小议决定系数R^2

江 苏

（Email: jiangsukust@163.com）

我们在近红外定量分析的模型建立及模型评估中经常会用到一个无量纲的统计指标——决定系数（Coefficient of Determination，$R\text{-}Squared$，R^2或COD），该统计量1921年由Wright[1]提出，其原始公式量化了因变量由自变量决定的程度，即方差比例；现在被广泛接受的定义是：因变量的全部变异能通过回归关系被自变量解释/预测的比例。决定系数可以用于评价模型拟合优度（基于校正集）及预测性能（基于验证集或测试集，此阶段有的文献和软件称其为Q^2）。

决定系数的计算方法不尽相同[2; 3]，教科书和文献中最常见的就有三种，分别记为R_1^2、R_2^2和R_3^2，假定一组数据包含n个样本，则公式分别为：

$$R_1^2 = \left(r_{y\hat{y}}\right)^2 = \left(\frac{\sum_{i=1}^{n}(y_i-\bar{y})\,(\hat{y}_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2}}\right)^2 = \left(\frac{cov(y,\hat{y})}{std(y)std(\hat{y})}\right)^2 \tag{1}$$

$$R_2^2 = \frac{SSR}{SST} = \frac{\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2}{\sum_{i=1}^{n}(y_i-\bar{y})^2} \tag{2}$$

$$R_3^2 = 1-\frac{SSE}{SST} = 1-\frac{\sum_{i=1}^{n}(\hat{y}_i-y_i)^2}{\sum_{i=1}^{n}(y_i-\bar{y})^2} \tag{3}$$

(1)式可以理解为相关系数的平方，其中：

y_i — 第i个样本的参考值（$i=1,2\ldots n$）

$\bar{y}$ — 所有n个样本参考值的平均值

$\hat{y}_i$ — 第i个样本经模型回归后的拟合值或经模型预测后得到的预测值

$r_{y\hat{y}}$ — $\boldsymbol{y}$与$\hat{\boldsymbol{y}}$之间的皮尔逊（泊松）相关系数，或称积矩相关系数

$cov(y,\hat{y})$ — $\boldsymbol{y}$与$\hat{\boldsymbol{y}}$之间的协方差

$std(y)$ — $\boldsymbol{y}$的标准偏差

$std(\hat{y})$ — $\hat{\boldsymbol{y}}$的标准偏差

(2)式可以理解为回归关系已经解释的$\boldsymbol{y}$值变异在其总变异中所占的比例，其中：

SSR — Sum of Squares Regression，回归平方和

SST — Sum of Squares Total，总平方和

(3)式可以理解为1减去回归关系不能解释的$\boldsymbol{y}$值变异在其总变异中所占的比例，其中：

SSE — Sum of Squares Error，残差平方和，或称$PRESS$

笔者发现一些教科书和文献或是由于篇幅限制、或是由于不够严谨、或是由于疏忽，对决定系数的描述或使用存在一些瑕疵，比如简单地将三者划上等号、错误地规定了其取值范围、仅使用决定系数进行模型选择或评估模型性能等，难免会给读者带来一些困惑，因此，我们尝试对这些问题逐一进行讨论。

1. $R_1^2=R_2^2=R_3^2$?

首先必须指出的是，这一等式仅适用于模型校正（拟合）阶段，不适用于交叉验证和预测阶段；其次，并非所有回归算法下该等式都成立。我们依次从单变量和多变量回归的角度来分析这一问题。

1.1 单变量回归

几乎所有的统计学教科书都会从最简单的一元线性回归开始介绍决定系数的计算，为了更清晰的阐述这个问题，我们模拟了一组包含 6 个样本的数据集：

表 1 模拟数据集

x_i	1.0	2.0	3.0	4.0	5.0	6.0
y_i	9.1	12.0	12.5	14.6	17.8	24.8

该数据集仅包含一个自变量$\boldsymbol{x}$，一个因变量$\boldsymbol{y}$。我们首先对其进行一元线性回归，模型参数的估计通常采用普通最小二乘法（Ordinary Least Squares，OLS），其基本原理可以参考教科书[4; 5]。

1.1.1 一元线性回归（$\hat{y} = b_1 x + b_0$）

对于没有编程经验的读者来说，我们可以利用 Excel 软件进行相应的操作和计算，而且步骤非常简单。将$\boldsymbol{x}$和$\boldsymbol{y}$以行或者列（更常见）的形式一一输入到单元格中，选择这些数据，然后在“图表”中选择“散点图”，单击任一数据点，右键菜单中选择“添加趋势线”，在“趋势线选项”中选择“线性”（默认项），同时勾选“显示公式”和“显示 R 平方值”，稍微做些格式修改即可得到下图：

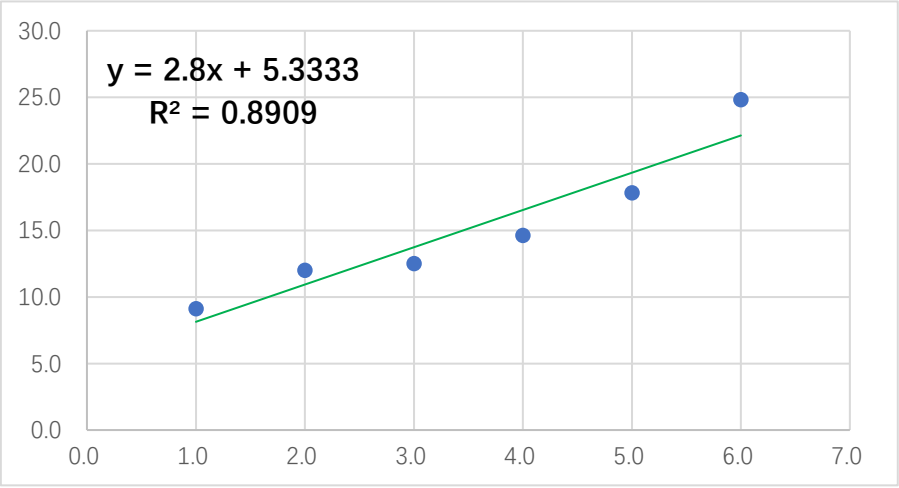

图 1 一元线性回归

图 1 中的回归方程是一条斜率为 2.8，截距不为 0（5.3333）的直线，该直线将通过点$(\bar{x}, \bar{y})$。为了减小小数位数带来的舍入误差，更准确的斜率和截距可以分别通过$SLOPE$和$INTERCEPT$函数（或$LINEST$函数）来得到。现在我们可以根据回归方程进行相应的计算，列表如下：

表 2 一元线性回归计算结果

序号	x_i	y_i	$\hat{y}_i$	$\hat{y}_i - y_i$	$\hat{y}_i - \bar{y}$	$y_i - \bar{y}$
1	1.0	9.1	8.13333	-0.96667	-7.00000	-6.03333
2	2.0	12.0	10.93333	-1.06667	-4.20000	-3.13333
3	3.0	12.5	13.73333	1.23333	-1.40000	-2.63333
4	4.0	14.6	16.53333	1.93333	1.40000	-0.53333
5	5.0	17.8	19.33333	1.53333	4.20000	2.66667
6	6.0	24.8	22.13333	-2.66667	7.00000	9.66667
Mean	**3.50000**	**15.13333**	**15.13333**	**0.00000**	**0.00000**	**0.00000**

$\boldsymbol{b_0 = 5.33333}$　　$\boldsymbol{b_1 = 2.80000}$

$\boldsymbol{cov(y, \hat{y}) = 27.44000}$　　$\boldsymbol{std(y) = 5.54965}$　　$\boldsymbol{std(\hat{y}) = 5.23832}$

$$\boldsymbol{SSE = \sum_{i=1}^{n} (\hat{y}_i - y_i)^2 = 16.79333}$$

$$\boldsymbol{SSR = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 = 137.20000}$$

$$\boldsymbol{SST = \sum_{i=1}^{n} (y_i - \bar{y})^2 = 153.99333}$$

那么有：

$$R_1^2 = \left(\frac{cov(y, \hat{y})}{std(y)std(\hat{y})}\right)^2 = \left(\frac{27.44000}{5.54965 \times 5.23832}\right)^2 = \mathbf{0.89095}$$

其实，在 Excel 里可以通过$CORREL$函数直接计算相关系数$r_{y\hat{y}} = 0.94390$，将其平方后可得$r^2 = 0.89095$，或使用RSQ函数直接计算，都将与上式结果一致；而且$\boldsymbol{x}$和$\boldsymbol{y}$的相关系数r_{xy}也等于0.94390。

类似地：

$$R_2^2 = \frac{SSR}{SST} = \frac{137.20000}{153.99333} = \mathbf{0.89095}$$

$$R_3^2 = 1 - \frac{SSE}{SST} = 1 - \frac{16.79333}{153.99333} = 1 - 0.10905 = \mathbf{0.89095}$$

注：如果在 Excel 里直接计算，几乎可以不考虑舍入误差

通过上面的计算，我们可以看到R_1^2、R_2^2和R_3^2的结果完全相同，与图 1 中显示的R^2也是一致的，$SSR + SSE = 137.20000 + 16.79333 = 153.99333 = SST$，一切事物都很美好！

然而，我们都知道，对于单变量回归而言，回归方程可以是线性也可以是任何其他形式，比如不含截距的线性回归、多项式回归、指数回归、对数回归等等，那么其他回归算法下还能维持这种完美的关系吗？我们分别通过不含截距的一元线性回归、一元二次多项式回归及指数回归来举例说明。

1.1.2 不含截距的一元线性回归（$\hat{y} = b_1 x$）

此时，回归线将通过原点$(0,0)$，该回归线斜率为 4.03077。

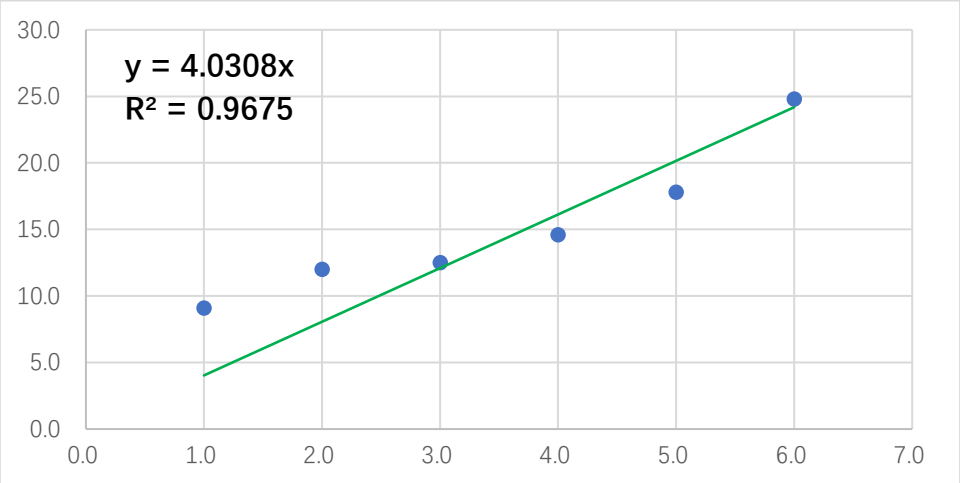

图 2 不含截距的一元线性回归

计算结果如表 3 所示：

表 3 不含截距的一元线性回归计算结果

序号	x_i	y_i	$\hat{y}_i$	$\hat{y}_i - y_i$	$\hat{y}_i - \bar{y}$	$y_i - \bar{y}$
1	1.0	9.1	4.03077	-5.06923	-11.10256	-6.03333
2	2.0	12.0	8.06154	-3.93846	-7.07179	-3.13333
3	3.0	12.5	12.09231	-0.40769	-3.04103	-2.63333
4	4.0	14.6	16.12308	1.52308	0.98974	-0.53333
5	5.0	17.8	20.15385	2.35385	5.02051	2.66667
6	6.0	24.8	24.18462	-0.61538	9.05128	9.66667
Mean	**3.50000**	**15.13333**	**14.10769**	**-1.02564**	**-1.02564**	**0.00000**
$b_1 = 4.03077$	$cov(y,\hat{y}) = 39.50154$		$std(y) = 5.54965$		$std(\hat{y}) = 7.54088$	
$SSE = \sum_{i=1}^{n}(\hat{y}_i - y_i)^2 = 49.61385$						
$SSR = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 = 290.63590$						
$SST = \sum_{i=1}^{n}(y_i - \bar{y})^2 = 153.99333$						

那么有：

$$R_1^2 = \left(\frac{cov(y,\hat{y})}{std(y)std(\hat{y})}\right)^2 = \left(\text{CORREL}(y,\hat{y})\right)^2 = \left(\frac{39.50154}{5.54965 \times 7.54088}\right)^2 = \mathbf{0.89095}$$

$$R_2^2 = \frac{SSR}{SST} = \frac{290.63590}{153.99333} = \mathbf{1.88733}$$

$$R_3^2 = 1 - \frac{SSE}{SST} = 1 - \frac{49.61385}{153.99333} = 1 - 0.32218 = \mathbf{0.67782}$$

我们可以看到计算出来的R_1^2、R_2^2、R_3^2都与图 2 中R^2不一致。实际上，对于无截距的一元线性回归模型，R^2的计算可以选择另外一个公式，记为R_4^2：

$$R_4^2 = 1 - \frac{\sum_{i=1}^{n} \hat{y}_i{}^2}{\sum_{i=1}^{n} y_i{}^2} = \frac{1478.48615}{1528.10000} = \mathbf{0.96753}$$

如此便符合 Excel 中的公式设定。

1.1.3 一元二次多项式回归（$\hat{y} = b_2x^2 + b_1x^1 + b_0$）

此时，回归线为一条二次曲线。

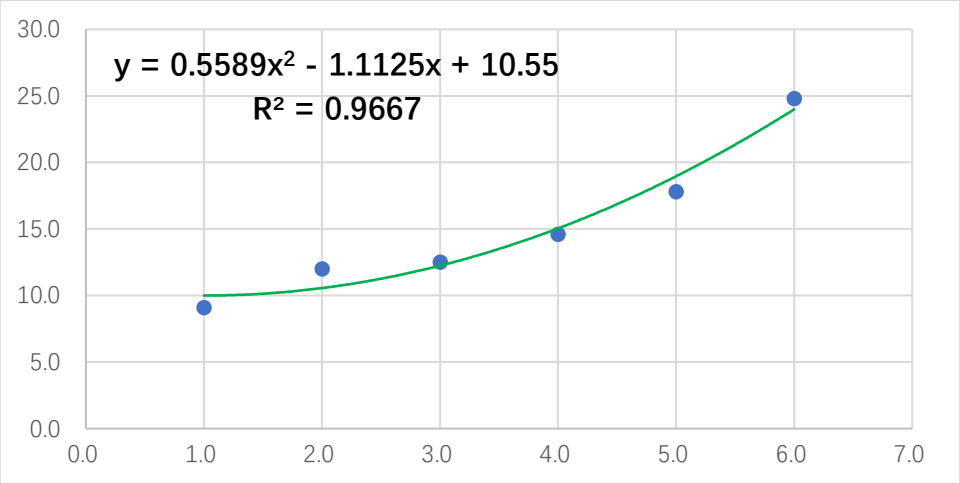

图 3 一元二次多项式回归

计算结果列表如下：

表 4 一元二次多项式回归计算结果

序号	x_i	y_i	$\hat{y}_i$	$\hat{y}_i - y_i$	$\hat{y}_i - \bar{y}$	$y_i - \bar{y}$
1	1.0	9.1	9.99643	0.89643	-5.13690	-6.03333
2	2.0	12.0	10.56071	-1.43929	-4.57262	-3.13333
3	3.0	12.5	12.24286	-0.25714	-2.89048	-2.63333
4	4.0	14.6	15.04286	0.44286	-0.09048	-0.53333
5	5.0	17.8	18.96071	1.16071	3.82738	2.66667
6	6.0	24.8	23.99643	-0.80357	8.86310	9.66667

Mean	**3.50000**	**15.13333**	**15.13333**	**0.00000**	**0.00000**	**0.00000**
$b_0 = 10.55000$		$b_1 = -1.11250$		$b_2 = 0.55893$		
$SSE = \sum_{i=1}^{n} (\hat{y}_i - y_i)^2 = 5.13036$ $SSR = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 = 148.86298$ $SST = \sum_{i=1}^{n} (y_i - \bar{y})^2 = 153.99333$						

那么有：

$$R_1^2 = \left(\text{CORREL}(y, \hat{y})\right)^2 = (0.98320)^2 = \mathbf{0.96668}$$

$$R_2^2 = \frac{SSR}{SST} = \frac{148.86298}{153.99333} = \mathbf{0.96668}$$

$$R_3^2 = 1 - \frac{SSE}{SST} = 1 - \frac{5.13036}{153.99333} = 1 - 0.03332 = \mathbf{0.96668}$$

显然，R_1^2、R_2^2和R_3^2的结果完全相同！

注 1：事实上，如果您使用的多项式阶数仅比数据点的数量少一个（$n-1$，对于上述数据集，采用 5 阶多项式），那么回归方程将完美拟合所有数据点，此时$R_1^2 = R_2^2 = R_3^2 = 1$，$SSE = 0$。

注 2：多项式回归是通过变量转换将其化为多元线性回归的问题，然后采用最小二乘法来估计未知参数，即令$x_1 = x$，$x_2 = x^2$，原公式转化为$\hat{y} = b_2 x_2 + b_1 x_1 + b_0$。

注 3：如果不含截距，回归方程为$\hat{y} = b_2 x^2 + b_1 x^1$，$R_1^2 = 0.82338$，$R_2^2 = 1.38880$，$R_3^2 = -8.66397$。

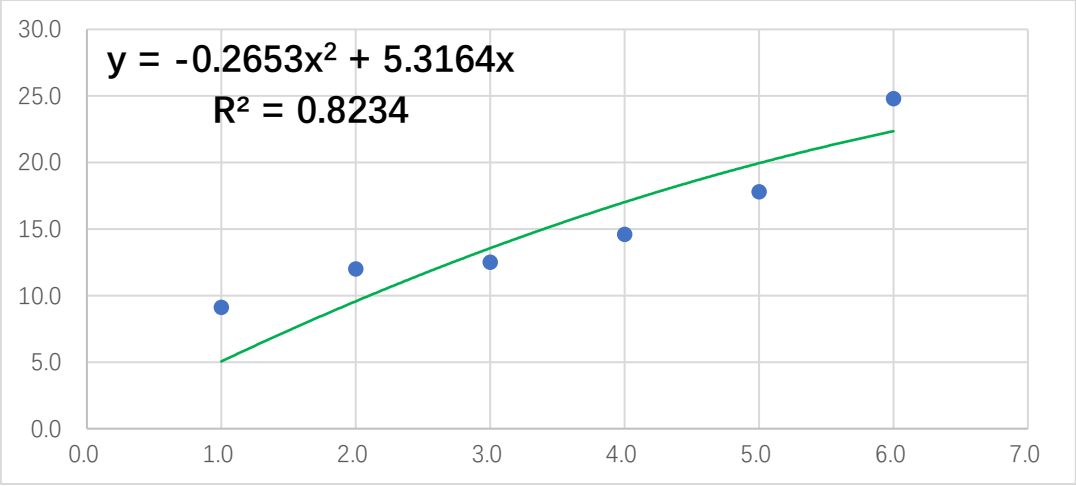

图 4 不含截距的一元二次多项式回归

值得注意的是，上图显示的$R^2 = R_1^2$，这一点与“不含截距的一元线性回归（使用R_4^2）”不同；使用$LINEST$函数计算出来的统计量$R^2 = R_4^2 = 0.97388$；对于不含截距的线性回归，$LINEST$始终坚持采用R_4^2，而散点图中的趋势线却不然，可以在不少问答网站看到相关讨论，比如 https://techcommunity.microsoft.com/t5/excel/difference-in-calculated-correlation-coefficient-value-between/m-p/2709119。

1.1.4 指数回归（$\hat{y} = a * e^{bx}$）

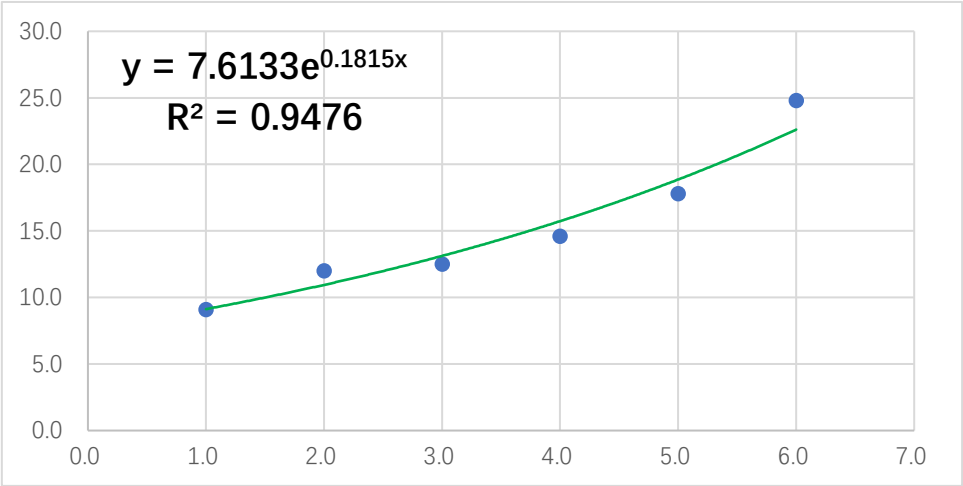

图 5 指数回归

对于这一类型（可线性化处理）的非线性回归模型，可以通过变量转换将其转化为线性模型。对原方程$\hat{y} = a * e^{bx}$的两边同时取自然对数，化为：

$$ln(\hat{y}) = ln(a) + bx$$

令$y' = ln(\hat{y})$，$a' = ln(a)$，则上式转化为：

$$y' = a' + bx$$

这不正是一元线性回归的公式嘛！所以，可以对原始因变量y取自然对数即$ln(y)$，利用最小二乘法求出a'和b，然后$EXP(a')$求出a。

计算结果见表 5：

表 5 指数回归计算结果

序号	x_i	y_i	y_i'	$\hat{y}_i$	$\hat{y}_i - y_i$	$\hat{y}_i - \bar{y}$	$y_i - \bar{y}$
1	1.0	9.1	2.20827	9.12805	0.028046	-6.00529	-6.03333
2	2.0	12.0	2.48491	10.94420	-1.055805	-4.18914	-3.13333
3	3.0	12.5	2.52573	13.12169	0.621693	-2.01164	-2.63333
4	4.0	14.6	2.68102	15.73243	1.132432	0.59910	-0.53333
5	5.0	17.8	2.87920	18.86261	1.062614	3.72928	2.66667
6	6.0	24.8	3.21084	22.61559	-2.184411	7.48226	9.66667
Mean	**3.50000**	**15.13333**	**2.66500**	**15.06743**	**-0.06590**	**-0.06590**	**0.00000**
$a' = 2.02989$			$a = EXP(a') = 7.61328$			$b = 0.18146$	
$SSE = \sum_{i=1}^{n} (\hat{y}_i - y_i)^2 = 8.68522$							

$$\boldsymbol{SSR = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 = 127.90966}$$

$$\boldsymbol{SST = \sum_{i=1}^{n} (y_i - \bar{y})^2 = 153.99333}$$

便有：

$$R_1^2 = \left(\mathrm{CORREL}(y, \hat{y})\right)^2 = (0.97347)^2 = \mathbf{0.94762}$$

$$R_2^2 = \frac{SSR}{SST} = \frac{127.90966}{153.99333} = \mathbf{0.83062}$$

$$R_3^2 = 1 - \frac{SSE}{SST} = 1 - \frac{8.68522}{153.99333} = 1 - 0.05640 = \mathbf{0.94360}$$

很明显，R_1^2、R_2^2和R_3^2的结果各不相同！

注 1：于本例而言，最小二乘法“最小化”的是新响应变量$ln(y)$的残差平方和，所以得到上述结果并不意外。

注 2：对数回归方程$\hat{y} = a + bln(x)$通过变量转换，原方程化为$\hat{y} = a + bx'$，可以采用最小二乘法进行参数估计，$R_1^2 = R_2^2 = R_3^2$；该方程无法将截距项设为 0。

注 3：幂函数$\hat{y} = ax^b$通过变量转换，原方程化为$\widehat{y'} = a' + bx'$，可以采用最小二乘法进行参数估计，R_1^2、R_2^2和R_3^2的数值各不相同；该方程同样无法将截距项设为 0。

注 4：指数回归和幂函数回归在进行变量转换的时候，对原始的因变量y进行了对数转换（非线性变换），这与对数回归是有区别的，应该是这一点造成R_1^2、R_2^2和R_3^2的数值各不相同。

※ **小结**

表 6 单变量回归结果统计表

	一元线性回归	不含截距的 一元线性回归	多项式回归	不含截距的 多项式回归	指数回归
$\boldsymbol{R_1^2}$	0.89095	0.89095	0.96668	0.82338	0.94762
$\boldsymbol{R_2^2}$	0.89095	1.88733	0.96668	1.38880	0.83062
$\boldsymbol{R_3^2}$	0.89095	0.67782	0.96668	-8.66397	0.94360
$\boldsymbol{SSE}$	16.79333	49.61385	5.13036	1488.18761	8.68522
$\boldsymbol{SSR}$	137.20000	290.6359	148.86298	213.86636	127.90966
$\boldsymbol{SST}$	153.99333	153.99333	153.99333	153.99333	153.99333

对于该数据集，通过上表及前述讨论我们可以得出以下结论：

1) 采用一元线性回归和一元二次多项式回归算法时，R_1^2=R_2^2=R_3^2，$SSR + SSE = SST$

2) 采用不含截距的一元线性回归时，R_1^2、R_2^2、R_3^2各不相同，$R_2^2 > 1$（$SSR > SST$），$SSR + SSE \neq SST$

3) 采用不含截距的多项式回归时，R_1^2、R_2^2、R_3^2各不相同，$R_2^2 > 1$（$SSR > SST$），$R_3^2 < 0$（$SSE > SST$）

4) 采用指数回归算法时，R_1^2、R_2^2、R_3^2各不相同，$SSR + SSE \neq SST$。

5) 采用一元线性回归和不含截距的一元线性回归时，R_1^2=$\big(CORREL(y,\hat{y})\big)^2 = \big(CORREL(x,y)\big)^2$。这其实不难理解，因为皮尔逊相关系数对位置变换（加一个常数）和尺度变换（乘一个正数）是不变的[6]。对于两个变量的相关系数$CORREL(X,Y)$来说，如果X和（或）Y经过线性变换：

$$X_{i*} = aX + b$$

$$Y_{i*} = cY + d$$

如果a和b是正数，c和d是常数，那么$\mathrm{CORREL}(X,Y) = \mathrm{CORREL}(X_*, Y_*)$。回到正题，因为$\hat{y}$是对$x$的线性转换（$\hat{y} = b_1 x + b_0$或$\hat{y} = b_1 x$），如果$b_1$为正数，那么$\mathrm{CORREL}(y,\hat{y}) = \mathrm{CORREL}(x,y)$；如果$b_1$为负数，$\mathrm{CORREL}(y,\hat{y}) = -\mathrm{CORREL}(x,y)$；两边平方后数值相等。

1.2 多变量回归

在分析化学的测量中，尤其是现代高通量分析仪器的广泛应用，我们遇到的自变量经常是一组，比如光谱、色谱、质谱等，这时候我们需要多变量校正（多元校正）方法来建立因变量与自变量之间的关系。多元校正方法丰富多样，如多元线性回归（Multiple Linear Regression, MLR）、主成分回归（Principal Component Regression, PCR）、偏最小二乘回归法（Partial Least Squares Regression, $PLSR$）等线性回归方法，以及支持向量回归（Support Vector Regression, SVR）、人工神经网络（Artificial Neural Network, ANN）等非线性回归方法。近年来，以卷积神经网络（Convolutional Neural Networks, CNN）为代表的深度学习算法也开始用于近红外光谱的定量和定性分析[7]。

1.2.1 多元线性回归

我们引用 James N. Miller 等所著《Statistics and chemometrics for analytical chemistry》（7th ed.）一书 P252 中 Table 8.4 的数据，该数据集含 10 个样本的紫外吸收光谱（100 个波长点中的 6 个，A_1，A_2…）及 3 种组分的含量（c_1，c_2，c_3），见表 7：

表 7 紫外吸收光谱数据

Specimen	c_1	c_2	c_3	A_1	A_2	A_3	A_4	A_5	A_6
A	0.89	0.02	0.01	18.7	26.8	42.1	56.6	70.0	83.2
B	0.46	0.09	0.24	31.3	33.4	45.7	49.3	53.8	55.3
C	0.45	0.16	0.23	30.0	35.1	48.3	53.5	59.2	57.7
D	0.56	0.09	0.09	20.0	25.7	39.3	46.6	56.5	57.8
E	0.41	0.02	0.28	31.5	34.8	46.5	46.7	48.5	51.1
F	0.44	0.17	0.14	22.0	28.0	38.5	46.7	54.1	53.6

G	0.34	0.23	0.20	25.7	31.4	41.1	50.6	53.5	49.3
H	0.74	0.11	0.01	18.7	26.8	37.8	50.6	65.0	72.3
I	0.75	0.01	0.15	27.3	34.6	47.8	55.9	67.9	75.2
J	0.48	0.15	0.06	18.3	22.8	32.8	43.4	49.6	51.1

采用最小二乘法建立$\boldsymbol{c_1}$与吸光度之间的回归方程，即$c_1 = b_0 + b_1A_1 + b_2A_2 \ldots + b_6A_6$ ，结果如下表所示：

表 8 紫外吸收光谱数据多元线性回归计算结果

Specimen	$\boldsymbol{c_1}$	$\boldsymbol{\hat{y}}$	$\boldsymbol{\hat{y}-y}$	$\boldsymbol{\hat{y}-\bar{y}}$	$\boldsymbol{y-\bar{y}}$
A	0.89	0.89699	0.00699	0.34499	0.33800
B	0.46	0.46574	0.00574	-0.08626	-0.09200
C	0.45	0.45484	0.00484	-0.09716	-0.10200
D	0.56	0.55214	-0.00786	0.00014	0.00800
E	0.41	0.41051	0.00051	-0.14149	-0.14200
F	0.44	0.45911	0.01911	-0.09289	-0.11200
G	0.34	0.33418	-0.00582	-0.21782	-0.21200
H	0.74	0.74641	0.00641	0.19441	0.18800
I	0.75	0.73249	-0.01751	0.18049	0.19800
J	0.48	0.46759	-0.01241	-0.08441	-0.07200
Mean	**0.55200**	**0.55200**	**0.00000**	**0.00000**	**0.00000**

$\boldsymbol{\bar{y} = 0.55200}$ $\boldsymbol{b_0 = 0.050095992}$ $\boldsymbol{b_1 = 0.002524674}$ $\boldsymbol{b_2 = -0.009387224}$

$\boldsymbol{b_3 = 0.003754205}$ $\boldsymbol{b_4 = -0.009196692}$ $\boldsymbol{b_5 = -0.001056312}$ $\boldsymbol{b_6 = 0.017880821}$

$$\boldsymbol{SSE = \sum_{i=1}^{n}(\hat{y}_i - y_i)^2 = 0.00107}$$

$$\boldsymbol{SSR = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 = 0.28949}$$

$$\boldsymbol{SST = \sum_{i=1}^{n}(y_i - \bar{y})^2 = 0.29056}$$

可得：

$$R_1^2 = \left(\mathrm{CORREL}(y,\hat{y})\right)^2 = (0.99816)^2 = \mathbf{0.99632}$$

$$R_2^2 = \frac{SSR}{SST} = \frac{0.28949}{0.29056} = \mathbf{0.99632}$$

$$R_3^2 = 1 - \frac{SSE}{SST} = 1 - \frac{0.00107}{0.29056} = 1 - 0.00368 = \mathbf{0.99632}$$

显然，R_1^2、R_2^2和R_3^2的结果完全相同！

如果执行不含截距的多元线性回归，情形会如何呢？计算结果显示：$R_1^2 = 0.99596$，$R_2^2 = 1.00441$，$R_3^2 = 0.99594$；Origin（OriginLab，美国）和$LINEST$函数计算出来的$R^2 = 0.99965$，这一数字与R_4^2相同。

分割线

分割线之前的计算都能在 Excel 中完成，且都能通过$LINEST$函数采用最小二乘法对数据进行拟合（有些回归算法需对原始变量进行线性转换，构造新的变量，需要注意传入$LINEST$函数的参数），并返回相应的参数及回归统计值。以一元二次多项式回归（$\hat{y} = b_2x^2 + b_1x^1 + b_0$）为例进行说明，见下表：

	A	B	C
1	x_i	x_i^2	y_i
2	1.0	1.0	9.1
3	2.0	4.0	12.0
4	3.0	9.0	12.5
5	4.0	16.0	14.6
6	5.0	25.0	17.8
7	6.0	36.0	24.8

LINEST(C2:C7, A2:B7, TRUE, TRUE)

	b_2	b_1	b_0
	0.558928571	-1.1125	10.55
	0.214025139	1.530441877	2.339312068
R^2	0.96668455	1.307715201	#N/A
	43.52415593	3	#N/A
	148.8629762	5.130357143	#N/A
	SSR	SSE	

详见: https://support.microsoft.com/en-us/office/linest-function-84d7d0d9-6e50-4101-977a-fa7abf772b6d

分割线

1.2.2 主成分回归

在MLR中只需知道样品中某些组分的浓度，就可以建立其定量模型，MLR的成功执行必须满足矩阵$\boldsymbol{X}^T\boldsymbol{X}$可逆，也就是该矩阵是满秩的。换言之$\boldsymbol{X}$矩阵必须满足：一是回归的变量数（波长点数）不能超过校正集的样本数目；二是光谱矩阵各列线性不相关。

现代近红外分析仪器的波长变量数从几十个（比如 Spectral Engines）到几千个（FT 型），不同样本之间相似度非常高，而且每个样本所对应的变量数远大于校正集样本数量（这也是现代近红外分析仪器的特点）。我们可以粗略地得到两个结论：1）光谱间相似度很高，即这些光谱的共线性很严重；2）模型变量数大大超过样本数，过拟合风险很高。此时，$\boldsymbol{X}'\boldsymbol{X}$将严重亏秩，此病态矩阵无法继续采用$MLR$求解。于是，数据压缩成了我们解决多重共线性问题的选择，在化学计量学发展过程中出现了两种流行的方法：一是选择少数变量；二是将原始变量经过线性组合，得到较少变量。后者成为目前的主流，也就是常见的主成分回归法（PCR）和偏最小二乘法（PLS），它们均为隐变量回归方法。

我们再次请出表 7 的数据，这一次尝试采用PCR来拟合。**值得说明的是：为了展示这一拟合过程，**

我们采用了两步法：即先进行主成分分析（PCA），然后进行MLR（如果在算法中直接进行，可以不考虑舍入误差）；同时，也不意味着该数据集存在MLR可能面临的问题，仅仅因为该数据集较为简单。

我们选择 The Unscrambler X（CAMO Software，挪威）来执行主成分分析，结果见表 9：

表 9 紫外吸收光谱数据主成分分析计算结果

Specimen	c_1	$Scores$					
		$PC1$	$PC2$	$PC3$	$PC4$	$PC5$	$PC6$
A	0.89	132.35614	18.49191	1.66104	-1.70683	0.55688	0.23052
B	0.46	111.53047	-11.00924	2.17465	-0.2664	-0.36296	-1.25558
C	0.45	118.57556	-9.96781	-0.9608	0.47507	0.63204	0.09028
D	0.56	106.26698	3.0792	-1.66384	0.88199	1.97054	-0.21466
E	0.41	106.10901	-15.30264	3.97106	-0.59769	0.18762	0.65871
F	0.44	103.55014	-1.5223	-2.34406	0.64057	-0.04642	0.22789
G	0.34	105.28655	-8.93355	-4.71764	-0.70175	-0.83198	0.50274
H	0.74	119.67226	13.61247	0.09768	1.1434	-1.16104	-0.07044
I	0.75	132.64796	3.46875	2.92244	1.03263	-0.58801	0.28462
J	0.48	94.17382	2.84456	-2.68575	-1.03288	-0.29872	-0.57474

注：在上述主成分分析的过程中，未选择“Mean center data”。

接下来，我们就能使用$Scores$代替原始吸光度值与c_1之间进行普通的多元线性回归（在 Excel 中使用$LINEST$函数完成），依据选用的主成分数不同，各自计算结果见表 10：

表 10 紫外吸收光谱数据 PCR（PCA-MLR）计算结果

	$1PCs$	$2PCs$	$3PCs$	$4PCs$	$5PCs$	$6PCs$
R_1^2	0.63176	0.95272	0.99506	0.99506	0.99592	0.99632
R_2^2	0.63176	0.95272	0.99506	0.99506	0.99592	0.99632
R_3^2	0.63176	0.95272	0.99506	0.99506	0.99592	0.99632
SSE	0.10699	0.01374	0.00144	0.00144	0.00118	0.00107
SSR	0.18357	0.27682	0.28912	0.28912	0.28938	0.28949
SST	0.29056	0.29056	0.29056	0.29056	0.29056	0.29056

显然，不管选择几个主成分，都有$R_1^2 = R_2^2 = R_3^2$。

我们接着尝试在多元线性回归的过程中不包含截距。以$2PCs$为例，计算结果显示：$R_1^2 = 0.93869$，$R_2^2 = 0.86129$，$R_3^2 = 0.93688$（Unscrambler 计算的结果与此相同）；$LINEST$函数计算出来的$R^2 = 0.99450$，

这一数字与R_4^2相同；选择其他主成分数，R_1^2、R_2^2和R_3^2的结果也将各不相同。

注：事实上，各个数据分析软件在处理不带截距的线性回归时，使用的决定系数公式并不相同。

看到这里，再结合表 6（单变量回归结果统计表），不知道大家有没有发现一个规律：是否包含截距“似乎”对$R_1^2 = R_2^2 = R_3^2$这一等式的成立很关键。

让我们接着往下看。

1.2.3 偏最小二乘回归

这一次我们以一组化学计量学领域常用的 Benchmark 数据集为例，该数据集包括 3 台 Foss 仪器（编号：M5、MP5、MP6）采集的 80 个玉米样品的近红外漫反射光谱数据及其对应的油分、淀粉、蛋白质和水分含量值。每条光谱包含 1100~2498 nm 波长范围内共 700 个波长通道下的响应数据，光谱数据点间隔为 2 nm。该数据集可以从以下网址免费下载 https://eigenvector.com/resources/data-sets/。

我们采用笔者团队开发的 SpecMC 光谱多元校正软件作为计算工具，选择 M5 采集的光谱数据，以蛋白含量为例进行说明。采用系统抽样方法从中选取 50%（40 个，序号为 01、03…77、79）作为校正集，剩余样本（序号为 02、04…78、80）作为测试集，采用“留一法”进行内部交叉验证（如此，用其他化学计量学软件便能重现计算结果）。为了方便显示，我们将最大因子数设为 10。

校正集校正的“模型结果图表”如图 6 所示，关键参数已做标记，**请忽略模型优化技巧**。

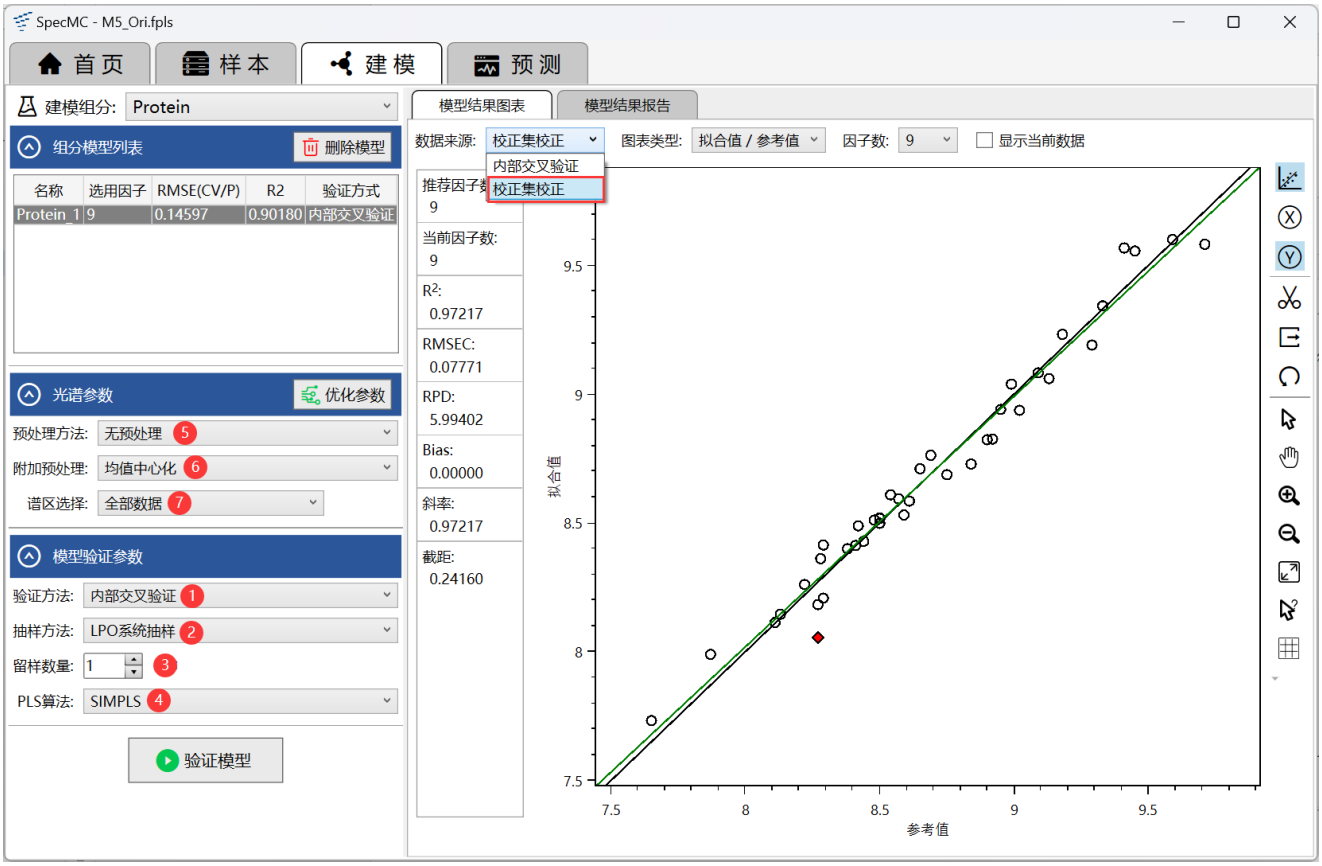

图 6 蛋白含量校正集校正结果图

从“模型结果报告”中一键导出“参考值”和 10 个因子数下的“拟合值”，进行R_1^2、R_2^2和R_3^2的计算，结果见表 11。

表 11 蛋白含量校正结果

	$1LVs$	$2LVs$	$3LVs$	$4LVs$	$5LVs$
R_1^2	0.06186	0.21286	0.62749	0.78174	0.90129
R_2^2	0.06186	0.21286	0.62749	0.78174	0.90129
R_3^2	0.06186	0.21286	0.62749	0.78174	0.90129
SSE	8.14169	6.83117	3.23281	1.89420	0.85670
SSR	0.53681	1.84733	5.44569	6.78430	7.82180
SST	8.67850	8.67850	8.67850	8.67850	8.67850
	$6LVs$	$7LVs$	$8LVs$	$9LVs$	$10LVs$
R_1^2	0.91763	0.95619	0.96112	0.97217	0.97606
R_2^2	0.91763	0.95619	0.96112	0.97217	0.97606
R_3^2	0.91763	0.95619	0.96112	0.97217	0.97606
SSE	0.71486	0.38022	0.33742	0.24155	0.20774
SSR	7.96363	8.29828	8.34108	8.43695	8.47076
SST	8.67850	8.67850	8.67850	8.67850	8.67850

由表 11 可知，不管选择的因子数是多少，对应的R_1^2、R_2^2和R_3^2的数值都相同，同时可以看到，选择 9 个因子时，$R_1^2 = R_2^2 = R_3^2 = 0.97217$，与图 6 中显示的$R^2$一致。

这个结果看起来似乎有些“意外”，要知道，笔者在设计 SpecMC 的PLS算法时，未曾考虑加入“截距项”，难不成前面提到的“关于截距的规律”在偏最小二乘法这里就失灵了？

其实，原因在于“均值中心化”！值得注意的是，均值中心化要求对光谱（自变量，$\boldsymbol{X}$）和浓度（因变量，$\boldsymbol{y}$）同时进行操作。

接下来，我们将“附加预处理”设为“无”，再次计算，其结果见表 12。

表 12 蛋白含量校正结果（无均值中心化）

	$1LVs$	$2LVs$	$3LVs$	$4LVs$	$5LVs$
R_1^2	0.06170	0.05563	0.22102	0.40660	0.66731
R_2^2	2.53338	0.60578	0.57939	1.01652	0.93101
R_3^2	-1.74475	-0.23879	0.13622	0.26901	0.64536
SSE	23.82030	10.75085	7.49633	6.34387	3.07774

$\boldsymbol{SSR}$	21.98595	5.25725	5.02826	8.82187	8.07979
$\boldsymbol{SST}$	8.67850	8.67850	8.67850	8.67850	8.67850
	$\boldsymbol{6LVs}$	$\boldsymbol{7LVs}$	$\boldsymbol{8LVs}$	$\boldsymbol{9LVs}$	$\boldsymbol{10LVs}$
$\boldsymbol{R_1^2}$	0.80956	0.84055	0.85965	0.87184	0.88265
$\boldsymbol{R_2^2}$	1.05642	0.98321	1.01145	1.00185	1.00440
$\boldsymbol{R_3^2}$	0.79312	0.83495	0.85346	0.86731	0.87871
$\boldsymbol{SSE}$	1.79544	1.43237	1.27171	1.15153	1.05260
$\boldsymbol{SSR}$	9.16813	8.53281	8.77791	8.69453	8.71670
$\boldsymbol{SST}$	8.67850	8.67850	8.67850	8.67850	8.67850

不难看出，这一次不管选择的因子数是多少，对应的R_1^2、R_2^2和R_3^2的数值都各不相同；而且R_2^2出现了大于 1 的情况，R_3^2出现了小于 0 的情况。

为了验证这一现象，我们以表 1 的模拟数据为例，同时将$\boldsymbol{x}$和$\boldsymbol{y}$减去各自的平均值，生成新的$\boldsymbol{x}'$和，$\boldsymbol{y}'$，对其进行不含截距的一元线性回归，结果见表 13。

表 13 不含截距的一元线性回归（均值中心化）

序号	x_i	y_i	x_i'	y_i'	$\hat{y}_i$	$\hat{y}_i - y_i$	$\hat{y}_i - \bar{y}$	$y_i - \bar{y}$
1	1.0	9.1	-2.50000	-6.03333	-7.00000	-0.96667	-7.00000	-6.03333
2	2.0	12.0	-1.50000	-3.13333	-4.20000	-1.06667	-4.20000	-3.13333
3	3.0	12.5	-0.50000	-2.63333	-1.40000	1.23333	-1.40000	-2.63333
4	4.0	14.6	0.50000	-0.53333	1.40000	1.93333	1.40000	-0.53333
5	5.0	17.8	1.50000	2.66667	4.20000	1.53333	4.20000	2.66667
6	6.0	24.8	2.50000	9.66667	7.00000	-2.66667	7.00000	9.66667
Mean	**3.50000**	**15.13333**	**0.00000**	**0.00000**	**0.00000**	**0.00000**	**0.00000**	**0.00000**

$\boldsymbol{b_1 = 2.80000}$

$$\boldsymbol{SSE = \sum_{i=1}^{n} (\hat{y}_i - y_i)^2 = 16.79333}$$

$$\boldsymbol{SSR = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 = 137.20000}$$

$$\boldsymbol{SST = \sum_{i=1}^{n} (y_i - \bar{y})^2 = 153.99333}$$

$$\boldsymbol{R_1^2 = R_2^2 = R_3^2 = 0.89095}$$

综合表 13 和表 2 的结果，可以观察到：回归线的斜率b_1、SSE、SSR、SST、R_1^2、R_2^2和R_3^2都未发生变化，一切重归美好！

注 1：如果在 1.2.2 主成分回归的第一步（PCA）选择“Mean center data”，即首先对X进行均值中心化，则有：

1) 执行不带截距的多元线性回归，得到的R_1^2、R_2^2和R_3^2也是不相等的；

2) 如果对y也进行均值中心化，再执行不带截距的多元线性回归，则$R_1^2 = R_2^2 = R_3^2$。

注 2：关于均值中心化对截距的消除作用可以参考 Stack Exchange 网站相关讨论：

- https://stats.stackexchange.com/questions/22329/how-does-centering-the-data-get-rid-of-the-intercept-in-regression-and-pca
- https://stats.stackexchange.com/questions/29781/when-conducting-multiple-regression-when-should-you-center-your-predictor-varia

分割线

一直没关注交叉验证的决定系数，不过，应该不难推断，R_1^2、R_2^2和R_3^2是不会相等的，见下表：

表 14 蛋白模型内部交叉验证结果（均值中心化）

	1LVs	**2LVs**	**3LVs**	**4LVs**	**5LVs**
R_1^2	0.00786	0.12627	0.49050	0.63784	0.79074
R_2^2	0.07334	0.25392	0.58874	0.77151	0.87811
R_3^2	-0.02534	0.10420	0.48572	0.63148	0.78839
	6LVs	**7LVs**	**8LVs**	**9LVs**	**10LVs**
R_1^2	0.85438	0.88817	0.88949	0.90394	0.91437
R_2^2	0.89959	1.01834	0.97399	0.99407	1.00176
R_3^2	0.85376	0.88371	0.88755	0.90180	0.91232

从前面的讨论我们还能观察到，对于线性回归来说，截距项的去除会带来模型性能的下降（***SSE*增大**），我们再来观察一下本例中内部交叉验证结果的$RMSECV$随因子数的变化，见图 7：

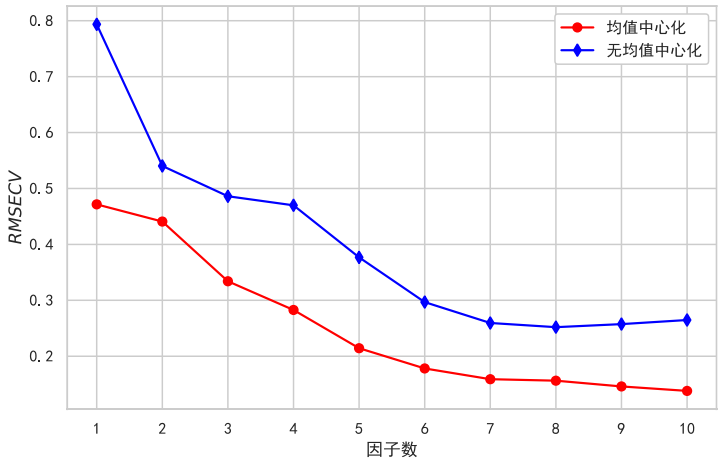

图 7 蛋白含量模型交叉验证均方根误差随因子数的变化

分割线

另一个被众多学者研究过的案例是 M5 的水分含量定量，研究表明：是否选择“均值中心化”对PLS模型的性能影响非常之大。我们同时比较校正和交叉验证的结果，其均方根误差（$RMSEC$和$RMSECV$）变化如图 8 所示，决定系数（SpecMC 采用R_3^2）的变化见图 9。**注意，两条虚线分别对应是否采用均值中心化的校正结果**。

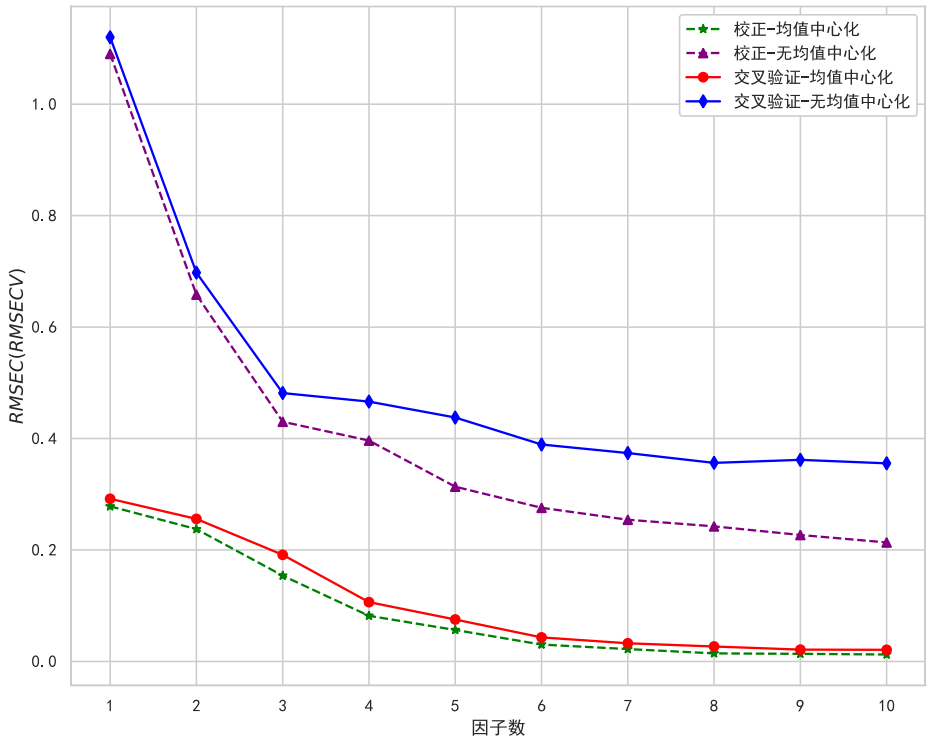

图 8 水分含量模型校正和交叉验证均方根误差变化图

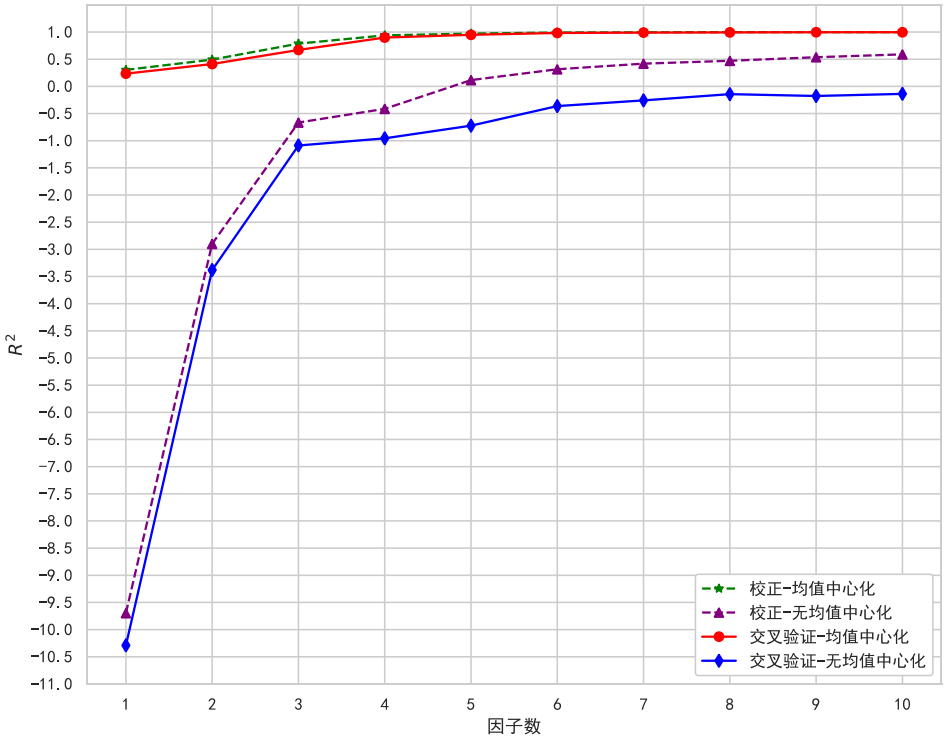

图 9 水分含量模型校正和交叉验证决定系数变化图

※ **关于均值中心化**

通常认为，在近红外光谱定量和定性分析中，采用均值中心化这一预处理方法具有很多益处[8; 9]，它能使建模算法聚焦于光谱变异，而非光谱吸光度值，能增强分辨，增大样品光谱之间的差异。

据笔者观察，MATLAB 自带的 plsregress 函数输入参数中的“Intercept”默认为“true”（返回的回归系数长度比波长变量数大 1），不过很少有其他化学计量学商业软件在*PLS*算法中加入“截距项”。

通过本文前面的讨论，笔者个人认为，从另一个角度来看，选用“均值中心化”能弥补线性回归算法中截距项的“缺失”，从而起到提高模型性能的作用。Camo 的 Unscrambler、Bruker 的 OPUS 以及我们的 SpecMC，都将均值中心化作为一个附加选项（光谱预处理的最后一步），且默认将其开启，意味着用户可以选择与常规的导数、标准正态变量变换、多元散射校正、矢量归一化等预处理方法联用。

※ **关于 Z-Score 标准化**

再聊一个有意思的事实，如果我们选择“Z-Score 标准化”对数据进行预处理，那么“参考值”与“拟合值”之间的回归线的斜率恰好与决定系数的数值相等，相关论述可参考文献[10]。

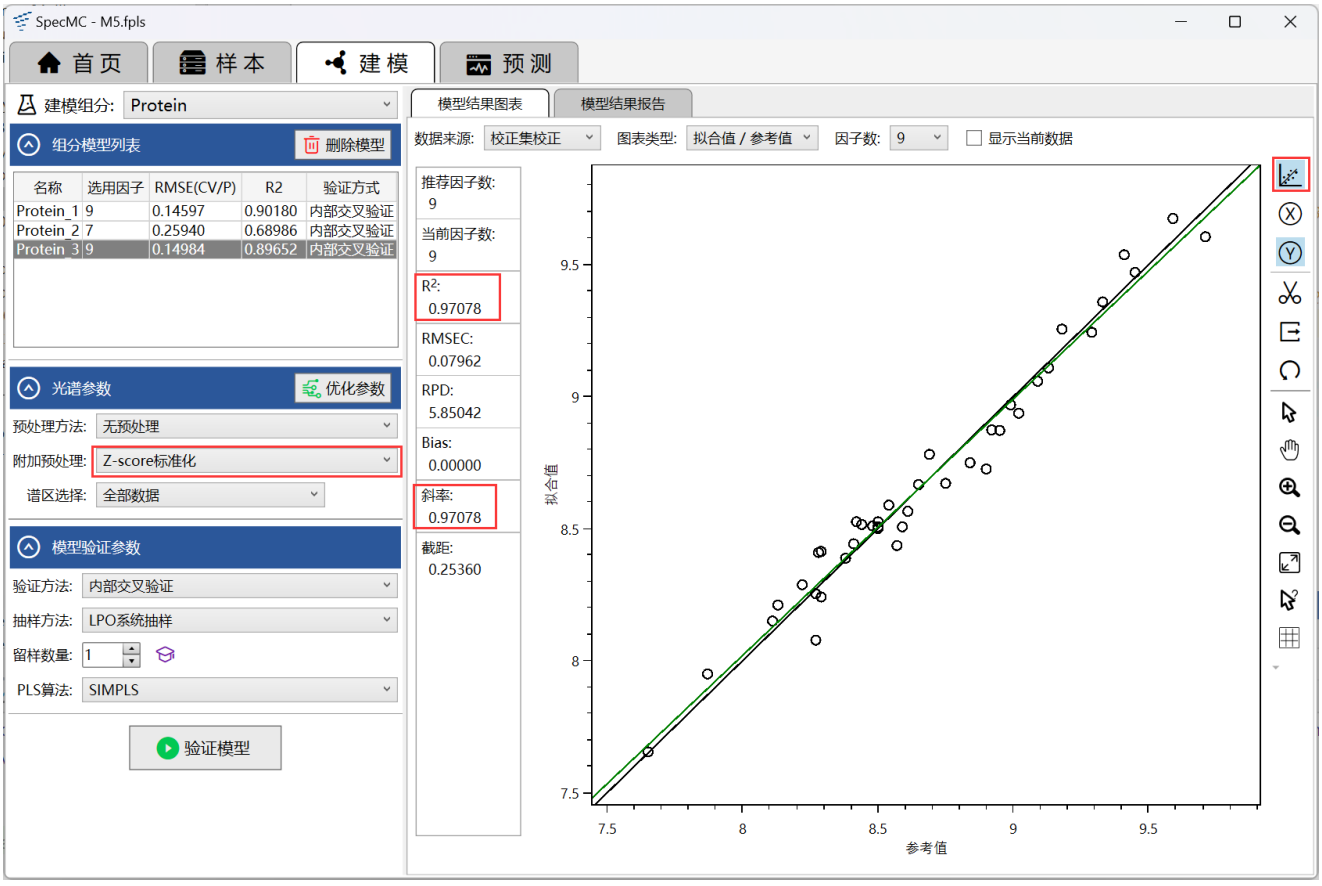

图 10 Z-Score 标准化的效应

注：均值中心化对未知样本的处理过程，不是将预测集减去其平均值，而是减去校正集的均值再代入模型中进行计算，最终的预测结果还需加上校正集浓度或性质的均值。标准化预处理与此类似。

1.2.4 预测阶段的R^2

前面讨论的都是回归分析中的拟合情况，我们再来看看模型预测的结果。

首先，我们引用 Tormod Naes 等所著《A user friendly guide to multivariate calibration and classification》一书 P166 中 Table13.1 的数据，作者假定已经建立了一个线性回归模型$\hat{y} = \sum_{i=1}^{k} b_k x_k + b_0$，随机选择了 7 个参考值已知的测试样本，其预测结果列表如下（前 4 列来自原文，后 2 列为计算所得）：

表 15 线性回归模型预测结果

Obs.	y	$\hat{y}$	$\hat{y}-y$	$\hat{y}-\bar{y}$	$y-\bar{y}$
1	5.2	5.7	0.5	0.3	-0.2
2	7.3	7.4	0.1	2.0	1.9
3	2.3	2.2	-0.1	-3.2	-3.1
4	4.8	5.3	0.5	-0.1	-0.6
5	8.4	8.9	0.5	3.5	3.0
6	6.1	6.5	0.4	1.1	0.7
7	3.7	4.2	0.5	-1.2	-1.7

$\bar{y} = \mathbf{5.4}$

$$\boldsymbol{SSE} = \sum_{i=1}^{n} (\hat{y}_i - y_i)^2 = \mathbf{1.18000}$$

$$\boldsymbol{SSR} = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 = \mathbf{29.24000}$$

$$\boldsymbol{SST} = \sum_{i=1}^{n} (y_i - \bar{y})^2 = \mathbf{26.00000}$$

可得：

$$R_1^2 = \left(\mathrm{CORREL}(y, \hat{y})\right)^2 = (0.99442)^2 = \mathbf{0.98887}$$

细心的读者可能会发现，如果按照 1.1.1 的方式将$\boldsymbol{y}$和$\hat{\boldsymbol{y}}$粘贴到 Excel 插入散点图，显示的R^2值将与R_1^2一致。实际上，这是因为这种方式对$\boldsymbol{y}$（假设为横坐标）和$\hat{\boldsymbol{y}}$（假设为纵坐标）进行了一元线性回归，得到了另外一组拟合值$\hat{y'}$，而一元线性回归的特点是$\left(\mathrm{CORREL}(y, \hat{y})\right)^2$与$\left(\mathrm{CORREL}(y, \hat{y'})\right)^2$的数值相等，这一点我们从前文单变量回归的结论部分可知。

注：部分化学计量学软件提供了显示参考值与拟合值（或预测值）之间一元线性回归关系的功能，比如 Unscrambler（Regression Line）、OPUS（Line）、SpecMC（显示回归线）以及 AQuant（Display regression line）。

$$R_2^2 = \frac{SSR}{SST} = \frac{29.2400}{26.0000} = \mathbf{1.12462}$$

$$R_3^2 = 1 - \frac{SSE}{SST} = 1 - \frac{1.1800}{26.0000} = 1 - 0.04538 = \mathbf{0.95462}$$

显而易见，R_1^2、R_2^2和R_3^2各不相同，$SSR + SSE \neq SST$，$SSR > SST$。

这会是一个孤例吗？

我们又选择了一篇基于支持向量回归的文献中表 4 的数据[11]，含 23 个样本的参考值和预测值（前 4 列来自原文，后 2 列为计算所得）：

表 16 支持向量回归模型预测结果

样本编号	y	$\hat{y}$	$\hat{y}-y$	$\hat{y}-\overline{y}$	$y-\overline{y}$
1	2.95	3.12	0.17	-0.62	-0.79
2	2.90	3.07	0.17	-0.67	-0.84
3	3.10	3.30	0.20	-0.44	-0.64
4	2.92	3.09	0.17	-0.65	-0.82
5	5.99	5.82	-0.17	2.08	2.25
6	2.90	3.07	0.17	-0.67	-0.84
7	2.95	2.78	-0.17	-0.96	-0.79
8	2.94	2.87	-0.07	-0.87	-0.80
9	5.93	5.85	-0.08	2.11	2.19
10	3.01	3.08	0.07	-0.66	-0.73
11	3.01	3.03	0.02	-0.71	-0.73
12	2.95	2.78	-0.17	-0.96	-0.79
13	3.07	3.24	0.17	-0.50	-0.67
14	3.08	2.91	-0.17	-0.83	-0.66
15	2.92	3.09	0.17	-0.65	-0.82
16	2.96	3.13	0.17	-0.61	-0.78
17	2.96	3.03	0.07	-0.71	-0.78
18	3.02	3.19	0.17	-0.55	-0.72
19	5.84	5.67	-0.17	1.93	2.10
20	5.87	5.70	-0.17	1.96	2.13
21	5.95	5.78	-0.17	2.04	2.21
22	2.94	3.11	0.17	-0.63	-0.80
23	2.99	3.16	0.17	-0.58	-0.75
$\overline{y} = 3.62$					

$$SSE = \sum_{i=1}^{n} (\hat{y}_i - y_i)^2 = \mathbf{0.55280}$$

$$SSR = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 = \mathbf{29.03330}$$

$$SST = \sum_{i=1}^{n} (y_i - \bar{y})^2 = \mathbf{33.89790}$$

计算可得：

$$R_1^2 = \left(\mathrm{CORREL}(y, \hat{y})\right)^2 = (0.99457)^2 = \mathbf{0.98917}$$

$$R_2^2 = \frac{SSR}{SST} = \frac{29.03330}{33.89790} = \mathbf{0.85649}$$

$$R_3^2 = 1 - \frac{SSE}{SST} = 1 - \frac{0.55280}{33.89790} = 1 - 0.01631 = \mathbf{0.98369}$$

最后，我们再来看看 1.2.3 所建蛋白含量模型对 40 个测试集样本的预测结果：

1) 均值中心化（因子数 9）：$R_1^2 = 0.95395$，$R_2^2 = 0.78190$，$R_3^2 = 0.94244$

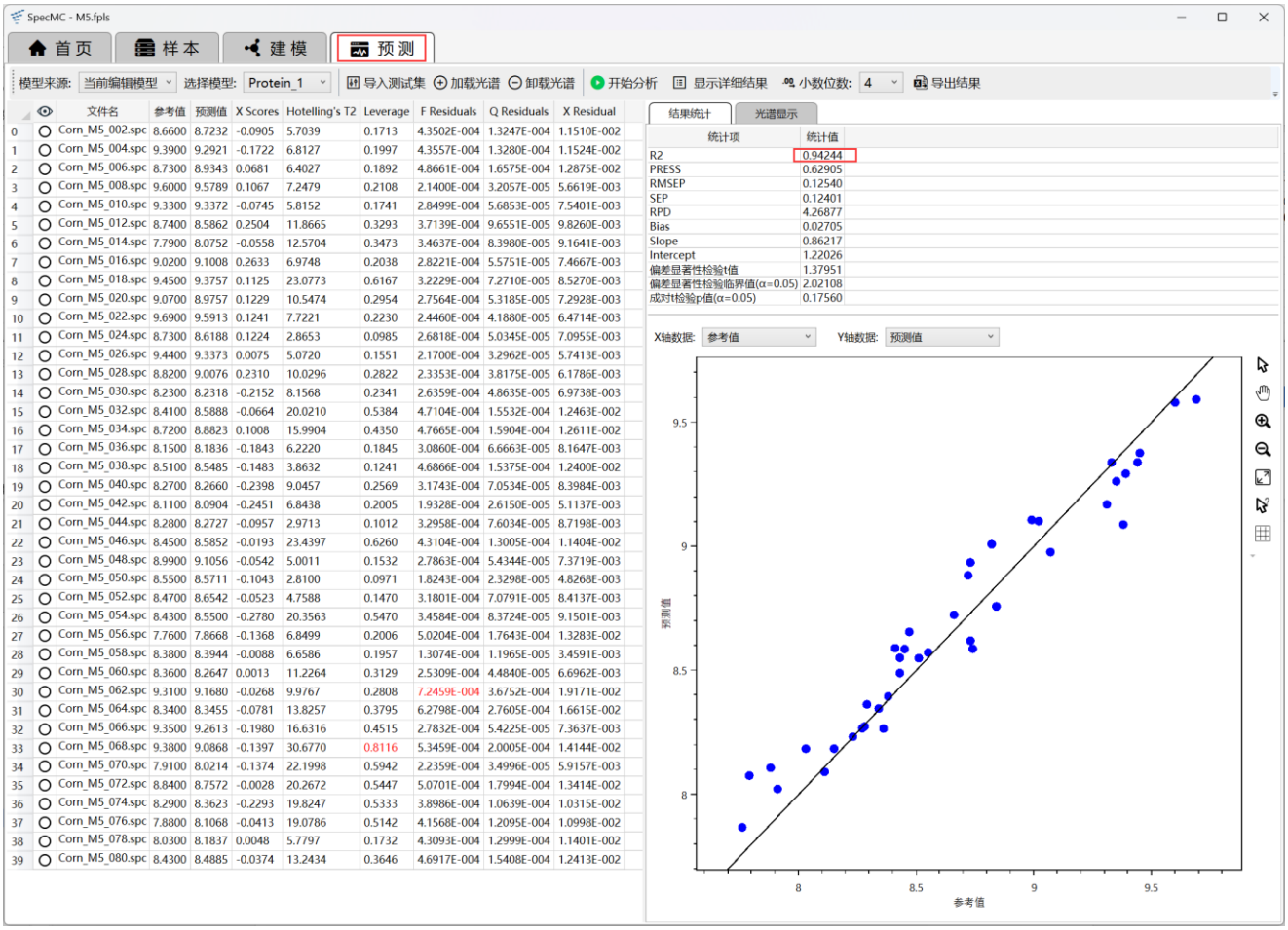

图 11 蛋白模型预测结果（均值中心化）

2) 未均值中心化（因子数 7）：$R_1^2 = 0.69679$，$R_2^2 = 0.85882$，$R_3^2 = 0.68794$

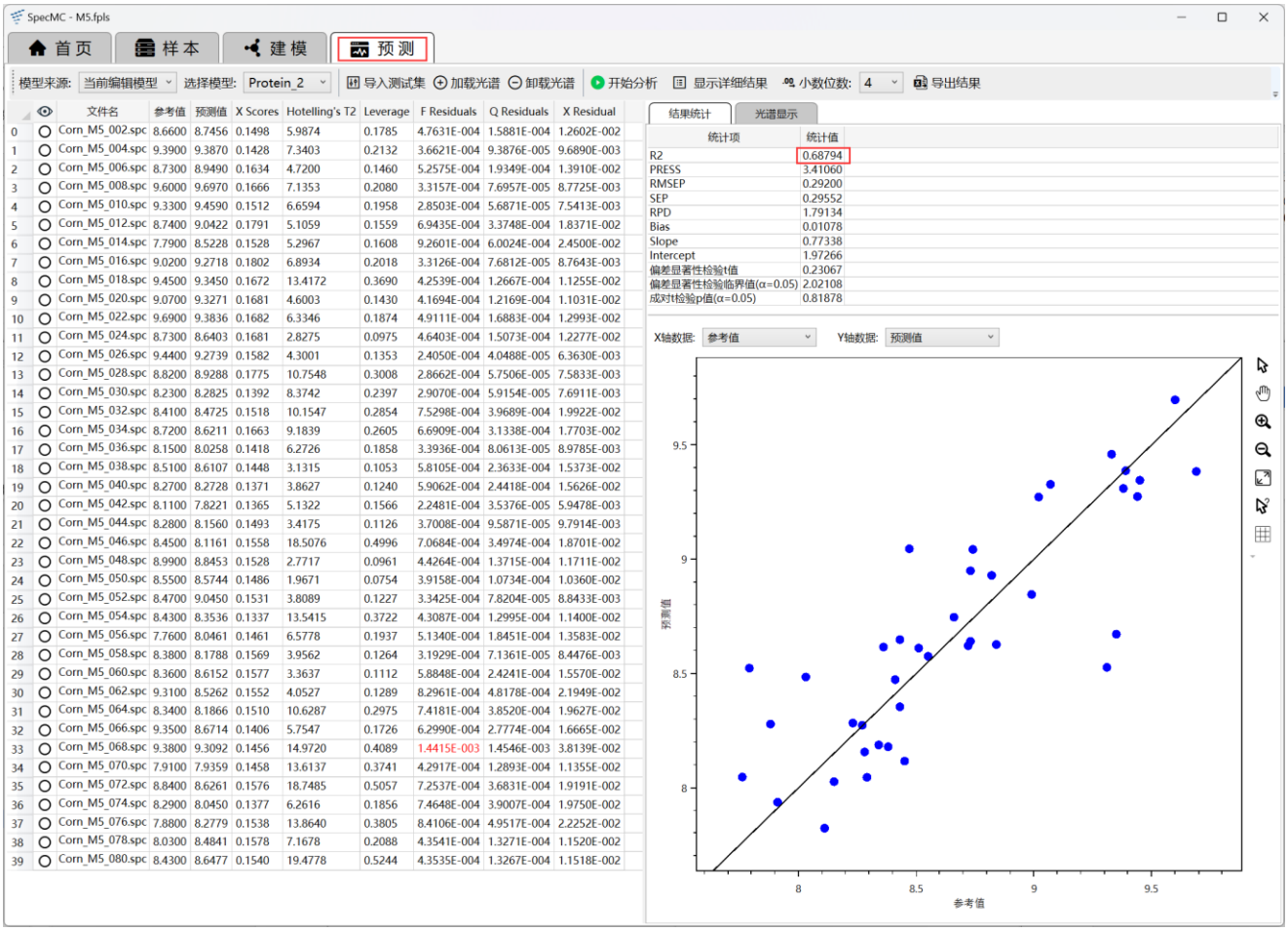

图 12 蛋白模型预测结果（未均值中心化）

1.3 其他非线性回归

笔者尝试了两种非线性回归方法：Unscrambler 的SVR和 Python 编写的$ELM-AE$（极限学习机自编码），结果表明：无论是否进行中心化，R_1^2、R_2^2和R_3^2都不相等。

注：本文主要讨论线性回归方法，感兴趣的读者可以观察其他机器学习和深度学习算法在这个问题上的表现。

1.4 结论

终于，我们可以对这一问题下个结论了：

1) R_1^2=R_2^2=R_3^2这一等式仅适用于模型校正（拟合）阶段，不适用于交叉验证和预测阶段;

2) 不同的回归算法及算法参数在这一问题上的表现是不同的；

3) 对于线性回归算法而言，是否含有截距（常数项）会显著影响这个等式的成立与否；

4) 均值中心化能够抵消线性回归算法中“截距项”缺失带来的影响，进而提高模型性能。

同时，笔者建议在公开发表的文章中应列出所用的决定系数的公式。

2. R^2的取值范围

有了前文的铺垫，这一问题解决起来变得简单了。

1) 从R_1^2的定义可知，它是简单相关系数或复相关系数的平方，所以$\boldsymbol{R_1^2 \in [0,1]}$

2) $R_2^2 = \frac{SSR}{SST}$，某些情况下（比如对某些数据集进行不含截距的回归）$SSR > SST$，因此$\boldsymbol{R_2^2 \in [0,+\infty)}$

3) $R_3^2 = 1 - \frac{SSE}{SST}$，通过前面的讨论，我们已经多次看到$SSE > SST$的情况，故而$\boldsymbol{R_3^2 \in (-\infty,1]}$。在长期的工作实践中，笔者也碰到过此类情况，比如：当光谱和参考值之间的相关很弱（不具备近红外定量分析的理论基础）、较差的参考值测量精密度、较差的光谱采集条件、参考值与样本的对应出错等等时，如果$R_3^2 < 0$是可以理解的。

R_3^2的取值范围“似乎”有些违背常理，读者可以参考文献[3; 12]、Stack Exchange 网站的讨论及 Origin 软件的帮助网页：

- https://stats.stackexchange.com/questions/183265/what-does-negative-r-squared-mean
- https://stats.stackexchange.com/questions/12900/when-is-r-squared-negative
- https://www.originlab.com/doc/Origin-Help/Details_of_R_square

3. R^2的选择

作为评估模型拟合优度或模型预测性能的统计量，好的决定系数定义最好能满足以下几个要求[3]：

- 具有好的可解释性
- 无量纲，与变量的单位无关
- 取值范围$\in [0,1]$，1 代表完美拟合，0 代表完全未拟合
- 能适用于所有类型的模型
- 不同模型拟合同一数据集的R^2具有可比性
- 兼容其它反映拟合优度的统计量，比如反映残差/误差的SSE、SEC/SEP、$RMSEC/RMSEP$等；也就是说当SSE增大时，R^2的值应当减小。

然而，遗憾的是，R_1^2、R_2^2和R_3^2都无法完全满足以上条件。

1) R_1^2的问题在于：当$\hat{y}$进行了线性变换后（亦即$\hat{y'} = b_1\hat{y} + b_0$），仍有$RSQ(y,\hat{y'}) = RSQ(y,\hat{y})$（尽管此时的$SSE$已经发生了变化），这一点可以参考本文 1.1 小结部分的第 5 条。

注：韩东海老师在今日头条发表的《模型精度评价》中提到的不能只用相关系数评价模型精度，与此同理，详见：https://www.toutiao.com/article/7205068518520324669/?log_from=312cabf81dd54_1695106205752

2) R_2^2有可能超过 1，没有明显的最大值，显然这不是一个好的性质。

3) R_3^2有可能小于 0，不能满足$0 \le R^2$，不过这不是一个大麻烦。因为当$R_3^2 < 0$时，提示着模型存在较大的问题，而且，我们不必对两个“垃圾”模型进行比较；反而正是由于R_3^2具备这一特性，能够对我们的建模过程起到很好的“示警”作用。因此，笔者个人倾向于使用R_3^2，这也体现在我们开发的光谱多元校正软件 SpecMC 和智能定量建模软 AQuant 中。

下面，我们以偏最小二乘回归为例，列举几个化学计量学软件选用的决定系数定义，见表 17：

表 17 不同化学计量学（统计学）软件选用的决定系数

软件	Unscrambler	OPUS	TQ Analyst	Origin	SpecMC	AQuant
$\boldsymbol{R^2}$	$R_1^2 or R_2^2 or R_5^2$	R_3^2	$\sqrt{R_1^2}$ **	R_3^2	R_3^2	R_3^2

注 1：Unscrambler 提供的R_1^2表现为$R2(Pearson)$；提供的$R-Squared$使用均值中心化时采用R_2^2，不使用均值中心化时采用$R_5^2 = 1 - SSE/\sum(y_i)^2$。

注 2：**ThermoFisher 的 TQ Analyst 选择使用相关系数

注 3：有些软件还提供R_{adjust}^2，用于消除线性回归模型中自变量（或隐变量）数目的影响，此处不讨论。

4. R^2的“欺骗性”

我们通过模型预测阶段的结果来展示R^2的“欺骗性”。选择来自 https://eigenvector.com/resources/data-sets/中的柴油样本的近红外数据集，该数据由 Soutwest Research Institute（SWRI）授权 Eigenvector 公开。光谱波长范围 750~1550 nm，波长间隔 2 nm，共 401 个数据点，包含 50%馏程、十六烷值、密度（@15℃）、凝点、总芳烃及粘度（@40℃）等 6 种组分（性质）。该网站事先已将光谱数据随机分成了训练集（**_ll_a**+20 个高杠杆值样本**_hl**）和测试集（**_ll_b**）供算法测试，同时也提供了经过一阶导数处理后的光谱数据，部分光谱如图 13 所示：

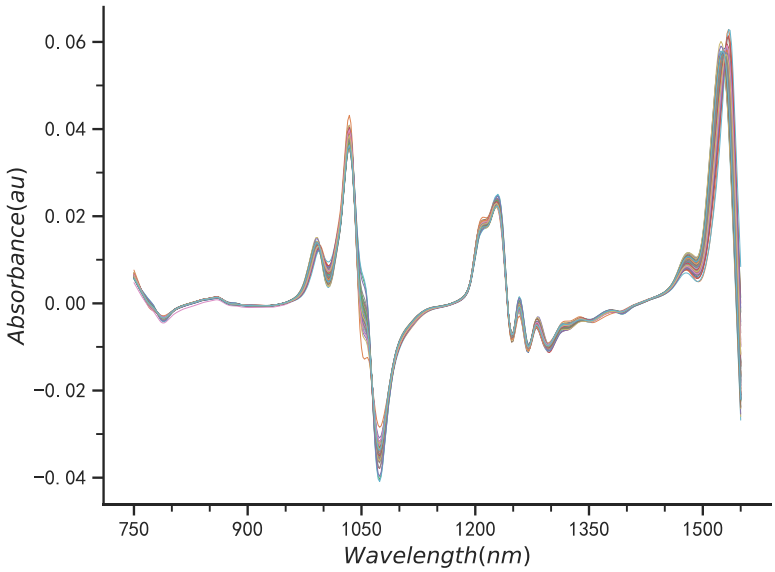

图 13 部分柴油样本近红外光谱图

我们以总芳烃（Total aromatics）质量百分数含量为例，方便起见（读者可以互换校正集和训练集进行测试），去除了其中的高杠杆值样本（**_hl**），最终包含 118 个校正集样本（**_ll_a**）和 118 个测试集样本（**_ll_b**），统计数据如下表所示：

表 18 柴油样本总芳烃数据统计表

	样本数	最小值	最大值	极差	平均值	标准偏差
校正集	118	14.3	44.2	29.9	31.32	5.93
测试集	118	13.7	42.8	29.1	31.60	6.07

由上表可知，校正集和测试集数据分布基本一致。我们采用 Python（numpy & scipy）编写相关代码，PLS采用$simpls$算法，为方便读者重现，仅对光谱（网站提供的数据已经过一阶导）进行均值中心化预处理，同时，不对波长变量进行优选，最大因子数设为 20，采用 10 折交叉验证（每一折的样本采用系统抽样来选择）来确定最佳因子数，**请忽略模型优化技巧**。$RMSEC$及$RMSECV$随因子数的变化见图 14，由图中$RMSECV$的趋势可知，选择 12 个因子是比较合理的。

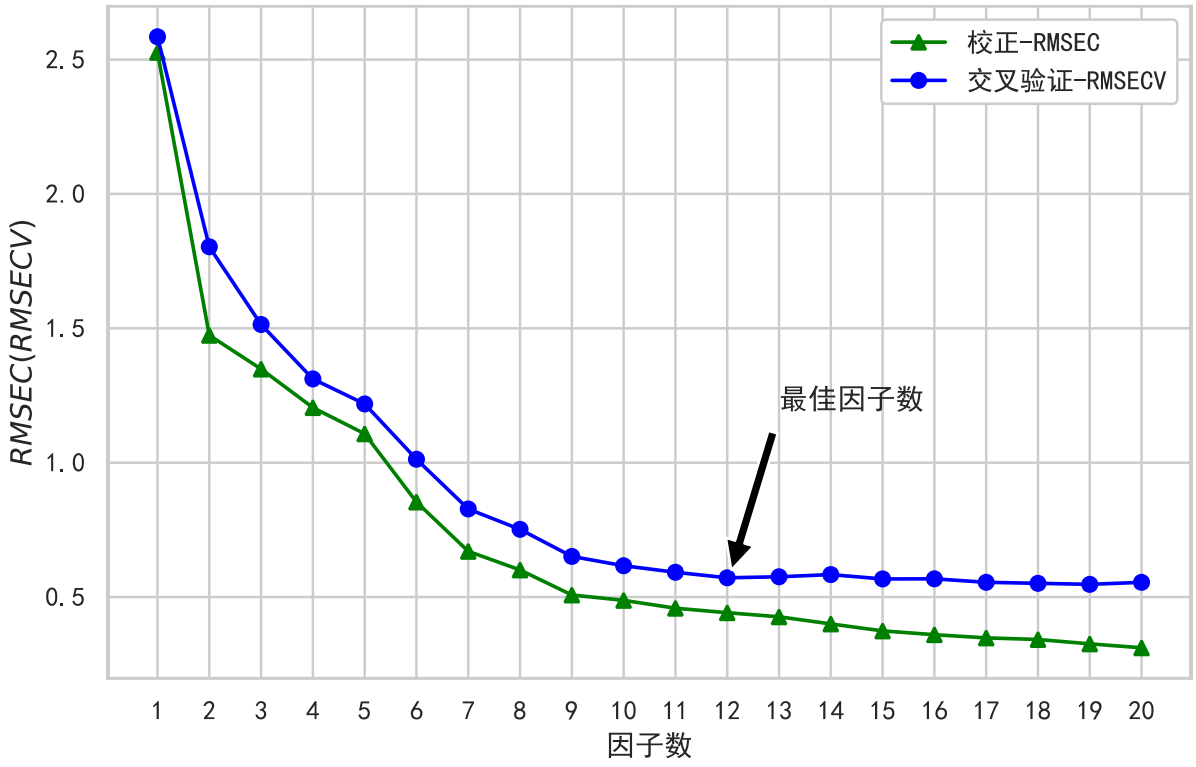

图 14 总芳烃 RMSEC 及 RMSECV 趋势图

内部交叉验证最佳因子数（$optimal\_nlv = 12$）下的预测值（y 轴）和参考值（x 轴）的散点图见图 15，R^2为 0.99062，$RMSECV$为 0.57174，预测值和参考值进行一元线性回归的斜率为 0.99054，截距为 0.30767，模型性能还是不错的。

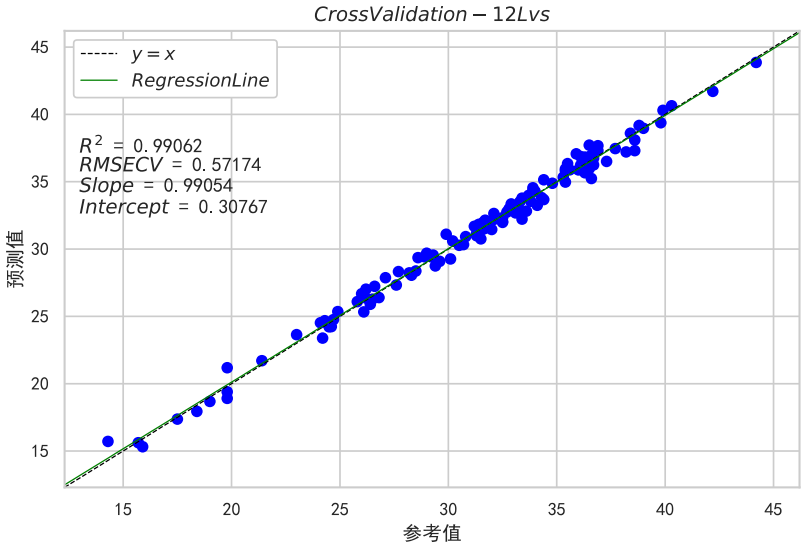

图 15 总芳烃内部交叉验证散点图

为阐述本节的观点，我们将测试集的参考值由低到高进行排序，然后选择排在中间的k个样本进行预测，k分别设为 30（Range-1），55（Range-2），75（Range-3）和 118（Range-Full，即全部测试集样本），其预测结果分别见图 16 的a)，b)，c)和d)。

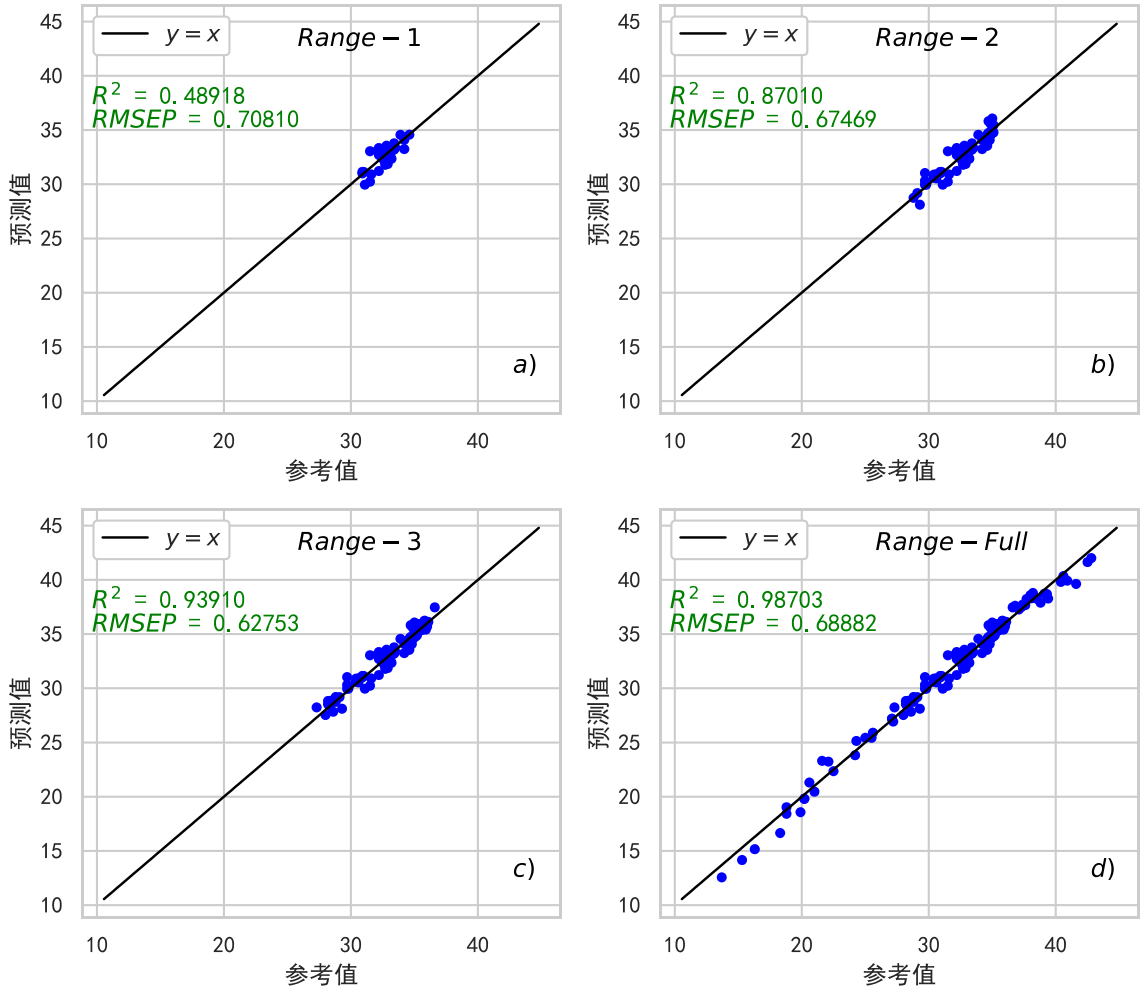

图 16 不同含量范围总芳烃预测结果

图 16 中的数据给我们的第一感觉是R^2差异甚大，且其数值依次增大，分别为 0.48918，0.87010，0.93910 和 0.98703，同一模型预测同一测试集中不同样本的结果迥异，到底发生了什么？这就要从决定系数R_3^2的公式说起，再来回顾一下：

$$R_3^2 = 1 - \frac{SSE}{SST} = 1 - \frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{3}$$

我们分别计算预测结果中的SSE，SST及其他指标，见表 19：

表 19 不同含量范围总芳烃预测结果

	Range-1	**Range-2**	**Range-3**	**Range-Full**
$\boldsymbol{SSE}$	15.04196	25.03622	29.53437	55.98760
$\boldsymbol{SST}$	29.44667	192.73745	484.93387	4317.27695
$\boldsymbol{k}$	30	55	75	118
$\boldsymbol{SSE/k}$	0.50140	0.45520	0.39379	0.47447
$\boldsymbol{SST/k}$	0.98156	3.50432	6.46578	36.58709
参考值的平均值	32.6	32.6	32.4	31.6
参考值的极差	3.7	6.3	9.3	29.1

由上表可以观察到，随着样本数量的增加，SSE和SST整体都呈现上升趋势，但SST的增长速度明显比SSE的增长速度快。这一点可以从SSE/k（又称MSE，均方误差）和SST/k（参考值的方差，准确地说应该叫总体方差，该式分母为k）的数值变化得知，SSE/k基本比较平稳，而SST/k则成倍地快速增长。

所以，问题的关键在于SST/k的区别。大家都知道，方差是衡量随机变量或一组数据离散程度的度量，反映了数据集的分布程度，数据分布越广，相对于均值的方差就越大，且与均值的大小无关。换句话说，在预测结果的均方误差差异不大的情况下，数据集参考值的变异越大（方差越大，离散程度越大，公式 3 中分式部分分母越大），则R^2越大——**$\boldsymbol{R^2}$的大小严重依赖于样本集参考值的分布情况**。

为了验证这一事实，我们人为将测试集中参考值最小的样本和参考值最大的样本加入 Range-1 样本集成为 Range-1-modified，修改后参考值的均值为 32.4（与 Range-1 的差异很小），极差为 29.1（与 Range-Full 相同），总体方差为 15.27741（为 Range-1 中SST/k的 15.6 倍）。采用原模型预测后，预测值和参考值的散点图如下：

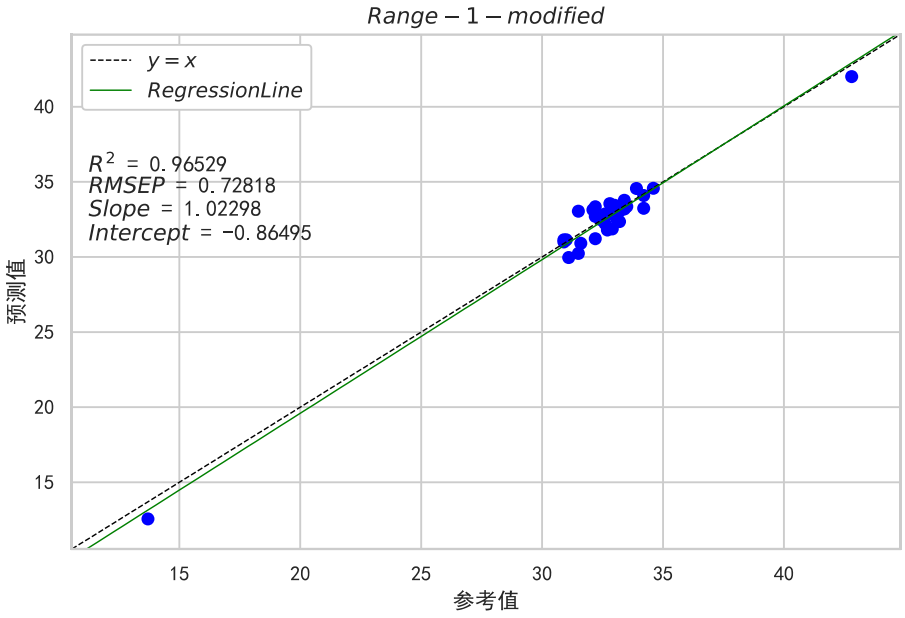

图 17 修改 Range-1 数据集后的预测结果

我们仅仅只是在测试集中增加了两个样本，决定系数瞬时就从 0.48918 变成了 0.96529，**是模型突然变好了吗？显然不是！！！**

我想，此刻读者们应该都能感觉到，数字有时候也会“撒谎”！我们能怎么办？是时候请出另一个常用的统计指标——预测均方根误差$RMSEP$，其计算公式如下：

$$RMSEP = \sqrt{\frac{\sum_{i=0}^{n}(\hat{y}_i - y_i)^2}{n}} = \sqrt{\frac{SSE}{n}} \tag{4}$$

式中：n为测试集样本数量。其实，表 15 已经帮我们完了计算$RMSEP$的大部分工作，只需取SSE/k的平方根即可。除此以外，我们还能计算其他一些统计量，比如平均绝对误差（Mean Absolute Error，MAE）、平均相对误差（Mean Relative Error，MRE），详情见表 20：

表 20 不同含量范围总芳烃预测结果统计表

	Range-1	Range-2	Range-3	Range-Full	Range-1-Modified
R^2	0.48918	0.87010	0.93910	**0.98703**	0.96529
$RMSEP$	0.70810	0.67469	**0.62753**	0.68882	0.72818
MAE	0.56543	0.53096	**0.49706**	0.54555	0.59045
MRE	1.74498%	1.63659%	**1.54526%**	1.88746%	1.95375%

由表 20 可知，上述 5 种不同含量范围样本集预测结果的$RMSEP$、MAE、MRE都基本相当，意味着模型预测性能相当，看来仅靠R^2来评估模型是不够的，应该综合观察多个统计量的结果。我们已将表 20 中表现最好的结果加粗显示，还可以观察到另外一个现象，Range-3 样本集除了R^2比 Range-Full 的R^2稍

小（毫不意外）外，其余指标均领先于其他样本集，这些都是值得研究的地方。

通过上面的讨论，我们还能意识到，测试集（或称验证集）样本的选择也须满足一定要求，即样本应当具有代表性，具体比如：其浓度范围应尽可能覆盖校正集样本参考值的范围、样本的数量不能过少等等。

…… 分割线 ……

本节所涉及的数据及 Python 代码均已上传至笔者在 GitHub 社区维护的光谱多元校正算法库 https://github.com/freesiemens/SpectralMultivariateCalibration，SpecMC 软件的核心算法也在库中，欢迎交流。

…… 分割线 ……

由于本人水平和经验有限，文中难免有错漏之处，偶有观点也仅代表个人意见，敬请读者批评指正。

参考文献

[1] Wright S. Correlation and Causation[J]. Journal of Agricultural Research, 1921, XX, No.7: 557-585.
[2] Marquardt D, Snee R. Test Statistics for Mixture Models[J]. Technometrics, 1974, 16: 533-537.
[3] Kvalseth T O. Cautionary Note about R2[J]. The American Statistician, 1985, 39(4): 279-285.
[4] Ron L, Betsy F. Elementary Statistics[M]. Pearson Education Limited, 2018.
[5] Strang G. Introduction to Linear Algebra[M]. 2009.
[6] Freedman D, Pisani R, Purves R. Statistics[M]. W.W. Norton & Company, 2007.
[7] 褚小立, 陈瀑, 李敬岩, et al. 近红外光谱分析技术的最新进展与展望[J]. 分析测试学报, 2020, 39(10): 8.
[8] Ciurczak E W, Igne B, Jerome Workman J, et al. Handbook of Near-Infrared Analysis[J], 2021.
[9] Ozaki Y, Huck C, Tsuchikawa S, et al. Near-Infrared Spectroscopy - Theory, Spectral Analysis, Instrumentation, and Applications[M]. Springer Singapore, 2021.
[10] Fearn T. r2 and R2[J]. NIR news, 2000, 11(1): 14-15.
[11] 王小亮, 杨静, 梁亚伟, et al. 基于近红外分析技术快速测定盐酸氨溴索口服液的含量[J]. 药品评价, 2022, 19(07): 403-406.
[12] Chicco D, Warrens M J, Jurman G. The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation[J]. PeerJ Comput Sci, 2021, 7: e623.