# Data Assignment 2

Shraddha Ganesh

2025-01-15

```
data = read_dta("C:/Users/shrad/OneDrive/Desktop/3_LASI_W1_Individual_v4.dta")
```

**DATA CLEANING**

1. Create a new age variable with two categories: individuals younger than 60 years (represented by 1) and individuals aged 60 years or older (represented by 2).
2. Creating an educational attainment variable with four categories: no schooling (represented by 1); less than 5 years (represented by 2); 5–9 years (represented by 3); 10 or more years of schooling (represented by 4).
3. Create a variable for ”Current working status” with the following categories: currently working (represented by 1); worked in the past but currently not working (represented by 2); never worked (represented by 3.

```
data$age = ifelse(data$dm005<60,1,
           ifelse(data$dm005>=60,2,NA))

table(data$age)
```

```
## 
##     1     2 
## 41494 31902
```

```
data$educational_attainment = ifelse(data$dm007==0,1,
                              ifelse(data$dm007>=1 & data$dm007<5,2,
                              ifelse(data$dm007>=5 & data$dm007<=9,3,
                              ifelse(data$dm007>=10,4,
                              ifelse(is.na(data$dm007),NA,NA)))))

table(data$educational_attainment)
```

```
## 
##     2     3     4 
##  8188 17192 14254
```

```
data$current_working_status = ifelse(data$we001==2,1,
                              ifelse(data$we004==1,2,
                              ifelse(data$we004==2,3,NA)))

table(data$current_working_status)
```

```
## 
##     1     2     3 
## 21948 33280 18096
```

**SUMMARY STATISTICS**

```
a=round(sum(data$age==1,na.rm=TRUE)/length(data$age),2)
b=round(sum(data$age==2,na.rm=TRUE)/length(data$age),2)

c=round(sum(data$educational_attainment==1,na.rm=TRUE)/
          length(data$educational_attainment),2)
d=round(sum(data$educational_attainment==2,na.rm=TRUE)/
          length(data$educational_attainment),2)
e=round(sum(data$educational_attainment==3,na.rm=TRUE)/
          length(data$educational_attainment),2)
f=round(sum(data$educational_attainment==4,na.rm=TRUE)/
          length(data$educational_attainment),2)

g=round(sum(data$current_working_status==1,na.rm=TRUE)/
          length(data$current_working_status),2)
h=round(sum(data$current_working_status==1,na.rm=TRUE)/
          length(data$current_working_status),2)
i=round(sum(data$current_working_status==1,na.rm=TRUE)/
          length(data$current_working_status),2)

j=round(sum(data$dm003==1,na.rm=TRUE)/length(data$dm003),2)
k=round(sum(data$dm003==2,na.rm=TRUE)/length(data$dm003),2)

l=round(sum(data$dm021==1,na.rm=TRUE)/length(data$dm021),2)
m=round(sum(data$dm021==2,na.rm=TRUE)/length(data$dm021),2)
n=round(sum(data$dm021>=3 & data$dm021 <=7,na.rm=TRUE)/length(data$dm021),2)

o=round(sum(data$hc102==1,na.rm=TRUE)/length(data$hc102),2)
p=round(sum(data$hc102==2,na.rm=TRUE)/length(data$hc102),2)

variable = c("Proportion of People Under the Age of 60",
             "Proportion of People 60 Years and Older",
             "Proportion of People with No Schooling",
             "Proportion of People with <5 Years of Schooling",
             "Proportion of People with 5-9 Years of Schooling",
             "Proportion of People with 10 or more Years of Schooling",
             "Proportion of those Currently Working",
             "Proportion of those who worked but no more work",
             "Proportion of those who never worked",
             "Proportion of Men",
             "Proportion of Women",
             "Proportion of People Currently Married",
             "Proportion of People Widowed",
             "Proportion of People Divorced/Seperated/Deserted/Others",
             "Proportion of People With Health Insurance",
             "Proportion of People Without Health Insurance")

data1 = c(a,b,c,d,e,f,g,h,i,j,k,l,m,n,o,p)

summary_stats_table = data.frame(variable,data1)

colnames(summary_stats_table) = c("Variable Name", "Statistic")

summary_stats_table
```

```
##                                               Variable Name Statistic
## 1                Proportion of People Under the Age of 60      0.57
## 2                 Proportion of People 60 Years and Older      0.43
## 3                  Proportion of People with No Schooling      0.00
## 4         Proportion of People with <5 Years of Schooling      0.11
## 5        Proportion of People with 5-9 Years of Schooling      0.23
## 6  Proportion of People with 10 or more Years of Schooling      0.19
## 7                   Proportion of those Currently Working      0.30
## 8         Proportion of those who worked but no more work      0.30
## 9                    Proportion of those who never worked      0.30
## 10                                      Proportion of Men      0.42
## 11                                    Proportion of Women      0.58
## 12                 Proportion of People Currently Married      0.77
## 13                           Proportion of People Widowed      0.20
## 14 Proportion of People Divorced/Seperated/Deserted/Others      0.03
## 15             Proportion of People With Health Insurance      0.23
## 16          Proportion of People Without Health Insurance      0.76
```

### DOCUMENTING PATTERNS IN HEALTH CARE UTILIZATION AND FINANCING

### PART 1

1. Proportion of Government-Provided Insurance Holders.
2. Proportion of Employer-Provided Insurance Holders.
3. Proportion of Private Insurance Holders.

```
#PART 1.



government_provided = sum(data$hc103s1==1,na.rm=TRUE)+
  sum(data$hc103s2==1,na.rm=TRUE)+
  sum(data$hc103s3==1,na.rm=TRUE)+
  sum(data$hc103s4==1,na.rm=TRUE)+
  sum(data$hc103s5==1,na.rm=TRUE)
round(government_provided/length(data$hc103s1),3)
```

```
## [1] 0.224
```

```
employer_provided = sum(data$hc103s7==1,na.rm=TRUE)+
  sum(data$hc103s8==1,na.rm=TRUE)
round(employer_provided/length(data$hc103s1),3)
```

```
## [1] 0.005
```

```
private_insurance = sum(data$hc103s9==1,na.rm=TRUE)
round(private_insurance/length(data$hc103s1),3)
```

```
## [1] 0.01
```

### PART 2

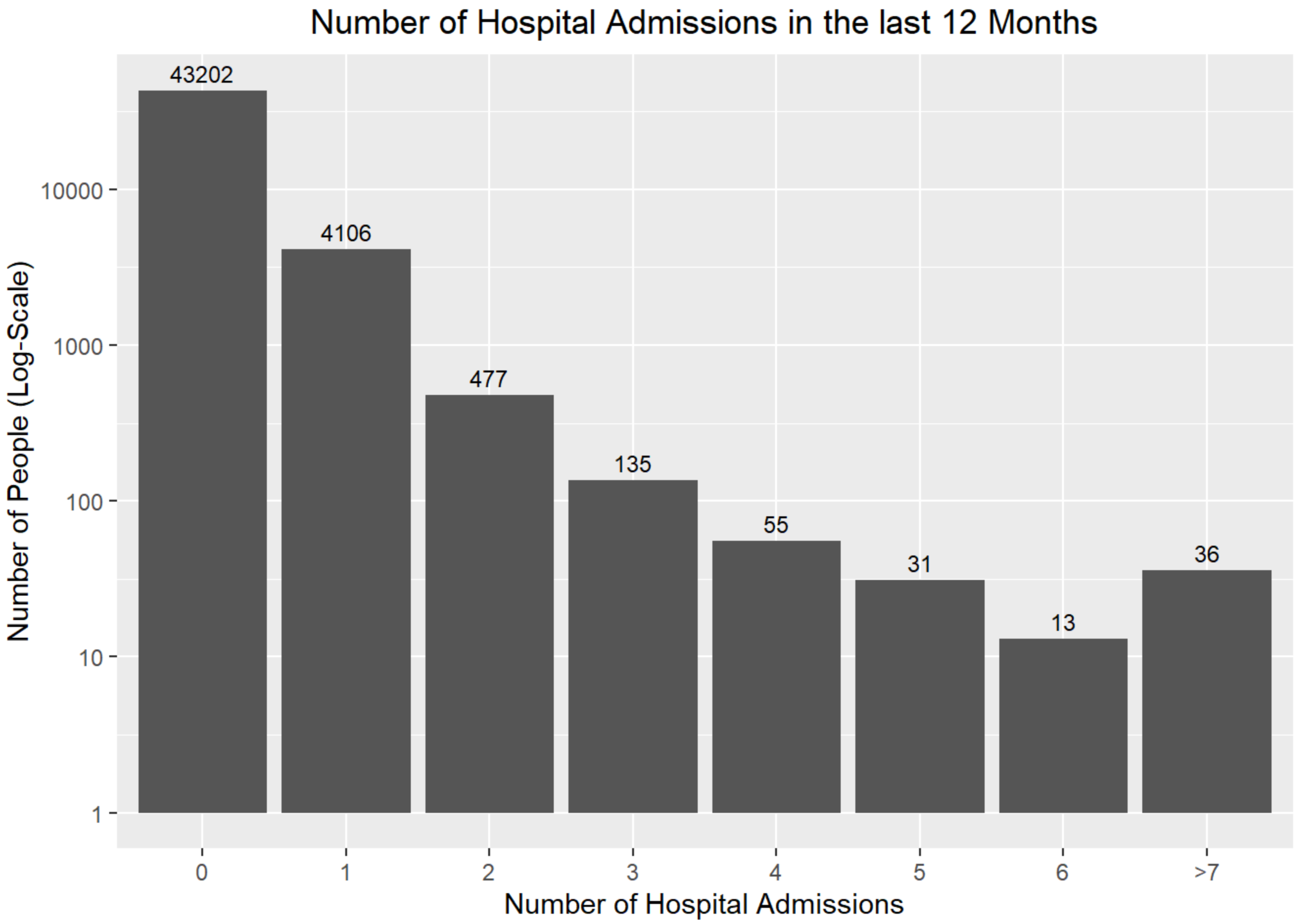
Plotting the frequency distribution of hospital admissions in the last 12 months.

```
q=sum(data$hc202==0,na.rm=TRUE)
r=sum(data$hc202==1,na.rm=TRUE)
s=sum(data$hc202==2,na.rm=TRUE)
t=sum(data$hc202==3,na.rm=TRUE)
u=sum(data$hc202==4,na.rm=TRUE)
v=sum(data$hc202==5,na.rm=TRUE)
w=sum(data$hc202==6,na.rm=TRUE)
x=sum(data$hc202>=7 & data$hc202<=24,na.rm=TRUE)

number = c(0,1,2,3,4,5,6,">7")
frequency = c(q,r,s,t,u,v,w,x)

new = data.frame(number,frequency)
colnames(new)=c("Number_of_Visits","Frequency")
new$Number_of_Visits <- factor(new$Number_of_Visits, levels =
                                 c("0", "1", "2", "3", "4", "5", "6", ">7"))

ggplot(new, aes(x=Number_of_Visits, y=Frequency))+
  geom_col()+
  labs(
    title="Number of Hospital Admissions in the last 12 Months",
    x = "Number of Hospital Admissions",
    y = "Number of People (Log-Scale)")+
  theme(plot.title = element_text(hjust=0.5))+
  scale_y_log10()+
  geom_text(aes(label = Frequency), vjust = -0.5, size = 3)
```



**PART 3**

1. Proportion of those cared for by spouse.

2. Proportion of those cared for by children, and grandchildren.

```
round(sum(data$hc212==1,na.rm=TRUE)/length(data$hc212),2)
```

```
## [1] 0.03
```

```
son = sum(data$hc212==2,na.rm=TRUE)
daughter = sum(data$hc212==3,na.rm=TRUE)
grandchild = sum(data$hc212==6,na.rm=TRUE)
total = son+daughter+grandchild
round(total/length(data$hc212),2)
```

```
## [1] 0.03
```

**PART 4**

Calculating total expenditure on the last outpatient visit.

```
data$total_expenditure = data$hc309a_1+data$hc309a_2+data$hc309a_3+
                         data$hc309a_4+data$hc309a_5+data$hc309a_6+
                         data$hc309a_7+data$hc309a_8+data$hc309a_9

summary(data$total_expenditure)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's 
##       0     500    4450    4245    6993  276000   48312
```

**PART 5**

1. t-test 1: outpatient expenditure and health insurance.
2. t-test 2: outpatient expenditure and gender.

```
insured = data$total_expenditure[data$hc102==1]
uninsured = data$total_expenditure[data$hc102==2]

t_test_insurance = t.test(insured,uninsured,
                          alternative="two.sided", var.equal = FALSE)
print(t_test_insurance)
```

```
## 
##  Welch Two Sample t-test
## 
## data:  insured and uninsured
## t = 8.1531, df = 13338, p-value = 3.861e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  381.4137 622.8563
## sample estimates:
## mean of x mean of y 
##  4629.366  4127.231
```

```
men = data$total_expenditure[data$dm003==1]
women = data$total_expenditure[data$dm003==2]

t_test_gender = t.test(men,women,
                       alternative="two.sided", var.equal = FALSE)
print(t_test_gender)
```

```
##
##  Welch Two Sample t-test
##
## data:  men and women
## t = 1.2208, df = 17541, p-value = 0.2222
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -49.10684 211.29328
## sample estimates:
## mean of x mean of y
##  4293.874  4212.780
```

**INTERPRETATION:**

For t-test1: The t-test reveals a statistically significant difference in outpatient care spending between those with health insurance and those without, with a p-value of 3. 861e-16, which is considerably less than the significance level of 0. 05. This indicates strong evidence to reject the null hypothesis of no difference.

For t-test2: The t-test indicates that there is no statistically significant difference in outpatient care expenditure between men and women, with a p-value of 0.2222, which is greater than the typical significance level of 0.05.

**REGRESSION ANALYSIS**

**PART 1** Regressing hospital admission on total outpatient expenditure.

```
data_new = data %>%
  filter(!is.na(hc202),
         !is.na(total_expenditure),
         total_expenditure > 0,
         !is.na(age),
         !is.na(dm003),
         !is.na(educational_attainment),
         !is.na(dm021),
         !is.na(hc102),
         !is.na(current_working_status))

regression1 = lm(log(total_expenditure) ~ hc202 + age + dm003 +
                   educational_attainment + dm021 + hc102 +
                   current_working_status, data = data_new)

summary(regression1)
```

```
##
## Call:
## lm(formula = log(total_expenditure) ~ hc202 + age + dm003 + educational_attainment +
##     dm021 + hc102 + current_working_status, data = data_new)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.9461 -1.0629  0.7434  1.0951  4.7077
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)             8.023239   0.116744  68.725  < 2e-16 ***
## hc202                   0.158742   0.031239   5.082 3.80e-07 ***
## age                     0.046007   0.029323   1.569  0.11668
## dm003                   0.095484   0.031016   3.079  0.00208 **
## educational_attainment  0.008659   0.018604   0.465  0.64165
## dm021                  -0.050838   0.016062  -3.165  0.00155 **
## hc102                  -0.208188   0.031290  -6.654 2.98e-11 ***
## current_working_status -0.038826   0.021538  -1.803  0.07147 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.518 on 12677 degrees of freedom
## Multiple R-squared:  0.007646,   Adjusted R-squared:  0.007098
## F-statistic: 13.95 on 7 and 12677 DF,  p-value: < 2.2e-16
```

**INTERPRETATION:**

Controlling for age, gender, education, marital status, insurance coverage, and work status, hospital admissions has a significant positive impact on outpatient cost. Insurance coverage correlates with reduced outpatient expenses. Gender and marital status have a substantial impact on spending habits, with males and married persons often spending more. The other factors were not significant.

**PART 2** Repeating the same with an interaction variable between the number of hospital visits and an indicator for insurance status.

```
data_clean = data %>%
  filter(!is.na(hc202),
         !is.na(total_expenditure),
         total_expenditure > 0,
         !is.na(age),
         !is.na(dm003),
         !is.na(educational_attainment),
         !is.na(dm021),
         !is.na(hc102),
         !is.na(current_working_status))

regression2 = lm(log(total_expenditure) ~ hc202 * hc102 + age + dm003 +
                   educational_attainment + dm021 + current_working_status,
                        data = data_clean)

summary(regression2)
```

```
##
## Call:
## lm(formula = log(total_expenditure) ~ hc202 * hc102 + age + dm003 +
##     educational_attainment + dm021 + current_working_status,
##     data = data_clean)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.9485 -1.0615  0.7445  1.0952  4.7052
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)             8.025700   0.117533  68.285  < 2e-16 ***
## hc202                   0.132473   0.148097   0.895  0.37107
## hc102                  -0.209651   0.032313  -6.488 9.01e-11 ***
## age                     0.046067   0.029326   1.571  0.11624
## dm003                   0.095469   0.031017   3.078  0.00209 **
## educational_attainment  0.008673   0.018605   0.466  0.64111
## dm021                  -0.050839   0.016063  -3.165  0.00155 **
## current_working_status -0.038787   0.021540  -1.801  0.07178 .
## hc202:hc102             0.014492   0.079863   0.181  0.85601
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.519 on 12676 degrees of freedom
## Multiple R-squared:  0.007648,   Adjusted R-squared:  0.007022
## F-statistic: 12.21 on 8 and 12676 DF,  p-value: < 2.2e-16
```

**INTERPRETATION:**

The interaction between hospital admissions and insurance status is not statistically significant, indicating that the impact of hospital admissions on outpatient expenditure does not vary by insurance coverage. Individuals with insurance, however, tend to spend less on outpatient care.