Estimating Regression Model Parameters: OLS and MLE

Ashley I Naimi

Spring 2022

Contents

1 Introduction

Regression is a cornerstone tool of any empirical analysis. It is arguably the most widely used tool in science. Regression models are often deployed for exploratory data analysis, to understand cause-effect relations between exposures and outcomes of interest, or to obtain predictions for an outcome of interest.

2 Least Squares Estimation

Consider a scenario in which we are interested in regressing an outcome Y against a set of covariates X. These covariates can be an exposure combined with a set of confounders needed for identification, or a set of predictors used to create a prediction algorithm via regression. In its most basic formulation, a regression model can be written as:

Note that, in this formulation, the function of interest on the left hand side of the equation is the conditional mean function, $E(Y \mid X)$. However, there are other options, including hazards, failure times, distribution quantiles, and many more.

$$E(Y \mid X) = f(X)$$

In principle, this model is most flexible in that it states that the conditional mean of Y is simply a *arbitrary function* of the covariates X. We are not stating (or assuming) precisely **how** the conditional mean is related to these covariates. Using this model, we might get predictions from to facilitate a decision making process, or obtain a contrast of expected means between two groups.

Because of the flexibility of this model, we may be interested in fitting it to a given dataset. But we can't. There is simply not enough information in this equation for us to quantify $f(X)$, even if we had all the data we could use. In addition to data, we need some "traction" or "leverage" to be able to quantify the function of interest.

The earliest attempt to find some "traction" to quantify $f(X)$ was proposed in the 1800s (Stigler, 1981, Shalizi (2019)). The approach starts by accepting a few tenets. First, we want the difference between the observed Y for any given individual and the fitted values $f(X)$ for that person to be "small." We also need a way to handle errors on both sides of the fitted values, so that two equal and opposite errors don't cancel out suggesting "zero" error. To address this, we can square the error and take it's average: $E[(Y - f(X))^2]$. Thus, we can define the "optimal" $f(X)$ as the function of X that minimizes the mean

Mean squared error can be be re-written as the sum of the squared bias and the variance:

$$E[(Y-f(X))^2] = [E(Y)-f(X)]^2+Var(Y),$$

which gives some insight as to what we are doing when we minimize mean squared error. In effect, we are finding the tradeoff between bias and variance for the model $f(X)$ given the data X.

squared error.

Recall from calculus that finding the $f(X)$ that minimizes mean squared error can be achieved by taking the derivative of the mean squared error with respect to $f(X)$, setting it to zero, then solving for $f(X)$.

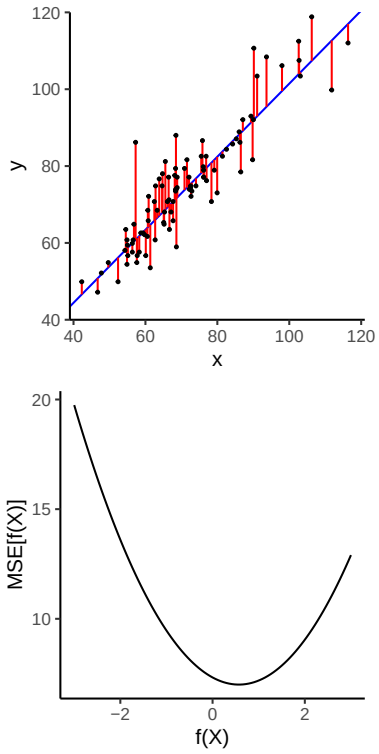


Figure 1: Line of 'best fit' (blue line) defined on the basis of minimizing the sum of squared residuals (red lines) displayed in the top panel; Partial representation of the mean squared error as a function of f(X) in the bottom panel. The lowest mean squared error value corresponds to the function f(X) represented by the blue line of 'best fit' in the top panel.

We've made some progress, but without a better sense of what $f(X)$ looks like, we still can't move forward. For example, there are several functions where either the derivative simply does not exist (e.g., if $f(X)$ is discontinuous), or where the derivative is still complex enough that we can't make progress with finding a unique solution for $f(X)$ that minimizes mean squared error (see technical note on nonlinear models).

Early on, it was recognized that if we select $f(X)$ to be *linear* (more technically, *affine*) the problem of finding the optimal $f(X)$ becomes much easier. That is, if we can simplify $f(X) = b_0 + b_1X$, then we can use calculus and simple algebra to find an optimal *linear* solution set $b_0 = \beta_0, b_1 = \beta_1$ that minimizes MSE.

Specifically, we can re-write the mean squared error as a function of $b_0 + b_1X$ to be $MSE(b_0, b_1) = E[(Y - (b_0 + b_1X))^2]$. Taking the partial

derivatives of $MSE(b_0, b_1)$ with respect to b_0 and b_1 gives us the ordinary least squares estimator for the coefficients in the model (Rencher, 2000, Shalizi (2019)):

$$\hat{b}_1 = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{x})}{\sum_i (x_i - \bar{x})^2}$$
$$\hat{b}_0 = \bar{y} - \hat{b}_1 \bar{x}$$

We can see these equations in action in an actual dataset. Let's conduct a simple exploratory analysis of the relation between weight (in kg) in 1982 and in 1971 among observations in the NHEFS data:

```
# locate the data on the website
file_loc <- url("https://bit.ly/47ECRcs")

# load the data
nhefs <- read_csv(file_loc) %>%
  dplyr::select(wt82, wt71) %>%
  na.omit(.)

# construct the numerator and denomiantor of the OLS estimator for b1
num <- sum((nhefs$wt71 - mean(nhefs$wt71))*(nhefs$wt82 - mean(nhefs$wt71)))
den <- sum((nhefs$wt71 - mean(nhefs$wt71))^2)

b1 <- num/den

# use the estimate of b1 to compute b0
b0 <- mean(nhefs$wt82) - b1*mean(nhefs$wt71)

# compare the b1 and b0 estimates above to what we get using OLS via
# the lm function in R
mod <- lm(wt82 ~ wt71, data=nhefs)

b0
```

```
## [1] 8.007219
```

```
b1
```

```
## [1] 0.9242009
```

```
summary(mod)$coefficients
```

```
##              Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) 8.0072185 0.93252917  8.586561 2.129193e-17
## wt71        0.9242009 0.01286839 71.819465 0.000000e+00
```

Technical Note:

Technically (almost to the point of pedantry), a nonlinear model is a model where the first derivative of the expectation taken with respect to the parameters is itself a function of other parameters (Seber and Wild, 1989). For example,

$$E(Y \mid X) = \beta_0 + \frac{X}{\beta_1}$$

is a nonlinear model, because it's first derivative taken with respect to β_1 is still a function of β_1. Recalling that the derivative of $\frac{1}{X} = \frac{1}{X^2}$:

$$\frac{dE(Y \mid X)}{d\beta_1} = \frac{d\left(\beta_0 + \frac{X}{\beta_1}\right)}{d\beta_1} = -\frac{X}{\beta_1^2}$$

The resolution to this specific example of nonlinearity is simple. We can define $\alpha = \frac{1}{\beta_1}$, and fit a linear model:

$$E(Y \mid X) = \beta_0 + \alpha X$$

and then obtain an estimate of β_1 as $\frac{1}{\hat{\alpha}}$.

However, this simple nonlinear model illustrates an important technical distinction between linear and nonlinear models. Why is this important? Solutions to these regression equations (which serve as our estimates), are obtained by finding where the slope of the tangent line of the parameters is zero. To do this, we need to set the first derivative of these regression equations to zero. But if there are still parameters in these first derivative equations, then there will not be a unique solution to the equation, and finding an estimate will require more complex approaches. This is the complication introduced by nonlinear models.

On the other hand, curvilinear models are linear models whose relationships can't be adequately captured by straight lines. These are easy to find solutions for, since their first derivatives are not functions of parameters. For instance, for a quadratic model such as:

$$E(Y \mid X) = \beta_0 + \beta_1 X + \beta_2 X^2$$

The first derivatives taken with respect to each parameter turn out to be:

$$\frac{dE(Y \mid X, C)}{d\beta_0} = 1$$

$$\frac{dE(Y \mid X, C)}{d\beta_1} = X$$

$$\frac{dE(Y \mid X, C)}{d\beta_2} = X^2$$

Thus, even though the regression function will not be a "straight line" on a plot, this model is still linear.

3 Maximum Likelihood Estimation

3.1 Basic Illustration

Maximum likelihood estimation is another more general technique that we can use to fit a model $f(X)$ to data. To illustrate, let's start with a simpler example where we want to estimate the 10 day risk of diarrhea ($Y \in [0, 1]$) among 30 infants infected with *Vibrio cholerae* who are also being breastfed. The scientific question of interest is the role that antibiotic concentrations in breastmilk ($X \in [0 = \text{low}, 1 = \text{high}]$) can play in reducing the incidence of diarrheal disease. The data are in Table 1 (taken from Cole et al., 2013):

Antibiotic Level	Cases ($Y = 1$)	non-Cases ($Y = 0$)	Total
Low ($X = 0$)	12	2	14
High ($X = 1$)	7	9	16
Total	19	11	30

If we assume that diarrhea was an independent occurrence across the 30 infants, we can define its probability mass function as:

$$P(Y = y) = \binom{n}{y} p^y (1-p)^{n-y} \tag{1}$$

In a direct problem, the known portion p is held fixed, and when we change the data inputs we obtain different probabilities of seeing y cases of diarrhea out of n observations.

In this case, p is the probability of observing a single diarrheal event. For the moment, let's ignore the fact that we have exposed and unexposed infants in the cohort that this equation is meant to represent. We are working in an indirect problem setting (we have the data, we don't know the underlying probability p). This will enable us to re-write equation 1 as:

$$P(Y = y) = \binom{30}{19} p^{19} (1-p)^{30-19}$$

which still doesn't help us, because all we have is data. However, if we guessed that $p = 0.6$, we could then get a predicted outcome:

$$P(Y = y; p = 0.6) = 0.14 = \binom{30}{19} 0.6^{19} (1-0.6)^{30-19}$$

This predicted value of 0.14 is no longer a probability, because the variable in the original equation is p, and not (as would usually be the case) the data

(i.e., the *variables*). Instead, we call this the **likelihood of the parameter** p. We can try another guess at the parameter, say $p = 0.4$:

$$P(Y = y; p = 0.4) = 0.005 = \binom{30}{19} 0.4^{19}(1 - 0.4)^{30-19}$$

Here, the likelihood of this parameter is much lower than the first. What does this suggest?: if the equation governing the data generating mechanism is correct, then the data are **more compatible** with $p = 0.6$ than they are with $p = 0.4$. Thus, it is more likely that the true p is closer to 0.6 than it is to 0.6, given the data we have.

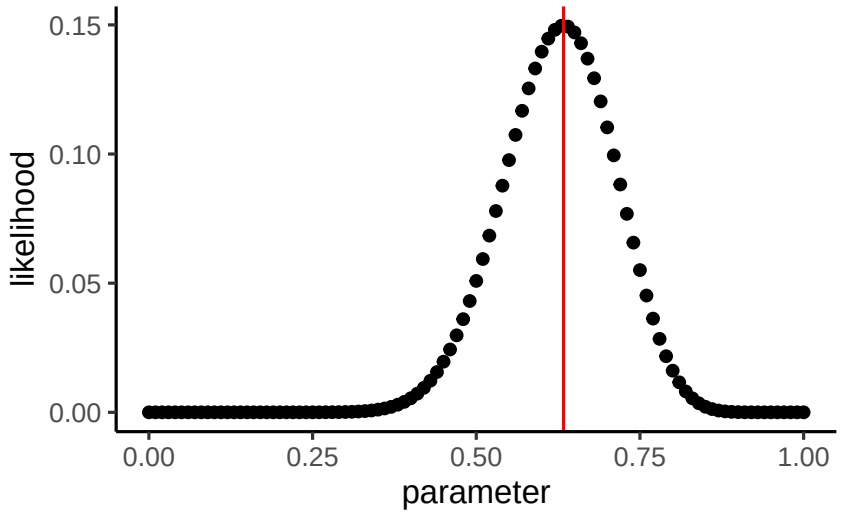
Let's look at this likelihood function a bit more systematically:

```
likelihood_res <- NULL
for (i in seq(0,1,.01)){
  likelihood_res <- rbind(
    likelihood_res,
    cbind(choose(30,19)*(i^19)*((1 - i)^(30 - 19)), i)
  )
}

likelihood_res <- data.frame(likelihood_res)

names(likelihood_res) <- c("likelihood", "parameter")

ggplot(likelihood_res) +
  geom_point(aes(x = parameter,
                 y = likelihood)) +
  geom_vline(xintercept = 19/30, color = "red")
```



This likelihood function tells us that the most likely parameter value is 0.633.

The key takeaway is that maximum likelihood estimation relies on a distribution function (probability mass function or probability density function) for the data generating mechanism. This distribution function is usually used in the context of direct problems to compute the probability of an event or set of events under a known set of parameters defining the distribution:

$$f(y;\theta)$$

However, in maximum likelihood estimation, this relationship is flipped. The likelihood function usually looks exactly the same as the distribution function, but it's used differently. Here, we treat the data as fixed, and try to find the parameters with the highest likelihood given the data:

$$L(\theta;y)$$

MLE for Regression

How might this work if we are interested in understanding the role that antibiotic concentrations in breastmilk play in the incidence of diarrhea?

Like in our OLS example, if we assume independence across observations, we can define the probability mass function for observing y cases of diarrhea among those with $X = x$ as:

$$P(Y = y \mid X = x) = \binom{n_x}{y_x} p_x^{y_x} (1 - p_x)^{n - y_x}$$

where (again) $X \in [0, 1]$. However, this time we can define $p_x = \text{expit}[\beta_0 + beta_1 x]$ where $\text{expit}(a) = \frac{1}{[1+\exp(-a)]}$. Now that we've made p_x a function of β_0 and $beta_1$, we can define the likelihood function here as:

$$L(\beta_0, \beta_1; y_x) = \prod_{x=0,1} \binom{n_x}{y_x} p_x^{y_x} (1 - p_x)^{n_x - y_x}$$

Once again, the task is to find values of β_0 and $beta_1$ that maximize this likelihood function, and that are thus most compatible with the data. We could plug different values in and evaluate the likelihood as we did above. However, in practice, we would once again use calculus to find where the slope of the likelihood function is equaled to zero.

To make the math easier, we often simplify the likelihood function by ignoring factors like $\binom{n}{y}$, since the derivative of the likelihood function will not depend on this scaling factor. Furthermore, if we take the log of the likelihood function, computing derivatives is much simpler:

$$\mathcal{L}(\beta_0, \beta_1; y_x) = \ln[L(\beta_0, \beta_1; y_x)] = \prod_{x=0,1} y_x \ln p_x + (n_x - y_x) \ln(1 - p_x)$$

The (partial) derivatives of this log-likelihood function with respect to the parameters is often referred to as the score function, the gradient, or (less commonly) the informant. If we set the score function for each parameter to zero and solve for the parameter values, we get the score equations for β_0 and β_1, which are our maximum likelihood estimators. In this simple example, solutions for the score equations are easy to compute, and become (Cole et al., 2013):

$$\hat{\beta}_0 = \ln\left(\frac{y_0}{n_0 - y_0}\right) = \ln(7/9) = -0.25$$

$$\hat{\beta}_1 = \ln\left[\frac{y_1(n_0 - y_0)}{y_0(n_1 - y_1)}\right] = \ln\left[\frac{12 \times (16 - 7)}{7 \times (14 - 12)}\right] = 2.04$$

We can again compare these to what we would get from an actual regression analysis of these data:

```
d <- data.frame(
  y = c(1, 1, 0, 0),
```

```
  x = c(1, 0, 1, 0),
  freq = c(12, 7, 2, 9)
)

mod_glm <- glm(y ~ x, data = d, weights = freq, family = binomial("logit"))

summary(mod_glm)$coefficients
```

```
##               Estimate Std. Error    z value   Pr(>|z|)
## (Intercept) -0.2513144  0.5039526 -0.4986866 0.61800018
## x            2.0430739  0.9150083  2.2328474 0.02555901
```

References

Stephen R. Cole, David B. Richardson, Haitao Chu, and Ashley I. Naimi. Analysis of occupational asbestos exposure and lung cancer mortality using the g formula. *Am J Epidemiol*, 177(9):989–996, 2013.

Alvin C. Rencher. *Linear Models in Statistics*. Wiley, New York, 2000.

G. A. F. Seber and C. J. Wild. *Nonlinear regression*. Wiley, New York, 1989.

Cosma Rohilla Shalizi. *The Truth About Linear Regression*. https://www.stat.cmu.edu/ cshalizi/TALR/TALR.pdf, 2019.

Stephen M. Stigler. Gauss and the invention of least squares. *The Annals of Statistics*, 9(3):465–474, 1981.