LEAST-SQUARES FITTING OF LINEAR MODELS

.

1. The Importance Of Linear Fits

Linear model fits are of tremendous importance in finance.

They are, approximately, the most complex model explainable to quantitatively inexperienced people, though the same observation does not apply once error bars are included. They are nontrivial, well-behaved in corner cases, robust and thoroughly studied. Calibration can be fast and easy, parameter counts are small, and they can be viewed as the local limit of all differentiable models.

Though we need to avoid complacency about nonlinear effects, and be aware of false positives, they serve well as initial models in most circumstances. We begin by examining fits using ordinary least squares regressions.

2. Perspective on Regression

Regression on a data set may be viewed as one of several sorts of exercise, including

- Minimizing a particular objective function characterizing the sizes of error in a model
- Maximizing statistical likelihood of a model's correctness
- Minimizing the expected error in conditional predictions made by the model

In special cases, particularly where we are confident that future errors and past residuals are independent draws from identical normal distributions, and where our objective function is a monotonic function of the sum of squared residuals, these exercises can turn out to be equivalent. This is particularly true of polynomial models and, *a forteriori*, linear models.

$$Y \sim \alpha + \beta X$$

Though the three exercises are philosophically different, our assumptions cause them to result in the same fitted model, i.e. identical values for α and β. Even in this special case of concomitance in the fitted models, though, these three approaches give us different views of what

the strengths and weaknesses might be in our modeling approach. The most important such views are, respectively,

- We have flexibility in defining goodness of fit
- We must consider whether the data population we work with exhibits unusual properties
- The goal of fitting a model is ultimately *not* to have a well-fitted model but rather to have a model good at making predictions.

We begin our course of study with the mathematics of least-squares linear models under the highly restrictive assumptions above. Later on, we will progress to considering the consequences of questioning our assumptions.

3. Fitting

Let's begin our mathematical tour by taking the statistical perspective. Assume we have a set of N observation times indexed by t, a dependent variable Y with N observations Y_t and a set of explanatory variables $X_i, i = 1, \ldots, K$ having observations $X_{i,t}$. Our model is

$$Y = \alpha + \beta X + u$$

where α and β are fixed coefficients and the u are random i.i.d symmetrically distributed about 0.

3.0.1. *Comment.* As noted by Craig Venables, our fitting process uses real data, and the "center" α of that data is unlikely to be near zero. Our fitting process is often really just a local linear approximation. Taylor's theorem tells us that

$$f(x) = f(x_0) + f'(x_0)(x - x_0) + O((x - x_0)^2)$$

Therefore our model above, while mathematically correct, is often more usefully considered in implementation as being the model

$$Y = \alpha + (\beta - \beta_0)X + u$$

or

$$Y = \alpha + \beta(X - X_0) + u$$

which, if we were to extend to a centered 2-variable version, would look like

$$Y = \alpha + \sum_{i=1}^{n} \beta_i (x_i - x_{i0}) + \sum_{i=1}^{n} \sum_{j=1}^{n} \beta_{ij} (x_i - x_{i0})(x_j - x_{j0}) + \left[\sigma + \sum_{i=1}^{n} \gamma_i (x_i - x_{i0}) \right] Z + \delta Z^2$$

Our errors in the linear model will therefore appear in four flavors, arising from our ignorance of various terms above

- curvature in the main effects (quadratic terms in one variable – β_{ii})
- linear-linear interactions (cross product terms in two variables – β_{ij})
- variance heterogeneity (terms in $(x_i - x_{i0})Z$), and
- skewness Z^2

Definition 3.1. *The* expectation *of a variable A is defined as the probability-weighted value over our entire state space Ω*

$$\mathbb{E}(A) := \int_{\Omega} A dP$$

where we are typically considering Ω to be standard euclidean space with metric P.

Now, surely if our model coefficients were chosen such that $\mathbb{E}(u) \neq 0$, then the presumed symmetry of the distribution of u would mean a better model is available by adding a further constant to α. Therefore, we have a primary statistical condition that the first moment of u be zero, and therefore

$$\mathbb{E}(Y) = \alpha + \beta \mathbb{E}(X)$$

Estimates of these expectations are easily available by computing the *sample mean* of Y

$$\hat{\mathbb{E}}(Y) = \frac{1}{N} \sum y_t$$

and likewise for $\hat{\mathbb{E}}(X)$.

We therefore have one equation with two unknowns,

$$\alpha = \beta \hat{\mathbb{E}}(X) - \hat{\mathbb{E}}(Y)$$

and find ourselves in need of further information, which will have to be derived from the second moment.

3.1. **Some Probability Theory.**

Definitions 3.2. *The* conditional probability $P(A|B)$*, the probability of* A *conditional on* B *is*

$$\frac{P(A \cap B)}{P(B)}.$$

If $P(A|B) = P(A)$ *we say that* A *and* B *are* independent*. Random variables* W, Z *are independent if, for all* $w, z \in \mathbb{R}$*, the events* $\{W < w\}, \{Z < z\}$ *are independent. We define the cumulative distribution function of* W *as*

$$\Psi_W(w) = P(W < w)$$

and note that by the Radon-Nikodym theorem there exists a (generalized) derivative ϕ_W *of* Ψ_W *such that*

$$\Psi_W(w) = \int_{-\infty}^{w} \phi_W(x)dx.$$

We define the conditional distribution of W *on* Z *as the function from* $\mathbb{R} \times \mathbb{R} \to \mathbb{R}$

$$\Psi_{W|Z}(w, z) = lim_{\Delta z \to 0} P(W < w | z \leq Z < z + \Delta z)$$

and its density function $\phi_{W|Z}(w, z)$ *in the obvious manner. We define the* p^{th} *moment of* W *as* $\mathbb{E}(W^p)$ *and designate the variance as the second moment. We analogize to two variables, labeling as covariance the quantity* $\mathbb{E}\left((W - \mathbb{E}(W))(Z - \mathbb{E}(Z))\right)$*.*

Lemma 3.3. *Law of Iterated Expectations For any two random variables* W, Z*,*

$$\mathbb{E}\left(\mathbb{E}\left(W|\ Z\right)\right) = \mathbb{E}(W).$$

Proof.

$$\begin{aligned}
\mathbb{E}\left(\mathbb{E}\left(W|\ Z\right)\right) &= \int_{z=-\infty}^{\infty} \mathbb{E}\left(W|\ Z\right) dP \\
&= \int_{z=-\infty}^{\infty} dz \int_{w=-\infty}^{\infty} w dw \\
&= \int_{w=-\infty}^{\infty} w dw \int_{z=-\infty}^{\infty} dz \\
&= \int_{w=-\infty}^{\infty} w dw \\
&= \mathbb{E}(W)
\end{aligned}$$

□

Lemma 3.4. *If two random variables W, Z are independent, then*

$$\mathbb{E}\left(W|\ Z\right) = \mathbb{E}(W).$$

and

$$\mathbb{E}\left(WZ\right) = \mathbb{E}(W)\mathbb{E}(Z).$$

Proof. Independence implies $\phi_{W|Z}(w|z) = \phi_W(w)$, so

$$\mathbb{E}(W|Z=z) = \int w\phi_{W|Z}(a|b) = \int w\phi_W(w) = \mathbb{E}(W).$$

□

3.2. **Second Moments For Regression.** Recall now that our model is

$$Y = \alpha + \beta X + u$$

and the u are random i.i.d symmetrically distributed about 0. Let us now assume also that the u are independent of X, which in reality is almost never true. Then we can compute that

$$\begin{aligned}\mathbb{E}(YX^*) &= \mathbb{E}(\alpha X) + \mathbb{E}(\beta XX^*) + \mathbb{E}(uX^*)\\ &= \alpha\mathbb{E}(X^*) + \beta\mathbb{E}(XX^*) + \mathbb{E}(u)\mathbb{E}(X^*)\end{aligned}$$

by independence of the constants and u from X. As we saw above, $\mathbb{E}(u) = 0$ and we will want to require that

$$\mathbb{E}(Y) = \alpha + \beta\mathbb{E}(X)$$

leaving us with

$$\alpha = \mathbb{E}(Y) - \beta\mathbb{E}(X)$$

and so

$$\mathbb{E}(YX^*) = \left(\mathbb{E}(Y) - \beta\mathbb{E}(X)\right)\mathbb{E}(X^*) + \beta\mathbb{E}(XX^*)$$

which solves to

$$\beta = \left(\mathbb{E}(YX^*) - \mathbb{E}(Y)\mathbb{E}(X)\right)\left(\mathbb{E}(X)\mathbb{E}(X^*) - \mathbb{E}(XX^*)\right)^{-1}.$$

For fitting actual data, we therefore estimate β as

$$\beta = \left(\hat{\mathbb{E}}(YX^*) - \hat{\mathbb{E}}(Y)\hat{\mathbb{E}}(X)\right)\left(\hat{\mathbb{E}}(X)\hat{\mathbb{E}}(X^*) - \hat{\mathbb{E}}(XX^*)\right)^{-1}.$$

or, in one dimension,

$$\beta = \frac{\hat{\mathbb{E}}(YX) - \hat{\mathbb{E}}(Y)\hat{\mathbb{E}}(X)}{\hat{\mathbb{E}}(X)^2 - \hat{\mathbb{E}}(XX^*)}.$$

Note that, due to the convexity of the square function and the triangle inequality, this equation is well-defined except when X_j is a constant

for some component j. When there exists a j such that X_j is a constant then the matrix $\left(\mathbb{E}(X)^2 - \mathbb{E}(XX^*)\right)^{-1}$ is degenerate and not invertible[1]. This computation, the so-called *Method of Moments*, is an extremely useful way of thinking about regression and is the basis of many convenient forms of mathematical proofs of regression's properties. However, except in one or two dimensions, matrix inversion is an expensive and numerically unstable operation, so we need to be careful about using it to actually compute our coefficients.

The method of moments is far more broadly useful than as a simple regression formula. It is widely used in time series analysis and derivatives pricing to form tractable equations for fitting highly nonlinear models.

4. Squared Errors

Let us now take another perspective on fitting the model

$$Y = \alpha + \beta X + u$$

where we consider each manifestation of u to be an "error" in the model's description of reality. For a parameter pair $\tilde{\alpha}, \tilde{\beta}$ and a given data set X_i, Y_i, we define $e_i = Y_i - \tilde{\alpha} + \tilde{\beta} X_i$ and the 2-parameter *objective function*

$$f(\tilde{\alpha}, \tilde{\beta}; \boldsymbol{e}) = \sum u_i^2$$

and refer to it as the *Sum of Squared Errors* or SSE. A pair α, β at which f takes on its minimal value are the "best available" set of parameters conditional on our choice of objective function. Generically, the problem of minimizing a 2-parameter function such as this is mathematically impossible in closed form and computationally difficult in practice. However, the extreme simplicity of the SSE allows us to minimize f quickly and easily in this special case.

It is unsurprising to find that a two-moment method of moments calibration of our model is mathematically equivalent to minimizing the sum of squared errors. Each sample of u contributes its square to the overall variance between the model and our measurements in the same manner it contributes its square to the SSE in the objective function. For convenience, we will momentarily depart from the CAPM-inspired choice of coefficients α, β and switch to statistical notation. Let us say

[1]Financially, this tends to happen if our regression involves numerous independent variables and only a few data points, as when an automated risk factor computation goes haywire.

that we want to solve for a *parameter vector*

$$\boldsymbol{\beta} = \{\beta_0, \beta_1, \ldots, \beta_n\}^*$$

where typically β_0 would be the constant term, or equivalently the coefficient on a constant (always taken to be 1.0) and the remaining dimensions would cover $n-1$ nontrivial variables. An exact solution to a linear model with independent variables $x^{(1)}, x^{(2)}$, if it existed, would find that we had a $\boldsymbol{\beta}$ such that

$$\begin{pmatrix} 1 & x_1^{(1)} & x_1^{(2)} \\ 1 & x_2^{(1)} & x_2^{(2)} \\ & \vdots & \\ 1 & x_m^{(1)} & x_m^{(2)} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_{n-1} \end{pmatrix} = \begin{pmatrix} y_0 \\ y_1 \\ y_2 \\ \vdots \\ y_m \end{pmatrix}$$

Let us label the columns of the initial *design matrix* $\mathbf{X}$ in Equation **??** as

$$\mathbf{X} = [\mathbf{X_1} \quad \mathbf{X_2} \quad \ldots \quad \mathbf{X_n}]$$

Then our goal is to choose β so as to minimize

$$g(\beta) = \sqrt{(\mathbf{X}\beta - \mathbf{y})(\mathbf{X}\beta - \mathbf{y})^*}.$$

Of course if β minimizes g then it also minimizes g^2, so right away we can see that it is sufficient to minimize

$$f(\beta) = (\mathbf{X}\beta - \mathbf{y})(\mathbf{X}\beta - \mathbf{y})^*.$$

Now suppose we have a solution, β, to the least-squares minimization problem, and a perturbation $\tilde{\beta}$ of it where

$$\tilde{\beta} - \beta = \alpha\mathbf{z}$$

and we consider letting $\alpha \to 0$. We can of course write the SSE corresponding to $\tilde{\beta}$ as

$$\begin{aligned}(\mathbf{X}(\beta + \alpha\mathbf{z}) - \mathbf{y})(\mathbf{X}(\beta + \alpha\mathbf{z}) - \mathbf{y})^* \\ = (\mathbf{X}\beta - \mathbf{y})(\mathbf{X}\beta - \mathbf{y})^* + 2\alpha\mathbf{z}^*\mathbf{X}^*(\mathbf{X}\beta - \mathbf{y}) + \alpha^{\mathbf{2}}(\mathbf{X}\beta)(\mathbf{X}\beta)^*\end{aligned}$$

Note that the third term on the right is a sum of squares, so

$$\begin{aligned}(\mathbf{X}(\beta + \alpha\mathbf{z}) - \mathbf{y})(\mathbf{X}(\beta + \alpha\mathbf{z}) - \mathbf{y})^* \\ < (\mathbf{X}\beta - \mathbf{y})(\mathbf{X}\beta - \mathbf{y})^* + 2\alpha\mathbf{z}^*\mathbf{X}^*(\mathbf{X}\beta - \mathbf{y})\end{aligned}$$

If we set $\mathbf{z}$ to

$$\boldsymbol{z} = -\mathbf{X}^*(\mathbf{X}\beta - \mathbf{y})$$

then the second term on the right becomes

$$-2\alpha(\mathbf{X}\beta - \mathbf{y})^*\mathbf{X}\mathbf{X}^*(\mathbf{X}\beta - \mathbf{y})$$

which is a negative number times a sum of squares. This yields

$$(\mathbf{X}(\beta + \alpha\mathbf{z}) - \mathbf{y})(\mathbf{X}(\beta + \alpha\mathbf{z}) - \mathbf{y})^* < (\mathbf{X}\beta - \mathbf{y})(\mathbf{X}\beta - \mathbf{y})^*,$$

contradicting our assumption that β minimized the SSE. How can this be? It must imply that the second term on the right was actually zero, i.e. that

$$\mathbf{X}^*(\mathbf{X}\beta - \mathbf{y}) = \mathbf{0}$$

or equivalently that

$$\mathbf{X}^*\mathbf{X}\beta = \mathbf{X}^*\mathbf{y}.$$

We call this the *normal equations*, and it seems to imply that the model coefficients β that we seek are available by computing

$$\beta = (X^*X)^{-1}X^*y$$

where we use the term *pseudoinverse* to describe the matrix $(X^*X)^{-1}X^*$. The quantity $\rho_{\text{LS}} = ||X\beta - y||_2$ characterizes the sizes of residuals.

4.1. **Weighted Regression.** If we wish to weight our observations, then we think of ourselves as having a diagonal matrix W containing those weights. In this case we obtain

$$\mathbf{X}^*\mathbf{W}\mathbf{X}\beta = \mathbf{X}^*\mathbf{W}\mathbf{y}.$$

Uses of weights include

- Handling different error estimates for different observations
- Fitting according to practical importance of points
- Local regressions

5. The Computational Complexity of Inverting a Matrix

Matrix inverses like the one above appear frequently in the theory of regression, in finite difference schemes for option pricing, and in risk computations. However, they should essentially never be computed, being both computationally expensive and unstable (sensitive to round-off error). For a scalar σ, viewed as a 1-dimensional matrix, there is no problem in computing $\sigma^{-1} = 1/\sigma$. However, in two dimensions the operation is noticeably more complex, with the matrix

$$\begin{pmatrix} \sigma_{1,1} & \sigma_{1,2} \\ \sigma_{2,1} & \sigma_{2,2} \end{pmatrix}$$

having a symbolic inverse

$$\begin{pmatrix} \frac{\sigma_{2,2}}{\sigma_{1,1}\sigma_{2,2}-\sigma_{1,2}\sigma_{2,1}} & -\frac{\sigma_{1,2}}{\sigma_{1,1}\sigma_{2,2}-\sigma_{1,2}\sigma_{2,1}} \\ -\frac{\sigma_{2,1}}{\sigma_{1,1}\sigma_{2,2}-\sigma_{1,2}\sigma_{2,1}} & \frac{\sigma_{1,1}}{\sigma_{1,1}\sigma_{2,2}-\sigma_{1,2}\sigma_{2,1}} \end{pmatrix}$$

involving (naively) 24 floating point operations. In 3 dimensions, the situation already becomes dire, with the matrix

$$\begin{pmatrix} \sigma_{1,1} & \sigma_{1,2} & \sigma_{1,3} \\ \sigma_{2,1} & \sigma_{2,2} & \sigma_{2,3} \\ \sigma_{3,1} & \sigma_{3,2} & \sigma_{3,3} \end{pmatrix}$$

having a symbolic inverse

$$\begin{pmatrix} \frac{\sigma_{2,2}\sigma_{3,3}-\sigma_{2,3}\sigma_{3,2}}{D} & \frac{\sigma_{1,3}\sigma_{3,2}-\sigma_{1,2}\sigma_{3,3}}{D} & \frac{\sigma_{1,2}\sigma_{2,3}-\sigma_{1,3}\sigma_{2,2}}{D} \\ \frac{\sigma_{2,3}\sigma_{3,1}-\sigma_{2,1}\sigma_{3,3}}{D} & \frac{\sigma_{1,1}\sigma_{3,3}-\sigma_{1,3}\sigma_{3,1}}{D} & \frac{\sigma_{1,3}\sigma_{2,1}-\sigma_{1,1}\sigma_{2,3}}{D} \\ \frac{\sigma_{2,1}\sigma_{3,2}-\sigma_{2,2}\sigma_{3,1}}{D} & \frac{\sigma_{1,2}\sigma_{3,1}-\sigma_{1,1}\sigma_{3,2}}{D} & \frac{\sigma_{1,1}\sigma_{2,2}-\sigma_{1,2}\sigma_{2,1}}{D} \end{pmatrix}$$

where

$$D = -\sigma_{1,3}\sigma_{2,2}\sigma_{3,1} + \sigma_{1,2}\sigma_{2,3}\sigma_{3,1} + \sigma_{1,3}\sigma_{2,1}\sigma_{3,2} - \sigma_{1,1}\sigma_{2,3}\sigma_{3,2} - \sigma_{1,2}\sigma_{2,1}\sigma_{3,3} + \sigma_{1,1}\sigma_{2,2}\sigma_{3,3}$$

with 252 naive FLOPS and 60 actual FLOPS. With larger matrices, the combinatorial explosion is staggering, going roughly as $(n!)^2$. If we look at the apparent computational needs for a matrix inverse, it quickly becomes obvious that a matrix inverse is never the end product of one of our calculations but rather an intermediate step. Therefore, we can be well-served by avoiding calculation of a matrix inverse and computing its *effects* in our formula instead.

6. Basics of Computation in Linear Algebra

We consider our matrices $A \in \mathbb{R}^{m\times n}$ as comprised of components in several ways. When not otherwise noted, the matrix elements are represented with lowercase equivalent letters and two indices over the natural numbers up to m, n, i.e. A is comprised of elements

$$a_{i,j}, i = 1, \ldots, m, j = 1, \ldots, n.$$

and its *transpose* as the matrix A^*

$$a^*_{i,j} = a_{j,i}, i = 1, \ldots, n, j = 1, \ldots, m.$$

We can also consider A as a stack of row vectors $\boldsymbol{r}_i \in \mathbb{R}^n$ or a row of column vectors defined similarly. The operation of transposition simply swaps row and column vectors. In the context of regression, we typically take the row count m to be the number of data points, so that we often have $m \gg n$. Considering scalars and row or column vectors

as small matrices, we have the *dot product* to define our typical form of multiplication in linear algebra, where if $A \in \mathbb{R}^{m\times p}$ and $B \in \mathbb{R}^{p\times n}$ then their product $C = A \cdot B \in \mathbb{R}^{m\times n}$ is computed as

$$C = AB \qquad \Longrightarrow \qquad c_{ij} = \sum_{k=1}^{p} a_{ik} b_{kj}$$

One consequence of this definition is that for vectors $x, y \in \mathbb{R}^n$, we have that $x^* y$ is a scalar. It is worth considering the special case of the *vector outer product* where $p = 1$. A common requirement is to compute

$$B = A + xy^*$$

where $A \in \mathbb{R}^{m\times n}, x \in \mathbb{R}^m, y \in \mathbb{R}^n$ and the quantity xy^* is the component we label as the outer product of x with y. These types of computations pop up frequently in practical fitting algorithms, in stability calculations, and in so-called "online" algorithms for live tracking of linear models. For notational purposes, we will often require *block notation* for our matrices, where we partition A to obtain

$$\begin{array}{c} \\ m_1 \\ \vdots \\ m_q \end{array} \begin{array}{c} \begin{array}{ccc} n_1 & \dots & n_r \end{array} \\ \begin{pmatrix} A_{11} & \dots & A_{1r} \\ \vdots & & \vdots \\ A_{q1} & \dots & A_{qr} \end{pmatrix} \end{array}$$

which represents $A = (A_{\alpha\beta})$ as a q-by-r block matrix with $\alpha = 1, \dots, q$, $\beta = 1, \dots, r$, component matrices $A_{\alpha\beta} \in \mathbb{R}^{m_\alpha \times n_\beta}$ and the m_α and n_β summing to m and n respectively.

6.1. **Linear Algebraic Terminology.**

Definitions 6.1. *We say that the* range *of a matrix $A \in \mathbb{R}^{m\times n}$ is*

$$ran(A) = \{y \in \mathbb{R}^m : y = Ax \text{ for some } x \in \mathbb{R}^n\}$$

and the null space *of A is*

$$null(A) = \{x \in \mathbb{R}^n : Ax = 0\}.$$

The rank

$$rank(A) = dim\big(ran(A)\big)$$

and we say that A is rank deficient *if $rank(A) < \min(m, n)$. The* identity matrix *in $\mathbb{R}^n$ is*

$$I_n = [e_1, \dots, e_n]$$

where the e_i are the standard orthonormal basis (or canonical*) vectors. When n is obvious, we just write I.*

$$e_\ell = (\underbrace{0, \dots, 0}_{\ell-1}, 1, 0, \dots, 0)^*.$$

The inverse *of square matrix $A \in \mathbb{R}^{n\times n}$ is a matrix A^{-1} such that $AA^{-1} = I_n$ if such a quantity exists. When it does we say A is* non-singular *or of* full rank*. Otherwise we call it* singular*, and this also implies A is rank deficient. The* determinant *of A is the (recursively defined) quantity*

$$det(A) = \sum_{j=1}^{n} = (-1)^{j+1} a_{1j} det(A_{1j})$$

with $A_{1j} \in \mathbb{R}^{n-1\times n-1}$ determined by deleting both the first row and the j^{th} column from A, and $det((a)) = a$.

6.1.1. *Determinant Effects. The determinant is invariant under transposition and homogeneous with respect to the dot product. For square $A \in \mathbb{R}^{n\times n}$, matrices $det(A) = 0$ if and only if the A is rank deficient.*

6.1.2. *Orthogonal. If a square matrix Q satisfies $QQ^* = Q^*Q = I$ we say the matrix is* orthogonal*.*

6.1.3. *Basic Properties.* Unlike with scalars, matrix multiplication is associative but intransitive, i.e. generically we have $AB \neq BA$. Thus

$$(AB)(B^{-1}A^{-1}) = A(BB^{-1})A^{-1} = AIA^{-1} = I$$

showing that $(AB)^{-1} = B^{-1}A^{-1}$. If we add a matrix C to A to form $B = A + C$, then

$$\begin{aligned} B(A^{-1} - B^{-1}CA^{-1}) &= BA^{-1} - CA^{-1} \\ &= (B - (B + A))A^{-1} \\ &= AA^{-1} = I \end{aligned}$$

so that we can conclude

$$\begin{aligned} B^{-1} &= A^{-1} - B^{-1}CA^{-1} \\ &= A^{-1} - B^{-1}(B - A)A^{-1} \end{aligned}$$

The special case where $C = UV^*$ with matrices $U, V \in \mathbb{R}^{n\times k}$ yields the *Sherman-Morrison Inversion Formula*

$$(A + UV^*)^{-1} = A^{-1} - A^{-1}U(I + V^*A^{-1}U)^{-1}V^*A^{-1}.$$

6.2. **Applications of Sherman-Morrison.** This formula is extremely useful in "online" algorithms, finite difference schemes and error propagation analysis. We most often consider the case $k = 1$ in which case the term $I + V^*A^{-1}U$ is a scalar and readily invertible. Furthermore, we often are able to assume U and V have just one or two nonzero elements, making the computation of $(A + UV^*)^{-1}$ as a perturbation of A^{-1} very simple in comparison to a full recalculation of an inverse.

Using this, the Sherman-Morrison inversion formula tells us how, if some new observation $\boldsymbol{x}, y$ arrives, we can update $\boldsymbol{\beta}$. For convenience we define the self-adjoint *prediction error matrix* or *dispersion matrix*

$$\boldsymbol{P} = (\boldsymbol{X}^*\boldsymbol{X})^{-1}$$

so that

$$\boldsymbol{\beta} = \boldsymbol{P}\boldsymbol{X}^*\boldsymbol{y}$$

Define the *prediction error* as

$$h = y - \boldsymbol{x}^*\boldsymbol{\beta}$$

and the *error dispersion* as the scalar

$$f = 1 + \boldsymbol{x}^*\boldsymbol{P}\boldsymbol{x}$$

Now we can compute the new dispersion matrix as

$$\begin{aligned}\boldsymbol{P}_{\text{new}} &= \left(\boldsymbol{P}^{-1} + \boldsymbol{x}\boldsymbol{x}^*\right)^{-1} \\ &= \boldsymbol{P} - \boldsymbol{P}\boldsymbol{x}\left(1 + \boldsymbol{x}^*\boldsymbol{P}\boldsymbol{x}\right)^{-1}\boldsymbol{x}^*\boldsymbol{P} \\ &= \boldsymbol{P} - \boldsymbol{P}\boldsymbol{x}f^{-1}\boldsymbol{x}^*\boldsymbol{P}\end{aligned}$$

and our new regression coefficients

$$\begin{aligned}\boldsymbol{\beta}_{\text{new}} &= \boldsymbol{\beta} + \boldsymbol{P}\boldsymbol{x}(1 + \boldsymbol{x}^*\boldsymbol{P}\boldsymbol{x})(y - \boldsymbol{x}^*\boldsymbol{\beta}) \\ &= \boldsymbol{\beta} + \boldsymbol{P}\boldsymbol{x}f^{-1}h.\end{aligned}$$

A similar formula applies, of course, when we are *subtracting* rather than adding some observation $\boldsymbol{x}, y$, whence

$$\boldsymbol{\beta}_{\text{reduced}} = \boldsymbol{\beta} - \boldsymbol{P}\boldsymbol{x}(\mathbf{1} - \boldsymbol{x}^*\boldsymbol{P}\boldsymbol{x})^{-1}(y - \boldsymbol{x}^*\boldsymbol{\beta})$$

which allows us to perform efficient *window regression*.

6.3. **Distance and Error.** We typically consider distance between vectors v_1, v_2 in terms of a *p-norm* on their difference $x = v_1 - v_2$. This is defined as

$$||x||_p = (|x_1|^p + \cdots + |x_n|^p)^{1/p}$$

and take a limit as $p \to \infty$ to define the ∞ norm

$$||x||_\infty = \max(|x_1|^p, \ldots, |x_n|^p).$$

For a matrix A we typically consider the standard norm $||\cdot||_F$ of A considered as a vector in $\mathbb{R}^{mn}$, along with p-norms defined by

$$||A||_p = \sup_{||x||_p \neq 0} \frac{||Ax||_p}{||x||_p}$$

and we define the *roundoff norm* $|\cdot| : \mathbb{R}^{m\times n} \to \mathbb{R}^{m\times n}$ by

$$C = |A| \qquad \Longrightarrow \quad c_{ij} = |a_{ij}|$$

and interpret inequalities written $A < B$ to mean $|A| < |B|$. We often make use of the *Hölder inequality*

$$|x^* y| \leq ||x||_p ||y||_q \quad \frac{1}{p} + \frac{1}{q} = 1$$

and its most important manifestation $|x^* y| \leq ||x||_2 ||y||_2$. We also often use the property that

$$||A||_2 \leq \sqrt{||A||_1 ||A||_\infty}$$

which relates the mathematically useful 2-norm to more easily computed 1- and ∞-norms. Given any norm $||\cdot||$ we say the *condition number* $\kappa(A)$ of A is

$$\kappa(A) = ||A||\ ||A^{-1}||.$$

Matrices with large κ are called *ill-conditioned* and ones with small κ are called *well-conditioned*. Orthogonal matrices neither contract nor expand any vectors, so for them we always have $\kappa = 1$. Given an approximation $\hat{x}$ to x, we say that the *absolute error* is

$$\epsilon_{\text{abs}} = ||\hat{x} - x||$$

and the *relative error* is

$$\epsilon_{\text{rel}} = \frac{||\hat{x} - x||}{||x||}.$$

Using the notation above we see that if we have a limit constant $\mathbf{u}$ on the relative error of coefficients of $\hat{A}$ approximating A then

$$|\hat{A} - A| \leq \mathbf{u}|A|.$$

Lemma 6.2. *If we have a small perturbation ΔA on the matrix $A \in \mathbb{R}^{n\times n}$ for which the linear system $Ax = b$ is solved, so that*

$$(A + \Delta A)\tilde{x} = b + \Delta b$$

characterized by both $||\Delta A|| < \epsilon||A||$ and $||\Delta b|| < \epsilon||b||$ with a sufficiently small condition number $\kappa < 1/\epsilon$ then the we define $r = \epsilon\kappa$ and

relative error in $\tilde{x}$ is limited by

$$\epsilon_{rel} = \frac{||\tilde{x} - x||}{||x||} \leq \frac{2\epsilon}{1-r}\kappa.$$

and $A + \Delta A$ is nonsingular.

Lemma **??** is useful for judging when an update to, or error in, linear model fitting inputs will result in acceptably sized effects on the resulting model.

7. Matrix Decomposition

Consider the case where we need to compute the inverse of a matrix A as applied to a vector $\boldsymbol{w}$. That is we wish to compute

$$\boldsymbol{z} = A^{-1}\boldsymbol{w}$$

but of course we can view this instead as *solving* the equation

$$\boldsymbol{w} = A\boldsymbol{z}.$$

So long as we properly compute $\boldsymbol{z}$ we don't necessarily mind that A^{-1} was never computed as an intermediate step. We accomplish this through matrix decompositions.

Definition 7.1. *A matrix $A \in \mathbb{R}^{n\times n}$ is* positive semidefinite *if $x^*Ax \geq 0$ for all nonzero $x \in \mathbb{R}^n$. We say it is* positive definite *if the inequality is strict.*

In most of finance, our matrices are (or ought to be) positive definite or at least semidefinite. When they are not, it implies we have poorly fit our risk matrix, or conceived an unstable finite differencing scheme.

7.1. Singular Value Decomposition.

Whether or not a matrix is square or positive semidefinite, $A \in \mathbb{R}^{m\times n}$ always admits a *singular value decomposition* (SVD) of matrices transforming it to the diagonal case. That is, there exist $U \in \mathbb{R}^{m\times m}$ and $V \in \mathbb{R}^{n\times n}$ such that the matrix $D = U^tAV$ is diagonal. It is easy to show that the rank of A is equal to the number of nonzero entries on the diagonal of D. These diagonal entries $\sigma_j, j = 1, \ldots, \min(m, n)$ are also called the *singular values* of A and their squares form the *eigenvalues* of A. It is useful to think about cases where some σ_j are nonzero but very small, in which case we say A is *nearly rank deficient*. For a matrix of full rank, the singular values of A^{-1} are proportional to $1/\sigma_j$, so small singular values σ_j can be troublesome in solving linear equations. It is relatively easy to use the SVD to show that the condition number satisfies

$$\kappa(A) = \lim_{\mathbf{u}\to 0} \sup_{||\Delta A|| \leq \mathbf{u}||A||} \frac{||(A+\Delta A)^{-1} - A^{-1}||}{\epsilon} \frac{1}{||A^{-1}||}.$$

7.2. **Triangular Matrices.** Some matrices L are *lower triangular*, that is they have zeros for every entry $\ell_{ij}, j > i$ above the diagonal. For these matrices, solving $Lx = b$ is a simple matter of applying a *back-substitution* algorithm,

$$x_i = \frac{1}{\ell_{ii}} \left(b_i - \sum_{j=1}^{i-1} \ell_{ij} x_j \right)$$

which in two-dimensional form reads as

$$\begin{pmatrix} \ell_{11} & 0 \\ \ell_{12} & \ell_{22} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} b_1 \\ b_2 \end{pmatrix}$$

This clearly enjoys the solution

$$x_1 = b_1/\ell_{11}$$
$$x_1 = (b_2 - \ell_{21} x_1)/\ell_{22}.$$

A similar way of solving *upper triangular* systems is now obvious. These two algorithms provide the basic idea behind the *tridiagonal algorithm* typically used in grid schemes for solving difference equations arising from PDE solvers for financial derivatives pricing. Note that if we break L up into blocks, an analogous algorithm applies block-by-block, i.e. for

$$\begin{pmatrix} L_{11} & 0 & \dots & 0 \\ L_{21} & L_{22} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ L_{N1} & L_{N2} & \dots & L_{NN} \end{pmatrix} \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_N \end{pmatrix} = \begin{pmatrix} B_1 \\ B_2 \\ \vdots \\ B_N \end{pmatrix}$$

we can iterate solving $L_{jj} X_j = B_j$ and then substituting

$$B_i \longrightarrow B_i - L_{ij} X_j \qquad \forall i > j.$$

note that singularities in any L_{jj} will cause errors. Now, since lower- and upper-triangular systems are so easy to solve, we see that if our normal equations

$$\mathbf{X}^* \mathbf{X} \beta = \mathbf{X}^* y.$$

involved a triangular, then we might have an easy time finding β without needing to invert anything.

7.3. **Cholesky Decomposition.** A symmetric positive definite matrix A may be broken down into a "matrix square root" C, called the *Cholesky Decomposition*, where

$$CC^* = A$$

and C is lower triangular with positive diagonal entries.

Having taken the square root of A, we may then solve the triangular systems $Cy = b$ and $C^*x = y$ to solve a corresponding system of equations.

7.4. **QR Decomposition.**

Theorem 7.2. *Any matrix $A \in \mathbb{R}^{m\times n}$ can be written in unitary-equivalent upper triangular form R. That is to say, there exists an orthogonal matrix Q and upper triangular matrix R such that $A = QR$.*

Consider this applied to the normal equations, where we the theorem tells us it is possible to have decomposed $\mathbf{X}$ into $\mathbf{X} = QR$. Then we can write

$$\begin{aligned} \mathbf{X}^*\mathbf{X}\beta &= \mathbf{X}^*y \\ R^*Q^*QR\beta &= R^*Q^*y \\ R^*R\beta &= R^*Q^*y \\ R\beta &= Q^*y \end{aligned}$$

which can easily be solved for β using the triangular algorithm if Q is known.

7.4.1. *Interesting Unitary Matrices.* A 2-by2 orthogonal matrix Q is a *rotation matrix* if it has the form

$$Q = \begin{pmatrix} \cos(\theta) & \sin(\theta) \\ -\sin(\theta) & \cos(\theta) \end{pmatrix}$$

If $y = Q^*x$ then y is obtained by rotating x counterclockwise through an angle θ. Changing cosine signs and sign on a sine,

$$Q = \begin{pmatrix} \cos(\theta) & \sin(\theta) \\ \sin(\theta) & -\cos(\theta) \end{pmatrix}$$

is a *reflection* and y is the reflection of x across the line defined by

$$Q = \begin{pmatrix} \cos(\theta/2) \\ \sin(\theta/2) \end{pmatrix}$$

7.4.2. *Householder Matrices.* Take a nonzero $v \in \mathbb{R}^n$. Let's note first of all that the vector $w = v/||v||$ is a unit length vector in the same direction as v. If we are using the 2-norm then we also have that $||v||^2 = v \cdot v$. We can form the outer product of a vector v with itself to get a full matrix, which has some useful properties. Consider the *projection*, computed as $(x \cdot v)v$, of a vector x in the direction v. This is a new vector, in the direction of v, of length equal to $x\cos(\theta)$ where θ is the angle between v and x.

If we subtract this projection from our original x, then we have the *image* of x in the hyperplane $v^\perp$ defined by all vectors perpindicular to v, known as the *null space* of v. Subtracting twice this projection will form the *reflection* of x through $v^\perp$. This reflection has a particularly convenient mathematical form, since

$$\begin{aligned} x - 2\frac{x \cdot v}{v \cdot v}v &= x - 2\frac{v \cdot x}{v \cdot v}v \\ &= x - 2\frac{2}{v^* \cdot v}(v^* x)v \\ &= x - 2\frac{2}{v^* \cdot v}v(v^* x) \\ &= x - 2\frac{2}{v^* \cdot v}(vv^*)x \\ &= \underbrace{\left(I - 2\frac{2}{v^* \cdot v}(vv^*)\right)}_{P_v} \cdot x \end{aligned}$$

The matrix P_v is called the *Householder reflection* in v. It is symmetric, unitary, and can be completely characterized by one vector v and a scalar $\beta = 2/||v||^2$. If we want to compute $P_v x$ for some vector x, we simply compute

$$x - \beta(x^* v) \times v.$$

For a matrix $A \in \mathbb{R}^{m \times n}$ we can precompute

$$z^* = \beta v^* A$$

and then calculate the product of A with P_v as

$$PA = (I - \beta vv^*)A = A - vz^*$$

thereby avoiding matrix-matrix multiplication. Now, if want to QR decompose a matrix A, we can start by trying to find a simple matrix P such that the product of P with A is triangular, at least as far as the first column. After that, we can repeat the operation to continue triangulating. The first column must have 1 in its first entry and zeros thereafter, so we want a matrix P that multiplies the first column $x = (a_{i1})$ to obtain $Px = \alpha \mathbf{e}_1$ for some α, which will end up in the top row. Let's assume we can somehow accomplish this with a Householder

reflection and solve for the corresponding vector v. Then

$$(I - \beta vv^*)x = \alpha \mathbf{e}_1$$
$$\beta(v^*x)v = x - \alpha \mathbf{e}_1$$
$$v = \frac{1}{\beta v^* x}(x - \mathbf{e}_1)$$

so therefore defining

$$v = x - \alpha \mathbf{e}_1$$

accomplishes our goal. It is relatively easy to compute that

$$\alpha = \pm ||x||.$$

This simple way to find v makes these Householder reflections very useful. They are also extremely accurate, having errors that do not propagate to degrade machine precision. In order to factor $A \in \mathbb{R}^{m \times n}$ we simply need to repeat this process n times on successive nontriangular blocks of A, at lower and lower cost as the blocks get smaller.

$$A_{j-1} = \begin{pmatrix} a_{11} & a_{12} & a_{13} & \dots & \dots & \dots & a_{1n} \\ & a_{22} & a_{23} & \dots & \dots & \dots & a_{2n} \\ & & a_{33} & \dots & \dots & \dots & a_{3n} \\ & & & \ddots & \dots & \dots & \vdots \\ & & & & a_{jj} & \dots & a_{jn} \\ & & & & a_{j+1,j} & \dots & a_{j+1,n} \\ & & & & \vdots & \vdots & \vdots \\ & & & & a_{mj} & \dots & a_{mn} \end{pmatrix}$$

Now to solve a least-squares problem

$$A^*Ax = A^*b$$

we partition A into

$$A = \begin{bmatrix} Q_1 & Q_2 \end{bmatrix} \begin{bmatrix} R_1 \\ 0 \end{bmatrix}$$

and we can equivalently solve

$$R_1 x = Q_1^* b.$$

We apply the successive Householder reflections P_i to b to compute the right hand side, and then use the triangular algorithm to complete our process of finding x.

7.4.3. *Givens Rotations.* Working with multidimensional versions of the rotations we previously observed, we can see that it is possible to

zero out the lower or upper triangle of A with a succession of operations aimed at eliminating just one component at a time.

$$\begin{array}{c} \\ i \\ \\ k \\ \end{array} \begin{pmatrix} 1 & \cdots & 0 & \cdots & 0 & \cdots & 0 \\ \vdots & \ddots & \vdots & & \vdots & & \vdots \\ 0 & \cdots & \cos(\theta) & \cdots & \sin(\theta) & \cdots & 0 \\ \vdots & & \vdots & & \vdots & & \vdots \\ 0 & \cdots & -\sin(\theta) & \cdots & \cos(\theta) & \cdots & 0 \\ 0 & \cdots & 0 & \cdots & 0 & \cdots & 1 \end{pmatrix}$$

Let us say that we want to zero out the element a_{ik} in A. We'll do so by rotating it to zero using the $i-1$ row above it.

We take

$$\theta = \cos^{-1}\left(\frac{a_{ik}}{\sqrt{a_{ik}^2 + a_{(i-1),k}^2}}\right)$$

so that the new value is zero.

The pattern for Givens QR decomposition is then that A is transformed as follows

$$\begin{bmatrix} \bullet & \bullet & \bullet \\ \bullet & \bullet & \bullet \\ \bullet & \bullet & \bullet \\ \bullet & \bullet & \bullet \end{bmatrix} \rightarrow \begin{bmatrix} \bullet & \bullet & \bullet \\ \bullet & \bullet & \bullet \\ \bullet & \bullet & \bullet \\ 0 & \bullet & \bullet \end{bmatrix} \rightarrow \begin{bmatrix} \bullet & \bullet & \bullet \\ \bullet & \bullet & \bullet \\ 0 & \bullet & \bullet \\ 0 & \bullet & \bullet \end{bmatrix}$$

$$\rightarrow \begin{bmatrix} \bullet & \bullet & \bullet \\ 0 & \bullet & \bullet \\ 0 & \bullet & \bullet \\ 0 & \bullet & \bullet \end{bmatrix} \rightarrow \begin{bmatrix} \bullet & \bullet & \bullet \\ 0 & \bullet & \bullet \\ 0 & \bullet & \bullet \\ 0 & 0 & \bullet \end{bmatrix} \rightarrow \begin{bmatrix} \bullet & \bullet & \bullet \\ 0 & \bullet & \bullet \\ 0 & 0 & \bullet \\ 0 & 0 & \bullet \end{bmatrix}$$

$$\rightarrow \begin{bmatrix} \bullet & \bullet & \bullet \\ 0 & \bullet & \bullet \\ 0 & 0 & \bullet \\ 0 & 0 & 0 \end{bmatrix}$$

7.4.4. *Which Factorization?* A least squares solution is sensitive to the quantity $\kappa + \rho_L D\kappa$. Error from using the method of normal equations with a Cholesky decomposition goes as κ^2. Cholesky decomposition will fail for small condition numbers $\kappa \gtrsim 1/\sqrt{u}$. For Householder QR decomposition, the failure boundary is $\kappa \gtrsim 1/u$. Gram-Schmidt may be constructed to achieve this same result with *pivoting* but at slightly higher computational cost. These solutions have approximate relative

error of

$$\mathbf{u}(\kappa + \rho_{\text{LS}})\kappa^2$$

The computational cost of all these techniques goes as $O(n^2 m)$.

8. Standard Error

The *standard error* of a regression assumes the residuals are uncorrelated. In this instance we write the square of standard error as a function of the squared residual

$$s^2 = \frac{\mathbf{e} \cdot \mathbf{e}^*}{m - n}$$

This can be particularly useful when looking at confidence envelopes of local regressions.

9. Intrinsically Linear Regression

We say that a function relating y to x is *intrinsically linear* if by means of transformations

$$x \xrightarrow{\psi} \check{x}$$

$$y \xrightarrow{\phi} \check{y}$$

the function can be expressed as $\check{y}_t = a + b\check{x}_t$ where $\check{x}_t$ is the transformed independent variable. In the multivariate case, we consider that we are solving the problem of minimizing

$$||\sum_{j-1}^{n} \beta\Psi(X_j) - \Phi(y)||^2.$$

From a computational point of view, this is easily accomplished by first transforming all dependent and independent variables, and then solving the system as usual. We sometimes also use the phrase *generalized linear regression*.

An example might be bond price versus yield.

10. Measuring Regression Model Performance

10.1. **Coefficient of Determination.** When the sum of squares is a good representation of the regression's accuracy, we can compute two interesting quantities. The first, the sum of squared residuals

$$RSS = \sum u_i^2$$

represents how close our model came to being truly ideal. The second quantity is the sum of squares inherent to the data, or *total sum of*

squares

$$TSS = \sum \big(y_i - \mathbb{E}(y)\big)^2$$

Linear regression can be thought of as a projection of the design matrix onto the subspace of purely linear data, i.e. residuals are orthogonal to the regression, so these two quantities add like the variances of uncorrelated variables. We can consider their relative size, and compute the correlation coefficient between model predictions and outcomes to be

$$\rho = \sqrt{\frac{TSS - RSS}{TSS}}$$

We can define the universal *coefficient of determination* or R^2, as

$$R^2 = 1 - \frac{RSS}{TSS}$$

It is quite easy to see that this quantity has the "right" asymptotics for representing regression quality.

Taking a more specific approach, we can see that for linear models orthogonality lets us write the *explained sum of squares*

$$\begin{aligned} ESS :=& \sum \big(\hat{y}_i - \mathbb{E}(y)\big)^2 \\ =& \sum \big(y_i - \mathbb{E}(y)\big)^2 - (y_i - \hat{y}_i)^2 \\ =& \sum \big(y_i - \mathbb{E}(y)\big)^2 - u_i^2 \\ =& TSS - RSS \end{aligned}$$

showing that in the linear case

$$R^2 = \frac{ESS}{TSS}$$

One of the most important metrics of model performance is its R^2 on *out-of-sample* data.

There is no technical limitation that R^2 may be computed only from linear models. Its construction makes it particularly relatable to ordinary least squares, but that need not limit us. Any set of residuals is fair game.

10.2. **Terminology of Data Points.**

- Recall that we use the term *residual* to describe the difference between the predicted value (based on the regression equation) and the actual, observed value.
- An *outlier* is an observation with large residual. In other words, it is an observation whose dependent-variable value is unusual

given its value on the predictor variables. An outlier may indicate a sample peculiarity or may indicate a data entry error or other problem.

- A data point near the "edge" of a distribution is said to have high *leverage*, and, being far from the mean will tend to be important.
- Outliers tend to have high *influence*, which means their inclusion in the data set greatly affects the estimate of regression coefficients.

One important characterization of outliers is *Cook's distance*, which combines the information of leverage and residual of the observation. Formally, leverage is based on the *fitted values* of a regression, where we begin by observing that our coefficients

$$\beta = (X^*X)^{-1}X^*\boldsymbol{y}$$

come from the pseudoinverse $(X^*X)^{-1}X^*$. We can see that (generically) our regression, if asked to predict values for the same set of independent y as it was fitted on, would not of course return its precise inputs – that's impossible – but rather would apply this coefficient set to y, obtaining

$$\begin{aligned}\hat{\boldsymbol{y}} &= X\beta \\ &= X(X^*X)^{-1}X^*\boldsymbol{y} \\ &= H\boldsymbol{y}\end{aligned}$$

where

$$H := X(X^*X)^{-1}X^*$$

is called the *hat matrix* because it operates on $\boldsymbol{y}$ to "put a hat on it".

The diagonal elements of the hat matrix are

$$h_{ii} = \boldsymbol{x}_i^*(X^*X)^{-1}\boldsymbol{x}_i$$

Cook's distance for an observation at i is the quotient of the squared Euclidean distance that the vector $\boldsymbol{y}$ moves when the ith point is omitted, by the parameter count and sum of squares. We can therefore think of it as the variation introduced by observation i, normalized to how much such variation we would expect given the overall data variation and parameter count.

We can compute the reduced model $\bar{X}\bar{\beta}$ where the ith row has been deleted, so one equation for Cook's distance is

$$c_i = \frac{1}{s^2 p}\|\boldsymbol{y} - \hat{\boldsymbol{y}}\|^2$$

where p is the parameter count (i.e. the rank of X). Now our residuals are obviously

$$\boldsymbol{u} = \boldsymbol{y} - \hat{\boldsymbol{y}}$$

so

$$\boldsymbol{u} = (I - H)\boldsymbol{y}$$

and the diagonal elements of H are all in $[0, 1]$ and satisfy

$$\sum h_{ii} = p.$$

If the residuals are indeed independent, then in particular they are are uncorrelated, so the residuals have the diagonal covariance matrix

$$\boldsymbol{S} = (I - H)\hat{\sigma}^2.$$

for some $\hat{\sigma}$. Define

$$\bar{\delta} = \bar{y} - \bar{X}\bar{\beta}$$

to describe the residuals from the reduced model and the leave-one-out residual

$$\bar{u}_i = y_i - \hat{\boldsymbol{\beta}}^* \boldsymbol{x}_i$$

If we choose

$$\bar{H} = \bar{X}(X^*X)^{-1}\bar{X}^*$$

then

$$\bar{u} = (I - \bar{H})\bar{\delta}$$

and

$$\bar{u}_i = \frac{y_i - \hat{\boldsymbol{\beta}}^* \boldsymbol{x}_i}{1 - h_{ii}}$$

The fitted value will change by an amount

$$\begin{aligned} \bar{y} - \bar{X} &= \frac{h_{ii}}{1 - h_{ii}} u_i \\ &= \frac{h_{ii}}{1 - h_{ii}} u_i \end{aligned}$$

showing the relative increase in the ith residual to be $\frac{h_{ii}}{1-h_{ii}}$. This leads us to use the term *leverage* to describe h_{ii}. By diagonality of I, J and $I - H$, we see that removing the data observation at index i, and forming the RSS

$$\bar{\delta}^*(I - \bar{H})\bar{\delta}$$

gives us the original RSS, but reduced by the missing row

$$\frac{u_i^2}{1-h_{ii}}$$

In this independent case we have that

$$\operatorname{var}(\hat{u}_i) = 1 - h_{ii}\hat{\sigma}^2$$

so the reduction is proportional to the residual size. A large value for h, near 1, will noticeably lever the fitted hyperplane in its direction.

The average leverage should be p/n where p is the parameter count and n the observation count. With parameter count p and s^2 as a reliable unbiased estimate of the error variance $\hat{\sigma}^2$ Cook's distance of residual u_i is the quantity

$$c_i = u_i^2 \frac{h_{ii}}{ps^2(1-h_{ii})^2}$$

In practice we usually take s^2 to be the mean residual square sum. A generalized extension of Cook's distance, due to Hampel (1974), is the *influence curve* for any functional T at any distribution F,

$$\operatorname{IC}(x; T, F) = \lim_{\epsilon \downarrow 0} \frac{T\left((1-\epsilon)F + \epsilon\Delta_x\right) - T(F)}{\epsilon}$$

where Δ_x is a distribution with unit mass at x, i.e. the probability distribution whose density is the Dirac (generalized) function δ_x. This curve defines, over the domain of x, what the incremental contribution of x is to an estimator F. We will see these in greater detail and more concrete form soon when we consider robust estimation schemes.

10.3. **Normal Quantile-Quantile Analysis.** Since in the real world, our data are rarely normally distributed, we would like to have some ways of determining when, and in what ways, our data (empirically) deviates from normality. The first metrics to use, of course, are the higher moments of the distribution, applied to the individual variables

It is obvious that for any symmetric distribution the odd moments

$$M_d = \mathbb{E}\left[(x - \bar{x})^n\right], \quad d \text{ odd}$$

disappear when $d \geq 3$. Other distributions may exhibit a nonzero *skew* and other evidence of asymmetry. Many empirical distributions in finance also have even moments that fail to match those expected from normal distributions. The standardized normal has fourth moment, *kurtosis*, equal to 3 so a higher fourth moment implies "fat tails". It is under-appreciated that high kurtosis also is associated with a "peaky" distributional center.

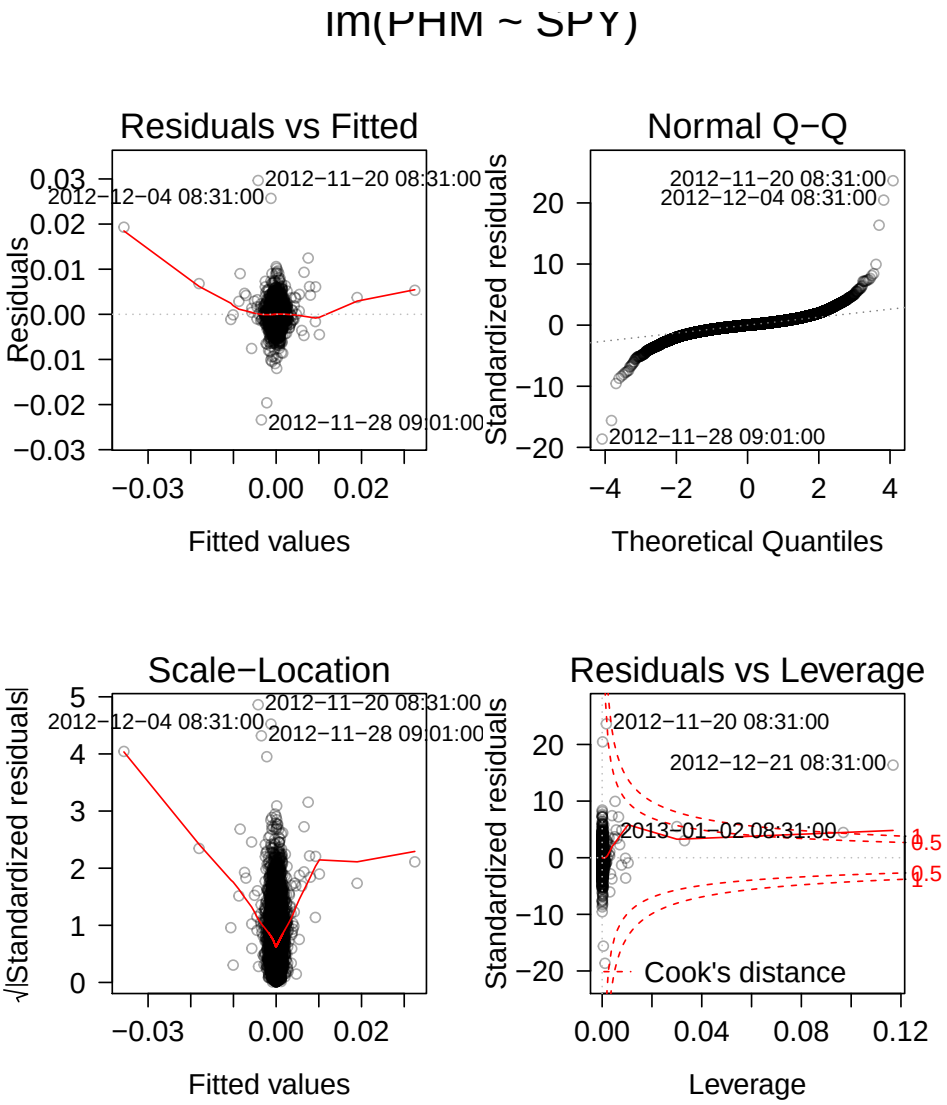

FIGURE 1. Quality measures of a linear fit. Strong evidence of fat tails is manifested, especially in the top metric plots.

Given a proportion p, we say that the pth *quantile* of a sample set $\boldsymbol{y}$ is

$$Q_p(\boldsymbol{y}) = \frac{1}{2}\left(\max\{y_i : y_i \leq p\} + \min\{y_i : y_i \geq p\}\right)$$

For a graphical picture of the difference between an empirical distribution and the normal, we can contemplate a *normal quantile-quantile*

plot. This plot is formed by obtaining the mean $\bar{x}$ and standard deviation s_x, then ordering our values $x_i, i = 1, \ldots, n$, and computing the CDF values

$$\phi_i = \Phi(i/n; \bar{x}, s_x)$$

A plot of (ϕ_i, x_i) points will be (approximately) linear with a slope of 1.0 and intercept at zero if $\{x_i\}$ came from a normal distribution. If the intercept is nonzero then odd moments are present. Skew is quite obvious as a varying slope. Fat tails are evident if the left side dives below the line and the right side rises above it. The deviation of a QQ plot from linearity, as measured by R^2, can serve along with moments as a rough-and-ready measure of distributional adherence to the normal.

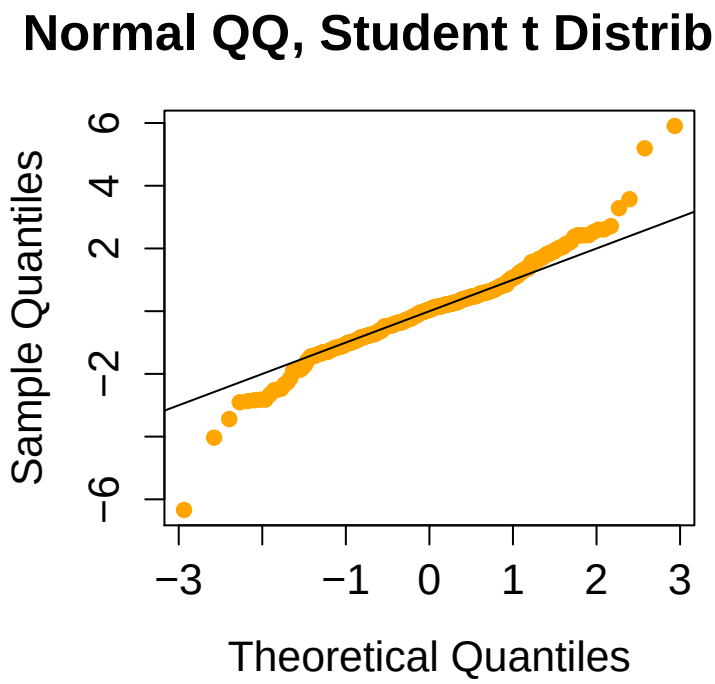

FIGURE 2. Fat tails from a Student t.

Note that visual examination of the residuals, along with leverage and distance metrics, can take us a long way to identifying when these issues have become significant. The `plot` function for `lm` in **R** is geared to this type of investigation.

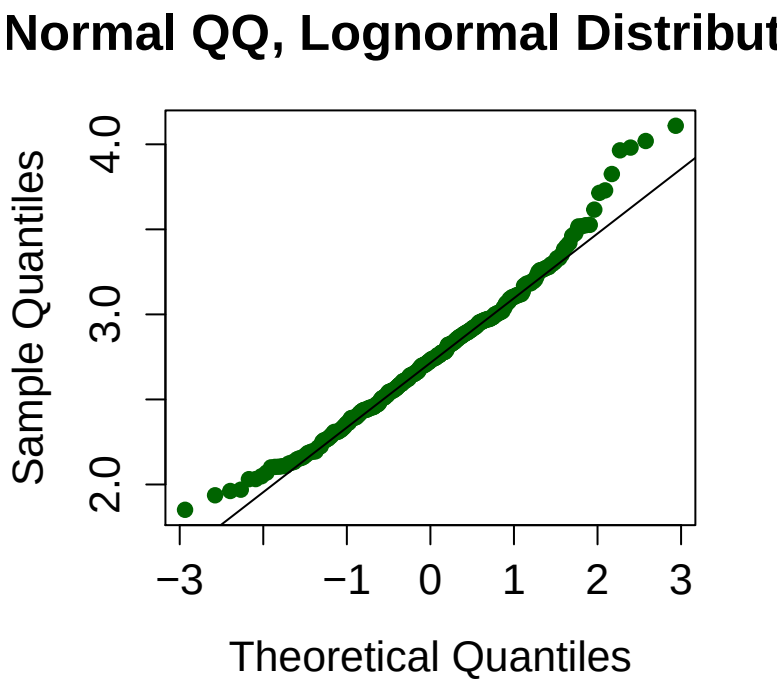

FIGURE 3. Fat tail from a Lognormal, skinny tail left

11. EXPONENTIALLY-WEIGHTED MOVING REGRESSIONS

Let's revisit the weighted regression, with a special weighting scheme.

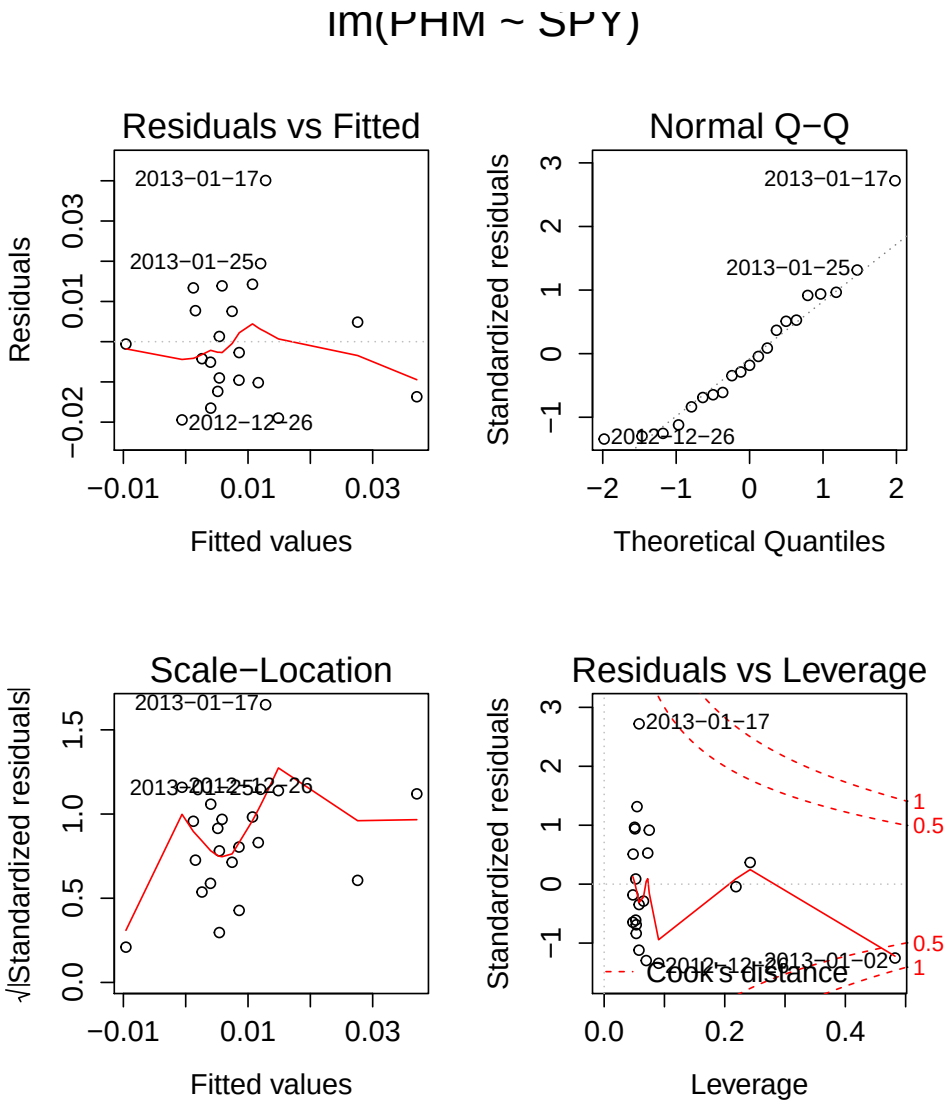

FIGURE 4. Linear fit quality on daily data.

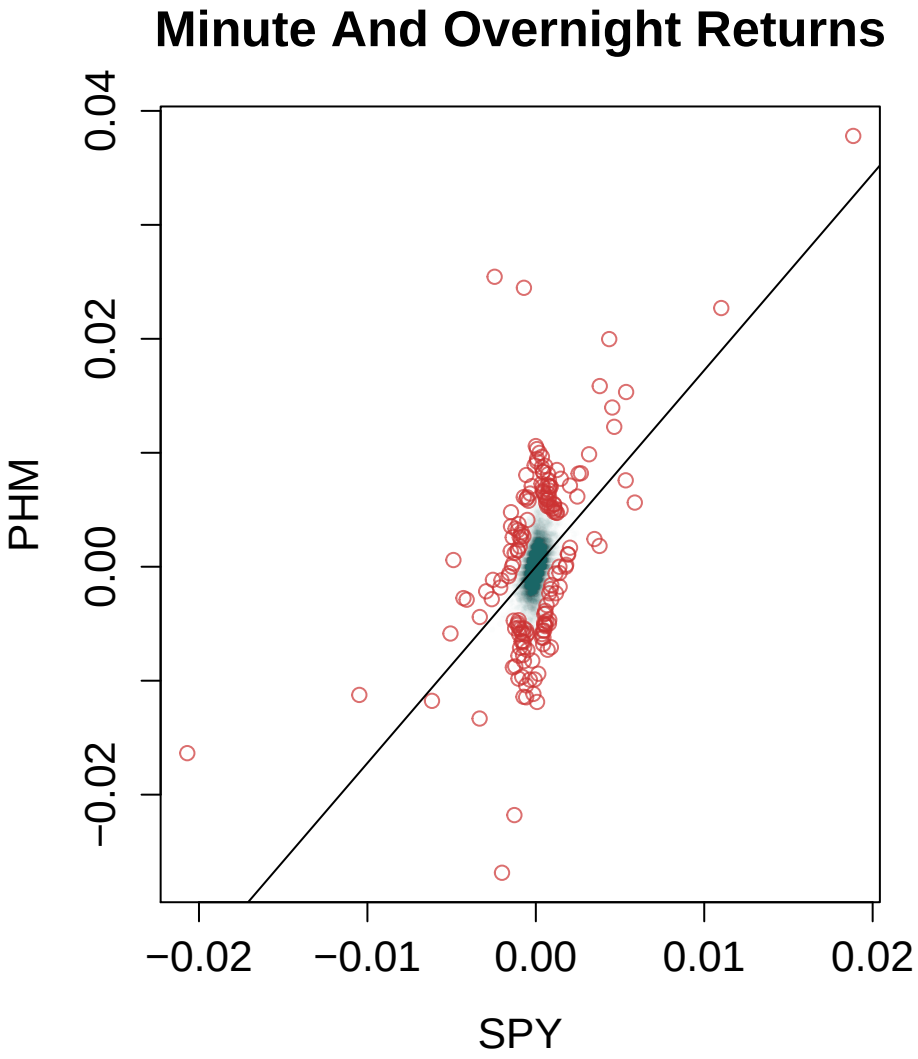

FIGURE 5. Equity returns with a linear fit. Large values of Cook's distance are circled in red.

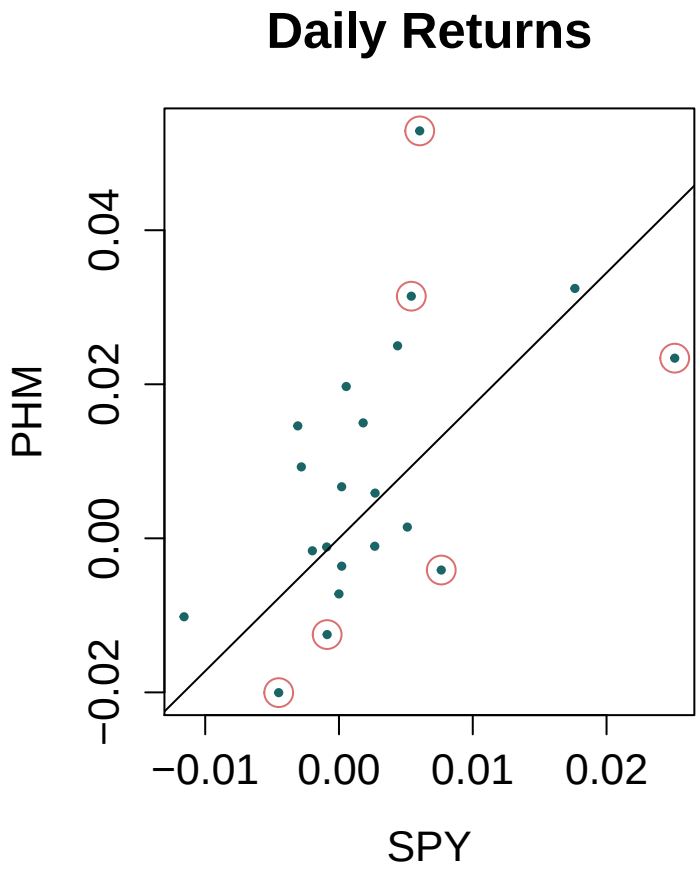

FIGURE 6. Equity returns with a linear fit. Large values of Cook's distance are circled in red.

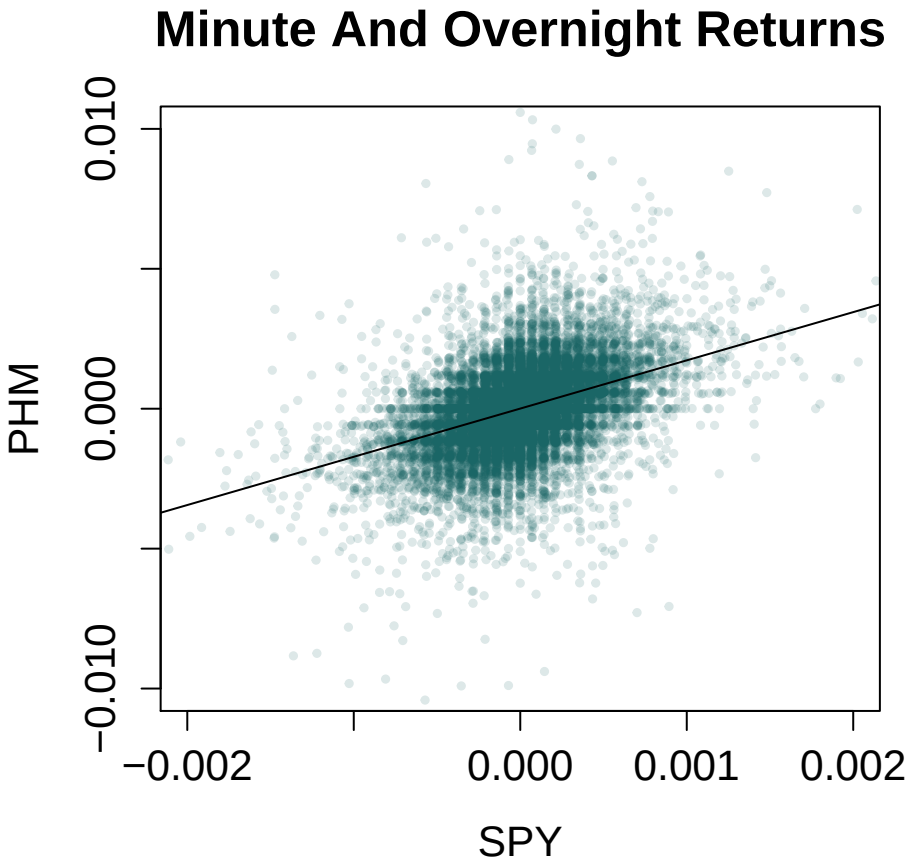

FIGURE 7. Equity returns with a linear fit, concentrating on high density region.

Recall that in ordinary least squares regression, we solve (in principle but not in practice) the normal equations to obtain a parameter estimate

$$\boldsymbol{\beta} = (\boldsymbol{X}^* \boldsymbol{X})^{-1} \boldsymbol{X}^* \boldsymbol{y}$$

The Sherman-Morrison inversion formula

$$(\boldsymbol{A} + \boldsymbol{U}\boldsymbol{V}^*)^{-1} = \boldsymbol{A}^{-1} - \boldsymbol{A}^{-1}\boldsymbol{U}(\boldsymbol{I} + \boldsymbol{V}^*\boldsymbol{A}^{-1}\boldsymbol{U})^{-1}\boldsymbol{V}^*\boldsymbol{A}^{-1}.$$

tells us how, if some new observation $\boldsymbol{x}, y$ arrives, we can update $\boldsymbol{\beta}$. For convenience we define the self-adjoint *prediction error matrix* or *dispersion matrix*

$$\boldsymbol{P} = (\boldsymbol{X}^* \boldsymbol{X})^{-1}$$

so that

$$\boldsymbol{\beta} = \boldsymbol{P}\boldsymbol{X}^* \boldsymbol{y}$$

Define the *prediction error* as

$$h = y - \boldsymbol{x}^* \boldsymbol{\beta}$$

and the *error dispersion* as

$$f = 1 + \boldsymbol{x}^* \boldsymbol{P} \boldsymbol{x}$$

Now we can compute the new dispersion matrix as

$$\begin{aligned} \boldsymbol{P}_{\text{new}} &= \left(\boldsymbol{P}^{-1} + \boldsymbol{x}\boldsymbol{x}^*\right)^{-1} \\ &= \boldsymbol{P} - \boldsymbol{P}\boldsymbol{x}\left(1 + \boldsymbol{x}^*\boldsymbol{P}\boldsymbol{x}\right)^{-1} \boldsymbol{x}^*\boldsymbol{P} \\ &= \boldsymbol{P} - \boldsymbol{P}\boldsymbol{x} f^{-1} \boldsymbol{x}^* \boldsymbol{P} \end{aligned}$$

and our new regression coefficients

$$\begin{aligned} \boldsymbol{\beta}_{\text{new}} &= \boldsymbol{\beta} + \boldsymbol{P}\boldsymbol{x}(1 + \boldsymbol{x}^*\boldsymbol{P}\boldsymbol{x})^{-1}(y - \boldsymbol{x}^*\boldsymbol{\beta}) \\ &= \boldsymbol{\beta} + \boldsymbol{P}\boldsymbol{x} f^{-1} h. \end{aligned}$$

A similar formula applies, of course, when we are *subtracting* rather than adding some observation $\boldsymbol{x}, y$, whence

$$\boldsymbol{\beta}_{\text{reduced}} = \boldsymbol{\beta} - \boldsymbol{P}\boldsymbol{x}(\boldsymbol{x}^*\boldsymbol{P}\boldsymbol{x} - \mathbf{1})^{-1}(y - \boldsymbol{x}^*\boldsymbol{\beta})$$

which allows us to perform efficient *window regression*.

Now let us say that we wish to discount our *old* data relative to new incoming information by some factor $\lambda \in (0, 1]$. We can therefore say that the dispersion matrix of the old data should be multiplied by λ,

allowing our update $\boldsymbol{x}, y$ to have full effect. That is to say we obtain

$$\begin{aligned}\boldsymbol{P}_{\text{new}} &= (\lambda \boldsymbol{P} + \boldsymbol{x}\boldsymbol{x}^*)^{-1} \\ &= \frac{1}{\lambda}\left(\boldsymbol{P} - \boldsymbol{P}\boldsymbol{x}\left(\boldsymbol{\lambda} + \boldsymbol{x}^*\boldsymbol{P}\boldsymbol{x}\right)^{-1}\boldsymbol{x}^*\boldsymbol{P}\right)\end{aligned}$$

Our new coefficients are then

$$\begin{aligned}\boldsymbol{\beta}_{\text{new}} &= \boldsymbol{\beta} + \boldsymbol{P}\boldsymbol{x}(\boldsymbol{\lambda} + \boldsymbol{x}^*\boldsymbol{P}\boldsymbol{x})^{-1}(y - \boldsymbol{x}^*\boldsymbol{\beta}) \\ &= \boldsymbol{\beta} + \boldsymbol{P}\boldsymbol{x} {f_\lambda}^{-1} h\end{aligned}$$

This approach is called *discounted least-squares regression* and is commonly seen in control theory[2].

[2] Discounted least-squares regression is equivalent to a subset of *Kalman filter* approaches to estimating state, with a trivial *transition equation* $\beta_{t+1} = \beta_t$ having no torsion and no state disturbance.

12. Epps Effect

Regressions, with their tight link to correlation, are obviously sensitive to the Epps effect.