1 Introduction

In this chapter, we introduce the concept of regression, which is a way to quantify the relationship between an outcome Y and a vector of features X. This relationship is expressed by the regression function, which is the mean of Y given X. This chapter discusses the main ideas and goals of regression analysis.

Regression is a broad term that refers to a collection of data analysis tools for relating an *outcome* Y and a vector of d *features* $X = (X_1, \ldots, X_d)$ also called *covariates*. Figure 1.1 shows data involving a single feature.

The data in Figure 1.1 come from Broadbent et al. (2004), which are also considered in later examples. The authors were relating the number of lesions X in the hippocampus (a part of the brain) and the score Y on a memory task in mice. We see that more lesions are associated with a lower memory score. The line on the plot is an estimate of the regression function, defined as follows:

$$\mu(x) = \mathbb{E}[Y|X = x] = \int y\, p(y|x)\, dy,$$

where $p(y|x) = p(x, y)/p(x)$ is the density of Y given X.

We can also write

$$Y = \mu(X) + \epsilon.$$

where $\mathbb{E}[\epsilon|X] = 0$. The function $\mu(x)$ can be a line, a quadratic, or any function of x. The line in Figure 1.1 is an estimate of $\mu(x)$. The method for computing this line is explained in Chapter 2. There are three reasons why we want to estimate the regression function:

1. Description: The regression function provides an interpretable summary of the relationship between X and Y.
2. Prediction: Given a new observation X, the regression function gives us a way to predict Y.
3. Causation: Regression, if used properly, allows one to estimate how Y would change if we *intervened* and changed X.

The third goal, *causal inference*, requires strong assumptions and a set of methods that are different from the methods used for the first two goals. There is often confusion about these three goals. We will deal with causal inference in Chapter 11. Until then, we focus on description and prediction.

https://doi.org/10.1017/9781009702799.002 Published online by Cambridge University Press

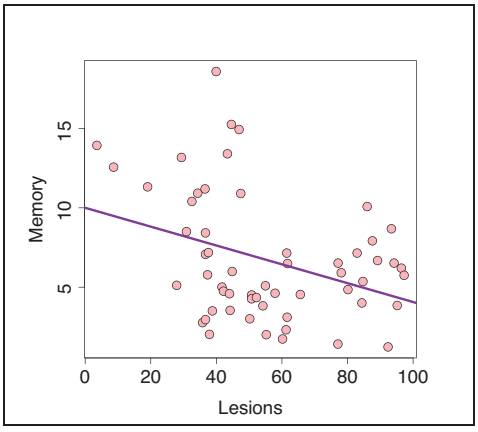


Figure 1.1 A scatterplot of data $(X_1, Y_1), \ldots, (X_n, Y_n)$. The line is a regression function which estimates the mean of Y given X.

Now we explain the second goal, *prediction*, in more detail. Suppose we observe X and we want to predict Y. Let $f(X)$ denote our prediction. The squared prediction error or *prediction risk* is defined by

$$R(f) = \mathbb{E}[(Y - f(X))^2] = \iint (y - f(x))^2 \, p(x, y) \, dx \, dy, \tag{1.1}$$

where $p(x, y)$ is the joint density of X and Y. The following result shows that the best choice for $f(x)$ is $\mu(x)$.

Theorem 1.1 *The prediction risk in* (1.1) *is minimized by choosing* $f(x) = \mu(x)$. *In other words, for any function* $f(x)$, $\mathbb{E}[(Y - \mu(X))^2] \leq \mathbb{E}[(Y - f(X))^2]$.

In practice, we don't know the distribution $p(x, y)$ so we can't compute $\mu(x)$. Instead, we must estimate $\mu(x)$ from data $(X_1, Y_1), \ldots, (X_n, Y_n)$ drawn from $p(x, y)$. We explain how to do this in later chapters.

Book Outline. Throughout the book, each chapter has an appendix that contains most proofs. Advanced optional sections are marked with an asterisk.

Chapter 2 introduces *linear regression*, where $\mu(x)$ is assumed to have the form

$$\mu(x) = \beta_0 + \beta_1 x_1 + \cdots + \beta_d x_d.$$

In this case, estimating $\mu(x)$ requires estimating the vector of parameters $\beta = (\beta_0, \ldots, \beta_d)$. Once we have an estimate $\widehat{\beta}$, our estimate of μ is

$$\widehat{\mu}(x) = \widehat{\beta}_0 + \widehat{\beta}_1 x_1 + \cdots + \widehat{\beta}_d x_d.$$

Given a new X, we predict Y by $\widehat{\mu}(X)$. As we explain in Chapter 2, it is unlikely that $\mu(x)$ would ever be linear. Instead, we should think of $\beta_0 + \beta_1 x_1 + \cdots + \beta_d x_d$ simply as a linear approximation to $\mu(x)$. We shall see that a good estimator of β is the *least squares estimator*, which is the vector $\widehat{\beta}$ that minimizes the average sums of squares

https://doi.org/10.1017/9781009702799.002 Published online by Cambridge University Press

$$\frac{1}{n}\sum_i (Y_i - [\beta_0 + \beta_1 X_{i1} + \cdots + \beta_d X_{id}])^2.$$

In Chapter 3, we consider estimating the prediction risk after estimating $\widehat{\beta}$, namely,

$$R = \mathbb{E}[(Y - [\widehat{\beta}_0 + \widehat{\beta}_1 X_1 + \cdots + \widehat{\beta}_d X_d])^2].$$

If the number of features d is large, the least squares estimator $\widehat{\beta}$ is not reliable. In these cases, we turn to other estimators described in Chapter 4.

In Chapter 5, we consider the case where Y is discrete. We discuss logistic regression, Poisson regression, and also generalized linear models.

In Chapter 6, we consider nonparametric regression, where we no longer assume that the regression function is linear. This includes kernel regression, local polynomial regression, nearest neighbor regression, splines, and reproducing kernel Hilbert space regression.

In Chapter 7, we turn to nonparametric regression when there is more than one feature. We cover kernel regression, local linear regression, additive models, trees, and random forests.

In Chapter 8, we consider *quantile regression*, where we estimate a quantile of Y given $X = x$. For example, we can estimate the median of Y given $X = x$.

In Chapter 9, we study classification. This is the case where we want to predict a variable Y that is discrete. The topics include linear classifiers, nonlinear classifiers, cross-validation, and receiver operating characteristic curves.

Chapter 10 considers prediction sets and conformal inference.

Up to this point, the focus is mainly on estimation and prediction. In Chapter 11, we turn to *causal inference* where we want to estimate the causal effect of a variable X on the outcome Y. This is very different than the prediction, and the difference creates much confusion. Prediction asks: what is my best guess of Y after observing X? Causation asks: what is my best guess of Y if I intervene and set X to a particular value? Answering such questions requires a very different set of tools.

This is the subject of Section 12.1. Chapter 12 discusses several extra topics that we only treat briefly. The topics include neural nets, deep learning, survival analysis, graphical models, and time series.

The book ends with two appendices, one on matrix theory (Appendix A) and the other on basic probability and statistics (Appendix B).

Book Resources. All the dataset mentioned in the book are available on the book's website (www.stat.cmu.edu/~larry/All_of_Regression).

Appendix: Proofs

Theorem 1.1 The prediction risk in (1.1) is minimized by choosing $f(x) = \mu(x)$. In other words, for any function $f(x)$, $\mathbb{E}[(Y - \mu(X))^2] \leq \mathbb{E}[(Y - f(X))^2]$.

https://doi.org/10.1017/9781009702799.002 Published online by Cambridge University Press

Proof. We show that for any function $f(x)$ we have that $R(\mu) \leq R(f)$. Note that

$$\begin{aligned} R(\mu) &= \iint (y-\mu(x))^2 p(x,y)\,dx\,dy \\ &= \iint (y-f(x)+f(x)-\mu(x))^2 p(x,y)\,dx\,dy \\ &= \iint (y-f(x))^2 p(x,y)\,dx\,dy + 2\iint (y-f(x))(f(x)-\mu(x))p(x,y)\,dx\,dy \\ &\quad + \iint (f(x)-\mu(x))^2 p(x,y)\,dx\,dy \\ &= R(f) + 2\iint (y-f(x))(f(x)-\mu(x))p(x,y)\,dx\,dy \\ &\quad + \iint (f(x)-\mu(x))^2 p(x,y)\,dx\,dy. \end{aligned}$$

Now, since $\int y\,p(y|x)\,dy = \mu(x)$,

$$\begin{aligned} \iint (y-f(x))(f(x)-\mu(x))p(x,y)\,dx\,dy & \\ &= \int \left\{ \int (y-f(x))(f(x)-\mu(x))p(y|x)\,dy \right\} p(x)\,dx \\ &= \int (\mu(x)-f(x))(f(x)-\mu(x))\,p(x)\,dx \\ &= -\int (f(x)-\mu(x))^2 p(x)\,dx. \end{aligned}$$

Hence,

$$R(\mu) = R(f) - \int (f(x)-\mu(x))^2 p(x)\,dx \leq R(f).$$

□

Exercises

1. Suppose that Y is a random variable with density $p(y)$. Assume that $p(y) > 0$ for all y. You have to predict Y. Let m be your prediction. We want to choose m to minimize
$$R(m) = \mathbb{E}(|Y-m|) = \int |y-m|\,p(y)dy.$$
Show that the best choice is to take m to be the median, that is, $P(Y < m) = P(Y > m) = 1/2$.
2. Recall that, if $\widehat{\theta}$ is an estimate of a parameter θ, then the *bias* is defined to be $\mathbb{E}[\widehat{\theta}] - \theta$. The mean squared error (MSE) is
$$\text{MSE} = \mathbb{E}[(\widehat{\theta}-\theta)^2].$$
Show that
$$\text{MSE} = \text{bias}^2 + \mathbb{V}[\widehat{\theta}].$$

https://doi.org/10.1017/9781009702799.002 Published online by Cambridge University Press

3. Let $Y_1, \ldots, Y_n \sim N(\mu, 1)$. Let $\overline{Y}_n = n^{-1} \sum_{i=1}^n Y_i$. Show that $\overline{Y}_n^2$ is a biased estimate of μ^2. Find an unbiased estimate.
4. Let $Y_1, \ldots, Y_n \sim N(\mu, \sigma^2)$. Suppose that σ^2 is known but μ is unknown.
   (a) Let $C = \overline{Y}_n \pm z_{\alpha/2}\, \sigma/\sqrt{n}$. Show that C is a $1-\alpha$ confidence interval for μ. In other words, show that $P(\mu \in C) = 1 - \alpha$.
   (b) Let $Y_{n+1} \sim N(\mu, \sigma^2)$ be a new observation. We want to construct a $1-\alpha$ prediction interval for Y_{n+1}. Let $C = [\overline{Y} - c, \overline{Y} + c]$. Find c such that $P(Y_{n+1} \in C) = 1 - \alpha$.
5. Suppose that $X_1, X_2, \ldots, X_m$ are random variables and let $a_1, \ldots, a_m$ be constants. Let $Z = \sum_{i=1}^m a_i X_i$. Show that

$$\mathbb{V}[Z] = \sum_{i=1}^m \mathbb{V}[X_i] + \sum_{i \neq j} \mathbb{C}[X_i, X_j].$$

6. Let

$$Y = \beta X + \epsilon,$$

   where $\epsilon \sim N(0, 1)$ and $X \sim N(0, 1)$. Assume that X and ϵ are independent.
   (a) Find the mean and variance of Y.
   (b) Find $\mathbb{E}[Y^2]$.
   (c) Find $\mathbb{E}[Y|X = x]$.
   (d) Find $\mathbb{E}[Y^3]$.
   (e) Find $\mathbb{C}(\epsilon, \epsilon^2)$. Are ϵ and ϵ^2 independent?
7. Suppose that $Y_1, Y_2, \ldots Y_n$ are independent Gaussian random variables, all with the same mean μ and variance σ^2. Let $\overline{Y}_n = n^{-1} \sum_{i=1}^n Y_i$.
   (a) Show that $\overline{Y}_n \sim N(\mu, \sigma^2/n)$. (You may use the fact that a sum of Gaussian random variables is Gaussian.)
   (b) What is the distribution of $W_i = (Y_i - \mu)/\sigma$?
   (c) What is the distribution of $\overline{W}_n$?
   (d) What is the mean and variance of $\sum_{i=1}^n W_i^2$?

https://doi.org/10.1017/9781009702799.002 Published online by Cambridge University Press