Laura Balzano

Estimation and Regression

A report[i] for Statistical Theory class by Laura Balzano, written January/February 2005

In science, engineering and statistics, we are often interested in predicting the future based on the past. When you have data and you would like to best fit a curve to that data, you do a regression. When you have collected samples and you would like to estimate parameterization of those samples, you do estimation. They boil down to the same thing: creating a model for the natural world from data samples such that you can interpolate and extrapolate more information.

Astronomy

During the 17th and 18th centuries, many prominent mathematicians were interested in understanding the orbit of the planets. Isaac Newton developed a now well-known equation for the gravitational pull between two masses, $G = g\frac{m_1 \cdot m_2}{r^2}$. The orbit of the planets can be fairly well described by using this equation to calculate the gravitational pull between the planet and the sun, which is the body in our solar system with the most mass. However, this equation did not give an exact orbit; the mass of some of the planets, particularly the largest two, Jupiter and Saturn, had an additional affect on how the orbit traced itself out in the sky.

In 1749 Euler, after setting out to solve the problem of Jupiter's influence on Saturn, came up with an equation solution. There were many variables and some of them could be calculated from observations.

$$\varphi = 23525'' \cdot \sin q + 168'' \cdot \sin 2q - 32'' \cdot \sin 2\omega - 257 \sin(\omega - q)$$
$$-243 \sin(2\omega - p) + m - x \sin q + y \sin 2q - z \sin(\omega - p)$$
$$-u(\alpha + 360\upsilon + p)\cos(\omega - p) + Nn - 114.5k \cos q + \frac{1}{600} k \cos 2q$$

$\varphi, \eta, q, \omega, p, N, \upsilon$ were given, and n, u could be easily calculated. There was a set of 75 observations that he used to measure these quantities. Still, that left him with $x, y, m, z, \alpha, k \ldots$ six more variables that remained unknown, and he wanted to estimate them. So with 75 equations and 6 unknowns, mathematicians were now faced with the problem of how best to solve for the 6 unknown variables.

Linear Regression and Combinations of Equations

It turned out that Euler's equation was linear in the unknown variables. We can therefore formulate the question in terms of a linear regression—how do we estimate those 6 unknowns in order to best fit the data?

Rewrite the equation as

Laura Balzano

$$y_i = \beta_1 X_{i1}, \beta_2 X_{i2} \ldots, \beta_d X_{id}$$

or, $y_i = \vec{\beta} X_i$ with

$$\vec{\beta} = \begin{matrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ \vdots \\ \beta_d \end{matrix}$$

Around this time, Laplace had a set of 24 equations with 4 unknown variables that he was trying to solve. His method for solving the equations was to choose 24 of the equations and combine them so as to get only 4 equations. His combinations were made up of additions and subtractions of various of the 24 equations in such a way that Laplace thought would find the best estimate for the unknown variables. It seems, though, that we could come up with many ways of combining equations to solve for the unknowns in a problem. What is the best way, and how do we know we are using the best way?

Least Squares

In 1805, Legendre decided to take a different approach. Instead of trying to find the best way to combine equations, he wanted to start from some criterion which he could optimize. He created an objective equation which finds the difference between what is observed and what is predicted. He decided to minimize this difference. Legendre designed this objective equation to be scalar-- it takes the observed equations and returns one number.

$$R(\beta) = \sum_{i=1}^{n} \left(y_i - \left(\beta_1 x_{i1} + \beta_2 x_{i2} \ldots, \beta_d x_{id} \right) \right)^2$$

$$\hat{\beta} = \arg\min_{\beta} R(\beta)$$

This method is named Least Squares. Gauss later claimed that he had been using this method since 1795.

LS Estimating Equation

Because Legendre's objective equation is convex, we can minimize this difference by taking the derivative and setting it to zero. This leaves us with the estimating equation for beta. Watch the matrix and vector notation very carefully—it is tricky to keep it straight.

$$R(\beta) = \sum_{i=1}^{n} \left(y_i - x_i^T \beta \right)^2$$

$$\frac{d}{d\beta} R(\beta) = 0 = -2 \sum_{i=1}^{n} \left(y_i - x_i^T \beta \right) \cdot x_i$$

$$\sum_{i=1}^{n} \left(y_i - x_i^T \beta \right) \cdot x_i = 0$$

$$\sum_{i=1}^{n} \left(x_i y_i - x_i x_i^T \beta \right) = 0$$

$$\sum_{i=1}^{n} \left(x_i y_i \right) = \sum_{i=1}^{n} \left(x_i x_i^T \beta \right)$$

$$\hat{\beta} = (\sum_{i=1}^{n} \left(x_i x_i^T \right))^{-1} (\sum_{i=1}^{n} \left(x_i y_i \right))$$

This last step is the trickiest matrix-wise and even I have not convinced myself of it …

$$\hat{\beta} = (x^T x)^{-1} (x^T y)$$

If you can be convinced of that… or even if not: We can generalize this estimating equation as follows. Assume we have n equations and d unknowns, and we would like to have d equations.

$$\underset{n \times 1}{Y} = \underset{n \times d}{X} \underset{d \times 1}{\beta}, \quad \underset{d \times n}{A}$$

$$AY = AX\beta$$

$$\beta = (AX)^{-1} AY$$

To turn our n equations into d, we can multiply both sides by a matrix A of dimensions d by n, and then solve for beta. Now we see that Least Square is of this general form with $A = x^T$.

Projection and the Cauchy Inequality

We can view the estimate, $X\beta$, as a projection of Y into the space of the model X. We can see that $\hat{Y}$ will always be smaller than Y from Cauchy's inequality.

$$\hat{Y} = X\hat{\beta} = X(x^T x)^{-1} x^T Y = HY, \quad H = X(x^T x)^{-1} x^T$$

$$H^2 = H$$

$$\|Y\|^2 \geq \|HY\|^2$$

$$Y^T Y \geq Y^T H^T HY = Y^T HY$$

Laura Balzano

I have never had a complete intuition for Cauchy-Schwartz, and thinking of things geometrically has never been my cup of tea. So this short section is not complete, but hopefully, if you are good with this sort of thing, it might give you some insight that I don't have.

Hypothetical Repeated Sampling

Take the following thought experiment: If we can fix the conditions for all n observations (X), and then measure Y repeatedly, the Y will be different from all previous observations due to random error. If we repeat this 1000 times, each time will yield a different Y. X is fixed and the Y we get is random.

If we then have new conditions x_0, how do we measure the accuracy? Recall that Y=X*beta. We can look at the difference between the new conditions with our estimated values for beta and the new conditions with the actual true value of beta.

$$y_{0true} = x_0^T \beta_{true}$$
$$\hat{y}_0 = x_0^T \hat{\beta}$$
$$E\{(x_0^T \hat{\beta} - x_0^T \beta_{true})^2\}$$

If we assume ergodicity, we can take the sample average and assume it will converge to the ensemble average over repeated sampling… that is, the sample average will converge to the statistical expected value. In order to then measure the accuracy, here we choose to look at the mean squared error. We have the MSE between a random variable $X_0^T \hat{\beta}$ and a constant $X_0^T \beta_{true}$. Let $E\{X_0^T \hat{\beta}\} = \mu$.

$$\begin{aligned}
&E\{\left|(X_0^T \hat{\beta} - \mu) + (\mu - X_0^T \beta_{true})\right|^2\} \\
&= E\{(X_0^T \hat{\beta} - \mu)^2 + (\mu - X_0^T \beta_{true})^2 - 2(X_0^T \hat{\beta} - \mu)(\mu - X_0^T \beta_{true})\} \\
&= E\{(X_0^T \hat{\beta} - \mu)^2\} + (\mu - X_0^T \beta_{true}) - 2(\mu - X_0^T \beta_{true})(E\{X_0^T \hat{\beta}\} - \mu) \\
&= Var\{X_0^T \hat{\beta}\} + (E\{X_0^T \hat{\beta}\} - X_0^T \beta_{true})^2
\end{aligned}$$

By first using the trick of adding and subtracting μ, and by then seeing that the cross-term goes to zero with the $E\{X_0^T \hat{\beta}\} - \mu$ term, we can see from this derivation that the MSE can be decomposed into the variance and the bias-squared.

If we then assume that the expectation of the noise error is 0, we can show that the Least Squares Estimator is unbiased, that is bias = 0. So in order to minimize the mean squared error, we simply have left to minimize the variance of the estimator.

Optimality of Least Squares

Laura Balzano

How can we discover whether Least Squares Estimator has an optimally minimal variance and minimizes the mean square error? Let's apply the variance + bias-squared decomposition to the general combination-of-equations form from above. This time assume that $Y = X\beta_{true} + \varepsilon$, where ε is measurement error.

$$\begin{aligned}\hat{\beta} &= (AX)^{-1}AY \\ &= (AX)^{-1}(AX\beta_{true} + A\varepsilon) \\ &= \beta_{true} + (AX)^{-1}A\varepsilon\end{aligned}$$

Now, let $\hat{y}_0 = x_0{}^T\hat{\beta}$

$$\begin{aligned}&\hat{y}_0 = x_0{}^T\beta_{true} + x_0{}^T(AX)^{-1}A\varepsilon \\ &E\{\hat{y}_0\} = x_0{}^T\beta_{true} \text{ because we assume } E\{\varepsilon\} = 0\end{aligned}$$

This concludes that the bias = 0, or that this matrix-A estimator is an unbiased estimator. Recall Least Square is therefore also in this category of unbiased estimators, because for Least Square $A = X^T$. Next we look at the variance. Recall that the first term on the right hand side is a constant. The variance is therefore as follows.

$$\begin{aligned}&Var\{\hat{y}_0\} = Var\{x_0{}^T\beta_{true}\} + Var\{x_0{}^T(AX)^{-1}A\varepsilon\} \\ &\text{Letting } a = x_0{}^T(AX)^{-1}A \text{ we have} \\ &Var\{a\varepsilon\} = (a_1{}^2 + a_2{}^2 + \ldots)\sigma^2 = aa^T\sigma^2 \\ &Var\{\hat{y}_0\} = (x_0{}^T(AX)^{-1}A)(A^T(X^TA^T)^{-1}x_0)\sigma^2 \quad \text{(Eqn 1)}\end{aligned}$$

For Least Square, $A = X^T$ and

$$\begin{aligned}&Var\{\hat{y}_0\} = (x_0{}^T(X^TX)^{-1}X^T)(X(X^TX)^{-1}x_0)\sigma^2 \\ &= x_0{}^T(X^TX)^{-1}(X^TX(X^TX)^{-1})x_0\sigma^2 \\ &= x_0{}^T(X^TX)^{-1}x_0\sigma^2 \quad \text{(Eqn 2)}\end{aligned}$$

Can we show that (Eqn 1) >= (Eqn 2), that is, in general any A will have equal or higher variance than the Least Square $A = X^T$? Again, we refer to Cauchy-Schwartz.

$Y^T Y \geq Y^T HY$

Let $H = X(X^T X)^{-1} X^T$

Let $Y = (X^T A)^{-1}$

$Y^T HY = x_0{}^T (A^T X)^{-1} A^T X (X^T X)^{-1} X^T A (X^T A)^{-1} x_0$

notice $(A^T X)^{-1} A^T X = I, \quad X^T A (X^T A)^{-1} = I$ (I is identity)

$Y^T HY = x_0{}^T (X^T X)^{-1} x_0$

And again, this is completely hand-wavy to me. Where did x_0 come from on the fourth line? I do not know. But still, we see that since $Y^T Y \geq Y^T HY$, Least Square achieves the lowest value here. In conclusion, under the assumption that we have a correct model, a constant variance, and independent error distributions, the Least Squares Estimator is optimal.

Model Capacity, Training and Testing errors, Bias and Variance

Orthonormal Regressors

Let us assume our $X_1, \ldots, X_d$ are orthonormal, that is, that the inner product between two X will give us zero, whereas the inner product of an X with itself will give us 1. Think: What does it mean for Xs to be orthonormal? I'm not sure, but let's go on. Now because we assume orthonormal basis, we can do individual projections of Y onto each X.

$$\hat{\beta}_i = \langle Y, X_i \rangle, \quad i = 1, \cdots, d$$

We can then see that Least Square can decompose our estimate into a signal part and a noise part.

$Y = f + \varepsilon, \quad f$ is true value, ε is random error

$\hat{\beta}_i = \langle f + \varepsilon, X_i \rangle$

$= \langle f, X_i \rangle + \langle \varepsilon, X_i \rangle$

$= b_i + \delta_i, \quad b$ is data or signal, $\quad \delta$ is random error or noise

The Model

At this point we would like to take a look at our model. Our best model can only approximate the truth. We will have training data and testing data, and from our training data we hope to predict our testing data accurately. Enter the trade-off between accuracy and simplicity. If we make our model too accurate to our training data, the model will be

very complex and will be a poor predictor of our testing data. On the other hand, if our model is simple but not accurate enough, we will miss the patterns in the training data.

$$\text{Observed Data: } Y_{training} = f + \varepsilon_{training}$$
$$\text{Future Data: } \quad Y_{testing} = f + \varepsilon_{testing}$$

To understand this let's first look at the training error of Least Square.

$$E\{\left\| Y_{training} - \hat{Y} \right\|^2\}$$
$$= E\{\left\| f + \varepsilon_{training} - (Xb + X\delta) \right\|^2\}$$
$$= E\{\left\| (f - Xb) + (\varepsilon_{training} - X\delta) \right\|^2\}$$

$(f - Xb)$ is the model bias, or the best the model can do.

$(\varepsilon_{training} - X\delta)$ is the noise or variation of the model.

$$= E\{\left\| (f - Xb) \right\|^2\} + E\{\left\| (\varepsilon_{training} - X\delta) \right\|^2\}$$
$$E\{\left\| (\varepsilon_{training} \right\|^2\} = n\sigma^2$$
$$E\{\left\| (X\delta) \right\|^2\} = d\sigma^2$$
$$\therefore \text{ Training Error} = \text{bias}^2 + (n - d)\sigma^2$$

Here we see that when examining the training data, if we increase the dimensionality d, both bias and variance of the estimator drop. My questions: Why does the cross-term cancel on line 6? And how do we get line 8 from line 6? The steps for the testing error are much clearer. Now let's examine the testing error.

$$E\{\left\| Y_{testing} - \hat{Y} \right\|^2\}$$
$$= E\{\left\| f + \varepsilon_{testing} - (Xb + X\delta) \right\|^2\}$$
$$= E\{\left\| (f - Xb) + (\varepsilon_{testing} - X\delta) \right\|^2\}$$

$(f - Xb)$ is the again the model bias.

But this time, $\varepsilon_{testing}$ and $X\delta$ are independent.

$$= E\{\left\| (f - Xb) \right\|^2\} + E\{\left\| \varepsilon_{testing} \right\|^2\} + E\{\left\| X\delta \right\|^2\}$$
$$E\{\left\| \varepsilon_{testing} \right\|^2\} = n\sigma^2$$
$$E\{\left\| X\delta \right\|^2\} = d\sigma^2$$
$$\therefore \text{ Testing Error} = \text{bias}^2 + (n + d)\sigma^2$$

Now we see that larger d will come back to hurt us. With the testing data, a higher d gives us a lower bias… but it also increases the variance of our estimator. We have two extremes. Too simple a model is dumb, whereas too complex a model is "superstitious," and captures pure coincidence.

In general, our testing error for estimator $\tilde{\beta}$ can be derived as follows.

Recall: $Y_{testing} = f + \varepsilon_{testing}$

$b = \langle f, X \rangle$

Error $= E\{\left\| Y_{testing} - X\tilde{\beta} \right\|^2\}$

$= E\{\left\| f + \varepsilon_{testing} - X\tilde{\beta} \right\|^2\}$

$= E\{\left\| f - Xb \right\|^2\} + E\{\left\| X(\tilde{\beta} - b) \right\|^2\} + E\{\left\| \varepsilon_{testing} \right\|^2\}$

and assuming X is orthonormal,

$= E\{\left\| f - Xb \right\|^2\} + E\{\left\| \tilde{\beta} - b \right\|^2\} + n\sigma^2$

= model bias^2 + (error in estimation - signal part of coefficients)^2 + observation error

We can reduce the center term with some intelligent manipulations on our estimator. One method is called "thresholding," where we set small $\tilde{\beta}_i$ to zero. This is like zeroing out the signal in places where it is much smaller than the noise.

$$E\{(\tilde{\beta}_i - b_i)^2\} = \sigma^2$$
$$E\{(0 - b_i)^2\} = b_i^2 \text{ for } \sigma^2 \square\ b_i^2$$

Another method is called "shrinkage," or the Ridge Estimator. A shrinking parameter λ for our estimator leads to a smaller error.

$$\tilde{\beta} = \frac{\hat{\beta}}{1+\lambda}, \quad \lambda = 0$$
$$\tilde{\beta} = \frac{b}{1+\lambda} + \frac{\delta}{1+\lambda} = b - \frac{\lambda}{1+\lambda} + \frac{\delta}{1+\lambda}$$

where $\frac{\lambda}{1+\lambda}$ is now our bias term and $\frac{\delta}{1+\lambda}$ is a smaller noise term

The Ridge Estimator is an example of the tradeoff between variance and bias. Instead of minimizing $\|Y - X\beta\|^2$, we are minimizing $\|Y - X\beta\|^2 + \lambda\|\beta\|^2$. With the first term, we do our best to explain the data and with the second term we curb the model from choosing noise. As an example, consider the spline regression. In linear spline, β represents a

change in slope. We then minimize $\lambda\|\beta\|^2$ to make the curve smooth. Wavelet regression is another example of a more complicated regressor.

Finally, here is a brief note on an interesting result that I have not explored in detail. In normal noise with $d \geq 3$, the Least Square method (i.e., Maximum Likelihood because of normal noise) can always be beaten by the Stein Estimator.

$$\tilde{\beta} = \hat{\beta}\left(1 - \frac{d-2}{\|\beta\|^2}\right)$$

From these last few sections we must take away that we have an important tradeoff between complexity of the model, degree d, and simplicity of the model in predicting future data. We can take the approach of reducing the actual degree of freedom in the model to make it simpler, even when we have a complex model with high degree d.

References:

My notes from Professor Wu's STAT 200B course, winter quarter 2005, University of California at Los Angeles. His webpage can be found here:
http://www.stat.ucla.edu/~ywu

This homework assignment page: http://www.stat.ucla.edu/~ywu/teaching/200BHW1.pdf

Wikipedia:
http://en.wikipedia.org/wiki/Gauss-Markov_theorem
http://en.wikipedia.org/wiki/Carl_Friedrich_Gauss
http://en.wikipedia.org/wiki/Least_squares

Kay, Steven. Fundamentals of Statistical Signal Processing: Estimation Theory. 1993.

[i] I would like to apologize for the incomplete proofs and lack of rigor in this report. I plan to use it as a basic reference and to improve the proofs and arguments as time passes. I welcome any suggestions to sunbeam@ee.ucla.edu.