## KTH

Dep. of Mathematics
Harald Lang \& Timo Koski 20/4-08
Linear Models 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 $x_{i 1}, \ldots, x_{i n}$.

More precisely, we take a model $m_{i}$ of of $y_{i}$ as a linear combination of the $x$-factors:

$$
m_{i}=\sum_{k=0}^{n} x_{i k} \theta_{k}
$$

where we for notational convenience introduce $x_{i 0}=1$. We think that this could predict the value of $y_{i}$ before it was observed. We set

$$
\mathbf{x}_{i}=\left(\begin{array}{c}
x_{i 0} \\
\vdots \\
x_{i n}
\end{array}\right), i=1, \ldots, N, \quad \theta=\left(\begin{array}{c}
\theta_{0} \\
\vdots \\
\theta_{n}
\end{array}\right) \quad \mathbf{m}=\left(\begin{array}{c}
m_{1} \\
\vdots \\
m_{N}
\end{array}\right)
$$

Thereby we can write

$$
\begin{equation*}
m_{i}=\mathbf{x}_{i}^{\prime} \theta, \quad i=1, \ldots, N \tag{1}
\end{equation*}
$$

where $\mathbf{x}_{i}^{\prime}$ denotes the transpose of the vector $\mathbf{x}_{i}$. We introduce additional matrix notations: let

$$
X=\left(\begin{array}{c}
\mathbf{x}_{1}^{\prime} \\
\vdots \\
\mathbf{x}_{N}^{\prime}
\end{array}\right)=\left(\begin{array}{ccc}
x_{10} & \cdots & x_{1 n} \\
\vdots & \ddots & \vdots \\
x_{N 0} & \cdots & x_{N n}
\end{array}\right)=\left(\begin{array}{cccc}
1 & x_{11} & \cdots & x_{1 n} \\
\vdots & \vdots & \ddots & \vdots \\
1 & x_{N 1} & \cdots & x_{N n}
\end{array}\right)
$$

We can now write (1) compactly as

$$
\begin{equation*}
\mathbf{m}=X \theta \tag{2}
\end{equation*}
$$

Next we introduce the residuals

$$
\epsilon_{i}=y_{i}-\mathbf{x}_{i}^{\prime} \theta, \quad i=1, \ldots, N
$$

so that with

$$
\epsilon=\left(\begin{array}{c}
\epsilon_{1} \\
\vdots \\
\epsilon_{N}
\end{array}\right), \quad Y=\left(\begin{array}{c}
y_{1} \\
\vdots \\
y_{N}
\end{array}\right)
$$

we can write

$$
\begin{equation*}
Y=X \theta+\epsilon \tag{3}
\end{equation*}
$$

## Least Squares Estimation

We want to estimate $\theta$ from the data $\left(y_{i}, \mathbf{x}_{i}^{\prime}\right)_{i=1}^{N}$. 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 $\theta$. We have using (3)

$$
\begin{equation*}
Q(\theta)=\frac{1}{2} \sum_{i=1}^{N} \epsilon_{i}^{2}=\frac{1}{2} \epsilon^{\prime} \epsilon=\frac{1}{2}(Y-X \theta)^{\prime}(Y-X \theta) \tag{4}
\end{equation*}
$$

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

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

Proposition If $\left(X^{\prime} X\right)$ is a positive definite matrix, then

$$
\begin{equation*}
\hat{\theta}=\left(X^{\prime} X\right)^{-1} X^{\prime} Y \tag{5}
\end{equation*}
$$

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

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

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

$$
\begin{align*}
Q(\theta)=\frac{1}{2}( & \left.\left(\theta-\left(X^{\prime} X\right)^{-1} X^{\prime} Y\right)^{\prime}\left(X^{\prime} X\right)\left(\theta-\left(X^{\prime} X\right)^{-1} X^{\prime} Y\right)\right) \\
& +\frac{1}{2}\left(Y^{\prime} Y-Y^{\prime} X\left(X^{\prime} X\right)^{-1} X^{\prime} Y\right) \tag{6}
\end{align*}
$$

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

From (6) we see that

$$
\begin{equation*}
Q(\hat{\theta})=\frac{1}{2} Y^{\prime} Y-Y^{\prime} X\left(X^{\prime} X\right)^{-1} X^{\prime} Y \tag{7}
\end{equation*}
$$

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 $\theta_{*}$ (unknown to us) so that residuals $\epsilon_{i}$ are independent and $N\left(0, \sigma^{2}\right)$ distributed random variables, or

$$
\begin{equation*}
Y=X \theta_{*}+\epsilon, \quad \epsilon \in N\left(0, \sigma^{2} I_{N}\right) \tag{8}
\end{equation*}
$$

where $\sigma$ is unknown, too.
This is of course a very strong assumption: we assume both normality and homoscedasticity, i.e., that the residuals $\epsilon_{i}$ 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

$$
Y \in N\left(X \theta_{*}, \sigma^{2} I_{N}\right)
$$

(When reading the text by Gut at this point one has to note the differences in notation. We must take $\mathbf{b}$ in Gut loc.cit. as $=X \theta_{*}$ and $\mathbf{X}$ in Gut loc.cit as $\epsilon$, $\boldsymbol{\Lambda}$ as $\sigma^{2} I_{N}$.) Then the probability density of $Y$ exists and the likelihood function $L(\theta)$ for $\theta$ 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 $\theta$ that maximizes $L(\theta)$. But clearly maximization of $L(\theta)$ is equivalent to minimization of $Q(\theta)$. Hence the ML-estimator of $\theta$ 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

$$
\begin{equation*}
\hat{\theta}=\left(X^{\prime} X\right)^{-1} X^{\prime} Y=\left(X^{\prime} X\right)^{-1} X^{\prime}\left(X \theta_{*}+\epsilon\right)=\theta_{*}+\left(X^{\prime} X\right)^{-1} X^{\prime} \epsilon \tag{9}
\end{equation*}
$$

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

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

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

$$
\begin{align*}
\operatorname{Cov}(\hat{\theta}) & =E\left[\left(\hat{\theta}-\theta_{*}\right)\left(\hat{\theta}-\theta_{*}\right)^{\prime}\right] \\
& =\left(X^{\prime} X\right)^{-1} X^{\prime} \operatorname{Cov}(\epsilon) X\left(X^{\prime} X\right)^{-1}  \tag{10}\\
& =\left(X^{\prime} X\right)^{-1} X^{\prime}\left(\sigma^{2} I_{N}\right) X\left(X^{\prime} X\right)^{-1} \\
& =\sigma^{2}\left(X^{\prime} X\right)^{-1}
\end{align*}
$$

Hence, we see by (9) that

$$
\begin{equation*}
\hat{\theta} \in N\left(\theta_{*}, \sigma^{2}\left(X^{\prime} X\right)^{-1}\right) \tag{11}
\end{equation*}
$$

In order to use this distribution for statistical purposes (e.g. testing of hypotheses on $\theta_{*}$ ), we obviously need an estimate for $\sigma$, 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 $\operatorname{Tr} A$ of $A$ is the sum of the entries in main diagonal:

$$
\operatorname{Tr}\left(\begin{array}{ccc}
a_{11} & \cdots & a_{1 k} \\
\vdots & \ddots & \vdots \\
a_{k 1} & \cdots & a_{k k}
\end{array}\right)=\sum_{1}^{k} a_{j j}
$$

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 $\operatorname{Tr}(A B)=\operatorname{Tr}(B A)$
2. In particular, if $a$ is a column-vector, then $a^{\prime} a=\operatorname{Tr}\left(a a^{\prime}\right)$.
3. $\operatorname{Tr}(C+D)=\operatorname{Tr} C+\operatorname{Tr} D$
4. Let $A$ be a $k \times n$-matrix (of full rank) where $n<k$. Define $P=A\left(A^{\prime} A\right)^{-1} A^{\prime}$. Then

- $\quad P$ is symmetric (i.e., $P^{\prime}=P$.)
- $P^{2}=P$
- $\operatorname{Tr} P=n$

Let us prove the last statement:

$$
\operatorname{Tr} P=\operatorname{Tr} A\left(A^{\prime} A\right)^{-1} A^{\prime}=(\text { by 1. })=\operatorname{Tr} A^{\prime} A\left(A^{\prime} A\right)^{-1}=\operatorname{Tr} I_{n}=n
$$

Estimation of $\sigma$, cont.
We now estimate $\sigma^{2}$. Let us denote by

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

the observed residuals under the LSE-prediction. Then

$$
\begin{aligned}
& \qquad \begin{aligned}
\hat{\epsilon} & =X \theta_{*}+\epsilon-X \hat{\theta} \\
& =X \theta_{*}+\epsilon-X\left(\theta_{*}+\left(X^{\prime} X\right)^{-1} X^{\prime} \epsilon\right) \\
& =\left(I_{N}-X\left(X^{\prime} X\right)^{-1} X^{\prime}\right) \epsilon \\
& =\left(I_{N}-P\right) \epsilon, \\
\text { where } P & \stackrel{\text { def }}{=} X\left(X^{\prime} X\right)^{-1} X^{\prime} .
\end{aligned}
\end{aligned}
$$

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

$$
\begin{aligned}
E\left[\sum_{i=1}^{N} \hat{\epsilon}_{i}^{2}\right] & =E\left[\hat{\epsilon}^{\prime} \hat{\epsilon}\right] \\
& =\operatorname{Tr} E\left[\hat{\epsilon}^{\prime}{ }^{\prime}\right] \\
& =\operatorname{Tr}\left(\left(I_{N}-P\right) E\left[\epsilon \epsilon^{\prime}\right]\left(I_{N}-P\right)\right) \\
& =\operatorname{Tr}\left(\left(I_{N}-P\right)\left(\sigma^{2} I_{N}\right)\left(I_{N}-P\right)\right) \\
& =\sigma^{2} \operatorname{Tr}\left(I_{N}-P\right) \\
& =\sigma^{2}\left(\operatorname{Tr} I_{N}-\operatorname{Tr} P\right)=\sigma^{2}(N-n-1)
\end{aligned}
$$

Our unbiased estimate of $\sigma^{2}$ is thus

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

Clearly it also holds that

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

## In summary:

$$
\begin{gathered}
Y=X \theta_{*}+\epsilon, \quad \epsilon \in N\left(0, \sigma^{2} I_{N}\right) \\
\hat{\theta}=\left(X^{\prime} X\right)^{-1} X^{\prime} Y \\
\hat{\theta} \in N\left(\theta_{*}, \sigma^{2}\left(X^{\prime} X\right)^{-1}\right) \\
\hat{\sigma}^{2}=\frac{1}{N-n-1} \sum_{i=1}^{N} \hat{\epsilon}_{i}^{2}
\end{gathered}
$$

## 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,

$$
Y \mid \mathbf{X}_{1}=\mathbf{x}_{1}, \ldots, \mathbf{X}_{N}=\mathbf{x}_{N} \quad \epsilon \in N\left(X \theta_{*}, \sigma^{2} I_{N}\right)
$$

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 $\theta$ to $\hat{\theta}$, then an unbiased prediction is

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

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

$$
y_{N+1}=\mathbf{x}_{N+1}^{\prime} \theta_{*}+\epsilon_{N+1}
$$

Note that $\epsilon_{N+1}$ 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

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

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

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

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

$$
\mathrm{MSE}=\mathbf{x}^{\prime}\left(X^{\prime} X\right)^{-1} X^{\prime} E\left[\sigma^{2} I_{N}\right] X\left(X^{\prime} X\right)^{-1} \mathbf{x}+\sigma^{2}=\sigma^{2}\left(\mathbf{x}^{\prime}\left(X^{\prime} X\right)^{-1} \mathbf{x}+1\right)
$$

Since we have to use an estimate of $\sigma^{2}$, the approximate MSE is

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

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

$$
\mathrm{RMSE}=\hat{\sigma} \sqrt{\mathbf{x}^{\prime}\left(X^{\prime} X\right)^{-1} \mathbf{x}+1}
$$

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 $\sigma$. 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 $\theta_{*}$, and test a hypothesis on their values. We know that $\hat{\theta}-\theta_{*} \in N\left(0, \sigma^{2}\left(X^{\prime} X\right)^{-1}\right)$. If $R$ is some $k \times n$-matrix, it follows that $R\left(\hat{\theta}-\theta_{*}\right) \in N\left(0, \sigma^{2} R\left(X^{\prime} X\right)^{-1} R^{\prime}\right)$

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

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

and hence approximately,

$$
\begin{equation*}
\hat{\sigma}^{-2}\left(\hat{\theta}-\theta_{*}\right)^{\prime} R^{\prime}\left(R\left(X^{\prime} X\right)^{-1} R^{\prime}\right)^{-1} R\left(\hat{\theta}-\theta_{*}\right) \in \chi^{2}(k) . \tag{12}
\end{equation*}
$$

The difference is that we have replaced $\sigma^{2}$ 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 $\theta$.

