NOTES ON LINEAR REGRESSION

September 2022

Nuno Sepúlveda & João Malato

NGSchool

Preface

Ikigai is a Japanese word that can be translated to the "purpose of getting up in the morning", "life's purpose", or even "joie de vivre". The practical utility of the word is given by its intimate connection with longevity, good health, and good quality of life, as demonstrated by many scientific studies. Given its existence but above all, its routine use among locals, it is no surprise that Japan is a nation that enjoys long-lived people.

Having a sense of Ikigai can be built up on five foundational pillars with no particular order of importance:

- start small;
- release your ego;
- harmony and sustainability;
- the joy of little things;
- being here and now.

These notes were created having these pillars in mind. In particular, we started small with simple linear regression upon which more complex models can be constructed and understood. In this way, we aim to achieve a certain sense of sustainability and harmony across concepts and models. Hopefully, this facilitates the learning process of the statistical concepts with increasing complexity. We also brought some practical examples whose scientific importance and personal interest of any reader can easily mingle. We also provided some footnotes along the text, which reflect our

own joy for specific details of the methodology. However, we avoided to getting too technical in our description. This decision aimed to release our egos for showing off how much we know — and, for sure we do not know a lot more — and not caring about who is going to read this. At the same time, it gives the chance of the reader to find her/his pathway (or other sources) in the topic. Finally, preparing these notes increased our attention and focus on being here and now. This focus was a good cure against some routine commitments that consume our days.

We hope then that these notes could be a joyful learning material and could motivate anyone of you to embark on a long but sustainable journey into statistical modelling and data analysis of any kind.

Contents

1 Introduction to Linear Regression

The grand purpose of a scientific inquiry is to increase current understanding or knowledge of the world by means of objective thinking; that is why post-truth rulers and their followers are somehow against scientific values. Ideally, this increase in the understanding would lead to better decision-marking by governments and each one of us in our daily life. Philosophically, it can be also seen as a leap forward towards the hidden truth of things.

Objective (or rationale) thinking is simply hardwired in the so-called scientific method as the *modus operandi* for exploring the possible causes of a given phenomenon. The scientific method entails the existence of an initial hypothesis for the explanation of a certain phenomenon. In other words, there is an initial suggestion for a certain cause-effect chain of events.

The next step of the method is to collect data by means of experimentation. These data are then used to compare formally between what is expected under the working hypothesis and what is actually observed in the data. This comparison leads to either the acceptance or the rejection of the working hypothesis.

However, the majority of scientific phenomena might have multiple causes. It is in this scenario that linear regression proves to be a key analytical tool. As we will see, it is a

statistical technique that helps in finding and prioritizing the investigation of competing causes for the same phenomenon. At the same time, linear regression is a pure conceptual construct from the field of Mathematics and Statistics and as such, it should be never invoked as a demonstrative agent of causality.

1.1 Why is it important to learn linear regression?

The question "why?" is in the heart and mind of any scientist or researcher; and, of course, it is also in our mind as a child while discovering the world. In this perspective, it is fair to ask the reason of why it is important to learn linear regression. Three reasons come to mind.

The first one is in the basis of what a scientist means. Being a scientist sets its grounds in an almost-eternal curiosity and persistent enthusiasm about to know how the world works. However, these two soft skills come short when applying the scientific method, namely, when confronting data with the working hypothesis. We need other, hard skills that can help doing exactly that. In this regard, linear regression is a fundamental technique that helps identifying possible causal factors and narrowing them down. Unsurprisingly, linear regression is on the top 5 of the analytical methods used scientific papers published in the last decade [1].

The second reason can be taken from a personal development perspective. For those who would like to embark in the long journey of statistical modelling, linear regression sets a solid foundation for learning more advanced statistical models, such as the generalised linear models, generalised additive models, and generalised linear mixed

models. Learning such advanced models also opens the horizon for the type of data that statisticians, modellers, or data scientists are able to analyse on a routine basis.

Finally, given the widespread application of linear regression in science and technology, investing time in learning it is investing in a tool that brings confidence in a job well done. Ultimately, many jobs-well-done might lead to promotions and a prosperous career development.

1.2 What is linear regression?

Statistical modelling is an iterative process by which different statistical models are tested against the data until the analyst finds one of them that is a right balance between good interpretation, goodness-of-fit, and simplicity. This balance can be simply put as follows: among the satisfactory statistical solutions to a given dataset, the simplest one should be preferred. This is on the basis of the parsimonious principle or the Occam's razor. One should bear in mind that, in the end of the modelling process, 'all models are wrong but some are useful' [2].

Linear regression is a type of statistical modelling in which the tested models represent a quantitative outcome (or a set of outcomes) by a linear combination of the so-called covariates. Here, we will focus our discussion on the basic case in which there is a single outcome variable, Y, and another set of variables $x_1, \ldots, x_p$ that might explain the variation of the latter. This case is usually called the "univariate analysis" in contrast with "multivariate analysis" in which there are multiple variables of primary interest, $Y_1, \ldots, Y_k$.

In the next Section, we will start off with the simplest linear regression model in which there is a single covariate and

a single outcome variable, both quantitative in nature. In the following chapter, we will increase the complexity of model in terms of the number of covariates. In the last chapter, we will discuss the alternative ways of estimating a regression models using penalized strategies. These alternative ways are particularly useful to studies in which there is a large number of covariates under evaluation.

1.3 How to use simple linear regression in practice?

1.3.1 Back to the basics on straight lines

Let Y and x be a quantitative outcome and a quantitative covariate, respectively. Mathematically, the simplest relationship between these two variables is given by a straight line. This simplicity is due to the fact that we only need two points to define a straight line. The respective formula is

$$Y = \beta_0 + \beta_1 x \,, \tag{1}$$

where β_0 and β_1 are called the intercept with the origin and the slope, respectively[1].

For a matter of interpretation, it is important to highlight which quantitative information is conveyed by β_0 and β_1. The parameter β_0 is the value of Y when $x = 0$ (Figure 1A). For example, if x and Y respectively represent a treatment dose and the quantity of a certain molecule, the parameter β_0 provides information about the quantity of that molecule in the absence of treatment. The parameter β_1 is the value

1 To make a clear distinction in the mathematical notation, the variables are typically written with roman letters while (unknown) parameters with greek letters.

by which Y changes per unit of x and it is typically the main interest of the analysis (Figure 1B). In the above example, the parameter β_1 represents how much the quantity of the molecule changes per dose unit. If β_1 is positive, then the outcome Y increases with x. In contrast, when β_1 is negative, then the outcome Y decreases with x. When β_1 is zero, then the outcome Y does not vary with x.

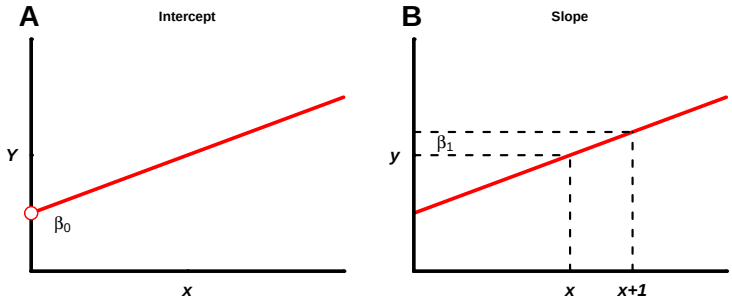

Figure 1: Visual interpretation of the parameters β_0 **(A)** and β_1 **(B)** in simple linear regression with a positive slope.

1.3.2 Sampling data from perfect and real world

Let's assume that we have collected a random sample of n individuals from the population in order to study the relationship between x and Y. The respective data are represented by the pairs $(x_i, Y_i, i = 1, \ldots, n)$.

In a perfect world where there is no uncertainty (or, in other words, where determinism occurs), the underlying straight line emerges clearly from the data without any deviations from its theoretical formulation (Figure 2A). In this situation, we predict the exact value of Y given x and the statistical exercise is facilitated.

Unsurprisingly, the real world is imperfect and, therefore, we cannot observe the straight line in its theoretical perfection (Figure 2B). The source of such imperfection might be the presence of experimental error when measuring the outcome. It might be other unmeasured or unaccounted covariates that influence the outcome. When we observe uncertainty in the outcome variable Y (and not in the covariate x), the basic formulation of the straight line should be somehow extended to account for such situation.

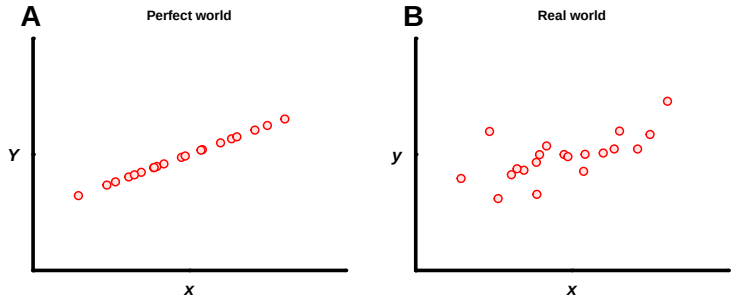

Figure 2: Scatter plots of $n = 20$ individuals from simple linear regression observed in a perfect and real world (**A** and **B**, respectively.

The simple extension is to add a random deviation ϵ to equation (1). This deviation is also known as the residual. We assume that each individual has its own residual resulting from a random draw of a Normal distribution with mean (or expected value) 0 and variance σ^2. This assumption leads to the following extended formulation of a straight line

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \; i = 1, \ldots, n \; , \tag{2}$$

where $\epsilon_i | \sigma \rightsquigarrow \mathcal{N}(0, \sigma^2)$. An important implication of using a Normal distribution for the residuals is that, in theory,

one can observe positive or negative values for the outcome variable Y_i. For the mathematical afficionados, another way of saying it is that the outcome variable Y_i should have a support or a domain in the real space. However, in practice, the above model can be applied to outcome variables defined in a subset of the real space as long as the respective model predictions outside this subset have a negligible probability.

By the basic properties of the Normal distribution, it is easy to confirm that the above formulation is equivalent to

$$Y_i|\mu_i, \sigma \rightsquigarrow \mathcal{N}(\mu_i = \beta_0 + \beta_1 x_i, \sigma^2) \ . \tag{3}$$

This alternative formulation motivates to re-write the simple linear regression model in terms of expected values, i.e.,

$$\mu_i = \beta_0 + \beta_1 x_i, \ i = 1, \ldots, n \ . \tag{4}$$

1.3.3 Guessing the unknowns

Given the data, the basic statistical question is to know which are the best guesses for the unknown parameters β_0 and β_1 in some statistical sense. A possible answer to this question is given by the (ordinary) least squares method. Intuitively, this method is based on the idea of determining the values of β_0 and β_1 (hereafter denoted as $\hat{\beta}_0$ and $\hat{\beta}_1$) that minimize the distance between the observed data and the hypothetical straight line passing through the data points. More precisely, the least squares method aims to determine $\hat{\beta}_0$ and $\hat{\beta}_1$ that minimize the following sum of squares

$$\sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2 \ . \tag{5}$$

If we define the residual related to i-th individual by $\epsilon_i = y_i - \beta_0 - \beta_1 x_i$, then the above minimisation problem is equiv-

alent to say that $\hat{\beta}_0$ and $\hat{\beta}_1$ are the values that minimises the sum of the squared residuals.

Using basic calculus, one can demonstrate that the above minimisation problem has the following solutions:

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \,, \tag{6}$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \,, \tag{7}$$

where $\bar{x}$ and $\bar{y}$ are the sampled means of $x_1, \ldots, x_n$ and $y_1, \ldots, y_n$. The estimate of σ^2, $\hat{\sigma}^2$ can be simply given by the variance of the estimated residuals, i.e.,

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2}{n-2} \,. \tag{8}$$

An alternative way to determine $\hat{\beta}_0$, $\hat{\beta}_1$, and $\hat{\sigma}^2$ is to use the maximum likelihood method. This method aims to maximise the so-called likelihood function for the data. Mathematically, the likelihood function is equivalent to the sampling distribution of the data but taking it as a function of the unknown parameters given the observed data.

Under the basic assumption that individuals are independent of each other, the sampling distribution can be written as sampling from a Normal distribution described by equation (3), i.e.,

$$f\left(\{y_i\}\right) | \{x_i\}, \beta_0, \beta_1, \sigma^2\right) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y_i - \beta_0 - \beta_1 x_i)^2}{2\sigma^2}} \,. \tag{9}$$

The likelihood function is then given by

$$\mathcal{L}\left(\beta_0, \beta_1, \sigma^2 | \{y_i, x_i\}\right) \equiv f\left(\{y_i\}\right) | \{x_i\}, \beta_0, \beta_1, \sigma^2\right) \,. \tag{10}$$

In this sense, the maximum likelihood method leads to $\hat{\beta}_0$, $\hat{\beta}_1$, and σ^2 that maximise the (exact) chance of observing the data at hand. Interestingly, in this specific model,

this method provides the same set of estimates as the least squared method. This theoretical coincidence is not true in general when using these methods for other type of models.

1.3.4 Making decisions about the unknowns

After having estimated the unknown parameters of a simple linear regression model, we can advance in the data analysis by performing hypothesis testing that can help simplifying the model. In general, we are interested in knowing whether there is sufficient evidence for a relationship between the covariate and the outcome (i.e., $\beta_1 \neq 0$). This gives rise to the following set of null and alternative hypothesis (H_0 and H_1, respectively)

$$H_0 : \beta_1 = 0 \text{ versus } H_0 : \beta_1 \neq 0 \, . \tag{11}$$

Note that $\beta_1 = 0$ is equivalent to say that the covariate does not influence the outcome[2]. Also note that the statistical formulation of the test does not allow to swap simply the null hypothesis with the alternative one. Therefore, we should know in advance which hypotheses are specified in a given test.

The basic idea is then to decide on what is the evidence conveyed by the observed data concerning the validity of H_0. To do that, we need a test statistic that measures the distance between what is predicted by H_0 and what is actually observed and whose probability distribution under the validity of H_0 is known upon hypothetical repeated sampling[3]. In

2 A more general test is $H_0 : \beta_1 = b_1$ versus $H_0 : \beta_1 \neq b_1$ where b_1 is a precise value for the unknown slope β_1. However, the use of this test only makes sense in very specific applications whose discussion is out of the scope of these notes.

3 Repeating sampling is on the basis of the classical or the frequentist approach to statistical inference. In this approach, the statistical analysis

most statistical software programs, it is implemented the so-called Wald's test which has the following statistic

$$T = \frac{\hat{\beta}_1}{se\left(\hat{\beta}_1\right)} \underset{H_0}{\rightsquigarrow} t_{(n-2)} \ , \tag{12}$$

where $se\left(\hat{\beta}_1\right)$ represents the standard error of the estimator $\hat{\beta}_1$[4] and $t_{(n-2)}$ is the resulting Student's t-distribution with $n-2$ degrees of freedom under H_0. In this context, very low (negative) values of the Wald's statistic are very highly improbable under H_0 (Figure 3A). The same happens for very high (positive) values; this rationale only applies when $H_1 : \beta_1 \neq 0$ where significant deviations of the slope can be either negative or positive. Therefore, the null hypothesis should be rejected if the test statistic calculated in your data lies in one of these two extremes.

The question is then to know which values of the Wald's statistic are deemed too low or too high. To answer this question, we invoke the so-called significance level of the test, hereafter denoted by α. This fundamental concept is the probability of rejecting the null hypothesis given that it is actually true. This is also called the type I error. A convention among the scientific community is to use $\alpha = 0.05$. However, there is no theoretical reason for not choosing any other value[5].

---

of a given dataset should be seen as one of an infinite statistical analyses that could be done if sampling was repeated from the population.

4 The standard error of an estimator is the standard deviation of all possible estimates for a given parameter under repeating sampling.

5 The limited understanding of what the p-value means has been widely discussed. Given its misuse and potential manipulation, some scientific circles have decided to outright ban its use. This decision led to an immediate reaction of the statistical community to gather together so to clarify the meaning and the purpose of the concept [3].

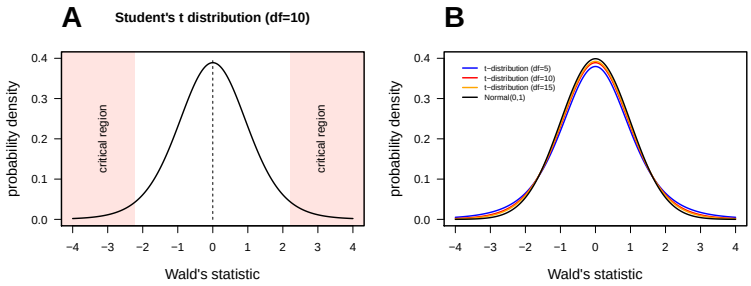

Figure 3: Wald's statistic and the respective t-distribution distribution with 10 degrees of freedom (df) under $H_0 : \beta_1 = 0$ for a sample size of n individuals **(A)**; the filled areas represent the values of the statistics that suggests the rejection of the null hypothesis at the usual 5% significance level. When the sample size is sufficiently large, the t-distribution can be approximated by a Normal distribution **(B)**.

Having the significance level specified, we can then equally distribute it to each of extreme tail of the above t-distribution. This creates the so-called critical region of a test whose values have probability of the test statistic under the null hypothesis lower than the significance level, thus, suggesting the rejection of that hypothesis (Figure 3A).

In practice, a simple decision is to use the p-value and compare it against the significance level. In general, the p-value can be seen as the probability of observing a value of the test statistic equal or more extreme than the one calculated for the data under analysis. The idea is that if p-value is too low, the value of the test statistic is on the extremes of the respective probability distribution under the null hypothesis. We can then take the following automated decision: reject H_0 if p-value $< \alpha$, otherwise do not reject it.

An important cautionary note is that if the p-value suggests the rejection of H_0, we might be making an error, albeit with a low probability, by following such suggestion. In contrast, if the p-value does not suggest the rejection of the null hypothesis, we might be also making an error but of a different kind: accepting the null hypothesis when it is actually false (also known as the type II error). Given these two opposing errors, we should avoid making strong statements such as

> *the covariate influences the outcome,*

or

> *the covariate does not influence the outcome.*

We should be careful with our (statistical) decision and its implication given that we are working with the unknown. Therefore, it is more advisable to say something along these lines:

> *there is not enough evidence to say that covariate influences the outcome,*

or

> *there is sufficient evidence for a significant (linear) effect of the covariate on the outcome.*

Another hypothesis test for the same data is to confront

$$H_0 : \beta_0 = 0 \text{ versus } H_0 : \beta_0 \neq 0 \, . \tag{13}$$

To do that, we can use the Wald's statistic again, but now defined in terms of $\hat{\beta}_0$, i.e.,

$$T = \frac{\hat{\beta}_0}{se\left(\hat{\beta}_0\right)} \underset{H_0}{\rightsquigarrow} t_{(n-2)} \, . \tag{14}$$

The decision for this test can be simply made using the p-value as described.

Before going to the real-world data, we make one quick remark about simple linear regression in the context of a covariate x that only takes values (0/1, all/nothing, standard/new treatment, etc) instead of being quantitative as previously assumed. To test the influence of this covariate on the outcome, the equation (4) of the simple linear regression model implies that there are only two possible means for Y_i, i.e.,

$$\mu_i = \begin{cases} \beta_0, & \text{if } x_i = 0 \\ \beta_0 + \beta_1, & \text{if } x_i = 1 \end{cases} \tag{15}$$

Therefore, testing $\beta_1 = 0$ is conceptually equivalent to testing the same mean of the outcome when dividing the sampled individuals into those with $x = 0$ and $x = 1$. Given that the residuals were assumed to be normally distributed with the same variance for all possible values of x, we are in the conditions to apply the popular t-test for comparing two independent means. We can even relax the assumption of the same variance across the groups under comparison and apply the so-called Welch-Satterthwaite's t-test as available in the R software.

1.4 Model Diagnostics or Validation

The final but often neglected stage of the analysis is the so-called model diagnostic (or model validation) that consists in verifying whether the assumptions are somehow in agreement with the data. This verification consists in a detailed analysis of the (estimated) residuals. In particular, the estimated residuals should follow a Normal distribution

approximately[6]. This assumption can be done by performing a visual inspection (e..g, boxplot, density plots or Q-Q plots) or assessing it formally via a statistical test (e.g., Lilliefors or Shapiro-Wilk tests). In the case that there is no evidence for a normal distribution in the residuals, one can apply a convenient transformation to the outcome variable (e.g., log-transformation) and re-estimate the model in these transformed data. A useful modelling strategy is to search the best Box-Cox transformation according to any goodness-of-fit test for the Normal distribution, as available in the `AID` package for the R software [4].

Another basic assumption is that the variation of the residuals should not vary with the covariate x. This assumption can be assessed visually by a scatter plot where the values of the covariate are plotted against the respective residuals. In this regard, the residuals should be placed along a symmetric horizontal band centered in zero. If the residuals show an increasing or decreasing trend with the covariate, we should consider a more complex model to the data. For example, we can construct a model in which the mean and variance of the outcome are described by two separate simple linear regression models.

Other diagnostic tools are available in the literature, such as calculating the influence of each observation in model estimation by leverage. A discussion about these alternative tools can be found elsewhere.

6 In general, it is better to use the standardized residuals in order to compare them to the Standard Normal distribution.

1.5 Are you constructing and interpreting a model responsibly?

Responsible statistical modelling presupposes a certain ethos during the process of finding a (reasonable) model to the data. The adjective 'responsible' integrates the following three ethical aspects while conducting data analysis.

The first aspect embodies the technical concepts that allow the construction of a model and the performance of the subsequent statistical inferences. More precisely, it is related to model assumptions and their validation as discussed in the previous section. The failure to provide sufficient evidence for reasonable model assumptions is similar to building scientific knowledge based on houses of cards, sandcastles, or flying pigs.

The second aspect is related to interpreting the model responsibly. In this regard, we should deeply aware that simple linear regression can only provide evidence of a significant relationship between the covariate and the outcome. In other words, we should never interpret this relationship as causality. We should also be aware that the model predictions for the outcome variable only make sense within the range of values for covariance x defined by the data. That is, interpolation of the model outcome is reasonable to do. As tempting it can be, extrapolation of the model outcome should be avoided at all cost.

The third one lies in the realm of model uncertainty or explainability (in a statistical sense). This aspect comes after the validation and interpretation stages when it is important to evaluate how much the data variation is in fact explained by the proposed model. A simple way to do this evaluation is to calculate the proportion of the variation explained by

the model. This can be done by the so-called coefficient of determination (R^2) defined as

$$R^2 = 1 - \frac{\sum_{i=1}^{n} \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right)^2}{\sum_{i=1}^{n} \left(y_i - \bar{y}\right)^2} . \tag{16}$$

Note that $\sum_{i=1}^{n} \left(y_i - \hat{\beta}_0 - \beta_1 x_i\right)^2$ is the sum of squared residuals estimated by the model while $\sum_{i=1}^{n} \left(y_i - \bar{y}\right)^2$ is the total sum of squares in the absence of the covariate. The interpretation of this measure is straightforward: higher the values of R^2 better the model is in explaining the variation of the data. The question is what is a reasonable level of explainability.

From a pure theoretical standpoint, we can decide if a model is or not significant by performing a hypothesis given by equation (11). In practice, we should go beyond the statistical significance of a model. Ultimately, there is a limited utility of a significant model with a low level of explainability. In this case, the wisest decision is to invest time and effort to find other covariates that might also influence the outcome. This investment motivates the use of multiple linear regression in which multiple covariates are included in the model. This topic will be discussed in the following Chapter.

2 Multiple Linear Regression

Multiple linear regression extends simple linear regression model to the situation in which there are more than one covariate in the model. The ultimate hope is to construct a model that improve the explainability of the data (or reduces model uncertainty).

To facilitate the discussion, we begin with a multiple linear regression model with two covariates, one quantitative and another binary. We will then dive into the case of two quantitative covariates and end up discussing the general case of constructing a model with p covariates.

2.1 Case I: one quantitative covariate and one qualitative

Let's start with the situation in which there are two covariates, x_1 (quantitative) and x_2 (binary), and the outcome of interest Y. This is a common situation in Epidemiology and Biomedical research where we have a treatment dose whose effect on the outcome can change with the gender or the presence of a given risk factor in a patient.

For a general individual i, the basic model is to consider the so-called main effect model in which the effects of the covariates are independent of each other, i.e.,

$$Y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \epsilon_i \,, \tag{17}$$

where ϵ_i is the residual of the i-th individual that should follow a Normal distribution defined as the simple linear model Given that x_2 can only take the values 0 and 1, we can expand the model as follows

$$Y_i = \begin{cases} \beta_0 + \beta_1 x_{1i} + \epsilon_i, & \text{if } x_{2i} = 0 \\ (\beta_0 + \beta_2) + \beta_1 x_{1i} + \epsilon_i, & \text{if } x_{2i} = 1 \end{cases} \quad (18)$$

Therefore, all the individuals will have the same effect (i.e., slope) of the covariate x_1 on the outcome, but those with $x_2 = 0$ and $x_2 = 1$ have distinct intercept with the origin (β_0 versus $\beta_0 + \beta_2$). This observation implies that the above model is similar to define two parallel straight lines with the covariate x_1 (Figure 4A). The distance between these two lines are simply given β_2.

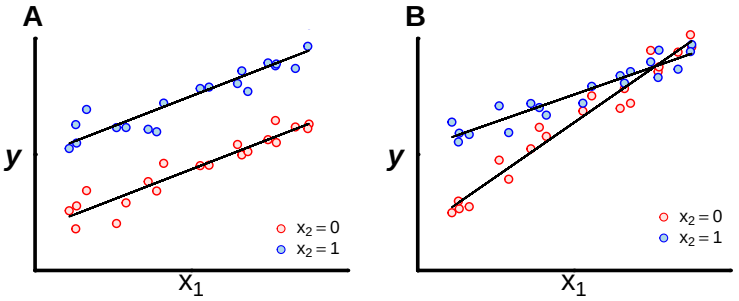

Figure 4: Two linear regression models including one quantitative covariate (x_1) and a binary one (x_2) for an outcome Y: **(A)** model in which there are two (parallel) lines relating x_1 to Y with the same slope but different intercepts with the origin according to $x_2 = 0$ and $x_2 = 1$ (equation (17)); **(B)** model in which there are two lines relating x_1 to Y with different slopes and intercepts with the origin according to $x_2 = 0$ and $x_2 = 1$ (equation (19).

A more general model can be constructed by considering not only different intercepts with the origin, but also differ-

ent slopes for individuals with $x_2 = 0$ or $x_2 = 1$. This can be done by adding a term that describes an interaction between x_1 and x_2 to the above model, i.e.,

$$Y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{1i} x_{2i} + \epsilon_i \,. \tag{19}$$

Again, we can write this model in a less-compact way where the effects of the parameters β_2 and β_3 on the slope and intercept are clearly seen:

$$Y_i = \begin{cases} \beta_0 + \beta_1 x_{1i} + \epsilon_i, & \text{if } x_{2i} = 0 \\ (\beta_0 + \beta_2) + (\beta_1 + \beta_3) x_{1i} + \epsilon_i, & \text{if } x_{2i} = 1 \end{cases} \tag{20}$$

In practice, this model is similar to have two distinct straight lines of Y as a function of x_1 and Y for the two possible values of x_2 (Figure 4B). Therefore, in practice, we could simply split the data set for individuals with $x_2 = 0$ and $x_2 = 1$ and estimate the straight lines separately.

The great advantage of this model is the integration of the whole data set in a single model in which we can draw all the inferences of interest, including possible model simplification. Another advantage is that, given that the distribution of residuals is shared among the individuals, we can reduce the uncertainty of our estimate association with the variation of the residuals. This increases the statistical power of the analysis.

A more general situation is to consider that the covariate x_2 is categorical with more than 2 categories[1]; for simplicity, $x_2 \in \{1, 2, \ldots, K\}$ with $K > 2$. For example, this covariate can represent different populations (e.g., Europeans, Americans, Africans, and Asians), the genotypes of a biallelic genetic marker (e.g., *AA*, *AB*, and *BB*), or different body mass

1 We avoided here the word "levels" because it gives the false impression of having a categorical covariate with different ordinal levels (e.g., mild, moderate, severe, extremely severe).

index categories (below 18.5 – underweight; 18.5 – 24.9 – normal weight; 25.0 – 29.9 – pre-obesity; 30.0 – 34.9 — Obesity class I; 35.0 – 39.9 — Obesity class II; above 40 – Obesity class III, according to the World Health Organization).

To model this situation, we first use a trick in which we re-coded data of this covariate into $K-1$ auxiliary binary variables (called dummy variables) defined as follows

$$x^*_{2li} = \begin{cases} 1, & \text{if } x_{2i} = l \\ 0, & \text{otherwise} \end{cases} \tag{21}$$

where x_{2i} denotes the category of the covariate x_2 observed in the i-the individual, and $l = 2, \ldots, K$. Note that, in this re-coding, we are using the category 1 has the reference category.

Similarly to equation (17) (model with main effects only), we can write

$$Y_i = \beta_0 + \beta_1 x_{1i} + \sum_{l=2}^{K} \beta_{2l} x_{2li} + \epsilon_i \, . \tag{22}$$

Given the binary nature of the auxiliary variables x_{2li}, the model formulation is equivalent to

$$Y_i = \begin{cases} \beta_0 + \beta_1 x_{1i} + \epsilon_i, & \text{if } x^*_{22i} = 0, \ldots, x^*_{2Ki} = 0 \\ (\beta_0 + \beta_{22}) + \beta_1 x_{1i} + \epsilon_i, & \text{if } x^*_{22i} = 1 \\ \cdots & \cdots \\ (\beta_0 + \beta_{2K}) + \beta_1 x_{1i} + \epsilon_i, & \text{if } x^*_{2Ki} = 1 \end{cases} \tag{23}$$

As for the binary case of x_2, we have now a model that describes K straight lines with the same slope but intercepts with the origin varying with the categories of x_2 (Figure 5A). In particular, β_0 represents the intercept with the origin for the reference category 1 of x_2, while β_1 represents the slope shared by all categories of x_2. The parameters $\beta_{2l}, l =$

$2, \ldots, K$ can be viewed as the increment (or decrement) in the intercept associated with the reference category of x_2. Interestingly, the above model includes $K+1$ covariates and, therefore, we are a step closer to the general case of p covariates.

We can finally consider a model which the straight line between x_1 and Y changes not only in the intercept with the original, but also the slope with the categories of x_2 (Figure 5B). This model can be written as follows

$$Y_i = \beta_0 + \beta_1 x_{1i} + \sum_{l=2}^{K} \beta_{2l} x_{2li} + \sum_{l=2}^{K} \beta_{3l} x_{1i} x_{2li} + \epsilon_i \,. \qquad (24)$$

Alternatively, it can also be written

$$Y_i = \begin{cases} \beta_0 + \beta_1 x_{1i} + \epsilon_i, & \text{if } x_{2i} = 1 \\ (\beta_0 + \beta_{22}) + (\beta_1 + \beta_{32}) x_{1i} + \epsilon_i, & \text{if } x_{2i} = 2 \\ \ldots & \ldots \\ (\beta_0 + \beta_{2K}) + (\beta_1 + \beta_{3K}) x_{1i} + \epsilon_i, & \text{if } x_{2Ki} = K \end{cases} \qquad (25)$$

where it is clear how the intercepts with the origin and the slopes are determined by the model. In this scenario, the interpretation of the model parameter is the following. The parameters β_0 and β_1 represent the intercept with the origin and the slope associated with the reference category 1 of x_2. The parameters $\beta_{2l}, l = 2, \ldots, K$ can be interpreted as in the previous model. Finally, the parameters $\beta_{3l}, l = 2, \ldots, K$ encapsulate the increment (or decrement) in the slope associated with the other categories of x_2 when comparing to the slope of the reference category. More precisely, $\beta_{3l} < 0$ indicates the slope of the line in individuals with $x_2 = l$ is decreased when compared to hat happens with the reference category $x_2 = 1$. In contrast, $\beta_{3l} > 0$ indicates that the slope of the line for individuals with $x_2 = l$ is decreased when compared to the one of the reference category $x_2 = 1$.

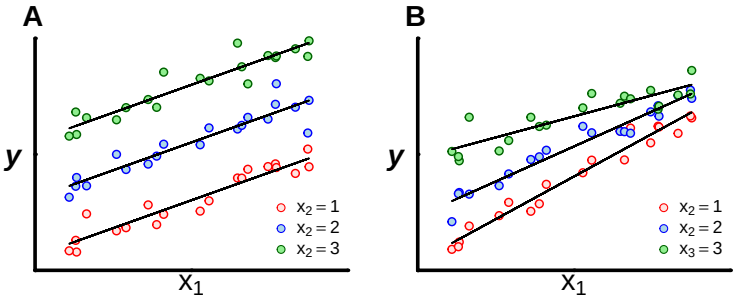

Figure 5: Two linear regression models including one quantitative covariate (x_1) and a categorical one (x_2) with three categories (1, 2, and 3) for describing a quantitative outcome Y : **(A)** model in which there are three parallel lines relating x_1 to Y with the same slope but different intercepts with the origin according to $x_2 = 1$, $x_2 = 2$, and $x_1 = 3$ (equation (22)); **(B)** model in which there are three lines relating x_1 to Y with different slopes and intercepts with the origin according to $x_2 = 1$, $x_2 = 2$, and $x_1 = 3$ (equation (24).

A cautionary note should be given at this point. In practice, one often defines the domain of the categorical covariate by integer values as assumed above. Such a definition opens the door to a possible but hopefully unlikely mistake. Given that the categorical covariate is defined by integer values, any statistical program accepts the definition of the linear regression model by equations (17) and (19). Therefore, it is extremely important to create a data dictionary[2] and consult it when needed.

2 A data dictionary is a supplementary document that details the information provided in a data set. Data dictionaries usually include the meaning and attributes of the contained variables as well as information about the creation, format, and usage of the data.

Before going to the case of two quantitative covariates, one might be wondering whether linear regression should be applicable to the case of having two categorical covariates. This situation can be easily modelled by

2.2 Case II: two quantitative covariates

The final case of two covariates is to consider them both quantitative. Again, we can construct one model with main effect only and another one that includes an interaction term additionally. From a mathematical point, we can use equations (17) (model with main effects only) and (19) as along as we see x_2 as a quantitative rather than a binary variable. As consequence, the models define regression planes (or response surfaces) in which we can see how the outcome Y varies with x_1 and x_2. Such planes can be represented by contour plots in which we can define isoclines in which the values of Y are held constant with the simultaneous variation of x_1 and x_2 (Figure 6). In the model with main effects only, these isoclines are parallel with each other (Figure 6A) similarly to the case of one quantitive covariate and a categorical one (Figure 5A). In the model with both main effects and interaction term, the isoclines become non-trivial (Figure 6B). Therefore, linear regression modelling is able to construct more complex relationships between the outcome and two quantitative covariates via the introduction of an interaction term.

2.3 General case: p covariates

The most general case of multiple regression is to consider a model in which there are $p > 2$ covariates of whatever

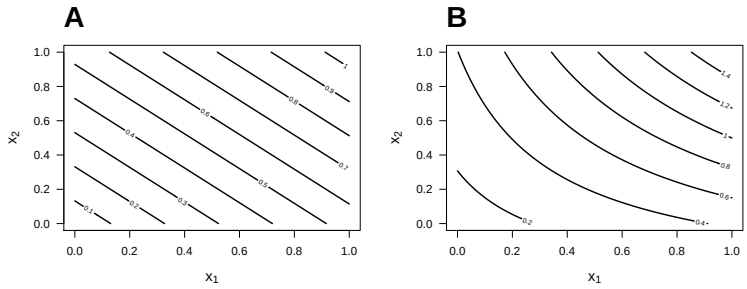

Figure 6: Examples of linear regression models with two quantitative covariates x_1 and x_2: **(A)** Contour plot of a model with main effects; **(B)** Contour plot of a model with main effects and an interaction term. In each plot, the solid lines represent the so-called isoclines in which the outcome is constant with the variation of x_1 and x_2.

nature (binary, categorical, or quantitative). In the case of having a mixture of binary and quantitative covariates only, the simplest model is to consider their main effects only, i.e.,

$$Y_i = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_p x_{pi} + \epsilon_i, i = 1, \ldots, n \ . \tag{26}$$

For mathematical (and also computational) convenience, the above equation can be written in the following matrix form

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \ , \tag{27}$$

where $\mathbf{Y} = \{Y_1, \ldots, Y_n\}$, $\mathbf{X}$ is a $n \times (p+1)$ matrix whose elements of the first column are equal to 1 and the rest are given by $x_{ij}, i = 1, \cdots, n$ and $j = 1, \ldots, p$, $\boldsymbol{\beta} = \{\beta_0, \beta_1, \ldots, \beta_p\}$, and $\boldsymbol{\epsilon} = \{\epsilon_1, \ldots, \epsilon_n\}$.

When the above model contemplates a mixture of p_c categorical covariates and p_q quantitative covariates ($p = p_c + p_q$), we should introduce as many dummy variables as categorical covariates in the equation. Therefore, the total

number of covariates is effectively $p^* = p_q + \sum_{i=1}^{p_c} k_c - p_c$ where k_i is the number of total categories of the i-th categorical covariate.

2.4 Guessing the unknowns

As done for estimating the simple linear regression model, we can use once again the least square method, which gives rise to the general minimisation problem

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \left\{ \sum_{i=1}^{n} \left(y_i - \beta_0 + \sum_{j=1}^{p} \beta_j x_i \right)^2 \right\} . \tag{28}$$

This problem can be written the following matrix form

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})^T (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) . \tag{29}$$

The respective solution is

$$\hat{\boldsymbol{\beta}} = \left(\mathbf{X}^T \mathbf{X} \right)^{-1} \left(\mathbf{X}^T \mathbf{Y} \right) . \tag{30}$$

A potential problem arises when $\mathbf{X}^T\mathbf{X}$ is not invertible[3] due to highly correlated covariates. This problem is usually referred to as multicollinearity. As a result, the ordinary least square estimates are deemed to be numerically unstable.

A way to overcome this problem is to make a prior assessment of the correlation between covariates and to remove those that are problematic. However, this task is not easy to perform when the number of covariates is large not only by the large number of correlations to evaluate but also by the

3 Mathematically, one says that the rank of this matrix is lower than the respective dimension.

different correlation patterns that emerge from the covariates. On the other hand, when there is a subset of covariates which are highly correlated with each other, one should ask which of these covariates should be picked up in order to maximise the sample information about the outcome. To address this question conveniently, we should discuss the problem of model selection and the use of penalised regression as a possible solution. This topic will be discussed in the following Chapter.

After estimating the model, we should consider to conduct hypothesis testing on the significance of each covariate. This testing is similar to the theoretical ground for drawing inferences over parameters in simple linear regression model. In particular, we are interested in testing the following pair of hypotheses

$$H_0 : \beta_j = 0 \text{ versus } H_1 : \beta_j \neq 0, j = 0, 1, \dots p \, . \tag{31}$$

The Wald's Score statistic is again used to measure the distance between what is expected under the null hypothesis and what we can obtain from the data directly, i.e.,

$$T = \frac{\hat{\beta}_j}{se\left(\hat{\beta}_j\right)} \underset{H_0}{\rightsquigarrow} t_{(n-p-1)} \, , \tag{32}$$

where large values of T indicate a marked deviation from the null hypothesis. Note that the distribution of T under H_0 is now a Student's t distribution with $n - p - 1$ degrees of freedom. The decision of the test follows the general rule based on p-value: if p-value $< \alpha$, accept H_0; otherwise do not accept H_0.

2.5 Model selection

The possibility of integrating multiple covariates in the same model gives rise to the basic question about how many of them should be used to describe a given outcome conveniently. This question is related to the problem of determining which model is the best one among all possible models that can be constructed with different subsets of covariates. These questions lie in the realm of variable selection or, in more fancy words, model selection.

In theory, a proper variable selection should contemplate the balance of four aspects:

1. good accuracy of the model predictions;
2. easy interpretation of the resulting model;
3. reduced chance of data overfitting;
4. reduced chance of having problems of multicollinearity.

In other words, the main objective of variable selection is to find the simplest possible model without redundant statistical information whose predictions are accurate but not too much. To achieve this goal, specially, when the number of covariates is large, we need to find (fast) automated procedures that can guide us throughout the process of variable selection.

A popular procedure to variable selection is known as the backward elimination strategy in which all covariates are added to an initial model and then they are successively dropped out according to a given measure for model comparison; this measure is usually defined by a term related to the goodness-of-fit of the model and an additional term that penalises the model by its statistical complexity measured

by the (effective) number of parameters (see for example, Akaike's information criterion or Bayesian information criterion). The procedure stops when the decrease in the quality of the model fit does not compensate the decrease in model complexity (and increase in model interpretation).

We can alternatively use a forward selection procedure. That is, we start the process with an empty model (i.e, without any covariates). We then add the most promising covariates to the model in an iterative manner. The procedure stops when the increase in model complexity does not compensate the increase in the corresponding goodness-of-fit to the data.

The forward selection and backward elimination strategies can be combined in the same procedure, the so-called stepwise regression. That is, we begin with forward selection but, after the inclusion of the second covariate, the algorithm tests at each step whether a covariate already included can be removed from the model without compromising the quality of the fit. This procedure and its forward/backward variants are available in most statistical softwares, such as the package `MASS` [5] for the R software.

Notwithstanding the popularity among users[4], the above variable selection procedures have been criticized on different grounds [6]. The most problematic aspect is that there is no warranty that only authentic covariates will be selected for the model [7]. Above all, even if all true covariates are included in the dataset, the final model selected by the above procedures might be far from the true one. The procedures also do not avoid the problem of multiple testing, thus, inflating the number of covariates included in the model

4 The use of stepwise regression ranged from 0.55% to 0.66% in papers published between 2009 and 2020 on biomedical, life and social sciences (66^{th} place in all statistical methods used in 2020) [1].

[8]. These problems motivated the research community to develop alternative approaches to model selection such as penalised regression, as discussed in the following Chapter.

3 Penalized Regression

The strength of the penalised regression arises as an ingenious solution to the problem of model selection. As we will see below, penalized regression combines estimation and variable selection in a single step, thus, avoiding the fitting of many different models during stepwise regression. The basic idea is to reduce (or shrink) some model coefficients or even to set them to zero. In this way, we can clearly identify the covariates that are not significant in the model. A prerequisite is that all the covariates should be standardised (with mean 0 and variance 1) beforehand, which might problematic when there are qualitative covariates to be included in the model.

In general, penalised regression can be formulated by the following maximisation problem

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\operatorname{argmax}} \left\{ \mathcal{L}(\boldsymbol{\beta}) - \lambda \times pen \right\} \ , \tag{33}$$

where $\boldsymbol{\beta} = \left\{ \beta_j, j = 0, \dots p \right\}$, $\mathcal{L}(\boldsymbol{\beta})$ is the log-likelihood function, $\lambda > 0$ is the regularisation or tuning parameter, and *pen* is as a penalty function that varies with the type of regularisation applied. Given that the maximum likelihood method and ordinary least squares provide the same estimates in multiple regression, the above equation can be alternatively recreated as follows

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}} \left\{ \sum_{i=1}^{n} \left(y_i - \beta_0 + \sum_{j=1}^{p} \beta_j x_i \right)^2 \right\} \ . \tag{34}$$

subject to the following constraint

$$pen \leq \lambda \, . \tag{35}$$

In this scenario, the popular Ridge, LASSO and Elastic Net regression can emerge as different penalty functions and tuning parameters. Note that, when $\lambda = 0$, the above equation converts to the classical ordinary least squares (or the maximum likelihood method). The tuning parameter λ is selected by a data-driven procedure such as cross-validation to maximise out-of-sample performance.

It is worth mentioning that penalised regression has been gaining traction in the scientific literature between 2009 and 2020 (Figure 7)[1]. The respective proponents have been also earning a stupendous number of citations since the publication of their original studies (Table 1). Such a trend suggests that this type of methodology is becoming more and more useful in this era of big data.

3.1 Ridge Regression

Ridge regression is the oldest penalised regression method dating back to 1970 with the seminal papers of Hoerl and Kennard [9, 10]. It defines the penalty function by the sum

1 However, LASSO, Ridge, and Elastic Net Regression are ranked on 91^{th}, 147^{th}, and 187^{th} places, respectively, among all statistical methods used in papers from biomedical, life, and social sciences published in 2020 [1]. These rankings are in contrast with the 66^{th} place for stepwise regression, despite all the theoretical problems already highlighted for this variable selection procedure.

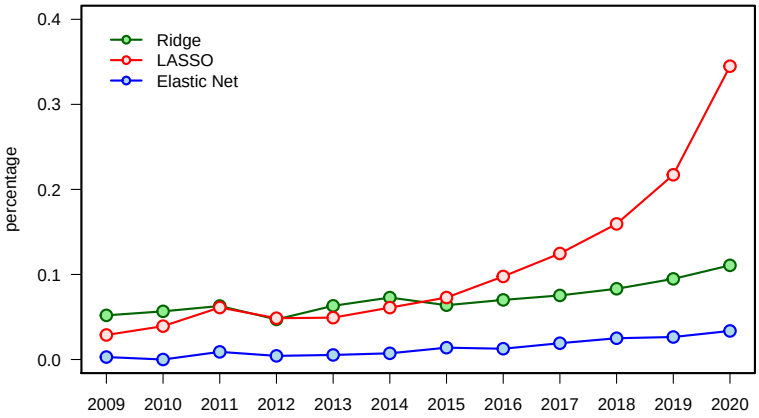

Figure 7: Percentages of scientific papers published between 2009 and 2020 citing the three most popular penalised regression methods [1].

Table 1: Bibliometrics (at the time of writing) of original papers proposing the three most popular penalised regression methods.

Method	**Number of Citations**	
	Google Scholar	Web of Knowledge
Ridge	13,360	5,793
LASSO	46,787	23,155
Elastic Net	16,652	8,592

of squared coefficients, which gives rise to the following minimization problem

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \left\{ \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_i \right)^2 \right\} , \tag{36}$$

subject to

$$\sum_{j=1}^{p} \beta_j^2 \leq \lambda_2 \text{ for some } \lambda_2 . \tag{37}$$

where λ_2 is the tuning parameter. Given the quadratic nature of the constraint, the amount of shrinkage is proportional to the size of each coefficient (according to the ordinary least squares). In other words, higher the coefficient greater the shrinkage. This implies that it is not possible to set some coefficients to zero. This theoretical impediment is the reason to say that, strictly speaking, Ridge regression is unable to perform variable selection, thus, not giving rise to simple and easy interpretation of the model[2].

The above minimisation problem generates the following solution for estimating $\boldsymbol{\beta}$

$$\hat{\boldsymbol{\beta}}^{ridge} = \left(\mathbf{X}^T \mathbf{X} + \lambda_2 \mathbf{I} \right)^{-1} \left(\mathbf{X}^T \mathbf{Y} \right) , \tag{38}$$

where $\mathbf{I}$ is the identity matrix with dimension p. Remember that, when $\lambda_2 = 0$, the above equation converts to the classical ordinary least solution for multiple linear regression (equation (30)).

As we discussed in previous Chapter, when there are highly correlated covariates, the matrix $\mathbf{X}^T\mathbf{X}$ might not be invertible (and therefore, it cannot be calculated). In this

2 Another disadvantage is that coefficient estimates under penalised regression are biased under a frequentist approach to statistical inference.

sense, the additional term $\lambda_2 \mathbf{I}$ implies that the Ridge regression is essentially an analytical tool that reduces the impact of multicollinearity in the respective model estimation.

The subsequent question is to know whether there is an optimal solution for the tuning parameter λ_2. A way to do is is to think about the accuracy (or predictive capacity) of the estimated model as a function of λ_2. In particular, accuracy can be determined the mean square error[3] defined as

$$\text{MSE} = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i(\lambda_2))^2}{n}\,, \tag{39}$$

where $\hat{y}_i(\lambda_2)$. In this scenario, one should choose the value of λ_2 that minimises the mean square error of the model. This minimisation problem can be done by estimating MSE for different values within a given interval where the solution should be located. The estimation of MSE is usually done via the traditional k-fold cross-validation.

To facilitate our search for this minimum, the tuning parameter can be alternative defined as the proportion of shrinkage, κ that should be applied to the ordinary least squares solution. In this way, the search can be focused on the interval $[0, 1]$ where 0 is no shrinkage and 1 is "complete" shrinkage. In practical, we search a grid of values for κ in which we estimate the MSE[4]. This procedure generates a plot of MSE as a function of κ (Figure 8A).

An important output of Ridge regression is also to understand the impact of κ in the coefficient estimates. This can be done by a plot called Ridge trace in which the different

3 Another possible measure is the predictive error as explained elsewhere [11].

4 This procedure is similar to the profile likelihood method where the likelihood maximisation is done successively by fixing one of the parameters at a given constant.

coefficient estimates are plotted against κ (Figure 8B). Therefore, this plot has as many lines represented as covariates in the model. As a consequence, this plot becomes hard to interpret when there are many covariates under analysis.

3.2 LASSO Regression

The problem of combining estimation with variable selection was eventually solved by the proposal of LASSO regression (from Least Absolute Shrinkage and Selection Operator) regression in 1996 [11]. In this type of regression, it is possible to shrink some coefficients and set others to zero. However, from a theoretical standpoint, LASSO might have some technical problems in generating valid standard errors for the coefficients set to zero.

Notwithstanding the increasing trend in the use of LASSO regression (Figure 7 and Table 1), the original publication did not immediately spark much interest in the research community. According to Tibshirani, the proponent of LASSO, the initial lack of interest on the methodology was simply due to six reasons [13]:

1. the slow computation back then;
2. its black-box nature;
3. no statistical motivation for the respective algorithm;
4. limited understanding of the statistical and numerical advantages of LASSO for dealing with sparse data;
5. small number of scientific problems where both n and p are large $n \ll p$;
6. inexistence of widespread statistical software (e.g., the R software) where new tools could be easily developed, shared, and optimised by the community.

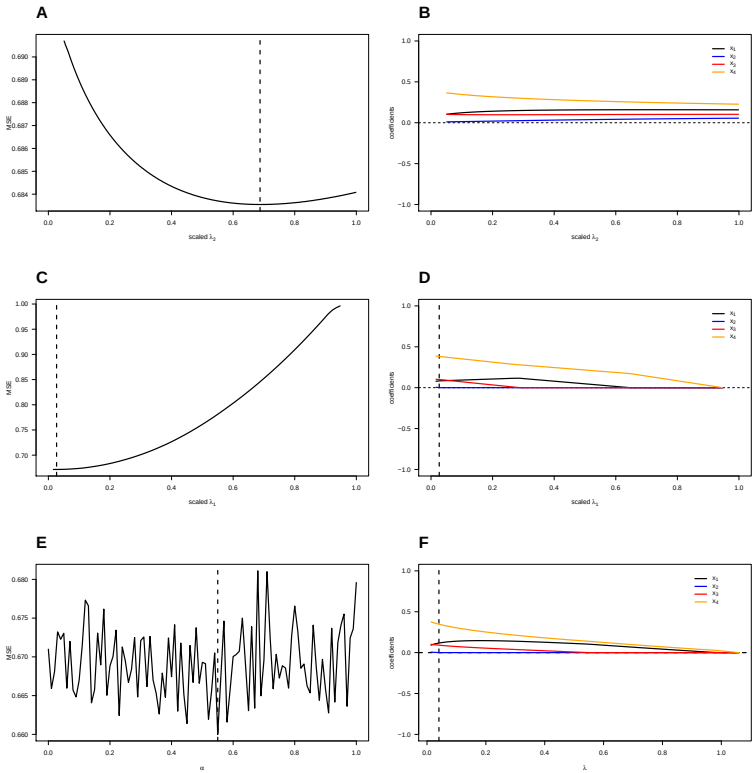

Figure 8: Examples of different penalised regression for the same simulated data concerning a quantitative outcome (Y) and four (correlated) covariates (x_1, x_2, x_3 and x_4). **(A)** Mean squared error (MSE) as a function of the scaled tuning parameter λ of Ridge regression; **(B)** Ridge coefficient estimates as a function of the scaled tuning parameter λ (i.e., Ridge Trace); **(C)** Mean squared error (MSE) as a function of the scaled tuning parameter λ of LASSO regression; **(D)** Coefficient estimates as a function of the scaled tuning parameter λ for LASSO regression; **(E)** Optimal mean squared error (MSE) as a function of the tuning α for Elastic-Net regression; **(F)** Coefficient estimates as a function of the scaled tuning parameter λ for Elastic-Net regression. In all plots, the vertical dashed lines represent the optimal parameter.

LASSO defines the penalty function by the sum of the absolute values of the coefficients (i.e., L1-norm). This function gives rise to the following minimisation problem

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \left\{ \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_i \right)^2 \right\} , \tag{40}$$

subject to

$$\sum_{j=1}^{p} \left| \beta_j \right| \leq \lambda_1 \text{ for some } \lambda_1 . \tag{41}$$

where λ_1 is the tuning parameter. The use of the L1-norm in the penalty implies that the same shrinkage will be applied to all coefficients (Figure 8D). This is in opposition to the Ridge regression in which the amount of shrinkage is proportional to the size of each coefficient (Figure 8B).

The estimation of LASSO regression follows similar steps to those described in Ridge Regression. First, it is more convenient to a scale tuning parameter related to the proportion of shrinkage, $\kappa \in [0, 1]$, that should be applied to the ordinary least squares solution. We then search a grid of values for κ in which we estimate the MSE and find its minimal value (Figure 8C). Again, for a given value of κ, the MSE is estimated via the k-fold cross-validation. The optimal tuning parameter is then the one that provides the minimum MSE or, in other words, the best model accuracy.

3.3 Elastic Net Regression

Elastic-net regression combines the advantages of both Ridge and LASSO regression [12]. To do that, the penalty term is

decomposed into a ridge component and a LASSO one. The respective maximisation problem is the following:

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\operatorname{argmax}} \left\{ \mathcal{L}(\boldsymbol{\beta}) - \lambda_1 \sum_{j=1}^{p} |\beta_j| - \lambda_2 \sum_{j=1}^{p} \beta_j^2 \right\} . \tag{42}$$

Again, this problem can be seen as a penalised version of the ordinary least square method

$$\hat{\beta} = \underset{\beta}{\operatorname{argmax}} |Y - X\beta|^2 , \tag{43}$$

subject to

$$\alpha|\beta|_1 + (1-\alpha)|\beta|^2 \leq \lambda \text{ for some } \lambda . \tag{44}$$

where $\alpha \in [0,1]$. Note that when $\alpha = 0$ and 1, we obtain Ridge or LASSO Regression, respectively. When $\alpha \in (0,1)$, the solution is a mixture between Ridge and LASSO regression. In this line of thought, the Elastic Net regression retains the advantages of both Ridge and LASSO regression.

Given the Elastic Net regression is expressed in terms of two tuning parameters (α and λ), the above estimation procedure for Ridge or LASSO needs to be adapted. We define a grid of values for α. For each value of α, we estimate the MSE via cross-validation for all possible values λ and find its minimal value. We then construct a graph where where the optimal MSE is plotted against each value of α (Figure 8E). Again, the best solution is the value of α that provides the minimum among all the optimal MSE. Finally, we construct a plot of MSE as a function of λ to determine the best λ (Figure 8F).

A very popular application of Elastic Net regression was the problem of estimating the biological age. The basic idea is to construct an elastic-net regression model that could be

used as a predictor of the biological data via DNA methylation demonstrated by the hallmark studies of Horvath [14] and Hannun et al. [15].

3.4 Practical recommendations

For the application of penalised regression, the covariates are typically required to be standardised (i.e., with mean zero and unit variance). Given that this pre-requisite, penalised regression is more adequate to be used when there are only quantitative covariates where data standardisation can be operated.

LASSO, Ridge and Elastic-Net approach for linear regression can be performed via the package `glmnet` for the R software [16]. The same package allows the application of these penalises approaches to other regression models such as the logistic and Multinomial regression, which are members of the generalized linear models.

Another package is called `caret` (short for Classification And REgression Training) [17]. The package comprises a set of functions that attempt to streamline the process for creating predictive models, including data splitting, preprocessing, feature selection, model, tuning using resampling, and variable importance estimation.

Bibliography

[1] Bolt T, Nomi JS, Bzdok D, Uddin LQ (2021). Educating the future generation of researchers: A cross-disciplinary survey of trends in analysis methods. PLoS Biol 19(7): e3001313.

[2] Box GEP (1976). Science and statistics. J Am Stat Assoc 71(356): 791–799.

[3] Wasserstein RL, Lazar NA (2016). The ASA Statement on p-Values: Context, Process, and Purpose. Am Stat 70:129–133.

[4] Asar O, Ilk O, Dag O (2017). Estimating Box-Cox Power Transformation Parameter Via Goodness-of-Fit Tests. Commun Stat - Simul Comput 46(1):91–105.

[5] Venables WN, Ripley BD (2002) Modern Applied Statistics with S. Fourth Edition, Springer, New York.

[6] Whittingham MJ, Stephens PA, Bradbury RB, Freckleton RP (2006). Why do we still use stepwise modelling in ecology and behaviour?. J Anim Ecol, 75(5), 1182–1189.

[7] Derksen S, Keselman HJ (1992). Backward, forward and stepwise automated subset selection algorithms: Frequency of obtaining authentic and noise variables. Br J Math Stat Psychol 45:265-282.

[8] Mundry R, Nunn CL (2009). Stepwise model fitting and statistical inference: turning noise into signal pollution. Am Nat 173(1):119–123.

[9] Hoerl AE, Kennard RW (1970). Ridge Regression: Biased Estimation for Nonorthogonal Problems. Technometrics 12(1): 55-–67.

[10] Hoerl AE, Kennard RW (1970). Ridge Regression: Applications to Nonorthogonal Problems. Technometrics, 12(1): 69—82.

[11] Tibshirani R (1996). Regression Shrinkage and Selection via the lasso. J Royal Stat Soc B 58(1):267-–88.

[12] Zou H, Hastie T (2005). Regularization and variable selection via the elastic net. J Royal Stat Soc B 67:301–320.

[13] Tibshirani R (2011). Regression shrinkage and selection via the lasso: a retrospective. J Royal Stat Soc B 73: 273–282.

[14] Horvath S (2013). DNA methylation age of human tissues and cell types. Genome Biol. 14(10):R115.

[15] Hannum G, Guinney J, Zhao L, Zhang L, Hughes G, Sadda S, Klotzle B, Bibikova M, Fan JB, Gao Y, Deconde R, Chen M, Rajapakse I, Friend S, Ideker T, Zhang K (2013). Genome-wide methylation profiles reveal quantitative views of human aging rates. Molecular cell 49(2):359-–367.

[16] Friedman J, Hastie T, Tibshirani R (2010). Regularization Paths for Generalized Linear Models via Coordinate Descent. J Stat Softw, 33(1), 1-–22.

[17] Kuhn M (2008). Building Predictive Models in R Using the caret Package. J Stat Softw, 28(5), 1—26.