STK4990 - Statistical Methods and Applications
How to write a mandatory assignment

Vinnie Ko

February 14, 2020

This document is an example of how one can write a mandatory assignment in a reasonable way. We solve *Exercise 6* from week 1 as if it is a mandatory assignment that has to be handed in.

For those of you who are not familiar with LaTeX, but wish to use it, this document also functions as a template that you can follow and modify. Please note that the use of LaTeX is not compulsory for the mandatory assignment. You are free to use any other word processor (e.g. Microsoft Word). Yet, regardless of which program you use, you have to make sure that your `R` code, its output and plots are clearly and properly displayed.

Dos and don'ts

- You are only allowed to hand in a single PDF file.
- The PDF file should contain your name, course and assignment number.
- All texts, tables and plots should be clearly readable.

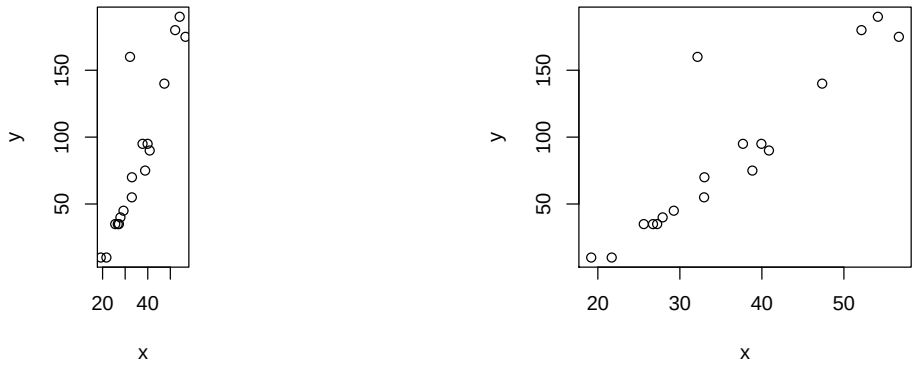

Figure 1: Left: A plot that is badly scaled. Right: Same plot as left, but now properly scaled.

- You should include all the codes that you used, including its output.
- Your code should be placed in the corresponding part of the problems. It should not be included as an appendix at the end of the document.
- Comment your code. So that other people can easily understand your code.
- "normal texts" (e.g. interpretation of the result) should not be written as a comment in `R` code.
- Answer all the questions that are asked.

Problem 1

a)

I used the following code to create a scatter plot.

```
> # Read the data into R from the web.
> HERS.data = read.table("http://www.uio.no/studier/emner/matnat/math/STK4900/v17/
    hers.sample.txt", header=T)
>
> # First observations in the data.
> head(HERS.data)
  age sbp
1  75 168
2  73 116
3  69  96
4  64 119
5  54 137
6  72 163
>
> # Last observations in the data.
> tail(HERS.data)
    age sbp
271  64 126
272  56 115
273  53 126
274  74 156
275  66 143
276  62 175
>
> # Create a scatter plot
> # Open a pdf device.
> pdf("./Exercise_6_a.pdf", width = 6, height = 6)
> # Create a scatter plot
> plot(
+   x=HERS.data[,"age"],
+   y=HERS.data[,"sbp"],
+   xlab="age",
+   ylab="sbp"
+ )
> # Turn off the device.
> dev.off()
null device
          1
```

Figure 2 shows the scatter plot between `sbp` (systolic blood pressure) and `age` generated by the code above. We can observe a weak positive correlation between `sbp` and age. However, at a first glance, the relationship looks quite weak. We need to run more analysis such as Pearson correlation coefficient and linear regression to determine whether there is a (positive) relationship between the two variables.

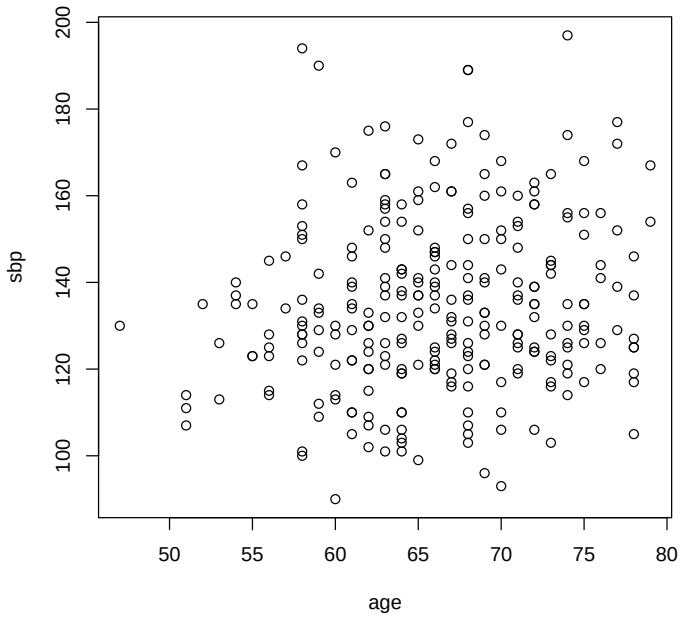

Figure 2: A scatter plot between `sbp` and `age` from a subset of `HERS` data.

b)

By using `lm()` we fitted a linear regression model to `HERS` data. As instructed in the problem text, we used `sbp` as outcome variable (y) and `age` as predictor (x).

```
> # Fit a linear model.
> lm.obj.1 = lm(sbp~age, data=HERS.data)
> # Show the result.
> summary(lm.obj.1)

Call:
lm(formula = sbp ~ age, data = HERS.data)

Residuals:
    Min      1Q  Median      3Q     Max
-43.550 -13.520  -2.431  12.578  62.736

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 105.7130    12.4024   8.524 1.05e-15 ***
age           0.4405     0.1865   2.363   0.0188 *
---
Signif. codes:

Residual standard error: 19.76 on 274 degrees of freedom
Multiple R-squared:  0.01997, Adjusted R-squared:  0.01639
F-statistic: 5.582 on 1 and 274 DF,  p-value: 0.01884

>
> # Extract estimated parameters
> beta.0.hat = lm.obj.1$coefficients[1]
> names(beta.0.hat) = NULL
> beta.1.hat = lm.obj.1$coefficients[2]
> names(beta.1.hat) = NULL
>
```

```
> # Create a scatter plot with fitted regression line
> # Open a pdf device.
> pdf("./Exercise_6_b.pdf", width = 6, height = 6)
> # Create a scatter plot
> plot(
+   x=HERS.data[,"age"],
+   HERS.data[,"sbp"],
+   xlab="age",
+   ylab="sbp"
+ )
> # Plot the fitted linear model.
> abline(a=beta.0.hat, b=beta.1.hat, col="blue")
> # Turn off the device.
> dev.off()
null device
          1
>
```

From the summary of the fitted linear model (`lm.obj.1`), we can read off the following estimated parameter values: $\widehat{\beta}_0 = 105.7130$, $\widehat{\beta}_1 = 0.4405$.
The fitted regression line is visualized with the original data points in Figure 3.

The slope ($\widehat{\beta}_1$) can be interpreted as "expected change in *systolic blood pressure* (`sbp`) when the age of a woman increases by 1 year". We also have to mention what the unit of `sbp` is, but that information is not given in the exercise text.

Further, we see, from the summary of the fitted linear model, that the p-value corresponding to $\widehat{\beta}_1$ is 0.0188. This p-value comes from the t-test with following null and alternative hypothesis:

$$H_0 : \beta_1 = 0, \quad H_A : \beta_1 \neq 0.$$

If we would use the significance level of $\alpha = 0.05$, we can reject H_0 and conclude that age has a significant (linear) effect on systolic blood pressure.

However, if we use the significance level of $\alpha = 0.01$, we have $0.0188 > 0.01$ and we cannot reject H_0. In this case, we conclude that age has no significant (linear) effect on systolic blood pressure.

The intercept ($\widehat{\beta}_0$) can be interpreted as "expected value of *systolic blood pressure* (y) when age (x) is set to 0".
However, this interpretation is not meaningful since age cannot be 0.

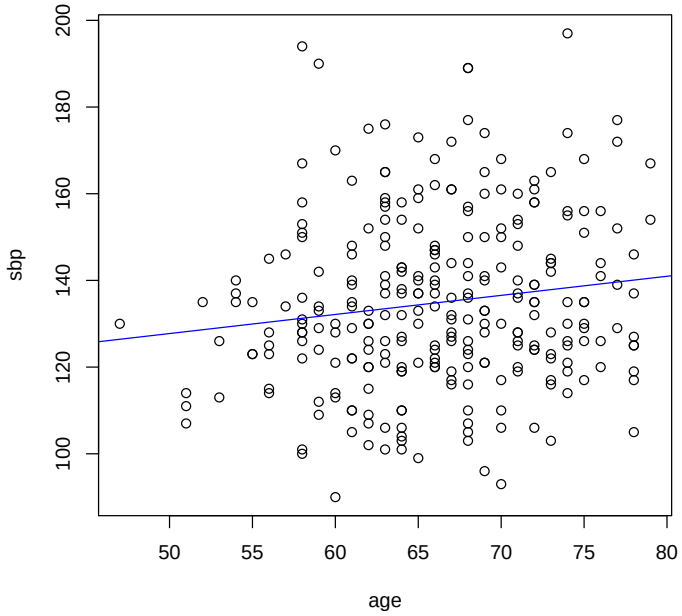

Figure 3: A scatter plot with fitted linear regression line from Problem 1 b).

c)

To get more meaningful interpretation of the intercept β_0, we subtract 67 from the predictor (age) and fit linear regression with the following code.

```
> # Fit a linear regression with "age - 67" as predictor.
> lm.obj.2 = lm(sbp~I(age-67), data=HERS.data)
> # Show the result.
> summary(lm.obj.2)

Call:
lm(formula = sbp ~ I(age - 67), data = HERS.data)

Residuals:
    Min      1Q  Median      3Q     Max
-43.550 -13.520  -2.431  12.578  62.736

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 135.2284     1.1985 112.829   <2e-16 ***
I(age - 67)   0.4405     0.1865   2.363   0.0188 *
---
Signif. codes:

Residual standard error: 19.76 on 274 degrees of freedom
Multiple R-squared:  0.01997, Adjusted R-squared:  0.01639
F-statistic: 5.582 on 1 and 274 DF,  p-value: 0.01884

>
> # Summary of the result.
> result.table = data.frame(
+   lm.1 = lm.obj.1$coefficients,
+   lm.2 = lm.obj.2$coefficients
+ )
> rownames(result.table) = c("beta.0.hat","beta.1.hat")
```

```
> result.table
                  lm.1         lm.2
beta.0.hat 105.7129691 135.2283886
beta.1.hat   0.4405286   0.4405286
```

From the output, we have following least squares estimates.

	not centered	centered
$\widehat{\beta}_0$	105.71	135.23
$\widehat{\beta}_1$	0.44	0.44

Tip: https://www.tablesgenerator.com/# is a handy website if you want to draw a table in LaTeX.

We can see that centering changed the intercept ($\widehat{\beta}_0$), but the slope ($\widehat{\beta}_1$) remains unchanged. This is logical since centering is nothing more than "shifting" all data points with the same amount.

The intercept ($\widehat{\beta}_0$) can now be interpreted as "expected value of *systolic blood pressure* (y) when centered age ($x - 67$) is set to 0".
In other words, the intercept can now be interpreted as "expected value of *systolic blood pressure* (y) when age (x) is set to 67".
This is a meaningful interpretation, without any biological nonsense like in a).

d)

We fit a new linear regression with $\frac{\text{age}}{10}$ as predictor by using the following code.

```
> # Fit a linear regression with age/10 as predictor.
> lm.obj.3 = lm(sbp~I(age/10), data=HERS.data)
> # Show the result.
> summary(lm.obj.3)

Call:
lm(formula = sbp ~ I(age/10), data = HERS.data)

Residuals:
    Min      1Q  Median      3Q     Max
-43.550 -13.520  -2.431  12.578  62.736

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  105.713     12.402   8.524 1.05e-15 ***
I(age/10)      4.405      1.865   2.363   0.0188 *
---
Signif. codes:

Residual standard error: 19.76 on 274 degrees of freedom
Multiple R-squared:  0.01997, Adjusted R-squared:  0.01639
F-statistic: 5.582 on 1 and 274 DF,  p-value: 0.01884

>
> # Summary of the result.
> result.table = cbind(result.table, lm.3 = lm.obj.3$coefficients)
> result.table
                  lm.1         lm.2       lm.3
beta.0.hat 105.7129691 135.2283886 105.712969
beta.1.hat   0.4405286   0.4405286   4.405286
>
```

From the output, we have following least squares estimates.

	age	age/10
$\widehat{\beta}_0$	105.71	105.71
$\widehat{\beta}_1$	0.44	4.41

We can see that intercept remains unchanged. This is because we only "scaled" the predictor (x) and did not "shift" it.
For the slope ($\widehat{\beta}_1$) , we can observe that it is increased by the factor of 10.

On slide 43 from Lecture 2. $\widehat{\beta}_1$ is defined as

$$\widehat{\beta}_1 = \frac{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{s_x^2}.$$

Dividing age by 10 is equal to: replacing x_i by $\dfrac{x_i}{10}$, replacing $\overline{x}$ by $\dfrac{\overline{x}}{10}$ and replacing s_x by $\dfrac{s_x}{10}$.
So, the slope becomes

$$\begin{aligned}
\frac{\frac{1}{n-1}\sum_{i=1}^{n}(\frac{1}{10}x_i - \frac{1}{10}\overline{x})(y_i - \overline{y})}{\left(\frac{1}{10}s_x\right)^2} &= \frac{\frac{1}{n-1}10^2\sum_{i=1}^{n}\frac{1}{10}(x_i - \overline{x})(y_i - \overline{y})}{s_x^2} \\
&= \frac{\frac{1}{n-1}10^2 \cdot \frac{1}{10}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{s_x^2} \\
&= 10 \cdot \frac{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{s_x^2} \\
&= 10\,\widehat{\beta}_1.
\end{aligned}$$

Hence, we can see that dividing the predictor (x) by 10 results in that the slope increases by 10.

THE END