社会统计学及SPSS软件应用
STATISTICS WITH SPSS

Instructor:王荣欣

Email: rxwang@qq.com

周一3-4节、单周周四3-4节，3A106-2

December 21, 2020

CONTENTS

1. Limited dependent variable

1 数据断尾

(1) 断尾回归

(2) 零断尾泊松回归和负二项回归

(3) 偶然断尾和样本选择

2 数据删失（归并）

(1) Tobit回归

(2) Quantile regression

(3) 事件史分析

- censored data：删失了那些尚未发生研究事件的人。

  (1) 左删失：$C > T$
  C为调查时间。有事件发生，但不知道是什么时候发生。

  (2) 右删失：$T > C$
  尚未经历所研究的事件，或在调查结束后仍未观测到事件的发生。

  (3) 左右都删失

- truncated data：断尾

CENSORED DATA

- For a continuous variable y*, if in the sample of size n, we observed:

  (1) $y_i = x_i\beta_i + \epsilon_i = y_i*$, if $y_i* > c$
  （真实值y_i*大于删失值c，y的观测值就等于y_i*）

  (2) $y_i = c$, otherwise
  （所有y_i都被归并为c）

- then the sample $(y_1, y_2 ... y_n)$ is said to be a censored sample, left-censored, to be specific.

TRUNCATED DATA

1. If the distribution of y* is cut off at the point y*=c before the sample is drawn, so that no observations are drawn for the cases in the range of y*$>$c,
2. then the sample is said to be truncated, in this case, right-truncated.

Example 2.1

y_i为所有企业的销售收入，而统计局只收集规模以上企业数据，比如$y_i \geq$*100,000*。被解释变量在 *100,000* 处存在“左端断尾”。

1 truncated regression只能观测到部分子样本信息，而另外一部分的个案无法观测到任何信息。

2 censored regression虽有全部的观测数据，但对于某些观测数据，因变量y_i被压缩在一个点上。

   (1) “收入高于10万元（top coding）”或“收入低于1000元”。

   (2) duration data:

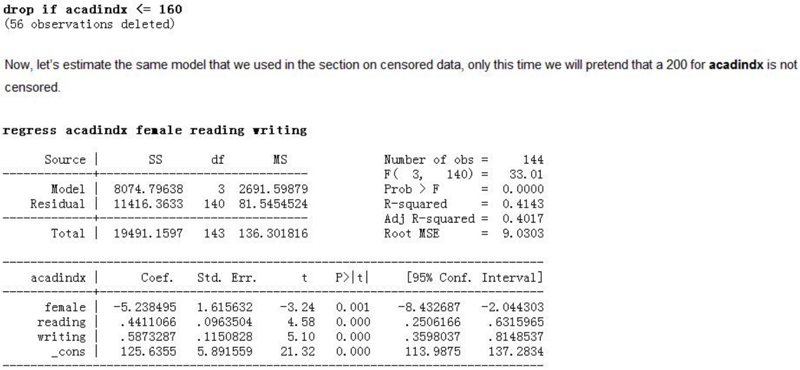
      1. in the study of the effects of unemployment insurance on the duration of employment, we might follow new claimants for up to 40 weeks.
      2. Anyone out of work for longer has an unemployment spell length that is censored at 40.

针对断尾因变量的断尾回归

```
truncreg y x1 x2 x3, ll(#) ul(#)
```

ll(#)表示lower limit，即左端断尾（截除）。ul(#)表示upper limit，即右端断尾。

Regression with Truncated Data

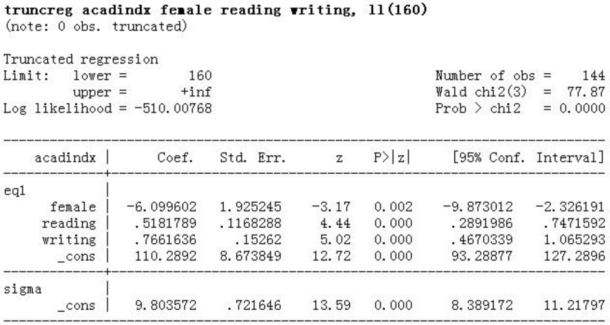


```
drop if acadindx <= 160
(56 observations deleted)
```

Now, let's estimate the same model that we used in the section on censored data, only this time we will pretend that a 200 for **acadindx** is not censored.

```
regress acadindx female reading writing

      Source |       SS       df       MS              Number of obs =     144
-------------+------------------------------           F(  3,   140) =   33.01
       Model |  8074.79638     3  2691.59879           Prob > F      =  0.0000
    Residual |  11416.3633   140  81.5454524           R-squared     =  0.4143
-------------+------------------------------           Adj R-squared =  0.4017
       Total |  19491.1597   143  136.301816           Root MSE      =  9.0303

------------------------------------------------------------------------------
    acadindx |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      female |  -5.238495   1.615632    -3.24   0.001    -8.432687   -2.044303
     reading |   .4411066   .0963504     4.58   0.000     .2506166    .6315965
     writing |   .5873287   .1150828     5.10   0.000     .3598037    .8148537
       _cons |   125.6355   5.891559    21.32   0.000     113.9875    137.2834
------------------------------------------------------------------------------
```

Regression with Truncated Data



```
truncreg acadindx female reading writing, ll(160)
(note: 0 obs. truncated)

Truncated regression
Limit:   lower =        160                     Number of obs =    144
         upper =       +inf                     Wald chi2(3)  =  77.87
Log likelihood = -510.00768                     Prob > chi2   = 0.0000

------------------------------------------------------------------------------
    acadindx |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
eq1          |
      female |  -6.099602   1.925245    -3.17   0.002    -9.873012   -2.326191
     reading |   .5181789   .1168288     4.44   0.000     .2891986    .7471592
     writing |   .7661636     .15262     5.02   0.000     .4670339    1.065293
       _cons |   110.2892   8.673849    12.72   0.000     93.28877    127.2896
-------------+----------------------------------------------------------------
sigma        |
       _cons |   9.803572    .721646    13.59   0.000     8.389172    11.21797
------------------------------------------------------------------------------
```

零断尾泊松回归与负二项回归

计数数据有时仅包括正整数，不包括取值为0的观测值，称为“零断尾”(zero-truncated)。

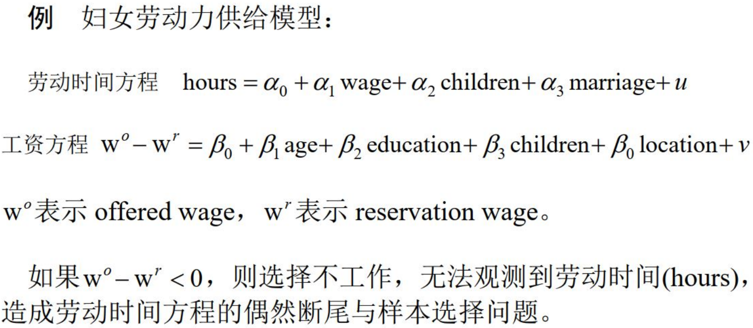
- 例：在商场发放问卷调查，研究消费者每周去商场的次数。
- 例：在公交车上发放问卷调查，研究乘车者每周坐公交的次数。

1. ztp y x1 x2 x3 /*zero-truncated Poisson regression*/
2. ztnb y x1 x2 x3
   /*zero-truncated negative binomial regression*/

样本选择

1 被解释变量y_i的断尾有时与另一变量z_i有关，称为“偶然断尾”（incidental truncation）或“样本选择”（sample selection）。

(1) z_i被称为选择变量。



例 妇女劳动力供给模型：

劳动时间方程 $\text{hours} = \alpha_0 + \alpha_1 \text{wage} + \alpha_2 \text{children} + \alpha_3 \text{marriage} + u$

工资方程 $\text{w}^o - \text{w}^r = \beta_0 + \beta_1 \text{age} + \beta_2 \text{education} + \beta_3 \text{children} + \beta_0 \text{location} + v$

w^o表示 offered wage， w^r表示 reservation wage。

如果$\text{w}^o - \text{w}^r < 0$，则选择不工作，无法观测到劳动时间(hours)，造成劳动时间方程的偶然断尾与样本选择问题。

Example 3.1

在美国的亚裔移民给人的整体印象是聪明能干。但在美国的亚裔并非亚洲人口的代表性样本。

通常只有受过高等教育或具有吃苦冒险精神的亚裔才会“自我选择”（*self selection*）移民。

决定移民与否的变量z_i便对被解释变量y_i产生了断尾作用，故“样本选择”将导致“选择性偏误”（*selection bias*）。

样本选择模型的Stata命令

1.
   ```
   heckman y x1 x2 x3, select(z1 z2)
     /*默认使用MLE，选择方程的被解释变量为y*/
   ```

针对删失因变量的TOBIT回归

- 删失因变量：记录的值不能够真实反映真正的潜在变量的全部取值。
- tobit模型：Tobin’s probit models as first studied by Tobin(1958), use MLE to get a set of unbiased β.

针对删失因变量的TOBIT回归

- The Tobit model is defined as follows:

$$y_i = x_i\beta' + \epsilon_i, \text{ if } y_i* > 0$$

$$y_i = 0 \text{ , otherwise}$$

Example 4.1

A leading example from labor economics is Current Population Survey earnings data, which topcodes (censors) very high values of earnings to protect respondent confidentiality.

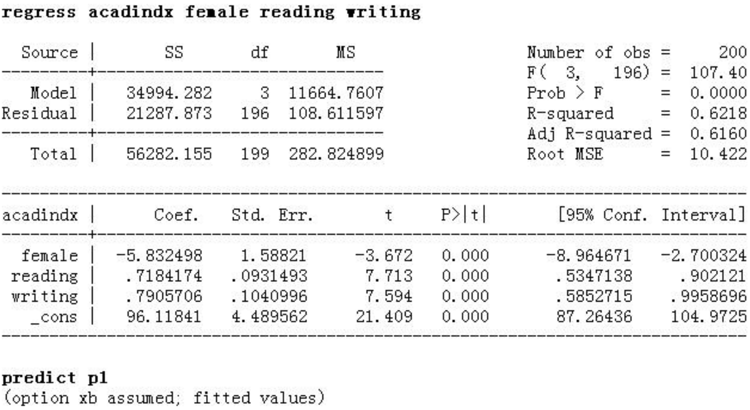
Typically, we are interested in the causal effect of schooling on earnings as it appears on respondents' tax returns, not their CPS-topcoded earnings.

Chamberlain(1994)shows that is some years, CPS topcoding reduces the measured returns to schooling considerably.

1 tobit y x1 x2 x3, ll(#) ul(#)

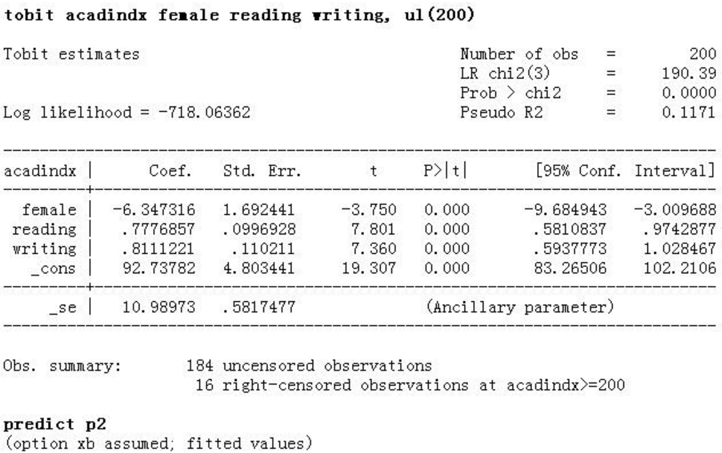
ll(#)表示lower limit，即左归并。ul(#)表示upper limit，即右归并。

REGRESSION WITH CENSORED DATA



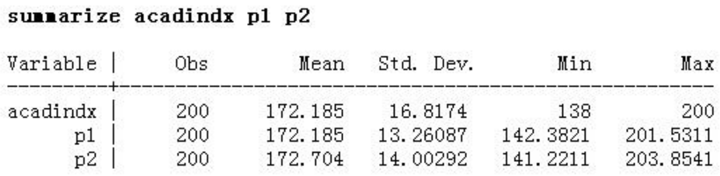
```
regress acadindx female reading writing

  Source |       SS       df       MS              Number of obs =     200
---------+------------------------------           F(  3,   196) =  107.40
   Model |  34994.282     3  11664.7607            Prob > F      =  0.0000
Residual |  21287.873   196  108.611597            R-squared     =  0.6218
---------+------------------------------           Adj R-squared =  0.6160
   Total |  56282.155   199  282.824899            Root MSE      =  10.422

------------------------------------------------------------------------------
acadindx |      Coef.   Std. Err.       t     P>|t|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
  female |  -5.832498    1.58821     -3.672   0.000      -8.964671   -2.700324
 reading |   .7184174   .0931493      7.713   0.000       .5347138     .902121
 writing |   .7905706   .1040996      7.594   0.000       .5852715    .9958696
   _cons |   96.11841   4.489562     21.409   0.000       87.26436    104.9725
------------------------------------------------------------------------------

predict p1
(option xb assumed; fitted values)
```

REGRESSION WITH CENSORED DATA



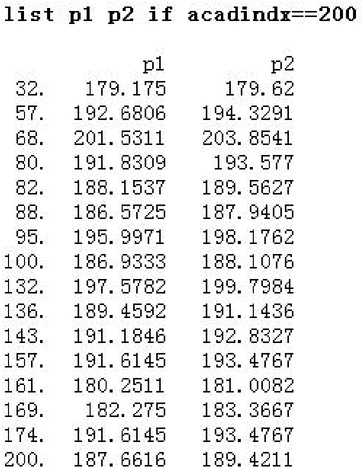
```
tobit acadindx female reading writing, ul(200)

Tobit estimates                                   Number of obs   =        200
                                                  LR chi2(3)      =     190.39
                                                  Prob > chi2     =     0.0000
Log likelihood = -718.06362                       Pseudo R2       =     0.1171

------------------------------------------------------------------------------
    acadindx |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      female |  -6.347316   1.692441    -3.750   0.000    -9.684943   -3.009688
     reading |   .7776857   .0996928     7.801   0.000     .5810837    .9742877
     writing |   .8111221    .110211     7.360   0.000     .5937773    1.028467
       _cons |   92.73782   4.803441    19.307   0.000     83.26506    102.2106
-------------+----------------------------------------------------------------
         _se |   10.98973   .5817477          (Ancillary parameter)
------------------------------------------------------------------------------

Obs. summary:        184 uncensored observations
                      16 right-censored observations at acadindx>=200

predict p2
(option xb assumed; fitted values)
```

Regression with Censored Data



```
summarize acadindx p1 p2

Variable |       Obs        Mean    Std. Dev.       Min        Max
---------+--------------------------------------------------------
acadindx |       200     172.185     16.8174         138        200
      p1 |       200     172.185    13.26087    142.3821   201.5311
      p2 |       200     172.704    14.00292    141.2211   203.8541
```

Regression with Censored Data



```
list p1 p2 if acadindx==200

              p1         p2
 32.     179.175     179.62
 57.    192.6806   194.3291
 68.    201.5311   203.8541
 80.    191.8309    193.577
 82.    188.1537   189.5627
 88.    186.5725   187.9405
 95.    195.9971   198.1762
100.    186.9333   188.1076
132.    197.5782   199.7984
136.    189.4592   191.1436
143.    191.1846   192.8327
157.    191.6145   193.4767
161.    180.2511   181.0082
169.     182.275   183.3667
174.    191.6145   193.4767
200.    187.6616   189.4211
```

1. Quantile regression can be used to estimate the effect of covariates on conditional quantiles that are below the censoring point.

2. If CPS topcoding affects relatively few people, censoring has no effect on estimates of the conditional median.

1. qreg y x1 x2 x3 /*默认为中位数回归*/
2. qreg y x1 x2 x3, q(#) /*#分位数回归*/

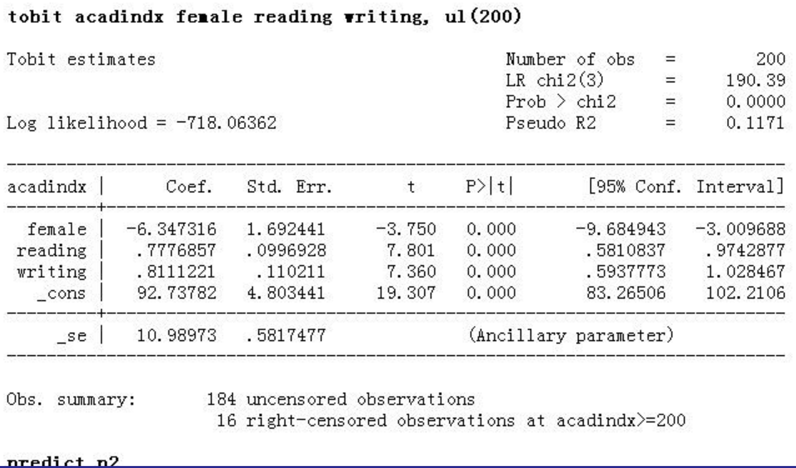
虚假相关（spurious correlation）（教材第150页）

1 explanation model（因果分析）：以事实鉴定因果的虚实，引入第三类变量

2 interpretation（阐明分析）：引入中介变量

3 conditional analysis（条件分析）:引入压抑变量（suppressor variable）

explanation model（因果分析）：引入第三类变量（前置变量）



```
tobit acadindx female reading writing, ul(200)

Tobit estimates                                   Number of obs   =        200
                                                  LR chi2(3)      =     190.39
                                                  Prob > chi2     =     0.0000
Log likelihood = -718.06362                       Pseudo R2       =     0.1171

------------------------------------------------------------------------------
    acadindx |      Coef.   Std. Err.       t     P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      female |  -6.347316   1.692441     -3.750   0.000     -9.684943   -3.009688
     reading |   .7776857   .0996928      7.801   0.000      .5810837    .9742877
     writing |   .8111221    .110211      7.360   0.000      .5937773    1.028467
       _cons |   92.73782   4.803441     19.307   0.000      83.26506    102.2106
-------------+----------------------------------------------------------------
         _se |   10.98973   .5817477            (Ancillary parameter)
------------------------------------------------------------------------------

Obs. summary:        184 uncensored observations
                      16 right-censored observations at acadindx>=200

predict p2
```

interpretation（阐明分析）：引入中介变量（mediator）

教育水平⟶生育子女数量

1 教育水平⟶结婚年龄⟶生育子女数量

2 教育水平⟶生育观念（重男轻女）⟶生育子女数量

3 教育水平⟶社会意识⟶生育子女数量

conditional analysis（条件分析）：引入压抑变量（suppressor variable）

人口密度$\longrightarrow$？ 精神病发病率

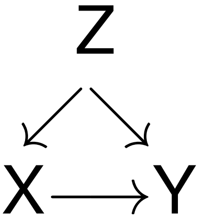
1 文化同质性高的社区

2 文化异质性低的社区$\longrightarrow$精神病发病率

$$X \longrightarrow Y$$

体育锻炼⟶身体健康

1 观察到的现象：爱锻炼身体的人更加健康。
2 我们要证明锻炼身体能促进身体健康，但得考虑年龄因素。
3 年轻人更爱锻炼身体，年轻人的身体也更加健康。
4 X 代表每天锻炼身体，Y是身体健康，Z是年龄。



1. The term ”confounding” originally meant ”mixing”.

2. The true causal effect X→Y is ”mixed” with the spurious correlation between X and Y induced by the fork X←Z→Y.

3. Z is a confounder of the proposed causal relationship between X and Y.

4. “控制变量”——控制了年龄变量，“controlling for Z”，再看X和Y之间有没有关系。

统计控制

1. 采用多元回归最大的优点是可以将对因变量有重要影响的自变量同时纳入分析，在控制其他自变量的条件下一一求解对应自变量的偏回归系数（教材第60-62页）。
2. 多元回归分析得到的回归系数表示在控制其他自变量的条件下，每个自变量对因变量单独的净作用（教材第138页）。

1. Sewall Wright(1889-1988)
   (1) Some correlations do imply causation.
2. path analysis
3. path coefficients

1. 原因变量（predictor variable）
2. 结果变量（response variable）

1 外生变量（exogenous variable）

2 内生变量（endogenous variable）

   (1) 最终反应变量（ultimate response variable）

1 递归模型（recursive model）

2 非递归模型（nonrecursive model）

1 total effects

(1) direct effects

(2) indirect effects

Total effect=Direct effect+Indirect effect

1. 模型设定
2. 模型识别与模型估计
3. 模型评价
4. 模型修正