PDHonline Course G429 (4 PDH)

Linear Least Squares 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

Linear Least Squares Parameter Estimation

Course Content

Introduction

Professional engineers are often asked to make customer recommendations based on a limited set of uncertain measurements of a physical system or process. Mathematical models and statistical techniques can be used to provide the theoretical foundation that enables reliable, supportable recommendations. The purpose of this course is to provide the student with the necessary understanding that enables such recommendations.

Parameters are constants found in mathematical models of systems or processes. Parameter estimation is a discipline that provides estimates of unknown parameters in a system or process model based on measurement data. Parameter estimation is a very broad subject that cuts a broad swath through engineering and statistical inference. Because parameter estimation is used in so many different academic and application areas, the terminology can be confusing to the uninitiated.

In this course we present an introductory overview of least squares estimation, the most widely applied area of parameter estimation, with a focus on linear system models.

Mathematical Models

Mathematics is the language of parameter estimation. We must have a mathematical model of the system or process in order to use parameter estimation.

General Case

We first develop the general mathematical framework.

Let the system or process model be described by

$$\mathbf{Y} = \boldsymbol{\eta}(\mathbf{X}, \boldsymbol{\beta}) \qquad (1.1)$$

where

- $\mathbf{Y}$ **is an m-dimensional vector of dependent variables**
- $\mathbf{X}$ **is an n-dimensional vector of independent variables**
- $\boldsymbol{\beta}$ **is a p-dimensional vector of unknown parameters**
- $\boldsymbol{\eta}$ **is a m-dimensional vector function of known form**

Parameters are unknown constants that appear in the model.

Now we measure Y with some additive error ε

$$\hat{Y} = Y + \varepsilon = \eta(X,\beta) + \varepsilon \qquad (1.2)$$

where

$\hat{Y}$ is an m-dimensional random vector of measured dependent variables, and

ε is an m-dimensional random vector of measurement errors

This general model is applicable to a wide variety of systems and processes. Unfortunately there are no closed-form parameter estimators available for this general case. There are closed-form parameter estimators available for the following more restrictive case.

Linear-in-the-Parameters Model

If Equation (1.2) can be written as

$$\hat{Y} = \eta(X)\beta + \varepsilon \qquad (1.3)$$

where

$\eta(X)$ is an mxp dimensional matrix of functions of the independent variables,

then the parameter estimation problem is said to be linear-in-the-parameters. The term $\eta(X)$ can be composed of nonlinear functions that relate the independent variables and parameter vector to the dependent variables.

A still further simplification is the linear algebraic model.

Linear Model

Sometimes $\eta(X)$ is composed of just the independent variables so that we can write

$$\hat{Y} = X\beta + \varepsilon \qquad (1.4)$$

where

X is an mxp dimensional matrix of independent variables

Equations (1.4) and (1.3) are the models we use in this introductory course.

Statistical Assumptions for Measurement Errors

Accurately characterizing the uncertainty in the parameter estimation problem is the key to producing (and defending) results to customers and other interested parties. The assumptions made about the measurement errors influence, and often dictate, the parameter estimation technique applied to a set of experimental measurements. Least squares estimation requires the fewest assumptions about the measurement errors and is the only parameter estimation technique presented in this course.

The following is a set of six statistical assumptions about the measurement errors that might be made in a parameter estimation problem. [Beck, J. V., & K. J. Arnold, *Parameter Estimation in Engineering and Science,* Wiley, 1977]

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

Equation (1.2) is repeated here for convenience

$$\hat{\mathbf{Y}} = \mathbf{Y} + \boldsymbol{\varepsilon} = \boldsymbol{\eta}(\mathbf{X}, \boldsymbol{\beta}) + \boldsymbol{\varepsilon} \qquad (1.5)$$

The measurement error vector, $\boldsymbol{\varepsilon}$, is composed of m random variables. The term $\boldsymbol{\eta}(\mathbf{X}, \boldsymbol{\beta})$ is not a random variable, but rather is a vector function of the independent variables, $\mathbf{X}$, and the parameter vector, $\boldsymbol{\beta}$. The independent variables are assumed known. The parameters, elements of the parameter vector, are unknown but constant during the experiment or measurement interval. The m-dimensional measurement vector, $\hat{\mathbf{Y}}$, is a random vector because of the measurement errors.

The necessary statistical characterization of $\boldsymbol{\varepsilon}$ for the purposes of parameter estimation as treated in this course consists of its expected value (often called the mean), covariance, and probability density function.

The expected value of the measurement error vector, $\boldsymbol{\varepsilon}$, is given by

$$\mathbf{E}(\varepsilon) = \begin{bmatrix} \mathbf{E}(\varepsilon_1) \\ \mathbf{E}(\varepsilon_2) \\ . \\ . \\ \mathbf{E}(\varepsilon_m) \end{bmatrix} \qquad (1.6)$$

The covariance matrix of the measurement error vector, ε, is given by

$$\mathbf{cov}(\varepsilon) = \mathbf{E}\left\{ \begin{bmatrix} \varepsilon_1 - \mathbf{E}(\varepsilon_1) \\ \varepsilon_2 - \mathbf{E}(\varepsilon_2) \\ . \\ . \\ \varepsilon_m - \mathbf{E}(\varepsilon_m) \end{bmatrix} \begin{bmatrix} \varepsilon_1 - \mathbf{E}(\varepsilon_1) & \varepsilon_2 - \mathbf{E}(\varepsilon_2) & . & . & \varepsilon_m - \mathbf{E}(\varepsilon_m) \end{bmatrix} \right\} \qquad (1.7)$$

$$= \begin{bmatrix} \mathbf{cov}(\varepsilon_1,\varepsilon_1) & \mathbf{cov}(\varepsilon_1,\varepsilon_2) & . & . & \mathbf{cov}(\varepsilon_1,\varepsilon_m) \\ \mathbf{cov}(\varepsilon_2,\varepsilon_1) & \mathbf{cov}(\varepsilon_2,\varepsilon_2) & . & . & \mathbf{cov}(\varepsilon_2,\varepsilon_m) \\ . & . & . & . & . \\ . & . & . & . & . \\ \mathbf{cov}(\varepsilon_m,\varepsilon_1) & \mathbf{cov}(\varepsilon_m,\varepsilon_2) & . & . & \mathbf{cov}(\varepsilon_m,\varepsilon_m) \end{bmatrix}$$

Note that $\mathbf{cov}(\varepsilon)$ is a $\mathbf{mxm}$ dimensional, square, symmetric matrix.

We now discuss each measurement error assumption sequentially.

1. **Additive**
   - **0 No, not additive**
   - **1 Yes, additive**

The measurement errors are additive by the formulation of the parameter estimation problem in Equation (1.2).

2. **Zero-mean**
   - **0 No, not zero-mean**
   - **1 Yes, zero-mean**

The measurement errors are zero-mean if

$$\mathbf{E}(\varepsilon) = \begin{bmatrix} \mathbf{E}(\varepsilon_1) \\ \mathbf{E}(\varepsilon_2) \\ . \\ . \\ \mathbf{E}(\varepsilon_m) \end{bmatrix} = \begin{bmatrix} \mathbf{0} \\ \mathbf{0} \\ . \\ . \\ \mathbf{0} \end{bmatrix} \qquad (1.8)$$

In situations where non-zero-mean errors are suspected, the mean (assumed to be constant for all m measurements) can be considered a parameter and estimated along with the other parameters in the parameter estimation problem.

3. Constant Variance
- **0 No, nonconstant variance**
- **1 Yes, constant variance**

The diagonal of any covariance matrix contains the variances of the elements of the associated random vector. Thus for our measurement error vector, the constant variance assumption implies that Equation (1.7) can be written

$$\mathrm{cov}(\varepsilon) = \begin{bmatrix} \sigma^2 & \mathrm{cov}(\varepsilon_1,\varepsilon_2) & . & . & \mathrm{cov}(\varepsilon_1,\varepsilon_m) \\ \mathrm{cov}(\varepsilon_1,\varepsilon_2) & \sigma^2 & . & . & \mathrm{cov}(\varepsilon_2,\varepsilon_m) \\ . & . & . & . & . \\ . & . & . & . & . \\ \mathrm{cov}(\varepsilon_1,\varepsilon_m) & \mathrm{cov}(\varepsilon_2,\varepsilon_m) & . & . & \sigma^2 \end{bmatrix} \tag{1.9}$$

where

$$\mathrm{cov}(X,Y) = E\left[(X-\mu_X)(Y-\mu_Y)\right] \tag{1.10}$$

and the mean (or expected) value is defined as

$$\mu_X = E(X) \tag{1.11}$$

Recalling that the correlation coefficient between two random variables, X and Y is defined by

$$\rho_{X,Y} = \frac{\mathrm{cov}(X,Y)}{\sigma_X \sigma_Y} \tag{1.12}$$

Equation (1.9) can be written

$$\mathrm{cov}(\varepsilon) = \sigma^2 \begin{bmatrix} 1 & \rho_{\varepsilon_1,\varepsilon_2} & . & . & \rho_{\varepsilon_1,\varepsilon_m} \\ \rho_{\varepsilon_1,\varepsilon_2} & 1 & . & . & \rho_{\varepsilon_2,\varepsilon_m} \\ . & . & . & . & . \\ . & . & . & . & . \\ \rho_{\varepsilon_1,\varepsilon_m} & \rho_{\varepsilon_2,\varepsilon_m} & . & . & 1 \end{bmatrix} \tag{1.13}$$

4. Uncorrelated
- **0 No, correlated errors**
- **1 Yes, uncorrelated errors**

It follows from Equation (1.12) that if the correlation coefficient between two random variables is zero, then the covariance between the two random variables is also zero. Using this fact, the covariance matrix for uncorrelated measurement errors is

$$\mathbf{cov}(\varepsilon) = \begin{bmatrix} \sigma_1^2 & 0 & . & . & 0 \\ 0 & \sigma_2^2 & . & . & 0 \\ . & . & . & . & . \\ . & . & . & . & . \\ 0 & 0 & . & . & \sigma_m^2 \end{bmatrix} \qquad (1.14)$$

If the measurement errors are both constant variance and uncorrelated the covariance matrix simplifies to

$$\mathbf{cov}(\varepsilon) = \sigma^2 \begin{bmatrix} 1 & 0 & . & . & 0 \\ 0 & 1 & . & . & 0 \\ . & . & . & . & . \\ . & . & . & . & . \\ 0 & 0 & . & . & 1 \end{bmatrix} = \sigma^2 \mathbf{I} \qquad (1.15)$$

where $\mathbf{I}$ is the identity matrix.

5. Normal Probability Distribution

- 0 No, nonnormal distribution
- 1 Yes, normal distribution

The normal probability distribution, often called the Gaussian distribution, is widely used in engineering. For a single random variable it is the "bell-shaped curve" described in the following description taken from the web site http://www.stat.yale.edu/Courses/1997-98/101/normal.htm.

A normal distribution has a bell-shaped density curve described by its mean μ and standard deviation σ. The density curve is symmetrical, centered about its mean, with its spread determined by its standard deviation. The height of a normal density curve at a given point x is given by (the vertical axis tic marks are 0.2 starting from 0)

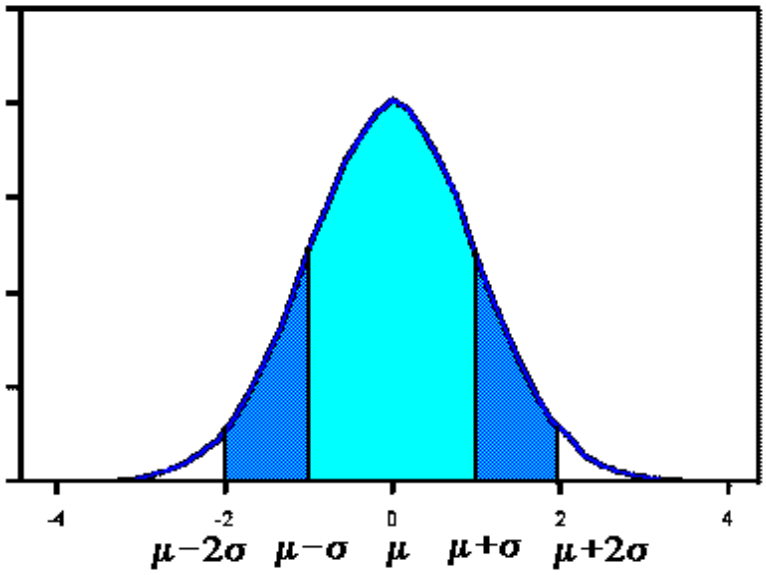


Figure 1. Standard Normal Density Function

The Standard Normal curve, shown here, has mean 0 and standard deviation 1. If a dataset follows a normal distribution, then about 68% of the observations will fall within σ of the mean μ, which in this case is within the interval (-1,1). About 95% of the observations will fall within 2 standard deviations of the mean, which is the interval (-2,2) for the standard normal, and about 99.7% of the observations will fall within 3 standard deviations of the mean, which corresponds to the interval (-3,3) in this case. Although it may appear as if a normal distribution does not include any values beyond a certain interval, the density is actually positive for all values, $(-\infty,+\infty)$. Data from any normal distribution may be transformed into data following the standard normal distribution by subtracting the mean μ and dividing by the standard deviation σ.

The normal distribution has a number of useful properties. An extremely important one is that the mean vector and covariance matrix of a vector of normally distributed random variables completely characterizes the random behavior of the random vector. This property allows us to make probability statements about our confidence in parameter estimates.

6. Known Covariance Matrix
- **0 No, covariance matrix of errors known only to within a multiplicative constant**
- **1 Yes, covariance matrix of errors known**

Let's define

$$\psi = \mathrm{cov}(\varepsilon) \qquad (1.16)$$

If ψ can be written as

$$\psi = \sigma^2 \Omega \qquad (1.17)$$

and σ^2 is unknown but Ω is known, the condition "0" applies.

An analyst should determine which of these six assumptions apply to his parameter estimation problem. A 6-member set is used to specify the assumptions. An assumption that is not or cannot be specified is denoted by "–".

The best-case situation is one in which the errors are additive, zero-mean, constant variance, uncorrelated, have a normal probability distribution, and known covariance matrix. We designate this set of assumptions as, (1 1 1 1 1 1). Note that for this set of assumptions, Equation (1.17) reduces to

$$\psi = \sigma^2 \mathbf{I} \qquad (1.18)$$

and σ^2 is known.

The worst-case situation is one in which we know nothing at all about the measurement errors. This set of assumptions is designated (– – – – – –). Note, however, that our model in Equation (1.2) assumes additive errors, which implies the worst-case situation for our formulation is (1 – – – – –).

An intermediate situation is where we can assume the errors are additive, zero-mean, uncorrelated with constant, but unknown variance. This set of assumptions is (1 1 1 1 – 0).

Least Squares Estimation

The fundamental notion of least squares estimation is that we choose an estimate of the unknown parameters that minimizes the sum of squares of the difference between the model and the measurements, hence the name "least squares." In matrix notation for a linear model we minimize

$$\mathbf{S}_{\mathrm{LS}} = \left(\hat{\mathbf{Y}} - \mathbf{X}\boldsymbol{\beta}\right)^{\mathrm{T}} \left(\hat{\mathbf{Y}} - \mathbf{X}\boldsymbol{\beta}\right) \qquad (1.19)$$

The least squares estimator for the parameter vector is

$$\hat{\boldsymbol{\beta}}_{\mathrm{LS}} = \left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1} \mathbf{X}^{\mathrm{T}}\hat{\mathbf{Y}} \qquad (1.20)$$

and the minimum sum of squares is

$$\mathbf{R}_{\mathrm{LS}} = \mathrm{Min}(\mathbf{S}_{\mathrm{LS}}) = \left(\hat{\mathbf{Y}} - \mathbf{X}\hat{\boldsymbol{\beta}}_{\mathrm{LS}}\right)^{\mathrm{T}} \left(\hat{\mathbf{Y}} - \mathbf{X}\hat{\boldsymbol{\beta}}_{\mathrm{LS}}\right) \qquad (1.21)$$

The term $\left(\hat{\mathbf{Y}} - \mathbf{X}\hat{\boldsymbol{\beta}}_{LS}\right)$ is called the residual vector. It is the difference between the model with the estimated parameters and the measurements.

For additive, zero-mean measurement errors, (1 1 – – – –), it is easy to show that the least squares parameter estimates are unbiased; i.e.

$$\mathbf{E}\left(\hat{\boldsymbol{\beta}}_{LS}\right) = \mathbf{E}\left(\left(\mathbf{X}^{T}\mathbf{X}\right)^{-1}\mathbf{X}^{T}\mathbf{Y}\right) = \left(\left(\mathbf{X}^{T}\mathbf{X}\right)^{-1}\mathbf{X}^{T}\mathbf{X}\boldsymbol{\beta}\right) = \boldsymbol{\beta} \qquad \textbf{(1.22)}$$

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

The least squares estimator can be computed without making any assumptions regarding the errors; i.e. when our error assumption set is (– – – – – –). In this case we can only make subjective assertions about the quality of the parameter estimates by analyzing the residual vector. The residual vector should have a random character. Evident trends or structure in the residual vector indicate that the chosen model is inadequate. An unmodeled phenomenon may be present. Consider using a different model or adding terms, with additional unknown parameters, to the present model.

If the errors are additive, zero-mean and have a known covariance matrix (1 1 – – – 1), we can compute the covariance matrix for the parameter estimate $\hat{\boldsymbol{\beta}}_{LS}$,

$$\begin{aligned}\mathbf{cov}(\hat{\boldsymbol{\beta}}_{LS} - \boldsymbol{\beta}) &= \mathbf{E}\left[(\hat{\boldsymbol{\beta}}_{LS} - \boldsymbol{\beta})(\hat{\boldsymbol{\beta}}_{LS} - \boldsymbol{\beta})^{T}\right] \\ &= \left(\mathbf{X}^{T}\mathbf{X}\right)^{-1}\mathbf{X}^{T}\boldsymbol{\psi}\mathbf{X}\left(\mathbf{X}^{T}\mathbf{X}\right)^{-1} \qquad \textbf{(1.23)}\end{aligned}$$

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

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

Knowing $\mathbf{cov}(\hat{\boldsymbol{\beta}}_{LS} - \boldsymbol{\beta})$ allows us to quantify the quality of our parameter estimates.

There may be a large amount of information in $\boldsymbol{\psi}$. Recall that $\boldsymbol{\psi}$ is symmetric and has dimension m x m where m is the number of measurements. The diagonal elements of $\boldsymbol{\psi}$ are the variances for each of the m measurement errors and the off diagonal elements are the pair-wise covariances of the measurement errors. For a problem with 100 measurements the number of elements of $\boldsymbol{\psi}$ that must be specified is 5,050!

It is rare that we know the detailed statistical structure of the measurement errors ($\boldsymbol{\psi}$) required by Equation (1.23). But there are many practical problems for which it is reasonable to assume that the measurement errors are pair-wise uncorrelated; i.e. the off-diagonal elements of $\boldsymbol{\psi}$ are 0. The corresponding

assumption set is (1 1 – 1– 1}. This reduces the number of non-zero elements of ψ to m. In the problem with 100 measurements we now must specify only 100 variances for the errors! Should the assumption of constant variance for all measurement errors also apply,
(1 1 1 1 – 1), Equation (1.23) reduces to

$$\mathrm{cov}(\hat{\beta}_{LS} - \beta) = \sigma^2 \left(X^T X\right)^{-1} \qquad (1.24)$$

where σ^2 is the known constant variance

The matrix $\left(X^T X\right)^{-1}$ contains the statistical characterization of the parameter estimates.

At this point we have not specified a probability density function for the measurement errors. What we know about the quality, or statistics, of the parameter estimates is their mean value and covariance matrix.

The covariance matrix, given above in Equation (1.23) for assumption set (1 1 – – – 1) and Equation (1.24) for assumption set (1 1 1 1 – 1), specifies the spread of the parameter estimates around the actual parameter vector, β. If the measurement errors are normally distributed as shown with a "1" in the fifth position of the measurement error assumption set, the parameter estimation errors are also normally distributed. For this case the mean vector and associated covariance matrix of the parameter estimation errors are sufficient to calculate the accuracy of the estimates. The random variables $\frac{\hat{\beta}_{LS,i} - \beta_i}{\sigma_i}$ for Equation (1.23) and $\frac{\hat{\beta}_{LS,i} - \beta_i}{\sigma}$ for Equation (1.24) have the standard normal density function given in Figure 1.

Parameter estimation accuracy is usually stated as a confidence interval; for example, for a single parameter estimate we can state that the probability that the actual parameter value is within $\pm\sigma$ of the estimated parameter with probability .68. For two or more parameters the standard deviation interval of the one-dimensional case becomes a 1-σ contour in the appropriate parameter space. Parameter space has the number of dimensions as there are parameters being estimated. The 1-σ contour in two dimensions is an ellipse centered on the estimated parameter vector. The 1-σ contour in three dimensions is an ellipsoid centered on the estimated parameter vector. An n-σ contour or interval is used to make a confidence statement appropriate to the needs of the analyst. In this course we consider only the one-parameter case since this is the most common case encountered in practice.

Frequently we can assume that the measurement errors have constant, but unknown variance. For the assumption set (1 1 1 1 – 0), the constant variance can be estimated from the sum of squares of the residuals by

$$\hat{\sigma}^2 = \frac{\mathbf{R}_{LS}}{\mathbf{m}-\mathbf{p}} = \frac{\left(\hat{\mathbf{Y}} - \mathbf{X}\hat{\boldsymbol{\beta}}_{LS}\right)^T \left(\hat{\mathbf{Y}} - \mathbf{X}\hat{\boldsymbol{\beta}}_{LS}\right)}{\mathbf{m}-\mathbf{p}} \qquad \textbf{(1.25)}$$

and we can compute a covariance matrix analogous to Equation (1.24) by using this estimated variance,

$$\mathrm{cov}(\hat{\boldsymbol{\beta}}_{LS} - \boldsymbol{\beta}) = \hat{\sigma}^2 \left(\mathbf{X}^T\mathbf{X}\right)^{-1} \qquad \textbf{(1.26)}$$

If the measurement errors are also normally distributed so that the assumption set is (1 1 1 1 1 0), we can use Equation (1.26) to compute confidence limits.

Unfortunately the random variable $\frac{\hat{\beta}_{LS,i} - \beta_i}{\hat{\sigma}}$ **is *not* normally distributed with zero mean and unit variance. Instead it has a Student's-t distribution with** $\mathbf{m}-\mathbf{p}$ **degrees of freedom. For a large number of measurements the Student's-t distribution approaches the standard normal distribution. [See http://www.itl.nist.gov/div898/handbook/eda/section3/eda3672.htm for a table of Student's-t distribution values.]**

It should be emphasized that least squares estimation should *not* be used when we know the ψ **matrix and the measurement errors are either correlated or not constant variance. Other parameter estimation techniques (Maximum Likelihood, Gauss-Markov, Maximum a posteriori) provide better parameter estimates for these cases. Use the following table to determine when least squares is the recommended parameter estimation technique.**

Use Least Squares Estimation
(– – – – – –)
(1 – – – – –)
(1 0 – – – –)
(1 1 1 1 1 0)
(1 1 1 1 1 1)
(1 1 1 1 – 0)
(1 1 1 1 – 1)
(1 0 1 1 1 0)
(1 0 1 1 1 1)
(1 0 1 1 – 0)
(1 0 1 1 – 1)

When the measurement errors are not zero-mean a parameter must be used in the model to estimate the mean value.

Next we will work through a detailed example.

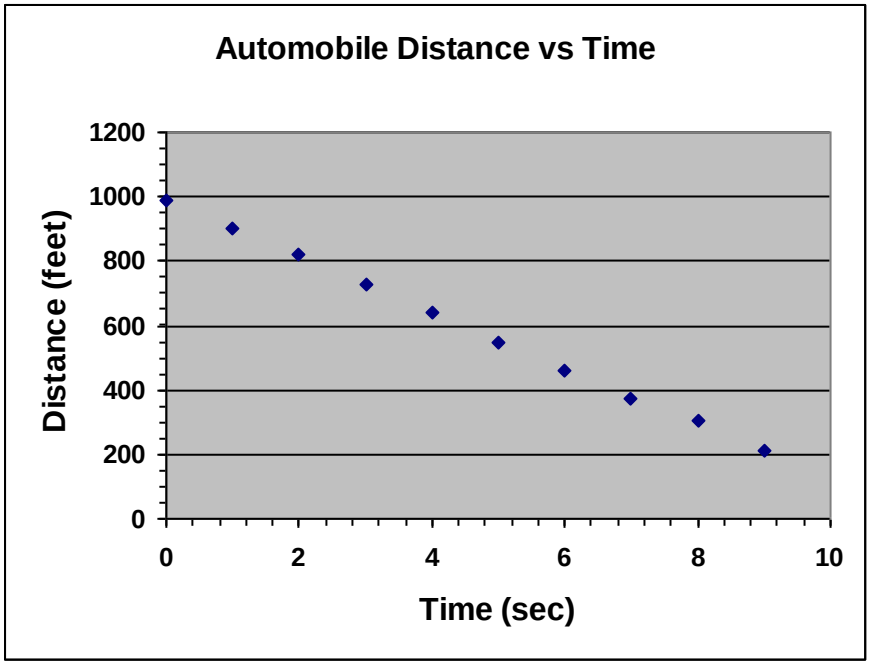
Example 1

Suppose a policeman uses a speed gun to determine the speed of an approaching automobile. The speed gun measures the distance from itself (the

sensor) to the automobile with an accuracy of 10 ft. Assume that the automobile is traveling at a constant velocity. Calculate the automobile's distance at t=0 and speed using least squares estimation.

First, we determine the appropriate assumptions for the measurement errors. The technology used in the speed gun is not specified but we know it has an accuracy of 10 ft. Unless there is more information available about the operation of the speed gun, we will assume that each time a distance measurement is made the error is zero-mean and normally distributed with a standard deviation, σ, of 10 ft. The standard deviation is considered constant for all measurements. This should be a good assumption if the speed gun is used within its specified minimum and maximum ranges. It is also reasonable to assume that the errors are uncorrelated measurement to measurement. Therefore, the measurement error assumption set for this problem is (1 1 1 1 1 1).

The measured distances are shown below.



Using Equation (1.4) we have,

$$\hat{\mathbf{Y}}_i = \mathbf{X}_i\boldsymbol{\beta} + \boldsymbol{\varepsilon}_i = \boldsymbol{\eta}_i + \boldsymbol{\varepsilon}_i; \quad i = 1,2,...10 \qquad (1.27)$$

where

$$\mathbf{X}_i = \begin{bmatrix} 1 & t_{i-1} \end{bmatrix}$$

and

$$\boldsymbol{\beta} = \begin{bmatrix} d_{t_0} \\ s \end{bmatrix}$$

where d_{t_0} is the distance to the automobile at t=0, and

s is the automobile's speed.

From Equation (1.20) we apply the equation for the least squares estimator.

$$\hat{\beta}_{LS} = \left(X^T X\right)^{-1} \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 0 & 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 \end{bmatrix} \begin{bmatrix} 988 \\ 899 \\ 819 \\ 726 \\ 637 \\ 550 \\ 460 \\ 374 \\ 307 \\ 210 \end{bmatrix} \quad (1.28)$$

This reduces to

$$\hat{\beta}_{LS} = \left(X^T X\right)^{-1} \begin{bmatrix} 5970 \\ 19737 \end{bmatrix} = \begin{bmatrix} .345455 & -.05455 \\ -.05455 & .012121 \end{bmatrix} \begin{bmatrix} 5970 \\ 19737 \end{bmatrix} \quad (1.29)$$

Finally we compute the parameter estimates,

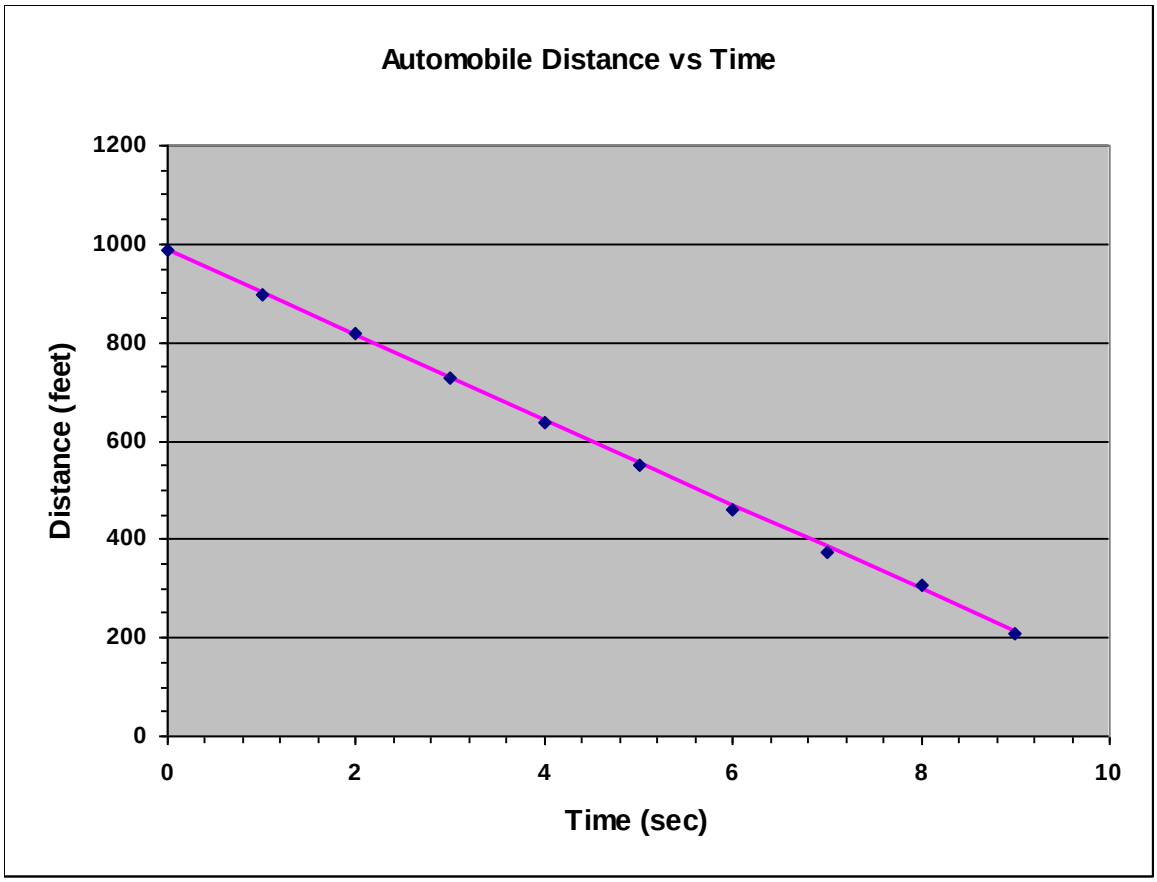
$$\hat{\beta}_{LS} = \begin{bmatrix} \hat{d}_{t_0} \\ \hat{s} \end{bmatrix} = \begin{bmatrix} 986 \\ -86 \end{bmatrix} \quad (1.30)$$

The estimated initial distance is 986 ft and the estimated speed is -86 ft/sec (-58.64 mph).

From Equation (1.27) we can write the solution for the estimation problem,

$$\hat{\eta}_i = \hat{d}_{t_0} + \hat{s}\, t_{i-1} = 986 - 86\, t_{i-1}; \quad i = 1, 2, \ldots 10 \quad (1.31)$$

Plotting this straight line on the measurements we have



We have assumed that the automobile is traveling at a constant speed over the 10 sec measurement interval. If we anticipated that the automobile would accelerate or decelerate we could include an acceleration term in our model η. In any application of parameter estimation we should use the simplest model (the one with the fewest unknown parameters) that captures the significant aspects of the problem. If we applied this "constant speed" model to a situation where the automobile was accelerating or decelerating over the measurement interval, the measurements would still be fit with a straight line and the residuals (the difference between the fit line and the measurements) would be significantly correlated. It is a good idea to visually inspect the residuals to verify that they have a random character. If there is a trend in the residuals there is likely an unmodeled phenomenon present. Consider using a different model or adding terms, with additional unknown parameters, to the present model.

What can we say about the accuracy of the parameter estimates? We know that the measurement error assumption set is (1 1 1 1 1 1).

From Equation (1.24) we have

$$\text{cov}(\hat{\beta}_{LS} - \beta) = \sigma^2 \left(X^T X\right)^{-1} = \begin{bmatrix} 34.5455 & -5.455 \\ -5.455 & 1.2121 \end{bmatrix} \qquad (1.32)$$

where $\sigma = 10$ ft.

From Equation (1.32) we know the initial distance estimate has an accuracy of

$$\sigma_{\hat{d}_{t_0}-d_{t_0}} = \sqrt{34.5455} = 5.88 \text{ ft} \qquad (1.33)$$

and the speed estimate has an accuracy of

$$\sigma_{\hat{s}-s} = \sqrt{1.2121} = 1.10 \text{ ft/sec} \qquad (1.34)$$

Since the measurement errors are normally distributed the parameter estimation errors are also normally distributed. Now we can make confidence statements about the parameter estimates. Using the properties of the standard normal density function given in Figure 1, we can state that

$$\text{Prob}\left\{\hat{s} - \sigma_{\hat{s}-s} \le s \le \hat{s} + \sigma_{\hat{s}-s}\right\} = .68 \qquad (1.35)$$

Substituting values for our problem,

$$\text{Prob}\left\{-87.1 \text{ ft/sec} \le s \le -84.9 \text{ ft/sec}\right\} = .68 \qquad (1.36)$$

This result states that the probability is .68 that the automobile's speed is between 84.9 and 87.1 ft/sec. (We drop the negative sign since the policeman is interested in magnitude, not direction!) This statement is possible because the measurement errors are normally distributed.

Similarly,

$$\text{Prob}\left\{\hat{s} - 2\sigma_{\hat{s}-s} \le s \le \hat{s} + 2\sigma_{\hat{s}-s}\right\} = .95 \qquad (1.37)$$

Again substituting values for our problem,

$$\text{Prob}\left\{-88.2 \text{ ft/sec} \le s \le -83.8 \text{ ft/sec}\right\} = .95 \qquad (1.38)$$

Notice that we have to widen the allowed interval when our confidence stated in probability goes from .68 to .95. We can use whatever multiplier of σ necessary to get the desired probability value from the standard normal density.

We can make analogous probability statements about the automobile's initial distance estimate, but we assume the automobile's speed is the focus of the policeman's attention!

In addition to the accuracy of the individual parameters, Equation (1.32) supplies information on how the errors in the estimated parameters are correlated.

From Equations (1.12) and (1.32) we have that

$$\rho_{\hat{d}_{t_0}-d_{t_0},\hat{s}-s} = \frac{\text{cov}(\hat{d}_{t_0} - d_{t_0}, \hat{s} - s)}{\sigma_{\hat{d}_{t_0}-d_{t_0}} \sigma_{\hat{s}-s}} = \frac{-5.455}{5.88 \times 1.10} = -0.84 \qquad (1.39)$$

The correlation coefficient, ρ, can have values in the interval [–1,1]. A value of "0" indicates that the estimation errors are uncorrelated. Since in this problem the correlation coefficient is -0.84 we have significant negative correlation between the estimated parameter errors.

To understand why the estimated parameter errors are negatively correlated consider the following figure. In addition to the measurements and least squares fit, we show two more possible fits to the measurements. Case 1 shown in blue is

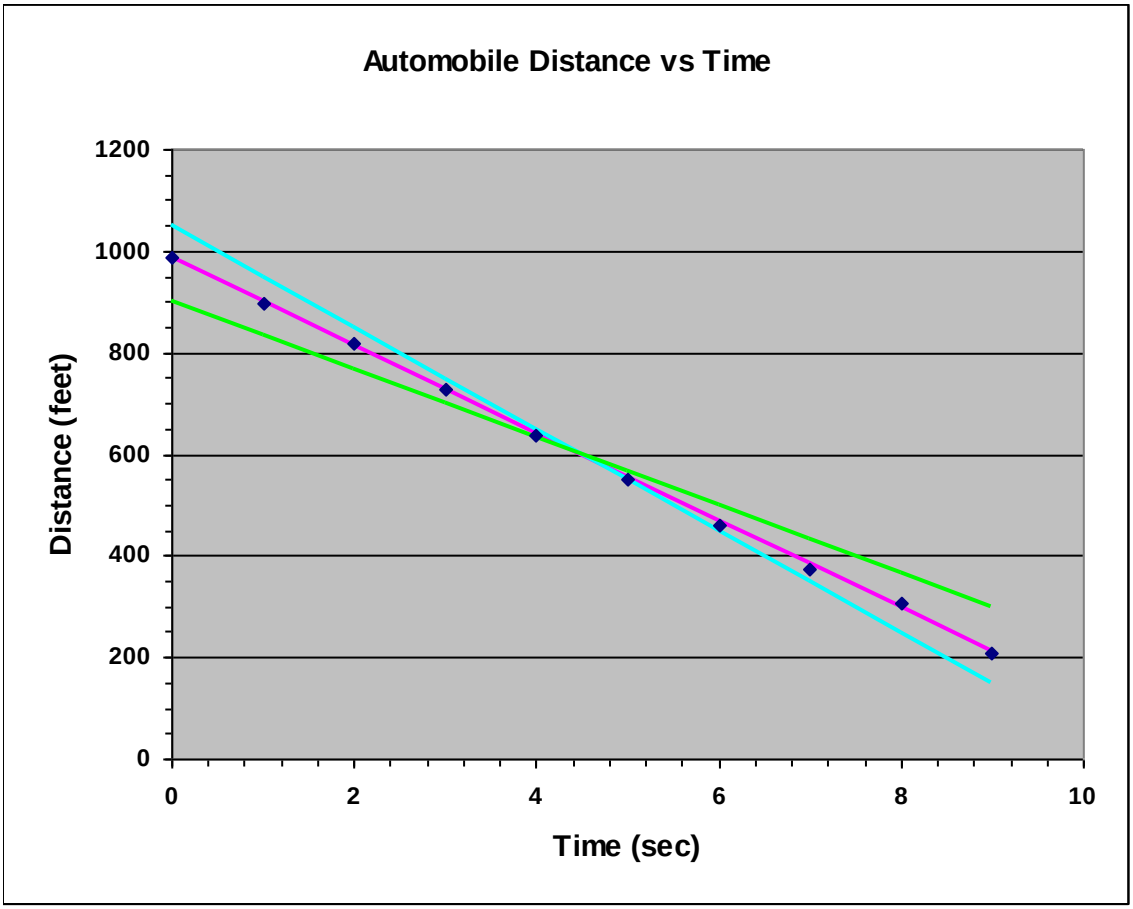
$$\hat{\eta}_i = \hat{d}_{t_0} + \hat{s}\ t_{i-1} = 1050 - 100.22\ t_{i-1};\quad i = 1,2,...10 \qquad (1.40)$$

and Case 2 shown in green is

$$\hat{\eta}_i = \hat{d}_{t_0} + \hat{s}\ t_{i-1} = 900 - 66.89\ t_{i-1};\quad i = 1,2,...10 \qquad (1.41)$$

Both Cases 1 and 2 are alternative fits to the measurements. Notice that the initial distance "estimate" for Case 1 (1050) is greater than the least square fit value of 986. On the other hand the speed "estimate"(-100.22) is less than the least squares fit value of –86. When the initial distance estimation error is positive the speed estimation error is likely to be negative.

Case 2 shows the situation should the initial position estimation error be negative, and then the speed estimation error is likely to be positive.



Suppose we don't know the accuracy of the speed gun. The assumption set becomes (1 1 1 1 1 0).

We can estimate the variance of the measurement errors from Equation (1.25),

$$\hat{\sigma}^2 = \frac{R_{LS}}{m-p} = \frac{380}{10-2} = 47.5 \text{ ft}^2 \quad (1.42)$$

or

$$\hat{\sigma} = \sqrt{47.5} = 6.9 \text{ ft} \quad (1.43)$$

The residuals indicate that our assumption of $\sigma = 10$ ft when using the assumption set (1 1 1 1 1 1) is conservative.

The estimator for σ^2 is unbiased,

$$E\left(\hat{\sigma}^2\right) = \sigma^2 \quad (1.44)$$

We can use $\hat{\sigma}^2$ in lieu of σ^2 in Equation (1.26) to obtain

$$\text{cov}(\hat{\beta}_{LS} - \beta) = \hat{\sigma}^2 \left(X^T X\right)^{-1} = \begin{bmatrix} 16.4091 & -2.591 \\ -2.591 & 0.5757 \end{bmatrix} \quad (1.45)$$

The 95% confidence limits taken from the t distribution with $m - p = 8$ degrees of freedom is

$$\text{Prob}\left\{\hat{s} - 2.3\hat{\sigma}_{\hat{s}-s} \le s \le \hat{s} + 2.3\hat{\sigma}_{\hat{s}-s}\right\} = .95 \quad (1.46)$$

which results in

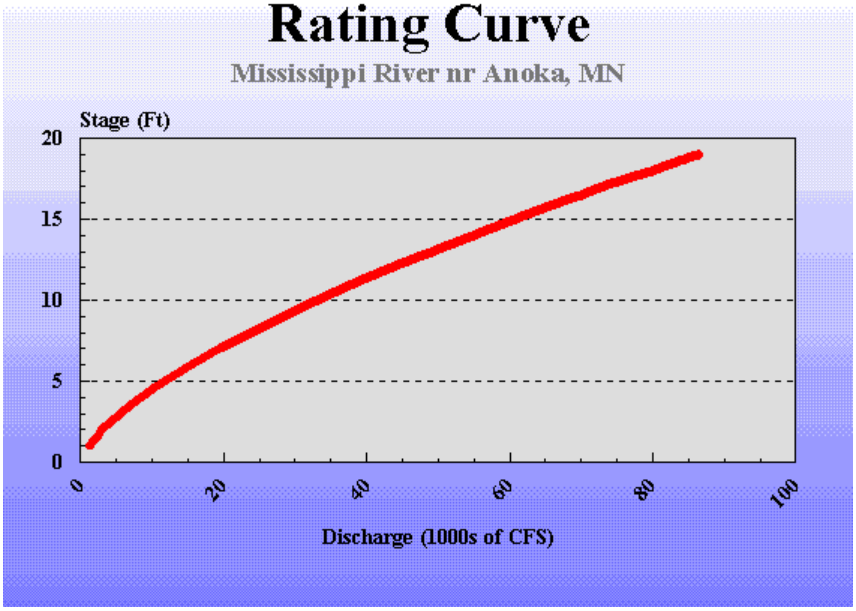
$$\text{Prob}\left\{-87.7 \text{ ft/sec} \le s \le -84.3 \text{ ft/sec}\right\} = .95 \quad (1.47)$$

Comparing Equations (1.47) and (1.38) we see that estimating the measurement error variance rather than using the given value for the variance has achieved a stronger result. This is not an unusual result because often the specified measurement error variance, in this case for the speed gun, is taken from a user's manual. Often sensor specifications prove to be conservative in use. On the other hand, the opposite situation occurs. In either case the estimated variance will reveal the discrepancy between the sensor specification and actual performance.

Example 2

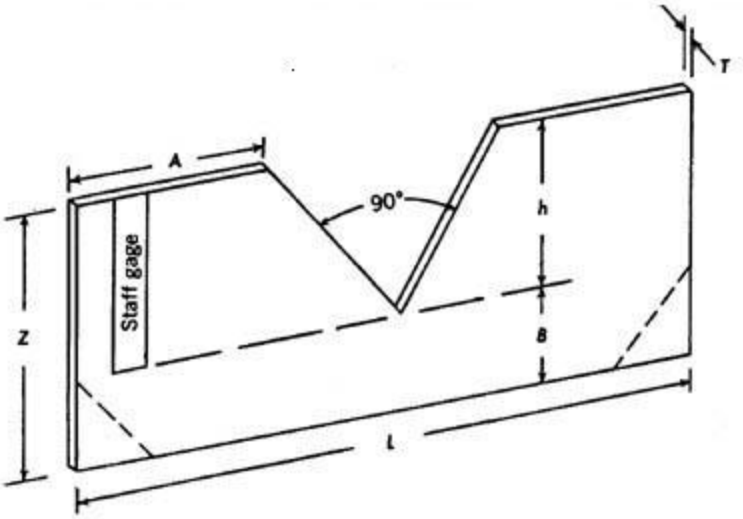
Our second example is taken from the field of hydrology.

Stream and river discharges are typically reported as flow rate in ft^3/s. A rating curve is developed that relates stage (water height) to flow rate. The following is an example rating curve for the Mississippi River near Anoka, MN.



Many velocity and depth measurements are required to develop a rating curve for a river. In this example we develop a rating curve for a weir, a structure that is used to measure water flow in small streams. An installed weir is shown in the following picture obtained from http://wwwrcamnl.wr.usgs.gov/sws/SWTraining/FlashFandR/Index.html

The design for a weir with a 90^{o} notch is illustrated in the following figure from http://wwwrcamnl.wr.usgs.gov/sws/SWTraining/FlashFandR/WSP2175/RnatzChpt8.htm



Corners may be trimmed if desired

Weir	Z	h	B	L	A	T	Weight (lb)
Large	1.75	1.00	0.75	4.0	1.0	16 ga.	24
Medium	1.25	.80	.45	3.0	.7	14 ga.	17
Small	.75	.47	.28	2.0	.53	10 ga.	8

All dimensions, other than T, are in feet

It is known from physics that the flow rate Q through the weir is related to the stage h by

$$Q = Kh^n \qquad (1.48)$$

This is a nonlinear model with parameters K and n. h is the independent variable. Several references give values of both K and n around 2.5 for English units; h in ft and Q in ft^3/s. [See Novak, Pavel; Hydraulic Structures, 2001, pg 310 & Avallone, Eugene A. ed; Marks' Standard Handbook for Mechanical Engineers (11^{th} ed), p3-58.]

We can transform this model to one that is linear-in-the-parameters by taking the logarithm of both sides of Equation (1.48). The logarithm to the base 10 is used to conform to standard practice for rating curves.

The transformed equation is,

$$\log_{10} Q = n \log_{10} h + \log_{10} K \qquad (1.49)$$

Now write this equation in standard form,

$$Y = \beta_1 X + \beta_2 \qquad (1.50)$$

where

$$\mathbf{Y} = \log_{10} \mathbf{Q}$$
$$\mathbf{X} = \log_{10} \mathbf{h}$$
$$\beta_1 = \mathbf{n}$$
$$\beta_2 = \log_{10} \mathbf{K}$$

Assume h is known; i.e. it can be measured without error. Now Equation (1.50) is linear-in-the-parameters with independent variable X. The rating curve h vs. Q is a straight line when plotted on a log-log scale.

Now we take flow rate measurements for a weir for specified values of stage h. What can we say about the measurement errors? Following Equation (1.3) for Equation (1.50) we have

$$\hat{\mathbf{Y}} = \log_{10} \hat{\mathbf{Q}} = \beta_1 \mathbf{X} + \beta_2 + \varepsilon' \qquad \textbf{(1.51)}$$

where ε' is the measurement error vector

Notice that the measurement vector $\hat{\mathbf{Y}}$ is the algorithm of the flow rate vector $\hat{\mathbf{Q}}$. The additive errors ε' cannot be strictly normally distributed because $\hat{\mathbf{Y}} \geq 0$ and the normal distribution extends to ± ∞. One can show however that for $|\varepsilon'| \ll 1$ the errors in Equation (1.51) are a percentage of the flow rate. Let

$$\hat{\mathbf{Q}} = (1+\varepsilon)\mathbf{Q} \qquad \textbf{(1.52)}$$

where ε is a zero mean, normally distributed error vector with $\sigma_\varepsilon \leq 0.02$

From Equations (1.48) and (1.52) we can write

$$\hat{\mathbf{Q}} = (1+\varepsilon)\mathbf{K}\mathbf{h}^{\mathbf{n}} \qquad \textbf{(1.53)}$$

Taking the logarithm of both sides of Equation (1.53) yields

$$\hat{\mathbf{Y}} = \log_{10} \hat{\mathbf{Q}} = \mathbf{n} \log_{10} \mathbf{h} + \log_{10} \mathbf{K} + \log_{10}(1+\varepsilon) \qquad \textbf{(1.54)}$$

Since $|\varepsilon| \ll 1$

$$\log_{10}(1+\varepsilon) \cong \frac{\varepsilon}{\ln 10}$$

Now Equation (1.54) becomes

$$\hat{\mathbf{Y}} = \log_{10} \hat{\mathbf{Q}} \cong \beta_1 \mathbf{X} + \beta_2 + \frac{\varepsilon}{\ln 10} \qquad \textbf{(1.55)}$$

which confirms Equation (1.51) to within a constant multiplier and a small percentage error approximation. Equations (1.54) and (1.55) also show that the additive errors are constant variance on a log-log plot when there are percentage errors in the original model, Equation (1.53). Notice that the variance of the multiplicative errors increases with flow rate as shown by Equation (1.53).

The errors in Equations (1.51) and (1.55) are zero mean, additive, constant variance, uncorrelated, not normally distributed, and have unknown variance; i.e. (1,1,1,1,0,0). Under the condition that $\sigma_\varepsilon \leq 0.02$ the errors are approximately normally distributed, (1,1,1,1,~1,0). Also,

$$\sigma_\varepsilon = \ln 10 * \sigma_{\varepsilon'} \qquad (1.56)$$

Suppose we have a laboratory experiment in which water is introduced into a simulated stream and flows out a weir. A flow rate meter is located in the pipe behind the inlet faucet. Flow rate is measured without bias (i.e. the errors have zero mean) and with an accuracy of ±3% of the flow rate. Treat this % as a 2σ number, thus $\sigma_\varepsilon = 0.015$.

We simulate a set of nine measurements $\hat{Q}_i$; $i = 1,2,..,9$ using Equation (1.52) with ε_i normally distributed with zero mean and $\sigma_\varepsilon = 0.015$, $K = n = 2.5$.

The stage h and flow rate measurements are

h_i (ft)	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
$\hat{Q}_i$ (ft^3/s)	0.008	0.044	0.126	0.251	0.430	0.704	1.022	1.445	1.946

Transforming both h and Q by log base 10 gives

$\log_{10} h_i$	-1.000	-0.699	-0.523	-0.398	-0.301	-0.222	-0.155	-0.097	-0.046
$\log_{10} \hat{Q}_i$	-2.103	-1.360	-0.901	-0.600	-0.367	-0.153	0.009	0.160	0.289

From Equation (1.20) we obtain parameter estimates,

$$\hat{\beta}_{LS} = \left(X^T X\right)^{-1} X^T \hat{Y} = \begin{bmatrix} 2.5070 \\ 0.4000 \end{bmatrix} \qquad (1.57)$$

where

$$\hat{Y}^T = (\log_{10} \hat{Q}_i)^T = \begin{bmatrix} -2.103 & -1.360 & -0.901 & -0.600 & -0.367 & -0.153 & 0.009 & 0.160 & 0.289 \end{bmatrix}$$

and

$$X^T = \begin{bmatrix} -1 & -0.6990 & -0.5229 & -0.3979 & -0.3010 & -0.2218 & -0.1549 & -0.0969 & -0.0458 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \end{bmatrix}$$

$$\hat{\beta}_1 = \hat{n} = 2.5070$$
$$\hat{\beta}_2 = \log_{10} \hat{K} = 0.4000 \qquad (1.58)$$

Therefore

$$\hat{K} = 10^{0.4000} = 2.5117$$

The linear-in-the-parameters model with estimated parameters rounded to two decimal places is

$$\log_{10} Q_m = 2.51 \log_{10} h + 0.40 \qquad (1.59)$$

Transforming back to the original nonlinear form, the model for the weir rating chart is

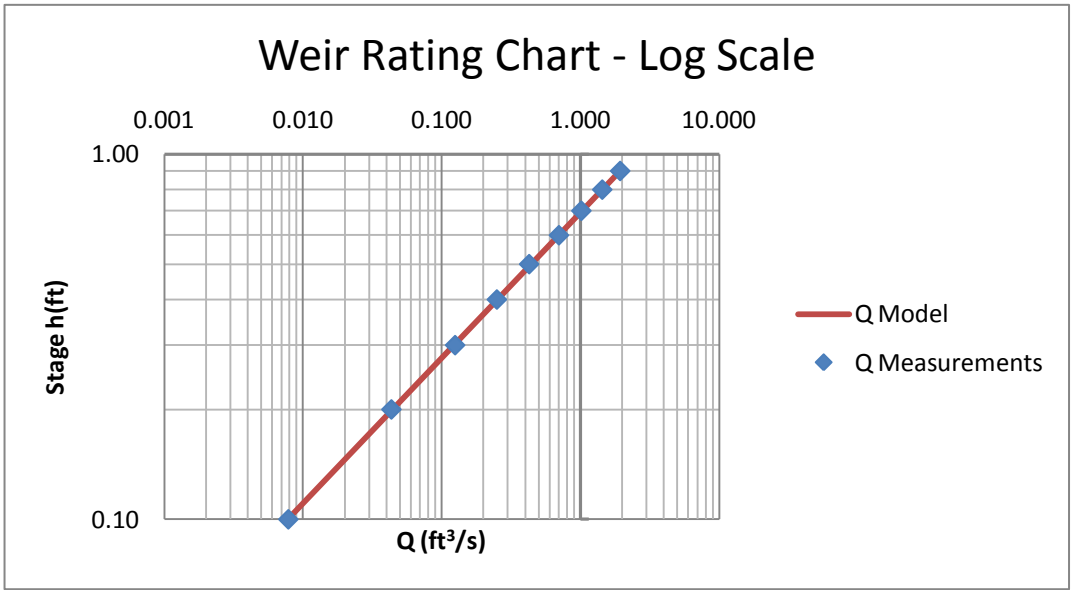
$$Q_m = \hat{K}h^{\hat{n}} = 2.51\,h^{2.51} \qquad (1.60)$$

where

h is the stage height in ft, and

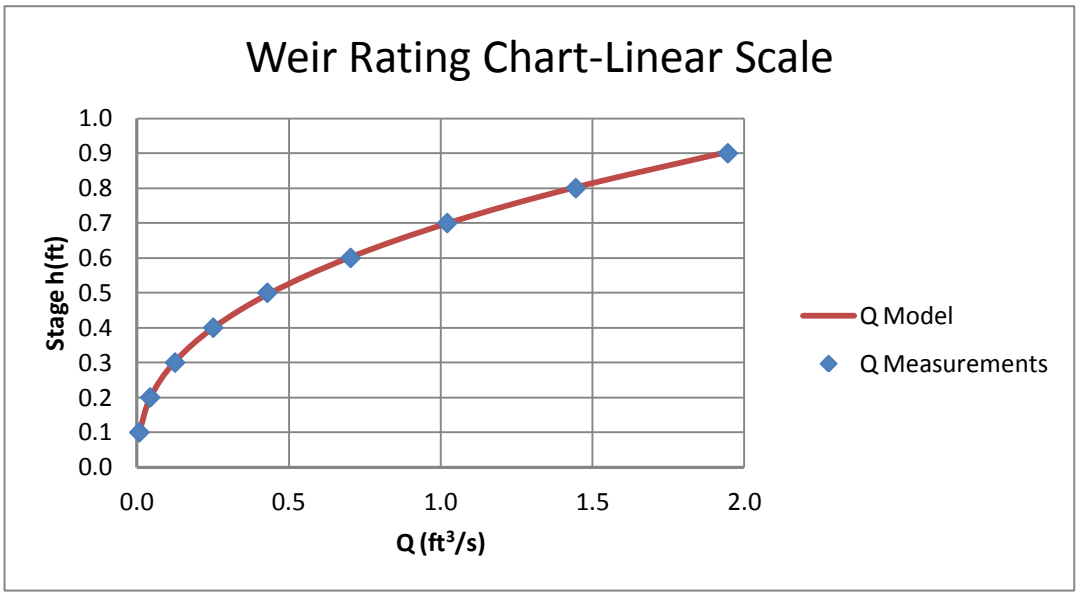
Q_m is the model flow rate in ft^3/s

The following figure is the rating curve described by either Equation (1.59) or Equation (1.60) with the measurements Q_i superimposed. The result is a straight line as expected.



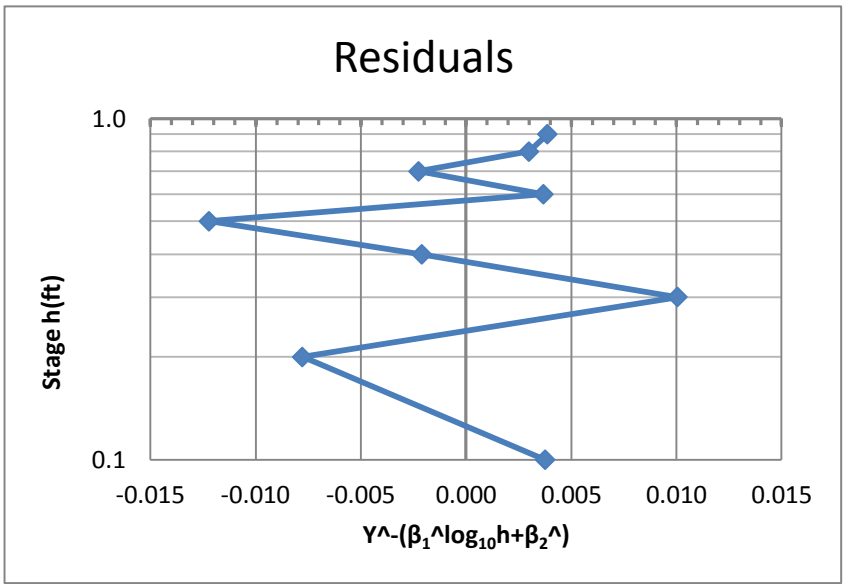
The errors are so small that the measurements overlay the model's predicted line.

The next figure shows a linear plot of the same data.



Again the measurements overlay the model's predicted curve. The measurements are bunched up for low flow rates and spread out at higher flow rates. The reverse is true in the log-log scale weir rating curve; the measurements are bunched up for high flow rates and spread out at the low ones.

The next figure shows the residuals, the difference between the model predictions and the measurements.



We showed in the discussion of Equation (1.55) that the errors in the linear-in-the-parameters model are constant variance. If our model is accurate, as we know it is in this simulated case, the residuals should also be constant variance and uncorrelated. The previous chart qualitatively confirms this assertion.

Now let's estimate the measurement error standard deviation from the residuals.

The estimated standard deviation of the measurement errors is computed from the residuals using Equation (1.25) since the errors are constant variance

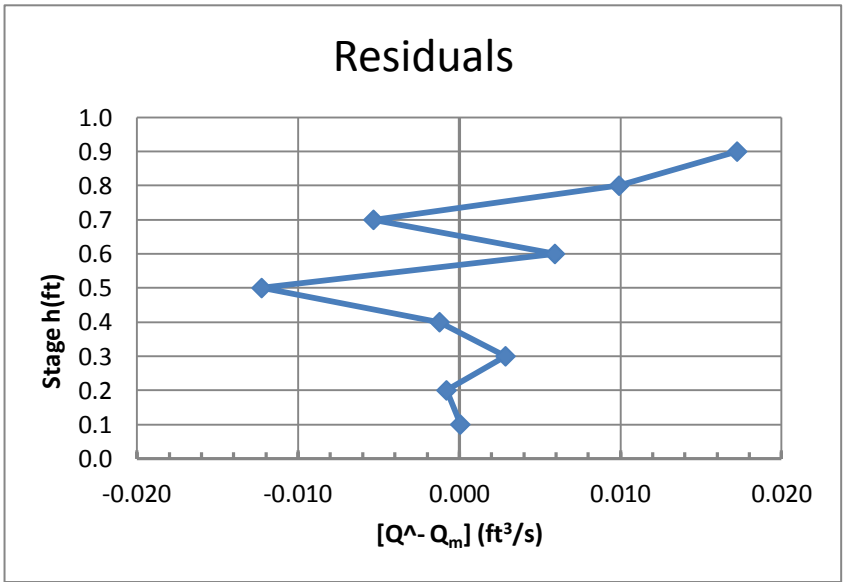
$$\hat{\sigma}_{\varepsilon'} = \sqrt{\frac{\left(\hat{\mathbf{Y}} - \mathbf{X}\hat{\boldsymbol{\beta}}_{LS}\right)^{T}\left(\hat{\mathbf{Y}} - \mathbf{X}\hat{\boldsymbol{\beta}}_{LS}\right)}{m-p}} = \sqrt{\frac{3.7214E-04}{9-2}} = \quad 0.0073 \qquad (1.61)$$

From Equation (1.56) we have that

$$\hat{\sigma}_{\varepsilon} = \ln(10) * \hat{\sigma}_{\varepsilon'} = \quad 2.3026 * 0.0073 = \quad 0.0168 \qquad (1.62)$$

The estimated standard deviation is slightly larger than the simulated standard deviation of ±1.5%, i.e. 0.015.

The variance of the residuals for the linear-scale weir rating curve increases with Q as shown in the following chart.



We cannot construct <u>exact</u> confidence limits for the estimated parameters from the material presented in this course because the probability distribution of the errors is not normal. However, for our assumption of $\sigma_{\varepsilon} \le 0.02$ the error distribution is approximately normal. Given the assumption set (1,1,1,1,~1,0), approximate confidence limits for the estimated parameters $\hat{K}$ & $\hat{n}$ can be established as they were in Example 1 <u>as long as the customer is informed that the confidence limits are approximate</u>.

Using Equation (1.26) with $\hat{\sigma}^2 = \hat{\sigma}^2_{\varepsilon'}$

$$\mathrm{cov}(\hat{\boldsymbol{\beta}}_{LS} - \boldsymbol{\beta}) = \hat{\sigma}^2_{\varepsilon'}\left(\mathbf{X}^{T}\mathbf{X}\right)^{-1} = (0.0073)^2 \begin{bmatrix} 1.2811 & 0.4897 \\ 0.4897 & 0.2983 \end{bmatrix} \qquad (1.63)$$

Continuing

$$\operatorname{cov}(\hat{\beta}_{LS} - \beta) = \operatorname{cov}\left(\begin{bmatrix} \hat{n} - n \\ \log_{10}\hat{K} - \log_{10} K \end{bmatrix}\right) = \begin{bmatrix} 6.810E-05 & 2.603E-05 \\ 2.603E-05 & 1.586E-05 \end{bmatrix} \quad (1.64)$$

We now construct 90% confidence limits to illustrate use of the Student's-t distribution. In practice the analyst should set confidence limits appropriate to his/her particular project and customer needs. Use the symbol "≅" to show that the confidence limits are approximate. The approximate 90% confidence limits taken from the Student's-t distribution with $m - p = 7$ degrees of freedom for the parameter n is

$$\text{Prob}\left\{\hat{n} - 1.9\hat{\sigma}_{\hat{n}-n} \le n \le \hat{n} + 1.9\hat{\sigma}_{\hat{n}-n}\right\} \cong .90 \quad (1.65)$$

$$\text{Prob}\left\{\hat{n} - 0.0157 \le n \le \hat{n} + 0.0157\right\} \cong .90 \quad (1.66)$$

$$\text{Prob}\left\{2.491 \le n \le 2.523\right\} \cong .90 \quad (1.67)$$

Similarly for $\log_{10}K$,

$$\text{Prob}\left\{0.3924 \le \log_{10} K \le 0.4075\right\} \cong .90 \quad (1.68)$$

$$\text{Prob}\left\{2.468 \le K \le 2.556\right\} \cong .90 \quad (1.69)$$

Finally we compute the correlation coefficient between the parameter estimates using Equations (1.12) and (1.64)

$$\rho = \frac{\operatorname{cov}(\hat{n} - n, \log_{10}\hat{K} - \log_{10} K)}{\sigma_{\hat{n}-n}\sigma_{\log_{10}\hat{K} - \log_{10} K}} = \frac{2.603E-05}{3.286E-05} = 0.792 \quad (1.70)$$

This large positive correlation coefficient means that the errors in estimating n and $\log_{10}K$ are likely to have the same sign; i.e. if $\hat{n}$ is too large, $\log_{10}\hat{K}$ and $\hat{K}$ are likely to be too large as well. Conversely if $\hat{n}$ is too small, $\log_{10}\hat{K}$ and $\hat{K}$ are likely to be too small as well.

This example illustrates how a linear-in-the-parameters model can be used to estimate parameters in a particular nonlinear model taken from hydrology.

Summary

This course has presented an overview of linear least squares parameter estimation theory with a focus on six basic assumptions that can be made about the measurement errors. The measurement error assumption sets for which least squares is the appropriate estimation technique were clearly delineated. How the quality of the parameter estimates can be communicated to customers was presented. Two comprehensive example problems were solved and explained.