统计学原理(Statistic)

胡华平

西北农林科技大学

经济管理学院数量经济教研室

huhuaping01@hotmail.com

2021-05-16

第五章 相关和回归分析

5.1 变量间关系的度量

5.2 回归分析的基本思想

5.3 OLS方法与参数估计

5.4 假设检验

5.5 拟合优度与残差分析

5.6 回归预测分析

5.7 回归报告解读

5.5 拟合优度与残差分析

拟合优度

残差分析

拟合优度：引子

怎么来判定OLS方法对特定样本数据拟合的好坏？

请大家思考如下几个问题：

- 样本数据不完全落在拟合的直线（或曲线）上，是经常发生的么？
- 怎么来表达或测量这种对样本数据拟合的不完全性？
- 在OLS方法和CLRM假设“双剑合璧”下，对特定样本数据的拟合不是已经证明最好的么（BLUE）？为什么还要说“拟合”有“好坏之分”？

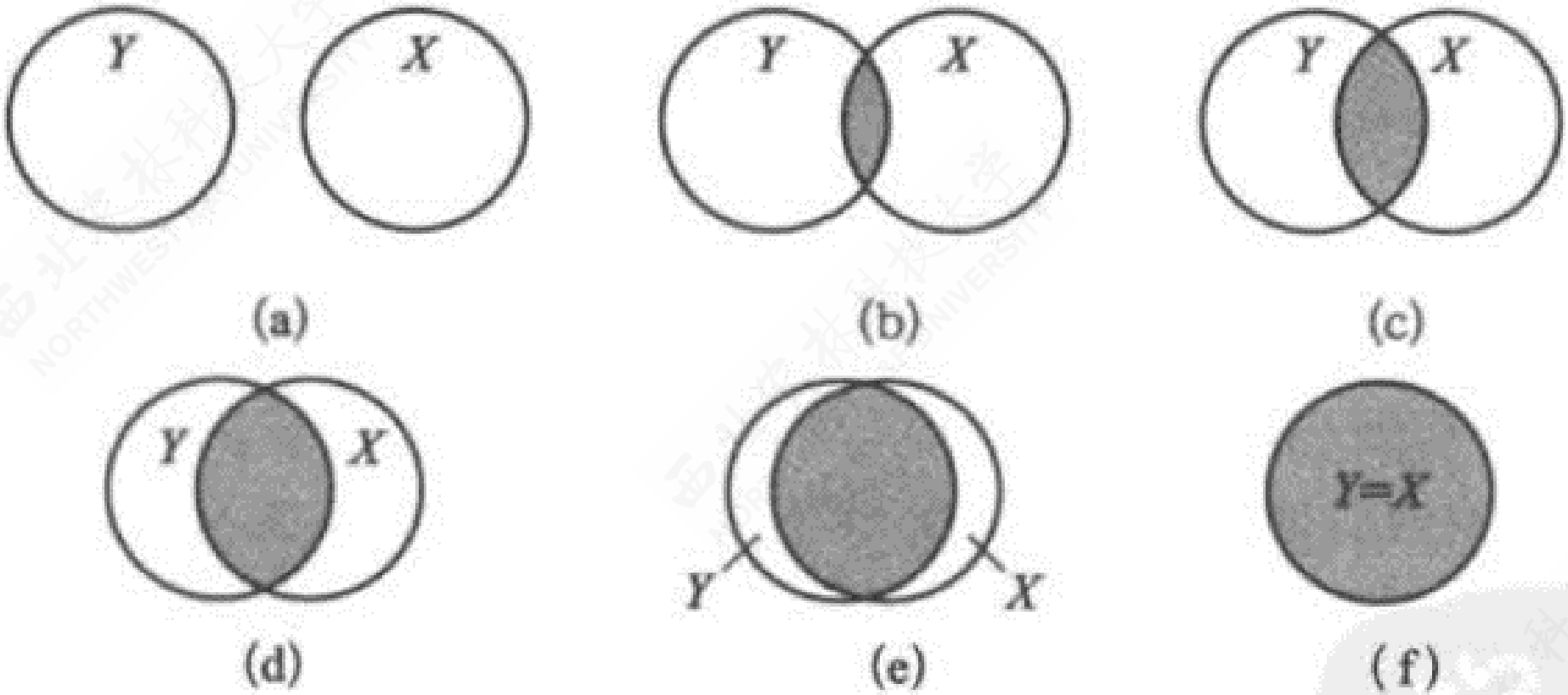
拟合优度：测量指标

拟合优度（Goodness of fit）：度量样本回归线对一组数据拟合优劣水平。

判定系数（coefficient of determination）：一种利用平方和分解，考察样本回归线对数据拟合效果的总度量。

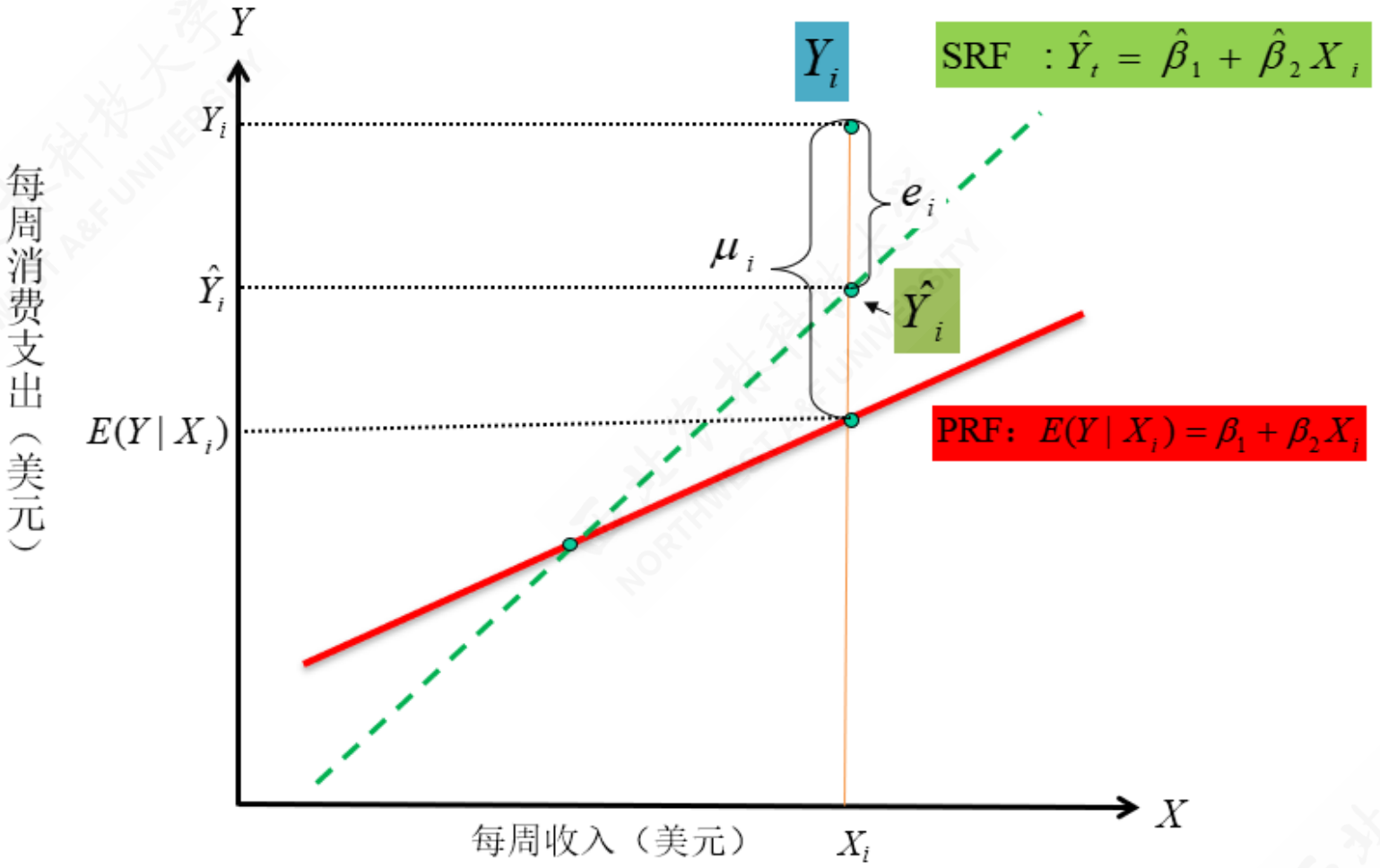
- 一元回归中，一般记为 r^2；
- 多元回归中，一般记为 R^2。

（示例）拟合优度的直观理解



维恩图看拟合优度

拟合优度：测量指标



平方和分解看拟合优度

拟合优度：判定系数

基于前述的方差分解ANOVA表，我们可以用如下公式计算判定系数。

判定系数 r^2计算公式1：

$$r^2 = \frac{ESS}{TSS} = \frac{\sum (\hat{Y}_i - \bar{Y})^2}{\sum (Y_i - \bar{Y})^2}$$

判定系数 r^2计算公式2：

$$r^2 = 1 - \frac{RSS}{TSS} = 1 - \frac{\sum e_i^2}{\sum (Y_i - \bar{Y})^2}$$

拟合优度：判定系数

判定系数 r^2计算公式3：

$$r^2 = \frac{ESS}{TSS} = \frac{\sum \hat{y}_i^2}{\sum y_i^2} = \frac{\sum (\hat{\beta}_2 x_i)^2}{\sum y_i^2} = \hat{\beta}_2^2 \frac{\sum x_i^2}{\sum y_i^2} = \hat{\beta}_2^2 \frac{S_{X_i}^2}{S_{Y_i}^2}$$

判定系数 r^2计算公式4：

$$r^2 = \hat{\beta}_2^2 \cdot \frac{\sum x_i^2}{\sum y_i^2} = \left(\frac{\sum x_i y_i}{\sum x_i^2}\right)^2 \cdot \left(\frac{\sum x_i^2}{\sum y_i^2}\right) = \frac{(\sum x_i y_i)^2}{\sum x_i^2 \sum y_i^2}$$

课堂讨论：

- 讨论1： r^2是一个非负量。为什么？
- 讨论2： $0 \le r^2 \le 1$，两个端值分别意味什么？

拟合优度：判定系数VS简单相关系数

判定系数与简单相关系数有什么区别与联系？

总体相关系数：是变量 X_i与变量 Y_i总体相关关系的参数，一般记为 ρ。

$$\rho = \frac{Cov(X,Y)}{\sqrt{Var(X_i)Var(Y_i)}} = \frac{E(X_i - EX)(Y_i - EY)}{\sqrt{E(X_i - EX)^2 E(Y_i - EY)^2}}$$

样本相关系数：是从总体中抽取随机样本，获得变量 X_i与变量 Y_i样本相关关系的统计量度量，一般记为 r。

$$r = \frac{S_{XY}^2}{S_X * S_Y} = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum (X_i - \bar{X})^2 \sum (Y_i - \bar{Y})^2}} = \frac{\sum x_i y_i}{\sqrt{\sum x_i^2 \sum y_i^2}}$$

拟合优度：判定系数VS简单相关系数

判定系数和简单相关系数的联系：

- 在一元回归中，判定系数 r^2等于样本相关系数 r的平方。

判定系数和简单相关系数的区别：

- 判定系数 r^2表明因变量变异由解释变量所解释的比例，而相关系数 r只能表明变量间的线性关联强度。
- 在多元回归中，这种区别会更加凸显！因为那时的相关系数r出现了偏相关的情形(交互关联)！

（案例）计算相关系数和判定系数

对于“教育程度与时均工资案例”，根据FF-ff计算表和方差分解ANOVA表，可以分别计算得到样本相关系数和模型判定系数。

样本相关系数 r:

$$r = \frac{S_{XY}^2}{S_X * S_Y} = \frac{10.9821}{3.8944 * 2.9597} = 0.9528$$

回归方程的判定系数 r^2:

$$r^2 = 1 - \frac{RSS}{TSS} = 1 - \frac{9.693}{105.1183} = 0.9078$$

二者关系

拟合优度：小结与思考

内容小结：

- 即使采用OLS方法，它对样本数据的拟合也是不完全的。意味着实际数据点在样本回归线附近，而不是在样本回归线上。我们可以把样本点行为的“变异”，划分为“回归”能解释的部分和“随机”的部分。并进一步获得变异平方和的分解。
- 判定系数 R^2是对OLS拟合程度的测量，它使用了变异平方和分解的思想。在一元线性回归（含截距）中，判定系数与相关系数存在如下关系 $R^2 = r^2_{(X_i,Y_i)}$。注意，在多元回归中则不存在这种关系。

问题思考：

- OLS方法的参数估计量，在CLRM假设满足情况下，就是最优线性无偏估计量（BLUE），为什么还要用判定系数来判断“拟合好还是不好？”。对此，你的回答是什么？
- 还有没有其他指标，来反映估计方法对样本数据的拟合好坏程度？请说出一两个。

残差分析：定义和作用

残差(residual)：是因变量的观测值与根据估计的回归方程求出的估计值之差，用 e_i表示。

$$e_i = Y_i - \hat{Y}_i$$

对模型的残差进行分析，主要目的包括：

- 反映用估计的回归方程去预测而引起的误差。
- 可用于确定有关随机干扰项 μ_i的假定是否成立。
- 用于检测有影响的观测值。

残差分析：皮尔逊标准化残差

标准化残差(standardized residual)：是对残差进行某种标准化变换。具体计算方法有皮尔逊标准化残差和学生化标准残差两种。

最常用的皮尔逊标准化残差（Pearson residual/internally studentized residuals）的计算公式如下：

$$e_{i,sd}^{*}=\frac{e_i}{s_{e_i}}=\frac{(Y_i-\hat{Y}_i)}{\sqrt{\frac{\sum(e_i-\bar{e})^2}{n-1}}}$$

残差分析：学生化标准化残差

学生化标准残差（Studentized Residuals/externally studentized residual/deleted Studentized residual/semi-studentized residuals/jackknifed residuals），是对残差的另一种特殊标准化变换（例如考虑到了X的影响力）。

残差分析：学生化标准化残差

学生化标准残差的计算公式有两个*：

$$e^*_{i,st} = \frac{e_i}{\sqrt{MSE_{(i)}(1-h_{ii})}} \quad \text{(eq.01)}$$

$$e^*_{i,st} = e^*_{i,sd}\left(\frac{n-m-2}{n-m-1-e^{*2}_{i,sd}}\right)^2 \quad \text{(eq.02)}$$

> 其中： $MSE_{(i)}$是指删除第 i个观测值进行建模的均方误差（MSE）； h_{ii}指删除第 i个观测值进行建模的第 i个影响权重（leverage）。 $m=k-1$为回归元个数。

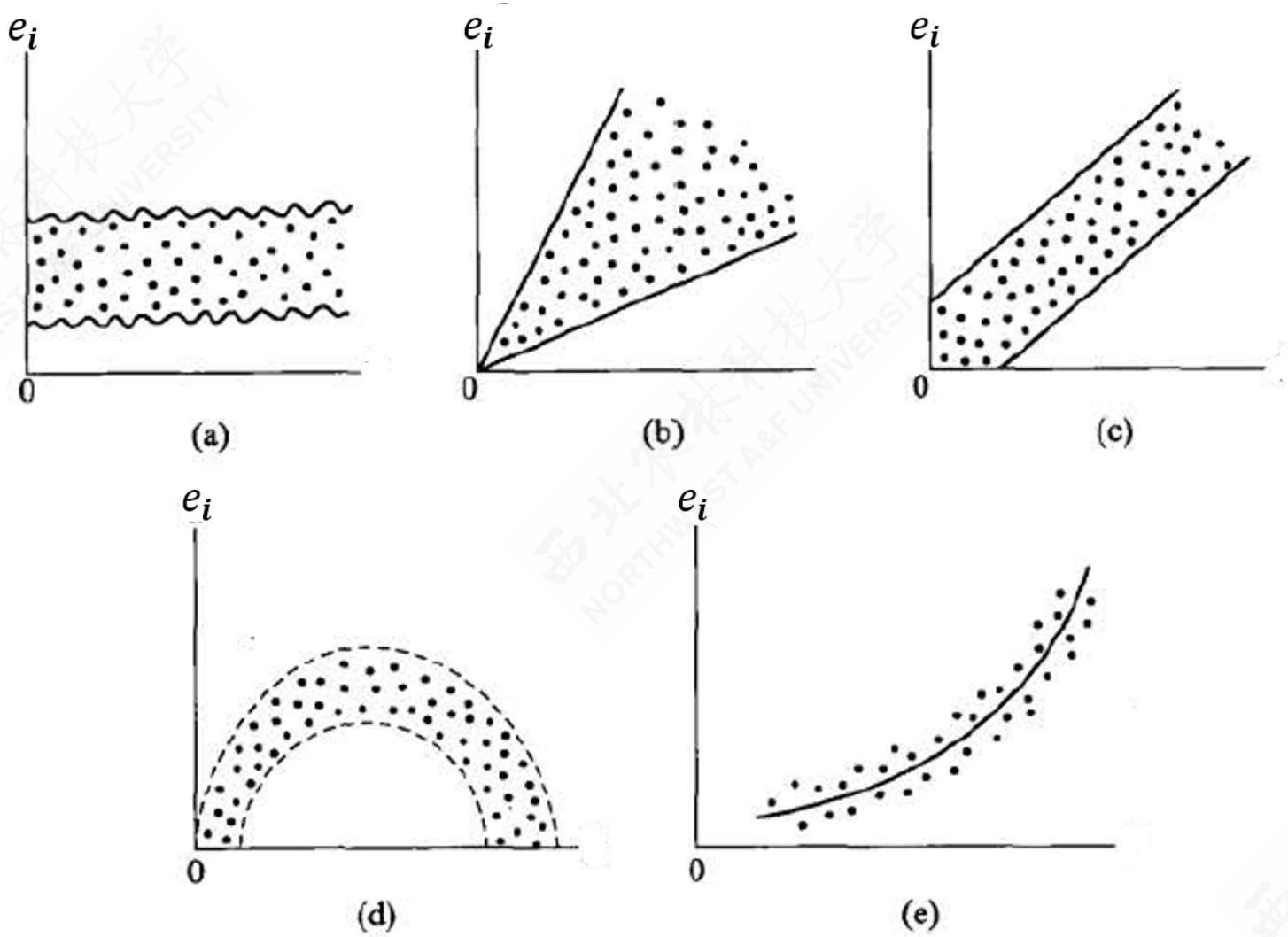
说明： 1）学生化残差的第一个计算公式计算起来比较麻烦和复杂。需要分别进行(n-1)次线性回归，然后依次计算相关$MSE_{(i)}$和 h_ii。2）学生化残差的第二个计算公式相对简单，只需要利用原来的回归模型及其标准化残差 $e^*_{i,sd}$。

残差分析：残差图

残差图(residual plot)：用于呈现残差数据 e_i的分布情况的统计图图形，主要包括：

- 关于 X_i的残差散点图。
- 关于 Y_i的残差散点图（或者关于 $\hat{Y}_i$）。
- 关于样本序号的残差散点图或标准化残差散点图。

（示例）残差图的模拟演示



（案例）皮尔逊标准化残差

obs	X_i	Y_i	$\hat{Y}_i$	e_i	$e^*_{i,sd}$
1	6	4.5	4.3	0.1266	0.1408
2	7	5.8	5.0	0.7158	0.7964
3	8	6.0	5.8	0.2004	0.2230
4	9	7.3	6.5	0.8293	0.9227
5	10	7.3	7.2	0.0917	0.1020
6	11	6.6	8.0	-1.3662	-1.5201
7	12	7.8	8.7	-0.8565	-0.9530
8	13	7.8	9.4	-1.5637	-1.7399
9	14	11.0	10.1	0.8994	1.0007
10	15	10.7	10.8	-0.1732	-0.1927
11	16	10.8	11.6	-0.7350	-0.8178
12	17	13.6	12.3	1.3198	1.4685
13	18	13.5	13.0	0.5117	0.5694
sum	156	112.8	112.8	0.0000	-0.0000

- 根据样本回归方程，可以计算得到 Y_i的回归拟合值 $\hat{Y}_i$，以及回归残差 e_i。

$$\hat{Y}_i = \hat{\beta}_1 + \hat{\beta}_2 X_i$$
$$e_i = Y_i - \hat{Y}_i$$

- 进一步地计算得到皮尔逊标准化残差 $e^*_{i,sd}$：

$$e^*_{i,sd} = \frac{e_i}{s_{e_i}} = \frac{(Y_i - \hat{Y}_i)}{\sqrt{\frac{\sum (e_i - \bar{e})^2}{n-1}}}$$

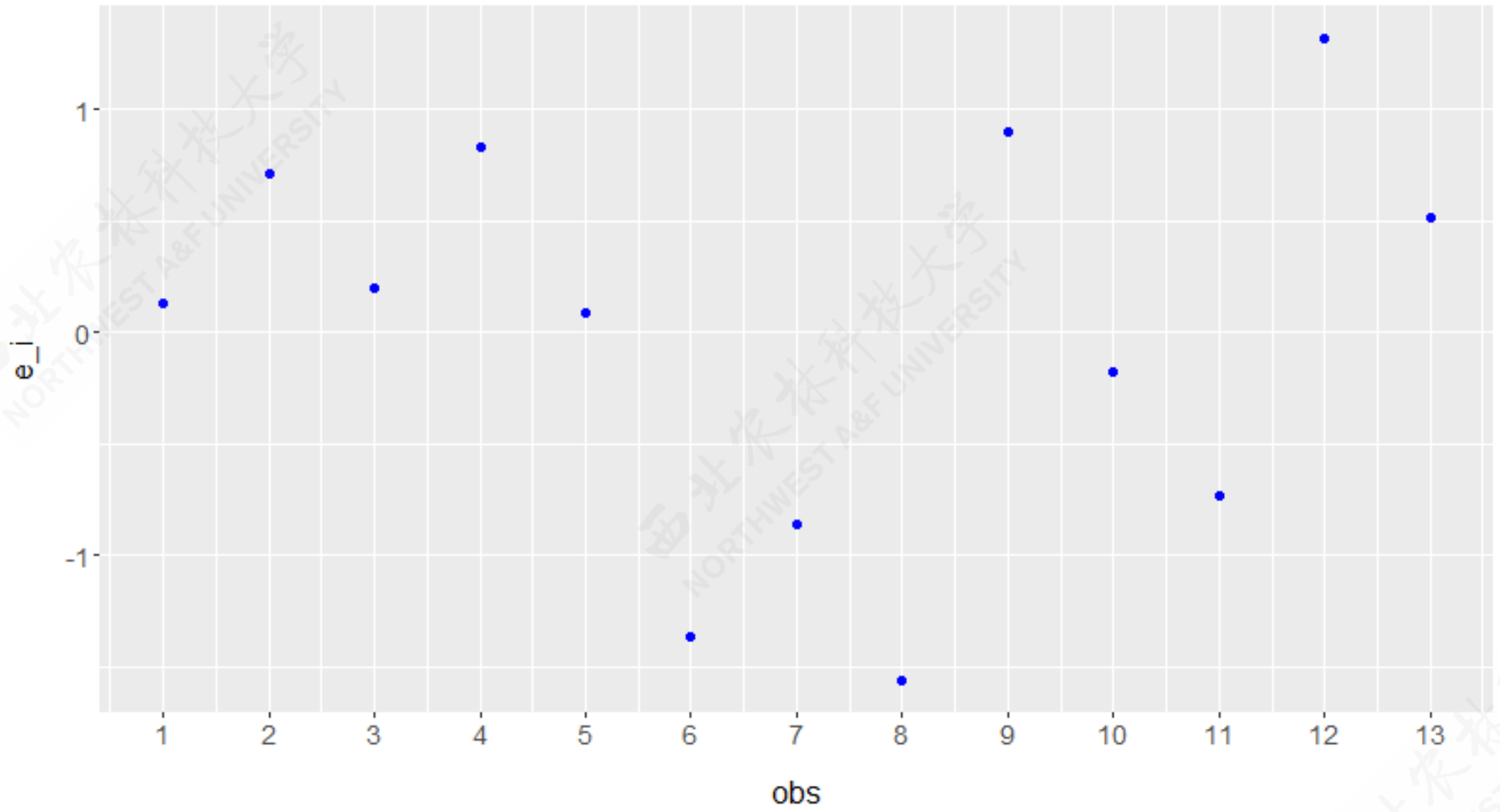
（案例）学生化标准残差

obs	X_i	Y_i	$\hat{Y}_i$	e_i	$e^*_{i,sd}$	$e^*_{i,st}$
1	6	4.5	4.3	0.1266	0.1408	0.1511
2	7	5.8	5.0	0.7158	0.7964	0.8493
3	8	6.0	5.8	0.2004	0.2230	0.2233
4	9	7.3	6.5	0.8293	0.9227	0.9402
5	10	7.3	7.2	0.0917	0.1020	0.0982
6	11	6.6	8.0	-1.3662	-1.5201	-1.6297
7	12	7.8	8.7	-0.8565	-0.9530	-0.9451
8	13	7.8	9.4	-1.5637	-1.7399	-1.9472
9	14	11.0	10.1	0.8994	1.0007	1.0103
10	15	10.7	10.8	-0.1732	-0.1927	-0.1885
11	16	10.8	11.6	-0.7350	-0.8178	-0.8456
12	17	13.6	12.3	1.3198	1.4685	1.7221
13	18	13.5	13.0	0.5117	0.5694	0.6220
sum	156	112.8	112.8	0.0000	-0.0000	0.0601

- 根据样本回归方程，可以计算得到Y_i的回归拟合值 $\hat{Y}_i$，以及回归残差 e_i，以及前述皮尔逊标准化残差 $e^*_{i,sd}$。
- 进而可以使用如下公式计算得到学生化标准残差 $e^*_{i,st}$：

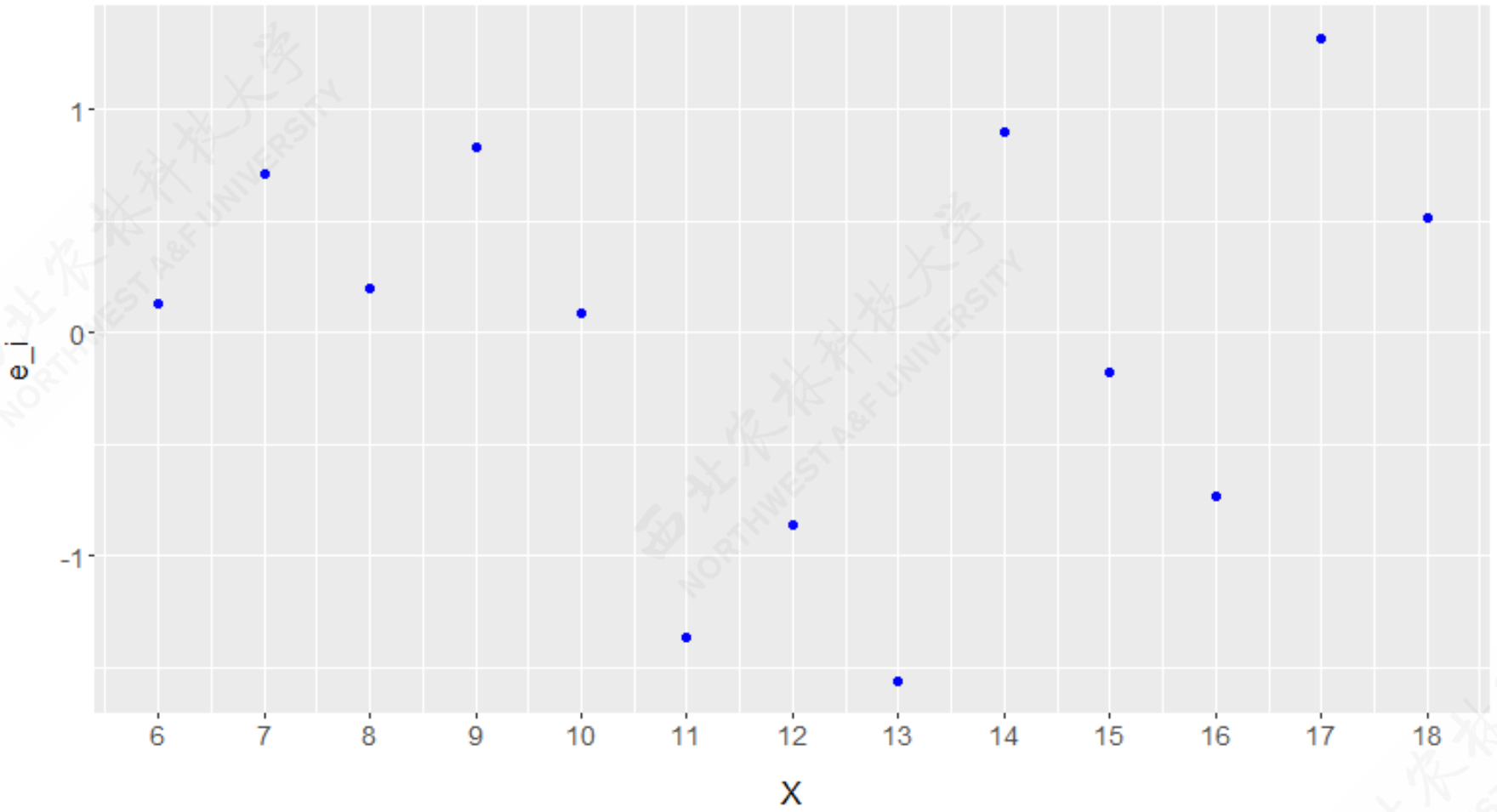
$$e^*_{i,st} = e^*_{i,sd}\left(\frac{n-m-2}{n-m-1-e^{*2}_{i,sd}}\right)^2$$

（案例）皮尔逊标准化残差散点图1



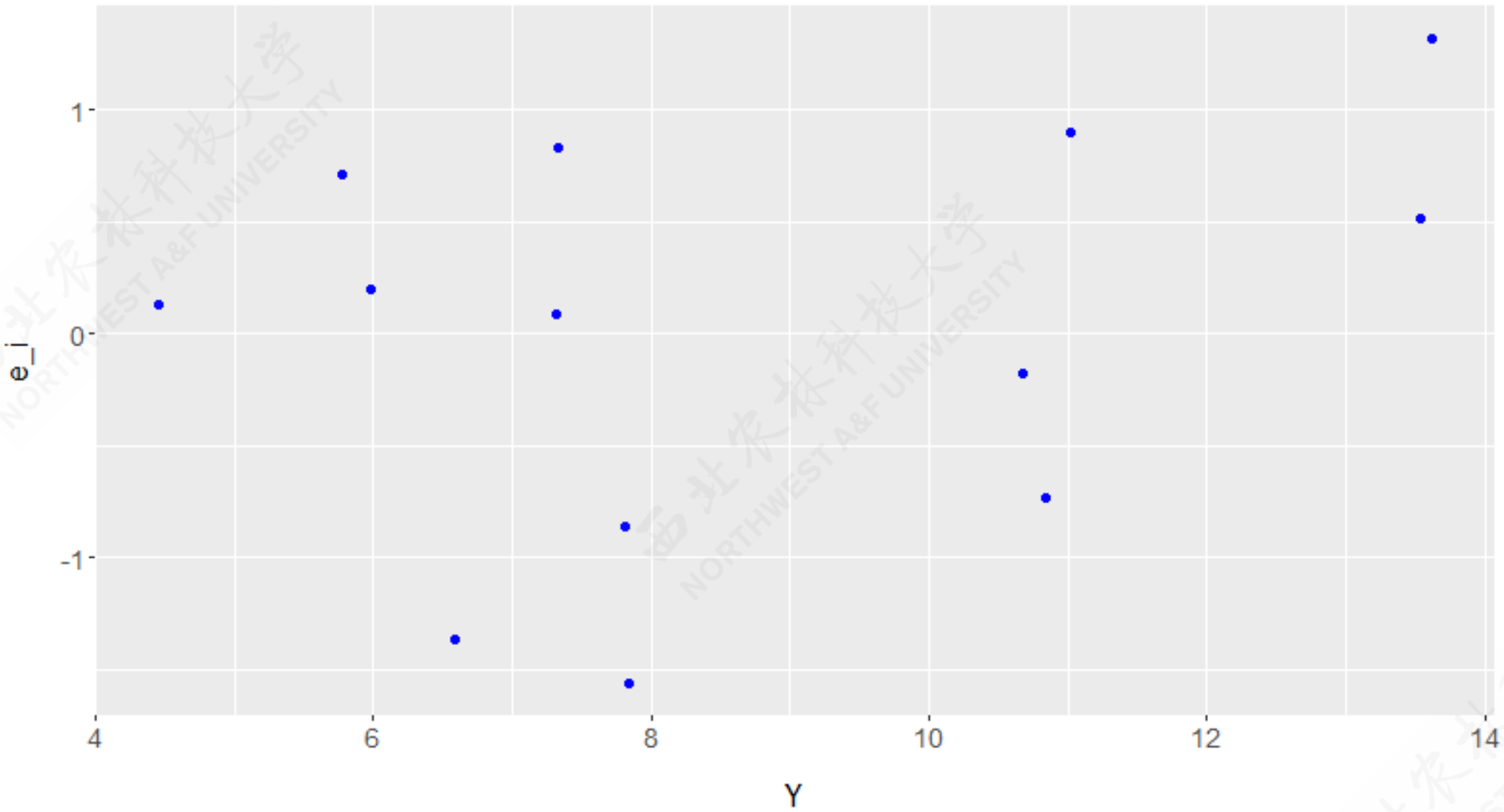
残差对样本编号作图

（案例）皮尔逊标准化残差散点图2



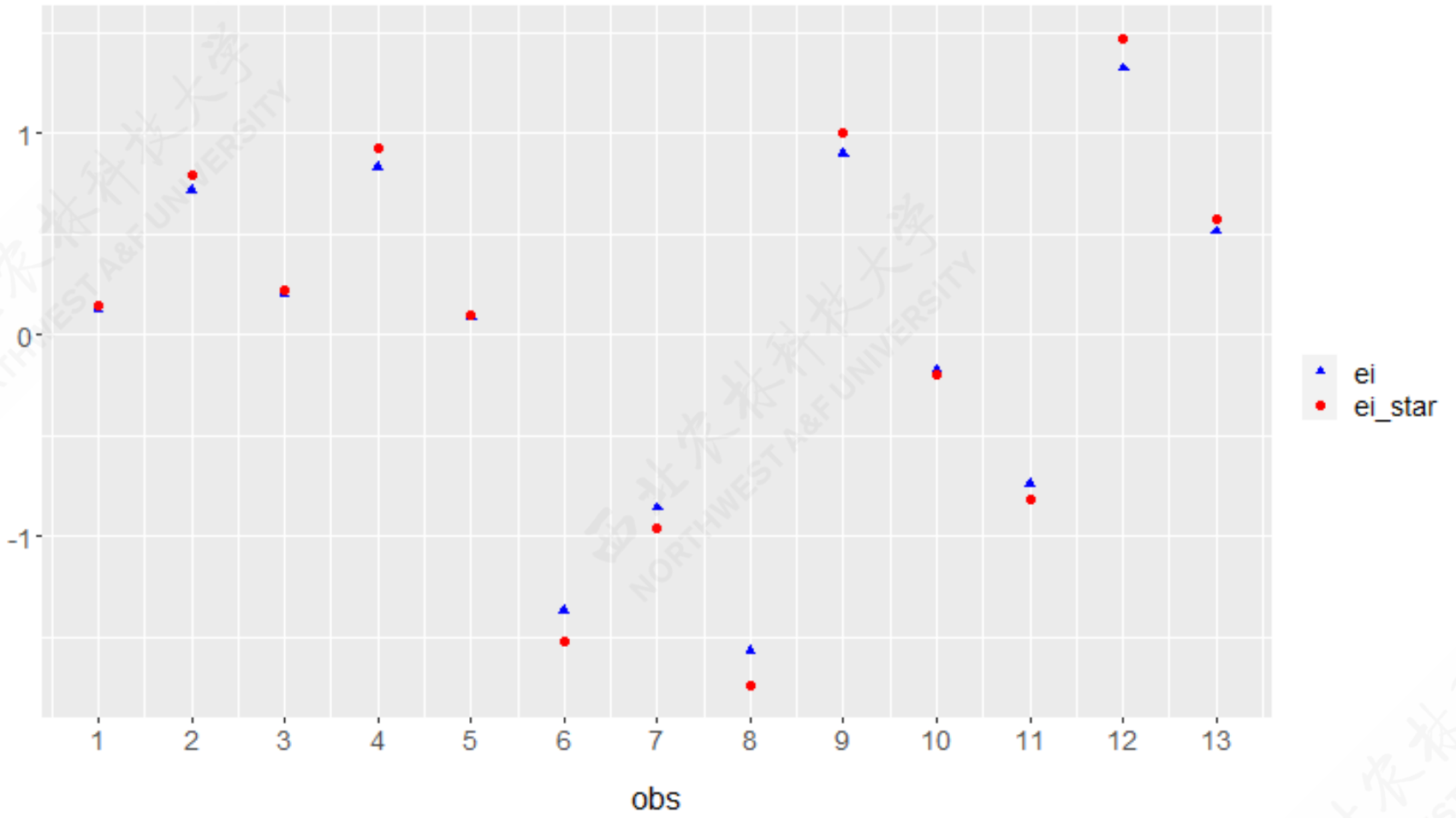
残差对自变量X作图

（案例）皮尔逊标准化残差散点图3



残差对因变量Y作图

（案例）皮尔逊标准化残差散点图4



标准化残差对样本编号作图

本节结束