社会统计学及SPSS软件应用

Instructor:王荣欣

Email: rxwang@qq.com

周一3-4节、单周周四3-4节，3A106-2

2020年10月19日

CONTENTS

CONTENTS

判定系数R^2

1. R^2：判定系数，是拟合优度（goodness of fit）的度量指标。（教材第106-107页）
   - R^2实际上是自变量与因变量的Pearson相关系数的平方。
2. adjusted R^2 （教材第108页）
   - 拟合优度会随着模型所包含自变量个数的增加而不断增加，所以要进行调整。

因变量的总变异（SST）=
被自变量所解释的变异（SSR）+未被自变量解释的变异（SSE）

$$\text{SST}=\sum_{i=1}^{i=n}(Y_i-\bar{Y})$$

$$\text{SSR}=\sum_{i=1}^{i=n}(\hat{Y}_i-\bar{Y})$$

$$\text{SSE}=\sum_{i=1}^{i=n}(Y_i-\hat{Y}_i)$$

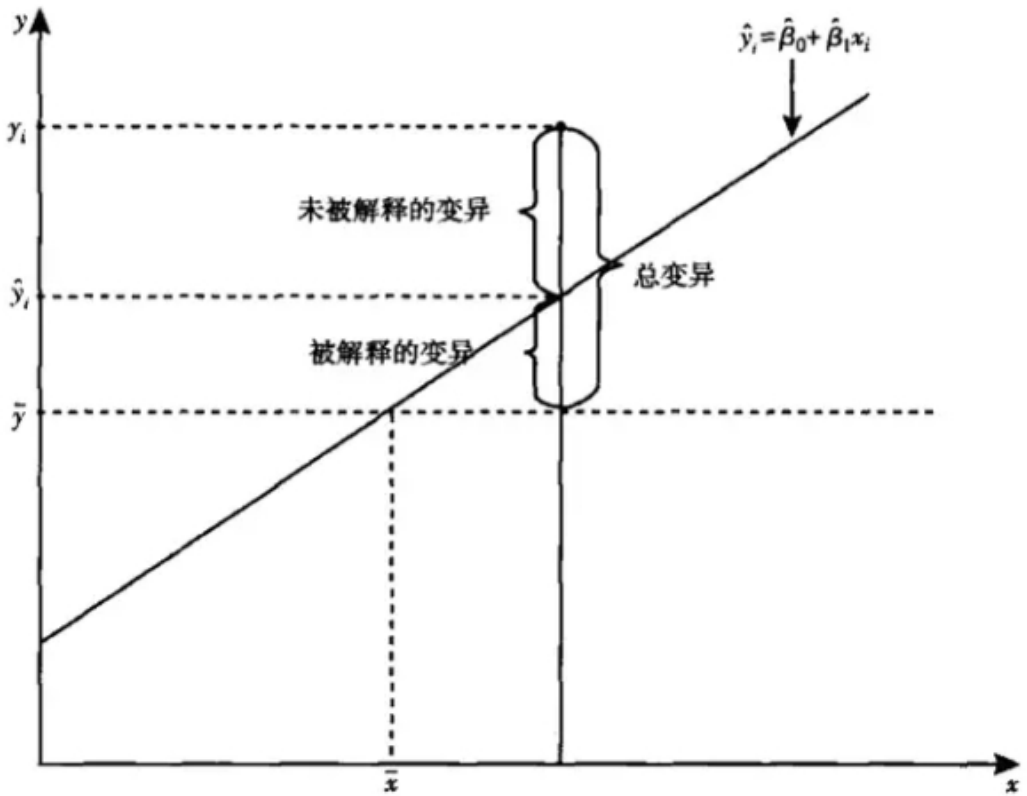
$$R^2=\frac{SSR}{SST}=1-\frac{SSE}{SST}$$

SST=SSR+SSE

1 SST=$\sum_{i=1}^{i=n}(Y_i - \bar{Y}) = 102.4$

2 SSR=$\sum_{i=1}^{i=n}(\hat{Y}_i - \bar{Y}) = 54.8$

3 SSE=$\sum_{i=1}^{i=n}(Y_i - \hat{Y}_i) = 47.6$



```
. reg edu fedu
```

Source	SS	df	MS	Number of obs	=	10
				F(1, 8)	=	9.23
Model	54.8496124 SSR	1	54.8496124	Prob > F	=	0.0161
Residual	47.5503876 SSE	8	5.94379845	R-squared	=	0.5356
				Adj R-squared	=	0.4776
Total	102.4 SST	9	11.3777778	Root MSE	=	2.438

edu	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
fedu	.6873385	.2262642	3.04	0.016	.1655722	1.209105
_cons	3.382429	1.822797	1.86	0.101	-.8209482	7.585806

1. 如果没有其他参考信息，只知道某个属性在一个群体中的平均值，若该属性在这个群体中呈正态分布，那么，我们在猜测一个个体的属性取值时，最好的猜测是该属性在群体中的平均值。
2. 猜测个体属性的时候，猜平均值得出的误差最小，因为大部分值都集中在平均值周围，68%的个案在平均值左右摇摆。
3. 对于正态分布，大约有68%的数据位于平均值附近±1 个标准差的范围内。

回归结果分析

如果用平均值（父亲平均受教育年限）猜，那么残差平方和就是102.4。

$$\sum_{i=1}^{i=n}(Y_i - \bar{Y}) = 102.4$$

- 如果根据回归模型猜，误差平方和是47.6。

$$\sum_{i=1}^{i=n}(Y_i - \hat{Y}_i) = 47.6$$

- 用回归模型猜所产生的残差的平方和比用平均值猜产生的残差平方和减少了54.8。

做最小二乘法回归分析，就是计算出一个回归系数，使用这个系数，根据自变量的值猜测因变量的值，能最大程度地减少根据平均值预测产生的误差。

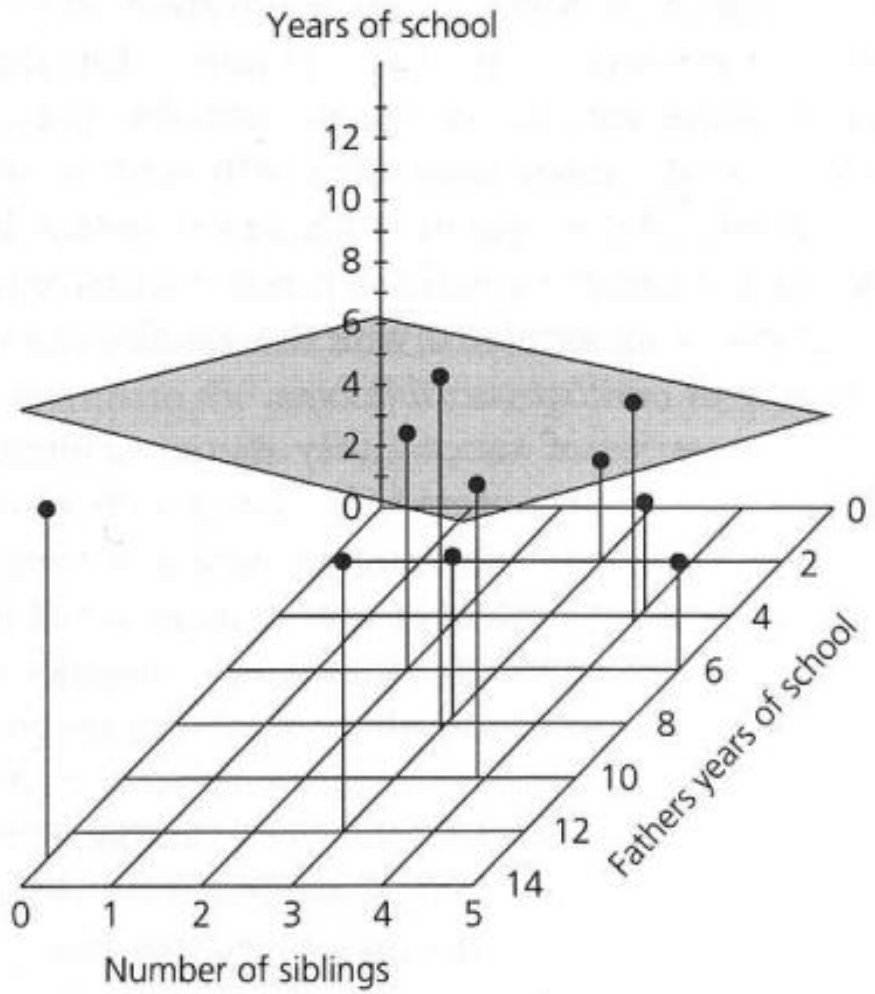
1. 减少的残差平方和54.8占102.4的53.6%，就是R^2等于0.536。
2. 减少的残差平方和54.8意味着我们猜测的准确率提高了53.6%。
3. 判定系数（R^2）显示预测的准确度提高了多少。

MSE (MEAN SQUARE ERROR)（教材第108-109页）

$$MSR = \frac{SSR}{degree\ of\ freedom} = \frac{SSR}{1} = SSR = 54.8$$

$$MSE = \frac{SSE}{n-2} = \frac{47.6}{10-2} = 5.94 = 2.438^2$$

Father's Years of Schooling	Respondent's Years of Schooling	Number of Siblings
2	4	3
12	10	3
4	8	4
13	13	0
6	9	2
6	4	5
8	13	3
4	6	4
8	6	3
10	11	4


Years of school
12
10
8
6
4
2
0
0
2
4
6
8
10
12
14
Fathers years of school
0
1
2
3
4
5
Number of siblings

```
. reg edu fedu siblings
```

Source	SS	df	MS	Number of obs	=	10
				F(2, 7)	=	4.96
Model	60.0144184	2	30.0072092	Prob > F	=	0.0456
Residual	42.3855816	7	6.05508309	R-squared	=	0.5861
				Adj R-squared	=	0.4678
Total	102.4	9	11.3777778	Root MSE	=	2.4607

edu	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
fedu	.5644457	.2643104	2.14	0.070	-.0605491	1.18944
siblings	-.6398143	.6927667	-0.92	0.386	-2.277947	.9983186
_cons	6.262971	3.621132	1.73	0.127	-2.299645	14.82559

1 一元线性回归模型：$Y = 3.38 + 0.687X$

2 多元线性回归模型：$Y = 6.26 + 0.564X_1 - 0.640X_2$

(1) A person who had no siblings and whose father had no education would be expected to have 6.26 years of schooling;

(2) Among those with the same number of siblings, whose whose father's schooling differed by one year would be expected to differ in their own schooling by 0.564 years.

(3) Among those whose fathers had the same amount of education, the expected cost of each additional sibling would be 0.64 years.

- poorly educated fathers tend to have more children, and those from large families tend to go less far in school.

增加一个自变量：兄弟姐妹数量

1. 父亲受教育年限对家庭兄弟姐妹数量产生负影响。
2. 家庭兄弟姐妹数量对个人受教育年限产生负影响。
3. 家庭兄弟姐妹数量作为中介变量，在控制该变量后，父亲受教育年限的估计系数会下降。
4. 在未控制的情况下，兄弟姐妹数量变量的影响并未消失，而是被包含在残差项中。此时存在遗漏变量偏差（omitted variable bias）。

标准化系数（STANDARDIZED COEFFICIENT）

1. fedu估计系数是0.56，sibling的估计系数是-0.64，能否说兄弟姐妹数量对个人教育水平的影响超过父亲教育年限变量？
2. 兄弟姐妹数量与父亲教育水平的单位不同，因此需对自变量进行标准化转换（将因变量和所有自变量都减去均值后再除以其标准差，转换后所有变量的均值都为0，标准差都为1），再进行回归和估计系数。
3. 标准化后的回归系数表示自变量X每变化一单位标准差，因变量Y会变化b个标准差。
4. Stata命令 reg y x_1 x_2 x_3, beta

```
. reg edu fedu siblings, beta
```

Source	SS	df	MS	Number of obs	=	10
				F(2, 7)	=	4.96
Model	60.0144184	2	30.0072092	Prob > F	=	0.0456
Residual	42.3855816	7	6.05508309	R-squared	=	0.5861
				Adj R-squared	=	0.4678
Total	102.4	9	11.3777778	Root MSE	=	2.4607

edu	Coef.	Std. Err.	t	P>|t|	Beta
fedu	.5644457	.2643104	2.14	0.070	.6010191
siblings	-.6398143	.6927667	-0.92	0.386	-.2599246
_cons	6.262971	3.621132	1.73	0.127	.

参考文献

1 李连江，2017，《戏说统计：文科生的量化方法》，北京：中国政法大学出版社。
2 谢宇，2013，《回归分析》（修订版）第17章，北京：社会科学文献出版社。