Project: Statistical Analysis of Prestige Dataset

Submitted in partial fulfillment of the requirements of the course

BANA 7031 Probability Models

By

Adrian Valles

Piyush Verma

MS Business Analytics Candidates 2018

Department of Operations, Business Analytics & Information Systems

Carl H Lindner School of Business: University of Cincinnati

Abstract

Prestige dataset contains information on 102 Canadian occupations from 1971. Following are the factors present:

Education	Average education of occupational incumbents
Income	Percentage of incumbents who are women
Women	Percentage of incumbents who are women
Prestige	Pineo-Porter prestige score for occupation, from a social survey conducted in the mid-1960s.
Census	Canadian Census occupational code
Type	Type of occupation. A factor with levels (note: out of order): bc, Blue Collar; prof, Professional, Managerial, and Technical; wc, White Collar

For the scope of this project we are going to ignore *Census*, as it is only an occupational code and doesn't add any valuable information to the dataset for any kind of analysis.

1. Introduction

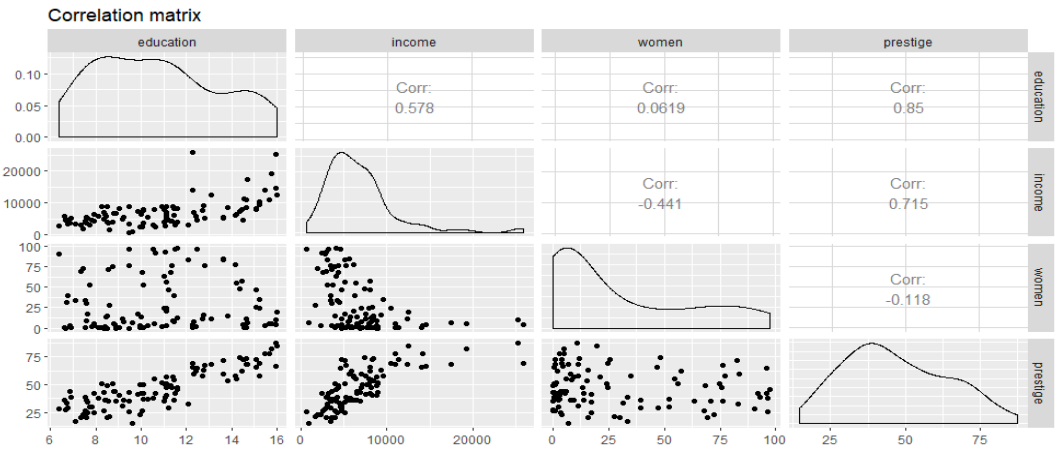
In this project we would be exploring the Prestige dataset specially focusing on the variables prestige and income levels of the occupations, which they have a positive correlation. For instance, occupations like physicians and university teachers have high prestige with higher income. For having a high-level picture of the dataset, we categorized the dataset into various income and prestige levels as shown in Table 1.1. Below contingency table depict this idea more clearly:

Table 1.1: Contingency table comparing Prestige level vs Income level

Levels	Low prestige	Mid prestige	High prestige
Low income	21	10	3
Mid income	13	10	10
High income	0	7	28

The above table shows that occupations with higher income have predominantly higher prestige. But there are indeed few occupations which have higher prestige score but still have low income levels (ministers, nurses and electronic workers). So, a higher prestige can't guarantee a higher income.

Figure 1.1 Correlation matrix showing how the 4 numerical variables are correlated



From the above figure, we can say that income seems to have strong positive association with prestige (r = 0.715) but also have some mild association with education score (r = 0.578) and a weak negative association with women's proportion (r = -0.441).

Therefore, in our analysis in this project we will try to:

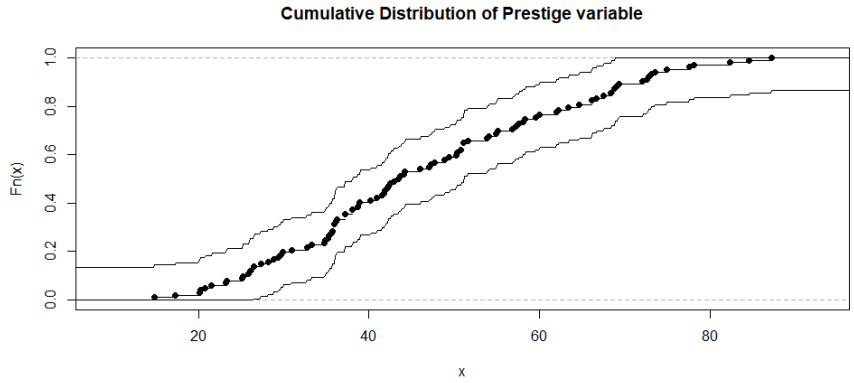
a. Establish a relationship between the income levels and other predictor variables
b. Test a few hypotheses to better understand how the data relates

2. Analysis

2.1 Investigation using Empirical Cumulative Distribution

Prestige: Empirical cumulative distribution function was plotted for the prestige variable along with a 95% confidence band.

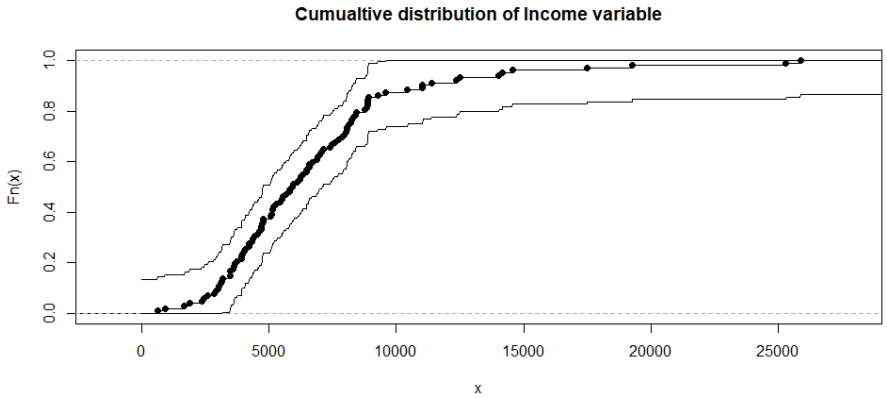
Figure 2.1 Cumulative distribution of Prestige variable



Using the ecdf function in R, it was found that an estimate 20% of the occupations in Canada are considered highly prestigious. Since it is only an estimate we also calculated a 95% confidence interval for this proportion. Thus, the true population proportion which can have prestige scores between 60 – 80 can be anywhere between: 12.74% – 28.44% with 95% confidence. Some of these occupations are architects, physicist and psychologists.

Income: Empirical cumulative distribution function was plotted for the income variable along with a 95% confidence band.

Figure 2.2 Cumulative distribution of income variable



It was found that 50% of the population was living with income between $5000 & $10,000. But since it's an estimate only, we calculated a 95% confidence band. Thus, the true population proportion which was living between income $5000 and $10,000 can be anywhere between 40.2% & 59.7%. In figure 2.3 we can observe that income is left skewed while prestige seems to be normally distributed.

Figure 2.3 Figure comparing distribution of income and prestige

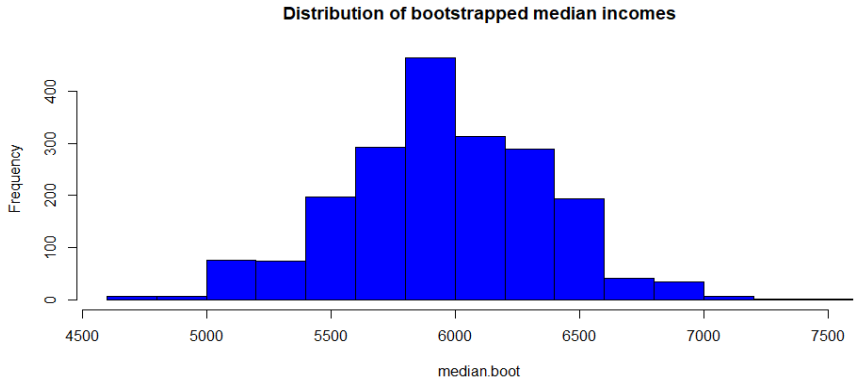
2.2 Investigation using Bootstrap

Median income: Since median income can be a good indicator of what the population in the middle was earning, we bootstrapped for the median income and calculated 95% confidence intervals from 3 different methods. Following were the results:

Table 2.1: 95% confidence interval for median income from different methods

Method	2.5% ile	97.5 % ile
Normal	$5,104.59	$6,756.40
Pivotal	$5,174.33	$6,727.00
Quantile	$5,134.00	$6,686.66

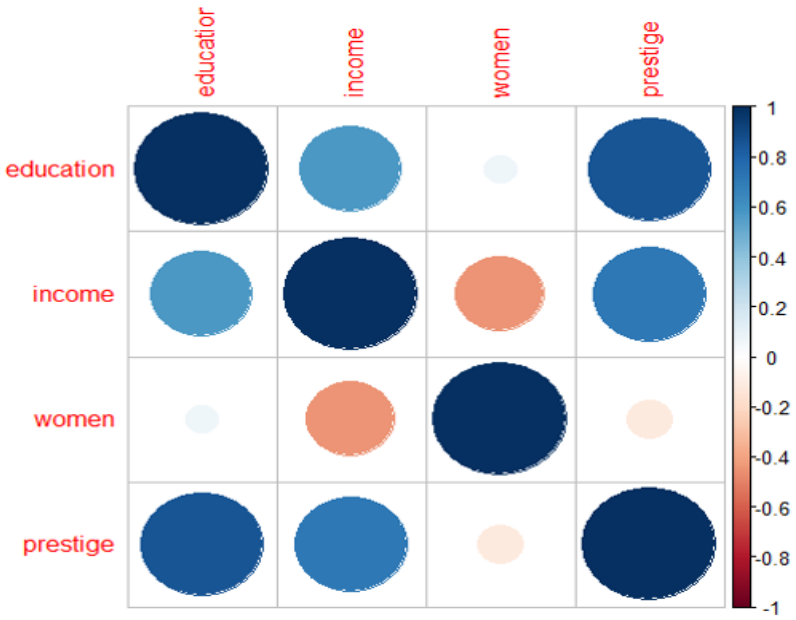
Figure 2.4 Below is the distribution of median incomes calculated from bootstrap



Median income estimate is $5,930.50 and bootstrap standard error is $408.64.

Correlation: Figure below shows the correlation among different numerical variables.

Figure 2.5 Correlation matrix among all numeric variables



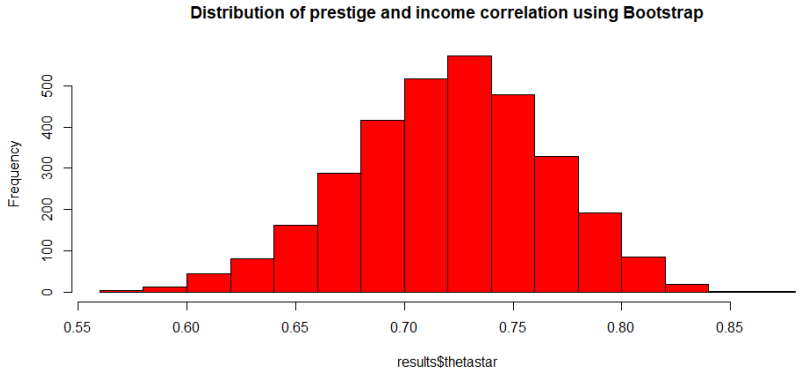
From figure 2.5 we can clearly see that income is associated with prestige score, education level and women proportion, in that order. But also notice that education is also highly correlated with prestige, meaning education and prestige can replace each other for analysis like regression, since a higher education generally brings in higher income. Since, prestige is highly correlated with income, we tried next to estimate this correlation with a 95% confidence interval. Following were the results from bootstrap:

Table 2.2: 95% confidence interval for correlation coefficient from different methods

Method	2.5% ile	97.5 % ile
Normal	0.62	0.81
Pivotal	0.63	0.80
Quantile	0.63	0.80

Median income estimate is $5,930.50 and bootstrap standard error is $408.64. Below is the distribution of prestige and income correlation numbers calculated from bootstrap (next page).

Figure 2.6 Distribution of bootstrapped prestige and income correlation



Regression coefficients

Before bootstrapping for regression coefficients, one must come up with a model with properties like lower residual standard errors, higher R squared and higher F-statistic value which make a regression model significant. We started with the following model:

Model A: Income = education + women + prestige

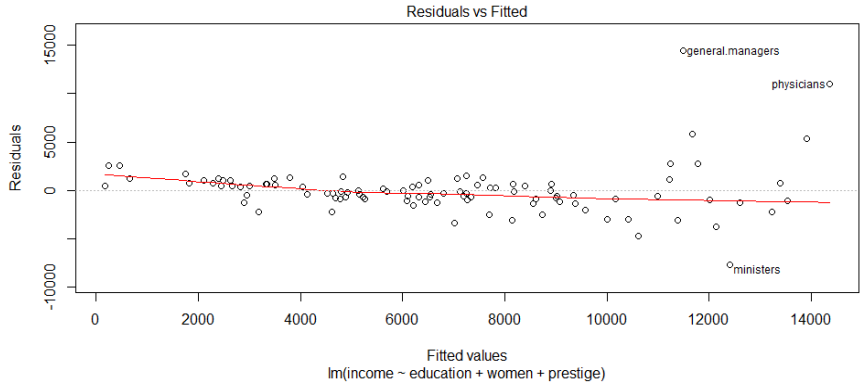
Following was the summary of the model A:

```
Call:
lm(formula = income ~ education + women + prestige, data = Prestige)

Residuals:
Min     1Q  Median     3Q    Max
-7715.3  -929.7  -231.2   689.7 14391.8

Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept) -253.850   1086.157  -0.234    0.816
education    177.199    187.632   0.944    0.347
women        -50.896      8.556  -5.948 4.19e-08 ***
prestige     141.435     29.910   4.729 7.58e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2575 on 98 degrees of freedom
Multiple R-squared:  0.6432,    Adjusted R-squared:  0.6323
F-statistic: 58.89 on 3 and 98 DF,  p-value: < 2.2e-16
```

We can observe that above Model A is not a good fit because of the following reasons:

a. Larger standard residual error (2575)
b. Violation of assumption of constant error variance

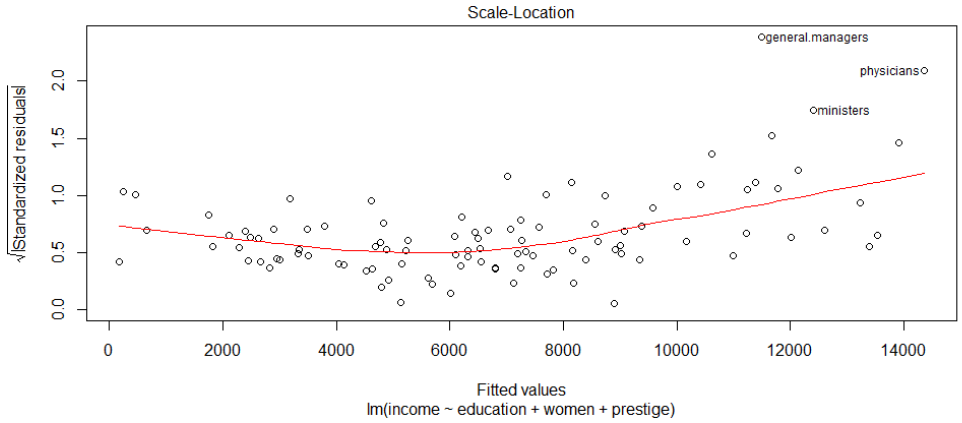
Figure 2.7 Residuals vs Fitted values: Constant variance assumption violated



So, we improvised our model and introduced a new model with:

a. Log transformation of income variable (because from Figure 2.3 we know that it is right skewed and is a big range).
b. Removal of insignificant variable education (education has good correlation with income but must be removed because it is also correlated to predictor prestige) and may inflate the residuals.
c. Adding square term for prestige because we found that a square term might had been missing by looking at residual vs fitted value diagnostics plot.

Figure 2.8 Quadratic term might have been missing from model A



So, we introduced a new Model B as follows:

Model B: log(Income) = women + prestige + prestige^2

Below (next page) are the summary results of model B. We can see that the residual standard errors and R-squared have been improved significantly along with an increase in F-statistic value, making our model B very significant when compared to model A.

```
Call:
lm(formula = log(income) ~ women + prestige + I(prestige^2),
   data = Prestige)

Residuals:
    Min      1Q  Median      3Q     Max
-1.12302 -0.09650  0.02764  0.13913  0.78303

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   7.286e+00  2.218e-01  32.851  < 2e-16 ***
women        -8.365e-03  9.376e-04  -8.922 2.64e-14 ***
prestige      4.701e-02  9.479e-03   4.959 2.97e-06 ***
I(prestige^2) -2.351e-04  9.392e-05  -2.503    0.014 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2963 on 98 degrees of freedom
Multiple R-squared:  0.7565,    Adjusted R-squared:  0.7491
F-statistic: 101.5 on 3 and 98 DF,  p-value: < 2.2e-16
```

So, we finalized our model B for estimating the regression coefficients using bootstrap. Following are the distribution of the coefficients for women and prestige from non-parametric bootstrap.

Figure 2.9 Bootstrap results: Distribution of women and prestige coefficients

Below table shows the 95% bootstrap confidence interval and the bootstrap estimate calculated for women and prestige coefficients:

Table 2.3: 95% confidence interval for regression coefficients using Bootstrap

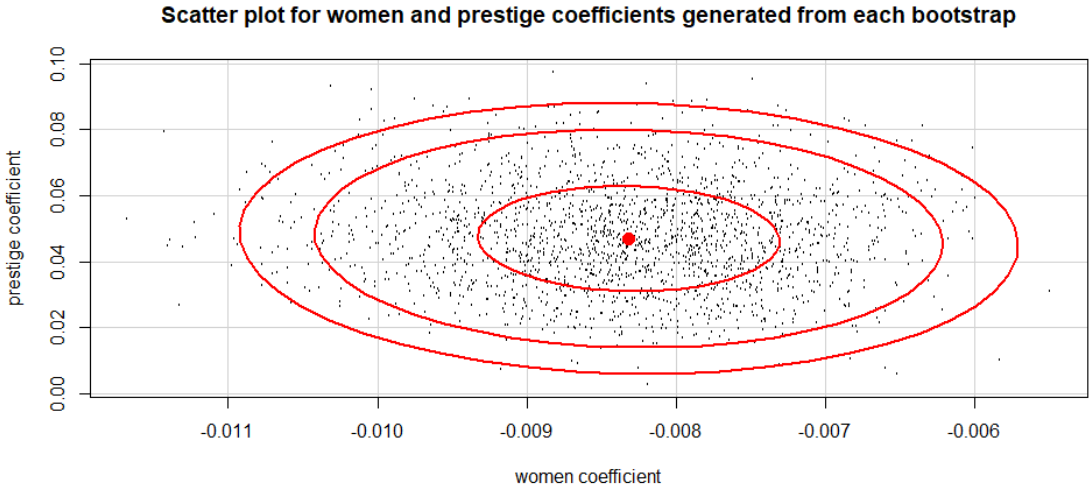
Coefficients	2.5%ile	Bootstrap estimate	97.5%ile
Women	-0.010	-0.008	-0.006
Prestige	0.016	0.048	0.077

From the above bootstrap coefficients, we can say that with every 1% increase in women proportion in the occupation, the income may decrease by an amount of exp(-0.008) = $0.99. Similarly, an increase of unit prestige score, the income increases by exp(0.048) = $ 1.04. Also since 0 is not in the 95% confidence interval for both, we can say that any difference in women proportion or prestige score will bring out some change in income level.

Below figure 2.10 is another way of visualizing the paired bootstrap coefficients. They are telling us:

a. how much positively the prestige score can affect the income, with 95% confidence
b. how much negatively a women's proportion can affect the income, with 95% confidence

Figure 2.10 Bootstrap results: Scatter plot of women and prestige coefficients



The outer most red line indicates 99%, middle 95% and innermost 5% confidence interval. As almost all of the coefficients for women lie on the -ve side, we can hypothesize that:

- Occupations with more women proportion (>50%) must have significantly different income level than occupations which have lesser women (<50%)

We would like to finish the analysis by digging on this topic: Discrimination of women in the workplace. For long time, women have had considerably lower salaries than men, and even if this difference is shortening, it is still a problem. So, was this a big issue in Canada in 1971? We will therefore analyze through hypothesis testing whether the difference in salaries between the jobs where women account for more than 50% of the workforce and the jobs where women account for less than 50% was women significant. For instance, 83.78% of the primary school teachers were women. We would use Wald Test as well as the permutation test to analyze this.

Wald test: Ho: μ1 = μ2 vs h1: μ1 not equal to μ2, where μ1 represents the average salary of the jobs with a majority of the women in the workforce and μ2 those with majority of men.

```
x1_hat <- mean(prestige_women$income )
x2_hat <- mean(prestige_man$income )

v1 <- var(prestige_women$income)/length(prestige_women$income)
v2 <- var(prestige_man$income)/length(prestige_man$income)

z <- (x1_hat-x2_hat)/sqrt(v1 + v2)
z
pnorm(z)
```

Z = (3940.26 – 7826.65)/ sqrt (56.62+59.23) = -6.57 which corresponds to a p-value very close to 0. Consequently, we can reject the null hypothesis and conclude that there is a significan ce difference in the salaries.

Permutation Test: Even if such a low p-value represented very strong evidence against Ho, we wanted to run a different test in order to support the significance difference in salaries. By running a permutation test, we obtain a p-value of 0, which suggests one more time the significant difference between the salaries of the jobs divided by gender workforce.

```
prestigeByWomen <- Prestige[order(Prestige$women, decreasing = TRUE),]
require(gtools)
 B = 1000
perm.matrix=replicate(B, sample(length(prestigeByWomen$income)))
data.vector<-prestigeByWomen$income
perm.T<-apply(perm.matrix, 1,function(x) {abs(mean(data.vector[x[1:27]])-mean(data.vector[x[28:102]]))})
p.value=mean(perm.T>abs(mean(data.vector[1:27])-mean(data.vector[28:102])))
p.value
```

3. Conclusions

Through the study of 102 different occupations in Canada during 1971, we have been able to find relevant insights. First, using a 95% C.I, (12.74, 28.44) % of the occupations were considered highly prestigious and (40.2, 59.7)% of the occupations reported salaries between $5,000 & $10,000 (with a median income ($5,104.59, $6736.3)95% C.I). Also, high levels of correlation have been found between Income and Prestige (0.62,0.81) 95%C.I. Related to this, the following linear regression model shows high income prediction capabilities:

Log(Income) = 7.29-0.008*women + 0.047 Prestige –0.0002 Prestige ^2

Finally, using hypothesis testing, a significant difference between the salaries of the jobs with predominant female workforce and the salaries of the jobs with predominant male workforce has been found.

4. Bibliography: Source of the dataset

Canada (1971) *Census of Canada*. Vol. 3, Part 6. Statistics Canada [pp. 19-1–19-21].

Personal communication from B. Blishen, W. Carroll, and C. Moore, Departments of Sociology, York University and University of Victoria.

5. Appendix:

R code used for the statistical analysis.

```
##### probability project

install.packages("car")
library("car")  # to get the package
install.packages("GGally")
library("GGally")


View(Prestige)
?Prestige   # run this to understand the dataset and variales


Prestige$Occupation<-rownames(Prestige)
ggpairs(Prestige[,c(1:4)], title = "Correlation matrix")

library("dplyr")  #Creating contingency tables revealing which type of jobs fetch higher income
Prestige<-Prestige %>% mutate(income_cat=cut(income,breaks = c(0,4690,7456,25879.00),labels = c("low_in","mid_in","high_in")))
Prestige$income_cat<-as.factor(Prestige$income_cat)

Prestige<-Prestige %>% mutate(prestige_cat=cut(prestige,breaks = c(0,36.53,50.26,87.200),labels = c("low_pr","mid_pr","high_pr")))
Prestige$prestige_cat<-as.factor(Prestige$prestige_cat)


table(Prestige$type,Prestige$income_cat)
table(Prestige$type,Prestige$prestige_cat)
table(Prestige$income_cat,Prestige$prestige_cat)
```

```
Prestige <- Prestige[,1:4]  # I have decided to delete the last 2 columns to make our work easier...

dim(Prestige)
str(Prestige)



# variable prestige
######## empirical cdf of the variable Prestige
prestige.prestige.ecdf <- ecdf(Prestige$prestige)
plot(prestige.prestige.ecdf, main = "Cumulative Distribution of Prestige variable")

### confidence band
Alpha=0.05
n=length(Prestige$prestige)
Eps=sqrt(log(2/Alpha)/(2*n))
grid<-seq(0,150, length.out = 1000)
lines(grid, pmin(prestige.prestige.ecdf(grid)+Eps,1))
lines(grid, pmax(prestige.prestige.ecdf(grid)-Eps,0))

### calculating what percentage of values have a prestige between 60 and 80
estimate.prestige <- prestige.prestige.ecdf(80)-prestige.prestige.ecdf(60)
estimate.prestige
## confidence interval for the estimate
right <- estimate.prestige + qnorm(0.975)*sqrt(estimate.prestige*(1-estimate.prestige) / n)
left <- estimate.prestige- qnorm(0.975)*sqrt(estimate.prestige*(1-estimate.prestige) / n)
c(left, right)


par(mfrow=c(1,2))
hist(Prestige$income, breaks = seq(0,26800.00,by = 2000), col = "blue")
hist(Prestige$prestige, breaks = seq(0,88,by = 11), col = "red")

################################################ median
```

```
############################################## median

## median of income (AVG income in dollars in 1971)
median <- median(Prestige$income)

## bootstrap
library(bootstrap)
B <- 2000

median.boot<-replicate(B, median(Prestige$income[sample(1:n,size=n, replace=T)]))
hist(median.boot, col = "blue", main = "Distribution of bootstrapped median incomes")

var(median.boot)
se.boot <- sd(median.boot)

### confidence intervals
# method 1
normal.ci<-c(median-2*se.boot, median+2*se.boot)
#method 2
pivatol.ci<-c(2*median-quantile(median.boot,0.975), 2*median-quantile(median.boot,0.025))
#method 3
quantile.ci<-quantile(median.boot, c(0.025, 0.975))
normal.ci
pivatol.ci
quantile.ci
```

```
##############################################correlation
install.packages("corrplot")
library(corrplot)
M <- cor(Prestige[,1:4])
corrplot(M, method = "circle")
## correlation between prestige and income

cor.hat <- cor(Prestige$prestige, Prestige$income)
cor.hat

### confidence interaval for the correlation using bootsrap
library(bootstrap)# load the package

n <- length(Prestige$prestige)
theta <- function(x,xdata){ cor(xdata[x,2],xdata[x,4]) }

results <- bootstrap(1:n,3200,theta, Prestige)
hist(results$thetastar, col = "red", main = "Distribution of prestige and income correlation using Bootstrap")

se.boot <- sqrt(var(results$thetastar))
cor.hat <- cor(Prestige$prestige, Prestige$income)

normal.ci <- c(cor.hat-2*se.boot, cor.hat+2*se.boot)
pivatol.ci <- c(2*cor.hat-quantile(results$thetastar,0.975), 2*cor.hat-quantile(results$thetastar,0.025))
quantile.ci <- quantile(results$thetastar, c(0.025, 0.975))

normal.ci
pivatol.ci
quantile.ci
```

```
#####################################women##  2017_12_06
#Bootstrapping of Regression Coefficients
library("boot")
library("car")
#MODEL 1
fit1<-lm(income ~ education + women + prestige,data = Prestige)
summary(fit1)
par(mfrow=c(2,2))
plot(fit1)


#Education is coming to be insignificant which doesn't make sense because income is highly dependent on education level
#Possible reason is the correlation between education and prestige which is 0.8501769

#Fitting 2nd model
#Below, we are removing Education, doing log transformation of income and adding square term for prestige
#Because residual plot suggested a missing square term in the model fit1
#MODEL2: With improved residual standard error, Multiple R-squared and F statistics
fit2<-lm(log(income) ~ women + prestige + I(prestige^2),data = Prestige)
summary(fit2)
par(mfrow=c(2,2))
plot(fit2)
```

```r
#Bootstrapping 2000 times for getting 95% confidence interval for regression coefficients
boot.fun<-function(data,indices,maxit){
  data<-data[indices,]
  fit.boot<-lm(log(income) ~ women + prestige + I(prestige^2),data=data)
  coefficients(fit.boot)
}
#Boot object
Prestige.boot<-boot(Prestige,boot.fun,2000,maxit = 100)
Prestige.boot

#Distribution of bootstrap coefficients for woemn and prestige
par(mfrow = c(1,2))
hist(Prestige.boot$t[,2], main = "Distribution of bootstrapepd women coefficients", col = "blue", xlab = "women coefficient")
hist(Prestige.boot$t[,3], main = "Distribution of bootstrapepd prestige coefficients", col = "red", xlab = "prestige coefficient")

#Calculating 95% CI for coefficients of women and prestige from above done bootstrap
women.boot.CI<-boot.ci(Prestige.boot, index=2, type=c("norm"), conf = 0.95)# women coef.
women.boot.estimate<-mean(Prestige.boot$t[,2]) # women coef. estimate

prestige.boot.CI<-boot.ci(Prestige.boot, index=3, type=c("norm"), conf = 0.95)# prestige coef.
prestige.boot.estimate<-mean(Prestige.boot$t[,3]) # prestige coef. estimate

CI<-data.frame(rbind(women.boot.CI$normal,prestige.boot.CI$normal))
colnames(CI)<-c("Conf","2.5%_ile","97.5%_ile")
CI

#Scatter plot of coefficients of women and prestige from each bootstrap
dataEllipse(Prestige.boot$t[,2], Prestige.boot$t[,3],
xlab="women coefficient", ylab="prestige coefficient",
cex=.3, levels=c(.5, .95, .99), robust=T, main = "Scatter plot for women and prestige coefficients generated from each bootstrap")

#Above scatter plot of coefficients reveals that the true popualtion fixed coefficients beta1 and beta0
#should lie somehwere inside the cluster marked by red lines
#where each line is a confidence level indicator
```

```r
##### wald tests and permutation test

## I want to know if there is a significance difference in salary between top jobs where more than 50% are women and the jo


median(Prestige$prestige) # even number of observations so median is (n+(n+1))/2..so in each top and bottom 50 we have 51 o
prestige_women <- Prestige[Prestige$women > 50,] # jobs where there are more woman
prestige_man <- Prestige[Prestige$women < 50,]

## test: ho mu1 = mu 2

x1_hat <- mean(prestige_women$income )
x2_hat <- mean(prestige_man$income )

v1 <- var(prestige_women$income)/length(prestige_women$income)
v2 <- var(prestige_man$income)/length(prestige_man$income)

z <- (x1_hat-x2_hat)/sqrt(v1 + v2)
z
pnorm(z)
# which corresponds to a p-value very close to 0.... and therefore we can conclude theat there is a significance difference
View(Prestige)



### permutation test...just to make sure both tests get the same results

prestigeByWomen <- Prestige[order(Prestige$women, decreasing = TRUE),]
require(gtools)
B = 1000
perm.matrix=replicate(B, sample(length(prestigeByWomen$income)))
data.vector<-prestigeByWomen$income
perm.T<-apply(perm.matrix, 1,function(x) {abs(mean(data.vector[x[1:27]])-mean(data.vector[x[28:102]]))})
p.value=mean(perm.T>abs(mean(data.vector[1:27])-mean(data.vector[28:102])))
p.value  # same result....there is a significant difference
```