Modern regression

Jean-Luc Bouchot

School of Mathematics and Statistics
Beijing Institute of Technology
jlbouchot@bit.edu.cn

Spring 2021

Outline

What is regression?

Definition 2.1

Regression, informally speaking, is the art of modeling certain phenomenon from the knowledge of others.

Example 2.1

Let f be the temperature values in degree Fahrenheit. Define c the temperature in degrees Celsius.
We have

$$f = 1.8c + 32 \Leftrightarrow c = \frac{5}{9}(f - 32).$$

In a sense this is a *trivial* regression.

But how did we come to find these values (beside utter randomness of the Fahrenheit scale)? How would we find these coefficients, should we not have access to them?

More examples

Example 2.2

Further examples include:

- Diameter $\leftrightarrow$ circumference
- Hooke's law for a spring: $Y = \alpha + \beta X$ with Y the stretch (displacement) and X the load/weight
- Ideal gas law: $PV = nRT$ for a pressure P, volume V, temperature T. R is a constant (called the gas constant) and n the amount of substance. Assuming the temperature is constant, we can derive Boyle's law: $P = \frac{\alpha}{V}$.

What's the difference?

Generally speaking the following is more or less agreed:

- Regression problems: find the parameters θ defining a function $f_\theta : \mathbb{R}^d \to \mathbb{R}$.
- Classification problems: find the parameters θ defining a function $f_\theta : \mathbb{R}^d \to \mathcal{A}$ with $|\mathcal{A}| < \infty$.

Basic variables

Definition 2.2

The output variable of the regression will carry the following (usual) names:

- *response variable or*
- *dependent variable or*
- *outcome variable.*

Similarly, the input variable of the regression will carry the following (usual) names:

- *predictor or*
- *independent variable or*
- *explanatory variable.*

Linear vs. nonlinear

Definition 2.3

We will talk about a **linear regression** *problem in the case that the function* f_θ *is linear with respect to the parameters, i.e.* $f_{\alpha+\beta} = f_\alpha + f_\beta$. *If this is not the case, we will talk about* **nonlinear regression** *problems.*
We say that the response variable depends linearly on the parameters (and linearly or not on the independent variables).

Examples

Example 2.3

Hooke's law is a linear regression. So is Boyle's law.

Example 2.4

Let $f_\theta : \mathbb{R} \to \mathbb{R}$ be a function defined as

$$f_\theta(t) = \alpha + \beta \exp(-t),$$

where the set of parameters θ is the pair (α, β).

Simple vs. multiple

Definition 2.4

We talk about **simple regression** *problems if the dimensionality of the independent variables is* 1*. Equivalently, we have that* $f_\theta : \mathbb{R} \to \mathbb{R}$ *(i.e.* $d = 1$ *in the general case above)*
In case of multivariate predictors we talk about a multiple regression problems.

Examples

Example 2.5

Hooke's law is a simple linear regression. So is Boyle's law.

Example 2.6

Let $f_\theta : \mathbb{R} \to \mathbb{R}$ be a function defined as

$$f_\theta(t) = \alpha + \beta \exp(-t),$$

where the set of parameters θ is the pair (α, β).

What's coming next

Remark 2.1

This course will cover the following topics:

- Simple linear regression (incl. basic examples and definitions, best fitting line, population model, Pearson correlation and hypothesis testing, ANOVA)
- Multiple linear regression (incl. review of linear and matrix algebra, some examples, model evaluation)
- Polynomial regression (modeling, extension to other basis)
- Regularization and some optimization (note that we will review some optimization throughout the course anyway), model fitting and constraints (e.g. non negativity)
- Non linear regression (Generalized Linear Models, Exponential model, Poisson regression)
- Categorical data: logistic regression, SVM, kernel SVM

Depending on progress and interest, we may cover some of the following aspects:

- BIC / AIC

Outline

SLR: Intro

Remember *simple* refers to the fact that only one independent variable is used (think of temperature at a place as a function of time). The *linear* part refers to the idea that the function used to model the phenomenon (or the function we are trying to fit, depending on the point of view) is linear with respect to the parameters.

Ohm' law

Example 3.1

Ohm's law relates intensity as a function of the resistance and a voltage. Assume you are given a small resistor for which you have no clue the resistance value. However, you know can control the voltage U in a circuit and can measure the intensity I. Then, using

$$U = rI,$$

you can (hopefully easily) obtain an estimation of the resistance value of your component.
Let's have a little sketch of that...

US Skin cancer survey

Example 3.2

What the simple linear regression is doing is pretty much finding the best fitting line. Of course, depending on how we measure quality, this *best fit* might not be easy to compute.
Let us look at an example. The following data look at the skin cancer mortality rate per state in the US (49 points, the data is fairly old and neither Hawaii nor Alaska was American at that time) and looks at how the relate to the latitude of their capitals.

SLR: Intro

A simple linear regression is pretty much this: simple and linear. Let us start with its simplest form: the dependent variable is expected to be linear with the independent one.
While simple this is not unheard of: most of the examples given in the introduction are simple linear regression examples which happen to be linear (affine to be more precise) in the independent variable.

Setup

We have the following formulation:

$$y = \alpha + \beta x.$$

We are given some samples: $\mathcal{D} = \{x_i, y_i\}_{1 \leq i \leq n}$. These samples may be noisy:

$$y_i = \alpha + \beta x_i + \varepsilon_i.$$

For simplicity, we will write the previous noise free relation as $y_i = f_{\alpha,\beta}(x_i) = \alpha + \beta x_i$. Moreover, we can also safely assume the noise components to be sampled i.i.d to ε and to be unbiased, i.e. $\mathbb{E}[\varepsilon] = 0$.

Least squares estimators

We are trying to find the parameters that best fit the data. How to define best or better (or even good!) is via a **least squares** criterion. This is saying nothing else than simply: minimize the sum of squared deviation. Formally speaking, we have

$$S(\alpha, \beta|\mathcal{D}) := \frac{1}{2} \sum_{(x,y)\in\mathcal{D}} |f_{\alpha,\beta}(x) - y|^2. \tag{3.1}$$

Exercise 3.1 (in-class)

Find the coefficients α and β which achieve the best (i.e. least) sum of squared error.

Lest squares estimators

Definition 3.1

The optimal $\widehat{\alpha}$ and $\widehat{\beta}$ found as minimizers to the sum of squares criteria are called the **least squares estimators**.

Computations for Least Squares Estimators

Proposition 3.1 (in-class)

Show that the following formulas hold:

$$\widehat{\alpha} = \bar{y} - \widehat{\beta}\bar{x}$$
$$\widehat{\beta} = \frac{\sum_i (x_i - \bar{x}) y_i}{\sum_i (x_i - \bar{x})^2},$$

where

$$\bar{x} = \frac{\sum_i x_i}{n}$$
$$\bar{y} = \frac{\sum_i y_i}{n}$$

Explain why we also have

$$\widehat{\beta} = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sum_i (x_i - \bar{x})^2}.$$

Definition 3.2

A **linear estimator** *is one such that the output is computed as a linear combination of the measured variables. More formally, an estimator* $\widehat{u}$ *of a variable* u *is linear if, given measurements* y_i, $1 \leq i \leq n$, *we can find coefficients* c_i *such that*

$$\widehat{u} = \sum_{i=1}^{n} c_i y_i.$$

Proposition 3.2

The least squares estimators are linear estimators.

Proposition 3.3

Assuming the noise in the measurements to have 0 *mean, the least squares estimators are unbiased.*

Proposition 3.4

The least squares estimator enjoys some interesting properties.

- *The least squares line passes through the centroid of the data points, i.e. the point whose coordinates are* $(\bar{x}, \bar{y})$.
- *The sum of the predictions is equal to the sum of the observations:* $\sum_{i=1}^{n} y_i = \sum_{i=1}^{n} \widehat{y}_i$ *with* $\widehat{y}_i := \widehat{\alpha} + \widehat{\beta} x_i$.

Remark 3.1

The prediction of the output variable y_i is denoted $\widehat{y_i}$ and is given by

$$\widehat{y_i} = \widehat{\alpha} + \widehat{\beta} x_i.$$

Similarly the prediction error, or residual, is

$$r_i = y_i - \widehat{y_i}.$$

Proposition 3.5

The variance of the estimators can be computed as

$$\mathrm{Var}(\widehat{\beta}) = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2},$$

$$\mathrm{Var}(\widehat{\alpha}) = \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \right).$$

Moreover, the estimators are unbiased. (admitted for now)

Remark 3.2

The linear regression model is valid under the following assumptions:

- **L** Linear function: the response dependent linearly on the predictor.
- **I** Independence: the noise terms in the measurements are independent.
- **N** Normally Distributed: the errors are distributed according to a normal distribution.
- **E** Equal variance: all noises have the same variance σ^2.

And that's how you have a LINEar regression.

Remark 3.3

Assuming the noise terms have variance σ^2, we know that they tell us the spread of the measured data about the predicted value. But we don't know this spread yet!

Definition 3.3

The **sample variance** *is given by*

$$s^2 := \frac{\sum_{i=1}^{n} (y_i - \bar{y})^2}{n-1}.$$

Definition 3.4

The **mean squared error** *is given by*

$$MSE = \frac{\sum_{i=1}^{n} (y_i - \widehat{y_i})^2}{n-2}.$$

It is also sometimes called **residual mean square** *or* **standard error of regression**.

Remark 3.4

How can we justify the $n-2$ instead of the n or $n-1$ expected?

Example 3.3 (in-class)

Assume the following data are given:

Obs. nb	Strength	Age	Obs. nb	Strength	Age
1	2158.70	15.50	2	1678.15	23.75
3	2316.00	8.00	4	2061.30	17.00
5	2207.50	5.50	6	1708.30	19.00
7	1784.70	24.00	8	2575.00	2.50
9	2357.90	7.50	10	2256.70	11.00
11	2165.20	13.00	12	2399.55	3.75
13	1779.80	25.00	14	2336.75	9.75
15	1765.30	22.00	16	2053.50	18.00
17	2414.40	6.00	18	2200.50	12.50
19	2654.20	2.00	20	1753.70	21.50

Table 1: Shear strength vs. age of propellant

Assuming the LINE assumptions are valid, compute the least squares estimators, the residuals, the variances of the estimators, the sample variance and the residual mean square.
How would you estimate the noise in this model (i.e. the σ^2 in the

Proposition 3.6

MSE is an unbiased estimator of the variance in the data σ^2.