MATH 577, ASSIGNMENT 1. Version: January, 2004

---

This is a very simple exercise in least squares. In fact, it appears in a variety of textbooks. Here we will focus less on the technique and more on the interpretation. The aims are: (1) Fill in the missing stepsa in the derivation of the technique, (2) use the output to think about what are the ways in which model error and noise affect the answer, (3) compute different statistic estimators and see what they are capable of saying about the outcome, (4) see how the results depend on the sample size. We will cover the material in class, so this exercise is also meant to generate questions and discussion for that time.

Please help each other and discuss this assignment. Use the listserv to formulate questions and venture answers.

Let's assume that $\psi(t)$ represents a physical scalar variable that depends on time $t > 0$. The data measured is noisy:

$$y(t) = \psi(t) + n(t), \tag{1}$$

where (assumed to be) normally distributed noise $n(t)$ has zero mean and variance $< n(t)n(t') >= \sigma_n^2 \delta_{t,t'}$. In fact, we will construct the data set synthetically: we will let $\psi(t)$ be a linear function in $t$ and generate a set of $M$ data points over $0 \leq t \leq 20$. We will be modifying the noise by changing $\sigma_n^2$, in order to see what happens when the data is noisier, and we will also be generating the data via 2 models, so that we can investigate what happens when we try to do a linear fit on a data set that is properly linear and one that is not.

For variance values, use $\sigma_n^2 = 9$ and $\sigma_n^2 = 100$. For models types, use $\psi_1(t) = 1 + 2t$ for the first set of experiments and $\psi_2(t) = 1 + 2t^2$ for the second set.

We have reason to believe that data obeys a "linear law"

$$\psi(t) = x_1 + x_2 t.$$

Sure, that's the right model for the case data generated with $\psi_1(t)$, but let's pretend we don't know that.

The idea here is to find the vector $\mathbf{x} = [x_1, x_2]^T$, where $T$ means "transpose", and hence find a model fit. We would also want to know how well we do, i.e. what the uncertainty in the estimates of $\mathbf{x}$ are. We use least squares here.

1. Let's get a feel for what the data and noise look like: First, a little bit of experimentation: what does a pdf look like if it is Gaussian and has variance 9? what does it look like when it has variance 100? Use analytical expressions of the normal distribution (always zero mean) and compare these two curves graphically (use the "eyeball" metric in these comparisons).

   Next, let's generate a normal distribution with zero mean and prescribed variance, a discrete and finite one: use matlab to generate a vector of size $M$ using the `randn` command. You'll need to learn to generate distributions with a specific mean and variance, if you do not know already how to do so. Now, compute the sample mean and variance of the vector as $M$ gets larger. Does the sample mean really equal zero? does it really have the prescribed variance? Use the command `mean` and `std`. Let $M = 2^p$ and try $p = 4, 5, 6...$, say.

   Finally, compute $\sum_{i=1}^{M} (\psi_2(t_i) - \psi_1(t_i))^2 / M$. To do so, take the interval in time to be fixed at $[0, 20]$. Make a grid of $M$ data points $t_i$. Try $M = 2^p$, $p = 4, 5, 6, ....$ I want you to simply to get an idea of what is the size of the errors you'd make if you had guessed the model for the data incorrectly in the best of circumstances, i.e. no noise.

2. The least squares:

   (a) Take $t_i$ to be the $i^{th}$ time of an observation. There will be $M$ observations. The vector $\mathbf{x}$ of unknowns has size $N = 2$. Write down explicit representations of the $M \times N$ matrix $E$, $\mathbf{x}$, and $\mathbf{y}$, the measurement vector, so that (1) can be written in the discrete case as

$$E\mathbf{x} + \mathbf{n} = \mathbf{y}.$$

where $\mathbf{n}$ is the noise vector. Let $J(\mathbf{x}) = \mathbf{n}^T\mathbf{n}$ a scalar functional of $\mathbf{x}$. Show that finding a minima of $J$ leads to the "normal equations"

$$E^T E\mathbf{x} = E^T\mathbf{y}$$

and hence an equation for the unknown $\mathbf{x}$. Let

$$\hat{\mathbf{x}} = (E^T E)^{-1}E^T\mathbf{y}.$$

We have to assume that we do not know the precise relationship between $\hat{\mathbf{x}}$ and $\mathbf{x}$. Show that the "residuals" are given by

$$\hat{\mathbf{n}} = (I - E(E^T E)^{-1}E^T)\mathbf{y}.$$

Here $I$ is the $M \times M$ identity matrix.
The uncertainty in the solution is the same as the variance about the mean and is

$$C_{\hat{\mathbf{x}}\hat{\mathbf{x}}} =< (\hat{\mathbf{x}} - \mathbf{x})(\hat{\mathbf{x}} - \mathbf{x})^T > .$$

Show that

$$C_{\hat{\mathbf{x}}\hat{\mathbf{x}}} = \sigma_n^2(E^T E)^{-1}, \tag{2}$$

if we assume that $< \hat{\mathbf{x}} >= \mathbf{x}$, and that $< \mathbf{n}\mathbf{n}^T >= \sigma_n^2 I$, there is no correlation between the noise in the different equations, *i.e.* white noise conditions.
If we are not confident that $< \hat{\mathbf{x}} >= \mathbf{x}$, (2) is still interpretable but as $C_{\hat{\mathbf{x}}\hat{\mathbf{x}}} = D^2(\hat{\mathbf{x}}- < \hat{\mathbf{x}} >)$, the covariance of $\hat{\mathbf{x}}$. The "standard error" might be defined as the $\pm\sqrt{|C_{\hat{\mathbf{x}}\hat{\mathbf{x}}}|}$.
Finally the uncertainty of the residual estimates is given by

$$C_{\hat{\mathbf{n}}\hat{\mathbf{n}}} =< (\hat{\mathbf{n}} - \mathbf{n})(\hat{\mathbf{n}} - \mathbf{n})^T > .$$

Show that

$$C_{\hat{\mathbf{n}}\hat{\mathbf{n}}} = \sigma_n^2(I - E(E^T E)^{-1}E^T)$$

when $< \mathbf{n} >= 0$ and white noise conditions are present.
Finally we define the "autocovariance"

$$R_{\hat{\mathbf{n}}\hat{\mathbf{n}}}(s) = \frac{1}{M}\sum_{i=1}^{M-|s|} \hat{\mathbf{n}}_i\hat{\mathbf{n}}_{i+s}$$

which is an estimate of $< \mathbf{n}_i\mathbf{n}_{i+s} >$, which has the property that the residuals would be truly uncorrelated if $< \mathbf{n}_i\mathbf{n}_{i+s} >= 0$ for $s \neq 0$. Further, one expects $R_{\hat{\mathbf{n}}\hat{\mathbf{n}}}(0)$ to approach a "delta function, " as $M$ gets larger.

3. Application of Least Squares. Write a matlab code that carries out the least squares calculation of $\mathbf{x}$. Let the interval in time be $[0, 20]$, assume that the measurements are equally spaced and that there are $M$ of these. Take as inputs to your code:

   - The data, i.e. either $\psi_1$ or $\psi_2$ with random noise of zero mean.
   - The number of data points $M = 2^p$. You will run the code for $p = 4, 5, 6...$
   - The variance in the noise.

   Use the above least squares procedure to find an estimate of $\mathbf{x}$ as you change the above input parameters. For each set of input parameters, plot the least squares fit over the data; the residuals; the standard error on one of the components of the variable. Also, compute the autocovariance of the residual. On this last one, remember that you are computing a convolution and that there's an assumption that the data is periodically extended. You can look at `conv` command in matlab or develop your own version. Also, for a given data set and given variance, plot the log of the norm of the covariance of $\hat{\mathbf{x}}$ and (2) against the log of $M$, for $p = 4, 5, 6, 7, ....$

4. Interpretation of the results. How does the outcome change as the noise variance is changed? How does it change when the model assumption is changed? Can you see that the fit is not terrible when using a small number of points and a linear model fit to $\psi_2$ plus noise data? What estimator tells you whether you are just seeing noise or an error in the model assumption? How about the actual nature of the finite-dimensional noise vector, as a function of $M$? Is it really randomly distributed with zero mean and specified variance? If the noise is really random and each time you do the experiment you get a different data set $y_i$, is it really legitimate to just present a single data set or an ensemble of data sets to make any conclusions on how the least squares calculation has proceeded?