lab 3

2025-11-24

1 multi-variable linear regression

1.0

```
hw=read.table("http://staff.ustc.edu.cn/~ynyang/2025/lab/height-weight.txt",head=T)
attach(hw)
plot(log(height),log(weight),col=sex+1)#since sex=0/1,col=sex+1 means coloring different sex

fit.all=lm(log(weight)~log(height),data=hw)
print(fit.all)
```

```
##
## Call:
## lm(formula = log(weight) ~ log(height), data = hw)
##
## Coefficients:
## (Intercept)  log(height)
##       2.599        2.930
```

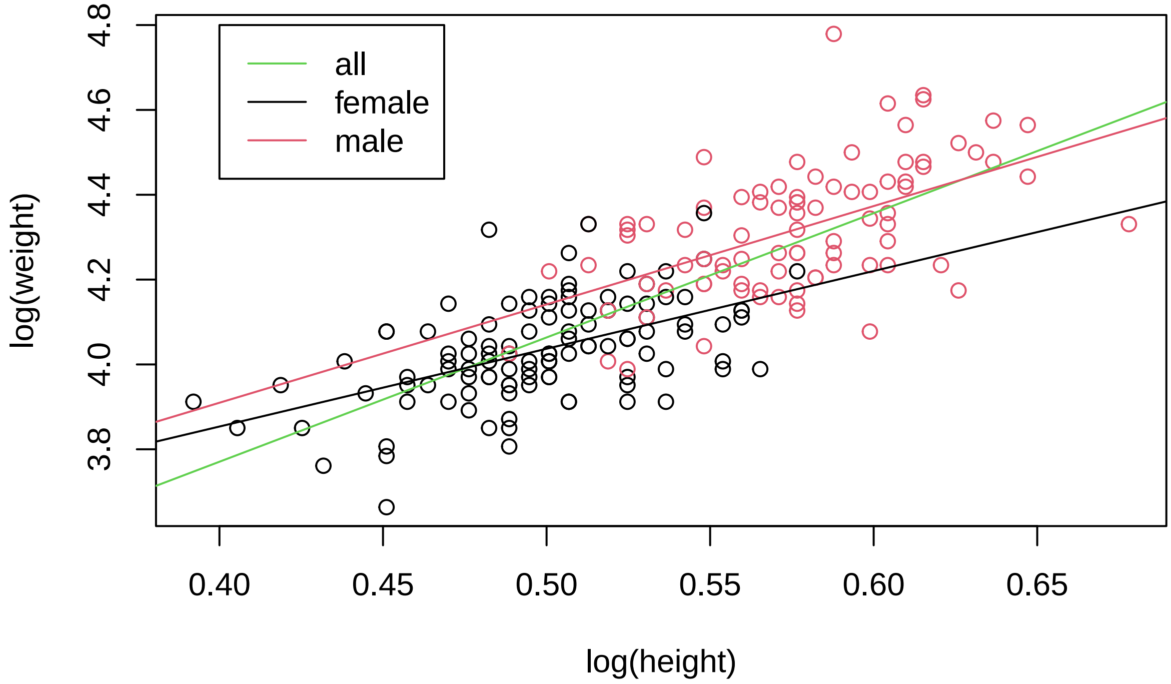
```
fit.female=lm(log(weight)~log(height),data=hw,subset=sex==0)
print(fit.female)
```

```
##
## Call:
## lm(formula = log(weight) ~ log(height), data = hw, subset = sex ==
##     0)
##
## Coefficients:
## (Intercept)  log(height)
##       3.120        1.833
```

```
fit.male=lm(log(weight)~log(height),data=hw,subset=sex==1)
print(fit.male)
```

```
##
## Call:
## lm(formula = log(weight) ~ log(height), data = hw, subset = sex ==
##     1)
##
## Coefficients:
## (Intercept)  log(height)
##       2.982        2.318
```

```
abline(fit.all,col=3)
abline(fit.female,col=1)
abline(fit.male,col=2)
legend(0.4,4.8,c("all","female","male"),col=c(3,1,2),lty=c(1,1,1))
```



```
summary(fit.female)
```

```
##
## Call:
## lm(formula = log(weight) ~ log(height), data = hw, subset = sex ==
##     0)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.28367 -0.05921 -0.01299  0.06503  0.31279
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   3.1203     0.1413   22.08  < 2e-16 ***
## log(height)   1.8333     0.2829    6.48 2.74e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.103 on 109 degrees of freedom
## Multiple R-squared:  0.2781, Adjusted R-squared:  0.2715
## F-statistic: 41.99 on 1 and 109 DF,  p-value: 2.743e-09
```

1.1

```
myfit=lm(log(weight)~log(height)+sex,data=hw)
names(myfit)
```

```
##  [1] "coefficients"  "residuals"     "effects"       "rank"
##  [5] "fitted.values" "assign"        "qr"            "df.residual"
##  [9] "xlevels"       "call"          "terms"         "model"
```

```
# myfit$coeff
# coef(myfit)
# myfit$residuals
# resid(myfit)
# myfit$fitted.balues
# fitted(myfit)
```

1.2

```
fit1=lm(log(height)~sex,data=hw)
height.perp=resid(fit1)
fit2=lm(log(weight)~height.perp,data=hw)
coef(fit2)
```

```
## (Intercept) height.perp
##    4.159456    2.057156
```

```
coef(myfit)
```

```
## (Intercept) log(height)         sex
##   3.0087053   2.0571556   0.1240784
```

1.3

```
summary(myfit)
```

```
##
## Call:
## lm(formula = log(weight) ~ log(height) + sex, data = hw)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.28715 -0.06964 -0.01143  0.07636  0.43717
##
```

```
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.00871    0.11560  26.028  < 2e-16 ***
## log(height)  2.05716    0.23092   8.909 3.55e-16 ***
## sex          0.12408    0.02427   5.113 7.51e-07 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 0.1145 on 196 degrees of freedom
## Multiple R-squared:  0.6601, Adjusted R-squared:  0.6567
## F-statistic: 190.4 on 2 and 196 DF,  p-value: < 2.2e-16
```

Check that $\hat{\sigma}^2 = RSS/(n-p)$

```
e=resid(myfit)
cat("RSS/(n-p):",sum(e^2)/(199-3),"\n")
```

```
## RSS/(n-p): 0.01312098
```

```
cat("sample variance:",(summary(myfit)$sigma)^2)
```

```
## sample variance: 0.01312098
```

Check that $R^2 = r^2_{\hat{y},y}$

```
y.hat=fitted(myfit)
y=log(hw[,"weight"])
cat("correlation:",cor(y.hat,y)^2,"\n")
```

```
## correlation: 0.6601487
```

```
cat("R squared:",summary(myfit)$r.squared)
```

```
## R squared: 0.6601487
```

Check that $F = \frac{n-1}{k} \times \frac{R^2}{1-R^2}$

```
R2=summary(myfit)$r.squared
n=nrow(hw)
p=ncol(hw)
k=2
F=(n-p)/k*R2/(1-R2)
F
```

```
## [1] 190.3614
```

```
summary(myfit)$fstatistic
```

```
##    value    numdf    dendf
## 190.3614   2.0000 196.0000
```

1.4

Recall the model:$log(weight) = a + b \times log(height) + c \times sex + \epsilon$

```
model.null=lm(log(weight)~1,data=hw)#b=c=0
model.full=lm(log(weight)~log(height)+sex,data=hw)
anova(model.null,model.full)
```

```
## Analysis of Variance Table
##
## Model 1: log(weight) ~ 1
## Model 2: log(weight) ~ log(height) + sex
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1    198 7.5672
## 2    196 2.5717  2    4.9955 190.36 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The result shows that the full model is significantly better than null model.

If we consider $H_0 : b = c$, the null model becomes: $log(weight) = a + b \times (log(height) + sex) + \epsilon$.

```
z=log(hw[,"height"]+hw[,"sex"])
model.null=lm(log(weight)~z,data=hw)
anova(model.null,model.full)
```

```
## Analysis of Variance Table
##
## Model 1: log(weight) ~ z
## Model 2: log(weight) ~ log(height) + sex
##   Res.Df    RSS Df Sum of Sq      F   Pr(>F)
## 1    197 3.2126
## 2    196 2.5717  1   0.64092 48.847 4.23e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The result shows that the full model is still significantly better than null model.

1.5

If we assume that there is an interaction between height and sex, consider the model:$log(height) = \beta_0 + \beta_1 log(height) + \beta_2 sex + \gamma log(height) \times sex + \epsilon$.

```
fit.intersection=lm(log(weight)~log(height)*sex,data=hw)
#fit.intersection=lm(log(height)~log(height)+sex+log(height):sex,data=hw)
summary(fit.intersection)
```

```
##
## Call:
## lm(formula = log(weight) ~ log(height) * sex, data = hw)
##
```

```
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.29310 -0.06623 -0.00553  0.07458  0.43410
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)       3.1203     0.1572  19.848  < 2e-16 ***
## log(height)       1.8333     0.3147   5.826 2.32e-08 ***
## sex              -0.1378     0.2513  -0.548    0.584
## log(height):sex   0.4848     0.4631   1.047    0.296
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 0.1145 on 195 degrees of freedom
## Multiple R-squared:  0.662,  Adjusted R-squared:  0.6568
## F-statistic: 127.3 on 3 and 195 DF,  p-value: < 2.2e-16
```

The result shows that p-value of $H_0 : \gamma = 0$ is 0.296. Therefore, we can accept that $\gamma = 0$.

2

2.1

```r
x=c(1,4,1,2,3,3,2,2)
is.numeric(x)
```

```
## [1] TRUE
```

```r
x1=as.factor(x)
is.numeric(x1)
```

```
## [1] FALSE
```

```r
is.factor(x1)
```

```
## [1] TRUE
```

```r
bpdata <- data.frame(
  BP     = c(112, 122, 133, 131, 127, 122),
  Race   = as.factor(c("White", "White", "Black", "Yellow", "Black", "White")),
  Weight = c(71, 82, 77, 68, 62, 79)
)
lm(BP~Weight+Race,data=bpdata)
```

```
##
## Call:
## lm(formula = BP ~ Weight + Race, data = bpdata)
##
## Coefficients:
## (Intercept)       Weight    RaceWhite   RaceYellow
##     87.5024       0.6115     -16.1232       1.9172
```

```
#change baseline
contrasts(bpdata[,"Race"])=contr.treatment(3,base=2)
#bpdata[,"Race"]
lm(BP~Weight+Race,data=bpdata)
```

```
##
## Call:
## lm(formula = BP ~ Weight + Race, data = bpdata)
##
## Coefficients:
## (Intercept)       Weight        Race1        Race3
##     71.3791       0.6115      16.1232      18.0405
```

2.2

When all the independent variables are factors, we can use aov instead of lm.

```
result=aov(BP~Race,data=bpdata)
summary(result)
```

```
##             Df Sum Sq Mean Sq F value Pr(>F)
## Race         2 204.83  102.42   3.629  0.158
## Residuals    3  84.67   28.22
```

Exercise

1

```
install.packages("alr4", repos="http://R-Forge.R-project.org")
```

```
## installing the source package 'alr4'
```

```
salary <- alr4::salary
```

(a)

```
t_result=t.test(salary~sex,data=salary,var.equal=T)
t_result
```

```
##
##  Two Sample t-test
##
## data:  salary by sex
## t = 1.8474, df = 50, p-value = 0.0706
## alternative hypothesis: true difference in means between group Male and group Female is not equal to
## 95 percent confidence interval:
##  -291.257 6970.550
## sample estimates:
##   mean in group Male mean in group Female
##             24696.79             21357.14
```

```
lm(salary~sex,data=salary)
```

```
##
## Call:
## lm(formula = salary ~ sex, data = salary)
##
## Coefficients:
## (Intercept)    sexFemale
##       24697        -3340
```

```
diff_mean <- unname(t_result$estimate[2] - t_result$estimate[1])
diff_mean# the difference of salary between male and female in t_test
```

```
## [1] -3339.647
```

We can see that both results show that the average salary of male is about 3340 higher than female. In the two sample t-test, p value:0.07<0.1.Therefore the result is significant.

I don't think it can lead to the conclusion that there is discrimination against female, since salary might also be affected by one's degree, year of working and year of having acquired degree.

(b)

```
summary(aov(salary~rank,data=salary))
```

```
##             Df    Sum Sq   Mean Sq F value   Pr(>F)
## rank         2 1.347e+09 673391900   75.17 1.17e-15 ***
## Residuals   49 4.389e+08   8958083
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Since the p-value is extremely small, it means rank has significant impact on salary. Therefore, rank and salary are correlated.

```
table <- table(salary$sex, salary$rank)
table
```

```
##
##          Asst Assoc Prof
##   Male     10    12   16
##   Female    8     2    4
```

```
fisher.test(table)
```

```
##
##  Fisher's Exact Test for Count Data
##
## data:  table
## p-value = 0.1564
## alternative hypothesis: two.sided
```

Since sex and rank are both factors, and the data size is not big, use fisher's test. The p-value,0.1564>0.05. Therefore, sex and rank are not correlated.

(c)

Consider $Salary = \beta_0 + \beta_1 Sex + \beta_2 Rank + \epsilon$(here,rank is numeric). The model is based on the assumption that rank doesn't affect the relation between salary and sex(β_1).

Verify the assumption:

```
salary$rank_num=as.numeric(factor(salary$rank,
                                      levels=c("Asst","Assoc","Prof")))#set rank as numeric


lm(salary~sex,data=salary,subset=(rank_num==1))
```

```
##
## Call:
## lm(formula = salary ~ sex, data = salary, subset = (rank_num ==
##     1))
##
## Coefficients:
## (Intercept)    sexFemale
##     17919.6       -339.6
```

```
lm(salary~sex,data=salary,subset=(rank_num==2))
```

```
##
## Call:
## lm(formula = salary ~ sex, data = salary, subset = (rank_num ==
##     2))
##
## Coefficients:
## (Intercept)    sexFemale
##       23444        -1874
```

```
lm(salary~sex,data=salary,subset=(rank_num==3))
```

```
##
## Call:
## lm(formula = salary ~ sex, data = salary, subset = (rank_num ==
##     3))
##
## Coefficients:
## (Intercept)    sexFemale
##       29872        -1067
```

We can see that the coefficients of sex in each subset is greatly different. Therefore, the assumption doesn't hold,i.e., the model is not appropriate.

(d)

Consider the model:$salary = \beta_0 + \beta_1 sex + \beta_2 rank + \beta_3 year + \beta_4 degree + \beta_5 ysdeg + \epsilon$. We hope to test $H_0 : \beta_1 = \beta_2 = 0$. Under H_0, the model becomes:$alary = \beta_0 + \beta_3 year + \beta_4 degree + \beta_5 ysdeg + \epsilon$

```
model.full=lm(salary~sex+rank+year+degree+ysdeg,data=salary)
model.null=lm(salary~year+degree+ysdeg,data=salary)
anova(model.null,model.full)
```

```
## Analysis of Variance Table
##
## Model 1: salary ~ year + degree + ysdeg
## Model 2: salary ~ sex + rank + year + degree + ysdeg
##   Res.Df       RSS Df Sum of Sq      F    Pr(>F)
## 1     48 672102002
## 2     45 258858365  3 413243637 23.946 2.053e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The result shows that p-value is much smaller than 0.05. Therefore, we should reject H_0.

2

Consider the model:$salary = \beta_0 + \beta_1 sex + \gamma_2 rank2 + \gamma_3 rank3 + \epsilon$.

```
rank=salary[,"rank"]
is.factor(rank)
```

```
## [1] TRUE
```

(a)

The model is based on the assumption that the influence of sex β_1 is invariant to ranks. To test the assumption, we can compare the original model with model including interaction.

```
model1=lm(salary~sex+rank,data=salary)#original model
model2=lm(salary~sex*rank,data=salary)#model with interaction
anova(model1,model2)
```

```
## Analysis of Variance Table
##
## Model 1: salary ~ sex + rank
## Model 2: salary ~ sex * rank
##   Res.Df       RSS Df Sum of Sq      F Pr(>F)
## 1     48 431871315
## 2     46 428769653  2   3101661 0.1664 0.8472
```

We can treat original model as null assumption: there is no interaction between sex and rank. The result shows that p-value is 0.8472. Therefore, we should accept the original model(null hypothesis), i.e., the assumption is reasonable.

(b)

In question 1, we treat rank as numeric 1,2, and 3.

To check whether it is reasonble, we only need to test $H_0 : \gamma_3 - \gamma_2 = \gamma_2 - \gamma_1 \iff \gamma_3 = 2\gamma_2$.

Under H_0, the model becomes $salary = \beta_0 + \beta_1 sex + \gamma_2(rank2 + 2 \times rank3) + \epsilon$.

```
encode_rank=function(rank){
  ifelse(rank == "Asst", 0,
       ifelse(rank == "Assoc", 1, 2))#ifelse can deal with vectors
}
rank.null=encode_rank(salary$rank)#rank.null=rank2+2*rank3
salary$rank.null=rank.null
null_model=lm(salary~sex+rank.null,data=salary)
full_model=lm(salary~sex+rank,data=salary)
anova(null_model,full_model)
```

```
## Analysis of Variance Table
##
## Model 1: salary ~ sex + rank.null
## Model 2: salary ~ sex + rank
##   Res.Df       RSS Df Sum of Sq      F Pr(>F)
## 1     49 436625638
## 2     48 431871315  1   4754323 0.5284 0.4708
```

We can see that the p-value is 0.4708. Therefore, we should accept null hypothesis,i.e., $\gamma_3 = 2\gamma_2$.

If model(10) is equivalent to the model in 1(c), then the results in 1(c) and 2(a) seem contradictory. However, I think since 1(c) is point estimate and each group only utilizes part of the data, the result from 2(a) is more reliable. That is, rank doesn't affect the relation between sex and salary.

3

```
expense=read.table("http://staff.ustc.edu.cn/~ynyang/2025/lab/edu.xls",header=T)
#show(expense)
```

(a)

```
options(repos = c(CRAN = "https://cloud.r-project.org"))
install.packages("maps")
```

```
## package 'maps' successfully unpacked and MD5 sums checked
```

```
## Warning: cannot remove prior installation of package 'maps'
```

```
## Warning in file.copy(savedcopy, lib, recursive = TRUE): problem copying
## D:\360Downloads\Software\R-4.5.1\library\00LOCK\maps\libs\x64\maps.dll to
## D:\360Downloads\Software\R-4.5.1\library\maps\libs\x64\maps.dll: Permission
## denied
```

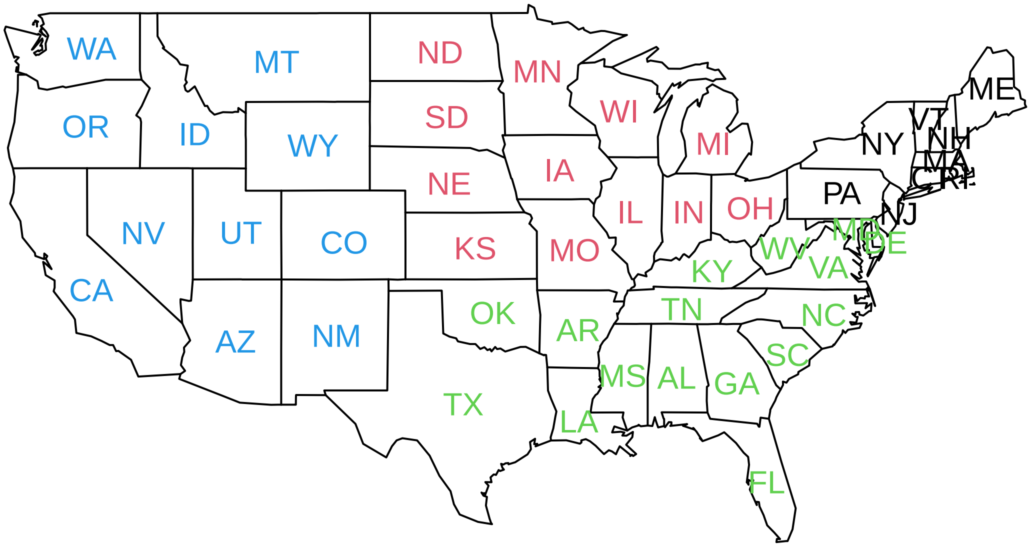
```
## Warning: restored 'maps'
```

```
##
## The downloaded binary packages are in
##  D:\Temp\RtmpszrfbB\downloaded_packages
```

```r
library(maps)
```

```
## Warning: package 'maps' was built under R version 4.5.2
```

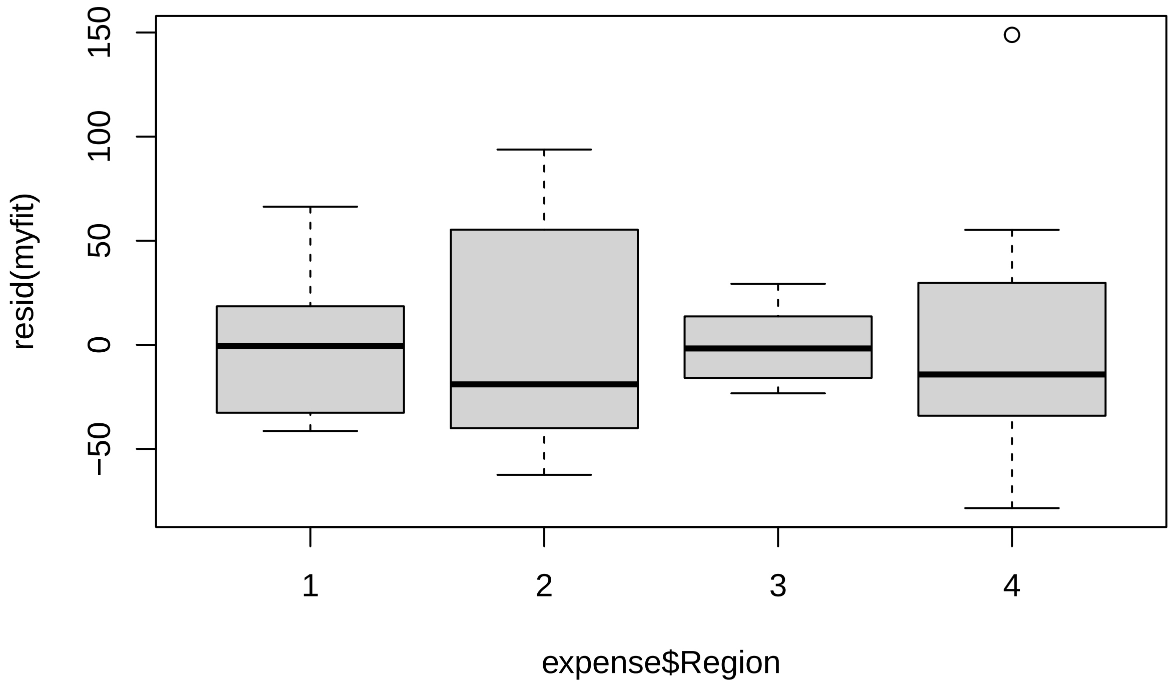
```r
#if the order of state.abb and expense don't match, the coloring will be wrong
expense_ordered <- expense[match(state.abb, expense$state), ]

map("state")
text(state.center, state.abb, col=expense_ordered$Region)
```



(b)

Consider the model$Expenditure = \beta_0 + \beta_1 \times Income + \beta_3 \times Urban + \alpha \times Region + \epsilon$.

```r
Region.factor=as.factor(expense$Region)
#is.factor(Region.factor)
expense$Region.factor=Region.factor
myfit=lm(Expenditure~Income+Urban+Region.factor,data=expense)
boxplot(resid(myfit)~expense$Region)
```



From the boxplot,the height of each box is widely different. Therefore, I think it is inappropriate to assume each region has the same variance.

4

According to the model, $y = \mu_1 \times G_1 + \mu_2 \times G_2 + \mu_3 \times G_3 + \epsilon, \epsilon \perp (G_1, G_2, G_3), \epsilon \sim \mathcal{N}(0, \sigma^2)$,where $G_i = I(y \in group\ i), i = 1, 2, 3$. We call it the full model.

Under $H_0 : \mu_1 = \mu_2 = \mu_3 = \mu$, the model becomes: $y = \mu + \epsilon, \mu \perp \epsilon, \epsilon \sim \mathcal{N}(0, \sigma^2)$, which we call the null model.

```
#write the table
data=data.frame(
  y=c(-1.7,-1.5,-0.4,-1.1,1.3,-0.3,2.0,1.2,0.6),
  G1=c(1,1,0,0,0,0,0,0,0),
  G2=c(0,0,1,1,1,1,0,0,0),
  G3=c(0,0,0,0,0,0,1,1,1)
)
data
```

```
##      y G1 G2 G3
## 1 -1.7  1  0  0
## 2 -1.5  1  0  0
## 3 -0.4  0  1  0
## 4 -1.1  0  1  0
## 5  1.3  0  1  0
## 6 -0.3  0  1  0
## 7  2.0  0  0  1
## 8  1.2  0  0  1
## 9  0.6  0  0  1
```

```
full_model=lm(y~G1+G2+G3,data=data)
z=data[,"G1"]+data[,"G2"]+data[,"G3"]
null_model=lm(y~z,data=data)
anova(full_model,null_model)
```

```
## Analysis of Variance Table
##
## Model 1: y ~ G1 + G2 + G3
## Model 2: y ~ z
##   Res.Df     RSS Df Sum of Sq      F  Pr(>F)
## 1      6  4.0942
## 2      8 14.0889 -2   -9.9947 7.3236 0.02454 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The result shows that the p value of H_0 is 0.02454<0.05, which suggests that we should reject H_0.