系数的解释及控制变量

司继春

2025年11月

核心解释变量与控制变量

- 外生性通常是一个非常难以成立的假设
- 而在微观实证中，通常我们仅仅关心一个核心解释变量对于被解释变量的影响，此时回归方程可以写为:

$$y = \gamma \cdot w + \tilde{x}'\tilde{\beta} + u$$

其中w为我们关心的核心解释变量，而$\tilde{x}$为其他控制变量（control variables），记$x = [w, \tilde{x}']'$。

- 如果只关注γ的识别和估计，我们可以放弃外生性假设这个比较强的假设，转而使用稍微弱一点的假设

条件外生性

条件外生性

在线性回归模型：

$$y = \gamma \cdot w + \tilde{x}'\tilde{\beta} + u$$

中，假设在给定$\tilde{x}$的条件下，误差项u与核心解释变量w均值独立，即：

$$\mathbb{E}\left(u|w,\tilde{x}\right) = \mathbb{E}\left(u|\tilde{x}\right)$$

条件外生性

- 以上假设强调了控制变量的重要性：
  - 如果在回归方程：
  $$y = \gamma \cdot w + e$$
  中，核心解释变量w和e相关
    - 然而如果额外控制$\tilde{x}$，在给定$\tilde{x}$的情况下回归方程中w和u不相关，那么仍然可以得到γ的一致估计。

条件外生性

锻炼身体与寿命

- 在辛普森悖论的例子中，回归方程式

$$y = \beta_0 + \beta \cdot \text{exer} + e$$

中的exer和e之间相关的，因为u中含有不可观测的性别因素，而性别因素与exer是相关的。

条件外生性

锻炼身体与寿命

- 然而，如果我们把性别sex作为控制变量：

$$y = \beta_0 + \beta \cdot \text{exer} + \beta_1 \cdot \text{sex} + u$$

  那么根据数据生成过程，e为一个纯随机的扰动，意味着给定sex，exer与u无关，即：

$$\mathbb{E}\left(u|\text{exer}, \text{sex}\right) = \mathbb{E}\left(u|\text{sex}\right)$$

  或者说exer对误差项u没有预测能力，从而条件外生性成立。
- 在这个问题中，条件外生性可以认为是给定性别相同的条件下，u与exer无关，即对于男性（女性），锻炼身体的人与不锻炼身体的人平均而言u是相同的，那么条件外生性即满足。

条件外生性

- 如果记$u^{\perp}=u-\mathbb{E}\left(u|\tilde{x}\right)$，即$u$中不与$\tilde{x}$相关的部分，那么以上假设意味着：

$$
\begin{aligned}
\mathbb{E}\left(u^{\perp}|w,\tilde{x}\right) &= \mathbb{E}\left[u-\mathbb{E}\left(u|\tilde{x}\right)|w,\tilde{x}\right] \\
&= \mathbb{E}\left(u|w,\tilde{x}\right)-\mathbb{E}\left(u|\tilde{x}\right) \\
&= \mathbb{E}\left(u|\tilde{x}\right)-\mathbb{E}\left(u|\tilde{x}\right) \\
&= 0
\end{aligned}
$$

从而我们有：$\mathbb{E}\left(w\cdot u^{\perp}\right)=0$，即$w$与$u^{\perp}$不相关。

条件外生性

- 进而，令$w^{\perp}=w-\mathbb{E}(w|\tilde{x})$，即$w$中与$\tilde{x}$无关的部分，可以得到：

$$\mathbb{C}\left(w^{\perp},u^{\perp}\right)=\mathbb{E}\left(w^{\perp}\cdot u^{\perp}\right)=0$$

  即在给定$\tilde{x}$的条件下，w与u不相关，或者可以写为：$u\perp w|\tilde{x}$。

- 条件外生性是比外生性更弱的假设，因为根据外生性假设，我们必须要求$\mathbb{E}(u|w,\tilde{x})=0$，从而$\mathbb{E}(u|\tilde{x})=\mathbb{E}[\mathbb{E}(u|w,\tilde{x})|\tilde{x}]=0$，从而必然有$\mathbb{E}(u|w,\tilde{x})=\mathbb{E}(u|\tilde{x})=0$成立。
- 条件外生性放松了对于$\tilde{x}$的外生性的假设，即可以允许有一些与$\tilde{x}$相关的因素存在于u中，部分允许$\tilde{x}$可以不是外生的，重要的是在给定$\tilde{x}$的情况下，w与u是不相关的。

条件外生性

能力的代理变量

- 在例教育回报中，如果我们可以观察到另外一个变量，如初中时某次考试分数s，并将其作为控制变量：

$$\ln \text{income} = \gamma \cdot \text{edu} + x'\beta + \delta \cdot s + u$$

其中$u = \text{ability} + \epsilon$，并假设$\mathbb{E}\left(\epsilon | \text{edu}, x, s\right) = 0$，即内生性只来自于不可观测的能力。

- 此时s被称为ability的**代理变量（proxy variable）**
- 如果我们将能力ability对s和其他特征做投影，即：

$$\text{ability} = \eta \cdot s + x'\zeta + e$$

其中$\mathbb{E}\left(e | s, x\right) = 0$

条件外生性

能力的代理变量

- 此时：

$$\begin{aligned}\mathbb{E}\left(u|\text{edu},x,s\right)&=\mathbb{E}\left(\text{ability}+\epsilon|\text{edu},x,s\right)\\&=\mathbb{E}\left(\text{ability}|\text{edu},x,s\right)\\&=\mathbb{E}\left(\eta\cdot s+x'\zeta+e|\text{edu},x,s\right)\\&=\eta\cdot s+x'\zeta+\mathbb{E}\left(e|\text{edu},x,s\right)\end{aligned}$$

  而注意到

$$\begin{aligned}\mathbb{E}\left(u|x,s\right)&=\mathbb{E}\left(\text{ability}+\epsilon|x,s\right)\\&=\mathbb{E}\left(\text{ability}|x,s\right)\\&=\eta\cdot s+x'\zeta\end{aligned}$$

条件外生性

能力的代理变量

- 从而条件独立性成立的条件为：$\mathbb{E}\left(e|\text{edu},x,s\right)=\mathbb{E}\left(e|s,x\right)=0$。
- 即：
    - 如果初中考试分数s和其他x相同的人，其能力的差异不会影响受教育程度，那么条件外生性就是成立的，我们仍然可以得到γ的识别。
    - 然而，即使s相同，能力的差异也许会影响受更高程度教育的可能，此时高考分数相同的人也会因为能力的不同而导致不同的edu，从而$\mathbb{E}\left(e|\text{edu},x,s\right)\neq 0$，那么条件外生性也就不满足了。

条件外生性下的估计

- 如果，我们仅仅假设条件外生性假设，不妨记$x = [w, \tilde{x}']'$, $\beta = \left[\gamma, \tilde{\beta}'\right]'$
- 两边分别对$\tilde{x}$做条件期望，得到：

$$\mathbb{E}(y|\tilde{x}) = \gamma \cdot \mathbb{E}(w|\tilde{x}) + \tilde{x}'\tilde{\beta} + \mathbb{E}(u|\tilde{x})$$

相减，得到：

$$y - \mathbb{E}(y|\tilde{x}) = \gamma \cdot [w - \mathbb{E}(w|\tilde{x})] + [u - \mathbb{E}(u|\tilde{x})]$$

简记为

$$y^{\perp} = \gamma \cdot w^{\perp} + u^{\perp}$$

- 在以上步骤中，我们分别将被解释变量y和核心解释变量w中有关$\tilde{x}$的部分剔除，从而也被称为“**排除**”（**partialing-out**）估计。

条件外生性下的估计

- 由于:

$$\mathbb{C}\left(w^{\perp}, u^{\perp}\right) = \mathbb{E}\left(w^{\perp} \cdot u^{\perp}\right) = 0$$

再次，我们可以使用矩条件:

$$\mathbb{E}\left(w^{\perp} \cdot u^{\perp}\right) = 0$$

对γ进行识别，进而可以使用矩估计方法对γ进行估计。

条件外生性下的估计

- 如果$\mathbb{E}\left(y|\tilde{x}\right)$和$\mathbb{E}\left(w|\tilde{x}\right)$可被观测，可以直接使用$y^{\perp}$对$w^{\perp}$做回归，就可以得到$\gamma$的估计。
- 然而在估计时，由于$\mathbb{E}\left(y|\tilde{x}\right)$和$\mathbb{E}\left(w|\tilde{x}\right)$都是不可观测的，因而我们必须将其替代为样本等价。
  - 前面我们已经介绍过，如果假设$\mathbb{E}\left(y|\tilde{x}\right)$和$\mathbb{E}\left(w|\tilde{x}\right)$都是$\tilde{x}$的线性函数，即假设：

$$\mathbb{E}\left(y|\tilde{x}\right)=\tilde{x}'\delta,\mathbb{E}\left(w|\tilde{x}\right)=\tilde{x}'\eta$$

  - 那么$\mathbb{E}\left(y|\tilde{x}\right)$和$\mathbb{E}\left(w|\tilde{x}\right)$可以分别使用$y$和$w$对$\tilde{x}$做线性回归进行估计。

条件外生性下的估计

线性函数假设

假设

$$\mathbb{E}\left(y|\tilde{x}\right)=\tilde{x}'\delta,\mathbb{E}\left(w|\tilde{x}\right)=\tilde{x}'\eta$$

此时，根据线性投影的性质，有

$$\mathbb{L}\left(y|\tilde{x}\right)=\mathbb{L}\left[\mathbb{E}\left(y|\tilde{x}\right)|\tilde{x}\right]=\tilde{x}'\delta$$
$$\mathbb{L}\left(w|\tilde{x}\right)=\mathbb{L}\left[\mathbb{E}\left(w|\tilde{x}\right)|\tilde{x}\right]=\tilde{x}'\eta$$

可以直接使用线性回归对$\mathbb{L}\left(y|\tilde{x}\right),\mathbb{L}\left(w|\tilde{x}\right)$进行估计。

条件外生性下的估计

- 结合以上，我们可以使用如下步骤对γ进行估计：
  1. 使用y对$\tilde{x}$做线性回归，得到残差$y^{\perp}$
  2. 使用w对$\tilde{x}$做线性回归，得到残差$w^{\perp}$
  3. 使用$y^{\perp}$对$w^{\perp}$做回归，得到$\hat{\gamma}$
- 注意以上步骤与分步回归完全一致。因而实际上，我们可以直接使用y对w和$\tilde{x}$做线性回归：

$$y = \gamma \cdot w + \tilde{x}'\tilde{\beta} + u = x'\beta + u$$

即可。只不过在这里，我们只关注γ的估计，所以我们使用的假设也更弱了。

系数的解释

系数的解释与模型设定实际上是一体两面的：

- 阅读论文时，其他人的模型结果该如何理解——系数解释
- 写作论文时，有一个想法该如何使用模型表达——模型设定

所以只有熟练掌握了系数的解释，自己写作论文时才知道如何设定模型从而讲自己的故事：

- 多读
- 多想
- 多写

系数的解释：总论

在回归方程

$$y = x'\beta + u$$

中，由于我们假设了外生性：$\mathbb{E}\left(u|x\right) = 0$，那么$y$给定$x$的条件期望：

$$\mathbb{E}\left(y|x\right) = x'\beta$$

我们可以使用条件期望函数将参数β表达出来，即对于参数的解释。

系数的解释：虚拟变量情形

- 在离散情况下，我们通常可以使用差分作为工具，将所关心的参数使用条件期望函数表示出来。
- 比如，在模型：

$$y = \gamma \cdot w + \tilde{x}'\tilde{\delta} + u$$

中若$w = 0/1$，且外生性满足，那么：

$$\begin{aligned}\gamma &= \left(\gamma + \tilde{x}'\tilde{\delta}\right) - \left(\tilde{x}'\tilde{\delta}\right) \\ &= \mathbb{E}\left(y|w=1,\tilde{x}\right) - \mathbb{E}\left(y|w=0,\tilde{x}\right)\end{aligned}$$

因而γ可以解释为：

- 其他条件（$\tilde{x}$）相同的条件下，w从0到1所导致的y的绝对值变化
- 或者其他条件（$\tilde{x}$）相同的条件下，$w=1$组比$w=0$组的均值高γ。

系数的解释：虚拟变量（对数）情形

- 很多时候回归中会加入取对数后的变量而非其绝对值。如果被解释变量取对数，即

$$\ln y = \gamma \cdot w + \tilde{x}'\tilde{\delta} + u$$

其中$w = 0/1$，那么γ可以表示为：

$$\gamma = \mathbb{E}(\ln y | w = 1, \tilde{x}) - \mathbb{E}(\ln y | w = 0, \tilde{x})$$

- 一般而言，对数之差可以解释为变动率。由于当$\frac{y_1 - y_0}{y_0} \approx 0$时：

$$\ln y_1 - \ln y_0 = \ln\left(\frac{y_1}{y_0}\right) = \ln\left(1 + \frac{y_1 - y_0}{y_0}\right) \approx \frac{y_1 - y_0}{y_0}$$

即y_0到y_1的百分比变化，因而通常我们也将γ解释为w从0到1导致的y的百分比变化的期望

- 这一解释的前提是$|\gamma|$比较小，如果$|\gamma|$较大，那么可以采用：

$$\ln\left(\frac{y_1}{y_0}\right) = \gamma \Rightarrow y_1 = e^{\gamma} y_0$$

进行解释。

系数的解释：虚拟变量（对数）情形

边界之谜

McCallum(1995)使用了如下回归解释美国和加拿大之间的贸易：

$$x_{ij} = a + b \times y_i + c \times y_j + d \times \mathrm{dist}_{ij} + e \times \mathrm{Dummy}_{ij} + u_{ij}$$

其中i, j为加拿大的省或者美国的州，x_{ij}为从i地区到j地区的出口额的对数，y_i, y_j为地区i, j的对数GDP，dist_{ij}为i和j之间的距离的对数，而$\mathrm{Dummy} = 1$代表加拿大的省份之间的贸易，$\mathrm{Dummy} = 0$代表加拿大的省和美国的州之间的贸易，数据中没有美国的州和州之间的贸易。

系数的解释：虚拟变量（对数）情形

边界之谜

	(1)	(2)
y_i	1.3	1.21
	(0.06)	(0.03)
y_j	0.96	1.06
	(0.06)	(0.03)
dist_{ij}	-1.52	-1.42
	(0.10)	(0.06)
Dummy_{ij}		3.09
		(0.13)
样本量	90	683
调整的R^2	0.890	0.811

系数的解释：虚拟变量（对数）情形

McCallum(1995)

在这个结果中，我们最关注的是第(2)列的结果，Dummy_{ij}的系数为3.09，且是显著的，那么意味着其他条件不变的情况下，加拿大的省内贸易比加拿大省份和美国州之间的跨国贸易额的对数大3.09，即加拿大省内贸易是跨国贸易额的$e^{3.09} \approx 22$倍。

系数的解释：虚拟变量（交叉项）情形

如果$d = 0/1$也为二元变量（$w = 0/1$依然为二元变量）比如：

- $w = 1$代表上过大学，$w = 0$代表没有上过大学
- $d = 1$代表男性，$d = 0$代表女性
- y代表对数收入

如果

$$y = \tau \cdot w + \eta \cdot d + \gamma \cdot w \cdot d + \tilde{x}'\tilde{\delta} + u$$

则

$$\begin{aligned}\gamma = & \left[\mathbb{E}\left(y|\tilde{x}, w=1, d=1\right) - \mathbb{E}\left(y|\tilde{x}, w=1, d=0\right)\right] \\ & - \left[\mathbb{E}\left(y|\tilde{x}, w=0, d=1\right) - \mathbb{E}\left(y|\tilde{x}, w=0, d=0\right)\right]\end{aligned}$$

why?

系数的解释：虚拟变量（交叉项）情形

由于：

$$\begin{cases} \mathbb{E}\left(y|w=1,d=1,\tilde{x}\right)=\tau+\eta+\gamma+\tilde{x}'\tilde{\delta} \\ \mathbb{E}\left(y|w=1,d=0,\tilde{x}\right)=\tau+\tilde{x}'\tilde{\delta} \\ \mathbb{E}\left(y|w=0,d=1,\tilde{x}\right)=\eta+\tilde{x}'\tilde{\delta} \\ \mathbb{E}\left(y|w=0,d=0,\tilde{x}\right)=\tilde{x}'\tilde{\delta} \end{cases}$$

从而：

$$\begin{cases} \eta & =\mathbb{E}\left(y|w=0,d=1,\tilde{x}\right)-\mathbb{E}\left(y|w=0,d=0,\tilde{x}\right) \\ \tau & =\mathbb{E}\left(y|w=1,d=0,\tilde{x}\right)-\mathbb{E}\left(y|w=0,d=0,\tilde{x}\right) \\ \gamma & =\left[\mathbb{E}\left(y|w=1,d=1,\tilde{x}\right)-\mathbb{E}\left(y|w=0,d=1,\tilde{x}\right)\right] \\ & \quad -\left[\mathbb{E}\left(y|w=1,d=0,\tilde{x}\right)-\mathbb{E}\left(y|w=0,d=0,\tilde{x}\right)\right] \end{cases}$$

γ度量了男性、女性的教育回报差异：**影响的异质性（heterogeneity）**

系数的解释：虚拟变量（交叉项）情形

- 而同时注意到，γ也可以表示为：

$$\begin{aligned}\gamma = & \left[\mathbb{E}\left(y|w=1,d=1,\tilde{x}\right)-\mathbb{E}\left(y|w=1,d=0,\tilde{x}\right)\right] \\ & -\left[\mathbb{E}\left(y|w=0,d=1,\tilde{x}\right)-\mathbb{E}\left(y|w=0,d=0,\tilde{x}\right)\right]\end{aligned}$$

- γ也度量了不同教育程度的性别歧视程度
- 两者是等价的：正是因为女性教育回报高，导致了教育降低女性歧视

系数的解释：虚拟变量（交叉项）情形

锻炼身体效果的异质性

在辛普森悖论的例子中，如果我们的数据生成过程为：

$$y_{1i} = 83 - 10d_i + u_{1i}$$
$$y_{0i} = 80 - 10d_i + u_{0i}$$

假设$u_{1i} = u_{0i} = u$，那么我们可以将其写为：

$$y = w \cdot y_1 + (1 - w) \cdot y_0 = 80 - 10 \times d + 3 \times w + u$$

然而，如果数据生成过程为：

$$y_{1i} = 83 - 8d_i + u_{1i}$$
$$y_{0i} = 80 - 10d_i + u_{0i}$$

即锻炼身体对男性和女性的效果不同，那么我们可以将其写为：

$$y = w \cdot y_1 + (1 - w) \cdot y_0 = 80 - 10 \times d + 3 \times w + 2 \times d \times w + u$$

系数的解释：虚拟变量（交叉项）情形

扶教育之贫

- 汪德华、邹杰和毛中根（2019）借助1995年的“义新工程”作为自然实验，将受到该项目影响的县作为处理组（$program = 1$），于此同时区分了1995年年龄为2-6岁的样本（$young = 1$）以及16-20岁（$young = 0$）的样本。
- 由于该项目主要针对义务教育，从而$young = 0$为不受项目影响的年龄组，而$young = 1$的年龄组在整个义务教育阶段都受到了该项目的影响。

系数的解释：虚拟变量（交叉项）情形

扶教育之贫

- 作者设定了如下模型：

$$y_{ic} = \beta \times \text{young}_i + \delta \times \text{program}_c + \lambda \times \text{program}_c \times \text{young}_i + x'_{ic}\delta + \epsilon_{ic}$$

  其中c为县，i为个体，y为受教育年限。
- 从而：
  - β度量了未受到项目影响的县中，95年2-6岁的样本相较于16-20岁的样本的受教育年限差异，或者“代际趋势”；
  - δ则度量了95年16-20岁的样本在受到和未受到项目影响两类地区的受教育年限差异；
  - 而λ实际上度量了受到和没有受到项目影响两类地区的“代际趋势”的差异。

系数的解释：连续情况

对于连续变量，可以使用微分代替差分：

$$\frac{\partial \mathbb{E}\left(y|x\right)}{\partial x_j}$$

即所谓偏效应（partial effects）。比如，对于模型：

$$y = \gamma \cdot w + \tilde{x}'\beta + u$$

那么：

$$\gamma = \frac{\partial \mathbb{E}\left(y|x\right)}{\partial w}$$

即当w增加一单位时，y平均增加γ。

系数的解释：对数情况

同样，如果模型为：

$$\ln y = \gamma \cdot w + \tilde{x}'\beta + u$$

那么：

$$\gamma = \frac{\partial \mathbb{E}\left(\ln y | w, \tilde{x}\right)}{\partial w} = \mathbb{E}\left(\frac{\partial \ln y}{\partial w}\bigg| w, \tilde{x}\right) = \mathbb{E}\left(\frac{1}{y}\frac{\partial y}{\partial w}\bigg| w, \tilde{x}\right)$$

即此时γ可以解释为w增加1单位时，y的百分比变化。

- 比如，如果$\gamma = 0.01$，意味着1单位的w增加会导致y的1%增加

系数的解释：对数的情况

同理，如果模型为：

$$\ln y = \gamma \cdot \ln w + \tilde{x}'\beta + u$$

那么此时

$$\gamma = \mathbb{E}\left(\frac{\partial \ln y}{\partial \ln w}\middle| w, \tilde{x}\right)$$

此时γ可以解释为w增加百分之一时，y的百分比变化，即所谓的弹性（elasticity）。

- 如果$\gamma = 1$，那么当w增加1%（从而$\ln w$增加0.01）时，$\ln y$增加$0.01 \times \gamma = 0.01$，从而$y$相应增加1%。
- 那么如果：

$$y = \beta_1 \ln x_1 + \tilde{x}'\delta + u$$

该如何解释？

实例：引力模型

边界之谜

其结果如下：

	(1)	(2)
y_i	1.3	1.21
	(0.06)	(0.03)
y_j	0.96	1.06
	(0.06)	(0.03)
dist_{ij}	-1.52	-1.42
	(0.10)	(0.06)
Dummy_{ij}		3.09
		(0.13)
样本量	90	683
调整的R^2	0.890	0.811

实例：引力模型

边界之谜

上例中，dist_{ij}为取过对数的地区间距离，而被解释变量为取过对数的贸易额，从而-1.42可以理解为当距离增加1%时，地区间的贸易额会减少1.42%。

偏效应

- 如果模型中存在着解释变量的非线性函数形式，我们通常也是使用偏导：

$$\frac{\partial \mathbb{E}\left(y|x\right)}{\partial w}$$

进行解释，此时通常该效应并非常数，而是存在着**异质性**的影响，我们通常将以上偏导称之为**偏效应**（partial effects）。

- 由于影响是异质性的，即对于每个个体都是不同的，我们通常会计算偏效应的期望：

$$\mathbb{E}\left[\frac{\partial \mathbb{E}\left(y|x\right)}{\partial w}\right]$$

我们称之为**平均偏效应**（average partial effects, APE）。

系数的解释：平方项

在模型中加入平方项一般可以建模影响的非线性性，特别是边际递增、递减等特征，有的时候甚至还能表达U型、倒U型关系。

- 加入平方项：

$$y = \gamma \cdot w + \eta \cdot w^2 + \tilde{x}'\beta + u$$

那么偏效应为：

$$\frac{\partial \mathbb{E}\left(y|w,\tilde{x}\right)}{\partial w} = \gamma + 2 \cdot \eta \cdot w$$

此时w对y的影响是：

- 异质性的
- 非线性的

系数的解释：平方项

- 异质性：
  - 每个人的w（比如年龄）不同，那么w增加一单位时对y的影响也不同。
  - 此时可以计算**平均偏效应**（average partial effects, APE）：

$$\text{APE} = \mathbb{E}\left(\frac{\partial \mathbb{E}\left(y|w,\tilde{x}\right)}{\partial w}\right) = \gamma + 2 \cdot \eta \cdot \mathbb{E}\left(w\right)$$

系数的解释：平方项

- 需要注意的是，一个常见的加入平方项的目的是查看是否存在U型或者倒U型的关系
- 然而，并非$\eta > 0$就是U型，或者反过来$\eta < 0$就是倒U型，我们还需要查看当我们查看对称轴$-\gamma/2\eta$与w的分布范围：
  - 只有当w的分布范围包含了对称轴时，才能做如上断言，否则只能认为具有边际递增/递减的特性。
- 需要额外注意的是，只有当对称轴两边有足够多的样本时才能认为w的分布范围包含了对称轴，如果对称轴的一侧样本量相对较少，仅仅能够说明w对y的影响存在非线性性，或者边际递增/递减的特性，而不能充分说明U型或者倒U型关系。

系数的解释：平方项

对称轴

在quadratic.do中，我们模拟了如下的函数形式：

$$y = 2 - e^{-\frac{x}{2}} + u$$

从而

$$\frac{\partial \mathbb{E}\left(y|x\right)}{\partial x} = \frac{1}{2} e^{-\frac{x}{2}} > 0$$

即一个单调递增但是边际递减的影响。然而如果我们假设

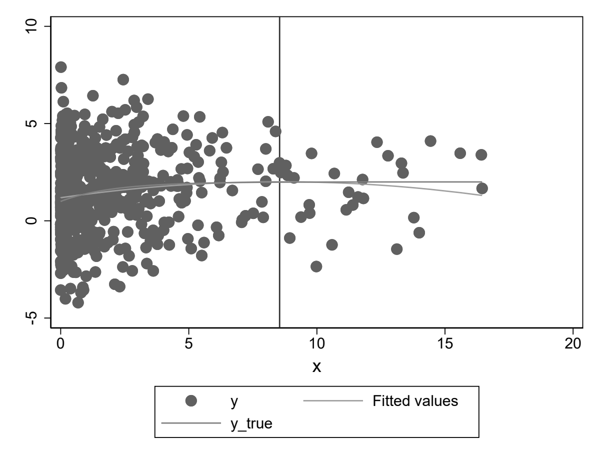
$$\frac{x}{2} \sim \chi^2\left(1\right), u \sim \mathcal{N}\left(0, 4\right)$$

的情况下做模拟，使用回归

$$y = \alpha + \beta x + \delta x^2 + u$$

系数的解释：平方项

对称轴



系数的解释：连续变量交乘项

- 有时我们也会设置交叉项：

$$y = \gamma \cdot w + \eta \cdot h + \tau \cdot w \cdot h + \tilde{x}'\beta + u$$

那么偏效应：

$$\frac{\partial \mathbb{E}\left(y|w, h, \tilde{x}\right)}{\partial w} = \gamma + \tau \cdot h$$

- w对y的影响受到另外一个变量h的影响，其平均偏效应为：

$$\mathbb{E}\left(\frac{\partial \mathbb{E}\left(y|w, \tilde{x}\right)}{\partial w}\right) = \gamma + \tau \cdot \mathbb{E}\left(h\right)$$

系数的解释：连续变量与虚拟变量交乘项

- 有时参与交乘的两项并非都是连续变量，而是将一个虚拟变量与连续变量相乘，即

$$y = \gamma \cdot w + \eta \cdot d + \tau \cdot w \cdot d + \tilde{x}'\beta + u$$

其中d为虚拟变量，w为连续变量，此时可以计算偏效应：

$$\frac{\partial \mathbb{E}\left(y|w,d,\tilde{x}\right)}{\partial w} = \gamma + \tau \cdot d = \begin{cases} \gamma & d = 0 \\ \gamma + \tau & d = 1 \end{cases}$$

从而：

  - τ度量了$d=1$组与$d=0$组之间w对y的影响的异质性，而γ则度量了$d=0$这一组w对y影响的偏效应。
  - 如果考虑d对y的影响，有

$$\mathbb{E}\left(y|d=1,w,\tilde{x}\right) - \mathbb{E}\left(y|d=0,w,\tilde{x}\right) = \eta + \tau \cdot w$$

  则意味着$d=1$组与$d=0$组之间y的平均差异随着w的变化而变化。

实例：交叉项的解释

财政教育支出与城乡收入差距

贾婷月、王笑涵和司继春（2024）通过匹配个体初中时的城市财政教育支出，讨论了财政教育支出对于城乡收入差距的影响，其主要结果为

$$\ln \text{Income}_{ic} = -0.256 \times \text{hukou}_i \times \ln \text{PEE}_{ic} + 0.507 \times \text{hukou}_i + 0.254 \times \ln \text{PEE}_{ic} + \tilde{x}'\beta + u_{ic}$$

其中交乘项的系数有两种解释：

1. 财政教育支出对于农村和城市的个体收入影响是不同的：教育公共支出每增加1%，会使得农村居民收入增加约0.254%，而对于城市这一效应为$(0.254-0.256)\,\% \approx 0$；
2. 城乡之间的收入差距应为$0.507 - 0.256 \times \ln \text{PEE}_{ic}$，随着财政教育支出的增加，城乡之间的收入差距会逐渐缩小。

估计平均偏效应：中心化

在实践中，将描述性统计中的均值带入到偏效应估计中虽然不复杂，但是繁琐，且无法直接获取APE的标准误。为了直接从回归中读出APE，一个实用的处理是把具有平方项、交叉项的变量去平均。比如在平方项和交叉项的设定中：

- 令$w_i^* = w_i - \bar{w}$，从而$\mathbb{E}(w^*) = 0$，如果使用回归方程：

$$y = \gamma^* \cdot w^* + \eta^* \cdot w^{*2} + \tilde{x}'\beta + u$$

那么平均偏效应为：

$$\mathbb{E}\left(\frac{\partial \mathbb{E}(y|w^*, \tilde{x})}{\partial w^*}\right) = \gamma^* + 2 \cdot \eta^* \cdot \mathbb{E}(w^*) = \gamma^*$$

从而如果将w**去平均（demean）**或者**中心化（centering）**，那么回归系数γ^*可以解释为w对y影响的平均偏效应。

估计平均偏效应：中心化

- 令$w_i^* = w_i - \bar{w}, h_i^* = h_i - \bar{h}$，如果使用回归方程：

$$y = \gamma^* \cdot w^* + \eta^* \cdot h^* + \tau^* \cdot w^* \cdot h^* + \tilde{x}'\beta + u$$

那么平均偏效应为：

$$\mathbb{E}\left(\frac{\partial \mathbb{E}\left(y|w^*, \tilde{x}\right)}{\partial w^*}\right) = \gamma^* + \tau^* \cdot \mathbb{E}\left(h^*\right) = \gamma^*$$

从而如果将h中心化，那么同样，回归系数γ^*可以解释为w对y影响的平均偏效应。

实例：中心化

企业进入退出与生产率

文件demean.do做了如下回归：

$$\ln \text{Pollution} = c + \beta \ln \text{pgdp} + \gamma (\ln \text{pgdp})^2 + u$$

即观察污染（二氧化硫）排放是否随着人均GDP先增加后减少。

- 我们首先进行了不去平均的回归，得到的回归系数为$\gamma = -0.342, \beta = 7.976$，而人均GDP的均值为0.702，可以计算平均偏效应应为$7.976 - 2 \times 0.342 \times 0.702 = 0.863$。
- 紧接着使用egen计算人均GDP的平均，然后进行去平均操作，得到的$\beta = 0.863, \gamma = -0.342$。注意到两组回归的$\gamma$估计是完全相同的，而去平均之后的$\beta$具有了平均偏效应的解释。

实例：中心化

老龄化

Cravino、Levchenko和Rojas（2022）使用国家层面数据，通过如下设定探索研究了人口老龄化与经济结构转变的关系：

$$\omega = \alpha + \beta \mathrm{Age} + \gamma_1 \ln \mathrm{GDP\_PC} + \gamma_2 \left(\ln \mathrm{GDP\_PC}\right)^2 + u$$

其中ω为农业、制造业、服务业的比例，Age为65岁以上人口比例，GDP_PC为人均GDP。以制造业为例，如下代码展示了使用中心化和没有中心化的回归：

实例：中心化

老龄化

代码 1: 老龄化与经济结构转变

```
// aging_transformation.do
use datasets/aging_transformation_macro.dta, clear
// 不
gen log_gdp_pc  = log(gdp_pc)
gen log_gdp_pc2 = log_gdp_pc^2
reghdfe hw_share_man log_gdp_pc log_gdp_pc2 share_65plus, a(code)
    vce(cl code)
egen mean_log_gdp_pc = mean(log_gdp_pc)
gen dm_log_gdp_pc2 = (log_gdp_pc-mean_log_gdp_pc)^2
reghdfe hw_share_man log_gdp_pc dm_log_gdp_pc2 share_65plus, a(
    code) vce(cl code)
```

实例：中心化

老龄化

- 如果不对人均GDP做中心化，得到的回归方程为

$$\omega = \alpha - 0.84 \times \text{Age} + 1.24 \times \ln \text{GDP\_PC} - 0.07 \times (\ln \text{GDP\_PC})^2 + u$$

而如果进行中心化，得到的回归方程为

$$\omega = \alpha - 0.84 \times \text{Age} - 0.07 \times \ln \text{GDP\_PC} - 0.07 \times \left(\ln \grave{\text{GDP}}\text{\_PC}\right)^2 + u$$

其中$\ln \grave{\text{GDP}}\text{_PC}$为中心化的对数GDP。

- 可以看到，如果没有做中心化处理，一次项系数为正；然而实际上，随着人均GDP的增加，制造业的就业比重平均而言应该是下降的。

固定效应

如果假设数据生成过程：

$$y_i = \alpha + \beta \cdot w_i + \gamma \cdot d_i + u_i$$

其中$d_i = 0/1$。按照分步回归的步骤：

1. 用w_i和y_i对d_i做回归，得到残差。注意由于d_i为0-1变量，因而实际结果是w_i和y_i被分组去均值。
2. 用e_y对e_w做回归，得到β的估计。

注意在以上步骤中，实际使用的variation是被去掉组均值的残差，因而不同组别之间没有variation，是一个「组内」估计量。

分步回归

实际上，以上β的最小二乘估计为：

$$
\begin{aligned}
\hat{\beta} &= \frac{\sum_{i=1}^{N} d_i\left(w_i-\bar{w}_1\right) y_i+\sum_{i=1}^{N}\left(1-d_i\right)\left(w_i-\bar{w}_0\right) y_i}{\sum_{i=1}^{N}\left[d_i\left(w_i-\bar{w}_1\right)^2+\left(1-d_i\right)\left(w_i-\bar{w}_0\right)^2\right]} \\
&= \frac{\hat{\beta}_1\left[\sum_{i=1}^{N} d_i\left(w_i-\bar{w}_1\right)^2\right]+\hat{\beta}_0\left[\sum_{i=1}^{N}\left(1-d_i\right)\left(w_i-\bar{w}_0\right)^2\right]}{\sum_{i=1}^{N}\left[d_i\left(w_i-\bar{w}_1\right)^2\right]+\sum_{i=1}^{N}\left[\left(1-d_i\right)\left(w_i-\bar{w}_0\right)^2\right]} \\
&\triangleq \hat{\beta}_1 \omega_1+\hat{\beta}_0 \omega_0
\end{aligned} \tag{1}
$$

其中$\omega_1+\omega_0=1$为一个权重，$\hat{\beta}_1$和$\hat{\beta}_0$分别为对应$d_i=1/0$两个组中，使用y对w做回归得到的回归系数。

组内估计量

- 最小二乘估计$\hat{\beta}$是在$d_i = 0/1$的不同组内进行最小二乘估计的一个加权平均，而不涉及不同组之间的比较
- 因而$\hat{\beta}$是一个组内（**within-group**）估计量。
- 实际上，以上结论对任意多个组别的虚拟变量以及多个解释变量都是成立的

固定效应

Simpson悖论

我们使用Stata模拟了Simpson悖论的数据:

- $y = 80 - 10 \times d + 3 \times w + u$
- $d = 1$：男性，$w = 1$：锻炼
- $\mathrm{P}\left(w = 1 | d = 1\right) = 0.8$
- $\mathrm{P}\left(w = 1 | d = 0\right) = 0.3$

固定效应

Simpson悖论

VARIABLES	(1) 全样本	(2) 男性	(3) 女性	(4) 全样本
exer	-2.266***	2.824***	3.134***	3.000***
	(0.277)	(0.113)	(0.0973)	(0.0741)
sex				-9.959***
				(0.0736)
Constant	77.82***	70.14***	79.91***	79.95***
	(0.207)	(0.102)	(0.0525)	(0.0496)
Observations	1,000	509	491	1,000
R-squared	0.063	0.550	0.680	0.952

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

固定效应

- 在上例中，我们通过控制性别得到了锻炼对寿命的真实影响。
- 实际上，以上通过虚拟变量控制分组进行比较的思想可以扩展到任意的分类变量
- 我们经常把这些虚拟变量称之为固定效应（fixed effects），因为通过这些虚拟变量，我们将比较限制在分组内部，相当于“固定”了分组内部那些固定不变的不可观测因素（组内估计量）。
- 比如，当我们控制了地区的固定效应时，即限制了所有的回归的结果都来自于地区内部不同个体之间比较的结果，不存在不同地区之间的个体的比较，由于同一地区的所有地区因素都相同，那么同一地区的x不同所导致的y的不同就不可能是地区因素所引起的。
    - 从而，可以想象的是，如果某个地区只有一个个体或者观测，那么该观测数据将无法与其他观测进行比较，这一地区的数据也就对最终回归结果没有任何贡献了。

固定效应

我们可以使用虚拟变量的形式写出回归方程式:

$$y_i = x_i'\beta + \sum_{r=1}^{R-1} (\gamma_r \times d_{ir}) + u_i$$

其中当i个体属于地区r时，$d_{ir} = 1$，否则为0。或者，我们可以简写为:

$$y_{ir} = x_{ir}'\beta + \gamma_r + u_{ir}$$

固定效应

NTV与俄罗斯大选

Enikolopov、Petrova和Zhuravskaya（2011）研究了媒体（特别是独立的NTV电视台）对于俄罗斯大选的影响，他们的设定如下：

$$\text{vote}_{sr,1999} = \beta_0 + \beta_1 \text{NTV}_{sr,1999} + x'_{sr}\beta + \delta_r + \epsilon_{sr}$$

其中$\text{vote}_{sr,1999}$为地区r的选区s在1999年选举中的结果，我们这里挑选对Unity党的投票结果；$\text{NTV}_{sr,1999}$为估计的电视台覆盖率，δ_r为地区固定效应，x_{sr}包含了经济社会控制变量以及4年前上一届选举的结果变量（ols_fe_ntv.do）

固定效应

- 注意到根据这一设定，所有的地区层面的因素，即只随着地区的不同而不同，但是同一个地区所有个体都相同的变量都会被地区的虚拟变量所完全线性表示出，从而所有这些地区层面的变量都无需加入到回归方程中，或者说地区固定效应控制了所有地区层面的因素。
  - 从而在上例中，控制变量中的所有变量都是选区层面的，没有地区层面的。

固定效应

扶教育之贫

- 在汪德华、邹杰和毛中根（2019）的设定中:

$$y_{ic} = \beta \times \text{young}_i + \delta \times \text{program}_c + \lambda \times \text{program}_c \times \text{young}_i + x_{ic}'\delta + \epsilon_{ic}$$

作者加入了县的固定效应，从而$program_c$被县的固定效应吸收了，其系数是无法估计的。

固定效应

- 有的时候不同的分类变量可能出现嵌套的情况:
  - 城市属于特定省份、细分行业属于特定的三类产业等。
  - 比如，在上例中，在选区和地区之间，还存在城市这一分类：每个地区有多个城市，每个城市有1-2个选区。
- 此时，我们可以控制更“高”层次的固定效应（比如地区），也可以控制更“低”层次的固定效应（比如城市）
- 而由于高层次的虚拟变量可以被低层次的虚拟变量表示出，从而低层次的固定效应是更加严格的控制。
- 当然，虽然理论上固定效应控制的越细越好，但是也需要考虑自由度（$N-K$）的问题
  - 更细致的固定效应可能会导致自由度快速耗散，能够使用的变异很小，导致回归结果不精确、不稳定，也更容易犯第II类错误。

固定效应

NTV与俄罗斯大选

- 在上例中，如果我们把absorb(region)换成更细的absorb(city_id)，可以发现在剔除了缺失值之后，只有137个城市、274个选区（如何观察？）的数据被使用，其他数据都因为一个城市只有一个选区，无法参与比较被剔除了，相较于absorb(region)，观测数少了1450个。
- 大量的样本被剔除导致自由度大大降低，这会大大增加第II类错误的概率，从而造成核心解释变量“错误的”不显著。可能由于这个原因，作者最终还是将聚类层级放在地区层面而非城市层面。

固定效应

- 此外，如果我们有两个分类变量（比如地区地区和行业），那么可以同时加入两个分类变量的虚拟变量进行控制:

$$y_i = x_i'\beta + \sum_{r=1}^{R-1} (\delta_r \times d_{ir}) + \sum_{m=1}^{M-1} (\gamma_m \times d_{im}) + u_i$$

其中d_{ir}为第一个分类变量的固定效应，d_{im}为第二个分类变量的固定效应。

- 以上回归方程经常被简写为:

$$y_{irm} = x_i'\beta + \delta_r + \gamma_m + u_{irm}$$

注意我们在加入这些固定效应时，都删掉了一个虚拟变量以保证矩阵可逆。

固定效应

边界之谜续

- 在之前的例子中介绍了边界之谜，Anderson和Wincoop（2003）提出之所以有边界之谜，是因为McCallum（1995）的设定中没有考虑“多边阻力”（multilateral resistance）。
- 他们通过理论推导得到了如下的回归方程：

$$x_{ij}=a+y_i+y_j+(1-\sigma)\,\mathrm{dist}_{ij}+(1-\sigma)\,\rho\mathrm{Dummy}_{ij}-(1-\sigma)\ln P_i-(1-\sigma)\ln P_j+u_{ij}$$

其中$\ln P_i,\ln P_j$为两个多边阻力，是一个由价格、距离等组成的非线性函数。

固定效应

边界之谜续

- 注意到P_i, P_j只是出口地或者目的地的特征，Redding和Venables（2004）建议直接使用固定效应处理。
- 实际上，出口地和目的地的固定效应不仅仅包括了以上的多边阻力，还包括了一些其他特征，比如是否为内陆国等特征都被吸收进固定效应中，从而解释变量中不需要额外加入其他地区层面的控制变量，只需要加入出口地-目的地层面的变量（如距离、是否有共同边界等）变量。
- 使用reghdfe的PPML版本：ppmlhdfe

```
ppmlhdfe trade lypex lypim lyex lyim ldist border comlang
    colony landl_ex landl_im lremot_ex lremot_im comfrt_wto
    open_wto, a( s1_im s2_ex )
```

可以看到回归中所有只与国家有关的变量，如对数GDP、是否为内陆国等变量都被吸收了，然而两国之间的距离、是否有共同边界、是否有共同语言等变量仍然可以保留

固定效应

- 或者，我们可以使用两个分类变量定义一个新的分类变量，只有当两个分类变量同时成立时才$=1$，并将其作为固定效应在回归中加以控制：

$$y_{irm} = x_i'\beta + \sum_{r,m} (\eta_{rm} \times d_{irm}) + u_{irm}$$

或者：

$$y_{irm} = x_i'\beta + \eta_{rm} + u_{irm}$$

- 这是一种比同时加入两组虚拟变量更为严格的控制方式，相当于将比较限制在同一地区且同一行业水平进行比较。

固定效应

- 当然，这么做的前提是在r, m的组合中，至少要有两个观测。
- 如果在所有的r, m组合中，都不足两个观测，那么意味着以上回归方程中的下角标i是多余的，此时样本量甚至等于d_{rm}的个数（$N = R \times M$），再考虑到x的维度，自由度$N - K < 0$被完全耗光，也就无法得到估计了。
- 但是此时，分别加入两种固定效应是完全可行的，只要$R \times M - R - M$大于x的维数即可。
  - 边界之谜的例子中，数据是出口国$\times$目的国层面的，无法加入出口国$\times$目的国的固定效应。

固定效应

- 比同时加入两组虚拟变量更为严格，之前的设定假设了：

$$\eta_{rm} = \delta_r + \gamma_m$$

- 该假设意味着不可观测的地区因素和行业因素的影响是独立的、可加，显然这一假设并不一定成立，因而使用η_{rm}控制的更加严格。
- 该假设既可以是对y的数据生成过程的，也可以是对x的（习题）

固定效应

固定效应的模拟

- 考虑如下数据生成过程：$\delta_r \sim \mathcal{N}(0,1), \gamma_m \sim \mathcal{N}(0,2)$，而解释变量

$$x_{irm} = h_i + \max(\delta_r, \gamma_m)$$

  其中$h_i \sim \mathcal{U}(0,1)$。考虑如下两种不同的设定：
  1. $y_{irm} = 1 + 2 \times x_{irm} + \delta_r + \gamma_m + u_{irm}$
  2. $y_{irm} = 1 + 2 \times x_{irm} + \delta_r \gamma_m + u_{irm}$
- 可见第1种设定中，$\eta_{rm} = \delta_r + \gamma_m$的设定，而第2种设定不满足。
- simulate_fe.do对以上两种不同的数据生成过程进行了模拟并使用两种不同的固定效用控制方法进行了回归

固定效应

固定效应的模拟

VARIABLES	(1) DGP1	(2) DGP1	(3) DGP2	(4) DGP2
x	1.989***	1.978***	-0.323***	1.877***
	(0.0425)	(0.0666)	(0.0691)	(0.0679)
Constant	1.851***	1.870***	5.063***	1.284***
	(0.0752)	(0.116)	(0.122)	(0.118)
Observations	3,000	3,000	3,000	3,000
R-squared	0.955	0.958	0.777	0.919
R固定效应	Yes	No	Yes	No
M固定效应	Yes	No	Yes	No
R×M固定效应	No	Yes	No	Yes

Standard errors in parentheses

固定效应

财政教育支出与城乡收入差距

jia2014.do重复了例贾婷月、王笑涵和司继春（2024）中的结果

代码 2: 固定效应使用示例

```
// jia2024.do
clear
use datasets/Jia_Wang_Si2024.dta
// 取对数
gen lnPEE=log(PEE)
// 去平均
egen m_lnPEE=mean(lnPEE)
gen dm_lnPEE=lnPEE-m_lnPEE
// 交乘项
gen hukou_dm_lnPEE=hukou*dm_lnPEE
// 回归
local controls "gender nation married parents_edu"
reg lnIncome hukou_dm_lnPEE hukou lnPEE `controls'
outreg2 using jia2024.tex, replace addtext(出生年份固定效应, 否, 城市固定
```

固定效应

VARIABLES	(1) lnIncome	(2) lnIncome	(3) lnIncome	(4) lnIncome
hukou_dm_lnPEE	-0.233***	-0.288***	-0.256***	-0.230***
	(0.0371)	(0.0374)	(0.0378)	(0.0414)
hukou	0.889***	0.789***	0.442***	0.398***
	(0.0508)	(0.0513)	(0.0517)	(0.0548)
lnPEE	0.340***	0.507***	0.207***	0.390***
	(0.0216)	(0.0242)	(0.0682)	(0.107)
gender	1.055***	1.053***	0.998***	1.024***
	(0.0441)	(0.0437)	(0.0407)	(0.0424)
nation	0.405***	0.337***	0.130*	0.0804
	(0.0728)	(0.0723)	(0.0703)	(0.0705)
married	0.389***	0.154***	-0.288***	-0.261***
	(0.0469)	(0.0503)	(0.0540)	(0.0572)
parents_edu	0.773***	0.681***	0.571***	0.597***
	(0.150)	(0.149)	(0.140)	(0.145)
Constant	3.727***	2.043***	6.070***	4.011***
	(0.267)	(0.287)	(0.789)	(1.228)

控制变量

- 在以上的讨论中，我们发现只要条件外生性$\mathbb{E}\left(u|w,\tilde{x}\right)=\mathbb{E}\left(u|\tilde{x}\right)$满足，就可以保证$w$的回归系数的一致
- 这也意味着我们部分允许控制变量$\tilde{x}$与误差项的相关性：只要条件外生性满足即可。

控制变量的内生性

条件外生性

考虑如下DGP:

$$x = u + e, w = x + h$$

其中u, e, h相互独立，那么此时

$$\mathbb{E}(u|w, x) = \mathbb{E}(u|x, h) = \mathbb{E}(u|x)$$

从而条件外生性是满足的，虽然x与u是相关的。

控制变量的内生性

- 以上例子中，在回归方程中加入“内生”控制变量x达到了条件外生性
- 即使x与u相关，但是加入x保证了w系数的一致性，这是一种“好”的控制变量。
- 然而，以上论述不意味着控制变量$\tilde{x}$的内生性无关紧要，实际上，即使$\mathbb{E}\left(u|w\right)=0$从而$w$是一个外生随机的变量，也不能保证

$$\mathbb{E}\left(u|w,\tilde{x}\right)=\mathbb{E}\left(u|\tilde{x}\right)$$

控制变量的内生性

坏控制

- 考虑如下DGP：$w \perp u$，$x = u - w$，那么：

$$\mathbb{E}\left(u|w,x\right) = \mathbb{E}\left(u|w,u-w\right) = x + w$$

显然$\mathbb{E}\left(u|x\right)$应该为$x$的函数而不应该包含$w$，从而

$$\mathbb{E}\left(u|x\right) \neq x + w = \mathbb{E}\left(u|w,x\right)$$

从而条件外生性不满足。

控制变量的内生性

坏控制

- 此时，如果我们使用回归：

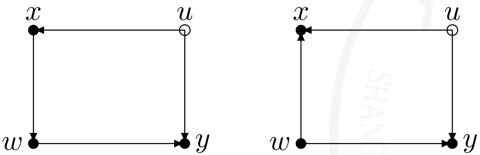
$$y = \alpha + \gamma \cdot w + \beta x + u$$

  那么自然有$\mathbb{L}(w|x) = -x$，从而$w - \mathbb{L}(w|x) = w + x = u$，一定与$u$是相关的。
- 此时不管y真实的数据生成过程是什么（从而γ的真值可以是任何值），都会得到$\text{plim}\hat{\gamma} = 1$，这显然是错误的结果。
- 此时，如果在回归方程中加入控制变量，反而会使得估计不再一致，这是一种“坏控制”（bad control）。

控制变量的内生性

- 比较以上两个例子，两者的区别在于：
  - 在第1个例子中，控制变量x是w的原因
  - 而在第2个例子中，x是w的一个结果。
- 我们可以使用Pearl（2009）的“因果图”（causal diagrams）来表达。
  - 因果图是一种表示因果影响的有向无环图（directed acyclic graphs），使用箭头表示因果的方向，实心点•代表可观测变量，空心点∘代表不可观测变量。

因果图



(a) (b)

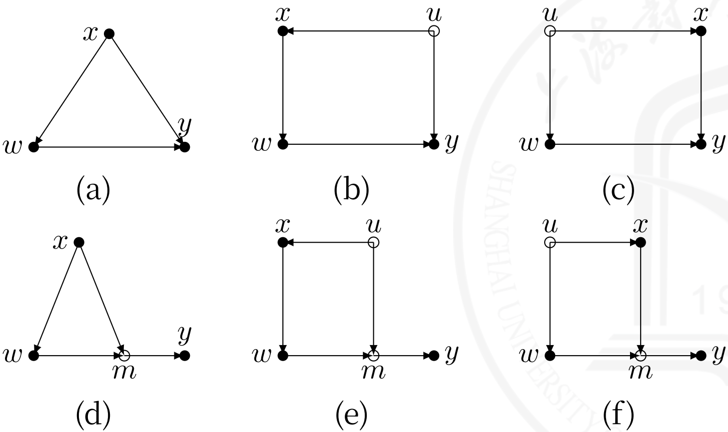
控制变量

- 一般而言，我们需要加入的控制变量应该是“事前”变量，即当可以影响w，但是不受w影响的“事后”变量。
  - 比如，当y是对数收入，w是教育年限，那么我们可以选取诸如性别、年龄、出生时的户口所在地作为控制变量，而不能选择当前的户口所在地作为控制变量
  - 因为性别、年龄、出生时的户口这些变量可能决定教育年限，然而当前的户口所在地可能是受教育年限的结果。
  - 而我们一旦控制了当前的户口所在地，相当于只是比较了工作地点在同一城市的不同个体之间的教育水平和收入
  - 而问题在于教育的回报可能刚好体现在工作地点上：大城市的工资水平更高，而较高的学历更容易留在大城市
  - 所以一旦只在相同的工作地点内部进行比较，就忽略了因为教育水平高从而可以留在大城市所导致的那部分收入的提高，从而低估了教育年限对于收入的影响。

控制变量

- 注意以上的“事前”、“事后”并非严格的时间意义上的前后，而是使用因果影响的方向定义的前后。
- 在因果图中，我们可以将指向核心解释变量的箭头称为“**后门**”（**back-door**）路径，从而好的控制应该是对后门路径的控制，但是不能控制核心解释变量指向的变量。
- 接下来我们介绍一些常见的情况。

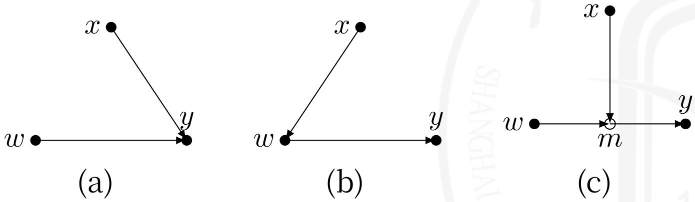
必须控制情形



必须控制情形

- (a)中的x会同时影响w和y，此时x作为混淆因素（confounder），即同时影响核心解释变量和被解释变量，必须作为控制变量放入到回归方程中。
  - 这也就是我们在“辛普森悖论”中提到的，为了获得“锻炼身体”对于“寿命”的影响，必须要控制性别，否则可能会得到错误的结果：如果不加入x作为控制变量，我们无法区分出来w的系数γ到底是由w带来的，还是x的影响。
- (b)中，作为后门路径，x也需要被控制，如果不控制x，那么误差项u就会与x相关，从而导致条件外生性不满足。
- (c)中的x并没有指向w，w, x都受到了u的影响，但是w并没有直接影响x，此时如果将w对x做投影，可以被预测的部分即共同的u的部分，从而不可预测的部分就排除了共同因素u的影响，从而给定$w \perp u|x$，也可以得到w系数的一致估计。
- (d)-(f)则分别代表了(a)-(c)的“中介”版本，即对于y的影响都是通过一个中间变量m实现的，此时原理与(a)-(c)相同

可控制可不控制



可控制可不控制

上图展示了一些控制变量不影响w系数的一致估计的情况。其中:

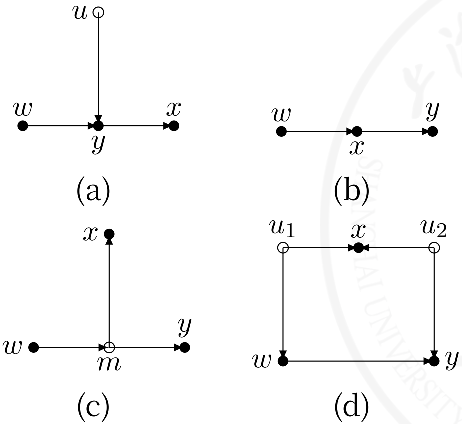
- (a)中的x与w不相关，从而控制与否都不影响w系数的一致性
- (b)中的x虽然与w相关，但是其本身不影响y，同样不影响w系数的一致性;
- (c)中x和w是中介变量m的共同原因，但是x与w不相关，这与(a)中的情况类似，也是可以不控制的。

可控制可不控制

虽然以上情况中的x从识别的角度来讲可控制可不控制，然而从推断的角度来讲，会影响w的系数的估计的精度

- (a)和(c)中如果控制了x，会使得y的残差变小，从而会增加w系数估计的精度（标准误降低）；
- (b)的情况，如果控制x，那么需要在w中排除x的影响，从而可用的变异减少，从而会降低w系数估计的精度（标准误变大）。

不能控制的情形



不能控制的情形

当x是w甚至y的结果时，x作为“坏”控制，控制x反而会导致给定x的情况下w与u相关，此时就不能控制了。

- (a)意味着x本身是结果变量y的一个结果，给定x会导致严重的偏误。

经济发展与身高

- Floud、Watcher和Gregory（1990）使用英国、美国19世纪部队的数据研究了儿童时期营养（w）与身高（y）的关系
- 他们发现随着人均GDP的增加，成年人的身高甚至在降低。
- 实际上，这很可能是由于坏控制的原因（Schneider，2020）。
  - 是否参与部队（x）本身可能是身高的一个结果，使用部队数据实际上控制了是否参与部队这一变量
  - 身高高的人本身有更多的就业机会，这会导致会参与部队的身高普遍比较矮，在经济增长时期更是如此，所以才会导致如此奇怪的结果。

不能控制的情形

- (b)展示的即x作为一个“中介变量”，同样也是w的结果，也是一个坏的控制。
- 图(c)与(b)类似，即存在一个不客观测的“中介”的情况。

不能控制的情形

出生体重悖论

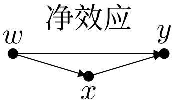
- Hernández-Díaz、Schisterman和Hernán（2006）讨论了“出生体重悖论”，即一般而言，吸烟者（w）生的婴儿死亡率（y）大于不吸烟者生的婴儿死亡率
- 然而，如果将样本限制在出生体重（x）比较低的样本，会发现以上的关系反转了。
- 为了解释这一现象，可以考虑图(b)的因果图：
  - 出生时的体重实际上是父母吸烟的一个结果
  - 而出生时体重以及婴儿死亡率肯定会受到同样的不可观测的因素u的影响。
  - 如果将样本限制在出生体重比较低的样本，实际上就是控制了x，从而得到了错误结果。
  - 可以想象，不吸烟的父母如果出生的小孩体重很轻，很可能是有其他的更加严重的问题（u），导致了更高的死亡率

不能控制的情形

- (d)展示了一种比较特殊的情况，虽然看起来x像是一个“事前”变量，然而此时控制x会导致w系数的不一致（M-bias）。
  - 比如，如果考虑
    $$x = u_1 + u_2, \quad w = h + u_1$$
    那么$\mathbb{L}(w|x) = x$，从而$w - \mathbb{L}(w|x) = h + u_1 - x = h - u_2$
  - 而u_2是y的误差项，从而$w - \mathbb{L}(w|x)$与y的误差项相关，导致条件外生性不满足。
- 该例子也可以看出，因果图只是作为一个思维导向作用，如果有可能，最好可以详细研究核心解释变量和被解释变量的数据生成过程。

中介效应与识别

有时x的控制与否也需要分情况讨论，是否加入x作为控制变量取决于模型所讲的故事，比如：



中介效应与识别

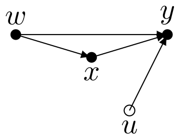
- 在以上情况中，x作为w的结果，同时影响y
- 此时x作为w影响y的中间渠道，如前所述，x是一个坏控制，是不能控制的。
- 然而：
  - 有的时候我们真正感兴趣的是w对于y的影响中排除了x这一渠道的“净效应”
  - 或者感兴趣的是x这一渠道的影响
- 而非包含所有渠道影响的“总效应”。
- 此时，干净的识别可能是一个非常困难的问题。

中介效应与识别

- 比如，当我们考虑教育的回报时，一般我们指的是教育的总回报，但是有时我们也关心羊皮纸效应，即有没有拿到学位证（x）对收入的影响。
- 再比如，如果我们需要研究私立高中对于收入的影响
  - 如果我们仅仅是对私立高中对于收入的“总效应”感兴趣，那么不需要控制学生的考试成绩；
  - 而如果我们对私立高中除了考试成绩之外的其他影响收入的渠道感兴趣，比如其他的兴趣爱好、交流能力的提升等，此时就需要排除高考成绩这一渠道了。

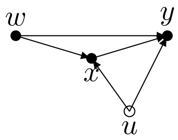
中介效应与识别

使用线性回归这一工具区分以上不同渠道的影响本身需要非常强的假定。在这里，我们不能仅仅依靠条件外生性假设了，而是需要w和x都要与误差项不相关，即如下图所示：



中介效应与识别

然而，实际中以上“干净”的外生性往往是不存在的，真实的情况往往是这样的：



此时，x与误差项相关，且x与w也相关，这会导致不仅x的系数不一致，而且w的系数也不一致。

中介效应与识别

- 在文献中，x可能会被当做一个“中介变量”进行一系列的检验
- 特别是在机制检验中，“中介效应”（mediating effect）经常被作为一种检验w对y影响的机制。
- 然而这些研究多应用在心理学等学科中，在经济学中其应用是有待商榷的。
- 江艇（2022）详细讨论了中介效应在经济学实证研究中所存在的问题，建议在实证研究中需要慎重使用这一类方法。