# Solutions to exercises

Listed below are the solutions to the exercises.

All solutions are found using RStudio, though **you should only do the exercises in RStudio if indicated in the list of exercises.** This may result in slight differences in numerical answers, which is due to rounding errors.

The solutions may often be computed in different ways and when two solutions are given it does not necessarily mean that more solutions does not exist. However, when two solutions are given we encourage you to think about why these two solutions are equivalent.

```
library(mosaic)
```

**9.4:**

**a)**

The intercept is 22 meaning that when x 'the social expenditure as a percentage of gross domestic product' is 0 then y 'the child poverty rate' is 22%.

The slope is -1.3 mean that whenever x 'the social expenditure as a percentage of gross domestic product' is increased by 1 percent point then y 'the child poverty rate' decrease by 1.3 percent points.

**b)**

For USA

```
22 - 1.3 * 2
```

```
## [1] 19.4
```

For DK

```
22 - 1.3 * 16
```

```
## [1] 1.2
```

**c)**

The correlation (measure of linear dependence) is between -1 and 1, where 0 implies no linear dependence while -1 and 1 implies perfect linear dependence. Since the correlation is negative, there is a negative association. That is, when x 'the social expenditure as a percentage of gross domestic product' increase y 'the child poverty rate' decrease. We **cannot** say anything about slope of the relationship between the two variables.

**9.11:**

**a.i)**

(20;90)

**a.ii)**

(37.5;40)

**b)**

```
pred <- -0.13 + 2.62 * 34.3
pred
```

```
## [1] 89.736
```

```
res <- 45.1 - pred
res
```

```
## [1] -44.636
```

The actual cell-phone usage is much lower than expected since the residual is negative.

**c)**

The correlation is positive since when the GDP increase the cellular usage seem to increase.

**9.13:**

The prediction equation is

$$\hat{y} = a + bx \qquad \Leftrightarrow \qquad a = \hat{y} - bx$$

and by the measure of correlation

$$r = b\left(\frac{s_x}{s_y}\right) \qquad \Leftrightarrow \qquad b = r\left(\frac{s_y}{s_x}\right)$$

Hence

```
b <- 0.6 * 120 / 80
b
```

```
## [1] 0.9
```

```
a <- 500 - b * 480
a
```

```
## [1] 68
```

and

$$\hat{y} = 68 + 0.9x$$

**9.27:**

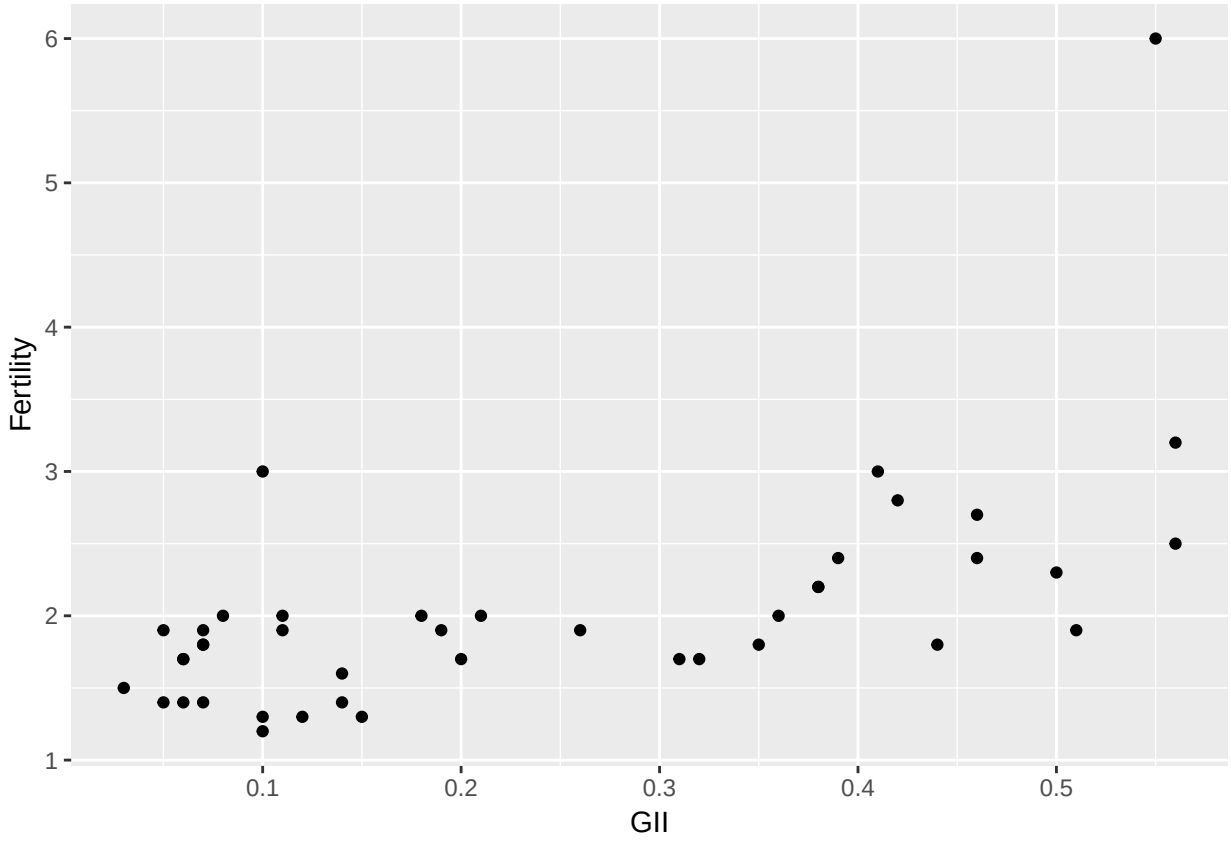
**Additional sub-exercises:**

Import data:

```
UN <- read.table("https://asta.math.aau.dk/datasets?file=UN2014.dat", header = TRUE)
```

Create relevant figures:

```
gf_point(Fertility ~ GII, data = UN,
         xlab = "GII", ylab = "Fertility")
```




Fit linear regression model:

```
fit <- lm(Fertility ~ GII, data = UN)
summary(fit)
```

```
##
## Call:
## lm(formula = Fertility ~ GII, data = UN)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.8723 -0.3619 -0.1284  0.2282  3.1183
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   1.3781     0.1717   8.027 7.27e-10 ***
## GII           2.7338     0.5795   4.717 2.90e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6448 on 40 degrees of freedom
## Multiple R-squared:  0.3575, Adjusted R-squared:  0.3414
## F-statistic: 22.25 on 1 and 40 DF,  p-value: 2.902e-05
```

**a)**

Is 'fertility rate' and 'the gender inequality index' (GII) dependent?

**b)**

$$\hat{y} = 1.378 + 2.734x.$$

A country with a score of 0 on the GII has a predicted fertility rate of 1.378 and each time the GII increase by 1 the fertility rate increase by 2.734.

**c)**

$R$ is the correlation between predictions and observations which for simple linear regression is equivalent to the correlation between the response and the explanatory variable. Thus an R value of 0.598 indicate a 'not so impressive' to moderate linear dependence between fertility rate and the GII. In addition, the correlation is positive indicating a positive relationship (when the GII increase the fertility rate increase).

$R^2$ is a similar measure giving how much of the variation in the response variable is explained by the explanatory variables relative to the total variation response variable; thus the interpretation is elegant. In this case with $R^2 = 0.357$, 35.7% of the total variation of fertility rate is explained by the GII.

**d)**

In conclusion the GII seems to be important in relation to modelling the fertility rate, however it seems reasonable to include more variables in the model that may explain fertility (we may look forward to this in the next lecture).

**9.33:**

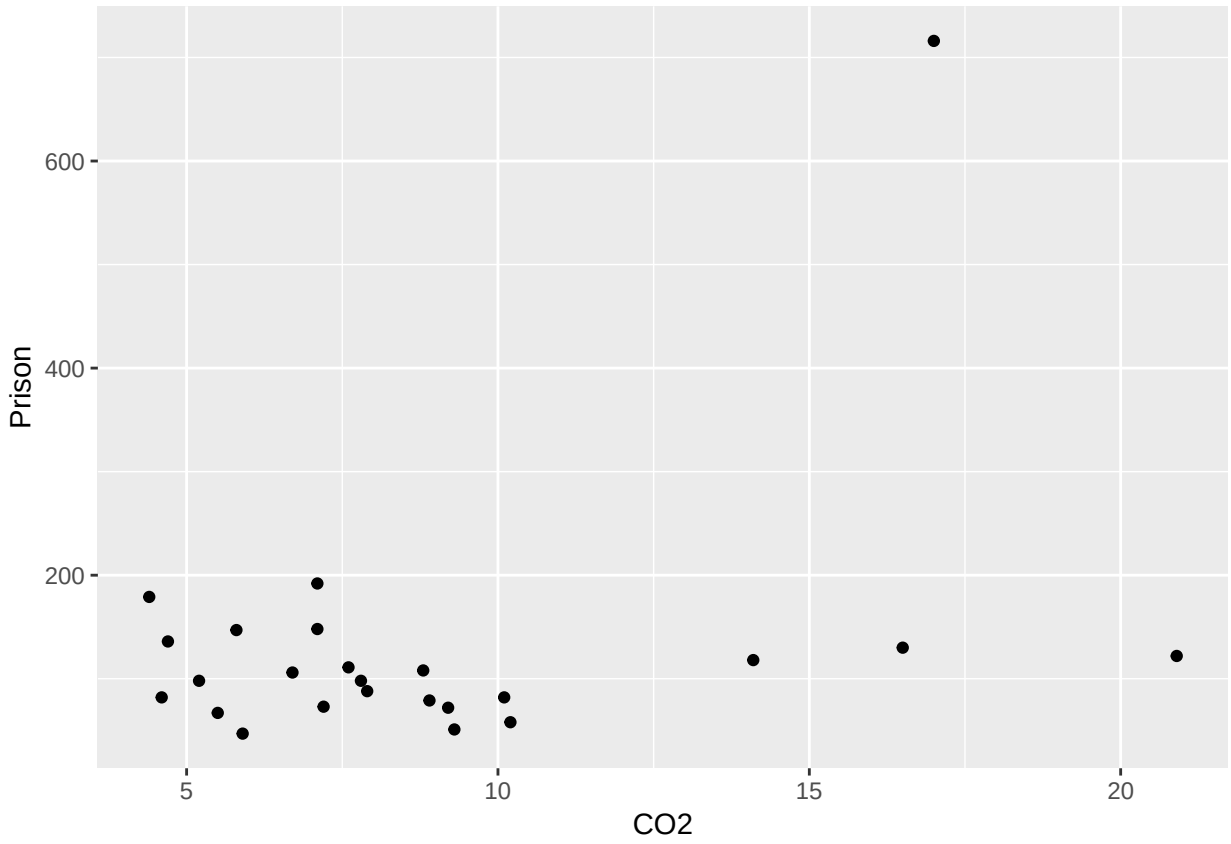
**Additional sub-exercises:**

Import data:

```
oecd <- read.table("https://asta.math.aau.dk/datasets?file=OECD_Agresti_ed5.dat", header = TRUE)
```

Create relevant figures:

```
gf_point(Prison ~ CO2, data = oecd,
         ylab = "Prison", xlab = "CO2")
```




Compute correlation:

```
cor(Prison ~ CO2, data = oecd)
```

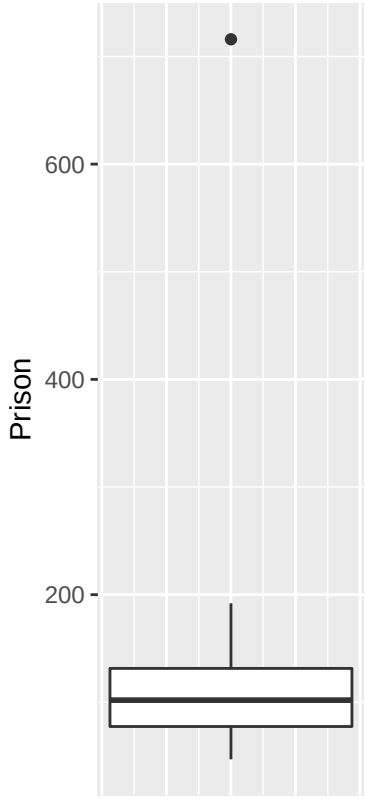
```
## [1] 0.390552
```

There is a weak positive linear relationship between carbon dioxide emissions and prison populations. The positivity means that increasing carbon dioxide emissions increases the prison population.

**a)**

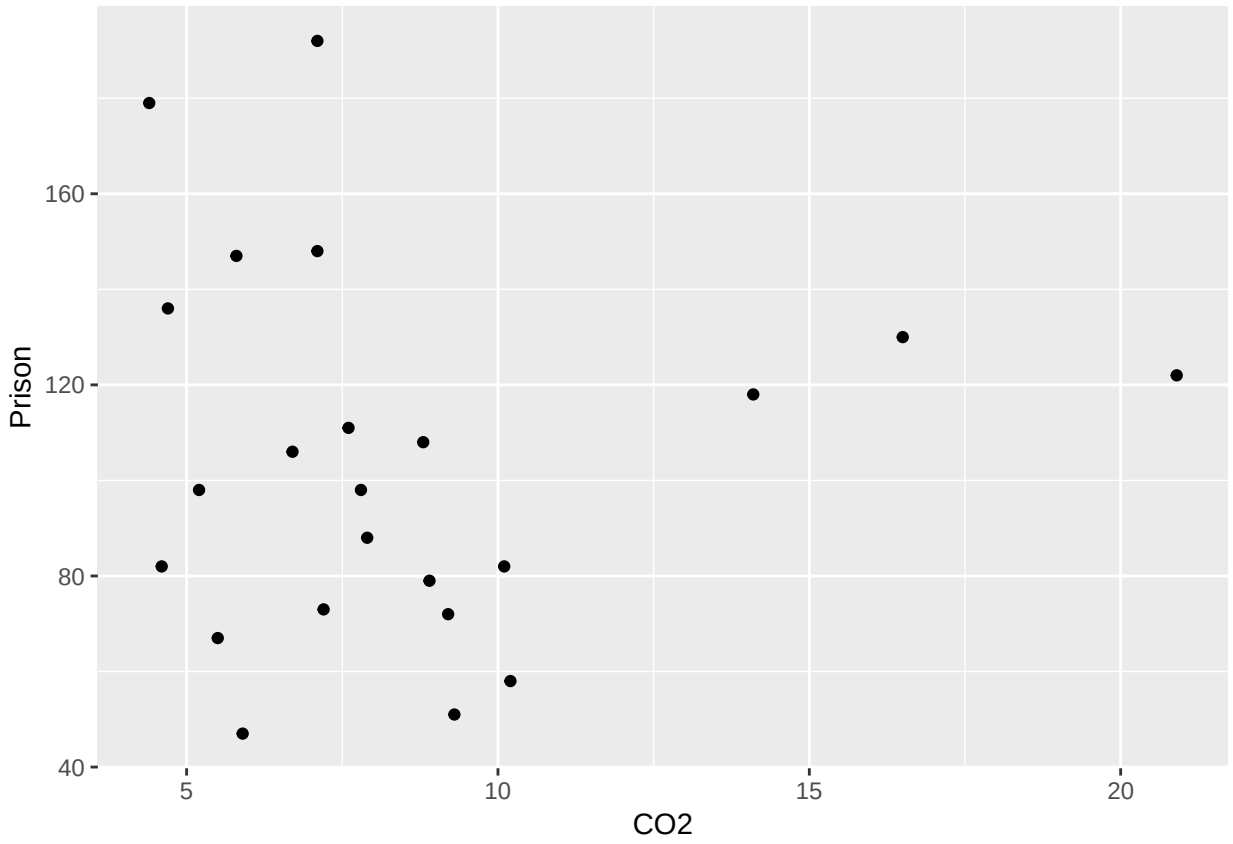
There is a nation with a very high prison population (over 700). When looking at a boxplot of prison population we also see that this is an outlier.

```
gf_boxplot(Prison ~ 1, data = oecd,
           xlab = "") +
  theme(axis.ticks.x = element_blank(),
        axis.text.x = element_blank())
```




Removing the point and computing correlation yields the following:

```
oecd2 <- subset(oecd, Prison != max(Prison))
gf_point(Prison ~ CO2, data = oecd2,
         ylab = "Prison", xlab = "CO2")
```




```
cor(Prison ~ C02, data = oecd2)
```

```
## [1] 0.0004736938
```

We see no correlation what so ever, showing us that one data point may have a huge influence on summary statistics (such as the mean, variance, correlation, and so on), thus it is important to always look at plots of the data.