Practice Session 3

The Relationship Between Two Quantitative Variables: Correlation and Regression

In this session you might use the formula of the correlation between two quantitative variables:

$r_{xy} = \frac{1}{(n-1)s_z s_y} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$

Remember that the fitted regression line is defined by the equation:

- $\hat{y} = a + bx$, or
- $\widehat{\text{Response}} = a + b \cdot (\text{Explanatory})$
- Residuals = observed - predicted = y - $\hat{y}$

Where:

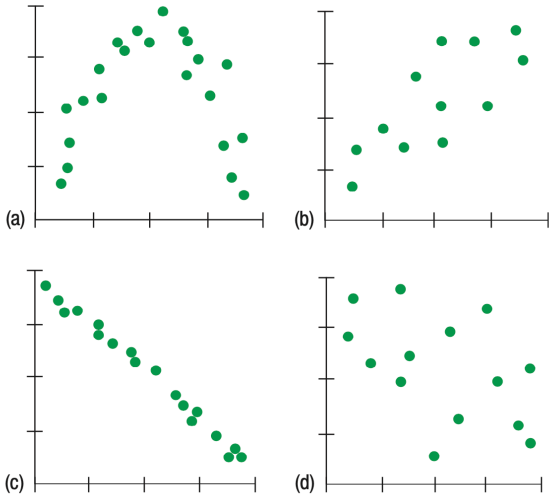
- Response: is the response variable or the dependent variable
- Explanatory: is the independent variable
- a: is the y-intercept
- b: is the slope of the regression line

You may use the following R functions: `plot()`, `lm()`, `cor()`, `abline()`. And you might need to download Lock5Data using `library(Lock5Data)`.

Part 1: Two quantitative variables/ Scatterplot and Correlation

Practice 1:

Here are several scatterplots. The calculated correlations are 0.006, - 0.977, - 0.487, and 0.777. Match each scatter plot with the appropriate correlation coeffecient.



Answers:

a.
b.
c.
d.

Practice 2:

Load the data `FloridaLakes` from library(Lock5Data).

1. Describe the type of each of the variables: `pH`, `Calcium`, and `Alkalinity`.
2. Create a three scatter plots for each pair of variables: pHvsCalcium, `pH` vs `Alkalinity`, and `Calcium` vs `Alkalinity` . Add the main title to each plot.
3. What is the correlation coefficient between `pH` and `Calcium`? Is it positive or negative?
4. What do these coefficients mean in the context of this data?
5. Try to calculate the correlation coefficient between `pH` and `Calcium` without using the `R` function for correlation.

Answers:

```
# download the data and load it into R
library(Lock5Data)
data(FloridaLakes)
```

1.
2.
3.
4.
5.

Part 2: Two quantitative variables/ Linear Regression

Practice 3:

State if the following sentences are true or false:

a. We choose the linear model that passes through the most data points on a scatter plot.

b. The residuals are observed y-values minus the y-value predicted by a linear model.

c. Least square means that the square of the largest residuals is as small as it could possibly be.

d. Some of the residuals from least linear model will be positive, and some will be negative.

e. Least squares means that some of the squares of the residuals are minimized.

f. We write $\hat{y}$ to denote the predicted value, and y to denote the observed value.

Answers:

a.
b.
c.
d.
e.
f.

Practice 4:

Use the previous data `FloridaLakes` and the two variables `Alkalinity` and `calcium`.

1. Using the appropriate R function, create a linear model to predict `Alkalinity` from `calcium`.
2. Find the coefficients of the regression and write the correct equation of this linear model.
3. Interpret the intercept and slope of the model within the context of the data.
4. Predict the `Alkalinity` level when Calcium= 2.5.
5. The actual `Alkalinity` was 8.5 when Calcium= 2.5. Find the residual for this data point.
6. Find the five number summary for the variable `Calcium`.
7. Predict the `Alkalinity` level when Calcium = 0.5
8. Do both of these predictions make sense given the data?

Answers:

```
# download the data and load it into R
library(Lock5Data)
data(FloridaLakes)
```

1.
2.
3.
4.
5.
6.
7.
8.

Part 3: Review

Practice 5:

From the data `ICUAdmissions` create descriptive statistics and visualizations for the variables `Infection`, `Age`, and `Race`, `HeartRate`.

Answers:

```
# download the data and load it into R
library(Lock5Data)
data(ICUAdmissions)
```