Prof. K. Nieselt, Faculty of Science

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

Lecture: Microarrays Bioinformatics — WS 2017/18

Assignment No. 5 (5 points)

Hand out:	Thursday, November 30
Hand in due:	Thursday, December 7, 10:00
Tutorial date:	Tuesday, December 5, 10:15-11:45
Direct inquiries to:	fabian.moertter@student.uni-tuebingen.de

Theoretical Assignments

1. Linear Regression (5p)

(a) Prove: After normalization using linear regression, the mean expression value of the genes in the red channel is equal to the mean expression value of the green channel.

(b) Given the following expression values:

Gene ID	Red	Green
1	9.5	8.5
2	8.5	9.5
3	8.0	7.0
4	7.0	6.5
5	6.0	7.0

i. Compute the linear regression by hand. Show all steps.

ii. The quantity

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

(a.k.a. Person correlation) measures the strength of linear dependency, its square, r^2 is called *coefficient of determination*. It measures the ratio of explained variance to total variance. Compute the coefficient of determination for the example.

iii. The variables $\epsilon_i = y_i - \hat{y}_i$ are called the *residuals* or *prediction errors*. Compute the residuals for the example.

(c) Bonus (+2.5P): Prove that

$$b = \frac{\sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

and

$$a = \bar{y} - b\bar{x}$$

are the solutions that minimize the error in the least squares equation to compute an estimate for the slope and intercept of

$$f(x) = \alpha + \beta x,$$

i.e. minimize $\sum(y_i - f(x_i))^2$.

Hint: Use the result of 1a) as well as the binomial and the Steiner Theorem (dt. Verschiebungssatz).

Practical Assignments

2. Linear Regression

In the last assignment we visualized raw background subtracted expression values. Now we want to normalize each pair of arrays by linear regression. Write a method to calculate the linear regression, do not use an R package/function! Apply your method to the same data set (`affy data.tsv`) as last week.

(a) Normalize the data using linear regression of the background-corrected expression values:

i. Use the data of Assignment 4 - Task 3(b). Implement a function `linearRegression` that performs a linear regression of the arrays (two columns of the data) as discussed in the lecture.

ii. Produce a scatterplot of the data and **add the regression line**.

iii. Use the derived linear regression to normalize the data.

iv. Produce a scatterplot of both arrays of the normalized data (also showing the regression line for the normalized data).

(b) Now repeat all steps, but this time compute the linear regression of the MA values:

i. Based on the calculated M and A values of Assignment 4 - Task 3(d), produce an MA-plot and **add the regression line**.

ii. Perform a linear regression of the MA values.

iii. Produce MA-plots of the normalized data (also showing the regression line for the normalized (M, A)-data).

(c) Repeat all steps from (a) and (b) for all pairs of arrays (these are 6 pairs).

i. Discuss your results: Compare the results of the linear regressions. What did you notice? Was linear regression appropriate for the expression values and/or respective (M, A) values?

Please read the questions carefully. If there are any questions, you may ask them during the tutorial session or via e-mail to your tutor. You will usually get an answer in time, but late e-mails (e.g. on Thursday morning before class) might not be answered in time. Please upload your solutions in the Ilias system. Please pack your source code, the plots, as well as the theoretical part into **one single archive file (zip)**. Source code should compile correctly.