应用回归分析第七章

邵智轩　　1400012141　　物理学院

7.11

7.11.1

原假设为 “main effect model”，即 RK 与 YR, DG, YD 没有交互项，

$$\text{NH: } SL = \beta_0 + \beta_{02}RK_2 + \beta_{03}RK_3 + \beta_1 YR + \beta_2 DG + \beta_3 YD + e$$

共 $p = 6$ 个参数

```
library(alr4)
lm.main <- lm(salary ~ rank + degree + year + ysdeg)
```

备择假设有 RK 与 YR, DG, YD 有交互项（共 $3 \times 2 = 6$ 项），

$$\begin{aligned}\text{AH: } SL \quad = \quad & \beta_0 + \beta_{02}RK_2 + \beta_{03}RK_3 + \beta_1 YR + \beta_2 DG + \beta_3 YD \\ & + \beta_{12}RK_2 \cdot YR + \beta_{13}RK_3 \cdot YR + \beta_{22}RK_2 \cdot DG + \\ & \beta_{23}RK_3 \cdot DG + \beta_{32}RK_2 \cdot YD + \beta_{33}RK_3 \cdot YD + e\end{aligned}$$

共 $p = 12$ 个参数。

```
lm.interaction <- lm(salary ~ rank * degree + rank * year + rank * ysdeg)
anova(lm.main, lm.interaction)
```

```
## Analysis of Variance Table
##
## Model 1: salary ~ rank + degree + year + ysdeg
## Model 2: salary ~ rank * degree + rank * year + rank * ysdeg
##   Res.Df       RSS Df Sum of Sq      F Pr(>F)
```

```
## 1     46 267993336
## 2     40 225487139  6  42506197 1.2567  0.299
```

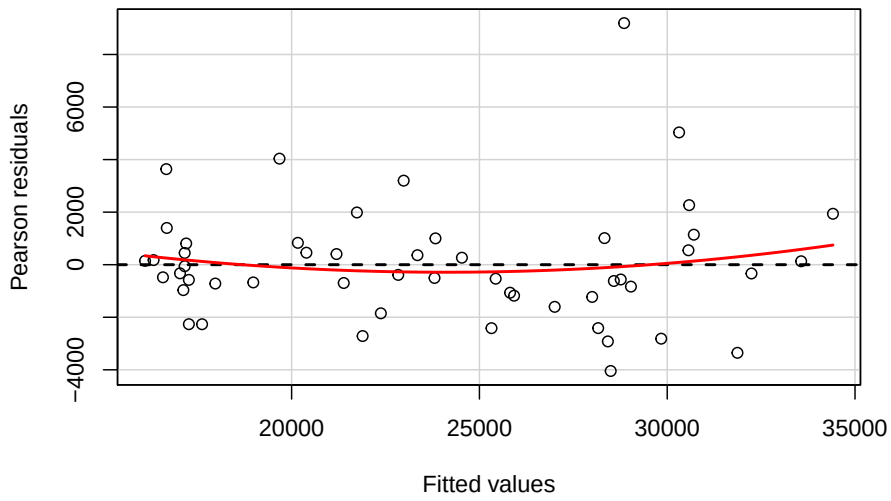
```
Anova(lm.interaction)
```

```
## Anova Table (Type II tests)
##
## Response: salary
##                 Sum Sq Df F value    Pr(>F)
## rank        404108665  2 35.8432 1.206e-09 ***
## degree        3212377  1  0.5699 0.4547383
## year         94884059  1 16.8318 0.0001949 ***
## ysdeg         3488722  1  0.6189 0.4361011
## rank:degree  11956591  2  1.0605 0.3558127
## rank:year     4837784  2  0.4291 0.6540589
## rank:ysdeg   19244381  2  1.7069 0.1943736
## Residuals   225487139 40
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

F 统计量并不显著（$p = 0.299$），每一项交互项也都不显著，故不拒绝原假设。应认为交互项并不显著，不拒绝原假设：对于每一组 rank，year、degree、ysdeg``的不同对 salary‘的调整是相同的。

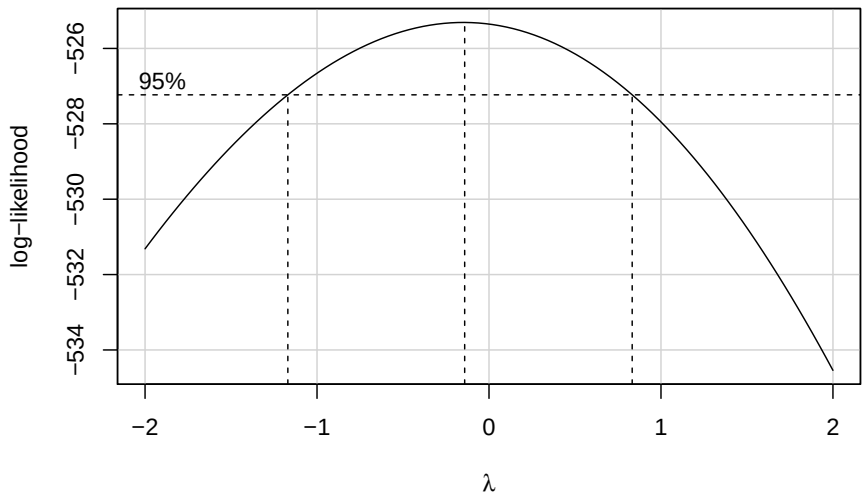
7.11.2 证明需要对响应变量 SL 变换，并找到该变换

```
# 使用所有的变量拟合模型
lm.all <- lm(salary ~ rank + degree + year + ysdeg + sex)
residualPlot(lm.all)
```



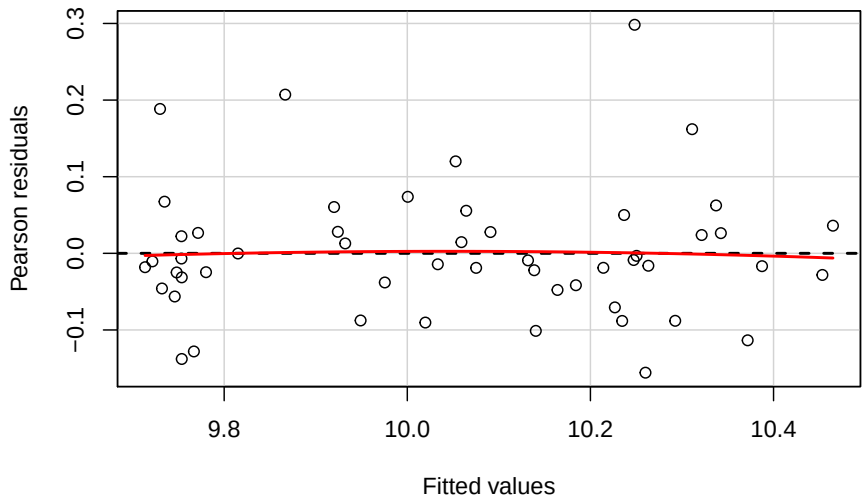
从残差图中看到明显的 U 型 pattern，提示我们对 Y 做变换。

```
boxCox(lm(salary ~ degree + year + ysdeg + sex))
```



$\hat{\lambda} \approx 0$，这与我们的直觉也是一致的。对 `salary` 作对数变换。

```
salary.log <- log(salary)
lm.log <- lm(salary.log ~ rank + degree + year + ysdeg + sex)
residualPlot(lm.log)
```



此时残差图不再有明显的非线性痕迹。

7.11.3 检验非常数方差

```
e <- lm.log$residuals
u <- e ^ 2 / mean(e ^ 2)
```

检验作为 salary 的一个函数

```
lm.ncon.salary <- lm(u ~ salary.log + 1)
SSreg.salary <- sum((lm.ncon.salary$fitted.values - mean(u)) ^ 2)
S.salary <- SSreg.salary / 2
S.salary
```

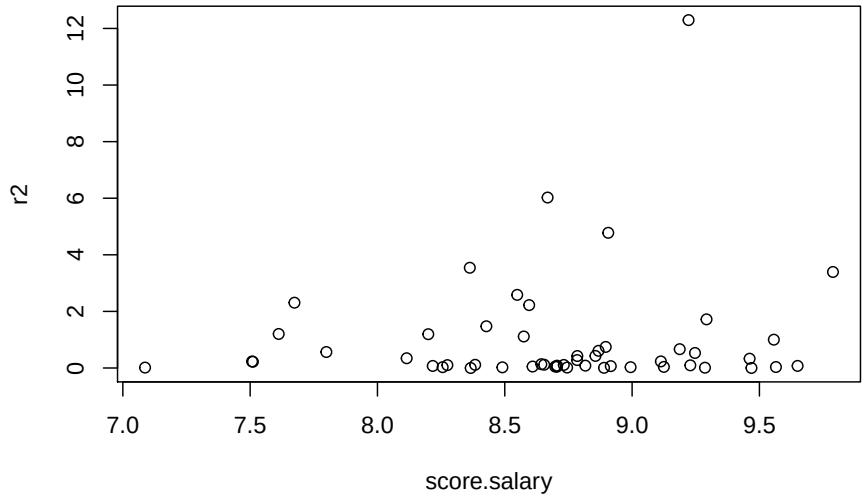
```
## [1] 5.706632
```

```
pchisq(S.salary, df = 1, lower.tail = FALSE)
```

```
## [1] 0.01690093
```

p 值是显著的，我们可以再考察 r_i^2 关于 $(1-h_{ii})z_i$ 的图

```
r2 <- (lm.log$residuals / sigma(lm.log)) ^ 2 / (1 - hatvalues(lm.log)) # 学生化内残差
score.salary <- (1 - hatvalues(lm.log)) * salary.log
plot(score.salary, r2)
```



有明显的楔形形状，表示有非常数方差。

对于变量 sex，可用同样的方法：

```
lm.ncon.sex <- lm(u ~ sex + 1)
SSreg.sex <- sum((lm.ncon.sex$fitted.values - mean(u)) ^ 2)
S.sex <- SSreg.sex / 2
S.sex
```

```
## [1] 5.785141
```

```
pchisq(S.sex, df = 1, lower.tail = FALSE)
```

```
## [1] 0.0161622
```

p 值也是显著的，支持拒绝常数方差的假设。然而变量 sex，不同于连续变量 salary，它是一个 binary factor，也许并不适合用刚才的方法检验，不如直接比较不同性别的残差的均方：

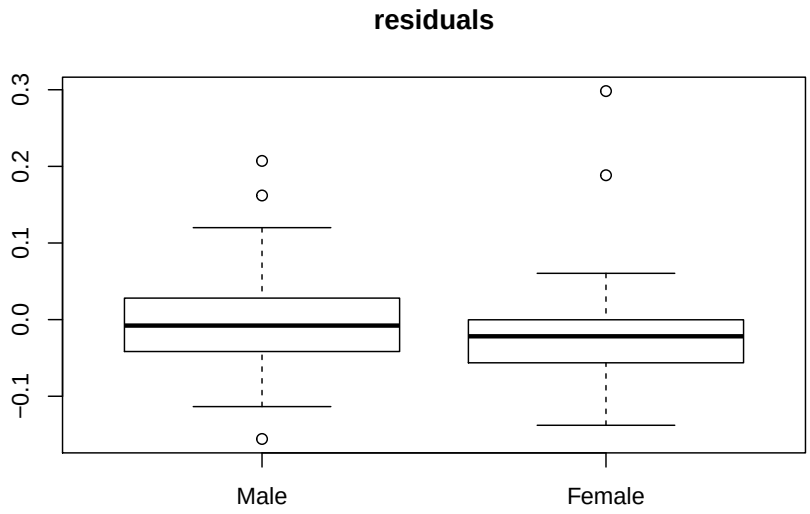
```
c(mean(e[sex=="Male"]),mean(e[sex=="Female"]))
```

```
## [1] -6.068768e-19  1.774413e-18
```

```
c(mean(e[sex=="Male"]^2),mean(e[sex=="Female"]^2,mean(e^2)))
```

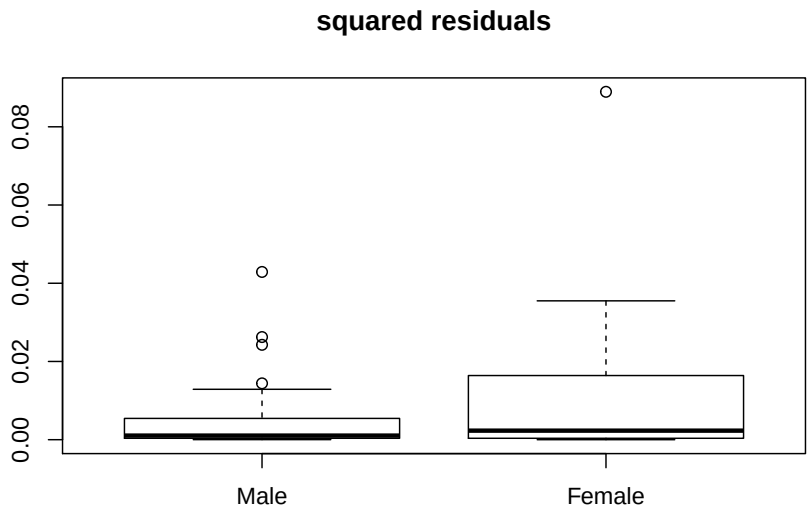
```
## [1] 0.005110096 0.012724565
```

```
boxplot(list("Male"=e[sex == "Male"], "Female"=e[sex == "Female"]),main="residuals")
```



可以看到，虽然两组的残差均值都为 0，但男性的均方残差比女性的均方残差比均方残差小。从残差 boxplot 中也可看出，女性的残差大部分都集中在小于 0 的部分。

```
boxplot(list("Male"=e[sex == "Male"]^2, "Female"=e[sex == "Female"]^2),main="squared re
```



基于以上分析，我认为，两组的残差有明显的不同，非常数方差不成立。

7.11.4 检验对变换后的薪水，在每种职位中，性别的差别是否一样。

即检验，职位和性别的交互项 `SX:RK` 是否显著。

$$\text{NH:} \quad \log SL = \beta_0 + \beta_{02} RK_2 + \beta_{03} RK_3 + \beta_1 YR + \beta_2 DG + \beta_3 YD + \beta_4 SX + e$$

$$\begin{aligned} \text{AH:} \quad \log SL \;=\; & \beta_0 + \beta_{02} RK_2 + \beta_{03} RK_3 + \beta_1 YR + \beta_2 DG \\ & + \beta_3 YD + \beta_4 SX + \beta_{42} SX \cdot RK_2 + \beta_{43} SX \cdot RK_3 + e \end{aligned}$$

```
lm.log.ia <- lm(salary.log ~ rank * sex + degree + year + ysdeg)
Anova(lm.log.ia)
```

```
## Anova Table (Type II tests)
##
```

```
## Response: salary.log
##             Sum Sq Df F value    Pr(>F)
## rank       0.74539  2 45.1346 2.741e-11 ***
## sex        0.01079  1  1.3069   0.25928
## degree     0.03252  1  3.9377   0.05362 .
## year       0.21137  1 25.5977 8.331e-06 ***
## ysdeg      0.02813  1  3.4064   0.07184 .
## rank:sex   0.01726  2  1.0451   0.36041
## Residuals 0.35507 43
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

交互项并不显著，故不拒绝原假设。应认为，在每种职位中，性别引起的薪水差别是几乎相同的。

7.11.5

```
summary(lm(salary~sex))$coef
```

```
##               Estimate Std. Error  t value     Pr(>|t|)
## (Intercept) 24696.789   937.9776 26.32983 5.761530e-31
## sexFemale   -3339.647  1807.7156 -1.84744 7.060394e-02
```

虽然男性平均工资比女性高，且差异是显著的，但是这时没有考虑其他因素的影响，有可能是其他因素导致的，并不能认为单纯由于性别引起了工资差异。

```
summary(lm.log) #sex 项不显著
```

```
##
## Call:
## lm(formula = salary.log ~ rank + degree + year + ysdeg + sex)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
```

```
## -0.15574 -0.04268 -0.01239  0.02783  0.29824
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  9.700334   0.030347 319.645  < 2e-16 ***
## rankAssoc    0.260419   0.043440   5.995 3.18e-07 ***
## rankProf     0.485491   0.051267   9.470 2.78e-12 ***
## degreePhD    0.073211   0.038636   1.895   0.0645 .
## year         0.018482   0.003600   5.134 5.88e-06 ***
## ysdeg       -0.005228   0.002939  -1.779   0.0820 .
## sexFemale    0.040089   0.035103   1.142   0.2595
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.09096 on 45 degrees of freedom
## Multiple R-squared:  0.881,  Adjusted R-squared:  0.8651
## F-statistic: 55.52 on 6 and 45 DF,  p-value: < 2.2e-16
```

```
confint(lm.log,c('sexFemale'))
```

```
##                 2.5 %    97.5 %
## sexFemale -0.03061121 0.1107896
```

模型 `logsalary ~ rank + degree + year + ysdeg + sex` 中，`sexFemale` 项系数不显著，其 95% 置信区间为 $[-3.1\%, 11.1\%]$，这一系数正是性别引起的薪水的相对变化：

$$\Delta \log SL \approx \frac{\Delta SL}{SL}$$

故在法庭上可以汇报如下：

”虽然男性平均工资比女性高，且差异是显著的，但是这没有考虑其他因素的影响，有可能是其他变量导致的，并不能认为单纯由于性别引起了薪水差异。

“在考虑了职称，最高学历，性别，在职年数，获最高学历后年数这五个变量的模型中，女性与男性的收入差距的 95% 置信区间是包含 0 的。也

就是说，我们没有充足的证据证明性别对薪水有直接影响；这一系数的点估计是正的，说明在调整了其他变量的影响后，其实女性有更高的平均薪水，虽然这一正效应并不显著。”

7.11.6

```
summary(update(lm.log, ~ . - rank))
```

```
##
## Call:
## lm(formula = salary.log ~ degree + year + ysdeg + sex)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.33620 -0.11026 -0.00152  0.10254  0.38411
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  9.771240   0.047289 206.629  < 2e-16 ***
## degreePhD   -0.122787   0.053657  -2.288   0.0267 *
## year         0.012402   0.005869   2.113   0.0399 *
## ysdeg        0.015247   0.003321   4.591  3.3e-05 ***
## sexFemale   -0.073968   0.054092  -1.367   0.1780
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1542 on 47 degrees of freedom
## Multiple R-squared:  0.6427, Adjusted R-squared:  0.6123
## F-statistic: 21.14 on 4 and 47 DF,  p-value: 5.035e-10
```

在去掉了认为“被污染”的 `rank` 后，女性的影响变成了负的，但仍不显著。