Regression: Least Squares and Statistical Inference... in a Nutshell

AT Patera, JD Penn, M Yano

October 9, 2014

1 Preamble

Least squares may be viewed as a best-fit procedure or as a statistical estimation procedure. There is much overlap between the two perspectives but the emphasis can be different: approximation in the best-fit context, and inference in the statistical estimation context. In this nutshell we summarize the intepretation of least-squares estimators from a statistical perspective. Note that we do not rederive the least-squares results for example as a maximum-likelihood estimator; rather, we take the estimators as given, and overlay the statistical interpretation.

In this nutshell:

> We provide an *a priori* analysis for the error in the least-squares parameter estimates as a function of the variance of the measurement noise and the regression design matrix, X. We identify the important role of independence in convergence of the parameter estimates as we increase the number of experiments, m.
>
> We provide an individual two-sided confidence interval for any least-squares parameter estimate under the assumptions of zero model error and zero-mean normal, homoscedastic, independent measurement noise. The confidence interval can also be applied, with less rigor, to non-normal and heteroscedastic noise.
>
> We translate our confidence interval into a simple hypothesis test on any parameter in our model. We introduce the concept of Type I error and we derive the probability of Type I error for the test proposed.

We refer to standard references for proofs of the results presented.

Prerequisite: operations on matrices and vectors; least squares minimization; probability (mean, variance, independence, normal distribution) and statistical estimation (sample mean, confidence intervals).

2 Recapitulation of Model, Measurements, and Estimators

We are given a system described by independent variable — p–vector — x and dependent variable y related through $y^{\text{truth}}(x)$. We provide a model of the form $y^{\text{model}}(x;\beta) = \sum_{j=0}^{n-1} \beta_j h_j(x)$ for prescribed $h_j(x), 0 \le j \le n-1$ (with $h_0(x) = 1$, by convention). We shall continue to assume that our model is adequate — there is no model error — in the sense that there exists a value for the parameter, β^{true}, such that $y^{\text{true}}(x) = y^{\text{model}}(x;\beta^{\text{true}})$. Finally, we are provided with measurements

$(x_i, y_i^{\text{meas}}), 1 \le i \le m$, where $y_i^{\text{meas}} = y^{\text{true}}(x_i) + \epsilon_i$, and ϵ_i is the measurement error, or "noise." We shall assume here that the noise is zero-mean normal, homoscedastic with unknown variance σ^2, and independent: for $i = 1, \ldots, m, E(\epsilon_i) = 0, E(\epsilon_i^2) = \sigma^2$, and $E(\epsilon_i \epsilon_k) = 0, k \ne i$.

Our least squares procedure shall then generate, for a particular realization of our experiment, an estimate $\hat{\beta}$ for β^{true}. We recall that $\hat{\beta}$ satisfies the normal equations,

$$(X^{\mathrm{T}} X)\hat{\beta} = X^{\mathrm{T}} y^{\text{meas}} \ , \tag{1}$$

where X is the design matrix given by $X_{ij} = h_j(x_i), 1 \le i \le m, 0 \le j \le n-1$. (Recall that the columns of X are indexed from 0.) We presume that the columns of X are independent such that (1) admits a unique solution, which we may thus write explicitly as

$$\hat{\beta} = (X^{\mathrm{T}} X)^{-1} X^{\mathrm{T}} y^{\text{meas}} \ . \tag{2}$$

Once armed with $\hat{\beta}$ we may also construct our estimate $\hat{y}(x) = y^{\text{model}}(x; \hat{\beta})$ for $y^{\text{true}}(x)$. We recall that, for $1 \le i \le m$,

$$(X\beta^{\text{true}})_i = y^{\text{true}}(x_i) \quad \text{and} \quad (X\hat{\beta})_i = \hat{y}(x_i) \ , \tag{3}$$

where $(X\beta^{\text{true}})_i$ (respectively, $(X\hat{\beta})_i$) refers to the i^{th} entry of the m-vector $X\beta^{\text{true}}$ (respectively, m-vector $X\hat{\beta}$).

In this nutshell we will understand the error in our estimator $\beta\,\hat{}$: we present in Section 3 *a priori* error bounds; we provide in Section 4 (*a posteriori*) confidence intervals. Although for the most part we retain our hypotheses — on the model and the measurement noise — in force, we shall also consider the extent to which the accuracy of our estimators, and our assessment of these estimators, degrades as our hypotheses are violated. In Section 3 we shall briefly discuss, in the context of *a priori* error analysis, the effect of model error, and also the effect of dependence (or correlation) in the measurement errors. In Section 4 we shall briefly discuss, in the context of confidence intervals, the effects of non-normality and heteroscedasticity — non-uniform variance for those with single-jointed tongues — of the measurement errors. We shall see that, in fact, some departure from ideal conditions can be readily accommodated, in particular in the case of a "well-designed" experiment.

3 A Priori Analysis

We provide here an analysis of the expected error. In the next section we translate these results into confidence intervals.

We first define the error in our parameter estimate, e: e is an n–vector with components $e_j \equiv \hat{\beta}_j - \beta_j^{\text{true}}, 0 \le j \le n-1$; we may thus write $\hat{\beta}_j = \beta_j^{\text{true}} + e_j, 0 \le j \le n-1$, or in vector form,

$$\hat{\beta} = \beta^{\text{true}} + e \ . \tag{4}$$

We also recall the error in our experimental measurements, ϵ: ϵ is an m–vector with components $\epsilon_i \equiv y_i^{\text{meas}} - y^{\text{true}}(x_i), 1 \le i \le m$; we may thus write $y_i^{\text{meas}} = y^{\text{true}}(x_i) + \epsilon_i, 1 \le i \le m$, or in vector form, from (3),

$$y^{\text{meas}} = X\beta^{\text{true}} + \epsilon \ . \tag{5}$$

We can now insert (4) and (5) into (1) to obtain

$$(X^{\mathrm{T}}X)(\beta^{\text{true}} + e) = X^{\mathrm{T}}(X\beta^{\text{true}} + \epsilon) \ , \tag{6}$$

which directly simplifies to

$$(X^{\mathrm{T}}X)e = X^{\mathrm{T}}\epsilon \ . \tag{7}$$

This "error equation" connects the errors in the measurements to the errors in the parameter estimate through the "intermediary" of the design matrix, X.

We first present a general bound for $\|e\|$, the norm of e — the sum of the square of the error in each of our n parameters – in terms of $\|\epsilon\|$, the norm of ϵ — the sum of the square of the error in each of our m measurements. In particular, it is not difficult to demonstrate that, for any realization of our experiment,

$$\|e\| \leq \frac{1}{\nu_{\min}} \|\epsilon\| \ , \tag{8}$$

where $\nu_{\min}$ is the minimum *singular value* of the matrix X. (This minimum singular value may also be expressed as the square root of the smallest eigenvalue of $(X^{\mathrm{T}}X)$.) This estimate is often quite pessimistic, as we shall see below. However, it demonstrates the connection between ϵ and e *through* X (and $\nu_{\min}$). The matrix X is determined by our model $h_j, 0 \leq j \leq n-1$, and the measurement points $x_i, 1 \leq i \leq m$. As we discussed earlier, we typically choose the (linearly independent) $h_j, 1 \leq j \leq n-1$, to approximate well the anticipated system behavior and to isolate the parameters we wish to estimate, and the measurement points $x_i, 1 \leq i \leq m$, to de-sensitize our parameter estimates to the measurement errors. We can now express the latter more quantitatively: we wish to maximize, through the design of the experiment, the singular value $\nu_{\min}$ (or other, related, stability factors). We note also that it is precisely $\nu_{\min}^{-1}$ that bounds the amplification of *model errors* in our parameter estimate: a good design can at least partially mitigate the effect of an inadequate model.

To develop a sharper and more enlightening *a priori* bound we must consider expectations rather than individual realizations. We shall first consider the illustrative and analytically simple case of a single parameter: $n = 1$. In this case we know that $X^{\mathrm{T}}X = m$, and hence (7) reduces to

$$e = \frac{1}{m} \sum_{i=1}^{m} \epsilon_i \ . \tag{9}$$

We note immediately that if the sum of the m measurement errors, $\epsilon_i, 1 \leq i \leq m$, vanishes, then the error in our parameter estimate, e, will also be zero. This provides an important clue: error cancellation plays a crucial role in the accuracy of our parameter estimates. In fact, there is little reason that the sum of the measurement errors in any realization should be (exactly) zero. But there is good reason that the sum of the measurement errors in any realization should be "small."

To demonstrate this point, we first directly square both sides of (9) — recall that, for $n = 1$, $e = \hat{\beta}_0 - \beta_0^{\text{true}}$ is a scalar — to obtain

$$e^2 = \frac{1}{m^2} \sum_{i=1}^{m} \sum_{i'=1}^{m} \epsilon_i \epsilon_{i'} \ . \tag{10}$$

We now recall that $\epsilon_i, 1 \leq i \leq m$, is a random variable of zero mean and variance σ^2. Thus $\hat{\beta}_0$ and e are also random variables: our parameter estimate and the error in our parameter estimate will both vary from realization to realization. We next take expectations (and a square root) to obtain

$$\sqrt{E(e^2)} = \frac{1}{m}\sqrt{\sum_{i=1}^{m}\sum_{i'=1}^{m} E(\epsilon_i \epsilon_{i'})}\ . \tag{11}$$

Note that $\sqrt{E(e^2)}$ is the expected error in our estimate $\hat{\beta}_0$ for β_0^{true}: the "root-mean-square" error over many realizations of our set of m measurements.

We note from (7) (and in our particular case, (9)) that, since $E(\epsilon) = 0$ — by which we mean that $E(\epsilon_i) = 0$, $1 \leq i \leq m$ — then $E(e) = 0$ — by which we mean that $E(e_j\ ) = 0$, $0 \leq j \leq n - 1$. It thus follows from the definition of e that $E(e) = E(\beta\hat{\ } - \beta^{\text{true}}) = 0$, or $E(\beta\hat{\ }\ ) = \beta^{\text{true}}$: $\beta\hat{\ }$ is an unbiased estimator for β^{true}. (Note that in the presence of model error, or model bias, $E(\hat{\beta}) \neq \beta^{\text{true}}$: our parameter estimator is now biased.) We may thus conclude that $e = \hat{\beta} - \beta^{\text{true}} = \hat{\beta} - E(\hat{\beta})$ and hence that $\sqrt{E(e^2)}$ is nothing more than the variance of our estimator. A small variance will of course signal a good estimator: an estimator for which large deviations from the mean — the parameter we wish to estimate, β^{true} — will be unlikely.

We now recall that the measurement errors are independent (in fact, uncorrelated suffices): for $1 \leq i \leq m$, $E(\epsilon_i \epsilon_k) = 0, k \neq i$; note also, from the assumption of homoscedasticity, that $E({\epsilon_i}^2) = \sigma^2$, $1 \leq i \leq n$. It follows that only m terms in the double sum of (11) are nonzero. More precisely, we obtain

$$\sqrt{E(e^2)} = \frac{1}{m}\sqrt{\mathop{}_{i=1}^{n} E(\epsilon_i^2)} = \frac{1}{m}\sqrt{m\sigma^2} = \frac{1}{\sqrt{m}}\sigma. \tag{12}$$

We observe that indeed, as m increases, the accuracy of our parameter estimate improves: cancellation of the measurement errors "helps" the parameter estimate find the middle of the data. Not unexpectedly, the convergence rate is the familiar $p = 1/2$ associated with statistical estimation.

It is simple to extend (12) to the case of general n: we obtain, for the expectation of $e_j \equiv \hat{\beta}_j - \beta_j^{\text{true}}$, the error in the j^{th} parameter,

$$\sqrt{E(e_j)^2} = \sqrt{(X^{\mathrm{T}}X)^{-1}_{jj}}\ \sigma\ ,\ \ 0 \leq j \leq n-1\ , \tag{13}$$

where $(X^{\mathrm{T}}X)^{-1}_{jj}$ refers to the j^{th} diagonal element of the inverse of the matrix $X^{\mathrm{T}}X$ associated with our normal equations. We again see the role, albeit less transparently, that the matrix X plays in the accuracy of our estimates: the jj entry of $(X^{\mathrm{T}}X)^{-1}$ relates the noise in the measurements to the error in the estimate $\hat{\beta}_j$ for the parameter β_j^{true}. It is possible to show, under certain general and plausible hypotheses, that

$$\sqrt{(X^{\mathrm{T}}X)^{-1}_{jj}} \sim m^{-1/2} \text{ as } m \to \infty\ , \tag{14}$$

just as we obtained by explicit calculation for the case of $n = 1$. In conclusion: our parameter estimate converge to true result (in expectation) as $m^{-1/2}$.

To better understand this result — and the importance of independence — we momentarily abandon our assumption that the errors $\epsilon_i, 1 \leq i \leq m$, are independent. Rather, we consider the opposite extreme: we assume *perfectly correlated* measurements such that $E(\epsilon_i \epsilon_{i'}) = \sigma^2, 1 \leq i, i' \leq m$. Now all m^2 terms in the double sum of (11) survive, and we arrive at

$$\sqrt{E(e^2)} = \sigma \; . \qquad (15)$$

In this case our error does not get smaller — out parameter estimate does not improve — as we take more measurements. This makes good sense: if all the measurements are correlated, then the m measurements are in fact equivalent to a single measurement; thus (15) should agree, and indeed does agree, with (12) for $m = 1$. A first conclusion is that independence is key to convergence. The second conclusion — clear from the derivation — is that some small correlation between measurement errors is not necessarily a disaster: our sum in (11) may still decrease with m, albeit with a larger constant or perhaps an (even) slower rate.

4 Confidence Intervals

4.1 Large Samples

We consider here confidence intervals which will be valid in the limit of many measurements, $m \to \infty$. In actual practice, infinity can be quite small in practice; in any event, we provide the necessary finite-sample corrections in the next section.

We know that, for $0 \leq j \leq n-1$, β_j is a random variable with mean β_j^{true} and variance $(X^{\mathrm{T}}X)^{-1}_{jj}\sigma$. It further follows from (2) that $\beta\,\hat{}$ is the sum of normal random variables, and hence $\beta\,\hat{}$ is also normally distributed. Thus

$$P\left(-z_\gamma \leq \frac{\hat{\beta}_j - \beta_j^{\text{true}}}{\overline{(X^{\mathrm{T}}X)^{-1}_{jj}}\sigma} \leq z_\gamma \right) = \gamma \; , \qquad (16)$$

where $\Phi(z_\gamma) = ((1+\gamma)/2)$ and Φ is the standard normal cumulative distribution function. We note that (19) is a probability statement for a particular j and *not* a statement valid jointly for all j, $0 \leq j \leq n-1$.

It remains to estimate the variance σ^2. Given that $\sigma^2 = E(\epsilon_i^2), 1 \leq i \leq m$, it is plausible to estimate σ^2 as the sample mean of the noise random variable squared, or

$$\frac{1}{m}\sum_{i=1}^{m} \epsilon_i^2 = \frac{1}{m}\sum_{i=1}^{m} (y_i^{\text{meas}} - (X\beta^{\text{true}})_i)^2 \; ; \qquad (17)$$

the second expression follows from the definition $\epsilon_i \equiv y_i^{\text{meas}} - y^{\text{true}}(x_i)$ and the identity $(X\beta^{\text{true}})_i = y^{\text{true}}(x_i)$. We exploit here, in a crucial way, (homoscedasticity and) independence of the measurement noise: we may enlist the m different measurements in a single estimate of σ^2. However, (17) is still not calculable, and thus we approximate β^{true} by $\hat{\beta}$ — recall that $\hat{\beta} \to \beta^{\text{true}}$ in our limit of large m — to arrive at an approximation $\hat{\sigma}^2$ to σ,

$$\hat{\sigma}^2 \equiv \frac{1}{m}\sum_{i=1}^{m} (y_i^{\text{meas}} - (X\hat{\beta})_i)^2 = \frac{1}{m}\|y^{\text{meas}} - X\hat{\beta}\|^2 \; , \qquad (18)$$

where the second expression follows from the definition of the norm. The quantity $\hat{\sigma}^2$ is denoted the (biased) *sample variance*. Note that our approximation $\sigma \approx \hat{\sigma}$ is only valid for large m and in particular for $m \gg n$ such that the dependence between $\hat{\beta}$ and σ is sufficiently weak. (For example, if we consider $m = n$, equivalent to interpolation, $\hat{\sigma}^2$ evaluates to zero — clearly not at all related to σ^2, the true variance of the measurement noise.) We also emphasize that σ^2 is an estimator for the variance of the measurement error, which through $(X^{\mathrm{T}}X)^{-1}_{jj}$ is then translated into an estimator for the variance in β_j. Finally, we caution that in the presence of non-zero model error, $\hat{\sigma}^2$ will not converge to σ^2 as m increases but rather will indiscriminately lump measurement noise and model error.

It remains to assemble our results. We first substitute our approximation $\sigma\hat{}$ for σ in (19) to obtain

$$P\left(-z_\gamma \le \frac{\hat{\beta}_j - \beta_j^{\mathrm{true}}}{\overline{(X^{\mathrm{T}}X)^{-1}_{jj}}\,\hat{\sigma}} \le z_\gamma\right) \sim \gamma\ ,\ m \to \infty\ . \tag{19}$$

We then pivot the resulting probability statement to arrive at two inequalities for β_j^{true} from which we can then form a confidence interval. The final result: for zero-mean normal, homoscedastic, independent noise, in the limit of sufficiently large m, for a given j, $0 \le j \le n-1$, the true parameter β_j^{true} resides in the confidence interval

$$[\mathrm{ci}_{\beta_j^{\mathrm{true}}}]^{\text{large sample}} \equiv [\hat{\beta}_j - z_\gamma\ \ \overline{(X^{\mathrm{T}}X)^{-1}_{jj}}\,\hat{\sigma}\ ,\ \hat{\beta}_j + z_\gamma\ \ \overline{(X^{\mathrm{T}}X)^{-1}_{jj}}\,\hat{\sigma}\ ] \tag{20}$$

with confidence level γ. We shall refer to (20) as the large-sample individual (two-sided) confidence interval for β_j^{true}: the adjective "individual" signals that (20) is not valid for *all* j, $0 \le j \le n-1$, simultaneously, but rather for any given *single* $j, 0 \le j \le n-1$. Note for n small we can directly form the $n \times n$ matrix $X^{\mathrm{T}}X$, next find the inverse, $(X^{\mathrm{T}}X)^{-1}$, and finally inspect the jj entry; for larger n, it is possible and preferrable to entirely avoid the formation of $X^{\mathrm{T}}X$ and $(X^{\mathrm{T}}X)^{-1}$.

In fact, our result (20) is also valid even for non-normal and heteroscedastic noise (with some limitations on the variance), thanks to certain versions of the central limit theorem which admit more generous hypotheses. The practical implications are perhaps more qualitative than quantitative: in general, we will not know for what values of m our large-sample confidence interval (20) will be accurate to (say) 10%; however, we do know that the (20), as well as the more precise result of the next section, is stable with respect to deviations from normality or homoscedasticity— some small non-normality or heteroscedasticity should have a commensurate small effect on our confidence intervals (or confidence level).

4.2 Finite Samples

We now present a more precise confidence interval valid for any $m > n$. We require only two changes.

First, we must replace $\sigma\hat{}^{\,2}$ in (20) with s^2 given by

$$s^2 = \frac{1}{m-n}\sum_{i=1}^{m} \|y^{\mathrm{meas}} - (X\hat{\beta})\|^2\ ; \tag{21}$$

s^2 is denoted the (unbiased) sample variance. We again emphasize that s^2 is an estimator for the variance of the measurement error, which through $(X^{\mathrm{T}}X)^{-1}_{jj}$ is then translated into an estimator for the variance of β_j. Clearly, for $m \gg n$, s^2 approaches $\hat{\sigma}^2$. However, for m close to n, s^2 and $\hat{\sigma}^2$ are very different: the $m - n$ in the denominator of s^2 reflects the dependence between our estimators for β^{true} and σ^2. We observe in particular that for $m = n$ we simply can not develop an estimate for s^2 — and hence we will not be able to develop a corresponding confidence interval — since for $m = n$ we interpolate the data and thus we can not possibly distinguish model from noise.

Second, we must replace z_γ in (20) with $t_{\gamma,m-n}$, where $t_{\gamma,m-n}$ is the solution of $T_{m-n}(t_{\gamma,m-n}) = ((1+\gamma)/2)$ for T_k the cumulative distribution function associated with the "Student t" density with k "degrees of freedom." More explicitly, we may write $t_{\gamma,m-n} = T^{-1}_{m-n}((1+\gamma)/2)$, where T^{-1}_k is the inverse cumulative distribution function of the Student t density with k degrees of freedom. We must consider T_k rather than Φ to reflect the additional uncertainty introduced by the variance approximation $s^2 \approx \sigma^2$; note that $t_{\gamma,m-n} > z_\gamma$ but that $t_{\gamma,m-n}$ tends to z_γ as $m \to \infty$ (for fixed n).

We can now state the result: for zero-mean normal, homoscedastic, independent noise, for any $m > n$, for a given j, $0 \le j \le n-1$, the true parameter β^{true}_j resides in the confidence interval

$$[\mathrm{ci}_{\beta^{\text{true}}_j}] \equiv [\hat{\beta}_j - t_{\gamma,m-m}\sqrt{(X^{\mathrm{T}}X)^{-1}_{jj}}\, s\ ,\ \hat{\beta}_j + t_{\gamma,m-n}\sqrt{(X^{\mathrm{T}}X)^{-1}_{jj}}\, s\ ] \tag{22}$$

with confidence level γ. We shall refer to (22) as the individual (two-sided) confidence interval for β^{true}_j. We note that our confidence interval depends of course on our data but also on the confidence level, γ, the model and the measurement points, through the design matrix X, and m and n through s. The frequentist interpretation of (22) is standard: β^{true}_j will reside in $[\mathrm{ci}_{\beta^{\text{true}}_j}]$ in a fraction γ of many replications of our experiment (note each experiment comprises m measurements).

There are many other confidence intervals possible: one-sided confidence intervals for the $\beta^{\text{true}}_j, 0 \le j \le n-1$; joint confidence intervals simultaneously on $\beta^{\text{true}}_j, 0 \le j \le n-1$; also confidence intervals on $y^{\text{true}}(x_0) = y^{\text{model}}(x_0; \beta^{\text{true}})$ for some value x_0 of the independent variable.

4.3 Hypothesis Testing

We consider here a simple introduction to Hypothesis Testing. We do not introduce much of the special language — very useful but also subtle — or mathematical technology associated with the general framework; instead, we directly build on the confidence intervals already developed to consider a relatively simple specific situation.

Let us say that we wish to decide if a particular parameter, $\beta^{\text{true}}_{j^*}$, takes on a particular value, say c^*. We may write this hypothesis as $\mathbf{H} : \beta^{\text{true}}_{j^*} = c^*$. We shall presume that it is our belief, based on some theoretical evidence, that indeed $\beta^{\text{true}}_{j^*} = c^*$. We thus wish to test, or challenge, this hypothesis with respect to data: Does the data suggest that the hypothesis might be false? If yes, we reject our hypothesis; if no, we accept our hypothesis. In the former case, data forces us to abandon a theory. (Purists prefer "not reject" rather than "accept": a theory can never be definitively verified, only definitively falsified.)

We must thus develop a test — a criterion — on the basis of which we will reject the hypothesis as inconsistent with observations. Clearly, given our *a priori* belief in the hypothesis, we might wish to be conservative: we want to reject a true hypothesis with low probability. Note the rejection of a true hypothesis is known as a Type I error. More quantitatively, our test should thus satisfy the following condition: the probability that we reject a true hypothesis — that we commit a Type I error — shall be equal to $1-\gamma$, for γ close to unity and hence $1-\gamma$ suitably close to zero. In effect, we assume the hypothesis is innocent (true) until proven guilty (false) beyond reasonable doubt $(1-\gamma)$.

Our test is very simple:

> if the confidence interval $[\mathrm{ci}_{\beta_{j^*}^{\text{true}}}]$ includes c^*, then we accept our hypothesis;
>
> if the confidence interval $[\mathrm{ci}_{\beta_{j^*}^{\text{true}}}]$ does not include c^*, then we reject our hypothesis.

We now show that we satisfy our Type I error condition. We suppose that our hypothesis is indeed true, and hence $\beta_{j^*}^{\text{true}} = c^*$. (Note this is not a probabilistic statement: the hypothesis is either true or false; our analysis considers the case in which the hypothesis is true.) Then by construction $[\mathrm{ci}_{\beta_{j^*}^{\text{true}}}]$ will contain c^* with probability, or confidence level, γ. Hence, $[\mathrm{ci}_{\beta_{j^*}^{\text{true}}}]$ will not contain c^* — and we will reject our hypothesis $\mathbf{H}$ — with probability, or confidence level, $1-\gamma$. (We recall that probability refers to random variables and confidence level to corresponding realizations.) Thus, we reject our true hypothesis with probability, or confidence level, $1-\gamma$: in a fraction $1-\gamma$ of experiments we perform (each of which constitutes m measurements), we reject the (true) hypothesis because we misinterpret an unlikely fluctuation as the statistically significant effect $\beta_{j^*}^{\text{true}} = c^*$.

Of course we could ensure no Type I errors if we simply accept our hypothesis independent of any data: we would then never reject a true hypothesis. But then also we could clearly accept a false hypothesis, and even a very false hypothesis — $\beta_{j^*}^{\text{true}}$ very different from c^*. The acceptance of a false hypothesis is known as a Type II error. Often, tests of hypotheses are constructed such that for a given tolerable probability of Type I error – we do not wish to reject a presumed true hypothesis unless the data rather unambiguously contradicts our assertion — we minimize the probability of Type II error — acceptance of false hypotheses $\beta_{j^*}^{\text{true}} = \overline{c}^*$ for some set of $\overline{c}^* = c^*$.

5 Perspectives

In this nutshell we only touch the surface of the rich topic of regression analysis, hypothesis testing, and statistical inference. For a more in-depth analysis of the basic mathematical assumptions and a complete derivation of the theoretical results we recommend AM Mood, FA Graybill, and DC Boes, "Introduction to the Theory of Statistics," McGraw-Hill, 1974. For a much more complete description of the many kinds of confidence intervals available, the interpretation of residuals and the identification of bias, and the development of more advanced regression methods and associated inferences techniques, we refer to NR Draper and H Smith, "Applied Regression Analysis," 3$^{\text{rd}}$ Edition, Wiley, 1998.

2.086 Numerical Computation for Mechanical Engineers
Fall 2014