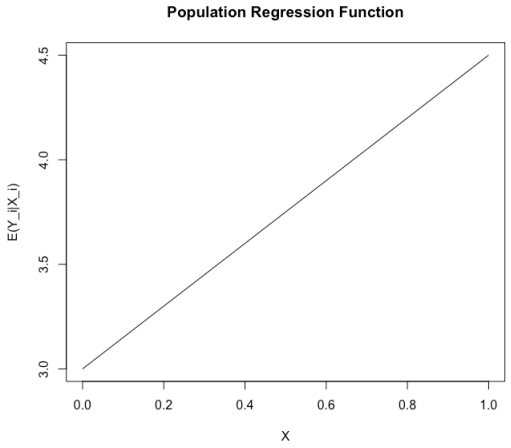
Problem Set 1

Pencil and Paper Questions

1. (a)

The population regression function is: $Y_i = E(Y_i|X_i) + \epsilon_i$.
In the following graph we will show the regression function without the random part (ϵ_i):



(b)

The following table presents the results of the required calculations:

	X	Y	X_i -mean(X)	Y_i -mean(Y)	x_i * y_i	x^2
	1	4	**-1.5**	**-3.5**	**5.25**	**2.25**
	4	10	**1.5**	**2.5**	**3.75**	**2.25**
	3	9	**0.5**	**1.5**	**0.75**	**0.25**
	2	7	**-0.5**	**-0.5**	**0.25**	**0.25**
mean(X/Y))	**2.5**	**7.5**				

(c)

The OLS estimators for β_1 and β_2 are estimated by minimizing the sum of squared residuals:

$$\min_{\hat{\beta}} \sum_{i=1}^{N} (y_i - x_i'\hat{\beta})^2$$

$$\frac{\partial \sum_{i=1}^{N} (y_i - x_i'\hat{\beta})^2}{\partial \hat{\beta}}$$

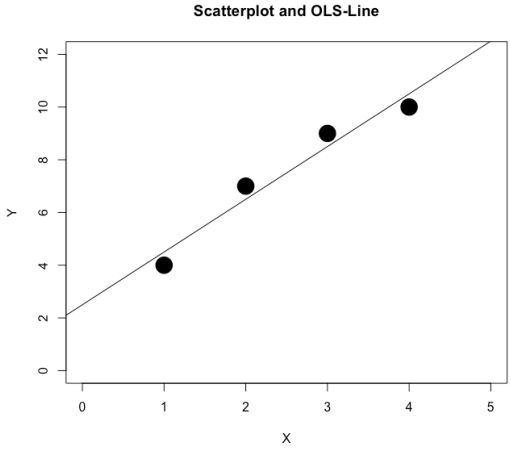
$$\widehat{\boldsymbol{\beta}}_{\mathbf{2}} = \frac{\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{N}(x_i - \bar{x})^2} = \frac{(-1.5) * (-3.5) + 1.5 * 2.5 + 0.5 * 1.5 - 0.5 * -0.5}{(-1.5)^2 + 1.5^2 + 0.5^2 + (-0.5)^2} = \mathbf{2}$$

$$\widehat{\boldsymbol{\beta}}_{\mathbf{1}} = \bar{y} - \hat{\beta}_2 \bar{x} = 7.5 - 2 * 2.5 = \mathbf{2.5}$$

The OLS estimates are $\hat{\beta}_1 = 2.5$ and $\hat{\beta}_2 = 2$.

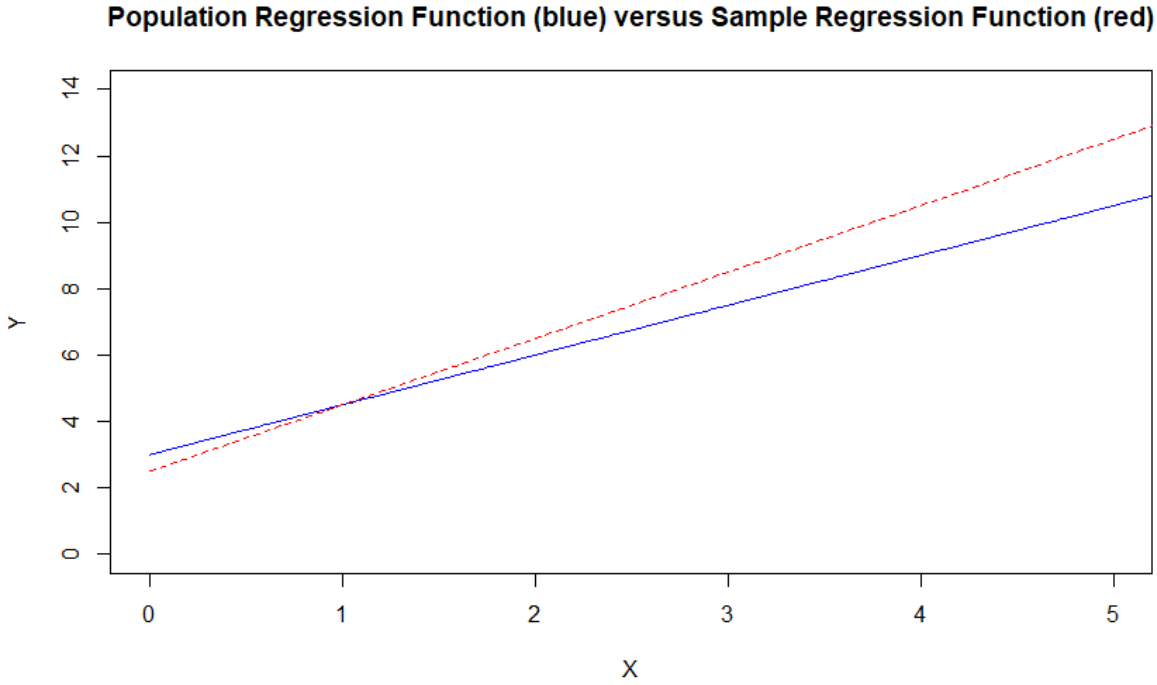
(d)

With the results under (c) we have drawn a scatterplot and the estimated OLS line.



(e)

The OLS line suggest a β_1 estimator $\hat{\beta}_1 = 2.5$ and a β_2 estimator $\hat{\beta}_2 = 2$, while the Population Regression Function has $\beta_1 = 3$ and $\beta_2 = 1.5$. Thus, the OLS estimation underestimates β_1 and overestimates β_2. Therefore, the y-intercept of the OLS line is smaller than the line drawn from the Population Regression Function, but the slope of the line is larger, i.e. the line is steeper.



(f)

The sample regression function crosses the population regression function, as can be seen in the graph above. It is possible that the two lines do not cross. This is the case if $\hat{\beta}_2$ equals β_2, thus the slopes are equal. The β_1 or the estimated $\hat{\beta}_1$ do not have an influence because they are only the intercepts.

(g)

The residuals e_i are calculated by the following equation:

$$e_i = y_i - \widehat{y_i} = y_i - x_i'\hat{\beta}$$

And the errors ϵ_i are calculated by the following equation:

$$\epsilon_i = y_i - E(y_i|x_i)$$

The results are summarized in the table below:

	X	Y	$\hat{Y}$	Y_{PopReg}	Residuals	Error
	1	4	4.5	4.5	**-0.5**	**-0.5**
	4	10	10.5	9	**-0.5**	**1**
	3	9	8.5	7.5	**0.5**	**1.5**
	2	7	6.5	6	**0.5**	**1**
mean(X/Y))	**2.5**	**7.5**		SUM	**0**	**3**

The residuals sum to 0 whereas the errors count up to 3. The OLS estimation process has minimized the sum of the squared residuals. With a small sample size (N=4) it is very unlikely to get the exact population regression function that is why the error term is large.

(h)

$$\sum_{\boldsymbol{i=1}}^{\boldsymbol{N}} (\boldsymbol{X_i} - \boldsymbol{\bar{X}}) = (1 - 2.5) + (4 - 2.5) + (3 - 2.5) + (2 - 2.5) = \mathbf{0}$$

The sum of $(X_i - \bar{X})$ equals 0. This is not just an idiosyncratic feature of this sample but holds in all samples. This is proven below:

$$\sum_{i=1}^{N} (X_i - \bar{X}) = x_1 - \bar{x} + x_2 - \bar{x} + \cdots + x_n - \bar{x} = x_1 - \sum_{i=1}^{N} \frac{x_i}{n} + x_2 - \sum_{i=1}^{N} \frac{x_i}{n} + \ldots + x_n - \sum_{i=1}^{N} \frac{x_i}{n}$$

$$= (x_1 + x_2 + \cdots + x_n) - n\left(\sum_{i=1}^{N} \frac{x_i}{n}\right) = \sum_{i=1}^{N} x_i - \sum_{i=1}^{N} x_i = 0$$

(i)

$$\sum_{i=1}^{N} x_i y_i = (-1.5 * -3.5) + (1.5 * 2.5) + (0.5 * 1.5) + (-0.5 * -0.5) = 10$$

$$\sum_{i=1}^{N} x_i Y_i = (-1.5 * 4) + (1.5 * 10) + (0.5 * 9) + (-0.5 * -7) = 10$$

$$\sum_{i=1}^{N} X_i y_i = (1 * -3.5) + (4 * 2.5) + (3 * 1.5) + (2 * -0.5) = 10$$

The results are summarized in the table below:

	X	Y	**X_i -mean(X)**	**Y_i -mean(Y)**	**x_i * y_i**	**x_i *Y**	**y_i *X**
	1	4	**-1.5**	**-3.5**	**5.25**	**-6**	**-3.5**
	4	10	**1.5**	**2.5**	**3.75**	**15**	**10**
	3	9	**0.5**	**1.5**	**0.75**	**4.5**	**4.5**
	2	7	**-0.5**	**-0.5**	**0.25**	**-3.5**	**-1**
mean(X/Y))	**2.5**	**7.5**		SUM	10	10	10

The sum of $(x_i y_i)$, $(x_i Y_i)$ and $(X_i y_i)$ equals 10. This is not just an idiosyncratic feature of this sample but holds in all samples. This is proven below:

$$\sum_{\boldsymbol{i=1}}^{\boldsymbol{N}} \boldsymbol{x_i y_i} = \sum_{i=1}^{N} (X_i - \bar{X})(Y_i - \bar{Y}) = \sum_{i=1}^{N} (X_i Y_i - X_i \bar{Y} - \bar{X} Y_i + \bar{X}\bar{Y})$$

$$= \sum_{i=1}^{N} X_i Y_i - \sum_{i=1}^{N} X_i \bar{Y} - \sum_{i=1}^{N} \bar{X} Y_i + \sum_{i=1}^{N} \bar{X}\bar{Y}$$

$$= \sum_{i=1}^{N} X_i Y_i - \bar{Y} \sum_{i=1}^{N} X_i - \bar{X} \sum_{i=1}^{N} Y_i + \sum_{i=1}^{N} \bar{X}\bar{Y} = \sum_{i=1}^{N} X_i Y_i - \bar{Y} n \bar{X} - \bar{X} n \bar{Y} + n \bar{X}\bar{Y}$$

$$= \sum_{i=1}^{N} X_i Y_i - n \bar{X}\bar{Y}$$

$$= \sum_{i=1}^{N} X_i Y_i - \sum_{i=1}^{N} \bar{X} Y_i = \sum_{i=1}^{N} X_i Y_i - \bar{X} Y_i = \sum_{i=1}^{N} (X_i - \bar{X}) Y_i = \sum_{\boldsymbol{i=1}}^{\boldsymbol{N}} \boldsymbol{x_i Y_i}$$

$$= \sum_{i=1}^{N} X_i Y_i - \sum_{i=1}^{N} X_i \bar{Y} = \sum_{i=1}^{N} X_i Y_i - X_i \bar{Y} = \sum_{i=1}^{N} (Y_i - \bar{Y}) X_i = \sum_{\boldsymbol{i=1}}^{\boldsymbol{N}} \boldsymbol{X_i y_i}$$

2. (a)

If the expected error term does not equal zero, then the beta estimators are subject to uncertainty. To prove this, we adapt the unbiased beta estimator:

$$y_i = \beta_1 + \beta_2 x_i + \varepsilon_i$$

$$\hat{\beta}_1 = \bar{y} - \hat{\beta}_2 \bar{x}$$

$$\hat{\beta}_2 = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2}$$

Auxiliary calculation:

$$\bar{y} = \frac{1}{N}\sum \beta_1 + \beta_2 x_i + \varepsilon_i = \frac{1}{N} * N * \beta_1 + \beta_2 \bar{x} + \bar{\varepsilon}$$

Use $y_i, \bar{y}$

$$\hat{\beta}_2 = \frac{\sum(x_i - \bar{x})(\beta_1 + \beta_2 x_i + \varepsilon_i - \beta_1 - \beta_2 \bar{x} - \bar{\varepsilon})}{\sum(x_i - \bar{x})^2}$$

$$\hat{\beta}_2 = \frac{\sum(x_i - \bar{x})\beta_2(x_i - \bar{x})}{\sum(x_i - \bar{x})^2} + \frac{\sum(x_i - \bar{x})\varepsilon_i}{\sum(x_i - \bar{x})^2} - \frac{\sum(x_i - \bar{x})\bar{\varepsilon}}{\sum(x_i - \bar{x})^2}$$

Auxiliary calculation:

$$\sum(x_i - \bar{x})\bar{\varepsilon} = \bar{\varepsilon}\sum(x_i - \bar{x}) = \bar{\varepsilon}(\sum x_i - N\bar{x}) = \bar{\varepsilon}(\sum x_i - N\left(\frac{1}{N}\sum x_i\right))$$
$$= \bar{\varepsilon}\left(\sum x_i - \sum x_i\right) = \bar{\varepsilon} * 0 = 0$$

Ergo:

$$\hat{\beta}_2 = \beta_2 + \frac{\sum(x_i - \bar{x})\varepsilon_i}{\sum(x_i - \bar{x})^2} - 0$$

$$\hat{\beta}_2 = \beta_2 + \frac{\sum(x_i - \bar{x})\varepsilon_i}{\sum(x_i - \bar{x})^2}$$

$$E(\hat{\beta}_2 | x_1, x_2, x_3 \dots x_n) = E\left(\beta_2 + \frac{\sum(x_i - \bar{x})\varepsilon_i}{\sum(x_i - \bar{x})^2} \middle| x_1, x_2, x_3 \dots x_n\right)$$
$$= \beta_2 + E\left(\frac{\sum(x_i - \bar{x})E(\varepsilon_i | x_1, x_2, x_3 \dots x_n)}{\sum(x_i - \bar{x})^2}\right)$$

Assumption 2: (Mean-Zero Error) $E(\varepsilon_i | x_1, x_2, x_3 \dots x_n) = 0$ → Unbiased

<u>BUT</u>: $E(\varepsilon_i | x_1, x_2, x_3 \dots x_n) \neq 0$

$$E(\hat{\beta}_2) = \beta_2 + E\left(\frac{\sum(x_i - \bar{x})E(\varepsilon_i | x_1, x_2, x_3 \dots x_n)}{\sum(x_i - \bar{x})^2}\right)$$

That is why the estimator is biased.

Further:

$$E(\hat{\beta}_1) = \bar{y} - E(\hat{\beta}_2)\bar{x} = \bar{y} - E\left(\frac{\sum(x_i - \bar{x})E(\varepsilon_i|x_1, x_2, x_3 \dots x_n)}{\sum(x_i - \bar{x})^2}\right)\bar{x}$$

Consequently, the estimator $\hat{\beta}_1$ is also biased.

(b)
The $\hat{\beta}$ and $\tilde{\beta}$ are calculated as follows ($\tilde{X} = 2X$):

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

$$\widetilde{\boldsymbol{\beta}} = \left(\tilde{X}'\tilde{X}\right)^{-1}\tilde{X}'y = \left((2X)'(2X)\right)^{-1}(2X)'y = (2X'2X)^{-1}2X'y = (4X'X)^{-1}2X'y$$
$$= 4^{-1}(X'X)^{-1}2X'y = 4^{-1} \cdot 2(X'X)^{-1}X'y = \frac{2}{4}(X'X)^{-1}X'y = \mathbf{\frac{1}{2}}\widehat{\boldsymbol{\beta}}$$

The transformation of X from X to $\tilde{X}$ = 2X causes a change of the beta estimator from $\hat{\beta}$ to $\frac{1}{2}\hat{\beta}$.

(c)
The $\hat{\beta}$ and β^* are calculated as follows ($y^* = 2y$):

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

$$\boldsymbol{\beta}^* = (X'X)^{-1}X'y^* = (X'X)^{-1}X'(2y) = (X'X)^{-1}X'2y = 2(X'X)^{-1}X'y = \mathbf{2}\widehat{\boldsymbol{\beta}}$$

The transformation of y from y to y*=2y causes a change of the beta estimator from $\hat{\beta}$ to $2\hat{\beta}$.

(d)
The solutions of (b) and (c) show that the units in which X and x are measured do not influence the conclusions drawn from the regressions because the beta estimators adjust to transformations of X and y. This can be shown mathematically:

$$\tilde{y} = \tilde{X}\tilde{\beta} = (2X)\left(\frac{1}{2}\beta\right) = 2 \cdot \frac{1}{2}X\beta = X\beta = y$$

(e) The variations $V(\hat{\beta})$ and $V(\tilde{\beta})$ are calculated using the following equations ($\tilde{X} = 2X$):

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

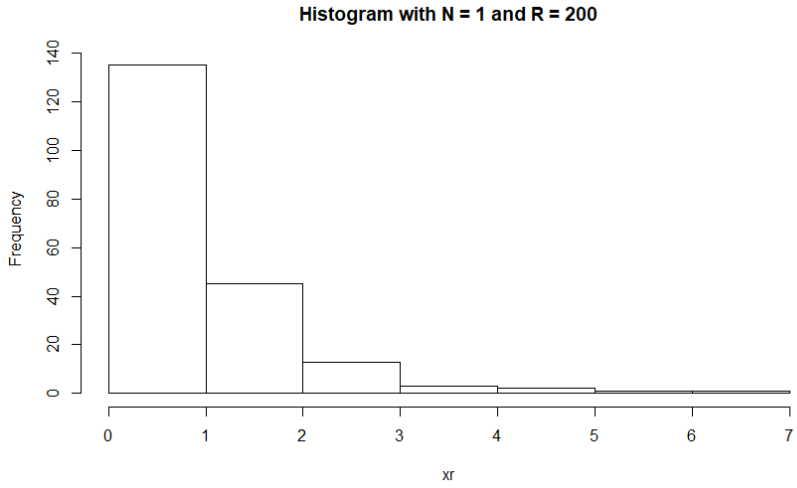
$$\boldsymbol{V}\left(\widetilde{\boldsymbol{\beta}}\right) = \sigma^2\left(\tilde{X}'\tilde{X}\right)^{-1} = \sigma^2((2X)'(2X))^{-1} = \sigma^2(2X'2X)^{-1} = \sigma^2 4^{-1}(X'X)^{-1} = \frac{1}{4}\sigma^2(X'X)^{-1}$$
$$= \frac{1}{4}V\left(\hat{\beta}\right) = \left(\mathbf{\frac{1}{2}}\right)^{\mathbf{2}}\boldsymbol{V}\left(\widehat{\boldsymbol{\beta}}\right)$$

The transformation of X from X to $\tilde{X}$ = 2X causes a change in the variance of the beta estimator from $V\left(\tilde{\beta}\right)$ to $V\left(\tilde{\beta}\right) = \left(\frac{1}{2}\right)^2 V\left(\hat{\beta}\right)$.

Computer Questions

1. (a)

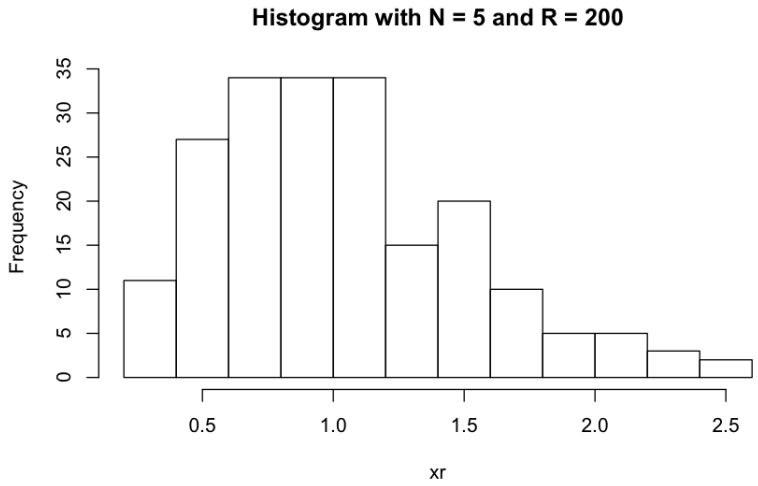
The following histogram displays the $\bar{x}^r$ for r = 1, 2, …, 200 for N=1, R=200:



The across-replication average $\bar{x}$ is 0.9424 and the sample variance $\bar{s}_x$ is 0.7330.

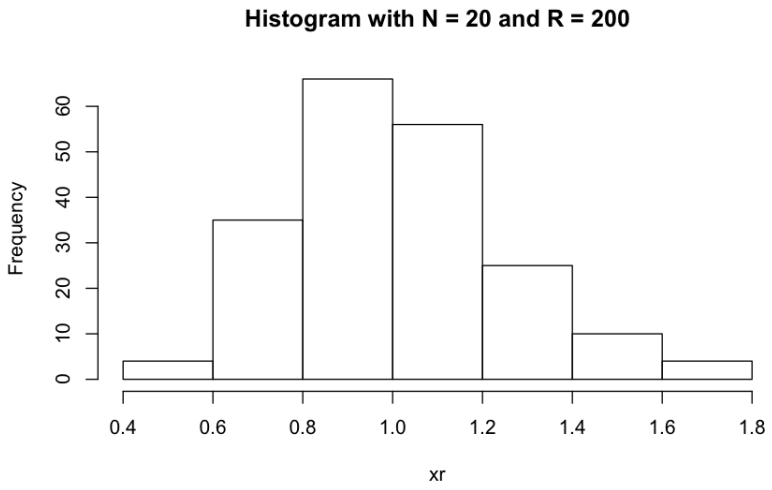
(b)

The following histogram displays the $\bar{x}^r$ for r = 1, 2, …, 200 for N=5, R=200:



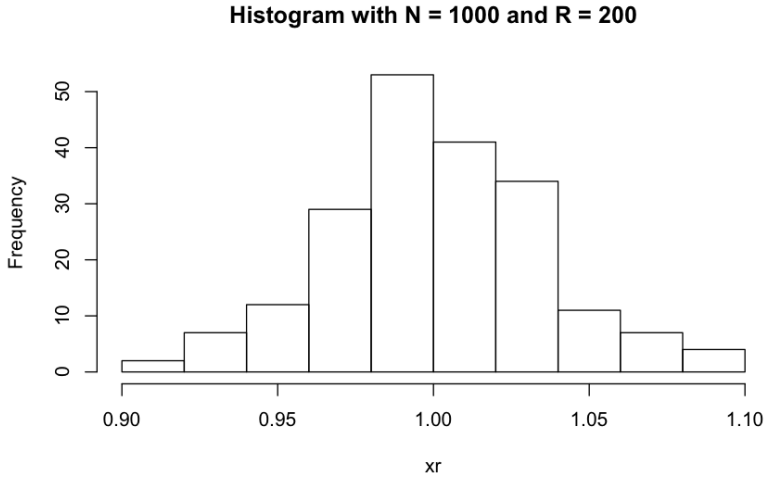
The across-replication average $\bar{x}$ is 0.9345 and the sample variance $\bar{s}_x$ is 0.1720.

(c)
The following histogram displays the $\bar{x}^r$ for r = 1, 2, …, 200 for N=20, R=200:



The across-replication average $\bar{x}$ is 1.0044 and the sample variance $\bar{s}_x$ is 0.0562.

(d)
The following histogram displays the $\bar{x}^r$ for r = 1, 2, …, 200 for N=1000, R=200:



The across-replication average $\bar{x}$ is 1.0011 and the sample variance $\bar{s}_x$ is 0.0012.

(e) i.
As presented in the examples (a) – (d), the distribution of $\bar{x}^r$ the distribution looks more like a normally distributed function. The bigger N, the more the histogram looks like a normally distributed function. For N = 1000, the mean is 1.0011 and the variance is $\sqrt{0.0012} = 0.0346$, which is very close to the values of the standard normal distribution.

e) ii.
The estimate of the $\bar{x}$ is close to $E(x_i) = 1$. By increasing N the $\bar{x}$ is getting closer to 1. Proof:

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$$

$$\boldsymbol{E(\bar{x})} = E[\frac{1}{N}\sum_{i=1}^{N} x_i] = \frac{1}{N}E[\sum_{i=1}^{N} x_i] = \frac{1}{N}\sum_{i=1}^{N} E[x_i] = \frac{1}{N}\sum_{i=1}^{N} \mu_x = \frac{1}{N}N\mu_x = \boldsymbol{\mu_x}$$

(e) iii.
The estimate of $s_{\bar{x}}$ is not close to $V(x_i) = 1$. By increasing N the $s_{\bar{x}}$ is getting more close to 0. Proof:

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$$

$$\boldsymbol{V_x} = V\left[\frac{1}{N}\sum_{i=1}^{N} x_i\right] = \frac{1}{N^2}V\left[\sum_{i=1}^{N} x_i\right] = \frac{1}{N^2}\sum_{i=1}^{N} V[x_i] = \frac{1}{N^2}\sum_{i=1}^{N} \sigma_x^2 = \frac{1}{N^2}(N\sigma_x^2) = \boldsymbol{\frac{\sigma_x^2}{N}}$$

It is the result underlying what is called a "Law of Large Numbers".

2. (a)
The dataset has 807 observations with 10 variables each.

(b)
The summary statistics over the asked variables are shown below:

Statistic	N	Mean	St. Dev.	Min	Max
cigs	807	8.686	13.722	0	80
educ	807	12.471	3.057	6.000	18.000
age	807	41.238	17.027	17	88
income	807	19,304.830	9,142.958	500	30,000
white	807	0.879	0.327	0	1
restaurn	807	0.247	0.431	0	1

The first variable "cigs" shows that over the sample in average 8.686 cigarettes are smoked every day. In this sample all individuals have attended school for at least 6 years. The average annual income is 19'304.830 with a minimum of only 500 and a maximum of 30'000 per year. The variable "white" as well as the variable "restaurn" are dummy variables. In the first column it is stated that all individual of the sample have answered all the questions.

(c) i.
The estimators for β_2 and β_1 are calculated using the following formulas:

$$\hat{\beta}_2 = \frac{cov(x,y)}{var(x)}$$

$$\hat{\beta}_1 = \bar{y} - \hat{\beta}_2\bar{x}$$

This leads to $\hat{\beta}_2 = 11.41203$ and a $\hat{\beta}_1 = -0.2185521$.

(c) ii.

The following table describes the results of the regression:

	Dependent variable:
	cigs
educ	-0.219
	(0.158)
Constant	11.412***
	(2.029)
Observations	807
R2	0.002
Adjusted R2	0.001
Residual Std. Error	13.714 (df = 805)
F Statistic	1.913 (df = 1; 805)
Note:	*p<0.1; **p<0.05; ***p<0.01

The estimation of the regression shows that the dependent variable "cigs" is weakly negatively correlated with the variable "educ". This relation is not significant on any common statistic level. Further the adjusted R^2 is very low with a value of only 0.001.

(c) iii.

The assumption 2 is mentioned below:

$$E(\epsilon_i|x_i) = 0$$

This leads to two implications:

1. $E(\epsilon_i) = 0$
2. $Cov(\epsilon_i, x_i) = 0$

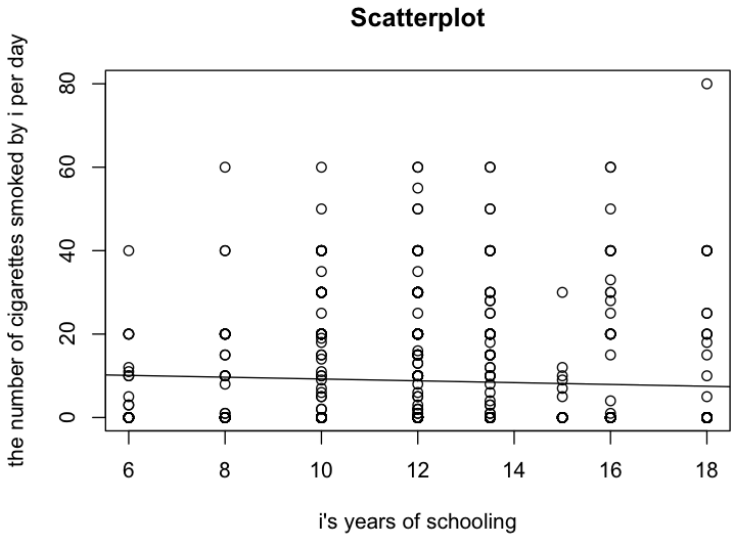
If the Assumption 2 (Mean-zero error) is satisfied, we can speak of unbiased estimators (see 2. (a)) between these two variables. That means if the level or the years spent in education increase by 1 the number of daily smoked cigarettes decreases by 0.219. A possible explanation for this negative relation is that with a higher level of education the individual is more aware of physical health and smokes less than people with a lower educational background. But this relation is very weak and not statistically significant.

$$\widehat{cigs} = 11.412 - 0.219 educ$$

This formula above represents the estimated regression under the assumption that the assumption 2 is satisfied.

(c) iv.

The following scatterplot shows the number of f cigarettes consumed by i against education and the estimated regression line:



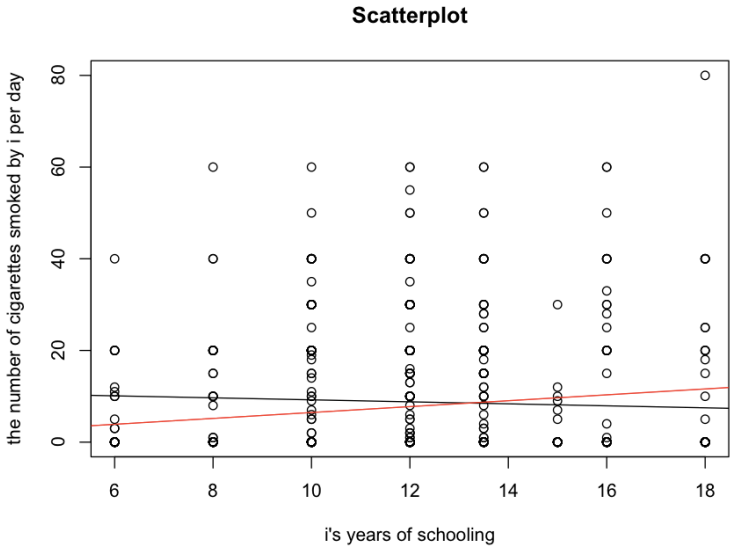
In the scatterplot above it is shown on the x lab the years that individual i spends in school and on the y lab the numbers of cigarettes that are smoked every day by i. Obviously, the years spent in school are discrete distributed. Further the estimated regression under (c) ii is presented with its negative slope.

(c) v.

The following table describes the results of the regression without a constant:

	Dependent variable:
	cigs
educ	0.645***
	(0.038)
Observations	807
R2	0.260
Adjusted R2	0.259
Residual Std. Error	13.972 (df = 806)
F Statistic	283.280*** (df = 1; 806)
Note:	*p<0.1; **p<0.05; ***p<0.01

Compared to the earlier regression line (black) this sample regression line (red) has a positive slope. We think we should include a constant because otherwise it is assumed that people who do not go to school (educ = 0) do not smoke any cigarettes (cigs = 0), which is very unlikely (Assumption 2 holds!).



(d) i.

The following table shows the results of the regression:

	Dependent variable:
	cigs
educ	-0.452***
	(0.162)
age	0.826***
	(0.154)
age_squ	-0.010***
	(0.002)
white	-0.624
	(1.456)
restaurn	-2.796**
	(1.104)
Constant	0.669
	(3.707)
Observations	807
R2	0.051
Adjusted R2	0.045
Residual Std. Error	13.407 (df = 801)
F Statistic	8.648*** (df = 5; 801)
Note:	*p<0.1; **p<0.05; ***p<0.01

The estimated regression with five independent variables has a higher adjusted R^2 (0.045) than with just one variable. Furthermore, the most variables are on a very high level statistically significant. If the Assumption 2 (Mean-zero error) is satisfied it is possible to give implication on the relation between the variables and the depended variable. In case of the variable "white" or in other terms the race it is a negative correlation calculated. That means if the individual is white it will smoke less than not-white individuals. But with this explanation you have to be careful due to no statistical significance. In case of the restaurants, that restricted smoking, you see a strong negative statistically significant relation. That means if a state restricted smoking in the restaurants the number of cigarettes been smoked every day has decreased.

(d) ii.
The regression function is:

$$\widehat{cigs} = 0.669 - 0.0452educ + 0.826age - 0.010age^2 - 0.624white - 2.796restaurn$$

To estimate the marginal effect of age, the derivative of this function with respect to age is used:

$$\widehat{cigs} = +0.826 - 2 * 0.010age$$

The marginal effect for the first case with 20 years:

$$\Delta\widehat{cigs} = 0.826 - 2 * 0.010 * \mathbf{20} = 0.426$$

The marginal effect for the first case with 40 years:

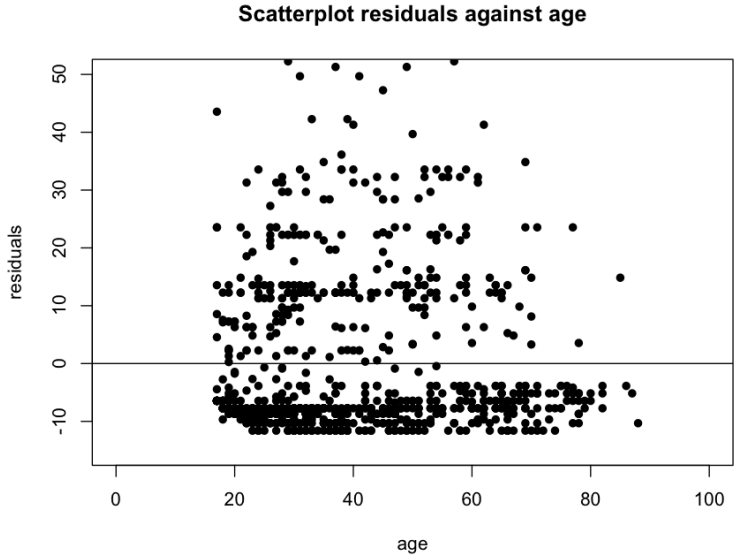
$$\Delta\widehat{cigs} = 0.826 - 0.010 * \mathbf{40} = 0.026$$

The marginal effect for the first case with 60 years:

$$\Delta\widehat{cigs} = 0.826 - 0.010 * \mathbf{60} = -0.374$$

After these calculations it is clear that the marginal effect of age follows a concave function. The older the individual gets, the more cigarettes it smokes. But this is only true up to a maximum and afterwards the number decreases again.

(d) iii. A
The following scatterplot shows the residuals against age:

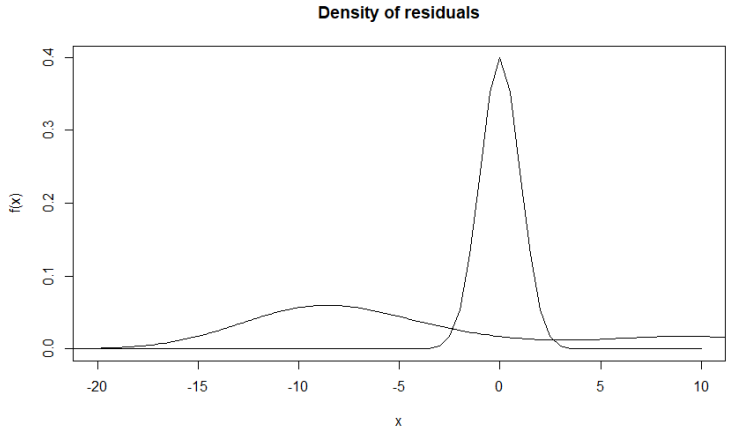


The scatterplot shows no discernible pattern, therefore the assumption 3 is likely satisfied. Thus, we can speak of homoscedasticity.

(d) iii. B
Assumption 4 assumes no correlation between any error term: $Cov(\epsilon_i, e_j) = 0$. For our regression, we get a correlation of 0.02257, which is close to 0. Therefore the assumption 4 is likely to be valid.

(d) iii. C

The following graph shows the density of the residuals and the density of a normal distribution:



The residuals are not normally distributed, and therefore, hypothesis testing is not recommended.

R Code

```
setwd("C:/Users/Markus/Dropbox/Empirical Methods/Problemset 1/Working
Process/R_Code")
library(lmtest)
library(stargazer)

#2 Computer Questions

#Task 1

#(a) N=1, R=200
exercisea <- replicate(200, rexp(n=1, rate = 1))
a <- exercisea
hist(a, main = "Histogram with N = 1 and R = 200", xlab="xr")
mean(a)
sd(a)
var(a)

#(b) N=5, R=200
exerciseb <- replicate(200, rexp(n=5, rate = 1))
b <- apply(exerciseb, MARGIN = 2, FUN = mean)
hist(b, main = "Histogram with N = 5 and R = 200", xlab="xr")
mean(b)
sd(b)
var(b)

#(c) N=20, R=200
exercisec <- replicate(200, rexp(n=20, rate = 1))
c <- apply(exercisec, MARGIN = 2, FUN = mean)
hist(c, main = "Histogram with N = 20 and R = 200", xlab="xr")
mean(c)
sd(c)
var(c)

#(d) N=1000, R=200
exercised <- replicate(200, rexp(n=1000, rate = 1))
d <- apply(exercised, MARGIN = 2, FUN = mean)
hist(d, main = "Histogram with N = 1000 and R = 200", xlab="xr")
mean(d)
sd(d)
var(d)

#Task 2

#(a) Download data by using wooldridge package
library(foreign)
library(wooldridge)
myxy <- wooldridge::smoke
head(myxy)

#(b) provide summary statistics
```

```
summaryI <- subset(myxy,select = c(cigs, educ, age, income, white, restaurn))
#if needed: library(stargazer)
stargazer(summaryI, type="text")

#(c)
#(i) calculation of B1 and B2, educ as dependent variable
y <- myxy$cigs
x <- myxy$educ
beta_1 <- cov(y,x) / var(x)
beta_0 <- mean(y) - beta_1 * mean(x)
beta_0
beta_1

#(ii) estimation of regression function
regression1 <- lm(myxy$cigs~myxy$educ)
stargazer(regression1, type = "text")

#(iii) no code needed

#(iv) prediction of consumed cigarettes, displays
plot(myxy$educ,myxy$cigs, xlab = "i's years of schooling",
    ylab = "the number of cigarettes smoked by i per day", main = "Scatterplot")
myfm <- data.frame(myxy$educ,myxy$cigs)
datamodell <- lm(myxy$cigs~myxy$educ, data= myfm)
abline(datamodell, col="black",lwd=1)

#(v) without constant
regression2 <- lm(myxy$cigs~myxy$educ-1)
stargazer(regression2, type = "text")
myfm2 <- data.frame(myxy$educ,myxy$cigs)
datamodell1 <- lm(myxy$cigs~myxy$educ-1, data= myfm2)
abline(datamodell1, col="red",lwd=1, xlim = c(0,18))

# (d) new regression
age_squ <-(myxy$age)^2
head(myxy$age)
head(age_squ)
regression2 <-
lm(myxy$cigs~myxy$educ+myxy$age+age_squ+myxy$white+myxy$restaurn)

#(i)

stargazer(regression2, type="text")
head(myxy$restaurn)
head(myxy$white)

#(ii) no code needed

#(iii)
##(A)
residuals <- residuals(regression2)
```

```
plot(myxy$age,residuals, pch=16,col="black",cex=1, xlab = "age", ylab = "residuals", main =
"Scatterplot residuals against age", xlim = c(0,100), ylim = c(-15,50))
abline(0,0)
plot(regression2)

#### B)
number <- (1:807)
cor(number, residuals)

#### C)
x<-seq(-20,10, by=0.5)
plot(x,dnorm(x),
    type="l", xlab="x", ylab="f(x)", main="Density of residuals and normal distribution")
lines(density(residuals))
```