Assignment 2

David Gyaraki, Thao Le

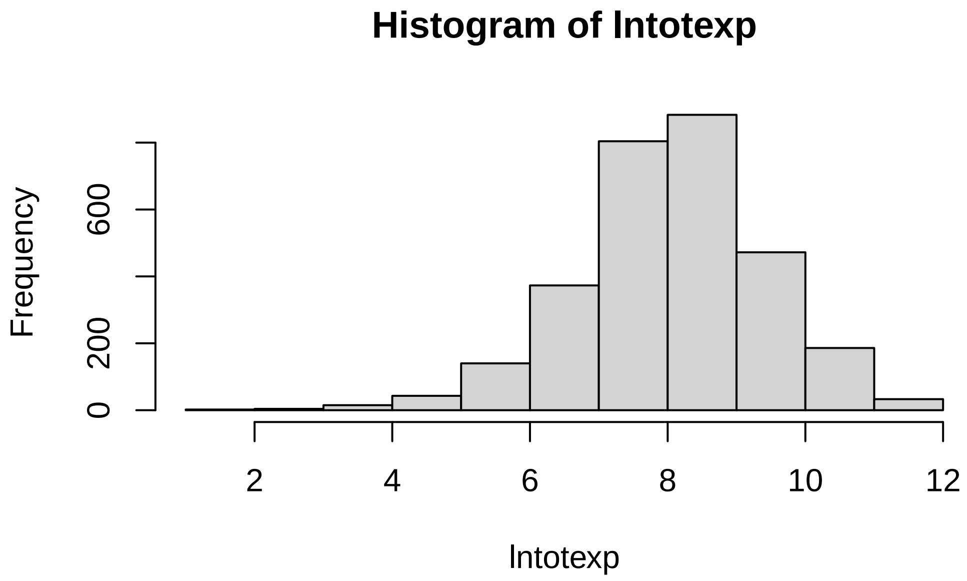
Contents

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

p_load(devtools,tidyverse,dplyr,ggplot2,latex2exp,
       sampleSelection, quantreg, plm, nlme)

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

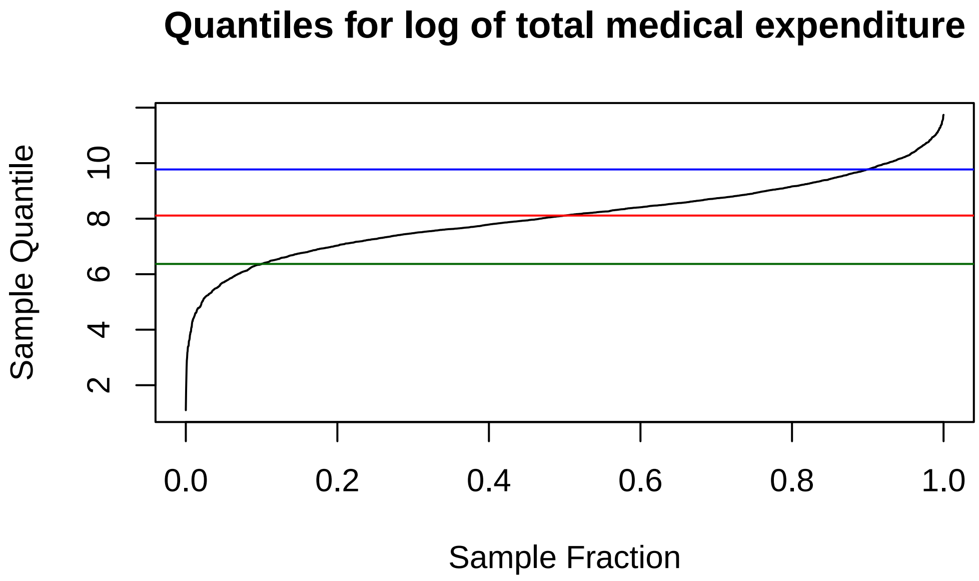
1 Question 1

1.1 (i)

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

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



```
# Quantile plot of log of total medical expenditure
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"))
```



```
integer(0)
```

In the quantile plot, the median is indicated by the red line, the 10th and 90th 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 in frequencies of observations, thus, the quantile plot increases slower. The slowest 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)

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

```r
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. We observe that for the 0.1 quantile, the female and white dummies are insignificant, for the 0.25 and 0.5 quantiles only the female dummy is insignificant, for the 0.75, interestingly the white dummy is insignificant while the female dummy turns out to be significant, and for the 0.9 quantile, only the chronic illness variable seems to be strongly significant with the female dummy slightly (at 10% level) significant too. These trends will lead to the conclusion that the different predictors likely have different dynamics across the groups of patients when ordered by medical expenditure. Being white significantly increases medical expenditure in the mid-groups but not in the tails of the expenditure distribution. Age and extra insurance are associated with significant increase in costs for low spending groups but not for the highest spenders, and gender comes into influence for the highest spenders only. Let us then look at the OLS results, coefficients and their significance levels.

```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
```

When one looks at the OLS regression results, the model shows that most variables are statistically significant for explaining the logarithm of medical expenditure, except for the female dummy variable. The variables *age*, *totchr* and *suppins* all have positive effect on medical expenditure with less than 0.001 significance, and the variable *white* has a positive effect as well on 5% significance level. The interpretation of the coefficients can also be given as one unit increase in the independent variables (keeping all else equal) increases the medical expenditure by $(exp(\beta_k) - 1) * 100)$ percentage. We can see below, that a year of age increase will result in an estimated 1.274% increase in medical expenses. Similarly, being female reduces the expenses by -7.366% (although this is only significant at 10% level in the OLS model), being white is associated with 37.412% increase in medical expenses, an additional chronic illness will increase expenditure by 56.091% and having a supplementary private insurance will result in 29.280% increase in medical expenses.

```r
(exp(OLS_reg$coefficients)-1)*100
```

```
 (Intercept)          age       female        white       totchr      suppins 
36336.450509     1.273661    -7.366254    37.411599    56.091416    29.280045 
```

1.3 (iii)

First we can re-estimate the quantile regressions from 0.05 to 0.95 in the same model as in Section 1.2.

```
# Quantile regression in increments of 0.05
q_005 = seq(0.05, 0.95, length.out=19)
quant_reg_005 = rq(lntotexp ~ . , tau = q_005, data = dfData)
qr_summary=summary(quant_reg_005)
qr_summary
```

```
Call: rq(formula = lntotexp ~ ., tau = q_005, data = dfData)

tau: [1] 0.05

Coefficients:
            Value    Std. Error t value  Pr(>|t|)
(Intercept)  3.36557  0.68439    4.91765  0.00000
age          0.01977  0.00893    2.21353  0.02694
female       0.12068  0.10803    1.11704  0.26407
white       -0.23365  0.23069   -1.01282  0.31123
totchr       0.63345  0.02977   21.27576  0.00000
suppins      0.41912  0.11495    3.64608  0.00027

Call: rq(formula = lntotexp ~ ., tau = q_005, 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_005, data = dfData)
```

```
tau: [1] 0.15

Coefficients:
            Value    Std. Error t value  Pr(>|t|)
(Intercept)  4.15640  0.41748    9.95605  0.00000
age          0.01865  0.00537    3.47031  0.00053
female       0.02271  0.07068    0.32138  0.74795
white        0.15737  0.13749    1.14459  0.25247
totchr       0.51204  0.02432   21.05569  0.00000
suppins      0.39942  0.06989    5.71491  0.00000

Call: rq(formula = lntotexp ~ ., tau = q_005, data = dfData)

tau: [1] 0.2

Coefficients:
            Value    Std. Error t value  Pr(>|t|)
(Intercept)  4.47890  0.34615   12.93916  0.00000
age          0.01734  0.00454    3.81746  0.00014
female      -0.01323  0.06120   -0.21618  0.82886
white        0.25032  0.09454    2.64763  0.00815
totchr       0.48030  0.02012   23.86793  0.00000
suppins      0.40203  0.06042    6.65370  0.00000

Call: rq(formula = lntotexp ~ ., tau = q_005, 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_005, data = dfData)

tau: [1] 0.3

Coefficients:
```

```
            Value    Std. Error t value  Pr(>|t|)
(Intercept)  5.18763  0.32873   15.78085  0.00000
age          0.01207  0.00428    2.82053  0.00483
female      -0.03342  0.05733   -0.58296  0.55996
white        0.47252  0.07958    5.93801  0.00000
totchr       0.42963  0.01802   23.84426  0.00000
suppins      0.28488  0.05991    4.75485  0.00000

Call: rq(formula = lntotexp ~ ., tau = q_005, data = dfData)

tau: [1] 0.35

Coefficients:
            Value    Std. Error t value  Pr(>|t|)
(Intercept)  5.14852  0.32570   15.80777  0.00000
age          0.01458  0.00420    3.46956  0.00053
female      -0.06382  0.05469   -1.16706  0.24328
white        0.52359  0.12196    4.29323  0.00002
totchr       0.41297  0.01906   21.66773  0.00000
suppins      0.29115  0.05391    5.40044  0.00000

Call: rq(formula = lntotexp ~ ., tau = q_005, data = dfData)

tau: [1] 0.4

Coefficients:
            Value    Std. Error t value  Pr(>|t|)
(Intercept)  5.34247  0.34784   15.35906  0.00000
age          0.01400  0.00414    3.38472  0.00072
female      -0.08100  0.05366   -1.50939  0.13131
white        0.54055  0.17574    3.07593  0.00212
totchr       0.41102  0.01960   20.97561  0.00000
suppins      0.28977  0.05397    5.36882  0.00000

Call: rq(formula = lntotexp ~ ., tau = q_005, data = dfData)

tau: [1] 0.45

Coefficients:
            Value    Std. Error t value  Pr(>|t|)
(Intercept)  5.53579  0.35381   15.64622  0.00000
age          0.01411  0.00407    3.46239  0.00054
female      -0.06450  0.05189   -1.24309  0.21393
```

```
white        0.49315  0.19768    2.49466  0.01266
totchr       0.40721  0.01893   21.50765  0.00000
suppins      0.25994  0.05275    4.92812  0.00000

Call: rq(formula = lntotexp ~ ., tau = q_005, 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_005, data = dfData)

tau: [1] 0.55

Coefficients:
            Value    Std. Error t value  Pr(>|t|)
(Intercept)  5.82910  0.39492   14.76022  0.00000
age          0.01416  0.00407    3.48048  0.00051
female      -0.09861  0.05257   -1.87593  0.06076
white        0.54989  0.26352    2.08671  0.03700
totchr       0.38758  0.01961   19.76391  0.00000
suppins      0.23471  0.05495    4.27124  0.00002

Call: rq(formula = lntotexp ~ ., tau = q_005, data = dfData)

tau: [1] 0.6

Coefficients:
            Value    Std. Error t value  Pr(>|t|)
(Intercept)  6.15907  0.44080   13.97262  0.00000
age          0.01506  0.00420    3.58836  0.00034
female      -0.10853  0.05583   -1.94395  0.05200
white        0.25683  0.31863    0.80602  0.42030
totchr       0.39562  0.02031   19.47553  0.00000
suppins      0.25798  0.05577    4.62590  0.00000
```

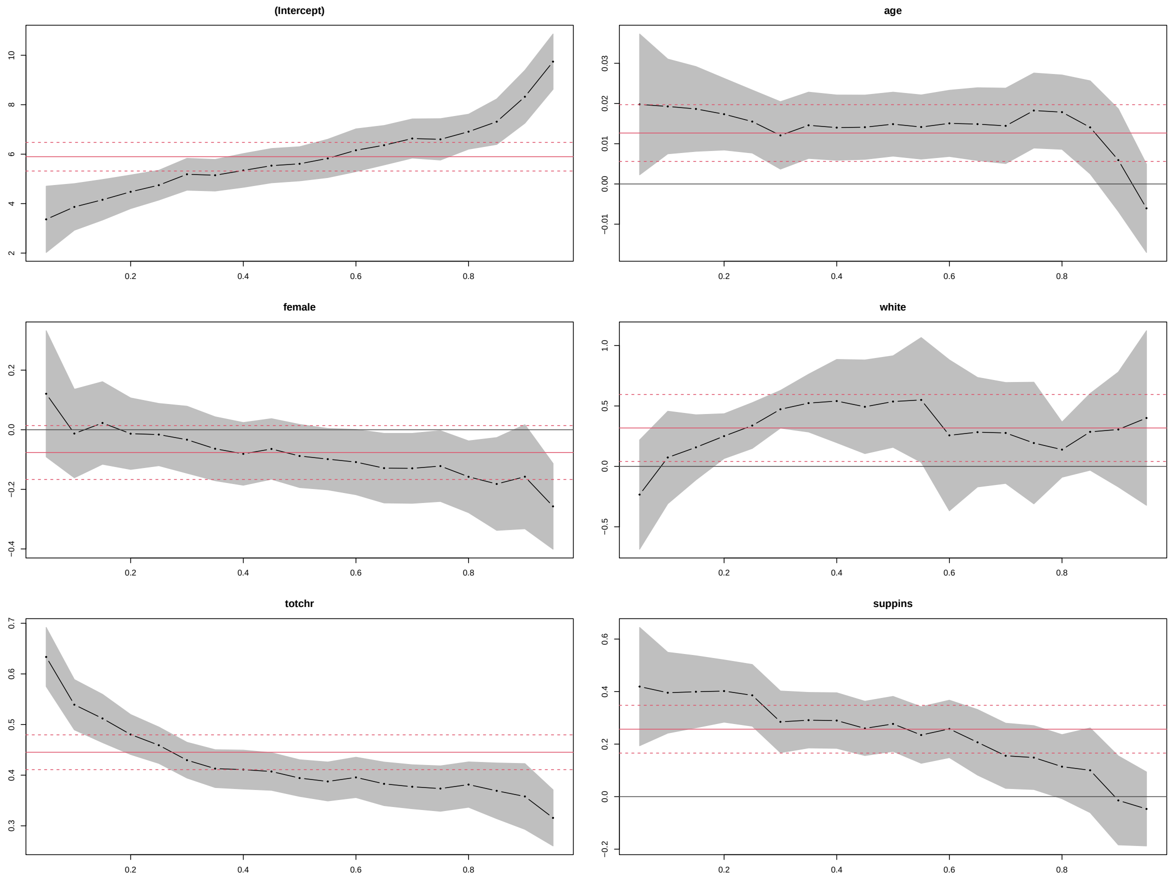
```
Call: rq(formula = lntotexp ~ ., tau = q_005, data = dfData)

tau: [1] 0.65

Coefficients:
            Value    Std. Error t value  Pr(>|t|)
(Intercept)  6.36258  0.40648   15.65275  0.00000
age          0.01487  0.00461    3.22352  0.00128
female      -0.12887  0.05958   -2.16293  0.03063
white        0.28299  0.23108    1.22462  0.22082
totchr       0.38288  0.02194   17.44947  0.00000
suppins      0.20693  0.06372    3.24745  0.00118

Call: rq(formula = lntotexp ~ ., tau = q_005, data = dfData)

tau: [1] 0.7

Coefficients:
            Value    Std. Error t value  Pr(>|t|)
(Intercept)  6.63358  0.40368   16.43281  0.00000
age          0.01444  0.00478    3.02030  0.00255
female      -0.12951  0.05988   -2.16259  0.03065
white        0.27653  0.21300    1.29828  0.19429
totchr       0.37716  0.02214   17.03824  0.00000
suppins      0.15564  0.06329    2.45903  0.01399

Call: rq(formula = lntotexp ~ ., tau = q_005, 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_005, data = dfData)

tau: [1] 0.8
```

```
Coefficients:
            Value    Std. Error t value  Pr(>|t|)
(Intercept)  6.90999  0.36065   19.15991  0.00000
age          0.01785  0.00471    3.78762  0.00016
female      -0.15788  0.06144   -2.56945  0.01023
white        0.13863  0.11657    1.18927  0.23443
totchr       0.38143  0.02285   16.69225  0.00000
suppins      0.11425  0.06222    1.83630  0.06641

Call: rq(formula = lntotexp ~ ., tau = q_005, data = dfData)

tau: [1] 0.85

Coefficients:
            Value    Std. Error t value  Pr(>|t|)
(Intercept)  7.31366  0.46945   15.57926  0.00000
age          0.01407  0.00590    2.38227  0.01727
female      -0.18200  0.07945   -2.29064  0.02205
white        0.28563  0.16208    1.76226  0.07813
totchr       0.36909  0.02806   13.15508  0.00000
suppins      0.10036  0.08226    1.21999  0.22257

Call: rq(formula = lntotexp ~ ., tau = q_005, 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

Call: rq(formula = lntotexp ~ ., tau = q_005, data = dfData)

tau: [1] 0.95

Coefficients:
            Value    Std. Error t value  Pr(>|t|)
(Intercept)  9.74213  0.57059   17.07369  0.00000
age         -0.00606  0.00560   -1.08127  0.27967
```

```
female       -0.25712  0.07341   -3.50255  0.00047
white         0.40026  0.36872    1.08554  0.27777
totchr        0.31566  0.02827   11.16644  0.00000
suppins      -0.04675  0.07189   -0.65032  0.51553
```

Then we can plot the resulting quantile regression coefficient estimates along with their 95% confidence intervals, and include the OLS linear regression estimates as well for a comparison.

```
plot(qr_summary, level= .95, ols= TRUE)
```



Finally, let's look at how the quantile regression coefficients compare to the OLS results and what their trend is. The graphs represent each coefficient estimate across quantiles (black dotted line) with their confidence intervals around them (shaded area). The OLS results are represented with the red continuous line along with the red dashed lines as the 95% CI. It seems that most coefficients have a relatively visible trend across the quantiles. From the lowest to the highest 0.05 incremental quantiles in terms of medical expenditure, gender, chronic illness and private insurance tends to have a decreasing coefficient and sometimes significance too. The age and the white variables seem to not be too different from the OLS estimates across the quantiles, apart from a few groups. This is the same pattern as seen before, where age is significantly positive across the lower quantiles as OLS, but deviates from the OLS when the highest spending quantiles are reached and actually becomes statistically insignificant. Similarly for white, the variable is not significant for most of the quantiles due to increased variance, but more or less follows the OLS estimate and has a statistically significant coefficient for the middle quantiles. The strongest deviations from the OLS estimates across the quantiles are exhibited by the chronic illness and private insurance variables. The number of chronic illness is a strong positive predictor of increased medical expenditure across all quantiles, but seems especially relevant for lower spending groups and has a less enhanced effect for the higher spending groups. Private insurance exhibits a similar effect on medical spending, with the exception that while the OLS shows the variable to be significant, the quantile regression reveals that this is not the case for the highest spending groups, only applies for the lower quantiles.

2 Question 2

2.1 (i)

Each individual has a different mean, likely due to individual-specific effects. When one takes the observation relative to the individual-level mean, we exclude the individual-specific information present in all the observations belonging to one individual. In this case, each observation's fitted value will exclude the effect of individual factors on its value, thus controlling for the individual effects.

2.2 (ii)

The idea between (1) “controlling for individual effects” and simply adding a polynomial/linear term for the time variable and (2) “controlling for individual and time effects” is that the first option controls for the individual effects and then includes the time dummy in the main regression model, while the second option considers the individual and time effects in a two-way model before including them in the main regression model. This latter can be especially helpful if the panel data is not balanced, i.e. some time periods have a lot more observations than others (or some periods are partially missing), and/or if the same applies to individual groups. In this case, one has to deal with this imbalance when using the first method, but if the data is not randomly missing (say one specific regressor quantile tends to be missing in the same period), or one does not want to deal with filling in the missing gaps, the two-way fixed effects control makes more sense as it deals with this imbalance in the individual effects estimation and not in the main model.

Furthermore, using simple individual effects and then adding a time dummy variable in the model can be problematic if the regressors exhibit a time-trend, in which case the time dummy will introduce multicollinearity to the model. In this case, it is better to include the time variable in the estimation of the individual effects term, since the two-way option will eliminate the risk of multicollinearity among regressors (however this will still require fixed effects estimation).

2.3 (iii)

Provided that the stronger assumptions of the random effects as compared to fixed effects hold, the random effects are better suited to estimate individual effects because the stochastic estimation of the individual effects. If the individual specific effects are uncorrelated to the regressors, the random effects estimator is consistent and more efficient than the fixed effects. However, if the individual effects are correlated to the regressors, the random effects is not consistent and it is better to use the fixed effects estimator which stays consistent in this scenario. Another argument for using random effects is the statistical “freedom” it provides. If the dataset contains a large number of individual groups, the fixed effects estimation will seriously impact the degrees of freedom in the model, thus affecting the statistical estimation of the models.

3 Question 3

```
dfData2 = read.csv("assignment2b_2023.csv")
attach(dfData2)

dfData2 <- na.omit(dfData2)
```

3.1 (i)

```
# Pooled OLS model including variable asvabc
pool_reg1 = plm(earnings ~  school + age + agesq  + ethblack + urban +
 ↪  regne + regnc + regw + regs + asvabc, data = dfData2, index =
 ↪  c("id","time"), model="pooling")

summary(pool_reg1)
```

```
Pooling Model

Call:
plm(formula = earnings ~ school + age + agesq + ethblack + urban +
    regne + regnc + regw + regs + asvabc, data = dfData2, model = "pooling",
    index = c("id", "time"))

Unbalanced Panel: n = 4765, T = 1-18, N = 40043

Residuals:
    Min.  1st Qu.   Median  3rd Qu.     Max.
-17.4600  -3.6407  -0.9170   2.3096 180.3323

Coefficients: (1 dropped because of singularities)
               Estimate  Std. Error  t-value  Pr(>|t|)
(Intercept) -1.6845e+01  8.7091e-01 -19.3415 < 2.2e-16 ***
school       7.8853e-01  1.9903e-02  39.6183 < 2.2e-16 ***
age          4.3574e-01  6.0898e-02   7.1552 8.498e-13 ***
agesq       -9.9785e-04  1.0391e-03  -0.9603    0.3369
ethblack    -1.2184e+00  1.2351e-01  -9.8649 < 2.2e-16 ***
urban        1.3013e+00  8.6368e-02  15.0672 < 2.2e-16 ***
regne        1.5878e+00  1.0400e-01  15.2667 < 2.2e-16 ***
regnc        7.9168e-02  9.0334e-02   0.8764    0.3808
regw         9.3336e-01  1.1488e-01   8.1246 4.615e-16 ***
asvabc       1.2281e-01  5.3880e-03  22.7938 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Total Sum of Squares:    2624500
Residual Sum of Squares: 2033300
R-Squared:      0.22528
Adj. R-Squared: 0.22511
F-statistic: 1293.48 on 9 and 40033 DF, p-value: < 2.22e-16
```

```
pool_reg1$coefficients
```

```
  (Intercept)        school           age         agesq      ethblack
-1.684458e+01  7.885342e-01  4.357356e-01 -9.978518e-04 -1.218389e+00
        urban         regne         regnc          regw        asvabc
 1.301322e+00  1.587785e+00  7.916813e-02  9.333638e-01  1.228122e-01
```

```r
# Pooled OLS model without variable asvabc
pool_reg2 = plm(earnings ~ school + age + agesq + ethblack + urban +
 ↪  regne + regnc + regw + regs, data = dfData2, index = c("id","time"),
 ↪  model="pooling")

summary(pool_reg2)
```

```
Pooling Model

Call:
plm(formula = earnings ~ school + age + agesq + ethblack + urban + 
    regne + regnc + regw + regs, data = dfData2, model = "pooling", 
    index = c("id", "time"))

Unbalanced Panel: n = 4765, T = 1-18, N = 40043

Residuals:
    Min.  1st Qu.   Median  3rd Qu.     Max. 
-18.4575  -3.6818  -0.9500   2.3581 180.8220 

Coefficients: (1 dropped because of singularities)
               Estimate  Std. Error  t-value  Pr(>|t|)    
(Intercept) -1.3400e+01  8.6323e-01 -15.5226 < 2.2e-16 ***
school       1.0419e+00  1.6616e-02  62.7079 < 2.2e-16 ***
age          3.8939e-01  6.1257e-02   6.3567 2.083e-10 ***
agesq       -2.3539e-04  1.0453e-03  -0.2252  0.821834    
ethblack    -2.2994e+00  1.1478e-01 -20.0339 < 2.2e-16 ***
urban        1.3540e+00  8.6895e-02  15.5823 < 2.2e-16 ***
regne        1.8092e+00  1.0422e-01  17.3599 < 2.2e-16 ***
regnc        2.4560e-01  9.0620e-02   2.7102  0.006727 ** 
regw         1.0853e+00  1.1543e-01   9.4029 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Total Sum of Squares:    2624500
Residual Sum of Squares: 2059600
R-Squared:      0.21523
Adj. R-Squared: 0.21507
F-statistic: 1372.45 on 8 and 40034 DF, p-value: < 2.22e-16
```

```
pool_reg2$coefficients
```

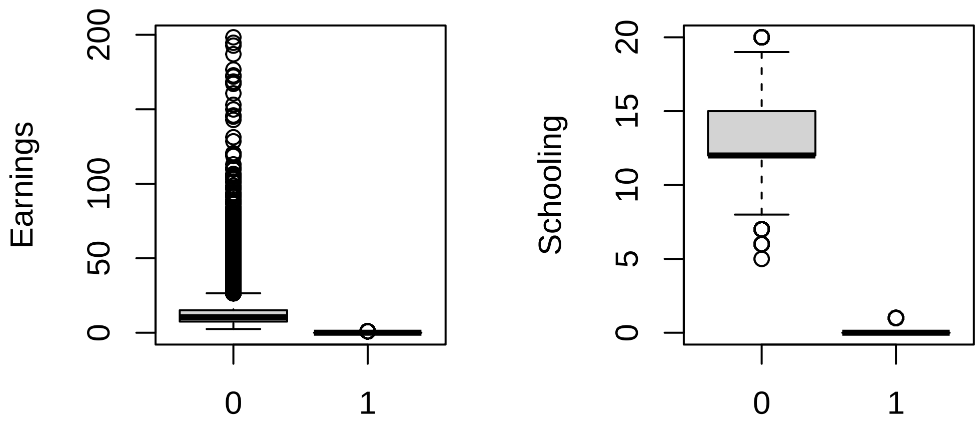
```
  (Intercept)        school           age         agesq      ethblack
-1.339954e+01  1.041934e+00  3.893922e-01 -2.353938e-04 -2.299405e+00
        urban         regne         regnc          regw
 1.354014e+00  1.809202e+00  2.455961e-01  1.085347e+00
```

β_1 is smaller when including the variable asvabc (index test score, constant over time for each individual). This means that when accounting for the individual effect of competence test result, the effect of years of schooling is mitigated on earnings, albeit still significant. This implies that differences in competence matter in terms of returns to schooling and that the competence drives between individual variance in the relationship between earnings and schooling.

3.2 (ii)

First, we check if there is a difference in mean in variability of returns to schooling and earnings between two groups of black and non-black people.

```
par(mfrow=c(1,2))
boxplot(dfData2$earnings, dfData2$ethblack, names=c(0,1),
 ↪  ylab="Earnings")
boxplot(dfData2$school, dfData2$ethblack, names=c(0,1),
 ↪  ylab="Schooling")
```



The boxplots show that in the non-black group, the earnings and years of schooling are much higher on average than the black group. There are also quite a few outliers for earnings in the non-black group and less so in the black group. Thus, we hypothesize that ethnicity might have at least some effect on earnings but it is unclear how this impact is influenced by crossing terms with schooling.

```
pool_reg3 = plm(earnings ~ school*ethblack + school + ethblack + urban +
 ↪  regne + regnc + regw + asvabc, data = dfData2, index =
 ↪  c("id","time"), model="pooling")
summary(pool_reg3)
```

```
Pooling Model

Call:
plm(formula = earnings ~ school * ethblack + school + ethblack +
    urban + regne + regnc + regw + asvabc, data = dfData2, model = "pooling",
    index = c("id", "time"))

Unbalanced Panel: n = 4765, T = 1-18, N = 40043

Residuals:
    Min.  1st Qu.   Median  3rd Qu.     Max.
```

```
-16.2613  -3.9618  -1.1982   2.4260 179.6043

Coefficients:
                   Estimate Std. Error  t-value  Pr(>|t|)    
(Intercept)      -7.9987473  0.2700092 -29.6240 < 2.2e-16 ***
school            1.0206171  0.0207928  49.0852 < 2.2e-16 ***
ethblack          0.2693862  0.7145561   0.3770    0.7062    
urban             0.8268955  0.0892940   9.2604 < 2.2e-16 ***
regne             1.4911444  0.1080799  13.7967 < 2.2e-16 ***
regnc             0.0411264  0.0938773   0.4381    0.6613    
regw              0.9267163  0.1193787   7.7628 8.502e-15 ***
asvabc            0.1189555  0.0055972  21.2525 < 2.2e-16 ***
school:ethblack -0.0940948  0.0549441  -1.7126    0.0868 .  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Total Sum of Squares:    2624500
Residual Sum of Squares: 2196100
R-Squared:      0.16323
Adj. R-Squared: 0.16306
F-statistic: 976.167 on 8 and 40034 DF, p-value: < 2.22e-16
```

According to the summary of the model, the cross term of ‘school’ and ‘ethblack’ is not significant on a 5% significance level. Thus, we cannot say that there is heterogeneity in schooling by ethnicity using pooled OLS regression. This is probably due to using pooled model instead of individual effects model, which fails to detect the heterogeneity of returns to schooling in the population.

3.3 (iii)

```r
ran_reg =plm(earnings ~  school +  ethblack + school*ethblack + urban +
 ↪   regne + regnc + regw + asvabc, data = dfData2, index =
 ↪   c("id","time"), model="random", effect="twoways")
summary(ran_reg)
```

```
Twoways effects Random Effect Model 
   (Swamy-Arora's transformation)

Call:
plm(formula = earnings ~ school + ethblack + school * ethblack + 
    urban + regne + regnc + regw + asvabc, data = dfData2, effect = "twoways",
    model = "random", index = c("id", "time"))

Unbalanced Panel: n = 4765, T = 1-18, N = 40043

Effects:
                   var std.dev share
idiosyncratic 31.0771  5.5747 0.607
individual    19.9836  4.4703 0.390
time           0.1694  0.4116 0.003
theta:
           Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
id    0.2198520 0.5965726 0.6480564 0.6248755 0.6838109 0.7179972
time  0.6243992 0.7206771 0.7372010 0.7259450 0.7409289 0.7513950
total 0.2038933 0.5496952 0.5920492 0.5741813 0.6197501 0.6592080

Residuals:
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
-14.973  -3.488  -0.757   0.498   2.882 179.984

Coefficients:
                  Estimate Std. Error  z-value  Pr(>|z|)
(Intercept)     -7.1326911  0.0934580 -76.3197 < 2.2e-16 ***
school           0.8621929  0.0063970 134.7804 < 2.2e-16 ***
ethblack         2.0252960  0.2233043   9.0697 < 2.2e-16 ***
urban            0.4985022  0.0185315  26.9003 < 2.2e-16 ***
regne            1.3155330  0.0316850  41.5191 < 2.2e-16 ***
regnc           -0.1685703  0.0275655  -6.1153 9.639e-10 ***
regw             0.9571009  0.0333374  28.7096 < 2.2e-16 ***
asvabc           0.1400315  0.0019082  73.3850 < 2.2e-16 ***
school:ethblack -0.2239882  0.0171540 -13.0575 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Total Sum of Squares:    2624500
Residual Sum of Squares: 2212100
R-Squared:      0.16178
Adj. R-Squared: 0.16161
Chisq: 62053.5 on 8 DF, p-value: < 2.22e-16
```

The results now show that all terms, including the ethnicity and schooling cross term are highly significant. This is in stark contrast with the pooled OLS solution of Section 3.2, where the ethnicity and the cross-sectional schooling are not significant. This suggests that when one controls for the heterogeneity among the racial groups in terms of returns to schooling, there is a significant disadvantage of black people in terms of earnings.

3.4 (iv)

This depends on whether the individual specific effects are correlated with the regressors. If the individual effects are correlated with the regressors, the fixed effects model makes more sense to be used since the random effects estimator will be inconsistent. We can reasonably assume that competence variable asvabc might be heavily correlated with the school variable or the inclusion of squared age will also introduce multicollinearity, therefore it would probably be more appropriate to use fixed effects.

3.5 (v)

```r
# Fixed effects estimation of the heterogeneity of returns to schooling
 ↪  by racial groups

fixed_reg <-  plm(earnings ~ school + ethblack + school*ethblack + age +
 ↪  agesq + urban + regne + regnc + regw + asvabc, data = dfData2, index
 ↪  = c("id","time"), model="within", effect = "individual")
summary(fixed_reg)
```

```
Oneway (individual) effect Within Model

Call:
plm(formula = earnings ~ school + ethblack + school * ethblack +
    age + agesq + urban + regne + regnc + regw + asvabc, data = dfData2,
    effect = "individual", model = "within", index = c("id",
        "time"))

Unbalanced Panel: n = 4765, T = 1-18, N = 40043

Residuals:
      Min.    1st Qu.     Median    3rd Qu.       Max.
-84.659043  -1.745915  -0.049047   1.458638 159.380450

Coefficients:
                   Estimate  Std. Error t-value  Pr(>|t|)
```

```
school           0.85196554  0.07070579 12.0494 < 2.2e-16 ***
age              0.42804890  0.05222973  8.1955 2.579e-16 ***
agesq           -0.00041708  0.00088502 -0.4713 0.6374503
urban            0.20409228  0.11758995  1.7356 0.0826386 .
regne            0.67237844  0.30015268  2.2401 0.0250892 *
regnc           -0.44728744  0.26207171 -1.7067 0.0878798 .
regw             1.06017614  0.29977886  3.5365 0.0004059 ***
school:ethblack -1.05406139  0.23104693 -4.5621 5.081e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Total Sum of Squares:    1269800
Residual Sum of Squares: 1099400
R-Squared:      0.13412
Adj. R-Squared: 0.016972
F-statistic: 682.917 on 8 and 35270 DF, p-value: < 2.22e-16
```

The output of the fixed effects regression shows that when we make the switch from the random effects, while some other variables become statistically insignificant such as the age-squared, the school and ethnicity cross term is still strongly significant and the coefficient decreases even more to show a stronger discrimination of black people in earnings.

3.6 (vi)

Null hypothesis: both $\widehat{\beta}_{ran\_effect}$ and $\widehat{\beta}_{fixed\_effect}$ are consistent but the former is more efficient. Alternative hypothesis: only $\widehat{\beta}_{fixed\_effect}$ is consistent.

```
# Perform Hausman test
phtest(ran_reg, fixed_reg, data= dfData2)
```

```
    Hausman Test

data:  earnings ~ school + ethblack + school * ethblack + urban + regne +  ...
chisq = 26.973, df = 6, p-value = 0.0001465
alternative hypothesis: one model is inconsistent
```

Since the p-value is significant, we reject the null hypothesis and accept the null hypothesis that only the fixed effect estimator is consistent. The Hausman test is especially useful therefore, since it shows us that the fixed effects estimator is more appropriate to be used than the random effects. Due to the heterogeneity between the regressors and the individual effects estimator, the fixed effects estimator will be consistent, while the random effects estimators are likely misleading.

3.7 (vii)

```
# Get average group mean of schooling per individual
dfData2=dfData2 %>% group_by(id) %>% mutate(group_m_school =
 ↪  mean(school))
# Get average group mean of age per individual
dfData2=dfData2 %>% group_by(id) %>% mutate(group_m_age = mean(age))
# Get average group mean of age squared per individual
dfData2=dfData2 %>% group_by(id) %>% mutate(group_m_agesq = mean(agesq))
# Get average group mean of urban per individual
dfData2=dfData2 %>% group_by(id) %>% mutate(group_m_urban = mean(urban))
# Get average group mean of regne per individual
dfData2=dfData2 %>% group_by(id) %>% mutate(group_m_regne = mean(regne))
# Get average group mean of regnc per individual
dfData2=dfData2 %>% group_by(id) %>% mutate(group_m_regnc = mean(regnc))
# Get average group mean of regw per individual
dfData2=dfData2 %>% group_by(id) %>% mutate(group_m_regw = mean(regw))

# Unrestricted model including all group means for Mundlak
Mundlak_gls_ur <- gls(earnings ~ group_m_school + group_m_age +
 ↪  group_m_agesq + group_m_urban + group_m_regne + group_m_regnc +
 ↪  group_m_regw + school + ethblack +age+agesq + urban + regne + regnc
 ↪  + regw + asvabc, data = dfData2)

# Restricted GLS without any group means
Mundlak_gls_r <- gls(earnings ~ school + ethblack +age+agesq + urban +
 ↪  regne + regnc + regw + asvabc, data = dfData2)

SSR_ur <- (t(Mundlak_gls_ur$residuals) %*% Mundlak_gls_ur$residuals)

SSR_r <- (t(Mundlak_gls_r$residuals) %*% Mundlak_gls_r$residuals)

nr_degf_ur = Mundlak_gls_ur[["dims"]][["N"]] -
 ↪  Mundlak_gls_ur[["dims"]][["p"]] - 1

F_teststat <- ((SSR_r - SSR_ur)/7)/(SSR_ur/nr_degf_ur)
```

```r
print(F_teststat)
```

```
         [,1]
[1,] 25.92072
```

```r
p_val <- 1-pf(F_teststat, 7, nr_degf_ur)
print(p_val)
```

```
     [,1]
[1,]    0
```

There is statistically significant evidence suggesting that the coefficients of the group_means γ are jointly different than 0. The null hypothesis is that the coefficients of the time-variant means are jointly 0, that is $H_0 : \gamma_1 = \gamma_2 = \gamma_3 = \gamma_4 = \gamma_5 = \gamma_6 = \gamma_7$. This hypothesis can easily be tested by estimating a restricted and an unrestricted model and investigating the improvement from the first to the second in terms of the sum of squared residuals (i.e. by how much the explanatory power of the model improves when including the hypothetically 0 coefficient variables). Our test rejects the null hypothesis at 5% significance level, which means that at least some of the grouped means are not zero-weighted in a GLS estimation model, which in turn implies that (a part of) the fixed effects are correlated with the regressors. Thus, the individual specific effects are correlated with the regressors and thus, we should use a fixed effect model.

3.8 (viii)

To sum up, there seems to be strong evidence of heterogeneity in terms of returns to schooling among racial groups. In this case, estimating the model of explaining earnings by schooling is better to be done by some individual effects estimation rather than just some simple pooling. Furthermore, the Hausman test suggests that there is significant correlation between the regressors and the fixed effects, which means that the random effects estimation will not be consistent. This is further confirmed by the Mundlak regression test (which is asymptotically equivalent to the Hausman test). Therefore, even though the fixed effects estimation is less efficient, we are left with this as our best model to control for the individual effects. Looking at the results, it seems that there is indeed a different return to schooling for black people versus non-black people.

3.9 (ix)

```
# Get frequency (number of waves) of each ID
df_freq = as.data.frame(table(dfData2$id))

# Filter out participants ID with frequency of less than 5
df_freq = df_freq[(df_freq$Freq >5),]

# Filter out participants with d=0 in the original dataset
df_balanced = dfData2[dfData2$id %in% df_freq$Var1,]

# Estimation on unbalanced panel
unbalanced = plm(earnings ~ school + ethblack +age+agesq + urban + regne
 ↪  + regnc + regw + asvabc, data = dfData2, index = c("id","time"),
 ↪  model="pooling")

# Estimation on balanced panel
balanced = plm(earnings ~ school + ethblack +age+agesq + urban + regne +
 ↪  regnc + regw + asvabc, data = df_balanced, index = c("id","time"),
 ↪  model="pooling")

# Verbeek and Nijman test
phtest(balanced,unbalanced, data= dfData2)
```

```
    Hausman Test

data:  earnings ~ school + ethblack + age + agesq + urban + regne +  ...
chisq = 58.813, df = 9, p-value = 2.269e-09
alternative hypothesis: one model is inconsistent
```

The Verbeek and Nijman test use a Hausman type test on $\widehat{\beta}_{balanced} - \widehat{\beta}_{unbalanced}$.

The null hypothesis is that there is no attrition bias, thus the unbalanced estimator is more efficient.

The alternative hypothesis is that there is attrition bias, thus, the balanced estimator is more efficient.

In this case, we reject the null hypothesis and conclude that the balanced estimator is more efficient. This implies that our analysis could be more efficient by dropping low observation number individuals from the dataset and estimating the model then, either by pooling or fixed effects.