KTH

Dep. of Mathematics
Harald Lang & Timo Koski 2/11-09

Computer Intensive Methods and the Least Squares Estimate

Introduction

The purpose of these notes is to give a brief introduction to linear models and the least squares method.

Linear models are treated in much more detail in the courses *Applied Mathematical Statistics* SF2950 and *Econometrics* SF2951.

Notation and Definitions

We observe a number of variables $y_1 \ldots, y_N$, where we assume that y_i is influenced by some **predictive** factors $x_{i1}, \ldots, x_{in}$.

More precisely, we take a model m_i of of y_i as a linear combination of the x-factors:

$$m_i = \textstyle\sum_{k=0}^{n} x_{ik}\, \theta_k,$$

where we for notational convenience introduce $x_{i0} = 1$. We think that this could predict the value of y_i before it was observed. We set

$$\mathbf{x}_i = \begin{pmatrix} x_{i0} \\ \vdots \\ x_{in} \end{pmatrix}, i = 1, \ldots, N, \quad \theta = \begin{pmatrix} \theta_0 \\ \vdots \\ \theta_n \end{pmatrix} \quad \mathbf{m} = \begin{pmatrix} m_1 \\ \vdots \\ m_N \end{pmatrix}.$$

Thereby we can write

$$m_i = \mathbf{x}_i' \theta, \quad i = 1, \ldots, N \tag{1}$$

where $\mathbf{x}_i'$ denotes the transpose of the vector $\mathbf{x}_i$. We introduce additional matrix notations: let

$$X = \begin{pmatrix} \mathbf{x}_1' \\ \vdots \\ \mathbf{x}_N' \end{pmatrix} = \begin{pmatrix} x_{10} & \cdots & x_{1n} \\ \vdots & \ddots & \vdots \\ x_{N0} & \cdots & x_{Nn} \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1n} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{N1} & \cdots & x_{Nn} \end{pmatrix}.$$

We can now write (1) compactly as

$$\mathbf{m} = X\theta \tag{2}$$

Next we introduce the **residuals**

$$\epsilon_i = y_i - \mathbf{x}_i' \theta, \quad i = 1, \ldots, N$$

so that with

$$\epsilon = \begin{pmatrix} \epsilon_1 \\ \vdots \\ \epsilon_N \end{pmatrix}, \quad Y = \begin{pmatrix} y_1 \\ \vdots \\ y_N \end{pmatrix},$$

we can write

$$Y = X\theta + \epsilon. \tag{3}$$

Least Squares Estimation

We want to estimate θ from the data $\left(y_i, \mathbf{x}_i^{'}\right)_{i=1}^{N}$. We will employ the **Least Squares Estimate** (LSE). First we define the least squares optimization criterion $Q(\theta)$. This is nothing but the sum of squared residuals, regarded as a function of θ. We have using (3)

$$Q(\theta) = \frac{1}{2}\sum_{i=1}^{N} \epsilon_i^2 = \frac{1}{2}\epsilon'\epsilon = \frac{1}{2}(Y - X\theta)'(Y - X\theta). \tag{4}$$

The LSE is the value of $\hat{\theta}$ that minimizes $Q(\theta)$, i.e.,

$$\hat{\theta} = \text{argmin}_{\theta} Q(\theta).$$

Proposition If $(X'X)$ is a positive definite matrix, then

$$\hat{\theta} = (X'X)^{-1}X'Y, \tag{5}$$

Proof : We expand $Q(\theta)$ from (4) to obtain

$$Q(\theta) = \frac{1}{2}(Y'Y - Y'X\theta - \theta'X'Y + \theta'X'X\theta).$$

Then we add and subtract $Y'X\hat{\theta} = Y'X(X'X)^{-1}X'Y$ inside the parenthesis in the right hand side. This gives us an opportunity to complete a square, and yields

$$Q(\theta) = \frac{1}{2}\left(\left(\theta - (X'X)^{-1}X'Y\right)'(X'X)\left(\theta - (X'X)^{-1}X'Y\right)\right)$$

$$+\frac{1}{2}\left(Y'Y - Y'X(X'X)^{-1}X'Y\right). \tag{6}$$

The second term in the right hand side does not depend on θ. Hence we can minimize the expression by minimizing the first term, which is a quadratic form. Since $X'X$ is a positive definite matrix, the quadratic form is zero if and only if we choose $\theta - (X'X)^{-1}X'Y$ to be the zero vector. Hence we have shown the proposition as claimed.*Q.E.D.*

From (6) we see that

$$Q\left(\hat{\theta}\right) = \frac{1}{2}Y'Y - Y'X(X'X)^{-1}X'Y. \tag{7}$$

Maximum Likelihood Estimation

In order to be able to analyze the properties of the least squares estimate $\hat{\theta}$ we need to make additional assumptions. We shall assume that there is a “true” value θ_* (unknown to us) so that residuals ϵ_i are independent and $N(0, \sigma^2)$ distributed random variables, or

$$Y = X\theta_* + \epsilon, \quad \epsilon \in N(0, \sigma^2 I_N), \tag{8}$$

where σ is unknown, too.

This is of course a very strong assumption: we assume both normality and *homoscedasticity,* i.e., that the residuals ϵ_i all have the same variance. (Situations when these assumptions do not hold are treated in the econometrics course, for instance.)

Then by (Gut Theorem 3.1 p. 124) the distribution of the random vector Y is

$$Y \in N(X\theta_*, \sigma^2 I_N).$$

(When reading the text by Gut at this point one has to note the differences in notation. We must take **b** in Gut loc.cit. as $= X\theta_*$ and **X** in Gut loc.cit as ϵ, $\mathbf{\Lambda}$ as $\sigma^2 I_N$.) Then the probability density of Y exists and the likelihood function $L(\theta)$ for θ becomes

$$L(\theta) = \frac{1}{(2\pi)^{N/2}\sigma^N} e^{-\frac{Q(\theta)}{\sigma^2}},$$

where $Q(\theta)$ is defined in (3). The **Maximum Likelihood Estimate** is defined as the value of θ that maximizes $L(\theta)$. But clearly maximization of $L(\theta)$ is equivalent to minimization of $Q(\theta)$. Hence the ML-estimator of θ coincides with the LSE in the current context.

Properties of the LSE

We continue with the statistical model in (8). Employing (5) and (8), we see that

$$\hat{\theta} = (X'X)^{-1}X'Y = (X'X)^{-1}X'(X\theta_* + \epsilon) = \theta_* + (X'X)^{-1}X'\epsilon \qquad (9).$$

From this we see that $\hat{\theta}$ is an **unbiased** estimator of θ, i.e. the mean vector of $\hat{\theta}$ is

$$E\left[\hat{\theta}\right] = \theta_* + (X'X)^{-1}X'E[\epsilon] = \theta_*.$$

The covariance matrix of $\hat{\theta}$ is using (9) (and theorem 2.2 in Gut p. 122)

$$\begin{aligned}
\mathrm{Cov}(\hat{\theta}) &= E\left[\left(\hat{\theta}-\theta_*\right)\left(\hat{\theta}-\theta_*\right)'\right] \\
&= (X'X)^{-1}X'\,\mathrm{Cov}(\epsilon)X(X'X)^{-1}\,. \\
&= (X'X)^{-1}X'(\sigma^2 I_N)X(X'X)^{-1} \\
&= \sigma^2(X'X)^{-1}
\end{aligned} \tag{10}$$

Hence, we see by (9) that

$$\hat{\theta} \in N(\theta_*, \sigma^2(X'X)^{-1}). \tag{11}$$

In order to use this distribution for statistical purposes (e.g. testing of hypotheses on θ_*), we obviously need an estimate for σ, which is typically unknown. Before addressing this, we need to point out some facts from matrix calculus.

Some Matrix Relations

Let A be a square matrix. The **trace** $\mathrm{Tr}\,A$ of A is the sum of the entries in main diagonal:

$$\mathrm{Tr}\begin{pmatrix} a_{11} & \cdots & a_{1k} \\ \vdots & \ddots & \vdots \\ a_{k1} & \cdots & a_{kk} \end{pmatrix} = \textstyle\sum_1^k a_{jj}$$

The following facts are easily established; the proofs are left as exercises:

1. If A is a $k \times n$-matrix, and B an $n \times k$-matrix, then $\mathrm{Tr}(AB) = \mathrm{Tr}(BA)$
2. In particular, if a is a column-vector, then $a'a = \mathrm{Tr}(aa')$.
3. $\mathrm{Tr}(C + D) = \mathrm{Tr}\,C + \mathrm{Tr}\,D$
4. Let A be a $k \times n$-matrix (of full rank) where $n < k$. Define $P = A(A'A)^{-1}A'$. Then
   - P is symmetric (i.e., $P' = P$.)
   - $P^2 = P$
   - $\mathrm{Tr}\,P = n$

   Let us prove the last statement:

$$\mathrm{Tr}\,P = \mathrm{Tr}\,A(A'A)^{-1}A' = \text{ (by 1.)} = \mathrm{Tr}\,A'A(A'A)^{-1} = \mathrm{Tr}\,I_n = n.$$

Estimation of σ, cont.

We now estimate σ^2. Let us denote by

$$\hat{\epsilon} \overset{\text{def}}{=} Y - X\hat{\theta}$$

the observed residuals under the LSE-prediction. Then

$$\begin{aligned}
\hat{\epsilon} &= X\theta_* + \epsilon - X\hat{\theta} \\
&= X\theta_* + \epsilon - X(\theta_* + (X'X)^{-1}X'\epsilon) \\
&= \big(I_N - X(X'X)^{-1}X'\big)\epsilon \\
&= \big(I_N - P\big)\epsilon, \\
\text{where } P &\overset{\text{def}}{=} X(X'X)^{-1}X'.
\end{aligned}$$

Note that P, by the previous section, is symmetric, $P^2 = P$ and $\operatorname{Tr} P = n + 1$. Hence

$$\begin{aligned}
E\big[\textstyle\sum_{i=1}^{N} \hat{\epsilon}_i^2\big] &= E[\hat{\epsilon}'\hat{\epsilon}] \\
&= \operatorname{Tr} E[\hat{\epsilon}\hat{\epsilon}'] \\
&= \operatorname{Tr}\big((I_N - P)E[\epsilon\epsilon'](I_N - P)\big) \\
&= \operatorname{Tr}\big((I_N - P)\big(\sigma^2 I_N\big)(I_N - P)\big) \\
&= \sigma^2 \operatorname{Tr}\big(I_N - P\big) \\
&= \sigma^2\big(\operatorname{Tr} I_N - \operatorname{Tr} P) = \sigma^2(N - n - 1)
\end{aligned}$$

Our unbiased estimate of σ^2 is thus

$$\hat{\sigma}^2 = \tfrac{1}{N-n-1} \textstyle\sum_{i=1}^{N} \hat{\epsilon}_i^2$$

Clearly it also holds that

$$\hat{\sigma}^2 = \tfrac{2}{N-n-1} Q\left(\hat{\theta}\right).$$

In summary:

$$Y = X\theta_* + \epsilon, \quad \epsilon \in N(0, \sigma^2 I_N)$$

$$\hat{\theta} = (X'X)^{-1}X'Y$$

$$\hat{\theta} \in N(\theta_*, \sigma^2(X'X)^{-1}).$$

$$\hat{\sigma}^2 = \frac{1}{N-n-1}\sum_{i=1}^{N} \hat{\epsilon}_i^2$$

On the Model

We can also think of the variables in X as outcomes of random variables (e.g., normal r.v.'s) jointly distributed with Y. The statistical model in the preceding would then be rewritten with a conditional statement,

$$Y \mid \mathbf{X}_1 = \mathbf{x}_1, \ldots, \mathbf{X}_N = \mathbf{x}_N \quad \epsilon \in N(X\theta_*, \sigma^2 I_N).$$

The rest of the properties of LSE are the same, when conditioned on the observed values of the predictor variables.

Prediction

A common situation is that we want to forecast a new value of y_{N+1} based on the values of the $\mathbf{x}$-parameters. If we have estimated θ to $\hat{\theta}$, then an unbiased prediction is

$$\hat{y}_{N+1} = \mathbf{x}'_{N+1}\hat{\theta}, \quad \text{where } \mathbf{x}'_{N+1} = (1, x_{N+11}, \ldots, x_{N+1n})$$

We can think of prediction in real time. We have observed the variables $y_1 \ldots, y_N$, up to time N and wish to predict the next value, y_{N+1}. We assume, of course, that the underlying "true" mechanism generating data is unchanged in the sense that

$$y_{N+1} = \mathbf{x}'_{N+1}\theta_* + \epsilon_{N+1}.$$

Note that ϵ_{N+1} is assumed to be independent of $\epsilon_1 \ldots, \epsilon_N,$.

In order to simplify writing (and to accomodate to other possible cases of prediction) we set

$$y = y_{N+1}, \hat{y} = \hat{y}_{N+1}, \mathbf{x} = \mathbf{x}_{N+1}, \quad e = \epsilon_{N+1}, y = \mathbf{x}'\theta_* + e.$$

We now proceed to calculate the prediction error $y - \hat{y}$ applying (9) from the above

$$y - \hat{y} = \mathbf{x}'\theta_* + e - \mathbf{x}'\hat{\theta} = \mathbf{x}'(\theta_* - \hat{\theta}) + e = -\mathbf{x}'(X'X)^{-1}X'\epsilon + e$$

Here ϵ and e are, as said, independent, so the **mean squared (prediction) error** MSE =
$E[(y - \hat{y})^2]$ is (by Theorem 2.2 in Gut p. 122)

$$\text{MSE} = \mathbf{x}'(X'X)^{-1}X'E[\sigma^2 I_N]X(X'X)^{-1}\mathbf{x} + \sigma^2 = \sigma^2\left(\mathbf{x}'(X'X)^{-1}\mathbf{x} + 1\right)$$

Since we have to use an estimate of σ^2, the approximate MSE is

$$\text{MSE} = \hat{\sigma}^2\left(\mathbf{x}'(X'X)^{-1}\mathbf{x} + 1\right).$$

The estimated **root mean squared error** RMSE $= \sqrt{E[(y - \hat{y})^2]}$ is thus

$$\text{RMSE} = \hat{\sigma}\sqrt{\mathbf{x}'(X'X)^{-1}\mathbf{x} + 1}$$

It is appropriate to assume that the prediction error is a Normal random variable with variance MSE, although the "true" distribution is a t-distribution, due to the estimate of σ. However, there is already in practice an approximation in the specification of the error term being Normally distributed, so why bother about t-distributions.

Hypothesis Testing

Another common situation is that we want to assess the values of θ_*, and test a hypothesis on their values. We know that $\hat{\theta} - \theta_* \in N\big(0, \sigma^2(X'X)^{-1}\big)$. If R is some $k \times n$-matrix, it follows that $R(\hat{\theta} - \theta_*) \in N\big(0, \sigma^2 R(X'X)^{-1}R'\big)$

We can now employ Theorem 9.1 on p. 139 in Gut and get:

$$\sigma^{-2}(\hat{\theta} - \theta_*)'R'\big(R(X'X)^{-1}R'\big)^{-1}R(\hat{\theta} - \theta_*) \in \chi^2(k)$$

and hence approximately,

$$\hat{\sigma}^{-2}(\hat{\theta} - \theta_*)'R'\big(R(X'X)^{-1}R'\big)^{-1}R(\hat{\theta} - \theta_*) \in \chi^2(k). \tag{12}$$

The difference is that we have replaced σ^2 by $\hat{\sigma}^2$. The strictly mathematically correct distribution is now an $F(k, N-n-1)$-distribution, but again, why bother, considering the unavoidable specification error. The fact in (12) can now be used in obvious ways to test hypotheses about the true values of the parameters θ.