Quant_II_hwk_04

吴温泉

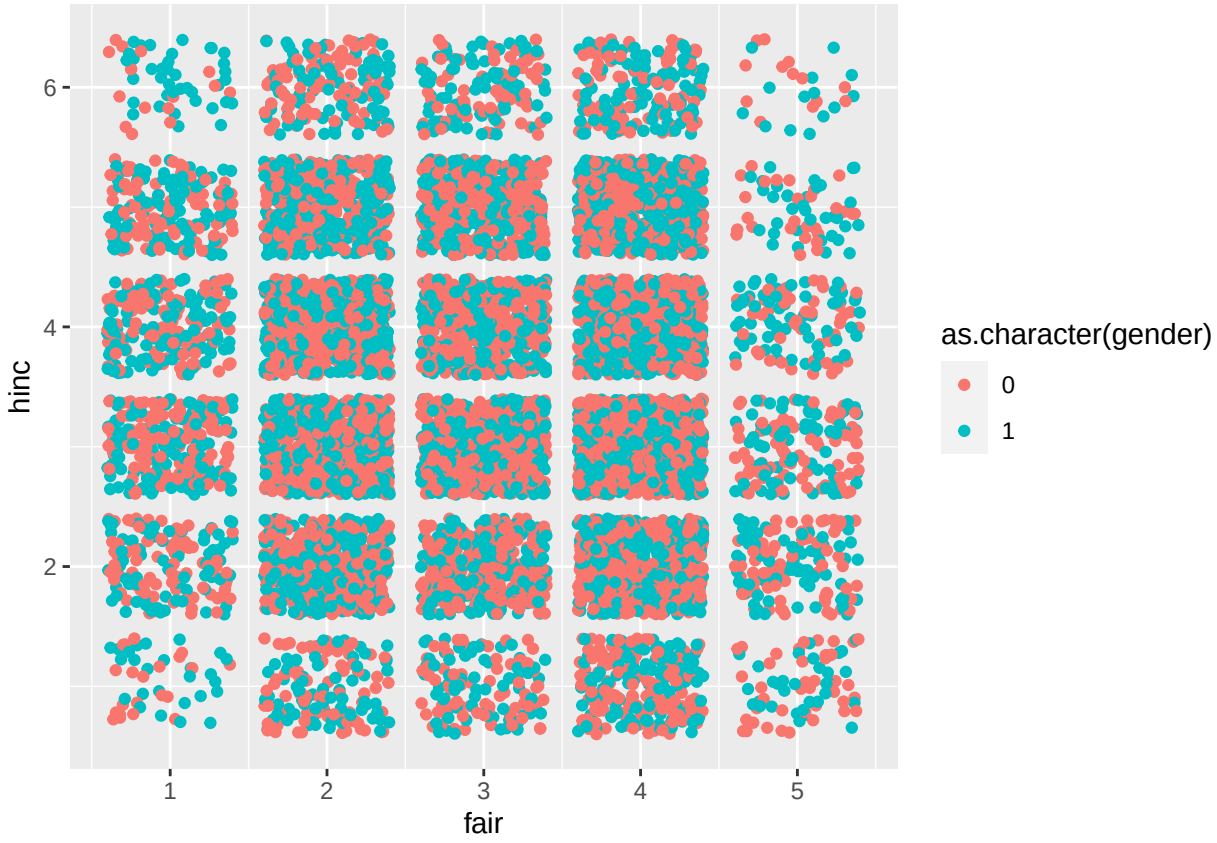
目录

1 定序、定类因变量分析：

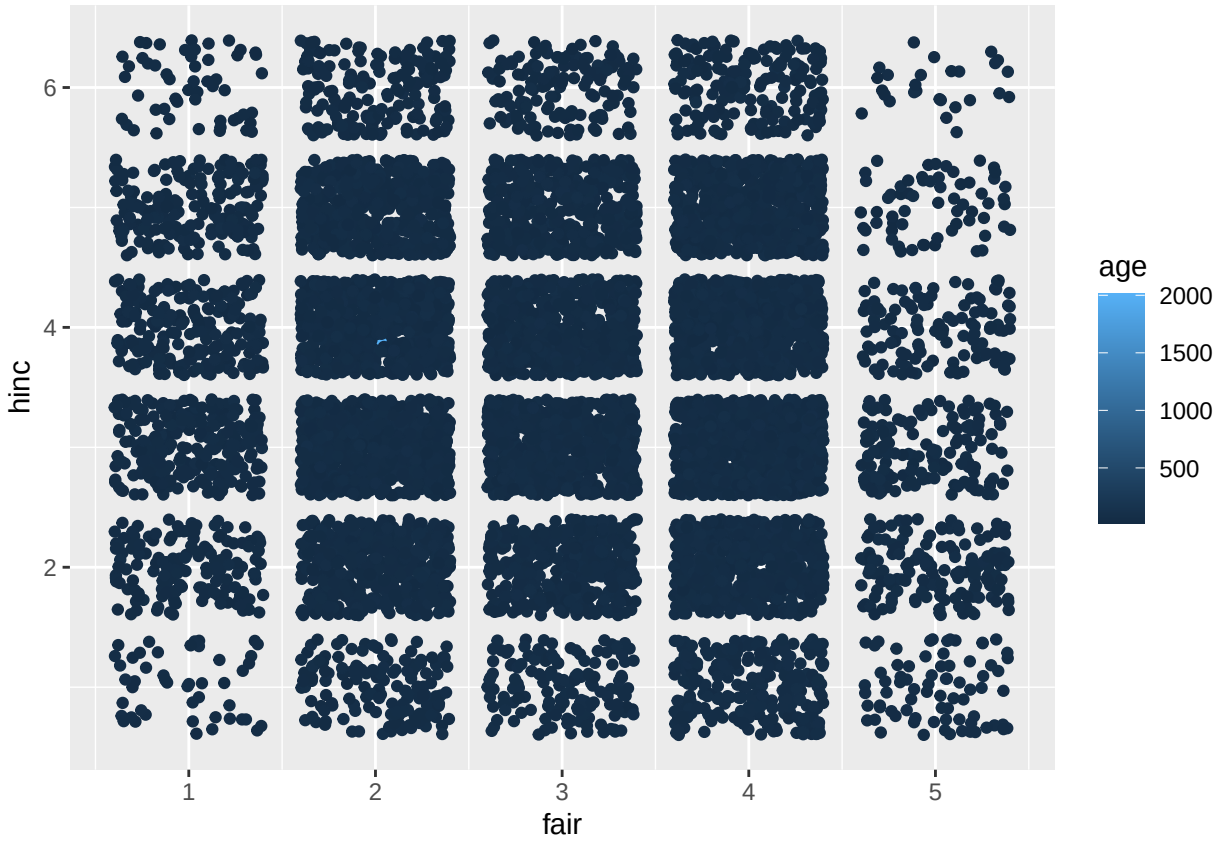
1.1 分析 CGSS2010 数据中 A35 问题："总的来说，您认为当今的社会是不是公平的？"探讨性别、年龄、收入对公平感知的关系。

```
data <- read_dta('./cgss2010_14.dta')


fair_data <- data %>%
  dplyr::select(c(a2,  # gender , 1 = male , 2 = female
           a3a,  # age
           # 'a8a', # income
           a62, # income
           a35  # fair
           )) %>%
  mutate(age = 2010-a3a) %>%
  filter(age > 17) %>%
  # filter(a62 < 9999996) %>%
  # filter(a35 > 0) %>%
  mutate(gender = a2, income = a62, fair = a35) %>%
  dplyr::select(c('gender','age','income','fair'))



age <- fair_data$age
gender <- ifelse(fair_data$gender == 1, 1, 0)  # 0 = female , 1 = male
# temp <- as.numeric(fair_data$fair)
fair <- ifelse(fair_data$fair < 0, NA, fair_data$fair)
income <- ifelse(fair_data$income > 9999996, NA, fair_data$income)
probs <- c(0, 0.07, 0.25, 0.5, 0.75, 0.93, 1)
kpts <- quantile(income, prob=probs, na.rm=TRUE)
hinc <- as.numeric(cut(income, breaks=kpts, labels = 1:6, right=TRUE))
loghinc <- log(income+1)

fair_data_na <- cbind.data.frame(fair, gender, age, hinc, loghinc)
fair_data <- na.exclude(cbind.data.frame(fair, gender, age, hinc, loghinc))


ggplot(fair_data) +
  geom_point(aes(x= fair, y = hinc, color = as.character(gender)),
```

```
            position = position_jitter())
```




```
ggplot(fair_data) +
  geom_point(aes(x= fair, y = hinc, color = age),
             position = position_jitter())
```



1.2 使用 **ordered logistic regression (polr())** 分析性别、年龄、收入（自变量）对公平感知（因变量）的关系，初步使用系数的正负关系解读因变量和自变量的关系，并说明 **Peudo-R2** 和似然值检验的结果。

```
library(MASS)
```

```
## Warning: package 'MASS' was built under R version 4.0.3
```

```
##
## Attaching package: 'MASS'
```

```
## The following object is masked from 'package:dplyr':
##
##     select
```

```r
M0 <- polr(ordered(fair)~1, method = "logistic", data = fair_data)
summary(M0)
```

```
##
## Re-fitting to get Hessian

## Call:
## polr(formula = ordered(fair) ~ 1, data = fair_data, method = "logistic")
##
## No coefficients
##
## Intercepts:
##     Value    Std. Error t value
## 1|2  -2.3172   0.0346   -67.0292
## 2|3  -0.5367   0.0205   -26.2168
## 3|4   0.4231   0.0202    20.9464
## 4|5   2.8286   0.0430    65.7364
##
## Residual Deviance: 29530.41
## AIC: 29538.41
```

```r
M1 <- polr(ordered(fair) ~ gender + age + hinc, data = fair_data,  method = "logistic", Hess=TRUE)
summary(M1)
```

```
## Call:
## polr(formula = ordered(fair) ~ gender + age + hinc, data = fair_data,
##     Hess = TRUE, method = "logistic")
##
## Coefficients:
##           Value Std. Error t value
## gender -0.01665    0.03571 -0.4661
## age     0.01135    0.00120  9.4599
## hinc   -0.05699    0.01386 -4.1103
##
## Intercepts:
##     Value    Std. Error t value
## 1|2  -2.0033   0.0889   -22.5323
## 2|3  -0.2135   0.0848    -2.5165
## 3|4   0.7554   0.0853     8.8588
```

```
## 4|5   3.1790   0.0942    33.7409
##
## Residual Deviance: 29401.14
## AIC: 29415.14
```

```
nullDev <- deviance(M0)
dev <- deviance(M1)
pR2 <- (nullDev-dev)/nullDev
pR2
```

```
## [1] 0.004377464
```

```
LR <- nullDev - dev
k <- length(coef(M1))
prob <- pchisq(LR, df=k-1, lower.tail = FALSE)
```

年龄与公正感正相关，年龄与公正感成负相关，女性比男性更加认为社会是公正的。Peudo-R2：有自变量的模型较没有自变数的模型可以解释的偏差比为 4% 似然值检验显著，有自变量的模型较没有自变数的模型可以解释 y 更多的偏差，拟合优度 (goodness of fit) 显著性改善。

1.3 使用胜算比 (odds ratio) 解释因变量和自变量的关系。

```
exp(coef(M1))
```

```
##    gender       age      hinc
## 0.9834923 1.0114176 0.9446062
```

比较除了性别以外相同的两个人，男性比女性少（$1 - 0.98 = 0.02$）2% 的概率认为社会是公正的。比较除了年龄以外相同的两个人，年龄大 1 岁的群体，他认为社会公正的概率多 1%。比较除了家庭收入以外相同的两个人，收入每增加 1%，有钱人比穷人少 6% 的概率认为社会是公正的。

1.4 比较 40 岁、收入为均值的男性和女性之间在各个公平感知类别的预测概率的差异 (提示：使用 predict())。

```
mean_inc <- mean(fair_data$loghinc)

newDat <- cbind.data.frame(gender = c(1, 0), age= rep(40,2), loghinc =  rep(mean_inc,2) )
```

```
M2 <- polr(ordered(fair) ~ gender + age + loghinc, data=fair_data_na,  method = "logistic")
phat <- predict(M2, newdata = newDat, type="probs")
library(reshape2)
```

```
## Warning: package 'reshape2' was built under R version 4.0.3
```

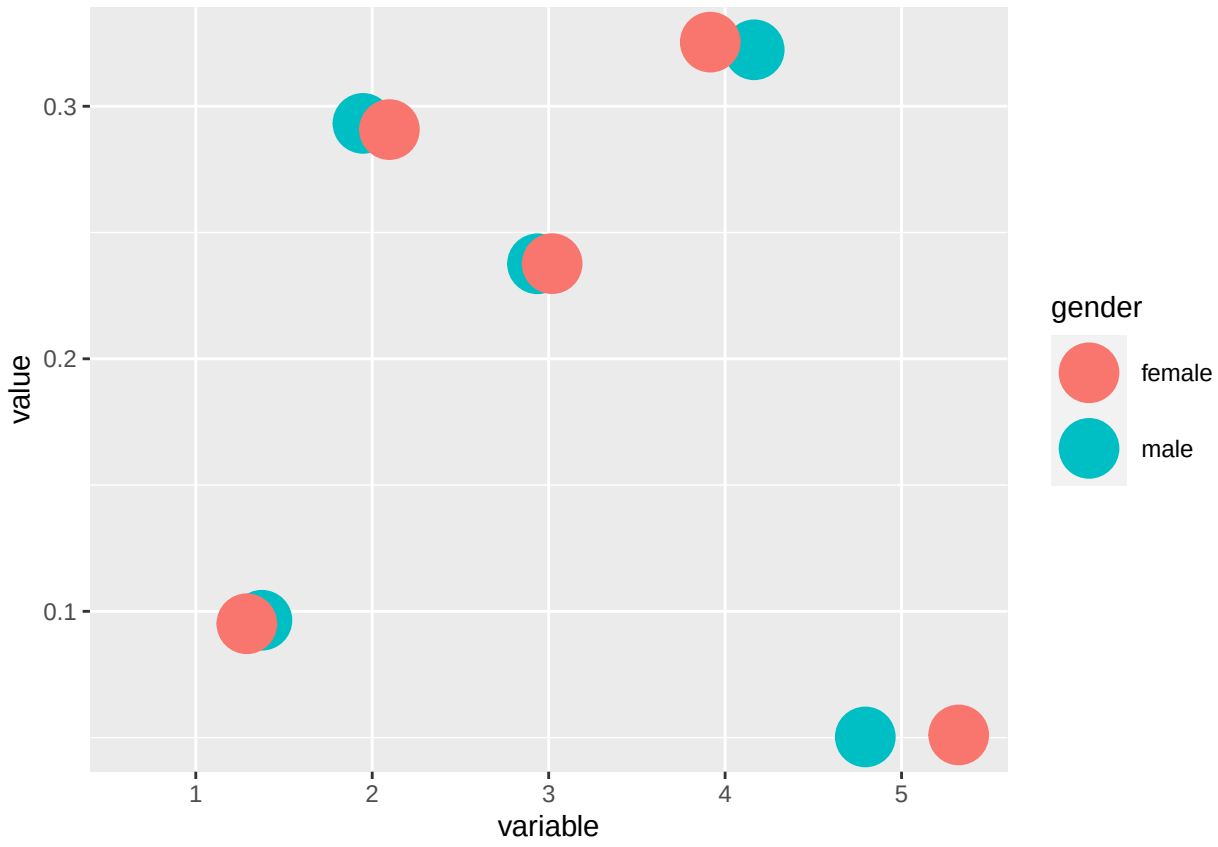
```
##
## Attaching package: 'reshape2'
```

```
## The following object is masked from 'package:tidyr':
##
##     smiths
```

```
df <- melt(as.data.frame(phat))
```

```
## No id variables; using all as measure variables
```

```
df$gender <- c('male', 'female')
ggplot(df) +
  geom_point(aes(x = variable, y = value , color = gender),
             size = 10,
             # shape = 1,
             position = position_jitter())
```

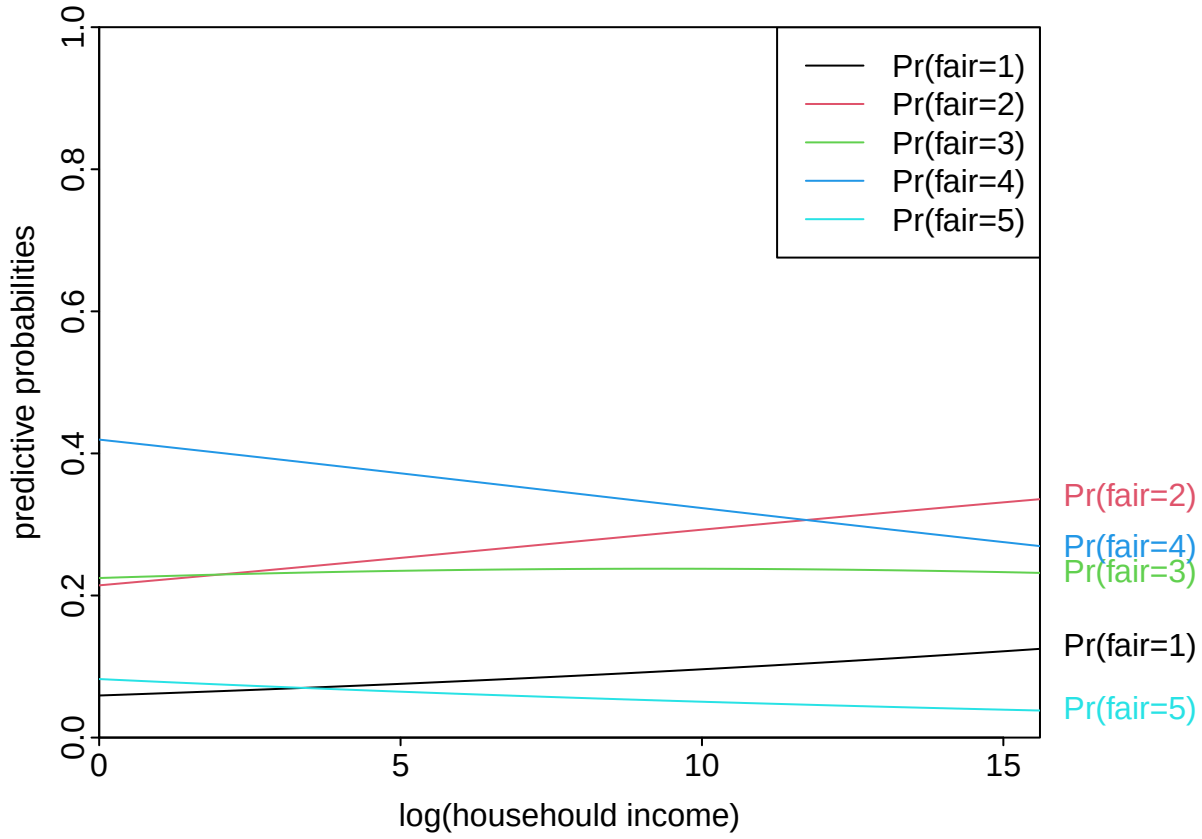


觉得社会比较不公平的男性比女性多，觉得社会比较公平的女性比男性多。

1.5 绘出 40 岁的男性收入和各个公平感知类别的预测概率的曲线图。从图中你观察到什么？

```
M2 <- polr(ordered(fair) ~ gender + age + loghinc, data=fair_data_na,  method = "logistic")

newDat <- cbind.data.frame(gender = rep(1, 1000), age=rep(40, 1000), loghinc = seq(0,16, length=10

phat <- predict(M2, newdata = newDat, type="probs")

par(mar=c(3,3,1,5), mgp=c(1.5,0.2,0), tcl=-0.2)
plot(0,0, ylim=c(0,1), xlim=range(loghinc, na.rm=TRUE), type="n",
    ylab="predictive probabilities", xlab="log(household income)", xaxs="i", yaxs="i")
for(i in 1:5){
    lines(x=newDat$loghinc, y=phat[,i], col=i)
}
for(i in 1:5){
    text(x=16, y=phat[1000, i], labels=paste("Pr(fair=", i, ")", sep=""), xpd=NA, adj=0, col=i)
```

```
}
legend("topright", col=1:5, lty=1, legend=paste("Pr(fair=", 1:5, ")", sep=""))
```



随着收入增加，比较公平（4）的概率下降，比较不公平（2）和完全不公平（1）的概率增加。

1.6 进行平行性检验 (提示：使用 brant() 命令)，上述模型是否通过检验。

```
library(brant)
```

```
## Warning: package 'brant' was built under R version 4.0.5
```

```
brant(M1)
```

```
## ------------------------------------------
## Test for X2  df  probability
## ------------------------------------------
## Omnibus      66.58   9   0
## gender       14.24   3   0
```

```
## age       12.31   3   0.01
## hinc      38.88   3   0
## ----------------------------------------
##
## H0: Parallel Regression Assumption holds
```

```
brant(M2)
```

```
## ----------------------------------------
## Test for X2  df  probability
## ----------------------------------------
## Omnibus      57.23   9   0
## gender       14.51   3   0
## age      13.88   3   0
## loghinc      27.52   3   0
## ----------------------------------------
##
## H0: Parallel Regression Assumption holds
```

差异显著，平行性检验不通过。

1.7 使用 multinomial logistic regression (multinom()) 分析性别、年龄、收入（自变量）对公平感知（因变量）的关系，初步使用系数的正负关系解读因变量和自变量的关系，并说明 Peudo-R2 和似然值检验的结果。

```
library(nnet)
M3 <- multinom(factor(fair) ~ 1, data = fair_data_na)
```

```
## # weights:  10 (4 variable)
## initial  value 18904.457719
## iter  10 value 16896.174562
## iter  10 value 16896.174472
## final  value 16896.174472
## converged
```

```
summary(M3)
```

```
## Call:
```

```
## multinom(formula = factor(fair) ~ 1, data = fair_data_na)
##
## Coefficients:
##   (Intercept)
## 2   1.1470631
## 3   1.0234597
## 4   1.3294266
## 5  -0.4989052
##
## Std. Errors:
##   (Intercept)
## 2  0.03563056
## 3  0.03619110
## 4  0.03490742
## 5  0.05050174
##
## Residual Deviance: 33792.35
## AIC: 33800.35
```

```
M4 <- multinom(factor(fair) ~ age + gender + loghinc, data=fair_data_na, Hess=TRUE)
```

```
## # weights:  25 (16 variable)
## initial  value 16594.914315
## iter  10 value 15856.942365
## iter  20 value 14784.785943
## final  value 14779.665799
## converged
```

```
summary(M4)
```

```
## Call:
## multinom(formula = factor(fair) ~ age + gender + loghinc, data = fair_data_na,
##     Hess = TRUE)
##
## Coefficients:
##   (Intercept)         age      gender      loghinc
## 2   0.6203024 0.007274175 -0.13888895  0.024826966
## 3   0.9485833 0.005976945 -0.26605762 -0.012738517
## 4   0.6806105 0.017077321 -0.13020815 -0.008784007
```

```
## 5   0.1536805 0.020879990 -0.06994934 -0.160985921
##
## Std. Errors:
##   (Intercept)         age      gender     loghinc
## 2   0.3638546 0.002582258 0.07577484 0.03152791
## 3   0.3675673 0.002640832 0.07753051 0.03180859
## 4   0.3517707 0.002521259 0.07425986 0.03041761
## 5   0.4266110 0.003465488 0.10661600 0.03587063
##
## Residual Deviance: 29559.33
## AIC: 29591.33
```

```
nullDev <- deviance(M3)
dev <- deviance(M4)

pR2 <- (nullDev- dev) / nullDev
pR2 <- 1 - logLik(M4)/logLik(update(M4, .~1))
```

```
## # weights:  10 (4 variable)
## initial  value 18904.457719
## iter  10 value 16896.174562
## iter  10 value 16896.174472
## final  value 16896.174472
## converged
```

```
LR <- nullDev - dev
k <- length(coef(M4))
prob <- pchisq(LR, df=k-2, lower.tail = FALSE)
```

1.8 使用相对曝险比 (relative risk ratio) 解释因变量和自变量的关系。

```
exp(coef(M4))
```

```
##   (Intercept)      age    gender   loghinc
## 2    1.859490 1.007301 0.8703247 1.0251377
## 3    2.582049 1.005995 0.7663950 0.9873423
## 4    1.975083 1.017224 0.8779127 0.9912545
## 5    1.166118 1.021100 0.9324411 0.8513041
```

1.8.1 相对于完全不公平人群，在比较不公平中人群：

比较除了性别以外相同的两个人，女性比男性少多 13% 的概率认为社会是公正的。比较除了年龄以外相同的两个人，年龄大 1 岁的群体，他认为社会公正的概率多 7%。比较除了家庭收入以外相同的两个人，收入每增加 1%，他认为社会公正的概率多 2.5%。

1.8.2 相对于完全不公平人群，在居中人群中：

比较除了性别以外相同的两个人，女性比男性少多 24% 的概率认为社会是公正的。比较除了年龄以外相同的两个人，年龄大 1 岁的群体，他认为社会公正的概率多 6%。比较除了家庭收入以外相同的两个人，收入每增加 1%，他认为社会公正的概率少 2%。

1.8.3 相对于完全不公平人群，在比较公平人群中：

比较除了性别以外相同的两个人，女性比男性少多 13% 的概率认为社会是公正的。比较除了年龄以外相同的两个人，年龄大 1 岁的群体，他认为社会公正的概率多 17%。比较除了家庭收入以外相同的两个人，收入每增加 1%，他认为社会公正的概率少 1%。

1.8.4 相对于完全不公平人群，在完全公平人群中：

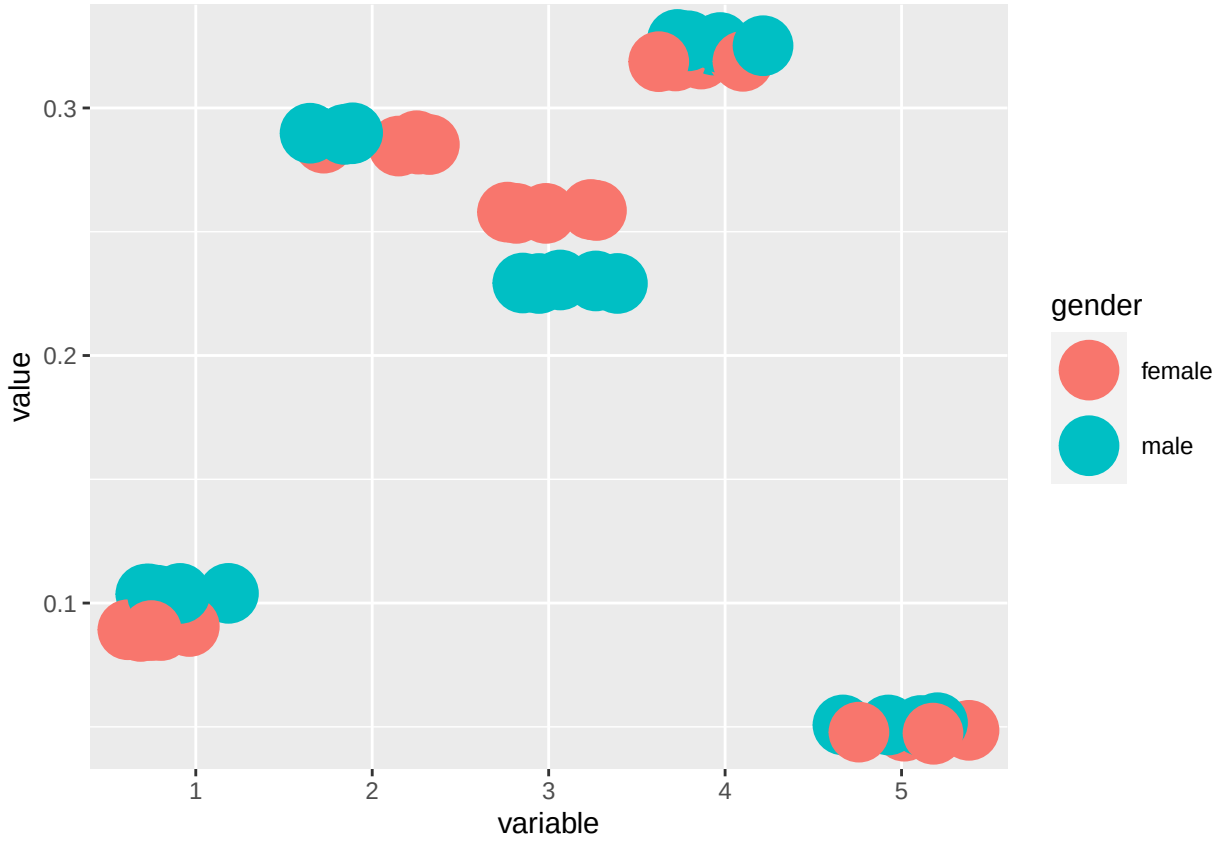
比较除了性别以外相同的两个人，女性比男性少多 13% 的概率认为社会是公正的。比较除了年龄以外相同的两个人，年龄大 1 岁的群体，他认为社会公正的概率多 21%。比较除了家庭收入以外相同的两个人，收入每增加 1%，他认为社会公正的概率少 15%。

1.9 比较 40 岁、收入为均值的男性和女性之间在各个公平感知类别的预测概率的差异。

```
mean_inc <- mean(fair_data$loghinc)

newDat <- cbind.data.frame(gender = rep(c(1, 0), 5), age= rep(40,10), loghinc =  rep(mean_inc,10 ))

phat <- predict(M4, newdata = newDat, type="probs")
library(reshape2)
df <- melt(as.data.frame(phat))
```

```
## No id variables; using all as measure variables
```

```
df$gender <- c('male', 'female')
ggplot(df) +
  geom_point(aes(x = variable, y = value , color = gender),
             size = 10,
             # shape = 1,
             position = position_jitter())
```

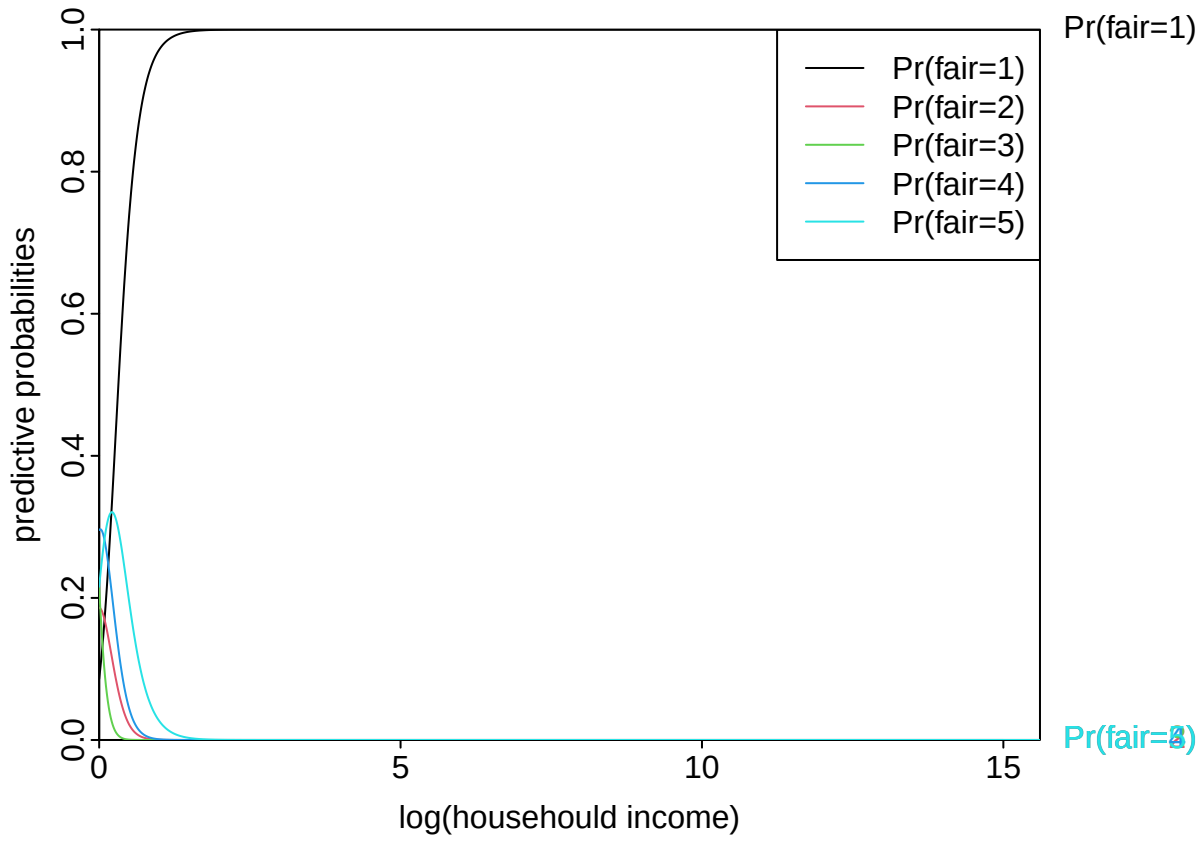


觉得社会比较不公平的男性比女性多，同时，觉得社会比较公平、居中的男性也比女性多。

1.10 绘出 40 岁的男性收入和各个公平感知类别的预测概率的曲线图。从图中你观察到什么？

```
newDat <- cbind.data.frame(gender = seq(1, 1000), age=rep(40, 1000), loghinc = seq(0,16, length=10

phat <- predict(M4, newdata = newDat, type="probs")

par(mar=c(3,3,1,5), mgp=c(1.5,0.2,0), tcl=-0.2)
plot(0,0, ylim=c(0,1), xlim=range(loghinc, na.rm=TRUE), type="n",
```

```
    ylab="predictive probabilities", xlab="log(household income)", xaxs="i", yaxs="i")
for(i in 1:5){
    lines(x=newDat$loghinc, y=phat[,i], col=i)
}
for(i in 1:5){
    text(x=16, y=phat[1000, i], labels=paste("Pr(fair=", i, ")", sep=""), xpd=NA, adj=0, col=i)
}
legend("topright", col=1:5, lty=1, legend=paste("Pr(fair=", 1:5, ")", sep=""))
```



看不出来。

1.11 进行公平感知各类别相互独立性检验 (Hausman Test 和 Small and Hsiao test)。

```
# Hausman McFaden IIA test

library(mlogit)
fair_data_hausman <- mlogit.data(fair_data, choice = "fair", shape = "wide", sep="_", alt.levels =
```

```r
M4 <- mlogit(fair ~ 1 | age + gender + loghinc, data=fair_data_hausman)
M5 <- mlogit(fair ~ 1 | age + gender + loghinc, data=fair_data_hausman, alt.subset = c(1,2,3,4))

mlogit:::hmftest(M3, M4)


# Small and Hsiao Test

smhsiaoTest <- function(formula, data, seed = 1234,
    choiceVar = "fair", keepLev = c(1,2,3,4)){
    set.seed(seed)
    library(mlogit)
    n <- dim(data)[1]
    idx <- sample(1:n, n, replace=FALSE)
    half <- floor(n/2)
    idx1 <- idx[1:half]
    idx2 <- idx[(half+1):n]
    datA <- data[idx1,]
    datB <- data[idx2,]
    levs <- levels(as.factor(data[,choiceVar]))
    datA <- mlogit.data(datA, choice = choiceVar, shape = "wide", sep="_")#, alt.levels = levs)
    datB <- mlogit.data(datB, choice = choiceVar, shape = "wide", sep="_")#, alt.levels = levs)
    fitA <- mlogit(formula, data=datA)
    fitB <- mlogit(formula, data=datB)
    betaF <- coef(fitA) * 1/sqrt(2) + coef(fitB)*(1-1/sqrt(2))
    fitC <- mlogit(formula, data=datB, alt.subset = keepLev)
    betaR <- coef(fitC)
    LL1 <- logLik(fitC)
    uu <- names(betaR)
    betaFR <- betaF[uu]
    K <- length(betaR)
    fitD <- mlogit(formula, data=datB, alt.subset = keepLev, start = betaFR, iterlim=0)
    LL0 <- logLik(fitD)
    SH <- -2*(LL0 - LL1)
    prob <- pchisq(SH, df=K, lower.tail=FALSE)
    out <- list("chisq" = SH[1], "df" = K, "pValue" = prob[1], "LLU" = LL0, "LLR" = LL1)
    if(prob > 0.05){
        cat("\nSmall and Hsiao test\nH0: IIA assumption is accepted\n")
```

```
    } else {
        cat("\nSmall and Hsiao test\nH0: IIA assumption is rejected\n")
    }
    return(out)
}

smhsiaoTest(shopping ~ 1 | age + gender + loghinc   , data=fair_data, seed = 11, keepLev = c(1,2))


smhsiaoTest(shopping ~ 1 | age + gender + loghinc + ccpmember, data=fair_data, seed = 111, keepLev


smhsiaoTest(shopping ~ 1 | age + gender + loghinc + ccpmember, data=fair_data, seed = 1111, keepLe

smhsiaoTest(shopping ~ 1 | age + gender + loghinc + ccpmember, data=fair_data, seed = 1111, keepLe
```

1.12 如果以上检验各类别相互并不独立，你会给出什么样的建模建议？