Оглавление

Теоретическое введение

Регрессионный анализ— метод моделирования измеряемых данных и исследования их свойств. Данные состоят из пар значений **зависимой переменной** (переменной отклика) и **независимой переменной** (объясняющей переменной).

Регрессионная модель есть функция независимой переменной и параметров с добавленной случайной переменной. Параметры модели настраиваются таким образом, что модель наилучшим образом приближает данные.

Критерием качества приближения (целевой функцией) обычно является среднеквадратичная ошибка: сумма квадратов разности значений модели и зависимой переменной для всех значений независимой переменной в качестве аргумента.

Регрессионный анализ — раздел математической статистики и машинного обучения. Предполагается, что зависимая переменная есть сумма значений некоторой модели и случайной величины. Относительно характера распределения этой величины делаются предположения, называемые гипотезой порождения данных. Для подтверждения или опровержения этой гипотезы выполняются статистические тесты, называемые анализом остатков. При этом предполагается, что независимая переменная не содержит ошибок. Регрессионный анализ используется для прогноза, анализа временных рядов, тестирования гипотез и выявления скрытых взаимосвязей в данных.

Регрессия — зависимость математического ожидания (например, среднего значения) случайной величины от одной или нескольких других случайных величин (свободных переменных), то есть $E(y|x) = f(x)$.

Регрессионным анализом называется поиск такой функции f, которая описывает эту зависимость. Регрессия может быть представлена в виде суммы неслучайной и случайной составляющих.

$$y = f(x) + \nu,$$

где f — функция регрессионной зависимости, а ν — аддитивная случайная величина с нулевым матожиданием. Предположение о характере распределения этой величины называется гипотезой порождения данных. Обычно предполагается, что величина ν имеет гауссово распределение с нулевым средним и дисперсией σ_ν^2.

Задача нахождения регрессионной модели нескольких свободных переменных ставится следующим образом. Задана выборка — множество $\{x_1,\dots,x_N | x \in \mathbb{R}^M\}$ значений свободных переменных и множество $\{y_1,\dots,y_N | y \in \mathbb{R}\}$ соответствующих им значений зависимой переменной. Эти множества обозначаются как D, множество исходных данных $\{(x,y)_i\}$. Задана регрессионная модель — параметрическое семейство

функций $f(w,x)$ зависящая от параметров $w \in \mathbb{R}$ и свободных переменных x. Требуется найти наиболее вероятные параметры $\overline{w}$:

$$\overline{w} = \arg\max_{w \in \mathbb{R}^W} p(y|x,w,f) = p(D|w,f).$$

Функция вероятности p зависит от гипотезы порождения данных и задается Байесовским выводом или методом наибольшего правдоподобия.

Линейная регрессия

Introduction to Linear Regression

Author(s)

David M. Lane

Prerequisites

Measures of Variability, Describing Bivariate Data

Learning Objectives

1. Define linear regression
2. Identify errors of prediction in a scatter plot with a regression line

In simple linear regression, we predict scores on one variable from the scores on a second variable. The variable we are predicting is called the *criterion variable*and is referred to as Y. The variable we are basing our predictions on is called the*predictor variable* and is referred to as X. When there is only one predictor variable, the prediction method is called *simple regression*. In simple linear regression, the topic of this section, the predictions of Y when plotted as a function of X form a straight line.

The example data in Table 1 are plotted in Figure 1. You can see that there is a positive relationship between X and Y. If you were going to predict Y from X, the higher the value of X, the higher your prediction of Y.

Table 1. Example data.

X	Y
1.00	1.00
2.00	2.00

3.00	1.30
4.00	3.75
5.00	2.25

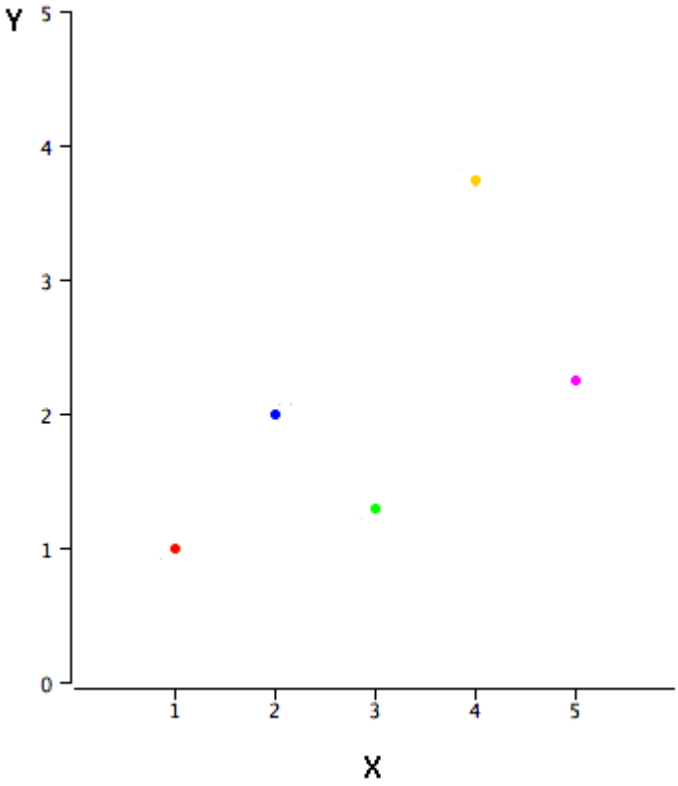


Figure 1. A scatter plot of the example data.

Linear regression consists of finding the best-fitting straight line through the points. The best-fitting line is called a *regression line*. The black diagonal line in Figure 2 is the regression line and consists of the predicted score on Y for each possible value of X. The vertical lines from the points to the regression line represent the errors of prediction. As you can see, the red point is very near the regression line; its error of prediction is small. By contrast, the yellow point is much higher than the regression line and therefore its error of prediction is large.

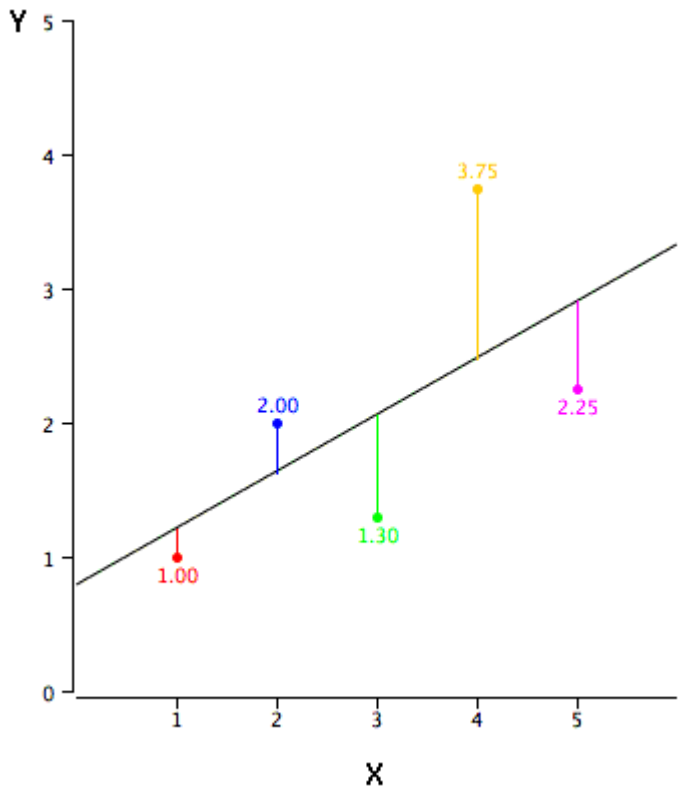


Figure 2. A scatter plot of the example data. The black line consists of the predictions, the points are the actual data, and the vertical lines between the points and the black line represent errors of prediction.

The error of prediction for a point is the value of the point minus the predicted value (the value on the line). Table 2 shows the predicted values (Y') and the errors of prediction (Y-Y'). For example, the first point has a Y of 1.00 and a predicted Y (called Y') of 1.21. Therefore, its error of prediction is -0.21.

Table 2. Example data.

X	Y	Y'	Y-Y'	(Y-Y')^2
1.00	1.00	1.210	-0.210	0.044
2.00	2.00	1.635	0.365	0.133
3.00	1.30	2.060	-0.760	0.578

4.00	3.75	2.485	1.265	1.600
5.00	2.25	2.910	−0.660	0.436

You may have noticed that we did not specify what is meant by "best-fitting line." By far, the most commonly-used criterion for the best-fitting line is the line that minimizes the sum of the squared errors of prediction. That is the criterion that was used to find the line in Figure 2. The last column in Table 2 shows the squared errors of prediction. The sum of the squared errors of prediction shown in Table 2 is lower than it would be for any other regression line.

The formula for a regression line is

$$Y' = bX + A$$

where Y' is the predicted score, b is the slope of the line, and A is the Y intercept. The equation for the line in Figure 2 is

$$Y' = 0.425X + 0.785$$

For X = 1,

$$Y' = (0.425)(1) + 0.785 = 1.21.$$

For X = 2,

$$Y' = (0.425)(2) + 0.785 = 1.64.$$

Computing the Regression Line

In the age of computers, the regression line is typically computed with statistical software. However, the calculations are relatively easy, and are given here for anyone who is interested. The calculations are based on the statistics shown in Table 3. M_X is the mean of X, M_Y is the mean of Y, s_X is the standard deviation of X, s_Y is the *standard deviation* of Y, and r is the *correlation* between X and Y.

Formula for standard deviation
Formula for correlation

Table 3. Statistics for computing the regression line.

M_X	M_Y	s_X	s_Y	r
3	2.06	1.581	1.072	0.627

The slope (b) can be calculated as follows:

$$b = r\ s_Y/s_X$$

and the intercept (A) can be calculated as

$$A = M_Y - bM_X.$$

For these data,

$$b = (0.627)(1.072)/1.581 = 0.425$$

$$A = 2.06 - (0.425)(3) = 0.785$$

Note that the calculations have all been shown in terms of sample statistics rather than population parameters. The formulas are the same; simply use the parameter values for means, standard deviations, and the correlation.

Standardized Variables

The regression equation is simpler if variables are *standardized* so that their means are equal to 0 and standard deviations are equal to 1, for then b = r and A = 0. This makes the regression line:

$$Z_{Y'} = (r)(Z_X)$$

where $Z_{Y'}$ is the predicted standard score for Y, r is the correlation, and Z_X is the standardized score for X. Note that the slope of the regression equation for standardized variables is r.

A Real Example

The case study "SAT and College GPA" contains high school and university grades for 105 computer science majors at a local state school. We now consider how we could predict a student's university GPA if we knew his or her high school GPA.

Figure 3 shows a scatter plot of University GPA as a function of High School GPA. You can see from the figure that there is a strong positive relationship. The correlation is 0.78. The regression equation is

$$\text{University GPA}' = (0.675)(\text{High School GPA}) + 1.097$$

Therefore, a student with a high school GPA of 3 would be predicted to have a university GPA of

$$\text{University GPA}' = (0.675)(3) + 1.097 = 3.12.$$

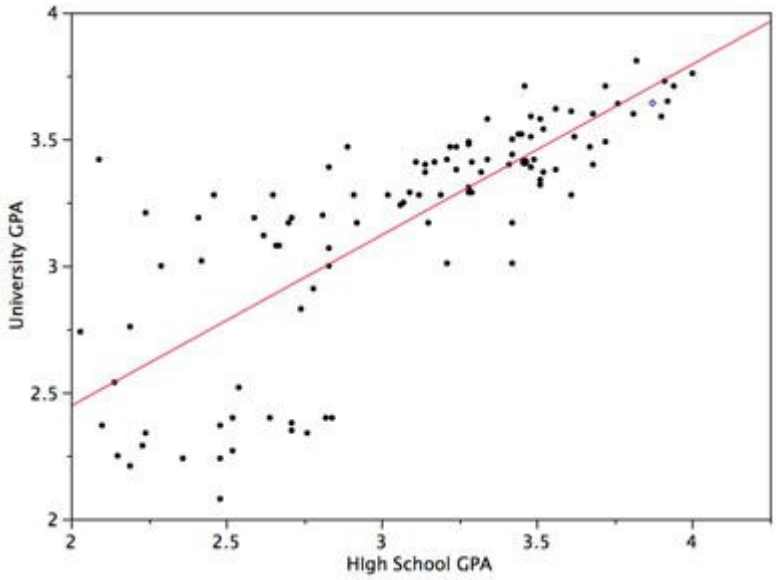


Figure 3. University GPA as a function of High School GPA.

Assumptions

It may surprise you, but the calculations shown in this section are assumption-free. Of course, if the relationship between X and Y were not linear, a different shaped function could fit the data better. *Inferential statistics* in regression are based on several assumptions, and these assumptions are presented in a later section of this chapter.

Логистическая регрессия

Выполнение

Матрица входных данных

Сгенерированная для 4 варианта

Id	Machine	OutPutVoltage	Amperage	Дата выпуска	Failuredate	TimeWork	isFailure
1	MACHINE1	5,04	1004	15.11.2013		365	0
2	MACHINE2	5,05	990	15.11.2013	03.06.2014	200	1
3	MACHINE3	5	1000	15.11.2013		365	0
4	MACHINE3	4,92	970	15.11.2013		365	0
5	MACHINE3	5,02	1000	15.11.2013	04.04.2014	140	1
6	MACHINE3	4,93	1005	15.11.2013		365	0
7	MACHINE2	4,97	1030	15.11.2013	31.10.2014	350	1
8	MACHINE2	5,03	1025	15.11.2013	11.09.2014	300	1
9	MACHINE2	5	1000	15.11.2013		365	0
10	MACHINE1	5	1030	15.11.2013		365	0
11	MACHINE1	5,1	1000	15.11.2013	25.03.2014	130	1
12	MACHINE1	5,1	1050	15.11.2013		365	0
13	MACHINE1	5,1	955	15.11.2013	13.07.2014	240	1
14	MACHINE2	5,05	1003	15.11.2013		365	0
15	MACHINE2	5,01	1002	15.11.2013	11.09.2014	300	1
16	MACHINE2	5,01	1005	15.11.2013		365	0
17	MACHINE3	5	956	15.11.2013		365	0
18	MACHINE1	5,02	1000	15.11.2013	12.08.2014	270	1
19	MACHINE3	5,03	1020	15.11.2013	22.08.2014	280	1
20	MACHINE3	5,07	1002	15.11.2013		365	0

Кодирование категориальной переменной

Сводная таблица отображает количество хороших и плохих (0 и 1 соответственно) изделий и, общее количество произведённых машиной изделий.

Количество по полю isFailure	**Названия столбцов**

Названия строк	0	1	Общий итог
MACHINE1	3	3	6
MACHINE2	3	4	7
MACHINE3	5	2	7
Общий итог	**11**	**9**	**20**

Используя данные сводной таблицы, рассчитал отношения хороших и плохих изделий, произведённых каждой машиной к общему числу хороших и плохих изделий соответственно. Используя эти значения для каждой машины рассчитал:

Весомость признака $WOE_j = \ln\left(\frac{P_j}{Q_j}\right)$

где P_j – отношение количества позитивных объектов в j-ой категории, к числу всех позитивных объектов

Оценка информативности переменных

$$P$$

$$(¿¿\, j - Q_j) * \ln\left(\frac{P_j}{Q_j}\right)$$

$$IV = \sum_{j=1}^{N} ¿$$

.

Machine	Отношение хороших изделий	Отношение плохих изделий	WoE	IV слагаемые
MACHINE 1	0,272727273	0,333333333	-20,0671	0,012162
MACHINE 2	0,272727273	0,444444444	-48,8353	0,083859
MACHINE 3	0,454545455	0,222222222	71,562	0,166255

IV = 0,262275576

В соответствии с критериями, определил значимость переменных.

Критерии:

IV < 0.02 — независимая переменная не обладает прогностической способностью;

$0.02 \le IV < 0.1$ — низкая прогностическая способность;

$0.1 \le IV < 0.3$ — средняя прогностическая способность;

$0.3 \le IV < 0.5$ — высокая прогностическая способность;

$IV \ge 0.5$ — превосходная прогностическая способность.

$IV > 1$ – требуется дополнительное исследование на независимость переменной.

Переменные имеют среднюю прогностическую способность.

Построение математической модели

Математическая модель позволяет определить надёжность изделий каждого станка в зависимости от производственных параметров и предсказать время работы его продукции. Это позволяет оценивать надёжность комплексов, использующих исследуемые изделия, а также позволяет более эффективно производить контроль качества продукции.

Для реализации был запущен инструмент Регрессия в модуле Анализ Данных. В качестве входного интервала Y выбрано время работы изделий, в качестве входного интервала X: выходное напряжение, силу тока и WoE станка. Получины лист регрессии и коэффициенты регрессии:

Коэффициенты регрессии:	
Y-пересечение	3235,567
WOE станка	-0,04904
Вольтаж	-717,728
Сила тока	0,678973

На основе листа регрессии получил линейную регрессионную модель:
$Y = 3235{,}567 - 0{,}04904\,x_1 - 717{,}728\,x_2 + 0{,}678973\,x_3;$

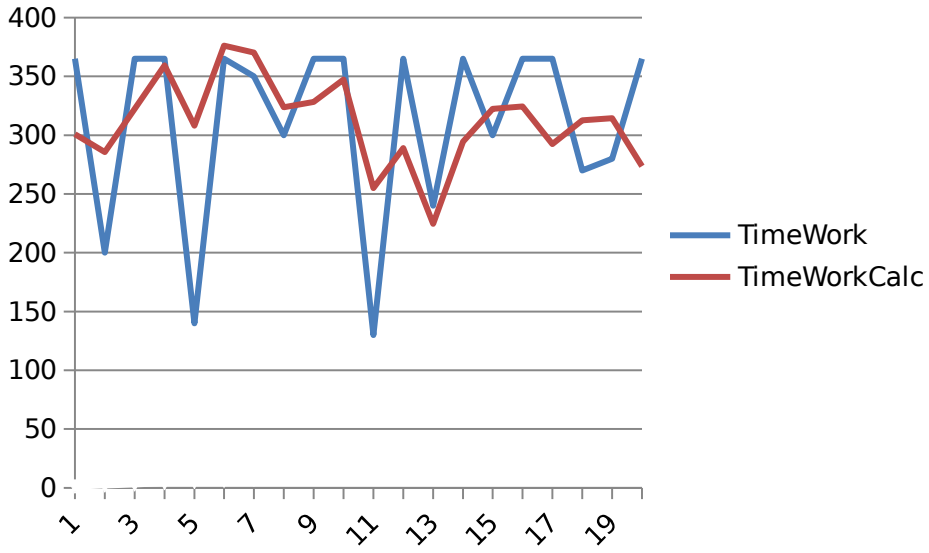
Оценка работы модели

В полученную модель были подставлены входные значения и получены следующие результаты:

WoE	OutPutVoltage	Amperage	β_0	$\beta_1 X_1$	$\beta_2 X_2$	$\beta_3 X_3$	TimeWorkCalc
-20,06706955	5,04	1004	3235,567252	0,984144	-3617,35	681,6886	300,8891
-48,835276	5,05	990	3235,567252	2,395015	-3624,5	672,1829	285,617

79					3		
71,56200364	5	1000	3235,567252	-3,5096	-3588,64	678,9727	322,3886
71,56200364	4,92	970	3235,567252	-3,5096	-3531,22	658,6035	359,4377
71,56200364	5,02	1000	3235,567252	-3,5096	-3603	678,9727	308,034
71,56200364	4,93	1005	3235,567252	-3,5096	-3538,4	682,3675	376,0244
-48,83527679	4,97	1030	3235,567252	2,395015	-3567,11	699,3418	370,1942
-48,83527679	5,03	1025	3235,567252	2,395015	-3610,17	695,947	323,7356
-48,83527679	5	1000	3235,567252	2,395015	-3588,64	678,9727	328,2932
-20,06706955	5	1030	3235,567252	0,984144	-3588,64	699,3418	347,2515
-20,06706955	5,1	1000	3235,567252	0,984144	-3660,41	678,9727	255,1095
-20,06706955	5,1	1050	3235,567252	0,984144	-3660,41	712,9213	289,0581
-20,06706955	5,1	955	3235,567252	0,984144	-3660,41	648,4189	224,5557
-48,83527679	5,05	1003	3235,567252	2,395015	-3624,53	681,0096	294,4437
-48,83527679	5,01	1002	3235,567252	2,395015	-3595,82	680,3306	322,4738
-48,83527679	5,01	1005	3235,567252	2,395015	-3595,82	682,3675	324,5108
71,56200364	5	956	3235,567252	-3,5096	-3588,64	649,0979	292,5138
-20,06706955	5,02	1000	3235,567252	0,984144	-3603	678,9727	312,5277
71,56200364	5,03	1020	3235,567252	-3,5096	-3610,17	692,5521	314,4362
71,56200364	5,07	1002	3235,567252	-3,5096	-3638,88	680,3306	273,5055

Сравнение со входными данными:



Среднеквадратичное отклонение результатов математического моделирования от входных данных составляет 53 дня.

Реализация модели

Реализация на языке c++:

```
double coefficients[4] = {3235,467, -0,04904, -717,728, 0,678973};

double timeWorkCalc ( double woe, double outputVoltage, double amperage) {
   double timeWork = 0;
   timework =  coefficient[0] + coefficient[1]*woe +
            coefficient[2]*outputVoltage + coefficient[3]*amperage;
}
```

Заключение

В данной работе была разработана математическая модель, позволяющая предсказывать время работы изделия в зависимости от ряда производственных параметров, включая конкретный станок, на котором изделие было изготовлено.

Модель была реализована как в среде MS Excel, так и в качестве функции на языке C++, которая может быть встроена в существующие информационные системы.