Linear Regression

Motivating Examples

Usually, regression models are used to determine the unknown relationship between one or multiple (independent) input variable(s) and a (dependent) target variable, the latter being influenced by the former. If a regression model can identify the dependency between these variables, then it can easily predict the value of the target variable for a new set of input values. The following examples illustrate some applications to give a better understanding of the relevance and functionalities of regression models.

Example 1

Determining the linear relationship between two variables (Gravity constant of the earth).

Let's think back to physics class in school. There, we learned that the gravitation of the earth accelerates the speed of every falling object by a constant g of about g =9:801m/s^2 near the equator. This means that a falling object freely increases its velocity by 9.801 m/s^2 per second it falls.

School was a long time ago for most of us; however, and so now, we check this basic constant again, with the help of regression analysis. According to physics, there is a linear relationship, h =gxS, where h is the height from which an object is falling, and S is the squared time in seconds that the object needs to reach the ground.

In this example, there is only a single input variable, and so these kinds of models are called univariate or simple; but regression models can also be used to estimate more complex linear variable connections. In these cases, the models are called multiple linear regression models.

Fig.1 Exemplary data from an experiment to determine earth's gravity constant

	height (m)	seconds squared
1	100	8.967
2	120	11.989
3	150	15.852
4	180	16.991
5	200	20.284
6	250	24.228
7	300	30.580

Example 2

Determination of variable relations and prediction of variables (Analysis of pretest and final exam results).

Students often prepare for exams by taking tests in advance. This gives students the opportunity to become familiar with the type and complexity of questions asked, as well as a chance to check their degree of readiness for the final exam.

To find out if pretesting is helpful, we can inspect the relationship between the performances in both exams. As well as measuring with an appropriate correlation coefficient, we can model the exact relationship using linear regression.

In addition, we are interested in a prediction of the final exam scores from future students, based on their pretest scores. This can be done by applying the build regression model on this new dataset of pretest results. We provides a dataset called "test_scores.sav" , which comprises data for the kind of analysis described here.

	school	school_setting	school_type	classroom	teaching_method	n_student	student_id	gender	lunch	pretest	posttest
1	ANKYI	1.000	2.000	6OL	0.000	20.000	2FHT3	1.000	2.000	62.000	72.000
2	ANKYI	1.000	2.000	6OL	0.000	20.000	3JIVH	1.000	2.000	66.000	79.000
3	ANKYI	1.000	2.000	6OL	0.000	20.000	3XOWE	0.000	2.000	64.000	76.000
4	ANKYI	1.000	2.000	6OL	0.000	20.000	556O0	1.000	2.000	61.000	77.000
5	ANKYI	1.000	2.000	6OL	0.000	20.000	74LOE	0.000	2.000	64.000	76.000
6	ANKYI	1.000	2.000	6OL	0.000	20.000	7YZO8	1.000	2.000	66.000	74.000
7	ANKYI	1.000	2.000	6OL	0.000	20.000	9KMZD	0.000	2.000	63.000	75.000
8	ANKYI	1.000	2.000	6OL	0.000	20.000	9USQK	1.000	2.000	63.000	72.000
9	ANKYI	1.000	2.000	6OL	0.000	20.000	CS5QP	0.000	2.000	64.000	77.000
10	ANKYI	1.000	2.000	6OL	0.000	20.000	D6HT8	1.000	2.000	61.000	72.000

Fig.2 test_scores.sav

Field name	Description
school	Name of the school
school_setting	School setting: 1 = Urban, 2 = Suburban, 3 = Rural
school_type	School type 1 = Public, 2 = Nonpublic
classroom	Classroom number
teaching_method	Teaching method 0 = Standard, 1 = Experimental
n_student	Number of students in the classroom
student_id	Student ID
gender	Gender of the student 0 = Male, 1 = Female
lunch	Reduced/Free lunch 1 = Qualifies for reduced/free lunch 2 = Does not qualify
pretest	Result of Pretest
Posttest	Result of Posttest

Exercise 3: Linear Regression with the Regression Node In this exercise, we build a regression model for the Boston housing data, housing.data.txt , with the Regression node.

1. Import the data and specify the variable types with the Type node.

2. Add a Regression node to the stream and select MEDV as the target variable and all other variables as the input.

3. Choose the Backwards method to find the significant input variables and then run the stream.

4. Inspect the model nugget and identify the estimated coefficients and the regression equation. Which variables are included in the final model, and which variable has a coefficient of 3.832?

5. What is the value of R^2 and the adjusted R^2 ?

The following exercise is optional and includes adding a cross-validation to the stream.

6. Include a Partition node in the stream and divide the dataset into 70 % training data and 30 % test data.

7. Select the partition field in the Fields tab of the Regression node, setting it to use only the training data in the model building procedure.

8. Add an Analysis node to the model nugget and run the stream again. Is the model suitable for processing unknown data?

Name of variable	Description
CIRM	Per capita crime rate by town
ZN	Proportion of residential land zoned for lots over 25,000 sq.ft.
INDUS	Proportion of non-retail business acres per town
CHAS	Charles River dummy variable (tract bounds river = 1; otherwise = 0)
NOX	Nitric oxides concentration (parts per 10 million)
RM	average number of rooms per dwelling
AGE	Proportion of owner-occupied units built prior to 1940
DIS	Weighted distances to five Boston employment centers
RAD	Index of accessibility to radial highways
TAX	Full-value property-tax rate per $10,000
PTRATIO	Pupil–teacher ratio by town
B	1000(Bk – 0.63)^2 where Bk is the proportion of African/Americans by town
LSTAT	percentage of population with low status
MEDV	Median value of owner-occupied homes in $1000's