3.4 自变量有分类变量的情况

徐平峰

长春工业大学

2021/3/17

工资数据

```
(w=read.csv("E:/teaching_plan_notes/msa11091083/吴喜之课件/《多元统计分析
```

```
#       Age  Edu Income Sex
# 1     Old Grad 520000   F
# 2  Middle Coll  51000   F
# 3   Young   Hs   1200   F
# 4     Old  Ele   1500   F
# 5  Middle Grad 200000   M
# 6   Young Coll  15000   M
# 7     Old   Hs   2000   M
# 8  Middle  Ele   1100   F
# 9   Young Grad 150000   M
# 10    Old Coll  10000   F
# 11 Middle   Hs   3000   F
# 12  Young  Ele    500   M
```

```
class(w$Age)
```

```
# [1] "character"
```

工资数据

```
a=lm(Income~., data=w);summary(a)#以Middle， coll, F作为参考水平
```

```
#
# Call:
# lm(formula = Income ~ ., data = w)
#
# Residuals:
#       1       2       3       4       5       6       7       8       9
#  134967   27744  -17956  -67056  -32900   55111   14944    2144 -102067  -8
#      11      12
#    3011   64911
#
# Coefficients:
#             Estimate Std. Error t value Pr(>|t|)
# (Intercept)    23256      71440   0.326   0.7580
# AgeOld         69600      70468   0.988   0.3687
# AgeYoung       19167      78786   0.243   0.8175
# EduEle        -24300      81369  -0.299   0.7772
# EduGrad       292178      84692   3.450   0.0182 *
# EduHs         -23267      81369  -0.286   0.7864
# SexM          -82533      70468  -1.171   0.2943
# ---
# Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#
# Residual standard error: 99660 on 5 degrees of freedom
# Multiple R-squared:  0.8089,    Adjusted R-squared:  0.5795
# F-statistic: 3.526 on 6 and 5 DF,  p-value: 0.09399
```

工资数据

- 变量Age变为三个哑变量 (AM, AO, AY)，其中只能由一个等于1，其余等于零。例如，Age=Middle, 则 $(AM, AO, AY) = (1, 0, 0)$。
- 变量Edu变为四个哑变量 (EC, EE, EG, EH)
- 变量Sex变为 (SM, SF)

回归方程为：

$$\begin{aligned} income = \beta_0 + \beta_{AM} x_{AM} + \beta_{AO} x_{AO} + \beta_{AY} x_{AY} \\ + \beta_{EC} x_{EC} + \beta_{EE} x_{EE} + \beta_{EG} x_{EG} + \beta_{EH} x_{EH} \qquad + \beta_{SM} x_{SM} + \beta_{SF} x_{SF} + \varepsilon \end{aligned}$$

为了保证可识别性， $\beta_{AM} = 0, \beta_{EC} = 0, \beta_{SM} = 0$回归方程为：

$$\begin{aligned} income = \beta_0 + 0 x_{AM} + \beta_{AO} x_{AO} + \beta_{AY} x_{AY} \qquad + 0 x_{EC} + \beta_{EE} x_{EE} + \beta_{EG} x_{EG} + \beta_{EH} x_{EH} \\ + 0 x_{SM} + \beta_{SF} x_{SF} + \varepsilon \end{aligned}$$

工资数据

- 哑变量

```
wf<-w;for(i in c(1,2,4))wf[,i]<-factor(wf[,i])
wd<-dummies::dummy.data.frame(wf[,-3], names= c("Age","Edu","Sex" ),
```

```
# Warning in model.matrix.default(~x - 1, model.frame(~x - 1), contrasts = FA
# non-list contrasts argument ignored

# Warning in model.matrix.default(~x - 1, model.frame(~x - 1), contrasts = FA
# non-list contrasts argument ignored

# Warning in model.matrix.default(~x - 1, model.frame(~x - 1), contrasts = FA
# non-list contrasts argument ignored
```

```
wd # 哑变量
```

```
#    AgeMiddle AgeOld AgeYoung EduColl EduEle EduGrad EduHs SexF SexM
# 1          0      1        0       0      0       1     0    1    0
# 2          1      0        0       1      0       0     0    1    0
# 3          0      0        1       0      0       0     1    1    0
# 4          0      1        0       0      1       0     0    1    0
# 5          1      0        0       0      0       1     0    0    1
# 6          0      0        1       1      0       0     0    0    1
# 7          0      1        0       0      0       0     1    0    1
# 8          1      0        0       0      1       0     0    1    0
# 9          0      0        1       0      0       1     0    0    1
# 10         0      1        0       1      0       0     0    1    0
# 11         1      0        0       0      0       0     1    1    0
# 12         0      0        1       0      1       0     0    0    1
```

```
w # 原始数据
```

```
#       Age  Edu Income Sex
# 1     Old Grad 520000   F
# 2  Middle Coll  51000   F
# 3   Young   Hs   1200   F
# 4     Old  Ele   1500   F
# 5  Middle Grad 200000   M
# 6   Young Coll  15000   M
# 7     Old   Hs   2000   M
# 8  Middle  Ele   1100   F
# 9   Young Grad 150000   M
# 10    Old Coll  10000   F
# 11 Middle   Hs   3000   F
# 12  Young  Ele    500   M
```

工资数据

- 哑变量

```
(X=cbind(rep(1,12), wd[,-c(1,4,8)]))
```

```
#    rep(1, 12) AgeOld AgeYoung EduEle EduGrad EduHs SexM
# 1           1      1        0      0       1     0    0
# 2           1      0        0      0       0     0    0
# 3           1      0        1      0       0     1    0
# 4           1      1        0      1       0     0    0
# 5           1      0        0      0       1     0    1
# 6           1      0        1      0       0     0    1
# 7           1      1        0      0       0     1    1
# 8           1      0        0      1       0     0    0
# 9           1      0        1      0       1     0    1
# 10          1      1        0      0       0     0    0
# 11          1      0        0      0       0     1    0
# 12          1      0        1      1       0     0    1
```

```
class(X)
```

```
# [1] "data.frame"
```

```
X=as.matrix(X)
```

工资数据

- 系数的最小二乘估计

```
(beta.fit<-solve(t(X)%*%X)%*%t(X)%*%(matrix(w$Income, ncol=1)))
```

```
#                  [,1]
# rep(1, 12)  23255.56
# AgeOld      69600.00
# AgeYoung    19166.67
# EduEle     -24300.00
# EduGrad    292177.78
# EduHs      -23266.67
# SexM       -82533.33
```

```
coef(a)
```

```
# (Intercept)      AgeOld    AgeYoung      EduEle     EduGrad       EduHs
#    23255.56    69600.00    19166.67   -24300.00   292177.78   -23266.67
#        SexM
#   -82533.33
```

工资数据

```
class(w$Age)
```

```
# [1] "character"
```

- R自动将字符串，转变为了哑变量
- Python需要将字符串转换为

回归系数显著么？

```
summary(a)
```

```
#
# Call:
# lm(formula = Income ~ ., data = w)
#
# Residuals:
#       1       2       3       4       5       6       7       8       9
#  134967   27744  -17956  -67056  -32900   55111   14944    2144 -102067  -8
#      11      12
#    3011   64911
#
# Coefficients:
#             Estimate Std. Error t value Pr(>|t|)
# (Intercept)    23256      71440   0.326   0.7580
# AgeOld         69600      70468   0.988   0.3687
# AgeYoung       19167      78786   0.243   0.8175
# EduEle        -24300      81369  -0.299   0.7772
# EduGrad       292178      84692   3.450   0.0182 *
# EduHs         -23267      81369  -0.286   0.7864
# SexM          -82533      70468  -1.171   0.2943
# ---
# Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#
# Residual standard error: 99660 on 5 degrees of freedom
# Multiple R-squared:  0.8089,    Adjusted R-squared:  0.5795
# F-statistic: 3.526 on 6 and 5 DF,  p-value: 0.09399
```

回归系数显著么？

必须用方差分析看变量是否显著

```
anova(a)
```

```
# Analysis of Variance Table
#
# Response: Income
#           Df     Sum Sq    Mean Sq F value  Pr(>F)
# Age        2 1.8322e+10 9.1610e+09  0.9224 0.45605
# Edu        3 1.7819e+11 5.9397e+10  5.9807 0.04149 *
# Sex        1 1.3624e+10 1.3624e+10  1.3718 0.29428
# Residuals  5 4.9657e+10 9.9315e+09
# ---
# Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

工资数据

```
(w1=read.csv("E:/teaching_plan_notes/msa11091083/吴喜之课件/《多元统计分
```

```
#    Age Edu Income Sex
# 1    1   1 520000   0
# 2    2   2  51000   0
# 3    3   3   1200   0
# 4    1   4   1500   0
# 5    2   1 200000   1
# 6    3   2  15000   1
# 7    1   3   2000   1
# 8    2   4   1100   0
# 9    3   1 150000   1
# 10   1   2  10000   0
# 11   2   3   3000   0
# 12   3   4    500   1
```

```
class(w1$Age)
```

```
# [1] "integer"
```

```
#此时，不能直接带入回归计算，因为0,1,...,4不代表数值，仅代表属性类别
```

工资数据

```
for(i in c(1,2,4))w1[,i]<-factor(w1[,i])#转换成因子型变量
a1=lm(Income~., data=w1);summary(a1)
```

```
#
# Call:
# lm(formula = Income ~ ., data = w1)
#
# Residuals:
#       1       2       3       4       5       6       7       8       9
#  134967   27744  -17956  -67056  -32900   55111   14944    2144 -102067  -8
#      11      12
#    3011   64911
#
# Coefficients:
#             Estimate Std. Error t value Pr(>|t|)
# (Intercept)   385033      78786   4.887  0.00452 **
# Age2          -69600      70468  -0.988  0.36866
# Age3          -50433      78786  -0.640  0.55026
# Edu2         -292178      84692  -3.450  0.01824 *
# Edu3         -315444      84692  -3.725  0.01365 *
# Edu4         -316478      84692  -3.737  0.01348 *
# Sex1          -82533      70468  -1.171  0.29428
# ---
# Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#
# Residual standard error: 99660 on 5 degrees of freedom
# Multiple R-squared:  0.8089,    Adjusted R-squared:  0.5795
```

工资数据

- 回归系数不同

```
coef(a)
```

```
# (Intercept)      AgeOld    AgeYoung      EduEle     EduGrad       EduHs
#    23255.56    69600.00    19166.67   -24300.00   292177.78   -23266.67
#        SexM
#   -82533.33
```

```
# 以Age = Middle, Edu = Coll, Sex = F作为参考水平，其系数为0
coef(a1)
```

```
# (Intercept)        Age2        Age3        Edu2        Edu3        Edu4
#   385033.33   -69600.00   -50433.33  -292177.78  -315444.44  -316477.78
#        Sex1
#   -82533.33
```

```
# 以Age = 1(Young), Edu = 1(Grad), Sex = 1(F)作为参考水平，其系数为0
```

工资数据

- 尽管回归系数不同，预测值是相同的

```
w[1,-3]
```

```
#   Age  Edu Sex
# 1 Old Grad   F
```

```
predict(a, newdata=w[1,-3])
```

```
#        1
# 385033.3
```

```
w1[1,-3]
```

```
#   Age Edu Sex
# 1   1   1   0
```

```
predict(a1, newdata=w1[1,-3])
```

```
#        1
# 385033.3
```

蟹蟹

本幻灯片由 R 包 **xaringan** 生成；

查克拉来自于 remark.js、**knitr**、以及 R Markdown。