第1章　从线性回归到模型评价

2025年7月

本讲关键问题

（1）OLS有哪些重要性质？

（2）什么是预测推断？什么是因果推断？

（3）如何评价估计量的“好坏”？

（4）拟合优度R^2与F统计量能否选出合适的模型？

（5）什么是样本内MSE与样本外MSE?什么是过度拟合？

（6）什么是交错鉴定法？

关键词：预测、因果推断、MSE、MSFE、过度拟合、交错鉴定法

一、回顾：线性回归与OLS

回顾多元线性回归的经典理论：OLS的BLUE性质、拟合优度R^2、调整后的拟合优度$\bar{R}^2$、回归的整体显著性检验等。思考如下问题：

（1）当介绍变量个数p逐渐增大时，OLS是否还能保持原来的优良性质？

（2）如何判别模型是否需要包含某个“看似”不太重要的解释变量？

（3）模型中处理包含某一解释变量的线性项，是否还要包含它的平方项、高次项及交互项？

（4）如何评判OLS方法的预测能力

线性回归模型

$$Y_i = X_i'\beta + U_i$$

系数估计量为

$$\hat{\beta} = (\sum_{i=1}^{n} X_i X_i')^{-1} (\sum_{i=1}^{n} X_i Y_i)$$

一、回顾：线性回归与OLS

在度量估计量的优劣时，可用均方误差（MSE, mean squared error）判断预测值$\hat{Y}_i$是否与Y_i足够接近，这是一种有监督的机器学习方法（supervised learning），MSE的表达式为：

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 = \frac{1}{n}\sum_{i=1}^{n}(Y_i - X_i\hat{\beta})^2 = \frac{1}{n}\sum_{i=1}^{n}\hat{U}_i^2. \quad (1.1)$$

(1.1)式也称为“样本内MSE”（或者训练集上的MSE）。

实际上，当我们评估一种估计方法的优劣时，关心的是它对新的（样本外的）数据的预测能力。设(X_{n+1}, Y_{n+1})是一组在原本样本之外的新观测值，样本外MSE也称为均方预测误差（MSFE, mean squared forecast error），表达式为：

$$MSFE = E(Y_{n+1} - X_{n+1}\hat{\beta})^2. \quad (1.2)$$

(1.2)式度量OLS估计量的预测能力。

一、回顾：线性回归与OLS

MSFE是否等于MSE?

$$MSFE = E(Y_{n+1} - X_{n+1}\hat{\beta})^2 = E(X_{n+1}^{'}(\beta - \hat{\beta}) + U_{n+1})^2$$
$$= E(X_{n+1}^{'}(\beta - \hat{\beta}))^2 + \sigma^2.$$

其中，

$$E(X_{n+1}^{'}(\beta - \hat{\beta}))^2 = E((\beta - \hat{\beta})^{'}X_{n+1}X_{n+1}^{'}(\beta - \hat{\beta}))$$
$$= tr[E((\beta - \hat{\beta})^{'}X_{n+1}X_{n+1}^{'}(\beta - \hat{\beta}))] = tr[E(X_{n+1}X_{n+1}^{'}(\beta - \hat{\beta})(\beta - \hat{\beta})^{'})]$$
$$= tr[E(X_{n+1}X_{n+1}^{'})E((\beta - \hat{\beta})(\beta - \hat{\beta})^{'})] = tr[E(X_{n+1}X_{n+1}^{'})Var(\hat{\beta})]$$
$$= E[tr(X_{n+1}X_{n+1}^{'}Var(\hat{\beta}))] = E[tr(X_{n+1}^{'}Var(\hat{\beta})X_{n+1})] = E[X_{n+1}^{'}Var(\hat{\beta})X_{n+1}].$$

思考题：MSFE总是大于MSE，这种说法是否正确？

二、预测推断

Infer: to form an opinion or guess that something is true because of the information that you have.

—— Cambirdge Dictionary

Y是想要解释的变量（目标变量、结果变量、因变量），X是一组数目众多，用于解释Y的变量。**预测推断的基本问题是：如何利用X预测因变量Y？**一般地，假设

$$Y = m(X) + U.$$

U是扰动项，它的条件期望等于零：$E(U|X) = 0$。

预测的目的是估计$m(x)$这一函数形式。记$m(x)$的估计量是$\hat{m}(x)$。它的含义是：如果我们知道某一个体（不一定在可观测的样本中）的$X = x$，那么我们预测它的Y等于$\hat{m}(x)$。

线性回归是预测的最简单方法之一。假设$m(x) = x'\beta$，预测关心的是$x'\beta$整体的估计值，而非β中某一分量的估计值。

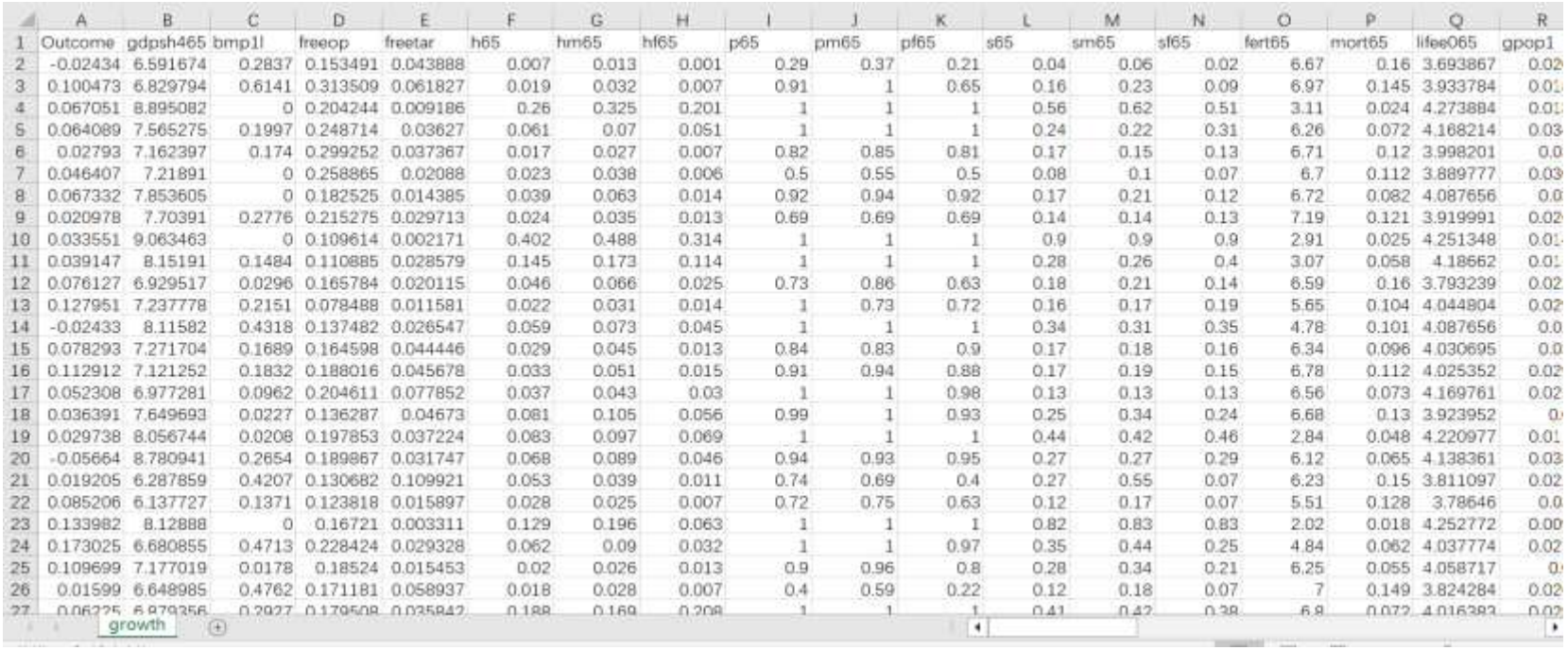
二、预测推断

例1：一国经济增长率的预测growth. dta

Y_i：第i个国家某一年的人均GDP增长率；

X_i：第i个国家的初始人均GDP、国家的人力资本水平、经济制度、法制、贸易开放度、政治稳定程度等。

参考文献： Robert Barro and Jong-Wha Lee, 2013, A new data set of educational attainment in the world, 1950-2010, Journal of Development Economics.



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
1	Outcome	gdpsh465	bmp1l	freeop	freetar	h65	hm65	hf65	p65	pm65	pf65	s65	sm65	sf65	fert65	mort65	lifee065	gpop1
2	-0.02434	6.591674	0.2837	0.153491	0.043888	0.007	0.013	0.001	0.29	0.37	0.21	0.04	0.06	0.02	6.67	0.16	3.693867	0.02
3	0.100473	6.829794	0.6141	0.313509	0.061827	0.019	0.032	0.007	0.91	1	0.65	0.16	0.23	0.09	6.97	0.145	3.933784	0.01
4	0.067051	8.895082	0	0.204244	0.009186	0.26	0.325	0.201	1	1	1	0.56	0.62	0.51	3.11	0.024	4.273884	0.01
5	0.064089	7.565275	0.1997	0.248714	0.03627	0.061	0.07	0.051	1	1	1	0.24	0.22	0.31	6.26	0.072	4.168214	0.03
6	0.02793	7.162397	0.174	0.299252	0.037367	0.017	0.027	0.007	0.82	0.85	0.81	0.17	0.15	0.13	6.71	0.12	3.998201	0.0
7	0.046407	7.21891	0	0.258865	0.02088	0.023	0.038	0.006	0.5	0.55	0.5	0.08	0.1	0.07	6.7	0.112	3.889777	0.03
8	0.067332	7.853605	0	0.182525	0.014385	0.039	0.063	0.014	0.92	0.94	0.92	0.17	0.21	0.12	6.72	0.082	4.087656	0.0
9	0.020978	7.70391	0.2776	0.215275	0.029713	0.024	0.035	0.013	0.69	0.69	0.69	0.14	0.14	0.13	7.19	0.121	3.919991	0.02
10	0.033551	9.063463	0	0.109614	0.002171	0.402	0.488	0.314	1	1	1	0.9	0.9	0.9	2.91	0.025	4.251348	0.01
11	0.039147	8.15191	0.1484	0.110885	0.028579	0.145	0.173	0.114	1	1	1	0.28	0.26	0.4	3.07	0.058	4.18662	0.01
12	0.076127	6.929517	0.0296	0.165784	0.020115	0.046	0.066	0.025	0.73	0.86	0.63	0.18	0.21	0.14	6.59	0.16	3.793239	0.02
13	0.127951	7.237778	0.2151	0.078488	0.011581	0.022	0.031	0.014	1	0.73	0.72	0.16	0.17	0.19	5.65	0.104	4.044804	0.02
14	-0.02433	8.11582	0.4318	0.137482	0.026547	0.059	0.073	0.045	1	1	1	0.34	0.31	0.35	4.78	0.101	4.087656	0.0
15	0.078293	7.271704	0.1689	0.164598	0.044446	0.029	0.045	0.013	0.84	0.83	0.9	0.17	0.18	0.16	6.34	0.096	4.030695	0.0
16	0.112912	7.121252	0.1832	0.188016	0.045678	0.033	0.051	0.015	0.91	0.94	0.88	0.17	0.19	0.15	6.78	0.112	4.025352	0.02
17	0.052308	6.977281	0.0962	0.204611	0.077852	0.037	0.043	0.03	1	1	0.98	0.13	0.13	0.13	6.56	0.073	4.169761	0.02
18	0.036391	7.649693	0.0227	0.136287	0.04673	0.081	0.105	0.056	0.99	1	0.93	0.25	0.34	0.24	6.68	0.13	3.923952	0.0
19	0.029738	8.056744	0.0208	0.197853	0.037224	0.083	0.097	0.069	1	1	1	0.44	0.42	0.46	2.84	0.048	4.220977	0.01
20	-0.05664	8.780941	0.2654	0.189867	0.031747	0.068	0.089	0.046	0.94	0.93	0.95	0.27	0.27	0.29	6.12	0.065	4.138361	0.03
21	0.019205	6.287859	0.4207	0.130682	0.109921	0.053	0.039	0.011	0.74	0.69	0.4	0.27	0.55	0.07	6.23	0.15	3.811097	0.02
22	0.085206	6.137727	0.1371	0.123818	0.015897	0.028	0.025	0.007	0.72	0.75	0.63	0.12	0.17	0.07	5.51	0.128	3.78646	0.0
23	0.133982	8.12888	0	0.16721	0.003311	0.129	0.196	0.063	1	1	1	0.82	0.83	0.83	2.02	0.018	4.252772	0.00
24	0.173025	6.680855	0.4713	0.228424	0.029328	0.062	0.09	0.032	1	1	0.97	0.35	0.44	0.25	4.84	0.062	4.037774	0.02
25	0.109699	7.177019	0.0178	0.18524	0.015453	0.02	0.026	0.013	0.9	0.96	0.8	0.28	0.34	0.21	6.25	0.055	4.058717	0.0
26	0.01599	6.648985	0.4762	0.171181	0.058937	0.018	0.028	0.007	0.4	0.59	0.22	0.12	0.18	0.07	7	0.149	3.824284	0.02
27	0.06225	6.879356	0.2927	0.179508	0.035842	0.188	0.169	0.208	1	1	1	0.41	0.42	0.38	6.8	0.072	4.016383	0.02

growth

图1.1 90个国家经济增长数据集growth.xls

表1.1 growth数据集中部分变量描述性统计

变量	均值	标准差	最小值	最大值
人均GDP增长率	0.0453	0.0513	-0.1010	0.1855
实际人均GDP（1980 international prices）	7.7029	0.8962	5.7621	9.2298
贸易开放程度	0.2201	0.0749	0.0785	0.4162
25岁及以上人口平均受教育年限	4.2149	2.5304	0.301	11.158
政治不稳定性	0.1142	0.2144	0	1.0685
劳动人口比重	0.3693	0.0695	0.2156	0.526
人口增长率	0.0213	0.0099	0.0026	0.039
政府实际消费支出占GDP比重	0.1552	0.0613	0.0355	0.3859
汇率（与美元比较）	42.6639	119.3351	0.003	652.85
65岁以上人口比重	0.0592	0.0377	0.0215	0.1511
女性平均生育孩子数量	4.742	1.9749	1.45	8
25岁及以上男女平均受教育年限之比	1.8769	1.6222	0.9691	13.2093

三、因果推断

在X的众多分量重，我们有时感兴趣其中一个分量（如初始人均GDP，记作X_1）对Y（人均GDP增速）产生的因果效应。所谓因果效应，是当X中其他因素不变时，仅仅变动X_1对Y带来的影响（的平均值）。X_1的变动存在多种形式，如：

1. $X_{1,\delta} = X_1 + \delta$，表示每个国家的初始人均GDP增加$\delta$时，变动对人均GDP增速带来的因果效应；

2. $X_{1,\delta} = \frac{X_1}{1+\delta}$，由于$E(X_{1,\delta}) = \frac{E(X_1)}{1+\delta}$，$Var(X_{1,\delta}) = \frac{Var(X_1)}{(1+\delta)^2}$，该变换形式同时在不同程度上降低了$X_1$的均值和方差。

三、因果推断

例 政府支出对产出的因果效应

假设存在以下凯恩斯模型：

$$\begin{cases} Y_t = C_t + I_t + G_t \\ C_t = \alpha + \beta Y_t + \varepsilon_t \end{cases}$$

其中，Y_t是总产出，C_t是私人消费，I_t是私人投资，G_t是政府支出，ε_t是消费的随机扰动项。参数α和β分别表示生存消费水平和边际消费倾向。政府支出的收入乘数取决于消费倾向β：

$$\frac{\partial Y_t}{\partial G_t} = \frac{1}{1-\beta}.$$

假设中国政府为了保证每年5%的经济增长率，将采取积极的财政政策，这需要确定每年应该发行多大规模的政府债券，政府支出取决于边际消费倾向β值的大小。由于文化差异可能导致消费行为的不同，可以预见，不同国家的β值可能不同；即使是同一经济体，在不同的经济发展阶段，β值也可能不同，从而导致政府支出对产出的拉动作用存在差异。

四、利用线性回归进行预测

- 为什么不能用样本内MSE作为评价估计量的指标？

当采用更为复杂的模型（如在线性回归中加入更多的解释变量），样本内MSE总是会不断缩小，但样本外MSE却不一定如此。

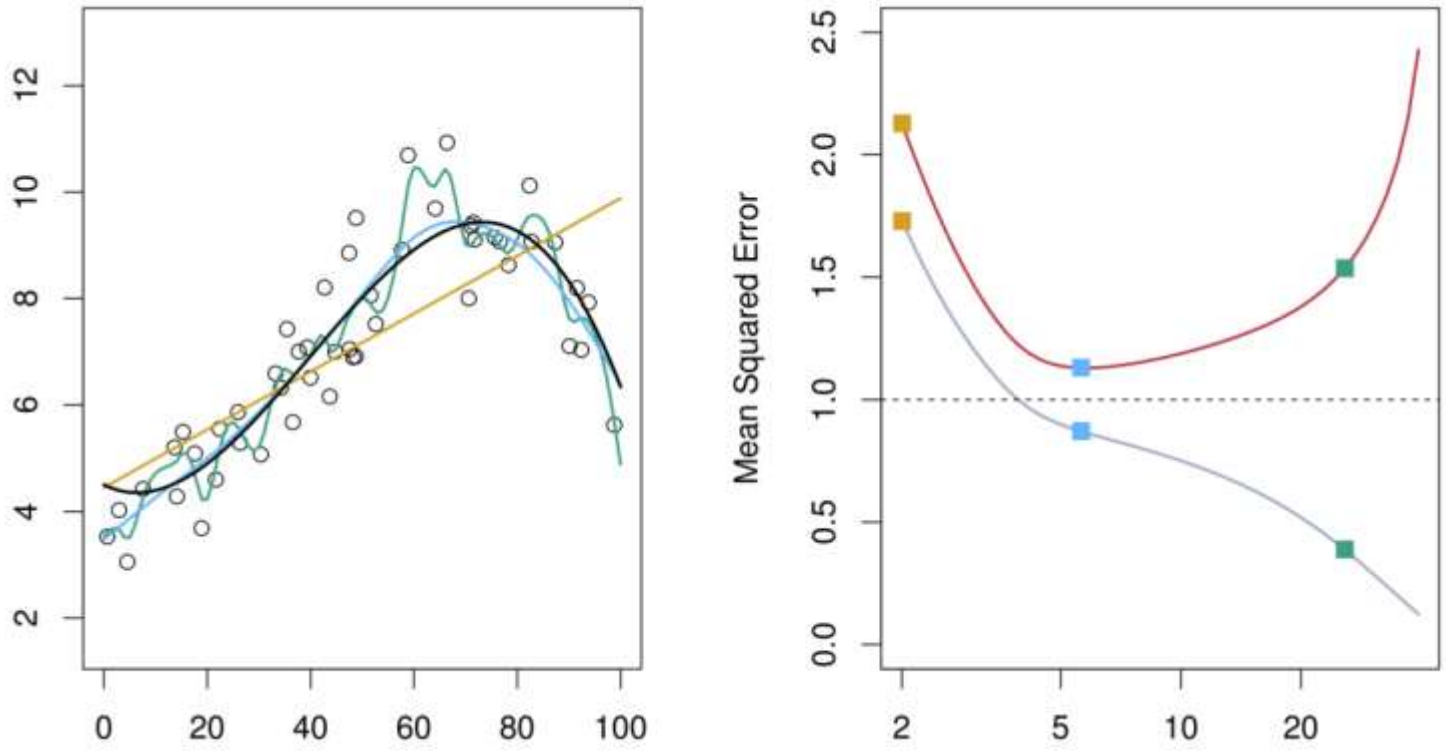


图1.2 样本内MSE与样本外MSE之间的关系（上）

四、利用线性回归进行预测

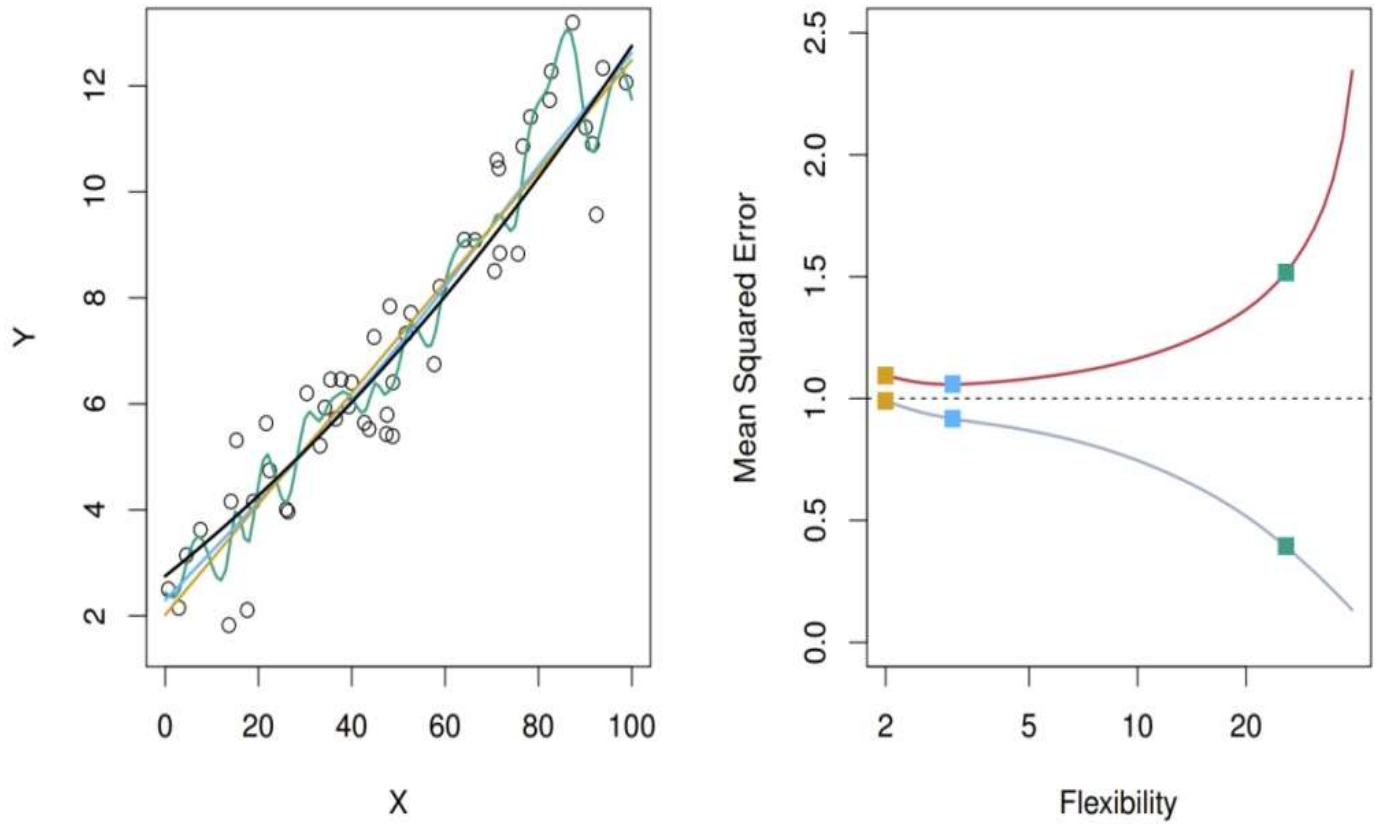


图1.3 样本内MSE和样本外MSE的关系（下）

四、利用线性回归进行预测

估计MSFE的步骤：

- 将全部样本“随机地”分成两组。不失一般性，每组的样本容量为$\frac{n}{2}$，第一部分称为训练集（training sample），第二部分称为测试集（testing sample）。
- 用训练集中样本估计模型。例如，得到线性回归预测$\hat{m}(x) = x^{'}\hat{\beta}$。
- 用测试集中样本去估计模型的预测能力。记测试集样本指标为J，其中样本容量为$\frac{n}{2}$，则

$$\widehat{MSFE} = \frac{2}{n}\sum_{j\in J}(Y_j - \hat{Y}_j)^2$$

其中$\hat{Y}_j = X_j^{'}\hat{\beta}$，$\hat{\beta}$是仅仅利用训练集中样本得到的回归系数。

五、过度拟合（overfitting）

- 当一种估计量的样本内MSE非常小，而此时样本外MSE非常大，就出现所谓的“过度拟合”。这是因为复杂度过高的模型会试图捕捉（拟合）实际上毫无规律的样本“噪声”（扰动项）。
- 在线性回归中，过度拟合出现在以下情形：解释变量个数p相对于n太大，例如$p \cong \frac{n}{2}$，甚至$p \cong n$;但同时并未采用惩罚等正则化方法（参见第二章中讲的Lasso）。
- 考虑以下数值模型：设Y服从标准正态分布N(0,1)，p维解释变量X服从p维正态分布$N(0, I_p)$。X和Y互相独立。由于X与Y之间独立，X理论上不能为预测Y提供任何信息。设n=10000，数值模拟的结果如下：

表1.2 过度拟合时R^2, $\bar{R}^2$ 与MSE的性质

p	250	500	1000	2000	5000	7500	10000
p/n	1/40	1/20	1/10	1/5	1/2	3/4	1
R^2	0.0256	0.0496	0.0955	0.1975	0.5080	0.7500	1
$\bar{R}^2$	0.00075	-0.00029	-0.0049	-0.0030	0.0161	9.02e-06	-Inf
样本内MSE	0.9781	0.9540	0.9080	0.8055	0.4939	0.2510	6.06e-21
样本外MSE	1.0243	1.0443	1.0780	1.1986	1.5021	1.7495	10038.0436

六、交错鉴定法

思考题：残差平方的样本均值$\frac{1}{n}\sum_{j=1}^{n}(Y_j-\hat{Y}_j)^2$是否适合作为MSFE的估计量？

Hold-out（留出）原理：

将整个样本$i=1,\ldots,n$分成$n_1+n_2=n$两部分。不失一般性，假设$n_1=n_2=\frac{n}{2}$。其中，n_1（测试集）用来估计参数，剩下的n_2用来预测Y与评估模型。参数估计量为：

$$\hat{\beta}_{n_1}=(\sum_{i=1}^{n_1}X_iX_i^{'})^{-1}(\sum_{i=1}^{n_1}X_iY_i).$$

预测误差（prediction error）为：

$$\widetilde{U}_i=Y_i-X_i\hat{\beta}_{n_1},i=n_1+1,\ldots,n_1+n_2.$$

MSFE的一个估计量是：

$$\tilde{\sigma}_{n_2}^2=\frac{1}{n_2}\sum_{i=n_1+1}^{n_1+n_2}\widetilde{U}_i^2\,.$$

$\tilde{\sigma}_{n_2}^2$是MSFE的一个无偏估计量。

思考题：为什么说$\tilde{\sigma}_{n_2}^2$是$E(Y-X^{'}\hat{\beta})^2$的无偏估计量？用$\tilde{\sigma}_{n_2}^2$作为MSFE的估计量有哪些缺陷？

六、交错鉴定法

$\tilde{\sigma}_{n_2}^2$作为MSFE的估计量<u>没有</u>有效利用样本信息。定义

$$\tilde{\sigma}_{n_1}^2 = \frac{1}{n_1}\sum_{i=1}^{n_1}\tilde{U}_i^2 .$$

则

$$\frac{\tilde{\sigma}_{n_1}^2 + \tilde{\sigma}_{n_2}^2}{2}$$

是一个比$\tilde{\sigma}_{n_2}^2$更好的评价模型优劣的指标。

思考题：为何$\frac{\tilde{\sigma}_{n_1}^2+\tilde{\sigma}_{n_2}^2}{2}$比$\tilde{\sigma}_{n_2}^2$能更好地估计MSFE?

如果将估计MSFE时所用的观测值个数取至最大（最大值为n），得到：

$$\widehat{MSFE} = \frac{1}{n}\sum_{i=1}^{n}\tilde{e}_i^2$$

$\tilde{e}_i = Y_i - X_i\hat{\beta}$中的$\hat{\beta}$应与$(Y_i, X_i)$无关，因此$\hat{\beta}$可以采用

$$\hat{\beta}_{-1} = (\sum_{j=1,j\neq i}^{n} X_j X_j^{'})^{-1}(\sum_{j=1,j\neq i}^{n} X_j Y_j)$$

六、交错鉴定法

- 把上述估计MSFE的思路进一步推广，可以得到一般的交错鉴定法（cross validation, CV）。CV是一种用途广泛的模型选择原理，可用于机器学习算法中调节参数（runing parameter）的选取。
- 交错鉴定法的一般步骤：

（1）将全部数据分成K组（folds）。例如，K=5, 表示把全部样本分成5组，每组中含观测值。

（2）每次留出K组中的1组作为测试集。剩下的K-1组样本用于训练模型。并计算训练出的模型在测试集上的均方预测误差。

（3）重复上述步骤K次，轮流使每一块成为测试集。

（4）把上述K次中得到的均方预测误差的平均值作为模型预测精度的指标。

（5）对不同的模型，或者不同调试参数重复上述步骤。采用使得均方预测误差取最小值的模型或者调试参数。

思考题：写出K=3时MSFE的CV估计量。

七、案例分析：采用线性模型预测经济增长

表1.3 不同预测模型设定描述

模型编号	模型描述
1	包含变量：（1）人均实际GDP、（2）贸易开放度、（3）25岁及以上人口平均上学年限、（4）政治不稳定性、（5）劳动人口比重、（6）人口增长率、（7）政府实际消费支出与实际GDP比重
2	包含变量：模型1变量+（8）汇率、（9）65岁以上人口比重、（10）女性平均生育孩子数量、（11）25岁及以上男女平均上学年限之比
3	包含变量：模型1变量+这些变量的平方项
4	包含变量：、模型2变量+这些变量的平方项
5	包含变量：模型3变量+这些变量的两两交互项
6	包含变量：模型4变量+这些变量的两两交互项

七、案例分析：采用线性模型预测经济增长

表1.4 不同线性模型对经济增长的预测能力

模型	解释变量个数p	p/n	样本内MSE	K=3时MSFE估计量	K=5时MSFE估计量	K=10时MSFE估计量	K=90时MSFE估计量
1	8	4/45	0.00219	0.0033	0.00306	0.00297	0.00307
2	12	2/15	0.00207	0.00356	0.00326	0.00312	0.00320
3	15	1/6	0.00199	0.00357	0.00329	0.00320	0.00328
4	23	23/90	0.00170	0.00408	0.00361	0.00349	0.00358
5	36	2/5	0.00152	0.00486	0.00398	0.00375	0.00375
6	78	13/15	0.00024	0.00594	0.00517	0.00485	0.00505

谢 谢!