Shenduo Zhang

Problem 1

1. (P92 2.2 题)对于**过原点**的简单线性回归模型

$$y_i=\beta x_i+\varepsilon_i,\ \ i=1,2,\cdots,n,$$

设 $\varepsilon_i(i=1,2,\cdots,n)$相互独立且服从 $N(0,\sigma^2)$分布.

(1) 求 β 的最小二乘估计，它是否是 β 的无偏估计？

(2) 求出误差方差 σ^2 的一个无偏估计.

(3) 写出回归关系显著性检验的统计量及其零分布，相应的方差分析表，它和具有常数项的简单线性回归模型的相应结果有何区别？

(4) 给出检验假设 $H_0:\beta=0$ 的 t 统计量及其零分布，它和(3)中的假设检验有何关系？

(5) 对于自变量的新观测值 x_0，给出相应的因变量 y_0 的预测值及其置信度为 $1-\alpha$ 的置信区间.

***Solution* 1.a** 首先将模型写为矩阵形式 $Y=X\beta+\epsilon$. 其中

$$Y=\begin{pmatrix}y_1\\y_2\\\vdots\\y_n\end{pmatrix},\quad X=\begin{pmatrix}x_1\\x_2\\\vdots\\x_n\end{pmatrix},\quad \epsilon\sim N(\mathbf{0},\mathbf{I}_n) \tag{1}$$

设 L 为 design matrix 列向量所张成的线性空间, 记从线性空间 R^n 到 L 的投影算子为 P_L. 由投影算子的性质, 去求 β 的最小二乘 $\hat\beta$ 估计等价于求解如下方程,

$$X\hat\beta=P_LY \tag{2}$$

既

$$\hat\beta=(X^TX)^{-1}X^TY=\frac{\sum_{i=1}^n x_iy_i}{\sum_{i=1}^n x_i^2}. \tag{3}$$

其是 β 的无偏估计, 因为

$$\mathbb{E}\hat{\beta} = \frac{\sum_{i=1}^{n} x_i \mathbb{E}(\beta x_i + \epsilon_i)}{\sum_{i=1}^{n} x_i^2} = \beta. \tag{4}$$

□

***Solution* 1.b** 设误差方差 σ^2 的一个无偏估计是 $\hat{\sigma}^2$, $P_{L^\perp}$ 是向 L 正交补空间的投影算子, 则

$$\hat{\sigma}^2 = \frac{||Y - X\hat{\beta}||}{\dim(R^n) - dim(L)} = \frac{||P_{L^\perp}Y||^2}{n-1} = \frac{\sum_{i=1}^{n}\hat{\epsilon}^2}{n-1} \tag{5}$$

是一个无偏估计.

□

***Solution* 1.c**

我们的假设为 $H_0 : \beta = 0 = L_0$, $\dim L_0 = 0$. 那么我们可以验证在这个模型下 $||P_L Y||^2$ 就是课上所说的 SSR, 而 $(I - H)$ 就是 P_{L^T}. 也就是说下式中 $SSR = ||P_L Y||^2 = Y^T HY, SSE = ||Y - P_L Y||^2 = Y^T(I - H)Y$

定义

$$T := \frac{||P_L Y - P_{L_0} Y||^2/(1-0)}{||Y - P_L Y||^2/(n-1)} = \frac{||P_L Y||^2}{||Y - P_L Y||^2/(n-1)} \sim \frac{\sigma^2\chi_1^2}{\sigma^2\chi_{n-1}^2/(n-1)} = F(1, n-1) \tag{6}$$

方差	自由度	平方和	均方	F 值	p 值
R	1	$\|\|P_L Y\|\|^2$	$\|\|P_L Y\|\|$	$F_0 = \frac{\|\|P_L Y\|\|^2}{\|\|Y - P_L Y\|\|^2/(n-1)}$	$\mathbb{P}\{F \geq F_0\}$
E	n-1	$\|\|Y - P_L Y\|\|^2$	$\|\|Y - P_L Y\|\|^2/(n-1)$		
T	n				

首先是维数上面少了 1, 另外我们还默认了当我们的输入与输出没有线性关系时, 我们的观测就是一个纯粹的均值为 0 的高斯噪声. 我们在计算 SSR 的时候不能直接 plug-in $\hat{y}$ 的平均值去作为 $\bar{y}$ 的估计, 因为在这个模型下 $\bar{y}$ 是已知的就是 0.

□

***Solution* 1.d**

首先我们观察到

$$\hat{\beta} \sim N(\beta, \frac{\sigma^2}{\sum_{i=1}^{n} x_i^2}) \tag{7}$$

其中 σ 为一未知量. 故根据 Studendized theorem, 统计量 T 与他的零分布为,

$$T := \frac{\hat{\beta} - \beta}{\sqrt{\hat{\sigma}^2(\sum_{i=1}^{n} x_i^2)^{-1}}} \sim \frac{N(0,1)}{\chi_{n-1}^2} = t(n-1) \tag{8}$$

而在这样的条件下, 这两个 test 是完全等价的. 因为他们的统计量所服从的分布是等价的, 既 $F(1, n-1) = t(n-1)^2$. 没有一个 test 比另一个更好. □

***Solution* 1.e**

其预测值为

$$y_0 = \hat{\beta} x_0. \tag{9}$$

为了给出他的一个置信区间, 我们考虑 $\hat{y}_0 - y_0$ 所服从的分布.

$$\hat{y}_0 - y_0 \sim N\left(0, \sigma^2\left(1 + \frac{x_0^2}{\sum_{i=1}^n x_i^2}\right)\right) \tag{10}$$

其中 σ 依旧未知, 故将其 studendized. 得到

$$\frac{\hat{y}_0 - y_0}{\sqrt{\hat{\sigma}^2\left(1 + \frac{x_0^2}{\sum_{i=1}^n x_i^2}\right)}} \sim t(n-1) \tag{11}$$

记 $t(n-1)$ 的 $\alpha/2$-**upper quantile** 为 $t_{\alpha/2}$ 所以我们可以构造一个 y_0 的 α 置信区间为

$$y_0 \in \left[\hat{y}_0 - t_{\alpha/2}\sqrt{\hat{\sigma}^2\left(1 + \frac{x_0^2}{\sum_{i=1}^n x_i^2}\right)}, \hat{y}_0 + t_{\alpha/2}\sqrt{\hat{\sigma}^2\left(1 + \frac{x_0^2}{\sum_{i=1}^n x_i^2}\right)}\right] \tag{12}$$

□

Problem 2

2. (P92 2.3 题)考察下列线性回归模型

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i1} x_{i2} + \beta_4 \sqrt{x_{i3}} + \varepsilon_i,\ \ i = 1, 2, \cdots, n,$$

并假定误差项独立同分布于 $N(0, \sigma^2)$.在下列情况下，写出**约简模型，相应的检验统计量和零分布：**

(1) $\beta_3 = \beta_4 = 0$；

(2) $\beta_1 = \beta_2$；

(3) $\beta_4 = 1$.

引入记号 $X_j = (x_{1j}, x_{2j}, \ldots, x_{nj})^T$ 当 $j = 1, 2$, $\sqrt{X_3} = (\sqrt{x_{13}}, \sqrt{x_{23}}, \ldots, \sqrt{x_{n3}})$, $Y = (y_1, y_2, \ldots, y_n)^T$ 与 $\epsilon = (\epsilon_1, \epsilon_2, \ldots, \epsilon_n)^T$. $X_1 \otimes X_2 = (x_{11}x_{12}, x_{21}x_{22}, \ldots, x_{n1}x_{n2})$, 既 X_1, X_2 逐坐标相乘的向量.

假设 X_1, X_2, X_3 是线性独立的并且都不是所有坐标都相同的向量, 既 design matrix 列满秩 (这样的假设对于课上讲的方法是必要的, 否则你根本没有办法去计算 $(X^T X)^{-1}$).

将模型改写为 $Y = X\beta + \epsilon$, 其中 design matrix $X = [\vec{1}, X_1, X_2, X_1 \otimes X_2, \sqrt{X_3}]$. 设 $L = \text{span}\{\vec{1}, X_1, X_2, X_1 \otimes X_2, \sqrt{X_3}\}$

Solution **2.a** 我们的约简模型为

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon_i, i = 1, 2, \ldots, n \tag{13}$$

我们的假设等价于一个对于 y 均值的假设 $\mu \in L_0 := \text{span}\{\vec{1}, X_1, X_2\}$.

检验统计量为 T,

$$T := \frac{||P_L Y - P_{L_0} Y||^2/2}{||Y - P_L Y||^2/(n-5)} \sim \frac{\chi_2^2/2}{\chi_{n-5}^2/(n-5)} = F(2, n-5) \tag{14}$$

□

***Solution* 2.b**
简约模型为

$$y_i = \beta_0 + \beta_1(x_{i1} + x_{i2}) + \beta_3 x_{i1} x_{i2} + \beta_4 \sqrt{x_{i3}} + \epsilon, i = 1, 2, \ldots, n. \tag{15}$$

我们的假设等价于一个对于 y 均值的假设 $\mu \in L_0 := \text{span}\{\vec{1}, X_1 + X_2, X_1 X_2, \sqrt{X_3}\}$

$$T := \frac{||P_L Y - P_{L_0} Y||^2/1}{||Y - P_L Y||^2/(n-5)} \sim \frac{\chi_1^2}{\chi_{n-5}^2/(n-5)} = F(1, n-5) \tag{16}$$

□

***Solution* 2.c**
记 $\bar{y}_i = y_i - \sqrt{x_{i3}}$, 则约简模型为

$$\bar{y}_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i1} x_{i2} + \epsilon, i = 1, 2, \ldots, n. \tag{17}$$

$$T := \frac{(SSE(R) - SSE(F))/2}{SSE(F)/(n-5)} \sim \frac{\chi_2^2/2}{\chi_{n-5}^2/(n-5)} = F(2, n-5) \tag{18}$$

□

Problem 3

3. (P93 2.4 题)某公司管理人员为了解某化妆品在一个城市的月销售量 Y (单位：箱)与该城市中适合使用该化妆品的人数 X_1 (单位：万人)以及他们的人均月收入 X_2(单位：元)之间的关系，在某个月中对 15 个城市做了调查，得上述各量的观测值如下表(可使用 exercise2_4.txt 的数据)所示：

表 1. 化妆品销售数据.

城市	销量(y)	人数(x_1)	收入(x_2)	城市	销量(y)	人数(x_1)	收入(x_2)
1	162	27.4	2450	9	116	19.5	2137
2	120	18.0	3254	10	55	5.3	2560

3	223	37.5	3802	11	252	43.0	4020
4	131	20.5	2838	12	232	37.2	4427
5	67	8.6	2347	13	144	23.6	2660
6	169	26.5	3782	14	103	15.7	2088
7	81	9.8	3008	15	212	37.0	2605
8	192	33.0	2450				

假如Y和X_1， X_2之间满足线性回归关系

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \varepsilon_i,\ \ i = 1, 2, \cdots, 15,$$

其中$\varepsilon_i (i = 1, 2, \cdots, n)$独立同分布于$N(0, \sigma^2)$.

(1) 求回归系数$\beta_0, \beta_1, \beta_2$的最小二乘估计和误差方差$\sigma^2$的估计，写出回归方程并对回归系数作解释；

(2) 给出**方差分析表**，解释对线性回归关系显著性检验的结果，求复相关系数的平方R^2的值并解释其意义；

(3) 分别求β_1, β_2的置信度为95%的**置信区间**；

(4) 对$\alpha = 0.05$，分别检验人数X_1及收入X_2对销量Y的影响是否显著，利用**与回归系数有关的一般假设检验方法**检验X_1和X_2的交互作用(即$X_1 X_2$)对Y的影响是否显著；

(5) 该公司欲在一个适宜使用该化妆品的人数$x_{01} = 220$，人均月收入$x_{02} = 2500$的新的城市中销售该化妆品，求其销量的**预测值及其置信度为95%的置信区间**；

(6) 求Y的拟合值、残差及学生化残差，根据对学生化残差正态性的频率检验及正态 QQ 图检验说明模型误差项的正态性假设是否合理，有序学生化残差与相应标准正态分布的分位数的相关系数是多少？做出各种残差图，分析模型有关假设的合理性.

***Solution* 3.a**

记 $\beta = (\beta_1, \beta_2, \beta_3)^T$. 设其最小二乘估计为 $\hat{\beta}$. 设 $\hat{\sigma}^2$ 为 σ^2 的一个无偏估计. 记 $X = (X_1, X_2)$ 则

$$\beta = (X^TX)^{-1}X^TY = (3.452, 4.960, 0.009)\sigma^2 = ||Y - X\hat{\beta}||^2/(n-3) = 4.74$$

故回归方程为

$$y_i = 3.452 + 4.960x_{i1} + 0.009x_{i2} + \epsilon_i, \epsilon_i \sim N(0, 4.74) \tag{19}$$

□

***Solution* 3.b**

```
> anova(lm3)
Analysis of Variance Table

Response: V1
          Df Sum Sq Mean Sq   F value    Pr(>F)
V2         1  53417   53417 11268.644 < 2.2e-16 ***
V3         1    428     428    90.289 6.201e-07 ***
Residuals 12     57       5
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

我们可以看到, Y 对 X_1, X_2 都有显著的相关性. 在 significance level 0.001 下我们都可以接受其 $R^2 = 0.9989$. 这意味着我们的线性模型能够很好的解释 Y 的变化. □

***Solution* 3.c**

β_1 置信区间为 [4.82,5.09], $beta_2$ 为 [0.007,0.011]

```
> alpha <-0.05
> df<- 12
> coe3<-summary(lm3)$coefficients
> interval <- matrix(c(coe3[,1]-coe3[,2]*qt(1-alpha/2,df),
+                 coe3[,1]+coe3[,2]*qt(1-alpha/2,df)), nrow = 3,ncol = 2)
> interval
             [,1]       [,2]
[1,] -1.843319690 8.74854527
[2,]  4.828134820 5.09196470
[3,]  0.007089742 0.01130842
```

□

***Solution* 3.d**

由 X_1, X_2 的 p 值可以看出其分别都有显著的影响. 而对于新建模型里的 X_1X_2 项的影响并不显著, 其 p 值有 0.862. (注意此时我们的统计量都是服从与 F(1,r) 分布的, 其与 t(r) 统计量等价.)

```
Call:
lm(formula = V1 ~ V2 + V3 + V4, data = problem34)

Residuals:
    Min      1Q  Median      3Q     Max 
-3.9094 -1.2010 -0.1811  1.5072  3.2141 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 4.901e+00  8.539e+00   0.574    0.578    
V2          4.911e+00  2.832e-01  17.344 2.45e-09 ***
V3          8.674e-03  3.124e-03   2.777    0.018 *  
V4          1.698e-05  9.556e-05   0.178    0.862    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.271 on 11 degrees of freedom
Multiple R-squared:  0.9989,	Adjusted R-squared:  0.9987 
F-statistic:  3481 on 3 and 11 DF,  p-value: < 2.2e-16
```

□

***Solution* 3.e**

预测值为 1117.661, 其 95% 置信区间为 [1090.812,1144.511].

```
> predict.lm(lm3,data.frame(V2=220,V3=2500),interval = "prediction")
       fit      lwr      upr
1 1117.661 1090.812 1144.511
```

□

***Solution* 3.f** 其拟合值, 残差与学生化残差分别为

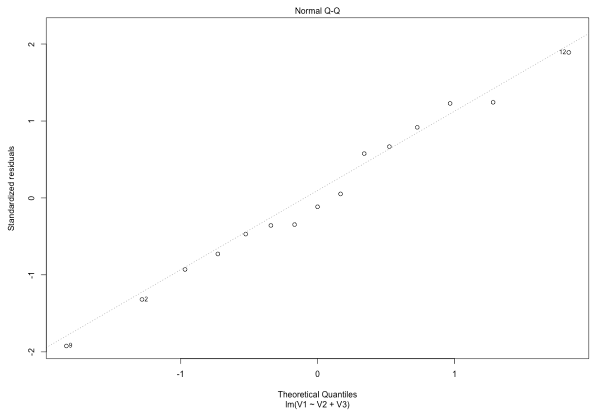
```
> fitted(lm3)
        1         2         3         4          5         6          7
161.89572 122.66732 224.42938 131.24062  67.69928 169.68486  79.73194 189.
       12        13        14        15
228.69079 144.97934 100.53307 210.93806
> residuals(lm3)
         1          2          3          4          5          6
 0.1042756 -2.6673176 -1.4293843 -0.2406244 -0.6992835 -0.6848553  1.26806
        11         12         13         14         15
-1.7150576  3.3092051 -0.9793423  2.4669251  1.0619404
> rstandard(lm3)
          1           2           3           4           5           6
 0.05194039 -1.31980863 -0.72772899 -0.11483379 -0.35782486 -0.34673628  0
         10          11          12          13          14          15
 0.91733453 -0.92965581  1.89099721 -0.46960171  1.24299305  0.57619385
```

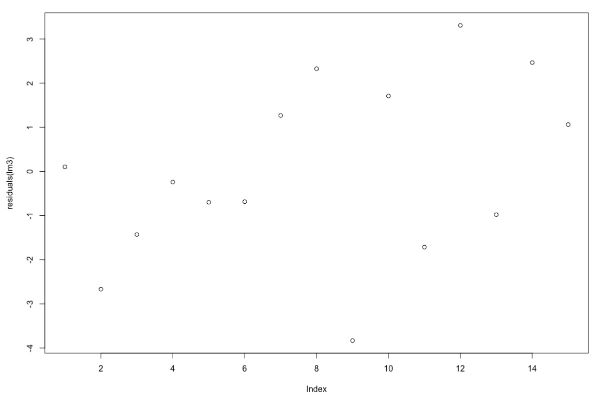
其 shapiro 检测的结果为

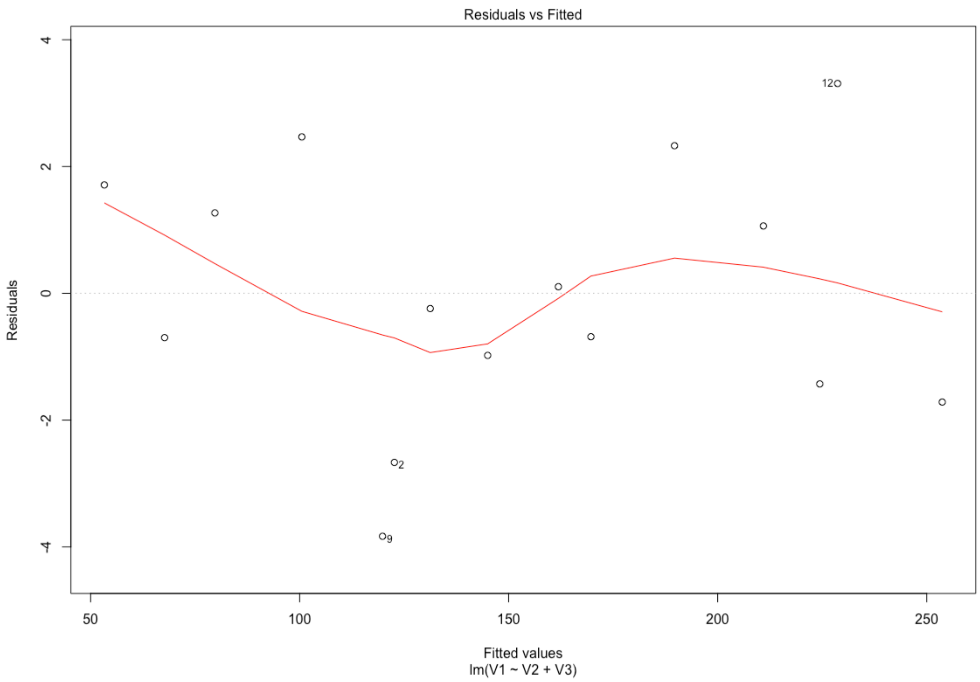
```
> shapiro.test(rstandard(lm3))

        Shapiro-Wilk normality test

data:  rstandard(lm3)
W = 0.98386, p-value = 0.9892
```

相关系数为 0.98386.

下面是各种残差图,







我们的学生化残差近似服从与正态分布, 我们认为我们的正态性假设是合理的.

□

Problem 4

4. 在林业工程中，研究树干的体积Y与离地面一定高度的树干直径X_1和树干高度X_2之间的关系具有重要的实用意义，下表 2 给出了 31 棵树的相关数据(可使用 exercise2_6.txt 的数据).

(1) 拟合线性回归模型$Y=\beta_0+\beta_1X_1+\beta_2X_2+\varepsilon$，通过**残差分析**考察模型的合理性，是否需要对数据作变换？

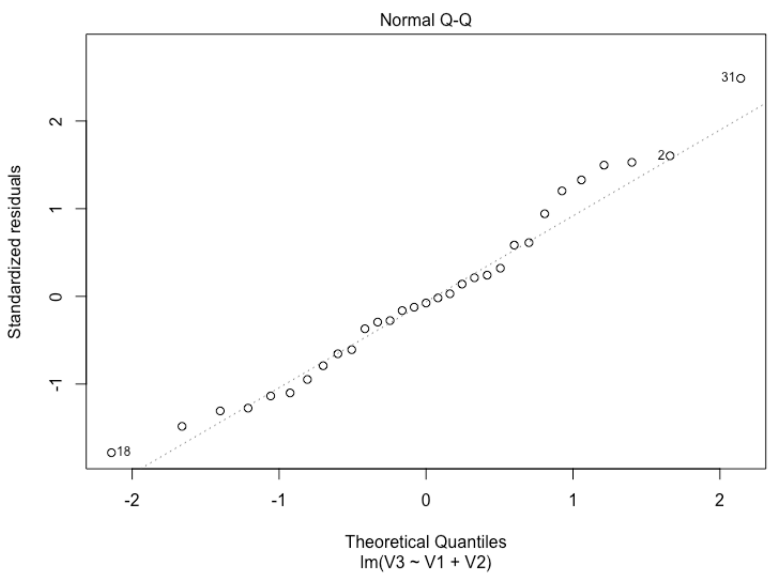
(2) 对因变量Y作 **Box-Cox 变换**，确定变换参数λ的值. 对变换后的因变量重新拟合与X_1，X_2的线性回归模型并作残差分析，Box-Cox 变换的效果如何？

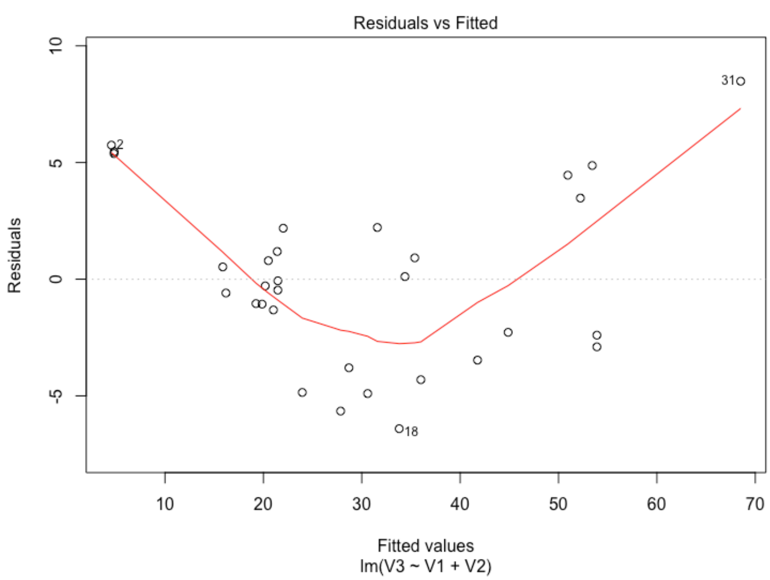
(3) 由于树干可近似看成圆柱或圆台，于是考虑线性回归模型$Y=\beta_0+\beta_1X_1^2+\beta_2X_2+\varepsilon$可能更合理，利用上述数据拟合此模型，进行与(1)、(2)相同的分析，并与前面结果进行比较.

Solution 4.a

```
> shapiro.test(rstandard(lm4))

        Shapiro-Wilk normality test

data:  rstandard(lm4)
W = 0.97414, p-value = 0.6389
```

回归模型为, $y = 10.816 - 0.0455X_1 + 0.169X_2$. 其中 X_1 无显著线性关系. 而通过残差分析的 QQ 图我们发现其有明显形状特征, 需要对其进行变换.

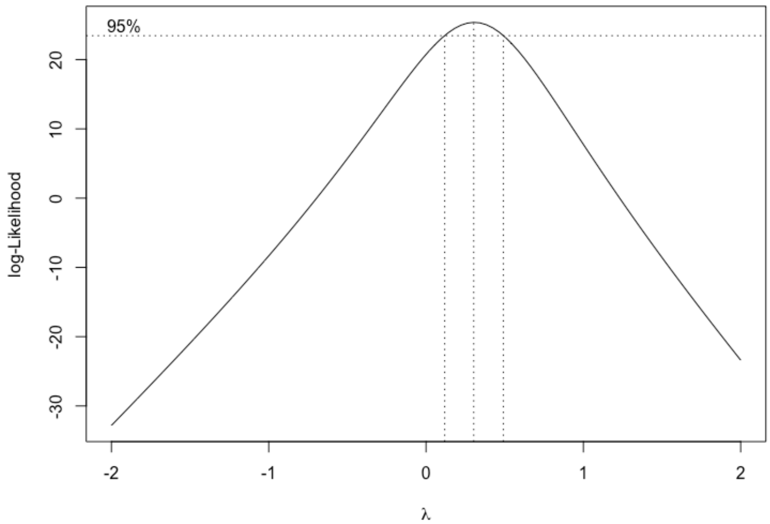




□

Solution **4.b**

确定 box-cox 变换 $\lambda = 0.3$.



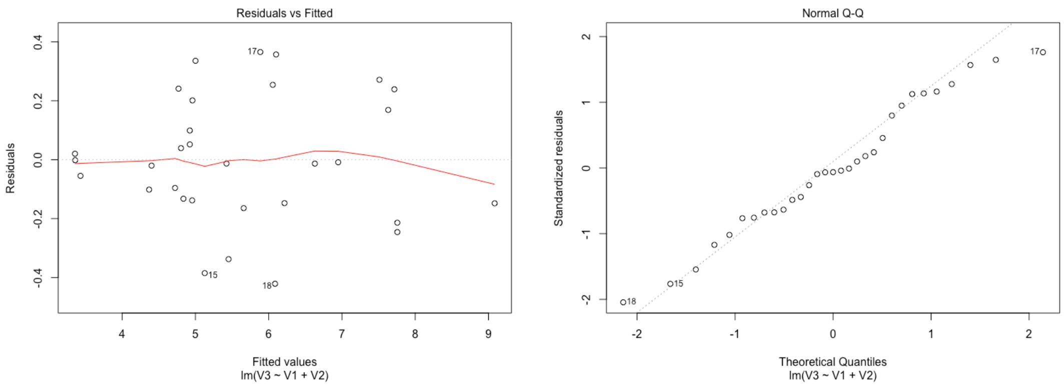
经过变换后的新拟合的模型如下

```
> summary(lm4)

Call:
lm(formula = V3 ~ V1 + V2, data = problem4)

Residuals:
    Min      1Q  Median      3Q     Max 
-6.4065 -2.6493 -0.2876  2.2003  8.4847 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -57.9877     8.6382  -6.713 2.75e-07 ***
V1            4.7082     0.2643  17.816  < 2e-16 ***
V2            0.3393     0.1302   2.607   0.0145 *  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.882 on 28 degrees of freedom
Multiple R-squared:  0.948,	Adjusted R-squared:  0.9442 
F-statistic:   255 on 2 and 28 DF,  p-value: < 2.2e-16
```

我们的 X_2 的影响变得显著了.

其残差分析如下

```
> shapiro.test(rstandard(lm41))

	Shapiro-Wilk normality test

data:  rstandard(lm41)
W = 0.9717, p-value = 0.5669
```



我们可以看到我们的 box-cox 变换在提升正态性上有一定效果.

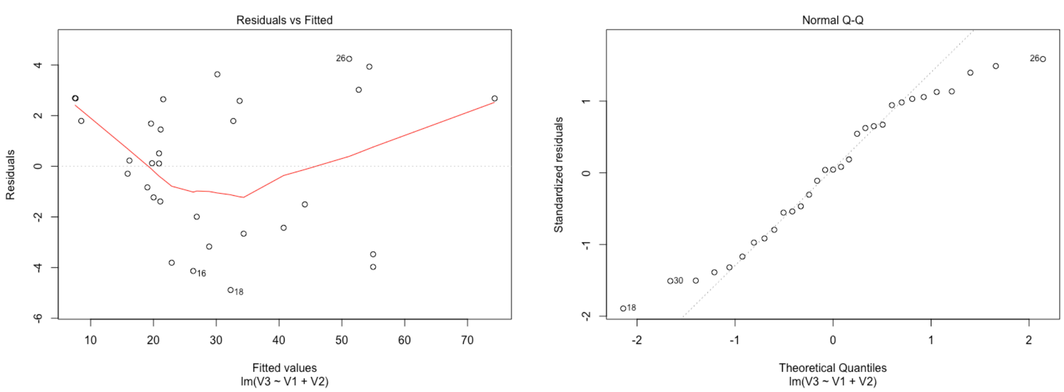
□

Solution **4.c**

拟合模型为 $Y = -27.5 + 0.17X_1^2 + 0.35X_2^2$.

```
Call:
lm(formula = V3 ~ V1 + V2, data = problem42)

Residuals:
    Min      1Q  Median      3Q     Max 
-4.8844 -2.2105  0.1196  2.6134  4.2404 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) -27.511603   6.557697  -4.195 0.000248 ***
V1            0.168458   0.006679  25.222  < 2e-16 ***
V2            0.348809   0.093152   3.744 0.000830 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.799 on 28 degrees of freedom
Multiple R-squared:  0.9729,	Adjusted R-squared:  0.971 
F-statistic: 503.2 on 2 and 28 DF,  p-value: < 2.2e-16
```

我们对其进行残差分析,



我们发现对比 (1)(2) 模型都有了一定的提升. 但是其残差相对正态分布来说具有厚尾性.

□

Problem 5

5. 某医院为了解患者对医院工作失误满意程度 Y 和患者的年龄 X_1，病情的严重程度 X_2 和患者的忧患程度 X_3 之间的关系，随机调查了该医院的 23 位患者，得数据如表 3(可使用 exercise2_9.txt 的数据)所示.

(1) 拟合线性回归模型 $Y=\beta_0+\beta_1X_1+\beta_2X_2+\beta_3X_3+\varepsilon$，通过**残差分析**考察模型及有关误差分布正态性假设的合理性.

(2) 若(1)中假设合理，分别在 $R_a^2(p), C_p, PRESS_p$ **准则**下选择最优回归方程，各准则下的最优方程是否一致？

(3) 对 $\alpha_E=\alpha_D=0.10$，用**逐步回归法**选择最优回归方程，其结果和(2)中的是否一致？

(4) 对选择的最优回归方程作残差分析，与(1)中的相应结果比较，有何变化？

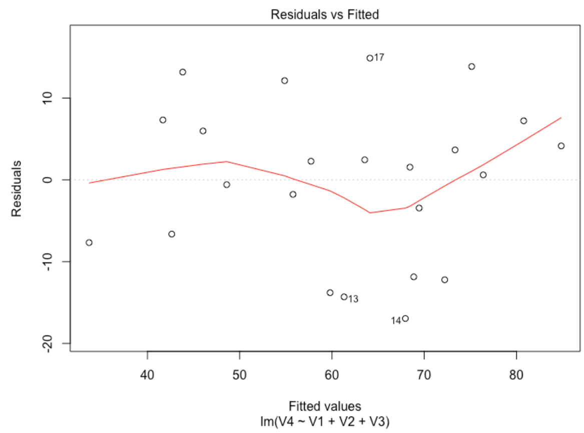
Solution 5.a

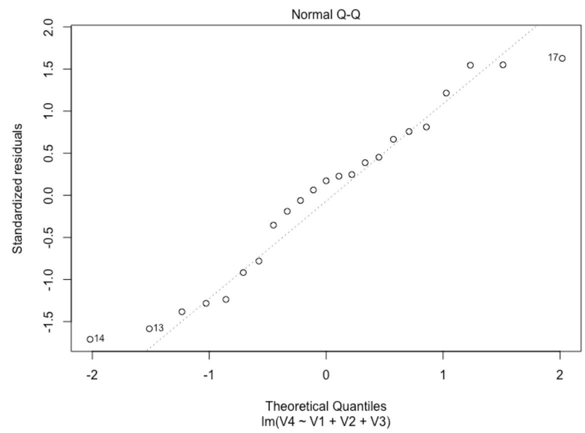
拟合的线性回归模型为 $Y = 162.876 - 1.210X_1 - 0.6659X_2 - 8.613X_3 + \epsilon$

```
Call:
lm(formula = V4 ~ V1 + V2 + V3, data = problem5)

Residuals:
    Min      1Q  Median      3Q     Max 
-16.954  -7.154   1.550   6.599  14.888 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 162.8759    25.7757   6.319 4.59e-06 ***
V1           -1.2103     0.3015  -4.015  0.00074 ***
V2           -0.6659     0.8210  -0.811  0.42736    
V3           -8.6130    12.2413  -0.704  0.49021    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 10.29 on 19 degrees of freedom
Multiple R-squared:  0.6727,	Adjusted R-squared:  0.621 
F-statistic: 13.01 on 3 and 19 DF,  p-value: 7.482e-05
```

通过残差分析我们认为正态性的假设是基本合理的. 可以参考图像与其 p-value

```
> shapiro.test(rstandard(lm5))

	Shapiro-Wilk normality test

data:  rstandard(lm5)
W = 0.95176, p-value = 0.3182
```





$\square$

Solution **5.b**

```
> summary(lm5f)
Subset selection object
Call: regsubsets.formula(V4 ~ V1 + V2 + V3, data = problem5, nbest = 3)
3 Variables  (and intercept)
   Forced in Forced out
V1     FALSE      FALSE
V2     FALSE      FALSE
V3     FALSE      FALSE
3 subsets of each size up to 3
Selection Algorithm: exhaustive
         V1  V2  V3
1  ( 1 ) "*" " " " "
1  ( 2 ) " " " " "*"
1  ( 3 ) " " "*" " "
2  ( 1 ) "*" "*" " "
2  ( 2 ) "*" " " "*"
2  ( 3 ) " " "*" "*"
3  ( 1 ) "*" "*" "*"
> summary(lm5f)$adjr2
[1] 0.5794702 0.3324340 0.3139047 0.6305423 0.6274569 0.3344295 0.6209731
> summary(lm5f)$cp
[1]  4.299472 17.986519 19.013133  2.495063  2.657873 18.119959  4.000000
```

我们可以看出在 $R_a^2(p)$ 我们应该选择 X_1, X_2.

在 C_p 准则下应该选择 X_1, X_2, X_3 作为我们的模型.

```
#V1 3024.209
#V1 V2 2714.105
#V2 V3 4966.428
#V1 V3 2693.434
#V2 4853.28
#V3 4652.835
#V1 V2 V3 3046.291
```

在 PRESS 准则下我们应该选择 V_1, V_3 作为我们的模型.

所以三个准则并不一致.

□

Solution **5.c**

```
> lm53 <- lm(V4~1,data = problem5)
> add1(lm53, scope = ~V1+V2+V3,test = "F")
Single term additions

Model:
V4 ~ 1
       Df Sum of Sq    RSS    AIC F value    Pr(>F)
<none>              6145.2 130.52
V1      1    3678.4 2466.8 111.53  31.315 1.489e-05 ***
V2      1    2120.7 4024.6 122.79  11.066  0.003205 **
V3      1    2229.3 3915.9 122.16  11.956  0.002356 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

选择 p 值最小的 X_1 加入模型.

```
> lm53 <- lm(V4~V1,data = problem5)
> add1(lm53, scope = ~V1+V2+V3,test = "F")
Single term additions

Model:
V4 ~ V1
       Df Sum of Sq    RSS    AIC F value  Pr(>F)
<none>              2466.8 111.53
V2      1    402.78 2064.0 109.43  3.9029 0.06216 .
V3      1    385.55 2081.2 109.62  3.7050 0.06859 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

选择 p 值最小的 X_2 加入模型.

```
> lm53 <- lm(V4~V1+V2,data = problem5)
> add1(lm53, scope = ~V1+V2+V3,test = "F")
Single term additions

Model:
V4 ~ V1 + V2
       Df Sum of Sq    RSS    AIC F value Pr(>F)
<none>              2064.0 109.43
V3      1    52.414 2011.6 110.84  0.4951 0.4902
```

X_3 的 p 值大于 0.1, 停止加入.

```
> lm53 <- lm(V4~1,data = problem5)
> add1(lm53, scope = ~V1+V2+V3,test = "F")
Single term additions

Model:
V4 ~ 1
       Df Sum of Sq    RSS    AIC F value    Pr(>F)
<none>              6145.2 130.52
V1      1    3678.4 2466.8 111.53  31.315 1.489e-05 ***
V2      1    2120.7 4024.6 122.79  11.066  0.003205 **
V3      1    2229.3 3915.9 122.16  11.956  0.002356 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

检验发现 p 值都小于 0.1 故停止.

最后的模型为

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon \tag{20}$$

□

Solution **5.d**

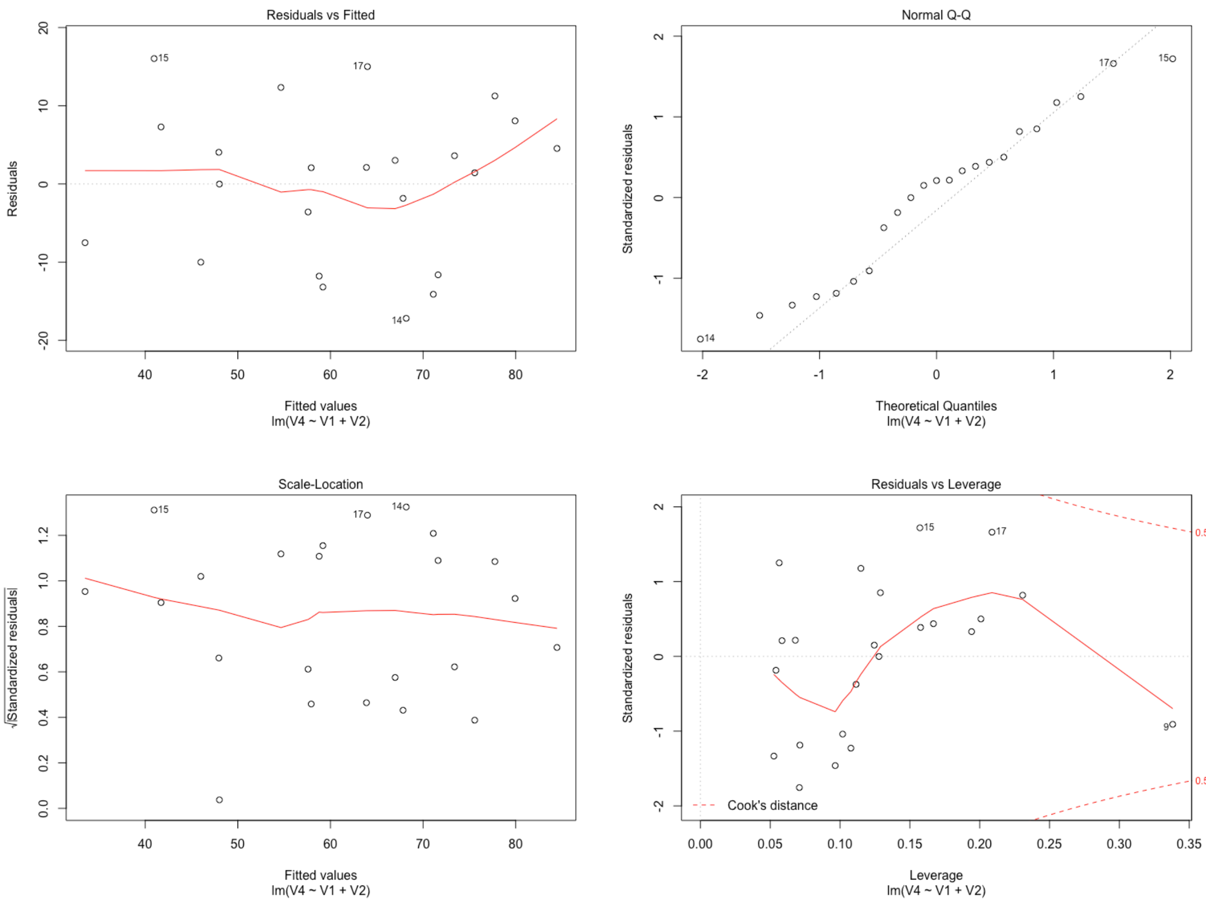
拟合的模型为 $Y = 166.591 - 1.261X_1 - 1.08X_2 + \epsilon$

```
> lm54 <- lm(V4~V1+V2, data = problem5)
> summary(lm54)

Call:
lm(formula = V4 ~ V1 + V2, data = problem5)

Residuals:
    Min      1Q  Median      3Q     Max
-17.180  -8.758   2.074   5.916  16.036

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 166.5913    24.9084   6.688 1.65e-06 ***
V1           -1.2605     0.2892  -4.359 0.000304 ***
V2           -1.0893     0.5514  -1.976 0.062163 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 10.16 on 20 degrees of freedom
Multiple R-squared:  0.6641,	Adjusted R-squared:  0.6305
F-statistic: 19.77 on 2 and 20 DF,  p-value: 1.827e-05
```

方差分析如下

```
> shapiro.test(rstudent(lm54))

        Shapiro-Wilk normality test

data:  rstudent(lm54)
W = 0.96093, p-value = 0.4825
```

我们的的正态性有了一定提升. 但是提升比较有限.



□

Problem 6

6. 表 4 是 66 家金融公司当时在财务运营指标 X_1, X_2 和 X_3 上的数据以及标示两年后各公司是否破产的变量 Y 的取值，其中

(1) 建立 P(Y=1)与 X_1, X_2和 X_3的 Logistic 回归模型，分析全局回归关系的显著性及各自变量对概率 P(Y=1)的影响.

(2) 利用似然比检验方法在显著性水平 $\alpha=0.05$ 下，检验自变量 X_3对 P(Y=1)的影响是否显著. 若 X_3的影响不显著，建立仅含 X_1, X_2的 Logistic 回归模型，分析全局回归关系的显著性，给出各公司关于 P(Y=1)的拟合值并分析有关结果.

(3) 假设某金融公司在 X_1，X_2 和 X_3 三个指标上的当前值为 $x_1=48.8, x_2=-10.5, x_3=1.8$，分别用(1)和(2)建立的模型预测该公司两年后不会破产的概率，二者的概率差别如何？

***Solution* 6.a**

模型为

$$\ln\frac{y}{1-y} = -20.002 + 0.607X_1 + 0.178X_2 + 8.986 \tag{21}$$

```
> summary(glm6)

Call:
glm(formula = V5 ~ V2 + V3 + V4, family = binomial(link = logit),
    data = problem6)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-1.54955   0.00000   0.00000   0.00003   1.40812

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -20.0022    32.7839  -0.610    0.542
V2            0.6069     0.9562   0.635    0.526
V3            0.1777     0.1243   1.429    0.153
V4            8.9862    13.6422   0.659    0.510

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 91.4954  on 65  degrees of freedom
Residual deviance:  5.2115  on 62  degrees of freedom
AIC: 13.211

Number of Fisher Scoring iterations: 14
```

全局显著性并不高, 其中只有 X_2 显著性比较高.

□

***Solution* 6.b**

```
> lrtest(glm62,glm6)

Model 1: V5 ~ V2 + V3
Model 2: V5 ~ V2 + V3 + V4

L.R. Chisq       d.f.          P
4.26044374 1.00000000 0.03900973
```

0.039<0.05, 不显著.

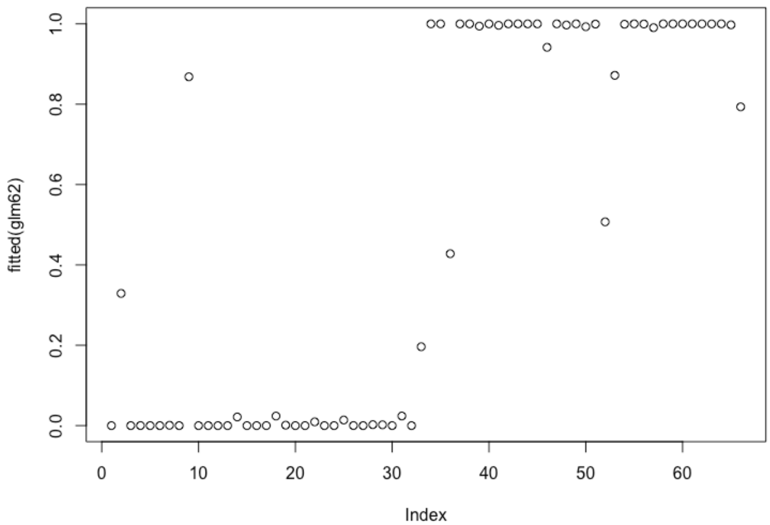
我们的模型为

$$\ln\frac{y}{1-y} = -0.550 + 0.157X_1 + 0.195X_2 + \epsilon. \tag{22}$$

```
Call:
glm(formula = V5 ~ V2 + V3, family = binomial(link = logit),
    data = problem6)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-2.01334  -0.00658   0.00095   0.01421   1.30309

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.55037    0.95098  -0.579   0.5628
V2           0.15737    0.07492   2.101   0.0357 *
V3           0.19475    0.12244   1.591   0.1117
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 91.4954  on 65  degrees of freedom
Residual deviance:  9.4719  on 63  degrees of freedom
AIC: 15.472

Number of Fisher Scoring iterations: 10
```

并且我们的 X_2 具有在 $\alpha = 0.05$ 时可以认为具有显著性.

根据下图我们的预测值只有个别分错的, 性能十分出色.



□

***Solution* 6.c**

这个公司两个模型预测得到的不会破产概率区别不大, 几乎都是 1.

```
> exp(predict(glm6,newdata = new))/(1+exp(predict(glm6,newdata = new)))
1
1

> exp(predict(glm62,newdata = new))/(1+exp(predict(glm62,newdata = new)))
        1
0.9938452
```

□