多元回归的应用与分析

蒋冰

学号：201322130002

1 多元线性回归模型介绍

1.1 问题引入

回归分析是利用大量的观测数据来确定变量与变量之间的统计相关关系的一种数理统计方法。

由于事物间的联系常常是多方面的，一个因变量的变化可能受到其它多个自变量的影响，如糖尿病患者的血糖变化可能受胰岛素、糖化血红蛋白、血清总胆固醇、甘油三脂等多种生化指标的影响。因此，多元回归分析，作为一种以多个自变量估计因变量的线性关系的方法，常常被用来解释和预报因变量的变化。

1.2 多元线性回归模型介绍

在多元回归分析中，因变量只有 1 个，常用变量 Y 表示，自变量往往不止 1 个，通常设为 m 个（$m \geq 2$），分别记作 $X_1, X_2, \cdots, X_m$，共 m+1 个变量。设观测样本含量为 n。观测数据格式如下图所示：

序号 i	X_1	X_2	...	X_m	Y
1	X_{11}	X_{12}	...	X_{1m}	Y_1
2	X_{21}	X_{22}	...	X_{2m}	Y_2
⋮	⋮	⋮		⋮	⋮
n	X_{n1}	X_{n1}	· · ·	X_{nm}	Y_n

假设条件：

1. Y 与 $X_1, X_2, \cdots, X_m$ 之间具有线性关系。
2. 各例观测值 $Y_i (i = 1, 2, \cdots, n)$ 相互独立。
3. 残差 ε 服从均值为 0，方差为 σ^2 的正态分布，它等价于对任意一组自变量 $X_1, X_2, \cdots, X_m$ 的值，因变量 Y 具有相同方差，并且服从正态分布。

回归模型的一般形式如下所示：$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_m X_m + \varepsilon$。式子的含义为数据中因变量 Y 可以近似地表示为自变量 $X_1, X_2, \cdots, X_m$ 的线性函数。β_0 为常数项，$\beta_1, \beta_2, \cdots, \beta_m$ 为偏回归系数，表示在其它自变量保持不变时，X_j 增加或减少一个单位时 Y 的平均变化量，ε 是去除 m 个自变量对 Y 影响后的随机误差（残差）。

1.3 多元线性回归系数的最小二乘估计

设 $\hat{\beta_0}, \hat{\beta_1}, \hat{\beta_2}, \cdots, \hat{\beta_m}$ 为 $\beta_0, \beta_1, \beta_2, \cdots, \beta_m$ 的最小二乘估计，则 y 的 n 次观测值可写为如下形式：

$$\begin{cases} Y_1 = \hat{\beta_0} + \hat{\beta_1} X_{11} + \hat{\beta_2} X_{12} + \cdots + \hat{\beta_m} X_{1m} + \hat{\varepsilon_1} \\ Y_2 = \hat{\beta_0} + \hat{\beta_1} X_{21} + \hat{\beta_2} X_{22} + \cdots + \hat{\beta_m} X_{2m} + \hat{\varepsilon_2} \\ \cdots \\ Y_n = \hat{\beta_0} + \hat{\beta_1} X_{n1} + \hat{\beta_2} X_{n2} + \cdots + \hat{\beta_m} X_{nm} + \hat{\varepsilon_n} \end{cases}$$

其中 $\beta_1, \beta_2, \cdots, \beta_m$ 是未知参数。其中$\hat{\varepsilon_i}$ 为误差 ε_i 的估计值，称为残差。

采用矩阵形式可以设

$$Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \ \beta = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_n \end{bmatrix}, \ X = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1m} \\ 1 & x_{21} & x_{22} & \cdots & x_{2m} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nm} \end{bmatrix}, \ \varepsilon = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix},$$

则多元模型可以表示为矩阵形式 $Y = X\beta + \varepsilon$

令 $\hat{Y}$ 为 Y 的估计值，则有 $\hat{Y_i} = \hat{\beta_0} + \hat{\beta_1} X_{i1} + \hat{\beta_2} X_{i2} + \cdots + \hat{\beta_m} X_{im}$，$\hat{\varepsilon_i} = Y_i - \hat{Y_i}$，为使得估计值与实际值拟合的最好, 则应使残差平方和 $Q(\hat{\beta_1}, \hat{\beta_2}, \cdots, \hat{\beta_m}) = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (Y_i - \hat{Y_i})^2 = \sum_{i=1}^{n} [Y_i - (\hat{\beta_0} + \hat{\beta_1} X_{i1} + \hat{\beta_2} X_{i2} + \cdots + \hat{\beta_m} X_{im})]^2$ 达到最小。用矩阵形式表达就是求使得残差平方和 $||\varepsilon||^2 = \varepsilon^T \varepsilon = (Y - X\beta)^2 (Y - X\beta)$ 达到最小的 β 值。因此，对上式对$\hat{\beta_0}, \hat{\beta_1}, \hat{\beta_2}, \cdots, \hat{\beta_m}$ 分别求导有：

$$\begin{cases} \frac{\partial Q}{\partial \beta_0} = -2 \sum_{i=1}^{n} (Y_i - \hat{Y_i}) = 0 \\ \frac{\partial Q}{\partial \beta_j} = -2 \sum_{i=1}^{n} (Y_i - \hat{Y_i}) X_{ij} = 0 \quad j = 1, 2, \cdots, m \end{cases}$$

即 $$\begin{cases} \sum_{i=1}^{n} (Y_i - \hat{\beta_0} - \hat{\beta_1} X_{i1} - \hat{\beta_2} X_{i2} - \cdots - \hat{\beta_m} X_{im}) = 0 \\ \sum_{i=1}^{n} (Y_i - \hat{\beta_0} - \hat{\beta_1} X_{i1} - \hat{\beta_2} X_{i2} - \cdots - \hat{\beta_m} X_{im}) X_{ij} = 0 \quad j = 1, 2, \cdots, m \end{cases}$$

整理并化简则得到下列的正规方程组

$$\begin{cases} n\beta_0 + (\sum_{i=1}^{n} X_{i1}) \hat{\beta_1} + (\sum_{i=1}^{n} X_{i2}) \hat{\beta_2} + \cdots + (\sum_{i=1}^{n} X_{im}) \hat{\beta_m} = \sum_{i=1}^{n} Y_i \\ (\sum_{i=1}^{n} X_{i1}) \beta_0 + (\sum_{i=1}^{n} X_{i1}^2) \hat{\beta_1} + (\sum_{i=1}^{n} X_{i1} X_{i2}) \hat{\beta_2} + \cdots + (\sum_{i=1}^{n} X_{i1} X_{im}) \hat{\beta_m} = \sum_{i=1}^{n} X_{i1} Y_i \\ \cdots \\ (\sum_{i=1}^{n} X_{im}) \beta_0 + (\sum_{i=1}^{n} X_{i1} X_{im}) \hat{\beta_1} + (\sum_{i=1}^{n} X_{i1} X_{im}) \hat{\beta_2} + \cdots + (\sum_{i=1}^{n} X_{im}^2) \hat{\beta_m} = \sum_{i=1}^{n} X_{im} Y_i \end{cases}$$

记方程的系数矩阵为 A，常数项矩阵为 B，则有

$$A=\begin{bmatrix} n & \sum_{i=1}^{n}X_{i1} & \sum_{i=1}^{n}X_{i2} & \cdots & \sum_{i=1}^{n}X_{im} \\ \sum_{i=1}^{n}X_{i1} & \sum_{i=1}^{n}X_{i1}^2 & \sum_{i=1}^{n}X_{i1}X_{i2} & \cdots & \sum_{i=1}^{n}X_{i1}X_{im} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \sum_{i=1}^{n}X_{im} & \sum_{i=1}^{n}X_{i1}X_{im} & \sum_{i=1}^{n}X_{i1}X_{im} & \cdots & \sum_{i=1}^{n}X_{im}^2 \end{bmatrix}$$

$$=\begin{bmatrix} 1 & 1 & \cdots & 1 \\ X_{11} & X_{21} & \cdots & X_{n1} \\ X_{12} & X_{22} & \cdots & X_{n2} \\ \cdots & \cdots & \cdots & \cdots \\ X_{1m} & X_{2m} & \cdots & X_{nm} \end{bmatrix}\begin{bmatrix} 1 & X_{11} & X_{12} & \cdots & X_{1m} \\ 1 & X_{21} & X_{22} & \cdots & X_{2m} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ 1 & X_{n1} & X_{n2} & \cdots & X_{nm} \end{bmatrix}=X^TX$$

$$B=\begin{bmatrix} \sum_{i=1}^{n}Y_i \\ \sum_{i=1}^{n}X_{i1}Y_i \\ \vdots \\ \sum_{i=1}^{n}X_{im}Y_i \end{bmatrix}=\begin{bmatrix} 1 & 1 & \cdots & 1 \\ X_{11} & X_{21} & \cdots & X_{n1} \\ X_{12} & X_{22} & \cdots & X_{n2} \\ \cdots & \cdots & \cdots & \cdots \\ X_{1m} & X_{2m} & \cdots & X_{nm} \end{bmatrix}\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}$$

因此正规方程的举证形式为 $(X^TX)\beta=X^TY$，如果系数矩阵 A 满秩，则 A^{-1} 存在，此时易求得，β 的最小二乘估计值为 $\hat{\beta}=\left(X^TX\right)^{-1}X^TY$

从而可得到经验回归方程为：$\hat{Y}=\hat{\beta_0}+\hat{\beta_1}X_1+\hat{\beta_2}X_2+\cdots+\hat{\beta_m}X_m$

2 模型的测算分析

2.1 数据来源

数据为 27 名糖尿病患者的血清总胆固醇、甘油三酯、空腹胰岛素、糖化血红蛋白和空腹血糖的测量值。数据如下图所示：

序号 i	Y	X_1	X_2	X_3	X_4
1	11.2	5.68	1.9	4.53	8.2
2	8.8	3.79	1.64	7.32	6.9
3	12.3	6.02	3.56	6.95	10.8
4	11.6	4.85	1.07	5.88	8.3
5	13.4	4.6	2.32	4.05	7.5
6	18.3	6.05	0.64	1.42	13.6
7	11.1	4.9	8.5	12.6	8.5
8	12.1	7.08	3	6.75	11.5
9	9.6	3.85	2.11	16.28	7.9
10	8.4	4.65	0.63	6.59	7.1
11	9.3	4.59	1.97	3.61	8.7
12	10.6	4.29	1.97	6.61	7.8
13	8.4	7.97	1.93	7.57	9.9
14	9.6	6.19	1.18	1.42	6.9
15	10.9	6.13	2.06	10.35	10.5
16	10.1	5.71	1.78	8.53	8
17	14.8	6.4	2.4	4.53	10.3
18	9.1	6.06	3.67	12.79	7.1
19	10.8	5.09	1.03	2.53	8.9
20	10.2	6.13	1.71	5.28	9.9
21	13.6	5.78	3.36	2.96	8
22	14.9	5.43	1.13	4.31	11.3
23	16	6.5	6.21	3.47	12.3
24	13.2	7.98	7.92	3.37	9.8
25	20	11.54	10.89	1.2	10.5
26	13.3	5.84	0.92	8.61	6.4
27	10.4	3.84	1.2	6.45	9.6

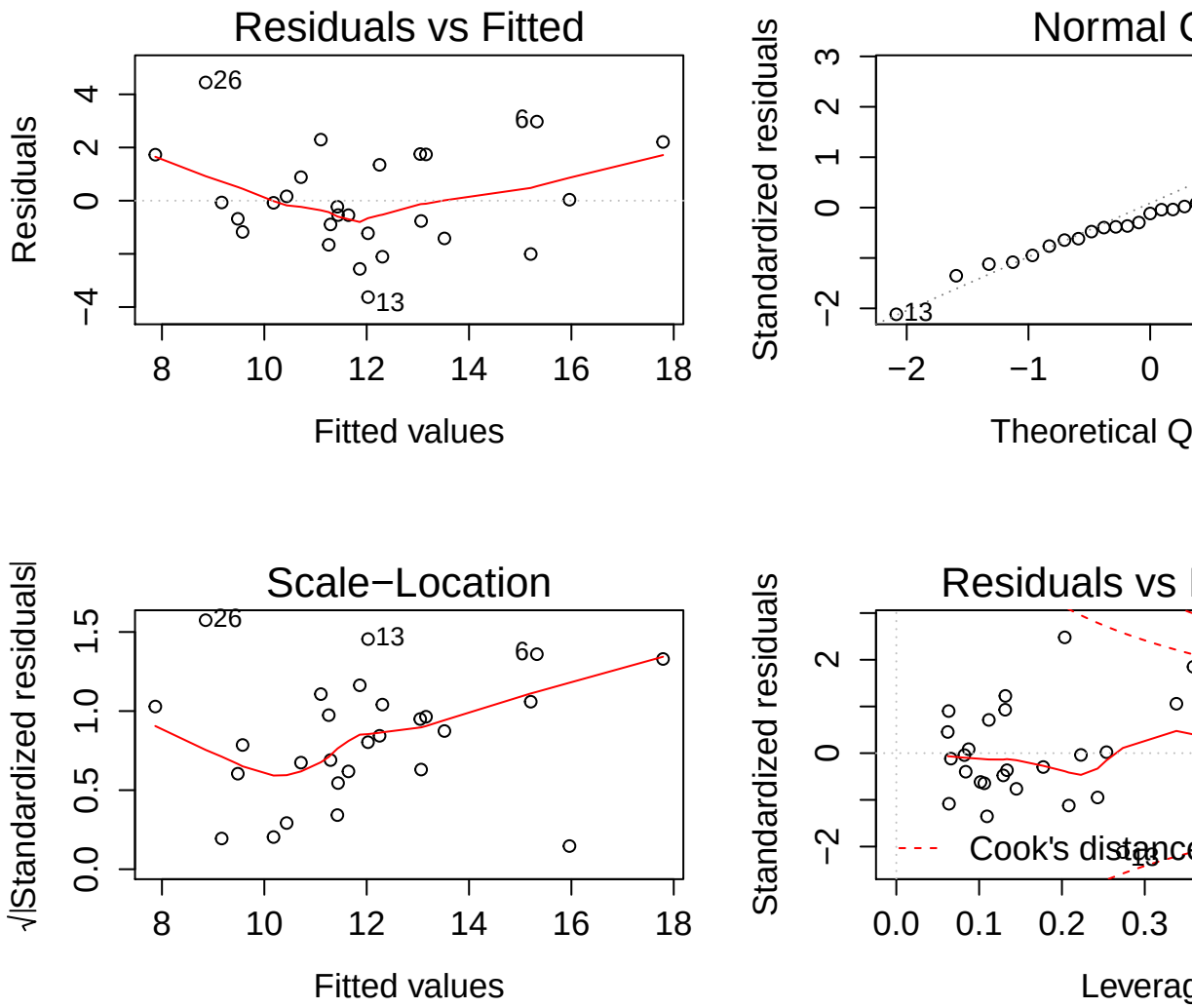
变量说明如下：
Y：空腹血糖的含量 (mmol/L)；
X_1：血清总胆固醇的含量 (mmol/L)；
X_2：甘油三酯的含量 (mmol/L)；
X_3：胰岛素的含量（ U/L）；
X_4：糖化血红蛋白（%）。

2.2 模型分析

2.2.1 多元线性回归方程

用 R 语言，可以求得回归方程为：$\hat{Y} = 5.9433 + 0.1424X_1 + 0.3515X_2 - 0.2706X_3 + 0.6382X_4$。从各项系数都为正数可知，糖尿病患者的血清总胆固醇、甘油三酯、空腹胰岛素、糖化血红蛋白都对空腹血糖有着正相关关系，即在其他情况不变的条件下，它们中任何一个升高，对应的平均血糖也会升高。

回归模型的诊断如下图所示：



如上图所示，回归模型得到的残差较为均匀地分布在零点两侧，没有显示出特定的函数形式。而且通过 Q-Q 图可以看出，残差较为符合正态分布。从图上也可以基本看出，可以否认异方差的存在性。

2.2.2 方差分析

方差分析是用于检验回归方程是否具有统计学意义。其零假设为 H_0：$\beta_1 = \beta_2 = \cdots = \beta_m = 0$，备择假设为 H_1：各 β_j(j=1,2,...,m) 不全为 0。当 H_0 成立时，统计量 $F = \frac{SS_{回}/m}{SS_{残}/(n-m-1)} \sim F(m, n-m-1)$，其中 $SS_{回} = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$，$SS_{残} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$，其置信度水平为 α 的拒绝域为 $F > F_\alpha(m, n-m-1)$。根据本文的多元回归模型做出方差分析表有：

变异来源	自由度	SS	MS	F	P
总变异	26	$SS_{总}$=222.55			
回归	4	$SS_{回}$=133.71	$MS_{回}$=33.43	$\frac{MS_{回}}{MS_{残}}$=8.278	0.0003121
残差	22	$SS_{残}$=88.84	$MS_{残}$=4.04		

由 R 语言计算易得 $F_{0.01}(4,22)=4.31$，求得多多元回归模型的 F=8.278>4.31，在 α=0.01 的置信度水平上拒绝 H_0, 接受 H_1，即认为所建的回归方程具有统计学意义，即可认为糖尿病患者的血糖含量与血清总胆固醇（X_1）、甘油三酯（X_2）、空腹胰岛素（X_3）、糖化血红蛋白（X_4）之间具有线性关系。

2.2.3 可决系数 R^2

可决系数 $R^2=\frac{SS_{回}}{SS_{总}}=1-\frac{SS_{残}}{SS_{总}}$，$0\le R^2\le 1$。$R^2$ 的大小说明自变量 $X_1,X_2,\cdots,X_m$ 能够解释 Y 的变化的百分比，其值越接近于 1，说明模型对数据的拟合程度越好。在本文多元线性回归模型的 $R^2=0.6008$，表明糖尿病患者血糖含量变异的 60% 可以由总胆固醇、甘油三酯、胰岛素和糖化血红蛋白的变化来解释。

2.2.4 t 检验法

t 检验法是检验多元线性回归方程的各项系数 $\hat{\beta}_j$ 是否具有统计学意义，即是否显著异于零。其原假设为 H_{j0}：$\beta_j=0$，备择假设为 H_{j1}：$\beta_j\neq 0$，$j=0,1,2,\cdots,m$。当 H_{j0} 成立时，统计量 $T_j=\frac{\hat{\beta}_j}{\hat{\sigma}\sqrt{c_{jj}}}\sim t(n-m-1)$，$j=0,1,2,\cdots,m$。其中 c_{ii} 是 $C=(X^TX)^{-1}$ 对角线上的第 i 个元素。其置信度水平为 α 的拒绝域为 $T_j>t_{\frac{\alpha}{2}}(n-m-1)$。 通过 R 语言，易求得下表：

	$\hat{\beta}_j$	$sd(\hat{\beta}_j)$	t value	Pr(>|t|)
β_0	5.9433	2.8286	2.101	0.0473
β_1	0.1424	0.3657	0.39	0.7006
β_2	0.3515	0.2042	1.721	0.0993
β_3	-0.2706	0.1214	-2.229	0.0363
β_4	0.6382	0.2433	2.623	0.0155

由 R 语言计算易得 $t_{0.05}(22)=2.074$，$t_4>|t_3|>2.074$，P 值均小于 0.05，所以拒绝零假设，接受备择假设，认为 $\hat{\beta}_3$ 和 $\hat{\beta}_4$ 有统计学意义。在 α=0.1 的置信度水平上可以认为 $\hat{\beta}_3$ 和 $\hat{\beta}_4$ 显著异于零，即可认为糖尿病患者的血糖含量的确受到空腹胰岛素（X_3）、糖化血红蛋白（X_4）的影响。

2.2.5 回归系数的置信区间

β_j 的估计值为 $\hat{\beta}_j$，估计值的标准差为:$se(\hat{\beta}_j)=\hat{\sigma}\sqrt{c_{jj}}, j=1,2,\cdots,m$，其中 c_{jj} 是 $C=(X^TX)^{-1}$ 对角线上的第 i 个元素。因此估计值 $\hat{\beta}_j$ 的置信度为 1-α 的置信区间为:$(\hat{\beta}_j-t_{\frac{\alpha}{2}}sd,\hat{\beta}_j+t_{\frac{\alpha}{2}}sd)$，取 α=0.05, 可以求得如下表所示，

	$\hat{\beta}_j$	Left	Right
β_0	5.9433	0.07713143	11.80940427
β_1	0.1424	-0.61587141	0.90076437
β_2	0.3515	-0.07202817	0.77495914
β_3	-0.2706	-0.52234058	-0.01882996
β_4	0.6382	0.13370175	1.14270073

2.2.6 偏回归平方和（sum of squares for partial regression）

偏回归平方和是一种与 t 检验法完全等价的方法。回归方程中某一变量 X_j 的偏回归平方和表示模型中含有其他 m-1 个自变量的条件下该自标量对 Y 的回归贡献，相当于从回归方程中剔除 X_j 后所引起的回归平方和的减少量，或在 m-1 个自变量的基础上新增加 X_j 引起的回归平方和的增加量。$SS_{回}(X_j) = SS_{回}(X_1, X_2, X_3, X_4) - SS_{回}(X_{i=1,2,3,4,i\neq j})$ ，$SS_{回}(X_j)$ 表示偏回归平方和，其值愈大说明相应的自变量愈重要。。

对某一自变量 X_j 进行偏回归平方和检验，其原假设为 H_0：$\beta_j = 0$，备择假设为 H_1：$\beta_j \neq 0$。当 H_0 成立时，$F_j = \frac{SS_{回}(X_j)/1}{SS_{残}/(n-m-1)} \sim F(1, n-m-1)$。其置信度水平为 α 的拒绝域为 $F_j > F_\alpha(1, n-m-1)$。

对回归方程和回归方程中剔除 X_j 后的 $SS_{回}$ 和 $SS_{残}$ 计算如下表所示：

	回归方程中包含的自变量	$SS_{回}$	$SS_{残}$	$SS_{回}(X_j)$	F
	X_1, X_2, X_3, X_4	133.7107	88.8412		
X_1	X_2, X_3, X_4	133.0978	89.4540	0.6129	0.152
X_2	X_1, X_3, X_4	121.7480	100.8038	11.9627	2.962
X_3	X_1, X_2, X_4	113.6472	108.9047	20.0635	4.968
X_4	X_1, X_2, X_3	105.9168	116.6351	27.7939	6.883

计算易得 $F_{0.05}(1, 22) = 4.30$，由于 F_3,F_4 大于 4.30，所以拒绝零假设，接受备择假设，认为血糖（ Y ）与胰岛素（ X_3 ）、糖化血红蛋白（ X_4 ）有线性回归关系。并且比较两个变量的偏回归平方和的大小有 $SS_{回}(X_4) > SS_{回}(X_3)$，即认为糖化血红蛋白（ X_4 ）的回归贡献更大。

2.2.7 多重共线性的诊断

设 $x_1, x_2, \cdots, x_m$ 是自变量 $X_1, X_2, \cdots, X_m$ 经过中心化和标准化得到的向量，记 $x = (x_1, x_2, \cdots, x_m)$，设 λ 为 X^TX 的一个特征值，φ 为对应的特征向量，其长度为 1，即 $\varphi^T\varphi = 1$ 若 $\lambda \approx 0$，则 $x^Tx\varphi = \lambda\varphi \approx 0$，用 φ^T 左乘可到 $\varphi x^T x\varphi = \varphi^T\lambda\varphi = \lambda \approx 0$, 所以有 $x\varphi \approx 0$，即 $\varphi_1 x_1 + \varphi_2 x_2 + \cdots + \varphi_m x_m \approx 0$，所以对应自变量 $X_1, X_2, \cdots, X_m$，存在 $c_1, c_2, \cdots, c_m$ 使得左式近似成立，即自变量之间存在着多重共线性。度量多重共线性严重程度的一个重要指标是矩阵 X^TX 的条件数，即 $\kappa(X^TX) = ||X^TX|| \cdot ||(X^TX)^{-1}|| = \frac{\lambda_{max}(X^TX)}{\lambda_{min}(X^TX)}$

其中$\lambda_{max}(X^TX)$，$\lambda_{min}(X^TX)$ 分别表示矩阵 X^TX 的最大最小特征值。从实际应用的经验角度，一般若 κ<100，则认为多重共线性的程度很小；若 $100 \leq \kappa \leq 1000$，则认为存在中等程度或较强的多重共线性；若 κ>1000，则认为存在严重的多重共线性。

利用 R 语言得到条件数 κ=11.42798, 认为多重共线性的程度很小。

另一种常见的度量多重共线性严重程度的一个重要指标是方程膨胀因子（ vif ），通过 R 语言计算得：

系数	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$
vif	2.1855	1.7799	1.2784	1.2667

一般 vif 大于 10 说明存在这变量可能与其他变量存在复共线性，而从上面的数值可以看出，可以认为多重共线性的程度很小。

2.2.8 标准化回归系数

变量标准化是将原始数据减去相应变量的均数，然后再除以该变量的标准差 $x_j = \frac{X_j - \bar{X}_j}{S_j}$。计算得到的回归方程称作标准化回归方程，对应的回归系数称为标准化回归系数。

因为回归系数有单位，用来解释各自变量对因变量的影响时，表示在其他自变量保持不变时，X_j 增加或减少 1 个单位时 Y 的平均变化量。不能用各自的 $|\hat{\beta}_j|$ 来比较各 X_j 对 $\hat{Y}$ 的影响大小。标准化回归系数没有单位，可以用来比较各个自变量 X_j 对 Y 的影响强度，通常在有统计学意义的前提下，标准化回归系数的绝对值愈大说明相应自变量对 Y 的作用越大。

对数据中的 Y, X_1, X_2, X_3, X_4 进行标准化得到 y, x_1, x_2, x_3, x_4，再进行回归得到标准化回归方程为：$y = 5.582 \times 10^{-17} + 0.00758x_1 + 0.3093x_3 - 0.3395x_4 + 0.3977x_5$

由标准化回归方程可以看出，对血糖影响大小的顺序依次为糖化血红蛋白（X_4）、胰岛素（X_3）、甘油三脂（X_2）和总胆固醇（X_1）。

2.2.9 自变量的选择

一般来说，如果在一个回归方程中忽略了对 Y 有显著影响的自变量，那所建立的回归方程与实际结果有较大的偏差，而如果自变量选得过多，则使用不方便。因此是当地选择变量以建立一个“最优”的回归方程十分重要。

选择自变量的一种重要方法是修正的可决系数 R_c^2 选择法。

修正的可决系数 $R_c^2 = 1 - (1 - R^2)\frac{n-1}{n-p-1} = 1 - \frac{MS_{残}}{MS_{总}}$，因为 R_c^2 具有这样的变化规律：当 R^2 相同时，自变量个数越多，R_c^2 越小。因此，可以选择不同的自变量，选其中 R_c^2 最大者作为“最优”回归方程。根据不同的自变量做出的回归的结果汇总如下：

方程中的自变量	R_c^2	方程中的自变量	R_c^2
X_2, X_3, X_4	0.546	X_2, X_3	0.408
X_1, X_2, X_3, X_4	0.528	X_1, X_3	0.375
X_1, X_3, X_4	0.488	X_4	0.347
X_1, X_2, X_4	0.447	X_1	0.284
X_1, X_4	0.441	X_1, X_2	0.275
X_2, X_4	0.44	X_3	0.231
X_3, X_4	0.435	X_2	0.179
X_1, X_2, X_3	0.408		

“最优”回归方程为 $\hat{Y} = 6.4996 + 0.4023X_2 - 0.2871X_3 + 0.6632X_4$。结果表明：血糖的变化与甘油三酯、胰岛素和糖化血红蛋白有线性回归关系，其中与胰岛素负相关。由标准化回归系数看出，糖化血红蛋白对血糖的影响最大。

另一种选择自变量的重要方法是后退法（backward elimination）。后退法先将全部自变量选入方程，然后逐步剔除误统计学意义的自变量。

剔除自变量的方法是在方程中选一个偏回归平方和最小的变量，作 F 检验决定它是否剔除，若无统计学意义则将其剔除，然后对剩余的自变量建立新的回归方程。重复这一过程，直至方程中所有的自变量都不能剔除为止。

当选择全部变量时，回归方程的 AIC 值为 42.16,。接下来，如果去掉 1 个变量，当去掉 X_1 时，回归方程的 AIC 值达到最小，为 40.34。如果再去掉 1 个变量，无论去掉哪个变量，回归方程的 AIC 值都会增大，因此“最优”回归方程选择的自变量为 X_2，X_3，X_4。“最优”回归方程为 $\hat{Y} = 6.4996 + 0.4023X_2 - 0.2871X_3 + 0.6632X_4$，这与修正的可决系数 R_c^2 选择法得到的结果相同。

2.2.10 交互作用的判断

因为两个自变量之间也可能存在交互作用，因此，在多元线性回归中添加自变量的乘积项，是一种常见的交互作用的判断方法。在上一步自变量选择中，已经选出甘油三酯（X_2），胰岛素（X_3），糖化血红蛋白（X_4）三个变量，要

检验胰岛素和糖化血红蛋白之间是否存在交互作用，可以在回归方程中添加显得变量 $X_5 = X_3X_4$，按照新的线性回归模型 $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_2X_2 + \hat{\beta}_3X_3 + \hat{\beta}_4X_4 + \hat{\beta}_5X_5$。对$X_5$ 的回归系数做显著性检验。利用 R 语言做回归得：

	$\hat{\beta}_j$	$sd(\hat{\beta}_j)$	t value	Pr(>|t|)
β_0	-0.7898	3.17202	-0.249	0.805679
β_2	0.3649	0.133	2.744	0.011855
β_3	1.22674	0.5101	2.405	0.02503
β_4	1.50974	0.34301	4.401	0.000226
β_5	-0.17855	0.05909	-3.022	0.006272

从上表可以看出 $\hat{\beta}_5$ 在统计学意义上显著（P<0.01），因此可以认为糖尿病患者的血糖含量受到胰岛素（X_3），糖化血红蛋白（X_4）之间的交互作用，即胰岛素对血糖的影响依赖于糖化血红蛋白的含量或糖化血红蛋白对血糖的影响依赖于胰岛素的含量。

3 其他多元回归模型

3.1 岭回归（ridge regression）

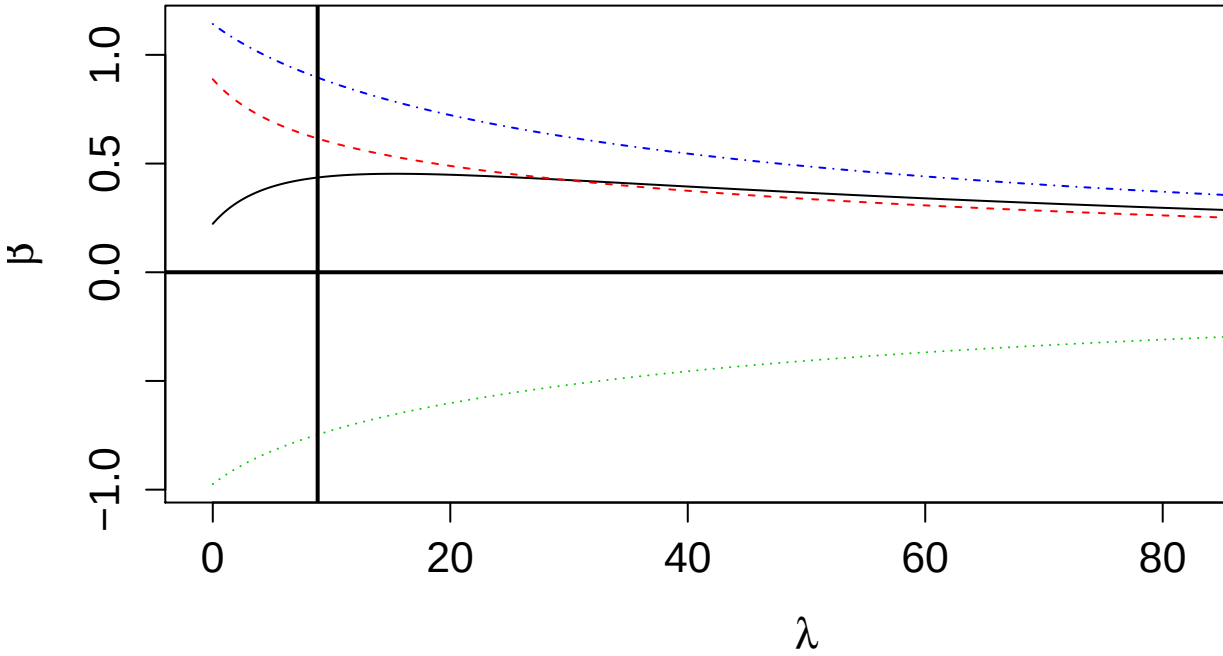
岭回归（ridge regression）可以用来处理下面两类问题：一是数据点少于变量个数；二是变量间存在共线性。

在最小二乘法中，当 X 列满秩时，有 $\hat{\beta} = \left(X^TX\right)^{-1}X^TY$. 当 X 不是列满秩，或者某些某些列之间的线性相关性比较大时，X^TX 的行列式接近于 0，即 X^TX 接近于奇异，计算 $\left(X^TX\right)^{-1}$ 时误差会很大。此时传统的最小二乘法缺乏稳定性与可靠性。

岭回归是对最小二乘回归的一种补充，它损失了无偏性，来换取高的数值稳定性，从而得到较高的计算精度。当 X^TX 的行列式接近于 0 时，我们将其主对角元素都加上一个数 k，可以使矩阵为奇异的风险大降低。于是：$\hat{\beta} = \left(X^TX + kI\right)^{-1}X^TY$.(I 是单位矩阵)

随着 k 的增大，B(k) 中各元素 bi(k) 的绝对值均趋于不断变小，它们相对于正确值 bi 的偏差也越来越大。k 趋于无穷大时，B(k) 趋于 0。b(k) 随 k 的改变而变化的轨迹，就称为岭迹。实际计算中可选非常多的 k 值，做出一个岭迹图，看看这个图在取哪个值的时候变稳定了，那就确定 k 值了。

在 R 语言中，MASS 包中的函数 lm.ridge() 可以很方便的完成岭回归，通过计算得到 k 值为 8.83。岭迹图如下所示：



有 R 语言分析得，bi(k) 的绝对值在 k=8.83 的时候变稳定了，对应的系数如下表所示：

	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$
OLS	0.223	0.888	-0.975	1.142
岭回归	0.436	0.615	-0.747	0.896

3.2 主成分分析（principal components analysis，PCA）

X 不满足列满秩，换句话就是说样本向量之间具有高度的相关性（如果每一列是一个向量的话）。遇到列向量相关的情形，岭回归是一种处理方法，也可以用主成分分析 PCA 来进行降维。

主成分分析（principal components analysis，PCA）是一种分析、简化数据集的技术。它把原始数据变换到一个新的坐标系统中，使得任何数据投影的第一大方差在第一个坐标（称为第一主成分）上，第二大方差在第二个坐标（第二主成分）上，依次类推。主成分分析经常用减少数据集的维数，同时保持数据集的对方差贡献最大的特征。

设 X 是 p 维随机变量，病假设 $\mu = E(x)$，$\Sigma = Var(X)$。并考虑如下线性变换

$$\begin{cases} Z_1 = a_1^T X \\ Z_2 = a_2^T X \\ \vdots \\ Z_p = a_p^T X \end{cases}$$

$Var(Z_i) = a_i^T \Sigma a_i$，$i = 1, 2, \cdots, p$，$Cov(Z_i, Z_j) = a_i^T \Sigma a_j$，$i, j = 1, 2, \cdots, p, i \neq j$. 为使 Z_1 方差达到最大，即 a_1 是约束优化问题

$max\ a^T \Sigma a$

$s.t.\ a^{Ta} = 1$

的解。

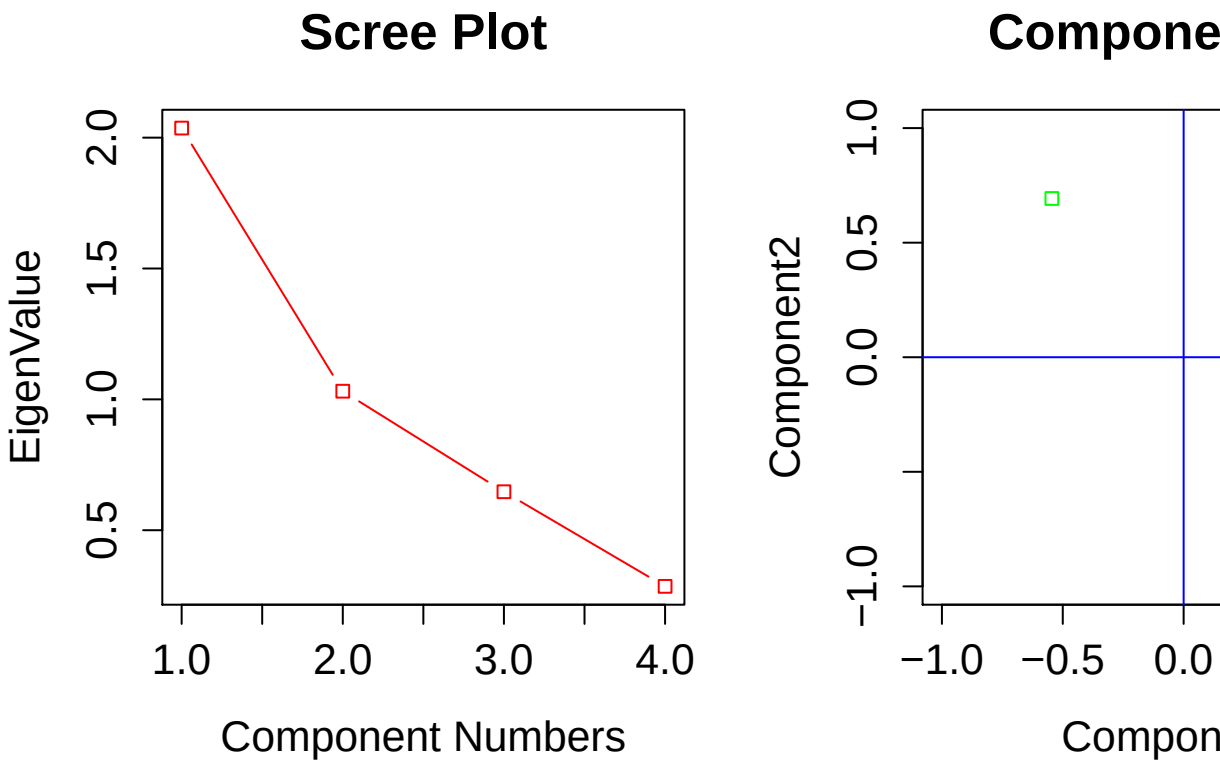
一般情况，对于协方差阵 Σ，可由正交阵 Q 将其化为对角阵，即

$$Q^T \Sigma Q = \begin{bmatrix} \lambda_1 & & & \\ & \lambda_2 & & \\ & & \ddots & \\ & & & \lambda_p \end{bmatrix}$$，且 $\lambda_1 > \lambda_2 > \cdots > \lambda_p$。

则 Q 的第 i 列就对应于 a_i，相应的 Z_i 为第 i 主成分。

R 语言中进行主成分分析可以采用基本的 princomp 函数，将结果输入到 summary 和 plot 函数中可分别得到分析结果和碎石图。

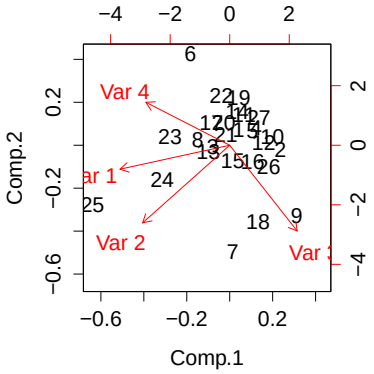
碎石图如下图所示。



各成分的载荷矩阵为：

	Comp.1	Comp.2	Comp.3	Comp.4
X_1	-0.621	-0.191	-0.18	0.738
X_2	-0.493	-0.617	-0.115	-0.602
X_3	0.382	-0.682	0.556	0.281
X_4	-0.475	0.342	0.803	-0.115

每个变量在因子向量中的载荷图如下所示（未作因子旋转）：



分析结果如下表：

	Comp.1	Comp.2	Comp.3	Comp.4
Standard deviation	1.4270707	1.0154272	0.8043559	0.53421759
Proportion of Variance	0.5091327	0.2577731	0.1617471	0.07134711
Cumulative Proportion	0.5091327	0.7669058	0.9286529	1

从上图和表中可以看出，前 2 个主成分的累计贡献率达到 76.6%，前 2 个主成分的累计贡献率达到 92.8%。所以可以舍去最后 1 个或最后 2 个主成分，达到降维的目的。

所以，可以舍去最后 1 个成分，用前 3 个成分做回归，得：$\hat{Y} = 11.9259 - 1.4903Z_1 + 0.4645Z_2 + 0.2327Z_3$，其中 Z_1, Z_2, Z_3 是前 3 个主成分。把 3 个主成分的表达式 $Z_1 = -0.621X_1 - 0.493X_2 + 0.382X_3 - X_4$，$Z_2 = -0.1911X_1 - 0.617X_2 - 0.682X_3 + 0.342X_4$，$Z_3 = -0.18X_1 - 0.115X_2 + 0.556X_3 - 0.803X_4$ 代入到回归表达式易得：

$\hat{Y} = 11.9259 + 0.7953X_2 + 0.4211X_2 - 0.7567X_3 + 1.0528X_4$ 将其与原最小二乘法得到的回归方程 $\hat{Y} = 5.9433 + 0.1424X_1 + 0.3515X_2 - 0.2706X_3 + 0.6382X_4$ 比较发现，只选取 3 个主成分做的回归虽然系数与最小二乘法得到回归方程的系数的不同，但系数的符号没有变化，系数的大小差距比较小。由此我们可以的得出，在适当的条件下，可以由主成分回归方程代替最小二乘回归方程。

4 总结

事实上，一种现象常常是与多个因素相联系的，由多个自变量的最优组合共同来预测或估计因变量，比只用一个自变量进行预测或估计更有效，更符合实际。因此多元线性回归应用地比一元线性回归更广。而且，多元回归分析还可以分析哪些因素有影响，哪些因素影响较大。所以，掌握多元线性回归的应用与分析，具有很重要的实际意义。