Case Study 3 - Group 4

Annika Janson h11829506
Jan Beck h11814291
Franz Uchatzi h1451890

29.11.2020

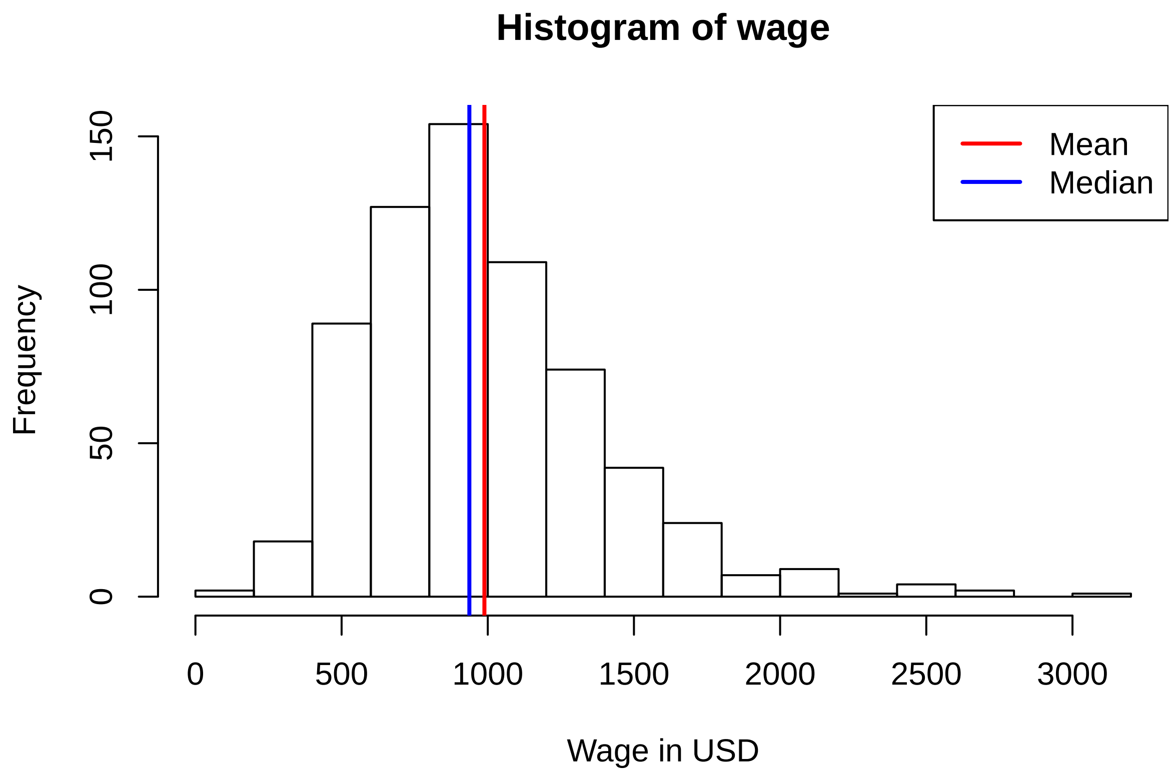
2 Descriptive statistics

Table 1: Summary statistics

Statistic	Mean	St. Dev.	Median	Min	Max
wage	988.475	406.512	937	115	3,078
hours	44.062	7.160	40	25	80
IQ	102.481	14.686	104	54	145
KWW	36.195	7.529	37	13	56
educ	13.680	2.231	13	9	18
exper	11.397	4.258	11	1	22
tenure	7.217	5.056	7	0	22
age	32.983	3.063	33	28	38
married	0.900	0.300	1	0	1
black	0.081	0.274	0	0	1
south	0.323	0.468	0	0	1
urban	0.719	0.450	1	0	1
sibs	2.846	2.241	2	0	14
brthord	2.178	1.488	2	1	10
meduc	10.828	2.823	12	0	18
feduc	10.273	3.288	11	0	18
lwage	6.814	0.412	6.843	4.745	8.032

2.1

The average wage is USD **988.48** and the median wage is USD **937**.



In the histogram we see that the distribution is right-skewed with a few observations exceeding USD 3000. Most observed values are concentrated around an interval of USD ±500 above and below the mean. Median and mean are fairly close to each other with the median being slightly higher due to the large outliers.

2.2

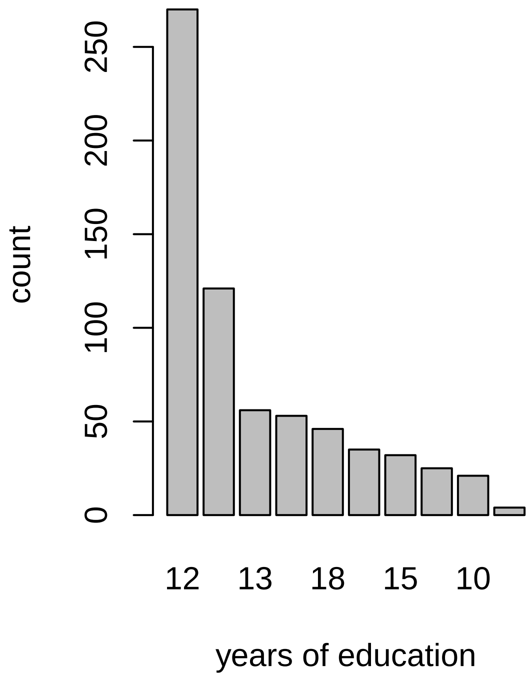
The proportion of workers working more than 40 hours a week is **42.081448%**.

2.3

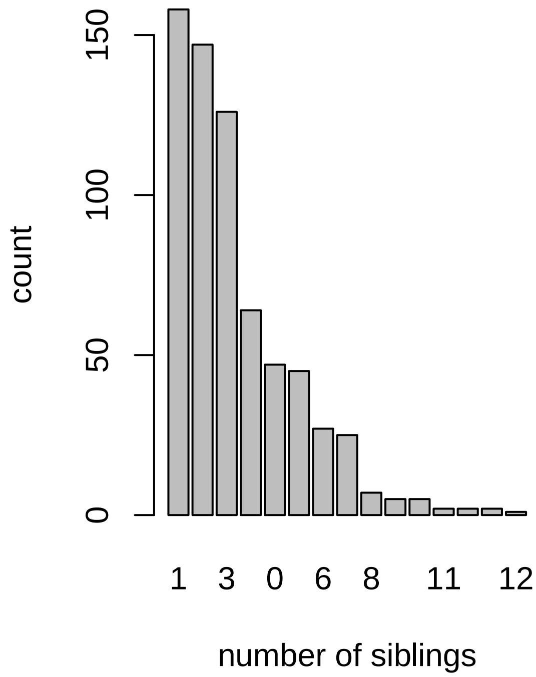
The most common number of years of education among the workers is **12**.

2.4

No. The most frequent sibling pattern is having **1** sibling.



count
0
50
100
150
200
250
12
13
18
15
10
years of education



count
0
50
100
150
1
3
0
6
8
11
12
number of siblings

3 Data modelling

3.1

Table 2: Model summary

	Dependent variable:
	log(wage)
hours	-0.006^{***} $(-0.010, -0.002)$
	$p = 0.007$
educ	0.042^{***} $(0.024, 0.060)$
	$p = 0.00001$
exper	0.012^{**} $(0.003, 0.021)$
	$p = 0.011$
tenure	0.008^{***} $(0.002, 0.014)$
	$p = 0.010$
age	0.015^{**} $(0.003, 0.026)$
	$p = 0.011$
iq	0.005^{***} $(0.002, 0.007)$
	$p = 0.0002$
sibs	0.005 $(-0.011, 0.021)$
	$p = 0.530$
brthord	-0.017 $(-0.041, 0.006)$
	$p = 0.149$
meduc	0.010 $(-0.003, 0.023)$
	$p = 0.117$
feduc	0.011^{*} $(-0.0005, 0.022)$
	$p = 0.061$
Constant	5.142^{***} $(4.714, 5.571)$
	$p = 0.000$
Observations	663
R^2	0.215
Adjusted R^2	0.203
Residual Std. Error	0.368 (df = 652)
F Statistic	17.895^{***} (df = 10; 652)
Note:	*p<0.1; **p<0.05; ***p<0.01

(Parenthesis show 95%-confidence intervals.)

The coefficient of determination is **0.2153517**. Therefore, **21.5352** % of the variation of the dependent variable wage is explained by variation of the independent variables of the current model.

3.2

If we increase X (education) by one year we expect wage to increase approximately[1] by **4.233**% (or by **4.324054**% exactly[2]), c.p.

3.3

For each additional older sibling we expect the wage to decrease by approximately[1] **1.735** % (or by **1.71971**% exactly[2]), c.p. However, the p-value of the variable `brthord` is **0.1487951**. This indicates that we can not reject the null hypothesis $\hat{\beta}_8 = 0$ when using a significance level of $\alpha = 0.05$.

3.4

If we increase X (education) by three years we expect wage to increase approximately[1] by **12.699**% (or by **12.972162**% exactly[2]), c.p.

Analogously, if we increase X (years of education of mother) by three years we expect wage to increase approximately[1] by **3.039**% (or by **3.0529718**% exactly[2]), c.p.

And if we increase X (years of education of father) by three year we expect wage to increase approximately[1] by **3.162**% (or by **3.1797132**% exactly[2]), c.p.

We can see that the largest average effect on wage with three additional years of education of the variables educ, meduc and feduc is achieved by the education of the workers themselves.

This is because the coefficient for `educ` (0.0423318) is larger than that of `meduc` (0.0101251) and `feduc` (0.0105433).

Note that the effects of `meduc` and `feduc` are not significant at the 5%-level.

3.5.1

$H_0 : \beta_{educ} = 0$
$H_1 : \beta_{educ} \neq 0$

The p-value is **0.000004570314** and the t-statistic is **4.622573**. Therefore, we **reject** the null hypothesis at the 5% significance level and assume that an additional year of education has influence on wage, c.p.

3.5.2

$H_0 : \beta_{brthord} = 0$
$H_1 : \beta_{brthord} \neq 0$

The p-value is **0.1487951** and the t-statistic is **-1.44551**. Therefore, we do **not reject** the null hypothesis at the 5% significance level and assume that the variable birth order has no influence on wages, c.p.

3.5.3

$H_0 : \beta_{sibs} = \beta_{brthord} = \beta_{meduc} = \beta_{feduc} = 0$
$H_1 : H_0$ is not true.

---

[1] 100 * $\hat{\beta}_j$%
[2] 100 * ($e_j^{\hat{\beta}}$ - 1)%

As we can see in the summary in 3.1 the variables with a p-value < 0.05 are `sibs`, `brthord`, `meduc`, `feduc` and, therefore, have on average no significant influence on wages individually.

We run a linear hypthesis test and find that the p-value is **0.0034094** and the F-statistic is **3.9732103**. Therefore, we **reject** the null hypothesis at the 5% significance level and assume that the coefficients for `sibs`, `brthord`, `meduc` and `feduc` can not be jointly excluded from the model c.p. as at least one of them is different from 0. However, we do not know which one.

Table 3: Linear hypothesis test: sibs=0 brthord=0 meduc=0 feduc=0

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	656	90.411				
2	652	88.260	4	2.151	3.973	0.003

3.5.4

$H_0 : \beta_{sibs} = \beta_{brthord} = \beta_{meduc} = 0$
$H_1 : H_0$ is not true.

We run a linear hypthesis test and find that the p-value is **0.1551325** and the F-statistic is **1.751907**. We find little evidence in the data that we should reject the null hypothesis that the coefficients for `sibs`, `brthord`, `meduc` and `feduc` are equal to 0 and therefore can not be jointly excluded from the model, c.p.

Table 4: Linear hypothesis test: sibs=0 brthord=0 meduc=0

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	655	88.971				
2	652	88.260	3	0.711	1.752	0.155

3.5.5

$H_0 : \beta_{meduc} - \beta_{feduc} = 0$ or: $H_0 : \beta_{meduc} = \beta_{feduc}$
$H_1 : \beta_{meduc} - \beta_{feduc} \neq 0$

We run a linear hypthesis test and find that the p-value is **0.9678076** and the F-statistic is **0.00163**. We find little evidence in the data that we should reject the null hypothesis that the coefficients for `meduc` and `feduc` are the same, c.p.

Table 5: Linear hypothesis test: meduc-feduc=0

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	653	88.260				
2	652	88.260	1	0.0002	0.002	0.968

3.5.6

Franz:

Annika:

Jan:

4 Simulation Study

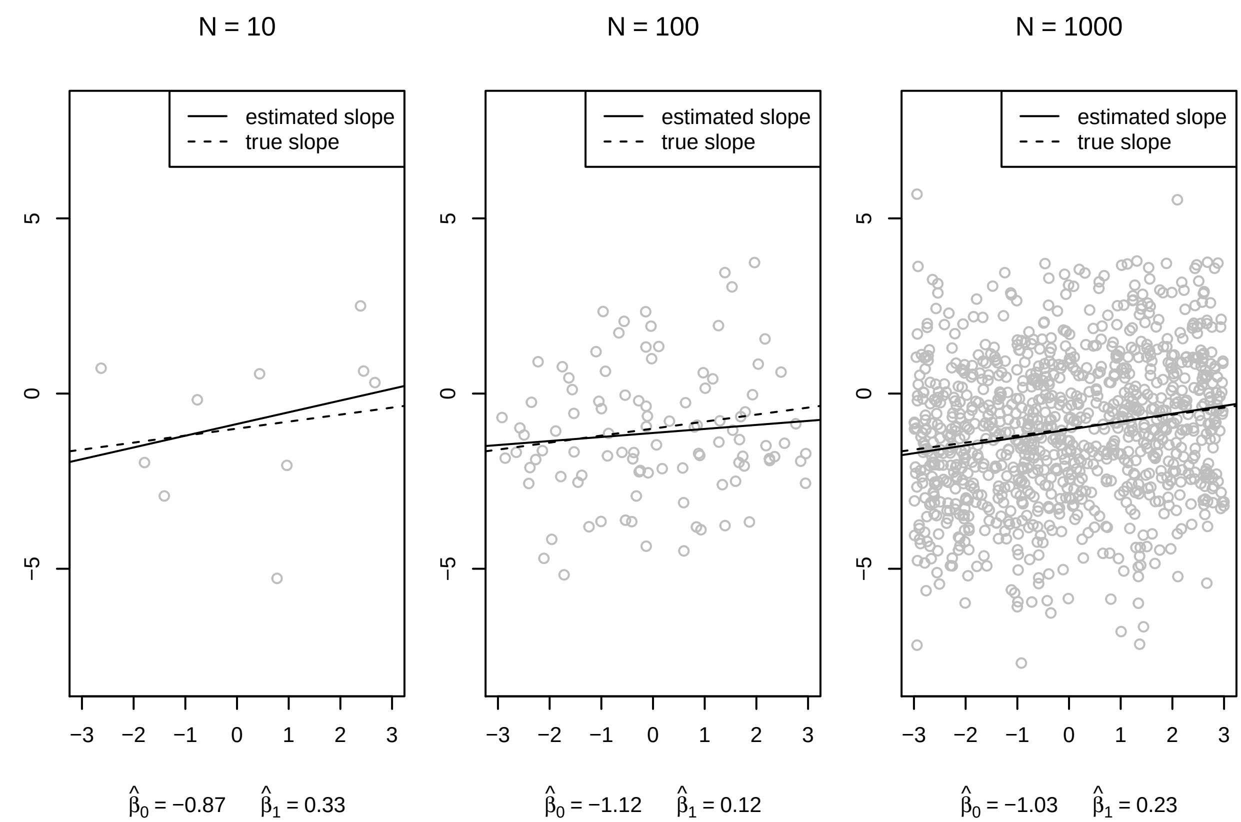
4.1

```r
set.seed(1)

# our parameters according to spec
N1 <- 10
N2 <- 100
N3 <- 1000
beta0 <- -1
beta1 <- 0.2
mu <- 0
sigma <- sqrt(4)
minX = -3
maxX = 3

# model 1
x1 <- x <- runif(N1, min = minX, max = maxX)
u1 <- rnorm(N1, sd = sigma, mean = mu)
y1 <- beta0 + beta1*x1 + u1
lm1 <- lm(y1 ~ x) # using x instead of x1 to show as one row in stargazer output

# model 2
x2 <- x <- runif(N2, min = minX, max = maxX)
u2 <- rnorm(N2, sd = sigma, mean = mu)
y2 <- beta0 + beta1*x2 + u2
lm2 <- lm(y2 ~ x) # using x instead of x2 to show as one row in stargazer output

# model 3
x3 <- x <- runif(N3, min = minX, max = maxX)
u3 <- rnorm(N3, sd = sigma, mean = mu)
y3 <- beta0 + beta1*x3 + u3
lm3 <- lm(y3 ~ x) # using x instead of x3 to show as one row in stargazer output
```

Table 6: Model comparison

	Dependent variable:		
	y1	y2	y3
	(1)	(2)	(3)
x	0.3346334	0.1150630	0.2255274***
	(0.4061975)	(0.1183332)	(0.0370705)
Constant	−0.8683543	−1.1244970***	−1.0266640***
	(0.7405519)	(0.1856205)	(0.0651777)
Observations	10	100	1,000
R^2	0.0782007	0.0095557	0.0357600
Adjusted R^2	-0.0370242	-0.0005509	0.0347938
Residual Std. Error	2.3079310 (df = 8)	1.8542300 (df = 98)	2.0608280 (df = 998)
F Statistic	0.6786786 (df = 1; 8)	0.9454931 (df = 1; 98)	37.0120100*** (df = 1; 998)

Note: *p<0.1; **p<0.05; ***p<0.01



4.2

Calculated 95%-confidence intervals and standard errors for beta0 and beta1 and different Ns:

Table 7:

N	beta0	beta1	se0	se1
10	(-2.5761, 0.8394)	(-0.6021, 1.2713)	0.741	0.406
100	(-1.4929, -0.7561)	(-0.1198, 0.3499)	0.186	0.118
$1,000$	(-1.1546, -0.8988)	(0.1528, 0.2983)	0.065	0.037

We observe that the standard errors decrease for greater Ns and therefore the confidence intervals get smaller. This means that we become more certain of our estimation. For example, only with N=1000 does the confidence interval for β_1 not include 0 and at the 5%-confidence interval, we could therefore reject the null hypothesis $\beta_1 = 0$.

This shows that the OLS estimator is **consistent** i.e. for ever larger Ns the estimator converges "in probability" to the true value of β.

The reason for this can be seen in the formula for the standard error:

$$\text{se}\left(\hat{\beta}_j|X\right) = \frac{\hat{\sigma}}{\sqrt{n}\,\text{sd}\left(x_j\right)\sqrt{1-R_j^2}}$$

All else equal, when $\sqrt{n}$ in the denominator approaches ∞, the overall result approaches 0.

4.3