Assignment 2

David Gyaraki, Thao Le

Contents

```r
# load packages
if(!require(pacman)){install.packages("pacman")}

p_load(devtools,tidyverse,dplyr,ggplot2,latex2exp,
       sampleSelection, quantreg)
```

```
  There is a binary version available but the source version is later:
         binary source needs_compilation
devtools  2.4.4  2.4.5             FALSE
```

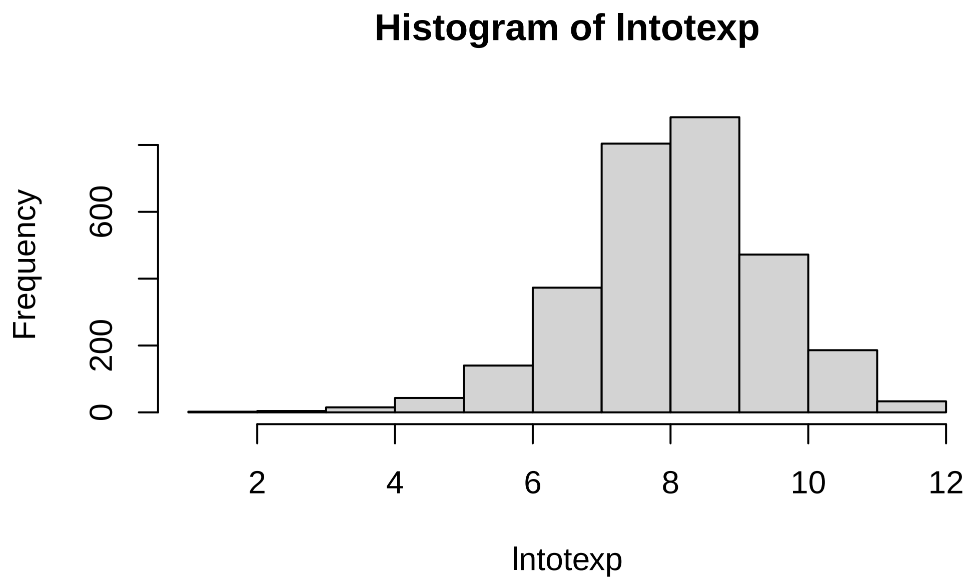
```r
#load data
dfData = read.csv("assignment2a_2023.csv")
attach(dfData)
```

1 Question 1

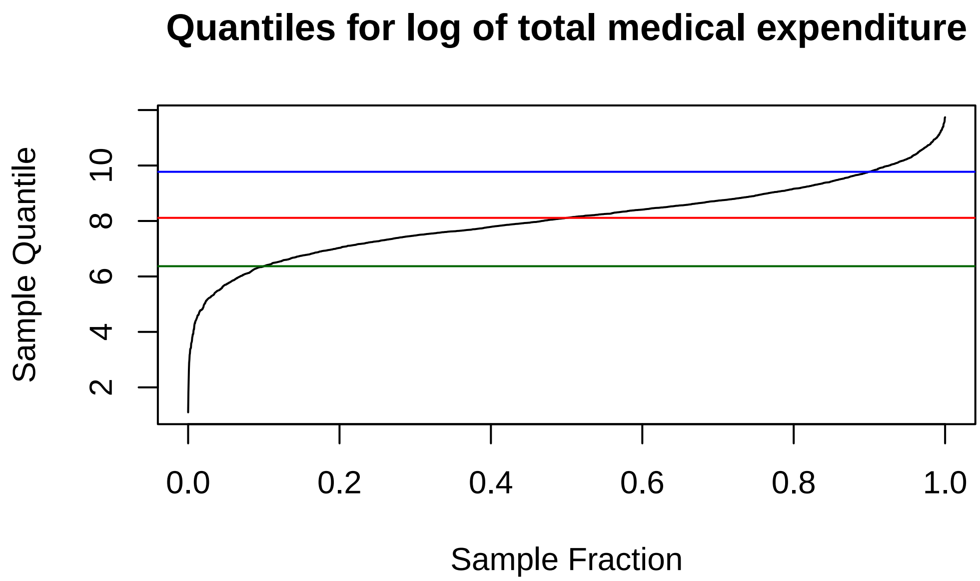
1.1 (i)

```r
# Get the quantile values
quant=quantile(lntotexp, seq(0.1, 0.9, by=.4))

# Histogram of log of total medical expenditure
hist(lntotexp)
```




```r
# Quantile plot of log of total medical expenditure
n = length(lntotexp)
plot((1:n - 1)/(n - 1), sort(lntotexp), type="l",
main = "Quantiles for log of total medical expenditure",
xlab = "Sample Fraction",
ylab = "Sample Quantile")
abline(h=quant, col = c("dark green","red", "blue"))
```



In the quantile plot, the median is indicated by the red line, the 10^{th} and 90^{th} quantile are indicated by the blue and green lines.

We can see from the distribution of log of total medical expenditure that there are few values from 0 to 4. Thus, the quantile plot increases quickly in this region. From 4 to 6, we see an increase frequencies of observations, thus, the quantile plot increases slower. The most rapid increase in the quantile plot is observed between 6 and 10, which makes sense because that is the region where most observations lie. After 10, there are less observations and the quantile plot increases rapidly again.

Although the quantile plot increases rapidly in both regions from 0 to 4 and 10 to 12, we observed a much steeper increase from 0 to 4, thus, we can say that the distribution of log total medical expenditure is left-skewed. This is confirmed by looking at its histogram.

1.2 (ii)

```r
# Quantile regression
q= c(0.1,0.25,0.5,0.75,0.9)
quant_reg = rq(lntotexp ~ . , tau = q, data = dfData)
summary(quant_reg)
```

```
Call: rq(formula = lntotexp ~ ., tau = q, data = dfData)

tau: [1] 0.1

Coefficients:
            Value    Std. Error t value  Pr(>|t|)
(Intercept)  3.86704  0.48065    8.04549  0.00000
age          0.01927  0.00601    3.20732  0.00135
female      -0.01273  0.07579   -0.16794  0.86664
white        0.07344  0.19533    0.37597  0.70697
totchr       0.53919  0.02534   21.27920  0.00000
suppins      0.39572  0.07851    5.04027  0.00000

Call: rq(formula = lntotexp ~ ., tau = q, data = dfData)

tau: [1] 0.25

Coefficients:
            Value    Std. Error t value  Pr(>|t|)
(Intercept)  4.74732  0.30724   15.45160  0.00000
age          0.01551  0.00399    3.88410  0.00010
female      -0.01623  0.05328   -0.30462  0.76068
white        0.33775  0.09662    3.49570  0.00048
totchr       0.45918  0.01833   25.04804  0.00000
suppins      0.38584  0.05992    6.43964  0.00000

Call: rq(formula = lntotexp ~ ., tau = q, data = dfData)

tau: [1] 0.5

Coefficients:
            Value    Std. Error t value  Pr(>|t|)
(Intercept)  5.61116  0.35187   15.94656  0.00000
age          0.01487  0.00406    3.66512  0.00025
female      -0.08810  0.05406   -1.62961  0.10329
white        0.53648  0.19319    2.77697  0.00552
totchr       0.39427  0.01846   21.35942  0.00000
suppins      0.27698  0.05347    5.18025  0.00000

Call: rq(formula = lntotexp ~ ., tau = q, data = dfData)

tau: [1] 0.75
```

```
Coefficients:
            Value    Std. Error t value  Pr(>|t|)
(Intercept)  6.59997  0.42690   15.46027  0.00000
age          0.01825  0.00475    3.83862  0.00013
female      -0.12194  0.06060   -2.01231  0.04428
white        0.19319  0.25684    0.75219  0.45200
totchr       0.37354  0.02286   16.33884  0.00000
suppins      0.14885  0.06203    2.39991  0.01646

Call: rq(formula = lntotexp ~ ., tau = q, data = dfData)

tau: [1] 0.9

Coefficients:
            Value    Std. Error t value  Pr(>|t|)
(Intercept)  8.32264  0.54599   15.24309  0.00000
age          0.00592  0.00651    0.91022  0.36278
female      -0.15763  0.08914   -1.76831  0.07711
white        0.30522  0.24260    1.25811  0.20845
totchr       0.35795  0.03310   10.81289  0.00000
suppins     -0.01428  0.08642   -0.16527  0.86874
```

Looking at the results, we observe different coefficients across the different quantiles. Quite expectedly, we have increasing intercept coefficients, however the interesting part is the different significance of the coefficients in the different quantile regressions. When one looks at the OLS regression results,

```r
# OLS Regression
OLS_reg = lm(lntotexp ~ . , data = dfData)
summary(OLS_reg)
```

```
Call:
lm(formula = lntotexp ~ ., data = dfData)

Residuals:
    Min      1Q  Median      3Q     Max
-6.2474 -0.7666 -0.0032  0.7827  3.8516

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  5.898155   0.295694  19.947  < 2e-16 ***
```

```
age          0.012656   0.003595   3.520 0.000437 ***
female      -0.076517   0.046110  -1.659 0.097132 .
white        0.317811   0.141360   2.248 0.024635 *
totchr       0.445272   0.017549  25.374  < 2e-16 ***
suppins      0.256811   0.046450   5.529 3.51e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.227 on 2949 degrees of freedom
Multiple R-squared:  0.1969,    Adjusted R-squared:  0.1955
F-statistic: 144.6 on 5 and 2949 DF,  p-value: < 2.2e-16
```