Maximum Likelihood Estimation

Professor V.M. Auken, PhD

Simple Linear Regression Analysis

- We consider the modeling between the dependent and one independent variable. When there is only one independent variable in the linear regression model, the model is generally termed as simple linear regression model. When there are more than one independent variables in the model, then the linear model is termed as the multiple linear regression model.

Simple Linear Regression Analysis

Consider a simple linear regression model

$$y = \beta_0 + \beta_1 X + \varepsilon$$

Simple Linear Regression Analysis

where y is termed as the **dependent** or **study variable** and X is termed as **independent** or **explanatory variable**.

The terms β_0 and β_1 are the parameters of the model. The parameter β_0 is termed as intercept term and the parameter β_1 is termed as slope parameter. These parameters are usually called as **regression coefficients**. The unobservable error component ε accounts for the failure of data to lie on the straight line and represents the difference between the true and observed realization of y. This is termed as **disturbance or error term**. There can be several reasons for such difference, e.g., the effect of all deleted variables in the model, variables may be qualitative, inherit randomness in the observations etc. We assume that ε is observed as independent and identically distributed random variable with mean zero and constant variance σ^2. Later, we will additionally assume that ε is normally distributed.

Simple Linear Regression Analysis

- The independent variable is viewed as controlled by the experimenter, so it is considered as non-stochastic whereas **y** is viewed as a random variable with

$$E(y) = \beta_0 + \beta_1 X$$

and

$$Var(y) = \sigma^2$$

Simple Linear Regression Analysis

- Sometimes **X** can also be a random variable. In such a case, instead of simple mean and simple variance of **y**, we consider the conditional mean of y given **X = x** as

$$E(y \mid x) = \beta_0 + \beta_1 x$$

and the conditional variance of y given $X = x$ as

$$Var(y \mid x) = \sigma^2.$$

When the values of β_0, β_1 and σ^2 are known, the model is completely described.

Simple Linear Regression Analysis

The parameters β_0, β_1 and σ^2 are generally unknown and ε is unobserved. The determination of the statistical model $y = \beta_0 + \beta_1 X + \varepsilon$ depends on the determination (i.e., estimation) of β_0, β_1 and σ^2.

In order to know the value of the parameters, n pairs of observations $(x_i, y_i)(i = 1,...,n)$ on (X, y) are observed/collected and are used to determine these unknown parameters.

Various methods of estimation can be used to determine the estimates of the parameters. Among them, the least squares and maximum likelihood principles are the popular methods of estimation.

The Likelihood Function

Suppose a sample of size n of a random vector y. Suppose the joint density of $Y = \begin{pmatrix} y_1 & \dots & y_n \end{pmatrix}$ is characterized by a parameter vector θ_0 :

$$f_{\mathcal{Y}}(Y, \theta_0).$$

This will often be referred to using the simplified notation $f(\theta_0)$.

The Likelihood Function

The *likelihood function* is just this density evaluated at other values θ

$$L(Y,\theta) = f_Y(Y,\theta), \theta \in \Theta,$$

where Θ is a *parameter space.*

- If the n observations are independent, the likelihood function can be written as

$$L(Y,\theta) = \prod_{t=1}^{n} f(y_t,\theta)$$

where the f_t are possibly of different form.

The Likelihood Function

- Even if this is not possible, we can always factor the likelihood into *contributions of observations*, by using the fact that a joint density can be factored into the product of a marginal and conditional (doing this iteratively)

$$L(Y,\theta) = f(y_1,\theta)f(y_2|y_1,\theta)f(y_3|y_1,y_2,\theta)\cdots f(y_n|y_1 y_2,\ldots y_{t-n},\theta)$$

The Likelihood Function

To simplify notation, define

$$x_t = \{y_1, y_2, \ldots, y_{t-1}\}, t \geq 2$$
$$= S, t = 1$$

where S is the sample space of Y. (With this, conditioning on x_1 has no effect and gives a marginal probability). Now the likelihood function can be written as

$$L(Y, \theta) = \prod_{t=1}^{n} f(y_t | x_t, \theta)$$

The Likelihood Function

The criterion function can be defined as the average log-likelihood function:

$$s_n(\theta) = \frac{1}{n} \ln L(Y, \theta) = \frac{1}{n} \sum_{t=1}^{n} \ln f(y_t | x_t, \theta)$$

The maximum likelihood estimator is defined as $\hat{\theta} = \arg\max s_n(\theta),$

where the set maximized over is defined below. Since $\ln(\cdot)$ is a monotonic increasing function, $\ln L$ and L maximize at the same value of θ. Dividing by n has no effect on $\hat{\theta}$.

Note that one can easily modify this to include exogenous conditioning variables in x_t in addition to the y_t that are already there. This changes nothing in what follows, and therefore it is suppressed to clarify the notation.

Consistency of MLE

To show consistency of the MLE, we need to make explicit some assumptions.

Compact parameter space $\theta \in \Theta$, a open bounded subset of $\Re^K$.

Maximixation is over $\bar{\Theta}$, which is compact.

This implies that θ is an interior point of the *parameter space* $\bar{\Theta}$.

Uniform convergence $s_n(\theta) \overset{u.a.s}{\rightarrow} \lim_{n\rightarrow\infty} \mathcal{E}_{\theta_0} s_n(\theta) \equiv s_\infty(\theta, \theta_0), \forall \theta \in \bar{\Theta}$.

We have suppressed Y here for simplicity. This requires that almost sure convergence holds for all possible parameter values.

Consistency of MLE

Continuity $s_n(\theta)$ is continuous in $\theta, \theta \in \overline{\Theta}$.

This implies that $s_\infty(\theta, \theta_0)$ is continuous in θ.

Identification $s_\infty(\theta, \theta_0)$ has a unique maximum in its first argument.

We will use these assumptions to show that $\hat{\theta} \xrightarrow{a.s.} \theta_0$. a.s. – **almost surely**

First, $\hat{\theta}$ certainly exists, since a continuous function has a maximum on a compact set.

Second, for any $\theta \neq \theta_0$ by Jensen's inequality ($\ln(\cdot)$ is a concave function).

$$\mathcal{E}\left(\ln\left(\frac{L(\theta)}{L(\theta_0)}\right)\right) \leq \ln\left(\mathcal{E}\left(\frac{L(\theta)}{L(\theta_0)}\right)\right)$$

Consistency of MLE

Now, the expectation on the RHS is

$$\mathcal{E}\left(\frac{L(\theta)}{L(\theta_0)}\right) = \int \frac{L(\theta)}{L(\theta_0)} L(\theta_0)dy = 1,$$

since $L(\theta_0)$ *is* the density function of the observations. Therefore, since $\ln(1) = 0$,

$$\mathcal{E}\left(\ln\left(\frac{L(\theta)}{L(\theta_0)}\right)\right) \leq 0,$$

Consistency of MLE

or

$$\mathcal{E}\left(s_n(\theta)\right) - \mathcal{E}\left(s_n(\theta_0)\right) \leq 0.$$

Taking limits, this is

$$s_\infty(\theta, \theta_0) - s_\infty(\theta_0, \theta_0) \leq 0$$

except on a set of zero probability (by the uniform convergence assumption).

Consistency of MLE

By the identification assumption there is a unique maximizer, so the inequality is strict if $\theta \neq \theta_0$:

$$s_\infty(\theta, \theta_0) - s_\infty(\theta_0, \theta_0) < 0, \forall \theta \neq \theta_0,$$

However, since $\hat{\theta}$ is a maximizer, independent of n, we must have

$$s_\infty(\hat{\theta}, \theta_0) - s_\infty(\theta_0, \theta_0) \geq 0.$$

Consistency of MLE

These last two inequalities imply that

$$\lim_{n \to \infty} \hat{\theta} = \theta_0, \text{ a.s.}$$

This completes the proof of strong consistency of the MLE. One can use weaker assumptions to prove weak consistency (convergence in probability to θ_0) of the MLE.

This is omitted here.
Note that almost sure convergence implies convergence in probability.

The score function

Differentiability

Assume that $s_n(\theta)$ is twice continuously differentiable in $N(\theta_0)$, at least when n is large enough.

To maximize the log-likelihood function, take derivatives:

$$\begin{aligned} g_n(Y,\theta) &= D_\theta s_n(\theta) \\ &= \frac{1}{n}\sum_{t=1}^{n} D_\theta \ln f(y_t|x_x,\theta) \\ &\equiv \frac{1}{n}\sum_{t=1}^{n} g_t(\theta). \end{aligned}$$

This is the *score vector* (with dim $K \times 1$). Note that the score function has Y as an argument, which implies that it is a random function. Y will often be suppressed for clarity, but one should not forget that it is still there.

The score function

The ML estimator $\hat{\theta}$ sets the derivatives to zero:

$$g_n(\hat{\theta}) = \frac{1}{n}\sum_{t=1}^{n} g_t(\hat{\theta}) = 0.$$

We will show that $\mathcal{E}_{\theta}[g_t(\theta)] = 0, \forall t$. *This is the expectation taken with respect to the density* $f(\theta)$, not necessarily $f(\theta_0)$.

$$\begin{aligned}
\mathcal{E}_{\theta}[g_t(\theta)] &= \int [D_{\theta} \ln f(y_t|x, \theta)] f(y_t|x, \theta) dy_t \\
&= \int \frac{1}{f(y_t|x_t, \theta)} [D_{\theta} f(y_t|x_t, \theta)] f(y_t|x_t, \theta) dy_t \\
&= \int D_{\theta} f(y_t|x_t, \theta) dy_t.
\end{aligned}$$

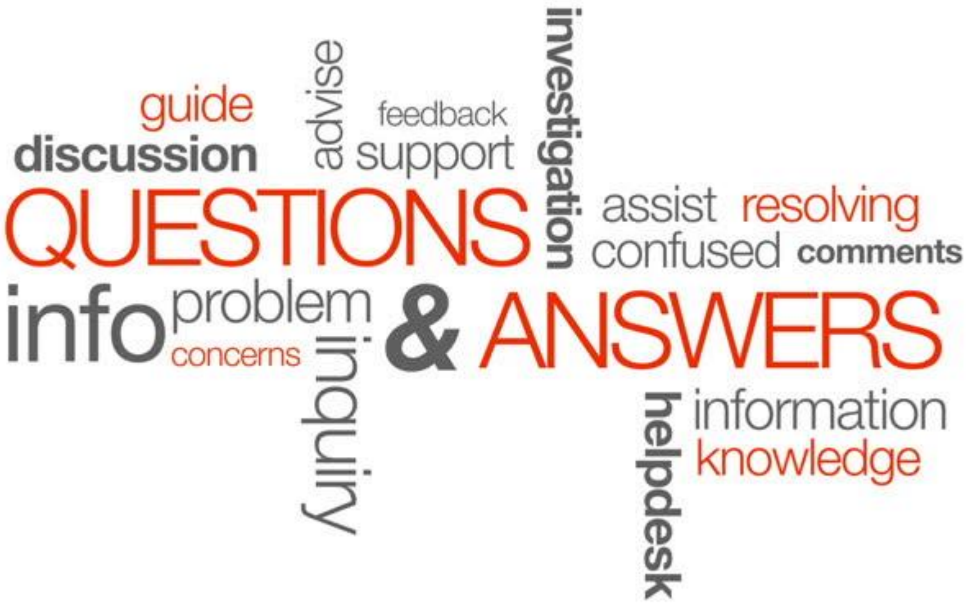
The score function

Given some regularity conditions on boundedness of $D_\theta f$, we can switch the order of integration and differentiation, by the dominated convergence theorem. This gives

$$\begin{aligned} \mathcal{E}_\theta[g_t(\theta)] &= D_\theta \int f t(y_t|x_t,\theta)dy_t \\ &= D_\theta 1 \\ &= 0. \end{aligned}$$

- So $\mathcal{E}_\theta(g_t(\theta) = 0$: *the expectation of the score vector is zero.*
- This hold for all t, so it implies that $\mathcal{E}_\theta g_n(Y,\theta) = 0$.

QUESTIONS!



guide
discussion
advise
feedback
support
investigation
QUESTIONS
assist
resolving
confused
comments
info
problem
concerns
inquiry
& ANSWERS
helpdesk
information
knowledge