Case Study 3

WS 2020/2021

Deadline: November 29nd, 2020@ 23:55

1 Data Description

The dataset `WD1` contains information on monthly wages (`wage`) of $N = 663$ US workers from 1980, reported in USD. The following variables are available as explanatory variables:

- `hours`: average weekly hours worked
- `educ`: years of education
- `exper`: years of work experience
- `tenure`: years with current employer
- `age`: age in years
- `IQ`: IQ score
- `sibs`: number of siblings
- `brthord`: birth order (`brthord` is one for a first-born child, two for a second-born child, and so on)
- `meduc`: years of education of mother
- `feduc`: years of education of father

2 Descriptive statistics

Answer the following questions by inspecting the data through histograms and/or summary statistics:

2.1

What is the average and median wage? What does the distribution of wages look like?

```
WD1 <- read.csv("WD1.csv")
```

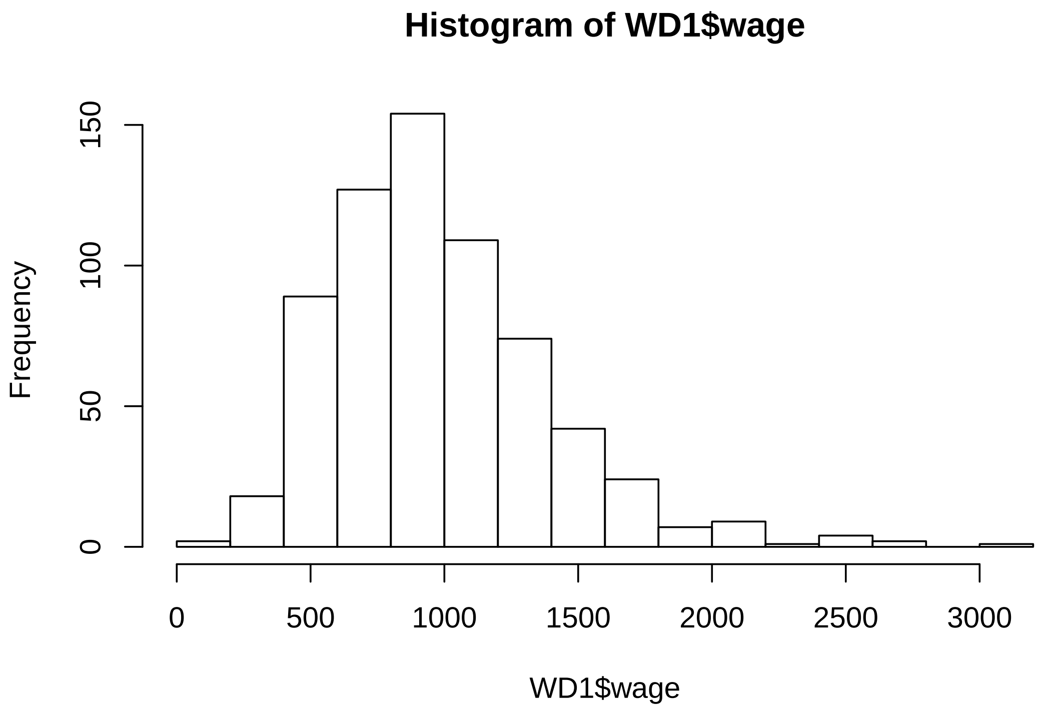
```
mean(WD1$wage)
```

```
## [1] 988.4751
```

```
median(WD1$wage)
```

```
## [1] 937
```

```
hist(WD1$wage)
```



The mean wage is 988 $, the median is 937 $. The distribuion of the wages is slightly skewed to the right.

2.2

What is the proportion of workers working more than 40 hours a week?

```
sum(WD1$hours > 40)/663
```

```
## [1] 0.4208145
```

42% of the workers work more than 40 hours a week.

2.3

What is the number of years of education that is most common among the workers?

```
tab <- table(WD1$educ)
tab[which.max(tab)]
```

```
##  12
## 270
```

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

2.4

Is *single child* the most frequent sibling pattern?

```
table(WD1$sibs)
```

```
##
##   0   1   2   3   4   5   6   7   8   9  10  11  12  13  14
##  47 158 147 126  64  45  27  25   7   5   5   2   1   2   2
```

No, the most frequent sibling pattern is one sibling.

3 Data modelling

Analyze the effect of the working conditions as well as the demographic variables on wage using a multiple regression model.

3.1

Now, regress the variables on the log(*wage*) (lwage) and report the output for the fitted model. How large is the coefficient of determination and how can it be interpreted qualitatively?

```
lm2<- lm(lwage ~ educ + exper + tenure + age +
           meduc + feduc + sibs + brthord + IQ + hours, WD1)
summary(lm2)
```

```
##
## Call:
## lm(formula = lwage ~ educ + exper + tenure + age + meduc + feduc +
##     sibs + brthord + IQ + hours, data = WD1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.81420 -0.21799  0.01747  0.23900  1.27419
##
## Coefficients:
```

```
##              Estimate Std. Error t value Pr(>|t|)    
## (Intercept)  5.142441   0.218199  23.568  < 2e-16 ***
## educ         0.042332   0.009158   4.623 4.57e-06 ***
## exper        0.011935   0.004643   2.571 0.010374 *  
## tenure       0.007870   0.003017   2.609 0.009297 ** 
## age          0.014637   0.005727   2.556 0.010819 *  
## meduc        0.010125   0.006446   1.571 0.116708    
## feduc        0.010543   0.005608   1.880 0.060534 .  
## sibs         0.005030   0.007990   0.630 0.529228    
## brthord     -0.017347   0.012000  -1.446 0.148795    
## IQ           0.004554   0.001195   3.810 0.000152 ***
## hours       -0.005539   0.002020  -2.742 0.006282 ** 
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3679 on 652 degrees of freedom
## Multiple R-squared:  0.2154, Adjusted R-squared:  0.2033 
## F-statistic: 17.89 on 10 and 652 DF,  p-value: < 2.2e-16
```

R^2 is 21.5%. That means, only 21.5% of the variance of the log wages is explained by the variation of the explanatory variables. The model does not fit well.

3.2

Interpret the effect of years of education.

An additional year of education increases the expected wage by 4.23 %, ceteris paribus.

3.3

Interpret the effect of `brthord`. Be particularly careful about the sign.

If the birth order increases by 1 position, the expacted wage will decrease by 1.73 %, ceteris paribus.
The negative sign indicates that the later a worker is in the sibling order, the smaller the expected wage, ceteris paribus.

3.4

Who achieves the largest average effect on `wage` with three additional years of education, ceteris paribus: the workers' mothers, their fathers or the workers themselves?

The effect of an additional year of education is largest for the workers themselves (4.23 %). It is roughly four times the corresponding effect of their parents' (1.01 % and 1.05 %).

3.5

Test the following hypotheses (report the hypotheses, significance level, test statistic, p-value and interpretation)

3.5.1

An additional year of education has no influence on wages, ceteris paribus.

$H_0 : \beta_{educ} = 0$, $H_1 : \beta_{educ} \neq 0$, $\alpha = 0.05$, $T = 4.623$, $p = 4.57e - 06$, H_0 is rejected. The impact of additional years of education on wages is statistically significant.

3.5.2

The birth order has no influence on wages, ceteris paribus.

$H_0 : \beta_{sib} = 0$, $H_1 : \beta_{sib} \neq 0$, $\alpha = 0.05$, $T = 0.630$, $p = 0.529228$. H_0 is not rejected. There is no statistical evidence that the birth order impacts wages.

3.5.3

The variables which have no significant influence on wages individually also do not influence wages jointly. That is, they can be jointly excluded from the model, ceteris paribus.

```r
library(car)
```

```
## Loading required package: carData
```

```r
hypothesis <- c("feduc=0","meduc=0","sibs=0","brthord=0")
test <- linearHypothesis(lm2, hypothesis)
test
```

```
## Linear hypothesis test
##
## Hypothesis:
```

```
## feduc = 0
## meduc = 0
## sibs = 0
## brthord = 0
##
## Model 1: restricted model
## Model 2: lwage ~ educ + exper + tenure + age + meduc + feduc + sibs +
##     brthord + IQ + hours
##
##   Res.Df    RSS Df Sum of Sq      F   Pr(>F)
## 1    656 90.411
## 2    652 88.260  4    2.1514 3.9732 0.003409 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

$H_0 : \beta_{feduc} = 0, \beta_{meduc} = 0, \beta_{sibs} = 0, \beta_{brthord} = 0$. H_1 : At least one of the coefficients is $\neq 0$. $\alpha = 0.05$, $F = 3.9732$, $p = 0.003409$. H_0 is rejected. There is evidence in the data that at least one of the variables has a non-zero impact on wages. Thus, not all 4 variables should be jointly excluded from the model.

3.5.4

The birth order, number of siblings and mother's education years can be jointly excluded from the model, ceteris paribus.

```
hypothesis <- c("sibs=0","brthord=0","meduc=0")
test <- linearHypothesis(lm2, hypothesis)
test
```

```
## Linear hypothesis test
##
## Hypothesis:
## sibs = 0
## brthord = 0
## meduc = 0
##
## Model 1: restricted model
## Model 2: lwage ~ educ + exper + tenure + age + meduc + feduc + sibs +
##     brthord + IQ + hours
##
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1    655 88.971
## 2    652 88.260  3   0.71145 1.7519 0.1551
```

$H_0 : \beta_{sibs} = 0, \beta_{brthord} = 0, \beta_{meduc} = 0$, H_1 : At east one if the parameters is $\neq 0$. $\alpha = 0.05$, $F = 1.7519$, $p = 0.1551$. H_0 is not rejected. There is no evidence that one or more of these variables has a non-zero effect on wages.

3.5.5

The effect of mother's and father's years of education on wages is the same, ceteris paribus.

```
hypothesis <- c("feduc=meduc")
test <- linearHypothesis(lm2, hypothesis)
test
```

```
## Linear hypothesis test
##
## Hypothesis:
## - meduc  + feduc = 0
##
## Model 1: restricted model
## Model 2: lwage ~ educ + exper + tenure + age + meduc + feduc + sibs +
##     brthord + IQ + hours
##
##   Res.Df   RSS Df  Sum of Sq      F Pr(>F)
## 1    653 88.26
## 2    652 88.26  1 0.00022065 0.0016 0.9678
```

$H_0 : \beta_{meduc} = \beta_{feduc}$. $H_1 : \beta_{meduc} \neq \beta_{feduc}$. $\alpha = 0.05$, $F = 0.0016$, $p = 0.9678$. H_0 is not rejected. There is no evidence that the effect on wages of years of education of the workers' mothers and fathers is different.

3.5.6

Do the results motivate you to continue your studies at WU? Why or why not?

The results show that there is a positive impact of years of education on wages. Therefore, if you value money, continuing your studies @WU is probably a good idea.

4 Simulation Study

The purpose of the following simulation study is to illustrate the asymptotic behaviour of a linear regression model. }

4.1

Simulate data $(y_i, x_i)_{i=1,\dots,N}$ with $N_1 = 10, N_2 = 100, N_3 = 1000$ from the simple linear regression model

$$Y = \beta_0 + \beta_1 X + u$$

where the error term u is independent of X and u has a normal distribution with mean 0 and variance $\sigma^2 = 4$. Moreover, X has a uniform distribution on $[-3, 3]$ and the true parameters are $\beta_0 = -1$, $\beta_2 = 0.2$.

Provide your code and the model output of the fitted regression models.

Remark: If you work in R, please set your seed to 1, using the command `set.seed(1)`.

```
set.seed(1)
N1 <- 10
N2 <- 100
N3 <- 1000
sig <- 2
beta0 <- -1
beta1 <- 0.2

u1 <- rnorm(n=N1, mean=0, sd=sig)
u2 <- rnorm(n=N2, mean=0, sd=sig)
u3 <- rnorm(n=N3, mean=0, sd=sig)

X1 <- runif(n=N1,-3,3)
X2 <- runif(n=N2,-3,3)
X3 <- runif(n=N3,-3,3)

Y1 <- beta0 + beta1*X1+u1
Y2 <- beta0 + beta1*X2+u2
Y3 <- beta0 + beta1*X3+u3

model1 <- lm(Y1~X1)
summary(model1)
```

```
##
```

```
## Call:
## lm(formula = Y1 ~ X1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.6530 -0.8674  0.1943  0.8422  1.8278
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -0.8726     0.4670  -1.868   0.0987 .
## X1           -0.2473     0.2872  -0.861   0.4142
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.45 on 8 degrees of freedom
## Multiple R-squared:  0.08483,    Adjusted R-squared:  -0.02957
## F-statistic: 0.7415 on 1 and 8 DF,  p-value: 0.4142
```

```
model2 <- lm(Y2~X2)
summary(model2)
```

```
##
## Call:
## lm(formula = Y2 ~ X2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.6307 -1.2850 -0.0465  1.1807  4.4896
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -0.7326     0.1852  -3.955 0.000145 ***
## X2            0.2304     0.1059   2.175 0.032004 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.838 on 98 degrees of freedom
## Multiple R-squared:  0.04607,    Adjusted R-squared:  0.03633
## F-statistic: 4.732 on 1 and 98 DF,  p-value: 0.032
```

```
model3 <- lm(Y3~X3)
summary(model3)
```

```
##
## Call:
## lm(formula = Y3 ~ X3)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.1209 -1.3943 -0.0389  1.4300  7.5358
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.05467    0.06597 -15.987   <2e-16 ***
## X3           0.12163    0.03771   3.225   0.0013 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.086 on 998 degrees of freedom
## Multiple R-squared:  0.01032,    Adjusted R-squared:  0.009323
## F-statistic:  10.4 on 1 and 998 DF,  p-value: 0.0013
```

4.2

Report 95%-confidence intervals for β_0 and β_1 for the three sample sizes. What effect can you observe? Make a link to the consistency result of the OLS-estimator.

```
confint(model1)
```

```
##                  2.5 %    97.5 %
## (Intercept) -1.9495649 0.2044022
## X1          -0.9094479 0.4148989
```

```
confint(model2)
```

```
##                   2.5 %     97.5 %
## (Intercept) -1.10010056 -0.3650086
## X2           0.02021979  0.4405044
```

```
confint(model3)
```

```
##                   2.5 %     97.5 %
## (Intercept) -1.18412651 -0.9252178
## X3           0.04762434  0.1956329
```

In line with the consistency results of the OLS estimator, the the confidence intervals become narrower with increasing sample size. This expresses that our uncertainty about the estimation error of the model parameter vanishes as the sample size increases.
What is also interesting to observe is that the reported t-tests reject if and only if 0 is not contained in the confidence intervals.

4.3

Perform similar experiments to observe the effects of the following on confidence intervals: (a) the variance of the error terms σ^2 varies, (b) the level of confidence, $1 - \alpha$, varies.

The confidence intervals become wider as σ^2 increases or as α decreases.