Introduction to Statistical Machine Learning for Bioinformaticians and Medical Informaticians

SoSe 2022

Tutor: Dana Petracek

1	2	$\sum$

Marina Dittschar & Clarissa Auckenthaler

Assignment 2

(Abgabe am 16.05.2022)

Task 1

Given:

Training data:

X	Y
0	0
9	8
3	4
6	4
2	2

Test data:

X	Y
7	7
4	3
10	7

a) Estimate the parameters of a simple linear regression model using the training data.

We used the following variables to describe the linear regression model:

$$\hat{Y} = b_0 + b_1 X + \epsilon \tag{1}$$

With data X, predicted values $\hat{Y}$, intercept b_0, slope b_1 and error ϵ.

We used the steps outlined in [2].

Step 1: Calculate $X * Y$, X^2, Y^2, $\sum X$, $\sum Y$, $\sum X * Y$, $\sum X^2$, $\sum Y^2$

X	Y	$X * Y$	X^2	Y^2
0	0	0	0	0
9	8	72	81	64
3	4	12	9	16
6	4	24	36	16
2	2	4	4	4
$\sum$	20	18	112	130	100

Step 2: Calculate b_0

$$b_0 = \frac{((\sum Y) * (\sum X^2) - (\sum X) * (\sum X * Y))}{n(\sum X^2) - (\sum X)^2}$$
$$b_0 = \frac{(18 * 130 - 20 * 112)}{5 * 130 - (20)^2} \tag{2}$$
$$b_0 = 0,4$$

Step 3: Calculate b_1

$$b_1 = \frac{(n(\sum X * Y) - (\sum X) * (\sum Y))}{n(\sum X^2) - (\sum X)^2}$$
$$b_1 = \frac{(5 * 112 - 20 * 18)}{5 * 130 - (20)^2} \tag{3}$$
$$b_1 = 0,8$$

Solution:

$$\hat{Y} = 0,4 + 0,8X \tag{4}$$

We obtain a linear regression formula with an intercept of 0.4 and a slope of 0.8. This means that the slope is positive and the regression line meets the y-axis just above zero.

b) Visualize the data set in a plot. Use different colors for training and test data points. Draw the regression line into the plot. Briefly state its interpretation. Highlight the residuals using dashed lines.

In Figure 1 we can see the resulting regression line has a positive slope and an intercept just above zero. The model is derived using the train data, but inserting the test data also seems to confirm the validity of the regression line.

c) Assess the fit of the model to the training data. Use a performance measure that does not depend on the range of Y.

We used R^2 as a performance measure because it is normalized, contrary to the residual standard error. Thus, it does not depend on the range of Y. We calculated the R^2 with the formulas in [3].

$$R^2 = (\frac{(n * (\sum X * Y) - ((\sum X) * (\sum Y))}{\sqrt{n(\sum X^2) - (\sum x)^2} * \sqrt{n * (\sum Y^2) - (\sum Y)^2}})^2$$
$$R^2 = (\frac{5 * 112 - 20 * 18}{\sqrt{5 * 130 - 20^2} * \sqrt{5 * 100 - 18^2}})^2 \tag{5}$$
$$R^2 = 0.909$$

The R^2 for this model is 0.909. This means that $\sim$ 91% of the variation in the variable y can be explained by variable x.

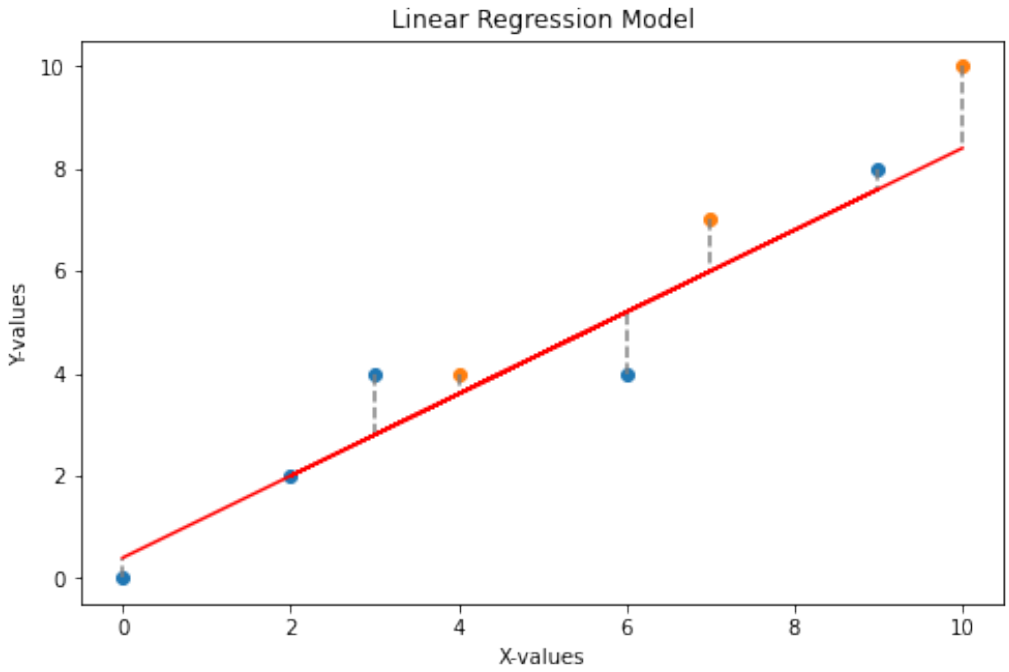

Figure 1: Plot of the training data (blue), test data (orange), the linear model (red) and the residuals (grey).

d) Compute and compare the training and test (mean square) error. What do you observe?

The formula for the mean squared error (MSE) is as follows:

$$MSE = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n} \tag{6}$$

With true single data points y_i, predicted single data points $\hat{y}_i$ and number of data points n.

We work with the following set of true values Y and predicted values $\hat{Y}$ for the training and test data sets:

Training data:	
Y	$\hat{Y}$
0	0.4
8	7.6
4	2.8
4	5.2
2	2.0

Test data	
Y	$\hat{Y}$
7	6.0
3	3.6
7	8.4

We insert the training and test data into the equation for the MSE:

– Training data MSE:

$$MSE = \frac{(0-0.4)^2 + (8-7.6)^2 + (4-2.8)^2 + (4-5.2)^2 + (2-2)^2}{5}$$
$$MSE = \frac{3.2}{5} \tag{7}$$
$$MSE = 0.64$$

– Test data MSE:

$$MSE = \frac{(7-6)^2 + (3-3.6)^2 + (7-8.4)^2}{5}$$
$$MSE = \frac{3.32}{3} \tag{8}$$
$$MSE = 1.1067$$

The resulting MSE is 0.64 for the training data and 1.1067 for the test data. We can observe the MSE of the training data is smaller than the MSE of the test data. That means our model fits better on the training data than to the test data. This is because the model was fit using the training data. The sample size for both (training and test data) is very low, so it is hard to make accurate predictions.

Task 2

a) Show that for simple linear regression $R^2 = cor(x,y)^2$ holds.

The formula for the squared correlation between x and y is:

$$cor(x,y)^2 = \frac{E[(x-\bar{x})(y-\bar{y})]^2}{\sqrt{var(x)var(y)}^2} = \frac{cov(x,y)^2}{var(x)var(y)} \tag{9}$$

For simple linear regression, the following formula is used [4]:

$$y_i = \alpha + \beta x_i + \widehat{\epsilon}_i \tag{10}$$

We want to minimize the sum of residuals, with the ordinary least squares approach. Therefore, we need to find $\widehat{\alpha}$ and $\widehat{\beta}$ such that $\sum_{i=1}^{n} \widehat{\epsilon}_i^2$ is minimal [1] [4]:

$$\widehat{\alpha} = \bar{y} - (\widehat{\beta}\bar{x}) \tag{11}$$

$$\widehat{\beta} = \sum_{i=1}^{n} \frac{(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{cov(x,y)}{var(x)} \tag{12}$$

If you insert $\widehat{\alpha}$ into the formula for linear regression [4]:

$$
\begin{aligned}
\widehat{y}_i &= \widehat{\alpha} + \widehat{\beta} x_i \\
\widehat{y}_i &= \bar{y} - (\widehat{\beta}\bar{x}) + \widehat{\beta} x_i \\
\widehat{y}_i &= \bar{y} + \widehat{\beta}(x_i - \bar{x}) \\
\widehat{y}_i - \bar{y} &= \widehat{\beta}(x_i - \bar{x}) \\
\widehat{\beta} &= \frac{\widehat{y}_i - \bar{y}}{x_i - \bar{x}}
\end{aligned}
\tag{13}
$$

Inserting equation 12 into equation 9 brings:

$$
\begin{aligned}
cor(x,y)^2 &= \frac{cov(x,y)^2}{var(x)var(y)} \\
&= \widehat{\beta}^2 \frac{var(x)}{var(y)} \\
&= \sum_{i=1}^{n} (\frac{\widehat{y} - \bar{y}}{x - \bar{x}})^2 \frac{var(x)}{var(y)} \\
&= \frac{\sum_{i=1}^{n} (\widehat{y} - \bar{y})^2}{\sum_{i=1}^{n} (x - \bar{x})^2} \frac{var(x)}{var(y)} \\
&= \frac{(\widehat{y} - \bar{y})^2}{var(y)} \\
&= \frac{\sum_{i=1}^{n} (\widehat{y} - \bar{y})^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2}
\end{aligned}
\tag{14}
$$

And comparing this with the formulas for the total sum off squares and the explained sum of squares:

$$
\begin{aligned}
&\frac{\sum_{i=1}^{n} (\widehat{y} - \bar{y})^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2} \\
&= \frac{ESS}{TSS} \\
&= R^2
\end{aligned}
\tag{15}
$$

Thus we see that $R^2 = cor(x, y)$.

b) Show that $R^2 = Cor(Y, \hat{Y})^2$ holds for simple linear regression as well.

To show that $R^2 = cor(y, \hat{y})^2$, we use the fact that we have already shown that $R^2 = cor(x, y)$.
We only need to show that $cor(x, y)^2 = cor(y, \widehat{y})^2$.

To do this, we again use the relationship

$$
\widehat{\beta} = \frac{cov(x,y)}{var(x)} \tag{16}
$$

$$
cov(x,y) = \widehat{\beta} \cdot var(x) \tag{17}
$$

as well as

$$\widehat{\beta}^2 = \sum_{i=1}^{n} \frac{(\widehat{y}_i - \bar{y})^2}{(x_i - \bar{x})^2} \tag{18}$$

and

$$cor(x,y)^2 = \frac{cov(x,y)^2}{var(x)var(y)} \tag{19}$$

We insert $\widehat{\beta}^2$ into equation 17 and the resulting covariance term in equation 19:

$$\begin{aligned} cor(x,y)^2 &= \frac{\widehat{\beta}^2 \cdot var(x)^2}{var(x)var(y)} \\ cor(x,y)^2 &= \frac{\sum_{i=1}^{n}(\widehat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \cdot \frac{(\sum_{i=1}^{n}(x_i - \bar{x})^2)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2 \cdot \sum_{i=1}^{n}(x_i - \bar{x})^2} \\ cor(x,y)^2 &= \frac{\sum_{i=1}^{n}(\widehat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \\ cor(x,y)^2 &= cor(\widehat{y},y)^2 \end{aligned} \tag{20}$$

We have shown that the squared correlation of x and y equals the squared correlation of $\hat{y}$ and y and therefore, $R^2 = cor(\hat{y},y)^2$.

References

[1] John F Kenney. Mathematics of statistics. D. Van Nostrand, 1939.

[2] Zach (Statology). How to perform linear regression by hand. https://www.statology.org/linear-regression-by-hand/?msclkid=f99f2c98d0f111eca2c1a374c4f6424f, 2020.

[3] Zach (Statology). How to calculate r-squared by hand. https://www.statology.org/calculate-r-squared-by-hand/#:~:text=R2%20%3D%20%5B%20%28n%CE%A3xy%20%E2%80%93%20%28%CE%A3x%29%20%28%CE%A3y%29%29%20%2F,given%20regression%20model.%20Step%201%3A%20Create%20a%20Dataset?msclkid=b14ed220d11d11ecb7cd3e4ed90cd5a7, 2021.

[4] Wikipedia. Simple linear regression. https://en.wikipedia.org/wiki/Simple_linear_regression, 2022.