Z_i^2 \le M + \max_{1\le i\le N} Y_i$ and hence

$$\frac{1}{N} \max_{1\le i\le N} Z_i^2 \le \frac{M}{N} + \frac{1}{N}\sum_{i=1}^{N} Y_i.$$

For fixed $M$ the first term on the right tends to zero as $N \to \infty$, and the second tends almost surely to $E(Y_i)$, by the strong law of large numbers. We conclude that the left side is bounded by any $\epsilon > 0$ as $N \to \infty$, almost surely. Hence the left side converges to zero almost surely. So does its root.

□

Now we have enough tools to prove the second part of theorem 2.2.

*Proof.* First we are going to look at the difference $\hat{\beta}_{2sls} - \beta$.

From equation (24) we know that:

$$\begin{aligned}\sqrt{N}(\hat{\beta}_{2sls} - \beta) &= \sqrt{N}(\beta + \frac{\sum_i (Z_i - \bar{Z})\epsilon_i}{\sum_i (Z_i - \bar{Z})X_i} - \beta) \\ &= \frac{\frac{1}{\sqrt{N}}\sum_i (Z_i - \bar{Z})\epsilon_i}{\frac{1}{N}\sum_i (Z_i - \bar{Z})X_i}\end{aligned} \tag{33}$$

By the law of large numbers we know that $\lim_{N\to\infty} \frac{1}{N}\sum_i (Z_i - \bar{Z})X_i \overset{\text{a.s}}{=} \text{Cov}(Z_i, X_i) > 0$. It is enough to prove that $\frac{1}{\sqrt{N}}\sum_i (Z_i - \bar{Z})\epsilon_i$ is asymptotically normally distributed.

For proving this we are using the Lindeberg-Feller Central limit theorem. We are now interested in the first requirement (31) of the theorem.

Let $X_{Ni} = (Z_i - \bar{Z})\epsilon_i$ with $E(\epsilon_i^2|Z_i) = \nu^2$ for $i = 1, 2, \ldots, N$ then:

$$\begin{aligned}\frac{1}{N}\sum_i E(X_{Ni}^2|Z_1, Z_2, \ldots, Z_N) &= \frac{1}{N}\sum_i E((Z_i - \bar{Z})^2\epsilon_i^2|Z_1, Z_2, \ldots, Z_N) \\ &= \frac{1}{N}\sum_i (Z_i - \bar{Z})^2 E(\epsilon_i^2|Z_1, Z_2, \ldots, Z_N) \\ &= \frac{1}{N}\sum_i (Z_i - \bar{Z})^2\nu^2\end{aligned} \tag{34}$$

By law of large numbers we see that

$$\lim_{N\to\infty} \sum_i (Z_i - \bar{Z})^2\nu^2 \overset{\text{a.s}}{=} \text{Var}(Z)\nu^2 = \tau^2 \tag{35}$$

Now we are proving the second requirement (32) of the Lindeberg-Feller central limit theorem.

In the proof we are using that $|(Z_i - \bar{Z})||\epsilon_i| \leq \max_{i\leq j\leq N} |(Z_j - \bar{Z})||\epsilon_j|$.

$$
\begin{aligned}
&\frac{1}{N}\sum_i E(X_{Ni}^2 \mathbb{1}_{\{|X_{Ni}|>\mu\sqrt{N}\}}|Z_1, Z_2, \ldots, Z_N) \\
&= \frac{1}{N}\sum_i E(((Z_i-\bar{Z})\epsilon_i)^2 \mathbb{1}_{\{|(Z_i-\bar{Z})\epsilon_i|>\mu\sqrt{N}\}}|Z_1, Z_2, \ldots, Z_N) \\
&= \frac{1}{N}\sum_i (Z_i-\bar{Z})^2 E(\epsilon_i^2 \mathbb{1}_{\{|(Z_i-\bar{Z})||\epsilon_i|>\mu\sqrt{N}\}}|Z_1, Z_2, \ldots, Z_N) \qquad (36) \\
&\leq \frac{1}{N}\sum_i (Z_i-\bar{Z})^2 E(\epsilon_i^2 \mathbb{1}_{\{\max_{i\leq j\leq N}|(Z_j-\bar{Z})||\epsilon_j|>\mu\sqrt{N}\}}|Z_1, Z_2, \ldots, Z_N) \\
&= \frac{1}{N}\sum_i (Z_i-\bar{Z})^2 E(\epsilon_i^2 \mathbb{1}_{\{|\epsilon_j|>\frac{\mu}{\frac{1}{\sqrt{N}}\max_{i\leq j\leq N}|(Z_j-\bar{Z})|}\}}|Z_1, Z_2, \ldots, Z_N)
\end{aligned}
$$

Using lemma 3.10 and $\max_{i\leq j\leq N}|(Z_j-\bar{Z})| \leq \max_{i\leq j\leq N}|Z_j|$, so:

$$
\frac{1}{\sqrt{N}}\max_{i\leq j\leq N}|(Z_j-\bar{Z})| \to 0 \qquad (37)
$$

almost surely when $N\to\infty$. So $\lim_{N\to N}\frac{\mu}{\frac{1}{\sqrt{N}}\max_{i\leq j\leq N}|(Z_j-\bar{Z})}=\infty$.

In this case, we can conclude that:

$$
\begin{aligned}
\frac{1}{N}\sum_i (Z_i-\bar{Z})^2 E(\epsilon_i^2 \mathbb{1}_{\{|\epsilon_j|>\frac{\mu}{\frac{1}{\sqrt{N}}\max_{i\leq j\leq N}|(Z_j-\bar{Z})|}\}}|Z_1, Z_2, \ldots, Z_N) &\to 0 \qquad (38) \\
\frac{1}{N}\sum_i E(((Z_i-\bar{Z})\epsilon_i)^2 \mathbb{1}_{\{|(Z_i-\bar{Z})\epsilon_i|>\mu\sqrt{N}\}}|Z_1, Z_2, \ldots, Z_N) &\to 0 \qquad (39)
\end{aligned}
$$

The two requirements are proved, so following the Linderberg-Feller central limit theorem, we can say that

$$
\frac{1}{\sqrt{N}}\sum_i (Z_i-\bar{Z})\epsilon_i \xrightarrow{\mathrm{d}} \mathbb{N}(0,\tau^2) \qquad (40)
$$

when $N\to\infty$.

To conclude, using (33) we proved that $\sqrt{N}(\hat{\beta}_{2SLS}-\beta)$ is normally distributed when $N\to\infty$. □

The variance of $\sqrt{N}(\hat{\beta}_{2SLS}-\beta)$ when $N\to\infty$ is

$$
\mathrm{Var}(\sqrt{N}(\hat{\beta}_{2SLS}-\beta)) = \frac{\mathrm{Var}(Z)\mathrm{Var}(\epsilon|Z_1, Z_2, \ldots, Z_N)}{\mathrm{Cov}(X_1, Z_1)^2} \qquad (41)
$$

When $N\to\infty$ using equations (40).

We also know that the $\sqrt{N}(\hat{\beta}_{OLS}-\beta)$ is normally distributed with variance:

$$
\mathrm{Var}(\sqrt{N}(\hat{\beta}_{OLS}-\beta)) = \frac{\mathrm{Var}(\epsilon|X_1, X_2, \ldots, X_N)}{\mathrm{Var}(X_1)} \qquad (42)
$$

## 3.4 Variance of OLS and 2SLS estimators

We have seen the ordinary least squares method (OLS) and the two stage least squares method (2SLS). In the case that $X_i$ for all $i$ is not endogenous, we can choose between the OLS method and the 2SLS method. In this part we are proving that the method to find the best estimator of $\beta$ in this case is the OLS method.

The parameter with the lowest variance of the difference $\sqrt{N}(\hat{\beta} - \beta)$ is the most precise estimator, so the best estimator.

In the previous subsection we have seen that:

$$\text{Var}(\sqrt{N}(\hat{\beta}_{OLS} - \beta)) \sim \frac{\text{Var}(\epsilon_i | X_1, X_2, \ldots, X_N)}{\text{Var}(X_1)} \tag{43}$$

$$\text{Var}(\sqrt{N}(\hat{\beta}_{2SLS} - \beta)) \sim \frac{\text{Var}(Z)\text{Var}(\epsilon_i | Z_1, Z_2, \ldots, Z_N)}{\text{Cov}(X_1, Z_1)^2} \tag{44}$$

when $N \to \infty$.

In this case we assume that $X_i$ is independent of the error term $\epsilon_i$, and that from the definition of the instrumental variable, $Z_i$ is independent of $\epsilon_i$. So $\text{Var}(\epsilon_i | X_1, X_2, \ldots, X_N) = \text{Var}(\epsilon_1 | Z_1, Z_2, \ldots, Z_N) = \text{Var}(\epsilon_i)$

The Cauchy-Schwarz inequality states:

$$(\sum_i a_i b_i)^2 \leq \sum_i a_i^2 \sum_i b_i^2 \tag{45}$$

By the Cauchy-Schwarz inequality [9]:

$$(\sum_i (Z_i - \bar{Z})(X_i - \bar{X}))^2 \leq \sum_i (Z_i - \bar{Z})^2 \sum_i (X_i - \bar{X})^2 \tag{46}$$

$$\frac{1}{\sum_i (X_i - \bar{X})^2} \leq \frac{\sum_i (Z_i - \bar{Z})^2}{(\sum_i (Z_i - \bar{Z})(X_i - \bar{X}))^2} \tag{47}$$

$$\frac{1}{\frac{1}{N}\sum_i (X_i - \bar{X})^2} \leq \frac{\frac{1}{N}\sum_i (Z_i - \bar{Z})^2}{(\frac{1}{N}\sum_i (Z_i - \bar{Z})(X_i - \bar{X}))^2} \tag{48}$$

Now using the law of large numbers and that $\{X_i, Z_i, Y_i\}$ are i.i.d distributed, we know that the limit of the equation (48) when $N \to \infty$ is :

$$\frac{1}{\text{Var}(X_1)} \leq \frac{\text{Var}(Z_1)}{\text{Cov}(X_1, Z_1)^2} \tag{49}$$

In other words when $N \to \infty$ by equations (43) and (44):

$$\frac{\text{Var}(\epsilon_1)}{\text{Var}(X_1)} \leq \frac{\text{Var}(Z_1)\text{Var}(\epsilon_1)}{\text{Cov}(X_1, Z_1)^2} \tag{50}$$

$$\text{Var}(\sqrt{N}(\hat{\beta}_{OLS} - \beta)) \leq \text{Var}(\sqrt{N}(\hat{\beta}_{2SLS} - \beta)) \tag{51}$$

Hence the variance of $\sqrt{N}(\hat{\beta}_{OLS} - \beta)$ is the lowest. So when the $X_i$ is independent of $\epsilon_i$, the best method to use is the least squares method.

To conclude the estimators found by OLS and 2SLS are consistent and the method that is the most accurate when the variable is not endogenous is the OLS method.

# 4 Testing for endogeneity and simulation

In this section we are interested in a test that tells us if the variable $X_i$ is endogenous. The Durbin-Wu-Hausman test is testing for endogeneity with the difference between the parameter $\hat{\beta}_{OLS}$ that you found by ordinary least squares and the parameter $\hat{\beta}_{2SLS}$ that you found by two stage least squares [2]. The test is looking at the standardized distribution of $\hat{\beta}_{OLS} - \hat{\beta}_{2SLS}$ using the value of the standard deviation found with the data points.

## 4.1 Explanation test

The test is using the null hypothesis with significance of 5%

$$\begin{aligned} H_0 &: X_i \text{ is independent of } \epsilon_i \\ H_1 &: X_i \text{ is endogenous} \end{aligned}$$

As told in the section introduction, this test is looking at the distribution of $\sqrt{N}(\hat{\beta}_{OLS} - \hat{\beta}_{2SLS})$. We have seen that $\sqrt{N}(\hat{\beta}_{OLS} - \beta)$ and $\sqrt{N}(\hat{\beta}_{2SLS} - \beta)$ are normally distributed .

We assume that $X_i$ is independent of $\epsilon_i$. So we can assume that $\hat{\beta}_{OLS}$ is not unbiased.

Using the two methods explained in section 2 we found:

$$\hat{\beta}_{OLS} = \frac{\sum_i^N (X_i - \bar{X})Y_i}{\sum_i^N (X_i - \bar{X})^2} \tag{52}$$

$$\hat{\beta}_{2SLS} = \frac{\sum_i^N (Z_i - \bar{Z})Y_i}{\sum_i^N (Z_i - \bar{Z})X_i} \tag{53}$$

$$\hat{\beta}_{OLS} - \hat{\beta}_{2SLS} = \sum_i^N \left( \frac{(X_i - \bar{X})}{\sum_i^N (X_i - \bar{X})^2} - \frac{(Z_i - \bar{Z})}{\sum_i^N (Z_i - \bar{Z})X_i} \right) Y_i \tag{54}$$

By similar arguments as before we can show taht $\hat{\beta}_{OLS} - \hat{\beta}_{2SLS} \sim N(0, \delta^2)$. Let $\hat{\delta}$ be the standard deviation found with the data set.

The zero mean in the limit distribution arises because both estimators are consistent for $\beta$ under $H_0$. On the other hand if $H_0$ is false, then $\hat{\beta}_{2SLS}$ is still consistent for $\beta$, but $\hat{\beta}_{OLS}$ has a different limit. In this case the distribution of $\hat{\beta}_{OLS} - \hat{\beta}_{2SLS}$ will not be centered at 0.

Reject $H_0$ if:

$$T = \left| \frac{\hat{\beta}_{OLS} - \hat{\beta}_{2SLS}}{\hat{\delta}} \right| > 1,96 \tag{55}$$

If $T > 1,96$ means that $P(T) < 0,05$ following the standard normal distribution. This is significant low (5%), so we reject $H_0$.

We will now calculate the standard deviation $\hat{\delta}$ found with the data.

$$\begin{aligned} &\text{Var}(\hat{\beta}_{OLS} - \hat{\beta}_{2SLS} | X_1, \ldots, X_N, Z_1, \ldots, Z_N) \\ &= \sum_i^N \left[ \frac{(X_i - \bar{X})}{\sum_i^N (X_i - \bar{X})^2} - \frac{(Z_i - \bar{Z})}{\sum_i^N (Z_i - \bar{Z}) X_i} \right]^2 \text{Var}(Y_i | X_1, \ldots, X_N, Z_1, \ldots, Z_N) \end{aligned} \tag{56}$$

with:

$$\begin{aligned} \text{Var}(Y_i | X_1, \ldots, X_N, Z_1, \ldots, Z_N) &= \text{Var}(\alpha + \beta X_i + \epsilon_i | X_1, \ldots, X_N, Z_1, \ldots, Z_N) \\ &= \text{Var}(\epsilon_i | X_1, \ldots, X_N, Z_1, \ldots, Z_N) \\ &= \text{Var}(\epsilon_i) \\ &= \sigma^2 \text{ Seen in theorem 2.1} \end{aligned} \tag{57}$$

We know that the approximation of $\sigma^2$ by filling in the equation values estimated from data: $\hat{\sigma}^2 = \frac{1}{N} \sum_i^N (Y_i - \hat{\alpha}_{OLS} - \hat{\beta}_{OLS} X_i)^2$.

We use the OLS method because we are under hypothesis $H_0$. We proved in subsection 3.4 that this method have a better approximation of $\beta$ when $X_i$ is not endogenous.

That means that:

$$\hat{\delta} = \sqrt{\sum_i^N \left[ \frac{(X_i - \bar{X})}{\sum_i^N (X_i - \bar{X})^2} - \frac{(Z_i - \bar{Z})}{\sum_i^N (Z_i - \bar{Z}) X_i} \right]^2 \hat{\sigma^2}} \tag{58}$$

# 5 simulation

In this section we test with a simulation the OLS and 2SLS methods and the Durbin-Wu-Hausman test for endogeneity. Moreover we simulate the these methods when the variance and the covariance of the instrument is changing.

## 5.1 OLS en 2SLS Method

The test is used on an equation with, as variable, a random generation of the normal distribution $X1$ with mean equal to 0 and standard deviation equal to 1.

$$Y = \alpha + \beta X1 + \epsilon \tag{59}$$

With $\epsilon$ a vector filled with random generation of the normal distribution with mean equal to 0 and standard deviation equal to $\frac{1}{2}$, $\alpha = 1$ and $\beta = 0.5$

First we look at the case that $X1$ is not endogenous with $\epsilon$.

For the ordinary least squares method $\hat{\alpha}$ and $\hat{\beta}$ are estimated with the R-function:

```
lm(y~x1)
```

For the two stage least squares method, we estimate $X1$ with a instrumental variable $Z$ that we defined as:

```
z=x1-rnorm(n,0,0.3)
```

$\hat{X1}$ ($X1$hat) is estimated with the function :

```
lm(x1~z)
```

Then we use this result to estimate $\hat{\alpha}$ and $\hat{\beta}$ with th R-function:

```
lm(y~x1hat)
```

For $X1$ endogenous we have to change the R-code. The error term must be dependent of $X1$ so we include the code:

```
eps=rnorm(n,0,0.50)
epstilde=x1+eps
```

Take 'eps' as $\epsilon$ and 'epstilde' as $\tilde{\epsilon}$ In this case we have to change the instrumental variable because $Z$ must be correlated with $X1$ but not with $\tilde{\epsilon}$.

$$\begin{aligned} 0 &= \text{Cov}(Z, \tilde{\epsilon}) \\ \text{Cov}(Z, \tilde{\epsilon}) &= \text{Cov}(Z, X1) + \text{Cov}(Z, \epsilon) \\ \text{Cov}(Z, X1) &= -\text{Cov}(Z, \epsilon) \end{aligned} \tag{60}$$

We are looking for $a, b, c$ so that $Z = aX1 + b\tilde{\epsilon} + c\epsilon$

$$\begin{aligned} \text{Cov}(Z, X1) &= \text{Cov}(aX1 + b\tilde{\epsilon} + c\epsilon, X1) \\ &= a\text{Var}(X1) + b\text{Cov}(X1 + \epsilon, X1) + c\text{Cov}(\epsilon, X1) \\ &= a + b\text{Var}(X1) + b\text{Cov}(\epsilon, X1) + c\text{Cov}(\epsilon, X1) \\ &= a + b \end{aligned} \tag{61}$$

$$\begin{aligned} \text{Cov}(Z, \epsilon) &= \text{Cov}(aX1 + b\tilde{\epsilon} + c\epsilon, \epsilon) \\ &= b\text{Cov}(X1 + \epsilon, \epsilon) + c\text{Var}(\epsilon) \\ &= b\text{Var}(\epsilon) + c\text{Var}(\epsilon) \\ &= b\frac{1}{4} + c\frac{1}{4} \end{aligned} \tag{62}$$

Using equation (60):

$$a + b = -b\frac{1}{4} - c\frac{1}{4} \tag{63}$$

We take as solution: $a = 1$, $b = -\frac{4}{5}$ and $c = 0$, because of:

$\text{Cov}(Z, X1) = \text{Cov}(X1 - \frac{4}{5}\tilde{\epsilon}, X1) = \text{Var}(X1) - \frac{4}{5}\text{Var}(X1) = \frac{1}{5} > 0$

So we have

$$Z = X1 - \frac{4}{5}\tilde{\epsilon} \tag{64}$$

To illustrate the difference between the two methods when $X1$ is endogenous and where it is not, we are simulating the methods and plot histograms of $|\hat{\beta}_{OLS} - \hat{\beta}_{2SLS}|$, $|\beta - \hat{\beta}_{OLS}|$ and $|\beta - \hat{\beta}_{2SLS}|$.

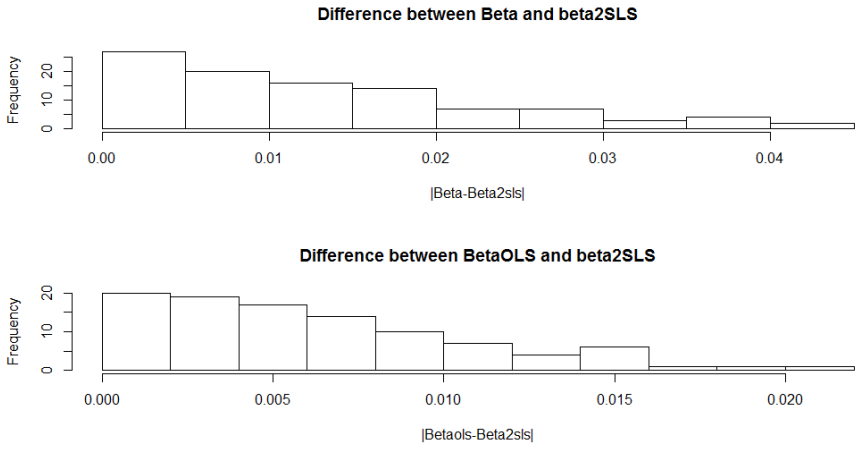



Figure 1: Histograms X1 exogenous

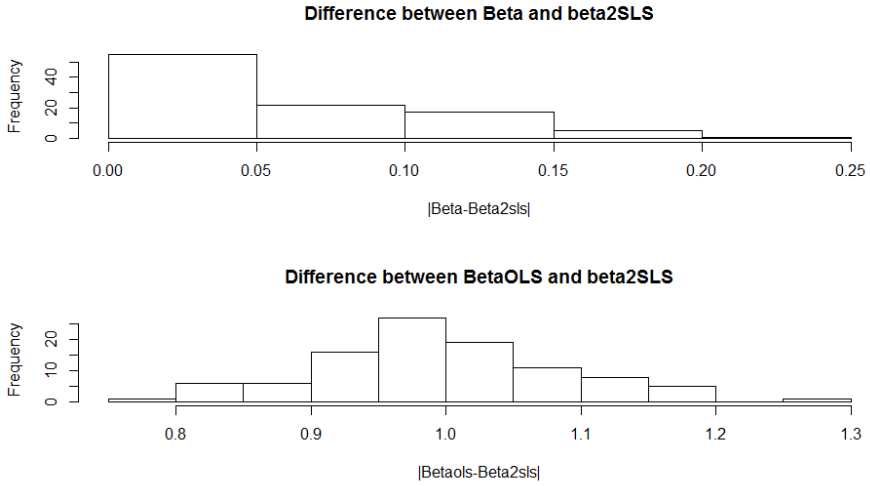



Figure 2: Histograms X1 endogenous

The difference $|\hat{\beta}_{OLS} - \hat{\beta}_{2SLS}|$ is small in the first histogram. But the second histogram this difference is much bigger. The difference $|\beta - \hat{\beta}_{2SLS}|$ stays practically the same for both cases. In the second case the estimator $\hat{\beta}_{OLS}$ estimates badly the parameter $\beta$. This affirm the theory explained in the previous sections: the estimator found with OLS is biased when $X1$ is endogenous.

## 5.2 Durbin-Wu-Hausman test

To simulate the test we calculate the T

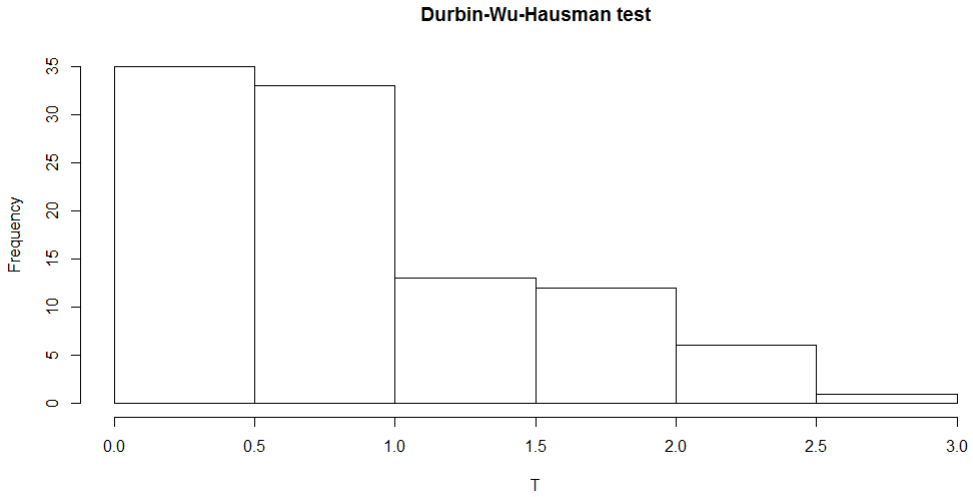
$$T = \left| \frac{\hat{\beta}_{OLS} - \hat{\beta}_{2SLS}}{\hat{\delta}} \right| > 1,96 \tag{65}$$

with

$$\hat{\delta} = \sqrt{\sum_{i}^{N} \left[ \frac{(X_i - \bar{X})}{\sum_i^N (X_i - \bar{X})^2} - \frac{(Z_i - \bar{Z})}{\sum_i^N (Z_i - \bar{Z}) X_i} \right]^2 \hat{\sigma}^2} \tag{66}$$

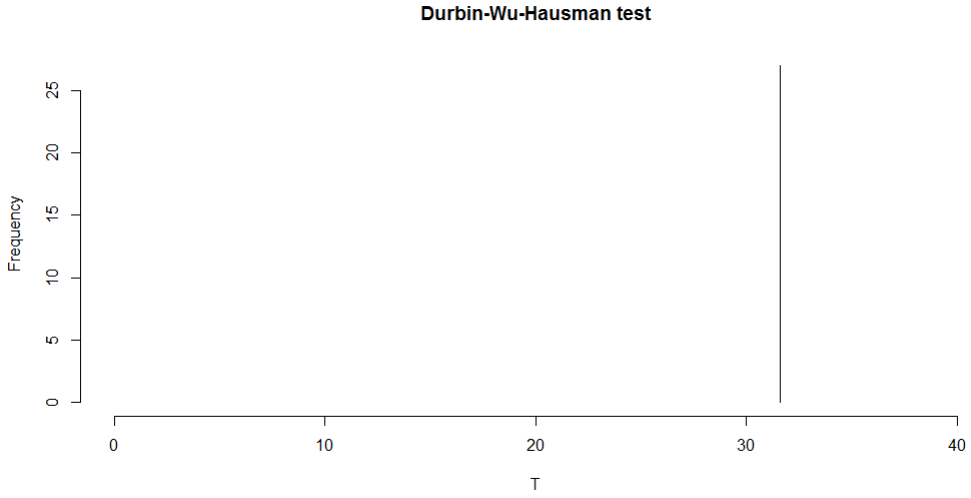
We did the simulation of T in R for a variable that is not correlated with the error term $\epsilon$ and one that is correlated with it.

The results for the model without correlation:




On the y-axe we can see the frequency out of 100 trials. Most of the values (95%) of $T$ are below $1,96$. We conclude that the value of $T$ are from the standard normal distribution.

The results for the models where $X1$ is correlated with $\epsilon$:




We can see that for the 100 trials, the value of T is equal to $31.62278 > 1,96$. $H_0$ is rejected.

## 5.3 Changing the variance of the instrument

We are now interested in the estimation of the $\hat{\beta}_{2SLS}$ when there is not correlation between $\epsilon$ and the variable $X1$. In our model we have standard instrumental variable defined by:

$$Z = X1 + \phi \tag{67}$$

where $\phi \sim N(0, \frac{1}{2})$.

We are changing the standard deviation of $\phi$ to 1, 10 and 100, then we are changing the variance of $Z$ to 2, 101, 10001.

On R we did for each of these variances the histogram of $|\beta - \hat{\beta}_{2SLS}|$,the histogram of $|\hat{\beta}_{OLS} - \hat{\beta}_{2SLS}|$.

For $Z = X1 + \phi$ where $\phi \sim N(0, \frac{1}{2})$ gives the histograms:

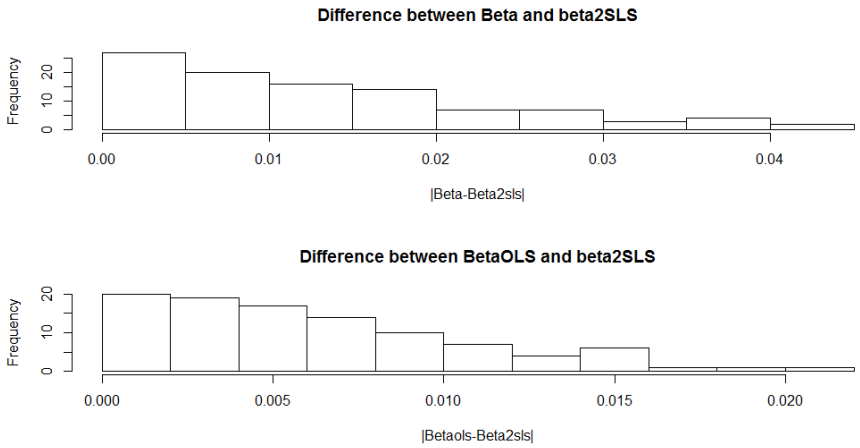



Figure 3: Histograms with $\text{Var}(Z) = \frac{5}{4}$

In these histograms, the difference $|\beta - \hat{\beta}_{2SLS}|$ is between the $0,00$ and $0,045$ and the difference $|\hat{\beta}_{OLS} - \hat{\beta}_{2SLS}|$ is between the $0,00$ and $0,025$

For $Z = X1 + \phi$ where $\phi \sim N(0, 10)$ gives the histograms:

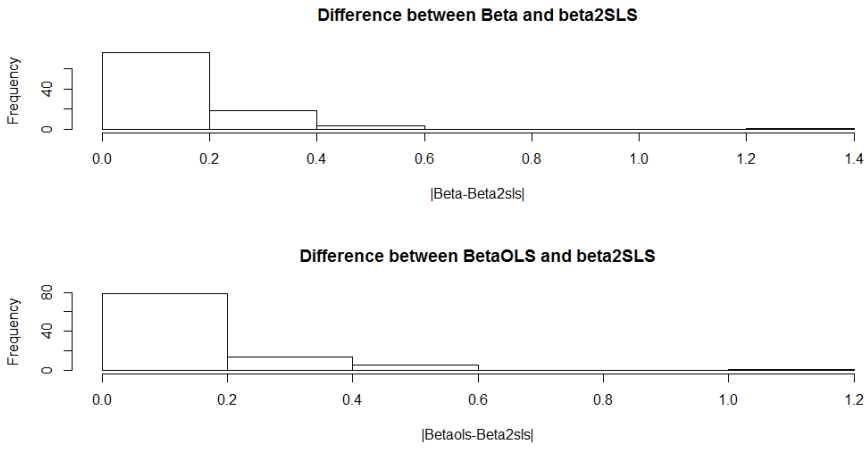



Figure 4: Histograms with $\text{Var}(Z) = 101$

In these histograms, the difference $|\beta - \hat{\beta}_{2SLS}|$ is between the $0,00$ and $1,4$ and the difference $|\hat{\beta}_{OLS} - \hat{\beta}_{2SLS}|$ is between the $0,00$ and $1,2$. We see that the scales of the differences become bigger that previous histogram.

For $Z = X1 + \phi$ where $\phi \sim N(0, 100)$ gives the histograms:

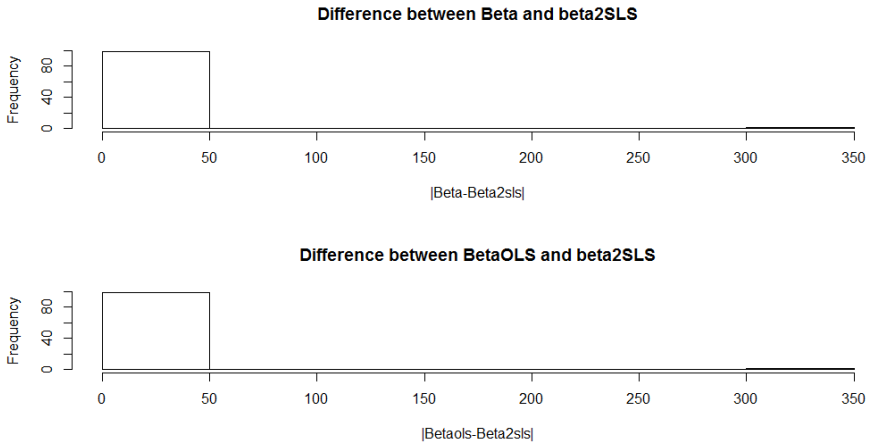



Figure 5: Histograms with $\text{Var}(Z) = 10001$

In these histograms, the difference $|\beta - \hat{\beta}_{2SLS}|$ is between the $0,00$ and $350$ and the difference $|\hat{\beta}_{OLS} - \hat{\beta}_{2SLS}|$ is between the $0,00$ and $350$. We see that the scales of differences become much bigger the two previous histograms.

The scales of the differences $|\beta - \hat{\beta}_{2SLS}|$ and $|\hat{\beta}_{OLS} - \hat{\beta}_{2SLS}|$ become bigger each time that we make the variance of $\phi$ bigger , implies also the growth of the variance of $Z$.

So the estimator found with an instrumental variable with a lower variance gives a better estimation of $\beta$

## 5.4 Changing the covariance between $X1$ and $Z$

In this part, we are interesting of the effect of changing the covariance between $X1$ and $Z$ on the quality of the estimator trough the two stage least square method.

In our model we have standard instrumental variable defined by:

$$Z = X1 + \phi \tag{68}$$

where $\phi \sim N(0, \frac{1}{2})$.

We are changing the coefficient of $X1$ to 1, 10, 100 an 1000. We did these simulations on R a we are again looking at the same histograms then previous part:

For $Z = X1 + \phi$, we did saw these histograms in figure 3

For $Z = 10X1 + \phi$, the covariance is:

$$\begin{aligned} \text{Cov}(Z, X1) &= \text{Cov}(10X1 + \phi, X1) \\ &= 10\text{Var}(X1) \\ &= 10 \end{aligned} \tag{69}$$

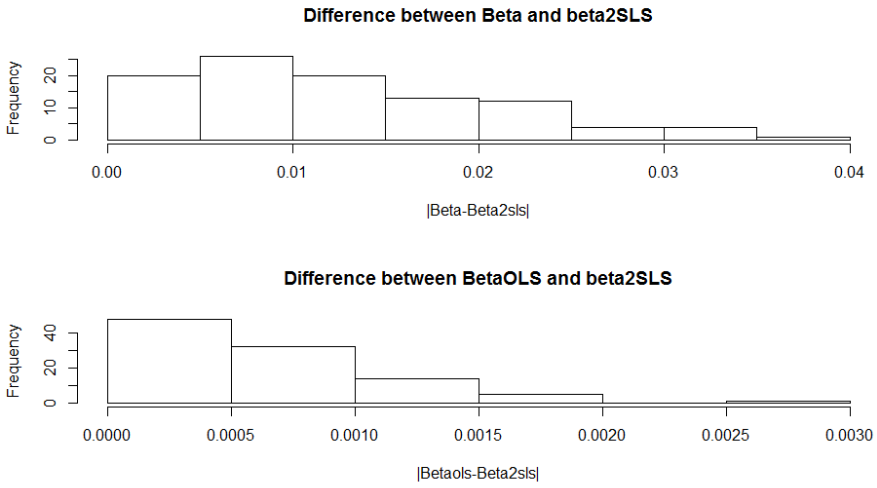



Figure 6: Histograms with $\text{Cov}(Z, X1) = 10$

In these histograms, the difference $|\hat{\beta}_{OLS} - \hat{\beta}_{2SLS}|$ is between the $0,00$ and $0,00025$. We see that the scale of this difference is smaller than when $\text{Cov}(Z, X1) = 1,100$

For $Z = 100X1 + \phi$, the covariance is:

$$\begin{aligned} \text{Cov}(Z, X1) &= \text{Cov}(100X1 + \phi, X1) \\ &= 100\text{Var}(X1) \\ &= 100 \end{aligned} \tag{70}$$

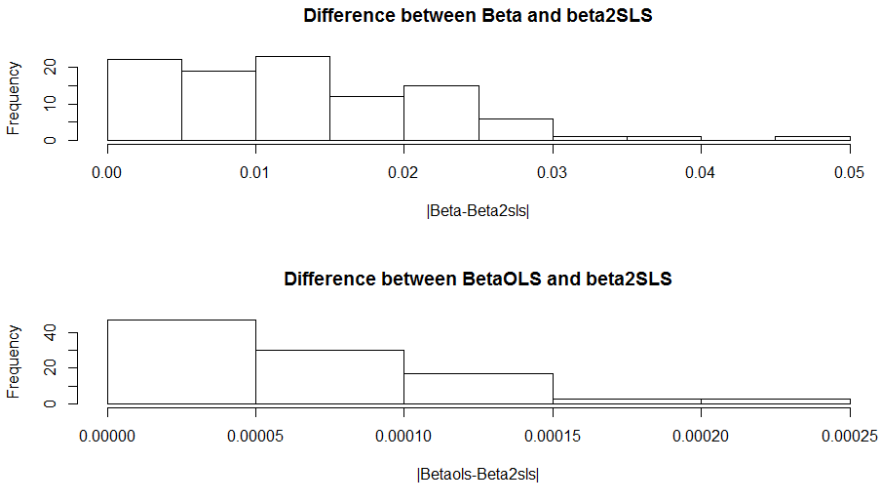



Figure 7: Histograms with $\text{Cov}(Z, X1) = 100$

In these histograms, the difference $|\hat{\beta}_{OLS} - \hat{\beta}_{2SLS}|$ is between the $0,00$ and $0,0030$. We see that the scale of this difference is smaller than the two previous ones.

For $Z = 1000X1 + \phi$, the covariance is:

$$\begin{aligned} \text{Cov}(Z, X1) &= \text{Cov}(1000X1 + \phi, X1) \\ &= 1000\text{Var}(X1) \\ &= 1000 \end{aligned} \tag{71}$$

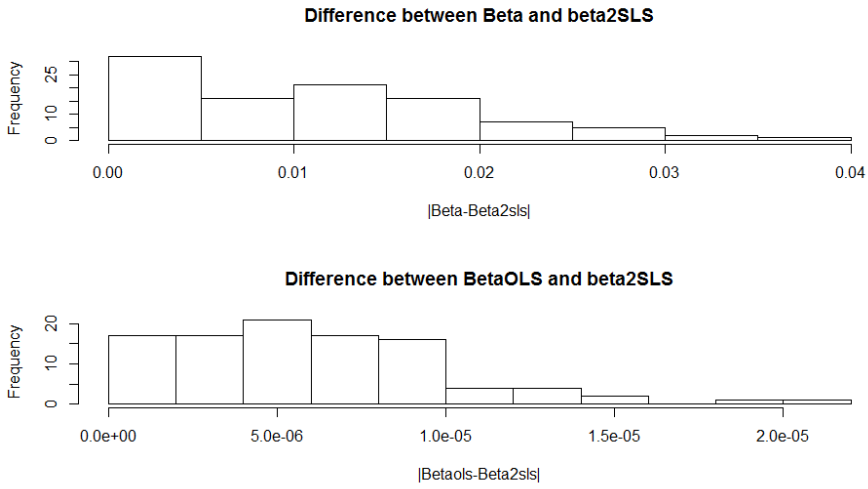



Figure 8: Histograms with $\text{Cov}(Z, X1) = 1000$

In these histograms, the difference $|\hat{\beta}_{OLS} - \hat{\beta}_{2SLS}|$ is between the $0,00$ and $2,0e-05$. We see that the scale of this difference is smaller than the three previous ones.

The difference $|\hat{\beta}_{OLS} - \hat{\beta}_{2SLS}|$ becomes smaller when the $\text{Cov}(Z, X1)$ is bigger. So the $\hat{\beta}_{2SLS}$ estimate like $\hat{\beta}_{OLS}$. So $\hat{\beta}_{2SLS}$ is a better estimator when $\text{Cov}(Z, X1)$ is bigger (explained in subsection 3.4).

The scale of the difference $|\beta - \hat{\beta}_{2SLS}|$, stays the same for the four different covariance. This is due to the fact that the $\hat{\beta}_{OLS}$ gets each time just a little bit closer to $\beta$, so the scale is to big to see that in the histogram.

These simulations give a good indication how the methods and the test works. Also the effect of the change of the variance of $Z$ and the covariance of $Z$ and $X1$.

# 6 Conclusion

In this thesis came forward that the instrumental variable estimation useful is to estimate parameters of a linear model when variables are endogenous. The use of the two stage least squares method is the best way to estimate the parameter $\beta$ under condition that the variable is endogenous. If the variable is not, then the ordinary least squares is the most accurate method. This theory is confirmed by the variances of the two estimators seen in subsection 3.4 and also by the simulation in subsection 5.1. The consistency of the two estimators by ordinary least squares and two stage least squares are proven. The distribution of the difference between the estimator and the parameter is normally distributed. In the simulation we have seen that the covariance of the instrument and the variable has a influence of the estimators. It could be interesting to research what the theory is behind the effect of the choice of a instrument on the estimators and the Durbin-Wu-Hausman test.

# References


[1] Unknown author, The Lindeberg-Feller Central Limit Theorem. Internet link:https://online.stat.psu.edu/∼dhunter/asymp/lectures/p93to100.pdf Date of visit: 08/07/2016.

[2] L. C. Adkins, R. C. Campbell, V. Chmelarova, and R. C. Hill. The Hausman Test, and Some Alternatives, with Heteroskedastic Data. *Essays in Honor of Jerry Hausman*, 29(May):515–546, 2015.

[3] F. Coolen, M. Troffaes, and T. Augustin. International Encyclopedia of Statistical Science. *International Encyclopedia of Statistical Science*, pages 645–648, 2011.

[4] P. Protter, J. Jacod. Probability Essentials. 2004.

[5] B. Lambert. Endogeneity and instrumental variables. Video link: https://www.youtube.com/watch?v=lLI-opK9MD8, Aug. 2013.

[6] B. Lambert. Two stage least squares - example. Video link: https://www.youtube.com/watch?v=54QIRrMkJsk, 09/2013.

[7] J. A. Rice. *Mathematical Statistics and Data Analysis*, volume 72. 2001.

[8] J. Lani (Owner Statistic Solution). Two-stage least squares (2sls) regression analysis. Link: http://www.statisticssolutions.com/two-stage-least-squares-2sls-regression-analysis/. Date of visit: June 2016.

[9] M. J. Steele. The Cauchy-Schwarz Master Class: An Introduction to the Art of Mathematical Inequalities. *American mathematical monthly*, 2005.

[10] A. Colin Cameron, P. Trivedi . Microeconometrics, Mehods and Applications. *Instrumental Variables*. (1975):255–260.

# 7 Appendix

R-code $X1$ not endogenous

```
#100 trials
for(j in 1:100){
n=1000
x1=rnorm(n)
alpha=1
beta=0.5
eps=rnorm(n,0,0.50)
phi=rnorm(n,0,0.5)
#Independent
epstilde=eps
z=1000*x1+phi
y=alpha+beta*x1+epstilde

#Beta alpha of  OLS and 2SLS

summary(lm(y~x1))
c1=coef(lm(y~x1))
betaols=c1[2]
alphaols=mean(y)-betaols*mean(x1)
b=coef(lm(x1~z))
x1hat=b[1]+b[2]*z
summary(lm(y~x1hat))
c2=coef(lm(y~x1hat))
beta2sls=c2[2]
alpha2sls=mean(y)-betaols*mean(x1hat)
#Denominator OLS en 2SLS
nols=0
for(i in 1:n){ nols= nols + (x1[i]-mean(x1))*(x1[i]-mean(x1))
}
n2sls=0
for(i in 1:n){ n2sls= n2sls + (z[i]-mean(z))*x1[i]
}

varbeta=0
for(i in 1:n){ varbeta= varbeta + ((x1[i]-mean(x1))/nols-(z[i]-mean(z))/n2sls)^2
}


#Variance of Y

vary=0
for(i in 1:n){ vary= vary+ 1/n*(y[i]-alphaols-betaols*x1[i])^2
```

```
}
R[j]=abs(betaols-beta2sls)
P[j]=abs(beta-beta2sls)
T[j]=abs((betaols-beta2sls)/(sqrt(varbeta*vary)))
}

par(mfrow=c(2,1))
hist(P, xlab = "|Beta-Beta2sls|", main = "Difference between Beta and beta2SLS")
hist(R, xlab = "|Betaols-Beta2sls|", main = "Difference between BetaOLS and beta2SLS")
hist(T, main = "Durbin-Wu-Hausman test")
```

R-code $X1$ endogenous

```
#100 trials
for(j in 1:100){
n=1000
x1=rnorm(n)
alpha=1
beta=0.5
eps=rnorm(n,0,0.50)
#dependent
epstilde=x1+eps
z=x1-(4/5)*epstilde
y=alpha+beta*x1+epstilde

#Beta and alpha OLS en 2SLS

summary(lm(y~x1))
c1=coef(lm(y~x1))
betaols=c1[2]
alphaols=mean(y)-betaols*mean(x1)
b=coef(lm(x1~z))
x1hat=b[1]+b[2]*z
summary(lm(y~x1hat))
c2=coef(lm(y~x1hat))
beta2sls=c2[2]
alpha2sls=mean(y)-betaols*mean(x1hat)
#denominator OLS en 2SLS
nols=0
for(i in 1:n){ nols= nols + (x1[i]-mean(x1))*(x1[i]-mean(x1))
}
n2sls=0
for(i in 1:n){ n2sls= n2sls + (z[i]-mean(z))*x1[i]
}

varbeta=0
for(i in 1:n){ varbeta= varbeta + ((x1[i]-mean(x1))/nols-(z[i]-mean(z))/n2sls)^2
```

```
}


#Variance of Y

vary=0
for(i in 1:n){ vary= vary+ 1/n*(y[i]-alphaols-betaols*x1[i])^2
}


T[j]=abs((betaols-beta2sls)/(sqrt(varbeta*vary)))
R[j]=abs(betaols-beta2sls)
P[j]=abs(beta-beta2sls)
}


par(mfrow=c(2,1))
hist(P, xlab = "|Beta-Beta2sls|", main = "Difference between Beta and beta2SLS")
hist(R, xlab = "|Betaols-Beta2sls|", main = "Difference between BetaOLS and beta2SLS")
hist(T, xlim = c(0,40), main = "Durbin-Wu-Hausman test")
```