Modern Regression

Jean-Luc Bouchot

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

Spring 2021

Outline

Example 2.1

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...

Exercise 2.1 (in-class)

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

Definition 2.1

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

Proposition 2.1 (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 2.2

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 2.2

The least squares estimators are linear estimators.

Proposition 2.3

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

Exercise 2.2

Show the followings

Proposition 2.4

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* $\widehat{y}_i := \widehat{\alpha} + \widehat{\beta} x_i$.

Verify these properties on a notebook/excel spreadsheet/by hand with some data (skin cancer or students' heights or anything of your choosing)

Remark 2.1

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 2.5

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 2.2

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 2.3

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 2.3

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

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

Definition 2.4

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 2.4

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

Example 2.2 (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 measurements)?

Proposition 2.6

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

Definition 2.5

We use the following clues:

- *We call the* **regression sum of squares** *the discrepancy between the average measurements and the predictions:*

$$SSR = \sum_{i=1}^{n} \left(\widehat{y_i} - \bar{y}\right)^2 .$$

  This is sometimes called the **explained sum of squares**.

- *We call the* **error sum of squares** *the discrepancy between measurements and predictions:*

$$SSE = \sum_{i=1}^{n} \left(\widehat{y_i} - y_i\right)^2 .$$

  This is sometimes called the **residual sum of squares**.

- *The* **total sum of squares** *corresponds to the variations of the measurements around the mean value:*

$$SSTO = \sum_{i=1}^{n} \left(y_i - \bar{y}\right)^2 .$$

Proposition 2.7 (Prove at home)

Under the conditions that the LINE assumptions are valid, the following partition of errors hold

$$SSTO = SSR + SSE.$$

Definition 2.6

The **coefficient of determination** *(also often called the r^2 coefficient) corresponds to the ratio of the total sum of squares explained by the regression sum of squares:*

$$r^2 = \frac{SSR}{SSTO} = \frac{\sum_{i=1}^{n} (\widehat{y}_i - \bar{y})^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2}$$

Example 2.3 (in-class)

Let us consider the following set of data:

Obs. nb	x_i	Measurement	Obs. nb	x_i	Measurement
1	0.538	0.982	2	2.234	0.209
3	4.566	-0.381	4	3.061	1.990
5	2.264	-1.303	6	1.826	-2.677
7	4.783	-1.457	8	2.945	-1.147
9	2.065	-1.013	10	2.289	0.958

Table 2: Some random data

Compute SSTO, SSE, SSR, r^2.

Example 2.4 (in-class)

Let us consider the following set of data:

Obs. nb	x_i	Measurement	Obs. nb	x_i	Measurement
1	0.538	1.196	2	2.234	-1.018
3	4.566	-4.398	4	3.061	-2.545
5	2.264	0.025	6	1.826	-0.699
7	4.783	-4.033	8	2.945	-1.641
9	2.065	-0.338	10	2.289	-3.581

Table 3: Some random data

Compute SSTO, SSE, SSR, r^2.

Definition 2.7 (Pearson correlation coefficient)

Define Pearson's correlation coefficient as

$$r = \text{sign}(\hat{\beta})\sqrt{r^2}.$$

This coefficient, on top of having a magnitude information, contains the upwards or downwards trend of the linear relation.

Proposition 2.8

Pearson's correlation coefficient can be computed as

$$r = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2 \sum_{i=1}^{n} (y_i - \bar{y})^2}}.$$

Proposition 2.9

We also have the following relation

$$r = \frac{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2}}{\sqrt{\sum_{i=1}^{n} (y_i - \bar{y})^2}} \hat{\beta}.$$

Remark 2.5

The r coefficient is telling us a lot about potential linear dependencies:

- r is close to 1: there is a strong evidence that the output is linearly correlated with the predictor and the larger the predictor, the larger the dependent variable.
- r is close to -1: there is a strong evidence that the output is linearly correlated with the predictor and the larger the predictor, the smaller the dependent variable.
- $r \equiv 0$: there is no strong evidence of a linear dependency.

Remark 2.6 (Linear dependencies)

r^2 is an information about linear dependencies.

Remark 2.7 (Trends, not optimality)

The r^2 coefficients is just a suggestion: it suggests that some trend is happening, but it does not mean that the linear model should be the model of choice!

Remark 2.8 (Unstability due to outliers)

The r^2 coefficient is very unstable when facing outliers... even a single one!

Remark 2.9 (Correlation $\neq$ causation)

Just because the r^2 shows some strong evidence towards a linear correlation does not mean there are any sort of causal relationship between the variables.

Definition 2.8 (Hypothesis testing (not a sound definition))

Hypothesis testing is the science of trying to assess a certain hypothesis (assumption we are trying to prove or disprove) based on a reduced sample size.

Definition 2.9

A t-test is a method for hypothesis testing defined as

1. *Define the* **null hypothesis**.
2. *Define your T_0, the* **test statistic**.
3. *Compute T_0 from your n samples .*
4. *Assess whether the value is within reasonable bounds for acceptance of the null hypothesis.*

Remark 2.10

This is a very abstract presentation of statistical testing. Besides practicing on numerous occasions, I have no better suggestions for grasping the whats and hows. Some comments though:

- The test is based on a confidence level θ which should be specified depending on the application.
- The t-test assumes normally distributed errors. If this is not the case, the expected distribution for T_0 is no longer a t distribution with $n-2$ degrees of freedom.

Proposition 2.10

Testing the null hypothesis $H_0 : \beta = \beta_0$, for a given β_0.

1. *Inputs: a set of points $(x_i, y_i)_{1 \leq i \leq n}$; a confidence level α (typically 0.05)*
2. *Compute the means $\bar{x}, \bar{y}$.*
3. *Compute the least squares estimators $\hat{\alpha}$ and $\hat{\beta}$.*
4. *Compute the estimator of the variance*

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

5. *Compute the test statistic*

$$T_0 := \frac{\hat{\beta} - \beta_0}{\sqrt{\frac{\hat{\sigma^2}}{\sum_{i=1}^{n} (x_i - \bar{x})^2}}}.$$

6. *Compare the T_0 with the tables for the t distribution with $n-2$ degrees of freedom and a prescribed confidence level*[a]

[a] See `https://homepage.divms.uiowa.edu/~mbognar/applets/t.html` or `https://stattrek.com/online-calculator/t-distribution.aspx` .

Exercise 2.3

Define the appropriate t tests for the following null-hypothesis:

1. $H_0 : \alpha = \alpha_0$ for a given value α_0.
2. $H_0 : \beta = 0$.

Discuss these results assuming the data in Table 3 with $\alpha_0 = 0$, then $\alpha_0 = -1$.

Exercise 2.4

Split into groups and analyze the dataset that you are given!
You are free to discuss pretty much anything and present anything. But here are some things you might be willing to look into:

- Average value(s)
- (Simple Linear) Least squares estimator
- Plotting the data (with or without predictions?)
- Some testing of parameters (say, are we expecting a slope of -1)
- Any comments on the data? Is the r^2 as expected?

Definition 2.10 (Confidence intervals)

Let $\widehat{\theta}$ be an estimator of the true variable θ_0. For a given $\alpha \in [0,1]$, the $(1-\alpha)100\%$ confidence interval is an interval with bounds θ^- and θ^+ such that

$$\mathbb{P}\left[\theta \in [\theta^-, \theta^+]\right] = 1 - \alpha.$$

The values θ^- and θ^+ are defined such that, for a random variable v, it holds

$$\mathbb{P}[v \leq \theta^-] = \alpha/2, \quad \textit{and} \quad \mathbb{P}[v \leq \theta^+] = 1 - \alpha/2.$$

Definition 2.11 (Confidence intervals for t distributions)

For a given confidence value $\alpha \in [0, 1]$ the $(1 - \alpha)100\%$ confidence interval of an estimator t distributed random variable $\widehat{\theta}$ is given by

$$\theta \in [\widehat{\theta} - t_{\alpha/2,df}\,\mathrm{se}(\widehat{\theta}), \widehat{\theta} + t_{\alpha/2,df}\,\mathrm{se}(\widehat{\theta})].$$

se denotes the standard error and df denotes the number of degree of freedoms. $t_{\alpha,df}$ denotes the value t for which the random variable

$$v = \frac{\widehat{\theta} - \theta_0}{\mathrm{se}(\widehat{\theta})}$$

satisfies

$$\mathbb{P}(v \leq t) = \alpha.$$

Remark 2.11

We have already seen some standard errors:

- The standard error of the estimate of the slope is given by

$$\mathrm{se}(\widehat{\beta}) := \sqrt{\frac{MSE}{\sum_{i=1}^{n}(x_i - \overline{x})^2}}.$$

- The standard error of the estimate of the intercept is given by

$$\mathrm{se}(\widehat{\alpha}) := \sqrt{MSE\left(\frac{1}{n} + \frac{\overline{x}^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2}\right)}$$

with $MSE := \frac{\sum_{i=1}^{n}(y_i - \widehat{y_i})^2}{n-2}$ denoting the Mean Squared (residual) Error.

Remark 2.12

Confidence intervals have an intrinsic frequentist point of view!

Exercise 2.5

Remembering that the (residual) Mean Squared Error is an (unbiased) estimator for the variance of the noise in the data. Compute what should be the confidence interval.