In a Nutshell...
Fitting a Model to Data: Least Squares Formulation

AT Patera, JD Penn, M Yano

October 10, 2014

1 Preamble

The predictive capability central to engineering science and design is provided both by mathematical models and by experimental measurements. In practice models and measurements best serve in tandem: the models can be informed by experiment; and the experiments can be guided by the models. In this nutshell we consider perhaps the most common approach to the integration of models and data: fitting a model to data by least-squares minimization of misfit.

In this nutshell we consider the following topics:

We introduce several "working examples" which exercise the contexts — calibration, parameter estimation, and hypothesis testing — in which fitting a model can play a central role.

We introduce parametrized models to represent the true behavior of a physical system. We provide an abstraction which can accommodate any particular problem of interest by suitable mapping of variables. We discuss the construction of these parametrized models: the physics–based approach; the smoothness approach. We illustrate the effects of underfit and overfit, and provide some practical guidelines to avoid these pitfalls.

We present standard (and plausible) assumptions on measurement error: zero-mean, homoscedastic, independent. We also discuss ways in which these requirements can be relaxed.

We introduce the notion of the residual and subsequently the sum-of-squares misfit function. We develop the misift minimization procedure which connects the (optimal) parameters of our model with the measurements. We derive the normal equations, solution of which yields our least-squares estimate for the true parameter.

We define and exercise the design matrix X, which summarizes the model and the experiment and plays a central role in the normal equations and hence estimation procedure.

We discuss and provide numerical evidence for the convergence of the least-squares parameter estimate (in the absence of model error) in the limit of either small noise or many measurements.

We do not in this nutshell present any theoretical error analysis.

Prerequisites: univariate calculus; operations on matrices and vectors; probability (mean, variance, independence, normal density).

2 Motivation

We describe here several examples in which we must fit a model to data.

Let us say that we measure the temperature at several points in a thin surface layer and we wish to predict the gradient of the temperature (from which we can then deduce the heat flux). We might propose a local model for the temperature T as a function of position x as $T^{\text{model}}(x) = \beta_0 + \beta_1 x$. Here β_0 and β_1 are coefficients, or *parameters*; in particular, β_1 is the desired derivative of the temperature with respect to the coordinate direction x. How can we deduce the "good" value of β_1 — and hence the temperature gradient — from noisy measurements of the temperature at several positions within the layer?

Let us say that we wish to incorporate an IR Distance Transducer into a robotic navigation system. We might know from physical considerations that distance D (from transducer to target) is related to the measured voltage V as $D^{\text{model}}(V) = CV^{-\gamma}$, where C and γ are calibration parameters. How can we deduce the good values of parameters C and γ from data such that subsequently, given a measured voltage, we can rapidly evaluate the distance?

Let us say that we we wish to infer the time and height at which a ballistic object was released. We know from Newtonian mechanics that, in the limit of negligible air drag, we can approximate the height z as a function of time t (in the simple one-dimensional context) as $z^{\text{model}}(t) = \beta_0 + \beta_1 t + \beta_2 t^2$, where $\beta_j, 0 \leq j \leq 2$, are parameters related to the initial conditions and the acceleration of gravity. How can we deduce the good values of $\beta_j, 0 \leq j \leq 2$, and subsequently the time and height of release, from measurements of the height at different times?

Amontons' Law states that the static friction force $F^{\text{max}}_{\text{f, static}}$ is proportional to the normal load F_{normal} and independent of superficial contact area, A_{surface}: $F^{\text{max}}_{\text{f, static}} = \mu F_{\text{normal}}$, independent of A_{surface}. Let us say that we measure the static friction force for a pair of materials for different normal loads and superficial contact areas and we wish to (i) estimate the friction coefficient, μ, for inclusion in a handbook, and also (ii) test the hypothesis "The friction force does not depend on the superficial contact area.". Towards that end, we might consider an expanded model

$$(F^{\text{max}}_{\text{f, static}})^{\text{model}}(F_{\text{normal}}, A_{\text{surface}}) = \beta_0 + \beta_1 F_{\text{normal}} + \beta_2 A_{\text{surface}} \ , \tag{1}$$

in which $\beta_1 \equiv \mu$ and — if Amontons is correct — β_0 and β_2 *should* be zero. How can we deduce the good value of β_1 (hence μ) from data? How can we decide if we must reject our hypothesis that the friction force is independent of superficial contact area?

Let us say that we administer a quiz (the same quiz) asynchronously to different groups of students over the course of a given week and we wish to confirm the hypothesis "Performance on the quiz is independent of the day on which the quiz is administered.". (You can easily envision many analogous situations in engineering; for example, we might wish to confirm that the performance of a part does not degrade with time.) We can postulate a simple model for average student grade g as a function of time t as $g^{\text{model}}(t) = \beta_0 + \beta_1 t$, where — if our hypothesis is correct — β_1 *should* be zero. How can we decide, based on the realized grades of the students, if we must reject our hypothesis on β_1 ?

Each of these examples is a particular case of a much more general setting which is ubiquitous in engineering practice. We are provided with, or propose, a model characterized by independent and dependent variables and unknown parameters. For example, for our IR transducer, the independent variable is voltage V, the dependent variable is distance D, and the unknown (calibration) parameters are C and γ. We then wish to deduce the parameters from measurements of

the dependent variable for different values of the independent variable. Once the model parameters are determined, we may subsequently exploit this information in a variety of fashions, as we now describe.

In the the "calibration" context, we wish to accurately predict system behavior at values of the independent variable different from those at which we performed experiments; our IR transducer example is an application of this variety. In the "parameter estimation" context, we wish to exploit the deduced parameter values — typically constants in a *constitutive relation* — in systems different from those of our experiment: our friction coefficient example is an application of this variety. Finally, in the "hypothesis testing" context (a variant of parameter estimation), we wish to make inferences from the form of our model; our asynchronous quiz example is an application of this variety.

3 An Example: Falling Ball

We shall provide a general model and fitting framework in the next sections of this nutshell. But we first motivate the key ingredients with a more discursive example. In particular, we shall invoke the example of a falling ball as already introduced in Section 2.

3.1 Models and Measurements

We consider the trajectory of a ball which is released at time t_{init} at height $z_{\text{init}} > 0$ (and zero initial velocity) and then falls for $t > t_{\text{init}}$ under the action of gravity. The height z of the ball is defined relative to the ground with respect to a coordinate axis oriented in the direction opposite to the acceleration of gravity. We restrict attention to times t prior to first contact with the ground — "pre-bounce." We denote the magnitude of the acceleration of gravity by g.

We shall denote by $z^{\text{true}}(t)$ the true behavior of the physical system: a perfect (noise-free) experiment will yield measurements of the height which agree exactly with $z^{\text{true}}(t)$. We shall assume for the moment that our experiment is conducted in a vacuum. In that case we know, from our initial conditions, our pre-bounce hypothesis, and elementary Newtonian mechanics, that $z^{\text{true}}(t) = z^{\text{true,vaccum}}(t)$, where

$$z^{\text{true,vacuum}}(t) = z_{\text{init}} - \frac{1}{2} g (t - t_{\text{init}})^2 \; . \tag{2}$$

In most cases — in the presence of an atmosphere — $z^{\text{true}}(t)$ will be more complicated than $z^{\text{true,vaccum}}(t)$ of (2): in addition to gravity, aerodynamic drag will also play a role in the force balance. We return to this issue shortly.)

We now postulate a parametrized *model* for the motion of the ball,

$$z^{\text{model}}(t; \beta) \equiv \beta_0 + \beta_1 t + \beta_2 t^2 \; . \tag{3}$$

We note that $z^{\text{model}}(t; \beta)$[1] represents a functional form — a quadratic polynomial in t — for any values of the β, where β denotes the 3×1 parameter vector $(\beta_0 \; \beta_1 \; \beta_2)^{\text{T}}$. (We shift our usual vector

[1] Strictly speaking, $z^{\text{model}}(t; \beta)$ refers to a real number, whereas $z^{\text{model}}(\cdot; \beta)$ refers to our function (or functional form): $z^{\text{model}}(t; \beta)$ is the evalution of our function $z^{\text{model}}(\cdot; \beta)$ at time t. For simplicity, we shall refer to both the function and the evaluation of our function as $z^{\text{model}}(t; \beta)$ and let context dictate the correct interpretation.

indices, so that the first element of β is denoted β_0, in order to conform to conventions associated with regression analysis.) We now relate (2) to (3): $z^{\text{true}}(t) = z^{\text{model}}(t;\, \beta^{\text{true}})$ for

$$\beta^{\text{true}} \equiv (\beta_0^{\text{true}}\ \beta_1^{\text{true}}\ \beta_2^{\text{true}})^{\text{T}} \equiv (\ (z_{\text{init}} - \frac{1}{2} g t_{\text{init}}^2)\quad t_{\text{init}} g\quad -\frac{1}{2} g)^{\text{T}}\ , \tag{4}$$

which we shall denote the "true" parameter value.

We now consider the scenario described in Section 2 in which t_{init} and z_{init}, and perhaps even (the local value of) g, are not known with the accuracy required for subsequent engineering purposes. We shall thus make m experimental observations $(t_i,\, z_i^{\text{meas}})$, $1 \le i \le m$, in order to determine an estimate $\hat{\beta} \equiv (\ \hat{\beta}_0\ \ \hat{\beta}_1\ \ \hat{\beta}_2\ )^{\text{T}}$ for the true parameter value $\beta^{\text{true}} \equiv (\beta_0^{\text{true}}\, \beta_1^{\text{true}}\, \beta_2^{\text{true}}\,)^{\text{T}}$. (We discuss the procedure by which we derive this estimate in the next section.). We may subsequently extract estimates for t_{init}, z_{init}, and g, $\hat{t}_{\ \text{init}}$, $\hat{z}_{\ \text{init}}$, and $\hat{g}$, respectively, by consideration of the relationship (4); for example, $\hat{g} = -2\hat{\beta}_{\ 2}$. We may furthermore construct an approximation to the ball trajectory, $\hat{z(t)}$, as

$$\hat{z}(t) = z^{\text{model}}(t;\hat{\beta}) = \hat{\beta}_0 + \hat{\beta}_1 t + \hat{\beta}_2 t^2\ . \tag{5}$$

Note that (5) is defined for all times t, not just the times t_i, $1 \le i \le m$, at which we perform our measurements.

We can summarize the many ingredients in terms of the cast of players, the roles of each, and the anticipated outcome.

> *The Cast*: $z^{\text{true}}(t)$ is the true trajectory of the ball; $z^{\text{model}}(t;\beta)$ is a parametrized model for the ball trajectory; β^{true} is the true value of the parameter; $\hat{\beta}$ is an estimate for the true value of the parameter; $\hat{z}(t)$ is an estimate for the true trajectory of the ball; $(t_i, z_i^{\text{meas}}), 1 \le i \le m$, are our experimental data.
>
> *The Roles*: the parametrized model, $z^{\text{model}}(t;\beta)$, includes the true behavior, $z^{\text{true}}(t) = z^{\text{model}}(t;\beta^{\text{true}})$, such that we may then "tune" this model — find $\hat{\beta} \approx \beta^{\text{true}}$ — to explain the experimental data, $(t_i, z_i^{\text{meas}}), 1 \le i \le m$.
>
> *The Aspirations*: the experimental data, $(t_i, z_i^{\text{meas}}), 1 \le i \le m$, reflect the true behavior, $z^{\text{model}}(t;\beta^{\text{true}})$, and hence the parameter value which well explains the experimental data, $\hat{\beta}$, should be close to the true parameter value, β^{true}, and thus also the estimated trajectory of the ball, $\hat{z}(t) = z^{\text{model}}(t;\hat{\beta})$, should be close to the true trajectory, $z^{\text{model}}(t;\beta^{\text{true}})$.

We now proceed to more establish a more direct connection between the parameters and the measurements.

To begin, we make more precise our characterization of the experimental measurements. We recall that our measurements are of the form $(t_i, z_i^{\text{meas}}), 1 \le i \le m$, where z_i^{meas} is the measured height of the ball at specified time t_i. We shall suppose that the measurement times are ordered, $t_1 \le t_2 \le \cdots \le t_m$. We shall define the measurement error as the difference between the measured height of the ball and the true height of the ball,

$$\epsilon_i \equiv z_i^{\text{meas}} - z^{\text{true}}(t_i)\ , 1 \le i \le m\ . \tag{6}$$

Our measurements of the height of the ball may then be expressed as a perturbation from the true result,

$$z_i^{\text{meas}} = z^{\text{true}}(t_i) + \epsilon_i\ ,\ 1 \le i \le m\ . \tag{7}$$

We shall suppose that the measurement error is random; random measurement error is often referred to, at least informally, as "noise." We shall assume the following of the noise $\epsilon_i, 1 \le i \le m$:

> *zero-mean*: $E(\epsilon_i) = 0, 1 \le i \le m$, where E refers to expectation; in words, there is no *systematic* measurement error.
>
> *homoscedastic*: $E(\epsilon_i^2) = \sigma^2, 1 \le i \le m$. Note since ϵ_i is zero-mean, $E(\epsilon_i^2)$ is the variance of the noise; homoscedasticity is the property that the measurement errors, $\epsilon_i, 1 \le i \le m$, all share the same variance. The standard deviation σ is a measure of the magnitude of the noise: $\sigma = 0$ corresponds to perfect measurements — no measurement error.
>
> *independent*, or *uncorrelated*: $E(\epsilon_i \ _k) = 0, 1 \le i \le n, k \ne i$; in words, on average, the product of the errors at two measurement locations vanishes.

We emphasize that since the $\ _i, 1 \le i \le m$, are random variables, so too are our measurements, $z_i^{\text{meas}}, 1 \le i \le m$; in each realization of the experiment, we will obtain different values for $z_i^{\text{meas}}, 1 \le i \le m$. Furthermore, we note from (7) that, since $E(\epsilon_i) = 0, 1 \le i \le m$, then $E(z_i^{\text{meas}}) = z^{\text{true}}(t_i), 1 \le i \le m$; in words, for any given time t_i, the expected value of the *measurement of the height* is equal to the *true value of the height*.

We note that (7) is not yet a useful equation for the estimation of β^{true} since in fact β^{true} does not appear. We thus now take advantage of the relation $z^{\text{true}}(t) = z^{\text{model}}(t; \beta^{\text{true}})$ to write

$$z_i^{\text{meas}} = z^{\text{model}}(t; \beta^{\text{true}}) + \epsilon_i \ , 1 \le i \le m \ . \tag{8}$$

The equation (8) now relates β^{true} to our measurements, and can serve, as we describe in the next section, to develop our estimate $\hat{\beta}$ for β^{true}. However, (8) also highlights a fundamental hypothesis implicit in our proceedings: there must exist a β, β^{true}, such that $z^{\text{true}}(t) = z^{\text{model}}(t; \beta^{\text{true}})$. In such cases we might say that our model is adequate, or consistent; equivalently, we might say that there is no model error, or that there is no model bias. In our model is not adequate, we must include an additional model error contribution on the right-hand side of (8); we may then interpret $z^{\text{model}}(t; \beta^{\text{true}})$ as the first few terms — and β^{true} as the first few coefficients — associated with a larger, complete, expansion for which there is no model error. It is important to admit that some model error is inevitable: rarely can we understand the full complexity of the physical world, and even more rarely can we represent this complexity with a relatively simple parametrized model. However, often the model error will be small, in which case the corresponding effect on $\hat{\beta}$ will also be small, at least for a well-designed experiment. We return to this point on several occasions.

Finally, we introduce the notion of *synthetic experiments*. In synthetic experiments we artificially create "data": we postulate an appropriate $z^{\text{true}}(t)$ associated with a particular choice for $\beta_j^{\text{true}}, 0 \le j \le 2$, and suitable noise $\epsilon_i, 1 \le i \le m$, consistent with our assumptions on measurement error. We then create realizations of our experiment based on pseudo-random variate generation. Synthetic experiments, often denoted simply *synthetic data*, are very useful in the development, analysis, and interpretation of estimation procedures. We shall often illustrate our formulation with synthetic data. However, it is important to always bear in mind that synthetic data are no substitute for real data: real data force us to assess the validity of, and if necessary update, our assumptions: is the model adequate — is the model error zero, or at least "small"? is the noise approximately homoscedastic? is the noise approximately independent?

We shall assume for the purposes of our numerical experiments that the synthetic noise $\epsilon_i, 1 \le i \le m$, is normal, (and perforce zero-mean, homoscedastic with variance σ^2, and independent, per

our general assumptions). Note that to create synthetic noise we must choose a particular, concrete, probability density. The normal density is often a good representation of real noise, however other choices are certainly also possible. Note that for each experiment the realization of the noise shall of course be different.

3.2 Least Squares Formulation

We now develop a procedure by which to find $\hat{\beta}$, our estimate for β^{true}. Our point of departure is equation (8) which relates the measurements and our parametrized model. Our goal of course is a technique which yields better and better results as we increase the number of measurements, m, or decrease the magnitude σ of the noise: as $m \to \infty$, $\hat{\beta}$ should approach β^{true}, and $\hat{z}(t) \equiv z^{\text{model}}(t;\hat{\beta})$ should approach $z^{\text{true}}(t) \equiv z^{\text{model}}(t;\beta^{\text{true}})$.

Let us say that we first consider only two measurements, $m = 2$. To determine β^{true} we shall invoke (8): since we do not know the experimental error $_i$, $1 \le i \le m$, we shall simply neglect this term in (8). (In fact, as we shall see, this obvious strategy is not, ultimately, the good strategy.) We shall thus look for a particular β, $\hat{\beta}$, that satisfies

$$(\hat{z}(t_1) \equiv) z^{\text{model}}(t_1;\hat{\beta}) = z_1^{\text{meas}} , \tag{9}$$

$$(\hat{z}(t_2) \equiv) z_{\text{model}}(t_2;\hat{\beta}) = z_2^{\text{meas}} . \tag{10}$$

In short, we ask for $\hat{\beta}$ such that our approximate trajectory agrees with our available experimental measurements at times t_1 and t_2. We can now apply (3) to expand these equations as

$$\hat{\beta}_0 + \hat{\beta}_1 t_1 + \hat{\beta}_2 t_1^2 = z_1^{\text{meas}} , \tag{11}$$

$$\hat{\beta}_0 + \hat{\beta}_1 t_2 + \hat{\beta}_2 t_2^2 = z_2^{\text{meas}} ; \tag{12}$$

note the equations are *linear* in $\hat{\beta}$. We observe that we have three unknowns — $\hat{\beta}_0$, $\hat{\beta}_1$, and $\hat{\beta}_2$ — but only two equations. As your intuition would suggest, there is no unique solution for $\hat{\beta}$, and hence (11), (12) are clearly insufficient: how would we choose the actual β^{true} from all the $\hat{\beta}$ consistent with (11), (12)? We say that (11), (12) are *underdetermined*.

CYAWTP 1. We ask you to illustrate the insufficiency of two measurements. You are given synthetic height data (associated with noise standard deviation $\sigma = 0.025$) of $z_1^{\text{meas}} = 1.927$ m and $z_2^{\text{meas}} = 1.307$ m associated with measurement times $t_1 = 0.2$ s and $t_2 = 0.4$ s, respectively. Draw a sketch of $\hat{z}(t) \equiv z^{\text{model}}(t, \hat{\beta})$ for two different values of $\hat{\beta}$ which satisfy (11), (12); note each curve represents a possible falling ball trajectory consistent with the measurements. (Of course, by definition, there is no unique answer — no unique sketch — for this question.)

We consider next three measurements, $m = 3$. Our equations for $\hat{\beta}$ would now read

$$(\hat{z}(t_1) \equiv) z^{\text{model}}(t_1;\hat{\beta}) = z_1^{\text{meas}} , \tag{13}$$

$$(\hat{z}(t_2) \equiv) z^{\text{model}}(t_2;\hat{\beta}) = z_2^{\text{meas}} , \tag{14}$$

$$(\hat{z}(t_3) \equiv) z^{\text{model}}(t_3;\hat{\beta}) = z_3^{\text{meas}} . \tag{15}$$

As before, we can expand these equations as

$$\hat{\beta}_0 + \hat{\beta}_1 t_1 + \hat{\beta}_2 t_1^2 = z_1^{\text{meas}} \ , \tag{16}$$

$$\hat{\beta}_0 + \hat{\beta}_1 t_2 + \hat{\beta}_2 t_2^2 = z_2^{\text{meas}} \ , \tag{17}$$

$$\hat{\beta}_0 + \hat{\beta}_1 t_3 + \hat{\beta}_2 t_3^2 = z_3^{\text{meas}} \ ; \tag{18}$$

we now have three linear equations in three unknowns. In this case we are squarely in the interpolation context: we wish to "put" a quadratic through three points; there is a unique solution for $\hat{\beta}$, presuming that the measurements are taken at distinct times, $t_1 < t_2 < t_3$.

Mathematical well-posedness notwithstanding, our solution to (16)-(18) is suspect due to the noise in the measurements: in general, $z_i^{\text{meas}} \neq z^{\text{true}}(t_i), 1 \leq i \leq m$, such that $\hat{\beta} \neq \beta^{\text{true}}$ and hence $z^{\text{model}}(t; \hat{\beta}) \neq z^{\text{model}}(t; \beta^{\text{true}})$; equivalently, at the three times t_1, t_2, t_3, we are interpolating not the (true) trajectory but rather the (true) trajectory *plus* measurement noise. The latter is known as *overfitting*; the "over" refers to excessive adherence to data which in fact are not exact. In short, only if the noise is small should we anticipate that our estimates $\hat{\beta}$ (solution of (16)-(18)) and $\hat{z}(t) = z^{\text{model}}(t; \hat{\beta})$ shall be close to β^{true} and $z^{\text{true}}(t)$, respectively.

CYAWTP 2. Invoke the Falling Ball GUI: truth $z^{\text{true}}(t) = 2 - (9.81/2)t^2$ over the interval $(0, 0.5)$; quadratic parametrized model as given in (3); zero-mean normal noise, homoscedastic with variance σ^2, independent; $m = 3$ measurements equispaced in time, $t_i = 0.5(i-1)/(m-1), 1 \leq i \leq m(= 3)$. Perform a synthetic experiment for $\sigma = 0.05$: Does $\hat{z}(t)$ well approximate $z^{\text{true}}(t)$? Is $\hat{\beta}_2$ close to $\beta_2^{\text{true}} = -(9.81/2)$? Now repeat the synthetic experiment for $\sigma = 0.01$: Does $\hat{z}(t)$ well approximate $z^{\text{true}}(t)$? Is $\hat{\beta}_2$ close to $\beta_2^{\text{true}} = -(9.81/2)$?

How can we incorporate additional data to improve our estimate? Let us say that we consider $m > 3$ measurements. We can no longer hope, in general, to pass a quadratic polynomial through $m > 3$ points (t_i, z_i^{meas}), $1 \leq i \leq m$; the equations are *overdetermined*. Our strategy of simply neglecting ϵ_i in (8) is thus not only flawed but also unsustainable as we increase the number of measurements, m. How then can we find $\hat{\beta}$ and subsequently $\hat{z}(t)$? We shall look for $\hat{\beta}$, an approximation to β^{true}, which minimizes — over all possible values of $\beta \equiv (\beta_0\, \beta_1\, \beta_2)^{\text{T}}$ — the goodness-of-fit function $J(\beta)$ given by

$$J(\beta) \equiv \sum_{i=1}^{m} (z_i^{\text{meas}} - z^{\text{model}}(t_i; \beta))^2 \ . \tag{19}$$

In words, we are looking for coefficients for our quadratic model that will minimize the sum of the squares of the deviations between the measurements and the model predictions: least–squares minimization.[2] Once we obtain $\hat{\beta}$ from the minmization of J, our approximation to the trajectory, $\hat{z}(t)$, is then given by $\hat{z}(t) \equiv z^{\text{model}}(t; \hat{\beta})$.

CYAWTP 3. Argue that it is *not* possible for $\beta = \hat{\beta}$ — the value of β which minimizes $J(\beta)$ of (19) — to satisfy either $\hat{z}(t_i)(\equiv z^{\text{model}}(t_i; \hat{\beta})) < z_i^{\text{meas}}$, $1 \leq i \leq m$, or $\hat{z}(t_i)(\equiv z^{\text{model}}(t_i; \hat{\beta})) > z_i^{\text{meas}}, 1 \leq i \leq m$. In other words, the approximate trajectory $\hat{z}(t)$ must "go through the data" in the sense that $\hat{z}(t)$ can not lie entirely outside the envelop of the data.

[2] We can also choose other norms for the goodness-of-fit function — for example, the *maximum* deviation. However, the sum of squares leads to a system of equations for the $\hat{\beta}$ — in particular, a *linear* system of equations — for which an efficient computational procedure exists.

We can now hope that, as we increase m, $\hat{\beta}$ which minimizes $J(\beta)$ will approach β^{true}, and hence also $\hat{z}(t) = z^{\text{model}}(t;\hat{\beta})$ will approach $z^{\text{true}}(t) = z^{\text{model}}(t;\beta^{\text{true}})$. Why? We first note that since $z_i^{\text{meas}}, 1 \le i \le m$, is a random variable, so too is our estimator, $\hat{\beta}$. We can also show that $E(\hat{\beta}) = \beta^{\text{true}}$ — the mean of $\hat{\beta}$ is the true parameter value β^{true} — and hence $\hat{\beta}$ is an unbiased estimator for β^{true}. Finally, we can interpret $\hat{\beta}$ as the sample mean of an appropriate function of our m measurements, $z_i^{\text{meas}}, 1 \le i \le m$. We might thus anticipate, correctly, that $\hat{\beta}$ will converge to its mean — our true parameter value, β^{true} — as $1/\sqrt{m}$. (This argument is irresponsibly oversimplified, but the underlying intuition and the final result is indeed correct.)

CYAWTP 4. Invoke the Falling Ball GUI: truth $z^{\text{true}}(t) = 2-(9.81/2)t^2$ over the interval $(0, 0.5)$; quadratic parametrized model as given in (3); zero-mean normal noise, homoscedastic with variance σ^2, independent; m measurements equispaced in time, $t_i = 0.5(i-1)/(m-1), 1 \le i \le m$. Consider noise standard deviation $\sigma = 0.05$ and $m = 10, 1000$, and 100000 measurements. As m increases, does $\hat{\beta}_2$ approach $\beta_2^{\text{true}} = (-9.81/2)$? How rapidly — what is the order of convergence? Does $\hat{z}(t)$ appear to approach $z^{\text{true}}(t)$?

4 A General Model Framework

4.1 Abstraction

Before we proceed further we provide an abstraction that will permit you to apply the least-squares methodology easily in many contexts.

Independent and Dependent Variables. We shall assume that we consider some physical system with p independent variables $x = (x_{(1)}, x_{(2)}, \ldots, x_{(p)})$ and a single dependent variable y. (Here of course "independent" does not denote the statistical significance of the word.) It can be advantageous to choose nondimensional independent and dependent variables so as to reduce p and also avoid disparate magnitudes.

Parametrized Model and Truth. We are given (or more typically, we choose) n prescribed functions $h_j(x), 1 \le j \le n-1$. We may then introduce an $n \times 1$ vector of parameters $\beta = (\beta_0\ \beta_1\ \cdots\ \beta_{n-1})$ and an associated parametrized model — a model "expansion" — of the form

$$y^{\text{model}}(x;\beta) = \beta_0 + \sum_{j=1}^{n-1} \beta_j h_j(x)\ . \tag{20}$$

We then assume that our dependent variable y depends on x as

$$y^{\text{true}}(x) = y^{\text{model}}(x;\beta^{\text{true}}) \tag{21}$$

for some value of β^{true}.

Measurements and Noise. We next introduce m observations, $(x_i, y_i^{\text{meas}}), 1 \le i \le m$, where x_i and y_i^{meas} are respectively given (more typically, chosen) values of the independent variable and corresponding measured values of the dependent variable. We define the experimental error as

$$\epsilon_i \equiv y_i^{\text{meas}} - y^{\text{true}}(x_i)\ ,\ 1 \le i \le m\ , \tag{22}$$

in terms of which we can express our measurements as

$$y_i^{\text{meas}} = y^{\text{true}}(x_i) + \epsilon_i \ , \ 1 \leq i \leq m \ . \tag{23}$$

We shall suppose that the measurement error is zero-mean, homoscedastic with unknown variance σ^2, and independent: for $i = 1, \ldots, m$, $E(\epsilon_i) = 0$, $E(\epsilon_i^2) = \sigma^2$, and, for $k \neq i$, $E(\epsilon_i \epsilon_k) = 0$.

Note that our formulation — in particular (21) — presumes *no* model error.

The selection of the $h_j, 1 \leq j \leq n-1$, is intended to ensure model adequacy (or equivalently, zero — in practice, small — model error); we discuss this point further below. We note that the form (20) assumes that our model always includes the constant function, multiplied by the coefficient β_0. Although not absolutely essential, this requirement ensures various useful properties in the resulting least-squares estimate. The remainder of the functions, $h_j(x), 1 \leq j \leq n-1$, can be chosen to suit the problem at hand. In total there are n functions in our expansion: we may also write our model as

$$y^{\text{model}}(x; \beta) = \sum_{j=0}^{n-1} \beta_j h_j(x) \tag{24}$$

for $h_0(x)$ defined as $h_0(x) \equiv 1$. We emphasize that although we permit *any form* for the functions $h_j(x), 1 \leq j \leq n-1$, we do require that the coefficient vector β appears *linearly*.

The selection of the x_i — the measurement points, or measurement sites — is known as the "design" of the experiment, and is intended to desensitize the estimate $\hat{\beta}$, to the extent possible, to the noise in the measurements. We note that the x_i need not be distinct: we can, and often will, take several measurements at the same value of the independent variable. On the notational side, we emphasize that each experiment i corresponds to a particular value of the independent variable, x_i, which we must interpret as $x_i \equiv (x_{(1)}, x_{(2)}, \ldots, x_{(p)})_i$: all p components of x_i are specified. (We provide the parentheses for the subscripts associated with the p components of x in order to avoid confusion with the subscript associated with the m experiments.)

We note that what we describe in this section are essentially the "inputs" to the estimation procedure. The "outputs" will be estimates $\hat{\beta}$ for β^{true} and $\hat{y}(x)$ for $y^{\text{true}}(x)$. Before discussion of the estimation procedure, we relate several of our previous examples to the canonical form.

4.2 Reduction to "Canonical" Form

The first step in the analysis of a new problem is the "mapping" of the model to the variables of our abstraction; once mapped, we may then apply the general procedure which we shall develop in the next section. Abstraction provides the advantages of encapsulation and re-use, and in this sense is equivalent to — and often embodied as — a *function* in a programming language.

4.2.1 Falling Ball

As a first example, let us consider the falling ball, as introduced in Section 3.1. We immediately identify $p \equiv 1$ (or $p \leftarrow 1$) and $x \equiv t$: we have a single independent variable, t, which we henceforth refer to as x. Next we identify $y \equiv z$: our dependent variable is height, z, which we henceforth

refer to as y. We next specify $n \equiv 3$ such that $\beta \equiv (\beta_0\ \beta_1\ \beta_2)^{\mathrm{T}}$. We further identify the n functions in our expansion as ($h_0(x) \equiv 1$, as always, and) $h_1(x) \equiv x, h_2(x) \equiv x^2$, such that

$$y^{\text{model}}(x;\beta) = \beta_0 + \beta_1 x + \beta_2 x^2 \ ; \tag{25}$$

note that if we replace in (25) y with z and x with t we exactly recover (3), as desired. Finally, we note that $y^{\text{true}}(x) = y^{\text{model}}(x;\beta^{\text{true}})$.

4.2.2 Block on Surface

Let us consider the friction example of Section 2. We identify $p \equiv 2$ independent variables: $x_{(1)} \equiv F_{\text{normal}}$, the normal force exerted by the block on the surface (for example, if the surface is oriented perpendicular to gravity, then in the absence of applied forces, F_{normal} would simply be the weight of the block); $x_{(2)} \equiv A_{\text{surface}}$, the superficial contact area between the block and the surface. Next we identify our dependent variable as $y \equiv F^{\text{max}}_{\text{f, static}}$, the maximum tangential force which can be applied to the the block such that *no slip* occurs. We next specify $n \equiv 3$ such that $\beta \equiv (\beta_0\ \beta_1\ \beta_2)$. We further identify the n functions in our expansion as ($h_0(x) = 1$, as always, and) $h_1(x) \equiv x_{(1)}$, $h_2(x) \equiv x_{(2)}$, such that

$$y^{\text{model}}(x;\beta) = \beta_0 + \beta_1 x_{(1)} + \beta_2 x_{(2)} \ . \tag{26}$$

note that if we replace in (26) y with $F^{\text{max}}_{\text{f, static}}$ and $x_{(1)}$ with F_{normal}, $x_{(2)}$ with A_{surface}, we exactly recover (1), as desired. Note that β_1 represents the coefficient of friction, μ, and furthermore we anticipate $\beta_2 = 0$ if Amontons' Law is correct.

4.2.3 IR Transducer

Let us consider the IR Transducer example of Section 2. In this case, we require some pre-processing if we are to accommodate the problem within the general framework. In particular, we observe that in the functional relationship $D^{\text{model}}(V) = CV^{-\gamma}$ the second parameter of interest, γ, does not appear linearly. However, if we apply a "log transformation" — take logs of both sides — we obtain

$$(\log D)^{\text{model}} = \log(C) - \gamma \log V) \ . \tag{27}$$

We now map this modified functional form to the general abstraction. In particular, we identify $p \equiv 1$ independent variable, $x_{(1)} \equiv \log V$. Next we identify our dependent variable as $y \equiv \log D$. We then specify $n \equiv 2$ and choose ($h_0(x) \equiv 1$, as always, and) $h_1(x) \equiv x_{(1)}$, such that

$$y^{\text{model}}(x;\beta) = \beta_0 + \beta_1 x_{(1)} \ . \tag{28}$$

Finally, we note $y^{\text{true}}(x) = y^{\text{model}}(x;\beta^{\text{true}})$ which, upon replacement of y with $\log D$ and $x_{(1)}$ with $\log V$ yields (27) if we further identify $\beta_0 \equiv \log(C)$ and $\beta_1 \equiv -\gamma$. Note from the latter we understand that once we deduce $\hat{\beta}_0$ and $\hat{\beta}_1$ from our least-squares procedure, we must subsequently form our estimates for the calibration constants C^{true} and γ^{true}, $\hat{C}$ and $\hat{\gamma}$, as $\exp(\hat{\beta}_0)$ and $-\hat{\beta}_1$, respectively.

CYAWTP 5. The relaxation of a first-order dynamical system perturbed from equilibrium is given by $u(t) = Ce^{-t/\tau}$, where C is the size of the perturbation and τ is the time constant. Apply the "log transformation" to reveal a *linear* regression problem and identify, for the transformed system, the reduction to canonical form: p, x, y, n, and $h_j(x), 0 \leq j \leq n-1$. Finally, indicate how estimates $\hat{C}$ and $\hat{\tau}$ can be derived from the least-squares parameter estimates.

CYAWTP 6. In fact, our additive model for the friction force of (1) is suspect, and inasmuch perhaps not a sufficiently severe test of Amontons' Law. In particular, (1) is plausible as a local (smoothness) model, but not as a more global (mechanistic) model: we anticipate that for $F_{\text{normal}} = 0$ we will perforce obtain $F^{\max}_{\text{f, static}} = 0$, and hence also β_0 and β_1 would vanish. As an alternative model, we might consider

$$(F^{\max}_{\text{f, static}})^{\text{model}}(F_{\text{normal}}, A_{\text{surface}}) = C F^{\alpha}_{\text{normal}} A^{\eta}_{\text{surface}} \ , \tag{29}$$

with parameters C, α, and η. if Amontons' Law is correct, then $\alpha = 1$ and $\eta = 0$ and C is the (dimensionless) coefficient of friction, μ. Apply the "log transformation" to reveal a *linear* regression problem and identify, for the transformed system, the reduction to canonical form: p, x, y, n, and $h_j(x), 0 \le j \le n-1$. Finally, indicate how estimates $\hat{C}$, $\hat{\alpha}$, and $\hat{\eta}$ for C^{true}, α^{true}, and η^{true}, respectively, may be deduced from the linear-regression least squares parameter estimates.

4.3 Model Considerations

There are perhaps two fundamental ways to develop a model: a "mechanistic" (or physics–based) approach; and a "smoothness" approach. In the mechanistic approach, the model is typically based on some underlying laws which govern the particular physical, or maybe social, phenomena of interest. Examples from Section 2 of this mechanistic approach include the falling ball — informed by Newton — and friction — informed by Amontons. In the smoothness approach we exploit only the continuity of the function and some low-order derivatives, as embodied (say) in a Taylor-series expansion; we can expect that these models — sometimes denoted "response surfaces" — will be broadly applicable but only in some rather limited range of the independent variable. Examples from Section 2 of this smoothness approach include the temperature gradient and the asynchronous quiz, both of which may be viewed as first-order Taylor series.

We of course wish to develop a model for which the model error is zero: we would like to ensure that there exists a β^{true} such that $E(y^{\text{meas}}_i) = y^{\text{true}}(x_i) = y^{\text{model}}(x_i; \beta^{\text{true}}), 1 \le i \le m$. We should not *underfit*: we should not omit independent variables, $x_{(1)}, x_{(2)}, \ldots, x_{(p)}$, which we believe will be relevant, or functional dependencies, $h_j, 1 \le j \le n$, which we believe will contribute significantly to $y^{\text{true}}(x)$. If we underfit, we will be unable to identify or represent the true behavior of our system. However, we should also not *overfit*: we should not choose p very large — include all possible "effects" — and n very large — include all possible functional dependencies — and then let the data decide what is important. If we overfit, we will effectively fit our model to noise rather than signal (in particular if n is on the order of m).

CYAWTP 7. Invoke the Least Squares Fit GUI: truth $y^{\text{true}}(x) = 2 - (9.81/2)x^2$ over the interval $(0, 1)$; zero-mean normal noise, homoscedastic with standard deviation $\sigma = 0.05$, independent; m measurements equispaced over the interval, $x_i = (i-1)/(m-1), 1 \le i \le m$.

> First, just right: consider $n = 3$ and $h_0(x) = 1, h_1(x) = x, h_2(x) = x^2$. Is $\hat{y}(x)$ a reasonable approximation to $y^{\text{true}}(x)$ for $m = 16$? Does $\hat{y}(x)$ converge to $y^{\text{true}}(x)$ as m increases? Does the least-squares estimate $\hat{\beta}$ converge to β^{true} as m increases?
>
> Second, significant underfit: consider $n = 2$ and $h_0(x) = 1, h_1(x) = x$. Is $\hat{y}(x)$ a reasonable approximation to $y^{\text{true}}(x)$ for $m = 16$? Does $\hat{y}(x)$ converge to $y^{\text{true}}(x)$ as m increases? Does the least-squares estimate $\hat{\beta}$ converge to β^{true} as m increases?

Third, significant overfit: consider $n = 16$ and $h_j(x) = x^j$, $0 \le j \le n-1$.[3] Is $\hat{y}(x)$ a reasonable approximation to $y^{\text{true}}(x)$ for $m = 16$? Does $\hat{y}(x)$ converge to y^{true} as m increases? Does the least squares estimate $\hat{\beta}$ converge to β^{true} as m increases?

You might also investigate the effect of smaller or larger noise.

We note from **CYAWTP 7** that significant underfit seriously degrades the accuracy of our estimates. However, we re-iterate that some small underfit is acceptable: as indicated earlier, a small model error typically results in a commensurately small perturbation to our least squares estimates.

CYAWTP 8. Invoke the Least Squares Fit GUI: truth $y^{\text{true}}(x) = 1 + x^2 + 0.025\sin(\pi x)$ over the interval $(0, 1)$; zero-mean normal noise, homoscedastic with standard deviation $\sigma = 0.05$, independent; $n = 3$ and $h_0(x) = 1, h_1(x) = x, h_2(x) = x^2$; m measurements equispaced over the interval, $x_i = (i-1)/(m-1), 1 \le i \le m$. Note that our model includes some underfit, due to the $0.025\sin(\pi x)$ term in the truth but absent from the model, and also some overfit, due to the x term in the model but absent from the truth. As m increases, does the least-squares procedure provide a reasonable estimate $\hat{\beta}_2$ for $\beta_2^{\text{true}} = 1$? A reasonable estimate $\hat{y}(x)$ for $y^{\text{true}}(x)$? Does $\hat{\beta}_2$ (respectively, $\hat{y}(x)$) converge to β_2^{true} (respectively, $y^{\text{true}}(x)$) as $m \to \infty$?

Finally, although in this section we emphasize the model, the accuracy of our estimates is determined both by the adequacy of $y^{\text{model}}(x; \beta)$ *and* the design of the experiment — the choice of the measurement points. A good design can mitigate the effect of both underfit and overfit.

CYAWTP 9. Invoke the Least Squares Fit GUI: truth $y^{\text{true}}(x) = 2 - (9.81/2)x^2$ over the interval $(0, 1)$; zero-mean normal noise, homoscedastic with standard deviation $\sigma = 0.1$, independent; $n = 16$ with $h_j(x) = x^j, 0 \le j \le n-1$; $m = 16$ measurements over the interval. Consider first a distribution of measurement points equispaced over the interval, $x_i = (i-1)/(m-1), 1 \le i \le m \equiv 16$: Is $\hat{y}(x)$ a reasonable approximation to $y^{\text{true}}(x)$? Consider now a distribution of measurement points clustered near the two ends of the interval, $x_i = (-\cos(\pi(i-1)/(m-1)) + 1)/2, 1 \le i \le m \equiv 16$: Is $\hat{y}(x)$ now a (more) reasonable approximation to $y^{\text{true}}(x)$?

5 Least Squares Formulation

5.1 Minimization Statement

We may now proceed to pose our least squares formulation for our general model formulation. To start, we define a *residual* associated with the i^{th} observation as

$$\begin{aligned} r_i(\beta) &\equiv y_i^{\text{meas}} - y^{\text{model}}(x_i; \beta) \\ &\equiv y_i^{\text{meas}} - \sum_{j=0}^{n-1} \beta_j h_j(x) \ , \end{aligned} \tag{30}$$

for $1 \le i \le m$; we can also then define the $m \times 1$ residual vector, $r(\beta) \equiv (r_1(\beta)\, r_2(\beta)\, \cdots r_m(\beta))^{\text{T}}$. In words, $r_i(\beta)$ is the difference, or discrepancy, between the i^{th} experimental measurement and the

[3] Note that the monomial functions $h_j(x) = x^j$, $0 \le j \le n-1$, in **CYAWTP 7** and also **CYAWTP 9**, are dangerously linearly dependent, which further aggravates the effects of overfitting on the accuracy of $\hat{\beta}$. Monomials are an extremely poor choice except for n quite small; they serve here only for illustrative purposes.

model prediction for parameter value β. We now wish to find that value of β, $\hat{\beta} = (\hat{\beta}_0\, \hat{\beta}_1\, \cdots, \beta_n)^{\mathrm{T}}$ — our "best" parameter estimate — which minimizes $J(\beta)$ given by

$$\begin{aligned} J(\beta) &\equiv \sum_{i=1}^{m} r_i^2(\beta) \\ &\equiv \|r(\beta)\|^2 \,. \end{aligned} \tag{31}$$

Once we obtain our parameter estimate $\hat{\beta}$ we may then form

$$\hat{y}(x) \equiv y^{\mathrm{model}}(x; \hat{\beta}) \tag{32}$$

as our approximation to $y^{\mathrm{true}}(x) \equiv y^{\mathrm{model}}(x; \beta^{\mathrm{true}})$. This least-squares formulation is simply a generalization of the procedure developed in Section 3.2.

We now express the residual more succinctly. Towards that end, we first note that

$$\begin{pmatrix} y^{\mathrm{model}}(x_1;\beta) \\ y^{\mathrm{model}}(x_2;\beta) \\ \vdots \\ y^{\mathrm{model}}(x_m;\beta) \end{pmatrix} \equiv \begin{pmatrix} \beta_0 + \sum_{j=1}^{n-1} h_j(x_1) \\ \beta_0 + \sum_{j=1}^{n-1} h_j(x_2) \\ \vdots \\ \beta_0 + \sum_{j=1}^{n-1} h_j(x_m) \end{pmatrix} \equiv \begin{pmatrix} 1 & h_1(x_1) & \cdots & h_{n-1}(x_1) \\ 1 & h_1(x_2) & \cdots & h_{n-1}(x_2) \\ \vdots & \vdots & \vdots & \\ 1 & h_1(x_m) & \cdots & h_{n-1}(x_m) \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_{n-1} \end{pmatrix} . \tag{33}$$

We next introduce the $m \times n$ matrix

$$X \equiv \begin{pmatrix} 1 & h_1(x_1) & \cdots & h_{n-1}(x_1) \\ 1 & h_1(x_2) & \cdots & h_{n-1}(x_2) \\ \vdots & \vdots & \vdots & \\ 1 & h_1(x_m) & \cdots & h_{n-1}(x_m) \end{pmatrix} . \tag{34}$$

The matrix X, which reflects both our model and the choice of measurement points, $x_i, 1 \leq i \leq m$, is often denoted the "design matrix," as it reflects the design of our experiment. We may develop an explicit expression for X: $X_{ij} = h_j(x_i), 1 \leq i \leq m, 0 \leq j \leq n-1$; note that we index the columns from zero, consistent with our enumeration of the functions $h_j(x)$, $0 \leq j \leq n-1$. We may now re-write (33) in terms of X as

$$\begin{pmatrix} y^{\mathrm{model}}(x_1;\beta) \\ y^{\mathrm{model}}(x_2;\beta) \\ \vdots \\ y^{\mathrm{model}}(x_m;\beta) \end{pmatrix} \equiv X\beta \,. \tag{35}$$

We thus see that $(X\beta)_i$ — the i^{th} entry of the column vector $X\beta$ — is the prediction of the model at $x = x_i$ for some given β.

As practice, we consider the formation of the X matrix in several cases. (We choose m small for convenience; we know that, in actual practice,we should choose m substantially larger than n.)

CYAWTP 10. Consider the falling ball example expressed in canonical form per Section 4.2.1. Find the matrix X — indicate each entry, $X_{ij}, 1 \leq i \leq m, 0 \leq j \leq n-1$ — for the particular case of $m = 5$ with $x_1 = 0.2, x_2 = 0.4, x_3 = 0.4, x_4 = 0.5$, and $x_5 = 0.6$ (which in fact correspond to times, in seconds). Given $\hat{\beta} = (1.9 \;\; -0.01 \;\; 5.0)^{\mathrm{T}}$, interpret and evaluate $X\hat{\beta}$.

CYAWTP 11. Consider the friction force example expressed in canonical form per Section 4.2.2. Find the matrix X — indicate each entry, $X_{ij}, 1 \leq i \leq m, 0 \leq j \leq n-1$ — for the particular case of $m = 4$ with $x_1 = (0.9810, 1.2903), x_2 = (0.9810, 2.5806), x_3 = (1.9620, 1.2903)$, and $x_4 = (1.9620, 2.5806)$. (Note in each case, in the pair of independent variables, the first entry is the value of the normal force, in Newtons, and the second entry is the value of the superficial contact area, in cm^2.)

We next invoke (35) to express the components of our residual vector (30) as $r_i(\beta) = y_i^{\text{meas}} - (X\beta)_i, 1 \leq i \leq m$; the residual vector may thus be expressed in terms of y^{meas}, X, and β as

$$r(\beta) = y^{\text{meas}} - X\beta \; . \tag{36}$$

We may then further express $J(\beta)$ of (31) as

$$\begin{aligned} J(\beta) \equiv \|r(\beta)\|^2 =& \|y^{\text{meas}} - X\beta\|^2 \\ =& (y^{\text{meas}} - X\beta)^{\text{T}}(y^{\text{meas}} - X\beta) \; . \end{aligned} \tag{37}$$

We next expand out the right-hand side to obtain

$$\begin{aligned} J(\beta) \equiv & \; (y^{\text{meas}} - X\beta)^{\text{T}}(y^{\text{meas}} - X\beta) \\ \equiv & \; (y^{\text{meas}})^{\text{T}}(y^{\text{meas}} - X\beta) - (X\beta)^{\text{T}}(y^{\text{meas}} - X\beta) \\ \equiv & \; (y^{\text{meas}})^{\text{T}}y^{\text{meas}} - (y^{\text{meas}})^{\text{T}}(X\beta) - (X\beta)^{\text{T}}y^{\text{meas}} + (X\beta)^{\text{T}}(X\beta) \; . \end{aligned} \tag{38}$$

We now note that $(y^{\text{meas}})^T(X\beta)$ is a real scalar: the product of the $1 \times m$ matrix $(y^{\text{meas}})^{\text{T}}$ and the $m \times 1$ matrix $X\beta$. But the transpose of a scalar is simply equal to the original scalar, and hence

$$(y^{\text{meas}})^T(X\beta) = ((y^{\text{meas}})^{\text{T}}(X\beta))^{\text{T}} \; . \tag{39}$$

We now recall the "product transpose rule": for two matrices A and B such that the product AB is permissible, $(AB)^{\text{T}} = B^{\text{T}}A^{\text{T}}$. Hence

$$\begin{aligned} (y^{\text{meas}})^T(X\beta) &= ((y^{\text{meas}})^{\text{T}}(X\beta))^{\text{T}} && \text{(the transpose of the scalar is the original scalar)} \\ &= (X\beta)^{\text{T}}((y^{\text{meas}})^{\text{T}})^{\text{T}} && \text{(the product transpose rule)} \\ &= (X\beta)^{\text{T}}y^{\text{meas}} && \text{(the transpose of the transpose is the original matrix).} \end{aligned}$$

We may thus write (38) as

$$J(\beta) = (y^{\text{meas}})^{\text{T}}y^{\text{meas}} - 2(X\beta)^{\text{T}}y^{\text{meas}} + (X\beta)^{\text{T}}(X\beta) \; . \tag{40}$$

We now apply the product transpose rule to $(X\beta)^{\text{T}}$ to obtain $(X\beta)^{\text{T}}y^{\text{meas}} = \beta^{\text{T}}X^{\text{T}}y^{\text{meas}}$ and also $(X\beta)^{\text{T}}(X\beta) = \beta^{\text{T}}X^{\text{T}}X\beta$. Assembling all these results, we arrive at

$$J(\beta) \equiv (y^{\text{meas}})^{\text{T}}y^{\text{meas}} - 2\beta^{\text{T}}X^{\text{T}}y^{\text{meas}} + \beta^{\text{T}}X^{\text{T}}X\beta \tag{41}$$

as the final expression for our goodness-of-fit function.

CYAWTP 12. Confirm that each of the three terms on the right-hand side of (41) is a scalar: for each term, identify the dimensions of each of the factors, and then consider the rules of matrix multiplication to identify the dimensions of the result.

5.2 Normal Equations

We now wish to find $\hat{\beta}$ which minimizes $J(\beta)$ of (41). More precisely, we wish to find an explicit *linear* equation the solution of which will yield $\hat{\beta}$.

To begin, we consider the simplest possible situation: $n = 1$. In this case, $y^{\text{model}}(x; \beta)$ of (20) reduces to

$$y^{\,\text{model}}(x; \beta) = \beta_0 \,, \tag{42}$$

and X of (34) reduces to the $m \times 1$ matrix

$$X \equiv \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix} \,. \tag{43}$$

Let us further define

$$\overline{y}^{\text{meas}} \equiv \frac{1}{m} \sum_{i=1}^{m} y_i^{\text{meas}} \,; \tag{44}$$

note that $\overline{y}^{\text{meas}}$ is simply the average of the m measurements of our independent variable. It then follows that

$$X^{\mathrm{T}} y^{\text{meas}} = m \overline{y}^{\text{meas}} \text{ (a scalar)} \,, \tag{45}$$

and

$$X^{\mathrm{T}} X = m \text{ (a scalar)} \,. \tag{46}$$

We may thus write (41) in our particular ($n = 1$) case as

$$J(\beta_0) = (y^{\text{meas}})^{\mathrm{T}} y^{\text{meas}} - 2m \overline{y}^{\text{meas}} \beta_0 + m \beta_0^2 \,; \tag{47}$$

note that β_0 is a scalar, and hence $\beta_0^{\mathrm{T}} = \beta_0$. We observe that, for the case $n = 1$, $J(\beta)$ is simply a quadratic polynomial of (the scalar) β_0.

It follows that to minimize the quadratic (47) we should set

$$\frac{dJ}{d\beta_0}(\hat{\beta}_0) = 0 \,; \tag{48}$$

$\hat{\beta}_0$ is defined as the particular value of β_0 at which the derivative of $J(\beta_0)$ vanishes. Performing the differentiation, we find that (48) reduces to

$$-2m \overline{y}^{\text{meas}} + 2m \hat{\beta}_0 = 0 \,, \tag{49}$$

or

$$m \hat{\beta}_0 = m \overline{y}^{\text{meas}} \,. \tag{50}$$

(We choose not to divide through by m for reasons which shall become clear shortly.) We thus obtain

$$\hat{y}(x)(\equiv y^{\text{model}}(x;\hat{\beta}_0)) = \hat{\beta}_0 = \overline{y}^{\text{meas}} \tag{51}$$

as our approximation to $y^{\text{true}}(x) = \beta^{\text{true}}$. In fact, we are not quite done: we must also confirm the second-order conditions for a minimum,

$$\frac{d^2J}{d\beta_0^2}(\hat{\beta}_0) > 0\ . \tag{52}$$

In our case we find from (47) that

$$\frac{d^2J}{d\beta_0^2}(\hat{\beta}_0) = m\ , \tag{53}$$

and since $m > 1 (> 0)$ we conclude that (52) is satisfied and that (50) is indeed a *minimizer*. Finally, it should be clear, since $J(\beta_0)$ is a parabola, that (50) yields not just a local minimizer but in fact a global minimizer: $J(\hat{\beta}_0) \leq J(\beta_0)$ for all $\beta_0 \neq \hat{\beta}_0$.

The result (51) makes good sense: to best represent m measurements by a constant $\hat{\beta}_0$ — alternatively, to find the constant $\hat{\beta}_0$ which goes "through the middle" of the data — we should choose $\hat{\beta}_0$ as the average (mean) of the measurements. Alternatively, since the noise is homoscedastic and independent, we may view $y_i^{\text{meas}} = \beta_0 + \ _i, 1 \leq i \leq m$, as a random sample associated with a normal random variable with mean β^{true} and variance σ^2; $\overline{y}^{\text{meas}}$ is then (a realization of) the sample mean, which we know will converge to the mean, β^{true}, as $1/\sqrt{m}$.

We see from (45) and (46) that our minimization condition (50) is given — before substitution of particular values — by the equation

$$(X^{\text{T}}X)\hat{\beta} = X^{\text{T}}y^{\text{meas}}\ . \tag{54}$$

In fact, this equation is valid not only for $n = 1$, but for any n. Note that the matrix $X^{\text{T}}X$ is of dimension $n \times n$, and the matrix $X^{\text{T}}y^{\text{meas}}$ is of dimension n: (54) is a *linear* system of n equations in n unknowns (the $n \times 1$ vector β); the system (54) is known as the "normal" equations (for reasons related to the interpretation of (54) as the projection onto a linear space spanned by the columns of X). It can be shown that, *as long as the columns of X are independent*, the equation (54) will have a unique solution $\hat{\beta}$ which is furthermore the global minimizer of (41) and hence also (37).

We can formally then write

$$\hat{\beta} = (X^{\text{T}}X)^{-1}X^{\text{T}}y^{\text{meas}}\ . \tag{55}$$

This equation is important from a theoretical perspective, and also quite practical for small n — an almost explicit expression for $\hat{\beta}$. However, we caution that, except for very small n, we *should not* form the inverse $(X^{\text{T}}X)^{-1}$: much more efficient and stable numerical procedures exist for solution of the the normal equations (54). (Indeed, for larger n, we need never even form $X^{\text{T}}X$.)

CYAWTP 13. Consider a truth $y^{\text{true}}(x) = 1 + x$ over the interval $(0, 1)$; zero-mean normal noise, homoscedastic with variance $\sigma^2 = 0.05^2$, independent; $n = 2$ with $h_0(x) = 1$, $h_1(x) = x$; $m = 5$ with $x_1 = 0.2, x_2 = 0.4, x_3 = 0.4, x_4 = 0.7$, and $x_5 = 1.0$, and associated measurements $y^{\text{meas}} = (1.23\ 1.49\ 1.29\ 1.74\ 2.02)^{\text{T}}$. Find X, $X^{\text{T}}X$, $X^{\text{T}}y^{\text{meas}}$, and $\hat{\beta} = (X^{\text{T}}X)^{-1}X^{\text{T}}y^{\text{meas}}$; for the latter, the simple formula for the inverse of a 2×2 matrix should prove convenient. Evaluate and interpret $y^{\text{meas}} - X\hat{\beta}$.

Finally, we note that if the columns of the design matrix X are "close" to linearly dependent, even if not exactly linearly dependent, we will be in trouble (in particular in the context in which m is not large compared to n): our parameter estimates will be very sensitive to the noise. We should thus choose, when possible, model functions $h_j(x), 0 \leq j \leq n-1$, which are not too similar. An example of a poor choice is $h_j(x) = x^j, 0 \leq j \leq n-1$, on the interval $(0,1)$, as these functions all "look like" 0 for n sufficiently large, and hence the columns of X — recall $X_{ij} = h_j(x_i), 1 \leq i \leq m, 0 \leq j \leq n-1$ – are "close" to linearly dependent. We can mitigate the effect by selection of different model functions $h_j(x), 0 \leq j \leq n-1$, for which the columns of X, are more orthogonal; note, however, that our functions $h_j(x), 0 \leq j \leq n-1$, are often constrained — *imposed* — by the parameters we wish to extract or the inferences we hope to draw.

5.3 Inspection and Exploration

We have now described a general procedure by which to fit models to data. In a subsequent nutshell we will provide *a priori* error estimates, *a posteriori* confidence intervals, and hypothesis tests to further quantify our conclusions. But in fact any fitting exercise should always be accompanied by some analysis secondary to the formal least-squares procedure. In particular, we must always return to the raw data to confirm our assumptions on model and measurements.

We should always plot and compare $\hat{y}(x) = y^{\text{model}}(x; \hat{\beta})$ and $(x_i, y_i^{\text{meas}}), 1 \leq i \leq m$. In this way we can identify by inspection possible outliers or trends which could signal either experimental difficulties or perhaps model error. Note that these studies should never serve the purpose of cleansing the data of measurements which either offend the eye or compromise a pet theory, but rather should be presented in tandem with the least-squares predictions as additional information.

We can also readily plot a histogram of the residuals, $y^{\text{meas}} - X\hat{\beta}$. This histogram may be interpreted as an empirical noise density function. We can again identify "skewness" or outliers related perhaps to experimental protocol or model error. However, we can also investigate (or not...) normality, or approximate normality, of the measurement error. Note that Gaussian noise is not a requirement for either the practice or theory of least-squares fitting procedures, however normality can facilitate the application of more quantitive confidence intervals.

Finally, we can undertake additional and more quantitative explorations if the data includes several measurements (of the dependent variable, y) at each of several measurement points (independent variable, x): repetitions. In particular, we can calculate the sample covariance of the residuals between different measurement points in order to test the hypotheses of homoscedasticity — is the variance the same at different measurement sites? — and also independence — are the errors correlated at different measurement sites? Small departures from either homoscedasticity or independence pose little threat, but large deviations in particular from independence can compromise both the accuracy of the estimates and the validity of confidence intervals. Note repetition can also serve to identify model error: the average of many repetitions at a particular value of the independent variable, x, will converge to $E(z^{\text{true}}(x))$; deviations of the latter from $\hat{y}(x)$ can be tested for statistical significance — is the discrepancy beyond anticipated fluctuations?

6 Perspectives

We motivate and discuss least squares here from a very particular viewpoint: the fitting of a parametrized model to (noisy) data. We may also consider least squares from the much more

general perspective of linear algebra and — for ultimate computations — of numerical linear algebra. For a presentation of the former we recommend G Strang, "Introduction to Linear Algebra," 4^{th} Edition, Wellesley-Cambridge Press and SIAM, 2009; for a presentation of the latter we suggest LN Trefethen and D Bau, III, "Numerical Linear Algebra," SIAM, 1997.

We touch in this nutshell on just a very few of the many issues which inform the choice, analysis, and assessment of models, measurements, and noise. For a much more complete discussion of these topics, in particular from the statistical perspective, we recommend NR Draper and H Smith, "Applied Regression Analysis," 3^{rd} Edition, Wiley, 1998.

7 Appendix: Small Residual *vs* Small Error

A small residual is perhaps a necessary condition for a good fit, but it is certainly *not* a sufficient condition for a good fit: a small residual need not imply either an accurate estimate $\hat{\beta}$ for β^{true} or an accurate estimate $\hat{y}(x)$ for $y^{\text{true}}(x)$. We consider here a simple example: $y^{\text{true}}(x) = 1 + x + x^2$; a specific set of measurements $(x_i, y_i^{\text{meas}}), 1 \le i \le m = 6$. The level of the noise is not crucial for our arguments.

First consider the model $n = 6$ and

$$y^{\text{model}}(x;\beta) = \sum_{j=0}^{5} \beta_j x^j \ .$$

We first form the design matrix, which we denote $X^{6\times 6}$. We next compute $\hat{\beta}^6$ from (54). Finally, we evaluate the residual, $r^6 \equiv \|y^{\text{meas}} - X^{6\times 6}\hat{\beta}^6\|$. Note here the superscript 6 serves to distinguish our $n = 6$ fit from the other candidates we now introduce.

Next we consider the model $n = 3$ and

$$y^{\text{model}}(x;\beta) = \sum_{j=0}^{2} \beta_j x^j \ .$$

We first form the design matrix, $X^{6\times 3}$; we note for future reference that $X^{6\times 3}$ is in fact just the first three columns of $X^{6\times 6}$. We next compute $\hat{\beta}^3$ from (54). Finally, we evaluate the residual $r^3 \equiv \|y^{\text{meas}} - X^{6\times 3}\,\hat{\beta}^3\|$. The superscript 3 shall refer to our $n = 3$ fit.

Finally, we define $r^{\text{true}} = \|y^{\text{meas}} - X^{6\times 3}\beta^{\text{true}}\|$, where for our particular truth, $\beta^{\text{true}} = (1\ 1\ 1)^{\text{T}}$. We can then readily demonstrate that

$$r^{\text{true}} \ge r^3 \ge r^6 \ . \tag{56}$$

In contrast, β^{true} is more accurate than $\hat{\beta}^3$, since of course $\|\beta^{\text{true}} - \beta^{\text{true}}\| = 0 \le \|\ \hat{\beta}\ - \beta^{\text{true}}\|$. Furthermore, from **CYAWTP 7**, $\hat{\beta}^3$ is almost certainly more accurate than $\hat{\beta}^6$. In short, for this particular example, the error in our parameter estimate increases as the residual decreases.

We prove (56). We first demonstrate that $r^{\text{true}} \ge r^3$. We recall that $\hat{\beta}$ minimizes $J(\beta)$ as defined in (37). Hence

$$\|y^{\text{meas}} - X^{6\times 3}\hat{\beta}\| \le \|y^{\text{meas}} - X^{6\times 3}\beta\| \tag{57}$$

for *any* $\beta \in \mathbb{R}^3$. The choice $\beta = \beta^{\text{true}}$ in (57), and the definitions of r^{true} and r^3, then yields $r^{\text{true}} \geq r^3$. In words, $\hat{\beta}^3$ is optimized for our particular data, y^{meas}, and hence even the truth can not further reduce the residual.

We next demonstrate that $r^3 \geq r^6$. We again recall that $\hat{\beta}$ minimizes $J(\beta)$ as defined in (37). Hence

$$\| y^{\text{meas}} - X^{6\times 6}\, \hat{\beta}^6 \,\| \leq \| y^{\,\text{meas}} - \tag{58}$$

$X^{6\times 6}\beta\|$ for *any* 6–vector β. We now choose in (58) $\beta = \bar{\beta}^3$, where

$$\bar{\beta}^3 \equiv \begin{pmatrix} \hat{\beta}^3_1 \\ \hat{\beta}^3_2 \\ \hat{\beta}^3_3 \\ 0 \\ 0 \\ 0 \end{pmatrix} .$$

But $X^{6\times 6}\bar{\beta}^3 = X^{6\times 3}\,\hat{\beta}^3$, and hence from (58) and the definition of r^6 and r^3 we conclude that $r^3 \geq r^6$. The small residual r^6 is of course achieved by overfit — not by honest representation of $y^{\text{true}}(x)$.

2.086 Numerical Computation for Mechanical Engineers
Fall 2014