2017

Topics In Statistical Inference For Treatment Effects

Yang Jiang
University of Pennsylvania, jiangyang1987@gmail.com

Topics In Statistical Inference For Treatment Effects

Abstract

This thesis unites three papers discussing different approaches for estimating treatment effects, either in observational study or randomized trial. The first paper presents an approach to sensitivity analysis for the instrumental variable(IV) method, which examines the sensitivity of inferences to violations of IV validity. Our approach is based on extending the Anderson-Rubin test and is robust to weak IVs. The second paper presents a unified \proglang{R} software \pkg{ivmodel} for analyzing instrumental variables with one endogenous variable. The package implements a general class of estimators, k-class estimators, and two confidence intervals that are fully robust to weak instruments. The package also provides power formulas. The sensitivity analysis discussed in the first paper is also included in the package. The third paper uses Hidden Markov Model to estimate the dynamic effects of lottery-based incentives towards patient's healthy behavior every day. The data is collected from randomized clinical trials.

Degree Type

Dissertation

Degree Name

Doctor of Philosophy (PhD)

Graduate Group

Statistics

First Advisor

Dylan Small

Second Advisor

Nancy Zhang

Keywords

instrumental variable, observational study, sensitivity analysis, statistical inference, treatment effects

Subject Categories

Statistics and Probability

TOPICS IN STATISTICAL INFERENCE FOR TREATMENT EFFECTS

Yang Jiang

A DISSERTATION

in

Statistics

For the Graduate Group in Managerial Science and Applied Economics

Presented to the Faculties of the University of Pennsylvania

in

Partial Fulfillment of the Requirements for the

Degree of Doctor of Philosophy

2017

Supervisor of Dissertation

Dylan Small

Professor of Statistics

Co-Supervisor of Dissertation

Nancy Zhang

Associate Professor of Statistics

Graduate Group Chairperson

Catherine Schrand, Celia Z. Moh Professor, Professor of Accounting

Dissertation Committee

Lawrence D. Brown, Professor of Statistics

Linda Zhao, Professor of Statistics

TOPICS IN STATISTICAL INFERENCE FOR TREATMENT EFFECTS

ACKNOWLEDGEMENT

I would like to thank my advisor Dylan Small and co-advisor Nancy Zhang, who guided me and advised me a lot in preparing this thesis. I also want to thank my parents Hong Yang and Ning Jiang, my girlfriend Alexandra He, who gave me support during this time and all people who helped me along the way.

ABSTRACT

TOPICS IN STATISTICAL INFERENCE FOR TREATMENT EFFECTS

Yang Jiang

Dylan Small

Nancy Zhang

This thesis unites three papers discussing different approaches for estimating treatment effects, either in observational study or randomized trial. The first paper presents an approach to sensitivity analysis for the instrumental variable(IV) method, which examines the sensitivity of inferences to violations of IV validity. Our approach is based on extending the Anderson-Rubin test and is robust to weak IVs. The second paper presents a unified R software **ivmodel** for analyzing instrumental variables with one endogenous variable. The package implements a general class of estimators, k-class estimators, and two confidence intervals that are fully robust to weak instruments. The package also provides power formulas. The sensitivity analysis discussed in the first paper is also included in the package. The third paper uses Hidden Markov Model to estimate the dynamic effects of lottery-based incentives towards patient's healthy behavior every day. The data is collected from randomized clinical trials.

TABLE OF CONTENTS

ACKNOWLEDGEMENT iii

ABSTRACT iv

LIST OF TABLES vii

LIST OF ILLUSTRATIONS ix

CHAPTER 1 : Introduction 1

CHAPTER 2 : Sensitivity Analysis and Power for Instrumental Variable Studies . 3

2.1 Introduction 3

2.2 Instrumental Variable Model with Possible Invalid Instruments 5

2.3 The 2SLS method and the Anderson-Rubin test 11

2.4 Sensitivity Analysis and Power of Sensitivity Analysis 14

2.5 Applications of Sensitivity Analysis and Power Calculation to Mendelian Randomization Studies 20

2.6 Discussion 27

CHAPTER 3 : ivmodel: An R Package for Inference and Sensitivity Analysis of Instrumental Variables Models with One Endogenous Variable . . . 29

3.1 Introduction 29

3.2 Instrumental Variables Model for One Endogenous Variable 31

3.3 k-Class Estimation and Inference 36

3.4 Dealing with Weak Instruments: Robust Confidence Interval Estimation . . 38

3.5 Dealing with Possibly Invalid Instruments: Sensitivity Analysis 40

3.6 Power 41

3.7 Application 44

3.8 Summary 50

CHAPTER 4 : Hidden Markov Model for Estimating Financial Incentive Effects towards Healthy Behavior 51

4.1 Introduction 51

4.2 Data source and trial protocols 52

4.3 HMM for dynamic incentive 54

4.4 Model fits and comparison 57

4.5 Discussion 60

APPENDIX 69

BIBLIOGRAPHY 69

LIST OF TABLES

TABLE 1 : Power for rare(common) variants under different sample size and sensitivity. We set $\sigma_1 = \sigma_2 = 1, \rho = 0.5, \lambda = 1$. For rare variant, $\gamma_r = 0.142, \mathrm{SD}(Z)_r = 0.071$ and for common variant, $\gamma_c = 0.046, \mathrm{SD}(Z)_c = 0.218$. In doing so rare and common variants have the same concentration parameter under the same sample size. The numbers in parentheses represent the power for common variants . 26

TABLE 2 : Application of instrumental variables methods based on source of instruments. Natural experiments/Mendelian randomization refer to instrumental variables studies where the instruments come from natural sources, such as genes or calendar years. Randomized experiments/encouragement designs refer to instrumental variables studies where the instruments represent actual randomization mechanisms. 30

TABLE 3 : Different types of k-class estimator 36

TABLE 4 : Given patient's hidden state and adherence, the probability to adhere in the next day under different lottery result. Parenthesis is the 95% confidence interval. 59

TABLE 5 : P-value of comparing different lottery result's effects toward patient's probability to adhere given patient's hidden state and adherence. 59

TABLE 6 : P-value for comparing different lottery results using the data from glucose reading trial. 60

TABLE 7 : P-value for comparing different lottery results using the data from warfarin trial. 60

LIST OF ILLUSTRATIONS

FIGURE 1 : A valid IV requires three conditions. The dash-dotted line suggests that the IV is associated with the exposure, which is IV-C1. The non-existing ("X" symbol in the figure means non-existing) dotted line suggests that the assignment of IV is independent of the unmeasured confounders, which is IV-C2. Similarly, the non-existing dashed line represents IV-C3 that the IV affects the outcome only through its effect on the exposure. 3

FIGURE 2 : Complete DAG for the model in Section 2.2. The dashed arrows represent the first stage model and the solid arrows represent the second stage model. 7

FIGURE 3 : Power of sensitivity analysis in simulated scenario where the base parameters are $\sigma_1^2 = 1, \sigma_2^2 = 4, \rho = 0.5, \gamma = 0.5, \lambda = -1, sd(Z) = 1, n = 200$. In each graph we vary one combination of parameters to observe the change of power. 18

FIGURE 4 : Data set is generated as $\sigma_1 = \sigma_2 = 1, \rho = 0.9(0.1), \beta = 0, \gamma \in [0, 0.1], \text{sd}(Z) = 1$ with sample size 5,000. Test with normal asymptotic distribution and standard AR test is performed with nominal significance level $\alpha = 0.05$. We calculate the average rejection rate among 20,000 simulated data sets. 23

FIGURE 5 : Sample size needed for achieving power>0.8 under different allowance of sensitivity. The vertical line stands for the design sensitivity 0.0499. Here $\lambda = 0.234, \delta = 0, \gamma = 0.158, \sigma_1^2 = 0.333, \sigma_2^2 = 1.0989, \rho = 0.548$. 25

FIGURE 6 : Complete. 48

FIGURE 7 : The HMM modeling the dynamic effect of lottery towards patient's tendency to adhere. 56

CHAPTER 1 : Introduction

This thesis is based on three papers, all of which address estimating treatment effects via different approaches. The first paper discusses the usage of current instrumental variable(IV) method in a less restricted scenario. In observational studies to estimate treatment effects, unmeasured confounding is often a concern. The IV method can control for unmeasured confounding when there is a valid IV. To be a valid IV, a variable needs to be independent of unmeasured confounders and only affect the outcome through affecting the treatment. When applying the IV method, there is often concern that a putative IV is invalid to some degree. We present an approach to sensitivity analysis for the IV method, which examines the sensitivity of inferences to violations of IV validity. Our approach is based on extending the Anderson-Rubin test and is robust to weak IVs. A power formula for this sensitivity analysis is presented. We illustrate its usage via examples about Mendelian randomization studies and its implications via a comparison of using rare vs. common genetic variants as instruments.

The second paper presents a unified R software **ivmodel** for analyzing instrumental variables with one endogenous variable. The package implements a general class of estimators, k-class estimators, and two confidence intervals that are fully robust to weak instruments. The package also provides power formulas. Finally, the package contains methods for sensitivity analysis to examine the sensitivity to the the instrumental variables assumptions. We demonstrate the software on the data set from Card (1995), looking at the causal effect of levels of education on log earnings where the instrument is the proximity to a four-year college. The approach developed in the first paper is also programmed in the package.

The third paper studies a treatment effects in randomized trials. Poor adherence to medical treatments can have large impact on health outcomes and health care costs. However, adherence is difficult to maintain, especially for long-term medication. Financial incentives are increasingly used as a method to improve medication adherence. Current literature

analyzes the overall effect of incentives applied to treatment group comparing to control group. In this paper, we focus on analyzing the dynamic effects of lottery-based incentives towards patient's healthy behavior every day. Hidden Markov Model is used in the modeling part and EM-algorithm and bootstrap methods are used for point estimation and confidence interval. The data is collected from 3 different clinical trials.

CHAPTER 2 : Sensitivity Analysis and Power for Instrumental Variable Studies

2.1 Introduction

In observational studies, it is challenging to make causal inference about treatment effects due to the potential presence of unmeasured confounding or reverse causation. One approach to address these challenges is the instrumental variable (IV) method, which uses an instrument to extract a quasi-random experimental study from an observational study. The method requires a valid IV, which is a variable that satisfies three conditions: (IV-C1) the IV is associated with the exposure; (IV-C2) there are no unmeasured confounders between the IV and outcome; (IV-C3) the IV affects the outcome only through its effect on the exposure. See Angrist et al. (1996a), Hernán and Robins (2006), Brookhart and Schneeweiss (2007), Baiocchi et al. (2014a) and Imbens (2014) for more discussions of IV.

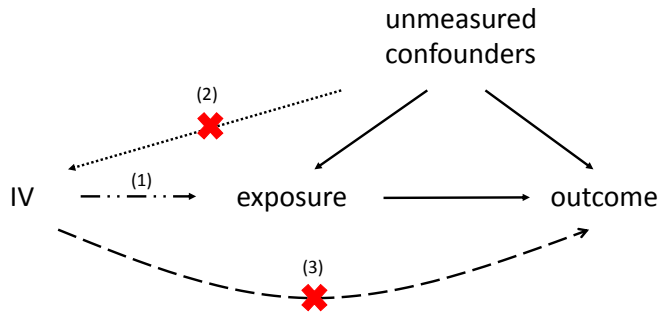

Figure 1: A valid IV requires three conditions. The dash-dotted line suggests that the IV is associated with the exposure, which is IV-C1. The non-existing ("X" symbol in the figure means non-existing) dotted line suggests that the assignment of IV is independent of the unmeasured confounders, which is IV-C2. Similarly, the non-existing dashed line represents IV-C3 that the IV affects the outcome only through its effect on the exposure.

Figure 1 depicts three conditions for a variable being a valid IV and the relationship between the IV, exposure, outcome and unmeasured confounders. When applying the IV method in a real study, investigators need to evaluate if there are any variables which satisfy the three conditions for being a valid IV. While (IV-C1) can be tested from the observed data, (IV-C2) and (IV-C3) cannot be completely tested, see Morgan and Winship (2007a). Therefore, it is often difficult to know whether an IV is perfectly valid in a study. Even when an IV is invalid, it may still be useful if the inferences from using the IV are not sensitive to plausible magnitudes of invalidity, which can be assessed through a sensitivity analysis (Angrist et al., 1996a; Imbens and Rosenbaum, 2005; Brookhart and Schneeweiss, 2007; Small and Rosenbaum, 2008). There is some previous work on sensitivity analysis for IV studies, see DiPrete and Gangl (2004a), Small (2007a), Kolesár et al. (2011a) and Conley et al. (2012a). These papers all use test statistics which are based on the two stage least squares estimator having an approximately normal distribution, which breaks down in the presence of weak instruments (instruments that are weakly associated with the exposure), see Nelson and Startz (1990a). This weak IV issue is very common in Mendelian randomization studies (Lawlor et al. (2008a), Section 4.10), in which genetic variants are used as IVs. See Section 2.5 for discussion of Mendelian randomization studies.

For settings with one IV that is either weak or strong, the Anderson-Rubin (AR) test (Anderson and Rubin, 1949a) has been shown to have good properties. Under the normal linear structural equation model setting that is reviewed in Section 2.2, the AR test is an exact test regardless of the strength of the IV. When the covariance matrix of the structural errors is known, Moreira (2001a) proved that AR test is uniformly most powerful among all unbiased tests; Andrews et al. (2006a) also proved that the AR test is uniformly most powerful among all invariant similar tests. When the covariance matrix is unknown, Andrews et al. (2006a) showed that AR test is asymptotically efficient for local alternatives under some regularity conditions.

Since the AR test has good performance for both weak and strong IVs, in this paper,

we develop a method of sensitivity analysis for instrumental variables based on the AR test. We demonstrate that the sensitivity analysis is robust to weak instruments unlike previous sensitivity analyses. Another contribution we make is that we give a power formula for the sensitivity analysis. The power formula enables researchers to decide how large a sample to collect if the goal is to find evidence for an exposure effect that is insensitive to a specified amount of invalidity of the IV. We show that when considering sensitivity analysis, the concentration parameter(F statistic), which is a commonly used criterion for measuring IV strength, is no longer a good measure for achieving a large power. Instead, it is better to focus on the IV effect size. This has important implications for the design of Mendelian randomization studies, in particular the choice between focusing on common vs. rare variants, which will be discussed in Section 2.5.

The remainder of this paper is organized as follows: In Section 2.2, we formulate a potential outcome model with a possibly invalid IV. In Section 2.3, we review the original 2SLS estimator and the AR test. In Section 2.4, we present our sensitivity analysis approach and provide the power formula for sensitivity analysis. In Section 2.5 we show how to do sensitivity analysis and calculate power in applications, including Mendelian randomization studies with common vs. rare variants. Section 2.6 provides conclusions.

2.2 Instrumental Variable Model with Possible Invalid Instruments

In this section, following Holland (1988a), we formulate a causal potential outcomes model (Rubin, 1974a; Splawa-Neyman et al., 1990) with a possibly invalid IV and connect it to the simultaneous equations model Hausman (1983). By doing so we can precisely define unit-level causal effects, obtain causal interpretations of certain regression coefficients and model possible violations of IV validity.

This paper will consider the setting of one IV and one exposure. For individual i, we use Z_i, D_i, Y_i to represent the observed IV, exposure and outcome accordingly. We use $Y_i^{(d,z)}$

to denote the potential outcome under the scenario where the individual i is assigned the exposure d and IV z. The variable $Y_i^{(d,z)}$ is the potential outcome individual i would have if we forcefully assign her/him to have exposure d and IV z. Similarly, $D_i^{(z)}$ is defined as the "potential exposure", which is the exposure individual i would have if we forcefully assign her/him to IV z. Notice that $D^{(z)}|(Z=z)$ might not equal $D^{(z)}|(Z=z')$ because even though we are comparing what would happen if two sets of subjects (those with $Z=z$ vs. those with $Z=z'$ got forcefully assigned the same level z of the IV, the set of subjects with observed $Z=z$ might have different levels of confounders than the set of subjects with observed $Z=z'$.

The first assumption we make is that the observed subjects $i=1,\ldots,n$ are an i.i.d. sample from a population. We also assume the effect of the IV on the exposure is linearly additive, i.e., for each extra unit of IV forcefully assigned, the potential exposure will increase by η units for every individual:

$$D_i^{(z)} - D_i^{(0)} = \eta z, \quad \forall i \tag{2.1}$$

in other words, η is the unit-level causal effect of the IV on exposure. For the potential exposure $D_i^{(0)}$, we can write it as:

$$\begin{aligned} D_i^{(0)} &= \mathbb{E}(D_i^{(0)}|Z_i=0) + (\mathbb{E}(D_i^{(0)}|Z_i) - \mathbb{E}(D_i^{(0)}|Z_i=0)) + (D_i^{(0)} - \mathbb{E}(D_i^{(0)}|Z_i)) \\ &= \mathbb{E}(D_i^{(0)}|Z_i=0) + \kappa(Z_i) + v_i, \quad \forall i \end{aligned} \tag{2.2}$$

where $\mathbb{E}(D_i^{(0)}|Z_i=0)$ is the population average of the potential exposure $D^{(0)}$ given that $Z=0$. The variations in $\{D_i^{(0)}; i=1,\ldots,n\}$ among different individuals come from two sources: 1) $\kappa(Z_i) = \mathbb{E}(D_i^{(0)}|Z_i) - \mathbb{E}(D_i^{(0)}|Z_i=0)$ measures the effect of unmeasured confounders between the IV and exposure. Different values of Z_i are associated with different levels of unmeasured confounders, which result in a difference of $\kappa(Z_i)$. Notice that $\kappa(0)=0$, if $\kappa(Z)=0$ for all possible values of Z, then there are no unmeasured confounders between the IV and exposure. We further assume this confounding effect is linear, so $\kappa(Z)=\kappa Z$. 2) The error term $v_i = D_i^{(0)} - \mathbb{E}(D_i^{(0)}|Z_i)$ can be understood as the individual error of the

potential exposure after removing the effect of all unmeasured confounders between the IV and exposure. v_i has a mean 0, $\mathbb{E}(v_i) = \mathbb{E}(D_i^{(0)}) - \mathbb{E}(\mathbb{E}(D_i^{(0)}|Z_i)) = 0$. We further assume v_i and Z_i are independent. Combining (2.1) and (2.2) and writing $\mathbb{E}(D_i^{(0)}|Z_i = 0) = \gamma_0$, we get the following "first stage" model that relates the observed D to the observed Z:

$$\begin{aligned} D_i = D_i^{(Z_i)} &= \left(D_i^{(Z_i)} - D_i^{(0)}\right) + \mathbb{E}(D_i^{(0)}|Z_i = 0) + \kappa(Z_i) + v_i \\ &= \gamma_0 + \eta Z_i + \kappa Z_i + v_i \\ &\forall i, \quad v_i \text{ is i.i.d. with mean 0}; \quad v_i \perp Z_i \end{aligned} \tag{2.3}$$

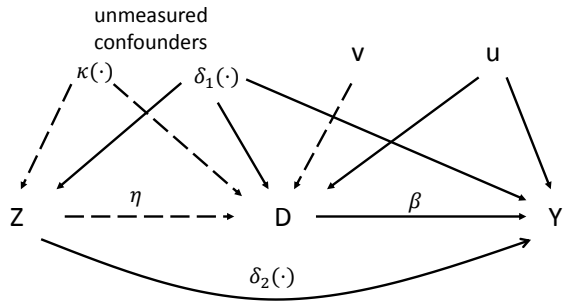

Figure 2: Complete DAG for the model in Section 2.2. The dashed arrows represent the first stage model and the solid arrows represent the second stage model.

(this is called the first stage model because it is the first stage in two stage least squares) The dashed arrows in Figure 7 show the causal relationship for our first stage model.

Now we consider the second stage model that relates the observed Y to the observed D. Assume that the causal effect for the exposure on the potential outcome is linear and it has no interaction with the IV, i.e., for each extra unit of exposure that is forcefully assigned,

the potential outcome will increase by β units for every individual:

$$Y_i^{(d,z)} - Y_i^{(0,z)} = \beta d, \quad \forall i \tag{2.4}$$

in other words, β is the unit-level causal effect of the exposure on outcome. Define $\delta_1^i(\cdot)$ as:

$$\delta_1^i(z) = Y_i^{(0,z)} - Y_i^{(0,0)}, \quad \forall i \tag{2.5}$$

Combining (2.4) and (2.5), we have:

$$Y_i^{(d,z)} - Y_i^{(d,0)} = (Y_i^{(0,z)} + \beta d) - (Y_i^{(0,0)} + \beta d) = \delta_1^i(z), \quad \forall i \tag{2.6}$$

Equation (2.6) says $\delta_1^i(\cdot)$ measures how much the IV affects the outcome through paths other than through the exposure for individual i. We assume this effect is linear and has no variation among different individuals, therefore $\delta_1^i(Z) = \delta_1 Z$. If $\delta_1 \equiv 0$, then $Y_i^{(d,z)} \equiv Y_i^{(d,0)}$, meaning that the IV affects the outcome only through its effect on the exposure, which is condition (IV-C3) for being a valid IV.

For the potential outcome term $Y_i^{(0,0)}$, we can write it as:

$$\begin{aligned} Y_i^{(0,0)} &= \mathbb{E}(Y_i^{(0,0)}|Z_i = 0) + (\mathbb{E}(Y_i^{(0,0)}|Z_i) - \mathbb{E}(Y_i^{(0,0)}|Z_i = 0)) + (Y_i^{(0,0)} - \mathbb{E}(Y_i^{(0,0)}|Z_i)) \\ &= \mathbb{E}(Y_i^{(0,0)}|Z_i = 0) + \delta_2(Z_i) + u_i, \quad \forall i \end{aligned} \tag{2.7}$$

where $\mathbb{E}(Y_i^{(0,0)}|Z_i = 0)$ is the population average of the potential outcome $Y^{(0,0)}$ given that $Z = 0$. The variations in $\{Y_i^{(0,0)}; i = 1, \ldots, n\}$ come from two sources: 1) $\delta_2(Z_i) = \mathbb{E}(Y_i^{(0,0)}|Z_i) - \mathbb{E}(Y_i^{(0,0)}|Z_i = 0)$ measures the effect of unmeasured confounders between the IV and outcome. Different values of Z_i are associated with different levels of unmeasured confounders, which result in a difference of $\delta_2(Z_i)$. Notice that $\delta_2(0) = 0$. If $\delta_2(Z) = 0$ for all possible values of Z, then there are no unmeasured confounders between the IV and outcome, which is condition (IV-C2) for being a valid IV. We further assume this

confounding effect is linear, so $\delta_2(Z) = \delta_2 Z$. 2) The error term $u_i = Y_i^{(0,0)} - \mathbb{E}(Y_i^{(0,0)}|Z_i)$ can be understood as the individual error of the potential outcome after removing the effect of all unmeasured confounders between the IV and outcome. u_i has a mean 0, $\mathbb{E}(u_i) = \mathbb{E}(Y_i^{(0,0)}) - \mathbb{E}(\mathbb{E}(Y_i^{(0,0)}|Z_i)) = 0$. We further assume u_i and Z_i are independent. Notice that although u_i is independent of the unmeasured confounders between the IV and the outcome, u_i may still be associated with the unmeasured confounder between the exposure and outcome, so u_i is not independent of D_i or v_i. Combining (2.4)-(2.7) and writing $\mathbb{E}(Y_i^{(0,0)}|Z_i = 0) = \beta_0$, we get the "second stage" model that relates the observed Y to the observed Z:

$$\begin{aligned} Y_i &= Y_i^{(D_i,Z_i)} \\ &= \left(Y_i^{(D_i,Z_i)} - Y_i^{(0,Z_i)}\right) + \left(Y_i^{(0,Z_i)} - Y_i^{(0,0)}\right) + \left(\mathbb{E}(Y_i^{(0,0)}|Z_i = 0) + \delta_2(Z_i) + u_i\right) \\ &= \beta_0 + \beta D_i + \delta_1 Z_i + \delta_2 Z_i + u_i \\ &\forall i, \quad u_i \text{ is i.i.d. with mean 0}, \quad u_i \perp Z_i \end{aligned} \tag{2.8}$$

The solid arrows in figure 7 show the causal relationship for our second stage model. Combining (2.3) and (2.8) and assuming v_i and u_i are bivariate normal our complete model can be written as:

$$\begin{aligned} Y_i &= \beta_0 + \beta D_i + (\delta_1 + \delta_2) Z_i + u_i \\ D_i &= \gamma_0 + (\eta + \kappa) Z_i + v_i \\ (u_i, v_i) \perp Z_i; \quad (u_i, v_i)^T &\sim N(\mathbf{0}, \Sigma); \quad \Sigma = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix} \end{aligned} \tag{2.9}$$

Here the parameter of interest is β, which represents the unit-level causal effect of the exposure on the outcome. δ_1 measures the violation of condition (IV-C3) for being a valid IV. δ_2 measures the violation of condition (IV-C2) for being a valid IV. The model (2.9) has the same structure as the models for sensitivity analysis in DiPrete and Gangl (2004a),

Kolesár et al. (2011a) and Conley et al. (2012a). As a summary, (2.9) relies on the following assumptions: *1) the observed subjects are an i.i.d. sample from a population; 2) the causal effects of the IV on the exposure and the exposure on the outcome are linearly additive; 3) the confounder function $\kappa(\cdot)$ and the IV violation functions $\delta_1^i(\cdot)$, $\delta_2(\cdot)$ are linear and $\delta_1^i(\cdot)$ is homogeneous for all individual i; 4) the error terms (u_i, v_i) are bivariate normal and independent of Z_i.*

We can reduce the number of parameters in (2.9) by defining $\gamma = \eta + \kappa$, $\delta = \delta_1 + \delta_2$. Here γ is the coefficient of IV Z in the first stage model. Z will satisfy condition (IV-C1) as long as $\gamma \neq 0$, i.e., the IV only needs to be associated with the exposure and could have no causal effect on the exposure ($\eta = 0$) (Hernán and Robins, 2006). The parameter δ_1 and δ_2 measures the violation of valid IV conditions and combined into $\delta = \delta_1 + \delta_2$, so that δ can be treated as the sensitivity parameter in (2.9), which describes the amount of invalidity of the IV.

There are other observed covariates (write as vector X_i of length k) for each individual i, they can be added into (2.9) as:

$$
\begin{aligned}
Y_i &= \beta_0 + \beta_X^T X_i + \beta D_i + \delta Z_i + u_i \\
D_i &= \gamma_0 + \gamma_X^T X_i + \gamma Z_i + v_i
\end{aligned}
$$

Writing the vector form of the observations as $Y_{n\times 1} = (Y_1, \ldots, Y_n)^T$, $D_{n\times 1} = (D_1, \ldots, D_n)^T$, $X_{n\times k} = (X_1, \ldots, X_n)^T$ etc, and also merging the intercept into the observed covariates X, we get an analogous model to (2.9):

$$
\begin{aligned}
Y &= X\beta_X + \beta D + \delta Z + u \\
D &= X\gamma_X + \gamma Z + v \\
(u, v) \perp Z; \quad (u_i, v_i)^T &\sim N(\mathbf{0}, \Sigma); \quad \Sigma = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}; \quad \operatorname{rank}(X) = k
\end{aligned}
\tag{2.10}
$$

Since β_X and γ_X are not of interest, we can use the Frisch-Waugh-Lovell theorem (Davidson and MacKinnon, 1993a; Wang and Zivot, 1998) to transform the model (2.10) by using the projection matrix $M_X = I_{N\times N} - X(X^TX)^{-1}X^T$. Also, to make the sensitivity parameter δ in the model more interpretable, we rescale δ as $\delta\sigma_1$. After this rescaling, a unit change in the invalid IV Z will lead to a change of δ standard deviations of the structural error $u = Y^{(0,0)} - \mathbb{E}(Y^{(0,0)}|Z)$. The final model after transforming and rescaling becomes:

$$\begin{aligned} Y^* &= \beta D^* + \delta\sigma_1 Z^* + u^* \\ D^* &= \gamma Z^* + v^* \\ Y^* = M_X Y;\quad D^* = M_X D;\quad Z^* = M_X Z;\quad u^* = M_X u;\quad v^* &= M_X v; \\ (u,v) \perp Z;\quad (u_i,v_i)^T \sim N(\mathbf{0},\Sigma);\quad \Sigma = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix} &\quad \operatorname{rank}(X) = k \end{aligned} \tag{2.11}$$

(A.1) is the model we will consider in the rest of the paper. In the following sections, our work will mainly focus on inference for β, given restrictions on the range of the sensitivity parameter δ.

2.3 The 2SLS method and the Anderson-Rubin test

In this section, we consider model (A.1) with $\delta = 0$, which is the usual two stage IV regression model with a valid IV. We will briefly review the two stage least squares estimator(2SLS) that has an asymptotic normal distribution and the standard AR test.

The 2SLS estimator of β is found by first regressing D^* on Z^* to find $\hat{D}^*$, and then regressing Y^* on $\hat{D}^*$. The 2SLS estimator can be writen as follows:

$$\hat{\beta}_{2SLS} = \frac{\operatorname{cov}(Z^*, Y^*)}{\operatorname{cov}(Z^*, D^*)} = \frac{Z^{*T}Y^*}{Z^{*T}D^*} \tag{2.12}$$

As the sample size increases to infinity, $\hat{\beta}_{2SLS} \to \beta$. Also the asymptotic variance for $\hat{\beta}_{2SLS}$

(Nelson and Startz, 1990a) is:

$$\text{Asymptotic Variance}(\sqrt{n} \times (\hat{\beta}_{2SLS} - \beta)) = \frac{\sigma^2_{u^*} \cdot Var(Z^*)}{Cov^2(Z^*, D^*)} \tag{2.13}$$

(2.12) and (2.13) can be used to construct an asymptotically valid t-test. However, when the IV is weak, the asymptotics of this test may provide a poor guide to the actual performance of the test even for moderately large sample sizes. (Nelson and Startz, 1990a; Staiger and Stock, 1997a)

From Anderson and Rubin (1949a), the AR test compares the null and alternative hypotheses:

$$H_0 : \beta = \beta_0 \qquad vs. \qquad H_1 : \beta \neq \beta_0$$

If the IV is valid, then under H_0, $Y^* - \beta_0 D^* = u^*$ is independent of Z^*. The coefficient of regressing $Y^* - \beta_0 D^* = u^*$ on Z^* should be 0. The AR test is an F test for this coefficient being 0. It has the following expression:

$$AR(\beta_0) = \frac{(Y^* - \beta_0 D^*)^T P_{Z^*} (Y^* - \beta_0 D^*)}{(Y^* - \beta_0 D^*)^T M_{Z^*} (Y^* - \beta_0 D^*)/(n - k - 1)} \tag{2.14}$$

where n is the number of samples, P_{Z^*}, M_{Z^*} are projection matrices $P_{Z^*} = Z^*(Z^{*T} Z^*)^{-1} Z^{*T}$ and $M_{Z^*} = I_n - P_{Z^*}$. Under $H_0 : \beta = \beta_0$, since $u^{*T} P_{Z^*} u^* \sim \sigma_1^2 \chi_1^2$, $u^{*T} M_{Z^*} u^*/(n-k-1) \sim \sigma_1^2 \chi^2_{n-k-1}$ and they are independent (see more details in the Supplementary Materials), we have:

$$AR(\beta_0) = \frac{u^{*T} P_{Z^*} u^*}{u^{*T} M_{Z^*} u^*/(n - k - 1)} \sim F_{1,n-k-1} \tag{2.15}$$

We reject H_0 when $AR(\beta_0) > F_{1,n-k-1;1-\alpha}$, where α is the significance level and $F_{1,n-k-1;1-\alpha}$ is the $1 - \alpha$ quantile of F distribution with degree of freedom 1 and $n - k - 1$. (Notice that since u^* is the error after projecting out the effect of the covariates which include an intercept, there are only $n - k - 1$ degree of freedom in the denominator.) In contrast with the t-test based on the 2SLS estimator, the Anderson-Rubin test has correct size regardless of the sample size and strength of the IV.

We can construct a $1-\alpha$ confidence interval (CI) from the AR test by solving the inequality:

$$CI_{1-\alpha} = \{\beta : \quad \frac{(Y^* - \beta D^*)^T P_{Z^*}(Y^* - \beta D^*)}{(Y^* - \beta D^*)^T M_{Z^*}(Y^* - \beta D^*)/(n-k-1)} \leq F_{1,n-k-1;1-\alpha}\} \tag{2.16}$$

We calculate the power functions for the AR test in the Supplementary Materials. Under the alternative hypothesis $H_1 : \beta - \beta_0 = \lambda$, the power formula is:

$$Power = P_\lambda(AR(\beta_0) > F_{1,n-k-1;1-\alpha}) = 1 - \Psi_{1,n-k-1,\frac{\gamma^2 Z^{*T} Z^*}{\sigma_2^2}\Lambda}(F_{1,n-k-1;1-\alpha}) \tag{2.17}$$

where $\Lambda = \frac{\lambda^2}{(\sigma_1/\sigma_2)^2+2\rho\sigma_1/\sigma_2\lambda+\lambda^2}$, $\Psi_{a,b,k}(\cdot)$ is the CDF of the non-central F-distribution with degree of freedom a, b and non-centrality parameter k. The term $\gamma^2 Z^{*T} Z^*/\sigma_2^2$ is called the concentration parameter, the larger the concentration parameter is, the larger the power is. The concentration parameter is the population value of the first stage F statistic for the IV when the treatment is regressed on it. It is used as a popular measure for instrument strength (Stock et al., 2002a). Large values of the concentration parameter indicate strong instruments. See Rothenberg (1984), Section 6.1 for more information about the concentration parameter.

Before closing this section, we want to point out that the strength of the IV poses a fundamental limit on the power in IV studies. Looking at equation (2.17), if the concentration parameter is fixed, then no matter how large the effect size λ/σ_1 is, Λ has a fixed upper bound $1/(1-\rho^2)$ and the power has an upper bound:

$$1 - \Psi_{1,n-k-1,\frac{\gamma^2 Z^{*T} Z^*}{\sigma_2^2(1-\rho^2)}}(F_{1,n-k-1;1-\alpha})$$

Therefore the power will not increase to 1 as the effect size increases to infinity. Thus, the concentration parameter imposes a fundamental limit on the power of the AR test which cannot be overcome with a large effect size of the treatment.

2.4 Sensitivity Analysis and Power of Sensitivity Analysis

For our sensitivity analysis, we assume that $\delta \in (\underline{\delta}, \bar{\delta})$ in model (A.1) and consider what inference we can make if we do not know the value of δ but know its range $(\underline{\delta}, \bar{\delta})$.

2.4.1 CI and power formula for sensitivity analysis using AR test

First suppose we know the true value of the sensitivity parameter δ in $Y^* = \beta D^* + \delta\sigma_1 Z^* + u^*$. The AR test statistic under $H_0 : \beta = \beta_0$ becomes:

$$\begin{aligned} AR(\beta_0) &= \frac{(Y^* - \beta_0 D^*)^T P_{Z^*}(Y^* - \beta_0 D^*)}{(Y^* - \beta_0 D^*)^T M_{Z^*}(Y^* - \beta_0 D^*)/(n-k-1)} \\ &= \frac{\left(\delta\sigma_1\sqrt{Z^{*T}Z^*} + \frac{Z^{*T}u^*}{\sqrt{Z^{*T}Z^*}}\right)^2}{u^{*T}M_{Z^*}u^*/(n-k-1)} \sim F_{1,n-k-1,\delta^2 Z^{*T}Z^*} \end{aligned} \tag{2.18}$$

where $F_{a,b,c}$ stands for the non-central F distribution with degree of freedom a, b and non-central parameter c. Therefore a $1-\alpha$ CI can be obtained as:

$$CI_{1-\alpha}(\delta) = \{\beta : \frac{(Y^* - \beta D^*)^T P_{Z^*}(Y^* - \beta D^*)}{(Y^* - \beta D^*)^T M_{Z^*}(Y^* - \beta D^*)/(n-k-1)} < F_{1,n-k-1,\delta^2 Z^{*T}Z^*;1-\alpha}\} \tag{2.19}$$

where $F_{a,b,c;1-\alpha}$ stands for the $1-\alpha$ quantile of the distribution $F_{a,b,c}$ defined as above. Now let's go back to the assumption where we only know $\delta \in (\underline{\delta}, \bar{\delta})$. Define $\Delta = \max(|\underline{\delta}|, |\bar{\delta}|)$. We can construct a CI for β which will provide at least $1-\alpha$ coverage by taking the union of $CI_{1-\alpha}(\delta)$, for every $\delta \in (\underline{\delta}, \bar{\delta})$:

$$\begin{aligned} CI_{1-\alpha} &= \cup_{\delta\in(\underline{\delta},\bar{\delta})}\, CI_{1-\alpha}(\delta) \\ &= \{\beta : \frac{(Y^* - \beta D^*)^T P_{Z^*}(Y^* - \beta D^*)}{(Y^* - \beta D^*)^T M_{Z^*}(Y^* - \beta D^*)/(n-k-1)} < F_{1,n-k-1,\Delta^2 Z^{*T}Z^*;1-\alpha}\} \end{aligned} \tag{2.20}$$

We now consider the power for being able to reject $H_0 : \beta = \beta_0$ for all $\delta \in (\underline{\delta}, \bar{\delta})$ when the true δ is δ^*. For calculating the power, details are derived in the Supplementary Materials

and here we only show the main results. Under $H_1 : \beta - \beta_0 = \lambda \neq 0$, we have,

$$AR(\beta_0) \sim F_{1,n-k-1,\frac{(\gamma+\delta^*\sigma_1/\lambda)^2 Z^{*T} Z^*}{\sigma_2^2}\Lambda} \tag{2.21}$$

therefore,

$$\begin{aligned} Power_{\delta^*} &= P_\lambda(AR(\beta_0) \notin CI_{1-\alpha}) \\ &= P_\lambda(AR(\beta_0) > F_{1,n-k-1,\Delta^2 Z^T Z;1-\alpha}) \\ &= 1 - \Psi_{1,n-k-1,\frac{(\gamma+\delta^*\sigma_1/\lambda)^2 Z^{*T} Z^*}{\sigma_2^2}\Lambda}\left(F_{1,n-k-1,\Delta^2 Z^{*T} Z^*;1-\alpha}\right) \end{aligned} \tag{2.22}$$

Equation (2.22) calculates the probability of correctly rejecting H_0 for a fixed value δ^* of δ.

For calculating the power of a sensitivity analysis, Rosenbaum (2010a) Chapter 14.2 suggests calculating the power for the "favorable situation" in which there is a treatment effect and in fact there is no bias from unmeasured confounding ($\delta^* = 0$), but we do not know that there is no unmeasured confounding and want to be able to reject the null hypothesis of no treatment effect given a certain magnitude of unmeasured confounding, i.e., $\delta \in (\underline{\delta}, \bar{\delta})$ in our setting. To calculate the power of sensitivity analysis under this favorable situation, we plug $\delta^* = 0$ into (2.22):

$$Power_0 = 1 - \Psi_{1,n-k-1,\frac{\gamma^2 Z^{*T} Z^*}{\sigma_2^2}\Lambda}\left(F_{1,n-k-1,\Delta^2 Z^{*T} Z^*;1-\alpha}\right) \tag{2.23}$$

Another type of power of sensitivity analysis calculation is to find the minimum power for rejecting the null hypothesis of no treatment effect under a sensitivity analysis that allows for unmeasured confounding in the range $\delta \in (\underline{\delta}, \bar{\delta})$. To calculate this minimum power, we take $\min_{\delta\in(\underline{\delta},\bar{\delta})}$ on the right hand side of (2.22).

$$Power \geq \min_{\delta\in(\underline{\delta},\bar{\delta})} Power_\delta = 1 - \Psi_{1,n-k-1,\frac{\min_{\delta\in(\underline{\delta},\bar{\delta})}(\gamma+\delta\sigma_1/\lambda)^2 Z^{*T} Z^*}{\sigma_2^2}\Lambda}\left(F_{1,n-k-1,\Delta^2 Z^{*T} Z^*;1-\alpha}\right) \tag{2.24}$$

We have written functions for the R package "ivpack", available on CRAN, that calculate the sensitivity analysis confidence interval (2.20) (function ARsensitivity.ci), the power of sensitivity analysis using formula (A.14) or (2.24) (function ARsensitivity.power) and the minimum sample size needed for reaching a certain power in a sensitivity analysis (function ARsensitivity.size).

2.4.2 Effect of Different Parameters on Power of Sensitivity Analysis

We consider the effect of different parameters on the power of sensitivity analysis. Here we will focus on analyzing the power formula (A.14) for the favorable situation.

The power formula (A.14) involves two different non-central F distributions with two non-centrality parameters:

$$\text{ncp}_1 = \frac{\gamma^2 Z^{*T} Z^*}{\sigma_2^2}\Lambda; \qquad \text{ncp}_2 = \Delta^2 Z^{*T} Z^*$$

To analyze the influence of different parameters in the power formula, we study how they affect the size of ncp_1 and ncp_2. The Supplementary Materials proves the following properties:

Proposition 1. *In the power formula* (A.14)*, we have*

(a) If $ncp_1 = ncp_2$, the power is always α.

(b) For fixed ncp_2, power increases as ncp_1 increases.

(c) For fixed ncp_1, power decreases as ncp_2 increases.

(d) If $ncp_1 > ncp_2$, the power is larger than α and will increase to 1 as the sample size increases.

(e) If $ncp_1 < ncp_2$, the power is smaller than α and will decrease to 0 as the sample size increases.

The $\text{ncp}_2 = \Delta^2 Z^{*T} Z^*$ is approximately equal to $n \cdot \Delta^2 \cdot \text{SD}(Z^*)^2$. Δ is determined by the allowance of sensitivity and n is the sample size. The ncp_1 is affected many parameters:

$$\text{ncp}_1 = \frac{\gamma^2 Z^{*T} Z^*}{\sigma_2^2}\Lambda = \frac{\gamma^2}{\sigma_2^2} \cdot Z^{*T} Z^* \cdot \frac{1}{(\frac{\sigma_1}{\sigma_2 \lambda} + \rho)^2 + 1 - \rho^2}$$

In Section 2.3 we stated that a large concentration parameter $\gamma^2 Z^{*T} Z^* / \sigma_2^2$ will lead to a large power in AR test, assuming that the IV is valid. However, if we are doing sensitivity analysis and considering the power formula (A.14), a large concentration parameter may be produced by a small $|\gamma|/\sigma_2$ and a large $Z^{*T} Z^*$ (or vice versa). This will result in both large ncp_1 and ncp_2, for which the power of sensitivity analysis may not be large. On the other hand, if the IV effect size $|\gamma|/\sigma_2$ increases and the other parameters keep the same values, then ncp_1 increases and ncp_2 stays the same, which leads to a larger power of sensitivity analysis. This suggests that if we want to have a large power of sensitivity analysis, we should focus on finding a large IV effect size $|\gamma|/\sigma_2$. We will discuss the implications of this for Mendelian randomization studies in Section 2.5.2.

For the effect size λ/σ_1, no matter how it varies, ncp_1 has an upper bound $\gamma^2 Z^{*T} Z^* / (\sigma_2^2 (1 - \rho^2))$ when $\lambda/\sigma_1 = -(\sigma_2 \rho)^{-1}$. Hence the power of sensitivity analysis cannot go to 1 as the effect size increases to infinity. This is similar to the discussions at the end of Section 2.3, which is about the power property in AR test.

If the effect size λ/σ_1 is very small or the IV effect size $|\gamma|/\sigma_2$ is very small, then we may have

$$\frac{1}{(\frac{\sigma_1}{\sigma_2 \lambda} + \rho)^2 + 1 - \rho^2} \cdot \frac{\gamma^2}{\sigma_2^2} < \Delta^2 \tag{2.25}$$

This will result in $\text{ncp}_1 < \text{ncp}_2$. Under such a situation, the sensitivity analysis cannot have power larger than α for any sample size.

For further illustration, we simulated a simple scenario and calculated the power of sensi-

tivity analysis by varying different parameters. We consider the following base parameters:

$$\sigma_1^2 = 1; \quad \sigma_2^2 = 4; \quad \rho = 0.5; \quad \gamma = 0.5; \quad \lambda = -1; \quad \mathrm{SD}(Z) = 1; \quad n = 200 \tag{2.26}$$

and we want to use the power formula (A.14) under different allowance of sensitivity parameter interval $\delta \in [-0.05, 0.05]$, $[-0.08, 0.08]$, or $[-0.1, 0.1]$. We will allow one combination of parameters to vary at a time, to see what's the effect upon the power.

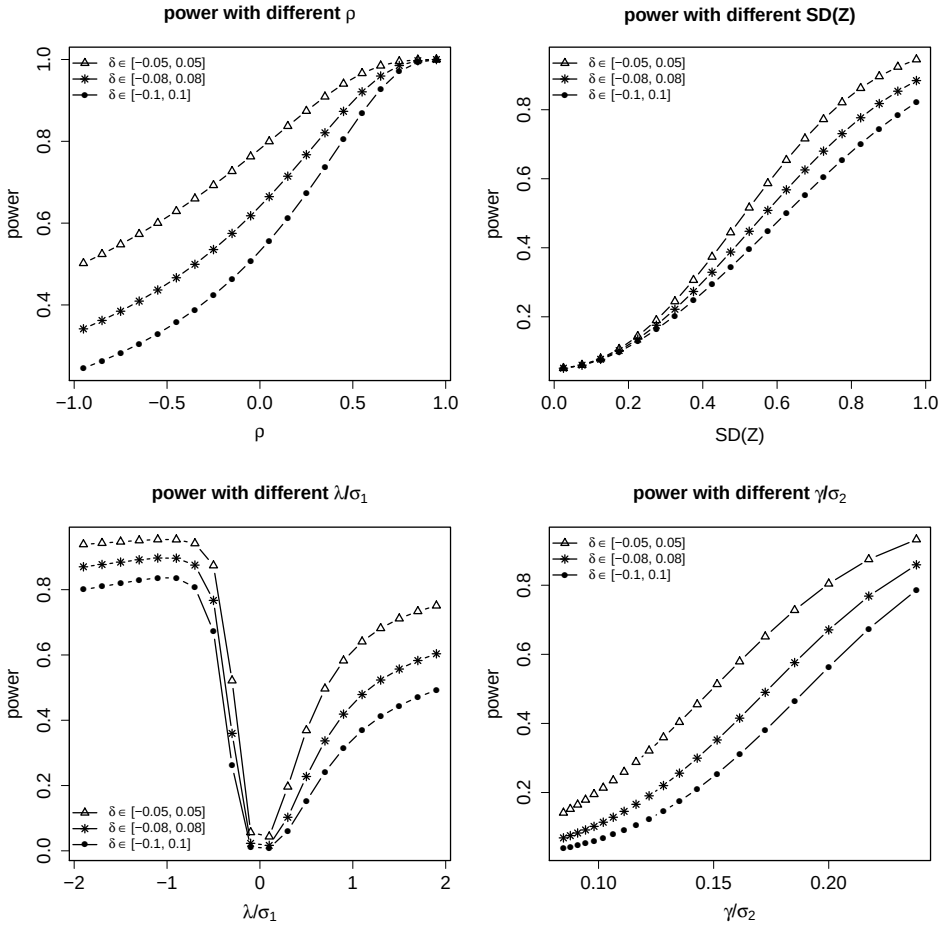

Figure 3: Power of sensitivity analysis in simulated scenario where the base parameters are $\sigma_1^2 = 1, \sigma_2^2 = 4, \rho = 0.5, \gamma = 0.5, \lambda = -1, sd(Z) = 1, n = 200$. In each graph we vary one combination of parameters to observe the change of power.

Figure 3 shows the result. In the top left graph, ρ varies between $(-1, 1)$ while other parameters are fixed. We can see the power increases as ρ increases when the effect size λ/σ_1 is negative. If the effect size λ/σ_1 is positive, then power increases if ρ decreases. In the top right graph, $\text{SD}(Z^*)$ varies between $(0, 1)$ while other parameters are fixed. We can see the power increases as $\text{SD}(Z^*)$ increases. If the inequality (2.25) holds, then the power is smaller than α and will decrease as $\text{SD}(Z^*)$ increases. In the bottom left graph, λ varies between $(-2, 2)$ while other parameters are fixed. This corresponds to the effect size λ/σ_1 varying between $(-2, 2)$. In general we can see the power is large when the effect size is substantial. However, the upper bound for the power is when $\lambda/\sigma_1 = -(\rho\sigma_2)^{-1} = -1$. As λ/σ_1 moves below -1, the power even starts to drop a little bit. This again corresponds to the previous discussion that no matter how large the effect size is, there's an upper bound for the power. In the bottom right graph, σ_2 varies between $(2, 6)$ while other parameters are fixed. This corresponds to the IV effect size γ/σ_2 varying between $(0.083, 0.25)$. In general we can see the power increases as the IV effect size increases.

2.4.3 Design sensitivity

The design sensitivity describes the asymptotic power of sensitivity analysis (Rosenbaum, 2004, 2010a). Here we calculate the design sensitivity in our sensitivity analysis model. Suppose the sensitivity range that δ lies is centered at 0: $\delta \in (-\Delta, \Delta)$. For large Δ, the power of sensitivity analysis using power (A.14) tends to 0 as the sample size increases. For small Δ, the power tends to 1 as the sample size increases. The switch point is defined as the design sensitivity Δ_{DS}.

To calculate the design sensitivity, we can use the property (d) and (e) in Proposition 3, which tells us that the switch point happens when $\text{ncp}_1 = \text{ncp}_2$, so we have:

$$\Delta_{DS} = \sqrt{\frac{\lambda^2\gamma^2}{\sigma_1^2 + 2\rho\sigma_1\sigma_2\lambda + \lambda^2\sigma_2^2}} \tag{2.27}$$

Notice that Δ_{DS} has an upper bound for any λ:

$$\Delta_{DS} = \sqrt{\frac{\gamma^2}{(\sigma_1/\lambda + \rho\sigma_2)^2 + \sigma_2^2(1-\rho^2)}} \leq \frac{|\gamma|}{\sigma_2\sqrt{1-\rho^2}}$$

Thus, no matter how large the effect size is, the design sensitivity is limited by the IV effect size $|\gamma|/\sigma_2$. This also suggests that the IV effect size is a better measure of IV strength than the concentration parameter when we are concerned that IV might not be perfectly valid and we would like to do a sensitivity analysis.

2.5 Applications of Sensitivity Analysis and Power Calculation to Mendelian Randomization Studies

An important application area of the IV method is Mendelian randomization studies (Smith and Ebrahim, 2003, 2004; Lawlor et al., 2008a; Ebrahim and Smith, 2008; Glymour et al., 2012). The basic idea of Mendelian randomization is to use inherited genetic variants as IVs to study the effect of an exposure on an outcome. By Mendel's second law, the transmission of genetic variants between generations is independent of possible confounders like environment and lifestyle factors. This makes it plausible that genetic variants satisfy the condition (IV-C2) for being a valid IV. If a generic variant is independent of unmeasured confounders and is also associated with the exposure and affects the outcome only through the exposure, then it is a valid IV. However, there are several ways that genetic variants could violate the conditions for being a valid IV, such as linkage disequilibrium, population stratification or pleiotropy (Didelez and Sheehan (2007) Section 7 and Lawlor et al. (2008a) Section 4). For example, if a genetic variant used as an IV is linked to another unmeasured genetic variant on the same chromosome that affects the outcome, there is linkage disequilibrium and the condition (IV-C2) is violated. Another way that a genetic variant could violate (IV-C2) is through population stratification (subpopulations which exhibit systematic differences in genotypes due to different ancestries) which is associated with both the IV and the outcome. Besides possibly violating (IV-C2), a genetic variant could violate (IV-C3) by

being pleiotropic in such a way that the genetic variant influences both the exposure and the outcome through a pathway other than the exposure. Consequently, in most studies using Mendelian randomization, there is some concerned about whether the proposed IVs are valid (e.g. see Nitsch et al. (2006)). It is useful to do a sensitivity analysis to examine how sensitive the analysis results are to the violation of (IV-C2) and (IV-C3).

2.5.1 Applications to a Mendelian Randomization Study

Here we will use the same example in Freeman et al. (2013a) to illustrate how to do sensitivity analysis and power calculation in a Mendelian randomization study.

The example concerns the causal effect of C-reactive protein (CRP), a marker of inflammation, on fibrinogen, a maker for coronary heart disease. The gene that makes CRP has several variations in the form of single nucleotide polymorphisms (SNPs). SNPs of the CRP gene have commonly been used to study the causal effects of CRP in Mendelian randomization studies, e.g. see Lawlor et al. (2008a). Although the CRP gene is believed to only directly affect CRP, it is possible that there is some unknown mechanism by which the CRP gene affects fibrinogen not through affecting CRP levels and it is also possible that there is population stratification. Consequently, we would like to consider a sensitivity analysis that allows for violations of the assumptions of the CRP gene being a valid IV. See Freeman et al. (2013a) for more details about the example.

We will use the same simulated data settings as Freeman et al. (2013a), the setting is:

$$Z = \{1, 2, 3\} \text{ with prob. } (1/9, 4/9, 4/9); \quad U \sim N(0, 1.11 * p * 0.99)$$

$$X \sim N(U + 0.1(Z-2)\sqrt{1.11 * 9/4}, 1.11 * (1-p) * 0.99); \quad Y = 0.234X + U/\sqrt{0.99}$$

which gives

$$\beta = 0.234; \quad \rho^2_{ZD} = 0.01; \quad \text{Var}(X) = 1.11; \quad \text{Var}(Z) = 4/9$$

These parameters are based on studies of CRP and fibrinogen from CRP-CHD-Genetics-Collaboration (CCGC), Burgess and Thompson (2010) and Burgess et al. (2012). We can rewrite this model in an equivalent form which fits our model setting (A.1) with:

$$\beta = 0.234;\ \ \delta = 0;\ \ \gamma = 0.1 * \sqrt{1.11 * 9/4};\ \ \sigma_1^2 = 1.11 * p;\ \ \sigma_2^2 = 1.11 * 0.99;\ \ \rho = \sqrt{p} \quad (2.28)$$

Freeman et al. (2013a) varied the parameter p in study, the larger p is, the more confounding there is. We will consider $p = 0.3$, a moderate confounding effect, in most of the analysis below.

First, without sensitivity analysis, we calculate the necessary sample size needed for rejecting the null hypothesis with power greater than 0.8. If we use the t test based on asymptotic normal distribution, then Freeman et al. (2013a) calculated the power formula as

$$n = \frac{(z_{\alpha/2} + z_\beta)^2 V}{\lambda^2 \rho_{ZD}^2} \quad (2.29)$$

where $V = \sigma_1^2/\mathrm{Var}(D)$. By formula (2.29), we need a sample size of 4301 if we want the power greater than 0.8. However, if we use the AR test and plug the parameters (2.28) into the power formula (2.17), then we would need a sample size of 7085, larger than the sample size of 4301 suggested by (2.29). Although this would seem to suggest that using the t-test based on the asymptotic normal distribution can reduce the sample size needed compared to the AR test, the sample size needed for the t-test based on the asymptotic normal distribution from (2.29) cannot be trusted while the sample size needed for the AR test from (2.17) can be trusted. There are two reasons for this: (i) the nominal level of the t-test based on the asymptotic normal distribution is not reliable for finite samples and can be much greater than the actual level while the nominal level of the AR test is exactly equal to the true level regardless of the sample size. (ii) the power formula (2.29) may not be accurate for finite samples while the formula (2.17) is an exact formula.

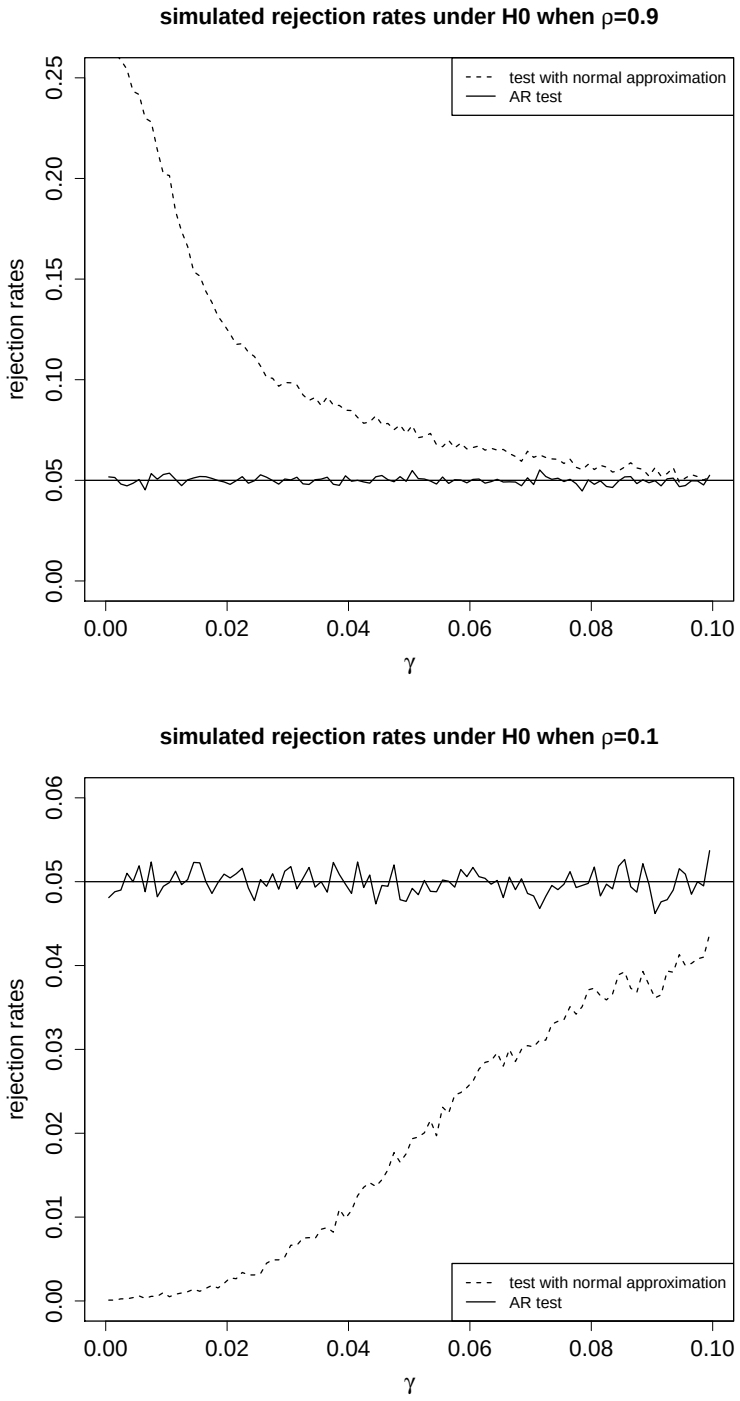

Figure 4: Data set is generated as $\sigma_1 = \sigma_2 = 1, \rho = 0.9(0.1), \beta = 0, \gamma \in [0, 0.1], \mathrm{sd}(Z) = 1$ with sample size 5,000. Test with normal asymptotic distribution and standard AR test is performed with nominal significance level $\alpha = 0.05$. We calculate the average rejection rate among 20,000 simulated data sets.

We illustrate point (i) in Figure 4. In this figure, the null hypothesis is tested both using the t-test based on the asymptotic normal distribution and the AR test where the null hypothesis is true. The top panel considers a situation of large confounding while the bottom panel considers a situation of small confounding. The strength of the IV is varied along the x-axis. 10,000 data sets are simulated with 5000 observations each and the rejection rate is displayed on the y-axis. The standard error of the rejection rate for a 0.05 level test from 10,000 simulations is smaller than $\sqrt{0.5 * 0.5/10000} = 0.005$. These simulations show that the AR test always has level about 0.05, equal to its nominal level, while the t-test based on the asymptotic normal distribution can have level way above its nominal level (an actual level of 0.25 compared to the nominal level of 0.05 in the top panel for a weak IV) or level way below its nominal level.

To illustrate point (ii) about the power formula (2.29) being less accurate than the power formula (2.17), we consider the setting in Freeman et al. (2013a) described above and simulated 10,000 data sets with 4301 observations from (2.28) with $p = 0.3$ and used the t-test based on the asymptotic normal distribution. The power in the simulated data sets is 0.7266 comparing to the number 0.8 that formula (2.29) said the power should be. In contrast, when we simulated 10,000 data sets with 7085 observations from (2.28) with $p = 0.3$ and used the AR test, the power in the simulated data sets was 0.8002 compared to 0.8, as formula (2.17) said the power should be.

Now we consider sensitivity analysis using the AR test for model (2.28) with $p = 0.3$. Suppose we would like to conduct a sensitivity analysis for $\delta \in [-0.01, 0.01]$, which means a one unit change in the IV could change up to a 0.01 standard deviation of the structural error. By the power formula (A.14), we need at least 8845 observations to achieve power at least 0.8 under the favorable situation $\delta = 0$.

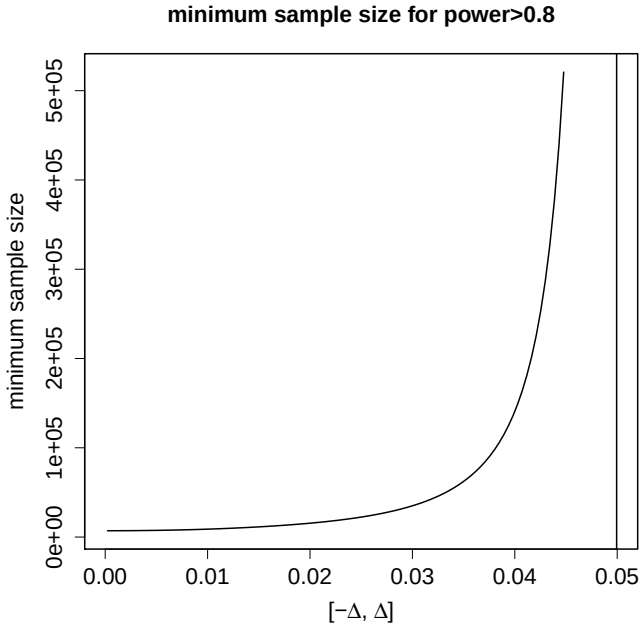

Figure 5: Sample size needed for achieving power>0.8 under different allowance of sensitivity. The vertical line stands for the design sensitivity 0.0499. Here $\lambda = 0.234, \delta = 0, \gamma = 0.158, \sigma_1^2 = 0.333, \sigma_2^2 = 1.0989, \rho = 0.548$.

Figure 5 further explores the relationship between the sample size needed and the allowance of sensitivity. We can see the curve starts flat for a while and then turns steep sharply. This suggests that if the allowance of sensitivity is within a small range and we want to perform sensitivity analysis, we do not need to increase the sample size much to achieve the same power as without sensitivity analysis ($\delta = 0$). However, after a certain threshold, even allowing an extra little amount of sensitivity will result in a large increase of the sample size. In this scenario the design sensitivity is 0.0499. If the range of sensitivity is greater than $(-0.0499, 0.0499)$, then the power will be close to zero no matter how large the sample size.

Table 1: Power for rare(common) variants under different sample size and sensitivity. We set $\sigma_1 = \sigma_2 = 1, \rho = 0.5, \lambda = 1$. For rare variant, $\gamma_r = 0.142, \text{SD}(Z)_r = 0.071$ and for common variant, $\gamma_c = 0.046, \text{SD}(Z)_c = 0.218$. In doing so rare and common variants have the same concentration parameter under the same sample size. The numbers in parentheses represent the power for common variants

	$n = 10^3$	$n = 10^4$	$n = 10^5$	$n = 10^6$
Concentration Parameter	0.1	10	100	1000
Δ=0	0.054 (0.054)	0.089 (0.089)	0.447 (0.447)	0.999 (0.999)
Δ=0.02	0.054 (0.052)	0.086 (0.03)	0.377 (0.116)	0.997 (0.409)
Δ=0.05	0.052 (0.042)	0.071 (0.016)	0.175 (0.001)	0.726 (0.000)

2.5.2 Sensitivity Analysis Using Rare vs. Common Variants as IVs

Two different types of genetic variants are common variants (the variant has frequency $> 1\%$) and rare variants (the variant has frequency $< 1\%$). Common variants tend to have small effects while rare variants can have larger effects. See Gibson (2012) and Zuk et al. (2014) for discussion about rare vs. common variants. Most Mendelian randomization studies have focused on using common variants but there is increasing opportunity for using rare variants by making use of next generation sequencing (Zuk et al., 2014). Here we compare the power of using a common variant vs. a rare variant. To make the comparisons, we assume $\sigma_1 = \sigma_2 = 1$, $\rho = 0.5$, $\beta = 1$. Suppose the rare variant takes 0/1 with probability 0.995/0.005 and the IV effect is $\gamma_r = 0.142$ while the common variant takes 0/1 with probability 0.95/0.05 and the IV effect is $\gamma_c = 0.046$. By choosing these IV effect sizes, the rare and common variants have the same concentration parameter under the same sample size. We can use (A.14) to calculate the power under different sample size and sensitivity. We investigate the scenarios where the sample size is $\{10^3, 10^4, 10^5, 10^6\}$ and the sensitivity allowance is $\{(0, 0), (-0.02, 0.02), (-0.05, 0.05)\}$.

Table 1 shows the result. We see that if there's no concern about the IV being invalid ($\Delta = 0$), i.e., no sensitivity analysis, the power for the rare and common variants are exactly the same across different sample sizes since they have the same concentration parameter. However, if we allow for some of amount of IV invalidity and calculate the power of sensitivity analysis, then the rare variant has better power than the common variant.

As discussed in Section 2.4.2 and 2.4.3, for power of sensitivity analysis, the IV effect size $|\gamma|/\sigma_2$ plays a more important role concentration parameter. The rare variant in Table 1 has a larger IV effect size and consequently a higher power of sensitivity analysis. Another thing to be noticed is that when $\Delta = 0.02$, the power of sensitivity analysis increases as sample size increases for both rare and common variants. However, when $\Delta = 0.05$, the power of sensitivity analysis decreases as sample size increases for common variant. This is because the allowance of sensitivity is too large here such that the inequality (2.25) holds and the power of sensitivity analysis goes to zero.

In summary, if a rare variant has a larger effect size than a common variant such that the rareness and effect size balance each other to result in the same concentration parameter for the rare and common variant, then the rare variant has larger power of sensitivity analysis if there is concern about the IV being invalid.

2.6 Discussion

We have developed a method of sensitivity analysis and a power formula of sensitivity analysis for causal studies using IVs based on the AR test. Compared to previously developed methods of sensitivity analysis for IVs, our method is robust to weak IVs. We have shown that when designing causal studies using IVs in which there is concern that the IV might not be perfectly valid, the key strength parameter one should consider about the IV is not the IV's concentration parameter, as has previously been done, but instead the effect size of the IV on the exposure. This suggests that IVs based on rare genetic variants with large effects will be less sensitive to bias than IVs based on common genetic variants with small effects.

Currently, almost all Mendelian randomization studies have used common variants, but next generation sequencing techniques such as whole exome sequencing facilitate the use of rare variants in Mendelian randomization studies. Next generation sequencing techniques also facilitate the possibility of using variation in the structure of a person's chromosome (e.g.,

deletions, duplications, copy-number variants, insertions, inversions and translocations) in Mendelian randomization studies; such structural variation often has larger effects than common SNP variants. Our findings suggest that Mendelian randomization studies can be made less sensitive to bias by harnessing rare variants with large effect sizes as IVs.

CHAPTER 3 : ivmodel: An R Package for Inference and Sensitivity Analysis of Instrumental Variables Models with One Endogenous Variable

3.1 Introduction

The instrumental variables (IV) method is a popular method to estimate the casual effect of a treatment, exposure, or policy on an outcome when there is concern about unmeasured confounding (Angrist et al., 1996b; Angrist and Krueger, 2001; Baiocchi et al., 2014b). IV methods have been widely used in many field including statistics (Angrist et al., 1996b), economics (Angrist and Krueger, 2001), genomics and epidemiology (Davey Smith and Ebrahim, 2003), sociology Bollen (2012), psychology (Gennetian et al., 2008), political science (Sovey and Green, 2011), and countless others. We also note that instrumental variables have been used to correct for measurement errors (see Fuller (2006) for a full treatment on measurement errors).

Informally speaking, IV methods rely on having variables called instruments which are related to the exposure and are exogenous. An instrument is exogenous if it only affects the outcome through the pathway of affecting the exposure (i.e. the instrument has no direct effect on the outcome) and is independent of unmeasured confounders (see Section 3.2.3 for details). Typically, instruments either come from (i) natural experiments whereby the instruments were naturally assigned to individuals at random or (ii) an actual randomized experiment whereby the actual randomization mechanism is used as an instrument. For example, in a field known as Mendelian randomization, natural genetic variations have been used as an instrument to answer causal questions in epidemiology (Davey Smith and Ebrahim, 2003, 2004; Lawlor et al., 2008b). Another example of the use of an instrument is in the study of the effect of pregnant mother's smoking on birth weight by Sexton and Hebel (1984) and Permutt and Hebel (1989). Here, the instrument was the actual randomized

encouragement assignment of mothers to one of the two groups, the first group where the healthcare provider encouraged the mothers to stop smoking and the second group where the healthcare provider did not provide such encouragement. Table 2 illustrates other examples of instrumental variables, divided based on the source of the instruments. For more examples, see Angrist and Krueger (2001) and Baiocchi et al. (2014b).

Outcome	Exposure	Instruments	Reference
		Natural experiments / Mendelian randomization	
Earnings	Years of schooling	Proximity to college when growing up	Card (1995)
Earnings	Years of schooling	Quarter of birth	Angrist and Krueger (1991)
Metabolic phenotypes	C-reactive protein (CRP)	SNPs rs1800947, rs1130864, rs1205	Timpson et al. (2005)
Blood pressure	Alcohol intake	Alcohol dehydrogenase (ALDH2) genotype	Chen et al. (2008)
		Randomized experiments / Encouragement designs	
Birth weight	Mother's smoking	Randomized encouragement to stop smoking	Sexton and Hebel (1984) and Permutt and Hebel (1989)
Test scores	Class size	Randomized assignment to different class sizes	Krueger (1999)

Table 2: Application of instrumental variables methods based on source of instruments. Natural experiments/Mendelian randomization refer to instrumental variables studies where the instruments come from natural sources, such as genes or calendar years. Randomized experiments/encouragement designs refer to instrumental variables studies where the instruments represent actual randomization mechanisms.

Software for running instrumental variables methods varies widely depending on the programming language. For example, in STATA, there are comprehensive and unified programs to handle the most popular instrumental variables methods, most notably **ivreg2** (Baum et al., 2003, 2007) and STATA's default program **ivregress**. In R, different instrumental variables methods are implemented across different packages, for instance **AER** by Kleiber and Zeileis (2008), **sem** by Fox et al. (2014), and **lfe** by Gaure (2013). Unfortunately, these packages do not include (i) modern instrumental variables methods that provide confidence intervals that are fully robust to weak instruments (see Section 3.4), (ii) sensitivity analysis

methods that examine sensitivity of inference to violations of IV assumptions (see Section 3.5), and (iii) power calculations for IV analysis (see Section 3.6).

The goal of the paper is to present a package **ivmodel** that integrates and unifies R functions to conduct a comprehensive instrumental variables analysis when there is one exposure/endogenous variable. These functions include a general class of estimators known as k-class estimators (see Section 3.3) and the corresponding standard errors, confidence intervals, and p-values. The functions also integrate more modern approaches to IV analysis, including two methods for confidence intervals that are fully robust to weak instruments (Stock et al., 2002b), the Anderson and Rubin confidence interval (Anderson and Rubin, 1949b) and the conditional likelihood ratio confidence interval (Moreira, 2003). The package includes functions to calculate power. Finally, the package includes methods to conduct sensitivity analysis to examine the sensitivity to the IV assumptions not holding. All these functions are integrated into an R software package called **ivmodel**.

3.2 Instrumental Variables Model for One Endogenous Variable

3.2.1 Notation

Let there be n individuals indexed by $i = 1, \ldots, n$. For each individual i, we observe the outcome $Y_i \in \mathbb{R}$, the exposure $D_i \in \mathbb{R}$, the L instruments $Z_{i\cdot} \in \mathbb{R}^L$, and the p covariates $X_{i\cdot} \in \mathbb{R}^p$. Let $Y = (Y_1, \ldots, Y_n) \in \mathbb{R}^n$ denote the vector of outcomes, $D = (D_1, \ldots, D_n) \in \mathbb{R}^n$ denote the vector of exposures, $Z \in \mathbb{R}^{n \times L}$ denote the matrix of instruments where the ith row corresponds to the vector $Z_{i\cdot}$, and $X \in \mathbb{R}^{n \times p}$ denote the matrix of covariates where the ith row corresponds to the vector $X_{i\cdot}$. Let $W = [Z : X]$ where W is an n by $L + p$ matrix that concatenates the matrices Z and X.

For any matrix M, denote its transpose M^T. Also, for any matrix M, let $P_M = M(M^T M)^{-1} M^T$ be the orthogonal projection matrix onto the column space of M and R_M be the residual

projection matrix so that $R_M + P_M = I$, where I is an n by n identity matrix. We assume that M has a proper inverse, so that $(M^T M)^{-1}$ is well-defined, unless otherwise stated.

3.2.2 Model

We assume the following linear structural model between the observed quantities, $Y_i, D_i, Z_{i.}$, and $X_{i.}$

$$Y_i = D_i\beta + X_{i.}^T\kappa + \epsilon_i, \quad \mathsf{E}(\epsilon_i|Z_{i.}, X_{i.}) = 0, \mathsf{VAR}(\epsilon_i|Z_{i.}, X_{i.}) = \sigma^2 \tag{3.1}$$

This is the standard, single equation homoscedastic linear structural model in econometrics (Wooldridge, 2010). Note that this is not the usual regression model in the sense that D_i is correlated with ϵ_i. The parameter of interest is β, which is the causal effect of the exposure D_i on the outcome Y_i (see next paragraph for more details on causal effects). The parameter κ relates the covariates to the outcome. Note that $X_{i.}$ can contain a value of 1 to represent the intercept.

The parameters in model (3.1) can be given a causal interpretation under the potential outcomes notation (Rubin, 1974b) where (3.1) represents the additive linear, constant effects (ALICE) model in Holland (1988b). Let $Y_i^{(d,z)}$ be the potential outcome if individual i were to have exposure d, a scalar value, and instruments z, an L dimensional vector. Let $D_i^{(z)}$ be the potential exposure if the individual had instruments z. For each individual, only one possible realizations of $Y_i^{(d,z)}$ and $D_i^{(z)}$ is observed, denoted as Y_i and D_i, respectively, based on his/her observed instrument values $Z_{i.}$ and exposure D_i. Then, for two possible values of the exposure d', d and instruments z', z, we assume the following potential outcomes model

$$Y_i^{(d',z')} - Y_i^{(d,z)} = (d' - d)\beta \quad \mathsf{E}(Y_i^{(0,0)} \mid Z_{i.}, X_{i.}) = X_{i.}^T\kappa \tag{3.2}$$

In model (3.2), β represents the causal effect (divided by $d' - d$) of changing the exposure from d' to d on the outcome. The parameter κ represents the impact of covariates on the

baseline potential outcome $Y_i^{(0,0)}$. If we further define $\epsilon_i = Y_i^{(0,0)} - \mathsf{E}(Y_i^{(0,0)} \mid Z_{i.}, X_{i.})$, we obtain the observed data model in (3.1), thus providing the parameters in the observed model in (3.1) a causal interpretation.

Note that in many works in econometrics literature, one makes additional assumptions about the relationship between the endogenous variable D_i, the instruments $Z_{i.}$, and the covariates $X_{i.}$, specifically

$$D_i = Z_{i.}^T \gamma + X_{i.}^T \tilde{\kappa} + \eta_i, \quad \mathsf{E}(\eta_i | Z_{i.}, X_{i.}) = 0, \mathsf{VAR}(\eta_i | Z_{i.}, X_{i.}) = \omega^2 \tag{3.3}$$

This "first stage" model in (3.3) is not necessary for all our methods in the **ivmodel** package. In particular, the k-class estimators in Section 3.3 and the confidence interval for the Anderson and Rubin test in Section 3.4 are valid without the first stage modeling assumption in (3.3). However, the other methods presented in the paper require this assumption and we introduce it in this section.

Similar to equation (3.2), we can provide a causal interpretation of the first stage model in (3.3) as follows.

$$D_i^{(z')} - D_i^{(z)} = (z' - z)\gamma \quad \mathsf{E}(D_i^{(0)} \mid Z_{i.}, X_{i.}) = X_{i.}^T \tilde{\kappa} \tag{3.4}$$

In model (3.4), γ represents the causal effect (divided by $z' - z$) of changing the IV from z' to z on the exposure. The parameter $\tilde{\kappa}$ represents the impact of covariates on the baseline potential outcome $D_i^{(0)}$. As before, if we further define $\eta_i = D_i^{(0)} - \mathsf{E}(D_i^{(0)} \mid Z_{i.}, X_{i.})$, we obtain the observed data model in (3.3).

Without loss of generality and throughout the paper, we will use the simplified version of the models in equations (3.1) and (3.3) where we project out the covariates X by the Frisch-Waugh-Lovell Theorem (Davidson and MacKinnon, 1993b). Specifically, models (3.1) and

(3.3) are equivalent to

$$Y_i^* = D_i^* \beta + \epsilon_i^* \tag{3.5}$$

$$D_i^* = Z_{i.}^* \gamma + \eta_i^* \tag{3.6}$$

where

$$Y^* = R_X Y, \quad D^* = R_X D, \quad Z^* = R_X Z, \quad \epsilon^* = R_X \epsilon, \quad \eta^* = R_X \eta$$

The superscripts Y^*, D^*, Z^* represent the outcome, the exposure, and the instruments after controlling for the covariates X by the residual orthogonal projection R_X defined in Section 3.2.1. The equivalent models (3.5) and (3.6) allow us to concentrate on the target parameter of interest, β, and simplify the derivations and expressions of the instrumental variables methods presented in the paper. Note that as before, the model in (3.6), the simplified version of the first stage model in (3.3), is not necessary for k-class estimators and the Anderson and Rubin confidence intervals.

3.2.3 Assumption of Instrumental Variables

Under the model in (3.1), we make the standard assumptions in the instrumental variables literature below (Wooldridge, 2010).

(A1) $\mathsf{E}(W^T W)$ is full rank.

(A2) Conditional on the covariates X, the instruments Z are associated with the exposure D, $\mathsf{E}(Z^T R_X D) \neq 0$

(A3) W is exogenous, $\mathsf{E}(W^T \epsilon) = 0$

Assumption (A1) is a standard moment condition on the matrix of exogenous variables that include the covariates and the instruments. Assumption (A2) states that conditional on the covariates X, the instruments are associated with the exposure. There are many ways to test this assumption in practice, the most popular being the F statistic for the

coefficients on the variables in Z being 0 when regressing D on X and Z. A strong association between the instruments Z and the exposure D is desired to reduce the precision of an IV estimator (Stock et al., 2002b). Instruments with strong associations are considered to be strong instruments while instruments with weak associations are considered to be weak instruments. For example, in the case of one instrument, an instrument is considered weak if the F statistic is less than 10 (Stock et al., 2002b).

For assumption (A3), in the ALICE model, (A3) is satisfied if Z has no direct on D and Z is independent of unmeasured confounders. Assumption (A3) is generally untestable in that it's impossible to check whether the exogenous variables Z and X are uncorrelated with the structural error ϵ_i, which is never observed. However, methods exist to partially test this assumption if there are more than one instruments, $L > 1$, the most popular being the Sargan's test (Sargan, 1958). Under all the three assumptions (A1)-(A3), standard econometric arguments show that the the model (3.1) is identified (Wooldridge, 2010).

Typically, practitioners assume that they have found instruments that satisfy (A1)-(A3) (Angrist and Krueger, 2001). However, violations of these assumptions occur, especially (A2) and (A3), and there has been progress in the literature to handle these violations (Angrist and Krueger, 2001; Murray, 2006). For (A2), even if it is satisfied, but only weakly, which is known as the weak instrument problem, the most commonly used instrumental variables estimation method, two stage least squares (TSLS), produces biased estimates of β in (3.1) (Nelson and Startz, 1990b; Staiger and Stock, 1997b; Stock et al., 2002b). Thankfully, many statistical methods exist to provide robust and honest estimates of the parameters in model (3.1) with weak instruments (Stock et al., 2002b) (see Section 3.4 for details). Violations of (A3), known as the invalid instrument problem, is the case where the instruments Z may have a direct effect on the outcome or when the instruments are correlated with ϵ_i. This problem has received less attention than the weak instrument problem (Murray, 2006), but has recently been considered by Kolesár et al. (2013), Kang et al. (2015), and Jiang et al. (2015).

Throughout the paper, we assume that our instruments Z satisfy assumptions (A1)-(A3). However, we discuss violations of (A2) and (A3) in Sections 3.4 and 3.5, respectively and provide methods that can handle these violations.

3.3 k-Class Estimation and Inference

3.3.1 Definitions and General Properties

A class of estimators for β, called the k-class estimator and denoted as $\hat{\beta}_k$, is defined as follows.

$$\hat{\beta}_k = (D^{*T}(I - kP_{Z^*})D^*)^{-1}D^{*T}(I - kP_{Z^*})Y^* \tag{3.7}$$

Table 3 lists all the estimators that are k-class estimators, including the ordinary least squares (OLS), two-stage least squares (TSLS), limited information maximum likelihood (LIML), and Fuller's estimator (FULL). In Table 3, k_{LIML} is the minimum value of k of the following equation

$$det \begin{pmatrix} Y^{*T}(I - kR_{Z^*})Y^* & Y^{*T}(I - kR_{Z^*})D^* \\ D^{*T}(I - kR_{Z^*})Y^* & D^{*T}(I - kR_{Z^*})D^* \end{pmatrix} = 0 \tag{3.8}$$

k	Name
$k = 0$	Ordinary least squares (OLS)
$k = 1$	Two-stage least squares (TSLS)
$k = k_{LIML}$	Limited information maximum likelihood (LIML)
$k = k_{LIML} - \frac{b}{n-L-p}, b > 0$	Fuller's estimator (FULL)

Table 3: Different types of k-class estimator

Each k yields an estimator with unique properties, which will be discussed in detail in Section 3.3.2. However, for all k-class estimators in equation (3.7), an estimate for the standard error of $\hat{\beta}_k$ is

$$\widehat{\mathsf{VAR}(\hat{\beta}_k)} = \hat{\sigma}^2(D^{*T}(I - kP_{Z^*})D^*)^{-1}, \quad \hat{\sigma}^2 = \frac{(Y^* - D^*\hat{\beta}_k)^T(Y^* - D^*\hat{\beta}_k)}{n - L - p} \tag{3.9}$$

As long as L and p are fixed, all k-class estimators are consistent so long as $k \to 1$ as $n \to \infty$ in probability (Davidson and MacKinnon, 1993b). In addition, for fixed L and p, as long as $\sqrt{n}(k-1) \to 0$ in probability as $n \to \infty$, the k-class estimator has the following asymptotic Normal distribution (Amemiya, 1985)

$$\frac{\hat{\beta}_k - \beta}{\sqrt{\widehat{\mathsf{VAR}(\hat{\beta}_k)}}} \to N(0,1) \tag{3.10}$$

The asymptotic distribution in (3.10) allows us to test the hypothesis

$$H_0 : \beta = \beta_0, \quad H_a : \beta \neq \beta_0 \tag{3.11}$$

by comparing the standardized deviate in (3.10) to the standard Normal (or the t distribution with degrees of freedom $n - L - p$). We can also create $1 - \alpha$ confidence intervals for β based on $\hat{\beta}_k$, i.e.

$$\left(\hat{\beta}_k - z_{1-\alpha/2}\sqrt{\widehat{\mathsf{VAR}(\hat{\beta}_k)}}, \quad \hat{\beta}_k + z_{1-\alpha/2}\sqrt{\widehat{\mathsf{VAR}(\hat{\beta}_k)}}\right)$$

where $z_{1-\alpha/2}$ is the $1 - \alpha/2$ quantile of the standard Normal distribution. We can alternatively use the $1 - \alpha/2$ quantile of the t distribution with degrees of freedom $n - L - p$.

3.3.2 Some Examples of k-class Estimators

The most well-known k-class estimator in instrumental variables is two-stage least squares (TSLS) where $k = 1$ in (3.7), i.e.

$$\hat{\beta}_1 = (D^{*T} M_{Z^*} D^*)^{-1} D^{*T} M_{Z^*} Y^*$$

In addition to being consistent and having an asymptotic Normal distribution, TSLS is efficient among all IV estimators using linear combination of instruments Z (Wooldridge, 2010). In fact, under the asymptotics rates of $\sqrt{n}(k-1) \to 0$ introduced in Section 3.3.1, all k-class estimators have the same asymptotic Normal distribution as TSLS. Also, when

$L = 1$, TSLS and LIML produce identical estimates of β (Davidson and MacKinnon, 1993b).

However, despite having the same asymptotic Normal distributions, TSLS and other types of k-class estimators, for instance LIML and FULL in Table 3, behave differently in finite-samples. With weak instruments (i.e. near violations of (A2)), TSLS tends to be biased towards OLS in finite sample. Even with large samples, TSLS can provide a very biased estimate of the causal effect in the presence of weak instruments (Bound et al., 1995). In contrast, LIML and FULL are more robust to violations of (A2) than TSLS (Stock et al., 2002b). However, LIML has no finite moments of any order while TSLS has moments of up to $L-1$. FULL corrects LIML's lack of moments by having moments so long as the sample size is large enough (Davidson and MacKinnon, 1993b).

Other types of k-class estimators exist and no single k-class estimator uniformly dominates another in all settings (Davidson and MacKinnon, 1993b). However, in practice, the most popular estimators are TSLS and LIML, with LIML having better robustness properties with regards to weak instruments (Stock et al., 2002b; Mariano, 2003; Chao and Swanson, 2005)

3.4 Dealing with Weak Instruments: Robust Confidence Interval Estimation

In this section, we discuss the case when the instruments Z may nearly violate (A2), also known as the weak instrument problem, and discuss two methods that are fully robust to near violations of (A2).

Let M be an n by 2 matrix where the first column contains Y^* and the second column contains D^*. Let $a_0 = (\beta_0, 1)$ and $b_0 = (1, -\beta_0)$ to be two-dimensional vectors and $\hat{\Sigma} = M^T R_{Z^*} M/(n - L - p)$. Let $\hat{S}$ and $\hat{T}$ be two-dimensional vectors defined as follows.

$$\hat{S} = \frac{(Z^{*T} Z^*)^{-1/2} Z^{*T} M b_0}{\sqrt{b_0^T \hat{\Sigma} b_0}}, \quad \hat{T} = \frac{(Z^{*T} Z^*)^{-1/2} Z^{*T} M \hat{\Sigma}^{-1} a_0}{\sqrt{a_0^T \hat{\Sigma}^{-1} a_0}}$$

We also define the following scalar values, $\hat{Q}_1, \hat{Q}_2$, and $\hat{Q}_3$.

$$\hat{Q}_1 = \hat{S}^T\hat{S}, \quad \hat{Q}_2 = \hat{S}^T\hat{T}, \quad \hat{Q}_3 = \hat{T}^T\hat{T}$$

Based on $\hat{Q}_1$, $\hat{Q}_2$, and $\hat{Q}_3$, we define two tests of the hypothesis in equation (3.11) that are fully robust to violations of (A2), the Anderson and Rubin test (Anderson and Rubin, 1949b), and the conditional likelihood test (Moreira, 2003).

$$AR(\beta_0) = \hat{Q}_1/L \tag{3.12}$$

$$CLR(\beta_0) = \frac{1}{2}(\hat{Q}_1 - \hat{Q}_3) + \frac{1}{2}\sqrt{(\hat{Q}_1 + \hat{Q}_3)^2 - 4(\hat{Q}_1\hat{Q}_3 - \hat{Q}_2^2)} \tag{3.13}$$

Much work has shown that these two tests are fully robust to weak instruments (Staiger and Stock, 1997b; Stock et al., 2002b; Moreira, 2003; Dufour, 2003; Andrews et al., 2006b). Between the two tests, there is no uniformly most powerful test under weak instruments, but Andrews et al. (2006b) and Mikusheva (2010) suggest using (3.13) due to its generally favorable power compared to (3.12) in most cases when weak instruments are present. However, the Anderson-Rubin test is the simplest of the two tests in that under a Normality error assumption (see next paragraph), it can be written as a standard F-test in regression where the outcome is $R_{Z^*}(Y - D\beta_0)$, the regressors are Z^*, and we are testing whether the coefficients associated with Z^* are zero or not with the standard F-test. Also, unlike the Anderson and Rubin test in (3.12), the conditional likelihood ratio test in (3.13) requires the first stage modeling assumption in (3.3) (Dufour, 2003).

We can invert both tests in equation (3.12) and (3.13) to obtain $1 - \alpha$ confidence intervals that are fully robust to weak instruments, i.e. $\{\beta : AR(\beta_0) \leq F_{L,n-L-p,1-\alpha}\}$ for the Anderson and Rubin confidence interval and $\{\beta : CLR(\beta_0) \leq q_{1-\alpha}\}$ for the conditional likelihood ratio test. Here, $F_{L,n-L-p,1-\alpha}$ is the $1 - \alpha$ quantile of the F distribution with degrees of freedom L and $n - L - p$ and $q_{1-\alpha}$ is the $1 - \alpha$ quantile of the the conditional likelihood ratio test. The F distribution for the Anderson and Rubin test is based on an

assumption about Normality of the errors in model (3.1) and our package **ivmodel** currently uses the F distribution. However, one can also use the χ^2 distribution as an asymptotic approximation should the Normality assumption be unreasonable in the data. As for the distribution that underlies the conditional likelihood ratio test and the details on $q_{1-\alpha}$, see Andrews et al. (2007).

3.5 Dealing with Possibly Invalid Instruments: Sensitivity Analysis

Morgan and Winship (2007b) showed that assumption (A3) cannot be completely tested. However, there is often concern that a putative IV is invalid in applications. In these cases, a sensitivity analysis can be used to examine the sensitivity of inferences to violations of (A3). Here we assume that there is only one IV ($L = 1$) in the study and this IV may be invalid to some degree.

There is some previous work on sensitivity analysis for IV studies, see DiPrete and Gangl (2004b), Small (2007b), Kolesár et al. (2011b) and Conley et al. (2012b). These papers all use test statistics which are based on the TSLS estimator having an approximately normal distribution, which breaks down in the presence of weak instruments (instruments that are weakly associated with the exposure), see Nelson and Startz (1990b). Our sensitivity analysis uses the AR test statistic because of the following properties: the AR test is robust to a weak instrument; the AR test is uniformly most powerful among all unbiased tests (Moreira, 2001b) and the AR test is uniformly most powerful among all invariant similar tests (Andrews et al., 2006b).

We revise the model in Section 4.3 to add the feature of invalid IV. For model (3.2), assume that Z_i violates the assumption (A3) so there is another term $\delta\sigma(z' - z)$ on the equation's right side:

$$Y_i^{(d',z')} - Y_i^{(d,z)} = (d' - d)\beta + \delta\sigma(z' - z), \quad \mathsf{E}(Y_i^{(0,0,0)} \mid Z_{i.}, X_{i.}) = X_{i.}^T\kappa \tag{3.14}$$

Here σ is the standard variation of $\epsilon_i = Y_i^{(0,0,0)} - \mathsf{E}(Y_i^{(0,0,0)} \mid Z_{i.}, X_{i.})$, which is a scaling parameter. δ measures how much the IV $Z_{i.}$ violates the assumption (A3). Further assume the sensitivity parameter is within a known range, $\delta \in (\underline{\delta}, \bar{\delta})$. Then the model for sensitivity analysis becomes:

$$Y_i = D_i\beta + X_{i.}^T\kappa + \delta\sigma Z_{i.} + \epsilon_i, \quad \mathsf{E}(\epsilon_i|Z_{i.}, X_{i.}) = 0, \quad \mathsf{VAR}(\epsilon_i|Z_{i.}, X_{i.}) = \sigma^2, \quad \delta \in (\underline{\delta}, \bar{\delta}) \tag{3.15}$$

If the error term has a normal distribution $\epsilon_i \sim N(0, \sigma^2)$, then hypothesis (3.11) can be tested by using the AR test statistic $AR(\beta_0)$ in equation (3.12). Under H_0, $AR(\beta_0)$ has a non-central F distribution :

$$AR(\beta_0) \sim F_{1,n-p-1,\delta^2 Z^{*T} Z^*} \tag{3.16}$$

Although δ is unknown and consequently we don't known exact the distribution of $AR(\beta_0)$ under H_0, we can define $\Delta = \max(|\underline{\delta}|, |\bar{\delta}|)$ and construct an interval that provides at least $1 - \alpha$ confidence:

$$CI_{1-\alpha} = \{\beta : AR(\beta_0) < F_{1,n-p-1,\Delta^2 Z^{*T} Z^*;1-\alpha}\} \tag{3.17}$$

For our sensitivity analysis, equation (3.17) is used for the hypothesis test. More details are provided in Jiang et al. (2015).

3.6 Power

If the research goal is to find evidence for an exposure effect, then we would like to know the power of rejecting the null hypothesis $H_0 : \beta = \beta_0$ when the true exposure effect is under the alternative $\beta - \beta_0 = \lambda \neq 0$. With a power formula, researchers can decide how large a sample to collect to achieve a certain power. Freeman et al. (2013b) presents a power formula for using the asymptotic normal distribution of TSLS estimator to do hypothesis test. Jiang et al. (2015) provides a power formula for the AR test and sensitivity analysis. These three different power formulas are included in **ivmodel**. By inverting the power formula, we can

calculate the sample size needed to achieve a certain power. These sample size calculation functions are also included in **ivmodel**. The three different approaches to a power formula rely on different assumptions, which will be discussed in the following paragraphs.

Freeman et al. (2013b) assumes there is only one IV($L = 1$) and there are no observed covariates $X(p = 0)$, which is model (3.1) with $\kappa = 0$. Asymptotically, the TSLS estimator has a normal distribution:

$$\hat{\beta}_{TSLS} \sim N\left(\beta, \quad \frac{\sigma^2}{n \cdot \mathsf{VAR}(D) \cdot \rho_{ZD}}\right) \tag{3.18}$$

If the true exposure effect is $\beta - \beta_0 = \lambda$, then the power of testing hypothesis (3.11) is:

$$\text{Power} = 1 + \Phi\left(-z_{\alpha/2} - \frac{\lambda\rho_{ZD}\sqrt{n \cdot \mathsf{VAR}(D)}}{\sigma}\right) - \Phi\left(z_{\alpha/2} - \frac{\lambda\rho_{ZD}\sqrt{n \cdot \mathsf{VAR}(D)}}{\sigma}\right) \tag{3.19}$$

where α is the desired significance level of the test (usually 0.05), Φ is the cumulative distribution function of the standard normal distribution, z_α is the value satisfies $\Phi(-z_\alpha) = \alpha$ and ρ_{ZD} is the correlation between Z and D.

The AR test is based on model (3.1). In order to calculate the exact power, we need to have another model between the exposure and IV, as stated in (3.3) and make the bivariate normality assumption for the error (ϵ_i, η_i). The extended model is summarized as follows.

$$\begin{aligned}
Y^* &= D^*\beta + \epsilon^* \\
D^* &= Z^*\gamma + \eta^* \\
Y^* = R_X Y, \quad D^* = R_X D, \quad Z^* &= R_X Z, \quad \epsilon^* = R_X \epsilon, \quad \eta^* = R_X \eta \\
(\epsilon, \eta) \perp Z; \quad (\epsilon_i, \eta_i)^T \sim N(\mathbf{0}, \Sigma); \quad \Sigma &= \begin{pmatrix} \sigma^2 & \rho\sigma\omega \\ \rho\sigma\omega & \omega^2 \end{pmatrix} \quad \text{rank}(X) = p
\end{aligned} \tag{3.20}$$

If the true exposure effect is $\beta - \beta_0 = \lambda$, then the power of testing hypothesis (3.11) is:

$$\text{Power} = 1 - \Psi_{1,n-p-L,\frac{(\gamma^T Z^{*T} Z^* \gamma)\lambda^2}{\sigma^2+2\rho\sigma\omega\lambda+\omega^2\lambda^2}}\left(F_{1,n-p-L;1-\alpha}\right) \tag{3.21}$$

where $F_{a,b;1-\alpha}$ is the $1-\alpha$ quantile of F distribution with degrees of freedom a and b. $\Psi_{a,b,k}(\cdot)$ is the cumulative distribution function of the non-central F distribution with degree of freedom a, b and non-central parameter k.

The sensitivity analysis relies on model (3.15) with one possibly invalid IV. To calculate its power, the model (3.3) and normality assumption is still needed. The extended model is similar to (3.20).

$$\begin{gathered} Y^* = D^*\beta + \delta\sigma Z^* + \epsilon^* \\ D^* = Z^*\gamma + \eta^* \\ Y^* = R_X Y; \quad D^* = R_X D; \quad Z^* = R_X Z; \quad \epsilon^* = R_X \epsilon; \quad \eta^* = R_X \eta; \\ (\epsilon, \eta) \perp Z; \quad (\epsilon_i, \eta_i)^T \sim N(\mathbf{0}, \Sigma); \quad \Sigma = \begin{pmatrix} \sigma^2 & \rho\sigma\omega \\ \rho\sigma\omega & \omega^2 \end{pmatrix} \quad \text{rank}(X) = p \end{gathered} \tag{3.22}$$

If the true exposure effect is $\beta - \beta_0 = \lambda$ and it is a favorable situation where the instrument is valid ($\delta = 0$) but we want to allow for the possibility that the instrument is invalid in the range $\delta \in (-\Delta, \Delta)$ (Rosenbaum, 2010b), then the power of being able to reject the null hypothesis for all $\delta \in (-\Delta, \Delta)$ is:

$$\text{Power} = 1 - \Psi_{1,n-p-1,\frac{\lambda^2\gamma^2 Z^{*T} Z^*}{\sigma^2+2\rho\sigma\omega\lambda+\omega^2\lambda^2}}\left(F_{1,n-p-1,\Delta^2 Z^{*T} Z^*;1-\alpha}\right) \tag{3.23}$$

where $F_{a,b,c;1-\alpha}$ is the $1-\alpha$ quantile of the non-central F distribution with degree of freedom a, b and non-central parameter c. (3.23) is called the power of sensitivity analysis for sensitivity Δ.

When the sample size is small or moderate and the IV is weak, the asymptotic test (3.18)

based on the two stage least squares estimator can have highly inflated Type I error and should be avoided (Jiang et al., 2015). For these settings, Jiang et al. (2015) recommend using the AR test and its associated power formula (3.21).

3.7 Application

In this section, we illustrate the application of **ivmodel** with the data set from Card (1995). The data is from the National Longitudinal Survey of Young Men (NLSYM), which has $n = 3010$ individual observations and 35 variables. We want to estimate the causal effect of education (`educ`) on log earnings (`lwage`). The IV is a binary variable indicating whether the individual grew up in a place with a nearby 4-year college (`nearc4`). There are also some exogenous variables included in the data (`exper`, `expersq`, `black`, `south`, etc).

3.7.1 Ivmodel Class and the Basic Usage

As discussed above, we specify the outcome Y is log earnings (`lwage`); the exposure D is (`educ`); the IV Z is (`nearc4`); other exogenous variables X is (`exper`, `expersq`, `black`, `south`). Then we can use `ivmodelFormula()` to generate an ivmodel class object which assembles various IV methods.

```
R> cardfit = ivmodelFormula(lwage ~ educ + exper + expersq + black + south |
R+                  nearc4 + exper + expersq + black + south, data=card.data)
```

However, when there are many exogenous variables, the formula style of input is not convenient, then we can use function `ivmodel()` instead of `ivmodelFormula()`:

```
R> Y = card.data[,"lwage"]
R> D = card.data[,"educ"]
R> Z = card.data[, "nearc4"]
R> Xname = c("exper", "expersq", "black", "south", "smsa",
R+           paste("reg", 661:668, sep=""), "smsa66")
R> X = card.data[, Xname]
```

```
R> cardfit = ivmodel(Y=Y, D=D, Z=Z, X=X)
```

After `ivmodel` object is generated, use `print.summary.ivmodel()` to display the information:

```
R> summary(cardfit)

Call:
ivmodel(Y = Y, D = D, Z = Z, X = X)
sample size: 3010
- - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -

First Stage Regression Result:


F=13.25579, df1=1, df2=2994, p-value is 0.00027634
R-squared=0.004407934,   Adjusted R-squared=0.004075405
Residual standard error: 1.940537 on 2995 degrees of freedom
- - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -

Coefficients of k-Class Estimators:

              k Estimate Std. Error t value Pr(>|t|)
OLS    0.000000 0.074693   0.003498  21.351   <2e-16 ***
Fuller 0.999666 0.127501   0.052708   2.419   0.0156 *
LIML   1.000000 0.131504   0.054964   2.393   0.0168 *
TSLS   1.000000 0.131504   0.054964   2.393   0.0168 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
- - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -
```

```
Alternative tests for the treatment effect under H_0: beta=0.


Anderson-Rubin test:
F=5.415279, df1=1, df2=2994, p-value=0.020028
95 percent confidence interval:
 [ 0.0248048359650698 , 0.284823593339102 ]


Conditional Likelihood Ratio test:
Test Stat=5.415279, p-value=0.020028
95 percent confidence interval:
 [0.0248043722947518, 0.284824550721994]
```

There are four sections in the display. The first section is a recall for the ivmodel expression and the sample size. The second section summarizes the first stage regression between the IV and exposure. Here the F statistic is 13.25579, which is greater than 10, indicating the IV is not weak (Stock et al., 2002b), so TSLS estimator will not be largely biased. The third section lists the results for several k-class estimator. The default k's are $k = 0$ (OLS), $k = 1$ (TSLS), and k's associated with LIML and Fuller. Here we only have one IV, so TSLS and LIML are the same. The estimated causal effect for TSLS is 0.1315, with a significant p value 0.0168. This can be interpreted as when increasing education by 1 level, the earnings will increase by 13%. The last section provides the AR and CLR confidence intervals, which are robust even for weak IVs (although in this case the IV is not weak). Here we can see both confidence intervals don't cover 0, so the causal effect is significant even under robust tests. The p value for robust test is 0.02, larger than 0.0168 in TSLS, we are trading the test power with robustness.

The method `confint.ivmodel()` calculates the confidence interval for various IV methods. Similarly, we also provide methods `coef.ivmodel()`, `fitted.ivmodel()`, `residuals.ivmodel()`.

```
R> confint(cardfit)
```

```
              2.5%      97.5%
OLS     0.06783385 0.08155266
Fuller  0.02415275 0.23084946
LIML    0.02373345 0.23927422
TSLS    0.02373345 0.23927422
AR      0.02480484 0.28482359
CLR     0.02485469 0.28472068
```

3.7.2 Power Calculation and Sample Size

Suppose the TSLS estimator is the real causal effect $\beta = \hat{\beta}_{TSLS}$. In this case, if we still test the null hypothesis that there is no causal effect $H_0 : \beta = 0$, then the probability to reject H_0 is defined as the power. **ivmodel** provides power calculation function `IVpower()` for TSLS, AR test and sensitivity analysis. See section 3.6 for details of the power formula. In this example, the power of TSLS is 0.668 and the power of AR test is 0.643.

```
R> IVpower(cardfit); IVpower(cardfit, type="AR")

[1] 0.6676418
[1] 0.6432517
```

We can compare the power of TSLS and AR under different sample size, Figure 6 is the output graph.

```
R> ngrid = (1:100)*100
R> plot(IVpower(cardfit, n=ngrid)~ngrid, type="l", lty=1, ylab="power")
R> points(IVpower(cardfit, n=ngrid, type="AR")~ngrid, type="l", lty=2)
R> legend("bottomright", legend=c("TSLS", "AR"), lty=c(1, 2))
```

Usually we want a power of 0.8 or higher. In experimental design, we can increase the sample size needed to achieve the power threshold. `IVsize()` calculates the minimum sample size needed for achieving a certain power threshold. In this example, we need 4,125 sample size

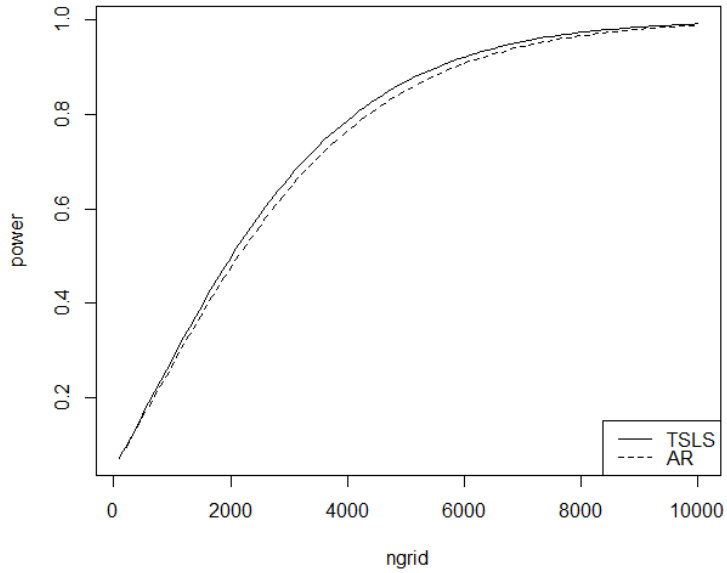

Figure 6: Complete.

for TSLS and 4,362 sample size for AR test.

```
R> IVsize(cardfit, power=0.8); IVsize(cardfit, power=0.8, type="AR")
```

```
[1] 4125
```

```
[1] 4362
```

3.7.3 Sensitivity Analysis

There is usually concern that the IV may be invalid, in this data, there may have confounding factors like geographic or social features that affects both the existence of a nearby 4-year college and the earnings, but not through education. A sensitivity analysis can be performed in such case. Assume that if we hold education and IV fixed, the variation remained in the earnings is σ. We assume the range for sensitivity is $\delta \in (-0.03, 0.03)$, which means a unit change in the invalid IV `near4c` will change the outcome `lwage` by up to 0.03σ in a way not related to the exposure `educ`. To perform the sensitivity analysis, just specify the sensitivity range in `ivmodel()`:

```
R> cardfit2=ivmodel(Y=Y, D=D, Z=Z, X=X, deltarange=c(-0.03, 0.03))
R> summary(cardfit2)

..................
..................
..................
- - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -

Alternative tests for the treatment effect under H_0: beta=0.

Anderson-Rubin test:
F=5.415279, df1=1, df2=2994, p-value=0.020028
95 percent confidence interval:
 [ 0.0248048359650698 , 0.284823593339102 ]

Sensitivity analysis with deltarange [ -0.03 ,  0.03 ]:
non-central F=5.415279, df1=1, df2=2994, ncp=0.4390019, p-value=0.049504
95 percent confidence interval:
 [ 0.000347142651197386 , 0.340944347177351 ]

Conditional Likelihood Ratio test:
Test Stat=5.415279, p-value=0.020028
95 percent confidence interval:
 [0.0248043722947518, 0.284824550721994]
```

The sensitivity analysis is reported in the last section when printing the summary information. The p-value is 0.049, suggesting that education still has a significant positive effects towards earnings even if the IV may be invalid at a certain degree. The power is 0.261 and

we need 17,869 sample size to have a power of at least 0.8.

```
R> IVpower(cardfit2, type="ARsens"); IVsize(cardfit2, power=0.8, type="ARsens")

[1] 0.2615532
[1] 17869
```

3.8 Summary

The package **ivmodel** provides a unified implementation of instrumental variables methods in the case of one endogenous variable. The package contains a general class of estimators, k-class estimators. The package also contains methods that can deal with violations of instrumental variables assumptions, (A2) and (A3). First, for violations of (A2), the package contains two confidence intervals that are fully robust to weak instruments. For (A3), the package contains methods for sensitivity analysis for the range of violation. The package also contains power formulas to guide designs of future instrumental variables studies. As our data example in Section 3.7 demonstrated, our package provides an easy and unified way of conducting a comprehensive instrumental variables analysis with data where there is one endogenous variable, along with ways to assess sensitivity to violations of instrumental variables assumptions and to compute power.

CHAPTER 4 : Hidden Markov Model for Estimating Financial Incentive Effects towards Healthy Behavior

4.1 Introduction

Patient's poor adherence to medications has large impact on their health outcomes and health costs (Peterson et al., 2003). Especially for chronic conditions where long-term medication is needed, the medication will be most efficient only within a short range and patient's consistent adherence is the key to keep the doses in that level. Besides medications, physicians may recommend people to have a minimum level of physical exercise (Go et al., 2013). However, it is difficult for people to maintain these healthy behaviors, either medications or exercises (Jackevicius et al., 2002; Bravata et al., 2007).

There are different methods to improve people's adherence of healthy behavior, such as implementing the support which comes from peers with the same chronic condition (Glasgow and Toobert, 1988); using an electronic medication monitoring system with a daily alarm to remind people take medication as scheduled (Kimmel et al., 2016a); using financial incentives to enhance adherence (Loewenstein et al., 2012; Ries, 2012). For the studies of financial incentives, previous literature only compares the overall effect among treatment group and control group. In this paper, we will focus on analyzing the dynamic effect of lottery-based financial incentive implemented with long-term medication. More specifically, every participant will be informed of a lottery result on a daily basis, if he does adhere on that day and he wins the lottery, he can take the money. If he doesn't adhere but wins the lottery, he cannot take the money. We want to study how does the outcome of lottery dynamically affect patient's adherence. Our study interest can better address questions like if we change the lottery result on a particular day, how much will this affect patient's behavior. Such result can be used for designing an efficient lottery incentive system with

limited budget.

We propose Hidden Markov Model(HMM) with random effects to model participant's longitudinal adherence and lottery results. The hidden state is the participant's tendency(probability) to adhere on each day and the combination of participant's outcome and lottery result determines the tendency to adhere next day. HMM is well-suited to model the dynamic effects in longitudinal data and there are similar works of HMM in behavioral economics, such as Shirley et al. (2010); Altman (2007). EM algorithm and Baum-Welch algorithm (McLachlan and Krishnan, 2007) are used to find the maximum likelihood estimation and 200 bootstrap data are created to construct the confidence interval. The estimation results are compared across 3 different trials with similar lottery incentive system.

The rest of the paper is organized as follows: Section 4.2 describes the data source and trail protocol. Section 4.3 introduces our assumptions and models. Section 4.4 compares the estimation results and Section 4.5 summaries the results and discussions.

4.2 Data source and trial protocols

Our study will be based on three different clinical trials, which share the similar protocols that patients are monitored by a remote device to check whether they adhere to the medication/exercise each day. Each study also has a similar structure of daily lottery system as the intervention of financial incentives. Notice that there is no control group selected in the trial, every patient takes the lottery. We will analysis the dynamic effect of financial incentives in each trail and compare the results among these three cohorts.

4.2.1 Walking steps

207 eligible participants from the University of Pennsylvania are selected to this 13-week trial which starts in Mar. 2014. They are all age 18 years or older and have a body mass index(BMI) of at least 27kg/m^2. The BMI threshold is chosen so the sample can represent

overweight people. Participants are given a goal of achieving at least 7,000 steps per day and they have a smartphone to track the adherence. The lottery system is designed such that every day each participant independently has 1% chance to win $50, 18% chance to win $5. Participant will be informed of the lottery result everyday, but they can collect the money only if they do adhere on that day. More details can be found from Patel et al. (2016), notice that we exclude the control group in the original paper.

4.2.2 Glucose reading

102 African American veterans with persistently poor glycemic control are identified from the Philadelphia VA Medical Center for this 24-week trial starting in Mar. 2011. They are from two arms, one arm receives financial incentives and the other receives both financial incentives and peer mentors. Participants use the provided glucometer to measure and report their fasting glucose values. If they adhere to healthy behavior, the reading values can be controlled within 80-140mg/dl. Lottery system will be implemented to both arms. In financial incentive arm, participant has 1% chance to win $100, 18% chance to win $10 everyday. They will be informed of the lottery result and they can take the money only if they adhere on that day. In financial incentive and peer mentor arm, the participant with split the winning lottery with their mentor, so he has 1% chance to win $50, 18% chance to win $5 everyday. More details can be found as the study NCT01125969 stated in Lorincz et al. (2013), notice that we exclude two arms which don't have financial incentives.

4.2.3 Warfarin dose

119 patients are recruited from the Hospital of the University of Pennsylvania and the Philadelphia Veterans Affairs Medical Center from Nov. 2009 to May 2012. They all have an expected duration of warfarin therapy of at least 6 months and they all have at least one international normalized ratio out of range within 90 days prior to enrollment. Patients are given an electronic medication monitoring system to measure their adherence to warfarin therapy everyday. Lottery system is implemented so that patient has 1% chance to win

$100 and 20% to win $10. Patient will be informed of the daily lottery result but they can take the money only if they adhere on that day. What different from previous two trials is that if the monitor gets disconnected, the lottery result cannot be delivered, which will be discussed in more details in the model section. More details can be found from Kimmel et al. (2016b), notice that we only use the two out of four groups which receive lottery in the original paper.

4.3 HMM for dynamic incentive

Current research (Patel et al., 2016; Lorincz et al., 2013; Kimmel et al., 2016b) studies the overall effect of financial incentive by comparing the treatment group and control group after the whole trial ends. In this paper, we will focus on the dynamic effect of the daily financial incentive, e.g., how does it affect participant's behavior along the timeline. In this section, the data from walking step trial will be used for illustration, the other two trials share the same model structure.

4.3.1 The hidden state and lottery effect

Suppose for patient i the trial lasts T_i days. The lottery result is observed as $\mathbf{L_i} = \{L_i^1, \ldots, L_i^{T_i}\}$, there are three possible outcomes of L_i^t, not win ($L_i^t = N$), win small \$5 ($L_i^t = S$) or win big \$50 ($L_i^t = B$). The patient's adherence is noted as $\mathbf{A_i} = \{A_i^1, \ldots, A_i^{T_i}\}, A_i^t \in \{A, N\}$. $A_i^t = A$ means patient i adhered on day t. On each day the patient has a certain probability to adhere, we will use $\mathbf{X_i} = \{X_i^1, \ldots, X_i^{T_i}\}$ to represent this. This X_i^t is the unobserved hidden state which describes patient's tendency to adhere on day t. The patient will first decide to adhere or not, or the value of A_i^t, according to X_i^t, then the lottery result L_i^t will be drawn and sent to the patient.

We assume that the combination of lottery result L_i^t and adherence A_i^t will affect the patient's tendency to adhere on the next day (X_i^{t+1}). More specifically, the patient will only receive money if he adheres and wins the lottery. This is a direct financial incentive. If he does not adhere but wins the lottery, he will be informed that he would have won XXX

if he adhered. In this scenario, there is still regret effect of sending out the lottery result (Chapman and Coups, 2006; Zeelenberg and Pieters, 2004). In total, there are six possible combinations of adherence and lottery result $\{NN, NS, NB, AN, AS, AB\}$ (e.g., NB means not adhere and win big), each combination has a different effect towards patient's tendency to adhere in the next day.

4.3.2 The models

Ideally, the tendency to adhere X_i^t would be best modelled as a probability in $[0, 1]$. Here for simplicity and practicality, we assume that X_i^t only has two states, $X_i^t = 1$ means patient i is in low adherence state and $X_i^t = 2$ represents the high adherence state. Parameters θ_1, θ_2 are used to model the probability of adherence via a logit function and δ_i is a heterogeneity parameter to modify the probability in the individual level. Therefore the probability function for patient's adherence can be written as:

$$P(A_i^t = A \mid X_i^t) = \frac{\exp(\theta_{X_i^t} + \delta_i)}{1 + \exp(\theta_{X_i^t} + \delta_i)}; \quad \theta_1 < \theta_2, \quad \sum_i \delta_i = 0. \tag{4.1}$$

To model the dynamic effects of lottery towards the hidden state, we assume that the hidden state X_i^t varies from day to day and has Markov property. Then the key is to model the transition probability from X_i^t to X_i^{t+1}. In the beginning of this section we discussed that each combination $A_i^t L_i^t$ has a different effect towards the transition of hidden states, since there are only two states for X_i^t, we will directly model the 2×2 probability transition matrix as

$$P(X_i^{t+1} \mid X_i^t, A_i^t, L_i^t) = Q_{A_i^t L_i^t}(X_i^t X_i^{t+1}) \tag{4.2}$$

where $(X_i^t X_i^{t+1})$ refer to the corresponding entry of matrix $Q_{A_i^t L_i^t}$. There are six probability transition matrices in total, $\{Q_{NN}, Q_{NS}, Q_{NB}, Q_{AN}, Q_{AS}, Q_{AB}\}$. The row sum of each matrix should be 1, so there are 2 free parameters in each matrix, and 12 degrees of freedom in total in this part of model. The difference between these probability transition matrices reflects the dynamic effect. Also, parameter p is used for the distribution of patient's hidden

state in the first day X_i^1 as:

$$P(X_i^1 = 1) = 1 - p; \quad P(X_i^1 = 2) = p. \tag{4.3}$$

Equation 4.1, 4.2 and 4.3 together are the complete HMM model we constructed, Figure 7 describes the probability relationship between variables and the corresponding parameters in the probability function are in parentheses.

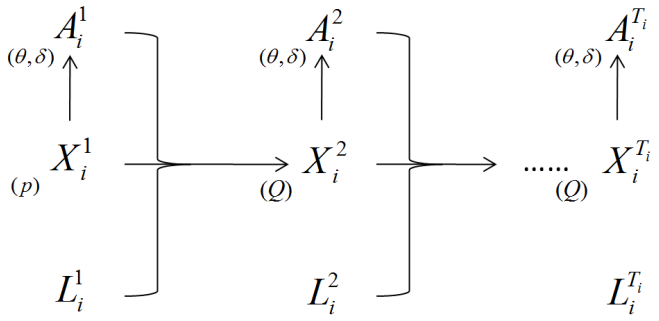

Figure 7: The HMM modeling the dynamic effect of lottery towards patient's tendency to adhere.

4.3.3 Model variation for the other trials

In glucose trial, the financial incentive and peer mentor arm only receive half lottery amount as the pure financial incentive arm. Here we will not differentiate these two arms, win \$100/\$50 are treated the same as win big , win \$10/\$5 are treated the same as win small. In doing so, the model for glucose trial is exactly the same as the walking step trial, so results can be compared across trials.

In warfarin trial, the electric monitor given to patient may get disconnected for a short while, patient can not receive any lottery result during those days. Therefore besides the six types of combination $A_i^t L_i^t$, we assume that the transition matrix is Q_{None} when patient does not hear from the lottery result. This can also be viewed as a different type of effect. Besides this, all other parts of the model are the same. There is also about 3% missing

data, we assume that they are missing at random.

4.4 Model fits and comparison

4.4.1 Fitting the model

We will use a standard approach to fit the HMM model derived in last section, which has the following steps:

(i) Initialize all the parameters.

(ii) Use Baum-Welch algorithm to calculate the conditional probability $P(X_i^t \mid \mathbf{L_i}, \mathbf{A_i})$ and $P(X_i^t, X_i^{t+1} \mid \mathbf{L_i}, \mathbf{A_i})$.

(iii) Use the above conditional probability and initialized parameter values to maximize the log likelihood function and update the parameter values.

(iv) Repeat step *ii* and *iii* until parameters converge.

The above steps give us a point estimation for the parameters and bootstrap will be used to construct the 95% confidence interval for each parameter. We sample the same number of patients with replacement to construct a bootstrap data and get a bootstrap estimation. This procedure is repeated for 200 times and the 95% confidence interval is constructed from the 200 bootstrap estimations.

4.4.2 Model Interpretation

The point estimation and 95% confidence interval for the parameters in equation 4.1, 4.2 and 4.3 are listed as below:

$$p = 0.23\ (0.04, 0.56); \quad \theta_1 = -1.88\ (-2.73, -1.81); \quad \theta_2 = -0.59\ (-1.12, 0.03); \qquad (4.4)$$

$$Q_{NN} = \begin{pmatrix} 1.00(0.48, 1.00) & 0.00(0.00, 0.52) \\ 0.40(0.17, 0.93) & 0.60(0.07, 0.83) \end{pmatrix}; \quad Q_{NS} = \begin{pmatrix} 1.00(0.61, 0.99) & 0.00(0.01, 0.39) \\ 0.40(0.20, 0.85) & 0.60(0.15, 0.80) \end{pmatrix} \tag{4.5}$$

$$Q_{NB} = \begin{pmatrix} 1.00(0.10, 1.00) & 0.00(0.00, 0.90) \\ 0.00(0.00, 0.99) & 1.00(0.01, 1.00) \end{pmatrix}; \quad Q_{AN} = \begin{pmatrix} 0.26(0.06, 0.62) & 0.74(0.38, 0.94) \\ 0.18(0.01, 0.41) & 0.82(0.59, 0.99) \end{pmatrix} \tag{4.6}$$

$$Q_{AS} = \begin{pmatrix} 0.40(0.00, 0.72) & 0.60(0.28, 1.00) \\ 0.08(0.00, 0.47) & 0.92(0.53, 1.00) \end{pmatrix}; \quad Q_{AB} = \begin{pmatrix} 0.00(0.00, 0.20) & 1.00(0.80, 1.00) \\ 0.00(0.00, 0.52) & 1.00(0.48, 1.00) \end{pmatrix} \tag{4.7}$$

We can see that when patient does not adhere, no matter what the lottery result is, the probability that he will transit from low to high hidden state is 0 (the upper left entries for transition matrix Q_{NN}, Q_{NS}, Q_{NB} are 1), therefore it's hardly to believe that regret has a significant effect towards patient's adherence in walking step trial. On the other hand, when patient does adhere, the probability of transiting from low/high to high state increases as the lottery outcome rises. This would suggest that the real financial incentive may have some effect towards patient's adherence.

However, it's hard to give a straightforward interpretation if we only look at the transition matrix. We need to interpret the result in an easier way. Notice that the hidden state X_i^t and adherence A_i^t will be determined first, then an independent lottery will be drawn. We can devide patients into 4 groups: the patient is in low state and adhered; in low state and did not adhere; in high state and adhered; in high state and did not adhere. Then for each group, given a lottery result on that day, we can calculate the patient's probability to adhere in the next day. Table 4 shows the probability and also the confidence interval. In doing so, we directly associate the effect of lottery result towards the probability to adhere next day. With this result, people can design an efficient lottery system that increases user's adherence within limited budget for lottery.

In table 4, for patients who did not adhere, no matter he was in low or high hidden state, his probability to adhere will not change much as lottery result varies. This also suggests that

Table 4: Given patient's hidden state and adherence, the probability to adhere in the next day under different lottery result. Parenthesis is the 95% confidence interval.

	not win	win small	win big
low & adhere	0.38(0.31, 0.42)	0.36(0.29, 0.44)	0.43(0.37, 0.52)
low & not adhere	0.25(0.21, 0.32)	0.25(0.21, 0.30)	0.25(0.20, 0.41)
high & adhere	0.40(0.36, 0.44)	0.42(0.35, 0.46)	0.43(0.36, 0.51)
high & not adhere	0.36(0.25, 0.43)	0.36(0.25, 0.41)	0.43(0.23, 0.52)

there is no significant regret effect in the trial. We further compare the p-value of testing whether different lottery result has significant effect towards the patient's future adherency. Table 5 shows the result.

Table 5: P-value of comparing different lottery result's effects toward patient's probability to adhere given patient's hidden state and adherence.

	win big - win small	win small - not win	win big - not win
low & adhere	0.015	0.515	0.020
low & not adhere	0.430	0.545	0.450
high & adhere	0.065	0.305	0.050
high & not adhere	0.255	0.430	0.215

There are 3 entries with p-values smaller than the threshold 0.05. For patient in low state and does adhere, win big is significantly better than win small or not win in terms of probability to adhere in next day. For patient in high state and does adhere, the only significance found is win big v.s. not win. In this walking step trial, we think the financial incentive will be useful if the patient actually adheres and receives the money. No regret effect is found.

4.4.3 Model comparison

We also perform the same model fitting procedure to the other two trials. Here we will just report the table of p-values as in Table 5.

Table 6 shows the p-value for glucose trial, which has two significant entries. The result here is very different from the result in walking trial. We find a significance in regret effect, but do not find significance in the effect of real financial incentive. Also the more frequent,

win small result is better than the less frequent, win big result.

Table 6: P-value for comparing different lottery results using the data from glucose reading trial.

	win big - win small	win small - not win	win big - not win
low & adhere	0.555	0.160	0.495
low & not adhere	0.865	0.020	0.780
high & adhere	0.290	0.395	0.245
high & not adhere	0.350	0.030	0.195

Table 7 shows the p-value for warfarin trial. In this trial we fail to find any significant effect. This may be that there are part of data missing in this trial and also patients in this trial do not have the same length of observation period, some patients have very few observations. These can lead to a wider confidence interval and less power to find the significance.

Table 7: P-value for comparing different lottery results using the data from warfarin trial.

	win big - win small	win small - not win	win big - not win
low & adhere	0.095	0.695	0.115
low & not adhere	0.650	0.130	0.575
high & adhere	0.805	0.220	0.750
high & not adhere	0.880	0.210	0.790

The different results from the 3 trials suggest us that such behaviors can be domain specific and we might not be powered to detect statistically significant patterns across studies yet, therefore future studies are needed to test the application of this model, or improved version of such model on a larger group of clinical trials in order to derive definitive evidences.

4.5 Discussion

In this paper, we constructed HMM to model the dynamic effect of financial incentives towards patient's tendency to adhere. Previous literature studies the overall financial incentive effect after the trial finishes while we studies the dynamic effect at different times, e.g., given patient's information on day t, how does the lottery result affect patient's tendency to adhere on day $t+1$.

We assume that everyday the patient is in a hidden state which determines his tendency(probability)

to adhere. If the patient does adhere and wins the lottery, he receives a real money incentive. If the patient does not adhere, he will still be informed of the lottery result, if he hears that he could have won the lottery if he adhered, there is also a regret effect. We assume these effects directly affect the transition probability of patient's hidden state.

We fit our model to 3 different trials, which all have a very similar procedure as well as the lottery system. We interpret the result by dividing patients into four groups, and in each group compare the probability to adhere in the next day if given different lottery results. Notice that in real life, the hidden state can not be observed so for one patient at one day, we can not exactly figure out his group. However, Baum-Welch algorithm provides the posterior distribution of the hidden state, which could be used as an inference of patient's group.

The results we get from the three trials are very different. In warfarin trial there is no significance at all. In walking trial, we find significance difference for real money incentive if the patient win big v.s. not win. However, in glucose trial, there is significance for regret effect if the patient win small v.s. not win. The different results from the 3 trials suggest us that such behaviors can be domain specific and we might not be powered to detect statistically significant patterns across studies yet, therefore future studies are needed to test the application of this model, or improved version of such model on a larger group of clinical trials in order to derive definitive evidences.

APPENDIX : Proof for Main Results in Chapter 2

A.1 The distribution of AR statistic and power calculation

Recall the model we constructed is

$$\begin{aligned} Y^* &= \beta D^* + \delta\sigma_1 Z^* + u^* \\ D^* &= \gamma Z^* + v^* \\ Y^* = M_X Y;\quad D^* = M_X D;\quad Z^* = M_X Z;\quad u^* = M_X u;\quad v^* &= M_X v; \\ (u,v) \perp Z;\quad (u_i, v_i)^T \sim N(\mathbf{0}, \Sigma);\quad \Sigma = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix} &\quad \text{rank}(X) = k \end{aligned} \tag{A.1}$$

and the AR test statistic is

$$AR(\beta_0) = \frac{(Y^* - \beta_0 D^*)^T P_{Z^*}(Y^* - \beta_0 D^*)}{(Y^* - \beta_0 D^*)^T M_{Z^*}(Y^* - \beta_0 D^*)/(n-k-1)} \tag{A.2}$$

In this section, we will prove the distribution of AR statistic under the alternative hypothesis $H_1 : \beta - \beta_0 = \lambda$ in sensitivity analysis:

$$AR(\beta_0) \sim F_{1,n-k-1,\frac{(\gamma+\delta^*\sigma_1/\lambda)^2 Z^{*T} Z^*}{\sigma_2^2}\Lambda} \tag{A.3}$$

If the IV is valid, then $\delta^* = 0$ becomes a specific case for no sensitivity. After (A.3) is proved, the power formula in original paper can be directly calculated.

Lemma 2. *In model* (A.1)*, we have*

$$r(M_X P_{Z^*} M_X) = 1; \qquad r(M_X M_{Z^*} M_X) = n - k - 1$$

where $r(A)$ is the rank of matrix A.

Proof of Lemma 2: Since $M_X = I_n - P_X$ is a projection matrix, it is also idempotent, so

$$Z^* = M_X Z = M_X M_X Z = M_X Z^*$$

therefore,

$$M_X P_{Z^*} = M_X Z^* (Z^{*T} Z^*)^{-1} Z^{*T} = Z^* (Z^{*T} Z^*)^{-1} Z^{*T} = P_{Z^*} \tag{A.4}$$

Similarly,

$$P_{Z^*} M_X = Z^* (Z^{*T} Z^*)^{-1} Z^{*T} M_X = Z^* (Z^{*T} Z^*)^{-1} Z^{*T} = P_{Z^*} \tag{A.5}$$

Since Z^* is a $n \times 1$ vector, by (A.4) and (A.5) we have

$$r(M_X P_{Z^*} M_X) = r(P_{Z^*}) = 1 \tag{A.6}$$

For $M_X M_{Z^*} M_X$, it's equivalent to

$$M_X M_{Z^*} M_X = M_X (I_n - P_{Z^*}) M_X = M_X - P_{Z^*}$$

From (A.4) and (A.5) we can also get $(M_X - P_{Z^*})(M_X - P_{Z^*}) = M_X - P_{Z^*}$, so $M_X - P_{Z^*}$ is also an idempotent matrix. The result 2.3.9 from Ravishanker and Dey (2001) shows

$$r(M_X M_{Z^*} M_X) = tr(M_X M_{Z^*} M_X) = tr(M_X - P_{Z^*}) = tr(M_X) - tr(P_{Z^*}) = n - k - 1 \tag{A.7}$$

(A.6) and (A.7) together proved this lemma.

Now we start proving (A.3).

Proof of (A.3)**:** First, the term $Y^* - \beta_0 D^*$ can be transformed as

$$\begin{aligned} Y^* - \beta_0 D^* &= \beta D^* + \delta^* \sigma_1 Z^* + u^* - \beta_0 D^* \\ &= (\lambda\gamma + \delta^*\sigma_1) Z^* + \lambda v^* + u^* \\ &= M_X w \end{aligned}$$

where

$$w = (\lambda\gamma + \delta^*\sigma_1) Z + \lambda v + u \sim N_n((\lambda\gamma + \delta^*\sigma_1) Z, \sigma^2 I_n)$$

$$\sigma^2 = \sigma_1^2 + 2\rho\sigma_1\sigma_2\lambda + \sigma_2^2\lambda^2$$

Consider $(w/\sigma)^T (M_X P_{Z^*} M_X)(w/\sigma)$ and $(w/\sigma)^T (M_X M_{Z^*} M_X)(w/\sigma)$, by Lemma 2 we know

$$r(M_X P_{Z^*} M_X) = 1; \qquad r(M_X M_{Z^*} M_X) = n - k - 1 \tag{A.8}$$

We also showed

$$M_X P_{Z^*} M_X = P_{Z^*} \text{ and } M_X M_{Z^*} M_X = M_X - P_{Z^*} \text{ are idempotent matrices} \tag{A.9}$$

We also have

$$M_X P_{Z^*} M_X \cdot M_X M_{Z^*} M_X = P_{Z^*}(M_X - P_{Z^*}) = 0 \tag{A.10}$$

(A.8)-(A.10) satisfy the conditions stated in result 5.4.8, Ravishanker and Dey (2001). This result directly leads to the following facts (Notice the different usage of non-central parameter between Ravishanker and Dey (2001) and this paper. Ravishanker and Dey (2001) has an extra constant 1/2.)

$$(w/\sigma)^T (M_X P_{Z^*} M_X)(w/\sigma) \sim \chi_1^2 \left(\frac{(\lambda\gamma + \delta^*\sigma_1)^2 Z^{*T} Z^*}{\sigma^2} \right) \tag{A.11}$$

$$(w/\sigma)^T (M_X M_{Z^*} M_X)(w/\sigma) \sim \chi_{n-k-1}^2 \tag{A.12}$$

$$(w/\sigma)^T(M_X P_{Z^*} M_X)(w/\sigma) \text{ and } (w/\sigma)^T(M_X M_{Z^*} M_X)(w/\sigma) \text{ are independent} \quad \text{(A.13)}$$

Finally, (A.11)-(A.13) lead to the proof of (A.3):

$$\begin{aligned} AR(\beta_0) &= \frac{(w/\sigma)^T(M_X P_{Z^*} M_X)(w/\sigma)}{(w/\sigma)^T(M_X M_{Z^*} M_X)(w/\sigma)/(n-k-1)} \\ &\sim F_{1,n-k-1,\frac{(\lambda\gamma+\delta^*\sigma_1)^2 Z^{*T} Z^*}{\sigma_1^2+2\rho\sigma_1\sigma_2\lambda+\lambda^2\sigma_2^2}} \end{aligned}$$

A.2 Power formula in relationship with the non-centrality parameter

For the power formula

$$Power_0 = 1 - \Psi_{1,n-k-1,\frac{\gamma^2 Z^{*T} Z^*}{\sigma_2^2}\Lambda}\left(F_{1,n-k-1,\Delta^2 Z^{*T} Z^*;1-\alpha}\right) \quad \text{(A.14)}$$

We will prove the following five propositions:

Proposition 3. *In the power formula* (A.14)*, we have*

(a) If $ncp_1 = ncp_2$, the power is always α.

(b) For fixed ncp_2, power increases as ncp_1 increases.

(c) For fixed ncp_1, power decreases as ncp_2 increases.

(d) If $ncp_1 > ncp_2$, the power is larger than α and will increase to 1 as the sample size increases.

(e) If $ncp_1 < ncp_2$, the power is smaller than α and will decrease to 0 as the sample size increases.

We start by proving the following Lemmas:

Lemma 4. *Suppose* $X \sim N(0,1), s > 0, t > 0$*, then*

$$P(X^2 < s) > P((X+t)^2 < s)$$

Proof of Lemma 4: Let $f_Z(\cdot)$ be the PDF of standard normal distribution, then

$$\begin{aligned} P(X^2 < s) - P((X+t)^2 < s) &= P(-\sqrt{s} < X < \sqrt{s}) - P(-\sqrt{s} - t < X < \sqrt{s} - t) \\ &= P(\sqrt{s} - t < X < \sqrt{s}) - P(-\sqrt{s} - t < X < -\sqrt{s}) \\ &= \int_0^t f_Z(\sqrt{s} - w)dw - \int_0^t f_Z(-\sqrt{s} - w)dw \end{aligned} \tag{A.15}$$

Since $\forall s > 0, w > 0$, we have $(\sqrt{s} - w)^2 < (-\sqrt{s} - w)^2$, therefore

$$f_Z(\sqrt{s} - w) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(\sqrt{s} - w)^2}{2}\right) > \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(-\sqrt{s} - w)^2}{2}\right) = f_Z(-\sqrt{s} - w)$$

Combining with (A.15), we proved Lemma 4.

$$P(X^2 < s) > P((X+t)^2 < s)$$

Lemma 5. *Suppose* $\Psi_{a,b,n}(x)$ *is the CDF of the non-central F distribution with degree of freedom a, b and non-centrality parameter n. Then for fixed* x_0*,* $\Psi_{a,b,n}(x_0)$ *increases as n increases.*

Proof of Lemma 5: We want to show that $\forall$ positive $n_1 < n_2$, $\Psi_{a,b,n_1}(x_0) < \Psi_{a,b,n_2}(x_0)$. Let $X \sim N(0,1)$, $Y \sim \chi^2_{a-1}(n_1)$, $Z \sim \chi^2_b$. X, Y, Z are independent.

$$S_1 = \frac{X^2 + Y}{Z} \sim F_{a,b,n_1}, \qquad S_2 = \frac{(X + \sqrt{n_2 - n_1})^2 + Y}{Z} \sim F_{a,b,n_2},$$

For fixed x_0, define random variable $W = Zx_0 - Y$, suppose $f_W(w)$ is the pdf of W.

Combined with the result from Lemma 4, we have:

$$\begin{aligned}
\Psi_{a,b,n_1}(x_0) &= P(X^2 < W) \\
&= \int_{t=0}^{\infty} P(X^2 < \sqrt{w}) f_W(w) dw \\
&> \int_{t=0}^{\infty} P((X + \sqrt{n_2 - n_1})^2 < \sqrt{w}) f_W(w) dw \\
&= P(X + \sqrt{n_2 - n_1})^2 < W) \\
&= \Psi_{a,b,n_2}(x_0)
\end{aligned}$$

Thus Lemma 5 is proved.

Now we come back to prove Proposition 3. (a) is very straightforward. (b) is equivalent to Lemma 5, which is proved. For (c), it's equivalent to prove

$$F_{a,b,n_1;1-\alpha} < F_{a,b,n_2;1-\alpha}, \quad \forall n_1 < n_2 \tag{A.16}$$

We have,

$$\Psi_{a,b,n_1}(F_{a,b,n_1;1-\alpha}) = 1 - \alpha = \Psi_{a,b,n_2}(F_{a,b,n_2;1-\alpha})$$

By Lemma 5 we also have

$$\Psi_{a,b,n_1}(F_{a,b,n_1;1-\alpha}) > \Psi_{a,b,n_2}(F_{a,b,n_1;1-\alpha})$$

Therefore,

$$\Psi_{a,b,n_2}(F_{a,b,n_1;1-\alpha}) > \Psi_{a,b,n_2}(F_{a,b,n_2;1-\alpha}) \quad \Rightarrow \quad F_{a,b,n_1;1-\alpha} < F_{a,b,n_2;1-\alpha}$$

Hence A.16 is proved and (c) holds.

Now we prove (d), as sample size n increases, ${Z^*}^T Z \sim n\text{Var}(Z^*)$, so let $\text{ncp}_1 = an$, $\text{ncp}_2 = bn$, $a > b > 0$ in power formula (A.14). Suppose $X_1 \sim N(\sqrt{an}, 1)$, $X_2 \sim N(\sqrt{bn}, 1)$, $Y \sim \chi^2_{n-k-1}/(n-k-1)$, X_1, X_2, Y are independent, then $X_1^2/Y \sim F_{1,n-k-1,\text{ncp}_1}$ and $X_2^2/Y \sim F_{1,n-k-1,\text{ncp}_2}$. Also let $q = F_{1,n-k-1,\text{ncp}_2;1-\alpha}$, then in order to prove (d) we just need to show

$$\lim_{n\to\infty} P(X_1^2/Y > q) = 1 \tag{A.17}$$

First for any $\epsilon > 0$, we have

$$\begin{aligned}
1-\alpha &= P(X_2^2/Y < q) \\
&= P(X_2^2 < qY, Y < 1-\epsilon) + P(X_2^2 < qY, Y > 1-\epsilon) \\
&> 0 + P(X_2^2 < q(1-\epsilon), Y > 1-\epsilon) \\
&= P(X_2^2 < q(1-\epsilon))P(Y > 1-\epsilon)
\end{aligned}$$

Since $Y \overset{a.s.}{\to} 1$, so for fixed $\epsilon = \frac{a-b}{a+b} > 0$, $\exists N$, such that $\forall n > N$, $P(Y > 1-\epsilon) > \frac{1-\alpha}{1-\alpha/2}$. Also notice $X_2 \sim N(\sqrt{bn}, 1)$. Let ψ_α represents the α-th quantile of standard normal distribution, then $\forall n > N$, we have

$$\begin{aligned}
& P(|X_2| < \sqrt{q(1-\epsilon)}) < 1-\alpha/2 \\
\Rightarrow \quad & \sqrt{q(1-\epsilon)} - \sqrt{bn} < \psi_{1-\alpha/4} \\
\Rightarrow \quad & \sqrt{q} - \sqrt{n(a+b)/2} < \psi_{1-\alpha/4}\sqrt{\frac{a+b}{2b}}
\end{aligned} \tag{A.18}$$

Now back to (A.17), for any $\delta > 0$, we have

$$\begin{aligned}
P(X_1^2/Y > q) &> P(X_1 > \sqrt{qY}) \\
&= P(X_1 > \sqrt{qY}, Y < 1+\delta) + P(X_1 < \sqrt{qY}, Y > 1+\delta) \\
&> P(X_1 > \sqrt{q(1+\delta)}) + 0 \\
&= 1 - \Psi_Z(\sqrt{q(1+\delta)} - \sqrt{an})
\end{aligned}$$

where $\Psi_Z(\cdot)$ is the CDF of standard normal distribution. Choose any $\delta < \frac{a-b}{a+b}$, such that

$$\sqrt{(1+\delta)(a+b)/2} - \sqrt{a} < 0 \tag{A.19}$$

then (A.18) and (A.19) tell us that

$$\begin{aligned}
\lim_{n\to\infty}(\sqrt{q(1+\delta)} - \sqrt{an}) &= \lim_{n\to\infty}\left(\sqrt{1+\delta}\left(\sqrt{q} - \sqrt{n\frac{a+b}{2}}\right) + \left(\sqrt{(1+\delta)\frac{a+b}{2}} - \sqrt{a}\right)\sqrt{n}\right) \\
&< \lim_{n\to\infty}\left(\sqrt{1+\delta}\psi_{1-\alpha/4}\sqrt{\frac{a+b}{2b}} + \left(\sqrt{(1+\delta)\frac{a+b}{2}} - \sqrt{a}\right)\sqrt{n}\right) \\
&= -\infty
\end{aligned}$$

which directly leads to

$$\lim_{n\to\infty} P(X_1^2/Y > q) = 1 - \lim_{n\to\infty}\Psi_Z(\sqrt{q(1+\delta)} - \sqrt{an}) = 1$$

Hence property (d) is proved. Similarly we can prove (e) by following a similar procedure.

BIBLIOGRAPHY

R. Altman. Mixed hidden markov models: An extension of the hidden markov model to the longitudinal data setting. *J Amer Statist Assoc*, 102:201–210, 2007.

T. Amemiya. *Advanced Econometrics*. Harvard University Press, 1985.

T. W. Anderson and H. Rubin. Estimation of the parameters of a single equation in a complete system of stochastic equations. *The Annals of Mathematical Statistics*, 20(1): 46–63, 1949a.

T. W. Anderson and H. Rubin. Estimation of the parameters of a single equation in a complete system of stochastic equations. *Annals of Mathematical Statistics*, 20:46–63, 1949b.

D. W. Andrews, M. J. Moreira, and J. H. Stock. Optimal two-sided invariant similar tests for instrumental variables regression. *Econometrica*, 74(3):715–752, 2006a.

D. W. K. Andrews, M. J. Moreira, and J. H. Stock. Optimal two-sided invariant similar tests for instrumental variables regression. *Econometrica*, 74(3):715–752, 2006b.

D. W. K. Andrews, M. J. Moreira, and J. H. Stock. Performance of conditional wald tests in iv regression with weak instruments. *Journal of Econometrics*, 139(1):116–132, 2007.

J. D. Angrist and A. B. Krueger. Does compulsory school attendance affect schooling and earnings? *The Quarterly Journal of Economics*, 106(4):979–1014, 1991.

J. D. Angrist and A. B. Krueger. Instrumental variables and the search for identification: From supply and demand to natural experiments. *The Journal of Economic Perspectives*, 15(4):69–85, 2001.

J. D. Angrist, G. W. Imbens, and D. B. Rubin. Identification of causal effects using instrumental variables. *Journal of the American statistical Association*, 91(434):444–455, 1996a.

J. D. Angrist, G. W. Imbens, and D. B. Rubin. Identification of causal effects using instrumental variables. *Journal of the American Statistical Association*, 91(434):444–455, 1996b.

M. Baiocchi, J. Cheng, and D. Small. Tutorial in biostatistics: Instrumental variable methods for causal inference. *Statistics in Medicine*, 33(in press):2297–2340, 2014a.

M. Baiocchi, J. Cheng, and D. S. Small. Instrumental variable methods for causal inference. *Statistics in Medicine*, 33(13):2297–2340, 2014b.

C. F. Baum, M. E. Schaffer, and S. Stillman. Instrumental variables and gmm: Estimation and testing. *Stata Journal*, 3(1):1–31, 2003.

C. F. Baum, M. E. Schaffer, and S. Stillman. Enhanced routines for instrumental variables/generalized method of moments estimation and testing. *Stata Journal*, 7(4):465–506, 2007.

K. A. Bollen. Instrumental variables in sociology and the social sciences. *Annual Review of Sociology*, 38:37–72, 2012.

J. Bound, D. A. Jaeger, and R. M. Baker. Problems with instrumental variables estimation when the correlation between instruments and the endogenous variable is weak. *Journal of the American Statistical Association*, 90:443–450, 1995.

D. Bravata, C. Smith-Spangler, and V. Sundaram. Using pedometers to increase physical activity and improve health: a systematic review. *JAMA*, 298:2296–2304, 2007.

M. A. Brookhart and S. Schneeweiss. Preference-based instrumental variable methods for the estimation of treatment effects: assessing validity and interpreting results. *The international journal of biostatistics*, 3:14, 2007.

S. Burgess and S. G. Thompson. Bayesian methods for meta-analysis of causal relationships estimated using genetic instrumental variables. *Statistics in medicine*, 29(12):1298–1311, 2010.

S. Burgess, S. G. Thompson, and CRP-CHD-Genetics-Collaboration(CCGC). Methods for meta-analysis of individual participant data from mendelian randomisation studies with binary outcomes. *Statistical methods in medical research*, 2012. (doi:10.1177/0962280212451882).

D. Card. Using geographic variations in college proximity to estimate the return to schooling. In L. N. Christofides, E. K. Grant, and R. Swidinsky, editors, *Aspects of Labor Market Behaviour: Essays in Honour of John Vanderkamp*. University of Toronto Press, 1995.

J. C. Chao and N. R. Swanson. Consistent estimation with a large number of weak instruments. *Econometrica*, 73(5):1673–1692, 2005.

G. Chapman and E. Coups. Emotions and preventive health behavior: worry, regret, and influenza vaccination. *Health psychology*, 25(1):82, 2006.

L. Chen, G. Davey Smith, R. M. Harbord, and S. J. Lewis. Alcohol intake and blood pressure: A systematic review implementing a mendelian randomization approach. *PLoS Medicine*, 5(3):e52, 2008.

T. G. Conley, C. B. Hansen, and P. E. Rossi. Plausibly exogenous. *Review of Economics and Statistics*, 94(1):260–272, 2012a.

T. G. Conley, C. B. Hansen, and P. E. Rossi. Plausibly exogenous. *Review of Economics and Statistics*, 94(1):260–272, 2012b.

CRP-CHD-Genetics-Collaboration(CCGC). Association between c reactive protein and coronary heart disease: mendelian randomisation analysis based on individual participant data. *BMJ*, 342:d548, 2011.

G. Davey Smith and S. Ebrahim. 'mendelian randomization': can genetic epidemiology contribute to understanding environmental determinants of disease? *International Journal of Epidemiology*, 32(1):1–22, 2003.

G. Davey Smith and S. Ebrahim. Mendelian randomization: prospects, potentials, and limitations. *International Journal of Epidemiology*, 33(1):30–42, 2004.

R. Davidson and J. G. MacKinnon. Estimation and inference in econometrics. *OUP Catalogue*, 1993a.

R. Davidson and J. G. MacKinnon. *Estimation and Inference in Econometrics*. Oxford University Press, New York, 1993b.

V. Didelez and N. Sheehan. Mendelian randomization as an instrumental variable approach to causal inference. *Statistical Methods in Medical Research*, 16(4):309–330, 2007.

T. A. DiPrete and M. Gangl. Assessing bias in the estimation of causal effects: Rosenbaum bounds on matching estimators and instrumental variables estimation with imperfect instruments. *Sociological Methodology*, 34(1):271–310, 2004a.

T. A. DiPrete and M. Gangl. Assessing bias in the estimation of causal effects: Rosenbaum bounds on matching estimators and instrumental variables estimation with imperfect instruments. *Sociological Methodology*, 34(1):271–310, 2004b.

J.-M. Dufour. Identification, weak instruments, and statistical inference in econometrics. *The Canadian Journal of Economics / Revue canadienne d'Economique*, 36(4):767–808, 2003.

S. Ebrahim and G. D. Smith. Mendelian randomization: can genetic epidemiology help redress the failures of observational epidemiology? *Human genetics*, 123(1):15–33, Feb. 2008. ISSN 1432-1203. doi: 10.1007/s00439-007-0448-6.

J. Fox, Z. Nie, and J. Byrnes. **sem***: Structural Equation Models*, 2014. URL `http://CRAN.R-project.org/package=sem`. R package version 3.1-5.

G. Freeman, B. J. Cowling, and C. M. Schooling. Power and sample size calculations for mendelian randomization studies using one genetic instrument. *International journal of epidemiology*, 42(4):1157–1163, 2013a.

G. Freeman, B. J. Cowling, and C. M. Schooling. Power and sample size calculations for mendelian randomization studies using one genetic instrument. *International journal of epidemiology*, 42(4):1157–1163, 2013b.

W. A. Fuller. *Measurement Error Models*. Wiley, 2006.

S. Gaure. **lfe**: Linear group fixed effects. *The R Journal*, 5(2):104–117, Dec 2013. URL `http://journal.r-project.org/archive/2013-2/gaure.pdf`.

L. A. Gennetian, K. Magnuson, and P. A. Morris. From statistical associations to causation: What developmentalists can learn from instrumental variables techniques coupled with experimental data. *Developmental Psychology*, 44(2):381, 2008.

G. Gibson. Rare and common variants: twenty arguments. *Nature Reviews Genetics*, 13 (2):135–145, 2012.

R. Glasgow and D. Toobert. Social environment and regimen adherence among type ii diabetic patients. *Diabetes Care*, 11:377–386, 1988.

M. M. Glymour, E. J. Tchetgen Tchetgen, and J. M. Robins. Credible mendelian randomization studies: Approaches for evaluating the instrumental variable assumptions. *American Journal of Epidemiology*, 175(4):332–339, Jan. 2012. ISSN 0002-9262, 1476-6256. doi: 10.1093/aje/kwr323.

A. Go, D. Mozaffarian, and V. Roger. Heart disease and stroke statistics-2013 update: a report from the american heart association. *Circulation*, 127(1):143–146, 2013.

J. Hausman. Specification and estimation of simultaneous equation models. *Handbook of econometrics*, 1:391–448, 1983.

M. A. Hernán and J. M. Robins. Instruments for causal inference: an epidemiologist's dream? *Epidemiology*, 17(4):360–372, 2006.

P. W. Holland. Causal inference, path analysis, and recursive structural equations models. *Sociological methodology*, 18(1):449–484, 1988a.

P. W. Holland. Causal inference, path analysis, and recursive structural equations models. *Sociological Methodology*, 18(1):449–484, 1988b.

G. W. Imbens. Instrumental variables: An econometrician's perspective. *Statistical Science*, 2014. in press.

G. W. Imbens and P. R. Rosenbaum. Robust, accurate confidence intervals with a weak instrument: quarter of birth and education. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 168(1):109–126, 2005.

C. Jackevicius, M. Mamdani, and J. Tu. Adherence with statin therapy in elderly patients with and without acute coronary syndromes. *JAMA*, 288(4):462–467, 2002.

Y. Jiang, N. Zhang, and D. S. Small. Sensitivity analysis and power for instrumental variable studies. *Biometrics, under review*, 2015.

H. Kang, A. Zhang, T. T. Cai, and D. S. Small. Instrumental variables estimation with

some invalid instruments and its application to mendelian randomization. *Journal of the American Statistical Association*, 2015.

S. Kimmel, B. Andrea, and B. French. A randomized trial of lotterybased incentives and reminders to improve warfarin adherence: the warfarin incentives (win2) trial. *Pharmacoepidemiol Drug Saf*, 25(11):1219–1227, 2016a.

S. Kimmel, A. Troxel, B. French, G. Loewenstein, J. Doshi, T. Hecht, M. Laskin, C. Brensinger, C. Meussner, and K. Volpp. A randomized trial of lotterybased incentives and reminders to improve warfarin adherence: the warfarin incentives (win2) trial. *Pharmacoepidemiology and Drug Safety*, 25(11):1219–1227, 2016b.

C. Kleiber and A. Zeileis. *Applied Econometrics with R*. Springer-Verlag, New York, 2008. URL `http://CRAN.R-project.org/package=AER`. ISBN 978-0-387-77316-2.

M. Kolesár, R. Chetty, J. N. Friedman, E. L. Glaeser, and G. W. Imbens. Identification and inference with many invalid instruments. Technical report, National Bureau of Economic Research, 2011a.

M. Kolesár, R. Chetty, J. N. Friedman, E. L. Glaeser, and G. W. Imbens. Identification and inference with many invalid instruments. Technical report, National Bureau of Economic Research, 2011b.

M. Kolesár, R. Chetty, J. N. Friedman, E. L. Glaeser, and G. W. Imbens. Identification and inference with many invalid instruments. *National Bureau of Economic Research*, page No. w17519, 2013.

A. B. Krueger. Experiemtnal estimates of education production functions. *the Quarterly Journal of Economics*, 114(2):497–532, 1999.

D. A. Lawlor, R. M. Harbord, J. A. Sterne, N. Timpson, and G. Davey Smith. Mendelian randomization: using genes as instruments for making causal inferences in epidemiology. *Statistics in medicine*, 27(8):1133–1163, 2008a.

D. A. Lawlor, R. M. Harbord, J. A. C. Sterne, N. Timpson, and G. Davey Smith. Mendelian randomization: Using genes as instruments for making causal inferences in epidemiology. *Statistics in Medicine*, 27(8):1133–1163, 2008b.

G. Loewenstein, K. Volpp, and D. Asch. Incentives in health: different prescriptions for physicians and patients. *JAMA*, 307(13):1375–1376, 2012.

I. Lorincz, B. Lawson, and J. Long. Provider and patient directed financial incentives to improve care and outcomes for patients with diabetes. *Current diabetes reports*, 13(2): 188–195, 2013.

R. S. Mariano. *Simultaneous Equation Model Estimators: Statistical Properties and Practical Implications*, pages 122–141. Blackwell Publishing Ltd, 2003.

G. McLachlan and T. Krishnan. *The EM algorithm and extensions*, volume 382. John Wiley and Sons, 2007.

A. Mikusheva. Robust confidence sets in the presence of weak instruments. *Journal of Econometrics*, 157(2):236–247, 2010.

M. J. Moreira. *Tests with correct size when instruments can be arbitrarily weak*. Center for Labor Economics, University of California, Berkeley, 2001a.

M. J. Moreira. *Tests with Correct Size When Instruments Can Be Arbitrarily Weak*. Center for Labor Economics, University of California, Berkeley, 2001b.

M. J. Moreira. A conditional likelihood ratio test for structural models. *Econometrica*, 71 (4):1027–1048, 2003.

S. L. Morgan and C. Winship. *Counterfactuals and causal inference: Methods and principles for social research*. Cambridge University Press, 2007a.

S. L. Morgan and C. Winship. *Counterfactuals and causal inference: Methods and principles for social research*. Cambridge University Press, 2007b.

M. P. Murray. Avoiding invalid instruments and coping with weak instruments. *The Journal of Economic Perspectives*, 20(4):111–132, 2006.

C. R. Nelson and R. Startz. The distribution of the instrumental variables estimator and its t-ratio when the instrument is a poor one. *The Journal of Business*, 63(1):S125–40, 1990a.

C. R. Nelson and R. Startz. The distribution of the instrumental variables estimator and its t-ratio when the instrument is a poor one. *The Journal of Business*, 63(1):S125–40, 1990b.

D. Nitsch, M. Molokhia, L. Smeeth, B. L. DeStavola, J. C. Whittaker, and D. A. Leon. Limits to causal inference based on mendelian randomization: a comparison with randomized controlled trials. *American journal of epidemiology*, 163(5):397–403, 2006.

M. Patel, D. Asch, R. Rosin, D. Small, S. Bellamy, J. Heuer, S. Sproat, C. Hyson, N. Haff, S. Lee, and L. Wesby. Framing financial incentives to increase physical activity among overweight and obese adults: A randomized, controlled trial. *Annals of internal medicine*, 164(6):385–394, 2016.

T. Permutt and J. R. Hebel. Simultaneous-equation estimation in a clinical trial of the effect of smoking on birth weight. *Biometrics*, 45(2):619–622, 1989.

A. Peterson, L. Takiya, and R. Finley. Meta-analysis of trials of interventions to improve medication adherence. *Am J Health Syst Pharm*, 60:657–665, 2003.

N. Ravishanker and D. K. Dey. *A first course in linear model theory*. CRC Press, 2001.

N. Ries. Financial incentives for weight loss and healthy behaviours. *Healthcare Policy*, 7 (3):23, 2012.

P. Rosenbaum. Design of observational studies. *Springer series in statistics*, 2010a.

P. R. Rosenbaum. Design sensitivity in observational studies. *Biometrika*, 91(1):153–164, 2004.

P. R. Rosenbaum. *Design of Observational Studies*. Springer Series in Statistics. Springer, New York, 2010b.

T. J. Rothenberg. Approximating the distributions of econometric estimators and test statistics. *Handbook of econometrics*, 2:881–935, 1984.

D. B. Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology*, 66(5):688, 1974a.

D. B. Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66(5):688–701, 1974b.

J. D. Sargan. The estimation of economic relationships using instrumental variables. *Econometrica*, pages 393–415, 1958.

M. Sexton and R. J. Hebel. A clinical trial of change in maternal smoking and its effect on birth weight. *Journal of the American Medical Association*, 251(7):911–915, 1984.

K. Shirley, D. Small, K. Lynch, S. Maisto, and D. Oslin. Hidden markov models for alcoholism treatment trial data. *The Annals of Applied Statistics*, pages 366–395, 2010.

D. S. Small. Sensitivity analysis for instrumental variables regression with overidentifying restrictions. *Journal of the American Statistical Association*, 102(479):1049–1058, 2007a.

D. S. Small. Sensitivity analysis for instrumental variables regression with overidentifying restrictions. *Journal of the American Statistical Association*, 102(479):1049–1058, 2007b.

D. S. Small and P. R. Rosenbaum. War and wages: The strength of instrumental variables and their sensitivity to unobserved biases. *Journal of the American Statistical Association*, 103(483):924–933, 2008.

G. D. Smith and S. Ebrahim. 'Mendelian randomization': can genetic epidemiology contribute to understanding environmental determinants of disease? *International journal of epidemiology*, 32(1):1–22, Feb. 2003. ISSN 0300-5771.

G. D. Smith and S. Ebrahim. Mendelian randomization: prospects, potentials, and limitations. *International Journal of Epidemiology*, 33(1):30–42, Feb. 2004. ISSN 0300-5771, 1464-3685. doi: 10.1093/ije/dyh132.

A. J. Sovey and D. P. Green. Instrumental variables estimation in political science: A readers guide. *American Journal of Political Science*, 55(1):188–200, 2011.

J. Splawa-Neyman, D. Dabrowska, and T. Speed. On the application of probability theory to agricultural experiments. essay on principles. section 9. *Statistical Science*, 5(4):465–472, 1990.

D. Staiger and J. H. Stock. Instrumental variables regression with weak instruments. *Econometrica*, 65(3):557–586, 1997a.

D. Staiger and J. H. Stock. Instrumental variables regression with weak instruments. *Econometrica*, 65(3):557–586, 1997b.

J. H. Stock, J. H. Wright, and M. Yogo. A survey of weak instruments and weak identification in generalized method of moments. *Journal of Business & Economic Statistics*, 20 (4):518–529, 2002a.

J. H. Stock, J. H. Wright, and M. Yogo. A survey of weak instruments and weak identification in generalized method of moments. *Journal of Business & Economic Statistics*, 20 (4), 2002b.

N. J. Timpson, D. A. Lawlor, R. M. Harbord, T. R. Gaunt, I. N. M. Day, L. J. Palmer, A. T. Hattersley, S. Ebrahim, G. Lowe, A. Rumley, and G. Davey Smith. C-reactive protein and its role in metabolic syndrome: Mendelian randomisation study. *The Lancet*, 366(9501):1954–1959, 2005.

J. Wang and E. Zivot. Inference on structural parameters in instrumental variables regression with weak instruments. *Econometrica*, pages 1389–1404, 1998.

J. M. Wooldridge. *Econometric Analysis of Cross Section and Panel Data*. MIT press, 2nd ed. edition, 2010.

M. Zeelenberg and R. Pieters. Consequences of regret aversion in real life: The case of the dutch postcode lottery. *Organizational Behavior and Human Decision Processes*, 93(2): 155–168, 2004.

O. Zuk, S. F. Schaffner, K. Samocha, R. Do, E. Hechter, S. Kathiresan, M. J. Daly, B. M. Neale, S. R. Sunyaev, and E. S. Lander. Searching for missing heritability: Designing rare variant association studies. *Proceedings of the National Academy of Sciences*, 111 (4):E455–E464, 2014.