第四讲：多元线性回归

软件工程《机器学习》课程

李济洪，王瑞波

山西大学 • 软件学院

wangruibo@sxu.edu.cn

2019 年 9 月 24 日

提纲

上讲回顾

- 一元线性回归算法: $\mathrm{y} = \beta_0 + \beta_1 \mathrm{x} + \epsilon$;
- 参数估计：

$$\hat{\beta}_1 = \frac{\sum_{\mathrm{i}=1}^{\mathrm{n}} (\mathrm{x_i} - \bar{\mathrm{x}})(\mathrm{y_i} - \bar{\mathrm{y}})}{\sum_{\mathrm{i}=1}^{\mathrm{n}} (\mathrm{x_i} - \bar{\mathrm{x}})^2} \tag{1}$$

$$\hat{\beta}_0 = \bar{\mathrm{y}} - \hat{\beta}_1 \bar{\mathrm{x}} \tag{2}$$

- 参数估计的准确性：偏差、方差、置信区间；
- 模型的准确性：残差标准误 RSE、拟合优度 R^2；

课程代码讲解

https://github.com/RamboWANG/statLearn.git

Tag

- git tag v0.1 -m ” 项目初始框架”
- git tag
- git push origin v0.1

以广告数据集为例

Table: Advertising 数据集示例

Id	TV	Radio	Newspaper	Sales
1	230.1	37.8	69.2	22.1
2	44.5	39.3	45.1	10.4
3	17.2	45.9	69.3	9.3
		.....		
200	232.1	8.6	8.7	13.4

思考

1. 实际情况：预算通常不会只分配到单个媒体上。
2. 若 TV、Radio、Newspaper 上均有预算，该如何预测对应的销量（Sales）？

思路一：多个一元线性回归模型

$$\widehat{\text{Sales}} = 9.312 + 0.203 \times \text{Radio} \tag{3}$$

$$\widehat{\text{Sales}} = 12.351 + 0.055 \times \text{Newspaper} \tag{4}$$

$$\widehat{\text{Sales}} = 7.0325 + 0.0475 \times \text{TV} \tag{5}$$

问题

如何将 Sales 的三个不同的预测值聚合成一个预测值？（本讲不做深入考虑）

- 难点：三个模型建立在同样的数据集上，有强相关性。聚合预测值需考虑这种相关性。

思路二：单个多元线性回归模型

$$\text{Sales} = \beta_0 + \beta_1 \times \text{TV} + \beta_2 \times \text{Radio} + \beta_3 \times \text{Newspaper} + \epsilon \tag{6}$$

其中：$\epsilon \sim \text{N}(0, \sigma^2)$.

基本问题

- $\beta_0, \ldots, \beta_\text{p}$ 如何估计？
- 预测变量 TV, Radio 和 Newspaper 如何影响销量 Sales?（推断）
- 给定预算，如何预测未来的销量？精度如何？（预测）

多元线性回归的一般形式

$$\mathrm{Y} = \beta_0 + \beta_1 \mathrm{X}_1 + \ldots + \beta_\mathrm{p} \mathrm{X}_\mathrm{p} + \epsilon = \boldsymbol{\beta}^\top \mathrm{X} + \epsilon \tag{7}$$

其中，$\epsilon \sim \mathrm{N}(0, \sigma^2)$.

$$\hat{\mathrm{Y}} = \hat{\beta}_0 + \hat{\beta}_1 \mathrm{X}_1^{\text{future}} + \ldots + \hat{\beta}_\mathrm{p} \mathrm{X}_\mathrm{p}^{\text{future}} + \epsilon = \hat{\boldsymbol{\beta}}^\top \mathrm{X}^{\text{future}} \tag{8}$$

一般概念

- 预测变量：X，为向量；
- 响应变量：Y；
- 误差项：ϵ；
- 回归系数（模型参数）：$\boldsymbol{\beta} = (\beta_1, \ldots, \beta_\mathrm{p})^\top$；
- 参数估计：$\hat{\boldsymbol{\beta}} = (\hat{\beta}_1, \ldots, \hat{\beta}_\mathrm{p})^\top$；
- 响应变量的预测值：$\hat{\mathrm{Y}}$；

多元线性回归的等价形式

$$\mathrm{Y} \sim \mathrm{N}(\boldsymbol{\beta}^{\top}\mathrm{X}, \sigma^2) \tag{9}$$

或者

$$\mathrm{E}[\mathrm{Y}] = \boldsymbol{\beta}^{\top}\mathrm{X} = \beta_0 + \beta_1\mathrm{X}_1 + \ldots + \beta_{\mathrm{p}}\mathrm{X}_{\mathrm{p}} \tag{10}$$

也就是说，我们希望找到的趋势是关于 Y 的均值的趋势。具体地，我们希望使用$\mathrm{p}+1$ 维空间中的一条直线 来拟合响应变量 Y 的均值。

多元线性回归算法中的 5 个理论问题

1. 给定训练数据 $\mathrm{D_n}$，如何估计多元线性回归算法的参数 $\beta_1, \ldots, \beta_\mathrm{p}$？
2. 预测变量 $\mathrm{X}_1, \ldots, \mathrm{X_p}$ 中是否至少有一个预测变量可以用来预测响应变量？
3. 所有预测变量都有助于解释响应变量 Y 吗？或仅仅是其一个子集对预测有用？
4. 模型对数据的拟合程度如何？
5. 给定一组预测变量的值，预测的精度如何？

问题 1：参数估计

问题描述

给定数据集 $\mathrm{D_n} = \{(\mathrm{x_i}, \mathrm{y_i})\}_{\mathrm{i}=1}^{\mathrm{n}}$，如何优化多元线性回归算法的系数 $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_{\mathrm{p}})^\top$？

- 可类比于高等数学中的给定 n 个线性方程，解方程组中的 p 个参数。
- 不同之处：很可能找不到一组解满足所有方程。

$$\hat{\boldsymbol{\beta}}^* = \arg\min_{\boldsymbol{\beta}} \mathrm{RSS} = \arg\min_{\boldsymbol{\beta}} \sum_{\mathrm{i}=1}^{\mathrm{n}} [\mathrm{y_i} - (\hat{\beta}_0 + \hat{\beta}_1 \mathrm{x_{i1}} + \ldots + \hat{\beta}_{\mathrm{p}} \mathrm{x_{ip}})]^2. \tag{11}$$

多元线性回归在广告数据集上的应用

$$\text{Sales} = 2.939 + 0.046 \times \text{TV} + 0.189 \times \text{Radio} - 0.001 \times \text{Newspaper} + \epsilon \tag{12}$$

一元线性回归

$$\widehat{\text{Sales}} = 9.312 + 0.203 \times \text{Radio} \tag{13}$$

$$\widehat{\text{Sales}} = 12.351 + 0.055 \times \text{Newspaper} \tag{14}$$

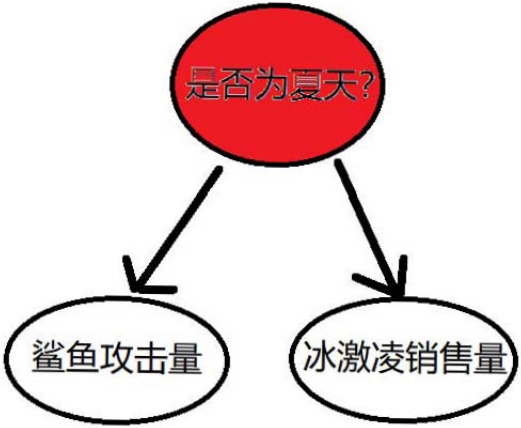
$$\widehat{\text{Sales}} = 7.0325 + 0.0475 \times \text{TV} \tag{15}$$

广告数据集上预测变量之间的相关性

Table: 预测变量间的相关性

	TV	radio	newspaper	sales
TV	1	0.0548	0.0567	0.7822
radio		1	0.3541	0.5762
newspaper			1	0.2283
sales				1

一个直观的解释



解释

- 预测变量：鲨鱼攻击量；
- 响应变量：冰激凌销售量；
- 根据一段时间内收集的海滩社区数据，预测变量和相应变量呈现正相关性。
- **思考：**是否可以通过减少冰激凌销售量，来降低鲨鱼攻击量？

问题 2：是否至少有一个预测变量可以用来预测响应变量？

假设检验

$$\mathrm{H}_0 : \beta_1 = \beta_2 = \ldots = \beta_\mathrm{p} = 0; \quad \text{v.s.} \quad \mathrm{H}_1 : \text{至少有一个}\beta_\mathrm{i}\text{不为 } 0; \tag{16}$$

$$\mathrm{F} = \frac{(\mathrm{TSS} - \mathrm{RSS})/\mathrm{p}}{\mathrm{RSS}/(\mathrm{n} - \mathrm{p} - 1)} \sim \mathrm{F}(\mathrm{p}, \mathrm{n} - \mathrm{p} - 1). \tag{17}$$

- F 分布的介绍：https://baike.baidu.com/item/F%E5%88%86%E5%B8%83/7917090?fr=aladdin

问题 3：所有变量都有助于解释响应变量 Y 吗？

问题设立

对算法中存在的一组变量 $\beta_{p-q+1}, \ldots, \beta_p$，做如下假设检验:

$$H_0 : \beta_{p-q+1} = \ldots = \beta_p = 0; \quad \text{v.s.} \quad H_1 : \text{这些变量中有一个不为 } 0; \tag{18}$$

$$F = \frac{(RSS_0 - RSS)/q}{RSS/(n-p-1)} \sim F(q, n-p-1). \tag{19}$$

- RSS_0 为不加上述系数对应预测变量的多元线性回归模型。
- 当 $q = 1$ 时，对应于上讲中的 t 检验。

如何选取重要变量？

准则及难点

- 准则：RSS, AIC, BIC, R^2；
- 难点：模型个数为 2^p，选择问题为 NP 完全问题；

常用选取方法

- 向前选取方法
- 向后选取方法
- 混合选取方法

1. 贪心方法的缺点；
2. 这些方法在高维情形下的表现；

问题 4：模型对数据的拟合程度？

两个准则

1. 准则一：残差标准误（RSE）
   - 取值范围 $[0, +\infty)$，越小越好；
   - $\mathrm{RSE} = \sqrt{\frac{1}{\mathrm{n-p-1}}\mathrm{RSS}}$；
2. 准则二：拟合优度 R^2
   - 取值范围 $[0, 1]$，越大越好；
   - $\mathrm{R}^2 = \mathrm{Corr}(\mathrm{Y}, \hat{\mathrm{Y}})$；

问题 5：模型的预测精度

两个区间

假定 $\mathrm{TV} = 10, \mathrm{Radio} = 2$，则可计算出：

1. $\hat{\mathrm{Y}}$ 的 95% 置信区间：[10985, 11528]；
2. Y 的 95% 预测区间: [7930, 14580]；

解释

- 预测区间比置信区间宽；
- Y 中含有随机误差项 ϵ；
- $\hat{\mathrm{Y}}$ 仅体现了 Y 的期望 E[Y] 的波动情况。

回看广告数据集

参数拟合的准确性

	系数	标准误	t 攻击量	p 值
截距项	2.939	0.3119	9.42	<0.0001
TV	0.046	0.0014	32.81	< 0.0001
Radio	0.189	0.0086	21.89	<0.0001
Newspaper	-0.001	0.0059	-0.18	0.8599

模型拟合的准确性

RSE: 1.69; R^2: 0.897; F 统计量：570;

何为定性预测变量

定性预测变量：取值为离散的类别值的预测变量；

信用卡债务数据

- 预测变量：信用卡的个人债务额度
- 定量预测变量：年龄、信用卡数量、受教育年限、收入、信用额度、信用评级；
- 定性预测变量：性别、是否为学生、婚姻状况、种族；

定型变量的水平

- 二水平：性别的值为“男”、“女”、是否为学生；
- 多水平：婚姻状况、种族；

以信用数据中的“性别”为例

哑变量

$$x_i = \begin{cases} 1 & \text{女性} \\ 0 & \text{男性} \end{cases} \qquad y_i = \beta_0 + \beta_1 x_i + \epsilon_i = \begin{cases} \beta_0 + \beta_1 + \epsilon_i & \text{女性} \\ \beta_0 + \epsilon_i & \text{男性} \end{cases} \tag{20}$$

性别对信用债务的影响

	系数	标准误	t 统计量	p 值
截距项	509.80	33.13	15.389	< 0.0001
性别	19.73	46.05	0.429	0.669

哑变量不同编码方式之间的等价性

哑变量的不同编码

$$x_i = \begin{cases} 1 & \text{女性} \\ 0 & \text{男性} \end{cases} \qquad x_i = \begin{cases} 1 & \text{女性} \\ -1 & \text{男性} \end{cases}$$

哑变量

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i = \begin{cases} \beta_0 + \beta_1 + \epsilon_i & \text{女性} \\ \beta_0 + \epsilon_i & \text{男性} \end{cases} \tag{21}$$

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i = \begin{cases} \beta_0 + \beta_1 + \epsilon_i & \text{女性} \\ \beta_0 - \beta_1 + \epsilon_i & \text{男性} \end{cases} \tag{22}$$

多水平预测变量的编码方式

做法：d 个水平需要扩展成 d − 1 个哑变量。

以信用数据中的“种族”为例，其取值为：亚洲人、白种人、其它；

“种族”变量的编码

$$x_{i1} = \begin{cases} 1 & \text{亚洲人} \\ 0 & \text{非亚洲人} \end{cases} \qquad x_{i2} = \begin{cases} 1 & \text{白种人} \\ 0 & \text{非白种人} \end{cases}$$

思考：为何不设置 3 个哑变量？

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon_i = \begin{cases} \beta_0 + \beta_1 + \epsilon_i & \text{亚洲人} \\ \beta_0 + \beta_2 + \epsilon_i & \text{白种人} \\ \beta_0 + \epsilon_i & \text{其它} \end{cases}$$

种族对于信用卡债务的影响

	系数	标准误	t 统计量	p 值
截距项 β_0	531.00	46.32	11.464	< 0.0001
亚洲人 β_1	-18.69	65.02	-0.287	0.7740
白种人 β_2	-12.50	56.68	-0.221	0.826

何为可加性假设？

可加性假设

预测变量 X_j 的变化对相应变量 Y 的影响与其它预测变量的取值无关。

例子

- 预测变量：工厂的生产线数 (lines)、工人总数 (workers)
- 响应变量：工厂生产的商品数 (units)

$\text{units} = \beta_0 + \beta_1 \times \text{workers} + \beta_2 \times \text{lines}$ 是否合理？

$$\text{units} = \beta_0 + \beta_1 \times \text{workers} + \beta_2 \times \text{lines} + \beta_3 \times (\text{lines} \times \text{workers}) \quad (23)$$

交互项建模

$$\mathrm{Y} = \beta_0 + \beta_1 \mathrm{X}_1 + \beta_2 \mathrm{X}_2 + \beta_3 \mathrm{X}_1 \mathrm{X}_2 + \epsilon \tag{24}$$

或者

$$\mathrm{Y} = \beta_0 + (\beta_1 + \beta_3 \mathrm{X}_2) \mathrm{X}_1 + \beta_2 \mathrm{X}_2 + \epsilon \tag{25}$$

也就是说，X_1 对 Y 的影响与 X_2 有关。

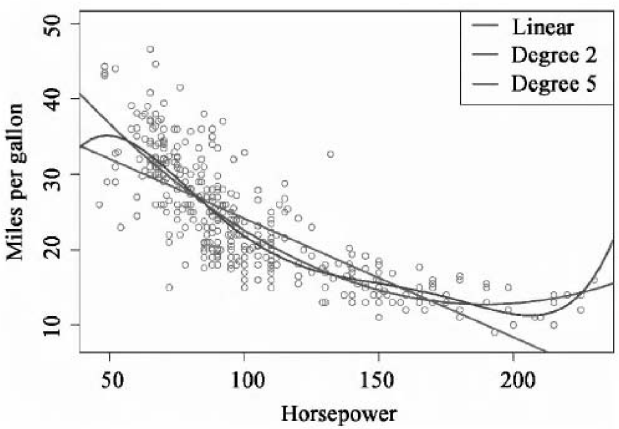
回看广告数据集例子

$$\text{Sales} = \beta_0 + \beta_1 \text{TV} + \beta_2 \text{Radio} + \beta_3 (\text{TV} \times \text{Radio}) + \epsilon \tag{26}$$

TV 与 Radio 间的交互对产品销量的影响

	系数	标准误	t 统计量	p 值
截距项	6.7502	0.248	27.23	<0.0001
TV	0.0191	0.002	12.70	<0.0001
Radio	0.0289	0.009	3.24	0.0014
TV × Radio	0.0011	0.111	20.73	<0.0001

多项式回归



$$\text{mpg} = \beta_0 + \beta_1 \times \text{horsepower} + \beta_2 \times \text{horsepower}^2.$$

多元线性回归的一些基本假设

基本假设

1. 线性趋势：响应变量的期望与预测变量间的关系是线性的；
2. 方差恒定假设：误差项的方差与预测变量无关；
3. 可加性假设：单个预测变量对响应变量的影响与其它预测变量无关；
4. 数据的独立同分布假设：训练集中的数据是独立同分布的。

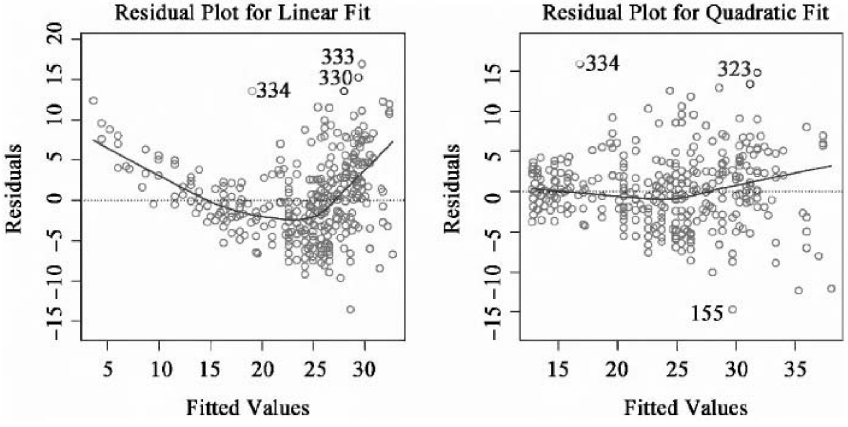
上述任何一条假设被破坏，都会使得多元线性回归算法失效。

潜在的问题

常见的 6 点潜在问题

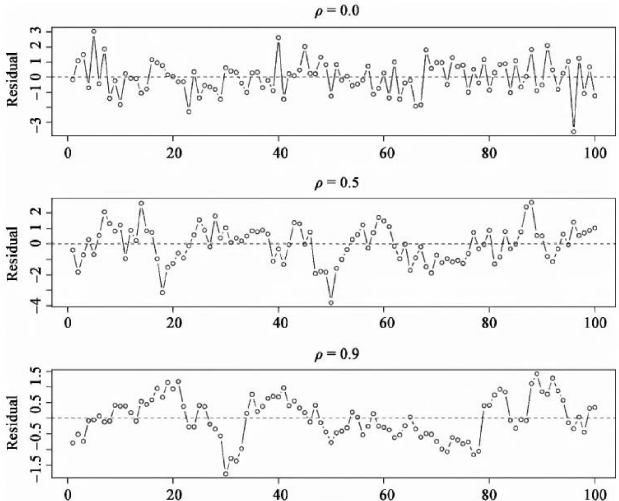
1. 数据非线性
2. 误差项自相关
3. 误差项方差非恒定
4. 离群点
5. 高杠杆点
6. 共线性

问题①：数据非线性

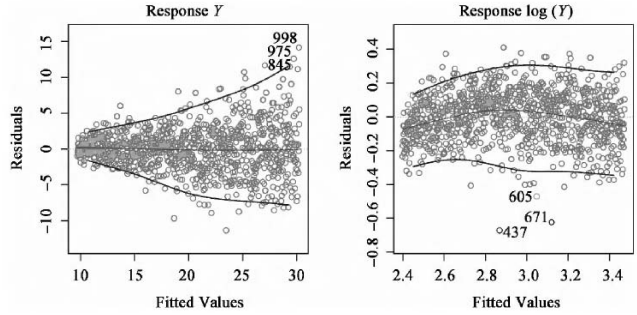


处理方法：对预测变量进行非线性变换，如 logX, $\sqrt{X}$ 或 X^2 等。

问题②: 误差项自相关

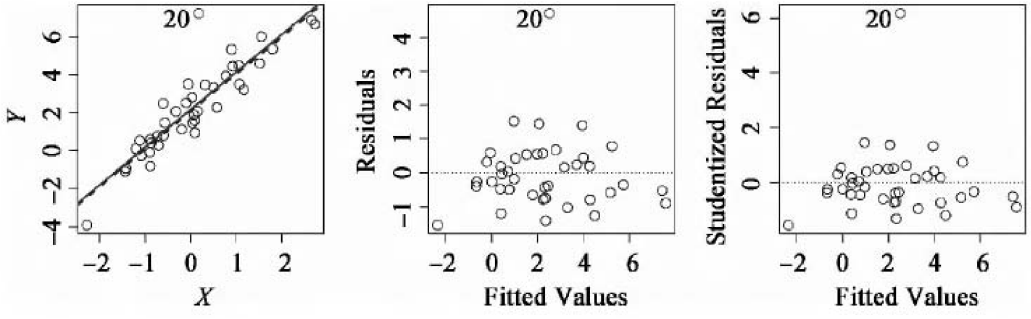

ρ = 0.0
Residual
−3 −1 0 1 2 3
0 20 40 60 80 100
ρ = 0.5
Residual
−4 −2 0 1 2
0 20 40 60 80 100
ρ = 0.9
Residual
−1.5 −0.5 0.5 1.5
0 20 40 60 80 100

问题③: 误差项方差非恒定



解决方法：对响应变量进行 log 变化（logY）或者开方变换（$\sqrt{Y}$）。

问题④: 离群点

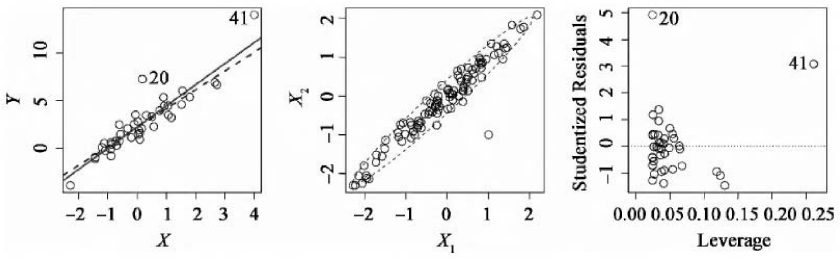


识别方法：学生化残差（残差/标准误）>3

处理方法：删除；

可能的原因：数据采集中的错误，或暗示模型存在缺陷。

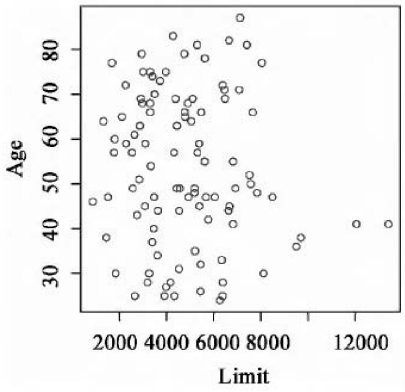
问题⑤: 高杠杆点

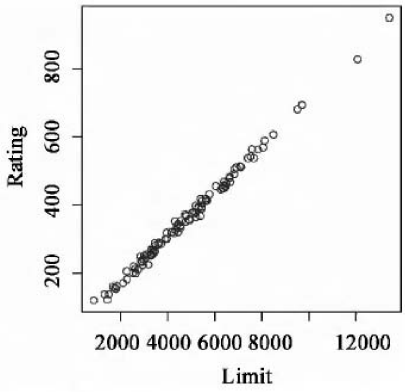


识别方法：杠杆统计量 $h_i = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{i'}(x_{i'} - \bar{x})^2}$ v.s. 学生化残差。

处理方法：删除

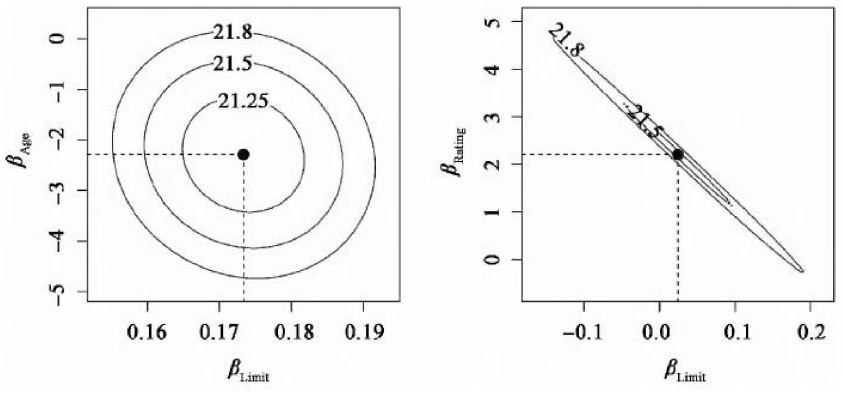
问题⑥: 共线性





处理方法：PCA 等。

问题⑥: 共线性（续）



复杂情形：多重共线性，即多个变量之间存在共线性。（不做讲解）

多元线性回归的 R 实现

充分掌握如下函数：

- lm
- summary
- predict

总结

- 多元线性回归算法的基本形式；
- 多元线性回归算法的一些理论问题；
- 定性预测变量的处理方法；
- 交互作用及非线性情形的处理方法；
- 多元线性回归算法的基本假设及潜在问题；

本讲作业

1. **推导**：多元线性回归算法参数的最小二乘估计的形式。
2. **思考**：在处理多水平的定性预测变量时，为何不将其形式化为取多值得单个变量，而要形式化为多个取二值的哑变量？
3. **应用**：分析清楚 Letter 数据集的预测变量和响应变量，及任务定义。进而，使用 R 语言的多元线性回归算法对 Letter 数据集进行建模。
   - Letter 数据集地址：http://archive.ics.uci.edu/ml/datasets/letter+recognition

谢谢！

Questions & Answering！