Homework 2

Yuwen(Suyi)Wu 5:45-7:45

2/28/2019

Question 1

Read the posted article, “Bordeaux wine vintage quality and weather,” by Ashenfelter, Ashmore, and LaLonde (CHANCE, 1995). Three regression models are considered in this article. Answer the following questions:

(a) What is a wine “vintage”?

A wine vintage is the specified year when the primarily grapes that uesed to make wine were harvested.

A wine’s vintage could largely represented taste and quality of a wine because the weather of that vintage would affects the grapes’ quality and mature throughout the growing season.

(b) What is the response variable for the three models described in this paper?

The response variable for the three models described in this paper is logarithm of the price relative to 1961 (LPRICE2).

(c) Which values of LPRICE2 are missing and, according to the article, why have they been omitted?

```r
# import data
winedata<- read_delim("wine.dat", col_names=TRUE,delim=',')
# find missing data
missingvalue <-filter(winedata, winedata$LPRICE2 == ".")
# The value of LPRICE2 missed:
missingvalue$VINT

##  [1] 1954 1956 1981 1982 1983 1984 1985 1986 1987 1988 1989
```

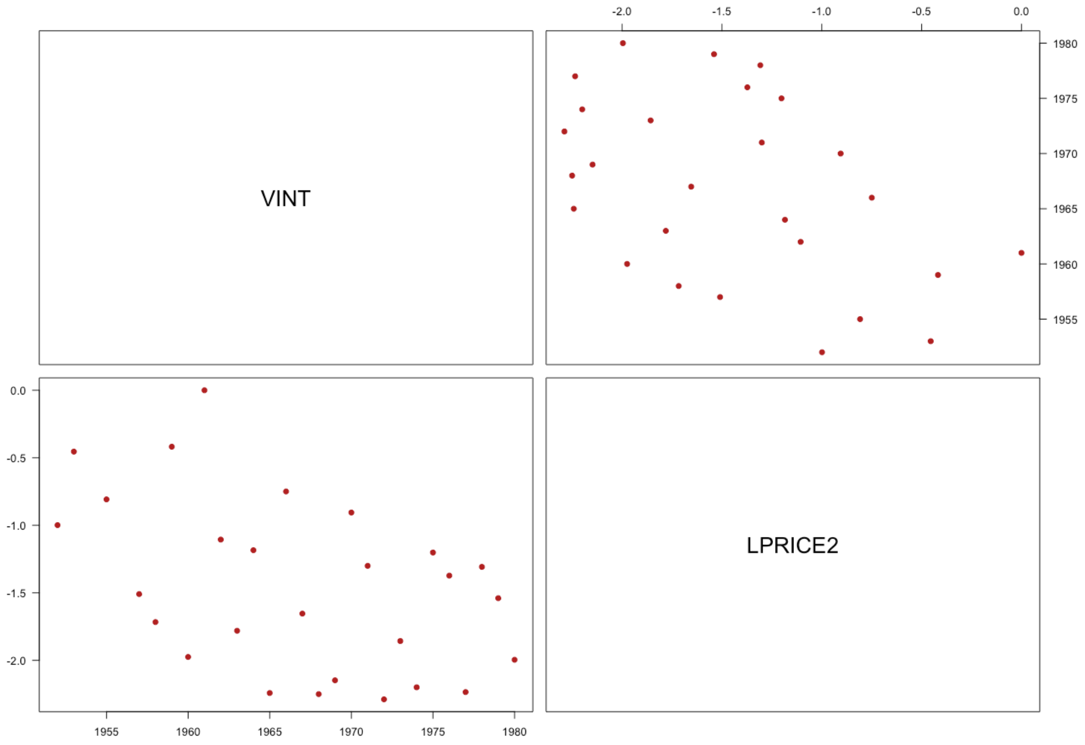
The vintages of 1954 and 1956 are not plotted because these wines were rarely sold. From the article, the wine needs 28 years to mature. Therefore, the vintages of 1981,1982,1983,1984,1985,1986,1987,1988 and 1989 are not ploted because these wines were not mature at that time.

(d) Make a scatterplot matrix of the variables (explanatory and response) included in the models. Describe what you see.

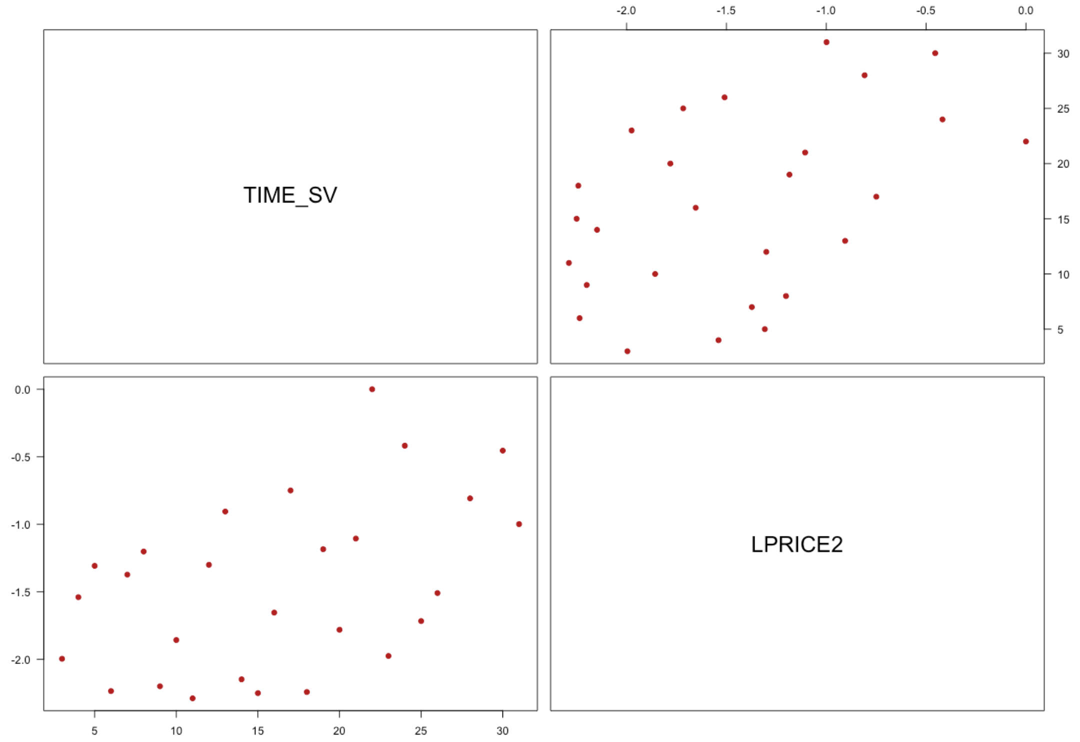
```r
# check type of column, in this case, the LPRICE2 and DEGREES are chr type.
winedata
```

```
## # A tibble: 38 x 6
##      VINT LPRICE2  WRAIN DEGREES HRAIN TIME_SV
##     <dbl> <chr>    <dbl> <chr>   <dbl>   <dbl>
##  1  1952 -0.99868   600 17.1167   160      31
##  2  1953 -0.4544    690 16.7333    80      30
##  3  1954 .          430 15.3833   180      29
##  4  1955 -0.80796   502 17.15     130      28
##  5  1956 .          440 15.65     140      27
##  6  1957 -1.50926   420 16.1333   110      26
##  7  1958 -1.71655   582 16.4167   187      25
##  8  1959 -0.418     485 17.4833   187      24
##  9  1960 -1.97491   763 16.4167   290      23
## 10  1961 0          830 17.3333    38      22
## # … with 28 more rows
```

```
# filter the na data
newwinedata <-filter(winedata, winedata$LPRICE2 != ".")
# make chr type to numeric type
newwinedata$LPRICE2 <- as.numeric(newwinedata$LPRICE2)
newwinedata$DEGREES <- as.numeric(newwinedata$DEGREES)

# make a scatterplot matrix of the varuables included in the models.

# first model
pairs(newwinedata[,c("VINT", "LPRICE2")], las = T, pch = 19, col =
"firebrick")
```

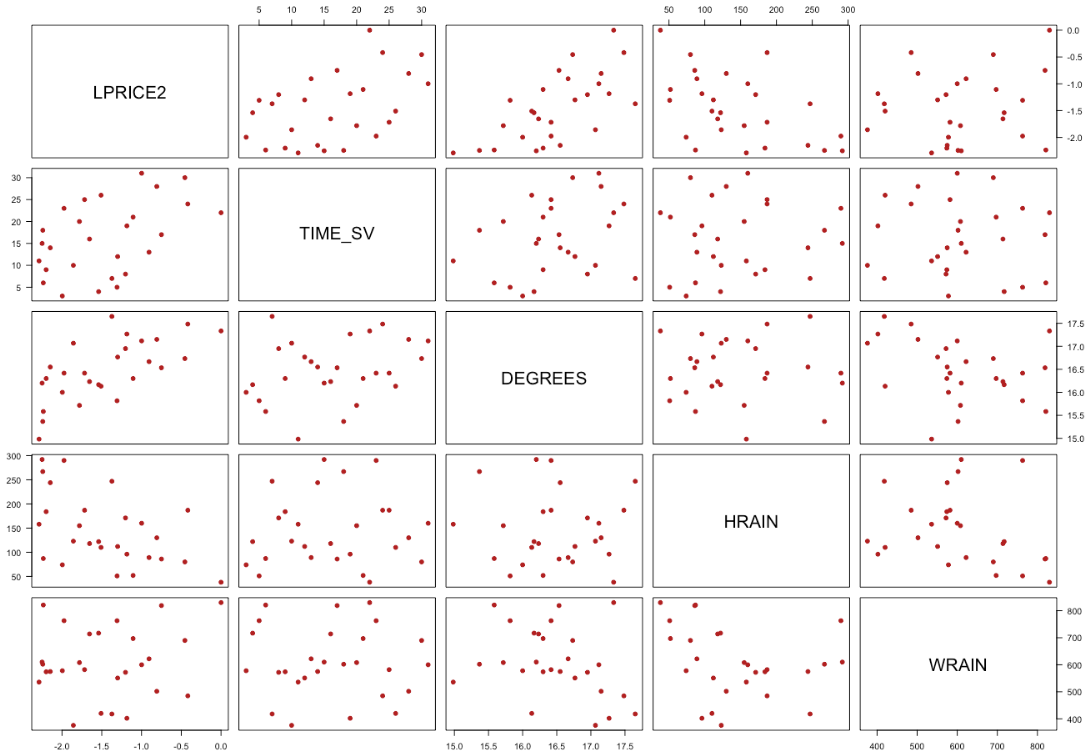


```
pairs(newwinedata[,c("TIME_SV", "LPRICE2")], las = T, pch = 19, col =
"firebrick") ## this is model mentioned in the article
```



From above scatterplot, all shows that with longer time wine stored, the price or return would be higher. And the vintage has negative linear relationship with LPRICE2.

```
# second model
pairs(newwinedata[,c("LPRICE2","TIME_SV","DEGREES", "HRAIN", "WRAIN" )], las
= T, pch = 19, col = "firebrick")
```



There is a negative linear relationship between VINT and LPRICE2, a positive linear relationship between TIME_SV and LNPRICE2, a pisitive linear relationship between LPRICE2 and DEGREES and a negative linear relationship between LPRICE2 and HRAIN.The relation between VINT and TIME_SV is a complete negative relation which is because of the time saving is the result of present year minus vintage.

(e) Fit the two regression models from the paper. Which is the best regression model?

Justify your answer and include relevant output (let α = 0.05). Did you choose the same model as the authors?

```
# linear regression model for one factor and multifactor

rm1 <- lm(LPRICE2~TIME_SV,data = newwinedata)
rm2 <- lm(LPRICE2~TIME_SV+DEGREES+HRAIN+WRAIN,data=newwinedata)

# check the result of regression models

summary(rm1)

##
## Call:
## lm(formula = LPRICE2 ~ TIME_SV, data = newwinedata)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.8545 -0.4788 -0.0718  0.4562  1.2457
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -2.02520    0.24723  -8.192 1.52e-08 ***
## TIME_SV      0.03543    0.01366   2.593   0.0157 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5745 on 25 degrees of freedom
## Multiple R-squared:  0.212,  Adjusted R-squared:  0.1804
## F-statistic: 6.725 on 1 and 25 DF,  p-value: 0.01567
```

```
summary(rm2)
```

```
##
## Call:
## lm(formula = LPRICE2 ~ TIME_SV + DEGREES + HRAIN + WRAIN, data =
newwinedata)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.46027 -0.23864  0.01347  0.18600  0.53446
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.215e+01  1.688e+00  -7.195 3.28e-07 ***
## TIME_SV      2.385e-02  7.167e-03   3.328  0.00306 **
## DEGREES      6.164e-01  9.518e-02   6.476 1.63e-06 ***
## HRAIN       -3.861e-03  8.075e-04  -4.781 8.97e-05 ***
## WRAIN        1.167e-03  4.820e-04   2.421  0.02420 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2865 on 22 degrees of freedom
## Multiple R-squared:  0.8275, Adjusted R-squared:  0.7962
## F-statistic: 26.39 on 4 and 22 DF,  p-value: 4.058e-08
```

```
# partial F-test

anova(rm1, rm2)
```

```
## Analysis of Variance Table
##
## Model 1: LPRICE2 ~ TIME_SV
## Model 2: LPRICE2 ~ TIME_SV + DEGREES + HRAIN + WRAIN
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1     25 8.2509
## 2     22 1.8058  3    6.4451 26.173 1.909e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# p-value = P(F > 26.173)
pf(26.173, df1=3, df2=22, lower.tail=FALSE)

## [1] 1.90895e-07
```

The model2 is better because the Adjusted R square of model2 (0.7962) is larger than Adjusted R square of model1 (0.1804).

Hypothesis: H0: β(WRAIN) = β(HRAIN) = β(DEGREES) = 0 H1: At least one of beta is not zero (y-intercept is not included) significance level : α = 0.05 test statistic: F = 26.173 of 4 and 22 degrees of freedom P-value of F test is 4.058e-08 < α = 0.05 At partial F-test, P-value is 1.90895e-07 < α = 0.05

we should reject null hypothesis which is all beta are equal to zero

at least one of betas is significantly different from 0. Therefore, the second model is better than the first model.I choose the same model as the author did.

(f) What is the sample size for your models?

```
# find out the sample size
nrow(newwinedata)

## [1] 27

# the sample size is 27
```

(g) Write out the regression equation of the model you chose in part (e). Remember to include the units of measurement. Interpret the partial slopes and the y-intercept. Does the y-intercept have a practical interpretation?

LPRICE2(per dozen bottle in $) = -12.15 - 0.02385(1/years)*TIME_SV(years) + 0.6164(1/℃)*DEGREES(°C) - 0.003861(1/ml)*HRAIN(ml) + 0.00167(1/ml)*WRAIN(ml)

beta for factor TIME_SV: if the age of vintage increase 1 year, holding other exploratory variables DEGREES, WRAIN and HRAIN constant, the log of average vintage price relative to 1961 decreases roughly 0.02385.

beta for factor DEGREES: if the average temperature over growing season increase 1°C, holding other exploratory variables TIME_SV, WRAIN and HRAIN constant, the log of average vintage price relative to 1961 increases roughly 0.6164.

beta for factor HRAIN: if the rain in September and August increase 1 ml, holding other exploratory variables TIME_SV, DEGREES and WRAIN constant, the log of average vintage price relative to 1961 decreases roughly 0.003861.

beta for factor WRAIN: if the rain in the months preceeding the vintage increase 1 ml, holding other exploratory variables TIME_SV, DEGREES and HRAIN constant, the log of average vintage price relative to 1961 increases roughly 0.00167.

Y-intercept: if the age of vintage is 0 year, the average temperature over the growing season is 0℃, the rain both in the months preceding the vintage and in September and August is 0 ml, the log of average vintage price relative to 1961 would be roughly -12.15.

The y-intercept does not have practical interpretation. Since the minimum values of DEGREES, WRAIN and HRAIN are 14.98℃, 376.0 ml and 38.0 m.

The y-intercept does not have a practical interpretation.

(h) Make a table with the following statistics for both models: SSE, RMSE, PRESS, and RMSE jackknife . Compare the relevant statistics. Based on this information, would you change your answer to part (e)? Justify your answers.

```
# Calcualte RMSE(Residual Mean Standard Error)
RMSE.1 <- summary(rm1)$sigma
# Model 1 has RMSE of 0.5745
RMSE.2 <- summary(rm2)$sigma
# Model 2 has RMSE of 0.2865

# Calculate SSE(Sum of Squared Error)
anova.1 <- anova(rm1)
SSE.1 <- anova.1$`Sum Sq`[2]
# Model 1 has SSE of 8.2509
anova.2 <- anova(rm2)
SSE.2 <- anova.2$`Sum Sq`[5]
# Model 2 has SSE of 1.8058

# Calculate PRESS
PRESS.1 <- press(rm1)
# Model 1 has PRESS of 9.395569
PRESS.2 <- press(rm2)
# Model 2 has PRESS of 2.816957

# Calculate RMSE jackknife
# Linear Regression Model One has only one variable
jack.1 <- sqrt(press(rm1)/(nrow(newwinedata)-1-1))
# Model 1 has RMSE jackknife of 0.6130438
jack.2 <- sqrt(press(rm2)/(nrow(newwinedata)-4-1))
# Linear Regression Model Two has 4 variables
# Model 2 has RMSE jackknife of 0.3578317

table <- data.frame("lm.1" = c(SSE.1, RMSE.1, PRESS.1, jack.1), "lm.2" =
c(SSE.2, RMSE.2, PRESS.2, jack.2))
rownames(table) <- c("SSE", "RMSE", "PRESS", "MSE(Jackknife)")
table

##                      lm.1      lm.2
## SSE            8.2508832 1.8058296
## RMSE           0.5744870 0.2865016
## PRESS          9.3955689 2.8169574
## MSE(Jackknife) 0.6130438 0.3578317
```

Based on the table, model 2 has smaller SSE, RMSE, PRESS and RMSE(jackknife) than model1, which means model 2 fits better. Therefore, I would not change my answer in part (e).

(i) Could we use these regression models to predict quality for wines produced in 2005? Justify your answer.

We cannot use these models to predict the quality for wines produced in 2005. Firstly, the data of VINT, DEGREES, HRAIN, WRAIN in 2005 is not included in the database. Secondly, in this case, we have to predict in extrapolation method, in this case, we cannot make sure that our prediction in the correct range for all x 0 s simultaneously and is feasible.

Question 2

(a) Do some internet research and write a short paragraph in your own words about how the Pineo-Porter prestige score is computed. Include the reference(s) you used. Do you think this score is a reliable measure? Justify your answer.

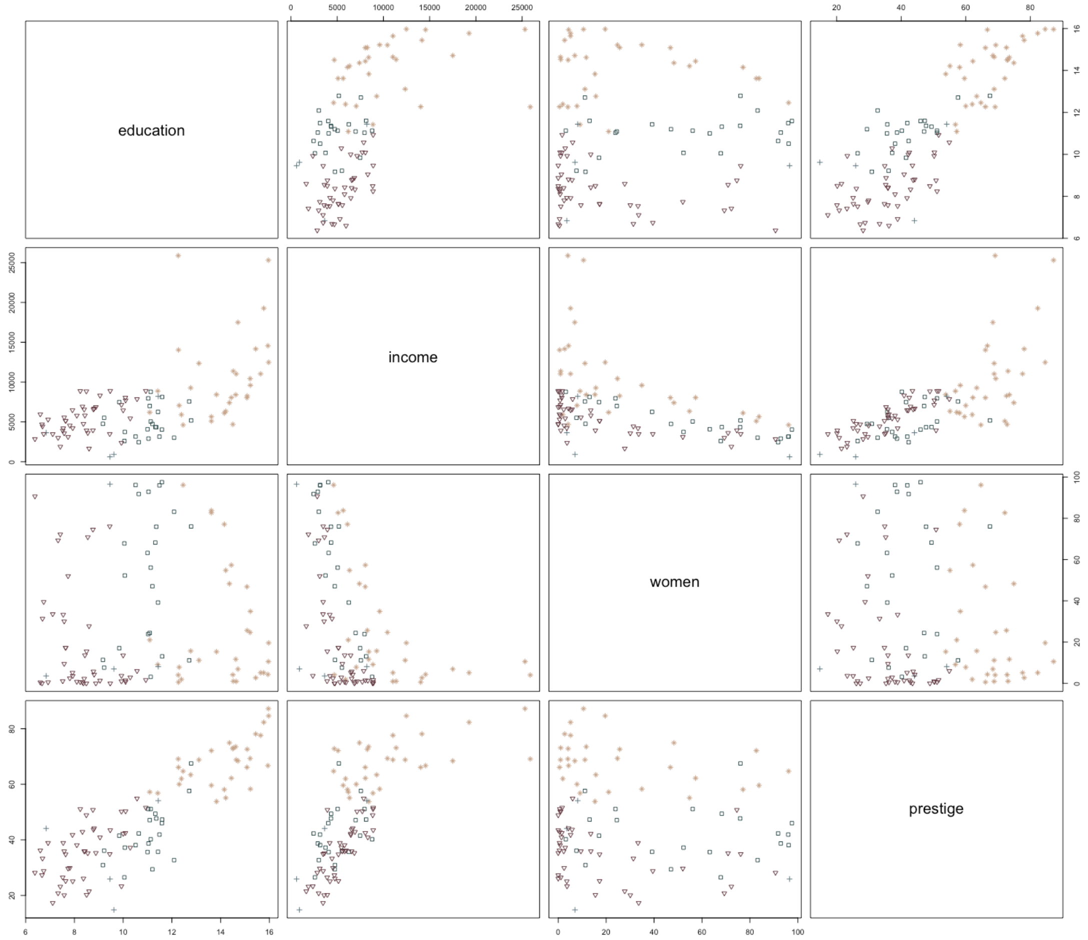
```
Pineo-Porter prestige score is a measurement for occupation, from a social
survey conducted in the mid-1960.

Occupational prestige (also known as job prestige) is a way for sociologists
to describe the relative social class positions people have. It refers to the
consensual nature of rating a job based on the belief of its worthiness.

Sociologists have identified prestige rankings for more than 700 occupations
based on results from a series of national surveys.

They created a scale with 0 being the lowest possible score to 100 being the
highest, and then ranked the occupations based on the results of the survey.

Pineo, Porter, McRoberts scale of 1977 is the one of prestige score that
ordered major groups.

Above mentioned scale score is the Pineo-Porter prestige score.

I think it is a reliable and valuabel measure because many researchers
prefered it and it ordered by major groups which reduced the dummy variables
to do linear regression analysis.
```

This is the Wikipedia,where we can search for the definition of Pineo-Porter presitge score.

(b) Create a scatterplot matrix of all the quantitative variables. Use a different symbol for each profession type: no type (pch=3), "bc" (pch=6), "prof" (pch=8), and "wc" (pch=0) when making your plot. For the remainder of this question, we will use the explanatory variables: income, education, and type. Does restricting our regression to only these variables make sense given your exploratory analysis? Justify your answer.

```r
# import data
prestige <- read_delim("prestige.dat", col_names=TRUE,delim=',')
```

```
# check data
head(prestige)

## # A tibble: 6 x 7
##   occupation.group    education income women prestige census type
##   <chr>                   <dbl>  <dbl> <dbl>    <dbl>  <dbl> <chr>
## 1 gov.administrators       13.1  12351 11.2      68.8   1113 prof
## 2 general.managers         12.3  25879  4.02     69.1   1130 prof
## 3 accountants              12.8   9271 15.7      63.4   1171 prof
## 4 purchasing.officers      11.4   8865  9.11     56.8   1175 prof
## 5 chemists                 14.6   8403 11.7      73.5   2111 prof
## 6 physicists               15.6  11030  5.13     77.6   2113 prof

# plot the pair plot
theme.info <- theme(plot.title = element_text(size=30, hjust=0.5),
                    axis.title = element_text(size=30),
                    axis.text = element_text(size=30),
                    legend.title = element_text(color = "black", size = 30),
                    legend.text = element_text(color = "black", size = 30))
# set pchs
pchs <- rep(NA,times = nrow(prestige))
prestige$type[prestige$type == ""] <- NA
pchs[prestige$type == "bc"] <- 6
pchs[is.na(prestige$type)] <- 3
pchs[prestige$type == "prof"] <- 8
pchs[prestige$type == "wc"] <- 0
# check pchs
pchs

##   [1] 8 8 8 8 8 8 8 8 8 8 8 8 8 8 8 8 8 8 8 8 8 8 8 8 8 8 8 6 8 8 0 8 0 3
0
##  [36] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 3 6 0 0 0 6 6 6 6 6 3 6 6 6 3 6 6
6
##  [71] 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 8 6 6 6 6 6 6

# set color
color.setting <- rep(NA,nrow(prestige))
color.setting [prestige$type == "bc"] <- "#663A44"
color.setting [is.na(prestige$type)] <- "#5F7880"
color.setting [prestige$type == "prof"] <- "#CAA78D"
color.setting [prestige$type == "wc"] <- "#2F4F4F"
# plot the pairs
pairs(prestige[,c("education","income","women","prestige")],pch = pchs, col =
color.setting)
```



```
Restricting our regression to only these variables make sense.
Firstly, we delete non-related column of census. Secondly, it could be seen
from scatterplot, education and income have linear relationship to prestige
but the woman does not. Therefore, we could delete woman data in regression
model.
```

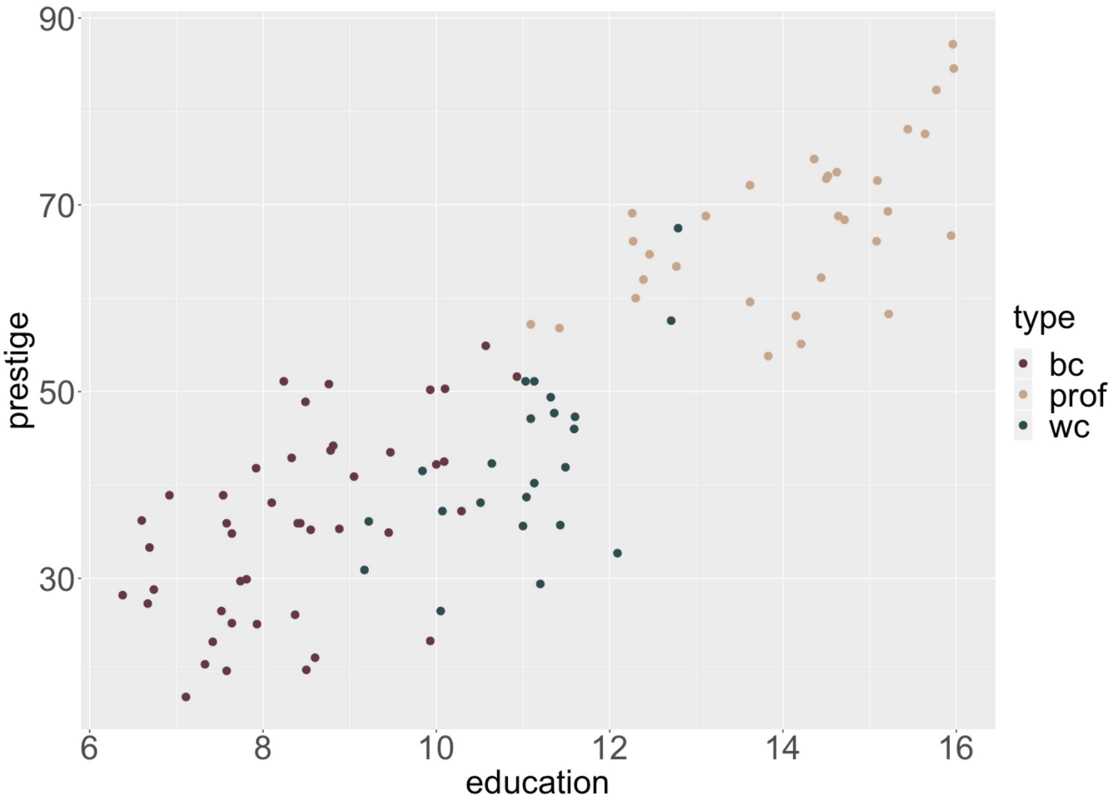
(c) Which professions are missing "type"? Since the other variables for these observations are available, we could group them together as a fourth professional category to include them in the analysis. Is this advisable or should we remove them from our data set? Justify your answer.

```
prestige$occupation.group[is.na(prestige$type)]

## [1] "athletes"    "newsboys"    "babysitters" "farmers"

# These four professions are  "athletes"    "newsboys"    "babysitters"
"farmers"
prestige <- prestige[!is.na(prestige$type),]
```

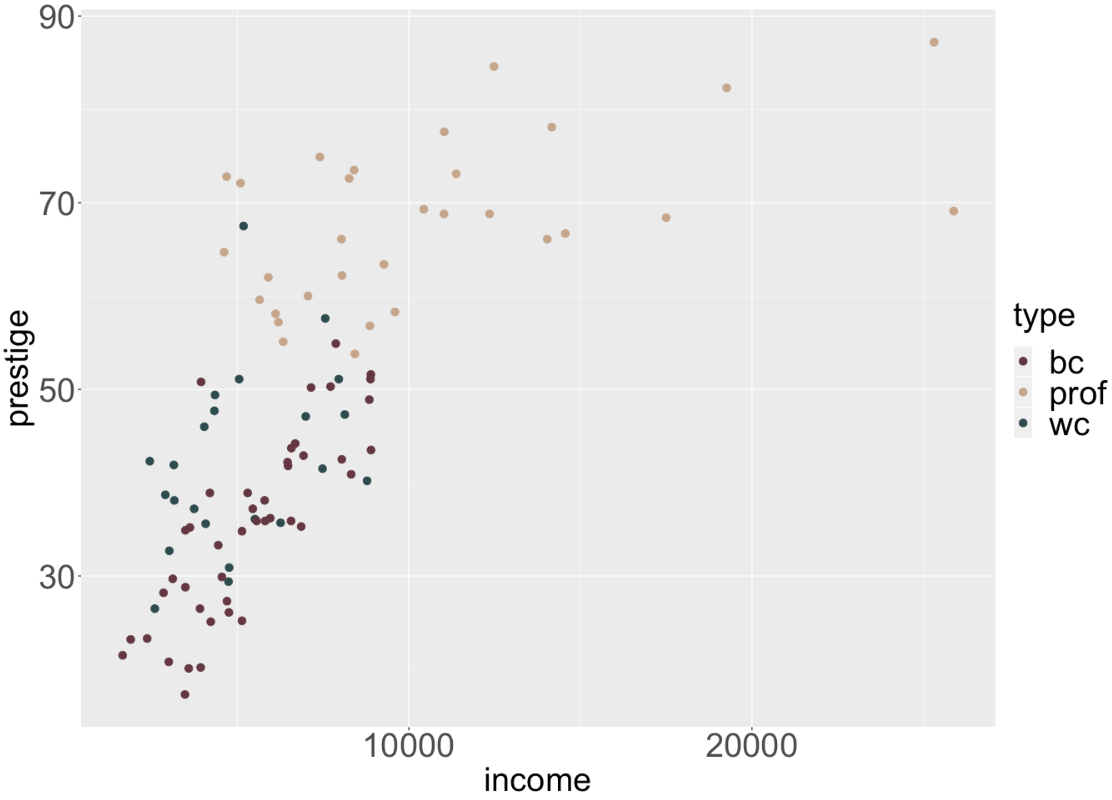
Since these four occupations are not belong to a specific profession type, we should remove it rather than group them together as a fourth professional category.

(d) Visually, does there seem to be an interaction between type and education and/or type and income? Justify your answer.

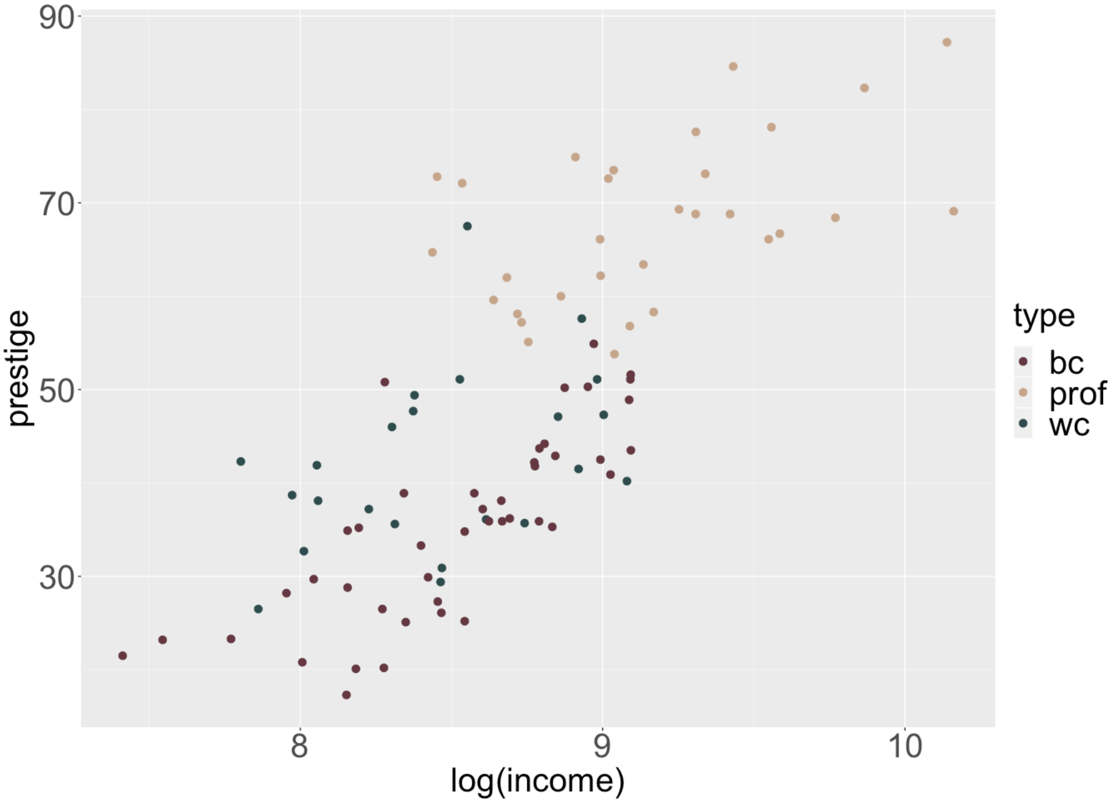
```
col.vector <- c("bc"="#663A44","prof"="#CAA78D","wc"="#2F4F4F")
s.1 <- prestige %>% ggplot(aes(x=education, y=prestige, col=type)) +
            geom_point(size=3) +
            scale_color_manual(values=col.vector) +
            theme.info
s.2 <- prestige %>% ggplot(aes(x=income, y=prestige, col=type)) +
            geom_point(size=3) +
            scale_color_manual(values=col.vector) +
            theme.info
s.3 <- prestige %>% ggplot(aes(x=log(income), y=prestige, col=type)) +
            geom_point(size=3) +
            scale_color_manual(values=col.vector) +
            theme.info
s.1
```



```
s.2
```


90
70
50
30
prestige
10000
20000
income
type
bc
prof
wc

```
s.3
```


90
70
50
30
prestige
8
9
10
log(income)
type
bc
prof
wc

From above scatterplot, it could be seen that there is an interaction between type and education and type and income.

The income have a curved relationship with prestige and log(income) have a linear relationship with prestige.

With the different shapes of point, the people who are in professional occupation have higher education level and higher income than white collar and blue collar.

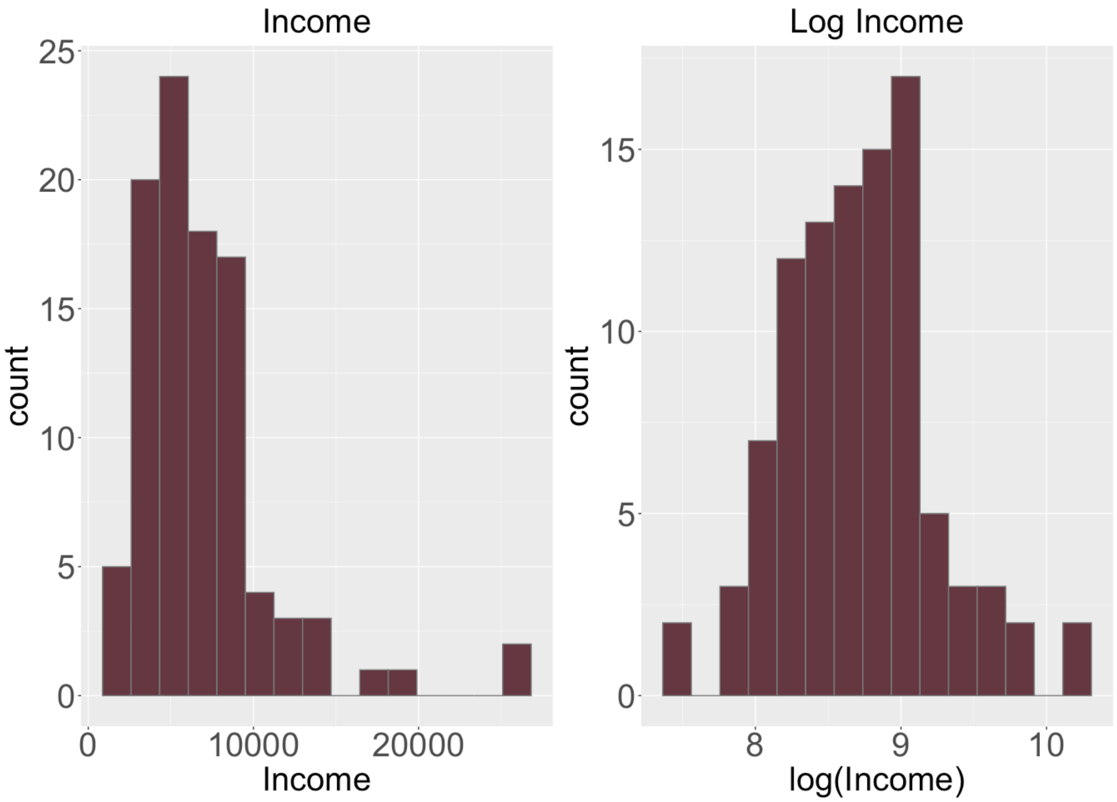
(e) Fit a model to predict prestige using: income, education, type, and any interaction terms based on your answer to part (d). Evaluate the model and include relevant output. Use your answer to part (c) to determine which observations to use in your analysis.

```
lm.2e <- lm(prestige~education+income+type+education*type+income*type, data =
prestige)
summary(lm.2e)

##
## Call:
## lm(formula = prestige ~ education + income + type + education *
##     type + income * type, data = prestige)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -13.462  -4.225   1.346   3.826  19.631
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)          2.276e+00  7.057e+00   0.323   0.7478
## education            1.713e+00  9.572e-01   1.790   0.0769 .
## income               3.522e-03  5.563e-04   6.332 9.62e-09 ***
## typeprof             1.535e+01  1.372e+01   1.119   0.2660
## typewc              -3.354e+01  1.765e+01  -1.900   0.0607 .
## education:typeprof  1.388e+00  1.289e+00   1.077   0.2844
## education:typewc    4.291e+00  1.757e+00   2.442   0.0166 *
## income:typeprof    -2.903e-03  5.989e-04  -4.847 5.28e-06 ***
## income:typewc      -2.072e-03  8.940e-04  -2.318   0.0228 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.318 on 89 degrees of freedom
## Multiple R-squared:  0.8747, Adjusted R-squared:  0.8634
## F-statistic: 77.64 on 8 and 89 DF,  p-value: < 2.2e-16

under the t test of beta = 0, with alpha = 0.05, assume regression
assumptions are satisfied, with 89 degrees of freedom.

slope of education, type of prof, type of white collar, education when type
of prof are not statistically significant.
```

```
slope of income, education when type of white collar, income when type of
prof, income when type of white colla are statistically significant.

p-value of overall F-statistic is small enough, and F test is significant.

Under adjusted R squared, about 86.34% of the variability in prestige can be
explained by a regression model all above mentioned
variables(education,income,type,education when prof, education when white
collar, income when prof, income when white collar)


Since income when type of prof is significant, we cannot delete the variable
of type even the dummy variable type is not statistically significant.

Since education when type of white collar is statistically significant, we
cannot delete the variable of education even education is not statistically
significant.

Final Model
When Type = bc:
Prestige(points) = 2.28(points) + 1.7(points/years) * education(years) +
0.00352(points/$) * income($)

When Type = prof:
Prestige (points) = 2.28(points) + 15.35(points) + 1.7(points/years) *
education(years) + 0.00352(points/$) * income($) + 1.39(points/years) *
education(years) - 0.0029(points/$) * income($)

Prestige (points) = 17.63(points) + 3.09(points/years)*education(years) +
0.00062(points/$) * income($)


When Type = white collar:

Prestige(points) = 2.28(points) - 33.54(points) + 1.7(points/years) *
education(years) + 0.00352(points/$) * income($) + 4.29(points/years) *
education(years) - 0.0021(points/$) * income($)

Prestige(points) = -31.26(points) + 5.99(points/years) * education(years) +
0.00142(points/$) * income($)
```

(f) Create a histogram of income and a second histogram of log(income) (i.e., natural logarithm). How does the distribution change?

```r
hist.1<- prestige %>% ggplot(aes(income)) +
  geom_histogram(bins=15, fill="#663A44", col="gray50") +
  ggtitle("Income") +
  labs(x = "Income") +
  theme.info

hist.2<- prestige %>% ggplot(aes(log(income))) +
  geom_histogram(bins=15, fill="#663A44", col="gray50") +
  ggtitle("Log Income") +
  labs(x = "log(Income)") +
  theme.info

grid.arrange(hist.1,hist.2,ncol=2)
```



```
## The distribution change from positive skewed to distribution looks
approximately symmetric.
```

(g) Fit the model in (e) but this time use log(income) (i.e., natural logarithm) instead of income. Evaluate the model and provide the relevant output.

```
prestige.2g <- prestige[,-3]
data.2g <- data.frame(prestige.2g,log(prestige$income))
lm.2g <-
lm(prestige~education+log.prestige.income.+type+education*type+log.prestige.i
ncome.*type,data = data.2g)
summary(lm.2g)

##
## Call:
## lm(formula = prestige ~ education + log.prestige.income. + type +
##     education * type + log.prestige.income. * type, data = data.2g)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -13.970  -4.124   1.206   3.829  18.059 
## 
## Coefficients:
##                                Estimate Std. Error t value Pr(>|t|)    
## (Intercept)                   -120.0459    20.1576  -5.955 5.07e-08 ***
## education                        2.3357     0.9277   2.518  0.01360 *  
## log.prestige.income.            15.9825     2.6059   6.133 2.32e-08 ***
## typeprof                        85.1601    31.1810   2.731  0.00761 ** 
## typewc                          30.2412    37.9788   0.796  0.42800    
## education:typeprof               0.6974     1.2895   0.541  0.58998    
## education:typewc                 3.6400     1.7589   2.069  0.04140 *  
## log.prestige.income.:typeprof   -9.4288     3.7751  -2.498  0.01434 *  
## log.prestige.income.:typewc     -8.1556     4.4029  -1.852  0.06730 .  
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 6.409 on 89 degrees of freedom
## Multiple R-squared:  0.871,  Adjusted R-squared:  0.8595 
## F-statistic: 75.15 on 8 and 89 DF,  p-value: < 2.2e-16
```

under the t test of beta = 0, with alpha = 0.05, assume regression assumptions are satisfied, with 89 degrees of freedom.

slope of education, log of income, type of prof, education when type of white collar and log income when type of prof are statistically significant.

slope of type of white collar, education when type of prof, log income when type of white collar are not statistically significant.

Under adjusted R squared, about 85.95% of the variability in prestige can be explained by a regression model all above mentioned variables(education,income,type,education when prof, education when white collar, income when prof, income when white collar)

p-value of overall F-statistic is small enough, and F test is significant.

Final model:
When Type = bc:

Prestige (points) = -120.0459(points) + 2.3357(points/years) * education(years) + 15.9825(points/$) * log.income($)

When Type = prof:
Prestige (points) = -120.0459(points) + 2.3357(points/years) *

```
education(years) + 15.9825(points/$) * log.income($) + -85.1601(points) +
0.6974(points/years) * education(years) - 9.4288(points/$) * log.income($)
Prestige (points) = -34.8858(points) + 3.0331(points/years) *
education(years) + 6.5537(points/$) * log.income($)

When Type = white collar:
Prestige(points) = -120.0459(points) + 2.3357(points/years) *
education(years) + 15.9825(points/$) * log.income($) + 30.2412(points) +
3.6400(points/years) * education(years) - 8.1556(points/$) * log.income($)
Prestige(points) = -89.8047(points) + 5.9757(points/years) * education(years)
+ 7.8269(points/$) * log.income($)
```

(h) Is the model in (e) or (g) better? Justify your answer. Why can't we use a partial F-test here?

```
# Calcualte RMSE(Residual Mean Standard Error)
RMSE.e <- summary(lm.2e)$sigma
# Model e has RMSE of 6.318211
RMSE.g <- summary(lm.2g)$sigma
# Model g has RMSE of 6.408734

# Calculate SSE(Sum of Squared Error)
anova.2e <- anova(lm.2e)
SSE.e <- anova.2e$`Sum Sq`[2]
# Model e has SSE of 1791.966
anova.2g <- anova(lm.2g)
SSE.g <- anova.2g$`Sum Sq`[5]
# Model g has SSE of 290.3298

## Calculate PRESS
PRESS.e <- press(lm.2e)
# Model e has PRESS of 4285.977
PRESS.g <- press(lm.2g)
# Model g has PRESS of 4399.257

# Calculate RMSE jackknife
# Linear Regression Model e has 8 variable
jack.e <- sqrt(press(lm.2e)/(nrow(prestige)-8-1))
# Model e has RMSE jackknife of 6.939528
jack.g <- sqrt(press(lm.2g)/(nrow(prestige.2g)-8-1))
# Linear Regression Model g has 8 variables
# Model g has RMSE jackknife of 7.030637

table2 <- data.frame("lm.e" = c(SSE.e, RMSE.e, PRESS.e, jack.e), "lm.g" =
c(SSE.g, RMSE.g, PRESS.g, jack.g))
rownames(table2) <- c("SSE", "RMSE", "PRESS", "MSE(Jackknife)")
table2

##                       lm.e        lm.g
## SSE            1791.965590  290.329761
## RMSE               6.318211    6.408734
## PRESS           4285.976747 4399.257220
## MSE(Jackknife)    6.939528    7.030637
```

The income have a curved relationship with prestige and log(income) have a linear relationship with prestige.

The linear regression model (e) has higher adjusted R-square(0.8634) than adjusted R-square(0.8595) in lineaer regression model (g).

Additionally, The model (e) has smaller SSE, RMSE, PRESS and RMSE(jackknife) than model in (g). Thus, model in (e) is better than model in (g).

We cannot use partial F-test here because these two models are not nested in each other.