应用统计II 第7讲 逻辑回归 (Logistic Regression)

Instructor: 郝壮

haozhuang@buaa.edu.cn
School of Economics and Management
Beihang University

May 29, 2022

逻辑回归 (Logistic Regression)

- 参考教材： Cameron, A. C., & Trivedi, P. K. (2005). Microeconometrics: methods and applications. Cambridge University Press.

逻辑回归 (Logistic Regression)

- Logistic 回归也叫 Logit model, 是判别分析(Discriminant Analysis)的一类, 也是广义线性模型, 是最简单的离散选择模型(discrete choice models)之一.

Logistic回归模型

- 工作目的：因变量为虚拟变量的回归模型
- 与判别分析的区别：可以给出样本点属于某一类的**概率**
- 例如：降雨概率, 违约概率, ...
- 如果因变量是2类: simple logit model (本节课的主要内容)
- 如果因变量是多类: multinomial logit.

Logistic回归的建模问题

- 工作目的：因变量为**虚拟变量**的回归模型

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \varepsilon$$

- 普通OLS回归：对回归模型中的自变量, 回归系数以及残差项的取值都没有任何限制, 作为自变量函数的因变量就必须能够在 $(-\infty, +\infty)$范围内自由取值.
- 如果因变量只取分类值, 或者只取两类值$(0, 1)$, 就会违反因变量为连续型变量的假设 (注意在经典高斯马尔可夫假设下, y应为正态分布).

回顾：一元线性回归模型

基本假设 (Basic Assumptions):

模型: $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \quad i = 1, 2, \cdots, n$

随机误差: ε_i

(1) ε_i 是一个独立的正态随机变量

(2) $\varepsilon_i \sim N\left(0, \sigma^2\right)$ $\left\{ \begin{array}{c} E\left(\varepsilon_i\right) = 0 \\ D\left(\varepsilon_i\right) = \sigma^2 \end{array} \right\}$ i.i.d.

(3) x_i 是非随机变量, 没有测量误差(固定设计)

结论: 如果 $\left(x_i, Y_i\right)$ 满足上述模型, 则 $Y_i (i = 1, 2, \ldots, n)$ 相互独立, 都服从正态分布, 并且:

$$E\left(y_i \mid x_i\right) = \beta_0 + \beta_1 x_i$$

因变量是虚拟变量时, 线性回归的含义

设因变量只取$0, 1$两个数值的虚拟变量, 是一个两点分布变量. 在给定的条件下, 记概率为:

$$\begin{aligned}
&\mathrm{P}\left(y_i = 1 \mid x_i\right) = p_i \\
&\mathrm{P}\left(y_i = 0 \mid x_i\right) = 1 - p_i = q_i \\
&\mathrm{E}\left(y_i \mid x_i\right) = 1 \times p_i + 0 \times (1 - p_i) = p_i
\end{aligned}$$

线性回归:

$$E(y_i|x_i) = \beta_0 + \beta_1 x_i$$

这时, 所建立的线性回归模型就是在分析当自变量变化时, 结果概率p_i是如何变化的.

但是由于 p_i的取值只能在 [0, 1] 范围内. 因此, 普通线性回归模型在预测概率大于1或小于0时存在解释困难 (可能是不适用的).

Logit模型构造过程

- **变换：** $p_i, \quad q_i \in [0,1] \quad p_i + q_i = 1$. 记:

$$\frac{p_i}{q_i} \in [0, +\infty),$$

则

$$z_i = \ln \frac{p_i}{q_i} = \ln \frac{p_i}{1 - p_i} \in (-\infty, +\infty).$$

可以构建模型

$$z_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

Logit模型构造过程

- **反变换：** 根据:

$$z_i = \ln \frac{p_i}{q_i} \Rightarrow p_i = q_i e^{z_i}$$

由于:

$$q_i e^i + q_i = 1$$

所以有:

$$q_i = \frac{1}{1+e^{z_i}}, \quad p_i = \frac{e^{z_i}}{1+e^{z_i}}$$

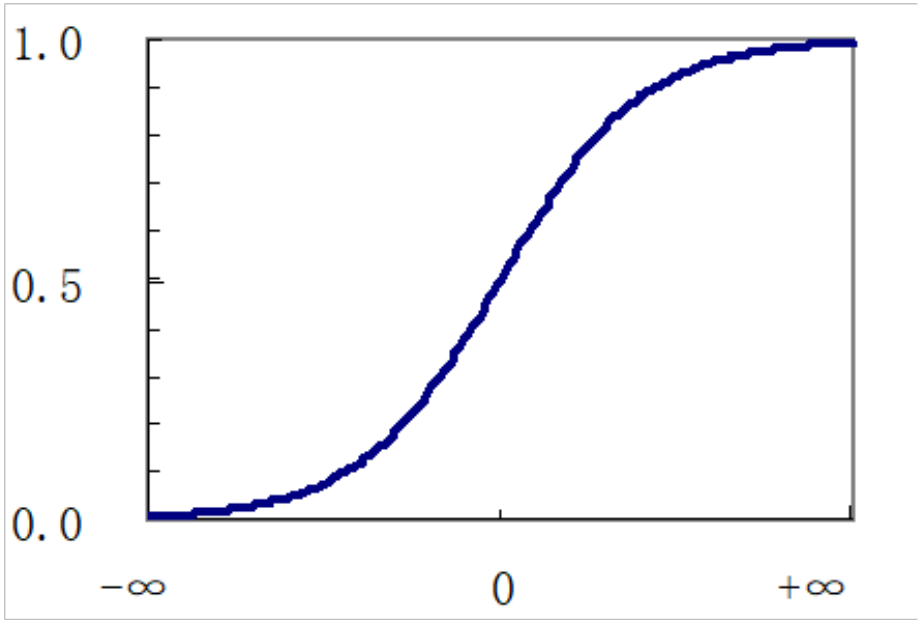
p_i可表示为(logit model)

$$p_i = \frac{\exp(\beta_0 + \beta_1 x_i)}{1 + \exp(\beta_0 + \beta_1 x_i)}$$

Logistic函数(S型曲线)

logistic函数

$$\Lambda(x) \equiv f(x) = \frac{e^x}{1 + e^x}$$



值域为$[0, 1]$, 定义域为： $x \in (-\infty, +\infty)$. (描述概率和自变量之间关系的曲线).

Logistic回归模型

Logit变换：

- 定义对数发生比(log odds)：
  $\text{Logit}\,(p_i) = \ln \frac{p_i}{1-p_i} \to (-\infty, \infty)$
- $\text{Logit}\,(p_i)$ 也叫“logit链接函数", 作用是完成了$(0,1)$到$(-\infty, \infty)$的映射.

设：

$$\text{Logit}\,(p_i) = \beta_0 + \beta_1 x_i,$$

所以logistic回归模型 (或称为logit模型) 可以表达为

$$E(y_i|x_i) \equiv p_i = \frac{\exp\,(\beta_0 + \beta_1 x_i)}{1 + \exp\,(\beta_0 + \beta_1 x_i)}$$

也可表达为

$$Logit(p_i) = \ln(p_i/(1-p_i)) = \beta_0 + \beta_1 x_i$$

Logistic回归模型

若有极大似然估计:

$$\hat{\beta}_0 \to \beta_0, \quad \hat{\beta}_1 \to \beta_1$$

则

$$\hat{p}_i = \frac{\exp\left(\hat{\beta}_0 + \hat{\beta}_1 x_i\right)}{1 + \exp\left(\hat{\beta}_0 + \hat{\beta}_1 x_i\right)} \in [0, 1]$$

- 克服了概率因变量的取值受到限制的困难.

$Logit(p)$对数发生比

发生比/优势比 (Odds)： 成功/发生的概率除以失败/未发生的概率

$$\text{Odds} = p/(1-p)$$

对数发生比： $\ln(p/(1-p))$

$$Logit(p) = \ln(p/(1-p)) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

- 发生比是非负数, 对数发生比是无界的.
- $\text{Logit}\,(p)$ 与 p 之间具有同向变化的规律.
- Logistic回归模型$\text{Logit}\,(p_i) = \beta_0 + \beta_1 x_i$ 解释了 “对数发生比" 与自变量集合之间的线性回归关系.

拓展：发生比 (Odds) vs 比值比/OR值 (odds ratio)

什么是发生比(Odds)? 发生概率比不发生概率.

$$\text{Odds} = p/(1-p)$$

什么是OR值/比值比 (Odds ratio)? 两个发生比的比值.

$$\text{Odds ratio} = \frac{Odds_1}{Odds_2}$$

拓展：发生比 (Odds) vs 比值比/OR值 (odds ratio)

- 比如成功的概率$p = 0.8$, 那么失败的概率$q = 1 - p = 0.2$. 所以, 成功的发生比为：

$$odds(success) = p/q = 0.8/0.2 = 4$$

成功概率是失败的4倍.

- 如果做作业成功的发生比(考试通过)是4, 不做作业成功的发生比是3, 那么做作业的OR值/比值比(Odds ratio)是

$$OR = 4/3 > 1,$$

代表做作业增加了成功率.

拓展：发生比 (Odds) vs 比值比/OR值 (odds ratio)

OR值优势: logistic regression中变量OR值为常量. 而边际概率影响随x不同会变化.

OR值劣势: 定量的直观意义较难解释, 不如边际概率影响(dy/dx)直观.

拓展： Alternative：概率比值(relative risk)或平均边际影响(average marginal effect)

拓展：发生比(Odds) vs 比值比/OR值(odds ratio)

注意：有的教材将发生比(Odds)说成是OR值(odds ratio)但实际表示的是$p/(1-p)$, 造成很大误解, 如Cameron and Trivedi 2005 就是如此定义(p470). 具体讨论见:
`https://www.stata.com/support/faqs/statistics/odds-ratio-versus-odds/`

Stata command logistic 默认报告 OR值/比值比 (Odds ratio). Stata报告的Odds ratio是传统意义上的(正确的)定义方式, 即两个Odds的比值 (永远为正, 和1相比).

Stata command logit 默认报告系数. (可正可负)

Logistic 回归模型中系数如何解释?

$$Logit(p) = \ln(p/(1-p)) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

- x_2 不变的情况下, x_1 每增加一个单位, 对数发生比增加 β_1 个单位.
- x_2 不变的情况下, x_1 每增加一个单位, 发生比(odds)增加 e^{β_1} 个单位.
- x_2 不变的情况下, x_1 每增加一个单位, 比值比(odds ratio)增加 $e^{\beta_1} - 1$ 个单位.
- 也可根据β_1点估计求出边际概率的变化加以解释(dy/dx)

极大似然估计

要估计以下logit模型

$$p_i = P(y_i = 1|x_i) = \frac{\exp(\beta_0 + \beta_1 x_i)}{1 + \exp(\beta_0 + \beta_1 x_i)}$$

因为是非线性模型, OLS不适用. 但由于我们有y_i的条件分布的表达, 所以可以立即采用极大似然估计 (MLE)

MLE步骤:

- (1) 求: 似然函数 $L(\beta_0, \beta_1)$
- (2) 求: 对数似然函数 $\ln[L(\beta_0, \beta_1)]$
- (3) 关于 $\ln[L(\beta_0, \beta_1)]$, 分别对 β_0, β_1 求偏导, 然后令之为0
- (4) 可求得 β_0, β_1 的极大似然估计: $\hat{\beta}_0, \hat{\beta}_1$

极大似然估计

$y_i \sim B(1, p_i)$, 得到一个观测值 y_i 的概率为:

$$P(y_i) = p_i^{y_i} (1 - p_i)^{1-y_i}$$

由于y_i相互独立, 所以似然函数可写为

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{n} p_i^{y_i} (1 - p_i)^{(1-y_i)} = \prod_{i=1}^{n} \left(\frac{p_i}{1 - p_i} \right)^{y_i} (1 - p_i)$$

对数似然函数 (将p_i和$ln(p_i/(1-p_i)$替换成x_i的函数):

$$\begin{aligned} \ln[L(\boldsymbol{\beta})] &= \sum_{i=1}^{n} \left[y_i (\beta_0 + \beta_1 x_i) + \ln \left(1 - \frac{\exp(\beta_0 + \beta_1 x_i)}{1 + \exp(\beta_0 + \beta_1 x_i)} \right) \right] \\ &= \sum_{i=1}^{n} \left[y_i (\beta_0 + \beta_1 x_i) - ln(1 + \exp(\beta_0 + \beta_1 x_i)) \right] \end{aligned}$$

极大似然估计

分别对 β_0, β_1 求导, 令之为0, 得到极大似然估计 b_0, b_1(非线性方程组, 迭代求解). FOCs:

$$\frac{\partial \ln\left[L\left(\beta_0\right)\right]}{\partial \beta_0} = \sum_{i=1}^{n}\left(y_i - \frac{\exp\left(\beta_0 + \beta_1 x_i\right)}{1+\exp\left(\beta_0 + \beta_1 x_i\right)}\right) = 0$$

$$\frac{\partial \ln\left[L\left(\beta_1\right)\right]}{\partial \beta_1} = \sum_{i=1}^{n} x_i\left(y_i - \frac{\exp\left(\beta_0 + \beta_1 x_i\right)}{1+\exp\left(\beta_0 + \beta_1 x_i\right)}\right) = 0$$

$\hat{\beta}_{MLE}$一般没有解析解, 但可以很方便求得数量解, 并可以证明logit (和probit)模型的似然函数是全局凸函数(利用Hassian矩阵正定可证). 所以:

$$\hat{p}_i = \frac{\exp\left(\hat{\beta}_0 + \hat{\beta}_1 x_i\right)}{1+\exp\left(\hat{\beta}_0 + \hat{\beta}_1 x_i\right)}$$

拓展知识：$\hat{\beta}$的方差和标准误

拓展知识： 由最大似然估计性质, 可推出

$$\widehat{\boldsymbol{\beta}}_{\mathrm{ML}} \overset{a}{\sim} \mathcal{N}\left[\boldsymbol{\beta}, \left(-\mathrm{E}\left[\partial^2 lnL/\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}'\right]\right)^{-1}\right]$$

可推出

$$\widehat{\mathrm{Var}}\left[\widehat{\boldsymbol{\beta}}_{\mathrm{ML}}\right] = \left(\sum_{i=1}^{N} \frac{1}{\Lambda\left(\mathbf{x}_i'\widehat{\boldsymbol{\beta}}\right)\left(1-\Lambda\left(\mathbf{x}_i'\widehat{\boldsymbol{\beta}}\right)\right)}\Lambda'\left(\mathbf{x}_i'\widehat{\boldsymbol{\beta}}\right)^2 \mathbf{x}_i\mathbf{x}_i'\right)^{-1}$$

其中

$$\Lambda\left(\mathbf{x}'\boldsymbol{\beta}\right) = \frac{e^{\mathbf{x}'\beta}}{1+e^{\mathbf{x}'\beta}}$$

有关$\widehat{\mathrm{Var}}\left[\widehat{\boldsymbol{\beta}}_{\mathrm{ML}}\right]$和$SE(\hat{\beta})$的推导, 感兴趣的同学可以阅读:

A. Colin Cameron Pravin K. Trivedi (2005) Chapter 14 of Microeconometrics Methods and Applications. Cambridge University Press

Stata实现logistic regression

根据任务不同, 选择不同的命令

- logistic 命令：报告发生比 (odds ratio) (生物统计、公共卫生学中常用)
- logit 命令: 报告系数 (coefficient)
  - post estimation: margins 命令：报告边际效应 (经济学中常用)

Stata实现logistic regression

数据来源：低出生重量low birth weight风险因素 (Hosmer et al., 2013)

```
. use https://www.stata-press.com/data/r16/lbw, clear
.  describe
  obs:           189                          Hosmer & Lemeshow data
 vars:            11                          15 Jan 2018 05:01
-------------------------------------------------------------------------------
              storage   display    value
variable name   type    format     label      variable label
-------------------------------------------------------------------------------
id                int     %8.0g                 identification code
low               byte    %8.0g                 birthweight<2500g
age               byte    %8.0g                 age of mother
lwt               int     %8.0g                 weight at last menstrual period
race              byte    %8.0g      race       race
smoke             byte    %9.0g      smoke      smoked during pregnancy
ptl               byte    %8.0g                 premature labor history (count)
ht                byte    %8.0g                 has history of hypertension
ui                byte    %8.0g                 presence, uterine irritability
ftv               byte    %8.0g                 number of visits to physician during 1st trimester
bwt               int     %8.0g                 birthweight (grams)
-------------------------------------------------------------------------------
```

Stata实现logistic regression

```
// 建立logit模型, 输出系数点估计
. logit low age lwt i.race smoke ptl ht ui

Iteration 0:   log likelihood =   -117.336
Iteration 1:   log likelihood = -101.28644
Iteration 2:   log likelihood = -100.72617
Iteration 3:   log likelihood =   -100.724
Iteration 4:   log likelihood =   -100.724

Logistic regression                               Number of obs     =        189
                                                  LR chi2(8)        =      33.22
                                                  Prob > chi2       =     0.0001
Log likelihood =   -100.724                       Pseudo R2         =     0.1416

------------------------------------------------------------------------------
         low |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |  -.0271003   .0364504    -0.74   0.457    -.0985418    .0443412
         lwt |  -.0151508   .0069259    -2.19   0.029    -.0287253   -.0015763
             |
        race |
       black |   1.262647   .5264101     2.40   0.016     .2309024    2.294392
       other |   .8620792   .4391532     1.96   0.050     .0013548    1.722804
             |
       smoke |   .9233448   .4008266     2.30   0.021      .137739    1.708951
         ptl |   .5418366    .346249     1.56   0.118     -.136799    1.220472
          ht |   1.832518   .6916292     2.65   0.008     .4769494    3.188086
          ui |   .7585135   .4593768     1.65   0.099    -.1418484    1.658875
       _cons |   .4612239    1.20459     0.38   0.702    -1.899729    2.822176
------------------------------------------------------------------------------
```

Stata实现logistic regression

```
// 求出平均边际效应  average marginal (partial) effects,
// which means effects are calculated for each observation in the data and then averaged.
//https://www.stata.com/features/overview/marginal-analysis/

.  margins, dydx(*)

Average marginal effects                        Number of obs     =        189
Model VCE    : OIM

Expression   : Pr(low), predict()
dy/dx w.r.t. : age lwt 2.race 3.race smoke ptl ht ui
------------------------------------------------------------------------------
             |            Delta-method
             |      dy/dx   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |  -.0048315   .0064747    -0.75   0.456    -.0175217    .0078587
         lwt |  -.0027011   .0011835    -2.28   0.022    -.0050207   -.0003816
             |
        race |
       black |   .2326941   .0995698     2.34   0.019     .0375409    .4278473
       other |   .1511004   .0760619     1.99   0.047     .0020217     .300179
             |
       smoke |   .1646164   .0681744     2.41   0.016     .0309971    .2982358
         ptl |   .0966001   .0602536     1.60   0.109    -.0214948    .2146951
          ht |   .3267063   .1148706     2.84   0.004     .1015641    .5518485
          ui |   .1352299   .0797297     1.70   0.090    -.0210375    .2914972
------------------------------------------------------------------------------
Note: dy/dx for factor levels is the discrete change from the base level.
```

Stata实现logistic regression

```
 // 建立logit模型, 输出发生比

. logistic low age lwt i.race smoke ptl ht ui
Logistic regression                               Number of obs     =        189
                                                  LR chi2(8)        =      33.22
                                                  Prob > chi2       =     0.0001
Log likelihood =   -100.724                       Pseudo R2         =     0.1416

------------------------------------------------------------------------------
         low | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   .9732636   .0354759    -0.74   0.457     .9061578    1.045339
         lwt |   .9849634   .0068217    -2.19   0.029     .9716834    .9984249
             |
        race |
       black |   3.534767   1.860737     2.40   0.016     1.259736    9.918406
       other |   2.368079   1.039949     1.96   0.050     1.001356    5.600207
             |
       smoke |   2.517698    1.00916     2.30   0.021     1.147676    5.523162
         ptl |   1.719161   .5952579     1.56   0.118     .8721455    3.388787
          ht |   6.249602   4.322408     2.65   0.008     1.611152    24.24199
          ui |     2.1351   .9808153     1.65   0.099     .8677528      5.2534
       _cons |   1.586014   1.910496     0.38   0.702     .1496092     16.8134
------------------------------------------------------------------------------
Note: _cons estimates baseline odds.
```

信息准则

信息准则：比较所含解释变量个数不同的多元回归模型的拟合优度常用的标准之一

常用的两个信息准则都是基于 “-2对数似然值(-2 log likelihood)" 构造的.

$$-2LL = -2\ln L$$

该值越小, 说明似然函数值越大, 模型拟合程度越好 (但需调整变量个数增多带来的“过拟合").

基于 “-2对数似然值," 产生了著名的AIC和BIC信息准则.

信息准则

基于 “-2对数似然值," 产生了著名的AIC和BIC信息准则.

- AIC 赤池信息准则 (Akaike’s information criteria)

$$\mathrm{AIC} = -2\ln L + 2k$$

k是待估参数数量.

- BIC 贝叶斯信息准则 (Bayesian information criteria)

$$\mathrm{BIC} = -2\ln L + k\ln N$$

k是待估参数数量, N是观察数量.
(Stata code: estat ic)

计算AIC和BIC

```
. estat ic

Akaike's information criterion and Bayesian information criterion

------------------------------------------------------------------------------
       Model |          N   ll(null)  ll(model)      df        AIC        BIC
-------------+----------------------------------------------------------------
           . |        189   -117.336   -100.724       9    219.448   248.6237
------------------------------------------------------------------------------
Note: BIC uses N = number of observations. See [R] BIC note.
```

变量的显著性检验

考察变量影响是否显著. 有两种等价的检验

- Wald检验 Wald统计量:

$$\text{Wald} = \left(\frac{\hat{\beta}_j}{SE(\beta_j)} \right)^2 \sim \chi^2(1)$$

Wald检验值越大,表明该自变量的作用显著. SPSS默认报告Wald test

- z-test Z统计量

$$\text{Z} = \left(\frac{\hat{\beta}_j}{SE(\beta_j)} \right) \sim N(0,1)$$

Stata默认报告z-test. 如想看Wald statistics, 可以用test命令

模型显著性检验: 似然比检验

模型显著性检验:

- 线性模型：F检验
- Logit模型：似然比检验.

评价一个含有 p 个自变量的模型, 定义:

- L_0 为截距模型 (只有截距项,不含自变量) (restrict model)的似然函数值
- L_x 为包含所有 p 个自变量的模型 (unrestricted model)的似然函数值

定义似然比 (Likelihood-Ratio)

$$\mathrm{LR} = \left(\frac{L_0}{L_x}\right)$$

如果p个自变量能够解释概率, 则L_x应该更大.

- $LR = 1$, 这些自变量完全没有解释效果
- $LR < 1$, 自变量对因变量变化的解释有显著的贡献

模型显著性检验: 似然比检验

然而要进行统计检验, 我们不知道LR的抽样分布. 但可以证明

$$-\ln\left(\frac{L_0}{L_x}\right)^2 \sim \chi^2(p)$$

所以模型显著性的似然比检验:

$$-\ln\left(\frac{L_0}{L_x}\right)^2 = -2\ln\left(\frac{L_0}{L_x}\right) = [-2LL_o]-[-2LL_x] = 2LL_x-2LL_0$$

哪里展示了似然比检验结果?

```
Logistic regression                               Number of obs     =        189
                                                  LR chi2(8)        =      33.22
                                                  Prob > chi2       =     0.0001
Log likelihood =   -100.724                       Pseudo R2         =     0.1416

------------------------------------------------------------------------------
         low |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |  -.0271003   .0364504    -0.74   0.457    -.0985418    .0443412
         lwt |  -.0151508   .0069259    -2.19   0.029    -.0287253   -.0015763
             |
        race |
       black |   1.262647   .5264101     2.40   0.016     .2309024    2.294392
       other |   .8620792   .4391532     1.96   0.050     .0013548    1.722804
             |
       smoke |   .9233448   .4008266     2.30   0.021      .137739    1.708951
         ptl |   .5418366    .346249     1.56   0.118     -.136799    1.220472
          ht |   1.832518   .6916292     2.65   0.008     .4769494    3.188086
          ui |   .7585135   .4593768     1.65   0.099    -.1418484    1.658875
       _cons |   .4612239    1.20459     0.38   0.702    -1.899729    2.822176
------------------------------------------------------------------------------
```

拓展知识：拟合优度: 伪 R^2, $pseudo - R^2$

伪 R^2 ($pseudo - R^2$): 与线性回归模型的R^2 相对应, 衡量模型拟合程度. 常用的伪 R^2:

- Cox-Snell's $pseudo - R^2$

$$R^2_{CS} = 1 - \left[\frac{L_0}{L_x}\right]^{2/n} < 1$$

其中, L_0 **截距模型**的似然值; L_x **当前模型**的似然值

- Nagelkerke's $pseudo - R^2$:

$$R^2_{\text{adj}} = \frac{R^2_{CS}}{R^2_{\max}} \in [0, 1]$$

其中, $R^2_{\max} = 1 - (L_0)^{2/n}$ 是 R^2_{CS} 的最大值

- (Stata默认值) McFadden (1974)'s $pseudo - R^2$

$$1 - LL_x/LL_0$$

拓展知识：有关伪R^2的进一步说明

WHAT ARE PSEUDO R-SQUAREDS?
UCLA IDRE Statistical Consulting.
https://stats.idre.ucla.edu/other/mult-pkg/faq/general/faq-what-are-pseudo-r-squareds/

哪里展示了拟合优度$pseudo - R^2$结果?

```
Logistic regression                               Number of obs     =        189
                                                  LR chi2(8)        =      33.22
                                                  Prob > chi2       =     0.0001
Log likelihood =   -100.724                       Pseudo R2         =     0.1416

------------------------------------------------------------------------------
         low |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |  -.0271003   .0364504    -0.74   0.457    -.0985418    .0443412
         lwt |  -.0151508   .0069259    -2.19   0.029    -.0287253   -.0015763
             |
        race |
       black |   1.262647   .5264101     2.40   0.016     .2309024    2.294392
       other |   .8620792   .4391532     1.96   0.050     .0013548    1.722804
             |
       smoke |   .9233448   .4008266     2.30   0.021      .137739    1.708951
         ptl |   .5418366    .346249     1.56   0.118     -.136799    1.220472
          ht |   1.832518   .6916292     2.65   0.008     .4769494    3.188086
          ui |   .7585135   .4593768     1.65   0.099    -.1418484    1.658875
       _cons |   .4612239    1.20459     0.38   0.702    -1.899729    2.822176
------------------------------------------------------------------------------
```

二分类的Logistic回归：预测

根据解释变量预测观测属于哪一类/事件发生不发生等. 分类规则：

- $\hat{p}_i > 0.5$ 属于第一类
- $\hat{p}_i < 0.5$属于第二类
- $\hat{p}_i = 0.5$待判

Stata实现: 预测 predicted probability

```
. predict y
(option pr assumed; Pr(low))

. sum y, de
                          Pr(low)
-------------------------------------------------------------
      Percentiles      Smallest
 1%     .0382711       .0272559
 5%     .0655031       .0382711
10%     .0827243        .040474       Obs                 189
25%     .1681123        .045619       Sum of Wgt.         189

50%     .2791996                      Mean           .3121693
                        Largest       Std. Dev.      .1913915
75%     .4124026       .7774884
90%     .5941253        .791357       Variance       .0366307
95%     .7065833       .8067943       Skewness       .7361157
99%     .8067943       .8391283       Kurtosis       2.856246
```

模型预测准确率的判断方法

分类表 (Classification Table): 建立一个 2×2 的交互表, 来比较预测情况和实际发生的情况. 默认用概率为0.5切割.

```
. estat classification
                -------- True --------
Classified |         D             ~D  |      Total
-----------+--------------------------+-----------
     +     |        21              12  |         33
     -     |        38             118  |        156
-----------+--------------------------+-----------
   Total   |        59             130  |        189
Classified + if predicted Pr(D) >= .5
True D defined as low != 0
-------------------------------------------------
Sensitivity                     Pr( +| D)   35.59%
Specificity                     Pr( -|~D)   90.77%
Positive predictive value       Pr( D| +)   63.64%
Negative predictive value       Pr(~D| -)   75.64%
-------------------------------------------------
False + rate for true ~D        Pr( +|~D)    9.23%
False - rate for true D         Pr( -| D)   64.41%
False + rate for classified +   Pr(~D| +)   36.36%
False - rate for classified -   Pr( D| -)   24.36%
-------------------------------------------------
Correctly classified                        73.54%
-------------------------------------------------
```

预测正确的概率 . di (21+118)/(21+118+38+12)
.73544974

拓展知识： Logit模型的另两个重要应用

- 指示函数模型 (Index function model) (潜变量模型 Latent variable model)
- 随机效用模型(Random Utility model)

根据随机误差项分布的假设不同, 上述模型均可化成Logit模型或Probit模型. Cameron and Trivedi (2005) p375-477

一元二分类的Logistic回归 vs 多元二分类的Logistic回归

一元二分类:

$$\text{Logit}\left(p_i\right) = \ln\left(\frac{p_i}{q_i}\right) = \beta_0 + \beta_1 x_i + \varepsilon_i$$

$\Rightarrow$

$$\hat{p}_i = \frac{\exp\left(b_0 + b_1 x_i\right)}{1 + \exp\left(b_0 + b_1 x_i\right)}$$

多元二分类:

$$\text{Logit}\left(p_i\right) = \ln\left(\frac{p_i}{q_i}\right) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \varepsilon_i$$

$\Rightarrow$

$$\hat{p}_i = \frac{\exp\left(b_0 + b_1 x_{i1} + b_2 x_{i2} + \cdots + b_p x_{ip}\right)}{1 + \exp\left(b_0 + b_1 x_{i1} + b_2 x_{i2} + \cdots + b_p x_{ip}\right)}$$

拓展知识：因变量为M类-多项Logistic回归

如果因变量是没有顺序的M类, 则使用多项的logistic 回归 (multinomial logit).

由于

$$P_1 + P_2 + \cdots + P_M = 1$$

所以

$$\frac{P_i}{P_M} \in (0, +\infty), i = 1, 2, \cdots, M-1$$

可以构造

$$\text{Logit}\,(P_i) = \ln\left(\frac{P_i}{P_M}\right) = \beta_i + \sum_{j=1}^{k} \beta_{ij} x_j \quad i = 1, 2, \ldots, M-1$$

其中

$$P_1 + P_2 + \cdots + P_M = 1$$

拓展知识：多项Logistic回归

$$P_i = \frac{\exp\left(\beta_i + \sum_{j=1}^{k} \beta_{ij} x_j\right)}{1 + \sum_{i=1}^{M-1} \exp\left(\beta_i + \sum_{j=1}^{k} \beta_{ij} x_j\right)} \quad i = 1, 2, \ldots, M-1$$

$$P_M = \frac{1}{1 + \sum_{i=1}^{M-1} \exp\left(\beta_i + \sum_{j=1}^{k} \beta_{ij} x_j\right)}$$

- 分类规则：在各类概率中, 若P_i取值最大, 则判为第 i 类

拓展知识： Stata做multinomial logit model

help mlogit

作业 (不交, 不计入平时成绩)

- 针对sysuse nlsw88.dta数据, 采用二分类logistic回归进行建模计算, 分析影响“是否加入工会(union=0,1)" 的影响因素, 给出回归结果和分析过程.