

政治经济学前沿方法论与量化分析

第七讲 质量型变量的建模

上课地点： 善斋306C
上课时间：周二第六大节

龙治铭
善斋307C
zhiminglong@tsinghua.edu.cn



清华大学
Tsinghua University

目录

CONTENTS



质量型变量的描述性统计



Logit和Probit模型的理论基础



Logit和Probit模型在Stata中的实现



参考文献

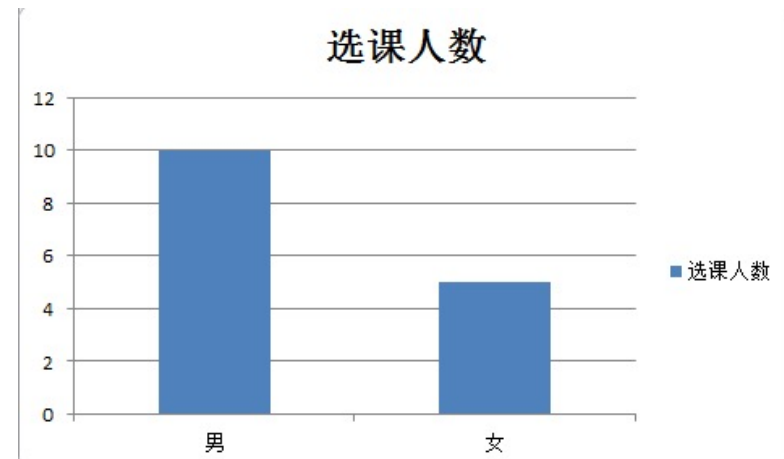
1

质量型变量 (Qualitative variable) 的描述性统计

※前面讲的随机变量都是数量型变量 (quantitative variable), 他们的随机试验结果本身是一个数, 如股价、分数、身高等等。数量型变量的取值可以是连续变化和连续记录的, 如温度, 也可以是离散变化和离散记录的, 如股价 (A股市场股价最小变动单位为分) 和成交量 (最低成交单位为100股, 不存在0.76股一类连续值)。

※一种随机变量的试验结果不是数, 但为了研究方便, 以数字记录, 如颜色、性别、级别等等, 我们称为质量型变量 (qualitative variable, 有的统计学教材也称为Categorical Variables或者nominal variable)。质量型变量的取值可以是连续变化和连续记录的, 如光谱的颜色 (有无穷多种红色, 只不过人类语言无法完全描述), 也可以是离散变化和离散记录的, 后者是常见的情况。

※质量型变量的取值可以只有两个, 通常用0,1表示, 这种情况叫做二元变量 (binary variable), 在回归分析中经常用作指示变量 (indicator variable) 或者哑变量 (Dummy variable), 也可以有多个取值如 “本科生年级” 这个变量取值可以为{一年级, 二年级, 三年级, 四年级}, 可以用数值{1,2,3,4}表示。



选课同学的性别是一个典型的 qualitative variable

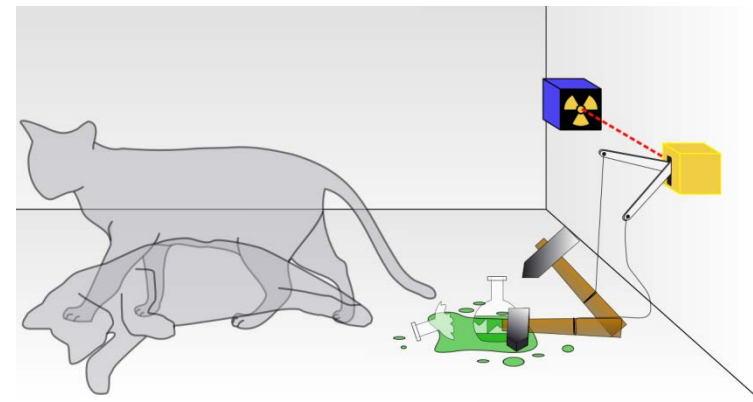
※质量型变量的取值尽管用数字表示，但**数字本身没有任何意义。因此数量型变量的数字特征一般不适用于质量变量**（二元变量的平均值有一定的意义，可以表示比例或者频率）。

例如：班上15名同学，其中5名女生。对于“性别”这个变量，如果用0表示女生，1表示男生，平均值为0.33；反过来如果用1表示女生，0表示男生，平均值为0.66。0.33和0.66表示什么性别？

又如，我院有助理教授3人，副教授18人，教授36人（包含退休），如果变量“职称”采取以下赋值：
助理教授为0，副教授为1，教授为2，那么均值为1.58，方差为0.36，此时的数字特征意义是什么？1.58是什么职称？标准差0.6还有没有意义？

如果改变赋值，助理教授为10，副教授为20，教授为40，均值为32，方差为113，此时的数字特征意义是什么？

※例子表明质量型变量的数字特征有其特殊性，需要专门的分析指标。



质量型随机变量均值的典型悖论——薛定谔猫：由于事件的随机性，猫会处于生存与死亡的叠加态，随机实验结果的空间并不存在这样的状态。

※ 二元变量的均值有一定的意义（尽管会出现一定的困惑），可以表示某一随机试验结果的出现频率。但如果质量型变量的取值（Modality）为三个以上，均值失去了直观的涵义，没有使用的意义。

※ 列联表（Contingency tables）用于显示变量的频率分布，适用于质量型变量的统计描述。大量的关于质量型变量的统计检验都基于列联表。

Sex \ Handed- ness	Handed- ness		Total
	Right handed	Left handed	
Male	43	9	52
Female	44	4	48
Total	87	13	100

Wiki上的二元列联表例子

tab rank			
rank	Freq.	Percent	Cum.
1	61	15.25	15.25
2	151	37.75	53.00
3	121	30.25	83.25
4	67	16.75	100.00
Total	400	100.00	

Stata自带数据库中的例子

※质量型变量的方差和标准差没有任何意义，标准差反映的本质是离散度（Dispersion），对于二元变量，例如掷硬币，掷50次和掷100次，理论上离散程度都是一样的，然而一个方差为0.2551，后者为0.1096。此外，均值±标准差反映了随机变量取值的分布区间，负的取值区间显然没有意义也没有必要。

大量经济学论文（甚至是等级较高的国际学术期刊）都对质量型变量运用了数量型变量的统计描述方法，这些论文都是错的！

※ Variation Ratio是一种较为简单和常用的，用于描述质量型变量离散程度的统计量。

$VR = 1 - (\text{proportion of cases in the mode})$

数值越大，表明分布越分散

理论上，掷硬币的VR=0.5

选课同学性别的VR=0.66或0.33

局限性：需要指明早哪种modality

Wilcox's indexes:
$$M = \sum_{i=1}^K (f_m - f_i)$$

※其他衡量方法：非常多，见. Kader (2007) , Variability for Categorical Variables, Journal of Statistics Education Volume 15, Number 2.

※我们知道，在简单线性回归中，OLS估计量本质上就是简单线性相关系数（Pearson's r），因此在做回归分析之前，我们会计算变量之间的相关系数来探索变量之间可能的关系。

然而两个质量型变量的之间皮尔逊相关系数没有任何意义（二元变量除外）。

想想为什么没有意义？计算皮尔逊相关系数需要使用标准差，质量型变量的标准差没有意义。

※参考文献：Pottier(1994), Mesures de la liaison entre deux variables qualitatives : relation entre un coefficient de corrélation généralisé et le χ^2 , Revue de Statistique Appliquée, Tome 42 no. 1, p. 41-61

※Phi coefficient：基于列联表，用于检验两个质量变量的相关性。二元变量的Phi coefficient与Pearson coefficient相等。

	$y = 1$	$y = 0$	总计
$x = 1$	n_{11}	n_{10}	$n_{1\bullet}$
$x = 0$	n_{01}	n_{00}	$n_{0\bullet}$
总计	$n_{\bullet 1}$	$n_{\bullet 0}$	n

$$\phi = \frac{n_{11}n_{00} - n_{10}n_{01}}{\sqrt{n_{1\bullet}n_{0\bullet}n_{\bullet 0}n_{\bullet 1}}}$$

※卡方检验：

H0：质量变量X与Y无关。 H1：质量变量X与Y相关

可以证明统计检验量 χ^2 渐进服从自由度为 (k-1)卡方分布（证明从略，见Pearson, 1990）

※Cramér 's V (Cramér, 1946) 基于卡方检验和Phi coefficient检验两个质量变量之间的相关性

$$V = \sqrt{\frac{\varphi^2}{\min(k-1, r-1)}} = \sqrt{\frac{\chi^2/n}{\min(k-1, r-1)}}$$

※质量变量的相关系数在回归分析中的应用:

1) 当被解释变量为数量型，而解释变量既有数量型又有质量型时：

二元变量的Phi coefficient与Pearson coefficient相同，因此二元变量可以直接用于OLS。但是如果质量变量有N个modality, 需要设计N-1个二元变量，会引起多重共线性。

这时需要使用logit regression 或者probit regression.

2) 当被解释变量为质量型变量时，我们无法按照通常的OLS估计量解释，“当自变量提高1%，应变量提高百分之多少”，此时必须使用logit regression 或者probit regression，并使用odds ratio进行解释（稍后详细解释）。

2

Logit和Probit模型的理论基础

※目的：基于样本容量 N 的一组自变量 X_1, X_2, \dots, X_p 解释质量型变量 Y 。
 Y ：质量型变量。数量型变量可以通过分组的办法转变为质量型变量。
 X_1, X_2, \dots, X_p ：可以是质量型变量也可以是数量型变量。

※分类：

如果 Y 是二元变量，模型是简单Logit 或Probit模型。

如果 Y 有两个以上的取值，模型可以分为：

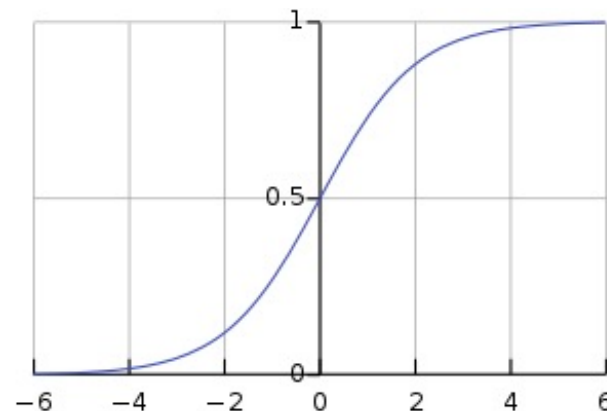
unordered polytomic model （ Y 的各个取值没有排序，例如颜色）

ordered polytomic model （ Y 的各个取值有排序，例如弱、中等、强）

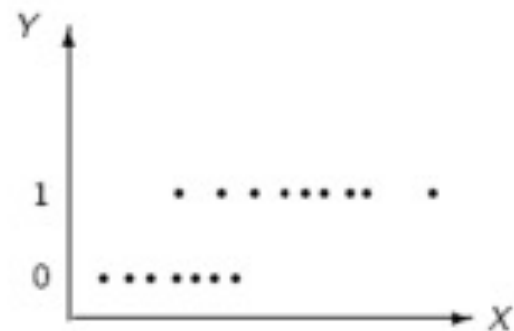
如果截去一部分信息，模型是Tobit 模型

※Probit model。Probit和Logit类似，只是使用的函数不同，Logit使用逻辑函数，Probit使用累积分布函数，Logit的推导更加简单，统计教材一般使用Logit作为推导示例。都使用最大似然估计。

最小二乘法不再适用，残差值通常不具备良好的性质（尤其是正态性和异方差性）



逻辑函数和累积分布函数是两个很好的，取值范围在0,1之间平滑变动的S型曲线，因此被发展出Logit 和Probit模型



※如何将连续变化的数值和0, 1二分取值相联系？需要引入潜在变量(variable inobservable)的概念：

$$Y_i^* = a + b_1 X_{1i} + b_2 X_{2i} + \dots + b_p X_{pi} + u_i = X_i \beta + u_i$$

$$\begin{cases} Y_i = 1 & \text{si } Y_i^* \geq 0 \\ Y_i = 0 & \text{sinon} \end{cases}$$

由于仅仅有变量Y可以观察到， $Y_i^* = X_i \beta + u_i$ 不能被直接估计。于是我们做这样的变换：

$$\begin{aligned} P(Y_i = 1) &= P(Y_i^* \geq 0) = P(X_i \beta + u_i \geq 0) = P(u_i \geq -X_i \beta) \\ &= P(u_i < X_i \beta) = F(X_i \beta) \end{aligned}$$

如果F是逻辑函数，那么模型是Logit

如果F是正态累积分布函数，那么模型是Probit

巧妙的将0,1取值和一个S型曲线联系起来了，这样可以使用参数估计。

※估计方法：最大似然估计（具体推导过程从略）

※将 $X_i\beta$ 代入逻辑函数：

$$F(X_i\beta) = \frac{1}{1 + \exp(-X_i\beta)} = \frac{\exp(X_i\beta)}{1 + \exp(X_i\beta)}$$

解反函数：

$$X_i\beta = F^{-1}[P(Y_i = 1)] = \log \frac{P(Y_i = 1)}{1 - P(Y_i = 1)} = \log \frac{P(Y_i = 1)}{P(Y_i = 0)}$$

因此我们得到了odds ratio 比值比：

基于解释变量 X_i 的比值比解释为：当 $X_i=1$ 时，相对于 $X_i=0$ ， $Y=1$ 的概率是 $Y=0$ 的 $\exp(b)$ 倍（假设 b 显著）

$$\begin{aligned} OR &= \frac{\frac{P(Y_i=1|X'_i)}{P(Y_i=0|X'_i)}}{\frac{P(Y_i=1|X_i)}{P(Y_i=0|X_i)}} = \exp \left[\log \frac{\frac{P(Y_i=1|X'_i)}{P(Y_i=0|X'_i)}}{\frac{P(Y_i=1|X_i)}{P(Y_i=0|X_i)}} \right] \\ &= \exp \left[\log \frac{P(Y_i = 1|X'_i)}{P(Y_i = 0|X'_i)} - \log \frac{P(Y_i = 1|X_i)}{P(Y_i = 0|X_i)} \right] = \exp [X'_i\beta - X_i\beta] \end{aligned}$$

显然，如果 X 也是一个二元变量，那么

OR的解释为：当 $X=1$ 时，相对于 $X=0$ ， $Y=1$ 的概率是 $Y=0$ 的 $\exp(b)$ 倍

$$\log \frac{P(Y_i = 1)}{P(Y_i = 0)} = a + bx_i$$

这样如果逻辑回归的系数是显著的（Wald Test），我们有了不同于OLS的解释

※一般情况：X有多个取值。

OR需要注明相对于X的哪一个的取值，意义是：

当 $X=k$ 时，相对于 $X=\text{reference}$ ， $Y=1$ 的概率是 $Y=0$ 的 $\exp(b)$ 倍

※特殊情况：X是数量型变量

OR的解释：X每增加一单位的效应

※估计质量：如何判断Logit模型估计得很好？

- 1) 信息准则：AIC, SIC, HQ, $\log \text{likelihood}$ (判断标准与前面介绍使用MLE的模型一致)
- 2) $\log \text{likelihood ratio test}$ (见VAR模型中对P的选择的介绍)
- 3) 伪 R^2 。使用MLE不存在 R^2 ，但可以计算一个Pseudo R^2
- 4) The number of matching pairs。模型可以预测（样本内预测），看Y的0,1取值是否符合观测值，正确的比例反映了模型的拟合能力。

unordered polytomic model的估

※如果Y有K个取值 $\begin{cases} Y_i = 1 \\ Y_i = 2 \\ \vdots \\ Y_i = K \end{cases}$ 只需要K-1个概率

$$\begin{aligned} p_{i1} &= P(Y_i = 1) \\ p_{i2} &= P(Y_i = 2) \\ &\vdots \\ p_{iK-1} &= P(Y_i = K - 1) \end{aligned}$$

$$\begin{aligned} Y_{i1}^* &= \beta_1' X_i + u_{i1} \\ &\vdots \\ Y_{ik}^* &= \beta_k' X_i + u_{ik} \\ &\vdots \\ Y_{iK-1}^* &= \beta_{K-1}' X_i + u_{iK-1} \end{aligned}$$

因此，可以选择某一个例如第K个取值作为reference，对于每一个latent variable $Y_{iK-1}^* = \beta_{K-1}' X_i + u_{iK-1}$

类似地推导过程我们得到了
其中假定 $\beta_K = 0$

$$\log \frac{P(Y_i = k)}{P(Y_i = K)} = \log \frac{\exp(\beta_k' X_i)}{\exp(\beta_K' X_i)} = \beta_k' X_i$$

系数可以解释为，Y=k相对于Y=K的概率

※优点：此时模型暗含的假设为Y=k相对于Y=K的概率只与Y=K相关，与其他取值无关，相当于分解为K-1个simple Logit.

否则需要Independance of Irrelevant Alternative假设，Logit模型中的每个取值都是独立的不相关的。后者是一个很强的假设，在Probit中更加难以估计。

※ordered polytomic model如果Y的取值有排序，例如我们希望知道，相对于弱，强的概率是多少

Latent variable 定义方式为：

其余推导过程类似（证明从略）

$$\begin{cases} Y_i = 1 & \text{si } Y_i^* < c_1 \\ Y_i = 2 & \text{si } c_1 \leq Y_i^* < c_2 \\ \vdots & \\ Y_i = K & \text{si } Y_i^* \geq c_{K-1} \end{cases}$$

$$\log \frac{Pr(Y > k)}{Pr(Y \leq k)}$$

※Tobit model的适用情景：Y是数量型变量但取值经常为0，这时对所有的观测值都应用线性回归是不合适的。于是我们把它分为两部分

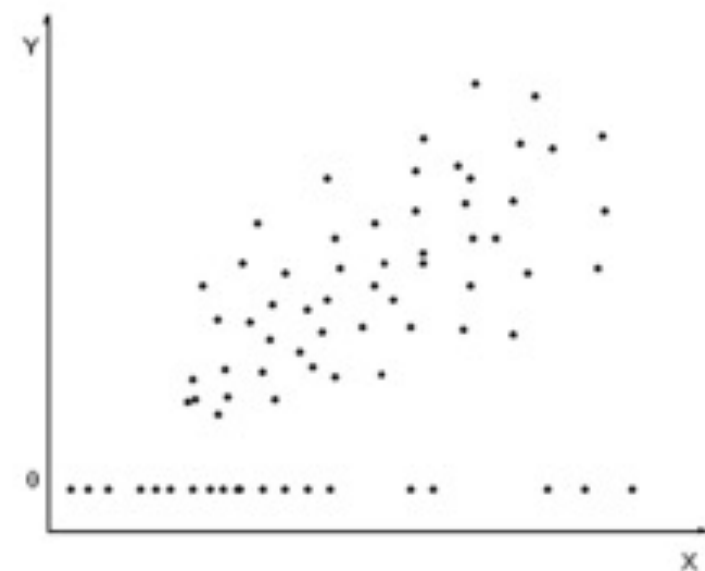
$$Y_i = \begin{cases} Y_i^* & \text{si } Y_i^* \geq 0 \\ 0 & \text{sinon} \end{cases} \quad Y_i^* = X_i\beta + u_i$$

其余推导过程类似（证明从略），使用MLE。

或者

$$Y_i = \begin{cases} Y_{1i}^* & \text{si } Y_{2i}^* > 0 \\ 0 & \text{si } Y_{2i}^* \leq 0 \end{cases} \quad \begin{aligned} Y_{1i}^* &= X_i'\beta_1 + u_{1i} \\ Y_{2i}^* &= X_i'\beta_2 + u_{2i} \end{aligned}$$

使用MLE或者Heckman 两步法（类似于工具变量）



3

Logit和Probit模型在Stata中的
实现

问题：研究龙治铭老师的给分规律

※GPA对清华同学非常重要，马原课作为4个学分的必修课，对GDP(x, GPA)具有重要的影响。那么龙治铭老师的给分是否存在规律，是科学还是艺术？

※假设1：龙老师是男老师，女同学是否更容易获得高分或者相反是否存在性别歧视？

美貌经济学：

Hamermesh (1994) The American Economic Review

Persico et. Al (2004). Journal of Political Economy

颜值难以量化，机器学习可能能够解决这个问题（我记得JPE2014年似乎有一篇用机器学习判断颜值对工作的影响（中国数据），没有找到原文）。

※假设2&3：龙老师是经济学博士，经管学院的同学是否更容易获得高分？龙老师是留学回国人员，外语系的同学是否更容易获得高分？

※假设4：龙老师教4个班，同样的内容需要重复四次（人类的本质是不是复读机？），更早的班级讲课更加有激情，课程安排在星期几对分数是否有显著影响？前八周后八周是否有影响？

※假设5：下午昏昏欲睡，上午起不了床，课程安排在上和下午是否对分数有显著影响？

※模型的选择：GPA虽然是数字，但本质是等级制成绩，是qualitative variable且有排序，因此应该使用ordered polytomic logit model

※数据来源：教务处&网络学堂

※微观数据的脱敏处理：删去一切可以识别个人特征的信息，保证同学们隐私

※数据的导入和描述：

```
. import excel "C:\Users\zhiminglong\Desktop\成绩单.xls",  
sheet("Sheet1") firstrow
```

```
. Describe
```

列出性别的列联表：

```
tabulate 总成绩 性别
```

```
. tabulate 总成绩 性别
```

总成绩	性别		Total
	女	男	
A	25	29	54
A+	9	8	17
A-	22	64	86
B	2	9	11
B+	52	156	208
C+	0	2	2
F	0	2	2
W	1	1	2
Total	111	271	382

```
-> tabulation of 性别 by 总成绩
```

性别	总成绩								Total
	A	A+	A-	B	B+	C+	F	W	
女	25	9	22	2	52	0	0	1	111
男	29	8	64	9	156	2	2	1	271
Total	54	17	86	11	208	2	2	2	382

```
Pearson chi2(7) = 17.3492 Pr = 0.015
```

```
Cramér's V = 0.2131
```

※计算所有可能的列联表及其卡方检验、Cramer's V

tab2 性别 院系 总成绩 绩点成绩 年级 星期 大节 前后, chi2 V

这个相当于系数矩阵，除了研究分布的情况，还可以发现各年级、各院系男女比例是否平衡

成绩和性别是否相关，介于显著和不显著之间，在5%水平上显著，在1%水平上不显著

Logit model1

※数据清洗：需要将汉字表述的变量取值赋予数值
generate sex= 0
replace sex = 1 if 性别 == "女"
destring 绩点成绩, replace force
destring 年级, replace force
generate grade= 2019-年级
generate week= 0
replace week = 1 if 前后 == "后"
院系较多，可以考虑设为group variable

※ 不考虑专业的ordered polytomic logit model
meoprobit 绩点成绩 sex grade week 星期 大节

结果解释: 1) 因共线性大节被删掉了
2) 模型具有整体显著性
3) 在控制年级、星期几、前八周后八周等因素情况下，性别是一个显著变量，女同学学习成绩更好。
4) /cut1 表示绩点成绩=0的系数（相对于绩点成绩=4.0），以此类推/cut2：2.3； /cut3：3.0； /cut4:3.3； /cut5:3.7.

```
. meoprobit 绩点成绩 sex grade week 星期 大节
note: 大节 omitted because of collinearity

Iteration 0:  log likelihood = -432.18358
Iteration 1:  log likelihood = -424.32647
Iteration 2:  log likelihood = -424.31991
Iteration 3:  log likelihood = -424.31991

Ordered probit regression                                Number of obs      =           380

Log likelihood = -424.31991                               Wald chi2(4)       =          15.61
                                                           Prob > chi2        =          0.0036
```

绩点成绩	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
sex	.4581722	.1284159	3.57	0.000	.2064816	.7098628
grade	-.2137731	.1562472	-1.37	0.171	-.520012	.0924657
week	-.1402222	.2313388	-0.61	0.544	-.5936378	.3131935
星期	-.0810244	.0765938	-1.06	0.290	-.2311456	.0690968
大节	0	(omitted)				
/cut1	-3.287253	.5474979			-4.360329	-2.214177
/cut2	-3.018003	.5169628			-4.031231	-2.004774
/cut3	-2.443779	.489131			-3.402458	-1.485099
/cut4	-.4264538	.4719919			-1.351541	.4986333
/cut5	.2619271	.4726879			-.6645241	1.188378

※ 考虑专业的Rank-ordered logistic regression
 rologit 绩点成绩 grade 星期 大节 sex,
 group(院系) ties(exactm)

※结果解释：在考虑院系专业的情况下，女同学也无法显著地获得高分，只跟年级有关。

院系跟性别显著相关，因此需要考虑IIA假设，
 将院系设为group variable 更简单

当然把院系作为解释变量，以经管学院或者外文系作为reference可以得出检验假设2&3的结论

这里院系较多，进行了简单处理。

```
. rologit 绩点成绩 grade 星期 大节 sex, group(院系) ties(exactm)
```

```
Iteration 0:   log likelihood = -274.52102
Iteration 1:   log likelihood = -271.34526
Iteration 2:   log likelihood = -271.26734
Iteration 3:   log likelihood = -271.26723
Refining estimates:
Iteration 0:   log likelihood = -271.26723
```

```
Rank-ordered logistic regression      Number of obs      =          380
Group variable: 院系                  Number of groups   =           33
```

```
Ties handled via the exactm method      Obs per group:
                                         min =           1
                                         avg =          11.52
                                         max =           34
```

```
LR chi2(4)          =           6.51
Prob > chi2         =          0.1643

Log likelihood = -271.2672
```

绩点成绩	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
grade	-.457965	.2223229	-2.06	0.039	-.8937099	-.0222202
星期	-.0176456	.0637353	-0.28	0.782	-.1425645	.1072733
大节	.0257066	.1181271	0.22	0.828	-.2058182	.2572314
sex	.2145116	.1908052	1.12	0.261	-.1594597	.588483

※更加细化的以4.0为目标的Multinomial logistic regression

mlogit 绩点成绩 week grade sex 大节 星期,
baseoutcome(4.0)

※结果解释：

1) 年级、性别、大节、排课都无法解释有的同学为什么会挂科，只有年级有微弱的显著性（容易理解）

2) 刚刚及格的同學，所有变量都不显著，只有年级有微弱的显著性

3) 中等的同学，所有变量都不显著，只有性别有微弱的显著性

4) 成绩良好的同学，性别变量显著

5) 成绩优秀的同学，性别变量显著

※结论：成绩越差，越跟年级有关，成绩越好越跟性别相关（女生成績更好）

Log likelihood = -415.65031

Pseudo R2

=

0.0383

绩点成绩		Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
0	week	-32.38069	2789.263	-0.01	0.991	-5499.236	5434.474
	grade	2.481189	1.511226	1.64	0.101	-.4807598	5.443137
	sex	-17.8171	3158.016	-0.01	0.995	-6207.416	6171.781
	大节	15.8463	1719.248	0.01	0.993	-3353.818	3385.51
	星期	0	(omitted)				
	_cons	-54.50526	5157.746	-0.01	0.992	-10163.5	10054.49
2.3	week	-17.24242	2602.286	-0.01	0.995	-5117.628	5083.143
	grade	2.196043	1.13449	1.94	0.053	-.0275164	4.419602
	sex	-17.22395	3329.937	-0.01	0.996	-6543.78	6509.332
	大节	.1873423	1.467253	0.13	0.898	-2.688421	3.063105
	星期	0	(omitted)				
	_cons	-7.518809	4.686046	-1.60	0.109	-16.70329	1.665673
3	week	-.237104	.6910406	-0.34	0.732	-1.591519	1.117311
	grade	.2614545	.9005158	0.29	0.772	-1.503524	2.026433
	sex	-1.440956	.8204906	-1.76	0.079	-3.049088	.1671758
	大节	.203641	.423839	0.48	0.631	-.6270682	1.03435
	星期	0	(omitted)				
	_cons	-2.385844	2.188347	-1.09	0.276	-6.674925	1.903238
3.3	week	-.2409701	.290869	-0.83	0.407	-.8110627	.3291226
	grade	.3187792	.4191986	0.76	0.447	-.502835	1.140393
	sex	-1.02235	.2887241	-3.54	0.000	-1.588239	-.4564614
	大节	.1036316	.1804444	0.57	0.566	-.2500329	.4572961
	星期	0	(omitted)				
	_cons	.6155557	.9863005	0.62	0.533	-1.317558	2.548669
3.7	week	-.5462942	.341579	-1.60	0.110	-1.215777	.1231883
	grade	.2167663	.4895865	0.44	0.658	-.7428057	1.176338
	sex	-.9515023	.3458085	-2.75	0.006	-1.629275	-.2737302
	大节	.0254523	.1122766	0.23	0.823	-.202756	.2536604
	星期	0	(omitted)				
	_cons	-.0254523	.1122766	-0.23	0.823	.202756	-.2536604

- ※ 没有其他影响成绩的根本变量：平时是否回答问题、考勤是否缺席、自评和他评相互影响效应、是否交加分作业等

- ※ 没有考虑刘震老师的慕课影响

- ※ Stata官方示例学习简单logit model:

<https://stats.idre.ucla.edu/stata/dae/logistic-regression/>

4

参考文献

站在巨人的肩膀上

"If I have seen further
it is by standing on
the shoulders of
Giants. "

by Isaac Newton in
1675

见网络学堂附件

下节课见

马克思主义学院

龙治铭



清华大学
Tsinghua University