University of Leiden

Master thesis

# Instrumental variable estimation with dichotomous outcomes

*Author:*
Hendrik Grondijs

*Supervisor:*
Dr. Saskia le Cessie

April 9, 2015

# Instrumental variable estimation with dichotomous outcomes

Hendrik Lodewijk Grondijs
hlgrondijs@gmail.com

April 9, 2015

**Abstract**

In many causal relationships there is a third factor that influences both the explanatory and the outcome variable. This is called a confounder and if left out of the model can cause bias in the parameter estimates. A solution to this problem are Instrumental Variable estimation methods. An instrument is a variable that correlates with the outcome only through its relation with the treatment. This means that it is not correlated with confounders or directly influencing the outcome. Many IV methods were designed to deal with continuous outcomes in a linear model. However, in medical research the outcome is often a dichotomous one and linear models may not always be appropriate. The research question in this thesis is to find out which IV methods are useful when dealing with binary outcome variables. There are two approaches to the estimation of the treatment effect: risk difference and the (log-)odds ratio. We compared methods within these two approaches by performing simulations in multiple scenarios as well as a case study on the effect of corticosteroids on 30 day mortality and infection. We found that the IV methods discussed work relatively well, but that their usefulness is constrained in the medical fields as datasets often lack the required number of observations. Other findings were that the two stage logistic regression estimator actually estimates a different quantity than the log odds ratio a logistic regression would normally produce.

Keywords: Instrumental variables, endogeneity, logistic regression, causality, non-collapsibility.

# Contents

# 1 Introduction

In epidemiology, exposure to some factor and its effect on health are subject of research. Mostly one wants to know whether, and, if so, to what extent, administering a treatment is causally related to the probability of a subject's survival, the occurrence of disease or to a change in his/her condition. In general, when dealing with causality one must be very careful about drawing conclusions. Correlation does not necessarily imply causation: another so-called confounding factor may well be the cause of the correlation. The instrumental variable methods discussed in this paper aim to prevent the classification of this kind of spurious relationships as a causal path.

A confounder is a factor that influences both the exposure variables and the outcome of our model. For example, the relation between smoking and an early death is confounded by socio-economic status. When confounding is present, standard methods for inference like Ordinary Least Squares (OLS) give biased estimates. One solution to the problem of confounding is to conduct a Randomized Controlled Trial (RCT). By distributing different treatments randomly across the subjects, possible confounders (e.g. in the case of medical treatment: age, disease severity) are expected to be evenly distributed across the treatment groups.

Unfortunately, there are many situations in which a randomized controlled trial is not feasible and the researchers will have to work with data from observational studies. These studies are always confounded to some degree and the researcher must decide which confounders are the most influential and measure them, to compensate for their effect. To this end the researcher ought to know all relevant confounding factors beforehand and be able to measure them at the same time, which is generally an unrealistic condition.

To deal with this statistical problem of unmeasured confounders, a scholar named Philip G. Wright developed a method in the early 20th century called Instrumental Variable (IV) analysis. He first documented this method when he used IV analysis for estimating supply elasticity.[14] Epidemiologists have only started to use IV-analysis over the last two decades.

To conduct IV analysis, the researcher has to identify an "instrument": a factor that influences the outcome only through the exposure and not directly. In the example where the effect of smoking on early death is studied, an instrument could be the price of cigarettes. It influences the outcome (early death) only through the exposure (smoking), because a high price of cigarettes only decreases the probability of early death, if it leads to decreased smoking. The diagram 1 illustrates the causal relations that are

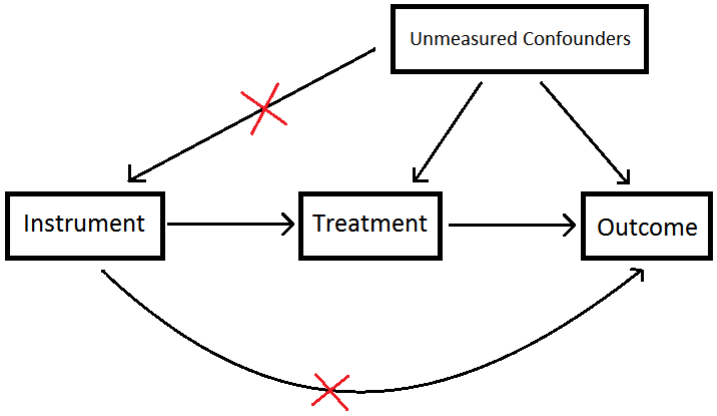



Figure 1: The red crosses indicate causal paths not allowed in an IV analysis.

assumed (not) to exist. It is assumed that there is a relation between the instrument and treatment, but the instrument is not affected by the unmeasured confounders. Furthermore the Instrument only affects the outcome via the treatment.

Instrumental variable analysis uses the relation between instrument and exposure, as well as the relation between instrument and outcome (via the treatment), for estimating the relation between this exposure and the outcome. We will show that, when exposure and outcome are continuous and we have a valid instrument, the often used Two Stage Least Squares (2SLS) estimator leads to asymptotically unbiased results under the instrumental variable assumptions, depicted in Figure 1 and discussed in detail in the next subsection. However this is done at the cost of some efficiency. Therefore in the situation of a continuous exposure and outcome, this procedure is the most common way to conduct IV analysis.

Epidemiologists often deal with dichotomous variables instead of continuous ones. Different IV analysis methods might be better suited to these kind of problems.

In this thesis we evaluate a number of estimation procedures to investigate whether we can obtain unbiased estimates in case of a dichotomous outcome and exposure.

## 1.1 IV in practice

To illustrate the method of IV analysis we present some examples of practical situations in which IV analysis can be useful.

### 1.1.1 Econometrics

In (macro-)economics it is rarely possible to conduct a controlled experiment. Researchers have to collect their data from observed real world events and draw their conclusions based on these observational data. Confounding is therefore often an issue. An example is given in the book of Heij et al. [8]. They want to analysed the relation between monthly changes in the AAA bond rate and in the short-term interest rate. Confounders here could be general financial conditions as they affect both the bond rates and interest rates. As instrumental variable they use the past changes in the short term interest rate as instrument. If these past changes are independent from current financial conditions and if they do only influence the current AAA bond rate via the current short term interest changes, this is a suitable IV .

### 1.1.2 Epidemiology

In epidemiology randomized controlled trials are, just as in econometrics, not always feasible. The effect of smoking for example can not be studied in a Randomized Controlled Trial. Even when randomized controlled trials are possible it is difficult to estimate the actual effect of treatment, because there is the problem of non-compliance. When treatments are assigned to the subjects, but the subjects fail to comply for reasons related to their outcome, we are essentially dealing with confounding when the actual treatment effect is estimated. An extensive framework has been developed which uses instrumental variable techniques to deal with non-compliance in randomized controlled trials. In this setting the instrument is the treatment assignment, whereas the exposure becomes the actual treatment received. It will be explored further later in this thesis.

In epidemiology, instrumental variables are becoming more popular in the last twenty years. The relation between the treatment and the outcome is often obscured by many (measurable or unmeasurable) confounders. An example of a confounder could be the age of a patient. An example of an unmeasurable confounder could be the appearance of a patient. Confounders influence the choice of treatment as well as the eventual outcome. To make inference about the effect of a treatment, one needs to extract the true effect of the treatment from this web of confounding. If all confounders are known and measured, standard statistical methods like regression analysis can be used to adjust for the confounders. However in practice confounders are often unknown or unmeasurable.

A way of circumventing this problem is by use of instrumental variables. An example of this solution is given in Brookhart et al [1]. They use physician

preference as an instrumental variable when trying to estimate the effect of a treatment. Intuition behind this is that some physicians are more inclined to prescribe a certain treatment to their patients than others. This means that this preference has some predictive value for the treatment selected and the selected treatment is not just determined by the patient characteristics.

Another application of instrumental variable analysis in epidemiology is Mendelian randomization. In this type of instrumental variable analysis, genes are used as instrumental variables. When it is assumed that the genes are allocated randomly across the subjects studied, this creates a naturally randomized controlled trial. [2] There is a large problem with this method, being that the genes may have other causal paths leading to the outcome in addition to the path that is being studied. This would violate the assumptions of IV and the unmeasured confounding will still obscure the parameter estimates.

Measurement error correction is one of the other applications of IV analysis. One can make the statement that we can never observe the true value of the exposure, but only a noisy version is measurable. One can use IV analysis to obtain a measurement error-free estimate of the exposure

## 1.2 Outline of this thesis

The outline of this thesis is as follows. First we will take a closer look at the assumptions that have to be made for an IV analysis to be relevant. This includes the models and assumptions that we have to make as well as an explanation of the estimation procedures in the general case. In the fourth chapter we will take a look at these models when both outcome and treatment are binary variables. This includes both linear and non-linear models and estimation methods. In chapter five we take a closer look at other potential application areas. In the sixth chapter we conduct a simulation experiment to see how the proposed methods perform in eight different scenarios. Finally we will discuss these results.

# 2 IV assumptions

There are some assumptions that have to be met for an variable $Z$ to be a valid instrument. For regularity we assume that $\text{Var}[Z] > 0$. Let $Y$ be the outcome and $X$ be the exposure of interest. The assumptions for an instrumental variable can be formulated in different ways. We will use the formulation of Clarke and Windmeijer. [3]

1. There is no correlation between the instrument $Z$ and the unmeasured confounder $U$ ($Z \perp\!\!\!\perp U$).

2. There is no direct relation between $Z$ and $Y$ ($Z \perp\!\!\!\perp Y | X, U$).

3. Furthermore there has to be enough correlation between X and Z model ($Z \not\perp\!\!\!\perp X$).

Now we describe how these assumptions could be violated in the physician preference example.

- When the physician's preference is somehow correlated with any of the unmeasured confounders, we violate assumption one. It is possible that some of the doctors in the study work in a specific country and that they have different preferences than other doctors that work in other countries. If living in a specific country has an influence on both the preference for a certain treatment and outcome (i.e. preference for Ebola treatment can differ between doctors in different countries) under investigation, we are then violating the first assumption.

- If doctors with a large preference for a certain treatment, are treating patients different in other ways (for example by prescribing other medication in addition to the treatment investigated), not all of the effect of $Z$ on $Y$ goes via $X$ and we violate the second assumption.

- The physician's preference is often measured by the last treatment prescribed by the physician for the disease being investigated. However this may be a bad predictor for the treatment prescribed to the current patient (meaning that $Z$ and $X$ can be very small).

In econometrics often multiple explanatory variables are modelled. When one has modelled multiple endogenous explanatory variables, at least as many instruments as there are endogenous explanatory variables are required. However, we will not explore the case with multiple instruments in this thesis.

# 3 A model for unmeasured confounding

In this section we introduce the reader to a simple model with unmeasured confounding. We first consider the situation with a continuous outcome and a continuous exposure, as is common in economics. Let $y_i$ denote element $i$ of the outcome vector $Y$. Let $x_i$ denote the $i$th observation of an endogenous

exposure $X$. We assume that there is a confounding variable $U$ which affects both $X$ and $Y$, where $u_i$ the $i$th observation of an unmeasured confounder $U$. We assume that there is an instrumental variable which affects $X$, but is independent of $U$ and does not affect $Y$ directly. We let $z_i$ be the $i$th observation of a instrumental variable $Z$. Instrumental variable analysis involves using the instrument $Z$ to estimate the values of $X$.

$$x_i = \alpha_0 + z_i\alpha_1 + \epsilon_{Xi} \tag{1}$$

In agreement with the diagram showed in the previous section we assume the following linear structural equation:

$$y_i = \beta_0 + x_i\beta_1 + \epsilon_{Yi} \tag{2}$$

and we assume:

$$\begin{pmatrix} \epsilon_{Xi} \\ \epsilon_{Yi} \end{pmatrix} \sim IIN \left( 0, \quad \begin{bmatrix} \sigma_X^2 & \sigma_{XY} \\ \sigma_{XY} & \sigma_Y^2 \end{bmatrix} \right). \tag{3}$$

As with any estimator, we need assumptions for the target parameter to be identified. Here we assume that, $\alpha_0$, $\alpha_1$, $\gamma_1$, $\beta_0$, $\beta_1$ and $\gamma_2$ are unknown scalar parameters. To model the unmeasured confounding we make the assumption that $U$ is a confounder by defining our residuals to be functions of that confounder.

$$\epsilon_{Xi} = u_i\gamma_1 + \eta_{Xi}$$

$$\epsilon_{Yi} = u_i\gamma_2 + \eta_{Yi}$$

and that $\eta_X$ and $\eta_Y$ are vectors of disturbances, which are IID as $\eta_X \sim N(0, \sigma_X^2 I)$ and $\eta_Y \sim N(0, \sigma_Y^2 I)$. The disturbances $\eta_X$ and $\eta_Y$ are assumed independent of $z_i$, $x_i$, $u_i$ and each other. This implies that we assume there are no other unmeasured confounders than those specified in our model.

Note that the linear models (1) and (2) described above satisfies the third IV condition if $\alpha_1 \neq 0$. That IV assumption 2 is true can be seen in (2) where $y_i$ does not depend on $z_i$, given $x_i$ and $u_i$. Assumption 1 is true because we assumed that $u_i$ and $z_i$ are independent.

## 3.1 Problems with Ordinary Least Squares

To illustrate the problems caused by unmeasured confounding we take a look at what would happen if we used Ordinary Least Squares estimation without

including the unmeasured confounders $U$. In that situation we consider the model

$$y_i = \tilde{\beta}_0 + x_i\tilde{\beta}_1 + \epsilon_i \tag{4}$$

The least squares method of obtaining estimates $b_0$ and $b_1$ of the unknown parameters $\tilde{\beta}_0$ and $\tilde{\beta}_1$ in model 4 is by minimizing the sum of squared residuals ($\sum_{i=1}^{n} e_i^2$ where $e_i$ are the residuals).

$$S = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - b_0 - x_i b_1)^2$$

To minimize the sum of squared error terms in the model, one can set the partial derivatives of the object function $S$ equal to zero, obtaining the first-order conditions. Taking a closer look at the second first-order condition we find the problem caused by unmeasured confounding.

$$\frac{\partial S}{\partial b_1} = -2 * \sum_{i=1}^{n} x_i(y_i - b_0 - b_1 x_i) = 0$$

If we cannot measure the confounder $U$, we cannot estimate $\gamma_2$ and $U$ will be ommitted from our model. This means that this estimation procedure, does not include the relation between $U$ and $Y$. If the structural equations 1 and 2 are true, then $\epsilon_i = \gamma_2 u_i + \eta_{Yi}$. This violates one of the assumptions of OLS estimation, namely the exogeneity of the treatment-outcome relationship $\mathbb{E}[x_i\epsilon_i] = 0$, because according to our model (1) and (2):

$$\mathbb{E}[x_i\epsilon_i] = \mathbb{E}[(\alpha_0 + \alpha_1 z_i + \gamma_1 u_i + \eta_{Xi})(\gamma_2 u_i + \eta_{Yi})] = \mathbb{E}\left[\alpha_0\gamma_2 u_i + \gamma_1\gamma_2 u_i^2\right] \neq 0$$

Because $X$ and $U$ are correlated, our estimate of $\beta_1$ will be subject to omitted variable bias. This means that the Least Squares estimate $b_1$ is not an unbiased estimate of $\beta_1$ in 2.

### 3.1.1 Omitted Variable Bias

When one leaves out a significant explanatory variable $U$ that correlates with the treatment of interest $X$ in an OLS procedure, the estimate of the target parameter $\beta_1$ in model 2 will be biased. This is because when we calculate OLS estimates we assume that our model is equal to model 4. One can show that estimating $\tilde{\beta}_1$ in this model is not equal to estimating $\beta_1$ from our model (2) [8]. The presence of this endogenous relationship between the treatment and outcome can be formally stated as

$$\mathbb{E}\left[x_i \epsilon_i\right] \neq 0 \tag{5}$$

The deviation from $\beta_1$ can be attributed to $\tilde{\beta}_1$ compensating for the effect ($\gamma_2$) of $U$ which is not included in the model, but is still correlated with both $X$ and $Y$. The Ordinary Least Squares estimate of $\tilde{\beta}_1$ in model (4) is

$$b_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

Using that $y_i - \bar{y} = (x_i - \bar{x})\tilde{\beta}_1 + u_i\gamma_2 + \eta_{Yi}$ yields:

$$b_1 = \beta_1 + \gamma_2 \frac{\sum_{i=1}^{n}(x_i - \bar{x})u_i}{\sum_{i=1}^{n}(x_i - \bar{x})^2} + \frac{\sum_{i=1}^{n}(x_i - \bar{x})\eta_{Yi}}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

Multiplying both the numerator and the denominator of the last term by $\frac{1}{n}$, we see that by applying the law of large numbers, the last term approaches zero as $n$ goes to infinity.

$$\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})\eta_{Yi} = \mathbb{E}\left[X\eta_{Yi}\right] - \mathbb{E}\left[X\right]\mathbb{E}\left[\eta_{Yi}\right] = 0 - \mathbb{E}\left[X\right] * 0 = 0$$

while,

$$\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \mathrm{Var}[X] > 0$$

The second term of the equation displays the observed covariance between $x$ and $u$. Multiplying both sides of this fraction by $\frac{1}{n}$ the numerator becomes:

$$\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})u_i = \mathbb{E}\left[XU\right] + \mathbb{E}\left[X\right]\mathbb{E}\left[U\right] = cov(X, U)$$

This shows that the probability limit of $b_1$ is

$$b_1 \xrightarrow{plim} \beta_1 + \gamma_2 \frac{cov(X, U)}{Var[X]}$$

which shows that the OLS estimate is no longer consistent. If we are able to measure $U$, we can include it in the OLS estimation procedure and the disturbances $\epsilon$ will no longer contain the unmeasured confounder $U$, which solves this violation of the exogeneity conditions of OLS. However, because we are interested in the case where we can not measure $U$, one has to find a different solution.

## 3.2 Simple Instrumental Variable Analysis

A possible way of consistently estimating the parameter of interest $\beta_1$ in model 2 is instrumental variable analysis. We try to isolate the variation in $Y$ that is due to $X$, from the effect due to $U$, by using an instrument $Z$. In the case of only one instrument, the (Wald) IV-estimator is defined as [3]

$$b_{IV} = \frac{\widehat{cov(z,y)}}{\widehat{cov(z,x))}}$$

, with $\widehat{cov(z,y)}$ and $\widehat{cov(z,x)}$ the observed covariance in the sample.

## 3.3 Two Stage Least Squares

When dealing with more than one instrument or more than one regressor, we have to use a technique called Two Stage Least Squares to obtain instrumental variable estimates. By splitting the estimation into two separate regressions we can easily calculate the 2SLS estimator, by regressing $Y$ on fitted values of $X$ to obtain an unconfounded estimate [3]. Here we will describe this estimation procedure. Intuition behind this method is that we try to "clean" the effects of the confounder $U$ on $X$ and only use the variation that is not caused by $U$. This is done by first making a prediction of $X$ by using the instrument $Z$ and then using this predicted $X$ to estimate the relation between $X$ and $Y$.

To describe the estimation procedure we introduce some matrix notation. $\boldsymbol{X}$ is a matrix with two columns, consisting of a vector of ones $(1 \times n)$ and the vector $X$ $(1 \times n)$ to include an intercept in our model $(\mathbf{X} = (1, X))$. The matrix $\boldsymbol{Z}$ contains a vector of ones and the instrument vector $(\boldsymbol{Z} = (1, Z))$. $\boldsymbol{Y}$ is the outcome vector $(1 \times n)$. Let $\boldsymbol{\beta} = [\beta_0, \beta_1]$.

First stage: Regress $\boldsymbol{X}$ on $\boldsymbol{Z}$ and store the fitted values $\hat{\boldsymbol{X}} = \mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X}$

Second stage: Regress $\boldsymbol{Y}$ on $\hat{\boldsymbol{X}}$ giving the parameter estimates $\boldsymbol{b}_{2SLS} = (\hat{\boldsymbol{X}}'\hat{\boldsymbol{X}})^{-1}\hat{\boldsymbol{X}}'\boldsymbol{Y}$. Rewriting this expression reduces it to $\boldsymbol{b}_{2SLS} = (\mathbf{Z}'\mathbf{X})^{-1}\mathbf{Z}'\boldsymbol{Y}$

We can plug in our model 2 in this expression of $\boldsymbol{b}_{2SLS}$ to further study the estimator.

$$\boldsymbol{b}_{2SLS} = (\boldsymbol{Z}'\boldsymbol{X})^{-1}\boldsymbol{Z}'(\boldsymbol{X}\boldsymbol{\beta} + U\gamma_2 + \eta_Y) = \boldsymbol{\beta} + (\boldsymbol{Z}'\boldsymbol{X})^{-1}\boldsymbol{Z}'U\gamma_2 + (\boldsymbol{Z}'\boldsymbol{X})^{-1}\boldsymbol{Z}'\eta_Y$$

Because we have assumed no correlation between the instrument $Z$ and the unmeasured confounder $U$, as well as no correlation between $Z$ and the residuals $\eta_Y$, t we have that $\mathbb{E}\,[(\boldsymbol{Z}'U] = \mathbf{0}$ and $\mathbb{E}\,[\boldsymbol{Z}'\eta_Y] = \mathbf{0}$. This implies that the expectation of $b_{\text{SLS}}$ is asymptotically equal to $\boldsymbol{\beta}$.

### 3.3.1 Small sample bias of 2SLS

In practice the number of observations $n$ does not approach infinity. This means that the 2SLS estimator is biased for small $n$.

Using the definition of $\boldsymbol{b}_{2SLS}$ above, one can show that

$$\boldsymbol{b}_{2SLS} - \boldsymbol{\beta} = \frac{\alpha' \boldsymbol{Z}' \epsilon_Y + \epsilon_X \boldsymbol{P_Z} \epsilon_Y}{\alpha' \boldsymbol{Z}' \boldsymbol{Z} \alpha + \alpha' \boldsymbol{Z}' \epsilon_X + \epsilon_X \boldsymbol{Z} \alpha + \epsilon_X \boldsymbol{P_Z} \epsilon_X}$$

where $\boldsymbol{P_Z} = Z(Z'Z)^{-1}Z'$.

Hahn and Hausman (2002) [10] show that, using asymptotic expansion techniques, the finite sample bias of the 2SLS estimator can be approximated as

$$\mathbb{E}\left[\boldsymbol{b}_{2SLS}\right] - \boldsymbol{\beta} \approx \frac{\sigma_{\epsilon_X \epsilon_Y}}{\sigma_{\epsilon_X}} \frac{k}{\mu + k}$$

where $\mu$ is the concentration parameter.

$$\mu = \frac{\alpha' \boldsymbol{Z}' \boldsymbol{Z} \alpha}{\sigma_X^2}$$

From this we can conclude

- 1 The finite sample bias increases when the covariance between $\epsilon_{Xi}$ and $\epsilon_{Yi}$ is high. This happens when the effect of the unmeasured confounding is large.
- 2 The finite sample bias also increases when the instrument does not predict our treatment very well. As $\alpha$ goes to zero, $\mu$ also approaches zero and the bias approximation becomes larger.

### 3.3.2 Asymptotic variance of 2SLS

We know from the Gauss-Markov theorem that, in the category of unbiased estimators, the OLS estimator has the smallest variance. So we know that the variance of our IV estimator will always be larger or equal to the variance of the least squares estimator. We can state that the 2SLS estimator is giving up efficiency in return for removing the bias that is due to unmeasured confounding. The asymptotic distribution of the 2SLS estimator $\boldsymbol{b}_{2SLS}$ has been derived in the book of Heij [8]. There it is shown that for large enough finite samples $\boldsymbol{b}_{2SLS}$ is asymptotically distributed as

$$\boldsymbol{b}_{2SLS} \xrightarrow{d} N(\beta_1, \sigma_Y^2 (X' P_Z X)^{-1})$$

This expression of the variance of the Two Stage Least Squares estimator implies that our loss in efficiency will be larger when the correlation between our instrument $Z$ and exposure $X$ is small. In other words, if we have a instrument that does not predict $X$ very well, our estimator could be far from the true value, even though it is an asymptotically unbiased estimation method.

### 3.4 Method of Moments

Another way of estimating $\boldsymbol{b}_{2SLS}$ is through the Method of Moments. This method of estimation uses the expressions for population moments implied by the model (1). Essentially we reformulate our model (1) into moment conditions. These population moments are replaced by sample moments to estimate the parameter of interest.

The sample moment conditions used to estimate $\beta_0$ and $\beta_1$ are given by

$$\boldsymbol{Z}'(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{b}) = \boldsymbol{0}$$

Where $\boldsymbol{0}$ is the null-vector in $\mathbb{R}^2$. Because we are in the two dimensional vector space, this equation entails two conditions. The first condition equates the mean of the residuals to zero, because our matrix $\boldsymbol{Z}$ contains an intercept, as is the case with any model with an intercept.

The second condition embodies the exogeneity assumption $\mathbb{E}[z'\epsilon] = 0$. Together they give us two equations to estimate the two unknown parameter vector $\boldsymbol{\beta}$. Rearranging this equation gives us the solution $\boldsymbol{b}_{2SLS} = (\boldsymbol{Z}'\boldsymbol{X})^{-1}\boldsymbol{Z}'\boldsymbol{Y}$. This is the exact same result as the two stage least squares procedure.

## 4 Binary outcomes

In this thesis the main focus is on using instrumental variables techniques to control for unmeasured confounding, when outcome and treatment are both dichotomous variables. In this section we will discuss the methods that we are going to evaluate in our simulation in chapter 6. Most of these methods are discussed in Rassen et al [13].

### 4.1 Linear probability models

One can assume that model (1) and (2) are still valid in the binary case, with the $\eta_{Xi}$ and $\eta_{Yi}$ variables independent of $X$, $U$ and $Z$ with mean 0. When $X$ and $Y$ are binary the model can be written as

$$P[x_i = 1|z_i, u_i] = \mathbb{E}\left[x_i|z_i, u_i\right] = \alpha_0 + \alpha_1 z_i + \gamma_1 u_i \tag{6}$$

$$P[y_i = 1|x_i, u_i] = \mathbb{E}\left[y_i|x_i, u_i\right] = \beta_0 + \beta_1 x_i + \gamma_2 u_i \tag{7}$$

The estimate for $\beta_1$ of linear probability models can be interpreted as the Risk Difference (RD) between being exposed (X=1) or not (X=0).

### 4.1.1 Two Stage Least Squares for binary outcome

We can estimate these probabilities by conducting for example the two stage least squares procedure as described in 2.1.3 to obtain an estimate of $\beta_1$ without knowing the values of the confounder $U$. This is done in Brookhart et al. [1].

In this approach our residuals ($e_i = y_i - b_0 - b_1 x_i$) are not normally distributed any more, as $Y$ is a binary variable. In addition, this will sometimes produce estimated probabilities greater than one or smaller than zero, as in this model the linear functions used to model $P[y_i = 1]$ and $P[x_i = 1]$ are not naturally bounded between 0 and 1. These probabilities can therefore be difficult to interpret. We are mostly interested in the Risk Difference estimate when using the linear probability model, so this interpretation difficulty should not be an issue. We do feel that we can do better, so we will also try fitting non-linear probability models.

### 4.1.2 Three Stage Approach

A three stage approach is appropriate when one is interested in risk difference estimates, but unsure whether the first stage model is correctly specified in (6).

Because the exposure variable is a dichotomous one, we might want to replace the first stage regression with its logistic regression equivalent, as this bounds its fitted values between 0 and 1, to deal with misspecification.

However, if this first stage logistic model is then misspecified, the second stage estimates will be biased. As a solution, we can use a logistic regression of $X$ on $Z$ to obtain the predicted probability $\hat{p}$ that $x_i = 1$. Then a linear first stage model is estimated using $\hat{p}$ as the instrument instead of $Z$. Lastly the linear probability model (7) is fitted to obtain our estimate of the risk difference. [13].

## 4.2 Non-linear probability models

To ensure that probabilities will fall between 0 and 1, often a non-linear model is employed.

$$P[x_i = 1|z_i, u_i] = F(\alpha_0 + \alpha_1 z_i + \gamma_1 u_i) \tag{8}$$

$$P[y_i = 1|x_i, u_i] = F(\beta_0 + \beta_1 x_i + \gamma_2 u_i) \tag{9}$$

with $F$ a Cumulative Distribution Function (CDF), because its values range between zero and one. Common choices for this CDF are the normal and the logistic density functions. These models are called respectively the probit and the logit model. Estimation of the parameters and their standard errors is done via maximum likelihood procedures.

An interesting notion is that these probit and logit models are non-collapsible. This means that in a model with two independent risk factors $X_1$ and $X_2$, leaving out $X_2$ will yield an estimate of the effect of $X_1$ which is on average closer to 0, because one of the , because $X_2 \not\perp\!\!\!\perp Y|X_1$ and/or $X_2 \not\perp\!\!\!\perp X_1|Y$ [7].

In this thesis we explore how non-linear models perform in a instrumental variable analysis setting. We will now take a look at the non-linear models that are investigated. These models were proposed in Rassen et al. [13]

## 4.3 Two Stage Logistic Model

An intuitive course of action when dealing with both dichotomous outcome $Y$ as well as dichotomous exposure $X$ is to assume a Two Stage Logistic Model. The Two Stage Logistic Model takes the following form:

$$P[x_i = 1|Z, U] = F(\alpha_0 + \alpha_1 z_i + \gamma_1 u_i)$$
$$P[y_i = 1|X, U] = F(\beta_0 + \beta_1 x_i + \gamma_2 u_i)$$

for $i = 1, .., n$ where $F(x) = \frac{\exp(x)}{1+\exp(x)}$ is the logistic cumulative distribution function.

### 4.3.1 Two Stage Logistic Estimation

The estimation of this model is similar to the estimation of Two Stage Least Squares, but using logistic instead of linear regression. This method is relatively easy to implement and has the benefit of providing us directly with an estimate of the log-odds ratio $\beta_1$. Wooldridge warns that this procedure will not yield consistent results [15].

The two stage logistic estimation procedure is as follows:

First stage: Fit a logistic model of X on Z through maximum likelihood estimation. Store the fitted probabilities $\hat{X}$.

Second stage: Fit another logistic model, explaining Y with the fitted values of the first stage ($\hat{X}$).

### 4.3.2 Method of Moments

Again the Method of Moments can be used to obtain an estimate of the treatment effect $\beta_1$. In similar fashion as with the linear model in 2.1.4 we derive the sample moment conditions in matrix form.

$$\boldsymbol{Z}'(\boldsymbol{Y} - F(\boldsymbol{X}\boldsymbol{b})) = \boldsymbol{0}$$

where $F(x) = \frac{\exp(x)}{1+\exp(x)}$ is the logistic distribution. Notably, the solution to this system of equations is not equal to the maximum likelihood estimator in the Two Stage Logistic Model (2SLM).

There are two equations and two unknown parameters, so we can numerically solve this optimization problem using for example a Newton-Raphson algorithm.

### 4.3.3 Residual Inclusion

Yet another way of obtaining our IV estimators is by residual inclusion [13]. This method follows a two stage procedure that is a bit different from the ones described earlier.

First stage: Regress $X$ on $Z$, obtaining a residual vector $e_x$.

Second stage: Regress $Y$ on $X$ and $e_X$

We can interpret this method as including the effect of $U$ on $X$ in our estimation. Even thought we did not measure $U$, the residuals of our first stage estimation are assumed to be $\epsilon_{xi} = \gamma_1 u_i + \eta_{xi}$ and so the expected value of this quantity $\mathbb{E}[\epsilon_x]$ is a linear combination of the unmeasured confounder $U$. ()

# 5 Causal Inference in a more general sense

When a researcher asks whether a certain treatment causes the condition of the patient to improve, he is asking a question about causality. Often these

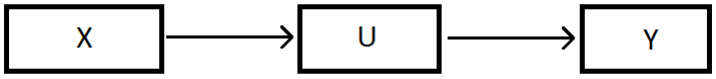



Figure 2: Here U is associated with X and Y but is no confounder.

type of questions seem deceptively simple, but are actually very hard to answer correctly. Just collecting data on the treatment and the outcome is often not enough. We need to know something about how these values of treatment and outcome interact with each other, this interaction between treatment and outcome is called the data-generating process (DGP). If we are to make any causal conclusions on this, merely studying the data by themselves is not enough. Behind every causal conclusions there lies a causal assumption that is not testable in observational studies. Statistical techniques such as Ordinary Least Squares merely evaluate the amount of association between two variables and not the causal path connecting these factors. [12] One way of separating causal from associative concepts is that associative relationships can be defined by a joint probability distribution of the observed variables, whereas a causal relationship requires additional assumptions to be defined.

Confounding is a causal concept, as it cannot be expressed by a joint distribution of observed variables. Some might think that when $U$ is dependent of $X$ and $U$ and $Y$ are dependent, than $U$ is a confounder. This is not correct however, as the causal path displayed below in Figure 5 has the same properties but $U$ is not a confounder in this case. There have to be additional assumptions about the direction of the arrows between variables to reach causal conclusions. Researchers often have to make these assumptions on a judgemental basis without any evidence.

## 5.1 The potential outcome framework

These causal relationships between observed variables require new mathematical notation. There is no general consensus yet about this notation, we will use the same notation as in [3]. Let $Y(x)$ be the potential outcome that would have been obtained if the treatment was set to $X = x$ by the researchers. Defining $Y(x)$ allows for some interesting quantities to be expressed, such as the Average Treatment Effect (ATE) which is defined as

$$ATE = \mathbb{E}\left[Y(1) - Y(0)\right]$$

For a binary outcome variable, this is equal to the marginal risk difference in the population:

$$marginalRD = P[Y(1) = 1] - P[Y(0) = 1]$$

This quantity is the expected difference in outcome between the treated and non-treated in the population. Usually it can only be reliably estimated in a randomized controlled trial, because we are concerned about confounding, i.e. violations of the exogeneity assumption, which is in the linear model equal to equation (5). We can use this potential outcome notation to write the definition of confounding in the potential outcome framework. If the treatment outcome relationship is confounded, the potential outcome Y(x) is not the same for treated or non-treated subjects. We have that

$$\mathbb{E}\left[Y(1)|X = 1\right] \neq \mathbb{E}\left[Y(1)|X = 0\right]$$

We can write our structural modelling equation (9) using the potential outcome framework. The first stage does not require to be a causal relationship between $Z$ and $X$ to remove the effect of the confounder on the parameter estimate. [3] This is because we are not interested in the question whether $Z$ is the cause of $X$ or the other way around, we only need to know the association between these two variables. In the potential outcome framework our model looks like this:

$$\mathbb{E}\left[Y(x)|U\right] = F(\beta_0 + x\beta_1 + U\gamma_2) \tag{10}$$

The most common choices for the distribution function F are the standard logistic and standard normal distributions.

The instrumental variable assumption stated in chapter 2, could also be stated in the potential outcomes framework. [3]

- Independence of potential outcomes and the instrument: Y(z,x) $\perp\!\!\!\perp$ Z
- $Z$ has no direct effect on Y. Y(z,x) = Y(x)
- Association between X and Z

The potential outcomes framework also allows for the estimation of the (causal) odds ratio's or risk differences of interest through 2SLS or GMM for example. The framework also allows the user to estimate causal effects of treatment regimes in randomized clinical trials affected by non-compliance, using "G-estimation". This method gives us access to quantities like the complier-specific causal effect.

# 6 Simulation

In this section we compare the described methods of Chapter 4 through simulations. We simulated a binary exposure $X$ and outcome $Y$, as well as a continuous instrument $Z$ and a confounder $U$. This is the same situation as we described in chapter 4.

Both $U$ and $Z$ were generated from a normal distribution with variance equal to 1 as follows:

$$U \sim N(0, 1)$$

$$Z \sim N(U * \delta, 1)$$

If $\delta = 0$, then $U$ and $Z$ are independent, which is one of the assumptions of IV analysis. By adjusting $\delta$ we can look at the consequences of violating this assumption.

The exposure and outcome are drawn from two Bernoulli distributions with parameters $p_1$ and $p_2$ respectively. For the treatment $p_1$ is a function of both the instrument and the confounder and for the outcome $p_2$ is a function of the treatment and the confounder.

$$X_i \sim Bernoulli(p_1 = \frac{exp(\alpha_0 + Z_i\alpha_1 + U_i\gamma_1)}{1 + exp(\alpha_0 + Z_i\alpha_1 + U_i\gamma_1)}) \qquad \forall i = 1, ..., n$$

$$Y_i \sim Bernoulli(p_2 = \frac{exp(\beta_0 + X_i\beta_1 + U_i\gamma_2)}{1 + exp(\beta_0 + X_i\beta_1 + U_i\gamma_2)}) \qquad \forall i = 1, ..., n$$

The parameters $\alpha, \beta_0, \beta_1, \gamma_1, \gamma_2, \delta$ are chosen to reflect specific scenarios a researcher may encounter in a real experiment. We investigated the effect of changes in the simulation parameters on the estimate and its standard deviation. Our selection of parameter settings for each scenario are displayed in table 1. The scenarios were chosen to resemble common practical issues, because our main focus is finding out whether these methods are useful in epidemiological practice. The first scenario serves as a sort of baseline, as the other scenario's are variations to the first one.

We designed the first scenario as follows:

1. – The exposure-outcome relation is confounded. The unmeasured confounder $U$ has a significant influence on both the treatment and outcome. $\gamma_1$ and $\gamma_2$ were both set to one.

- The IV assumptions are met; there is no correlation between the instrument $Z$ and $U$, $Z$ has no direct influence on $Y$ and lastly $Z$ is significantly correlated with $X$.
- The effect of the exposure is set at an odds-ratio of 2 ($b_1 = log(2)$). This means that when a patient is treated he or she is twice as likely to experience the outcome $Y = 1$ ceteris paribus, than if he is not treated.
- Prevalence of the outcome is around 50%.

Next, we varied the parameters as follows:

2. In the second scenario we reduced the prevalence of the outcome to around 1% by setting $\beta_0$ to -5 instead of 0.

3. The third scenario has a much higher prevalence in Y than in the first scenario. About 90% of the $Y$'s will have the value one.

4. In the fourth scenario we introduce a violation of the assumptions of IV analysis. We let $Z$ correlate with $U$ by setting $\delta = .75$.

5. In the fifth scenario the effect of the unmeasured confounder $U$ on the outcome $Y$ was reduced to $\gamma_2 = 0.1$. This will reduce the omitted variable bias in the OLS estimates and we can see whether the IV method still provides better estimates.

6. In the sixth scenario the correlation between the treatment $X$ and the instrument $Z$ was reduced. If there is no (or very small) correlation between $X$ and $Z$ one of our IV assumptions is violated.

7. In the seventh simulation we enhanced the relative magnitude of the confounding effect, by increasing both $\gamma_1$ and $\gamma_2$ to 2.

8. In the eighth scenario the odds-ratio of the treatment effect was increased to 4. This simulates a more impactful treatment. To keep the prevalence of $Y$ at about 50% the value of $\beta_0$ was decreased.

Table 1: Parameter settings for each scenario

| Scenario | $\alpha_0$ | $\alpha_1$ | $\beta_0$ | $\beta_1$ | $\gamma_1$ | $\gamma_2$ | $\delta$ | n | N |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 0 | 1.5 | 0 | log(2) | 1 | 1 | 0 | 10000 | 1000 |
| 2 | 0 | 1.5 | -5 | log(2) | 1 | 1 | 0 | 10000 | 1000 |
| 3 | 0 | 1.5 | 2 | log(2) | 1 | 1 | 0 | 10000 | 1000 |
| 4 | 0 | 1.5 | 2 | log(2) | 1 | 1 | .75 | 10000 | 1000 |
| 5 | 0 | 1.5 | 0 | log(2) | 1 | 0.1 | 0 | 10000 | 1000 |
| 6 | 0 | 0.5 | 0 | log(2) | 1 | 1 | 0 | 10000 | 1000 |
| 7 | 0 | 1.5 | 0 | log(2) | 2 | 2 | 0 | 10000 | 1000 |
| 8 | 0 | 1.5 | -3 | log(4) | 1 | 1 | 0 | 10000 | 1000 |

In each scenario we performed 1000 simulations of 10000 observations each. For each simulation, we estimated the target parameter $\beta_1$ with the methods described in chapter 3 and 4. Mean and standard deviation of the $N$ estimates were calculated. All simulations and estimation procedures were performed using the statistical package R.

The estimated parameters were compared to the theoretical quantities. The log odds ratio estimates were compared to $\beta_1$ and to the marginal log odds ratio, which is defined as

$$MarginallogOR = ln(\frac{P[Y(1) = 1]/P[Y(1) = 0)]}{P[Y(0) = 1]/P[Y(0) = 0]}),$$

with $Y(1)$ and $Y(0)$ the potential outcomes as defined in chapter 5.

We calculated $P[Y(1) = 1]$ and $P[Y(0) = 1]$, as follows. If we let everyone in the population have the exposure by setting $X = 1$ for each subject, denoted by Y(1), the probability that $Y(1) = 1$ is

$$P[Y(1) = 1] = \mathbb{E}[Y(1)] = \mathbb{E}[\mathbb{E}[Y(1)|U]] = \int_u \frac{exp(\beta_0 + \beta_1 + \gamma_2 u)}{1 + exp(\beta_0 + \beta_1 + \gamma_2 u)} f(u)du$$

It is difficult to analytically calculate this integral. One can approximate the integral by averaging out the confounder $U$. Therefore, the expression $P[Y(1) = 1]$ is estimated in each simulation by the proportion of subjects that would have $Y = 1$ when the exposure for each of the simulated subjects is set to 1.

$$P[Y(1) = 1] \approx \frac{1}{n} \sum_{i=1}^{n} \frac{exp(\beta_0 + \beta_1 + \gamma_2 u_i)}{1 + exp(\beta_0 + \beta_1 + \gamma_2 u_i)}$$

In each simulation we calculated the Marginal log odds ratio directly from the values of $U$ in the simulated population and the true values of the

parameters. Using this probability to calculate the Marginal log-Odds Ratio tells us what parameters a randomized controlled trial on our population would estimate. In a randomized control trial the researcher artificially sets the exposure value to either 0 or 1, so the $MarginallogOR$ is what would be the estimation target in the analysis of a randomized controlled trial on all simulated individuals. In the tables we report the average $MarginallogOR$ across all simulations.

The estimates of risk difference in the linear methods, were compared with the marginal risk difference, defined as

$$MarginalRD = P[Y(1) = 1] - P[Y(1) = 0]$$

Table 2: Point estimates of $\beta_1$ using the non-linear methods in various scenario's with their corresponding standard deviation between brackets. $\beta_1 = log(2) \approx 0.693$,except in Scenario 8 where $\beta_1 = log(4) \approx 1.38$

| Scenario | Logistic | 2SLM | GMM | 2SCM | Marginal log OR |
|---|---|---|---|---|---|
| 1 | 1.1261 (0.0422) | 0.5842 (0.0811) | 0.5861 (0.0831) | 0.5729 (0.0855) | 0.5748 |
| 2 | 1.3191 (0.1883) | 0.5608 (0.3091) | 0.5935 (0.3505) | 0.7275 (0.3306) | 0.6776 |
| 3 | 1.2169 (0.0666) | 0.5772 (0.1176) | 0.5903 (0.1250) | 0.6536 (0.1241) | 0.6223 |
| 4 | 1.5456 (0.0428) | 2.0851 (0.0663) | 2.3012 (0.0773) | 2.1264 (0.0684) | 0.5748 |
| 5 | 0.7550 (0.0407) | 0.6858 (0.0811) | 0.6903 (0.0841) | 0.6909 (0.0828) | 0.6915 |
| 6 | 1.2684 (0.0432) | 0.5806 (0.1980) | 0.5858 (0.2027) | 0.5582 (0.2170) | 0.5748 |
| 7 | 1.8707 (0.0441) | 0.4921 (0.0994) | 0.4923 (0.1014) | 0.3265 (0.1193) | 0.4210 |
| 8 | 1.8436 (0.0708) | 1.0599 (0.1095) | 1.1201 (0.1304) | 1.2891 (0.1239) | 1.2421 |

Table 3: Point estimates of Risk Differences in various scenario's with their corresponding standard deviation between brackets.

| Scenario | OLS | 2SLS | 3Stage | marginal RD |
|---|---|---|---|---|
| 1 | 0.2688 (0.0096) | 0.1427 (0.0200) | 0.1430 (0.0197) | 0.1399 |
| 2 | 0.0198 (0.0026) | 0.0096 (0.0053) | 0.0095 (0.0052) | 0.0102 |
| 3 | 0.1264 (0.0066) | 0.0650 (0.0134) | 0.0648 (0.0132) | 0.0655 |
| 4 | 0.3608 (0.0091) | 0.5069 (0.0136) | 0.4834 (0.0138) | 0.1399 |
| 5 | 0.1812 (0.0096) | 0.1659 (0.0198) | 0.1661 (0.0195) | 0.1663 |
| 6 | 0.3008 (0.0096) | 0.1424 (0.0486) | 0.1424 (0.0485) | 0.1399 |
| 7 | 0.4312 (0.0088) | 0.1212 (0.0247) | 0.1218 (0.0245) | 0.1037 |
| 8 | 0.2001 (0.0069) | 0.1340 (0.0141) | 0.1340 (0.0139) | 0.1357 |

## 6.1 Simulation Results

### 6.1.1 Log Odds Ratio estimates

In Table (2) the results for the methods which estimate an odds-ratio are reported. In the first scenario we see that the instrumental variable estimates are much closer to both the $MarginallogOR$ and $\beta_1$ than the standard logistic regression estimates. The standard deviations of the IV estimates are larger than those obtained from standard logistic regression. The IV methods seem to approximate the $MarginallogOR$ reasonably well in scenario 1, whereas the logistic regression estimate is completely off. Straightforward logistic regression also performs poorly compared to the IV methods in each of the other simulated scenario's as well, even when the confounding effect is smaller (scenario 5).

The standard deviations in scenario 2 and 3 are relatively large. They show the influence of having a very low or high prevalence of the outcome in the study sample. These results suggest that a more extreme prevalence leads to a higher variance of the treatment effect estimator.

While the estimates of linear regression models are not affected by the mean of the outcome, this is not true for models with a binary outcome. When dealing with rare event data, the probabilities of having an outcome $Y = 1$ estimated by logit models are too small, even with a sample size in the thousands. [11] This means that the variance of logistic regression estimates explodes when $P[Y_i = 1|x_i]$ is either very large or very small.

One can see this by deriving the expression of the variance of logistic regression estimates $\beta_{logit}$.

$$\text{Var}[\beta_{logit}] = -\mathbb{E}\left[\frac{\delta^2 \mathcal{L}(\beta)}{\delta^2 \beta}\right] = \sum_{i=1}^{n} \pi_i(1-\pi_i)x_i' x_i$$

where $\pi_i = P[Y_i = 1|x_i]$ and $\mathcal{L}(\beta) = \sum_{i:y=1} log(\pi_i) + \sum_{i:y=0} log(1-\pi_i)$ is the log-likelihood function of a logistic regression. This expression shows that if $\pi_i$ is close to zero or one, the variance of the estimators will be large. Scenario 2 and 3 show that the prevalence of the outcome in the sample is important for getting a good estimate of the effect of the treatment. Not only do the standard deviations blow up, the point estimate is also further away from the $MarginallogOR$ than in other scenarios.

In scenario 4 the estimates are very far from both the Marginal log OR and $\beta_1$. This result is explained by the correlation between $Z$ and $U$, which violates the IV assumptions as stated at the start of chapter 2. It displays the importance of this assumption in IV analysis.

Scenario 6 stands out with a high standard deviation for the GMM and

2SLM estimates. In this situation we simulated a weak instrument that has very low correlation with $X$. It seems that a weak correlation between $Z$ and $X$ inflates the variance of the estimators. This result is similar to the linear case as described in (3.3.2).

In scenario 7 we see that when the effect of the unmeasured confounder is large relative to the actual treatment effect, our IV estimates are more biased and the standard deviations increase. Prevalence is on average 50% here so this is most likely not the cause of the increase in standard deviation.

In scenario 8 we find that a larger treatment effect does not lead to very different results. As long as the prevalence of $Y$ is kept around 50% the standard deviations are of similar size to the standard deviations in the scenarios.

When comparing the IV methods with each other, we see that the GMM, 2SLM and 2SCM methods perform very similarly. The $marginallogOR$ is within 2 standard deviations of the mean estimate for each method in every scenario except scenario 4. The estimates are not close to the true parameter value $\beta_1$ however. This is because the model is not collapsible over the unmeasured confounder $U$, as it does not meet the necessary conditions for collapsibility [7] [6]. Another notable observation is that 2SCM overestimates the $MarginallogOR$ in scenario 2 and 3, which indicates that this method could suffer even more from either a very high or low prevalence.

### 6.1.2 Risk Difference estimates

The first thing noticed while looking at the table 3 is that, similar to the odds-ratio estimates in the previous section, the confounded relation between treatment and outcome in our data leads to biased estimates when estimating the treatment effect by using OLS. We see that the mean of the estimated risk difference using the different IV methods are very close to the marginal Risk Difference every time, except in scenario 4, where the instrument is correlated with the unmeasured confounding factor. Even though the data was simulated from non-linear probability models, the risk difference estimates still perform well. This indicates that the estimation methods are robust to violations of the model assumptions.

In scenario 2 and 3 we see the opposite pattern of increase in standard deviation as with the odds-ratio estimates in the same scenarios. The variance of the 2SLS estimator decreases as the variance of $Y$ decreases, which is exactly what happens in scenario 2 and 3. This is an interesting result, because this means that when the prevalence of the outcome $Y$ is either very high or low, the risk difference estimates are more precise, whereas for the odds ratio estimates the opposite is true. Thus, when the variation in $Y$ is

small, the variance of the risk difference estimates is also small.

$$\text{Var}[\beta_{OLS}] = \sigma_Y^2 (X'X)^{-1}$$

Similarly to the odds-ratio estimates, a small correlation between $X$ and $Z$ (scenario 6) increases the variance of the IV-estimators.

The 2SLS and 3Stage methods perform comparably with very minimal differences. It seems that model misspecification does not have a large influence on the risk difference estimates. Scenario 4 is again where the IV analysis goes wrong, as here the instrument is correlated with the unmeasured confounder. It is clear that violation of this assumption will have detremental effects on the validity of the inference.

# 7 Case Study

In this part of the thesis we apply the instrumental variable methods to real data with binary outcomes. We will use the methods discussed in this thesis to assess whether the administration of Prophylactic Corticosteroids before cardiac surgery has a beneficial effect on the patients recovery after surgery. This is done using data from an observational study in the LUMC with 476 patients. This part of the thesis can be viewed as an extension of the study done by Boef et al. [4], whom I worked with at my department.

When this study was carried out it was not known whether steroids should be administered before cardiac surgery. Some anaesthetists believe that there is a benefit to using steroids, when trying to reduce adverse outcomes.

Because of the observational nature of the data collected, instrumental variable analysis could be appropriate as we expect unmeasured confounding to be prevalent. It is quite likely that steroids are more or less often prescribed to more severe patients.

We would like to obtain estimates of risk difference and odds ratio between the patients with and without steroids for several binary outcomes. The main question of interest for us as statisticians is how the results of standard analysis methods (like OLS and logistic regression) differ from IV analyses. We compare the results of both ways of analysis to results of a recently published randomised clinical trial [5].

## 7.1 Data Summary

The data consist of $n = 476$ observations on medical records of patients that underwent cardiac surgery in Leiden University Medical Centre in 2005. Of these patients 115 received prophylactic corticosteroids in anticipation

of their surgery. The type of surgery varied from coronary artery bypass grafting to heart failure surgery (see Boef et al. for details). Patients were treated by ten different anaesthetists, who decided whether to administrate corticosteroids or not, with each anaesthetist being observed from 20 up to 91 times. Other covariates in the study are:

- EuroSCORE. This is a prognostic score on the mortality of the patient. A high EuroSCORE indicates a high risk of dying during hospitalization. Most patients in the study (75% of the study population) have a EuroSCORE lower than 9, with the highest EuroSCORE being 49 on one of the patients.

- BMI. This well known measure of relative weight, indicates a patients weight in relationship to the his or her height.

- Age. The age of the patient at the time of surgery. Age ranges from 16 to 88 year old patients, the distribution is skewed however, with a median of 67, indicating that most heart surgery is done on the elderly.

- Gender. The gender of the patient. There are twice as much males in the study, as men have cardiac problems more often than women.

- Diabetes. Whether the patient has diabetes or not.

Outcomes in this situation are manifold and can be measured in multiple ways. The main outcome variable considered here is thirty-day mortality. This is a binary variable set to 1 if the patient died within 30 days after heart surgery. A second outcome variable which we consider in this thesis is a binary variable indicating infection after surgery.

In Table 4, we describe the covariates measured at baseline separately for the patients who received steroids and who did not receive the steroids. We tested whether any of the covariates have a significant influence on the choice of administration of corticosteroids. For continuous variables we used a unpaired t-test, the categorical variable EuroSCORE was tested with a chi-squared test for independence and the binary variables were tested with a Z-test. This way we find out which of these variables are potential confounders.

Table 4: Comparison of the variables measured before surgery, between patients treated with steroids and those without steroids.

| | Prophylactic Corticosteroids | | |
|---|---|---|---|
| | Yes (n=113) | No (n=348) | P-values |
| | Covariates | | Student's t-Test |
| Age (Years) | 64.0 (13.0) | 64.7 (13.2) | 0.61 |
| BMI (kg/m$^2$) | 26.4 (4.2) | 26.6 (4.1) | 0.76 |
| | | | Z-test |
| Gender (Males) | 69 (61%) | 237 (68%) | 0.17 |
| Diabetes | 15 (13%) | 49 (14%) | 0.82 |
| EuroSCORE category | | | Chi-square contingency table test |
| 1-2% | 23 (20%) | 113 (33%) | |
| 3-5% | 35 (31%) | 106 (30%) | 0.03 |
| >6% | 55 (49%) | 129 (37%) | |

Summaries are given as means (standard deviation) or numbers (percentages)

From the table we can see that there is a significant difference in the administration of Corticosteroids for patients with different EuroSCORE levels. Patients in the high EuroSCORE category are the most likely to receive Prophylactic Corticosteroids before surgery. This is important to note when we examine the relation between steroids and outcomes as we have shown that the EuroSCORE is a likely confounder.

### 7.1.1 Different Anaesthetists

The anaesthetist is involved in the decision to administer steroids. Some anaesthetists might be more prone to prescribe steroids before surgery than others. This preference may provide us with an instrument to deal with confounding as described in chapter 3 and 4. Table 7.1.1 shows the prescription of steroids for each of the anaesthetists. The difference in anaesthetist prescription preference is tested for statistical significance by conducting a Chi-squared test of independence (p-value $<0.001$).

Table 5: Comparing the anaesthetists steroid prescription behaviour.

| Anesthesist | Prophylactic Corticosteroids | | Cases prescribed (in %) |
|---|---|---|---|
| | Yes (n=113) | No (n=348) | |
| 1 | 8 | 12 | 40% |
| 2 | 0 | 35 | 0% |
| 3 | 30 | 18 | 62.5% |
| 4 | 1 | 53 | 1.8% |
| 5 | 17 | 74 | 18.7% |
| 6 | 31 | 20 | 60.8% |
| 7 | 1 | 27 | 3.6% |
| 8 | 3 | 48 | 5.9% |
| 9 | 7 | 18 | 28% |
| 10 | 15 | 43 | 25.9% |

We see that there is a large difference between anaesthetists when it comes to prescribing steroids. Where some of the anaesthetists never prescribe steroids, some of them prescribe steroids very often. The anaesthetist preference therefore seems to be a good candidate instrument, as it significantly influences the choice of steroid administration. Because the correlation between these two variables is high, it already fulfils the third IV assumption mentioned in chapter 2.

## 7.2 Anaesthetists preference

It is difficult to assign a quantitative value to the preference of a anaesthetist, because it cannot be measured directly. One commonly used method to estimate the preference is to use the prescription for the last patient as the preference for the current one [13]. With our data we take it a step further and use all previous treatment decisions of the same anaesthetist to calculate his or her preference for prescribing steroids (as a percentage of total patients treated). For each patient we calculated the percentage of previous steroid prescriptions of his or her anaesthetist. This preference was used as the instrument in our instrumental variable analyses. This way of calculating the instrument makes the most use of the data available, concerning the anaesthetist's preference.

## 7.3 Instrumental Variable Assumptions

In order to use the anaesthetists preference as an instrument, it has to fulfil the assumptions stated in chapter 2.

The significance of the chi-squared test of independence in the previous section showed that the physician preference fulfils the third instrumental variable assumption in 2.

The other two assumptions are impossible to prove, because they are not testable using the data, but we can discuss whether they are plausible. The first assumption is that there is no correlation between our instrument $Z$ and unmeasured confounders. It may be that a certain anaesthetist treats patients that are more severely ill and prone to die within 30 days. However, these confounding pathways are unlikely because the anaesthetists work by a (semi-)randomized schedule in the LUMC. Because of this we don't expect confounding pathways to be present. We checked this further by testing whether any of the measured variables at baseline were correlated with our instrument. Results are shown in table 7.3 below. No statistically significant associations were found.

Table 6: Results of regressing different variables on the anaesthetist's preference.

| Covariate | Raw correlation | OLS estimate | P-value |
|---|---|---|---|
| EUROscore | 0.025 | 0.0008 | 0.59 |
| Age | -0.022 | -0.0004 | 0.63 |
| BMI | 0.036 | 0.002 | 0.49 |
| Diabetes | 0.039 | 0.028 | 0.40 |

The second assumption is that there can be no direct pathway between the instrument and the outcome. We are confident that there is no direct effect of administration of corticosteroids on other patients on the 30-day mortality of the current patient.

This assumption could be violated if anaesthetists who frequently prescribe steroids are also more inclined to make other treatment decisions. However, the clinical researchers involved in this project thought this was not likely.

There may still be a different unmeasured factor in the pathway of our instrument to the outcome. There is no way to exclude this possibility from the available data.

## 7.4 Estimation

We used the following methods discussed in chapter 4 to assess the effect of steroids on the two outcome measures. To estimate risk differences, we

used OLS without covariates, OLS with corrections for the (measured) confounders, 2SLS without covariates, 2SLS with correction for (measured) confounding and the 3-stage approach with and without covariates. To estimate the odds ratio's, we used logistic regression, 2 stage logistic regression, the generalized method of moments approach and the 2 stage logistic regression with control function, all of them with and without the measured confounders.

For the IV estimators we calculated the 95%-confidence intervals using bootstrap methods except for the 2SLS method, because the analytical derivation of the variance of those estimators is very difficult. For 2-stage logistic regression it would involve differentiating a linear approximation of a logistic likelihood function within another logistic likelihood function.

Bootstrapping involves resampling the available data with replacement to simulate multiple experiments. This way we can estimate the parameters of interest multiple times and if the bootstrap sampling procedure is consistent, we can deduce the distribution of this parameter in this way.[9] The GMM method adjusted for covariates was left out of the analysis due to time constraints.

## 7.5 Results

In the steroids group 4 out of 113 patients died within 30 days after surgery (3.5%) and in the non-steroids group 10 out of the 348 patients died (2.9%). These numbers would indicate that the patients using steroids have a slightly increased risk of dying, but this difference is not statistically significant at the 5%-significance level (p=0.72).

For the infection outcome, patients in the steroid group had an infection in 15 out of 113 cases (13.3%) and in the non-steroid group 52 out of 348 (14.9%) obtained an infection (p=0.662).

In the figure 7.5 below we report the results of the different estimation methods and their corresponding percentile confidence intervals. The first thing we notice is that none of the estimates excludes zero from it's 95%-confidence interval, meaning that these results are not significant on the 95% confidence level.

When comparing the ordinary methods for estimating risk difference and odds ratio's (OLS and Logistic Regression) to the IV methods, we find that the estimates are indeed different. The graphs show that the IV methods estimate a slightly larger effect of steroids on reducing both infection and 30-day mortality. This would suggest that there indeed is a small effect of steroids, that is beneficial to the patient. However we cannot be certain about this, because the variance of the IV estimators is too large for the

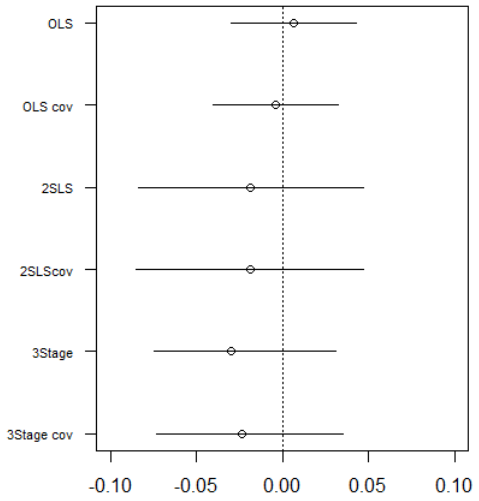
result to be statistically significant.

The bootstrapping posed problems for some of the non-linear models estimated. Because of the medium sample size and the low prevalence of the 30 day mortality outcome, bootstrapping this data 10.000 times sometimes lead to datasets that have complete separation between treated and non-treated patients. In this case the logistic regression would estimate $P[Y = 1|X = 1] \approx 0$, which is associated with an infinitely large negative value of $\beta$ in (9). This also leads to an incorrect estimate of the standard deviation of the parameter distribution. For the logistic regression, the OLS and the 2SLS methods the analytical variance of the parameter estimates is known, so we used those. The infection outcome did not pose this problem as there were no fully separated datasets in the bootstrap concerning these outcomes.

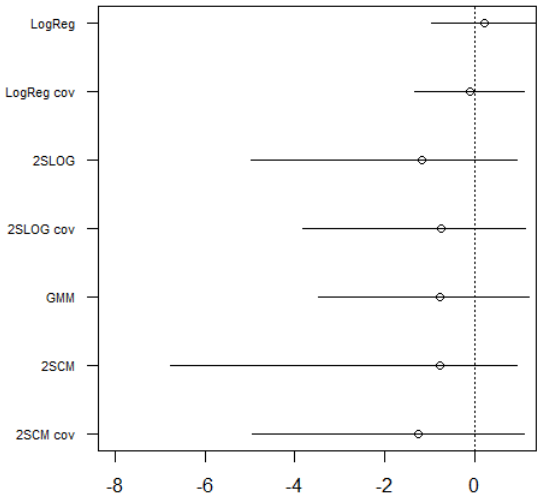
Comparing the OLS result without covariates to the OLS results that include the measured confounders in the analysis, we see that correcting for them switches the sign of the measured risk difference when considering 30 day mortality as the outcome. For the infection outcome there is a shift in the same direction. We actually see this type of shifting to larger negative risk differences or (log-)odds ratios for both outcomes and for all methods. This indicates omitted variable bias in the ordinary regression methods.

The 2SLS and 3Stage approach have similar results for both outcomes, with much larger confidence intervals than the non-IV methods.

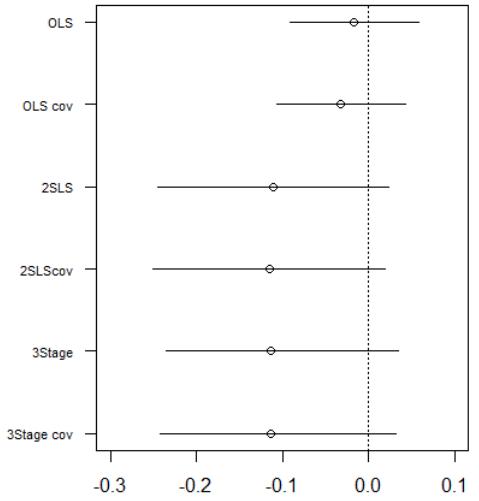
Finally we can see that looking at the results for 2SLS in the analysis with infection as the outcome, that the result is very close to being significant. It would be statistically significant on the 90% confidence level. This means that we can be around 80-90% sure that the true risk difference of using steroids vs not using steroids is non-zero and negative, meaning it benefits the patients, if we are willing to assume the assumptions stated in (2).



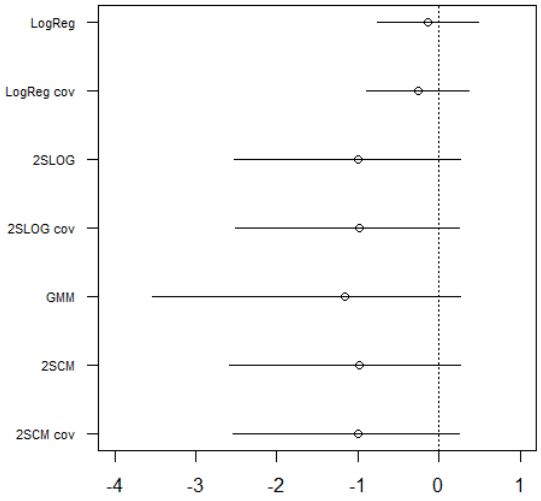

(a) Mortality. Risk Difference




(b) Mortality. log-Odds Ratio




(c) Infection. Risk Difference




(d) Infection. log-Odds Ratio

Figure 3: Comparison of the estimated effect of steroids on mortality and infections.

Table 7: Results of conventional and instrumental variable analysis. Outcome is 30 days mortality. Estimates of risk difference and odds ratios are given for the different analysis methods

| | Unadjusted | Adjusted[1] |
|---|---|---|
| **Risk difference** | | |
| Ordinary Least Squares | 0.007 (-0.030, 0.043) | -0.004 (-0.040, 0.032) |
| Two Stage Least Squares | -0.018 (-0.084, 0.047) | -0.019 (-0.085, 0.047) |
| 3 Stage IV Approach | -0.030 (-0.075, 0.031) | -0.023 (-0.073, 0.035) |
| **Odds Ratio** | | |
| Logistic Regression | 0.215 (-0.964, 1.395) | -0.127 (-1.342, 1.089) |
| Two Stage Logistic Regression | -1.171 (-4.976, 0.927) | -0.750 (-3.833, 1.110) |
| General Methods of Moments | -0.771 (-3.474, 1.189) | - |
| Control Function | -1.258 (-6.776, 0.930) | -0.778 (-4.948, 1.092) |

Table 8: Results of conventional and instrumental variable analysis. Outcome is infections . Estimates of risk difference and odds ratios are given for the different analysis methods

| | Unadjusted | Adjusted[1] |
|---|---|---|
| **Risk difference** | | |
| Ordinary Least Squares | -0.017 (-0.092, 0.058) | -0.032 (-0.107, 0.043) |
| Two Stage Least Squares | -0.111 (-0.245, 0.023) | -0.115 (-0.250, 0.020) |
| 3 Stage IV Approach | -0.114 (-0.236, 0.035) | -0.114 (-0.243, 0.033) |
| **Odds Ratio** | | |
| Logistic Regression | -0.138 (-0.756, 0.480) | -0.261 (-0.895, 0.373) |
| Two Stage Logistic Regression | -0.992 (-2.531, 0.263) | -0.977 (-2.509, 0.255) |
| General Methods of Moments | -1.154 (-3.529, 0.267) | - |
| Control Function | -0.994 (-2.581, 0.262) | -0.981 (-2.543, 0.251) |

## 7.6 Conclusion

Because we are looking for a small effect in a heavily confounded relationship, where millions of factors influence the event of dying in the 30 days after surgery, the non-significant results of a standard logistic regression analysis do not exclude the possibility of a positive effect of corticosteroids on survival or infection. However, the results of the instrumental variable analysis also did not show any statistically significant effects. Therefore these results

[1]Instrumental variable analysis adjusted for age, sex and EUROscore.

do not give us a clear answer to the question whether corticosteroids are beneficial to the patient or not. The difference between the treated and non-treated patients being not statistically significant on the 5% significance level was the same result as found in the randomized controlled trial conducted by Dieleman et al. [5] on the same subject. They found a small, not statistically significant reduction in 30 days mortality for steroids (relative risk 0.92 95% confidence interval (0.57-1.49)) comparing the use of corticosteroids to placebo. For infections Dieleman et al. did find a significant reduction for steriods users (relative risk 0.64 95% CI (0.54-0.75)). Our IV points estimates for infection were also in favour of steroids, but the confidence intervals were too wide to reject the null hypothesis of no effect of steroids.

Comparison our results to the randomized controlled trial of Dieleman et al. is not really possible because as the confidence intervals of our IV estimates are very wide, they contain both the value of the null hypothesis, but also contain the result of the Dieleman et al. trial. This means we can not really interpret our results in a meaningful way.

When looking at the difference between the statistical methods, we find that the results do display the bias-variance trade-off which the IV methods make compared to ordinary estimation methods. We see that variance increases for all of the IV methods. Even with an instrument as strong as ours (the treatment and physician preference are highly correlated). One can wonder whether this increase in variance is worth it, even if we believe that the IV assumptions are met. For every situation this question has a different answer, as it depends on the sample size and the strength of the instrument. A larger observational study on the effect of prophylactic corticosteroids reduces the variance of our estimates and could reveals the significance of this treatment.

# 8 Discussion

In this thesis we studied the effectiveness of instrumental variable methods for the estimation of risk differences and odds-ratios in the situation were both the exposure and the treatment are binary. By simulating multiple scenarios we were able to see how the estimators behave, without having to find an analytical solution to the bias of, for example, the two stage logistic regression estimator. We found that the (log-)odds ratio estimations never appeared to be asymptotically estimating $\beta_1$ unless the effect of the unmeasured confounding was very small. Estimates of the risk difference on the contrary behaved as we expected them to, even if the simulated model was not a risk difference model.

Woolridge [15] shows that the two stage logistic regression approach we use in (8) and (9) is flawed. Substituting fitted values of the endogenous variable $X$ inside the non-linear function in the second stage does not estimate $\beta_1$, because the conditional expectation operator does not pass through non-linear functions. He also shows that $\beta_1$ in these models is estimable, but require estimating the full likelihood of the structural equation models rather than mimicking the 2SLS approach.

Burgess [6] in turn shows that the estimand of two stage logistic regression is interpretable and possibly even more useful to the researcher than the value of $\beta_1$. The parameter $\beta_1$ isthe oddsratio, conditional on $U$, while the IV methods seem to estimate the Marginal (log-)Odds Ratio or the Population (log-)Odds Ratio. This oddsratio is comparable to the (log-)odds ratio of the control group against the treatment group in a RCT when the number of participants goes to infinity.

An important lesson to be learned here is that the researcher has to decide which quantities he or she is truly interested in. Depending on the answer to this, one can select the proper (instrumental variable) estimation method.

When the researcher is only interested in the difference between treated and non-treated patients, we recommend using 2SLS to estimate linear structural equation model, even when the outcome and treatment are both binary. While this model is "misspecified" it allows us to adequately estimate the risk difference between two groups while controlling for unmeasured confounding, given that the instrument is valid. Using the risk difference one can calculate quantities such as the Number Needed to Treat (NNT), which an estimate of how many individuals should be treated to prevent one adverse outcome.

An important question for users of Instrumental Variable methods is whether the reduction in bias is worth the increase in variance of the estimator. When many observations (10.000+) are available to estimate the effect of one treatment, using IV methods may be well worth the search for an instrument. On the other hand, it is impossible to be completely sure that a chosen instrument is valid, as well as that most clinical studies do not have this kind of magnitude in sample size. For medium sized studies (+/- 500 observations) the use of IV estimation may be a bad idea, leading the researcher to the wrong conclusions.

Instrumental variable estimation is originally a technique designed by economists, who have more easy access to large amounts of data, because the observations are not linked to treatments on individuals, which removes a lot of restrictions for the researcher.

Concluding, I would only suggest using IV methods when the sample size is large and in that case to just compare the treated to the non-treated using risk difference estimates from a 2SLS procedure.

# References


[1] M. Alan Brookhart, P. S. Wang, D. H. Solomon, and S. Schneeweiss. Evaluating short-term drug effects using a physician- specific prescribing preference as an instrumental variable. *Epidemiology*, 17(3):268–275, 2006.

[2] S. Burgess, A. Butterworth, A. Malarstig, and S. G. Thompson. Use of mendelian randomization to assess potential benefit of clinical intervention. *The BMJ*, 345, 2012.

[3] P. S. Clarke and F. Windmeijer. Instrumental variable estimators for binary outcomes. *Journal of the American Statistical Association*, 107(500):1638–1652, 2012.

[4] A. G. C. Boef et al. Physicians preference-based instrumental variable analysis. *Epidemiology*, 25(6):923–927, 2014.

[5] Dieleman et al. Intraoperative high-dose dexamethasone for cardiac surgery. *Journal of the American Medical Association*, 308(17):1761–1767, 2012.

[6] S. Burgess et al. Identifying the odds ratio estimated by a two-stage instrumental variable analysis with a logistic regression model. *Statistics in Medicine*, 32:4726–4747, 2013.

[7] J. Guo and Z. Geng. Collapsibility of logistic regression coefficients. *Journal of the Royal Statistical Society*, 57(1):263–267, 1994.

[8] C. Heij, P. de Boer, P. H. Franses, T. Kloek, and H. K. van Dijk. *Econometric Methods with Applications in Business and Economics*. Oxford University Press, 2004.

[9] J. L. Horowitz. *Handbook of Econometrics, Vol. 5*. Elsevier Science B.V., 2001.

[10] J. Hausman J. Hahn. Note on bias in estimators for simultaneous equation models. *Economics Letters*, 75:237–241, 2002.

[11] G. King and L. Zeng. Logistic regression in rare events data. *Policital Analysis*, 9(2):137–163, 2001.

[12] J. Pearl. Causal inference in statistics: An overview. *Statistics Surveys*, 3:96–146, 2009.

[13] J. A. Rassen, S. Schneeweiss, R. J. Glynn, M. A. Mittleman, and M. Alan Brookhart. Instrumental variable analysis for estimation of treatment effects with dichotomous outcomes. *American Journal of Epidemiology*, 169(3):273–284, 2009.

[14] J. H. Stock and F. Trebbi. Who invented instrumental variable regression? *Journal of Economic Perspectives*, 17(3):177–194, 2003.

[15] J. Wooldridge. *Econometric Analysis of Cross Section and Panel Data*. MIT Press, 2002.

# Appendices

## A R code Simulations

```
1  inverse_logit <- function(x) {
2    exp(x) / (1 + exp(x))
3  }
4
5
6  #Main simulation function. Returns b_0 and b_1 for each
       method
7  #and some  other parameters of interest.
8  Simulate <- function(a,b0,b1,c1,c2,cz,z,n,seed) {
9    i<<-i+1
10   cat("run numero",i)
11   set.seed(seed)
12
13
14
15   C <- rnorm(n,0,1)
16
17   Z <- rnorm(n,C*cz,1)
18
19   X <- rbinom(n,1,inverse_logit(a + Z*z + C*c1))
20
21   Y <- rbinom(n,1,inverse_logit(b0 + X*b1 + C*c2))
22
23   #P1
24   P_1 <- 1/n * sum( inverse_logit(b0 + b1 + C*c2))
25   #P0
26   P_0 <- 1/n * sum( inverse_logit(b0 + C*c2))
27   #PY
28   P_Y <- length(Y[Y==1]) / n
29   #Population Odds Ratio
30   ORLY <- log((P_1 / (1-P_1))/(P_0 / (1-P_0)))
31
32   #Least Squares
33   lm.mdl <- lm(Y~X)
34
```

```
35  beta.lm <- coef(lm.mdl)
36
37  #Logistic Regression
38  logreg <- glm(Y~X,binomial)
39
40  beta.logreg <- coef(logreg)
41
42  #Two-stage Least Squares
43  step.one <- lm(X ~ Z)
44  step.two <- lm(Y ~ fitted(step.one))
45
46  beta.2sls <- coef(step.two)
47
48  #Two-stage Least Squares Control Function (equivalent
        aan TSLS)
49  step.one <- lm(X ~ Z)
50  step.two <- lm(Y ~ X + residuals(step.one))
51
52  beta.2slsc <- coef(step.two)
53
54  #2 Stage Logistic Model
55  step.one <- glm(X ~ Z, binomial)
56  step.two <- glm(Y ~ fitted(step.one), binomial)
57
58  beta.2slog <- coef(step.two)
59
60  #Probit Structural Equation Models
61  #step.one <- glm(X ~ Z, binomial(link="probit"))
62  #step.two <- glm(Y ~ fitted(step.one), binomial(link
        ="probit"))
63
64  #beta.2sprob.1.6 <- coef(step.two) * 1.6
65
66  #3 Stage Least Squares Model
67  step.one <- glm(X ~ Z, binomial)
68  step.two <- lm(X ~ fitted(step.one))
69  step.three <- lm(Y ~ fitted(step.two))
70
71  beta.3slog <- coef(step.three)
72
73  #Control 2S Logistic Model
```

```
74  step.one <- glm(X ~ Z, binomial)
75  residuals <- X - fitted(step.one)
76
77  step.two <- glm(Y ~ X + residuals, binomial)
78
79  beta.2cslm <- coef(step.two)[1:2]
80
81
82  #GMM
83  J <- matrix(nrow=2,ncol=2)
84
85  MGF <- function(beta) {
86    f <- sum(Y - ( exp(beta[1] + X * beta[2]) / (1 +
        exp(beta[1] + X * beta[2])) ) )
87    g <- sum(Z * (Y - ( exp(beta[1] + X * beta[2]) / (1
         + exp(beta[1] + X * beta[2])) )))
88    return(c(f,g))
89  }
90
91  Jacobian <- function(beta) {
92
93    #f
94    J[1,1] <- -sum(exp(beta[1] + beta[2] * X) / (1 +
        exp(beta[1] + beta[2] * X))^2)
95    J[1,2] <- -sum(X * exp(beta[1] + beta[2] * X) / (1
        + exp(beta[1] + beta[2] * X))^2)
96    #g
97    J[2,1] <- -sum(Z * exp(beta[1] + beta[2] * X) / (1
        + exp(beta[1] + beta[2] * X))^2)
98    J[2,2] <- -sum(Z * X * exp(beta[1] + beta[2] * X) /
         (1 + exp(beta[1] + beta[2] * X))^2)
99    return(J)
100 }
101
102 beta0 <- beta.hat <- c(0,0)
103 i <- 0
104
105 while(eval > 0.00001) {
106   i <- i + 1
107   #Error handling
108   tester <- try(beta.hat <- beta0 - solve(Jacobian(
```

```
        beta0)) %*% MGF(beta0))
109     beta0 <- beta.hat
110     eval <- sum(abs(MGF(beta0)))
111
112     if(inherits(tester,"try-error")){
113       eval <- 0.00001
114       beta0 <- c(NA,NA)
115     }
116
117     if(is.na(eval)){
118       eval <- 0.00001
119       beta0 <- c(NA,NA)
120     }
121   }
122
123   beta.gmm <- beta0
124
125
126
127
128   #Risk difference Marginal approach
129   rd.marginal <- beta.2slog[2] * exp(beta.2slog[1] +
        beta.2slog[2]) / ((1 + exp(beta.2slog[1] + beta.2
        slog[2]))^2)
130
131   #True RD
132   #true.rd <- inverse_logit(b0 + b1) - inverse_logit(b0
        )
133   true.rd <- mean(inverse_logit(b0 + b1 + C*c2)) - mean
        (inverse_logit(b0 + C*c2))
134   sample.rd <- length(Y[X==1 & Y==1]) / length(Y[X==1])
        - length(Y[X==0 & Y==1]) / length(Y[X==0])
135
136   #True OR
137   true.or <- exp(b1)
138   sample.or <- (length(Y[X==1 & Y==1]) / length(Y[Y==0
        & X==1])) / (length(Y[X==0 & Y==1]) / length(Y[X
        ==0 & Y==0]))
139
140   output <- c(beta.lm,beta.2sls,beta.2slog,beta.3slog,
        beta.gmm,beta.2cslm,rd.marginal,true.rd,true.or,
```

```
        sample.rd,sample.or,seed,P_1,P_0,ORLY,beta.logreg)
141     names(output) <- c("Intercept (OLS)","Treatment (OLS)
        ","Intercept (2sls)","Treatment (2sls)","Intercept
        (2slog)","Treatment (2slog)","Intercept (3sls)","
        Treatment (3sls)","Intercept (GMM)","Treatment (
        GMM)","Intercept (2cslm)","Treatment (2cslm)","rd
        (marginal approach)","True RD","True OR","Sample
        RD","Sample OR","seed","p1","p0","ORLY","Intercept
        (logreg)","Treatment (logreg)")
142
143     return(output)
144
145   }
146
147   OUTPUT <- list()
148   parameters <- matrix(ncol=8,nrow=6)
149
150   N <- 100
151   n <- 1000
152
153   ##############################################
154   #Scenario 1
155   #-Veel confounding
156   #-Odds ratio = 2
157   #-Sterk instrument
158   #-Prevalentie 50%
159   #
160   # P(X=1) = inverse_logit(a + z * Z + c1 * U)
161   # P(Y=1) = inverse_logit(b0 + b1 * X + c2 * U)
162   a0 <- rep(0,N)
163   a1 <- rep(1.5,N)
164   b0 <- rep(0,N)
165   b1 <- rep(log(2),N)
166   c1 <- rep(1,N)
167   c2 <- rep(1,N)
168   cz <- rep(0,N)
169   seed <- seq(10,9+N,1)
170
171   i<-0
172
173   OUTPUT[[1]] <- mapply(Simulate,a=a0,b0=b0,b1=b1,c1=c1,
```

```
    c2=c2, cz=cz, z=a1, n=n, seed=seed)
parameters[,1] <- c(a0[1], a1[1], b0[1], b1[1], c1[1], c2
    [1])
##############################################
#Data set Scenario 1
a0=0
a1=5
b0=0
b1=log(2)
c1=1
c2=1

seed=123

C <- rnorm(n,0,1)
Z <- rnorm(n,0,1)
X <- rbinom(n,1,inverse_logit(a0 + Z*a1 + C*c1))
Y <- rbinom(n,1,inverse_logit(b0 + X*b1 + C*c2))

##############################################
#Scenario 2
#-Veel confounding
#-Odds ratio = 2
#-Sterk instrument
#-Prevalentie 10%
#
# P(X=1) = inverse_logit(a + z * Z + c1 * U)
# P(Y=1) = inverse_logit(b0 + b1 * X + c2 * U)
a0 <- rep(0,N)
a1 <- rep(1.5,N)
b0 <- rep(-5,N)
b1 <- rep(log(2),N)
c1 <- rep(1,N)
c2 <- rep(1,N)
cz <- rep(0,N)
seed <- seq(1223,1222+N,1)

i<-0

OUTPUT[[2]] <- mapply(Simulate, a=a0, b0=b0, b1=b1, c1=c1,
    c2=c2, cz=cz, z=a1, n=n, seed=seed)
```

```
212 parameters[,2] <- c(a0[1],a1[1],b0[1],b1[1],c1[1],c2
    [1])
213 ################
214 #Data set Scenario 2
215 a0=0
216 a1=1.5
217 b0=-5
218 b1=log(2)
219 c1=1
220 c2=1
221
222 seed=1024
223
224 C <- rnorm(n,0,1)
225 Z <- rnorm(n,0,1)
226 X <- rbinom(n,1,inverse_logit(a0 + Z*a1 + C*c1))
227 Y <- rbinom(n,1,inverse_logit(b0 + X*b1 + C*c2))
228
229 ###################################
230 #Scenario 3
231 #-Veel confounding
232 #-Odds ratio = 2
233 #-Sterk instrument
234 #-Prevalentie 80-90%
235 #
236 # P(X=1) = inverse_logit(a + z * Z + c1 * U)
237 # P(Y=1) = inverse_logit(b0 + b1 * X + c2 * U)
238 a0 <- rep(0,N)
239 a1 <- rep(1.5,N)
240 b0 <- rep(2,N)
241 b1 <- rep(log(2),N)
242 c1 <- rep(1,N)
243 c2 <- rep(1,N)
244 cz <- rep(0,N)
245 seed <- seq(1,N,1)
246
247 i<-0
248
249 OUTPUT[[3]] <- mapply(Simulate,a=a0,b0=b0,b1=b1,c1=c1,
    c2=c2,cz=cz,z=a1,n=n,seed=seed)
250 parameters[,3] <- c(a0[1],a1[1],b0[1],b1[1],c1[1],c2
```

```
    [1])
#####################
#Data set Scenario 3
a0=0
a1=1.5
b0=2
b1=log(2)
c1=1
c2=1

seed=12

C <- rnorm(n,0,1)
Z <- rnorm(n,0,1)
X <- rbinom(n,1,inverse_logit(a0 + Z*a1 + C*c1))
Y <- rbinom(n,1,inverse_logit(b0 + X*b1 + C*c2))

######################
#Scenario 4 Z U CORR
a0 <- rep(0,N)
a1 <- rep(1.5,N)
b0 <- rep(0,N)
b1 <- rep(log(2),N)
c1 <- rep(1,N)
c2 <- rep(1,N)
cz <- rep(.75,N)

seed <- seq(3,2+N,1)

i<-0

OUTPUT[[4]] <- mapply(Simulate,a=a0,b0=b0,b1=b1,c1=c1,
    c2=c2,cz=cz,z=a1,n=n,seed=seed)
parameters[,4] <- c(a0[1],a1[1],b0[1],b1[1],c1[1],c2
    [1])
#####################
#Data set Confounded Instrument (cor(z,c)!=0)
a0=0
a1=1.5
b0=2
b1=log(2)
```

```
289 c1=1
290 c2=1
291 cz=.75
292
293 seed=1
294
295 Z1 <- rnorm(n,0,1)
296 C <- rnorm(n,Z1*cz,1)
297 X <- rbinom(n,1,inverse_logit(a0 + Z1*a1 + C*c1))
298 Y <- rbinom(n,1,inverse_logit(b0 + X*b1 + C*c2))
299
300 step.one <- glm(X~Z1,binomial)
301 step.two <- glm(Y~fitted(step.one),binomial)
302 summary(step.two)
303
304 ####################################
305 #Scenario 5
306 #-Weinig confounding
307 #-Odds ratio = 2
308 #-Sterk instrument
309 #-Prevalentie 50%
310 #
311 # P(X=1) = inverse_logit(a + z * Z + c1 * U)
312 # P(Y=1) = inverse_logit(b0 + b1 * X + c2 * U)
313 a0 <- rep(0,N)
314 a1 <- rep(1.5,N)
315 b0 <- rep(0,N)
316 b1 <- rep(log(2),N)
317 c1 <- rep(1,N)
318 c2 <- rep(.1,N)
319 cz <- rep(0,N)
320
321 seed <- seq(1,N,1)
322
323 i<-0
324
325 OUTPUT[[5]] <- mapply(Simulate,a=a0,b0=b0,b1=b1,c1=c1,
        c2=c2,cz=cz,z=a1,n=n,seed=seed)
326 parameters[,5] <- c(a0[1],a1[1],b0[1],b1[1],c1[1],c2
        [1])
327 ####################
```

```
#Data set Scenario 5
a0=0
a1=1.5
b0=0
b1=log(2)
c1=1
c2=.1

seed=1

C <- rnorm(n,0,1)
Z <- rnorm(n,0,1)
X <- rbinom(n,1,inverse_logit(a0 + Z*a1 + C*c1))
Y <- rbinom(n,1,inverse_logit(b0 + X*b1 + C*c2))

#####################################
#Scenario 6
#-Veel confounding
#-Odds ratio = 2
#-Prevalentie 50%
#-Lage correlatie X en Z
#
# P(X=1) = inverse_logit(a + z * Z + c1 * U)
# P(Y=1) = inverse_logit(b0 + b1 * X + c2 * U)
a0 <- rep(0,N)
a1 <- rep(.5,N)
b0 <- rep(0,N)
b1 <- rep(log(2),N)
c1 <- rep(1,N)
c2 <- rep(1,N)
cz <- rep(0,N)

seed <- seq(1,N,1)

i<-0

OUTPUT[[6]] <- mapply(Simulate,a=a0,b0=b0,b1=b1,c1=c1,
    c2=c2,cz=cz,z=a1,n=n,seed=seed)
parameters[,6] <- c(a0[1],a1[1],b0[1],b1[1],c1[1],c2
    [1])
####################
```

```
367 #Data set Scenario 6
368 a0=1
369 a1=.5
370 b0=0
371 b1=log(2)
372 c1=1
373 c2=1
374
375 seed=244
376
377 C <- rnorm(n,0,1)
378 Z <- rnorm(n,0,1)
379 X <- rbinom(n,1,inverse_logit(a0 + Z*a1 + C*c1))
380 Y <- rbinom(n,1,inverse_logit(b0 + X*b1 + C*c2))
381
382 ####################################
383 #Scenario 7
384 #-Veel confounding
385 #-Odds ratio = 2
386 #-Prevalentie normaal
387 #-Sterk Instrument
388 #
389 # P(X=1) = inverse_logit(a + z * Z + c1 * U)
390 # P(Y=1) = inverse_logit(b0 + b1 * X + c2 * U)
391 a0 <- rep(0,N)
392 a1 <- rep(1.5,N)
393 b0 <- rep(0,N)
394 b1 <- rep(log(2),N)
395 c1 <- rep(2,N)
396 c2 <- rep(2,N)
397 cz <- rep(0,N)
398
399 seed <- seq(22355,22354+N,1)
400
401 i<-0
402
403 OUTPUT[[7]] <- mapply(Simulate,a=a0,b0=b0,b1=b1,c1=c1,
        c2=c2,cz=cz,z=a1,n=n,seed=seed)
404 parameters[,7] <- c(a0[1],a1[1],b0[1],b1[1],c1[1],c2
        [1])
405 ####################
```

```
406 #Data set Scenario 7
407
408 a0=0
409 a1=1.5
410 b0=0
411 b1=log(2)
412 c1=2
413 c2=2
414 seed=1
415
416 C <- rnorm(n,0,1)
417 Z <- rnorm(n,0,1)
418 X <- rbinom(n,1,inverse_logit(a0 + Z*a1 + C*c1))
419 Y <- rbinom(n,1,inverse_logit(b0 + X*b1 + C*c2))
420
421 ####################################
422 #Scenario 8
423 #-Veel confounding
424 #-Odds ratio = 2
425 #-Prevalentie normaal
426 #-Ander Treatment effect
427 #
428 # P(X=1) = inverse_logit(a + z * Z + c1 * U)
429 # P(Y=1) = inverse_logit(b0 + b1 * X + c2 * U)
430 a0 <- rep(0,N)
431 a1 <- rep(1.5,N)
432 b0 <- rep(-3,N)
433 b1 <- rep(log(4),N)
434 c1 <- rep(1,N)
435 c2 <- rep(1,N)
436 cz <- rep(0,N)
437
438 seed <- seq(1,N,1)
439
440 i<-0
441
442 OUTPUT[[8]] <- mapply(Simulate,a=a0,b0=b0,b1=b1,c1=c1,
        c2=c2,cz=cz,z=a1,n=n,seed=seed)
443 parameters[,8] <- c(a0[1],a1[1],b0[1],b1[1],c1[1],c2
        [1])
444 ####################
```

```
#Data set Scenario 8
a0=0
a1=1.5
b0=-3
b1=log(4)
c1=1
c2=1
seed=1

C <- rnorm(n,0,1)
Z <- rnorm(n,0,1)
X <- rbinom(n,1,inverse_logit(a0 + Z*a1 + C*c1))
Y <- rbinom(n,1,inverse_logit(b0 + X*b1 + C*c2))
```

## A.1 Organizing Simulation Results

```
#Functions to sort output in OR's and RD's
OR <- function(OUTPUT) {
  or <- rbind(OUTPUT[5:6,],OUTPUT[9:12,],OUTPUT[21,],
     OUTPUT[23,])
  or <- or[-c(1,3,5),]
  return(or)
}
RD <- function(OUTPUT) {
  rd <- rbind(OUTPUT[1:4,],OUTPUT[7:8,],(OUTPUT[19,]-
     OUTPUT[20,]),OUTPUT[13,])
  rd <- rd[-c(1,3,5),]
  return(rd)
}

#Functions to calculate our the sd's of our estimates
ErrorEstimate <- function(OUTPUT) {
  means <- rowMeans(OUTPUT,na.rm=T)
  sd <- sqrt(rowSums((OUTPUT-means)^2) / (dim(OUTPUT)
     [2] - 1))
  return(sd)
}

a <- matrix(nrow=8,ncol=1)
b <- matrix(nrow=8,ncol=1)
parameters <- t(parameters)
```

```
23 parameterz <- matrix(nrow=8,ncol=1)
24
25 for (i in 1:8) {
26   rd.output <- RD(OUTPUT[[i]])
27   rd.output <- t(na.omit(t(rd.output)))
28   or.output <- OR(OUTPUT[[i]])
29   or.output <- t(na.omit(t(or.output)))
30   rd.means  <- rowMeans(rd.output,na.rm=T)
31   rd.sds    <- ErrorEstimate(rd.output)
32   or.means  <- rowMeans(or.output,na.rm=T)
33   or.sds    <- ErrorEstimate(or.output)
34
35   or.means <- round(or.means,digits=4)
36   or.sds <- round(or.sds,digits=4)
37   rd.means <- round(rd.means,digits=4)
38   rd.sds <- round(rd.sds,digits=4)
39
40
41   a[i,] <- paste("Scenario",i,"&",or.means[5],or.sds
         [5],"&",or.means[1],or.sds[1],"&",or.means[2],or.
         sds[2],"&",or.means[3],or.sds[3],"&",or.means[4])
42   b[i,] <- paste("Scenario",i,"&",rd.means[1],rd.sds
         [1],"&",rd.means[2],rd.sds[2],"&",rd.means[3],rd.
         sds[3],"&",rd.means[5],rd.sds[5],"&",rd.means[4])
43   parameters[i,] <- paste("Scenario",i,"&",parameters[i
         ,1],"&",parameters[i,2],"&",parameters[i,3],"&",
         parameters[i,4],"&",parameters[i,5],"&",parameters
         [i,6],"&",n,"&",N)
44 }
45
46 a
47 b
48 parameterz
```

# B R code Case Study

```
1 require(nnet)
2 require(foreign)
3
4 setwd("C:/Users/hlgrondijs/Downloads/Scriptie Part 2")
5
```

```
6  sdat <- read.dta("Data Steroids Hendrik stata 12.dta",
       convert.dates = TRUE, convert.factors = TRUE,
7                   missing.type = FALSE,
8                   convert.underscore = FALSE, warn.
                        missing.labels = TRUE)
9
10 sdat$steroids <- as.numeric(sdat$steroids)-1
11 sdat$low <- sdat$euroscore_cat == 1
12 sdat$med <- sdat$euroscore_cat == 2
13 sdat$high <- sdat$euroscore_cat == 3
14
15 sdat$mort_30 <- as.numeric(sdat$mort_30)-1
16 sdat$infection <- as.numeric(sdat$infection)-1
17
18 #Data summary table
19 #30 day Mortality
20 table(sdat$mort_30[sdat$steroids==0])
21 table(sdat$mort_30[sdat$steroids==1])
22
23 summary(glm(mort_30~steroids,family=binomial,data=sdat)
       )
24
25 #Infection
26 table(sdat$infection[sdat$steroids==0])
27 table(sdat$infection[sdat$steroids==1])
28
29 summary(glm(infection~steroids,family=binomial,data=
       sdat))
30
31 #EuroSCORe
32 table(sdat$euroscore_cat[sdat$steroids==0])
33 table(sdat$euroscore_cat[sdat$steroids==1])
34
35 chisq.test(table(sdat$euroscore_cat,sdat$steroids))
36 summary(glm(steroids ~ low + med,data=sdat,family=
       binomial))
37
38 #Diabetes
39 table(sdat$Diabetes[sdat$steroids==1])
40 table(sdat$Diabetes[sdat$steroids==0])
41
```

```
42 #Gender
43 table(sdat$gender[sdat$steroids==0])
44 table(sdat$gender[sdat$steroids==1])
45
46 summary(glm(gender~steroids,family=binomial,data=sdat))
47
48 #BMI
49 sd(sdat$bmi[sdat$steroids==0],na.rm=T)
50 sd(sdat$bmi[sdat$steroids==1],na.rm=T)
51 mean(sdat$bmi[sdat$steroids==0],na.rm=T)
52 mean(sdat$bmi[sdat$steroids==1],na.rm=T)
53
54 t.test(x=sdat$bmi[sdat$steroids==1],y=sdat$bmi[sdat$
       steroids==0])
55
56 #Age
57 sd(sdat$age[sdat$steroids==0])
58 mean(sdat$age[sdat$steroids==0])
59 sd(sdat$age[sdat$steroids==1])
60 mean(sdat$age[sdat$steroids==1])
61
62 t.test(x=sdat$age[sdat$steroids==1],y=sdat$age[sdat$
       steroids==0])
63
64 #Anesthesist
65 table(sdat$anesthesist[sdat$steroids==0])
66 table(sdat$anesthesist[sdat$steroids==1])
67
68 tabel<-c(table(sdat$anesthesist[sdat$steroids==1])
       [1],0,table(sdat$anesthesist[sdat$steroids==1])
       [2:9])
69
70 contingency <- cbind(tabel,table(sdat$anesthesist[sdat$
       steroids==0]))
71
72 chisq.test(contingency)
73
74 #Calculate physician preference
75 instrument <- numeric(length(sdat$steroids))
76 pos <- 0
77 for(i in 1:10){
```

```
78    for(j in 1:length(sdat$anesthesist[sdat$anesthesist==
          i])){
79      pos <- pos+1
80      instrument[pos] <- sum(sdat$steroids[sdat$
            anesthesist==i][1:j]) / j
81    }
82  }
83
84  #Preliminary regressions
85  #OLS
86  olsmdl <- lm(mort_30~steroids+age+gender+low+med+bmi,
        data=sdat)
87  olsmdl2 <- lm(infection~steroids,data=sdat)
88
89  #2sls
90  first.stage <- lm(steroids~instrument,data=sdat)
91  second.stage <- lm(mort_30 ~fitted(first.stage)+age+
        gender+low+med+bmi,data=sdat)
92  second.stage2 <- lm(infection~fitted(first.stage)+age+
        gender+low+med+bmi,data=sdat)
93  summary(second.stage)
94  summary(second.stage2)
95
96  #logistic regression
97  logreg <- glm(mort_30~steroids+age+gender+low+med+bmi,
        family=binomial,data=sdat)
98  summary(logreg)
99
100 Y.original <- sdat$mort_30
101 X.original <- sdat$steroids
102 Z.original <- instrument
103
104 Y <- Y.original
105 X <- X.original
106 Z <- Z.original
107
108 #testing assumptions
109 cor(Z,sdat$euroscore)
110 cor(Z,sdat$age)
111 cor(Z,sdat$bmi,use="complete.obs")
112 cor(Z,sdat$Diabetes,use="complete.obs")
```

```
113
114
115 #Analyze-function computing all 8 methods and returning
        b_1 for each
116
117 analyze <- function(X,Y,Z) {
118   #Least Squares
119   lm.mdl <- lm(Y~X)
120
121   beta.lm <- coef(lm.mdl)
122
123   #Logistic Regression
124   logreg <- glm(Y~X, binomial)
125
126   beta.logreg <- coef(logreg)
127
128   #Two-stage Least Squares
129   step.one <- lm(X ~ Z)
130   step.two <- lm(Y ~ fitted(step.one))
131
132   beta.2sls <- coef(step.two)
133
134   #Two-stage Least Squares Control Function (equivalent
        aan TSLS)
135   step.one <- lm(X ~ Z)
136   step.two <- lm(Y ~ X + residuals(step.one))
137
138   beta.2slsc <- coef(step.two)
139
140   #2 Stage Logistic Model
141   step.one <- glm(X ~ Z, binomial)
142   step.two <- glm(Y ~ fitted(step.one), binomial)
143
144   beta.2slog <- coef(step.two)
145
146   #Probit Structural Equation Models
147   #step.one <- glm(X ~ Z, binomial(link="probit"))
148   #step.two <- glm(Y ~ fitted(step.one), binomial(link
        ="probit"))
149
150   #beta.2sprob.1.6 <- coef(step.two) * 1.6
```

```
151
152   #3 Stage Model
153   step.one <- glm(X ~ Z, binomial)
154   step.two <- lm(X ~ fitted(step.one))
155   step.three <- lm(Y ~ fitted(step.two))
156
157   beta.3s <- coef(step.three)
158
159   #Control 2S Logistic Model
160   step.one <- glm(X ~ Z, binomial)
161   residuals <- X - fitted(step.one)
162
163   step.two <- glm(Y ~ X + residuals, binomial)
164
165   beta.2cslm <- coef(step.two)[1:2]
166
167   #GMM
168   J <- matrix(nrow=2,ncol=2)
169
170   MGF <- function(beta) {
171     f <- sum(Y - ( exp(beta[1] + X * beta[2]) / (1 +
          exp(beta[1] + X * beta[2])) ) )
172     g <- sum(Z * (Y - ( exp(beta[1] + X * beta[2]) / (1
           + exp(beta[1] + X * beta[2])) )))
173     return(c(f,g))
174   }
175
176   Jacobian <- function(beta) {
177     #f
178     J[1,1] <- -sum(exp(beta[1] + beta[2] * X) / (1 +
          exp(beta[1] + beta[2] * X))^2)
179     J[1,2] <- -sum(X * exp(beta[1] + beta[2] * X) / (1
          + exp(beta[1] + beta[2] * X))^2)
180     #g
181     J[2,1] <- -sum(Z * exp(beta[1] + beta[2] * X) / (1
          + exp(beta[1] + beta[2] * X))^2)
182     J[2,2] <- -sum(Z * X * exp(beta[1] + beta[2] * X) /
           (1 + exp(beta[1] + beta[2] * X))^2)
183     return(J)
184   }
185
```

```
186     beta0 <- beta.hat <- c(0,0)
187     eval <- sum(abs(MGF(beta0)))
188     i <- 0
189
190     while(eval > 0.00001) {
191       i <- i + 1
192       #Error handling when algorithm does not converge.
193       tester <- try(beta.hat <- beta0 - solve(Jacobian(
              beta0)) %*% MGF(beta0))
194       beta0 <- beta.hat
195       eval <- sum(abs(MGF(beta0)))
196
197       if(inherits(tester,"try-error")){
198         eval <- 0.00001
199         print(beta0)
200         beta0 <- c(NA,NA)
201      }
202     }
203
204     beta.gmm <- beta0
205
206
207
208     #Risk differnce marginal approach
209     rd.marginal <- beta.2slog[2] * exp(beta.2slog[1] +
          beta.2slog[2]) / ((1 + exp(beta.2slog[1] + beta.2
          slog[2]))^2)
210
211
212     output <- c(beta.lm[2],beta.2sls[2],beta.3s[2],rd.
          marginal,beta.logreg[2],beta.2slog[2],beta.gmm[2],
          beta.2cslm[2])
213     names(output) <- c("OLS","2SLS","3 Stage","Marginal
          Approach","LogReg","2SLOG","GMM","2SCM")
214     return(output)
215   }
216
217   #Analyze-function including measured confounders and/or
          covariates.
218
219   analyze.covariates <- function(X,Y,Z,data) {
```

```
220  #Least Squares
221  lm.mdl <- lm(Y~X+age+gender+low+med,data=data)
222
223  beta.lm <- coef(lm.mdl)
224
225  #Logistic Regression
226  logreg <- glm(Y~X+age+gender+low+med,data=data,
         binomial)
227
228  beta.logreg <- coef(logreg)
229
230  #Two-stage Least Squares
231  step.one <- lm(X ~ Z+age+gender+low+med,data=data)
232  step.two <- lm(Y ~ fitted(step.one)+age+gender+low+
         med,data=data)
233
234  beta.2sls <- coef(step.two)
235
236  #Two-stage Least Squares Control Function (equivalent
         aan TSLS)
237  step.one <- lm(X ~ Z+age+gender+low+med,data=data)
238  step.two <- lm(Y ~ X + residuals(step.one)+age+gender
         +low+med,data=data)
239
240  beta.2slsc <- coef(step.two)
241
242  #2 Stage Logistic Model
243  step.one <- glm(X ~ Z+age+gender+low+med,data=data,
         binomial)
244  step.two <- glm(Y ~ fitted(step.one)+age+gender+low+
         med,data=data, binomial)
245
246  beta.2slog <- coef(step.two)
247
248  #Probit Structural Equation Models
249  #step.one <- glm(X ~ Z, binomial(link="probit"))
250  #step.two <- glm(Y ~ fitted(step.one), binomial(link
         ="probit"))
251
252  #beta.2sprob.1.6 <- coef(step.two) * 1.6
253
```

```
254  #3 Stage Model
255  step.one <- glm(X ~ Z+age+gender+low+med,data=data,
         binomial)
256  step.two <- lm(X ~ fitted(step.one)+age+gender+low+
         med,data=data)
257  step.three <- lm(Y ~ fitted(step.two)+age+gender+low+
         med,data=data)
258
259  beta.3s <- coef(step.three)
260
261  #Control 2S Logistic Model
262  step.one <- glm(X ~ Z+age+gender+low+med,data=data,
         binomial)
263  residuals <- X - fitted(step.one)
264
265  step.two <- glm(Y ~ X + residuals+age+gender+low+med,
         data=data, binomial)
266
267  beta.2cslm <- coef(step.two)[1:2]
268
269  #GMM
270  J <- matrix(nrow=2,ncol=2)
271
272  MGF <- function(beta) {
273    f <- sum(Y - ( exp(beta[1] + X * beta[2]) / (1 +
           exp(beta[1] + X * beta[2])) ) )
274    g <- sum(Z * (Y - ( exp(beta[1] + X * beta[2]) / (1
           + exp(beta[1] + X * beta[2])) )))
275    return(c(f,g))
276  }
277
278  Jacobian <- function(beta) {
279    #f
280    J[1,1] <- -sum(exp(beta[1] + beta[2] * X) / (1 +
           exp(beta[1] + beta[2] * X))^2)
281    J[1,2] <- -sum(X * exp(beta[1] + beta[2] * X) / (1
           + exp(beta[1] + beta[2] * X))^2)
282    #g
283    J[2,1] <- -sum(Z * exp(beta[1] + beta[2] * X) / (1
           + exp(beta[1] + beta[2] * X))^2)
284    J[2,2] <- -sum(Z * X * exp(beta[1] + beta[2] * X) /
```

```
            (1 + exp(beta[1] + beta[2] * X))^2)
285       return(J)
286     }
287
288     beta0 <- beta.hat <- c(0,0)
289     eval <- sum(abs(MGF(beta0)))
290     i <- 0
291
292     while(eval > 0.00001) {
293       i <- i + 1
294       #Error handling when algorithm does not converge.
295       tester <- try(beta.hat <- beta0 - solve(Jacobian(
              beta0)) %*% MGF(beta0))
296       beta0 <- beta.hat
297       eval <- sum(abs(MGF(beta0)))
298
299       if(inherits(tester,"try-error")){
300         eval <- 0.00001
301         print(beta0)
302         beta0 <- c(NA,NA)
303       }
304     }
305
306     beta.gmm <- beta0
307
308
309
310     #Risk differnce marginal approach
311     rd.marginal <- beta.2slog[2] * exp(beta.2slog[1] +
          beta.2slog[2]) / ((1 + exp(beta.2slog[1] + beta.2
          slog[2]))^2)
312
313
314     output <- c(beta.lm[2],beta.2sls[2],beta.3s[2],rd.
          marginal,beta.logreg[2],beta.2slog[2],beta.gmm[2],
          beta.2cslm[2])
315     names(output) <- c("OLS","2SLS","3 Stage","Marginal
          Approach","LogReg","2SLOG","GMM","2SCM")
316     return(output)
317   }
318
```

```
319 #Analyze original data
320
321 results <- analyze(X,Y,Z)
322 results.infection <- analyze(X,sdat$infection,Z)
323
324 results.covariates <- analyze.covariates(X,Y,Z,sdat)
325 results.covariates.infection <- analyze.covariates(X,
      sdat$infection,Z,sdat)
326
327
328 #create bootstrapdatasets
329 N <- length(sdat$steroids)
330 case <- list()
331 bigN <- 10000
332
333 for(i in 1:bigN){
334   temp <- numeric(N)
335   #Setting random seeds to be able to replicate the
        same "randomized" data later.
336   set.seed(i)
337   for(j in 1:N){
338     #Resampling with replacement
339     temp[j] <- sample(1:461,1)
340   }
341   case[[i]] <- temp
342
343   #perform analysis with data = sdat[case[[i]],] with i
        =1...bigN to compute standard errors of b_IV
344
345 }
346
347 #calculate bootstrap estimates
348 bootstrap.estimates <- matrix(nrow=8,ncol=bigN,0)
349 bootstrap.estimates2 <- matrix(nrow=8,ncol=bigN,0)
350 bootstrap.estimates.infection <- matrix(nrow=8,ncol=
      bigN,0)
351 bootstrap.estimates.infection2 <- matrix(nrow=8,ncol=
      bigN,0)
352 for(K in 1:bigN) {
353   b.dat <- sdat[case[[K]],]
354   #Order by anesthesist and then by date to allow
```

```
        calculation of instrument
355     b.dat <- b.dat[order(b.dat$anesthesist,b.dat$date_
            surgery),]
356     #30day mortality as main outcome
357     Y <- b.dat[,7]
358     #Infection as secondary outcome
359     Y2 <- b.dat[,6]
360     #Steroids as treatment
361     X <- b.dat[,10]
362     #Calculate instrument for dataset K
363     Z <- numeric(dim(b.dat)[1])
364     pos <- 0
365     for(i in 1:10){
366       for(j in 1:length(b.dat$anesthesist[b.dat$
              anesthesist==i])){
367         pos <- pos+1
368         Z[pos] <- sum(b.dat$steroids[b.dat$anesthesist==i
                ][1:j]) / j
369       }
370     }
371     print(K)
372     #analyze data with the resampled dataset and store
            estimates of b_1
373     bootstrap.estimates[,K] <- analyze(X,Y,Z)
374     bootstrap.estimates2[,K] <- analyze.covariates(X,Y,Z,
            b.dat)
375
376     bootstrap.estimates.infection[,K] <- analyze(X,Y2,Z)
377     bootstrap.estimates.infection2[,K] <- analyze.
            covariates(X,Y2,Z,b.dat)
378 }
379
380 #GMM correctie. Mislukte GMM procedures als NA invoeren
        .
381 gmm.correctie <- bootstrap.estimates[7,]==0
382  bootstrap.estimates[7,gmm.correctie] <-NA
383 gmm.correctie <- bootstrap.estimates2[7,]==0
384  bootstrap.estimates2[7,gmm.correctie] <- NA
385 gmm.correctie <- bootstrap.estimates.infection[,K]==0
386  bootstrap.estimates.infection[7,gmm.correctie] <- NA
387 gmm.correctie <- bootstrap.estimates.infection2[,K]==0
```

```
388 bootstrap.estimates.infection2[7,gmm.correctie] <- NA
389
390 #95-Percentile confidence interval calculation function
391 Percentile <- function(estimates){
392   quant <- matrix(nrow=8,ncol=2,0)
393   for(i in 1:8) {
394     quant[i,] <- quantile(estimates[i,],probs=c
          (0.025,0.975),na.rm=T)
395   }
396   return(quant)
397 }
398
399 #90Percentile confidence interval function
400 Percentile.90 <- function(estimates){
401   quant <- matrix(nrow=8,ncol=2,0)
402   for(i in 1:8) {
403     quant[i,] <- quantile(estimates[i,],probs=c
          (0.05,0.95),na.rm=T)
404   }
405   return(quant)
406 }
407
408 #Calculate 95 and 90 percentiles
409 Q <- Percentile(bootstrap.estimates)
410 Q.90 <- Percentile.90(bootstrap.estimates)
411
412 Q.cov <- Percentile(bootstrap.estimates2)
413 Q.cov.90 <- Percentile.90(bootstrap.estimates2)
414
415 Q.infection <- Percentile(bootstrap.estimates.infection
      )
416 Q.infection.90 <- Percentile.90(bootstrap.estimates.
      infection)
417
418 Q.infection.cov <- Percentile(bootstrap.estimates.
      infection2)
419 Q.infection.cov.90 <- Percentile.90(bootstrap.estimates
      .infection2)
420
421 #Plot functions 2
422 plot.rd <- function(results,Q,results.cov,Q.cov,xmin,
```

```
    xmax){
423   plot(cbind(c(results[c("OLS","2SLS","3 Stage")],
        results.cov[c("OLS","2SLS","3 Stage")]),c
        (6,4,2,5,3,1)),xlim=c(xmin,xmax),yaxt="n",xlab="",
        ylab="")
424   axis(cex.axis=.7,labels=c("OLS","OLS cov","2SLS","2
        SLScov","3Stage","3Stage cov"),side=2,at=c
        (6,5,4,3,2,1),las=1)
425   lines(c(Q[1,1],Q[1,2]),c(6,6))
426   lines(c(Q.cov[1,1],Q.cov[1,2]),c(5,5))
427   lines(c(Q[2,1],Q[2,2]),c(4,4))
428   lines(c(Q.cov[2,1],Q.cov[2,2]),c(3,3))
429   lines(c(Q[3,1],Q[3,2]),c(2,2))
430   lines(c(Q.cov[3,1],Q.cov[3,2]),c(1,1))
431   abline(v=0,lty=3)
432 }
433
434 plot.OR <- function(results,Q,results.cov,Q.cov,xmin,
    xmax){
435   plot(cbind(c(results[c("LogReg","2SLOG","GMM","2SCM")
        ],results.cov[c("LogReg","2SLOG","2SCM")]),c
        (7,5,3,1,6,4,2)),xlim=c(xmin,xmax),yaxt="n",xlab="
        ",ylab="")
436   axis(cex.axis=.7,labels=c("LogReg","LogReg cov","2
        SLOG","2SLOG cov","GMM","2SCM","2SCM cov"),side=2,
        at=c(7,6,5,4,3,2,1),las=1)
437   lines(c(Q[5,1],Q[5,2]),c(7,7))
438   lines(c(Q.cov[5,1],Q.cov[5,2]),c(6,6))
439   lines(c(Q[6,1],Q[6,2]),c(5,5))
440   lines(c(Q.cov[6,1],Q.cov[6,2]),c(4,4))
441   lines(c(Q[7,1],Q[7,2]),c(3,3))
442   #lines(c(Q.cov[7,1],Q.cov[7,2]),c(3,3))
443   lines(c(Q[8,1],Q[8,2]),c(2,2))
444   lines(c(Q.cov[8,1],Q.cov[8,2]),c(1,1))
445   abline(v=0,lty=3)
446 }
447
448
449 #Use analytical confidence intervals instead of
      bootstrap where the standard error is known.
450 #OLS
```

```
451 ols.mdl <- lm(Y.original ~ X.original)
452 se.ols <- summary(ols.mdl)$coef[2,2]
453 Q[1,1] <- results[1] - 1.96 * se.ols
454 Q[1,2] <- results[1] + 1.96 * se.ols
455
456 ols.cov.mdl <- lm(Y.original ~ X.original+sdat$age+sdat
      $gender+sdat$low+sdat$med)
457 se.ols.cov <- summary(ols.cov.mdl)$coef[2,2]
458 Q.cov[1,1] <- results.covariates[1] - 1.96 * se.ols.cov
459 Q.cov[1,2] <- results.covariates[1] + 1.96 * se.ols.cov
460
461 ols.inf <- lm(sdat$infection ~ X.original)
462 se.ols.inf <- summary(ols.inf)$coef[2,2]
463 Q.infection[1,1] <- results.infection[1] - 1.96 * se.
      ols.inf
464 Q.infection[1,2] <- results.infection[1] + 1.96 * se.
      ols.inf
465
466 ols.inf.cov <- lm(sdat$infection ~ X.original+sdat$age+
      sdat$gender+sdat$low+sdat$med)
467 se.ols.cov.inf <- summary(ols.inf)$coef[2,2]
468 Q.infection.cov[1,1] <- results.covariates.infection[1]
      - 1.96 * se.ols.cov.inf
469 Q.infection.cov[1,2] <- results.covariates.infection[1]
      + 1.96 * se.ols.cov.inf
470
471
472 #logreg
473 lr.mdl <- glm(Y.original ~ X.original, family=binomial)
474 se.lr <- summary(lr.mdl)$coefficients[2,2]
475 Q[5,1] <- results[5] - 1.96 * se.lr
476 Q[5,2] <- results[5] + 1.96 * se.lr
477
478 lr.mdl.cov <- glm(Y.original ~ X.original+sdat$age+sdat
      $gender+sdat$low+sdat$med, family=binomial)
479 se.lr.cov <- summary(lr.mdl.cov)$coefficients[2,2]
480 Q.cov[5,1] <- results.covariates[5] - 1.96 * se.lr.cov
481 Q.cov[5,2] <- results.covariates[5] + 1.96 * se.lr.cov
482
483 lr.mdl.inf <- glm(sdat$infection ~ X.original, binomial
      )
```

```
484 se.lr.inf <- summary(lr.mdl.inf)$coefficients[2,2]
485 Q.infection[5,1] <- results.infection[5] - 1.96 * se.lr
    .inf
486 Q.infection[5,2] <- results.infection[5] + 1.96 * se.lr
    .inf
487
488 lr.mdl.inf.cov <- glm(sdat$infection ~ X.original+sdat$
    age+sdat$gender+sdat$low+sdat$med, binomial)
489 se.lr.inf.cov <- summary(lr.mdl.inf.cov)$coefficients
    [2,2]
490 Q.infection.cov[5,1] <- results.covariates.infection[5]
     - 1.96 * se.lr.inf.cov
491 Q.infection.cov[5,2] <- results.covariates.infection[5]
     + 1.96 * se.lr.inf.cov
492
493
494 #2SLS mortality
495 require(systemfit)
496 SEM <- Y.original ~ X.original
497 inst <- ~ Z.original
498 fit2sls <- systemfit(SEM,"2SLS",inst=inst)
499 summary(fit2sls)
500 #systemfit fails, compute by hand!
501
502 mdl.1 <- lm(X.original~Z.original)
503 mdl.2 <- lm(Y.original~fitted(mdl.1))
504
505 u <- residuals(mdl.2)
506 ssr <- t(u)%*%u / 461
507 Z.2 <- cbind(rep(1,461),Z.original)
508 X.2 <- cbind(rep(1,461),X.original)
509
510 var.2sls <- as.numeric(ssr) * solve(t(Z.2) %*% X.2) %*%
    t(Z.2)%*%Z.2 %*% solve(t(Z.2)%*%X.2)
511 Q[2,1] <- results[2] - 1.96 * sqrt(var.2sls[2,2])
512 Q[2,2] <- results[2] + 1.96 * sqrt(var.2sls[2,2])
513 sqrt(var.2sls[2,2])
514
515
516 #2SLS mortality with covariates
517 X.cov <- cbind(X.2,sdat$age,sdat$gender,sdat$low,sdat$
```

```
    med)
518 Z.cov <- cbind(Z.2, sdat$age, sdat$gender, sdat$low, sdat$
    med)
519 SEM.cov <- Y.original ~ X.cov
520 fit2sls.cov <- systemfit(SEM.cov,"2SLS",inst=inst)
521
522
523 mdl.1.c <- lm(X.original~Z.original)
524 mdl.2.c <- lm(Y.original~fitted(mdl.1.c))
525
526 u <- residuals(mdl.2.c)
527 ssr <- t(u)%*%u / 461
528
529 var.2sls <- as.numeric(ssr) * solve(t(Z.cov) %*% X.cov)
    %*% t(Z.cov)%*%Z.cov %*% solve(t(Z.cov)%*%X.cov)
530 Q.cov[2,1] <- results.covariates[2] - 1.96 * sqrt(var.2
    sls[2,2])
531 Q.cov[2,2] <- results.covariates[2] + 1.96 * sqrt(var.2
    sls[2,2])
532 sqrt(var.2sls[2,2])
533
534
535 #2SLS infection
536 SEM.inf <- sdat$infection ~ X.2
537 fit2sls.inf <- systemfit(SEM.inf,"2SLS",inst=inst)
538
539 mdl.2.inf <- lm(sdat$infection~fitted(mdl.2))
540
541 u <- residuals(mdl.2.inf)
542 ssr <- t(u)%*%u / 461
543
544 var.2sls <- as.numeric(ssr) * solve(t(Z.2) %*% X.2) %*%
    t(Z.2)%*%Z.2 %*% solve(t(Z.2)%*%X.2)
545 Q.infection[2,1] <- results.infection[2] - 1.96* sqrt(
    var.2sls[2,2])
546 Q.infection[2,2] <- results.infection[2] + 1.96* sqrt(
    var.2sls[2,2])
547 sqrt(var.2sls[2,2])
548
549
550 #2SLS infection with covariates
```

```
551 SEM.cov.inf <- sdat$infection ~ X.cov
552 fit2sls.cov.inf <- systemfit(SEM.cov.inf,"2SLS",inst=
      inst)
553
554 mdl.1.c.i <- lm(X.cov~Z.original)
555 mdl.2.c.i <- lm(sdat$infection~fitted(mdl.1.c.i))
556
557 u <- residuals(mdl.2.c.i)
558 ssr <- t(u)%*%u / 461
559
560 var.2sls <- as.numeric(ssr) * solve(t(Z.cov) %*% X.cov)
      %*% t(Z.cov)%*%Z.cov %*% solve(t(Z.cov)%*%X.cov)
561 Q.infection.cov[2,1] <- results.covariates.infection[2]
      - 1.96 * sqrt(var.2sls[2,2])
562 Q.infection.cov[2,2] <- results.covariates.infection[2]
      + 1.96 * sqrt(var.2sls[2,2])
563 sqrt(var.2sls[2,2])
564
565
566
567
568 plot.rd(results,Q,results.covariates,Q.cov,-.1,.1)
569 plot.OR(results,Q,results.covariates,Q.cov,-8,1)
570
571 plot.rd(results.infection,Q.infection,results.
      covariates.infection,Q.infection.cov,-.3,.1)
572 plot.OR(results.infection,Q.infection,results.
      covariates.infection,Q.infection.cov,-4,1)
573
574 #Histograms of the bootstrap estimates
575 hist(bootstrap.estimates[1,],breaks="fd")
576 hist(bootstrap.estimates[2,],breaks="fd")
577 hist(bootstrap.estimates[3,],breaks="fd")
578 hist(bootstrap.estimates[4,],breaks="fd")
579 hist(bootstrap.estimates[5,],breaks="fd")
580 hist(bootstrap.estimates[6,],breaks="fd")
581 hist(bootstrap.estimates[7,],breaks="fd")
582 hist(bootstrap.estimates[8,],breaks="fd")
583
584 hist(bootstrap.estimates2[1,],breaks="fd")
585 hist(bootstrap.estimates2[2,],breaks="fd")
```

```
586 hist(bootstrap.estimates2[3,], breaks="fd")
587 hist(bootstrap.estimates2[4,], breaks="fd")
588 hist(bootstrap.estimates2[5,], breaks="fd")
589 hist(bootstrap.estimates2[6,], breaks="fd")
590 hist(bootstrap.estimates2[7,], breaks="fd")
591 hist(bootstrap.estimates2[8,], breaks="fd")
592
593 hist(bootstrap.estimates.infection[1,], breaks="fd")
594 hist(bootstrap.estimates.infection[2,], breaks="fd")
595 hist(bootstrap.estimates.infection[3,], breaks="fd")
596 hist(bootstrap.estimates.infection[4,], breaks="fd")
597 hist(bootstrap.estimates.infection[5,], breaks="fd")
598 hist(bootstrap.estimates.infection[6,], breaks="fd")
599 hist(bootstrap.estimates.infection[7,], breaks="fd")
600 hist(bootstrap.estimates.infection[8,], breaks="fd")
601
602 hist(bootstrap.estimates.infection2[1,], breaks="fd")
603 hist(bootstrap.estimates.infection2[2,], breaks="fd")
604 hist(bootstrap.estimates.infection2[3,], breaks="fd")
605 hist(bootstrap.estimates.infection2[4,], breaks="fd")
606 hist(bootstrap.estimates.infection2[5,], breaks="fd")
607 hist(bootstrap.estimates.infection2[6,], breaks="fd")
608 hist(bootstrap.estimates.infection2[7,], breaks="fd")
609 hist(bootstrap.estimates.infection2[8,], breaks="fd")
610
611
612
613
614
615 #Comparing bootstrap to analytical SE's
616 #mort_30 with covariates
617 mdl <- glm(Y.original ~ X.original+sdat$age+sdat$gender
      +sdat$low+sdat$med, binomial)
618 summary(mdl)$coefficients[2,2]
619 sd(bootstrap.estimates2[5,])
620
621 mdl <- lm(Y.original ~ X.original+sdat$age+sdat$gender+
      sdat$low+sdat$med)
622 summary(mdl)$coefficients[2,2]
623 sd(bootstrap.estimates2[1,])
624
```

```
#infection no covariates
mdl <- glm(sdat$infection ~ X.original, binomial)
summary(mdl)$coefficients[2,2]
sd(bootstrap.estimates.infection[5,])

mdl <- lm(sdat$infection ~ X.original)
summary(mdl)$coefficients[2,2]
sd(bootstrap.estimates.infection[1,])

#infection with covariates
mdl <- glm(sdat$infection ~ X.original+sdat$age+sdat$
    gender+sdat$low+sdat$med, binomial)
summary(mdl)$coefficients[2,2]
sd(bootstrap.estimates.infection2[5,])

mdl <- lm(sdat$infection ~ X.original+sdat$age+sdat$
    gender+sdat$low+sdat$med)
summary(mdl)$coefficients[2,2]
sd(bootstrap.estimates.infection2[1,])

#Compare to 2SLS SE while using package
require(systemfit)
SEM <- Y ~ X+sdat$age+sdat$gender+sdat$low+sdat$med
inst <- ~ Z
fit2sls <- systemfit(SEM,"2SLS",inst=inst)
fit3sls <- systemfit(SEM,"OLS",inst=inst)
hausman.systemfit(fit2sls,fit3sls)
#seems ok.
summary(fit2sls)
```