Problem Set 1 : Empirical

Romain Fernex

2025-02-08

Contents

1 Problem 1 : 2
- 1.1 **Loading the data** . . . 2
- 1.2 **Question 1 : Summarize the dataset.** . . . 2
- 1.3 **Question 2 : Summarize variable salfr (monthly earnings) using the option "detail".** 3
- 1.4 **Question 3 : Tabulate the education variable ddipl1.** . . . 3
- 1.5 **Question 4 : Use the command "label" to associate a label to variable ddipl1.** . . . 4
- 1.6 **Question 5 : Use the command "tabulate" to calculate mean wage by education. Do the same using "by var1: sum var2". Comment on the differences.** . . . 4
- 1.7 **Question 6 : Tabulate adfe and ddipl1 as a two-entry table. How would you interpret adfe=0 and adfe=99** . . . 5
- 1.8 **Question 7 : Calculate the mean of adfe by ddipl1.** . . . 6
- 1.9 **Question 8 : Produce a scatterplot of salfr and adfe for adfe different from 0 and 99.** . . . 6
- 1.10 **Question 9 : Generate a variable lw that is the log of salfr.** . . . 7
- 1.11 **Question 10 : Produce a scatterplot of lw and adfe for adfe different from 0 and 99. Does that suggest to you that some additional trimming would be useful to introduce?** . . . 7
- 1.12 **Question 11 : Calculate the 1st and the 99th percentiles of lw using command "summarize".** . . . 8
- 1.13 Question 12 : Use "_pctile lw, p(0.5, 99.95)" and "return list" to calculate the 0.5 and 99.5 percentiles (say p0050 and p9995). Explain what they mean. . . . 9
- 1.14 **Question 13 : How would you use command "global" to store these values.** . . . 9
- 1.15 **Question 14 : Calculate the variances of lw and adfe and the covariance between lw and adfe for 0 < adfe < 99 and p0050 < lw < p9995 using command "correlate".** 9
- 1.16 **Question 15 : Calculate the total sum of squares (SST) for lw.** . . . 10
- 1.17 **Question 16 : Deduce the OLS estimator of the regression of lw on adfe.** . . . 10
- 1.18 **Question 17 : Regress lw on adfe without selection and repeat the exercise with the selection. Briefly discuss the effect of the trimming.** . . . 10
- 1.19 **Question 18 (a,b,c) : regression with the trimming** . . . 11
- 1.20 **Question 19 : Calculate predicted values and residuals using command predict. Show that the these two variables are uncorrelated.** . . . 12
- 1.21 **Question 20 :** . . . 12

2 Problem 2 : 13
- 2.1 **Question 1 : Please discuss the direction of the relationship. Does the intercept have a useful inter- pretation here? Explain. How much higher is the GPA predicted to be if the ACT score is increased by five points?** . . . 13
- 2.2 **Question 2 : Compute the fitted values and residuals for each observation, and verify that the residuals sum to zero (approximately).** . . . 14
- 2.3 **Question 3 : What is the predicted value for GPA when ACT = 20?** . . . 14

2.4 **Question 4 : For the eight students whose data are provided above, how much of the variation in GPA is explained by ACT? Explain.** 14

1 Problem 1 :

1.1 Loading the data

```
root_dir =  "/Users/rfernex/Documents/Education/SciencesPo/Courses"
Sub_dir = "/S2/Econometrics/TD/Psets/Pset1/Data/ee2002ext.dta"
fullpath = paste0(root_dir,Sub_dir)
employment_data = read_dta(fullpath)
```

1.2 Question 1 : Summarize the dataset.

```
summary <- skim(employment_data)
summary
```

Table 1: Data summary

Name	employment_data
Number of rows	143978
Number of columns	20
Column type frequency:	
character	1
numeric	19
Group variables	None

Variable type: character

skim_variable	n_missing	complete_rate	min	max	empty	n_unique	whitespace
nafg4	0	1	0	2	72849	5	0

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
s	0	1.00	1.53	0.50	1	1.00	2.0	2	2	
agd	0	1.00	46.39	19.47	15	31.00	45.0	61	99	
noi	0	1.00	1.85	1.14	1	1.00	2.0	2	16	
fi	0	1.00	2.93	2.26	1	1.00	2.0	5	8	
pub	72896	0.49	4.20	1.41	1	4.00	5.0	5	5	
hh	82213	0.43	36.24	10.44	0	35.00	35.0	39	99	
salfr	89004	0.38	9065.08	7159.79	0	6000.00	8000.0	10500	524766	
cspp	4362	0.97	45.52	20.50	10	22.00	52.0	63	69	
extri	0	1.00	332.57	93.03	95	260.00	350.0	403	841	
csei	62669	0.56	49.35	14.02	10	41.00	51.0	61	83	
dipl1	26	1.00	6.41	5.38	1	2.00	4.0	12	16	
ddipl1	26	1.00	3.24	2.11	1	1.00	3.0	5	7	

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
adfe	46	1.00	27.02	26.57	0	15.00	18.0	22	99	
enf3	0	1.00	0.07	0.27	0	0.00	0.0	0	3	
enf6	0	1.00	0.16	0.45	0	0.00	0.0	0	4	
enf18	0	1.00	0.66	1.03	0	0.00	0.0	1	9	
np	0	1.00	2.86	1.45	1	2.00	3.0	4	16	
tymen90r	0	1.00	3.87	1.42	1	4.00	4.0	5	5	
nomen	80048	0.44	16570.49	9449.83	1	8486.25	16541.5	24807	32740	

1.3 Question 2 : Summarize variable salfr (monthly earnings) using the option "detail".

```
summary_salfr <- skim(employment_data$salfr)
summary_salfr
```

Table 4: Data summary

Name	employment_data$salfr
Number of rows	143978
Number of columns	1
Column type frequency:	
numeric	1
Group variables	None

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
data	89004	0.38	9065.08	7159.79	0	6000	8000	10500	524766	

1.4 Question 3 : Tabulate the education variable ddipl1.

```
table_ddipl1 <- table(employment_data$ddipl1)
knitr::kable(table_ddipl1, caption = "Tabulation of Education Variable (ddipl1)")
```

Table 6: Tabulation of Education Variable (ddipl1)

Var1	Freq
1	48842
2	9303
3	30663
4	14734
5	11681
6	11111
7	17618

1.5 Question 4 : Use the command "label" to associate a label to variable ddipl1.

```
var_label(employment_data$ddipl1) <- "Highest Level of Education Attained"
```

Adds the label "Highest Level of Education Attained" to the variable titled "ddipl1".

1.6 Question 5 : Use the command "tabulate" to calculate mean wage by education. Do the same using "by var1: sum var2". Comment on the differences.

```
# method 1 : using the R built-in function
mean_wage_by_education <- employment_data %>%
  filter(!is.na(ddipl1) & !is.na(salfr)) %>%
  group_by(ddipl1) %>%
  summarise(mean_income = mean(salfr))

# method 2 : using a hand-made mean function
average <- function(x) {
  mean = sum(x)/length(x)
  return(mean)
}

mean_wage_by_education_manual <- employment_data %>%
  filter(!is.na(ddipl1) & !is.na(salfr)) %>%
  group_by(ddipl1) %>%
  summarise(mean_income = average(salfr))

knitr::kable(mean_wage_by_education, caption = "Mean Wage by Education (Built-in Function)
             ")
```

Table 7: Mean Wage by Education (Built-in Function)

ddipl1	mean_income
1	6885.303
2	8395.550
3	8262.101
4	8981.970
5	10687.947
6	15081.231
7	4573.195

```
knitr::kable(mean_wage_by_education_manual, caption = "Mean Wage by Education (Manual Function)
             ")
```

Table 8: Mean Wage by Education (Manual Function)

ddipl1	mean_income
1	6885.303
2	8395.550
3	8262.101
4	8981.970

ddipl1	mean_income
5	10687.947
6	15081.231
7	4573.195

We find the same result independently of whether we use the built in function or create the function by ourself.

1.7 Question 6 : Tabulate adfe and ddipl1 as a two-entry table. How would you interpret adfe=0 and adfe=99

```
two_way_table <- employment_data %>%
  filter(!is.na(adfe) & !is.na(ddipl1)) %>%
  group_by(adfe, ddipl1) %>%
  summarise(count = n(), .groups = 'drop') %>%
  pivot_wider(names_from = adfe,
              values_from = count,
              values_fill = list(count = 0))

#splits the two-way table in half for improved legibility upon render
n <- ncol(two_way_table)
first_half <- two_way_table[, 1:17]
second_half <- two_way_table[, c(1,18:n)]

# Display the two parts as separate tables
kable(first_half, caption = "Two-Way Table of adfe with ddipl1") %>%
  kable_styling(bootstrap_options = c("striped", "hover", "condensed"),
                full_width = FALSE, position = "center")

kable(second_half) %>%
  kable_styling(bootstrap_options = c("striped", "hover", "condensed"),
                full_width = FALSE, position = "center")
```

Table 9: Two-Way Table of adfe with ddipl1

ddipl1	0	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	1661	15	30	96	104	612	860	3703	4162	19004	3211	7635	3169	2779	778	540
2	3	0	0	0	0	2	2	7	20	165	498	2288	2023	2346	916	606
3	4	0	2	1	1	8	16	78	184	2443	1013	4019	7204	9307	3198	1883
7	41	0	0	1	0	0	2	3	4	4	39	157	127	173	105	100
4	0	0	0	1	2	5	3	10	15	91	65	282	685	3101	2720	3253
5	0	0	0	0	1	0	1	7	7	55	41	105	213	570	626	2036
6	0	0	0	0	0	0	0	0	5	21	16	40	61	168	137	307

ddipl1	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	99
1	180	93	45	30	20	9	2	2	1	2	2	0	2	1	10	67
2	183	85	37	17	23	4	6	2	3	1	3	0	0	1	2	56
3	634	300	119	55	46	18	14	8	9	3	3	2	2	2	8	78
7	102	67	65	41	34	18	10	11	8	3	5	2	1	0	2	16490

4	1804	1300	593	346	153	57	45	20	15	8	8	3	5	2	7	134
5	2358	2294	1454	865	491	191	121	76	29	24	16	17	6	7	27	43
6	544	1196	1893	2067	1696	941	644	390	263	238	110	97	69	44	117	45

The variable adfe corresponds, broadly, to the age at which the respondent completed his studies. adfe = 0 and adfe = 99 are both special cases :

- respondents for which adfe = 0 are respondent who never studied and thus the age at which he/she completed his study is assumed to be zero.
- respondents for which adfe = 99 are respondent who have yet to complete their studies. In other words, they were still students at the time this survey was made.

1.8 Question 7 : Calculate the mean of adfe by ddipl1.

```
mean_adfe_by_ddipl1 <- employment_data %>%
  filter(!is.na(ddipl1) & !is.na(adfe)) %>%
  group_by(ddipl1) %>%
  summarise(mean_income = mean(adfe))

knitr::kable(mean_adfe_by_ddipl1, caption = "Mean of adfe by ddipl1")
```

Table 11: Mean of adfe by ddipl1

ddipl1	mean_income
1	14.28864
2	17.92064
3	17.62987
4	20.47512
5	21.81474
6	24.63804
7	93.87931

1.9 Question 8 : Produce a scatterplot of salfr and adfe for adfe different from 0 and 99.

```
employment_data %>%
  filter(adfe != "0" & adfe != "99") %>%  # Filter out 0 and 99
  ggplot(aes(x = adfe, y = salfr)) +
  geom_point() +
  theme_minimal() +
  labs(title = "Scatter plot of salary by employment status",
       x = "Last year of study  (adfe)",
       y = "Monthly wage (salfr)")
```

```
Warning: Removed 71873 rows containing missing values or values outside the scale range
(`geom_point()`).
```

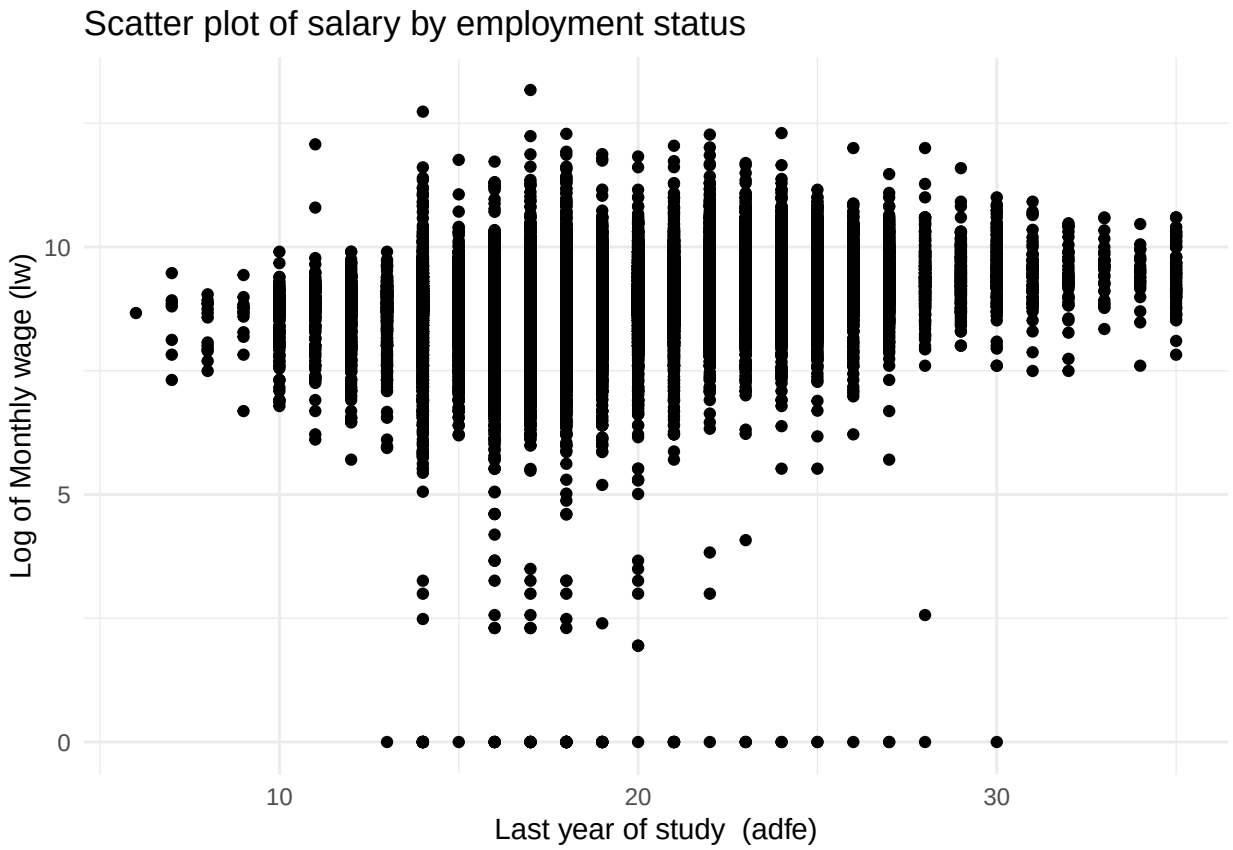
1.10 Question 9 : Generate a variable lw that is the log of salfr.

```
# Exclude values of 'salfr' that are equal to 0 to avoid having -infinite values  in log wages
employment_data$lw <- ifelse(employment_data$salfr!=0, log(employment_data$salfr), NA)

# Remove rows where 'lw' is NA
employment_data <- employment_data[!is.na(employment_data$lw), ]
```

1.11 Question 10 : Produce a scatterplot of lw and adfe for adfe different from 0 and 99. Does that suggest to you that some additional trimming would be useful to introduce?

```
employment_data %>%
  filter(adfe != "0" & adfe != "99") %>%  # Filter out 0 and 99
  ggplot(aes(x = adfe, y = lw)) +
  geom_point() +
  theme_minimal() +
  labs(title = "Scatter plot of salary by employment status",
       x = "Last year of study  (adfe)",
       y = "Log of Monthly wage (lw)")
```



As R does not exclude them automatically, unlike stata, we also filter away the log of null salaries which result in -∞ and bias the data. Based on this new graph, we can see that the relation between the log of wage and the end-of-study age is not linear. It also appears that there remains some extreme values with a decent share of wages that lie around 1.

1.12 Question 11 : Calculate the 1st and the 99th percentiles of lw using command "summarize".

```
percentiles_lw <- employment_data %>%
  summarise(
    q01 = quantile(lw, probs = 0.01, na.rm = TRUE),
    q99 = quantile(lw, probs = 0.99, na.rm = TRUE)
  )

knitr::kable(percentiles_lw, caption = "1st and 99th Percentiles of lw")
```

Table 12: 1st and 99th Percentiles of lw

q01	q99
6.907755	10.34174

1.13 Question 12 : Use "_pctile lw, p(0.5, 99.95)" and "return list" to calculate the 0.5 and 99.5 percentiles (say p0050 and p9995). Explain what they mean.

```
percentiles_lw_extreme <- employment_data %>%
  summarise(
    q0050 = quantile(lw, probs = 0.005, na.rm = TRUE),
    q9995 = quantile(lw, probs = 0.995, na.rm = TRUE)
  )

knitr::kable(percentiles_lw_extreme, caption = "0.5 and 99.5 Percentiles of lw")
```

Table 13: 0.5 and 99.5 Percentiles of lw

q0050	q9995
6.39693	10.59663

Based on observed values we can say the following :

- from q0050 : Exactly 0.5% of monthly log wages observed are inferior or equal to 6.4
- from q9995 : Exactly 99.5% of monthly log wages observed are inferior or equal to 10.6

1.14 Question 13 : How would you use command "global" to store these values.

```
p0050 <- quantile(employment_data$lw, probs = 0.005, na.rm = TRUE)
p9995 <- quantile(employment_data$lw, probs = 0.995, na.rm = TRUE)
```

As R does not really have a stricto sensu equivalent of the Stata "global" command, I store individually the value of each of the computed quartiles in the R environment.

1.15 Question 14 : Calculate the variances of lw and adfe and the covariance between lw and adfe for 0 < adfe < 99 and p0050 < lw < p9995 using command "correlate".

```
# Computes the variance of log wages
variance_lw <- employment_data %>%
  filter(adfe > 0 & adfe < 99 & lw > p0050 & lw < p9995) %>%
  pull(lw) %>%
  var(na.rm = TRUE)
cat("Variance of wages : ", variance_lw, "\n")
```

```
Variance of wages :  0.2827326
```

```
# Computes the variance of end-of-study age
variance_adfe <- employment_data %>%
  filter(adfe > 0 & adfe < 99 & lw>p0050 & lw<p9995 ) %>%
  pull(adfe) %>%
  var(na.rm = TRUE)
cat("Variance of end-of-study age : ", variance_adfe, "\n")
```

```
Variance of end-of-study age :  11.72327
```

```r
# Computes the covariance between lw and adfe
covariance_lw_adfe <- employment_data %>%
  filter(adfe > 0 & adfe < 99 & lw > p0050 & lw < p9995) %>%
  with(cov(lw, adfe, use = "complete.obs"))
cat("Covariance between log-wage and end-of-study year : ", covariance_lw_adfe, "\n")
```

```
Covariance between log-wage and end-of-study year :  0.5669017
```

1.16 Question 15 : Calculate the total sum of squares (SST) for lw.

```r
SST <- employment_data %>%
  with(sum((lw - mean(lw, na.rm = TRUE))^2, na.rm = TRUE))
cat("SST : ", SST, "\n")
```

```
SST :  26663.24
```

1.17 Question 16 : Deduce the OLS estimator of the regression of lw on adfe.

```r
beta = covariance_lw_adfe / variance_adfe
cat("OLS estimator   : ", beta, "\n")
```

```
OLS estimator   :  0.04835695
```

1.18 Question 17 : Regress lw on adfe without selection and repeat the exercise with the selection. Briefly discuss the effect of the trimming.

```r
reg_not_trimmed <- employment_data %>%
  lm(lw ~ adfe, data = .)

# Create the trimmed regression model
reg_trimmed <- employment_data %>%
  filter(adfe > 0 & adfe < 99 & lw > p0050 & lw < p9995) %>%
  lm(lw ~ adfe, data = .)

# Use stargazer to display the regression results
library(stargazer)
stargazer(reg_not_trimmed, reg_trimmed, type = "text",
          title = "Regression Results",
          column.labels = c("Untrimmed", "Trimmed"),
          dep.var.labels = "Log of Monthly Wage (lw)",
          covariate.labels = "Age at End of Study (adfe)",
          omit.stat = c("f"))
```

```
Regression Results
================================================================
                                Dependent variable:
                            -----------------------------------
                                Log of Monthly Wage (lw)
                              Untrimmed          Trimmed
                                 (1)               (2)
----------------------------------------------------------------
Age at End of Study (adfe)     -0.006***          0.048***
```

```
                                (0.0002)          (0.001)

Constant                        9.054***          8.057***
                                (0.006)           (0.012)

---------------------------------------------------------------
Observations                     54,768            52,701
R2                               0.011             0.097
Adjusted R2                      0.011             0.097
Residual Std. Error        0.694 (df = 54766) 0.505 (df = 52699)
===============================================================
Note:                                *p<0.1; **p<0.05; ***p<0.01
```

The R-squared coefficient goes down a little as we reduce the number of observations. Besides the estimate of the impact of the age of the last year of study (adfe) on monthly wages is more pronounced in the trimmed model and has the opposite sign. The untrimmed model indicates that an increase in adfe incurs a decrease in wage. Conversely, The trimmed model indicates an increase in wage with a one year increase in adfe. This is due to the removal of outliers for both the regressor and the dependent variable.

1.19 Question 18 (a,b,c) : regression with the trimming

1.19.1 18.a

```
beta_coeff <- coef(reg_trimmed)["adfe"]
percentage_increase <- (exp(beta_coeff) - 1) * 100
cat("Increasing the last year of education by 1 yields a", sprintf("%.2f%%", percentage_increase), "inc
```

```
Increasing the last year of education by 1 yields a 4.95% increase in wage.
```

1.19.2 18.b

```
#creates the regression model
model <- lm(lw ~ adfe,
            data = employment_data %>%
              filter(adfe > 0 & adfe < 99 & lw > p0050 & lw < p9995))

#sum of squared residuals
SSR2 <- employment_data %>%
  filter(adfe > 0 & adfe < 99 & lw > p0050 & lw < p9995) %>%
  with(sum((lw - predict(model))^2, na.rm = TRUE))

#sum of squares total
SST2 <- employment_data %>%
  filter(adfe > 0 & adfe < 99 & lw>p0050 & lw<p9995) %>%
  with(sum((lw - mean(lw, na.rm = TRUE))^2, na.rm = TRUE))

# sum of squares explained
SSE2 <- SST2 - SSR2



cat("SSE : ", SSE2, "\n")
```

```
SSE :  1444.699
```

```r
cat("SSR : ", SSR2, "\n")
```

```
SSR :  13455.31
```

```r
cat("SST : ", SST2, "\n")
```

```
SST :  14900.01
```

1.19.3 18.c

```r
R_squared <- SSE2/SST2
cat("R-squared : ", R_squared, "\n")
```

```
R-squared :  0.0969596
```

1.20 Question 19 : Calculate predicted values and residuals using command predict. Show that the these two variables are uncorrelated.

```r
predicted_values <- predict(model)
residuals <- residuals(model)
correlation <- cor(predicted_values, residuals, use = "complete.obs")
cat("The correlation coefficient is :", correlation, "\n")
```

```
The correlation coefficient is : -4.449037e-16
```

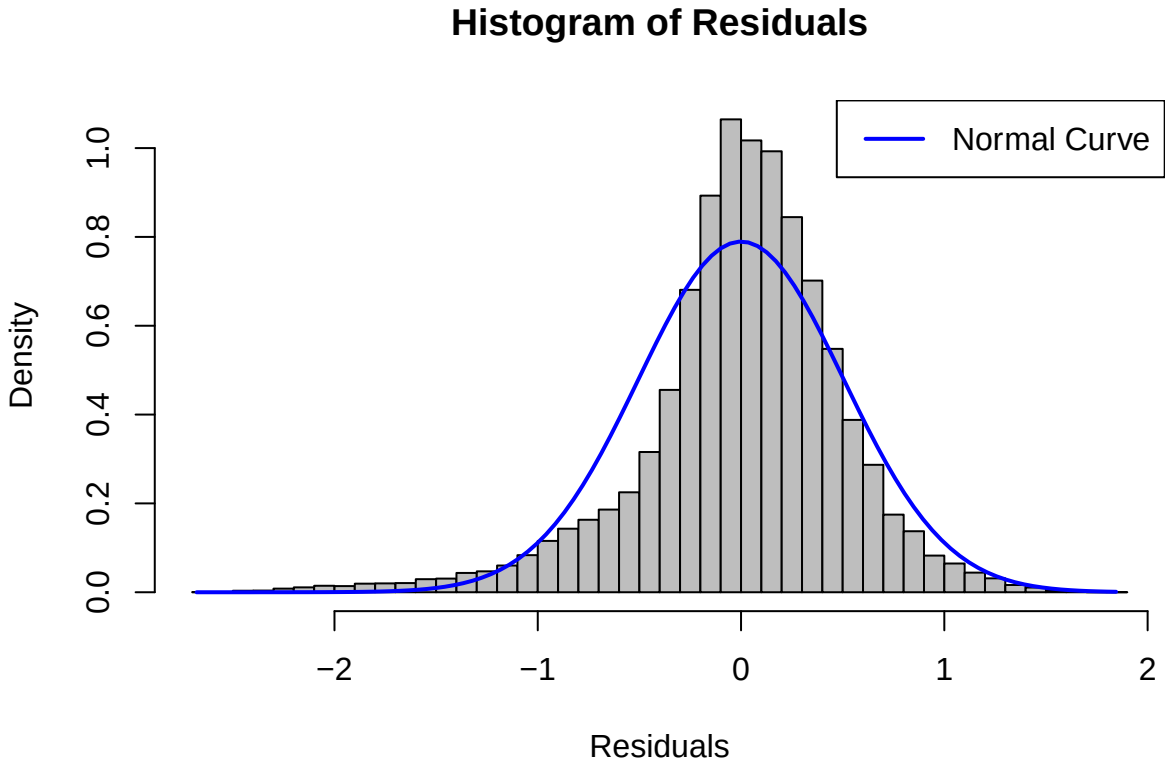
The correlation coefficient strongly tends towards zero which shows that these two variable are indeed not correlated.

1.21 Question 20 :

```r
histogram = hist(residuals,
     freq = FALSE,
     breaks = 50,
     col = "grey",
     main = "Histogram of Residuals",
     xlab = "Residuals",
     ylab = "Density",
     border = "black")

# Add a normal distribution curve with adjusted range
curve(dnorm(x, mean = mean(residuals), sd = sd(residuals)),
      from = min(residuals), to = max(residuals),
      add = TRUE, col = "blue", lwd = 2)

# Add a legend for clarity
legend("topright", legend = c("Normal Curve"),
       col = c("blue"), lwd = 2)
```



2 Problem 2 :

2.1 Question 1 : Please discuss the direction of the relationship. Does the intercept have a useful inter- pretation here? Explain. How much higher is the GPA predicted to be if the ACT score is increased by five points?

```
data_students <- tibble(
  ACT = c(21, 24, 26, 27, 29, 25, 25, 30),
  GPA = c(2.8, 3.4, 3.0, 3.5, 3.6, 3.0, 2.7, 3.7)
)

model <- lm(GPA ~ ACT,
            data = data_students)

stargazer(model, type = "text",
          title = "Regression Summary: GPA vs ACT",
          dep.var.labels = "GPA",
          covariate.labels = "ACT Scores",
          omit.stat = c("f"))
```

```
Regression Summary: GPA vs ACT
================================================
                         Dependent variable:
                     ---------------------------
                                GPA
------------------------------------------------
ACT Scores                    0.102**
```

```
                               (0.036)

Constant                        0.568
                               (0.928)

-----------------------------------------------
Observations                      8
R2                               0.577
Adjusted R2                      0.507
Residual Std. Error       0.269 (df = 6)
===============================================
Note:               *p<0.1; **p<0.05; ***p<0.01
```

If the ACT score is increased by 5 the predicted GPA is higher by 5*0.1022. The intercept tells us the value of the predicted GPA when the ACT score is equal to 0, thus we know predicted GPA will never go below 0.56.

2.2 Question 2 : Compute the fitted values and residuals for each observation, and verify that the residuals sum to zero (approximately).

```r
residuals <- residuals(model)
fitted_values <- fitted(model)
cat("The sum of residuals is : ", sum(residuals), "\n")
```

```
The sum of residuals is :  -4.163336e-17
```

2.3 Question 3 : What is the predicted value for GPA when ACT = 20?

```r
predicted_gpa <- predict(model, newdata = data.frame(ACT = 20))
cat("The predicted GPA is :", predicted_gpa, "\n")
```

```
The predicted GPA is : 2.612088
```

2.4 Question 4 : For the eight students whose data are provided above, how much of the variation in GPA is explained by ACT? Explain.

Based on the R-squared coefficient obtained with the regression, about 58% of the GPA is explained by ACT scores.