MOEC0021 - Empirical Methods

Group BlancSchneiderMazidi

Fabienne Blanc (15-732-142)

Flavio Schneider (15-716-202)

Manuel Mazidi (15-704-984)

Course

Empirical Methods

Prof. Greg Crawford

University of Zurich

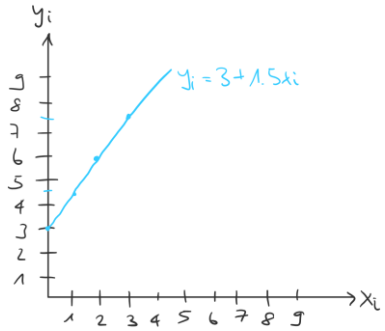
Submitted on October 22, 2018

1 Pencil and Paper Questions

Exercise 1

(a) Population regression function: $\boldsymbol{y_i = E(y_i|\, x_i) + \epsilon_i = 3 + 1.5 * x_i + \epsilon_i}$

Picture of $E(y_i|\, x_i)$, the non-random part of the population regression function:



(b) 4 observations:

	X	Y
1	1	4
2	4	10
3	3	9
4	2	7

Mean of X and Mean of Y:

$\bar{X}$	$\bar{Y}$
2.5	7.5

X and Y in mean-deviation form:

	$x_i = (X_i - \bar{X})$	$y_i = (Y_i - \bar{Y})$
1	-1.5	-3.5
2	1.5	2.5
3	0.5	1.5
4	-0.5	-0.5

Cross-product $x_i y_i$ and square of x_i in mean-deviation form:

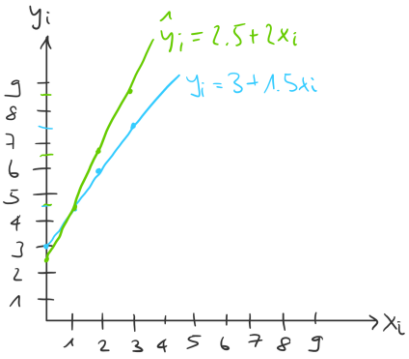
	$x_i y_i$	x_i^2
1	5.25	2.25
2	3.75	2.25
3	0.75	0.25
4	0.25	0.25

(c) Calculation of the OLS estimates of β_1 and β_2 (i.e. $\hat{\beta}_1$ and $\hat{\beta}_2$):

$$\hat{\boldsymbol{\beta}}_{\mathbf{2}} = \frac{\sum(x_i - \bar{X}) * (y_i - \bar{Y})}{\sum(x_i - \bar{X})^2} = \frac{\sum x_i y_i}{\sum {x_i}^2} = \frac{5.25 + 3.75 + 0.75 + 0.25}{2.25 + 2.25 + 0.25 + 0.25} = \mathbf{2}$$

$$\hat{\boldsymbol{\beta}}_{\mathbf{1}} = \bar{Y} - \hat{\beta}_2 * \bar{X} = 7.5 - 2 * 2.5 = \mathbf{2.5}$$

(d) Plot OLS line of the sample regression function: $\hat{\boldsymbol{y}}_{\boldsymbol{i}} = \mathbf{2.5} + \mathbf{2} * \boldsymbol{x}_{\boldsymbol{i}}$ **(in green)**



(e) The OLS line of the sample regression function has a different intercept than the population regression function. The sample regression cuts the y-axis at 2.5, whereas the population regression function cuts the y-axis at 3. However, the sample regression is determined by a $\hat{\beta}_2$ of 2.5 and is therefore steeper than the population regression function with a β_2 1.5.

(f) The OLS line of the sample regression function does cross the population regression function because it starts at a lower intercept and is at the same time stepper than the population regression function. It follows that they will cross each other at any point.

If another sample is selected from the same population, it is likely that different OLS estimates of β_1 and β_2 would result. Hence, it would be possible that the line of the first sample (sample A) and the line of the second sample (sample B) would <u>not</u> cross. This is the case in the following three scenarios:

(1) $\begin{cases} \hat{\beta}_{1,SampleA} > \hat{\beta}_{1,SampleB} \\ \hat{\beta}_{2,SampleA} > \hat{\beta}_{2,SampleB} \end{cases}$ i.e. line of sample A starts higher and is steeper than the line B

(2) $\begin{cases} \hat{\beta}_{1,SampleA} < \hat{\beta}_{1,SampleB} \\ \hat{\beta}_{2,SampleA} < \hat{\beta}_{2,SampleB} \end{cases}$ i.e. line of sample A starts lower and is less steep than the line B

(3) $\begin{cases} \hat{\beta}_{1,SampleA} \neq \hat{\beta}_{1,SampleB} \\ \hat{\beta}_{2,SampleA} = \hat{\beta}_{2,SampleB} \end{cases}$ i.e. both lines are parallel, but do not have the same intercept

The two lines would fall together, if $\hat{\beta}_{1,SampleA} = \hat{\beta}_{1,SampleB}$ and $\hat{\beta}_{2,SampleA} = \hat{\beta}_{2,SampleB}$.

(g) The errors are calculated by subtracting the Y_i of the population regression function from the observation Y. The residuals are calculated by subtracting the $\hat{y}_i$ from the observation.

X	Y	$Y_i = 3 + 1.5 * X_i$	Errors: $\epsilon_i = Y - Y_i$
1	4	4.5	-0.5
4	10	9	1
3	9	7.5	1.5
2	7	6	1

X	Y	$\hat{y}_i = 2.5 + 2 * x_i$	Residuals: $e_i = Y - \hat{y}_i$
1	4	4.5	-0.5
4	10	10.5	-0.5
3	9	8.5	0.5
2	7	6.5	0.5

The sum of the errors is 3. If all data points of this population would be observed, the errors would sum up to zero since the population regression function is a reflection of the entire population. However, in this example, only 4 observations are available.

The sample regression function does not correspond to the population regression function. Consequently, the residuals do not correspond to the errors. The sum of the residuals is 0. Therefore, the mean-zero error assumption (assumption 2 of the lecture) is valid.

(h) Show that $\sum_{n=1}^{4}(x_i - \bar{X}) = 0$. Intuition: This property should hold in every sample (sample with size n = 1, …, N) because the mean is exactly the middle of the sample distribution, i.e. the sum of all differences between the observation and the mean should equal to zero.

$$\sum_{n=1}^{N}(x_i - \bar{X}) = \sum_{n=1}^{N}\left(x_i - \frac{\sum x_i}{n}\right)$$

$$= \left(x_1 - \frac{\sum x_i}{n}\right) + \left(x_2 - \frac{\sum x_i}{n}\right) + \cdots + \left(x_N - \frac{\sum x_i}{n}\right)$$

$$= (x_1 + x_2 + \cdots + x_N) - N * \left(\frac{\sum x_i}{n}\right)$$

$$= \sum_{n=1}^{N}(x_i) - \sum_{n=1}^{N}(x_i)$$

$$= 0$$

(i) Show that $\sum(x_i y_i) = \sum(x_i Y_i) = \sum(X_i y_i)$. This property should hold in every sample.

$$\sum(x_i y_i) = \sum(X_i - \bar{X}) * (Y_i - \bar{Y})$$

$$= \sum(X_i Y_i) - \sum(X_i \bar{Y}) - \sum(\bar{X} Y_i) + \sum(\bar{X}\bar{Y})$$

$$= \sum(X_i Y_i) - n * \bar{X}\bar{Y} - n * \bar{X}\bar{Y} + n * \bar{X}\bar{Y}$$

$$= \sum(X_i Y_i) - n * \bar{X}\bar{Y}$$

$$= \sum(X_i Y_i) - n * \bar{X}\frac{\sum Y_i}{n} \qquad = \sum(X_i Y_i) - n * \bar{Y}\frac{\sum X_i}{n}$$

$$= \sum(Y_i) * \sum(X_i - \bar{X}) \qquad = \sum(X_i) * \sum(Y_i - \bar{Y})$$

$$= \sum[(X_i - \bar{X}) * (Y_i)] \qquad = \sum[(Y_i - \bar{Y}) * (X_i)]$$

$$= \sum(x_i Y_i) \qquad = \sum(X_i y_i)$$

Exercise 2

(a) We use the same notation as in exercise 1.1 (b) where $\boldsymbol{x_i}$ is defined as $(X_i - \bar{X})$ and $\boldsymbol{y_i}$ is defined as $(Y_i - \bar{Y})$. Thus, the formula on slide 4 (in *The Classical Linear Regression Model Foundations*) can be written as follows:

$$\hat{\beta}_2 = \frac{\sum x_i y_i}{\sum x_i^2}$$

In exercise 1.1 (i) we showed that $\Sigma x_i y_i = \Sigma x_i Y_i$. There fore we subsitute this in the formula above.

$$\hat{\beta}_2 = \frac{\sum x_i Y_i}{\sum x_i^2}$$

Y_i is defined as the population regression function $\beta_1 + \beta_2 X_i + \epsilon_i$.

$$\hat{\beta}_2 = \frac{\sum x_i(\beta_1 + \beta_2 X_i + \epsilon_i)}{\sum x_i^2}$$

$$= \beta_1 \frac{\sum x_i}{\sum x_i^2} + \beta_2 \frac{\sum x_i X_i}{\sum x_i^2} + \frac{\sum x_i \epsilon_i}{\sum x_i^2}$$

where $\frac{\sum x_i}{\sum x_i^2} = 0$ and $\frac{\sum x_i X_i}{\sum x_i^2} = 1$

$$= \beta_2 + \frac{\sum x_i \epsilon_i}{\sum x_i^2}$$

$$E[\hat{\beta}_2] = E\left[\beta_2 + \frac{\sum x_i \epsilon_i}{\sum x_i^2}\right]$$

$$= E[\beta_2] + E\left[\frac{\sum x_i \epsilon_i}{\sum x_i^2}\right]$$

$$= \beta_2 + \frac{\sum x_i [\epsilon_i | X_i]}{\sum x_i^2}$$

$$= \beta_2 + \frac{\sum x_i \mu_\epsilon}{\sum x_i^2}$$

Therefore, we see that $\hat{\beta}_2$ is biased $E[\hat{\beta}_2] \neq \beta_2$ as the mean of the error is not zero anymore.

For $\hat{\beta}_1$ we assume that the average $\bar{Y}$ is given as $\beta_1 + \beta_2 \bar{X} + \bar{\epsilon}$. The estimate $\hat{\beta}_1$ is given as

$$\hat{\beta}_1 = \bar{Y} - \hat{\beta}_2 \bar{X}$$

We substitute $\bar{Y}$:

$$\hat{\beta}_1 = \beta_1 + \beta_2 \bar{X} + \bar{\epsilon} - \hat{\beta}_2 \bar{X}$$

which can be written as

$$\hat{\beta}_1 = \beta_1 + (\beta_2 - \hat{\beta}_2)\bar{X} + \bar{\epsilon}$$

$$E[\hat{\beta}_1 | X_i] = E[\beta_1 + (\beta_2 - \hat{\beta}_2)\bar{X} + \bar{\epsilon}]$$

$$= E[\beta_1] + E[\beta_2 - \hat{\beta}_2 | X_i]\bar{X} + E[\bar{\epsilon} | X_i]$$

$$= \beta_1 + (E[\beta_2] - E[\hat{\beta}_2])\bar{X} + E[\bar{\epsilon} | X_i]$$

$$= \beta_1 + (\beta_2 - E[\hat{\beta}_2])\bar{X} + E[\bar{\epsilon} | X_i]$$

If $E[\hat{\beta}_2] \neq \beta_2$ as we derived above and the mean error is not zero, then $\hat{\beta} \neq \beta_1$.

Therefore $\hat{\beta}$ is biased.

(b) We use the formula on slide 5 (in *The Classical Linear Regression Model Foundations*) for $\widehat{\beta}$:

$$\hat{\beta} = (X'X)^{-1}X'y$$

We define $\tilde{X} = 2X$

$$\tilde{\beta} = (\tilde{X}'\tilde{X})^{-1}\tilde{X}'y$$

$$= (2X'2X)^{-1}2X'y$$

$$= \frac{1}{2}(X'X)^{-1}X'y$$

whereas we can substitute $(X'X)^{-1}X'y$ from the first equation and therefore derive

$$\tilde{\beta} = \frac{1}{2}\hat{\beta}$$

(c) We start with the same equation as in (b) ($\hat{\beta} = (X'X)^{-1}X'y$) and define $y^* = 2y$

$$\beta^* = (X'X)^{-1}X'y^*$$

$$= (X'X)^{-1}X'2y$$

$$= 2(X'X)^{-1}X'y$$

whereas we again substitute and the following equation results:

$$\beta^* = 2\,\hat{\beta}$$

(d) A change in unit is a linear transformation which does not affect the results as they only scale the estimates but do not change the underlying relationship between y_i and x_i.

(e) We use the formula on slide 5 (in *The Classical Linear Regression Model Foundations*) for $\widehat{\beta}$:

$$V(\hat{\beta}) = \sigma^2(X'X)^{-1}$$

and we define $\tilde{X} = 2X$.

$$V(\tilde{\beta}) = \sigma^2(\tilde{X}'\tilde{X})^{-1}$$

$$= \sigma^2(2X'2X)^{-1}$$

$$\frac{1}{4}\sigma^2(X'X)^{-1}$$

which can be written as

$$V(\tilde{\beta}) = \frac{1}{4}\,V(\hat{\beta})$$

2 Computer Questions

Exercise 1

(a)-(d) For each different sample size of N, we will provide the corresponding histogram first and thereafter show the across-replication average, $\overline{x}$, and sample standard deviation, $s_{\overline{x}}$.

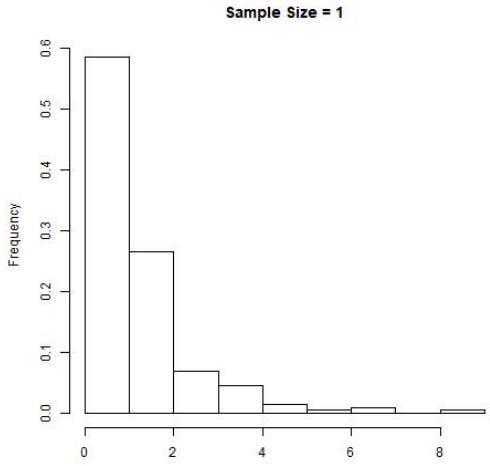


Table 1: histogram if N = 1

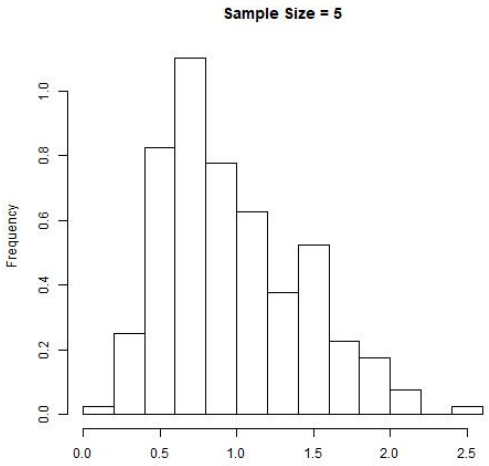


Table 2: histogram if N = 5

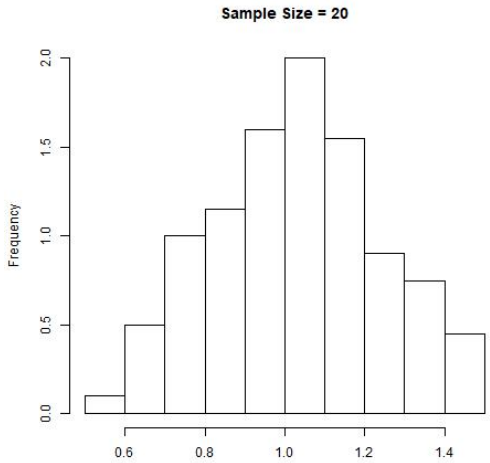


Table 3: histogram if N = 20

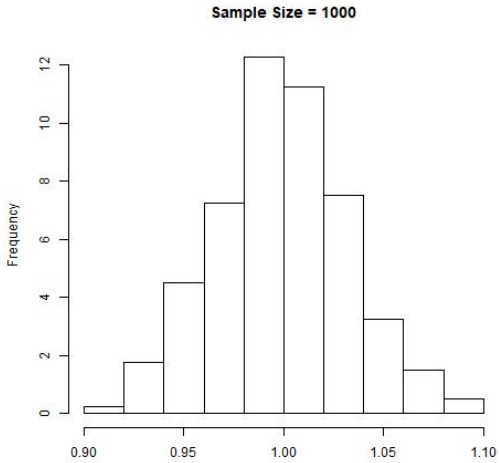


Table 4: histogram if N = 1000

sample size	*mean* $\bar{x}$	*standard deviation* $s_{\bar{x}}$
1	**1.1579**	**1.2793**
5	**1.0397**	**0.4179**
20	**1.0115**	**0.2306**
1000	**0.9985**	**0.0336**

Table 5: means and standard deviation across different sample sizes

(e) i) For *N = 1* we can clearly see an exponential distribution showing the first quadrant of a mathematical function like $\frac{1}{x}$.

For *N = 5* the distribution looks more normally distributed but still shows clear signs of positive skewness.

For *N = 20* the distribution clearly resembles a normal distribution.

For *N = 1000* the distribution looks like a normal distribution.

The generalized central limit theorem in accordance with the law of large numbers states that as the number of variables grows, the average tends towards a normal distribution.

(e) ii) The calculated estimate mean $\bar{x}$ tends to get closer to its expected value of 1. With increasing numbers our estimate gets closer to its expected value since we can apply the law of large numbers.

(e) iii)

The calculated estimate standard deviation $s_{\bar{x}}$ is only in the first experiment close to 1. With increasing values of *N* the standard deviation gets lower. With the experiments resembling a normal distribution more, the standard deviation shrinks towards 0. Again, we can apply the law of large numbers in accordance with the generalized central limit theorem, which states that as we increase *N* the standard deviation and variance move closer to 0.

Exercise 2

(a) In the whole dataset there are 807 observations.

(b)

	cigs	*educ*	*age*	*income*	*white*	*restaurn*
Minimum	0	6	17	500	0	0
1st Quartile	0	10	28	12500	1	0
Median	0	12	38	20000	1	0
Mean	8.7	12.5	41.2	19305	0.88	0.25
3rd Quartile	20	13.5	54	30000	1	0
Maximum	80	18	88	30000	1	1

Table 6: summary of selected variables

Immediately some things can be interpreted from this summary. For example, consisted the group from which the data is observed only of mostly middle-aged adults. Since the consummation of cigarettes is only allowed above 18 years of age, the variable *educ* may be restricted by the variable *age* if the people are too young and most underage people don't have an observable income, this does make sense. Since the maximum and 3rd quartile for *income* are the same, we can assume that this variable has been observed by creating buckets of a certain range of income and the observations have been sorted accordingly. This way the common outliers in observations of an individual's income are eliminated or smoothed out.

(c) i) $\widehat{\boldsymbol{\beta}}_0 = 11.41203$ *and* $\widehat{\boldsymbol{\beta}}_1 = -0.21855$

(c) ii) The estimators show the same values as calculated before. We can also see, that the estimator for β_o is statistically significant different from 0 on an α-level of 1%, while the estimator of β_1 is not.

(c) iii) If the mean-zero error assumption holds, we speak of an unbiased estimator. Therefor we can say that an average person smokes 0.21855 cigarettes less with every additional of school ut attended. We do think that such a relation does make sense, but is rather small, considering the maximum value of *educ* is not very big. Additionally, the result is not statistically significant different from 0, so we can't make a causal statement in regard of this variable.

(c) iv) $\widehat{cigs}_i = 11.41203 - 0.21855 * educ_i$

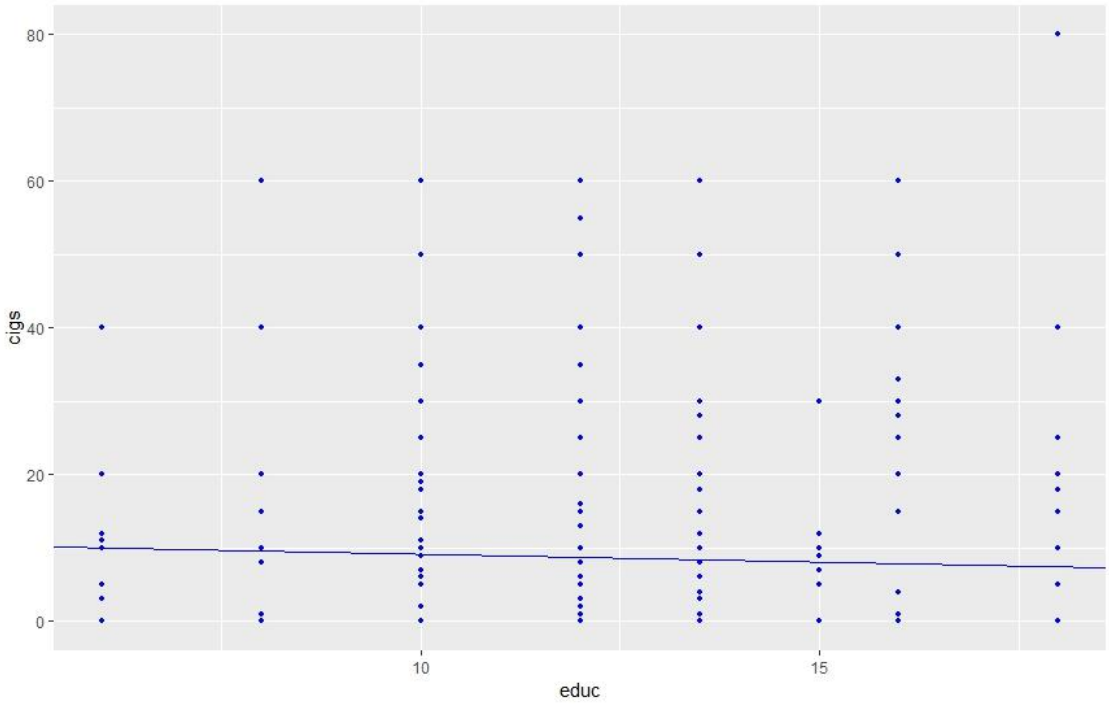


Table 7: scatterplot and regression iv

(c) v) Not using a constant, changes the algebraic sign of the estimator of β_1 which is now positive and has a value of 0.64473. We do think it makes sense to include a constant in the regression. If we don't use one, we would exclude the hypothetical possibility of an individual who never attended school to smoke because the regression must cross the origin point. This restriction does not necessarily make sense and wo don't let the data tell its story but force it to tell ours.

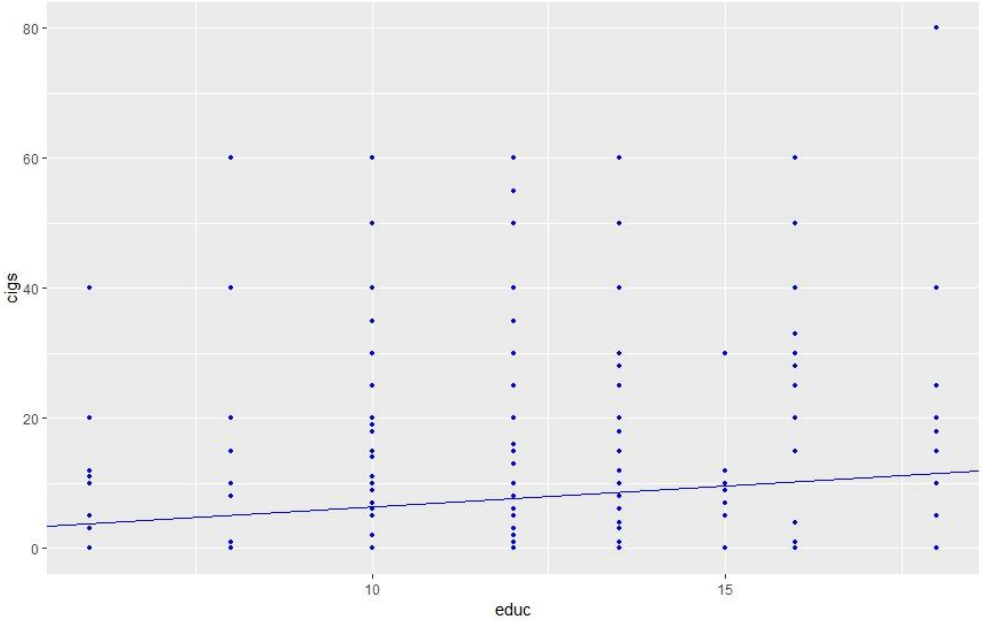


Table 8: scatterplot and regression v

(d) i) Estimated coefficient *white* = -0.624 → This would indicate a person smokes 0.624 cigarettes less per day if its from a white ethnicity, holding every other variable constant. We don't think this makes sense in general. The coefficient is not statistically significant different from 0 and in respect to the mean consumed cigarettes consumed per day very small and does not have a large impact on an explicit estimation.

Estimated coefficient *restaurn* = -2.796 → This would indicate a person smokes 2.796 cigarettes less per day if it lives in a state where it is prohibited to smoke in restaurants, holding every other variable constant. We think this relation makes sense economically. The impact is large compared to the mean of consumed cigarettes and statistically significant on an α-level of 10%.

(d) ii) The marginal effect according to varying values of *age* is calculated by deriving the regression formula in respect to the age-variable: $cigs' = 0.826 + 2 * (-0.01) * age$

age	20	40	60
marginal effect	**0.426**	**0.026**	**-0.374**

(d) iii) A)

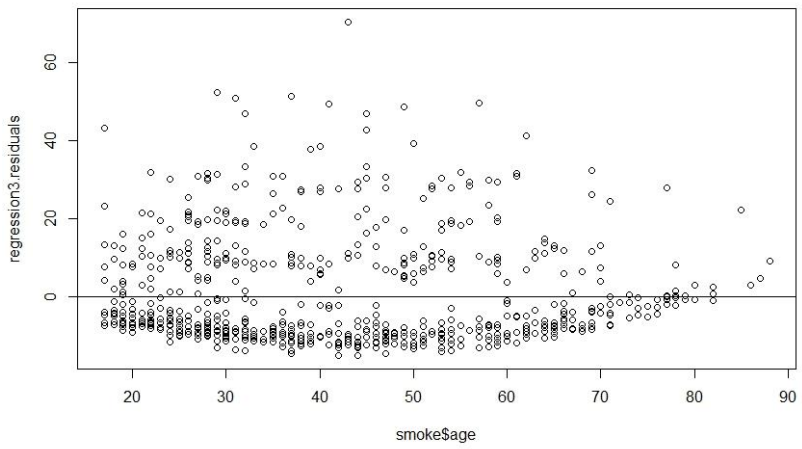


Table 9: scatterplot of residuals$age

As we can see the variance of the residuals is more or less constant throughout the observations. Towards the upper end of the spectrum, the observations narrow down and the residuals don't scatter as much as in the middle section. But this can be explained due to the fact, that we have less observations for higher ages. Therefore, we can state that our assumption 3 holds.

(d) iii) B) The calculated correlation across individuals equals to -2.103e^{-17}. Since the correlation is very close to 0, we can state that our assumption 4 holds.

(d) iii) C) The histogram of the residuals shows a positive skewness and compared to a normal distribution (black line) a negative mean. Therefore, the assumption 5 does not hold. The residuals are not normally distributed which means that in this case we would mostly overestimate the consumption of cigarettes for specific values of *age*.

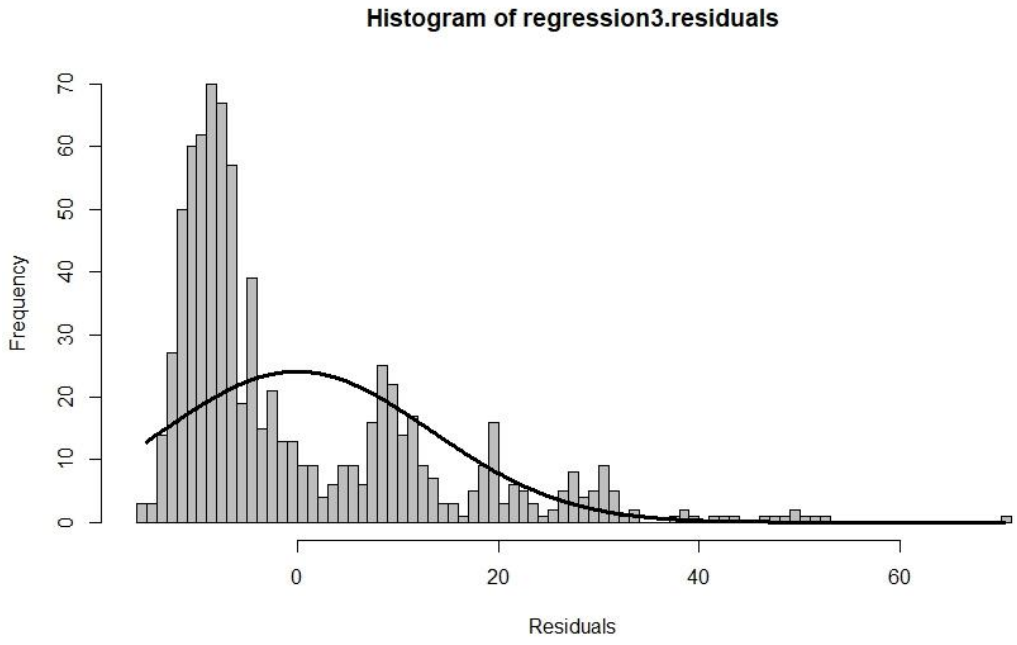


Table 10: histogram of residuals$age

3 R Code

```
setwd("C:/Users/Flavio Schneider/Desktop/uzh/Unterlagen UZH/Master/Semeste
r VII/Empirical Methods/PS1")

#Computer Question 1

set.seed(42)
sim <- 200 #number of repetitions
size <- c(1,5,20,1000) #given sample sizes N we want to use
means <- rep(NA, length(size))#containers for means
sd <- rep(NA, length(size))#containers for standard deviations

#Matrix to store the replicated means in each row for each of the sample s
izes N
MMatrix <- matrix(NA, nrow = sim, ncol = length(size))

for (j in 1:length(size)) {
```

```
  for (i in 1:sim) {
    x <- rexp(n = size[j], rate=1) #replicate single values of exponential
function
  MMatrix[i,j] <- mean(x)
  }
means[j] <- mean(MMatrix[ , j])
sd[j] <- sd(MMatrix[ , j])
}

par(mfcol = c(1,2))

for (i in 1:4) {
  jpeg(paste("graphs_",i,".jpg"))
  histogram <- hist(MMatrix[ ,i],
                    freq=FALSE,
                    ylab="Frequency", xlab="",
                    main=paste("Sample Size =", size[i]))
  dev.off()
}

print(means)

print(sd)



#Computer Question 2


library(foreign)

smoke <- read.dta("http://fmwww.bc.edu/ec-p/data/wooldridge/smoke.dta")

# 2b) summary of variables and defining variables for later
variables <- subset(smoke, select=c(cigs, educ, age, income, white, restau
rn))
summary(variables)

cigs <- subset(smoke, select=c(cigs))

educ <- subset(smoke, select=c(educ))

age <- subset(smoke, select=c(age))

income <- subset(smoke, select=c(income))

white <- subset(smoke, select=c(white))

restaurn <- subset(smoke, select=c(restaurn))


#2c) estimate B1
cor(cigs,educ)

sqrt(var(cigs))

sqrt(var(educ))
```

```
B1 <- cor(cigs,educ)*(sqrt(var(cigs))/sqrt(var(educ)))
print(B1)

# estimate Bo
colMeans(cigs, na.rm = FALSE, dims = 1)

colMeans(educ, na.rm = FALSE, dims = 1)

B0 <- colMeans(cigs, na.rm = FALSE, dims = 1)-(B1*colMeans(educ, na.rm = F
ALSE, dims = 1))
print(B0)

#2c) ii
regression1 <- lm(cigs ~ educ, data = smoke)
summary(regression1)

#2c) iv plot
library(ggplot2)
ggplot(smoke, aes(x=educ, y=cigs)) + geom_point(size=1, col="blue") + geom
_abline(slope = -0.2185521, intercept = 11.41203, col="blue")
```

```
#2c) v
regression2 <- lm(formula = cigs ~ educ -1, data = smoke)
summary(regression2)

par(mfrow = c(1, 2))
ggplot(smoke, aes(x=educ, y=cigs)) + geom_point(size=1, col="blue") + geom
_abline(slope = 0.64473, col="blue")
```

```
#2d)

regression3 <- lm(cigs ~ educ+age+agesq+white+restaurn, data = smoke)
summary(regression3)

#iii
#A
regression3 <- lm(cigs ~ educ+age+agesq+white+restaurn, data = smoke)
regression3.residuals <- resid(regression3)
plot(smoke$age, regression3.residuals)
abline(0,0)

#B
cor(regression3.residuals, smoke$age)

#C
h <-hist(regression3.residuals, breaks = 100,
     col = "gray", xlab = "Residuals")
xfit<-seq(min(regression3.residuals),max(regression3.residuals),length=100
0)
yfit<-dnorm(xfit,mean=mean(regression3.residuals),sd=sd(regression3.residu
als))
yfit <- yfit*diff(h$mids[1:2])*length(regression3.residuals)

lines(xfit, yfit, col="black", lwd=3)
```