PDHonline Course G160 (3 PDH)

Maximum Likelihood & Gauss-Markov Parameter Estimation

Instructor: Drayton D. Boozer, Ph.D, PE

2020

PDH Online | PDH Center

5272 Meadow Estates Drive
Fairfax, VA 22030-6658
Phone: 703-988-0088
www.PDHonline.com

An Approved Continuing Education Provider

Maximum Likelihood & Gauss-Markov Parameter Estimation

Course Content

Introduction

In PDHonline Course G429, Linear Least Squares Parameter Estimation, the general subject of parameter estimation was introduced. A mathematical framework and set of error assumptions were developed. Linear least squares parameter estimation was then developed using the framework and assumption set.

In this course we develop maximum likelihood parameter and Gauss-Markov estimation using the same framework and assumption set. Both methods enable use of known correlation structure in the measurement errors. The maximum likelihood method as developed in this course requires the assumption of normally distributed measurement errors whereas the Gauss-Markov method does not.

For the assumed linear measurement model, both the maximum likelihood and Gauss-Markov estimators have a closed form and are structurally similar to the least squares estimator.

A concise summary of the least squares, maximum likelihood, and Gauss-Markov estimators concludes the course.

Measurement Model

For the linear parameter estimation model from G429 we have

$$\hat{\mathbf{Y}} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \qquad \textbf{(2.1)}$$

where

- $\hat{\mathbf{Y}}$ is an m-dimensional random vector of measured dependent variables
- $\mathbf{X}$ is an (mxp)-dimensional matrix of independent variables
- $\boldsymbol{\beta}$ is a p-dimensional vector of unknown parameters
- $\boldsymbol{\varepsilon}$ is an m-dimensional random vector of measurement errors

Statistical Assumptions for Measurement Errors

The measurement error assumption table from G429 is repeated here for convenience.

Table 1
Statistical Assumptions for Measurement Errors

1. **Additive**
   - 0 No, not additive
   - 1 Yes, additive
2. **Zero-mean**
   - 0 No, not zero-mean
   - 1 Yes, zero-mean
3. **Constant Variance**
   - 0 No, nonconstant variance
   - 1 Yes, constant variance
4. **Uncorrelated**
   - 0 No, correlated errors
   - 1 Yes, uncorrelated errors
5. **Normal Probability Distribution**
   - 0 No, nonnormal distribution
   - 1 Yes, normal distribution
6. **Known Covariance Matrix**
   - 0 No, covariance matrix of errors known only to within a multiplicative constant
   - 1 Yes, covariance matrix of errors known

Additive errors are assumed in the model given in equation (2.1). Zero-mean errors is often a reasonable assumption. When nonzero-mean errors are suspected, estimation of the nonzero-mean is accomplished simply by adding another parameter to the model in equation (2.1). To illustrate, append the nonzero-mean parameter, β_{p+1}, to the parameter vector $\boldsymbol{\beta}$. Next append an m-dimensional unity column to the $\mathbf{X}$ matrix. Now we can write an equation analogous to equation (2.1),

$$\hat{\mathbf{Y}} = \mathbf{X}_{new}\boldsymbol{\beta}_{new} + \boldsymbol{\varepsilon} = \begin{bmatrix} \mathbf{X} & \vdots & 1 \\ & & 1 \\ & & \vdots \\ & & 1 \end{bmatrix} \begin{bmatrix} \boldsymbol{\beta} \\ \cdots \\ \beta_{p+1} \end{bmatrix} + \boldsymbol{\varepsilon} \qquad \textbf{(2.2)}$$

where

$\mathbf{X}_{new}$ is an $\left(\text{m x}\left(\text{p+1}\right)\right)$-dimensional matrix of independent variables

$\boldsymbol{\beta}_{new}$ is a $\left(\text{p+1}\right)$-dimensional vector of unknown parameters

$\beta_{\text{p+1}}$ is the parameter that represents the nonzero-mean of the measurement vector

Measurement errors are not required to be either constant variance or uncorrelated in maximum likelihood and Gauss-Markov estimation.

We now develop the multivariate normal probability density function. The probability density function for a normally distributed random variable with mean value μ and variance σ^2 is given by

$$f(x)=\frac{1}{\sqrt{2\pi}\sigma}\exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right] \qquad \textbf{(2.3)}$$

Denote this density function by $N(\mu,\sigma^2)$.

For the corresponding multivariate measurement error case of equation (2.1) there are m normally distributed random variables. Define the mean value of $\boldsymbol{\varepsilon}$ by

$$\bar{\boldsymbol{\varepsilon}}=E\left(\boldsymbol{\varepsilon}\right) \qquad \textbf{(2.4)}$$

and the covariance matrix of $\boldsymbol{\varepsilon}$ by

$$\boldsymbol{\psi}=cov(\boldsymbol{\varepsilon})=E\left[(\boldsymbol{\varepsilon}-\bar{\boldsymbol{\varepsilon}})(\boldsymbol{\varepsilon}-\bar{\boldsymbol{\varepsilon}})^{\mathbf{T}}\right] \qquad \textbf{(2.5)}$$

From this definition we see that $\boldsymbol{\psi}$ is an m x m symmetric matrix.
The normal multivariate probability density function is given by

$$f(\boldsymbol{\varepsilon})=\left(2\pi\right)^{-\frac{m}{2}}\left|\boldsymbol{\psi}\right|^{-\frac{1}{2}}\exp\left[-\frac{1}{2}(\boldsymbol{\varepsilon}-\bar{\boldsymbol{\varepsilon}})^{\mathbf{T}}\boldsymbol{\psi}^{-1}(\boldsymbol{\varepsilon}-\bar{\boldsymbol{\varepsilon}})\right] \qquad \textbf{(2.6)}$$

where

$\left|\boldsymbol{\psi}\right|$ is the determinant of $\boldsymbol{\psi}$

Denote this multivariate density function by $N(\bar{\boldsymbol{\varepsilon}},\boldsymbol{\psi})$. When the measurement errors are normally distributed the errors are completed characterized by $\bar{\boldsymbol{\varepsilon}}$ and $\boldsymbol{\psi}$. Notice in the formulation of equation (2.2) that all elements of $\bar{\boldsymbol{\varepsilon}}$ are the same, namely $\beta_{\text{p+1}}$.
If the covariance matrix $\boldsymbol{\psi}$ is known to within a multiplicative constant, $\boldsymbol{\psi}$ is written as

$$\boldsymbol{\psi}=\sigma^2\boldsymbol{\Omega} \qquad \textbf{(2.7)}$$

and $\boldsymbol{\Omega}$ is also a covariance matrix.

Maximum Likelihood Estimation

The fundamental idea of maximum likelihood estimation is that the probability density function is useful in describing measurements before they are collected, whereas likelihoods are useful in estimating parameters after the measurements are collected. Although maximum likelihood estimation can be applied to a wide range of practical problems, we limit our development in this course to the case where the measurement errors are normally distributed.

Suppose the measurement errors are additive, zero mean, and normally distributed with known covariance matrix ψ ; (1 1 – – 1 1) in our shorthand notation.

The conditional probability density function for the measurements $\hat{\mathbf{Y}}$ given parameter values β is the multivariate normal density function

$$f(\hat{\mathbf{Y}}/\boldsymbol{\beta}) = (2\pi)^{-\frac{m}{2}} |\boldsymbol{\psi}|^{-\frac{1}{2}} \exp\left[-\frac{1}{2} (\hat{\mathbf{Y}} - \mathbf{X}\boldsymbol{\beta})^{\mathrm{T}} \boldsymbol{\psi}^{-1} (\hat{\mathbf{Y}} - \mathbf{X}\boldsymbol{\beta}) \right] \quad \textbf{(2.8)}$$

Before the measurements are taken this function gives the distribution of the measurements for specified values of β. After the measurements are taken we seek the β that best explains the measurements. The likelihood function for the given assumption set is

$$L(\boldsymbol{\beta}/\hat{\mathbf{Y}}) = (2\pi)^{-\frac{m}{2}} |\boldsymbol{\psi}|^{-\frac{1}{2}} \exp\left[-\frac{1}{2} (\hat{\mathbf{Y}} - \mathbf{X}\boldsymbol{\beta})^{\mathrm{T}} \boldsymbol{\psi}^{-1} (\hat{\mathbf{Y}} - \mathbf{X}\boldsymbol{\beta}) \right] \quad \textbf{(2.9)}$$

Maximum likelihood estimation requires that we choose as the parameter estimate the β that maximizes the likelihood function. Taking the natural logarithm we have

$$l(\boldsymbol{\beta}/\hat{\mathbf{Y}}) = \ln\left[L(\boldsymbol{\beta}/\hat{\mathbf{Y}}) \right] = -\frac{1}{2} \left[m \ln(2\pi) + \ln|\boldsymbol{\psi}| + \mathbf{S}_{ML} \right] \quad \textbf{(2.10)}$$

where

$$\mathbf{S}_{ML} = (\hat{\mathbf{Y}} - \mathbf{X}\boldsymbol{\beta})^{\mathrm{T}} \boldsymbol{\psi}^{-1} (\hat{\mathbf{Y}} - \mathbf{X}\boldsymbol{\beta}) \quad \textbf{(2.11)}$$

Since ψ is known, maximizing $L(\boldsymbol{\beta}/\hat{\mathbf{Y}})$ is equivalent to minimizing the weighted sum of squares function $\mathbf{S}_{ML}$.

The value of β that minimizes equation (2.11) is the maximum likelihood estimate of β, $\hat{\boldsymbol{\beta}}_{ML}$. The maximum likelihood estimator for the parameter vector is

$$\hat{\boldsymbol{\beta}}_{ML} = \left(\mathbf{X}^T \boldsymbol{\psi}^{-1} \mathbf{X} \right)^{-1} \mathbf{X}^T \boldsymbol{\psi}^{-1} \hat{\mathbf{Y}} \quad \textbf{(2.12)}$$

The minimum of the weighted sum of squares from equation (2.11) is

$$\mathbf{R}_{ML} = Min(\mathbf{S}_{ML}) = \left(\hat{\mathbf{Y}} - \mathbf{X}\hat{\boldsymbol{\beta}}_{ML} \right)^T \boldsymbol{\psi}^{-1} \left(\hat{\mathbf{Y}} - \mathbf{X}\hat{\boldsymbol{\beta}}_{ML} \right) \quad \textbf{(2.13)}$$

The term $\left(\hat{\mathbf{Y}} - \mathbf{X}\hat{\boldsymbol{\beta}}_{ML}\right)$ is called the residual vector.

For the measurement error assumptions, (1 1 – – 1 1), it is easy to show that the maximum likelihood parameter estimates are unbiased; i.e.

$$E\left(\hat{\boldsymbol{\beta}}_{ML}\right) = E\left(\left(\mathbf{X}^T\boldsymbol{\psi}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}^T\boldsymbol{\psi}^{-1}\hat{\mathbf{Y}}\right) = \left(\left(\mathbf{X}^T\boldsymbol{\psi}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}^T\boldsymbol{\psi}^{-1}\mathbf{X}\boldsymbol{\beta}\right) = \boldsymbol{\beta} \qquad \textbf{(2.14)}$$

This result says that if we conducted a large number of identical parameter estimation experiments, the expected (or mean) value of the estimated parameter vectors is the actual parameter vector.

The covariance matrix for the parameter estimate $\hat{\boldsymbol{\beta}}_{ML}$, denoted by $\mathbf{P}_{ML}$, is given by

$$\mathbf{P}_{ML} = \text{cov}(\hat{\boldsymbol{\beta}}_{ML} - \boldsymbol{\beta}) = E\left[(\hat{\boldsymbol{\beta}}_{ML} - \boldsymbol{\beta})(\hat{\boldsymbol{\beta}}_{ML} - \boldsymbol{\beta})^T\right] = \left(\mathbf{X}^T\boldsymbol{\psi}^{-1}\mathbf{X}\right)^{-1} \qquad \textbf{(2.15)}$$

where $\hat{\boldsymbol{\beta}}_{ML} - \boldsymbol{\beta}$ is the parameter estimation error vector, and

$\boldsymbol{\psi}$ is the covariance matrix of the measurement errors, $\text{cov}(\boldsymbol{\varepsilon})$.

Knowing $\hat{\boldsymbol{\beta}}_{ML}$ and $\mathbf{P}_{ML}$ enables us to write the probability density function for the parameter estimates as the multivariate normal density

$$f(\hat{\boldsymbol{\beta}}_{ML} - \boldsymbol{\beta}) = (2\pi)^{-\frac{p}{2}}\left|\mathbf{P}_{ML}\right|^{-\frac{1}{2}}\exp\left[-\frac{1}{2}(\hat{\boldsymbol{\beta}}_{ML} - \boldsymbol{\beta})^{\mathbf{T}}\mathbf{P}_{ML}{}^{-1}(\hat{\boldsymbol{\beta}}_{ML} - \boldsymbol{\beta})\right] \qquad \textbf{(2.16)}$$

Having the probability density function for the parameter estimates enables us to quantify the quality of the parameter estimates by constructing confidence limits based on the known density function.

The random variable $\dfrac{\hat{\boldsymbol{\beta}}_{ML,i} - \boldsymbol{\beta}_i}{\sqrt{\mathbf{P}_{i.i}}}$ has the standard normal density function $N(0,1)$. A typical confidence limit is

$$\text{Prob}\left[-2 < \frac{\hat{\boldsymbol{\beta}}_{ML,i} - \boldsymbol{\beta}_i}{\sqrt{\mathbf{P}_{i.i}}} < 2\right] = .95 \qquad \textbf{(2.17)}$$

or

$$\text{Prob}\left[\hat{\boldsymbol{\beta}}_{ML,i} - 2\sqrt{\mathbf{P}_{i.i}} < \boldsymbol{\beta}_i < \hat{\boldsymbol{\beta}}_{ML,i} + 2\sqrt{\mathbf{P}_{i.i}}\right] = .95 \qquad \textbf{(2.18)}$$

The confidence limit statement for this case is; "The probability that the true parameter lies within two standard deviations of the estimated parameter is .95." The number of standard deviations of the standard normal density function can be used to obtain any confidence limit desired. This confidence limit example is an approximation. One finds from a standard normal density function table that a more accurate value for .95 probability is 1.96 standard deviations. The student is encouraged to consult a standard normal density function table or appropriate software to verify this result. A table is found in most statistical textbooks and the

standard normal density and distribution is available in software such as Microsoft Excel.

Recall that the least squares estimator of Course G429, equation (1.20), is,

$$\hat{\boldsymbol{\beta}}_{LS} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\hat{\mathbf{Y}} \qquad \textbf{(2.19)}$$

Comparing the maximum likelihood estimator of equation (2.12) with the least squares estimator of equation (2.19) we see that the maximum likelihood estimator weights the measurements by the appropriate combinations of both ψ and $\mathbf{X}$ matrices. In the case where the measurement errors are constant variance and uncorrelated, (1 1 1 1 1 1),

$$\boldsymbol{\psi} = \sigma^2\mathbf{I} \qquad \textbf{(2.20)}$$

and the maximum likelihood and least squares estimates are identical.

Suppose the measurement errors are uncorrelated but the variances are nonconstant. For this case ψ is the diagonal matrix

$$\boldsymbol{\psi} = diag\left[\sigma_1^2 \quad \sigma_2^2 \quad . \quad . \quad . \quad \sigma_m^2\right] \qquad \textbf{(2.21)}$$

and

$$\boldsymbol{\psi}^{-1} = diag\left[\frac{1}{\sigma_1^2} \quad \frac{1}{\sigma_2^2} \quad . \quad . \quad . \quad \frac{1}{\sigma_m^2}\right] \qquad \textbf{(2.22)}$$

Substituting equation (2.22) in equation (2.13) we have

$$\mathbf{R}_{ML} = Min(\mathbf{S}_{ML}) = \left(\hat{\mathbf{Y}} - \mathbf{X}\hat{\boldsymbol{\beta}}_{ML}\right)^T \begin{bmatrix} \frac{1}{\sigma_1^2} & 0 & . & . & 0 \\ 0 & \frac{1}{\sigma_2^2} & . & . & 0 \\ . & . & . & . & . \\ . & . & . & . & . \\ 0 & 0 & . & . & \frac{1}{\sigma_m^2} \end{bmatrix} \left(\hat{\mathbf{Y}} - \mathbf{X}\hat{\boldsymbol{\beta}}_{ML}\right) \qquad \textbf{(2.23)}$$

We see that each squared residual is multiplied by the inverse variance of the corresponding measurement error.

Sometimes the term "weighted least squares" is used to describe the suboptimal approach of assigning subjective weights to the diagonal elements of ψ (or ψ^{-1}) and applying the maximum likelihood estimator of Equation (2.12).

Estimation of Measurement Error Variance

If ψ can be written as

$$\boldsymbol{\psi} = \sigma^2 \boldsymbol{\Omega} \tag{2.24}$$

where σ^2 is unknown but Ω is known, the assumption set becomes (1 1 – – 1 0).

Equation (2.9) becomes

$$L(\boldsymbol{\beta}/\hat{\mathbf{Y}}) = (2\pi)^{-\frac{m}{2}} |\boldsymbol{\Omega}|^{-\frac{1}{2}} \sigma^{-2} \exp\left[-\frac{\sigma^{-2}}{2} (\hat{\mathbf{Y}} - \mathbf{X}\boldsymbol{\beta})^{\mathbf{T}} \boldsymbol{\Omega}^{-1} (\hat{\mathbf{Y}} - \mathbf{X}\boldsymbol{\beta}) \right] \tag{2.25}$$

and equations (2.10) and (2.11) become

$$l(\boldsymbol{\beta}/\hat{\mathbf{Y}}) = \ln L(\boldsymbol{\beta}/\hat{\mathbf{Y}}) = -\frac{1}{2}\left[m \ln(2\pi) + \ln\left(\sigma^{2m} |\boldsymbol{\Omega}|\right) + \mathbf{S}_{ML} \right] \tag{2.26}$$

where

$$\mathbf{S}_{ML} = \sigma^{-2} (\hat{\mathbf{Y}} - \mathbf{X}\boldsymbol{\beta})^{\mathbf{T}} \boldsymbol{\Omega}^{-1} (\hat{\mathbf{Y}} - \mathbf{X}\boldsymbol{\beta}) \tag{2.27}$$

respectively.

By taking the derivative of equation (2.26) with respect to the parameters and σ^2; then setting the result to zero yields the maximum likelihood estimators,

$$\hat{\boldsymbol{\beta}}_{ML} = \left(\mathbf{X}^T \boldsymbol{\Omega}^{-1} \mathbf{X} \right)^{-1} \mathbf{X}^T \boldsymbol{\Omega}^{-1} \hat{\mathbf{Y}} \tag{2.28}$$

$$\hat{\sigma}^2_{ML} = \frac{1}{m} (\hat{\mathbf{Y}} - \mathbf{X}\hat{\boldsymbol{\beta}}_{ML})^{\mathbf{T}} \boldsymbol{\Omega}^{-1} (\hat{\mathbf{Y}} - \mathbf{X}\hat{\boldsymbol{\beta}}_{ML}) \tag{2.29}$$

The latter estimator is biased; that is,

$$E(\hat{\sigma}^2_{ML}) = \left(1 - \frac{p}{m}\right) \sigma^2 \neq \sigma^2 \tag{2.30}$$

Note that the maximum likelihood estimator of σ^2 is asymptotically unbiased, i.e. it becomes unbiased as the number of measurements becomes large.

An unbiased estimator for σ^2 is,

$$s^2 = \frac{1}{m - p} (\hat{\mathbf{Y}} - \mathbf{X}\hat{\boldsymbol{\beta}}_{ML})^{\mathbf{T}} \boldsymbol{\Omega}^{-1} (\hat{\mathbf{Y}} - \mathbf{X}\hat{\boldsymbol{\beta}}_{ML}) \tag{2.31}$$

where p is the number of parameters being estimated. The latter estimator is preferred even though it is <u>not</u> the maximum likelihood estimator for σ^2. Note that p is the dimension of the β vector and does not include the unknown variance σ^2.

The covariance matrix for the parameter estimate for this case is given by

$$\mathbf{P}_{ML} = \text{cov}(\hat{\boldsymbol{\beta}}_{ML} - \boldsymbol{\beta}) = \sigma^2 \left(\mathbf{X}^T \boldsymbol{\Omega}^{-1} \mathbf{X} \right)^{-1} \tag{2.32}$$

We don't know σ^2 but we can estimate it using Equation (2.31). Let

$$\tilde{\mathbf{P}} = s^2\left(\mathbf{X}^T\boldsymbol{\Omega}^{-1}\mathbf{X}\right)^{-1} \qquad \textbf{(2.33)}$$

The estimated standard error for parameter β_i is $\sqrt{\tilde{\mathbf{P}}_{i,i}}$. Since we are estimating σ^2, the t distribution, rather than the normal distribution, must be used to construct confidence limits. For the $100(1-\alpha)\%$ confidence interval we have

$$\text{Prob}\left[\hat{\boldsymbol{\beta}}_i - \sqrt{\tilde{\mathbf{P}}_{i,i}}t_{1-\alpha/2}(m-p) < \boldsymbol{\beta}_i < \hat{\boldsymbol{\beta}}_i + \sqrt{\tilde{\mathbf{P}}_{i,i}}t_{1-\alpha/2}(m-p)\right] = 1-\alpha \qquad \textbf{(2.34)}$$

where

$t_{1-\alpha/2}(m-p)$ is the t statistic for m-p degrees of freedom

m is number of measurements

p is the number of estimated parameters (does not include s^2)

α is the level of confidence

For 95% confidence and m-p=10, we find from a table of the t distribution that

$$t_{1-\alpha/2}(m-p) = t_{1-.05/2}(10) = t_{.975}(10) = 2.23 \qquad \textbf{(2.35)}$$

Then substituting in Equation (2.34) we have

$$\text{Prob}\left[\hat{\boldsymbol{\beta}}_i - 2.23\sqrt{\tilde{\mathbf{P}}_{i,i}} < \boldsymbol{\beta}_i < \hat{\boldsymbol{\beta}}_i + 2.23\sqrt{\tilde{\mathbf{P}}_{i,i}}\right] = .95 \qquad \textbf{(2.36)}$$

Comparing Equations (2.18) and (2.36) we see that the normal distribution gives tighter confidence limits than the t distribution.

See http://www.itl.nist.gov/div898/handbook/eda/section3/eda3672.htm for a table of t distribution values.

Thermal Conductivity Example

The thermal conductivity, k, of Armco iron is measured over temperature at two different input heater power levels. The temperature and power can be measured much more accurately than the thermal conductivity. From previous results we know that the variance in the measurements at low power is four times the variance at high power. Assume that the measurement errors have a normal probability distribution. Given the data below determine the maximum likelihood estimates of the parameters in the model $k = \beta_1 + \beta_2 T$ and $\hat{\sigma}^2_{ML}$.

Run	Temp (°F)	Power Level	k(Btu/hr-ft-°F)
1	100	High	41.60
2	90	Low	42.35
3	227	High	36.50
4	206	Low	37.35
5	362	High	34.53
6	352	Low	33.92

The assumption set is (1 1 0 1 1 0) and the measurements are given by

$$\hat{Y}_i = k_i + \varepsilon_i = \beta_1 + \beta_2 T_i + \varepsilon_i; \quad i = 1,6 \tag{2.37}$$

In matrix notation

$$\hat{\mathbf{Y}} = \begin{bmatrix} 41.60 \\ 42.35 \\ 36.50 \\ 37.35 \\ 34.53 \\ 33.92 \end{bmatrix} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} = \begin{bmatrix} 1 & 100 \\ 1 & 90 \\ 1 & 227 \\ 1 & 206 \\ 1 & 362 \\ 1 & 352 \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix} + \boldsymbol{\varepsilon} \tag{2.38}$$

$$\boldsymbol{\psi} = \sigma^2 \boldsymbol{\Omega} = \sigma^2 diag\begin{bmatrix} 1 & 4 & 1 & 4 & 1 & 4 \end{bmatrix} \tag{2.39}$$

$$\boldsymbol{\Omega}^{-1} = diag\begin{bmatrix} 1 & .25 & 1 & .25 & 1 & .25 \end{bmatrix} \tag{2.40}$$

Performing the mathematical operations gives

$$\hat{\boldsymbol{\beta}}_{ML} = \left(\mathbf{X}^T \boldsymbol{\Omega}^{-1} \mathbf{X}\right)^{-1} \mathbf{X}^T \boldsymbol{\Omega}^{-1} \hat{\mathbf{Y}} = \begin{bmatrix} 43.927 \\ -.02784 \end{bmatrix} \tag{2.41}$$

and

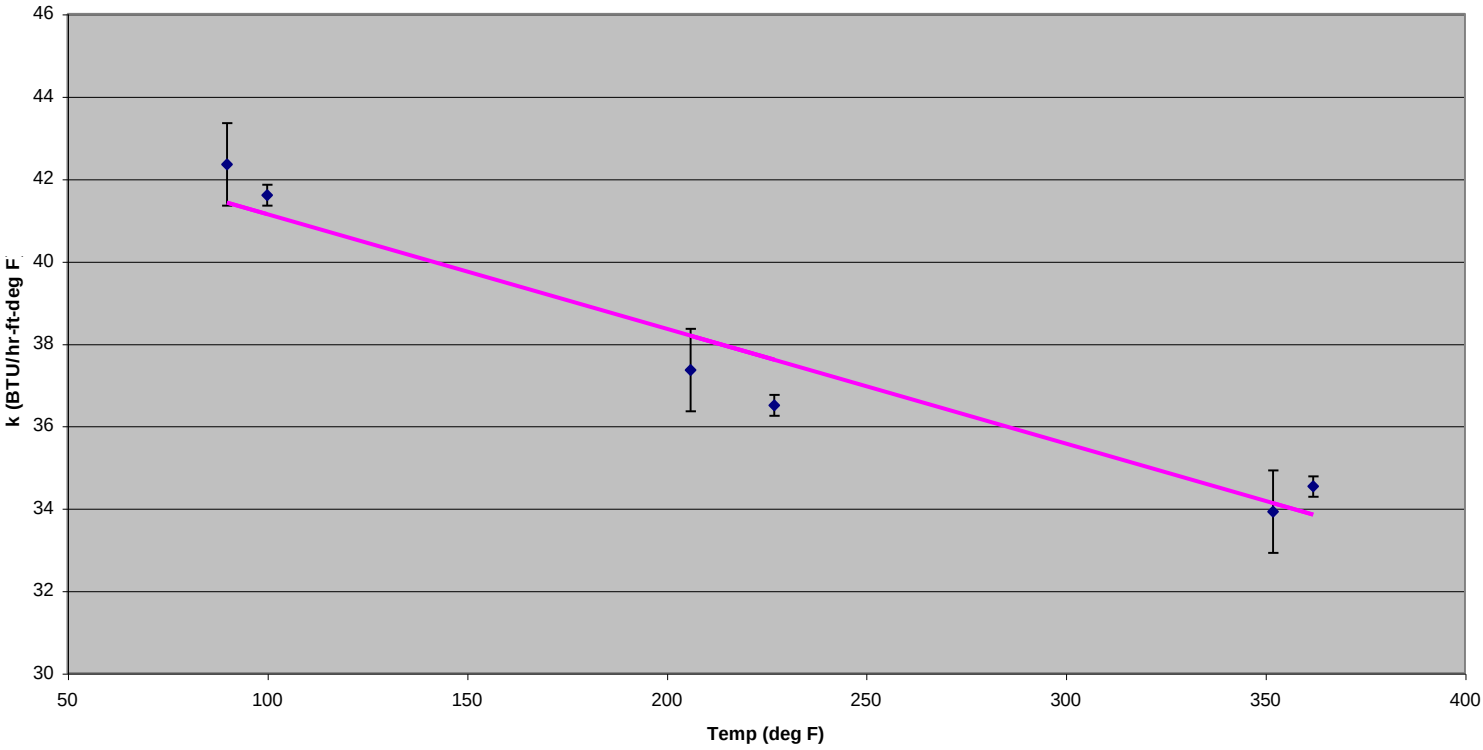
$$\hat{\sigma}^2_{ML} = \frac{1}{m} (\hat{\mathbf{Y}} - \mathbf{X}\hat{\boldsymbol{\beta}}_{ML})^{\mathbf{T}} \boldsymbol{\Omega}^{-1} (\hat{\mathbf{Y}} - \mathbf{X}\hat{\boldsymbol{\beta}}_{ML}) = .3837 \tag{2.42}$$

Plotting the measurements along with the line

$$Y = k_{ML} = \beta_{1,ML} + \beta_{2,ML} T = 43.927 - .02784T \tag{2.43}$$

reveals that the residuals are not random.

Thermal Conductivity of Armco Iron



The length of the vertical lines through the measurements indicates the relative errors between the measurements taken at high and low power levels. The estimated parameters are affected more by the measurements taken at high power than the ones at low power.

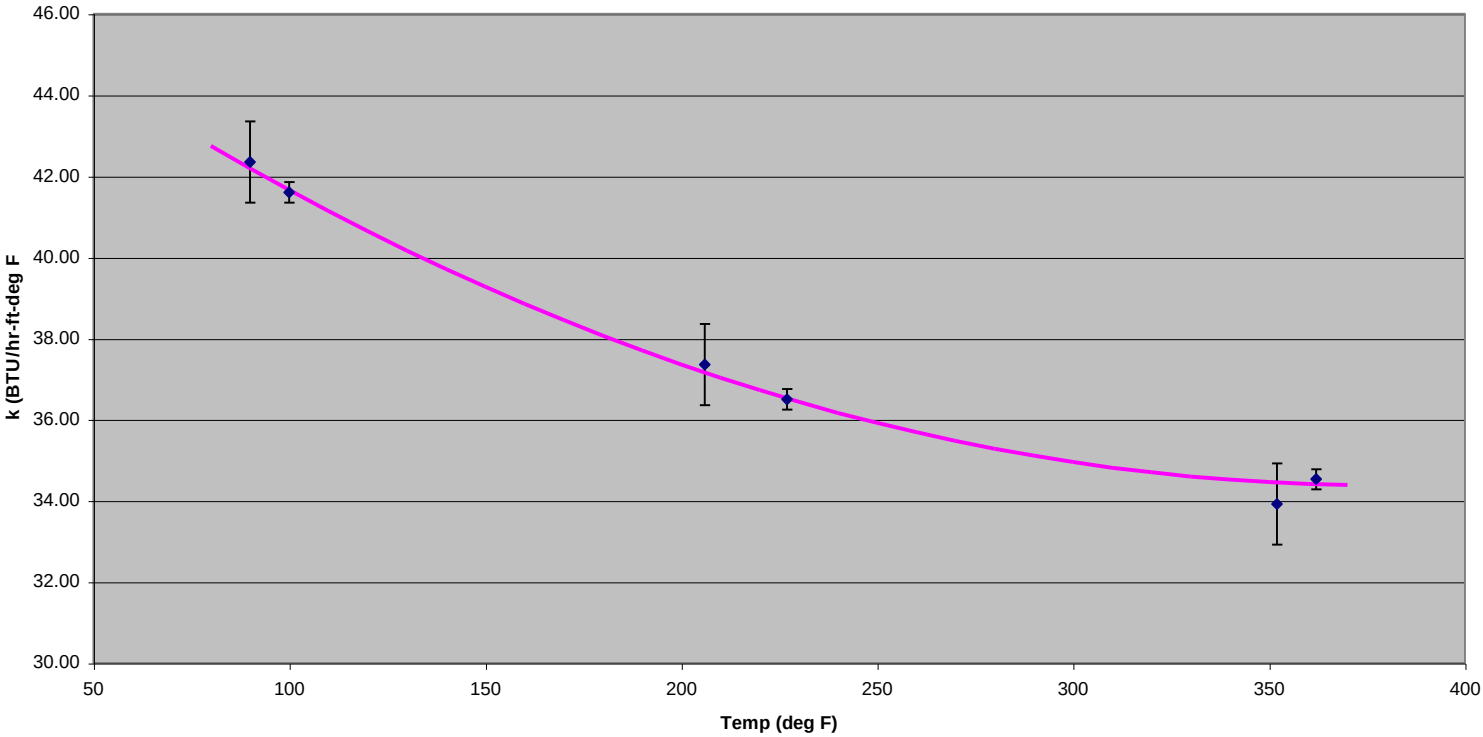
Because the residual corresponding to the measurement taken at high power at Temp=227°F is not "close" to the line, we add another term to our model to obtain $k = \beta_1 + \beta_2 T + \beta_3 T^2$. Carrying out analogous operations to those above yields

$$\hat{\boldsymbol{\beta}}_{ML} = \begin{bmatrix} 47.8467 \\ -7.1386E-2 \\ 9.4668E-5 \end{bmatrix} \qquad \textbf{(2.44)}$$

and

$$\hat{\sigma}^2_{ML} = .017195 \qquad \textbf{(2.45)}$$

The chart is



The residuals are much more consistent for the three parameter model and the estimate of the measurement error variance is reduced by a factor of 22!

When more rigor than the residual "eyeball" test is desired, the F distribution can be used to guide choosing among competing models. Use of the F distribution for model building depends on the normal probability assumption and is beyond the scope of this introductory course.

We now construct 90% confidence limits for β_1. First use Equation (2.31) to compute an unbiased estimate of σ^2.

$$s^2 = \frac{1}{m-p}(\hat{\mathbf{Y}} - \mathbf{X}\hat{\boldsymbol{\beta}}_{ML})^{\mathrm{T}}\boldsymbol{\Omega}^{-1}(\hat{\mathbf{Y}} - \mathbf{X}\hat{\boldsymbol{\beta}}_{ML}) = \frac{0.103169}{6-3} = 0.03439 \qquad \textbf{(2.46)}$$

From Equation (2.33) we have

$$\tilde{\mathbf{P}} = s^2\left(\mathbf{X}^{\mathrm{T}}\boldsymbol{\Omega}^{-1}\mathbf{X}\right)^{-1} = 0.03439\begin{bmatrix} 8.45\text{E+0} & -8.29\text{E-2} & 1.69\text{E-4} \\ -8.29\text{E-2} & 8.85\text{E-4} & -1.87\text{E-6} \\ 1.69\text{E-4} & -1.87\text{E-6} & 4.08\text{E-9} \end{bmatrix} \qquad \textbf{(2.47)}$$

Substituting into Equation (2.34) gives

$$\mathrm{Prob}\left[\hat{\beta}_1 - .539t_{.95}(3) < \beta_1 < \hat{\beta}_1 + .539t_{.95}(3)\right] = .9 \qquad \textbf{(2.48)}$$

From t distribution tables we determine that

$$t_{.95}(3) = 2.353 \qquad \textbf{(2.49)}$$

so finally we have

$$\mathrm{Prob}\left[47.85 - 1.27 < \beta_1 < 47.85 + 1.27\right] = \mathrm{Prob}\left[46.58 < \beta_1 < 49.12\right] = .9 \qquad \textbf{(2.50)}$$

Confidence limits for the remaining parameters are constructed in the same manner.

We now turn to an estimation technique that does not require the assumption of normally distributed measurement errors.

Gauss-Markov Estimation

Let the measurement errors be additive, zero mean with a covariance matrix of known form but with an unknown multiplier. The assumption set is (1 1 – – – 0).

Once again the model is

$$\hat{\mathbf{Y}} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \qquad \textbf{(2.51)}$$

and

$$\boldsymbol{\psi} = \sigma^2\boldsymbol{\Omega} \qquad \textbf{(2.52)}$$

The Gauss-Markov Theorem states that the minimum variance, linear, unbiased estimator of $\boldsymbol{\beta}$ is given by

$$\hat{\boldsymbol{\beta}}_{GM} = \left(\mathbf{X}^T\boldsymbol{\Omega}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}^T\boldsymbol{\Omega}^{-1}\hat{\mathbf{Y}} \qquad \textbf{(2.53)}$$

and the associated covariance matrix for the parameter estimation errors is

$$\mathbf{P}_{GM} = \text{cov}(\hat{\boldsymbol{\beta}}_{GM} - \boldsymbol{\beta}) = \sigma^2\left(\mathbf{X}^T\boldsymbol{\Omega}^{-1}\mathbf{X}\right)^{-1} \qquad \textbf{(2.54)}$$

The minimum variance property of the Gauss-Markov Theorem means that the variances along the diagonal of the $\mathbf{P}_{GM}$ matrix are the smallest possible. (Recall that the sum of the diagonal elements of a square matrix is called its trace.)

Note that the form of the Gauss-Markov estimator and associated covariance matrix, equations (2.53) and (2.54), are identical in form to those of the maximum likelihood estimator given in equations (2.28) & (2.32).

An unbiased estimator of σ^2 is

$$s^2 = \frac{1}{m-p}(\hat{\mathbf{Y}} - \mathbf{X}\hat{\boldsymbol{\beta}}_{GM})^{\mathrm{T}}\boldsymbol{\Omega}^{-1}(\hat{\mathbf{Y}} - \mathbf{X}\hat{\boldsymbol{\beta}}_{GM}) \qquad \textbf{(2.55)}$$

The statistical distribution of the measurement errors is not required to use the Gauss-Markov estimator. We are assured that the Gauss-Markov estimates have minimum variance of all linear estimators. Since we don't know the probability distribution of errors in the parameter estimates, we cannot make confidence statements like we can when normally distributed measurement errors are assumed. Neither can we use the F distribution for model building.

It is important to understand that there may exist nonlinear estimators that have lower sum-of-variances than the Gauss-Markov estimator.

Comparison of Three Estimation Methods

Table 1 summarizes and compares the parameter estimators for least squares, maximum likelihood, and Gauss-Markov estimation. The table covers the content of both this course and Course G429.

Least squares is the most basic of the three methods. It can be applied without any assumptions being made about the measurement errors. The estimated parameters minimize the sum of squares of the residuals; that is the difference between the measurements and the model with the estimated parameters. If we assume the measurement errors are additive and zero mean the covariance of the estimates can be computed IF the covariance of the measurement errors is known. This is the case only if the sixth position in our assumption set is "1." If ψ is known and the measurement errors are normally distributed the maximum likelihood estimator should be used instead of least squares. In the special case of $\psi = \mathbf{I}$, the maximum likelihood and least squares estimators are the same.

If the measurement errors are additive, zero-mean, constant variance, and uncorrelated, the structure of the covariance matrix of the parameter estimates is known to be of the form $\psi = \sigma^2\mathbf{I}$. The constant variance can be estimated from the sum-of-squares function if it is not known.

The advantage of maximum likelihood and Gauss-Markov estimation over least squares is the flexibility to include a general measurement error covariance matrix in the estimation process. As we have learned, the maximum likelihood method is based on the maximization of a likelihood function. The likelihood function describes how likely the parameter estimates are, given the specific measurements. The minimization of the weighted sum-of-squares function depends on the assumption of normally distributed measurement errors. The covariance matrix of the maximum likelihood parameter estimates is straightforward to calculate when ψ is completely known. When $\psi = \sigma^2\Omega$, $\hat{\sigma}^2$ is computed by either the maximum likelihood estimator or the alternative unbiased estimator.

Since the measurement errors are normally distributed for maximum likelihood estimation, confidence limits can be constructed around the parameter estimates to quantify their accuracy. The standard normal density function is used to construct confidence limits when ψ is known. The t distribution is used when the measurement error variance is estimated. The unbiased estimate of the measurement error variance must be used rather than the biased maximum likelihood estimate.

Finally, the Gauss-Markov estimator can be used when the normal density assumption cannot be made. It is identical in form to the maximum

likelihood estimator; however, confidence limits and model building extensions that are available for the maximum likelihood estimator are not available for the Gauss-Markov method.

Table 1. Summary of Estimators for Linear Model $\boldsymbol{\eta = X\beta}$

Name	Assumptions	Estimator	$\text{cov}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})$	Estimator of σ^2; $\boldsymbol{\psi} = \sigma^2\boldsymbol{\Omega}$
Least Squares	**– – – – – –**	$\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\hat{\mathbf{Y}}$		
Least Squares	**1 1 – – – –**	$\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\hat{\mathbf{Y}}$	$\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\boldsymbol{\psi}\mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}$	
Least Squares	**1 1 1 1 – –**	$\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\hat{\mathbf{Y}}$	$\sigma^2\left(\mathbf{X}^T\mathbf{X}\right)^{-1}$	$\frac{1}{m-p}(\hat{\mathbf{Y}} - \mathbf{X}\hat{\boldsymbol{\beta}}_{LS})^{\mathrm{T}}(\hat{\mathbf{Y}} - \mathbf{X}\hat{\boldsymbol{\beta}}_{LS})$
Maximum Likelihood	**1 1 – – 1 1**	$\left(\mathbf{X}^T\boldsymbol{\psi}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}^T\boldsymbol{\psi}^{-1}\hat{\mathbf{Y}}$	$\left(\mathbf{X}^T\boldsymbol{\psi}^{-1}\mathbf{X}\right)^{-1}$	
Maximum Likelihood	**1 1 – – 1 0**	$\left(\mathbf{X}^T\boldsymbol{\Omega}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}^T\boldsymbol{\Omega}^{-1}\hat{\mathbf{Y}}$	$\sigma^2\left(\mathbf{X}^T\boldsymbol{\Omega}^{-1}\mathbf{X}\right)^{-1}$	$\frac{1}{m}(\hat{\mathbf{Y}} - \mathbf{X}\hat{\boldsymbol{\beta}}_{ML})^{\mathrm{T}}\boldsymbol{\Omega}^{-1}(\hat{\mathbf{Y}} - \mathbf{X}\hat{\boldsymbol{\beta}}_{ML})$
Gauss-Markov	**1 1 – – – 0**	$\left(\mathbf{X}^T\boldsymbol{\Omega}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}^T\boldsymbol{\Omega}^{-1}\hat{\mathbf{Y}}$	$\sigma^2\left(\mathbf{X}^T\boldsymbol{\Omega}^{-1}\mathbf{X}\right)^{-1}$	$\frac{1}{m-p}(\hat{\mathbf{Y}} - \mathbf{X}\hat{\boldsymbol{\beta}}_{GM})^{\mathrm{T}}\boldsymbol{\Omega}^{-1}(\hat{\mathbf{Y}} - \mathbf{X}\hat{\boldsymbol{\beta}}_{GM})$

Summary

This course develops maximum likelihood and Gauss-Markov estimators for a linear measurement model. These estimators are valuable extensions to the least squares estimator presented in Course G429, Linear Least Squares Parameter Estimation. The maximum likelihood estimator accommodates any valid measurement error covariance matrix when the measurement errors are normally distributed. In addition, the measurement error variance can be estimated if the covariance structure of the errors is known. The maximum likelihood estimator for the measurement error variance is developed. It is biased; however, an unbiased estimator is presented. Confidence limits for the parameter estimates are derived for both the known error covariance matrix and unknown variance with known covariance structure cases.

A detailed thermal conductivity example is presented that applies the maximum likelihood estimation method to a practical problem.

The Gauss-Markov estimator does not require the normality assumption for measurement errors yet is the minimum variance unbiased linear estimator for the parameters in the linear measurement model. The Gauss-Markov estimator also accommodates any valid measurement error covariance matrix. Many practical problems for which measurement errors are not normally distributed can be solved using Gauss-Markov estimation. Unfortunately confidence limits cannot be constructed for the general measurement error distribution case. The measurement error variance can be estimated if the covariance structure of the errors is known.

A concise table is presented that compares the least squares, maximum likelihood, and Gauss-Markov parameter estimators. It should prove valuable to those tasked with applying parameter estimation to engineering and scientific problems.