第 10 章　含定性变量的回归问题

李　杰

数据科学学院，浙江财经大学

2017 年 8 月 12 日

内容提要

10.1 自变量含定性变量的回归模型

☞ 定性数据

- 某交易日上证综指的涨或跌 $SH\_index\_up = 1$，$SH\_index\_up = 0$;
- 在职人员的性别，男或女 $female = 1$，$female = 0$;
- 是否拥有杭州市的户口 $HZ\_residence = 1$，$HZ\_residence = 0$;
- 改革前，还是改革后 $reform = 1$，$reform = 0$;
- 战争还是和平 $war = 1$，$war = 0$.

☞ 定性数据的量化

person	wage	educ	exper	female	married
1	3.10	11	2	1	0
2	3.24	12	22	1	1
3	3.00	11	2	0	0
4	6.00	8	44	0	1
5	5.30	12	7	0	1
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.
525	11.56	16	5	0	1
526	3.50	14	5	1	0

☞ 简单模型

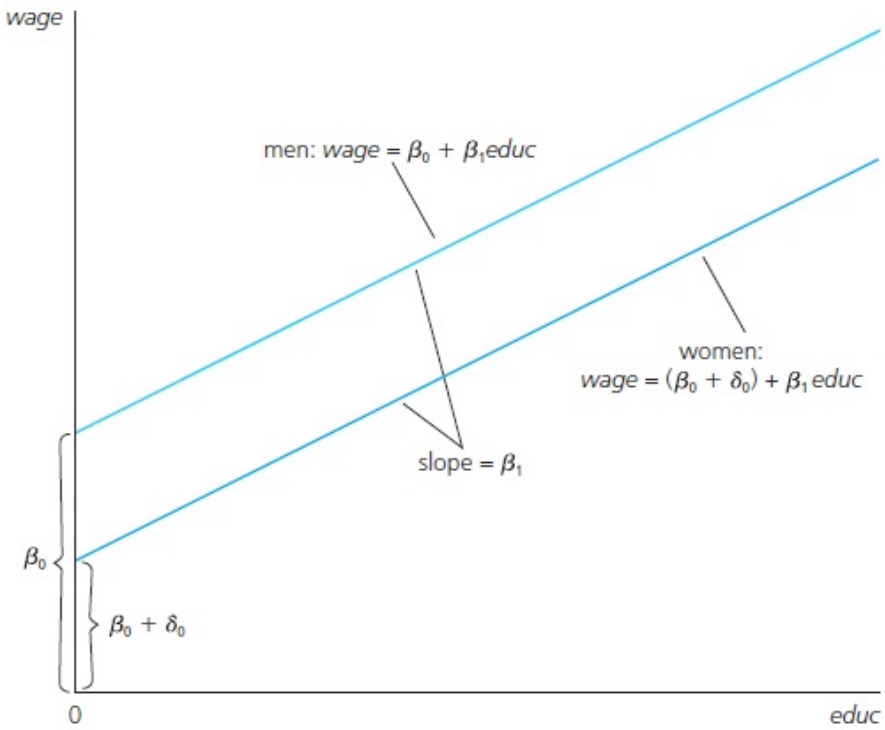
- 在教育 (*educ*) 对工资 (*wage*) 的回报中, 有没有性别 (*female*) 歧视?

$$wage = \beta_0 + \beta_1 educ + \beta_2 female + \varepsilon$$

其中 *female* 是虚拟变量 (dummy variable), $female = 1$ 表示女性, $female = 0$ 表示男性.

- 虚拟变量系数的涵义

$$\beta_2 = \mathrm{E}(wage|educ, female = 1) - \mathrm{E}(wage|educ, female = 0)$$



☞ 案例：时薪的性别差异

- **模型：** 利用数据 WAGE1.DTA 估计教育回报模型，得到

$$\widehat{wage} = \underset{(0.72)}{-1.57} \underset{(0.26)}{-1.81}\, female + \underset{(0.049)}{0.572}\, educ + \underset{(0.012)}{0.025}\, exper + \underset{(0.021)}{0.141}\, tenure \qquad (1)$$

$$n = 526, \quad R^2 = 0.364$$

- **对 *female* 系数的解释：** 在给定的相同 *educ*, *exper*, *tenure* 下，女性的平均时薪比男性的平均时薪低 1.81 美元。
- **估计各组均值以组间差异：** 做 *wage* 对虚拟变量 *female* 的回归，得到

$$\widehat{wage} = \underset{(0.21)}{7.10} \underset{(0.30)}{-2.51}\, female \qquad (2)$$

$$n = 526, \quad R^2 = 0.116$$

由式 (2) 可知，数据集中男性的平均时薪为 7.10 美元，女性的平均时薪为 $7.10 - 2.51 = 4.59$ 美元，女性平均时薪比男性低 2.51 美元。式 (1) 与式 (2) 中对性别歧视的度量的差异，主要是因式 (2) 中控制变量较少，估计的方差较大造成。比较说来，式 (1) 的结果更可靠。

案例代码

```
library(foreign)

mydata <- read.dta("E:/statafiles/WAGE1.DTA", convert.factors = FALSE)
wage_lm <- lm(wage~female+educ+exper+tenure,data=mydata)

group_lm <- lm(wage~female,data=mydata)
```

☞ 多个定性变量的情形

① 研究的个体是人的时候, 有很多属性, 譬如性别, 婚否, 种族, 是不是外国人等很多定性变量, 则可量化这些定性变量为

$$male = 1|0, \quad married = 1|0, \quad black = 1|0, \quad foreign = 1|0, \cdots$$

② 零售行业的数据具有季节性, 一年有四季, 每个观测数据必属于某一个季度, 则可设定

$$spring = 1|0, \quad summer = 1|0, \quad autumn = 1|0, \quad winter = 1|0$$

但在模型中引用这些虚拟变量时候, 只能引用其中的三个, 如果全部引用, 则将落入**虚拟变量陷阱** (dummy variables trap).

③ 一般情况, 当一个个体必属于 k 个类之一, 则模型中引入这些虚拟变量时只能引入 $k-1$ 个.

案例：工资方程 I

- 总体回归模型中加入新的虚拟变量，譬如表示是否已婚的变量 *married*, 模型变为

$$\log(wage) = \beta_0 + \delta_1 female + \delta_2 married + \beta_1 educ + \beta_2 exper + \beta_3 exper^2 + \beta_4 tenure + \beta_5 tenure^2 + \varepsilon \tag{3}$$

估计的 *married* 系数为 0.053 (0.041). 这表示“婚姻溢价”为 5.3%, 该变量并不是统计显著的，如此设定的模型存在重要局限：婚姻溢价对男女都一样。这与现实相悖。

- 模型的改进：引入 *female* 和 *married* 两个虚拟变量后，样本中的个体可分为“已婚男性”、“已婚女性”、“单身男性”、“单身女性”，选定一个基组，譬如单身男性，为其他每个组设定一个虚拟变量，譬如 *marrmale*, *marrfem*, *singfem*, 估计得模型为

$$\widehat{\log(wage)} = \underset{(0.100)}{0.321} + \underset{(0.055)}{0.213} marrmale - \underset{(0.058)}{0.198} marrfem - \underset{(0.056)}{0.110} singfem + \underset{(0.007)}{0.079} educ + \underset{(0.005)}{0.027} exper - \underset{(0.00011)}{0.00054} exper^2 + \underset{(0.007)}{0.029} tenure - \underset{(0.00023)}{0.00053} tenure^2 \tag{4}$$

$n = 526, \quad R^2 = 0.461$

案例：工资方程 II

☞ **注解：**

① 如果回归模型中有 g 组，则需选定一个基组且设定 $g-1$ 个虚拟变量。如果设置 g 个虚拟变量，则将导致虚拟变量陷阱。模型中的截距是基组的截距，其他组的截距为基组截距加上相应的虚拟变量系数。如果在模型中设置 g 个虚拟变量，则在模型中不设置截距，可以避免虚拟变量陷阱，但这不是 建模的好方法。上例中有 4 个组，所以需要设置 3 个虚拟变量。

② 解释虚拟变量系数时，注意该系数表示相应的组与基组的差别。譬如，*marrmale* 前的系数表示在其它因素不变的情况下，已婚男性比单身男性多挣 21.3%.

③ 非基组之间的差异可以比较，但不方便检验。譬如，由模型可知，在其他因素不变的情况下， 单身女性比已婚女性多挣 $100\times(-0.110+0.198)\%=8.8\%$. 解决办法：选择单身女性或已婚女性为基组，重做回 归，譬如选择已婚女性为基组后，估计得到的模型为

$$\widehat{\log(wage)} = \underset{(0.106)}{0.123} + \underset{(0.056)}{0.411}marrmale + \underset{(0.058)}{0.198}singmale + \underset{(0.052)}{0.088}singfem + \cdots \qquad (5)$$

检验的 $t=0.088/0.052\approx 1.69$，该值不够大，要在较高的显著性水平才能拒绝原假设。

案例代码

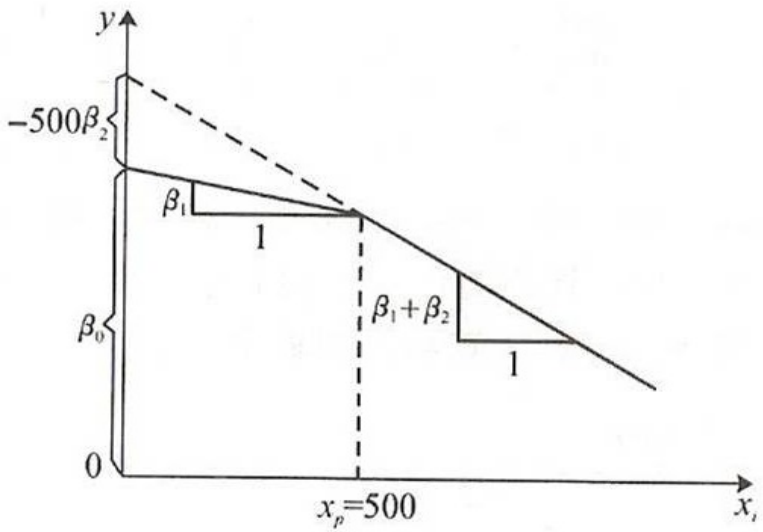
```
mydata <- read.dta("E:/statafiles/WAGE1.DTA", convert.factors = FALSE)

mydata$marrmale <- (!mydata$female)*mydata$married # 生成新的组: 已婚男性
mydata$marrfem <- mydata$female * mydata$married   # 生成新的组: 已婚女性
mydata$singfem <- mydata$female * (!mydata$married)# 生成新的组: 单身女性

wage_lm <- lm(log(wage)~marrmale+marrfem+singfem+educ+exper+I(exper^2)
                    +tenure+I(tenure^2),data=mydata)
```

10.2 自变量含定性变量的回归模型的应用

☞ 分段回归

- **问题:** 在有些实际问题中, 某自变量在不同范围内的波动对因变量的影响不同. 曲线拟合也不能达到理想效果.
- **模型**

$$y = \beta_0 + \beta_1 x + \beta_2 (x - 500) D + \varepsilon$$



- **估计:** 最小二乘法
- **检验:** 检验系数 β_2 估计值的显著性.

☞ 回归系数相等的检验

- **问题:** 员工的工资方程中存在性别歧视, 此外, 教育的回报中是否男女有别呢? 即偏导数 $\dfrac{\partial wage}{\partial educ}$ 是常数还是关于性别的函数呢?

- **模型**

$$wage = \beta_0 + \beta_1 educ + \beta_2 female + \beta_3 female * educ + \varepsilon$$

$$wage = \beta_0 + \beta_1 educ + \beta_2 female + \beta_3 female * (educ - \overline{educ}) + \varepsilon$$

- **估计:** 最小二乘估计
- **检验:** 检验系数 β_3 估计值的显著性.

10.3 因变量是定性变量的回归模型

☞ **问题背景**

- 上证综指在某个交易日的涨跌情况;
- 某个成年劳动力是否曾接受高等教育;
- 企业的并购行为是否成功;
- 政府的某项政策是否成功;
- 大学毕业生是否成功找到工作;
- ……

在这类问题中，因变量 $y = 1$ 表示某个结果发生，而 $y = 0$ 表示对立的结果发生

☞ 线性概率模型

- **0-1 分布随机变量的数学期望：** 设随机变量 y 服从 0-1 分布，且 $\mathrm{P}\{y=1\}=p$, $\mathrm{P}\{y=0\}=1-p$. 则 $\mathrm{E}(y)=p=\mathrm{P}\{y=1\}$.
- **模型：** 设多元回归模型为

$$y=\beta_0+\beta_1 x_1+\cdots+\beta_k x_k+\varepsilon \tag{6}$$

其中因变量 y 仅取 0，1 两个值。假定多元回归模型假设 MRL.1∼MRL.4 都成立，则

$$\mathrm{E}(y|\mathbf{x})=\beta_0+\beta_1 x_1+\cdots+\beta_k x_k=\mathrm{P}\{y=1|\mathbf{x}\} \tag{7}$$

$\mathrm{P}\{y=1|\mathbf{x}\}$ 称为**响应概率** (response probability)，式 (7) 表明响应概率是以 β_j $(j=0,1,\cdots,k)$ 为参数，以 x_j $(j=1,\cdots,k)$ 为变量的线性函数，故二元因变量多元线性回归模型又称为**线性概率模型** (linear probability model, LPM).

- **注解：**
  ① 模型 (6) 中的参数 β_j **不应理解**为“控制其他因素不变，x_j 变化一单位导致 y 变化 β_j”，而应理解为 x_j 变化导致概率 $\mathrm{P}\{y=1|\mathbf{x}\}$ 的变化，即

$$\Delta\mathrm{P}\{y=1|\mathbf{x}\}=\beta_j\Delta x_j \tag{8}$$

  ② 模型 (6) 的拟合值 $\hat{y}=\hat{\beta}_0+\hat{\beta}_1 x_1+\cdots+\hat{\beta}_k x_k$ 是对概率 $\mathrm{P}\{y=1|\mathbf{x}\}$ 的估计。

☞ 案例：已婚女士是否进入劳动力市场

- **估计的模型：**

$$\widehat{inlf} = \underset{(0.154)}{0.586} - \underset{(0.0014)}{0.0034}nwifeinc + \underset{(0.007)}{0.038}educ + \underset{(0.006)}{0.039}exper - \underset{(0.00018)}{0.00060}exper^2$$
$$- \underset{(0.002)}{0.016}age - \underset{(0.034)}{0.262}kidslt6 + \underset{(0.0132)}{0.0130}kidsge6$$
$$n = 753, \quad R^2 = 0.264$$

- **变量含义：** 二值因变量 $inlf = 1$ 表示某女士为获取工资而参与家庭外工作，$nwifeinc$ 表示其丈夫的收入，$educ$ 表示所受的教育程度，$exper$ 表示工作经验，age 表示年龄，$kidslt6$ 表示年龄不足 6 岁的孩子数量，$kidsge6$ 表示年龄介于 $6-18$ 岁之间孩子的数量。

- **对参数的解释：**

  ① 变量 $educ$ 前面的系数为 0.038，表示控制其他因素不变，已婚女士的所受教育增加一年，则进入劳动力市场的概率增加 0.038.

  ② 变量 $kidslt6$ 前面的系数为 -0.262，表示控制其他因素不变，已婚女士 6 岁以下孩子每增加一个，则进入劳动力市场的概率下降 0.262.

- **代码：**

```
mydata <- read.dta("E:/statafiles/MROZ.DTA", convert.factors = FALSE)
jobmkt_lm <- lm(inlf~nwifeinc+educ+exper+I(exper^2)+age+kidslt6+kidsge6,
                data=mydata)
```

☞ 线性概率模型存在的问题

① 概率值的取值范围为 $[0,1]$，但线性概率模型的**拟合值可能越界**；

② **拟合值不可能与所有自变量都呈现线性关系**。譬如，已婚女士年龄小于 6 岁的孩子从 0 增加到 1，进入劳动力市场的概率下降 0.262，但该女士年幼的孩子从 1 增加到 2 时，进入劳动力市场的概率下降却不止 0.262。

③ 由于 y 的二值性，**线性概率模型必定会违背同方差假设**，因为

$$\mathrm{Var}(y|\mathbf{x}) = \mathrm{P}(y|\mathbf{x}) \cdot [1 - \mathrm{P}(y|\mathbf{x})] \tag{9}$$

而同方差性即便在大样本情形对 t 统计量和 F 统计量的有效性至关重要。

10.4 Logistic 回归模型

☞ 响应概率函数

- 令响应概率函数为

$$P(y=1|x_1,\cdots,x_k)=G(\beta_0+\beta_1x_1+\cdots+\beta_kx_k)=G(\mathbf{X}\beta)$$

其中 $0\leq G(z)\leq 1,\ \forall z\in R.$

☞ Logistic 模型

- 如果 $G(z)$ 是 logistic 函数，即

$$G(z)=\frac{exp(z)}{1+exp(z)}$$

则称该二值响应模型为**Logistic 模型**.

☞ Probit 模型

- 如果 $G(z)$ 是标准正态分布函数，即 $G(z)=\Phi(z)=\int_{-\infty}^{z}\phi(v)dv$，其中 $\phi(z)=\frac{1}{\sqrt{2\pi}}e^{-\frac{z^2}{2}}$,则称该二值响应模型为**Probit 模型**.

☞ 解释变量的偏效应

- **连续型解释变量的偏效应：** 设 x_j 为连续型变量，则 x_j 的偏效应为

$$\frac{\partial \mathrm{P}(y=1|X)}{\partial x_j} = \frac{\partial G(X\beta)}{\partial x_j} = g(X\beta)\beta_j$$

  其中 $g(X\beta) = \frac{\partial G(X\beta)}{\partial (X\beta)}$.
- **二值解释变量的偏效应：** 设 x_1 为二值变量，则在其他变量保持不变的情况下，x_1 从 0 到 1 的偏效应是

$$G(\beta_0 + \beta_1 + \beta_2 x_2 + \cdots + \beta_k x_k) - G(\beta_0 + \beta_2 x_2 + \cdots + \beta_k x_k)$$

- **离散型解释变量的偏效应：** 设 x_k 为离散型变量，则 x_k 取值从 c_k 变化到 $c_k + 1$ 时的偏效应为

$$G(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k (c_k + 1)) - G(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k c_k)$$

☞ Logistic 和 Probit 模型的极大似然估计

① 求出 y_i 的分布函数

$$\mathrm{F}(y_i|X_i) = [G(X_i\beta)]^{y_i}[1 - G(X_i\beta)]^{1-y_i}, \quad i = 1, \cdots, n$$

② 计算 $y_1, \cdots, y_n$ 的似然函数

$$L(\beta; X_1, \cdots, X_n) = \prod_{i=1}^{n} \mathrm{F}(y_i|X_i) = \prod_{i=1}^{n} \left\{ [G(X_i\beta)]^{y_i}[1 - G(X_i\beta)]^{1-y_i} \right\}$$

③ 计算对数似然函数

$$\mathscr{L}(\beta; X_1, \cdots, X_n) = \sum_{i=1}^{n} \{y_i \log[G(X_i\beta)] + (1 - y_i) \log[1 - G(X_i\beta)]\}$$

④ 求对数似然函数关于每一个参数的偏导数，并令其等于 0，得到对数似然方程组；

⑤ 求解对数似然方程组，得到模型中参数的对数似然估计量。

☞ 注解：

- 由于似然方程组是非线性方程组，所以一般写不出对数单位或者概率单位模型极大似然估计量的表达式；

- 理论证明，极大似然估计量是一致的，渐近正态的，以及渐近有效的估计量.

☞ 似然比检验：

- 适用场合：约束模型和无约束模型都比较容易估计，则似然比检验适用。
- 检验思想：极大似然估计是最大化了对数似然函数，所以去掉一个或一些变量之后一般会导致一个较小的对数似然值，对数似然值的下降程度是否大到足以断定去掉的变量是否是重要的。
- **似然比统计量** (likelihood ratio statistic):

$$LR = 2(\mathscr{L}_{ur} - \mathscr{L}_r) \sim \chi^2(q)$$

其中 q 是受约束条件的个数，$\mathscr{L}_{ur}$ 是无约束模型的对数似然值，$\mathscr{L}_r$ 是约束模型的对数似然值，且 $\mathscr{L}_{ur} \geq \mathscr{L}_r$.

☞ 概率单位与对数单位模型的拟合优度

- **正确预测百分比** (percent correctly predicted)
  - 选定临界值 0.5，当 $G(X_i\hat{\beta}) \geq 0.5$ 时，$\tilde{y}_i = 1$, 否则 $\tilde{y}_i = 0$. 如果预测 $\tilde{y}_i = 1$ 时观测值 $y_i = 1$，预测 $\tilde{y}_i = 0$ 时观测值 $y_i = 0$，则称预测正确，否则预测错误。正确预测百分比就是 $\tilde{y}_i = y_i$ 占总观测数的比例；
  - 使用样本中成功的比例作为临界值，计算正确预测百分比；
  - 搜索临界值 τ, 使得在该临界值下正确预测百分比等于样本成功百分比。
- **伪 R^2 度量** (pseudo R-squared)
  - 伪 R^2 由 McFadden 提出，

$$PR^2 = 1 - \frac{\mathscr{L}_{ur}}{\mathscr{L}_0}$$

其中 $\mathscr{L}_{ur}$ 是被估模型的对数似然值，而 $\mathscr{L}_0$ 是只有截距项的模型的对数似然值。一般情况下 $\mathscr{L}_0 \leq \mathscr{L}_{ur} < 0$, 所以 $0 < \frac{\mathscr{L}_{ur}}{\mathscr{L}_0} \leq 1$, 只有当解释变量完全没有解释能力时，$\mathscr{L}_0 = \mathscr{L}_{ur}$，此时 $PR^2 = 0$, 解释变量解释能力越强，$\frac{\mathscr{L}_{ur}}{\mathscr{L}_0}$ 越小，而 PR^2 越接近 1.

☞ 解释概率单位与对数单位模型的估计值

- **当 x_j 为连续型解释变量时**，它的偏效应为 $g(X\beta)\beta_j$，相应的估计值为 $g(X\hat{\beta})\hat{\beta}_j$，可以看到偏效应的值取决于 X。两种调整方法
  - 方法一：将偏效应调整为 $g(\bar{X}\hat{\beta})\hat{\beta}_j$ (这样的处理也存在问题);
  - 方法二：计算**平均偏效应** (average partial effect)，将偏效应调整为

$$\left[\frac{1}{n}\sum_{i=1}^{n} g(X_i\hat{\beta})\right]\hat{\beta}_j$$

- **当 x_k 为离散型解释变量时**，它的偏效应为

$$G(\hat{\beta}_0+\hat{\beta}_1x_1+\cdots+\hat{\beta}_{k-1}x_{k-1}+\hat{\beta}_k(c_k+1))-G(\hat{\beta}_0+\hat{\beta}_1x_1+\cdots++\hat{\beta}_{k-1}x_{k-1}\hat{\beta}_kc_k)$$

它的值也取决于 X。两种调整方法

- ✧ 方法一：将偏效应调整为

$$G(\hat{\beta}_0+\hat{\beta}_1\bar{x}_1+\cdots+\hat{\beta}_{k-1}\bar{x}_{k-1}+\hat{\beta}_k(c_k+1))-G(\hat{\beta}_0+\hat{\beta}_1\bar{x}_1+\cdots+\hat{\beta}_{k-1}\bar{x}_{k-1}+\hat{\beta}_kc_k)$$

- ✧ 方法二：计算平均偏效应，将偏效应调整为

$$\frac{1}{n}\sum_{i=1}^{n}\Big\{G(\hat{\beta}_0+\hat{\beta}_1x_{i1}+\cdots+\hat{\beta}_{k-1}x_{ik-1}+\hat{\beta}_k(c_k+1)) - G(\hat{\beta}_0+\hat{\beta}_1x_{i1}+\cdots++\hat{\beta}_{k-1}x_{ik-1}\hat{\beta}_kc_k)\Big\}$$

☞ 案例 已婚妇女的劳动市场参与

```
library(foreign)
library(nlme)
mroz <- read.dta("E:/statafiles/MROZ.DTA", convert.factors = FALSE)
mroz_lpm <- lm(inlf~nwifeinc+educ+exper+I(exper^2)+age+kidslt6
                                                  +kidsge6,data=mroz)
summary(mroz_lpm)
mroz_log <- glm(inlf~nwifeinc+educ+exper+I(exper^2)+age+kidslt6
                    +kidsge6,data=mroz,family=binomial(link="logit"))
summary(mroz_log)
mroz_log0 = update(mroz_log,formula=.~1)
1-as.vector(logLik(mroz_log)/logLik(mroz_log0))  # 计算伪 R^2
mroz_prob <- glm(inlf~nwifeinc+educ+exper+I(exper^2)+age+kidslt6+
                    kidsge6,data=mroz,family=binomial(link="probit"))
summary(mroz_prob)
mroz_prob0 = update(mroz_prob,formula=.~1)
1-as.vector(logLik(mroz_prob)/logLik(mroz_prob0))  # 计算伪 R^2
```