HW

Applied regression

Isabelle

00/00/2020

## Question 1

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.

Answer: The Pineo-Poter prestige score is used to measure the socioeconomic status of the people in Canada at the time of its creation. It is calculated by regression analysis based on social attributes including education and income. Pineo Poter can calculate the prestige index of 16 works, and its calculation formula is

This was a reliable measure when it was first released, since the study was led by professor using the data from census. It was created to help in comparing and measure occupations in the 1971 and 1981 censuses. However, in the current situation, due to the emergence of more new occupations, the definition of occupation type has changed a lot. Moreover, many previous occupations have been redefined, and their income levels and social status are totally different. Therefore, in the current environment, the reputation score model, Pineo Poter, is not reliable.

reference:<http://rstudio-pubs-static.s3.amazonaws.com/425420_448c3a57871f4ac3a98f7b7781ffc91e.html>

<http://deonandan.com/pdf/lcdce.pdf>

## 

## Question 2

Create a scatterplot matrix of all the quantitative variables. Use a different symbol (or color) for each profession type: no type, “bc”, “prof”, and “wc” 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.

Solution:

#read the data  
x <- read\_delim("prestige.dat", col\_names=TRUE, delim=",")

## Parsed with column specification:  
## cols(  
## occupation.group = col\_character(),  
## education = col\_double(),  
## income = col\_double(),  
## women = col\_double(),  
## prestige = col\_double(),  
## census = col\_double(),  
## type = col\_character()  
## )

head(x)

## # 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

# check the data type  
glimpse(x)

## Observations: 102  
## Variables: 7  
## $ occupation.group <chr> "gov.administrators", "general.managers", "accountan…  
## $ education <dbl> 13.11, 12.26, 12.77, 11.42, 14.62, 15.64, 15.09, 15.…  
## $ income <dbl> 12351, 25879, 9271, 8865, 8403, 11030, 8258, 14163, …  
## $ women <dbl> 11.16, 4.02, 15.70, 9.11, 11.68, 5.13, 25.65, 2.69, …  
## $ prestige <dbl> 68.8, 69.1, 63.4, 56.8, 73.5, 77.6, 72.6, 78.1, 73.1…  
## $ census <dbl> 1113, 1130, 1171, 1175, 2111, 2113, 2133, 2141, 2143…  
## $ type <chr> "prof", "prof", "prof", "prof", "prof", "prof", "pro…

# one-variable summaries   
summary(x[,c("women", "income", "education","prestige","census")])

## women income education prestige   
## Min. : 0.000 Min. : 611 Min. : 6.380 Min. :14.80   
## 1st Qu.: 3.592 1st Qu.: 4106 1st Qu.: 8.445 1st Qu.:35.23   
## Median :13.600 Median : 5930 Median :10.540 Median :43.60   
## Mean :28.979 Mean : 6798 Mean :10.738 Mean :46.83   
## 3rd Qu.:52.203 3rd Qu.: 8187 3rd Qu.:12.648 3rd Qu.:59.27   
## Max. :97.510 Max. :25879 Max. :15.970 Max. :87.20   
## census   
## Min. :1113   
## 1st Qu.:3120   
## Median :5135   
## Mean :5402   
## 3rd Qu.:8312   
## Max. :9517

round(apply(x[,c("women", "income", "education","prestige","census")], 2, sd), digits=3)

## women income education prestige census   
## 31.725 4245.922 2.728 17.204 2644.993

#create a scatterplot matrix for all quantative variables  
x %>% ggpairs(aes(color=type), columns=c(2,3,4,5,6),title = "Scatterplot Matrix of all the quantitative variables")

## Warning in (function (data, mapping, alignPercent = 0.6, method = "pearson", :  
## Removed 4 rows containing missing values

## Warning in (function (data, mapping, alignPercent = 0.6, method = "pearson", :  
## Removed 4 rows containing missing values

## Warning in (function (data, mapping, alignPercent = 0.6, method = "pearson", :  
## Removed 4 rows containing missing values

## Warning in (function (data, mapping, alignPercent = 0.6, method = "pearson", :  
## Removed 4 rows containing missing values

## Warning in (function (data, mapping, alignPercent = 0.6, method = "pearson", :  
## Removed 4 rows containing missing values

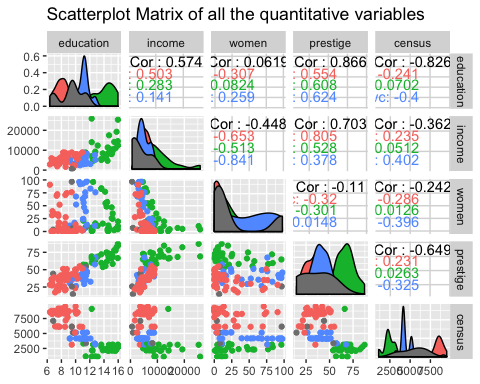
## Warning in (function (data, mapping, alignPercent = 0.6, method = "pearson", :  
## Removed 4 rows containing missing values

## Warning in (function (data, mapping, alignPercent = 0.6, method = "pearson", :  
## Removed 4 rows containing missing values

## Warning in (function (data, mapping, alignPercent = 0.6, method = "pearson", :  
## Removed 4 rows containing missing values

## Warning in (function (data, mapping, alignPercent = 0.6, method = "pearson", :  
## Removed 4 rows containing missing values

## Warning in (function (data, mapping, alignPercent = 0.6, method = "pearson", :  
## Removed 4 rows containing missing values



theme.info

## List of 3  
## $ axis.title:List of 11  
## ..$ family : NULL  
## ..$ face : NULL  
## ..$ colour : NULL  
## ..$ size : num 14  
## ..$ hjust : NULL  
## ..$ vjust : NULL  
## ..$ angle : NULL  
## ..$ lineheight : NULL  
## ..$ margin : NULL  
## ..$ debug : NULL  
## ..$ inherit.blank: logi FALSE  
## ..- attr(\*, "class")= chr [1:2] "element\_text" "element"  
## $ axis.text :List of 11  
## ..$ family : NULL  
## ..$ face : NULL  
## ..$ colour : NULL  
## ..$ size : num 14  
## ..$ hjust : NULL  
## ..$ vjust : NULL  
## ..$ angle : NULL  
## ..$ lineheight : NULL  
## ..$ margin : NULL  
## ..$ debug : NULL  
## ..$ inherit.blank: logi FALSE  
## ..- attr(\*, "class")= chr [1:2] "element\_text" "element"  
## $ plot.title:List of 11  
## ..$ family : NULL  
## ..$ face : NULL  
## ..$ colour : NULL  
## ..$ size : num 16  
## ..$ hjust : num 0.5  
## ..$ vjust : NULL  
## ..$ angle : NULL  
## ..$ lineheight : NULL  
## ..$ margin : NULL  
## ..$ debug : NULL  
## ..$ inherit.blank: logi FALSE  
## ..- attr(\*, "class")= chr [1:2] "element\_text" "element"  
## - attr(\*, "class")= chr [1:2] "theme" "gg"  
## - attr(\*, "complete")= logi FALSE  
## - attr(\*, "validate")= logi TRUE

#only use the explanatory variables: income, education, and type   
# why not use independent variable women?  
# Correlation analysis for linear related variabels  
x %>%   
 dplyr::select(2,3,5)%>%  
 cor(method = "pearson", use = "complete.obs") %>%  
 round(digits=4)

## education income prestige  
## education 1.0000 0.5776 0.8502  
## income 0.5776 1.0000 0.7149  
## prestige 0.8502 0.7149 1.0000

Answer: Yes, only these variables are significant enough. According to the scatter diagram, there is no linear relationship between the proportion of professional women and the prestige score, as well as between census and prestige score. So the proportion of professional women and census will not be used as explanatory variables. Explanatory variables, including income, education and type, can be used for the rest of the problem. According to correlation analysis, there is a strong positive correlation between income, education and reputation. The correlation coefficient was 0.8502 and 0.7149, respectively. So it is reasonable to believe that adding education and income can provide a reasonable explanation ability.

## Question 3

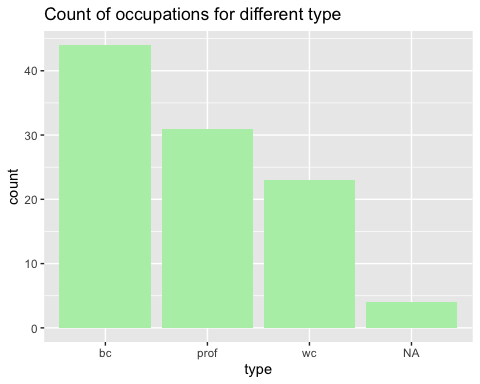
Which professions are missing “type”? Since the other variables for these observations are available, we could group them together as a fourth professional category to include them in the analysis. Is this advisable or should we remove them from our data set? Justify your answer.

Solution:

#Find the missing type  
x[which(is.na(x$type)),]

## # A tibble: 4 x 7  
## occupation.group education income women prestige census type   
## <chr> <dbl> <dbl> <dbl> <dbl> <dbl> <chr>  
## 1 athletes 11.4 8206 8.13 54.1 3373 <NA>   
## 2 newsboys 9.62 918 7 14.8 5143 <NA>   
## 3 babysitters 9.46 611 96.5 25.9 6147 <NA>   
## 4 farmers 6.84 3643 3.6 44.1 7112 <NA>

#check the frequency of each type  
x%>% ggplot(aes(x=type))+  
 geom\_bar(fill="darkseagreen2")+  
 ggtitle("Count of occupations for different type")



#Change value of NA to others  
x[which(is.na(x$type)),]$type<-"others"  
  
# linear regression model with the fourth type  
lm.1<-lm(prestige ~ education + income + type, data=x)  
  
# linear regression model without the fourth type  
lm.2<-lm(prestige ~ education + income + type, data=x[-which(x$type=="others"),])  
  
summary(lm.1)

##   
## Call:  
## lm(formula = prestige ~ education + income + type, data = x)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -18.0864 -4.8662 0.1436 5.3524 19.2652   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 1.7693934 5.3361820 0.332 0.7409   
## education 3.3059733 0.6537085 5.057 2.04e-06 \*\*\*  
## income 0.0011392 0.0002305 4.942 3.28e-06 \*\*\*  
## typeothers -1.7322264 4.0258430 -0.430 0.6680   
## typeprof 7.4877370 3.9698331 1.886 0.0623 .   
## typewc -1.7190573 2.6174653 -0.657 0.5129   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 7.512 on 96 degrees of freedom  
## Multiple R-squared: 0.8188, Adjusted R-squared: 0.8093   
## F-statistic: 86.75 on 5 and 96 DF, p-value: < 2.2e-16

summary(lm.2)

##   
## Call:  
## lm(formula = prestige ~ education + income + type, data = x[-which(x$type ==   
## "others"), ])  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -14.9529 -4.4486 0.1678 5.0566 18.6320   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -0.6229292 5.2275255 -0.119 0.905   
## education 3.6731661 0.6405016 5.735 1.21e-07 \*\*\*  
## income 0.0010132 0.0002209 4.586 1.40e-05 \*\*\*  
## typeprof 6.0389707 3.8668551 1.562 0.122   
## typewc -2.7372307 2.5139324 -1.089 0.279   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 7.095 on 93 degrees of freedom  
## Multiple R-squared: 0.8349, Adjusted R-squared: 0.8278   
## F-statistic: 117.5 on 4 and 93 DF, p-value: < 2.2e-16

Answer:

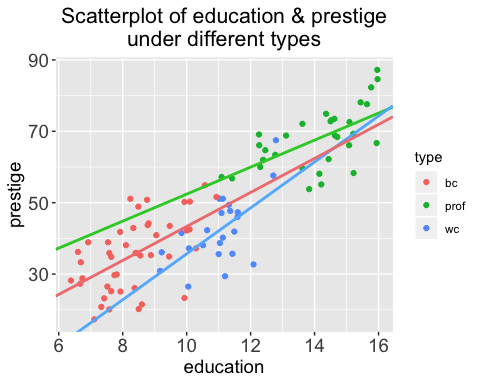
First of all, no-types occupations are athletes, newsboys, nannies, and farmers. From a practical point of view, these jobs do not have the same professional requirements, nor belong to the same occupational category. Therefore, there is no reason for us to combine these three professions.

Secondly, in 102 observations, the missing value is 4. Because the proportion of missing values is less than 5%, the four missing observations can be deleted. Finally, according to the results of the linear regression model with or without the fourth category, the fourth category has no statistical necessity (P > 0.05). Therefore, the lost data should be deleted from the dataset.

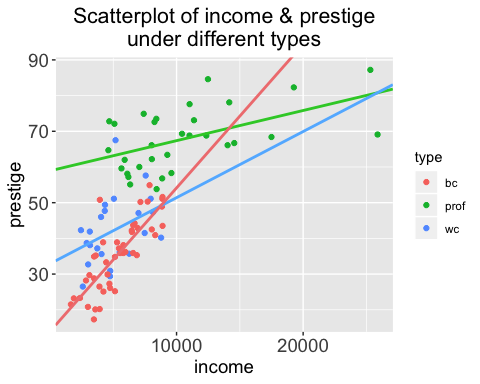
## Question 4

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

#Data cleaning  
#delete the missing value  
x.1<-x[-which(x$type=="others"),]  
#delete the unrelated attributes  
x.clean<-x.1[,c("education","income","type","prestige")]  
  
#linear regression for one variable and the inter-variable  
lm.edu<-lm(prestige ~ education+ type + education\*type, data=x.clean)  
lm.inc<-lm(prestige ~ income+ type + income\*type, data=x.clean)  
  
#Scatterplot between education and prestige  
x.clean %>% ggplot( aes(x=education,y=prestige,color=type))+  
 geom\_point()+  
 #prof  
 geom\_abline(slope = coef(lm.edu)[2]+coef(lm.edu)[5], intercept=coef(lm.edu)[1]+coef(lm.edu)[3], size=1, col="limegreen")+   
 #wc  
 geom\_abline(slope = coef(lm.edu)[2]+coef(lm.edu)[6], intercept=coef(lm.edu)[1]+coef(lm.edu)[4], size=1, col="steelblue1")+   
 #bc  
 geom\_abline(slope = coef(lm.edu)[2], intercept=coef(lm.edu)[1], size=1, col="lightcoral")+  
 ggtitle("Scatterplot of education & prestige\nunder different types")+  
 theme.info



#Scatterplot between income and prestige  
x.clean %>% ggplot( aes(x=income,y=prestige,color=type))+  
 geom\_point()+  
 ggtitle("Scatterplot of income & prestige\nunder different types")+  
#prof  
 geom\_abline(slope = coef(lm.inc)[2]+coef(lm.inc)[5], intercept=coef(lm.inc)[1]+coef(lm.inc)[3], size=1, col="limegreen")+   
 #wc  
 geom\_abline(slope = coef(lm.inc)[2]+coef(lm.inc)[6], intercept=coef(lm.inc)[1]+coef(lm.inc)[4], size=1, col="steelblue1")+   
 #bc  
 geom\_abline(slope = coef(lm.inc)[2], intercept=coef(lm.inc)[1], size=1, col="lightcoral")+  
 theme.info



Visually, since all the lines are not parallel, every line has a different slope, there should be an interaction between type and education as well as between income and type. Therefore, including interactive variables will provide more explanatory capability for the model.

#linear regression after considering interaction   
#education \* type  
lm.3<-lm(prestige ~ education + income + type +education\*type, data=x.clean)  
summary(lm.3)

##   
## Call:  
## lm(formula = prestige ~ education + income + type + education \*   
## type, data = x.clean)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -15.1168 -4.1751 0.4384 5.1625 15.2362   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -2.331e+00 7.783e+00 -0.299 0.765   
## education 3.852e+00 9.406e-01 4.096 9.12e-05 \*\*\*  
## income 1.052e-03 2.201e-04 4.782 6.66e-06 \*\*\*  
## typeprof 2.209e+01 1.520e+01 1.454 0.149   
## typewc -2.822e+01 1.959e+01 -1.440 0.153   
## education:typeprof -1.227e+00 1.304e+00 -0.941 0.349   
## education:typewc 2.270e+00 1.872e+00 1.213 0.228   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 7.036 on 91 degrees of freedom  
## Multiple R-squared: 0.8411, Adjusted R-squared: 0.8306   
## F-statistic: 80.27 on 6 and 91 DF, p-value: < 2.2e-16

# according to the p-value, since it of all education \* type > 0.05, this interacted variable is not statistically significant.  
  
#income \* type  
lm.4<-lm(prestige ~ education + income + type +income\*type, data=x.clean)  
summary(lm.4)

##   
## Call:  
## lm(formula = prestige ~ education + income + type + income \*   
## type, data = x.clean)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -12.8720 -4.8321 0.8534 4.1425 19.6710   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -6.7272633 4.9515480 -1.359 0.1776   
## education 3.0396961 0.6003699 5.063 2.14e-06 \*\*\*  
## income 0.0031344 0.0005215 6.010 3.79e-08 \*\*\*  
## typeprof 25.1723873 5.4669586 4.604 1.34e-05 \*\*\*  
## typewc 7.1375093 5.2898177 1.349 0.1806   
## income:typeprof -0.0025102 0.0005530 -4.539 1.72e-05 \*\*\*  
## income:typewc -0.0014856 0.0008720 -1.704 0.0919 .   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 6.455 on 91 degrees of freedom  
## Multiple R-squared: 0.8663, Adjusted R-squared: 0.8574   
## F-statistic: 98.23 on 6 and 91 DF, p-value: < 2.2e-16

# according to the p-value, since it of income\*typeprof < 0.05, this interacted variable is significant, therefore income \* type is statistically significant  
  
#(income+education)\*type  
lm.5<-lm(prestige ~ education + income + type + (income+education)\*type, data=x.clean)  
summary(lm.5)

##   
## Call:  
## lm(formula = prestige ~ education + income + type + (income +   
## education) \* type, data = x.clean)  
##   
## 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 .   
## 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 \*   
## education:typeprof 1.388e+00 1.289e+00 1.077 0.2844   
## education:typewc 4.291e+00 1.757e+00 2.442 0.0166 \*   
## ---  
## 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

# according to the p-value, since it of income\*typeprof,income\*typeprof and education:typewc < 0.05, they should be statistically significant, this model should be kept.

## 

## Question 5

Fit a model to predict prestige using: income, education, type, and any interaction terms based on your answer to question (4). Evaluate the model and include relevant output. Use your answer to question (3) to determine which observations to use in your analysis.

Answer: According to the answer of Question(4), the model used for prediction is lm.5<-lm(prestige ~ education + income + type + (income+education)*type, data=x.clean) #Checking the regression assumption:(skip)*

*#Model Evaluation: lm.5<-lm(prestige ~ education + income + type + (income+education)*type, data=x.clean)

lm.5

##   
## Call:  
## lm(formula = prestige ~ education + income + type + (income +   
## education) \* type, data = x.clean)  
##   
## Coefficients:  
## (Intercept) education income typeprof   
## 2.275753 1.713275 0.003522 15.351896   
## typewc income:typeprof income:typewc education:typeprof   
## -33.536652 -0.002903 -0.002072 1.387809   
## education:typewc   
## 4.290875

#summary of model  
summary(lm.5)

##   
## Call:  
## lm(formula = prestige ~ education + income + type + (income +   
## education) \* type, data = x.clean)  
##   
## 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 .   
## 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 \*   
## education:typeprof 1.388e+00 1.289e+00 1.077 0.2844   
## education:typewc 4.291e+00 1.757e+00 2.442 0.0166 \*   
## ---  
## 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

#1. RMSE  
summary(lm.5)$sigma

## [1] 6.318211

#The estimate variability of response variable around the regression line is 6.318211  
  
#2. R^2 and adjusted R^2  
r1<-lapply(summary(lm(prestige ~ education, data=x.clean))[c("r.squared", "adj.r.squared")], round, digits=4)  
r1<-rbind(r1,lapply(summary(lm(prestige ~ education+income, data=x.clean))[c("r.squared", "adj.r.squared")], round, digits=4))  
r1<-rbind(r1,lapply(summary(lm(prestige ~ education+income+type, data=x.clean))[c("r.squared", "adj.r.squared")], round, digits=4))  
r1<-rbind(r1,lapply(summary(lm(prestige ~ education+income+type+income\*type, data=x.clean))[c("r.squared", "adj.r.squared")], round, digits=4))  
r1<-rbind(r1,lapply(summary(lm.5)[c("r.squared", "adj.r.squared")], round, digits=4))  
r1<-cbind(r1,c("prestige vs. education","prestige vs. education+income","prestige vs. education+income+type","prestige vs. education+income+type+income\*type","prestige vs. education+income+type+(income+education)\*type"))  
r1

## r.squared adj.r.squared  
## r1 0.7508 0.7482   
## 0.814 0.8101   
## 0.8349 0.8278   
## 0.8663 0.8574   
## 0.8747 0.8634   
##   
## r1 "prestige vs. education"   
## "prestige vs. education+income"   
## "prestige vs. education+income+type"   
## "prestige vs. education+income+type+income\*type"   
## "prestige vs. education+income+type+(income+education)\*type"

The adjusted R square shows that adding the interact variable increases the accuracy of the model About 87.47% of the variability in prestige can be explained by a regression model containing education, income, type and (income+education)\*type.

#3 Overall F-test  
anova(lm.5)

## Analysis of Variance Table  
##   
## Response: prestige  
## Df Sum Sq Mean Sq F value Pr(>F)   
## education 1 21282.5 21282.5 533.1309 < 2.2e-16 \*\*\*  
## income 1 1792.0 1792.0 44.8892 1.828e-09 \*\*\*  
## type 2 591.2 295.6 7.4044 0.00106 \*\*   
## income:type 2 890.0 445.0 11.1476 4.783e-05 \*\*\*  
## education:type 2 238.4 119.2 2.9859 0.05557 .   
## Residuals 89 3552.9 39.9   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

a=0.05

hypotheses:

H0: ßeducation = ßincome = ßtype = ß(income+education)\*type=0

H1:at least one slope is not zero test statistic:

Fc = (21282.5+1792+591.2+890+238.4)/8/(3552.9/89) = 77.6364

with p=8 and n-(p+1) = 98-(8+1)=89 degrees of freedom

Since p-value of education: type is 0.05557>a=0.05 the null hypothesis cannot be rejected, the model is not adequate

#partial F-test for income\*type and (income+education)\*type  
anova(lm.4,lm.5)

## Analysis of Variance Table  
##   
## Model 1: prestige ~ education + income + type + income \* type  
## Model 2: prestige ~ education + income + type + (income + education) \*   
## type  
## Res.Df RSS Df Sum of Sq F Pr(>F)   
## 1 91 3791.3   
## 2 89 3552.9 2 238.4 2.9859 0.05557 .  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Since the p-value of partial F-test is larger than 0.05, the hypothesis zero cannot be rejected. Therefore, (income+education)\**type is not statistically significants. Only keep the interactive variable income*type is enough

Keep the model lm.4<-lm(prestige ~ education + income + type +income\*type, data=x.clean)

lm.4

##   
## Call:  
## lm(formula = prestige ~ education + income + type + income \*   
## type, data = x.clean)  
##   
## Coefficients:  
## (Intercept) education income typeprof   
## -6.727263 3.039696 0.003134 25.172387   
## typewc income:typeprof income:typewc   
## 7.137509 -0.002510 -0.001486

#summary of model  
summary(lm.4)

##   
## Call:  
## lm(formula = prestige ~ education + income + type + income \*   
## type, data = x.clean)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -12.8720 -4.8321 0.8534 4.1425 19.6710   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -6.7272633 4.9515480 -1.359 0.1776   
## education 3.0396961 0.6003699 5.063 2.14e-06 \*\*\*  
## income 0.0031344 0.0005215 6.010 3.79e-08 \*\*\*  
## typeprof 25.1723873 5.4669586 4.604 1.34e-05 \*\*\*  
## typewc 7.1375093 5.2898177 1.349 0.1806   
## income:typeprof -0.0025102 0.0005530 -4.539 1.72e-05 \*\*\*  
## income:typewc -0.0014856 0.0008720 -1.704 0.0919 .   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 6.455 on 91 degrees of freedom  
## Multiple R-squared: 0.8663, Adjusted R-squared: 0.8574   
## F-statistic: 98.23 on 6 and 91 DF, p-value: < 2.2e-16

#1. RMSE  
summary(lm.4)$sigma

## [1] 6.454624

#The estimate variability of response variable around the regression line is 6.454624  
  
#2. R^2 and adjusted R^2  
r1<-lapply(summary(lm(prestige ~ education, data=x.clean))[c("r.squared", "adj.r.squared")], round, digits=4)  
r1<-rbind(r1,lapply(summary(lm(prestige ~ income, data=x.clean))[c("r.squared", "adj.r.squared")], round, digits=4))  
r1<-rbind(r1,lapply(summary(lm(prestige ~ education+income, data=x.clean))[c("r.squared", "adj.r.squared")], round, digits=4))  
r1<-rbind(r1,lapply(summary(lm(prestige ~ education+income+type, data=x.clean))[c("r.squared", "adj.r.squared")], round, digits=4))  
r1<-rbind(r1,lapply(summary(lm.4)[c("r.squared", "adj.r.squared")], round, digits=4))  
r1<-cbind(r1,c("prestige vs. education","prestige vs. income","prestige vs. education+income","prestige vs. education+income+type","prestige vs. education+income+type+income\*type"))  
r1

## r.squared adj.r.squared   
## r1 0.7508 0.7482 "prestige vs. education"   
## 0.4946 0.4894 "prestige vs. income"   
## 0.814 0.8101 "prestige vs. education+income"   
## 0.8349 0.8278 "prestige vs. education+income+type"   
## 0.8663 0.8574 "prestige vs. education+income+type+income\*type"

About 86.63% of the variability in prestige can be explained by the regression model containing education, income, type and income\*type.

#3 Overall F-test  
anova(lm.4)

## Analysis of Variance Table  
##   
## Response: prestige  
## Df Sum Sq Mean Sq F value Pr(>F)   
## education 1 21282.5 21282.5 510.8344 < 2.2e-16 \*\*\*  
## income 1 1792.0 1792.0 43.0118 3.236e-09 \*\*\*  
## type 2 591.2 295.6 7.0947 0.00137 \*\*   
## income:type 2 890.0 445.0 10.6814 6.809e-05 \*\*\*  
## Residuals 91 3791.3 41.7   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

a=0.05

hypotheses:

H0:ßeducation = ßincome = ßtype = ßincome\*type = 0

H1:at least one slope is not zero test statistic: Fc=(21282.5+1792+591.2+890)/6/(3791.3/91)=98.2323

with p=6 and n-(p+1) = 98-(6+1)=91 degrees of freedom

Since p-value<a=0.05 the null hypothesis can be rejected, the model is adequate

#4 t-test

Hypotheis 1: H0:ßeducation=0 H1:ßeducation≠0

test statistic: t education = 5.063 Which has t distribution with n-(p+1)=98-(6+1)=91 degree of freedom

p-value< a=0.05 The null hypothesis is rejected, education is a statistically significant when others are already in the model

Hypotheis 2: H0:ßincome=0 H1:ßincome≠0

test statistic: t income = 6.010

Which has t distribution with n-(p+1)=98-(6+1)=91 degree of freedom

p-value< a=0.05 The null hypothesis is rejected, income is a statistically significant when others are already in the model

Hypotheis 3: H0:ßtypeprof=0 H1:ßtypeprof≠0

test statistic: t typeprof = 4.604 Which has t distribution with n-(p+1)=98-(6+1)=91 degree of freedom

p-value< a=0.05 The null hypothesis is rejected, typeprof is a statistically significant when others are already in the model

Hypotheis 4: H0:ßtypewc=0 H1:ßtypewc≠0

test statistic: t typewc = 1.349 Which has t distribution with n-(p+1)=98-(6+1)=91 degree of freedom

p-value> a=0.05 The null hypothesis cannot be rejected, typewc is not statistically significant when others are already in the model

Hypotheis 5: H0:ßincome:typeprof=0 H1:ßincome:typeprof≠0

test statistic: t income:typeprof = -4.539 Which has t distribution with n-(p+1)=98-(6+1)=91 degree of freedom

p-value< a=0.05 The null hypothesis is rejected, income:typeprof is a statistically significant when others are already in the model

Hypotheis 6: H0:ßincome:typewc=0 H1:ßincome:typewc≠0

test statistic: t education = -1.704 Which has t distribution with n-(p+1)=98-(6+1)=91 degree of freedom p-value > a=0.05

The null hypothesis cannot be rejected, income:typewc is not statistically significant when others are already in the model

#5 Partial F-test(with or without the interactive variable)  
anova(lm.2)

## Analysis of Variance Table  
##   
## Response: prestige  
## Df Sum Sq Mean Sq F value Pr(>F)   
## education 1 21282.5 21282.5 422.8056 < 2.2e-16 \*\*\*  
## income 1 1792.0 1792.0 35.5999 4.355e-08 \*\*\*  
## type 2 591.2 295.6 5.8721 0.003966 \*\*   
## Residuals 93 4681.3 50.3   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

anova(lm.2,lm.4)

## Analysis of Variance Table  
##   
## Model 1: prestige ~ education + income + type  
## Model 2: prestige ~ education + income + type + income \* type  
## Res.Df RSS Df Sum of Sq F Pr(>F)   
## 1 93 4681.3   
## 2 91 3791.3 2 890.02 10.681 6.809e-05 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

a=0.05

H0:ß=0 H1:ß≠0

test statistic: Fc = (4681.3-3791.3)/2/50.3 = 8.846918

p-value< a=0.05 The null hypothesis is rejected; the slope of interactive variable is statistically significant.

The final model is lm.4 <-lm(prestige ~ education + income + type +income\*type, data=x.clean) which is statistically significant and have enough explanatory capability.

## Question 6

Interpret the slopes and y-intercept in the context of the data for your model. Solution:

summary(lm.4)

##   
## Call:  
## lm(formula = prestige ~ education + income + type + income \*   
## type, data = x.clean)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -12.8720 -4.8321 0.8534 4.1425 19.6710   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -6.7272633 4.9515480 -1.359 0.1776   
## education 3.0396961 0.6003699 5.063 2.14e-06 \*\*\*  
## income 0.0031344 0.0005215 6.010 3.79e-08 \*\*\*  
## typeprof 25.1723873 5.4669586 4.604 1.34e-05 \*\*\*  
## typewc 7.1375093 5.2898177 1.349 0.1806   
## income:typeprof -0.0025102 0.0005530 -4.539 1.72e-05 \*\*\*  
## income:typewc -0.0014856 0.0008720 -1.704 0.0919 .   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 6.455 on 91 degrees of freedom  
## Multiple R-squared: 0.8663, Adjusted R-squared: 0.8574   
## F-statistic: 98.23 on 6 and 91 DF, p-value: < 2.2e-16

Model summary:

Whole model:

1. Regression line when type is prof:

Slope: If the job type of candidate is professional, every additional year of education will increase the prestige 3.0396961, and every additional dollar of income will increase prestige 0.0006242

Y-intercept: If the type of the candidate is professional, and he/she doesn’t have any education or income, he/she will still have 18.445124 in prestige.

1. Regression line when type is wc:   
   Slope: If the job type of candidate is a white collar , every additional year of education will increase the prestige 3.0396961, and every additional dollar of income will increase prestige 0.0016488

Y-intercept: If the type of the candidate is a white collar, and he/she doesn’t have any education or income, he/she will still have 0.410246 in prestige.

1. Regression line when type is bd:

Slope:If the job type of candidate is a blue collar, every additional year of education will increase the prestige 3.0396961, and every additional dollar of income will increase prestige 0.0031344

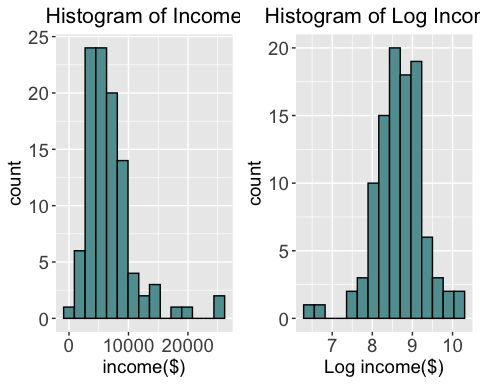
Y-intercept: If the job type of the candidate is a blue collar, and he/she doesn’t have any education or income, he/she will have -6.7272633 in prestige.

## Question 7

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

Solution:

h1 <- x %>% ggplot(aes(income)) +  
 geom\_histogram(bins=15, col="black", fill="cadetblue") +  
 ggtitle("Histogram of Income") +  
 labs(x="income($)") +  
 theme.info  
  
h2 <- x %>% ggplot(aes(log(income))) +  
 geom\_histogram(bins=15, col="black", fill="cadetblue") +  
 ggtitle("Histogram of Log Income") +  
 labs(x="Log income($)") +  
 theme.info  
  
grid.arrange(h1, h2,ncol=2)



The distribution changed from a slightly right-skewed to a more normally distribution.

## Question 8

Fit the model in question (5) but this time use log(income) (i.e., natural logarithm) instead of income. Evaluate the model and provide the relevant output.

Solution:

lm.4.log<-lm(prestige ~ education + log(income) + type +log(income)\*type, data=x.clean)  
lm.4.log

##   
## Call:  
## lm(formula = prestige ~ education + log(income) + type + log(income) \*   
## type, data = x.clean)  
##   
## Coefficients:  
## (Intercept) education log(income)   
## -118.432 3.211 14.934   
## typeprof typewc log(income):typeprof   
## 82.776 51.372 -8.569   
## log(income):typewc   
## -6.193

summary(lm.4.log)

##   
## Call:  
## lm(formula = prestige ~ education + log(income) + type + log(income) \*   
## type, data = x.clean)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -13.484 -4.453 1.122 4.123 18.737   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -118.4325 20.3728 -5.813 8.97e-08 \*\*\*  
## education 3.2107 0.5993 5.357 6.31e-07 \*\*\*  
## log(income) 14.9336 2.4928 5.991 4.12e-08 \*\*\*  
## typeprof 82.7757 31.5059 2.627 0.0101 \*   
## typewc 51.3717 36.8521 1.394 0.1667   
## log(income):typeprof -8.5690 3.5251 -2.431 0.0170 \*   
## log(income):typewc -6.1925 4.3172 -1.434 0.1549   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 6.491 on 91 degrees of freedom  
## Multiple R-squared: 0.8647, Adjusted R-squared: 0.8558   
## F-statistic: 96.96 on 6 and 91 DF, p-value: < 2.2e-16

lm.4.log<-lm(prestige ~ education + log(income) + type +log(income)\*type, data=x.clean)  
summary(lm.4.log)

##   
## Call:  
## lm(formula = prestige ~ education + log(income) + type + log(income) \*   
## type, data = x.clean)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -13.484 -4.453 1.122 4.123 18.737   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -118.4325 20.3728 -5.813 8.97e-08 \*\*\*  
## education 3.2107 0.5993 5.357 6.31e-07 \*\*\*  
## log(income) 14.9336 2.4928 5.991 4.12e-08 \*\*\*  
## typeprof 82.7757 31.5059 2.627 0.0101 \*   
## typewc 51.3717 36.8521 1.394 0.1667   
## log(income):typeprof -8.5690 3.5251 -2.431 0.0170 \*   
## log(income):typewc -6.1925 4.3172 -1.434 0.1549   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 6.491 on 91 degrees of freedom  
## Multiple R-squared: 0.8647, Adjusted R-squared: 0.8558   
## F-statistic: 96.96 on 6 and 91 DF, p-value: < 2.2e-16

#Evaluation

#summary of model  
summary(lm.4.log)

##   
## Call:  
## lm(formula = prestige ~ education + log(income) + type + log(income) \*   
## type, data = x.clean)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -13.484 -4.453 1.122 4.123 18.737   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -118.4325 20.3728 -5.813 8.97e-08 \*\*\*  
## education 3.2107 0.5993 5.357 6.31e-07 \*\*\*  
## log(income) 14.9336 2.4928 5.991 4.12e-08 \*\*\*  
## typeprof 82.7757 31.5059 2.627 0.0101 \*   
## typewc 51.3717 36.8521 1.394 0.1667   
## log(income):typeprof -8.5690 3.5251 -2.431 0.0170 \*   
## log(income):typewc -6.1925 4.3172 -1.434 0.1549   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 6.491 on 91 degrees of freedom  
## Multiple R-squared: 0.8647, Adjusted R-squared: 0.8558   
## F-statistic: 96.96 on 6 and 91 DF, p-value: < 2.2e-16

#1. RMSE  
summary(lm.4.log)$sigma

## [1] 6.49104

#The estimate variability of response variable around the regression line is 6.49104  
  
#2. R^2 and adjusted R^2  
r2<-lapply(summary(lm(prestige ~ log(income), data=x.clean))[c("r.squared", "adj.r.squared")], round, digits=4)  
r2<-rbind(r2,lapply(summary(lm(prestige ~ education+log(income), data=x.clean))[c("r.squared", "adj.r.squared")], round, digits=4))  
r2<-rbind(r2,lapply(summary(lm(prestige ~ education+log(income)+type, data=x.clean))[c("r.squared", "adj.r.squared")], round, digits=4))  
r2<-rbind(r2,lapply(summary(lm.4.log)[c("r.squared", "adj.r.squared")], round, digits=4))  
r2<-cbind(r2,c("prestige vs.Log income","prestige vs. education+Log income","prestige vs. education+Log income+type","prestige vs. education+Log income+type+Log income\*type"))  
r2

## r.squared adj.r.squared  
## r2 0.5644 0.5598   
## 0.8389 0.8356   
## 0.8555 0.8493   
## 0.8647 0.8558   
##   
## r2 "prestige vs.Log income"   
## "prestige vs. education+Log income"   
## "prestige vs. education+Log income+type"   
## "prestige vs. education+Log income+type+Log income\*type"

About 86.47% of the variability in prestige can be explained by a regression model containing education, log(income), type and log(income\*type) .

#3 Overall F-test  
anova(lm.4.log)

## Analysis of Variance Table  
##   
## Response: prestige  
## Df Sum Sq Mean Sq F value Pr(>F)   
## education 1 21282.5 21282.5 505.1188 < 2.2e-16 \*\*\*  
## log(income) 1 2499.1 2499.1 59.3125 1.583e-11 \*\*\*  
## type 2 469.1 234.5 5.5664 0.00524 \*\*   
## log(income):type 2 262.1 131.1 3.1107 0.04934 \*   
## Residuals 91 3834.2 42.1   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

a=0.05

hypotheses: H0:ßeducation = ßlog(income) = ßtype = ßlog(income)\*type = 0

H1:at least one slope is not zero test statistic: Fc=

(21282.5+2499.1+469.1+262.1)/6/(3834/91)

## [1] 96.96856

with p=6 and n-(p+1) = 98-(6+1)=91 degrees of freedom

since p-value<a=0.05 the null hypothesis can be rejected, the model is adequate

#4 T-test  
summary(lm.4.log)

##   
## Call:  
## lm(formula = prestige ~ education + log(income) + type + log(income) \*   
## type, data = x.clean)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -13.484 -4.453 1.122 4.123 18.737   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -118.4325 20.3728 -5.813 8.97e-08 \*\*\*  
## education 3.2107 0.5993 5.357 6.31e-07 \*\*\*  
## log(income) 14.9336 2.4928 5.991 4.12e-08 \*\*\*  
## typeprof 82.7757 31.5059 2.627 0.0101 \*   
## typewc 51.3717 36.8521 1.394 0.1667   
## log(income):typeprof -8.5690 3.5251 -2.431 0.0170 \*   
## log(income):typewc -6.1925 4.3172 -1.434 0.1549   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 6.491 on 91 degrees of freedom  
## Multiple R-squared: 0.8647, Adjusted R-squared: 0.8558   
## F-statistic: 96.96 on 6 and 91 DF, p-value: < 2.2e-16

a=0.05

Hypotheis 1: H0:ßeducation=0 H1:ßeducation≠0

test statistic: t education = 5.357 Which has t distribution with n-(p+1)=98-(6+1)=91 degree of freedom

p-value< a=0.05 The null hypothesis is rejected, education is a statistically significant when others are already in the model

Hypotheis 2: H0:ßlog(income)=0 H1:ßlog(income)≠0

test statistic: t log(income) = 5.991 Which has t distribution with n-(p+1)=98-(6+1)=91 degree of freedom

p-value< a=0.05 The null hypothesis is rejected, log(income) is a statistically significant when others are already in the model

Hypotheis 3: H0:ßtypeprof=0 H1:ßtypeprof≠0

test statistic: t typeprof = 2.627 Which has t distribution with n-(p+1)=98-(6+1)=91 degree of freedom

p-value< a=0.05 The null hypothesis is rejected, typeprof is a statistically significant when others are already in the model

Hypotheis 4: H0:ßtypewc=0 H1:ßtypewc≠0

test statistic: t typewc = 1.394 Which has t distribution with n-(p+1)=98-(6+1)=91 degree of freedom

p-value> a=0.05 The null hypothesis cannot be rejected, typewc is not statistically significant when others are already in the model

Hypotheis 5: H0:ßlog(income):typeprof=0 H1:ßlog(income):typeprof≠0

test statistic: t log(income):typeprof = -2.431 Which has t distribution with n-(p+1)=98-(6+1)=91 degree of freedom

p-value< a=0.05 The null hypothesis is rejected, log(income):typeprof is a statistically significant when others are already in the model

Hypotheis 6: H0:ßlog(income):typewc=0 H1:ßlog(income):typewc≠0

test statistic: t log(income):typewc = -1.434 Which has t distribution with n-(p+1)=98-(6+1)=91 degree of freedom

p-value > a=0.05 The null hypothesis cannot be rejected, log(income):typewc is not statistically significant when others are already in the model

## Question 9

Is the model in question (5) or (8) better? Justify your answer. Why can’t we use a partial F-test here?

Solution:

#RMSE of the model lm.4 and lm.4.log  
summary(lm.4)$sigma

## [1] 6.454624

summary(lm.4.log)$sigma

## [1] 6.49104

By comparing the RMSE of both model, since it of the model lm.4 is lower, lm.4 is better.

#Comparing the R^2  
lapply(summary(lm.4)[c("r.squared", "adj.r.squared")], round, digits=4)

## $r.squared  
## [1] 0.8663  
##   
## $adj.r.squared  
## [1] 0.8574

lapply(summary(lm.4.log)[c("r.squared", "adj.r.squared")], round, digits=4)

## $r.squared  
## [1] 0.8647  
##   
## $adj.r.squared  
## [1] 0.8558

By comparing, the lm.4 has higher accuracy, therefore it is better.

The partial F test cannot be used since lm.4 and lm.4.log are not nested model.The partial F-test can only be used when the same observations are used in both models. And it is used to determine whether the extra variables provide enough extra explanatory power as a group. In other words, the partial F-test tests whether the full model is significantly better than the reduced model.