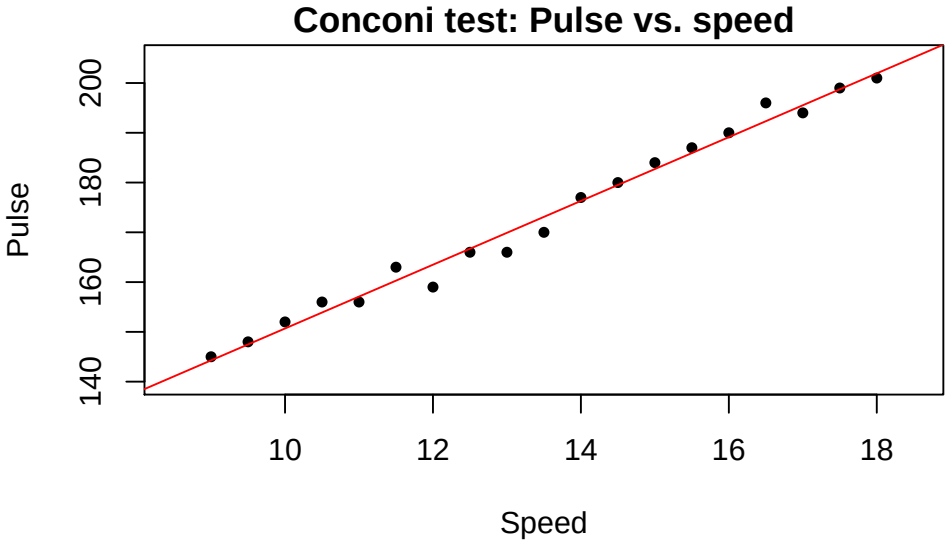
# Series 3

**1.** The *Conconi* test measures the endurance performance of a person. It takes place on the 400m-track where one starts running slowly (9 km/h). Every 200 meters the speed is increased by 0.5 km/h. At the end of every 200m section the pulse is measured. The test continues until the speed can no longer be increased. The lecturer did this test in summer 2012. The data is contained in the file *conconi.rda*. The scatter plot with the OLS regression line looks as follows:




**a)** Visualize the data in a scatter plot as shown above. Fit the regression line with the command `lm()` and add it to the plot. Create the summary output with R and answer the following questions:

- To what extent can we explain the scatter in the pulse by the increase in speed?
- By what amount does the pulse increase on average when the speed is increased by 1 km/h? What other values are also plausible?
- How large is the resting heart rate (i.e. when there is no movement)? In what interval do you expect this value to be? Does it seem plausible?

**b)** We now consider the values from Dani who also took the *Conconi* test:

```
> summary(fit.dani)
 Coefficients:   Estimate   Std. Error   t value   Pr(>|t|)
 (Intercept)     xxx        xxx          xxx       xxx        xxx
 Speed           4.09323    0.09972      41.05     <2e-16     ***
```

Whose pulse is increasing more slowly when the speed is increased? Can we say whether there is a significant difference between the two increases of the pulse? Can we deduce who of the two is in better shape?

**c)** Plot the residuals against the predictor as well as the normal plot of the residuals. Decide which of the following four assumptions are fulfilled:

- The regression line captures the relation correctly, i.e. $\mathcal{E}(E_i) = 0$.
- The variance of the error is constant, i.e. $\text{Var}(E_i) = \sigma_E^2$.
- The errors follow a Normal distribution, i.e. $\mathcal{N}(0, \sigma_E^2)$.
- The errors are uncorrelated, i.e. $\text{Corr}(E_i, E_j) = 0 \quad \forall i \neq j$.

How do we detect potential violations of these assumptions and how could these be explained? What statements made in the previous subproblems are still valid, which ones aren't?

**2.** While fitting and visualizing simple linear regression models as well as conducting the corresponding tests becomes a routine after a while, assessing whether a model fits remains a challenging task. We will practice this with two additional data sets:

**a)** The file `gas.rda` contains the gas consumption (in kWh) and the differences of temperature (in ℃) inside and outside of 15 houses which are heated with gas. The measurements were collected over a long time span and then averaged. The goal is to explain the gas consumption with the temperature difference. Plot the regression line and perform a residual analysis.

**b)** The file `antikeUhren.rda` contains the age and the price of antique clocks that are auctioned.The goal is to predict the price with the age of the clock. Plot the regression line and perform a residual analysis.

**3.** Which of the following statements are false and why?

**a)** When $R^2 > 0.8$, a residual analysis is not important as the model already fits well.

**b)** When the $p$-value for the null hypothesis $H_0 : \beta_1 = 0$ is smaller than 0.001, then there is a causal relation between the predictor and the response.

**c)** When the residuals do not follow a Normal distribution exactly but show a symmetric, moderately long-tailed distribution instead, then the model still produces useful predictions.

**d)** When we see a deviation of the smoother from the x-axis in the "residuals vs. predictor" plot while the variance seems constant and the Normal plot does not show any abnormalities, then the model and its results can still be used with some care.

**e)** The residual standard error indicates to what extent the observations vary around the regression line. If the model assumptions are fulfilled, then approximately 95% of the data points lie in an interval of $\pm 2\sigma$ around the regression line.

**4.** The article "Characterization of Highway Runoff in Austin, Texas, Area" gave a scatter plot of x=rainfall volume and y=runoff volume for a particular location. The values are:

| x | 5 | 12 | 14 | 17 | 23 | 30 | 40 | 47 | 55 | 67 | 72 | 81 | 96 | 112 | 127 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| y | 4 | 10 | 13 | 15 | 15 | 25 | 27 | 46 | 38 | 46 | 53 | 70 | 82 | 99 | 100 |

**a)** Produce a scatterplot of runoff volume vs. rainfall volume. Fit a simple linear regression, add the regression line to the plot and generate the summary output.

**b)** How much of the observed variation in runoff volume can be attributed to the simple linear association between runoff and rainfall volume?

**c)** Is there a significant linear association between runoff and rainfall volume? Provide an illustrative interpretation of the regression coefficient.

**d)** Use the regression fit for predicting the runoff volume when the rainfall volume takes the value 50. Also compute the 95% prediction interval for this case.

**e)** Check the model assumptions with the model diagnostics tools you have seen so far.

**f)** Assess whether a log-transformation is appropriate by plotting the corresponding histograms.

**g)** Fit a new regression model with the log-transformed variables and add it to the scatter plot.

**h)** Assess the strength and the significance of a linear relation between runoff and rainfall. Interpret the regression coefficient.

**i)** Predict the expected runoff when rainfall takes a value of 50. Create a 95% prediction interval and compare it to the original solution. Add the prediction interval for arbitray $x$ to the scatter plot.

**j)** Perform a residual analysis. Which model is more appropriate and why?

**5.** In an experiment marine bacteria were exposed to x-rays during 15 intervals of six minutes. The following table contains the amount of bacteria *after* each interval

| Interval | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Amount | 255 | 211 | 197 | 166 | NA | 106 | 104 | 60 | 56 | 38 | 36 | 32 | 21 | 19 | 15 |

**a)** Show the relation between the number of surviving bacteria and the number of radiation intervals. Does it make sense to fit a OLS regression to the data?

**b)** Fit a simple linear regression model and check the model assumptions.

**c)** Improve the model by transforming the target variable or/and the predictor. Hint: The theory suggests that per radiation interval the proportion of bacteria that is killed remains constant.

**d)** Predict the missing value for the fifth interval and compute a 95% prediction interval. In addition, compute the estimate for the relative decrease in the number of surviving bacteria, together with a 95% confidence interval. Lastly, provide the expected number of bacteria at the beginning, i.e. before the first radiation interval. Also compute a 95% confidence interval for this value.

**Preliminary discussion:** Monday, October 5.

**Deadline:** Monday, October 19.