

Linear regression with one variable.

Univariate linear regression

To establish notation for future use, we’ll use *x*(*i*) to denote the “input” variables (living area in this example), also called input features, and *y*(*i*) to denote the “output” or target variable that we are trying to predict (price). A pair (*x*(*i*),*y*(*i*)) is called a training example, and the dataset that we’ll be using to learn—a list of m training examples (*x*(*i*),*y*(*i*));*i*=1,...,*m*—is called a training set. Note that the superscript “(i)” in the notation is simply an index into the training set, and has nothing to do with exponentiation. We will also use X to denote the space of input values, and Y to denote the space of output values. In this example, X = Y = ℝ.

To describe the supervised learning problem slightly more formally, our goal is, given a training set, to learn a function h : X → Y so that h(x) is a “good” predictor for the corresponding value of y. For historical reasons, this function h is called a hypothesis. Seen pictorially, the process is therefore like this:



When the target variable that we’re trying to predict is continuous, such as in our housing example, we call the learning problem a regression problem. When y can take on only a small number of discrete values (such as if, given the living area, we wanted to predict if a dwelling is a house or an apartment, say), we call it a classification problem.

Df: 3368

X: 3334 (因为舍弃了最后的34个)

Y: 3334

X\_train: 2667

X\_test: 667

y\_train: 2667

y\_test: 667

一共有3368个数据，然后先拿出来3334个作为已有的数据集来训练模型测试模型

剩下的这34个其实是当做新来的数据来预测的

Lastday 2017-11-14

（去掉了学习动力和学习成绩之间的关系，变为测试两个自变量都和因变量之间存在相关性）

本课题采用线性回归(Linear Regression) 的方法得出学习动机、学业成绩是如何影响职业规划的。应用分析出的模型，预测未来学生对职业规划，并提早矫正学生已到达让更多的学生有明确的职业规划的目标。

线性回归的目的是选取自变量和目标变量，分析自变量如何影响目标变量，给出合理的拟合关系式。



a. 目标变量

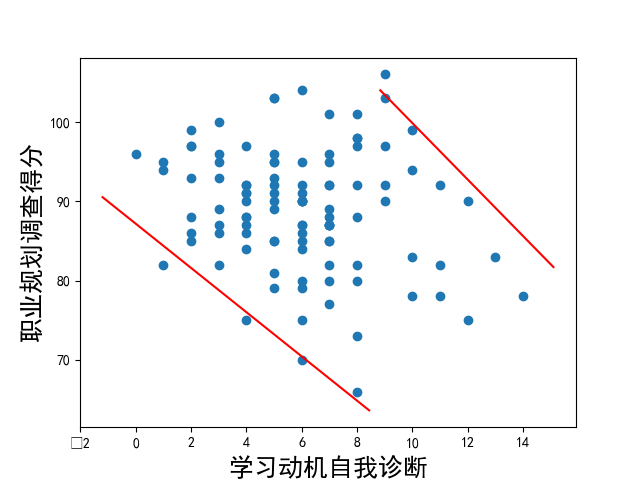
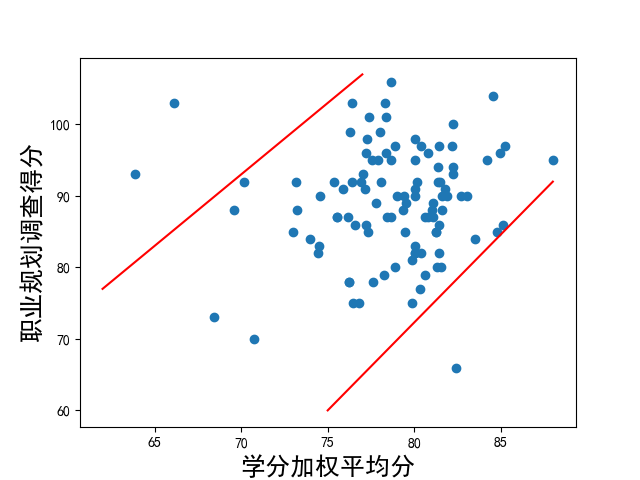
职业规划：职业规划变量可以用被采访者所占有的属性个数来量化，比如被采访者拥有主动性，现实性，自信心则用3表示。用y表示。

b. 自变量选取

学习成绩采用该受访者到目前为止的加权平均成绩表示，是0到100之间的一个整数，如89.5，用x\_1表示；根据研究方法的调查问卷结果，学习动机变量可以用1-20的量化数字表示，分值越大表明动机越弱，用x\_2表示。

c. 目标变量和自变量之间的相关性

首先，我们画出这两组变量之间的分布图来直观的理解两个自变量和目标变量之间的关系。



另一方面，我们采用皮尔森相关系数来计算，得出corrcoef(x\_1, y) = 0.0877293

corrcoef(x\_2, y) = -0.19287287

从计算可以看出是具有明显的相关性的。

c. 线性回归

通过我们分别计算两个因变量和目标变量之间的皮尔逊相关系数，我们可以发现，两个变量和学生的职业规划都是相关的。为了进一步确定这两个变量是如何影响职业规划的，我们假设其服从线性关系，即. 我们采用最小二乘法进行拟合，得出a b c三个参数的拟合值。

通过计算我们得到：*a= 0.043 ，b=-0.542，c=88.851*

d. 模型准确性

为了检测我们得出的的有效性，我们把这个数据集分为训练集合和测试集。在此项目里，我们随机选取百分之八十的数据作为训练集，在此进行线性回归分析，训练模型。在剩余的百分之二十上进行测试，这里我们引入误差error来衡量模型的准确性。计算出模型预测的职业规划值和真实职业规划之间的误差来描述模型的准确性。进而分析我们提出的方案的有效性。

其中表示测试集中的样本个数，表示样本的真实结果，表示通过我们的模型的预测结果。计算结果显示表明我们的模型的预测结果和真实值的误差只有0.09. 可以很好的预测学生职业规划。

e. 预测

通过模型的准确性分析，表明我们的模型在预测学生职业规划的有效性。在未来，假设我们已知学生的学习成绩为75.89，学生的学习动机为4， 那么我们可以预测出其职业规划的数值为：

0.043 \* 75.89 + 0.542 \* 4 + 88.851 = 94.28. 此学生具有良好的职业规划，不需要进行校正。