基于惩罚拟似然方法的孟德尔随机化分析

摘要

推断变量之间的因果效应是很多学科都关心的话题，但是实际问题中因果效应容易受混杂因素和反因果性的影响。计量经济学中最先提出了工具变量这一方法来解决这些问题。生物学中用于推断性状之间因果效应的孟德尔随机化就是工具变量法的一个应用。孟德尔随机化常用的估计方法有 Wald、Two-stage 和 Adjusted two-stage 方法，但是这些方法在实际使用中偏差较大。

本文针对二元结果，基于广义线性混合模型，利用惩罚拟似然方法（PQL）来进行因果效应的推断。本文考虑了单个工具变量和多个工具变量两种情况，并根据理论推导设计算法进行了模拟实验，来测试 PQL 的估计效果，并和其他三种方法进行了比较。

关键词：孟德尔随机化；二元结果；广义线性混合模型；惩罚拟似然

Abstract

Inferring causal effects between variables is a topic of concern for many disciplines, but in practice causal effects are susceptible to confounding factors and anti-causality. Instrumental variables were the first to be proposed in econometrics to solve these problems. Mendelian randomization, which is used in biology to infer causal effects between traits, is an application of the instrumental variable method. Mendelian randomization commonly used estimation methods include Wald, Two-stage, and Adjusted two-stage methods. However, these methods deviate greatly in actual use.

In this paper, based on the generalized linear mixed model, we use penalised quasi-likelihood(PQL) to estimate causal effects. In this paper, we mainly consider the two cases of single instrumental variable and multiple instrumental variables, and design the simulation experiment according to the theoretical derivation to test the estimation effect of PQL and compare it with the other three methods.

Keywords: Mendelian Randomization; binary outcome; generalized linear mixed model; penalised quasi-likelihood

1 问题背景与现状

1.1 MR 的基本原理

在生物学中，推断相关变量之间的因果方向是很普遍的问题，然而简单的回归分析无法做到这一点。两个变量之间的联系可能是因果关系，但是因果性的方向（A 导致 B 或者 B 导致 A）并不能确定。而且还可能存在无法观测的因素会同时影响这两个变量，从而导致它们的联系，称之为混杂因素（confounding factor）。在第二种情形下，单独变量对结果的因果效应可能是 0。即使假设的因果方向是正确的，如果单独变量和某个未观测的混杂相关，那么它对结果的因果效应的估计也很有可能是有偏的。孟德尔随机化（Mendelian Randomization，MR）就是一项希望对因果效应进行无偏估计的技术。

假设性状 A 和性状 B 相关，并且它们的相关性是因为 A 导致 B，那么任何影响性状 A 的变量应该也会影响性状 B。于是推断 A 和 B 之间因果关系的关键就在于找到一个“工具”，它和 A 之间以一种已知的方向相联系。生物学家在这一点上很有优势，因为几乎所有感兴趣的性状都至少部分地受遗传效应的影响，并且遗传效应有多个原因支持它充当这种工具。首先，在遗传关联中，因果性的方向一定是从遗传多态性指向感兴趣的性状。其次，通常测量的环境暴露往往和广泛的行为、社会和生理因素有关，这会混淆与结果之间的关联。而另一方面，遗传变异对特定性状的影响不受环境等混杂因素的干扰。第三，遗传变异及其效应的测量误差相对较小。最后，在如今全基因组关联分析（GWAS）的时代，基因数据不难得到。

1.2 MR 和 RCT 的类比

为了理解 MR 如何用于推断因果性，一个直观的方式就是类比随机对照试验（RCT）。在 RCT 中，被试者被随机分配到两组治疗的其中一个，以避免治疗和结果之间潜在的混杂，这样因果推断就很清晰了。MR 就为我们创造了一个类似的情景。假设一个特定的等位基因稳定的和性状 A 相关，并且性状 A 导致了性状 B。等位基因从父母传给后代，很大程度上与环境独立，并且遗传了等位基因的人被分配到性状 A 平均效应更高的组，而没有遗传到等位基因的人被分配到平均效应低的组。和 RCT 一样，由基因型定义的组别将会在性状 A 的暴露上产生平均上的差异，而在混杂因素方面不会有差异。因此，按基因型分析等同于 RCT 中的意向性分析，即按照他们被随机分到的组别去分析每个个体，与他们是否真正接受治疗无关。这样就确保了不会再引入混杂因素。

1.3 MR 的假设和 IV 分析

用基因型来做暴露对结果的因果推断是工具变量（IV）分析的一个应用。一般来说，IV 是满足如下假设的变量：

1. 工具变量 Z 和感兴趣的暴露 X 相关联；
2. 混杂因素 U 混淆 X 和结果 Y，Z 和 U 独立；
3. 给定 X 和混杂 U 时，Z 和结果 Y 独立.

这些假设可以用图 1 所示的有向无环图来表示。

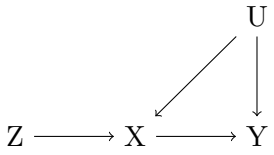


图 1: MR 的假设

1.4 常用的估计方法

考虑二元结果 Y，根据因果图我们建立如下的模型：

$$\mathbf{logit}\mathbb{E}[\mathbf{Y}|\mathbf{X}=x,\mathbf{U}=u]=\beta_0+x\beta_1+u,$$

$$\mathbf{X}=\mathbf{Z}\gamma+\mathbf{V}.$$

为了估计暴露对结果的因果效应 β_1，通常有下面三种做法。

1.4.1 Wald 方法

将 X 的表达式代入 Y 中，我们得到

$$\mathbf{logit}\mathbb{E}[\mathbf{Y}|\mathbf{X},\mathbf{U}]=\beta_0+\beta_1(\mathbf{Z}\gamma+\mathbf{V})+\mathbf{U}.$$

于是我们可以通过下面的 Wald 方法来估计 β_1

$$\hat{\beta}_1=\frac{\hat{\Gamma}}{\hat{\gamma}},$$

这里 $\hat{\gamma}$ 是 X 对 Z 的回归系数，$\hat{\Gamma}$ 是忽略 $\beta_1\mathbf{V}+\mathbf{U}$，$\mathbf{Y}$ 对 $\mathbf{Z}$ 的回归系数。

1.4.2 Two-stage 方法

Two-stage 方法包含两个阶段：第一阶段是 $\mathbf{X}$ 对 $\mathbf{Z}$ 做回归，得到

$$\widehat{\mathbf{X}} = \mathbf{Z}\hat{\gamma},$$

第二阶段是把 $\mathbf{X} = \widehat{\mathbf{X}} + \widehat{\mathbf{V}}$ 代入 $\mathbf{Y}$ 中，即

$$\mathbf{logit}\mathbb{E}[\mathbf{Y}|\mathbf{X}, \mathbf{U}] = \beta_0 + \beta_1(\widehat{\mathbf{X}} + \widehat{\mathbf{V}}) + \mathbf{U}.$$

忽略 $\beta_1\widehat{\mathbf{V}} + \mathbf{U}$，并让 $\mathbf{Y}$ 对 $\widehat{\mathbf{X}}$ 做回归，就得到估计值 $\hat{\beta}_1$。

1.4.3 Adjusted two-stage 方法

上述两种方法由于忽略了包含 $\mathbf{U}$ 和 $\mathbf{V}$ 的项，所以会有偏差。为了对偏差进行调整，我们可以通过第一阶段中残差的估计值 $\widehat{\mathbf{V}} = \mathbf{X} - \widehat{\mathbf{X}}$，来粗略估计混杂项 $\mathbf{U}$。于是第二阶段就变成

$$\mathbf{logit}\mathbb{E}[\mathbf{Y}|\mathbf{X}, \mathbf{U}] = \beta_0 + \beta_1\widehat{\mathbf{X}} + \beta_V\widehat{\mathbf{V}}.$$

2 背景理论介绍

2.1 指数族

考虑随机变量 Y，期望为 $E(Y) = \mu$，方差为 $Var(Y) = \sigma^2$. 如果 Y 的概率密度函数可以写成如下形式就称它的分布属于指数族：

$$f(y|\theta) = m(y)r(\theta)e^{s(\theta)t(y)}$$

这里 θ 表示正规参数（canonical parameter）。为了方便，也可以写成下面的形式：

$$f(y|\theta) = exp[\frac{y\theta - b(\theta)}{a(\phi)} + c(y, \phi)]$$

或者

$$\log[f(y|\theta)] = l(\theta; y, \phi) = \frac{y\theta - b(\theta)}{a(\phi)} + c(y, \phi) \tag{1}$$

这里 ϕ 表示分散参数，它是期望的函数。

举些例子来看，考虑正态、二项以及泊松分布的密度函数：

- 正态分布：

$$\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(y-\mu)^2}{2\sigma^2}}$$

- 二项分布：

$$\binom{n}{y}p^y(1-p)^{n-y}$$

- 泊松分布：

$$\frac{e^{-\lambda}\lambda^y}{y!}$$

对应的对数似然函数为：

- 正态分布：

$$-\log(\sigma\sqrt{2\pi})-\frac{(y-\mu)^2}{2\sigma^2}=\frac{y\mu-\frac{\mu^2}{2}}{\sigma^2}-[\frac{y}{2\sigma^2}+\log(\sigma\sqrt{2\pi})]$$

- 二项分布：

$$\log\binom{n}{y}+y\log(p)+(n-y)\log(1-p)=y\log(\frac{p}{1-p})+n\log(1-p)+\log\binom{n}{y}$$

- 泊松分布：

$$-\lambda+y\log(\lambda)-\log(y!)=+y\log(\lambda)-\lambda-\log(y!)$$

根据对数似然函数，我们可以看出这些分布都属于指数族。

分布	正规参数 θ	$b(\theta)$	$a(\phi)$	$c(y,\phi)$
正态分布	μ	$\frac{\mu^2}{2}$	σ^2	$\frac{y}{2\sigma^2}+\log(\sigma\sqrt{2\pi})$
二项分布	$\log(\frac{p}{1-p})$	$-n\log(1-p)$	1	$\log\binom{n}{y}$
泊松分布	$\log(\lambda)$	λ	1	$-\log(y!)$

我们遇到的很多分布都属于指数族，包括负二项分布、多元正态分布、伽马分布以及贝塔分布。

2.2 拟似然函数

最大似然估计依赖于对数似然函数的偏导 $\frac{\partial l(\theta;y,\phi)}{\partial\theta}$. 在 y 是一维的情况下，这等于 $\frac{y-\mu}{a(\phi)}$. 因此对于指数族，尽管似然函数需要 $c(y,\phi)$ 才能写出来，但是在做最大似然估计

的时候并不需要，只需要 $\frac{y\theta-b(\theta)}{a(\phi)}$. 这就是拟似然想法的来源。

拟似然函数定义为 $\int_y^\mu \frac{y-t}{a(\phi)v(t)}dt$, 这里 $v(t)$ 是方差函数，$a(\phi)$ 表示分散参数函数。举个例子，令 $v(t)=t, a(\phi)=1$, 得到 $\int_y^\mu \frac{y-t}{t}dt = y\log\mu - \mu - (y\log y - y)$, 和泊松分布的对数似然函数相差 $c(y,\phi) = y\log y - y - \log y!$. 类似地，令 $v(t)=1, a(\phi)=\phi^2$ 得到 $\int_y^\mu \frac{y-t}{\phi^2}dt = -\frac{(y-\mu)^2}{2\phi^2}$, 即为正态分布对数似然函数的拟似然部分。

下面解释一下分散参数 ϕ 及其函数 $a(\phi)$ 的含义。以计数数据（count data）为例，我们通常假设其服从泊松分布。泊松分布要求期望和方差相等，即 $E(y)=\lambda=Var(y)$. 然而在实际数据中常常会有超扩散的现象，即样本方差远大于均值，因此远大于理论方差。对具有超扩散现象的计数数据建模的常用方法就是在假设的方差上乘一个分散参数函数 $a(\phi)$. 在上面拟似然的例子中，我们定义 $v(t)=t$, 这和泊松分布的假设一样，现在令 $a(\phi)=\phi$. 得到拟似然函数为 $\int_y^\mu \frac{y-t}{\phi t}dt = \frac{y\log\mu-\mu-(y\log y-y)}{\phi}$. 因此 $E(y)=\mu$, 而 $Var(y)=\phi\mu$. 虽然不存在某个概率分布满足这个条件，但是在对计数数据建模时，利用这样的模型往往效果很好。

2.3 广义线性混合模型

假设有 n 组观测值，第 i 组观测值由一个响应变量 y_i，以及两个解释变量 $\mathbf{x_i}, \mathbf{z_i}$ 组成。其中 y_i 是一维的，$\mathbf{x_i}$ 是 p 维列向量，对应 p 维的固定效应 $\boldsymbol{\alpha}$，$\mathbf{z_i}$ 是 q 维列向量，对应 q 维的随机效应 $\mathbf{b}$. 我们假设：给定随机效应 $\mathbf{b}$ 时，y_i 条件独立，且条件均值为 $E(y_i|\mathbf{b})=\mu_i^b$, 条件方差为 $var(y_i|\mathbf{b})=\phi a_i v(\mu_i^b)$, 这里 $v(\cdot)$ 是指定的方差函数，a_i 是已知的常数，ϕ 是分散参数，可能知道也可能不知道。条件均值通过一个连接函数 $g(\mu_i^b)=\eta_i^b$ 来和线性预测子 $\eta_i^b = \mathbf{x_i}^T\boldsymbol{\alpha} + \mathbf{z_i}^T\mathbf{b}$ 关联。记观测值 $\mathbf{y}=(y_1,\cdots,y_n)^T$，设计矩阵 $\mathbf{X}=(\mathbf{x_1},\cdots,\mathbf{x_n})^T, \mathbf{Z}=(\mathbf{z_1},\cdots,\mathbf{z_n})^T$. 于是条件均值满足

$$E(\mathbf{y}|\mathbf{b}) = g^{-1}(\mathbf{X}\boldsymbol{\alpha} + \mathbf{Z}\mathbf{b})$$

模型还假定随机效应 $\boldsymbol{b}$ 服从多元正态分布，均值为 $\mathbf{0}$，协方差矩阵为 $\boldsymbol{D}=\boldsymbol{D}(\boldsymbol{\theta})$，其依赖于未知向量 $\boldsymbol{\theta}$。在大多数例子中，比如 y_i 条件服从二项分布、泊松分布或者超几何分布，散布参数 ϕ 都被预先设定。然而在其他一些应用当中，它可能会和 $\boldsymbol{\theta}$ 一起当做参数估计。

可以通过极大化下面的积分拟似然函数来估计这里的参数 $(\boldsymbol{\alpha},\boldsymbol{\theta})$：

$$e^{ql(\alpha,\theta)} \propto |\mathbf{D}|^{-1/2}\int exp[-\frac{1}{2\phi}\sum_{i=1}^n d_i(y_i;\mu_i^b) - \frac{1}{2}\mathbf{b}^T\mathbf{D}^{-1}\mathbf{b}]d\mathbf{b}$$

这里

$$d_i(y,\mu) = -2\int_y^{\mu} \frac{y-\mu}{a_i v(u)} du$$

3 本文的模型和理论介绍

3.1 单个 IV 的 MR 推断

考虑下面的广义线性混合模型：

$$\mathbf{logit}(\mu_i|x_i,u_i) = \beta_0 + \beta_1 x_i + u_i,$$

$$x_i = z_i\gamma + v_i, \quad i = 1,2,\cdots,n.$$

这里 x_i, y_i, u_i 分别表示个体 i 的暴露，结果以及混杂项。其中 y_i 是取值为 0,1 的随机变量，x_i, u_i 是连续随机变量。$\mu_i|x_i,u_i$ 表示给定 x_i,u_i 时 y_i 的条件均值。由于 u_i 是混杂项，它与 x_i 也相关，即 x_i 有所谓的“内生性”。因而引入工具变量 z_i，并假设随机误差 v_i 与 z_i 独立。

为了便于对感兴趣参数 β_1 进行推断，我们这里假设 u_i, v_i 相互独立并且服从正态分布，即：

$$\begin{pmatrix} u_i \\ v_i \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}\right)$$

于是有

$$\begin{pmatrix} u_i \\ x_i \end{pmatrix} | z_i \sim N\left(\begin{pmatrix} 0 \\ z_i\gamma \end{pmatrix}, \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}\right)$$

进而得到给定 x_i, z_i 时 u_i 的条件分布

$$u_i|x_i,z_i \sim N(\frac{\sigma_1}{\sigma_2}\rho(x_i - \gamma z_i), \sigma_1^2(1-\rho^2)).$$

为了方便计算，用矩阵形式表示上述结果，并记 $\Theta = (\beta_0, \beta_1, \sigma_1, \sigma_2, \rho, \gamma)$：

$$\mathbf{logit}(\boldsymbol{\mu}|\mathbf{x},\mathbf{u}) = \begin{pmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} + \mathbf{I_n u}$$

$$\mathbf{u}|\mathbf{x},\mathbf{z};\Theta \sim N(\boldsymbol{\mu}_1, \Sigma_1), \quad \mathbf{x}|\mathbf{z};\Theta \sim N(\boldsymbol{\mu}_2, \Sigma_2)$$

这里

$$\boldsymbol{\mu}_1 = \frac{\sigma_1}{\sigma_2}\rho(\mathbf{x} - \gamma\mathbf{z}), \quad \Sigma_1 = diag(\sigma_1^2(1-\rho^2))$$

$$\boldsymbol{\mu}_2 = \gamma\mathbf{z}, \quad \Sigma_2 = diag(\sigma_2^2)$$

下面计算模型的积分拟似然函数：（可能需要补充拟似然函数，以及解释这里 ai 的含义。）

$$\begin{aligned}
Ql(\Theta) &= \int e^{ql(\mathbf{y}|\mathbf{x},\mathbf{u})} p(\mathbf{x},\mathbf{u}|\mathbf{z};\Theta) d\mathbf{u} \\
&= \int exp[-\frac{1}{2\phi}\sum_{i=1}^{n} d_i(y_i;\mu_i|x_i,u_i)] p(\mathbf{u}|\mathbf{x},\mathbf{z};\Theta) p(\mathbf{x}|\mathbf{z};\Theta) d\mathbf{u} \\
&\propto |\Sigma_2|^{-\frac{1}{2}} exp[-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_2)^T\Sigma_2^{-1}(\mathbf{x}-\boldsymbol{\mu}_2)] \\
&|\Sigma_1|^{-\frac{1}{2}} \int exp[-\frac{1}{2\phi}\sum_{i=1}^{n} d_i(y_i;\mu_i|x_i,u_i) - \frac{1}{2}(\mathbf{u}-\boldsymbol{\mu}_1)^T\Sigma_1^{-1}(\mathbf{u}-\boldsymbol{\mu}_1)] d\mathbf{u}
\end{aligned}$$

这里 $d_i(y,\mu) = -2\int_y^\mu \frac{y-t}{a_i v(t)} dt$ 表示拟合的偏差度量。

于是对数拟似然函数

$$\begin{aligned}
ql(\Theta) &= \log Ql(\Theta) \\
&= -\frac{1}{2}\log|\Sigma_2| - \frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_2)^T\Sigma_2^{-1}(\mathbf{x}-\boldsymbol{\mu}_2) - \frac{1}{2}\log|\Sigma_1| + \log\int exp[-\kappa(\mathbf{u})] d\mathbf{u}
\end{aligned} \tag{2}$$

其中

$$\kappa(\mathbf{u}) = \frac{1}{2\phi}\sum_{i=1}^{n} d_i(y_i;\mu_i|x_i,u_i) + \frac{1}{2}(\mathbf{u}-\boldsymbol{\mu}_1)^T\Sigma_1^{-1}(\mathbf{u}-\boldsymbol{\mu}_1)$$

为了处理积分部分，我们用 Laplace 方法对其进行近似计算。

将 $\kappa(\mathbf{u})$ 在 $\tilde{\mathbf{u}}$ 处二阶泰勒展开：

$$\kappa(\mathbf{u}) \approx \kappa(\tilde{\mathbf{u}}) + (\mathbf{u}-\tilde{\mathbf{u}})^T\kappa'(\tilde{\mathbf{u}}) + \frac{1}{2}(\mathbf{u}-\tilde{\mathbf{u}})^T\kappa''(\tilde{\mathbf{u}})(\mathbf{u}-\tilde{\mathbf{u}})$$

将上式代入2中得到

$$ql(\Theta) \approx -\frac{1}{2}\log|\Sigma_2| - \frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_2)^T\Sigma_2^{-1}(\mathbf{x}-\boldsymbol{\mu}_2) - \frac{1}{2}\log|\Sigma_1| - \frac{1}{2}\log|\kappa''(\tilde{\mathbf{u}})| - \kappa(\tilde{\mathbf{u}}) \tag{3}$$

这里 $\tilde{\mathbf{u}}$ 表示 $\kappa(\mathbf{u})$ 的极小值点，即 $\kappa'(\mathbf{u})=\mathbf{0}$ 的解：

$$\kappa'(\mathbf{u})=-\sum_{i=1}^{n}\frac{(y_i-\mu_i|x_i,u_i)\mathbf{e}_i}{\phi a_i v(\mu_i|x_i,u_i)\mathbf{logit}'(\mu_i|x_i,u_i)}+\Sigma_1^{-1}(\mathbf{u}-\mu_1)=\mathbf{0}$$

再对 $\mathbf{u}$ 求导得到

$$\begin{aligned}\kappa''(\mathbf{u})&=\sum_{i=1}^{n}\frac{\mathbf{e}_i\mathbf{e}_i^T}{\phi a_i v(\mu_i|x_i,u_i)[\mathbf{logit}'(\mu_i|x_i,u_i)]^2}+\Sigma_1^{-1}+\mathbf{R}\\&=\mathbf{I_n}^T\mathbf{W}\mathbf{I_n}+\Sigma_1^{-1}+\mathbf{R}\\&=\mathbf{W}+\Sigma_1^{-1}+\mathbf{R}\end{aligned}$$

这里 $\mathbf{W}$ 是 n 阶对角阵，对角元为 $w_i=\{\phi a_i v(\mu_i|x_i,u_i)[\mathbf{logit}'(\mu_i|x_i,u_i)]^2\}^{-1}$. 余项

$$\mathbf{R}=-\sum_{i=1}^{n}(y_i-\mu_i^b)\mathbf{z}_i\frac{\partial}{\partial\mathbf{b}}[\phi a_i v(\mu_i|x_i,u_i)\mathbf{logit}'(\mu_i|x_i,u_i)]$$

对于二项分布,方差函数 $v(\mu)=\mu(1-\mu)$,而 $\mathbf{logit}'(\mu)=\frac{1}{\mu(1-\mu)}$,所以 $v(\mu_i|x_i,u_i)\mathbf{logit}'(\mu_i|x_i,u_i)=1$，从而 $\mathbf{R}=\mathbf{0},\kappa''(\mathbf{u})=\mathbf{W}+\Sigma_1^{-1}$.

将 $\kappa'(\mathbf{u}),\kappa''(\mathbf{u})$ 代入3，得到

$$\begin{aligned}ql(\Theta)\approx&-\frac{1}{2}\log|\Sigma_2|-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_2)^t\Sigma_2^{-1}(\mathbf{x}-\boldsymbol{\mu}_2)-\frac{1}{2}\log|\mathbf{I}+\mathbf{W}\Sigma_1|\\&-\frac{1}{2\phi}\sum_{i=1}^{n}d_i(y_i;\mu_i|x_i,\tilde{u_i})-\frac{1}{2}(\tilde{\mathbf{u}}-\boldsymbol{\mu}_1)^t\Sigma_1^{-1}(\tilde{\mathbf{u}}-\boldsymbol{\mu}_1)\end{aligned}\tag{4}$$

$\mathbf{W}$ 作为均值的函数，我们假设其随参数 Θ 变化的幅度很小，因而将第三项忽略。我们寻找 $\Theta,\mathbf{u}$ 来最大化如下目标函数，称为 PQL（penalized quasi-likelihood）：

$$\begin{aligned}ql(\Theta,\mathbf{u})\approx&-\frac{1}{2}\log|\Sigma_2|-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_2)^t\Sigma_2^{-1}(\mathbf{x}-\boldsymbol{\mu}_2)\\&-\frac{1}{2\phi}\sum_{i=1}^{n}d_i(y_i;\mu_i|x_i,u_i)-\frac{1}{2}(\mathbf{u}-\boldsymbol{\mu}_1)^t\Sigma_1^{-1}(\mathbf{u}-\boldsymbol{\mu}_1)\end{aligned}\tag{5}$$

下面我们分别令 $\Theta, \mathbf{u}$ 的一阶偏导等于 0，并化简得到如下方程组：

$$
\begin{aligned}
&u_i : \frac{y_i-\mu_i}{\phi} = \frac{1}{\sigma_1^2(1-\rho^2)}(u_i - \frac{\sigma_1}{\sigma_2}\rho(x_i-\gamma z_i)), \quad i=1,\cdots,n\\
&\beta_0 : \sum_{i=1}^{n}(y_i-\mu_i)=0,\\
&\beta_1 : \sum_{i=1}^{n}x_i(y_i-\mu_i)=0,\\
&\sigma_1 : \sum_{i=1}^{n}u_i[\frac{u_i}{\sigma_1} - \frac{\rho(x_i-\gamma z_i)}{\sigma_2}]=0,\\
&\sigma_2 : -\frac{n}{\sigma_2} + \frac{1}{\sigma_2^3}\sum_{i=1}^{n}(x_i-\gamma z_i)^2 - \frac{\rho}{\sigma_1(1-\rho^2)\sigma_2^2}\sum_{i=1}^{n}[u_i - \frac{\sigma_1}{\sigma_2}\rho(x_i-\gamma z_i)](x_i-\gamma z_i)=0,\\
&\rho : \frac{\rho}{1-\rho^2}\sum_{i=1}^{n}[u_i - \frac{\sigma_1}{\sigma_2}\rho(x_i-\gamma z_i)]^2 - \frac{\sigma_1}{\sigma_2}\sum_{i=1}^{n}[u_i - \frac{\sigma_1}{\sigma_2}\rho(x_i-\gamma z_i)](x_i-\gamma z_i)=0,\\
&\gamma : \frac{1}{\sigma_2}\sum_{i=1}^{n}(x_i-\gamma z_i)z_i - \frac{\rho}{\sigma_1(1-\rho^2)}\sum_{i=1}^{n}[u_i - \frac{\sigma_1}{\sigma_2}\rho(x_i-\gamma z_i)]z_i=0.
\end{aligned}
$$

解之即得各参数的估计值。

3.2 多个 IV 的 MR 推断

在 MR 中我们经常会遇到多个基因做工具变量的情形，对3.1中的模型稍作修改：

$$\mathbf{logit}(\mu_i|x_i,u_i) = \beta_0 + \beta_1 x_i + u_i,$$

$$x_i = \mathbf{z_i}^T\boldsymbol{\gamma} + v_i, \quad i=1,2,\cdots,n.$$

这里 $\mathbf{z_i}$ 是 q 维列向量，表示第 i 个个体 q 个工具变量的取值，$\boldsymbol{\gamma} = (\gamma_1,\gamma_2,\cdots,\gamma_q)^T$ 是对应的效应。记 $(\mathbf{z_1},\mathbf{z_2},\cdots,\mathbf{z_n}) = \mathbf{z}$.

同样假设 u_i, v_i 相互独立并且服从正态分布，即：

$$\begin{pmatrix}u_i\\v_i\end{pmatrix} \sim N\left(\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}\sigma_1^2 & \rho\sigma_1\sigma_2\\ \rho\sigma_1\sigma_2 & \sigma_2^2\end{pmatrix}\right)$$

于是有

$$\begin{pmatrix}u_i\\x_i\end{pmatrix}|z_i \sim N\left(\begin{pmatrix}0\\\mathbf{z_i}^T\boldsymbol{\gamma}\end{pmatrix}, \begin{pmatrix}\sigma_1^2 & \rho\sigma_1\sigma_2\\ \rho\sigma_1\sigma_2 & \sigma_2^2\end{pmatrix}\right)$$

进而得到给定 x_i, z_i 时 u_i 的条件分布

$$u_i|x_i, z_i \sim N(\frac{\sigma_1}{\sigma_2}\rho(x_i - \mathbf{z_i}^T\boldsymbol{\gamma}), \sigma_1^2(1-\rho^2)).$$

记 $\Theta = (\beta_0, \beta_1, \sigma_1, \sigma_2, \rho, \gamma_1, \gamma_2, \cdots, \gamma_q)$, 将上面结果写成矩阵形式：

$$\mathbf{u}|\mathbf{x}, \mathbf{z}; \Theta \sim N(\boldsymbol{\mu}_1, \Sigma_1), \quad \mathbf{x}|\mathbf{z}; \Theta \sim N(\boldsymbol{\mu}_2, \Sigma_2)$$

这里

$$\boldsymbol{\mu}_1 = \frac{\sigma_1}{\sigma_2}\rho(\mathbf{x} - \mathbf{z}^T\boldsymbol{\gamma}), \quad \Sigma_1 = diag(\sigma_1^2(1-\rho^2))$$

$$\boldsymbol{\mu}_2 = \mathbf{z}^T\boldsymbol{\gamma}, \quad \Sigma_2 = diag(\sigma_2^2)$$

不难看出，多个 IV 的结果和单个 IV 类似，$\mathbf{y}|\mathbf{x}, \mathbf{u}$ 的拟似然函数不变，只是上面这两个分布的均值稍作修改。因此多个 IV 时模型的拟似然函数还是5式：

$$\begin{aligned} ql(\Theta, \mathbf{u}) \approx & -\frac{1}{2}\log|\Sigma_2| - \frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_2)^t\Sigma_2^{-1}(\mathbf{x} - \boldsymbol{\mu}_2) \\ & - \frac{1}{2\phi}\sum_{i=1}^{n} d_i(y_i; \mu_i|x_i, u_i) - \frac{1}{2}(\mathbf{u} - \boldsymbol{\mu}_1)^t\Sigma_1^{-1}(\mathbf{u} - \boldsymbol{\mu}_1) \end{aligned} \tag{6}$$

仍然令 $\Theta, \mathbf{u}$ 的一阶偏导等于 0，并化简得到如下方程组：

$$u_i : \frac{y_i - \mu_i}{\phi} = \frac{1}{\sigma_1^2(1-\rho^2)}(u_i - \frac{\sigma_1}{\sigma_2}\rho(x_i - \mathbf{z_i}^T\boldsymbol{\gamma})), \quad i = 1, 2, \cdots, n$$

$$\beta_0 : \sum_{i=1}^{n}(y_i - \mu_i) = 0,$$

$$\beta_1 : \sum_{i=1}^{n} x_i(y_i - \mu_i) = 0,$$

$$\sigma_1 : \sum_{i=1}^{n} u_i[\frac{u_i}{\sigma_1} - \frac{\rho(x_i - \mathbf{z_i}^T\boldsymbol{\gamma})}{\sigma_2}] = 0,$$

$$\sigma_2 : -\frac{n}{\sigma_2} + \frac{1}{\sigma_2^3}\sum_{i=1}^{n}(x_i - \mathbf{z_i}^T\boldsymbol{\gamma})^2 - \frac{\rho}{\sigma_1(1-\rho^2)\sigma_2^2}\sum_{i=1}^{n}[u_i - \frac{\sigma_1}{\sigma_2}\rho(x_i - \mathbf{z_i}^T\boldsymbol{\gamma})](x_i - \mathbf{z_i}^T\boldsymbol{\gamma}) = 0,$$

$$\rho : \frac{\rho}{1-\rho^2}\sum_{i=1}^{n}[u_i - \frac{\sigma_1}{\sigma_2}\rho(x_i - \mathbf{z_i}^T\boldsymbol{\gamma})]^2 - \frac{\sigma_1}{\sigma_2}\sum_{i=1}^{n}[u_i - \frac{\sigma_1}{\sigma_2}\rho(x_i - \mathbf{z_i}^T\boldsymbol{\gamma})](x_i - \mathbf{z_i}^T\boldsymbol{\gamma}) = 0,$$

$$\gamma_j : \frac{1}{\sigma_2}\sum_{i=1}^{n}(x_i - \mathbf{z_i}^T\boldsymbol{\gamma})z_{ij} - \frac{\rho}{\sigma_1(1-\rho^2)}\sum_{i=1}^{n}[u_i - \frac{\sigma_1}{\sigma_2}\rho(x_i - \mathbf{z_i}^T\boldsymbol{\gamma})]z_{ij} = 0, \quad j = 1, 2, \cdots, q.$$

解之即得各参数的估计值。

4 模拟实验

4.1 单个 IV

我们把 $\beta_0, \beta_1, \sigma_1, \rho, \gamma$ 的真值固定在（2,1,1,0.7,1），让 σ_2 取不同的值（1,2,3）来测试本文的算法对感兴趣参数 β_1 的估计效果，并和1.4中介绍的常用估计方法进行比较。

首先生成 $\mathbf{z}: z_i \sim b(2, 0.3), i = 1, \cdots, n$，以及 $\mathbf{u}, \mathbf{v}$：

$$\begin{pmatrix} u_i \\ v_i \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}\right), i = 1, \cdots, n$$

再通过 $x_i = z_i\gamma + v_i$ 来生成 $\mathbf{x}$。根据 $E(y_i|x_i, u_i) = \mathbf{logit}^{-1}(\beta_0 + \beta_1 x_i + u_i)$，所以可以通过 $y_i \sim b(1, \frac{1}{1+e^{-(\beta_0+\beta_1 x_i+u_i)}})$ 来生成 $\mathbf{y}$。

4.1.1 $\sigma_2 = 1$

在样本量 n=1000 以及同一组参数 $\Theta = (2, 1, 1, 1, 0.7, 1)$ 下，调整不同的 ϕ 来估计参数，对每一个 ϕ 我们都做了 500 次模拟试验来分析 β_1 的估计效果，结果如下表：

ϕ	4	8	10	12	14	16
MEAN	1.1839449	1.0822218	1.0606009	1.0413668	1.0205299	1.0125709
MSE	0.0578864	0.0280463	0.0256339	0.0221890	0.0215251	0.0221387
RMSE	0.2405960	0.1674703	0.1601059	0.1489599	0.1467145	0.1487907

ϕ	18	20	22	24	26
MEAN	1.0040174	1.016287	0.9976137	1.0073227	0.9862698
MSE	0.0213881	0.026540	0.0256022	0.0247455	0.0254040
RMSE	0.1462466	0.162911	0.1600068	0.1573070	0.1593862

表 1: 不同 ϕ 对估计效果的影响

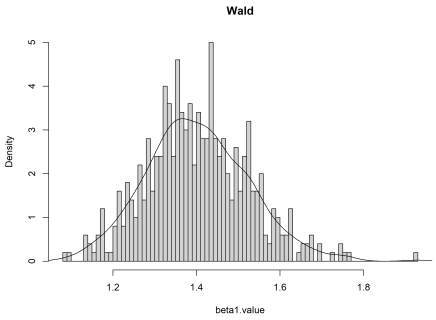
从上表可以看出 ϕ 大于 14 的时候均值都比较接近真值 1。再结合 MSE 的角度来看，ϕ 在 14 到 18 之间的时候 MSE 较小，因此我们在可以这之间选取 ϕ，这里我们选择 ϕ=16。ϕ=16 时 PQL 对各个参数的估计以及四种方法下 β_1 的估计值的直方图（图 2）和箱线图（图 3）如下：

	beta0	beta1	sigma1	sigma2	rho
MEAN	1.796992	1.0125709	0.6626897	0.9993237	0.8150630
MSE	0.057976	0.0221387	1.7978139	0.0005378	0.0253831
RMSE	0.240782	0.1487907	1.3408258	0.0231905	0.1593208

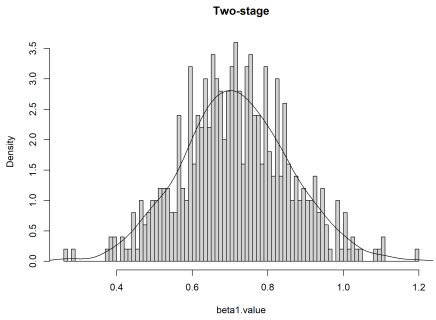
表 2: ϕ=16 时 PQL 对各参数的估计值

	Wald	Twostage	Adjusted	PQL
Mean	1.4025735	0.7151835	0.9374527	1.0125709
MSE	0.1769805	0.1014669	0.0368553	0.0221387
RMSE	0.4206905	0.3185387	0.1919773	0.1487907

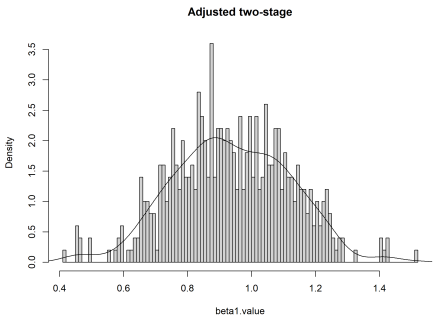
表 3: 四种方法对 β_1 的估计



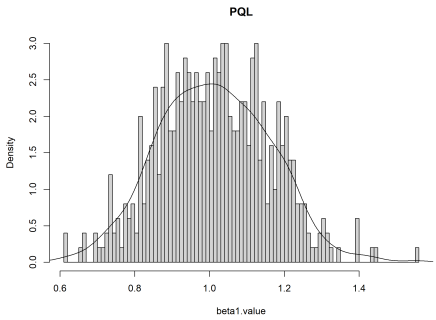
(a) Wald



(b) Two-stage



(c) Adjusted two-stage



(d) PQL

图 2: 四种估计方法的直方图

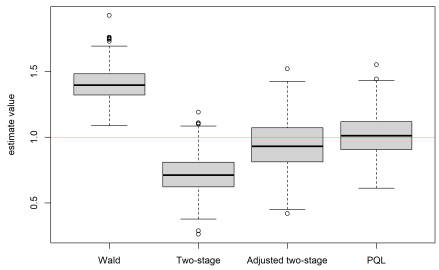


图 3: 四种估计方法的箱线图

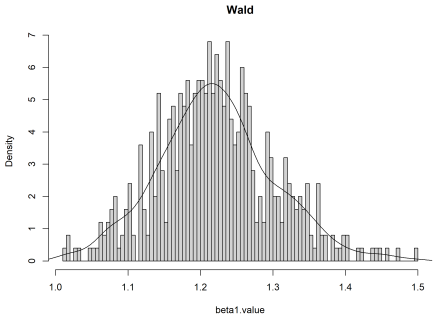
从表 2 可以看到，PQL 对 β_1, σ_2 的估计偏差很小，β_0, σ_1 和 ρ 偏差较大。从表 3 可以看出：Wald 方法的均值偏大，Two-stage 和 Adjusted two-stage 方法均值偏小，PQL 方法的均值很接近真值 1，并且 PQL 的 MSE 最小。从图 2 的直方图来看，除了 Adjusted two-stage 方法，其他三种方法的估计结果都比较接近正态分布。再来看图 3 的箱线图，可以看出 PQL 的中位数也很接近真值 1，而 Wald 方法偏大，Two-stage 和 Adjusted two-stage 方法偏小。

4.1.2 $\sigma_2 = 2$

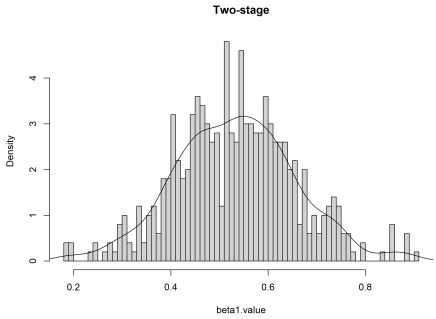
保持其他参数不变，将 σ_2 增加到 2，根据4.1.1的结果，我们选择 ϕ=16 来估计。结果如下：

	Wald	Twostage	Adjusted	PQL
Mean	1.2213233	0.5326495	0.9392972	0.9900014
MSE	0.0554318	0.2336879	0.0400902	0.0175840
RMSE	0.2354395	0.4834128	0.2002253	0.1326046

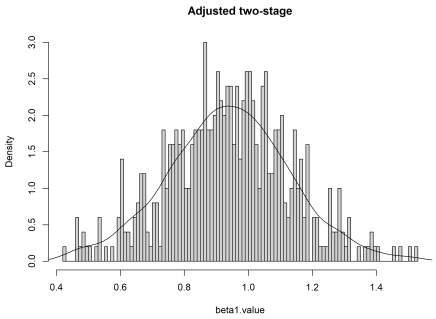
表 4: 四种方法对 β_1 的估计



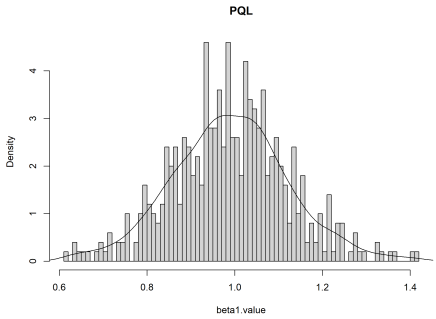
(a) Wald



(b) Two-stage



(c) Adjusted two-stage



(d) PQL

图 4: 四种估计方法的直方图

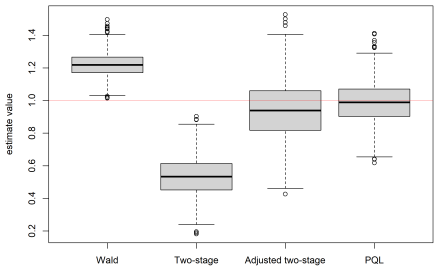


图 5: 四种估计方法的箱线图

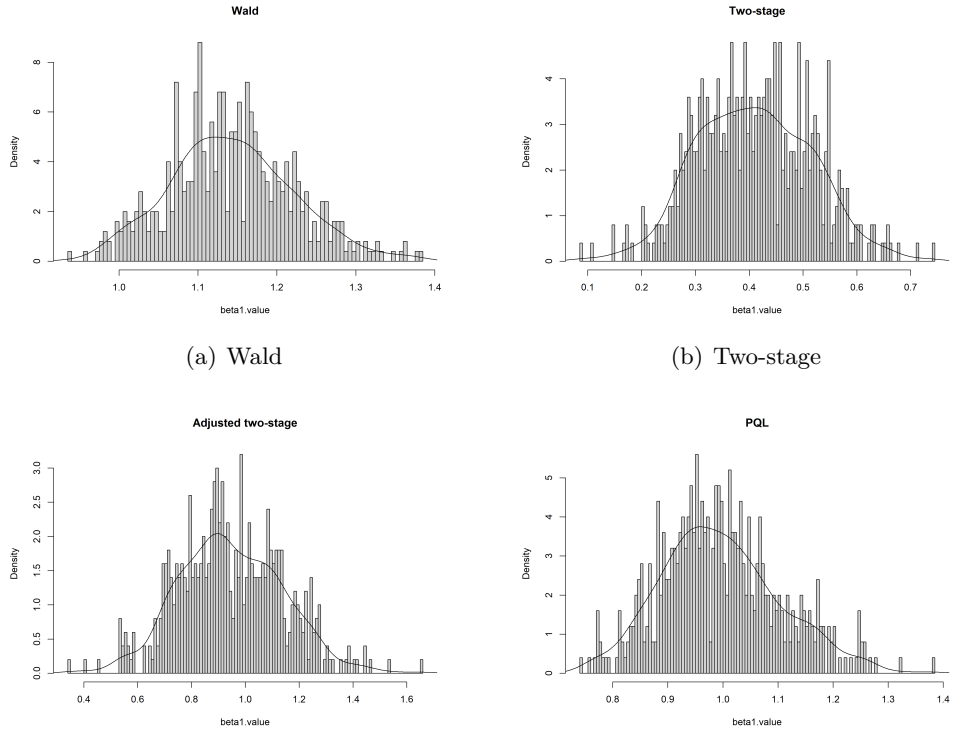
从表 4 可以看出，Wald 方法均值偏大，Two-stage 和 Adjusted two-stage 方法偏小，PQL 的均值很接近真值 1，并且 PQL 的 MSE 最小。图 4 的直方图中，四种方法都比较接近正态分布。图 5 的箱线图可以看出，PQL 的中位数很接近真值 1，Wald 方法偏大，Two-stage 和 Adjusted two-stage 方法偏小。

4.1.3 $\sigma_2 = 3$

保持其他参数不变，继续增加 σ_2 到 3，仍然采用4.1.1的结果选择 ϕ=16 来估计。结果如下：

	Wald	Twostage	Adjusted	PQL
Mean	1.1425917	0.4088278	0.9464101	0.9953955
MSE	0.0264979	0.3606139	0.0398811	0.0116271
RMSE	0.1627819	0.6005113	0.1997025	0.1078291

表 5: 四种方法对 β_1 的估计



(a) Wald

(b) Two-stage

(c) Adjusted two-stage

(d) PQL

图 6: 四种估计方法的直方图

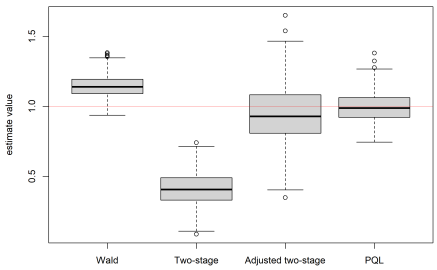


图 7: 四种估计方法的箱线图

表 5 的结果和之前类似：Wald 方法均值偏大，Two-stage 和 Adjusted two-stage 方法偏小，PQL 的均值很接近真值 1 且 MSE 最小。图 6 的直方图中，Wald 和 PQL 方法比较接近正态分布。图 7 的箱线图也和之前结果类似：PQL 的中位数很接近真值 1，Wald 方法偏大，Two-stage 和 Adjusted two-stage 方法偏小。

4.1.4 样本量 n 对估计效果的影响

最后我们看看样本量 n 对 PQL 估计效果的影响。选取上面 $\sigma_1 = 1$ 中 ϕ=16 的情况，让 n 从 1000 逐渐增加到 2000，对 β_1 的估计结果如下表：

n	1000	1200	1400	1600	1800	2000
MEAN	1.0125709	1.0121556	1.0089284	1.0065202	1.0008515	0.9927272
MSE	0.0221387	0.0194035	0.0149723	0.0136570	0.0123448	0.0117533
RMSE	0.1487907	0.1392965	0.1223615	0.1168632	0.1111072	0.1084126

表 6: 样本量 n 对估计效果的影响

从表 6 可以看出，随着 n 的增加，均值越来越接近真值 1，MSE 也逐渐减小，说明估计越来越稳定。

4.2 多个 IV

和单个 IV 类似，我们把 $\beta_0, \beta_1, \sigma_1, \rho$ 固定在（2,1,1,0.7），不同的是，由于 $\boldsymbol{\gamma}$ 是 q 维向量，我们假设其 q 个分量独立同分布于正态分布，因此在生成数据时，先通过标准正态生成 $\boldsymbol{\gamma}$ 的 q 个分量 $\gamma_1, \cdots, \gamma_q$。然后让 σ_2 取不同的值来测试多个 IV 情形下 PQL 对 β_1 的估计效果。

$\mathbf{z}, \mathbf{u}, \mathbf{v}$ 的生成方式和前面一样：

$$z_i \sim b(2, 0.3), \quad \begin{pmatrix} u_i \\ v_i \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}\right), i = 1, \cdots, n$$

再通过 $x_i = \mathbf{z_i}^T\boldsymbol{\gamma} + v_i$ 生成 $\mathbf{x}$，以及 $y_i \sim b(1, \frac{1}{1+e^{-(\beta_0+\beta_1 x_i+u_i)}})$ 来生成 $\mathbf{y}$。 以下的模拟实验都取 q=10，样本量 n=1000。

4.2.1 $\sigma_2 = 1$

首先让 $\sigma_2 = 1$，和单个 IV 的做法一样，设置不同的 ϕ 来估计参数，并对每个 ϕ 做 500 次模拟实验来分析 β_1 的估计效果。结果如下：

	ϕ=1	ϕ=2	ϕ=3	ϕ=4	ϕ=5	ϕ=6
MEAN	0.9905775	1.0121207	1.0370737	1.0278325	1.0253565	1.0316098
MSE	0.0096428	0.0122881	0.0159599	0.0144816	0.0126092	0.0128875
RMSE	0.0981977	0.1108516	0.1263323	0.1203396	0.1122909	0.1135232

表 7: 不同 ϕ 对估计效果的影响

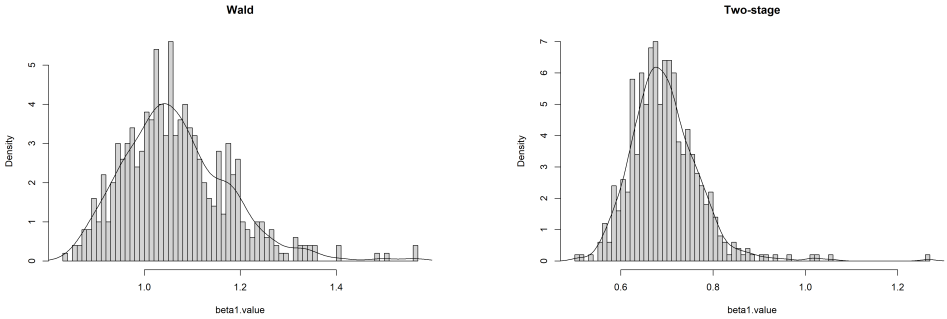
从上表可以看出，ϕ=1 的时候均值最接近真值 1，MSE 也最小。因此在这组参数下，选取 ϕ=1 来进行估计是相对比较好的。为了进一步评价 PQL 的估计效果，表 2 列出了 ϕ=1 时 PQL 对各个参数的估计值，以及四种方法下 β_1 的估计值的直方图（图 1）和箱线图（图 2）。

	beta0	beta1	sigma1	sigma2	rho
MEAN	1.7810870	0.9905775	0.3391456	0.9927311	0.9822052
MSE	0.0882860	0.0096428	2.7672623	0.0006079	0.0845528
RMSE	0.2971296	0.0981977	1.6635090	0.0246547	0.2907797

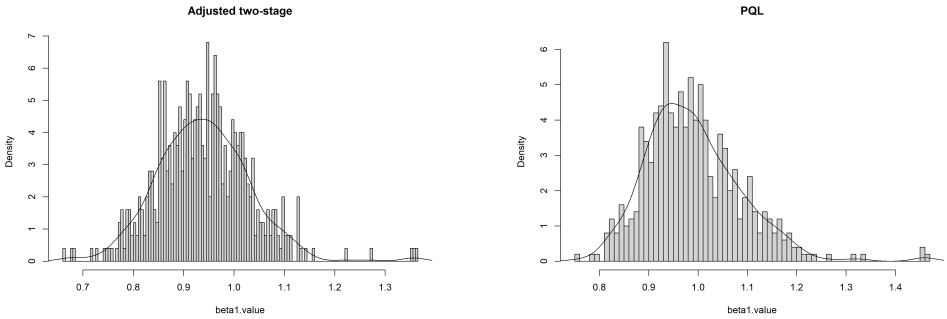
表 8: ϕ=1 时 PQL 对各参数的估计值

	Wald	Twostage	Adjusted	PQL
Mean	1.0669129	0.6952067	0.9394912	0.9905775
MSE	0.0169122	0.0987110	0.0122521	0.0096428
RMSE	0.1300468	0.3141831	0.1106891	0.0981977

表 9: 四种方法对 β_1 的估计



(a) Wald (b) Two-stage



(c) Adjusted two-stage (d) PQL

图 8: 四种估计方法的直方图

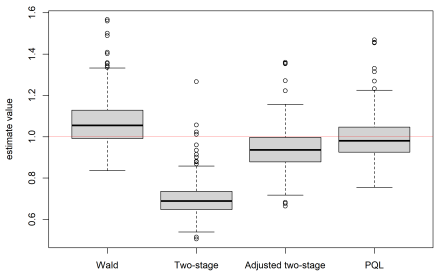


图 9: 四种估计方法的箱线图

从表 3 可以看出，PQL 对 β_1, σ_2 的估计较准，对其他参数的估计有偏差。从表 4 可以看出，Wald 方法的均值偏大，Two-stage 和 Adjusted two-stage 方法均值偏小，PQL 方法的均值很接近真值 1，并且 PQL 的 MSE 最小。从图 5 的直方图来看，Two-stage,Adjusted two-stage 和 PQL 都比较接近正态分布。图 6 的箱线图中，其他三种方法的中位数都离

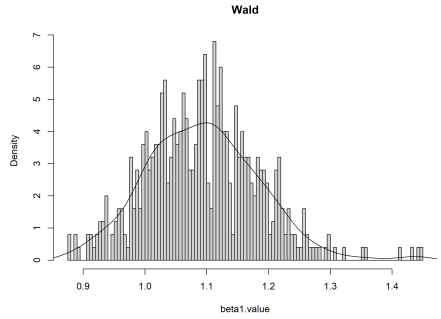
真值偏差较大，PQL 比较接近真值 1。

4.2.2 $\sigma_2 = 2$

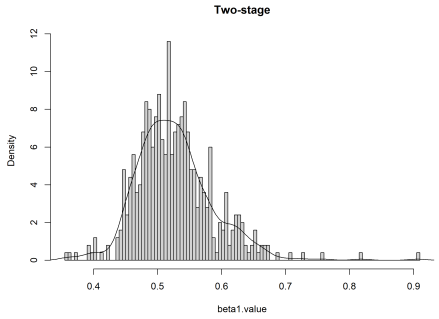
将 σ_2 增加到 2，根据4.2.1的结果，我们选择 ϕ=1 来进行估计。结果如下：

	Wald	Twostage	Adjusted	PQL
Mean	1.0938302	0.5267848	0.9373422	1.0101945
MSE	0.0171949	0.2276324	0.0131613	0.0084185
RMSE	0.1311295	0.4771084	0.1147229	0.0917521

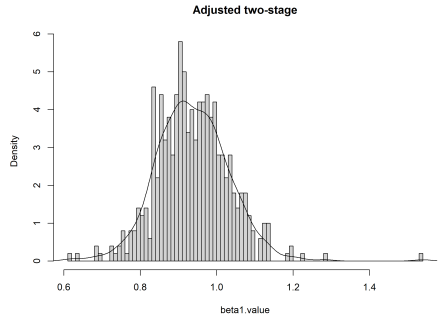
表 10: 四种方法对 β_1 的估计



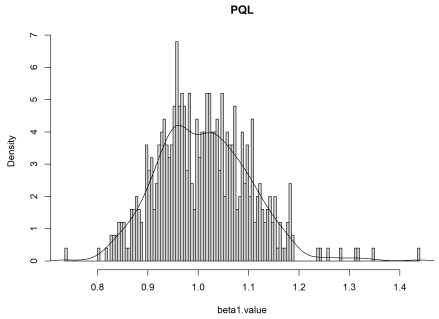
(a) Wald



(b) Two-stage



(c) Adjusted two-stage



(d) PQL

图 10: 四种估计方法的直方图

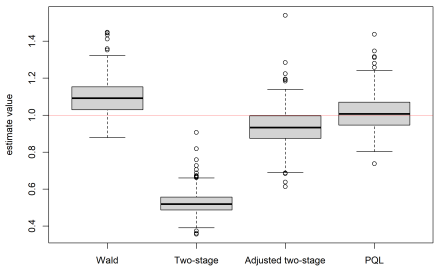


图 11: 四种估计方法的箱线图

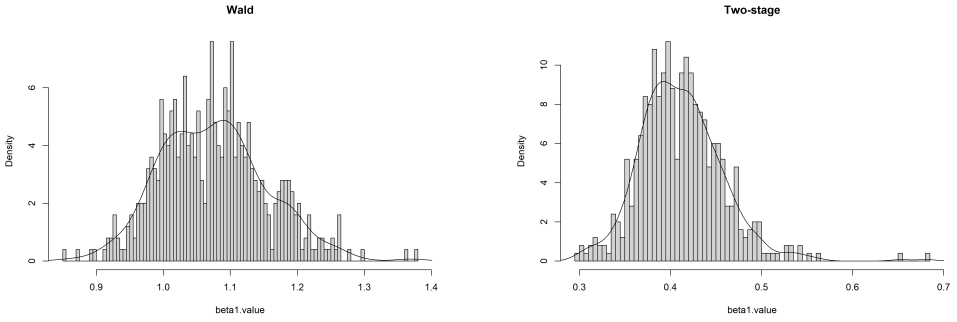
从表 4 可以看出，PQL 对 β_1 的估计较准，MSE 相比其他三种方法也最小。图 5 的直方图中，Two-stage,Adjusted two-stage 和 PQL 都比较接近正态分布。图 6 的箱线图可以看出，PQL 的中位数和真值 1 很接近，而其他三种方法都偏差较大。

4.2.3 $\sigma_2 = 3$

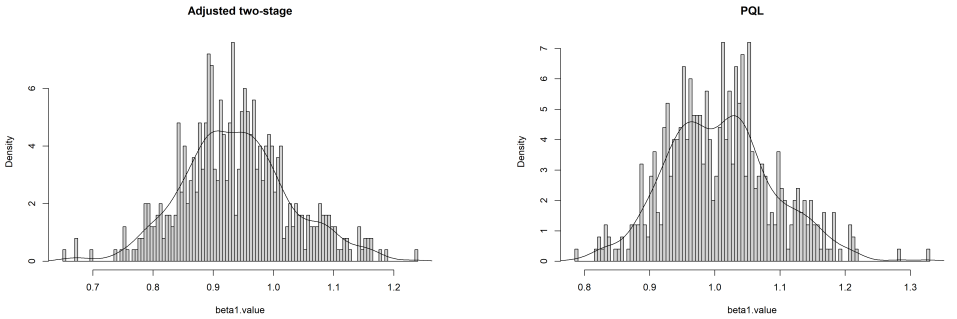
继续增加 σ_2 到 3，仍然设置 ϕ=1，结果如下：

	Wald	Twostage	Adjusted	PQL
Mean	1.0750116	0.4116525	0.9380382	1.0084602
MSE	0.0119141	0.3482657	0.0119621	0.0067936
RMSE	0.1091518	0.5901404	0.1093715	0.0824234

表 11: 四种方法对 β_1 的估计



(a) Wald (b) Two-stage



(c) Adjusted two-stage (d) PQL

图 12: 四种估计方法的直方图

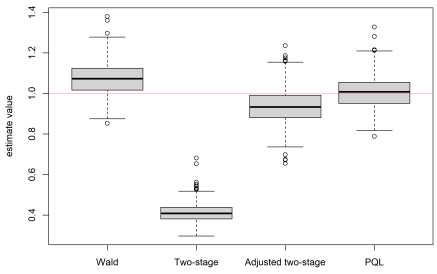


图 13: 四种估计方法的箱线图

表 9 的结果和前面类似：Wald 方法均值偏大，Two-stage 和 Adjusted two-stage 方法均值偏小，PQL 方法的均值很接近真值 1，并且 PQL 的 MSE 最小。而图 9 的直方图呈现出和之前不一样的情况：四种方法的估计都出现了“双峰”。图 10 的箱线图可以看出：这种情况下 Two-stage 方法的偏差极大，而 PQL 仍然是比较准的。

4.2.4 样本量 n 对估计效果的影响

最后我们看看样本量 n 对 PQL 估计效果的影响。选取上面 $\sigma_1 = 1$ 中 ϕ=1 的情况，让 n 从 1000 逐渐增加到 2000，对 β_1 的估计结果如下表：

	n=1000	n=1200	n=1400	n=1600	n=1800	n=2000
MEAN	0.9905775	0.9862202	0.9839118	0.9872714	0.9842450	0.9850806
MSE	0.0096428	0.0092393	0.0072887	0.0071503	0.0064342	0.0064669
RMSE	0.0981977	0.0961213	0.0853740	0.0845596	0.0802134	0.0804172

表 12: 样本量 n 对估计效果的影响

从表 13 可以看出，随着 n 的增加，均值变化不大，MSE 逐渐减小，说明估计越来越稳定。

5 讨论

变量之间的因果关系是很多学科都很关心的问题，包括生物学、经济学、教育学等等。传统的回归模型不能直接用来推断因果效应，这是因为变量之间的相关性可能是由“混杂”因素引起的，而并不是因果性；即便是有因果性，但是因果关系的方向也不能确定。于是最早在经济学里引入了“工具变量”来解决这个问题。它的想法就是找到一个工具变量 Z，它和暴露 X 的因果关系已知，并且只能通过 X 来影响结果 Y，这样就能排除混杂因素 U 的影响，从而对 X，Y 之间的因果效应进行推断。孟德尔随机化就是工具变量分析法在生物学中的应用。MR 利用遗传变异作为工具变量来推断性状之间的因果效应，这是因为遗传变异有很多优势：它和性状之间的因果关系是确定的，并且它基本不受环境等混杂因素干扰，测量误差也很小。

MR 常用的估计方法有 Wald、Two-stage 和 Adjusted two-stage 方法，它们在估计的时候都忽略了某些项，因此在实际估计中偏差较大。本文提出的 PQL 估计方法是基于广义线性混合模型的拟似然函数来进行因果效应的推断。之所以不用似然函数，是因为对于离散的响应变量 Y，似然函数的形式比较复杂，在求积分的时候比较困难。而拟似然函数形式简单，结合 Laplace 方法可以避免积分的计算。

在模拟实验中，我们分两种情况：单个 IV 和多个 IV。两种情况的做法类似：固定 $\beta_0, \beta_1, \sigma_1, \rho, \gamma$（在多个 IV 中是固定 $\boldsymbol{\gamma}$ 的分布）不动，让 σ_2 从小到大变化，观察 PQL 对感兴趣参数 β_1 的估计效果。在实验中我们发现，ϕ 的选取对参数的估计影响很大，但是选好 ϕ 的取值可以对参数进行很好的估计。在单个 IV 的模拟实验中我们发现，ϕ 取 16 的时候估计效果很好，而多个 IV 中 ϕ 取 1 的时候效果很好。而其他三种方法在均值

上都表现出了较大程度的偏差：Wald 方法往往偏大，Two-stage 和 Adjusted two-stage 方法往往偏小。并且随着 σ_2 的逐渐增大，我们发现 Two-stage 方法的偏差也越来越大，而 Adjusted two-stage 的偏差比较稳定。从 MSE 的角度来看：PQL 估计的 MSE 往往也是最小的，并且随着样本量的增加，其 MSE 也是逐渐减小，说明提高样本量也在一定程度上增加了估计的稳定性。

在实验过程中也有一些值得思考的问题：首先是关于 ϕ 的选取，由于 ϕ 对估计效果的影响很大，所以必须对每组参数选择一个合适的 ϕ。这使得对每组参数我们都要不断调整 ϕ 的取值重复做模拟实验，比较麻烦。但是在实验中我们发现，σ_1 的变化对 ϕ 的选取影响比较大，而改变 σ_2 的取值对 ϕ 的选取影响较小。第二个问题是关于 σ_2 变大时，直方图中出现的双峰现象。

最后是本课题后续可能的研究方向。一是可以将 PQL 方法推广到工具变量具有多效性的情形：即工具变量 $\mathbf{Z}$ 不仅能通过暴露 X 影响结果 Y，也可以直接影响 Y。具体地，在3.2中考虑模型

$$\mathbf{logit}(\mu_i|x_i,u_i)=\beta_0+\beta_1 x_i+\mathbf{z_i}^T\boldsymbol{\alpha}+u_i,$$

$$x_i=\mathbf{z_i}^T\boldsymbol{\gamma}+v_i,\quad i=1,2,\cdots,n.$$

其中 $\boldsymbol{\alpha}$ 是多效性的效应值。我们同样可以用 PQL 的方法去估计该模型下的因果效应 β_1。

第二个是关于参数求解的算法。以单个 IV 的情况为例，本文是对 PQL（5 式）中的所有参数同时求解，需要解一个 n+6 元的方程组，计算起来并不快。Green（1987）[6] 发展了 Fisher scoring 算法，可以迭代地求解该方程组。大致思路是通过定义一个工作向量 $Y,Y_i=\eta_i^b+(y_i-\mu_i^b)g'(\mu_i^b),i=1,\cdots,n$，每次迭代之前先通过某个分布随机生成 Y，然后依次算出 $\hat{\boldsymbol{\beta}},\hat{\mathbf{u}},(\hat{\sigma_1},\hat{\sigma_2},\hat{\rho},\hat{\gamma})$，如此不断迭代。这种做法的好处在于不需要同时解 n+6 个方程组，事实上每次迭代只需要解一个二元方程组算出 $\hat{\boldsymbol{\beta}}$，然后可以直接表示出 $\hat{\mathbf{u}}$，再解一个四元方程组求出剩下的参数即可。

最后是 PQL 偏差的校正问题。从前面的实验可以看出，可以通过适当的选取 ϕ 来减小偏差，但是依据 MSE 来选 ϕ 十分麻烦。Breslow(1995) [3] 通过比较 PQL 和真正的对数似然函数，给出了 PQL 在数据分散程度较小时的一阶和二阶校正公式，模拟实验结果显示校正的 PQL 估计偏差显著减小。另外 Kuk（1995）[8] 提出一种基于模拟实验的方法来校正 PQL 中的偏差，大致思路是先根据观测值给出一个有偏的估计 $\theta^0=\tilde{\theta}$, 然后根据这组参数去模拟生成 H 组数据，再用这 H 组数据分别估计参数 θ，求出平均值 $\bar{\theta}$，那么下一次迭代的参数就是 $\theta^1=\theta^0+(\tilde{\theta}-\bar{\theta})$。如此迭代下去直至达到收敛条件。这种方法可以有效改善 PQL 估计的偏差，并且不需要考虑 ϕ 的选取。

参考文献

[1] Baiocchi, M., Cheng, J., and Small, D. S. Instrumental variable methods for causal inference. *Statistics in medicine 33*, 13 (2014), 2297–2340.

[2] Breslow, N. E., and Clayton, D. G. Approximate inference in generalized linear mixed models. *Journal of the American statistical Association 88*, 421 (1993), 9–25.

[3] Breslow, N. E., and Lin, X. Bias correction in generalised linear mixed models with a single component of dispersion. *Biometrika 82*, 1 (1995), 81–91.

[4] Burgess, S., Small, D. S., and Thompson, S. G. A review of instrumental variable estimators for mendelian randomization. *Statistical methods in medical research 26*, 5 (2017), 2333–2355.

[5] Davey Smith, G., and Hemani, G. Mendelian randomization: genetic anchors for causal inference in epidemiological studies. *Human molecular genetics 23*, R1 (2014), R89–R98.

[6] Green, P. J. Penalized likelihood for general semi-parametric regression models. *International Statistical Review/Revue Internationale de Statistique* (1987), 245–259.

[7] Hinde, J., and Demétrio, C. G. Overdispersion: models and estimation. *Computational statistics & data analysis 27*, 2 (1998), 151–170.

[8] Kuk, A. Y. Asymptotically unbiased estimation in generalized linear models with random effects. *Journal of the Royal Statistical Society: Series B (Methodological) 57*, 2 (1995), 395–407.

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

[10] Lin, X., and Breslow, N. E. Analysis of correlated binomial data in logistic-normal models. *Journal of Statistical Computation and Simulation 55*, 1-2 (1996), 133–146.

[11] McCullagh, P. Quasi-likelihood functions. *The Annals of Statistics 11*, 1 (1983), 59–67.

[12] McCullagh, P., and Nelder, J. A. *Generalized linear models*. Routledge, 2019.

[13] Morrison, J., Knoblauch, N., Marcus, J. H., Stephens, M., and He, X. Mendelian randomization accounting for correlated and uncorrelated pleiotropic effects using genome-wide summary statistics. *Nature genetics 52*, 7 (2020), 740–747.

[14] Ng, E. S., Carpenter, J. R., Goldstein, H., and Rasbash, J. Estimation in generalised linear mixed models with binary outcomes by simulated maximum likelihood. *Statistical Modelling 6*, 1 (2006), 23–42.

[15] Stroup, W. W. *Generalized linear mixed models: modern concepts, methods and applications*. CRC press, 2012.

[16] 于天琦, 徐文涛, 苏雅娜, and 李静. 孟德尔随机化研究基本原理, 方法和局限性. 中国循证医学杂志 *21*, 10 (2021), 8.

[17] 费宇. 线性和广义线性混合模型及其统计诊断. 线性和广义线性混合模型及其统计诊断, 2013.

[18] 陈希孺. 广义线性模型的拟似然法. 广义线性模型的拟似然法, 2011.

致谢