Problem Set 1 Empirical Methods

Patrick Glettig (13-252-143), Carmine Ragone (14-995-021)

21.10.2018

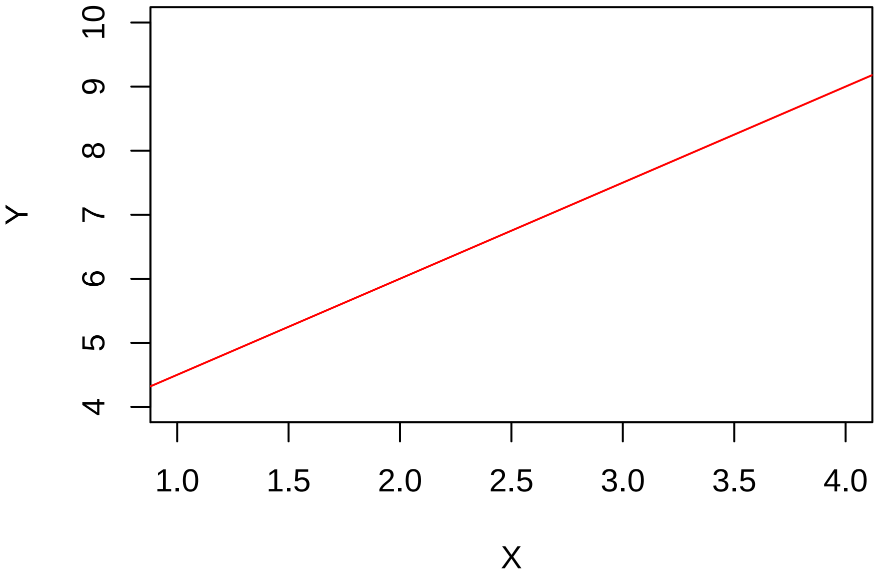
Pencil and Paper Questions

Exercise 1

(a) Write down the population regression function. Draw a picture of $E(Y_i|X_i)$, the non-random part of the PRF.

Population Regression Function:

$$\begin{aligned} Y_i &= E(Y_i|X_i) + \epsilon_i \\ &= \beta_1 + \beta_2 X_i + \epsilon_i \\ &= 3 + 1.5X_i + \epsilon_i \end{aligned}$$



(b) Construct a table:

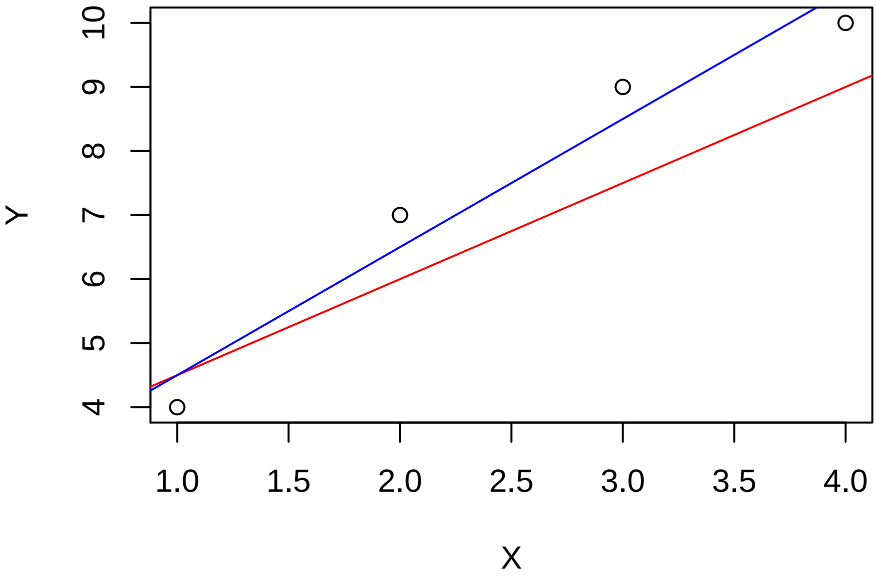
N	X	Y	mean_X	mean_Y	x_i	y_i	x_iy_i	x_square_i
1	1	4	2.5	7.5	-1.5	-3.5	5.25	2.25
2	4	10	2.5	7.5	1.5	2.5	3.75	2.25
3	3	9	2.5	7.5	0.5	1.5	0.75	0.25
4	2	7	2.5	7.5	-0.5	-0.5	0.25	0.25

(c) Calculate OLS estimates: $\widehat{\beta}_1, \widehat{\beta}_2$

$$\widehat{\beta}_2 = \frac{\sum_{n=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2} = \frac{10}{2} = 2$$

$$\widehat{\beta}_1 = \bar{y} - \widehat{\beta}_2\bar{x} = 7.5 - 2 * 2.5 = 2.5$$

(d) Plot the 4 points and OLS line



(e) How does the OLS line compare to the line you drew from your Population Regression Function?

The lines are close to each other, however they are not the same. This is because of the residual part of the regression function. This results in a different intercept and different slope conditioned by the sample.

(f) Does your sample regression function cross the population regression function? Suppose you were to select another sample from the same population. Is it possible that the two lines would not cross? Why or why not?

Yes it does and the crossing point is in $P = (1, 4.5)$ However, this is not always the case, in fact if we had another sample with $\widehat{\beta}_1 = 1.5$ and $\widehat{\beta}_2 = 4$ the line would not cross and this would not mean that it is not an estimate. As final remark the fitted OSL estimate of the PRF only depends on the sample we collect and if we had the possibility to draw an infinite sample the estimate and the real PRF would be identical.

(g) Calculate the error, ϵ_i, for the data points in your sample. Also calculate the residuals, e_i, for these data points. Do the errors sum to zero? Do the residuals? Do your answers differ? If so, explain why in your own words.

Errors:

$$\begin{aligned}\epsilon_i &= Y_i - E(X_i|Y_i)\\ \epsilon_i &= Y_i - 3 - 1.5X\\ \epsilon_i &= (0.5, 1, 1.5, 1)\end{aligned}$$

$$\sum \epsilon_i = 4$$

Residuals:

$$\begin{aligned}e_i &= Y_i - \widehat{\beta}_1 - \widehat{\beta}_2 X\\ e_i &= Y_i - 2.5 - 2X\\ e_i &= (-0.5, -0.5, 0.5, 0.5)\end{aligned}$$

$$\sum e_i = 0$$

The residuals have to sum up to zero because the OLS estimates $\widehat{\beta}_1$ and $\widehat{\beta}_2$ are chosen to to make the residuals add up to zero. The fitted line is created in a way that it has the "same distance"" from every point of our sample. In fact, the residuals are defined as:

$$e_i = y_i - \widehat{\beta}_1 - \widehat{\beta}_2$$

For the errors, they can sum up to any number as we cannot be sure of what the real β_1 and β_2 are, they are given by the true data generating process which we do not have.

(h) Show that $\sum_{n=1}^{4}(X_i - \bar{x}) = 0$ Is this an idiosyncratic feature of this sample or would you expect it to hold in every sample?

$$\sum_{n=1}^{4}(X_i - \bar{x}) = -1.5 + 1.5 + 0.5 - 0.5 = 0$$

$$\sum(x_i - \bar{x}) = \sum X_i - n\bar{x} = \sum x_i - n\frac{1}{n}\sum x_i = 0$$

Given by the equation above we can see that this holds in every sample.

(i)Show that $\sum x_i y_i = \sum x_i Y_i$. Also show that this equals $\sum X_i y_i$ Is this an idiosyncratic feature of this sample or would you expect it to hold in every sample?

$$\sum_{n=1}^{N}(X_i - \bar{x})(Y_i - \bar{y}) = 5.25 + 3.75 + 0.75 + 0.25 = 10$$

$$\sum_{n=1}^{N}(X_i - \bar{x})Y_i = -1.5(4) + 1.5(10) + 0.5(9) - 0.5(7) = 10$$

$$\sum_{n=1}^{N}(Y_i - \bar{y})X_i = -3.5(1) + 2.5(4) + 1.5(3) - 0.5(2) = 10$$

This is not an idiosyncratic feature of this sample, but a general one. In fact, we can see that for the general cases is true that:

$$\begin{aligned}
\sum_{n=1}^{N}(X_i-\bar{x})(Y_i-\bar{y}) &= \sum_{n=1}^{N}(X_i-\bar{x})Y_i - \sum_{n=1}^{N}(X_i-\bar{x})\bar{y} \\
&= \sum_{n=1}^{N}(X_i-\bar{x})Y_i - \bar{y}\sum_{n=1}^{N}(X_i-\bar{x}) \\
&= \sum_{n=1}^{N}(X_i-\bar{x})Y_i - \bar{y}*0 \\
&= \sum_{n=1}^{N} x_iY_i
\end{aligned}$$

and as well that:

$$\begin{aligned}
\sum_{n=1}^{N}(X_i-\bar{x})(Y_i-\bar{y}) &= \sum_{n=1}^{N}(Y_i-\bar{y})X_i - \sum_{n=1}^{N}(Y_i-\bar{y})\bar{x} \\
&= \sum_{n=1}^{N}(Y_i-\bar{y})X_i - \bar{x}\sum_{n=1}^{N}(Y_i-\bar{y}) \\
&= \sum_{n=1}^{N}(Y_i-\bar{y})X_i - \bar{x}*0 \\
&= \sum_{n=1}^{N} y_iX_i
\end{aligned}$$

This holds in every sample where the OLS assumptions hold.

Exercise 2

Consider the simple linear regression model

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

(a) Suppose that the unconditional expectation $E(\epsilon_i) = \mu_\epsilon \neq 0$ Using the formulas for $\hat{\beta}_1$ and $\hat{\beta}_2$ on slide 4 of the lecture notes, evaluate $E(\hat{\beta}_1)$ and $E(\hat{\beta}_2)$. What does your answer tell you about the robustness of OLS estimation to ϵ_i having a non-zero error?

Defined that

$$mxx = \sum_{n=1}^{N}(x_i-\bar{x})^2 \quad W_i = \frac{(x_i-\bar{x})}{mxx} \quad v_i = (\frac{1}{n} - \bar{x}W_i)$$

We can write the value of $\widehat{\beta}_2$ as follows:

$$\widehat{\beta}_2 = \frac{\sum_{n=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{n=1}^{N}(x_i - \bar{x})^2} = \sum_{n=1}^{N} \frac{(x_i - \bar{x})}{mxx}\bar{y} = \sum_{n=1}^{N} W_i y_i$$

Let's calculate the expected value of $\widehat{\beta}_2$:

$$\begin{aligned}
E(\widehat{\beta}_2) &= E(\sum_{n=1}^{N} W_i y_i) \\
&= \sum_{n=1}^{N} W_i E(y_i) \\
&= \sum_{n=1}^{N} W_i(\beta_1 + \beta_2 X_1 + E(\epsilon_i)) \\
&= \beta_1 \sum_{n=1}^{N} W_i + \beta_2 \sum_{n=1}^{N} W_i X_1 + \sum_{n=1}^{N} W_i E(\epsilon_i) \\
&= \beta_1 * 0 + \beta_2 * 1 + \sum_{n=1}^{N} W_i E(\epsilon_i) \\
&= \beta_2 + \sum_{n=1}^{N} W_i E(\epsilon_i)
\end{aligned}$$

As we just demonstrated $E(\beta_2) \neq \beta_2$. As $\beta_1 = \bar{y} - \beta_2\bar{x}$, $E(\beta_1) \neq \beta_1$ which is telling us that the OLS estimation is not robust.

Consider now the general CLRM with multiple regressors

$$y = X\beta + \epsilon$$

where β includes the constant β_1.

(b) Suppose you decided to measure all of your X variables in different units such that your new X variable, call it $\tilde{X}$, is exactly double your old one, i.e. $\tilde{X} = 2X$. Suppose you run the regression of y on $\tilde{X}$; call the resulting estimate $\tilde{\beta}$. What is the relationship between $\tilde{\beta}$ and $\hat{\beta}$, the regular OLS estimator from the regression of y on X?

If the measures of a variable are doubled the new estimate $\tilde{\beta}$ will have the same costant β_1, all the other estimates will be the half of $\widehat{\beta}$. This is true because $Y = \beta_1 + \beta_2 X + \epsilon_i = \beta_1 + \frac{\beta_2}{2}(2 * X) + \epsilon_i$ and the standard error will be also half.

c) Suppose you decided now to measure y in different units such that your new y variable, call it y^*, is exactly double your old one, i.e. $y^* = 2 * y$. Suppose you run the regression of y^* on X; call the resulting estimate β^*. What is the relationship between β^* and $\hat{\beta}$?

If the dependent variable doubles, β^* will exactly be double of $\hat{\beta}$:

$$Y = X\beta + \epsilon$$

$$2Y = 2X\beta + 2\epsilon$$

(d)Given your answers to the last two questions, how meaningful are the units in which X and y are measured for the conclusions you draw from an OLS regression?

Given the previous results, nature is indifferent to units of measurment, which means that there will be not inappropriate effects changing the units of the explanatory variable in a OLS regression model. Changes of units will not affect the goodness of fit of the model nor its accuracy. It will just change the units when you interprete the coefficients. Thus, the units in which X or Y are measured doesn't really matter.

*(e) Return to the scenario in part (2b) above with $\tilde{X} = 2X$.** Calculate* *$V(\tilde{\beta})$. What is the relationship between $V(\tilde{\beta})$ and $V(\hat{\beta})$?*

The bigger X it gets, the smaller the variance.

$$V(\widehat{\beta}) = \frac{1}{(\sum_{n=1}^{N}(x_i - \bar{x}))^4}\sum_{n=1}^{N}(x_i - \bar{x})^2 V(Y_i) = \frac{\sigma^2}{\sum_{n=1}^{N}(x_i - \bar{x})^2}$$

Given that

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

We have that

$$V(\tilde{\beta}) = V(\frac{1}{2}\widehat{\beta}) = \frac{1}{4}V(\widehat{\beta})$$

Computer Questions

Empirical Exercise 1

As questions (a) to (d) essentially ask the same but with different parameters, it makes sense to create a function that does the calculation for us.

```r
solveTask <- function(N=1,R=200){#define the default parameters as for the first task
  set.seed(42)
  library(ggplot2)#for the histogram
  #Create empty DF first to be filled
  samples <- data.frame(matrix(data=NA,nrow=R,ncol=N))

  for (i in 1:R){#R is number of replications
    samples[i,] <- rexp(n=N) #N is observations of the exponential distribution
  }
  #Rename column names
  names(samples) <- paste0('x_',1:N)

  samples['x_bar_avg']<-apply(samples,MARGIN = 1,FUN = mean)#calculate average per sample

  #Now we do the plot
  plot <- ggplot(samples, aes(x=x_bar_avg)) +     geom_histogram(binwidth=0.1)+
    theme(text = element_text(size=10, family="LM Roman 10"))+
    ggtitle(paste0('Histogram with N=',N,' and R=',R))+
    xlab(TeX('$\\bar{x}^r$'))+
    ylab('Count')
  x_bar <- mean(samples$x_bar_avg)
  s_x_bar <- var(samples$x_bar_avg)
  result <- list(plot,x_bar,s_x_bar)
  names(result) <- c('Histogram','x_bar','s_x_bar')
  return(result)
}
```

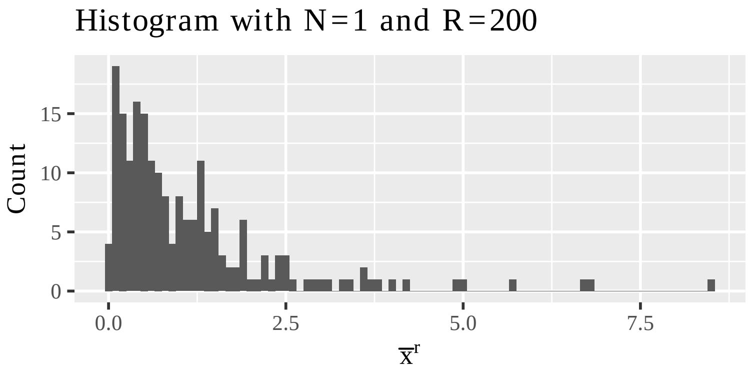
Note that we can adress the asked solutions with the following code:

```r
N1R200[1] #for the histogram
N1R200[2] #for the sample average
N1R200[3] #for the sample variance
```

(a) Let $N = 1$ and $R = 200$. *Calculate* $\bar{x}^r$ *for* $r = 1, ..., 200$ *show them in a histogram. Also calculate the across-replication average,* $\bar{x}$ *, and sample variance,* s_{barx}.

```r
N1R200 <- solveTask(N=1,R=200)
```

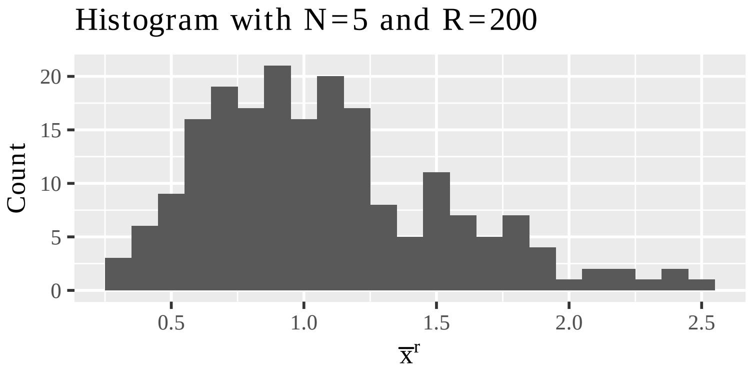
Where $\bar{x}^r = 1.157854$ and $s_{barx} = 1.636663$. This yields the following histogram:



(b) Let $N = 5$ and $R = 200$. *Calculate* $\bar{x}^r$ *for* $r = 1, ..., 200$ *show them in a histogram. Also calculate the across-replication average,* $\bar{x}$ *, and sample variance,* s_{barx}.

```
N5R200 <- solveTask(N=5,R=200)
```

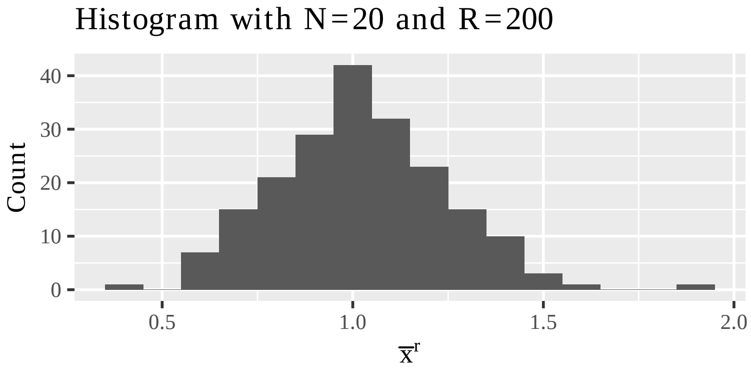
Where $\bar{x}^r = 1.0779428$ and $s_{barx} = 0.2138821$. This yields the following histogram:



(c) Let $N = 20$ and $R = 200$. *Calculate* $\bar{x}^r$ *for* $r = 1, ..., 200$ *show them in a histogram. Also calculate the across-replication average,* $\bar{x}$ *, and sample variance,* s_{barx}.

```
N20R200 <- solveTask(N=20,R=200)
```

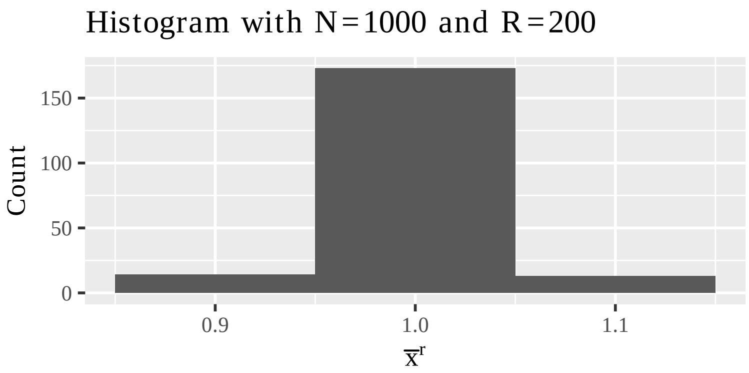
Where $\bar{x}^r = 1.0167628$ and $s_{barx} = 0.049094$. This yields the following histogram:



(d) Let $N = 200$ and $R = 200$. *Calculate* $\bar{x}^r$ *for* $r = 1, ..., 200$ *show them in a histogram. Also calculate the across-replication average,* $\bar{x}$ *, and sample variance,* s_{barx}.

```
N1000R200 <- solveTask(N=1000,R=200)
```

Where $\bar{x}^r = 0.999437$ and $s_{barx} = 0.0010919$. This yields the following histogram:



(e) Based on your answers to the previous parts of this question,

i. For each of $N = 1, N = 5, N = 20$, and $N = 1000$: Does the distribution of $\bar{x}^r$ look more like an exponential distribution or a normal distribution?

The plots above show that the higher N, the more the distribution of $\bar{x}^r$ looks like a normal distribution. As we are taking the sample average, extreme values are smoothed more the higher N.

ii. Is your estimate of $\bar{x}$ close to $E(x_i) = 1$ in each experiment? If not, why not?

	N1	N5	N20	N1000
x__bar	1.157854	1.0779428	1.016763	0.9994370
s__x__bar	1.636663	0.2138821	0.049094	0.0010919

We can see how $\bar{x}$ converges to $E(x_i) = 1$ the larger N. However, $s_{\bar{x}}$ becomes smaller the larger N. This is because extreme values carry less weight with a larger sample size. Hence, we observe a tendency to the middle. Therefore, we come closer to a normal distribution with $N \sim (0, 1)$.

*iii. Is your estimate of $s_{\bar{x}}$ close to $V(x_i) = 1$** in each experiment? If not, why not?*

The table above shows the convergence of $s_{\bar{x}}$ towards 0 as N increases. The underlying effect is the same as above: extreme values carry less weight, $\bar{x}^r$ are closer to each other, variance is reduced. This comes from the central limit theorem.

Empirical Exercise 2

(a) Download the data and import them into Stata or R. How many observations are there?

```
smoke <- read.dta13('smoke.dta')
dim(smoke) #dim() returns the dimensions of the dataframe.
```

There are 807 observations and 10 variables.

(b) Provide a table of summary statistics for the variables cigs, educ, age, income, white, restaurn. Briefly describe patterns you find particularly interesting (if any).

```
summary(smoke[,c('cigs', 'educ', 'age', 'income', 'white', 'restaurn')])
##       cigs             educ            age            income     
##  Min.   : 0.000   Min.   : 6.00   Min.   :17.00   Min.   :  500  
##  1st Qu.: 0.000   1st Qu.:10.00   1st Qu.:28.00   1st Qu.:12500  
##  Median : 0.000   Median :12.00   Median :38.00   Median :20000  
##  Mean   : 8.686   Mean   :12.47   Mean   :41.24   Mean   :19305  
##  3rd Qu.:20.000   3rd Qu.:13.50   3rd Qu.:54.00   3rd Qu.:30000
##  Max.   :80.000   Max.   :18.00   Max.   :88.00   Max.   :30000
##      white           restaurn
##  Min.   :0.0000   Min.   :0.0000
##  1st Qu.:1.0000   1st Qu.:0.0000
##  Median :1.0000   Median :0.0000
##  Mean   :0.8786   Mean   :0.2466
##  3rd Qu.:1.0000   3rd Qu.:0.0000
##  Max.   :1.0000   Max.   :1.0000
```

- *cigs*: With a median of 0, the majority of the sample does not smoke, however, it seams that the ones who smoke have a high daily cigarette consumption as indicated by the 3rd quantile.
- *educ*: There is nobody in the sample that does not have any school education, we would need to check if this is a representative sample then, asking the question if everyone in the population received an education. The mean and median values seam quite reasonable.
- *age*: A minimum age of 17 seems really high, based on personal experience, people start smoking earlier.
- *income*: It is questionable if in a representative sample, the minimum income is 500. It would surprise if the lowest income might be lower if not zero.
- *white*: with a mean of 0.88, a large fraction of the sample is white. This is a clear hint of a selection bias of the sample. Also, we might as well want to include other ethnicities if possible.
- *restaurn*: Most of the restaurants do not seem to forbid smoking. This is ok because in reality it is not normally distributed.

(c) We want to estimate the relationship between number of cigarettes smoked and education, measured as i's years of schooling.

$$cigs_i = \beta_0 + \beta_1 educ_i + \epsilon_i$$

i. Compute β_1 and β_0.

$$\hat{\beta}_1 = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

```
beta_1 <- sum((smoke$educ-mean(smoke$educ))*
                (smoke$cigs-mean(smoke$cigs)))/sum((smoke$educ-mean(smoke$educ))^2)

beta_0 <- mean(smoke$cigs)-beta_1*mean(smoke$educ)
```

$\hat{\beta}_0$ is 11.4120303 and $\hat{\beta}_1$ is -0.2185521.

ii. Run the regression in equation (2c) *How do the computer's estimates of β_0* and *β_1 relate to the ones you have just computed?*

```
my_lm <- lm(cigs ~ educ, data = smoke)
my_lm$coefficients
## (Intercept)        educ
##  11.4120303  -0.2185521
```

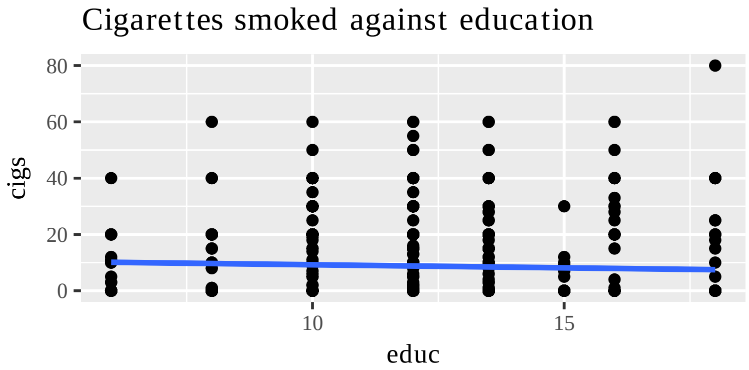
The estimates do not differ, not surprising as they use the same OLS formula.

iii. Suppose that Assumption 2 (Mean-zero error) is satisfied. How do you interpret the coefficient of educ? Is this a big or small effect?

A one-unit increase in *educ* (school years), decreases daily cigarette consumption by 0.21. As the mean of *cigs* is around 8.6, this is a 0.3136881% decrease for an individual at the average number of school years (~12.5), a relatively large effect.

iv. Using your estimates, predict the number of cigarettes consumed by i and denote this $\widehat{cigs}$. In a graph, display both the scatterplot of cigarettes smoked against education and your regression line.

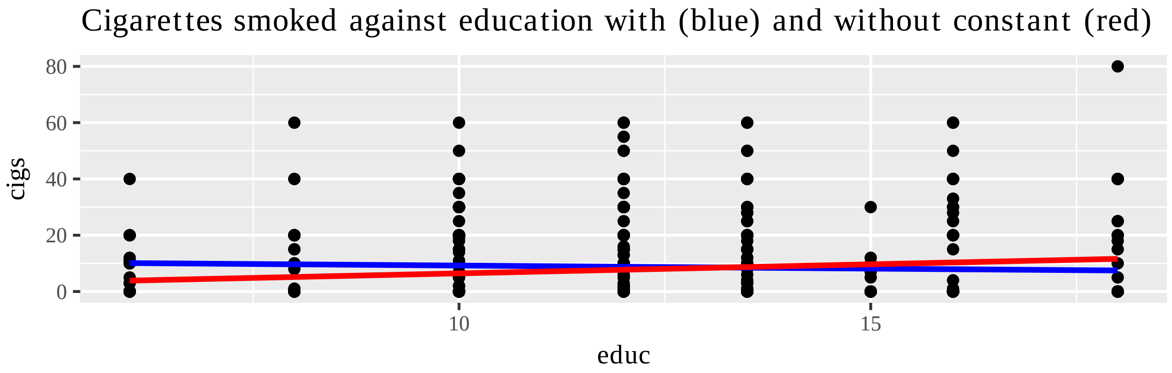
```
cigs_hat <- predict(my_lm)#Predict values
print(head(cigs_hat))
##         1         2         3         4         5         6
##  7.915197  7.915197  8.789405  8.461577  9.226509 10.100718
#Create Scatterplot with regression line
ggplot(smoke, aes(x=educ, y=cigs)) +
  theme(text = element_text(size=10, family="LM Roman 10"))+
  geom_point() +     # Scatters
  geom_smooth(method=lm, se=FALSE)+
  ggtitle('Cigarettes smoked against education')
```



v. Now regress cigarettes on education without including a constant. Generate predicted values and add the new regression line to the previous graph. What changes compared to the earlier regression line? Do you think you should include a constant or not?

```
my_lm_no_constant <- lm(cigs ~ educ + 0, data = smoke)
my_lm_no_constant$coefficients
##      educ
## 0.6447271
cigs_hat_no_constant <- fitted(my_lm_no_constant)#calculates the fitted y values
print(head(cigs_hat_no_constant))
##         1         2         3         4         5         6
## 10.315634 10.315634  7.736726  8.703817  6.447271  3.868363

ggplot(smoke, aes(x=educ, y=cigs)) +
  theme(text = element_text(size=10, family="LM Roman 10"),
        legend.position = "bottom")+
  geom_point() +     # Scatters
  geom_smooth(method=lm, se=FALSE, color='blue')+
  geom_smooth(method=lm,
              formula = y ~ 0 + x, #remove constant
              se=FALSE, color='red')+
  ggtitle('Cigarettes smoked against education with (blue) and without constant (red)')
```



We can see that the relationship between *educ* and *cigs* changes from negative to positive, indicating that more years of education increase cigarette consumption. By removing the constant, we would assume that the regression line goes through the origin, which is clearly not the case. Therefore it is important to include a constant.

(d) Now regress cigs on educ, age, age^2*, white and restaurant and assume again that Assumption 2 (Mean-zero error) is satisfied.*

$$cigs_i = \beta_0 + \beta_1 educ_i + \beta_2 age_i + \beta_3 {age_i}^2 + \beta_4 white_i + \beta_5 restaurant_i + \epsilon_i$$

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

i. What are the coefficients of race (i.e. white) and of the dummy restaurant? How would you interpret them?

```
summary(third_lim) #gives us information about every coefficient
##
## Call:
## lm(formula = cigs ~ educ + age + agesq + white + restaurn, data = smoke)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -15.029  -9.256  -6.175   8.035  70.382
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.668834   3.706849   0.180  0.85686
## educ        -0.451501   0.161588  -2.794  0.00533 **
## age          0.825764   0.154474   5.346 1.18e-07 ***
## agesq       -0.009631   0.001682  -5.727 1.45e-08 ***
## white       -0.623739   1.456110  -0.428  0.66850
## restaurn    -2.796182   1.103552  -2.534  0.01147 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13.41 on 801 degrees of freedom
## Multiple R-squared:  0.05122,    Adjusted R-squared:  0.0453
## F-statistic: 8.648 on 5 and 801 DF,  p-value: 5.387e-08
```

The coefficent of race, *white*, is -0.6237386. This means that white people smoke on average -0.6237386 less cigarettes per day. However, the coefficient is not significant with a p-value of 0.6685046 and should therefore not be interpreted. On the contrary, the dummy *restaurant* is highly significant with a p-value of 0.0114728. On average, people living in states where smoking in restaurants is forbidden smoke -0.009631 less per day.

ii. Calculate the marginal effect of age on cigarette consumption. What is the value of this marginal effect at age 20? At age 40? At age 60?

First, one extracts the necessary coefficients from the model:

```r
linear_coef <- summary(third_lim)$coefficients[3,1]#extract linear coefficient
squared_coef <- summary(third_lim)$coefficients[4,1]#extract squared coefficient
```

Then, to calculate the marginal effects, one needs to take the derivate with respect to age of the regression equation:

$$cigs_i = \beta_0 + \beta_1 educ_i + \beta_2 age_i + \beta_3 {age_i}^2 + \beta_4 white_i + \beta_5 restaurant_i + \epsilon_i$$

$$\frac{\delta cigs_i}{\delta age} = \beta_2 + 2\beta_3 age_i$$

To implement this in R, one can first extract the coefficients β_2 and β_3 and then store the above derivate in a function to calculate the marginal effect at different ages:

```r
me_age <- function(age=1){
  x <- linear_coef + 2*squared_coef*age
  return(x)
}
me_age(20)
me_age(40)
me_age(60)
```

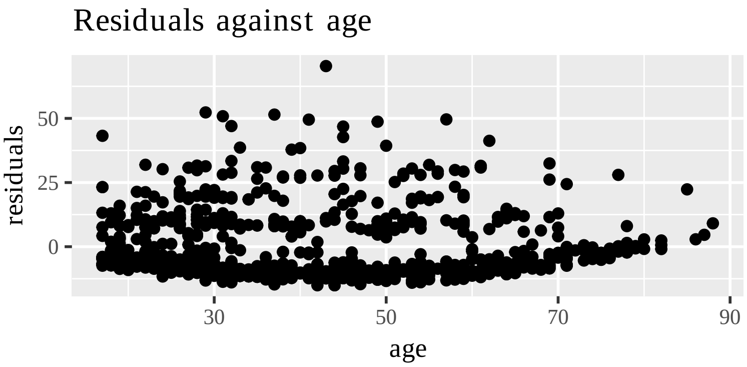
The marginal effect at age 20 is 0.4405238, at 40 0.0552836 and -0.3299567 at 60.

iii. Predict the residuals from your model, $e_i = cigs_i - \hat{cigs}_i$, where $\hat{cigs}_i$ is the fitted value of $cigs_i$ from your regression.

```r
head(third_lim$residuals)#calculates residuals. Only head otherwise too many observations.
##          1          2          3          4          5          6
## -10.427013 -10.442028  -7.122691 -10.054845  -6.784711   2.879127
smoke$residi <- third_lim$residuals #add it to dataframe to plot
```

A. Construct a scatter plot of these residuals against age. What does this tell you about the likely validity of our Assumption 3?

```r
ggplot(smoke,aes(x=age,y=residi))+
  theme(text = element_text(size=10, family="LM Roman 10"))+
  geom_point()+
  ggtitle('Residuals against age')+
  labs(y='residuals')
```



Assumption 3 states that the variance across residuals is constant. The plot clearly shows how the variance is larger between ages of 30 to 70, so there is an element of heteroskedasticity. It might be sensible to perform a formal test to verify if assumption 3 is fulfilled.

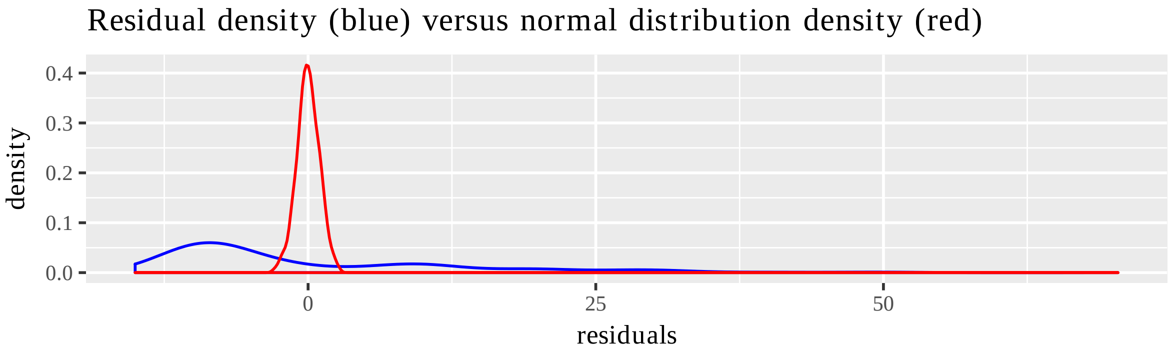
B. Calculate the correlation of these residuals across individuals. What does this tell you about the likely validity of our Assumption 4?

```
dwtest(third_lim)
##
##  Durbin-Watson test
##
## data:  third_lim
## DW = 2.0074, p-value = 0.5258
## alternative hypothesis: true autocorrelation is greater than 0
```

We can reject the H_1 that the autocorrelation among residuals is greater than 0. In other words, assumption 4 is fulfilled.

C. Plot the density of the residuals together with the density of a normal distribution. What does this tell you about the likely validity of our Assumption 5?

```
#add a "normal" series to benchmark:
smoke$benchmark <- rnorm(807)
ggplot(smoke)+
    geom_density( aes(residi), color="blue")+
    geom_density(aes(benchmark), color="red")+
  theme(text = element_text(size=10, family="LM Roman 10"))+
  ggtitle('Residual density (blue) versus normal distribution density (red)')+
  labs(x='residuals')
```



Assumption 5 stands for a normal distribution of the residuals. We can see that the distribution is far away from normal and hence the assumption is not fulfilled.