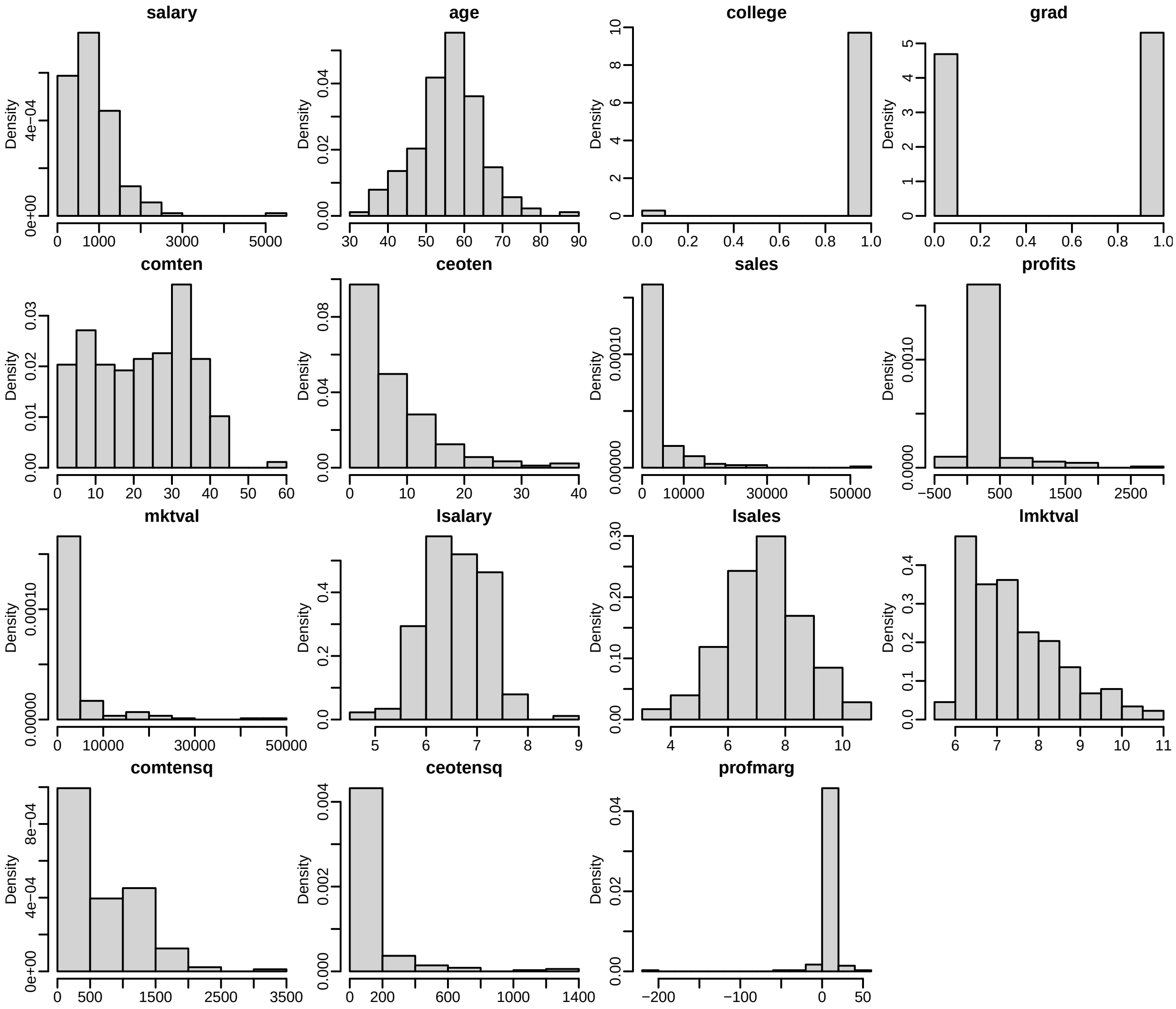
WS 2021/22 1048 - Econometrics I
Case Study 2

Anja Kulagic (h12100543), Jovana Mileusnic (h12100542), Ema Vargova (h11914081)

1 Data Description



The provided contains 15 variables where the ones used in our models below are called *salary*, *sales* and *mktval* which do not seem to be normally distributed. However, these variables on logarithmic scale named *lsalary*, *lsales* and *lmktval* seem to be much closer to a normal distribution with much smaller variance. Moreover, the variable *profits* seems to follow normal distribution as well while *ceoten* does not follow normal distribution (rather a Pareto-like distribution). The rest of the variables are not going to be used in our models.

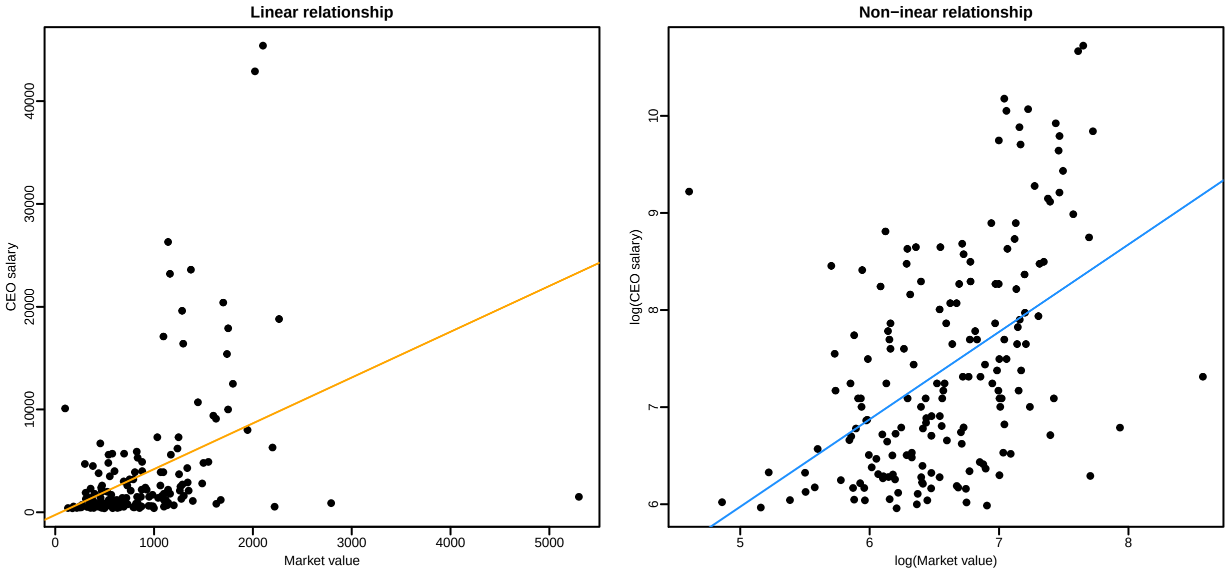
2 Model

Model 1: $Y = \beta_0^* \cdot X_1^{\beta_1} \cdot X_2^{\beta_2} \cdot e^u, \quad \mathbb{E}[u[X_1, X_2]], \quad \beta_0^* > 0$
where Y is the *CEO salary*, X_1 is the *market value* of the company and X_2 are the *sales* of the company.

3 Data Analysis

3.1

Investigating the type of relationship between the *market value* and *CEO salary* and *log(market value)* and *log(CEO salary)*:



From the scatter plots above we can conclude that *salary* and *market value* of the firm have strong **non-linear relationship**. The graph on the left illustrates the presence of non-linearity by showing that for each market value there is a wide range of different salaries. Moreover, even after taking the logarithmic form of the given data, there is still a sign of non-linearity between the variables. On the contrary, relying on the economic theory and economic intuition we gained so far, we would expect a positive linear relation between them. However, this data gives the impression of the existence of other variables beside the market value that have an influence on the dependent variable.

3.2

Model 1: $\log Y = \beta_0 + \beta_1 \log X_1 + \beta_2 \log X_2 + u, \quad \mathbb{E}[u[X_1, X_2]], \quad \beta_0^* > 0$
where $\beta_0 = \log \beta_0^*$.

```
model1 <- lm(lsalary ~ lmktval + lsales, data=data)
summary(model1)
```

```
##
## Call:
## lm(formula = lsalary ~ lmktval + lsales, data = data)
##
```

```
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.28060 -0.31137 -0.01269  0.30645  1.91210
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.62092    0.25441  18.163  < 2e-16 ***
## lmktval      0.10671    0.05012   2.129   0.0347 *
## lsales       0.16213    0.03967   4.087 6.67e-05 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 0.5103 on 174 degrees of freedom
## Multiple R-squared:  0.2991, Adjusted R-squared:  0.2911
## F-statistic: 37.13 on 2 and 174 DF,  p-value: 3.727e-14
```

Computed estimator $\hat{\beta}_1 = 0.10671$ suggests that after a 1% change in the market value, salary increases only by approximately 0.11%, holding other variables constant, which is less than we expected following the economic theory. Not surprisingly, its magnitude could make us suspect that u and x are correlated, thus producing our regression model to be biased, what we will prove in the following tasks (*log-log model* $\longrightarrow \%\Delta y = \beta_1 \%\Delta x$).

3.3

Model 2: $\log Y = \beta_0 + \beta_1 \log X_1 + \beta_2 \log X_2 + \beta_3 X_3 + u, \quad \mathbb{E}[u[X_1, X_2, X_3]], \quad \beta_0^* > 0$
where $\beta_0 = \log \beta_0^*$, Y is the *CEO salary*, X_1 is the *market value* of the company, X_2 are the *sales* and X_3 are the *profits*.

```
model2 <- lm(lsalary ~ lmktval + lsales + profits, data=data)
summary(model2)
```

```
##
## Call:
## lm(formula = lsalary ~ lmktval + lsales + profits, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.27002 -0.31026 -0.01027  0.31043  1.91489
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 4.687e+00  3.797e-01  12.343  < 2e-16 ***
## lmktval     9.753e-02  6.369e-02   1.531    0.128
## lsales      1.614e-01  3.991e-02   4.043 7.92e-05 ***
## profits     3.566e-05  1.520e-04   0.235    0.815
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 0.5117 on 173 degrees of freedom
## Multiple R-squared:  0.2993, Adjusted R-squared:  0.2872
## F-statistic: 24.64 on 3 and 173 DF,  p-value: 2.53e-13
```

a) As we know from the fundamentals of mathematics, logarithms of negative numbers are not defined in the real numbers. With regard to this rule, we are not allowed to consider variable $profits$ in its logarithmic form having in mind that our variable can take negative numbers as well (losses).

b) To check whether the variables included in the Model 2 explain most of the variation in CEO salaries, we look at the R^2 value. By definition, R^2 is a proportion of the total variation of the dependent variable Y, which is explained by the regressors. In our case, R^2 is equal to 0.2993, which means that 30% of variation of the variable $log(salary)$ is explained by these regressors, meaning that 70% of the variation in $log(salary)$ stays unexplained. However, while using the multiple regression model, R^2 can not be reliable measure as it increases whenever a new regressor is added to the model ($R^2 = 0.2991$ in Model 1, whereas in Model 2 $R^2 = 0.2993$). Adjusted R^2, denoted as $\bar{R}^2$, takes it into consideration by "punishing" the additional regressors using correction factor. Here, we face the situation where the **adjusted R-squared** value actually decreases after adding the additional regressor as it does not improve the model fit by a sufficient amount ($\bar{R}^2 = 0.2911$ in *Model 1*, whereas in *Model 2* $\bar{R}^2 = 0.2872$).

3.4

a) In order to examine a significance of variables, we should start with hypothesis testing about the coefficient of the regressor that we are interested in (*profits* in our case).

The **null hypothesis** claims that β_3 is equal to 0. In other words, null hypothesis states no relationship between the variable $log(salary)$ and the variable $profits$.

$$H_0 : \beta_3 = 0$$

Then, we need to use data from our sample to see whether there is enough evidence to reject the null hypothesis in favor of the alternative hypothesis H_A.

$$H_A : \beta_3 \neq 0$$

If we manage to reject null hypothesis, the variable is **significant**.

For checking the significance, we will use a **p-value**. P-value is the probability of obtaining results at least as extreme as the observed results of a statistical hypothesis test, assuming that the null hypothesis is correct. At 5% level of significance, which means that we have to compare p-value with 0.05, we will reject null hypothesis only if ⇒ *P-value < 0.05.*

In our particular example, p-value for the variable $profits$ is far higher than 0.05, hence we can conclude that we **cannot reject null hypothesis**. Therefore, our sample did not provide sufficient evidence to conclude that the relation between the variable $log(salary)$ and the variable $profits$ exists. However, at the same time, that lack of evidence doesn't prove that the relationship does not exist.

b) Comparing p-value with the 5% level of significance, we can come to a conclusion that, again, we **cannot reject null hypothesis**.

In other words, due to the fact that p-value $0.128 > 0.05$, the strength of our evidence falls short of being able to reject the null and prove there is relationship between the variable $log(salary)$ and the variable $log(mktval)$.

Note that the confidence intervals (shown in the table below) of the coefficients for *log(mktval)* and *profits* include both positive and negative values, hence, we cannot say whether there is a positive or negative relationship since the intervals include 0 which refers to the non-existence of significant log-linear relationship between the two regressors and the response variable.

```r
library(knitr)
kable(round(confint(model2), 4))
```

	2.5 %	97.5 %
(Intercept)	3.9374	5.4364
lmktval	-0.0282	0.2232
lsales	0.0826	0.2401
profits	-0.0003	0.0003

3.5

Model 3: $\log Y = \beta_0 + \beta_1 \log X_1 + \beta_2 \log X_2 + \beta_3 X_3 + \beta_4 X_4 + u, \quad \mathbb{E}[u[X_1, X_2, X_3, X_4]], \quad \beta_0^* > 0$ where $\beta_0 = \log \beta_0^*$, Y is the *CEO salary*, X_1 is the *market value* of the company, X_2 are the *sales*, X_3 are the *profits* and X_4 is the *CEO's tenure.*

```r
model3 <- lm(lsalary ~ lmktval + lsales + profits + ceoten, data=data)
summary(model3)
```

```
## 
## Call:
## lm(formula = lsalary ~ lmktval + lsales + profits + ceoten, data = data)
## 
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -2.48726 -0.29259  0.00808  0.29959  1.85560 
## 
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)    
## (Intercept) 4.558e+00  3.804e-01  11.982  < 2e-16 ***
## lmktval     1.018e-01  6.304e-02   1.615   0.1081    
## lsales      1.622e-01  3.949e-02   4.108 6.17e-05 ***
## profits     2.903e-05  1.504e-04   0.193   0.8471    
## ceoten      1.165e-02  5.354e-03   2.176   0.0309 *  
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
## 
## Residual standard error: 0.5063 on 172 degrees of freedom
## Multiple R-squared:  0.3181, Adjusted R-squared:  0.3022 
## F-statistic: 20.06 on 4 and 172 DF,  p-value: 1.42e-13
```

The coefficient β_4 means that, holding all the other variables fixed, another year of CEO tenure is predicted to increase log(salary) by 0.01165, which translates into a roughly 1.17% [100(.092)] increase in salary (*log-level model* $\longrightarrow \%\Delta y = (100\beta_1)\Delta x$).

3.6

```r
prediction <- as.numeric(exp(predict(model3,newdata=data.frame(
  lsales=log(15E3), lmktval=log(3E3), profits=700,ceoten=10))))
print(prediction)
```

```
## [1] 1175.49
```

The predicted CEO salary is 1,175 thousand USD, **on average**.

3.7

```
correlation <- cor(data$lmktval,data$profits)
print(correlation)
```

```
## [1] 0.7768976
```

The correlation is 0.78.

While computing the correlation between two independent variables, we are searching for **multicollinearity** and **omitted variable bias**.

Correlation coefficient could be classified as follows:

Stregth of correlation	Range
Value weak	0.00-0.19
Weak	0.20-0.39
Moderate	0.40-0.59
Strong	0.60-0.79
Very strong	0.80-1.00

When the **imperfect multicollinearity** is present, OLS struggles to precisely estimate $\hat{\beta}_i$ since $\hat{\beta}_i$ still stands for a consistent and unbiased estimator of β_i, but it has a larger variance due to the fact that X_i and X_j are highly correlated. Namely, if the least square assumption holds (the error homoskedacity and $\mathbb{E}[u|X_1, X_2, X_3] = 0$), then the OLS estimators in the regression will be unbiased, but it will have larger variance as the $\sigma_{\hat{\beta}_i}$ is inversely proportional to $1 - \rho^2_{XY}$. Formally, this problem can be expressed as follows:

$$\sigma^2_{\hat{\beta}_i} = \frac{\sigma^2_u}{n(1 - \rho^2_{X_iXj})\sigma^2_{X_i}}$$

Those particular variables are highly correlated, with the correlation coefficient value of 0.78, which leads us to the high variance value of β_1 and imprecisely estimated value of this coefficient.

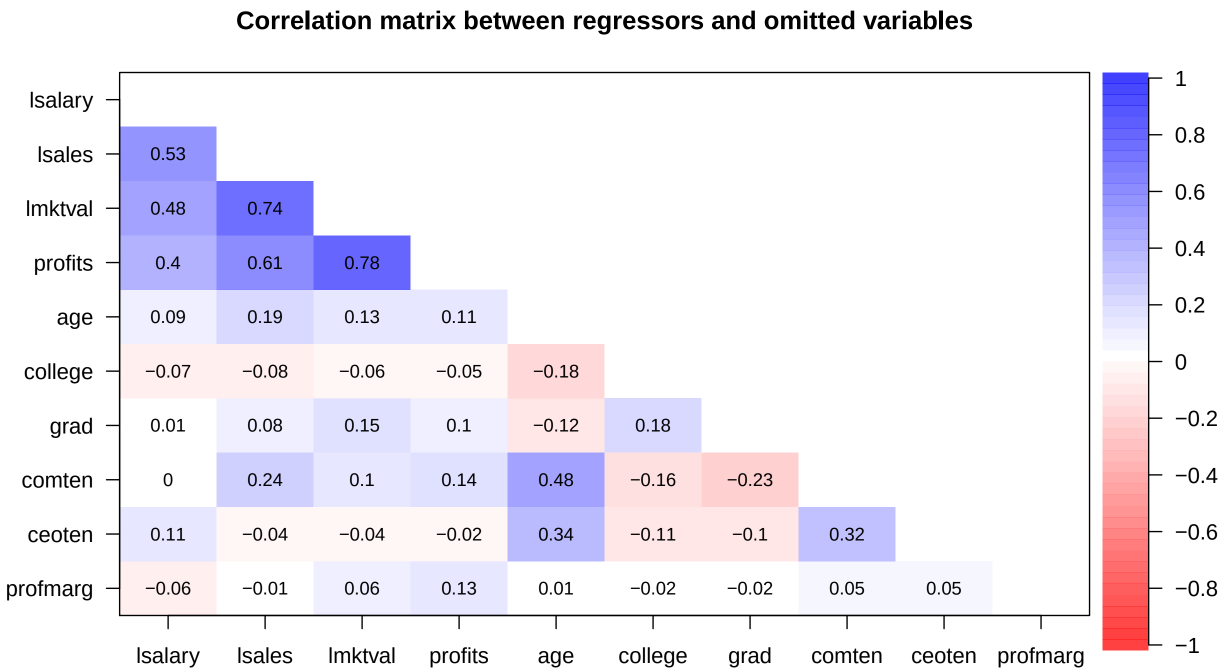
However, excluding the variable $profit$ might decrease the variance of $\hat{\beta}_1$, but it might lead to the issue of omitted variable bias.

Omitted variable bias (OVB): "The influences on the dependent variable which are not captured by the model are collected in the error term, which we so far assumed to be uncorrelated with the regressor. However, this assumption is violated if we exclude determinants of the dependent variable which vary with the regressor. This might induce an estimation bias, i.e., the mean of the OLS estimator's sampling distribution is no longer equal to the true mean."

As we can see in the correlation matrix below, our regressors face mostly week correlation with omitted variables of the Model 3, hence, there is no OVB committed with the given dataset. However, there could be other variables which are not captured in the dataset for which we would need to check the OMV to make sure that our model is not highly impacted by the bias.

Nevertheless, the fact that $\hat{\rho}_{lmktval,profts} = 0.78$ is a cause for concern that omitting variable $profits$ leads to a positively biased estimate of *lmktval* since the variable is very strongly correlated with the regressor causing **omitted variable bias**. Furthermore, the correlation between $lsales$ and $profits$ is also strong.

```r
library("psych")
cor.mat <-cor(data[c(10:12,8,2:6,15)])
corPlot(cor.mat, diag=F, upper=F, cex = 0.8, MAR=4,
        main="Correlation matrix between regressors and omitted variables")
```



As a consequence we expect $\hat{\beta}_1, \hat{\beta}_2$, the coefficients on the logarithmic value of the firm's sales (X_1) and the firm's market value (X_2), to be too large. Put differently, the OLS estimate of suggests that higher market value and sales leads up to higher CEO salary, but their effect is overestimated as it captures the effect of making higher profit, too.

Following the reasoning above, we end up with a positive but much smaller estimates of the coefficient $\hat{\beta}_1, \hat{\beta}_2$ in Model 3, when compared to the Model 1, where $\hat{\beta}$ of *lmktval* is no longer significant after including the strongly correlated variable $profits$ which is not significant either due to their high variance of betas.

Hence, including the variable $profits$ in the model helps to tackle OVB but causes the problem of imperfect multicollinearity which presents us with **bias-variance trade-off**.

3.8

i) Finding an estimate of the variance of the residuals using the following formula:

$$\tilde{\hat{\sigma}}^2 = \frac{SSR}{N}$$

where
$SSR = \sum_{i=1}^{N} \hat{u}_i^2$ and N is the number of observations.

```r
model1.res.var <- sum(model1$residuals^2)/nrow(data)
print(model1.res.var)
```

```
## [1] 0.2559867
```

ii) Computing the value of the unbiased estimator of the error variance σ^2 using the following formula:

$$\hat{\sigma}^2 = \frac{SSR}{df}$$

where
$SSR = \sum_{i=1}^{N} \hat{u}_i^2$, $df = N - K - 1$ while N is the number of observations and K is the number of predictors.

```r
df <- model1$df.residual
model1.error.var <- sum(model1$residuals^2)/df
print(model1.error.var)
```

```
## [1] 0.2604003
```

iii) Estimating the covariance matrix of the OLS estimators of the parameters β_1, β_2:

```r
require(knitr)
model1.coeff.cov <- vcov(model1)
kable(round(model1.coeff.cov,4))
```

	(Intercept)	lmktval	lsales
(Intercept)	0.0647	-0.0080	-0.0006
lmktval	-0.0080	0.0025	-0.0015
lsales	-0.0006	-0.0015	0.0016

4 Theory

4.1

a) The model studying college grade point average (GPA) is given as follows:

$$GPA = \beta_0 + \beta_1 study + \beta_2 sleep + \beta_3 work + \beta_4 leisure + u$$

where

$$study + sleep + work + leisure = 168$$

which is the sum of hours spent on all the activities per week.

This implies that the variables are mutually exclusive which means that **we cannot change one variable and hold the other constant** as they face the problem of **perfect multicollinearity**.

b) Rewriting the model given the condition of the sum of the regressors:

$$\begin{aligned} GPA =& \beta_0 + \beta_1(168 - sleep - work - leisure) + \beta_2(168 - study - work - leisure) + \\ & \beta_3(168 - study - sleep - leisure) + \beta_4(168 - study - sleep - work) + u \end{aligned}$$

This results in the $\hat{\beta}_i$ being defined as follows:
$study = 168 - sleep - work - leisure$
$sleep = 168 - study - work - leisure$
$work = 168 - study - sleep - leisure$
$leisure = 168 - study - sleep - work$

As we can see, all the values are interdependent on each other which results in the correlation matrix very similar to the one below.

	study	sleep	work	leisure
study	1	1	1	1
sleep	1	1	1	1
work	1	1	1	1
leisure	1	1	1	1

Hence, this model faces the issue of **perfect multicollinearity** meaning that the predictors $X_1, ..., X_K$ may be expressed as linear function of the remaining predictors which is a violation of assumption about the predictors in multiple OLS regression model which exposes the model to the following issues:

1) The OlS estimator cannot be found as **$\mathbf{X}'\mathbf{X}$ is not invertible**. This is due the reason that $\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$ where unique estimator for β can be obtained only when $\mathbf{X}'\mathbf{X}$ is invertible.

2) The amount of values γ having the same $SSR(\gamma)$ is infinite.

3) **The parameters are not identified** in the regression model as the model does not contain only dependent variables about students which makes R^2 value to be very close to 1 causing the variance of the corresponding β estimator to have infinite variance.

$$\mathbb{V}(\hat{\beta}_j|\mathbf{X}) = \lim_{R_j^2 \to 1} \frac{\sigma^2}{N s_{x_j}^2 (1 - R_j^2)} \Rightarrow \infty$$

4.2

Considering the fact that the temperature in degree Kelvin can be computed by adding the number 273.15 (absolute value) to the temperature in degree Celsius, it will affect our solution by changing the intercept of the regression function.

$$Y_{celsius} = \beta_0 \ + \beta_1 X$$
$$\Downarrow$$
$$Y_{kelvin} = \beta_0 + 273.15 + \beta_1 X$$

The new value of our intercept is consisted of the old value β_0 plus 273.15 which represents the difference caused by the change in the unit of measure of our predictor. Note that the goodness-of-fit of the model should not depend on the units of measurement of our variables, thus it can be shown that R^2 is invariant to changes in the units of y or x.

Bonus: In order to see the influence of adding new variable → *precipitation* (X_2) to our model, firstly we have to check potential correlation between the regressors (*air pressure and precipitation*).

Science has proved that air pressure and precipitation are negatively correlated which means when the pressure is low, the air is free to rise into the atmosphere where it cools and condenses. Consequently, condensation forms clouds, causing the rain. **For more on this topic, see the article in this link.** This fact draws a conclusion that existing correlation indicates the presence of **imperfect multicollinearity**.

"If an omitted variable is a determinant of the dependent variable and is correlated with the regressor, the OLS estimator of the slope coefficient will be biased and will reflect both the effect of the regressor and the effect of the omitted variable."

The fact that $\rho_{Y,X_1} > 0$, points to the conclusion that we can expect OLS estimate for the $\hat{\beta}_1$ (the coefficients on the air pressure X_1) to be positive. However, after computing the correlation $\rho_{X_1,X_2} < 0$ and due to an inverse relationship between pressure and temperature, we can claim the existence of omitted variable bias before adding *precipitation*(X_2). As a consequence, our expected $\hat{\beta}_1$ should be negatively biased. Put differently, the OLS estimate suggests that higher air pressure indicates higher temperature, but their effect is underestimated as it captures the effect of the precipitation, too.

$$\hat{\beta}_i \xrightarrow{p} \beta_i + \rho_{X_i Y} \frac{\sigma_u}{\sigma_{X_i}}$$

OVB prevents the estimator from converting in probability to the parameter value, whereas the strength and direction of the bias are determined by ρ_{XY}.

Following the reasoning above, we should still end up with a positive but higher coefficient estimate $\hat{\beta}_1$ than before and a negative value of estimate $\hat{\beta}_2$.

As mentioned previously, if the correlation coefficient is closer to 1, presence of multicollinearity makes it impossible to compute a considered model. However, presence of the perfect linear combination between observed variables is rare to find in practice, while the imperfect multicollinearity does not necessarily represent an error, but rather just a feature of OLS estimators, their imprecisely estimated values and high variances. In practice, this problem is called *"bias-variance trade-off* which is especially noticeable when we are dealing with small samples.