Econometrics II - Assignment 1

Uncensored sloths

10 Jan 2022

Question 1

a) Run an OLS regression for log-earnings on schooling, age, and age squared. Present the results and comment on the estimates.\

```
# Load data
data <- read.csv("assignment1a.csv")
```

```
# Run regression
model1 <- lm(logwage ~ schooling + age + agesq, data = data)
stargazer(model1)
```

% Table created by stargazer v.5.2.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu
% Date and time: Mo, Jan 10, 2022 - 22:20:53

Table 1:

	Dependent variable:
	logwage
schooling	0.216^{***}
	(0.032)
age	-0.342
	(0.521)
agesq	-0.011
	(0.008)
Constant	26.409^{***}
	(8.057)
Observations	416
R^2	0.815
Adjusted R^2	0.813
Residual Std. Error	1.499 (df = 412)
F Statistic	604.261^{***} (df = 3; 412)
Note:	*p<0.1; **p<0.05; ***p<0.01

As the results of the regression show the estimate for schooling is highly significant and positive which means an additional year of schooling increases the log wage by around 0.22, i.e. that an additional year of schooling is associated with an $(\exp(0.22) - 1) \cdot 100\% \approx 24.61\%$ increase in wage. The estimates for age and age squared

are negative which means that older individuals are associated with lower log-wages. However, both estimates are not significant. Thus we do not expand on the interpretation of these coefficients.\

b) Briefly discuss the sample selection problem that may arise in using these OLS estimates for the purpose of predicting the potential earnings of the non-employed. Formulate the sample selection model. In your answer, include an explanation why OLS may fail in this context.\

The researcher's aim is to gain insight on the potential earnings of unemployed people. However, there is the (likely) possibility of a non-random (self) selection process of unemployed individuals. Reasons for self selection might be (non-exclusively) unwillingness to work and/or inaccessibility to the labor market, and both of these aspects may relate to certain characteristics of individuals.\ To be clear, we do not exactly know why these individuals are unemployed and whether these reasons could have an effect on their earnings if they worked. If for instance the reason why they do not work is due to inaccessibility to the job market as a result of a low education level, we can assume that in case of employment they would have a lower wage than individuals who are already employed. Put differently, the estimation of the sign and magnitude of variable effects on wages, on which a prediction of the potential earnings of the non-employed would ultimately be based on, needs to account for the fact that conditional on certain individual characteristics such a wage is not observed, because the individual is not employed. OLS is omitting all observations for which the wage is not observed, and is thus a conceptually inappropriate technique to apply in the context at hand. Moreover, if the selection process is assumed to be non-random, i.e. that certain characteristics determine the (self) selection into unemployment, this implies non-random sampling which renders the OLS estimates inconsistent. Both the conceptual and technical inadequacy of OLS estimation is revisited below after a sample selection model has been explicitly specified.\

A sample selection model helps to address the issue of self selection. Therefore, we formulate (i) a selection equation and (ii) a regression equation.

In our case, the (i) selection equation ($I_i^* = Z_i\prime\gamma + V_i$) models the latent variable I_i^*, which depends on a set of observed variables Z_i and unobserved characteristics V_i, and based on which individuals' labor market participation is determined according to the following indicator function:

$$I_i = \begin{cases} 1 \text{ (employed)} & \text{if } I_i^* > 0 \\ 0 \text{ (unemployed)} & \text{if } I_i^* \leq 0 \end{cases}$$

One interpretation of this selection modulation could be that I_i^* corresponds to individuals' utility from working, so that individuals only decide to work if this utility is strictly positive. However, considerations of the selection process being also determined by individuals' accessibility to the labor market complicate the matter of interpretation and arguably conflict with its designation as self selection.

The (ii) regression equation specifies the latent variable Y_i^*, i.e. the log wage of individuals, which depends on a set of observed regressors X_i and unobserved characteristics U_i:

$$Y_i^* = X_i\prime\beta + U_i$$

This latent log wage is, however, only observed for individuals being (self) selected into working by the above specified selection process, that is if $I_i = 1$. Specifically, observed log wages Y_i are specified as follows.

$$Y_i = \begin{cases} Y_i^* & \text{if} I_i = 1 \\ missing & \text{if} I_i = 0 \end{cases}$$

So one needs to find characteristics that determine employment I_i to be included in Z_i. For the selection regression, we include schooling, age, age squared, marriage, subsidy and distance in Z_i as they all could have potential impact on the employment decision/selection. Schooling, subsidy and distance may have an impact on the educational level of individuals. Since we would expect chances of employment to increase in an individual's educational level, these factors may significantly enter the selection process. Furthermore, distance (and subsidy) could be viewed as proxies for the geographical remoteness of individuals' location of living and

as such may be associated with a lower accessibility to the labor market.Moreover, age is likely to determine the employment decision as usually very young an old individuals do not work. However, as this is obviously not a linear relationship, one needs to include the squared age to account for this non-linearity. Lastly, marriage is included as couples, especially in traditional value settings, may tend to divide responsibilities between each other resulting in one partner only working while the other takes care of housework. Along these argumentative lines, for all but one of these variables an assertion towards their impact on wages can be easily made, but is omitted here for sake of textual efficiency. Hence, these should be included in X_i of the regression equation. We would exclude marriage as marriage saliently affects the employment decision while it cannot be intuitively rationalized to affect wages (see below).\

As alluded to above, OLS estimation fails in this context and given the purpose of predicting the potential earnings of the non-employed for both conceptual and technical reasons. Conceptually, it is intuitive to conclude that OLS is unsuitable since it disregards all data of the non-employed and thus ignores the entire sample drawn from the population for which inferences constitute the empirical goal. Furthermore, parameter estimation under OLS does not honor the fact that sample selection is non-random and under non-random and observable conditions wages are not observed. Technically, it can be shown that OLS only provides consistent parameter estimates if either (i) the error terms U_i and V_i are independent or, i.e. to assume random sampling, or (ii) if X_i and Z_i are uncorrelated, i.e. $\mathrm{E}(U_i|X_i) = 0$. To make this point more explicit, we define the conditional expectation with which the empirical task at hand is concerned with, that of the unobserved earnings of the non-employed:\

$$\mathrm{E}(Y_i^*|I_i = 0, X_i, Z_i) = \mathrm{E}(X_i'\beta + U_i|Z_i'\gamma + V_i \leq 0, X_i, Z_i) = X_i'\beta + \mathrm{E}(U_i|V_i \leq -Z_i'\gamma, X_i, Z_i)$$

If either of the above specified conditions are satisfied it follows that the last term equals $\mathrm{E}(U_i) = 0$, such that OLS gives consistent estimates of $/beta$, on which predictions of the potential wage the non-employed would be based, otherwise, as can be explicitly shown, OLS estimates are both biased and inconsistent. Because we assume sampling to be non-random as well as because Z_i and X_i have common components, both conditions very likely do not hold in the context at hand, so that OLS estimation is clearly inadequate.\

c) Choose a suitable candidate as an exclusion restriction for the sample selection model. Motivate your choice. Estimate the sample selection model with the Heckman two-step estimator both with and without the exclusion restriction and compare the outcomes.\

As stated above, we use marriage as the exclusion restriction. Especially in traditional settings, when married, the partner who earns less often decides to stay unemployed and take responsibility for house work. However, marriage does not necessarily directly affect the amount individuals earn (although research shows that this is not necessarily the case for married men (Pollmann-Schult 2011; Chun and Lee 2001)).

```
data$work = ifelse(is.na(data$logwage), 0, 1)
```

```
heck1 = heckit(work ~ age + agesq + married + schooling + distance + subsidy,
               logwage ~ age + agesq + married+ schooling + distance + subsidy, data=data)
```

```
## Warning in heckit2fit(selection, outcome, data = data, weights = weights, :
## Inverse Mills Ratio is (virtually) collinear to the rest of the explanatory
## variables
```

```
heck2 = heckit(work ~ age + agesq + married + schooling + distance + subsidy,
               logwage ~ age + agesq + schooling + distance + subsidy, data=data)

stargazer(heck1)
```

% Table created by stargazer v.5.2.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu
% Date and time: Mo, Jan 10, 2022 - 22:20:53

```
stargazer(heck2)
```

Table 2:

	Dependent variable:
	logwage
age	−3.314
	(3.007)
agesq	0.035
	(0.049)
married	−4.037
schooling	−0.027
	(0.204)
distance	0.076
	(0.368)
subsidy	0.216
	(0.178)
Constant	85.831**
	(40.726)
Observations	666
R^2	0.817
Adjusted R^2	0.814
ρ	−1.341
Inverse Mills Ratio	−16.551
Note:	*p<0.1; **p<0.05; ***p<0.01

% Table created by stargazer v.5.2.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu
% Date and time: Mo, Jan 10, 2022 - 22:20:53

Table 3:

	Dependent variable:
	logwage
age	−0.353
	(0.538)
agesq	−0.011
	(0.009)
schooling	0.192***
	(0.034)
distance	0.049
	(0.058)
subsidy	0.057*
	(0.030)
Constant	26.205***
	(8.450)
Observations	666
R^2	0.817
Adjusted R^2	0.814
ρ	−0.082
Inverse Mills Ratio	−0.122 (0.606)
Note:	*p<0.1; **p<0.05; ***p<0.01

Comparing both models, we can see that for indicator function the estimates, their standard errors and t-values are the same. In this regression only marriage has a positive significant effect on the indicator which means that a married person is more likely to work. This is not in line with the sign of the relationship proposed a priori. However, a reason why married individuals are more likely to work may be that they bear a higher social responsibility for their household while singles generally only have to take care of a one-person household.\

For the outcome equation, schooling is positive and significant for the estimation where we exclude marriage in the outcome equation, indicating that an additional school year increases the log wage by approximately 0.19, i.e. is associated with an $(\exp(0.19) - 1) \cdot 100\% \approx 20.09$ increase in earnings. In the nother estimation, where we include marriage in the outcome equation, no estimator is significant (except for the intercept). Furthermore, no values for the standard error, t-value and p-value were estimated for the marriage estimator in the regression equation owing to the identification problem of the Heckman two-step procedure which is not resolved by means of an exclusion restriction in this case.\

More specifically, this identification problem is due to perfect collinearity in the second step of the Heckman two-step procedure. If $Z_i = X_i$ the regression equation in this step reads as follows and suffers from perfect collinearity because the inverse Mills ratio $\lambda(x) = \frac{\phi(x)}{\Phi(x)}$ is near linear for most of its domain:\

$$Y_i = X_i'\beta + \rho\sigma\frac{\phi(Z_i'\hat{\gamma})}{\Phi(Z_i'\hat{\gamma})} + U_i^* = X_i'\beta + \rho\sigma\lambda(Z_i'\hat{\gamma}) + U_i^* = X_i'\beta + \rho\sigma\lambda(X_i'\hat{\gamma}) + U_i^*$$

The described identification problem is resolved by means of the exclusion restriction adopted in the model above, in which the variable marriage is excluded from X_i. Hence, only this Heckit model is considered for subseding analysis.\

The fact that all other regressors, i.e. except schooling, are not significant in the model above prompts considerations of model selection which are disregarded here since they lay outside the methodologically concerned scope of this assignment. More specifically, more parsimonious specifications of the (2nd step) regression equation could be considered.\

As all other estimators are not significant, we should estimate another selection model where we include only the significant estimators in the next step. (COMMENT: if we exclude age and agesq then schooling becomes insignificant)!\

d) Estimate the sample selection model with maximum likelihood both with and without the exclusion restriction and compare the outcomes.

```
ml1 = selection(work ~ age + agesq + married + schooling + distance + subsidy,
                logwage ~ age + agesq + married+ schooling + distance + subsidy, data=data)
```

```
## Warning in heckit2fit(selection, outcome, data = data, printLevel =
## printLevel, : Inverse Mills Ratio is (virtually) collinear to the rest of the
## explanatory variables
```

```
ml2 = selection(work ~ age + agesq + married + schooling + distance + subsidy,
                logwage ~ age + agesq + schooling + distance + subsidy, data=data)

stargazer(ml1)
```

% Table created by stargazer v.5.2.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu
% Date and time: Mo, Jan 10, 2022 - 22:20:53

```
stargazer(ml2)
```

% Table created by stargazer v.5.2.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu
% Date and time: Mo, Jan 10, 2022 - 22:20:53

When estimating the equations with the maximum likelihood method we can see that there are only minor differences between the estimation including and excluding the exclusion restriction.\

In both models, marriage is the only significant estimator in the probit selection equation while in the outcome equation, schooling is the only significant estimator, having a positive effect on log wage (an additional school year is associated with a $(\exp(0.19) - 1) \cdot 100\% \approx 20.92\%$ in earnings).\

Moreover, note that no exclusion restriction is necessary in the case of estimation by means of maximum likelihood because here no two-step approach is taken which gives rise to a collinearity problem. However, normalization of the variance of V_i is necessary also in this case to ensure identifiability. To be explicit, the distributional assumption on the error terms made both in Hackman estimation as well as here is bivariate normality according to:

$$\begin{bmatrix} U_i \\ V_i \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma^2 \rho\sigma \\ \rho\sigma 1 \end{bmatrix} \right)$$

To further carve out the benefit and effect of ML estimation that no exclusion restriction is required, the model specification including marriage in X_i is considered for further analysis in e).

e) On the basis of your results, how would you specify the distribution of potential earnings for the non-employed?\

To specify the the distribution of potential earnings for the non-employed, histograms of fitted values for all observed non-employed individuals are contrasted across all 3 models considered above: (i) sample selection

Table 4:

	Dependent variable:
	logwage
age	−0.353
	(0.564)
agesq	−0.011
	(0.009)
married	−0.009
	(0.341)
schooling	0.192***
	(0.037)
distance	0.049
	(0.058)
subsidy	0.057*
	(0.032)
Constant	26.213***
	(9.191)
Observations	666
Log Likelihood	−1,184.034
ρ	−0.089 (0.838)
Note:	*p<0.1; **p<0.05; ***p<0.01

Table 5:

	Dependent variable:
	logwage
age	−0.348
	(0.534)
agesq	−0.011
	(0.009)
schooling	0.193***
	(0.034)
distance	0.049
	(0.058)
subsidy	0.057*
	(0.030)
Constant	26.107***
	(8.363)
Observations	666
Log Likelihood	−1,184.034
ρ	−0.068 (0.368)
Note:	*p<0.1; **p<0.05; ***p<0.01

model with maximum likelihood, (ii) sample selection model with the Heckman two-step estimator, (iii) OLS.\

Here fitted values should read as the expectation towards the unobserved log wage Y_i^* conditional on the fact that the log wage is not observed, i.e. that the individual is non-employed ($I_i = 0$), and on the respective individual characteristics X_i and Z_i:\

$$\begin{aligned}
\mathrm{E}(Y_i^*|I_i = 0, X_i, Z_i) &= \mathrm{E}(X_i'\beta + U_i|Z_i'\gamma + V_i \leq 0, X_i, Z_i) \\
&= X_i'\beta + \mathrm{E}(U_i|V_i \leq -Z_i'\gamma, X_i, Z_i) \\
&= X_i'\beta + \rho\sigma\mathrm{E}(V_i|V_i \leq -Z_i'\gamma, X_i, Z_i) \\
&= X_i'\beta + \rho\sigma\frac{\phi(Z_i'\gamma)}{\Phi(-Z_i'\gamma)}
\end{aligned}$$

While ρ and σ are directly estimated along with β and γ in the case of (i) maximum likelihood estimation, this is not the case for the two-step Heckman estimation. Details on the estimation technique of these structural parameters deployed by the customary pre coded function used here are outlined by Greene (2003, p. 784).\

```
beta <- ml1$estimate[8:14]
X <- cbind(1, data$age, data$agesq, data$married, data$schooling, data$distance, data$subsidy)
sigma <- ml1$estimate[15]
rho <- ml1$estimate[16]
Z <- cbind(1, data$age, data$agesq, data$married, data$schooling, data$distance, data$subsidy)
gamma <- ml1$estimate[1:7]
Ystarhat = X %*% beta + rho * sigma * dnorm( Z %*% gamma ) / pnorm( -(Z %*% gamma) )
datanew <- cbind(data, Ystarhat)
```

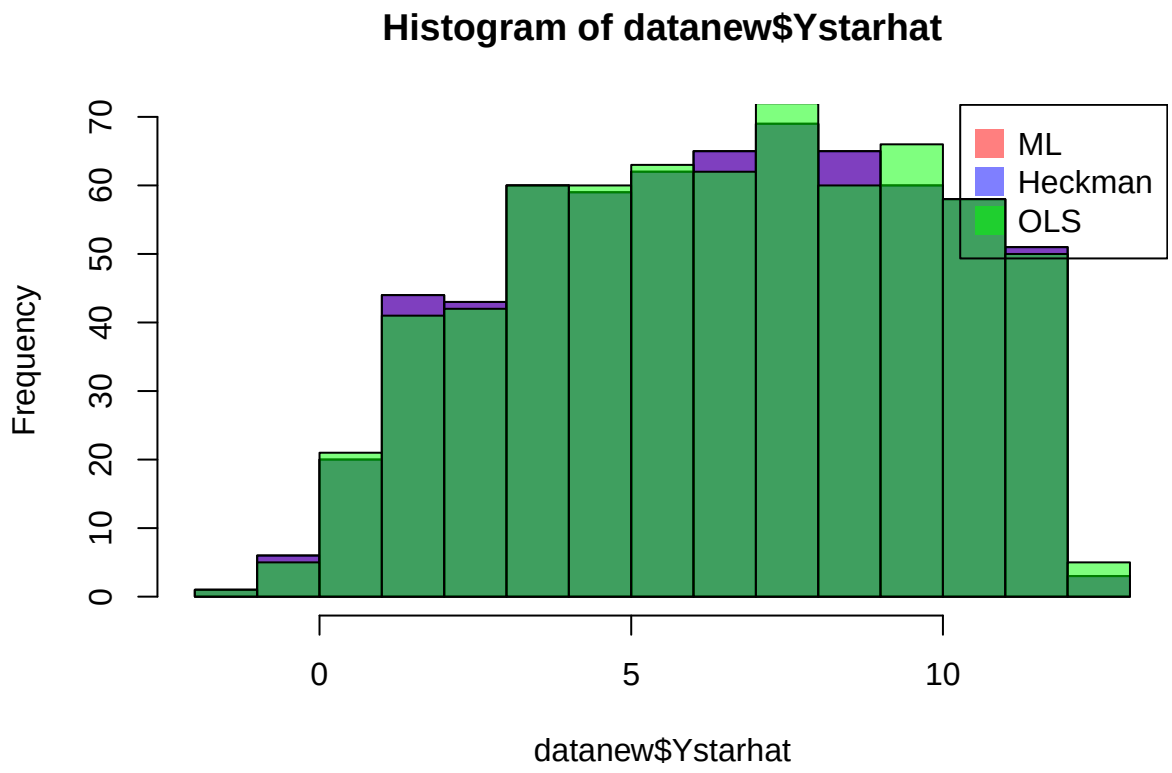
```
beta2 <- heck2$coefficients[8:13]
sigma2 <- heck2$coefficients[15]
rho2 <- heck2$coefficients[16]
gamma2 <- heck2$coefficients[1:7]
X2 <- cbind(1, data$age, data$agesq, data$schooling, data$distance, data$subsidy)
Ystarhat2 = X2 %*% beta2 + rho2 * sigma2 * dnorm( Z %*% gamma2 ) / pnorm( -(Z %*% gamma2) )
datanew <- cbind(datanew, Ystarhat2)
```

Note also that in the case of (iii) OLS estimation the last term of the expression for the conditional expectation of Y_i^* above vanishes since random sampling is assumed under OLS such that $\mathrm{E}(U_i|V_i \leq -Z_i'\gamma, X_i, Z_i) = \mathrm{E}(U_i) = 0$.\

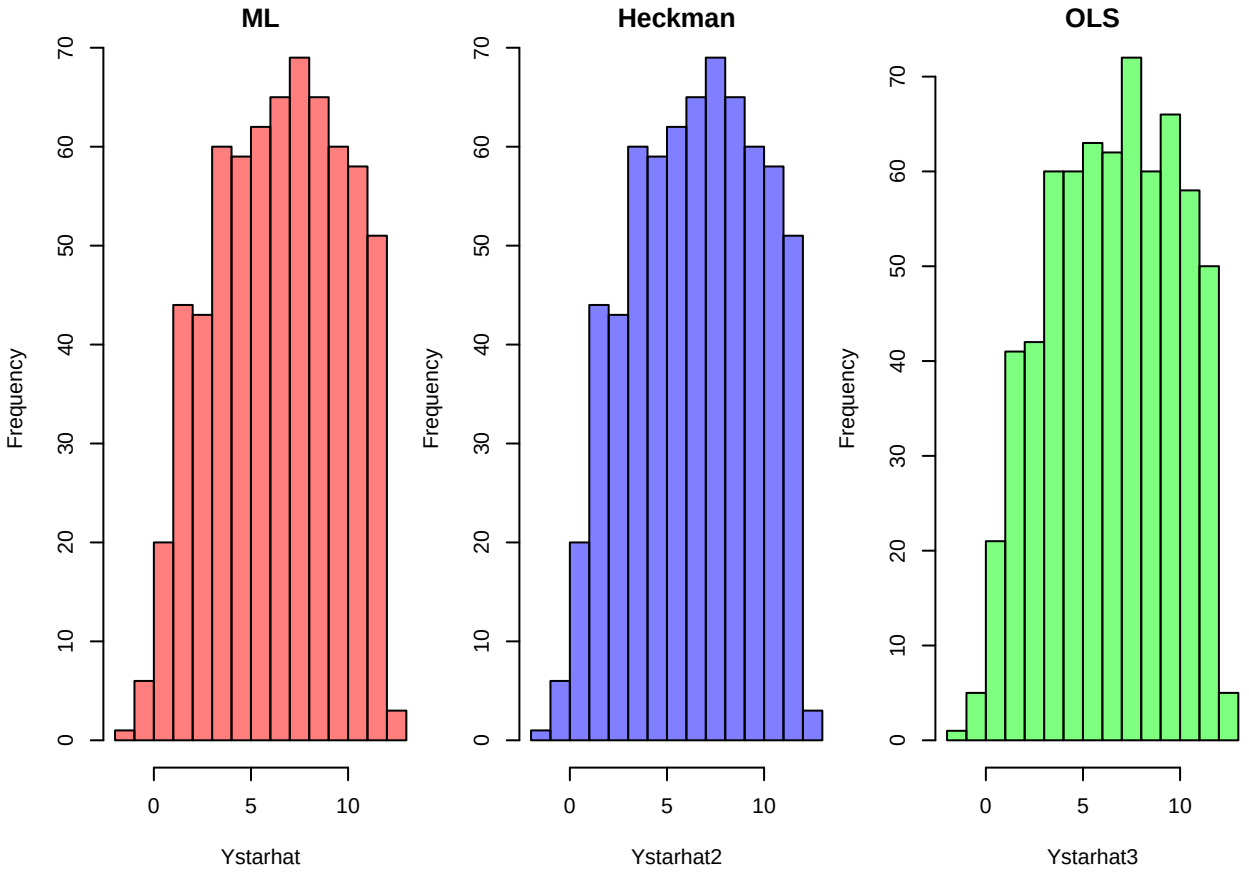
```
model2 <- lm(logwage ~ age + agesq + married + schooling + distance + subsidy, data = data)
beta3 <- model2$coefficients
Ystarhat3 = X %*% beta3
datanew <- cbind(datanew, Ystarhat3)
```

```
# Plot histogramms for conditional expected log wages of unemployed individuals

df <- datanew %>% filter_at(vars(logwage),all_vars(is.na(.)))

hist(datanew$Ystarhat, col=rgb(1,0,0,0.5))

hist(datanew$Ystarhat2, col=rgb(0,0,1,0.5), add=T)

hist(datanew$Ystarhat3, col=rgb(0,1,0,0.5),add=T)

# Add legend
```

```
legend("topright", legend=c("ML","Heckman", "OLS"), col=c(rgb(1,0,0,0.5),
     rgb(0,0,1,0.5), rgb(0,1,0,0.5)), pt.cex=2, pch=15 )
```




```
par(
  mfrow=c(1,3),
  mar=c(4,4,1,0)
)

hist(datanew$Ystarhat, col=rgb(1,0,0,0.5),xlab="Ystarhat", main="ML" )

hist(datanew$Ystarhat2, col=rgb(0,0,1,0.5),xlab="Ystarhat2", main="Heckman" )

hist(datanew$Ystarhat3, col=rgb(0,1,0,0.5),xlab="Ystarhat3", main="OLS" )
```



Comparison of the histograms obtained yields the conclusion that the distributions of potential earnings for the non-employed implied by the three models are in general very close, while those of the two sample selection models are as one might expect virtually identical. Obviously, the exclusion restriction applied to the Heckman model results in a small difference in estimated parameters and thus in expected conditional values. This difference is, however, unargubly bigger towards the OLS model, which not only yields a further diverging estimate for β (for reasons explained above), but also generates values for the conditional expectation which are further off owing to omission of the correction term $\mathrm{E}(U_i|V_i \leq -Z_i'\gamma, X_i, Z_i)$.

Question 2

a) Discuss the estimation of the causal effect of schooling on earnings by OLS. In particular, address whether or not it is plausible that regularity conditions for applying OLS are satisfied.\

```
model3 <- lm(logwage ~ schooling, data = data)
stargazer(model3)
```

% Table created by stargazer v.5.2.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu
% Date and time: Mo, Jan 10, 2022 - 22:20:53

Table 6:

	Dependent variable:
	logwage
schooling	0.101
	(0.073)
Constant	5.778***
	(0.512)
Observations	416
R^2	0.005
Adjusted R^2	0.002
Residual Std. Error	3.467 (df = 414)
F Statistic	1.950 (df = 1; 414)
Note:	*p<0.1; **p<0.05; ***p<0.01

Earnings usually depend on the educational level of individuals which is why it is sensible to theorize and estimate the causal effect of schooling on earnings. OLS estimation performed for all employed observed individuals suggests that this causal effect is not significant though. However, in order for OLS to yield unbiased, consistent and efficient estimation results the following regularity conditions need to hold:\

1. x_i is stochastic and $plim(\frac{1}{n}X\prime X) = Q$ of rank k\
2. ϵ_i is random with $E(\epsilon_i|x_i) = 0$ (zero conditional mean)\
3. $E[\epsilon_i^2|x_i] = Var(\epsilon_i|x_i) = \sigma^2$ (homoskedasticity)\
4. $E[\epsilon_i, \epsilon_j] = 0$ for $i \neq j$ (no autocorrelation)\
5. α, β, σ^2 are fixed and unknown.\
6. $y_i = \alpha + \beta x_i + \epsilon_i$ (linear model)\
7. $\epsilon_i \sim N(0, \sigma^2 I)$\

So it is essential that X_i is an exogenous regressor. The issue that can arise in the estimation of the causal effect of schooling on earnings is that schooling itself is an endogenous regressors. For instance, the distance to the school, the amount of financial support for schools, the educational background of the individual's parents are all potential determining factors of an individual's education. The distance is often regarded as a determining factor as individuals who live closer to a college or school tend to attain higher education levels. Subsidies for schools are meant to improve the educational quality of institutions which is why one might presume that students from highly subsidized school tend to study more and longer. Moreover, children whose parents have an academic background statistically study more often than children whose parents have a non-academic background. All these aspects support the argument that schooling actually is not an exogenous but an endogenous regressors. Consequently, the assumption of the zero conditional mean does not hold and the OLS estimation can be assumed to be both biased and inconsistent.

b) Perform instrumental variables estimation for these three options. Which option do you prefer? Include in your answer the necessary analyses and numbers on which you base your choice.\

```r
IV1 <- feols(logwage ~ 1 | schooling ~ distance, data, se = 'hetero')
```

```
## NOTE: 250 observations removed because of NA values (LHS: 250).
```

```r
IV2 <- feols(logwage ~ 1 | schooling ~ subsidy, data, se = 'hetero')
```

```
## NOTE: 250 observations removed because of NA values (LHS: 250).
```

```r
IV3 <- feols(logwage ~ 1 | schooling ~ distance + subsidy, data, se = 'hetero')
```

```
## NOTE: 250 observations removed because of NA values (LHS: 250).
```

```r
summary(IV1)
```

```
## TSLS estimation, Dep. Var.: logwage, Endo.: schooling, Instr.: distance
## Second stage: Dep. Var.: logwage
## Observations: 416
## Standard-errors: Heteroskedasticity-robust
##               Estimate Std. Error  t value Pr(>|t|)
## (Intercept)   5.788948   3.965473 1.459838  0.14509
## fit_schooling 0.099726   0.595939 0.167343  0.86718
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## RMSE: 3.45834   Adj. R2: 0.002282
## F-test (1st stage), schooling: stat = 5.64531 , p = 0.017956, on 1 and 414 DoF.
##                     Wu-Hausman: stat = 7.131e-6, p = 0.997871, on 1 and 413 DoF.
```

```r
summary(IV2)
```

```
## TSLS estimation, Dep. Var.: logwage, Endo.: schooling, Instr.: subsidy
## Second stage: Dep. Var.: logwage
## Observations: 416
## Standard-errors: Heteroskedasticity-robust
##               Estimate Std. Error  t value Pr(>|t|)
## (Intercept)   1.598811   1.687915 0.947211 0.344084
## fit_schooling 0.730369   0.252730 2.889922 0.004056 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## RMSE: 3.75869   Adj. R2: -0.178544
## F-test (1st stage), schooling: stat = 41.2    , p = 3.769e-10, on 1 and 414 DoF.
##                     Wu-Hausman: stat =  7.58581, p = 0.006143 , on 1 and 413 DoF.
```

```r
summary(IV3)
```

```
## TSLS estimation, Dep. Var.: logwage, Endo.: schooling, Instr.: distance, subsidy
## Second stage: Dep. Var.: logwage
## Observations: 416
## Standard-errors: Heteroskedasticity-robust
##               Estimate Std. Error t value  Pr(>|t|)
## (Intercept)   2.077843   1.546767 1.34335 0.1798958
## fit_schooling 0.658272   0.231424 2.84444 0.0046697 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## RMSE: 3.69585   Adj. R2: -0.139465
## F-test (1st stage), schooling: stat = 23.0     , p = 3.369e-10, on 2 and 413 DoF.
##                     Wu-Hausman: stat =  6.64186 , p = 0.010307 , on 1 and 413 DoF.
##                         Sargan: stat =  0.808893, p = 0.368448 , on 1 DoF.
```

For the first instrumental variable estimation, which considers distance as the only instrumental variable, we can see that the first-stage F-test yields a F-statistic of around 5.65. As this is below the customary threshold of 10, below which the IV estimator can be shown to possess (approximately) more than 10% of the OLS bias, this gives a clear indication that the instrument at hand is not relevant, which is why it will not be consider any further.\

The second instrumental variable estimation considers subsidy as the only instrumental variable and gives a F-statstic of around 41.2 in the first-stage regression. Hence, there is a strong indication that the instrument considered here is relevant and should be preferred over the instrument discussed before. This can also be seen by comparing the efficiency of the obtained estimators, i.e. of the coefficients' standard errors, since the loss in precision associated with IV estimation increases in the weakness of instruments.\

Analyzing the IV estimation results, when instrumented by the variable subsidy, we see that schooling has a significant and positive effect on wages. Hence, the results obtained here exemplify that inappropriate treatment of edogeneity problems can greatly affect the inferential conclusions drawn towards causal relationships studied.\

The third instrumental variable estimation includes distance and subsidy as instrumental variables and yields a F statistic of around 23. Since more instrumental variables than (potentially) endogenous variables are considered here,i.e. the model is overidentified, the Sargan test for instrument validity can be applied. Based on this test we cannot reject the null hypothesis of joint instrument exogeneity as it gives a p-value of around 0.37.\

This leaves the question whether using both variables as instrumental variables should be preferred over using only subsidy as such. Since the Sargan test only tests the joint validity of the instruments, it does not contradict the characterization of distance as a weak instrument. Since it can be shown that dropping weak instrumental variables, reduces the bias in the case of overidentification, we argue that adopting soley subsidy as an instrument is the most suitable model specification here.\

Hence, we would choose subsidy as the only instrumental variable for our estimation. \

```
IV4 <- feols(logwage ~ age + agesq | schooling ~ subsidy, data, se = 'hetero')
```

```
## NOTE: 250 observations removed because of NA values (LHS: 250).
```

```
IV5 <- feols(logwage ~ age + agesq | schooling ~ distance, data, se = 'hetero')
```

```
## NOTE: 250 observations removed because of NA values (LHS: 250).
```

```
IV6 <- feols(logwage ~ age + agesq | schooling ~ distance + subsidy, data, se = 'hetero')
```

```
## NOTE: 250 observations removed because of NA values (LHS: 250).
```

```
summary(IV4)
```

```
## TSLS estimation, Dep. Var.: logwage, Endo.: schooling, Instr.: subsidy
## Second stage: Dep. Var.: logwage
## Observations: 416
## Standard-errors: Heteroskedasticity-robust
##                Estimate Std. Error   t value   Pr(>|t|)
## (Intercept)   23.694293   8.079595  2.932609 0.00354881 **
## fit_schooling  0.400797   0.104840  3.822955 0.00015223 ***
## age           -0.233059   0.508248 -0.458554 0.64679636
## agesq         -0.013087   0.008069 -1.621749 0.10562211
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## RMSE: 1.55267   Adj. R2: 0.797916
## F-test (1st stage), schooling: stat = 43.3    , p = 1.416e-10, on 1 and 412 DoF.
##                     Wu-Hausman: stat =  3.63415, p = 0.057303 , on 1 and 411 DoF.
```

```
summary(IV5)
```

```
## TSLS estimation, Dep. Var.: logwage, Endo.: schooling, Instr.: distance
## Second stage: Dep. Var.: logwage
## Observations: 416
## Standard-errors: Heteroskedasticity-robust
##                Estimate Std. Error   t value Pr(>|t|)
## (Intercept)   22.680558   9.418521  2.408081 0.016475 *
## fit_schooling  0.469799   0.315271  1.490141 0.136952
## age           -0.192424   0.554561 -0.346984 0.728781
## agesq         -0.013813   0.008929 -1.546960 0.122641
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## RMSE: 1.60473   Adj. R2: 0.784136
## F-test (1st stage), schooling: stat = 5.37382 , p = 0.02093 , on 1 and 412 DoF.
##                     Wu-Hausman: stat = 0.844602, p = 0.358623, on 1 and 411 DoF.
```

```
summary(IV6)
```

```
## TSLS estimation, Dep. Var.: logwage, Endo.: schooling, Instr.: distance, subsidy
## Second stage: Dep. Var.: logwage
## Observations: 416
## Standard-errors: Heteroskedasticity-robust
##                Estimate Std. Error   t value   Pr(>|t|)
## (Intercept)   23.589211   8.094800  2.914119 3.7614e-03 **
## fit_schooling  0.407949   0.101388  4.023628 6.8191e-05 ***
## age           -0.228847   0.509668 -0.449012 6.5366e-01
## agesq         -0.013162   0.008092 -1.626573 1.0459e-01
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## RMSE: 1.55737   Adj. R2: 0.796689
## F-test (1st stage), schooling: stat = 23.9     , p = 1.524e-10, on 2 and 411 DoF.
##                     Wu-Hausman: stat =  4.34336 , p = 0.037771 , on 1 and 411 DoF.
##                         Sargan: stat =  0.052355, p = 0.819015 , on 1 DoF.
```

Going further, we also perform instrumental variable estimations where age and squared age are included as presumably exogenous regressors (i.e. instrumented by themselves). Although estimated values expectedly change across estimations, the obtained results yield the same conclusion that distance is not a valid instrumental variable while subsidy is. Therefore, we draw the equivalent conclusion that dropping distance as a weak instrument should be preferred to reduce the biasedness of the IV estimation.\

c) Compare the IV estimates with the OLS outcomes. Under which conditions would you prefer OLS over IV? Perform a test and use the outcome of the test to support your choice between OLS and IV. Motivate your choice.

```
stargazer(model1)
```

% Table created by stargazer v.5.2.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu
% Date and time: Mo, Jan 10, 2022 - 22:20:54

```
stargazer(model3)
```

% Table created by stargazer v.5.2.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu
% Date and time: Mo, Jan 10, 2022 - 22:20:54

```
summary(IV2)
```

TSLS estimation, Dep. Var.: logwage, Endo.: schooling, Instr.: subsidy Second stage: Dep. Var.: logwage

Table 7:

	Dependent variable:
	logwage
schooling	0.216^{***}
	(0.032)
age	-0.342
	(0.521)
agesq	-0.011
	(0.008)
Constant	26.409^{***}
	(8.057)
Observations	416
R^2	0.815
Adjusted R^2	0.813
Residual Std. Error	1.499 (df = 412)
F Statistic	604.261^{***} (df = 3; 412)
Note:	*p<0.1; **p<0.05; ***p<0.01

Table 8:

	Dependent variable:
	logwage
schooling	0.101
	(0.073)
Constant	5.778^{***}
	(0.512)
Observations	416
R^2	0.005
Adjusted R^2	0.002
Residual Std. Error	3.467 (df = 414)
F Statistic	1.950 (df = 1; 414)
Note:	*p<0.1; **p<0.05; ***p<0.01

Observations: 416 Standard-errors: Heteroskedasticity-robust Estimate Std. Error t value Pr(>|t|)
(Intercept) 1.598811 1.687915 0.947211 0.344084
fit_schooling 0.730369 0.252730 2.889922 0.004056 ** — Signif. codes: 0 ‘’ ***0.001*** ’’ *0.01* ’’ 0.05 ‘.’ 0.1 ’ ’ 1 RMSE: 3.75869 Adj. R2: -0.178544 F-test (1st stage), schooling: stat = 41.2 , p = 3.769e-10, on 1 and 414 DoF. Wu-Hausman: stat = 7.58581, p = 0.006143 , on 1 and 413 DoF.

```
summary(IV4)
```

TSLS estimation, Dep. Var.: logwage, Endo.: schooling, Instr.: subsidy Second stage: Dep. Var.: logwage Observations: 416 Standard-errors: Heteroskedasticity-robust Estimate Std. Error t value Pr(>|t|)
(Intercept) 23.694293 8.079595 2.932609 0.00354881 ** fit_schooling 0.400797 0.104840 3.822955 0.00015223 *** age -0.233059 0.508248 -0.458554 0.64679636
agesq -0.013087 0.008069 -1.621749 0.10562211
— Signif. codes: 0 ‘’ ***0.001*** ’’ *0.01* ’’ 0.05 ‘.’ 0.1 ’ ’ 1 RMSE: 1.55267 Adj. R2: 0.797916 F-test (1st stage), schooling: stat = 43.3 , p = 1.416e-10, on 1 and 412 DoF. Wu-Hausman: stat = 3.63415, p = 0.057303 , on 1 and 411 DoF. Our first OLS model includes schooling, age and age squared as regressors. Schooling is highly significant and has a positive effect on earnings. Age and age squared have a negative effect on wages but are not significant. The F test is around 604.3 so the estimator are jointly significant. The adjusted R^2 is around 0.8135.

The second OLS model include only schooling as a regressor. It has a positive effect on wages, although lower than in the other OLS estimation, and is not significant. In comparison to the previous OLS estimation, the R^2 is substantially lower at around 0.0023.

Comparing these two models, the first OLS estimation seems more sensible as its explanatory power is substantially higher.\

The instrumental variable estimation that include schooling as the only regressors and subsidy as its instrumental variable has a F statistic of 41.2. Schooling has a positive effect on wages and is significant at 0.001. The Wu-Hausman test rejects the null hypothesis of exogeneity at 0.01. Hence, there is substantial evidence for endogeneity of the regressor schooling. R^2 does not have any informative power in an instrumental variable estimation as it is only meaningful when the Gauss-Markov Theorem assumptions are satisfied which is why we do not consider it here any further.

The last estimation includes schooling as its endogenous regressors with subsidy as its instrumental variable and age and age squared as its exogenous regressors. The F statistic is around 43.3, so slightly higher than in the previous estimation. Schooling is significant at 0.0001 and has a positive effect on wages, although lower than in the previous instrumental variable estimation. The Wu-Hausman test is not as significant as in the previous estimation but we are still able to reject the null hypothesis at 10%. \

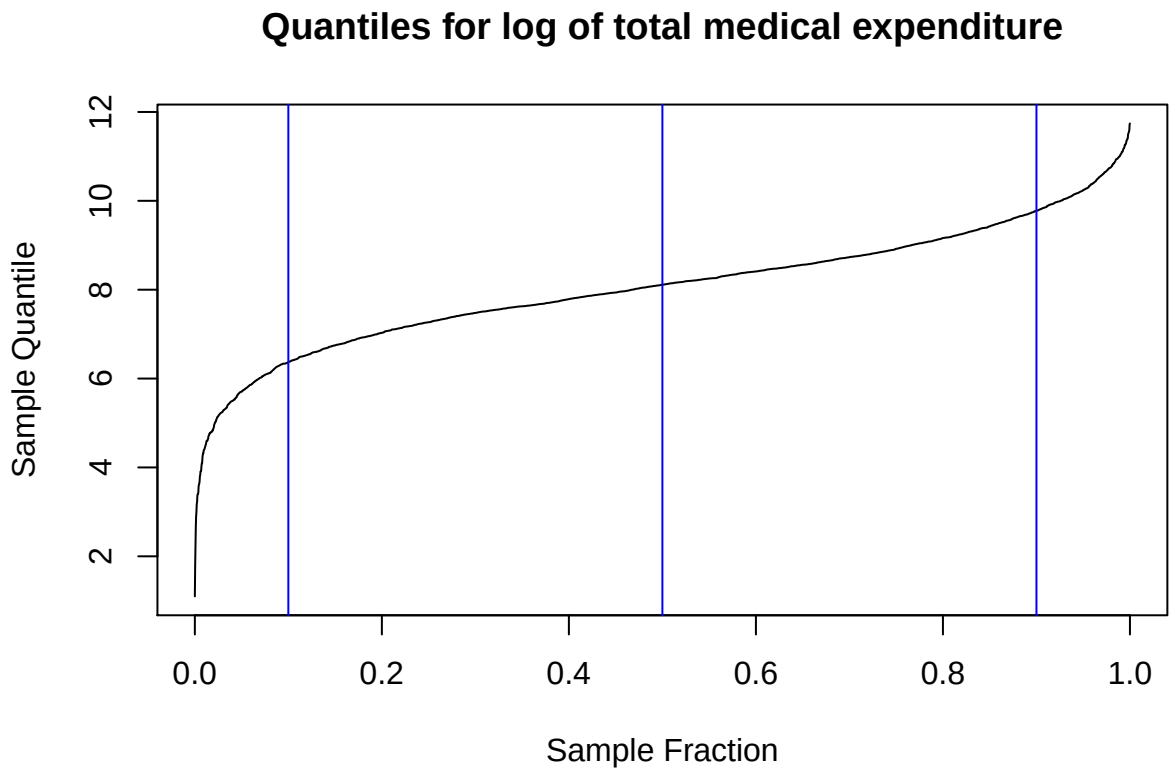
Given our results, we would prefer the IV estimation as there is evidence for the endogeneity of schooling. Further, we would choose the estimation that also includes the exogenous regressors age and age squared as the first-stage F statistic is higher than in the estimation where we only included schooling and analysis of the adjusted R^2 for OLS estimation indicates that the overall model fit can be thus considerably improved.\

Question 3

a) Create a quantile plot for the log of total medical expenditure. Draw vertical lines to indicate the median, the 10th percentile, and the 90th percentile. Describe your plot.

```
# Load data
data2 <- read.csv("assignment1b.csv")
```

```
x = data2$lntotexp
n = length(x)
plot((1:n - 1)/(n - 1), sort(x), type="l",
main = "Quantiles for log of total medical expenditure",
xlab = "Sample Fraction",
ylab = "Sample Quantile")
abline(v=c(0.1,0.5,0.9),col="blue")
```



The graph plots every sample quantile for each sample fraction which is why we have the sample fraction on the x axis and the sample quantile on the y axis. Additionally, we added a vertical line at the 10th percentile, the median and the 90th percentile. As the data is ordered for the plot, it is not surprising that the sample quantiles are increasing monotonically. Between the 0th and 10th percentile, the sample quantiles increase exponentially. From te 10th until the 90th percentile, the sample quantile only increase from approximately 6 to 10. Afterwards, they increase exponentially until 12. Hence, we find a clear indication of asymmetry, namely left-skewness, in the distribution of the considered variable of log medical expenditures. This makes intuitive sense if one presumes that there is some upper limit to total medical expenditures.\

b) Estimate the model for the quantiles $q = 0.1, 0.25, 0.5, 0.75 \text{and} 0.9$. Briefly explain your result. Compare quantile regression results to OLS estimates.

```
ols <-lm(lntotexp~ totchr + suppins + age + female + white, data = data2)
qr <- rq(lntotexp~ totchr + suppins + age + female + white, data = data2, tau = c(0.1, 0.25, 0.5, 0.75,

## Warning in rq.fit.br(x, y, tau = tau, ...): Solution may be nonunique

## Warning in rq.fit.br(x, y, tau = tau, ...): Solution may be nonunique
stargazer(ols)
```

% Table created by stargazer v.5.2.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu
% Date and time: Mo, Jan 10, 2022 - 22:20:54

Table 9:

	Dependent variable:
	lntotexp
totchr	0.445***
	(0.018)
suppins	0.257***
	(0.046)
age	0.013***
	(0.004)
female	−0.077*
	(0.046)
white	0.318**
	(0.141)
Constant	5.898***
	(0.296)
Observations	2,955
R^2	0.197
Adjusted R^2	0.196
Residual Std. Error	1.227 (df = 2949)
F Statistic	144.598*** (df = 5; 2949)
Note:	*p<0.1; **p<0.05; ***p<0.01

```
summary(qr)

## Warning in summary.rq(xi, U = U, ...): 4 non-positive fis
```

Call: rq(formula = lntotexp ~ totchr + suppins + age + female + white, tau = c(0.1, 0.25, 0.5, 0.75, 0.9), data = data2)

tau: [1] 0.1

Coefficients: Value Std. Error t value Pr(>|t|) (Intercept) 3.86704 0.48065 8.04549 0.00000 totchr 0.53919 0.02534 21.27920 0.00000 suppins 0.39572 0.07851 5.04027 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

Call: rq(formula = lntotexp ~ totchr + suppins + age + female + white, tau = c(0.1, 0.25, 0.5, 0.75, 0.9), data = data2)

tau: [1] 0.25

Coefficients: Value Std. Error t value Pr(>|t|) (Intercept) 4.74732 0.30724 15.45160 0.00000 totchr 0.45918 0.01833 25.04804 0.00000 suppins 0.38584 0.05992 6.43964 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

Call: rq(formula = lntotexp ~ totchr + suppins + age + female + white, tau = c(0.1, 0.25, 0.5, 0.75, 0.9), data = data2)

tau: [1] 0.5

Coefficients: Value Std. Error t value Pr(>|t|) (Intercept) 5.61116 0.35187 15.94656 0.00000 totchr 0.39427 0.01846 21.35942 0.00000 suppins 0.27698 0.05347 5.18025 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

Call: rq(formula = lntotexp ~ totchr + suppins + age + female + white, tau = c(0.1, 0.25, 0.5, 0.75, 0.9), data = data2)

tau: [1] 0.75

Coefficients: Value Std. Error t value Pr(>|t|) (Intercept) 6.59997 0.42690 15.46027 0.00000 totchr 0.37354 0.02286 16.33884 0.00000 suppins 0.14885 0.06203 2.39991 0.01646 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

Call: rq(formula = lntotexp ~ totchr + suppins + age + female + white, tau = c(0.1, 0.25, 0.5, 0.75, 0.9), data = data2)

tau: [1] 0.9

Coefficients: Value Std. Error t value Pr(>|t|) (Intercept) 8.32264 0.54599 15.24309 0.00000 totchr 0.35795 0.03310 10.81289 0.00000 suppins -0.01428 0.08642 -0.16527 0.86874 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 Just as for variables such as wealth and earnings (with which medical expenditures might correlate to some degree), one can saliently assert that causal effects of considered factors determining the level of total medical expenditures are unstable across quantiles of the distribution of medical expenditures. This is exactly what the considered quantile regression results suggest for several regressors which do not only differ in the magnitude of estimated effects (e.g. that of the number of chronic problems) but also in their significance of effects (e.g. that of the existence of supplementary private insurance). Since discussion of these differences is greatly aided by the graphical presentation produced in c) further discussions of these findings are provided alongside below.\

These differences also manifest in the differences between quantile regression results and OLS estimates. By construction OLS estimates disregard any differences in effects across the distrubtion of the dependent variable and would only resemble quantile regression results if such differences would not exist. Thus, a more detailed discussion of the exact differences in parameter estimates seems little insightful. However, it may be noteworthy to point out the difference in OLS estimates, concerned with the conditional mean, to those of the 0.5 quantile, concerned with the conditional median. Clearly the difference in results is due to the left-skewness of the dependent variable data outlined above, namely that the median of the log of total medical expenditures is bigger than the mean.\

```
mean(data2$lntotexp)
```

```
## [1] 8.059866
```

```
median(data2$lntotexp)
```

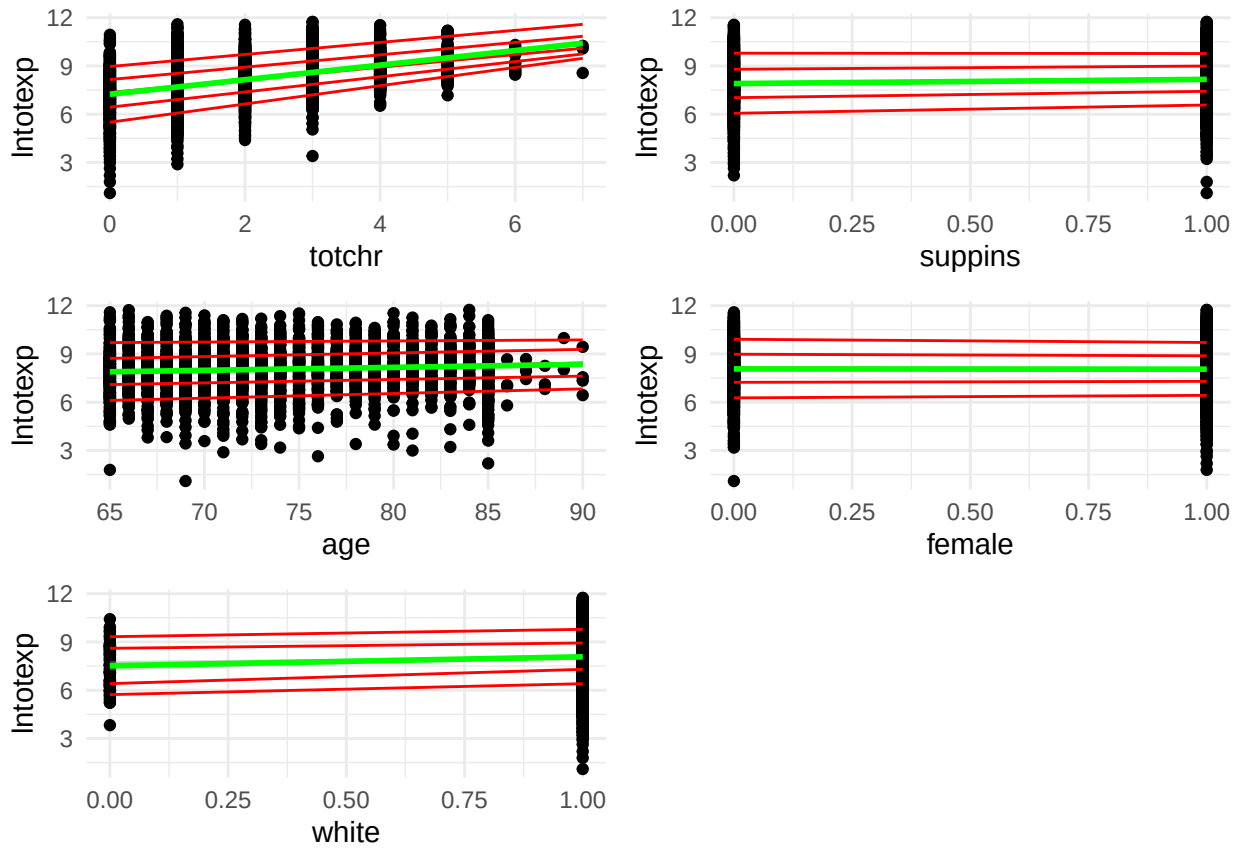
```
## [1] 8.111928
```

```
ggp1 <- ggplot(data2, aes(totchr, lntotexp)) +
  theme_minimal() +
```

```
  geom_point() +
  geom_quantile(quantiles = c(0.1, 0.25, 0.5, 0.75, 0.9), col="red") +
  geom_smooth(method = "lm", col="green")

ggp2 <- ggplot(data2, aes(suppins, lntotexp)) +
  theme_minimal() +
  geom_point() +
  geom_quantile(quantiles = c(0.1, 0.25, 0.5, 0.75, 0.9), col="red") +
  geom_smooth(method = "lm", col="green")

ggp3 <- ggplot(data2, aes(age, lntotexp)) +
  theme_minimal() +
  geom_point() +
  geom_quantile(quantiles = c(0.1, 0.25, 0.5, 0.75, 0.9), col="red") +
  geom_smooth(method = "lm", col="green")

ggp4 <- ggplot(data2, aes(female, lntotexp)) +
  theme_minimal() +
  geom_point() +
  geom_quantile(quantiles = c(0.1, 0.25, 0.5, 0.75, 0.9), col="red") +
  geom_smooth(method = "lm", col="green")

ggp5 <- ggplot(data2, aes(white, lntotexp)) +
  theme_minimal() +
  geom_point() +
  geom_quantile(quantiles = c(0.1, 0.25, 0.5, 0.75, 0.9), col="red") +
  geom_smooth(method = "lm", col="green")

grid.arrange(ggp1, ggp2, ggp3, ggp4, ggp5, ncol = 2)
```

```
## Smoothing formula not specified. Using: y ~ x
## Warning in rq.fit.br(wx, wy, tau = tau, ...): Solution may be nonunique
## Warning in rq.fit.br(wx, wy, tau = tau, ...): Solution may be nonunique
## Warning in rq.fit.br(wx, wy, tau = tau, ...): Solution may be nonunique
## `geom_smooth()` using formula 'y ~ x'
## Smoothing formula not specified. Using: y ~ x
## Warning in rq.fit.br(wx, wy, tau = tau, ...): Solution may be nonunique
## Warning in rq.fit.br(wx, wy, tau = tau, ...): Solution may be nonunique
## Warning in rq.fit.br(wx, wy, tau = tau, ...): Solution may be nonunique
## `geom_smooth()` using formula 'y ~ x'
## Smoothing formula not specified. Using: y ~ x
## Warning in rq.fit.br(wx, wy, tau = tau, ...): Solution may be nonunique
## `geom_smooth()` using formula 'y ~ x'
## Smoothing formula not specified. Using: y ~ x
## Warning in rq.fit.br(wx, wy, tau = tau, ...): Solution may be nonunique
## `geom_smooth()` using formula 'y ~ x'
```

```
## Smoothing formula not specified. Using: y ~ x
## Warning in rq.fit.br(wx, wy, tau = tau, ...): Solution may be nonunique
## `geom_smooth()` using formula 'y ~ x'
```



Though plotting fitted values for quantile regression vs. OLS regression estimates for all considered regressors gains little insights into the discusssion above, plots are provided here to informally conclude that fitted quantiles do not cross.\

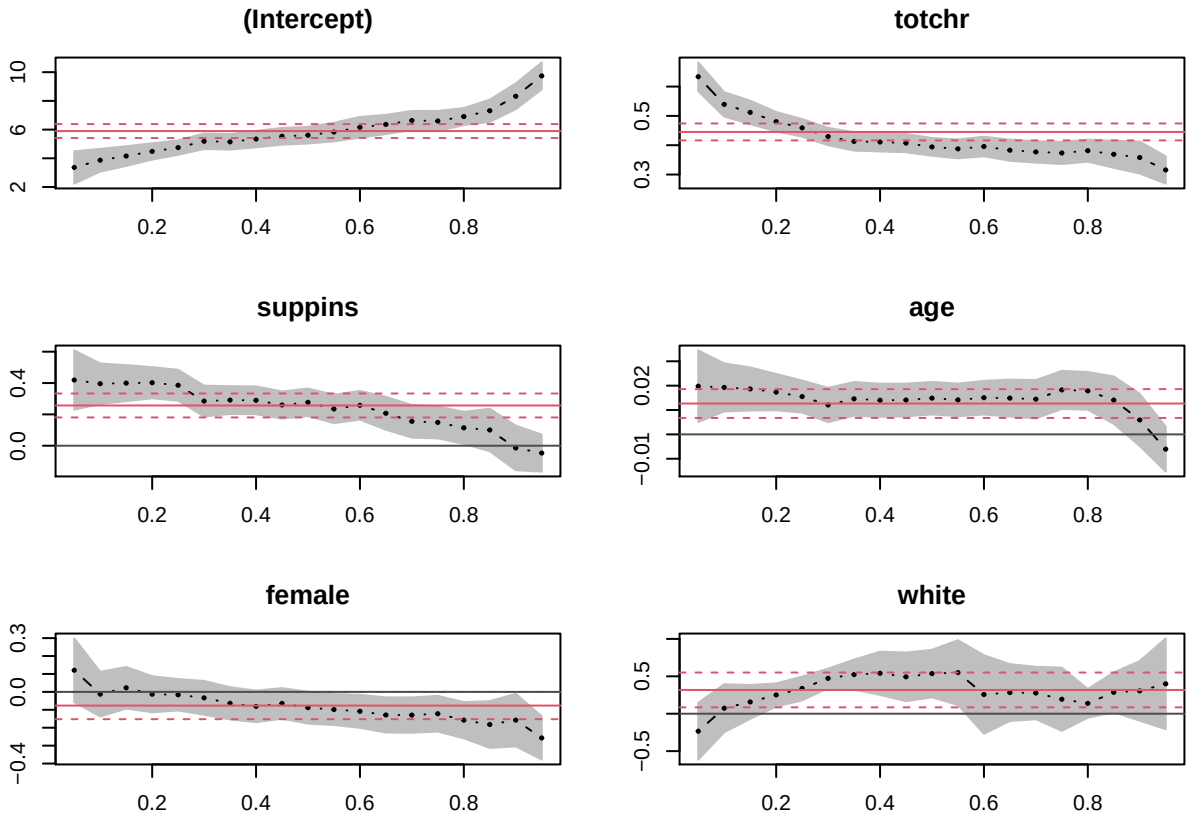
c) Graph the estimated coefficients from the quantile regressions for $q$from 0.05to0.95 in increments of 0.05, together with their 95% confidence interval and the corresponding estimates from a linear regression (and their 95 confidence interval). Discuss your findings.

```
qr2 <- rq(lntotexp~ totchr + suppins + age + female + white, data = data2,  tau = seq(0.05, 0.95, by = 
```

```
## Warning in rq.fit.br(x, y, tau = tau, ...): Solution may be nonunique

## Warning in rq.fit.br(x, y, tau = tau, ...): Solution may be nonunique

## Warning in rq.fit.br(x, y, tau = tau, ...): Solution may be nonunique
```

```
plot(summary(qr2))
```

```
## Warning in summary.rq(xi, U = U, ...): 1 non-positive fis
## Warning in summary.rq(xi, U = U, ...): 4 non-positive fis

## Warning in summary.rq(xi, U = U, ...): 4 non-positive fis
## Warning in summary.rq(xi, U = U, ...): 2 non-positive fis
```



The graphs show that the estimates for different quantiles differ across the quantiles and therefore do not align with the OLS estimator although the magnitude differs among the various regressors.\

The impact of the number of chronic problems is higher for lower sample fractions than estimated by OLS. It monotonically decreases for increasing sample fractions and is lower than estimated by OLS from around the 40th percentile on. As one would expect, this effect is estimated to be (very) significant and positive across the entire distribution. Chronic problems expectedly result in higher medical expenditures unconditional on whether general expenditures are high. To make sense of the fact that the estimated coeficient is decreasing in the considered percentile, we note that total medical expenditures are measured in logs so that they have to be interpreted as indicating the relative increase in expenditues associated with an increase in regressor variables. Hence, if expenditures are already very high, the effect of chronic medical problems is lower in relative terms.

A similar dynamic can be observed for the supplementary private insurance. For lower sample fractions a supplementary private insurance has a higher positive effect on the total medical expenditures than for higher sample fractions. From the 80th percentile on the effect is even negative. Until around the 40th percentile the effect is higher than estimated by OLS. Moreover, the significance of the effect decreases in the considered percentile so that the effect is no longer sigificant from the 0.85 quantile onwards. The insignificance for higher percentiles can be explained by the fact that the impact of an increasing probability that individuals have a supplementary private insurance is lower when the probability is already high.

In case of the regressor age, the OLS and quantile estimation align from the 30th to the 70th percentile. Most natably, the effect is not estimated to be efficient for the top 15 percentiles. This may be due to the fact that very high levels of medical expenditure are associated with very severe medical diseases whose costs are likely to be greatly invariant towards the individual's age.

Along these lines, one could also make sense of the finding that being female only has a significant negative effect on medical expenditures for high percentiles if one assumes that costs associated with severe and thus costly diseases systemically differs across sex.

Lastly, the quantile regression output suggests the variable white to have a significantly positive effect on medical expenditures for intermediate percentiles. In contrast to the other effects, it seems rather unintuitive to construct an argument for this notion and is left to an econometrician who is more familiar with the pecularities of systemic ethnic differences in the country considered here.

References

Chun, H., and Lee, I. (2001). Why do married men earn more: Productivity or marriage selection?, Economic Inquiry, 39(2), 307-319.\

Greene, W. H. (2003). Econometric analysis. 5th edition, Pearson Education India.\

Pollmann-Schult, M. (2011). Marriage and Earnings: Why Do Married Men Earn More than Single Men?, European Sociological Review, Volume 27, Issue 2,April 2011, Pages 147–163.\