Model Selection

Bryan S. Graham, UC - Berkeley & NBER

February 25, 2026

Let $Y_i = \mathbb{Y} \subset \mathbb{R}^1$ be a scalar outcome of interest, generated according to

$$Y_i = m\left(x_i\right) + U_i,\ U_i \overset{iid}{\sim} \mathcal{N}\left(0, \sigma^2\right),$$

with $x_i \in \mathbb{X} \subset \mathbb{R}^K$ a $K \times 1$ vector of regressors and $m : \mathbb{X} \rightarrow \mathbb{Y}$ an unknown function. For example, Y_i might be a measure of academic performance in college (e.g., cumulative GPA) and X_i a vector of student attributes measured at the time of application (e.g., high school GPA, class rank, SAT score etc.). Available to the econometrician is the training sample $\left(X_1', Y_1\right)', \left(X_2', Y_2\right)', \ldots, \left(X_N', Y_N\right)$. We consider the so-called *fixed design* case, where the observed $\mathbf{X} = \mathbf{x} = \left(x_1, \ldots, x_N\right)'$ *design matrix* is assumed non-stochastic. This could literally be true, as when the $\mathbf{X}$ corresponds to different experimental conditions controlled by the researcher (or when $\mathbf{X}$ is held fixed across repeated samples via stratification), or it could be a convenient simplifying 'as if' assumption.

Our goal is to construct estimates of the $m\left(x\right)$ function at each of the N design points (i.e., to estimate the vector $\mathbf{m} = \left(m\left(x_1\right), \ldots, m\left(x_N\right)\right)'$). We shall see that this goal is, in a particular sense, isomorphic to that of predicting new values of Y at the observed $\mathbf{X} = \mathbf{x}$:

$$Y_i^* = m\left(x_i\right) + U_i^*,\ U_i^* \overset{iid}{\sim} \mathcal{N}\left(0, \sigma^2\right),$$

for $i = 1, \ldots, N$ and $\mathbf{x} = \left(x_1, \ldots, x_N\right)'$. For example, we might wish to predict academic performance in college for a new cohort of admits whose attributes coincide with those in our training sample. We will use the training sample to construct the estimate $\hat{\mathbf{m}}$; we seek a method for choosing among a collection of possible estimators. Elements of this collection are typically index by a *tuning parameter*.

Let $\left\|\mathbf{m}\right\| = \left[\sum_{i=1}^{N} m_i^2\right]^{1/2}$ denote the Euclidean norm of a vector. As in Efron (2004), the *apparent error* of any estimate $\hat{\mathbf{m}}$ is

$$\begin{aligned}\mathbb{E}\left[\left\|\mathbf{Y}-\hat{\mathbf{m}}\right\|^{2}\right]= & \mathbb{E}\left[\left\|\left(\mathbf{Y}-\mathbf{m}\right)-\left(\hat{\mathbf{m}}-\mathbf{m}\right)\right\|^{2}\right]\\
= & \mathbb{E}\left[\left\|\left(\mathbf{Y}-\mathbf{m}\right)\right\|^{2}\right]+\mathbb{E}\left[\left\|\left(\hat{\mathbf{m}}-\mathbf{m}\right)\right\|^{2}\right]\\
 & -2\mathbb{E}\left[\left(\mathbf{Y}-\mathbf{m}\right)'\left(\hat{\mathbf{m}}-\mathbf{m}\right)\right]\\
= & N\sigma^{2}+\mathbb{E}\left[\left\|\left(\hat{\mathbf{m}}-\mathbf{m}\right)\right\|^{2}\right]\\
 & -2\mathbb{E}\left[\left(\mathbf{Y}-\mathbf{m}\right)'\hat{\mathbf{m}}\right]\\
= & N\sigma^{2}+\sum_{i=1}^{N}\mathbb{E}\left[\left(\hat{m}\left(x_{i}\right)-m\left(x_{i}\right)\right)^{2}\right]\\
 & -2\sum_{i=1}^{N}\mathbb{E}\left[\left(Y_{i}-m\left(x_{i}\right)\right)\hat{m}\left(x_{i}\right)\right]\\
= & N\sigma^{2}+\sum_{i=1}^{N}\mathbb{E}\left[\left(\hat{m}\left(x_{i}\right)-m\left(x_{i}\right)\right)^{2}\right]\\
 & -2\sum_{i=1}^{N}\mathbb{C}\left(Y_{i},\hat{m}\left(x_{i}\right)\right),\end{aligned}$$

where $\mathbf{Y}=\left(Y_{1},\ldots,Y_{N}\right)'$. Observe that the sum-of-squared residuals associated with $\hat{\mathbf{m}}$ equals

$$\left\|\mathbf{Y}-\hat{\mathbf{m}}\right\|^{2}=\sum_{i=1}^{N}\left(Y_{i}-\hat{m}\left(x_{i}\right)\right)^{2},$$

such that apparent error coincides with *expected* in sample fit (i.e., the expected sum of squared residuals). Apparent error is a pre-sample measure of how good we *expect* our estimation procedure to be at fitting the training sample (in hand).

Let

$$\mathrm{df}\left(\hat{\mathbf{m}}\right)\overset{def}{\equiv}\sum_{i=1}^{N}\frac{\mathbb{C}\left(Y_{i},\hat{m}\left(x_{i}\right)\right)}{\sigma^{2}}$$

be the *degrees-of-freedom* of $\hat{\mathbf{m}}$; a measure of the ‘complexity’ of our estimator. The reason for the degrees-of-freedom parlance will become clear as we proceed. With this notation we can write apparent error as

$$\mathbb{E}\left[\left\|\mathbf{Y}-\hat{\mathbf{m}}\right\|^{2}\right]=N\sigma^{2}+\sum_{i=1}^{N}\mathbb{E}\left[\left(\hat{m}\left(x_{i}\right)-m\left(x_{i}\right)\right)^{2}\right]-2\sigma^{2}\mathrm{df}\left(\hat{\mathbf{m}}\right). \qquad (1)$$

We will call $\sum_{i=1}^{N}\mathbb{E}\left[\left(\hat{m}\left(x_{i}\right)-m\left(x_{i}\right)\right)^{2}\right]$ the integrated mean squared error (IMSE) of $\hat{\mathbf{m}}$ for $\mathbf{m}$. This is a measure of the expected overall quality of our estimates of the $m\left(x\right)$ at each of the design points in $\mathbf{x}$. I will also call IMSE the *risk* of the estimate $\hat{\mathbf{m}}$; risk measures how well any estimation procedure does *on average* (i.e., across many replications of the experiment in hand). In this case across many training samples (all of which have identical design matrices for $\mathbf{X} = \mathbf{x}$).

Solving (1) for IMSE yields

$$\sum_{i=1}^{N} \mathbb{E}\left[\left(\hat{m}\left(x_i\right) - m\left(x_i\right)\right)^2\right] = -N\sigma^2 + \sum_{i=1}^{N} \mathbb{E}\left[\left(Y_i - \hat{m}\left(x_i\right)\right)^2\right] + 2\sigma^2 \mathrm{df}\left(\hat{\mathbf{m}}\right). \tag{2}$$

Risk, measured by integrated mean squared error, is decreasing in apparent error and increasing model degrees-of-freedom / complexity. Complicated models tend to have lots of free parameters and hence fit the training sample well, unfortunately these numerous parameters are difficult to precisely estimate. This variability is what drives the risk penalty associated with degrees of freedom.

Consider the ‘connect-the-dots’ estimator: $\hat{m}\left(x_i\right) = Y_i$. The apparent error of this estimator is zero (its fits the sample perfectly!), but it’s degrees-of-freedom is $\mathrm{df}\left(\hat{\mathbf{m}}\right) = N$. It has the same number of parameters as observations! Alternatively consider the ‘always-guess-the-mean’ estimator $\hat{m}\left(x_i\right) = \bar{Y}$ for all $i = 1, \ldots, N$ (with $\bar{Y} = \frac{1}{N}\sum_{i=1}^{N} Y_i$ the sample mean). The apparent error of this estimate is

$$\begin{aligned}
\sum_{i=1}^{N} \mathbb{E}\left[\left(Y_i - \bar{Y}\right)^2\right] &= \sum_{i=1}^{N} \mathbb{E}\left[\left(\left(Y_i - m\left(x_i\right)\right) - \left(\bar{Y} - m\left(x_i\right)\right)\right)^2\right] \\
&= \sum_{i=1}^{N} \left\{\sigma^2 - \frac{2\sigma^2}{N}\right\} + \sum_{i=1}^{N} \mathbb{E}\left[\left(\bar{Y} - m\left(x_i\right)\right)^2\right] \\
&= \sigma^2\left(N-2\right) + \sum_{i=1}^{N} \mathbb{E}\left[\left(\frac{1}{N}\sum_{i=1}^{N}\left(Y_i - m\left(x_i\right)\right) - \left(m\left(x_i\right) - \frac{1}{N}\sum_{i=1}^{N} m\left(x_i\right)\right)\right)^2\right]. \\
&= \sigma^2\left(N-1\right) + \sum_{i=1}^{N}\left(m\left(x_i\right) - \frac{1}{N}\sum_{i=1}^{N} m\left(x_i\right)\right)^2
\end{aligned}$$

The expected in sample fit of ‘always-guess-the-mean’ is poor, especially in the underlying data are noisy (large σ^2) and/or $m\left(x_i\right)$ varies a lot with x_i. In contrast the ‘always-guess-the-mean’ is a simple, easy to estimate model, with low degrees of freedom:

$$\mathrm{df}\left(\hat{\mathbf{m}}\right) = \sum_{i=1}^{N} \frac{\mathbb{C}\left(Y_i, \bar{Y}\right)}{\sigma^2} = 1.$$

Hence the risk of the 'always-guess-the-mean' estimator equals

$$\sum_{i=1}^{N}\mathbb{E}\left[\left(\bar{Y}-m\left(x_i\right)\right)^2\right]=\sum_{i=1}^{N}\left(m\left(x_i\right)-\frac{1}{N}\sum_{i=1}^{N}m\left(x_i\right)\right)^2+\sigma^2.$$

Ideally we'd like a way to choose between 'connect-the-dots', 'always-guess-the-mean' and/or some intermediate estimator.

Let $\mathbf{Y}^*$ be a vector of new draws of Y_i, holding the design points fixed. The out-of-sample *prediction risk* of $\hat{\mathbf{m}}$ is

$$\begin{aligned}\mathbb{E}\left[\left\|\mathbf{Y}^*-\hat{\mathbf{m}}\right\|^2\right]=&\mathbb{E}\left[\left\|\left(\mathbf{Y}^*-\mathbf{m}\right)-\left(\hat{\mathbf{m}}-\mathbf{m}\right)\right\|^2\right]\\=&\mathbb{E}\left[\left\|\left(\mathbf{Y}^*-\mathbf{m}\right)\right\|^2\right]+\mathbb{E}\left[\left\|\left(\hat{\mathbf{m}}-\mathbf{m}\right)\right\|^2\right]\\&-2\mathbb{E}\left[\left(\mathbf{Y}^*-\mathbf{m}\right)'\left(\hat{\mathbf{m}}-\mathbf{m}\right)^2\right]\\=&N\sigma^2+\mathbb{E}\left[\left\|\left(\hat{\mathbf{m}}-\mathbf{m}\right)\right\|^2\right]+0\end{aligned}$$

Since $\mathbb{E}\left[\left\|\mathbf{Y}^*-\hat{\mathbf{m}}\right\|^2\right]\propto\mathbb{E}\left[\left\|\left(\hat{\mathbf{m}}-\mathbf{m}\right)\right\|^2\right]$ if we want to minimize prediction risk, then we might as well just minimize IMSE risk for the function itself.

Before we continue, it is worth observing that IMSE can be decomposed into a bias and variance component:

$$\begin{aligned}\mathbb{E}\left[\left\|\left(\hat{\mathbf{m}}-\mathbf{m}\right)\right\|^2\right]=&\mathbb{E}\left[\left\|\hat{\mathbf{m}}-\mathbb{E}\left[\hat{\mathbf{m}}\right]+\mathbb{E}\left[\hat{\mathbf{m}}\right]-\mathbf{m}\right\|^2\right]\\=&\mathbb{E}\left[\left\|\hat{\mathbf{m}}-\mathbb{E}\left[\hat{\mathbf{m}}\right]\right\|^2\right]+\left\|\mathbb{E}\left[\hat{\mathbf{m}}\right]-\mathbf{m}\right\|^2\\&+2\mathbb{E}\left[\left(\hat{\mathbf{m}}-\mathbb{E}\left[\hat{\mathbf{m}}\right]\right)'\right]\left(\mathbb{E}\left[\hat{\mathbf{m}}\right]-\mathbf{m}\right)\\=&\mathbb{E}\left[\left\|\hat{\mathbf{m}}-\mathbb{E}\left[\hat{\mathbf{m}}\right]\right\|^2\right]+\left\|\mathbb{E}\left[\hat{\mathbf{m}}\right]-\mathbf{m}\right\|^2\\=&\sum_{i=1}^{N}\mathbb{V}\left(\hat{m}\left(x_i\right)\right)+\sum_{i=1}^{N}\left(\mathbb{E}\left[\hat{m}\left(x_i\right)\right]-m\left(x_i\right)\right)^2.\end{aligned}$$

Risk for linear estimators

Let $\lambda_J\left(x_i\right)$ be a $J\times 1$ vector of linearly-independent basis functions in x_i. For example with x_i scalar we might choose the power series $\lambda_J\left(x_i\right)=\left(1,x_i,x_i^2,\ldots,x_i^{J-1}\right)'$. If $X_i\in[0,1]$, we might use the cosine basis $\lambda_J\left(x_i\right)=\left(1,\sqrt{2}\cos(\pi x_i),\sqrt{2}\cos(\pi 2x_i),\ldots,\sqrt{2}\cos(\pi\left(J-1\right)x_i)\right)'$. Many options are possible.

Let $\mathbf{W}_J=\left(\lambda_J\left(x_1\right),\ldots,\lambda_J\left(x_N\right)\right)'$ and consider the OLS fit of $\mathbf{Y}$ onto $\mathbf{W}_J$

$$\hat{\pi}_J=\left(\mathbf{W}_J'\mathbf{W}_J\right)^{-1}\mathbf{W}_J'\mathbf{Y}.$$

Our estimate of $\hat{m}_J(x_i)$ will be the corresponding fitted value $\lambda_J(x_i)'\hat{\pi}_J$:

$$\hat{m}_J(x_i) = \lambda_J(x_i)'\hat{\pi}_J = \lambda_J(x_i)'\left(\mathbf{W}_J'\mathbf{W}_J\right)^{-1}\mathbf{W}_J'\mathbf{Y}.$$

Hence we have $\hat{\mathbf{m}}_J = \left(\hat{m}_J(x_1), \ldots, \hat{m}_J(x_N)\right)'$ equal to

$$\begin{aligned}\hat{\mathbf{m}} &= \mathbf{W}_J\left(\mathbf{W}_J'\mathbf{W}_J\right)^{-1}\mathbf{W}_J'\mathbf{Y} \\ &= \mathbf{H}_J\mathbf{Y}.\end{aligned}$$

This is a linear estimator; $\mathbf{H}_J = \mathbf{W}_J\left(\mathbf{W}_J'\mathbf{W}_J\right)^{-1}\mathbf{W}_J$ is an $N \times N$ matrix and the elements of $\hat{\mathbf{m}}$ are linear combinations of the elements of the outcome vector $\mathbf{Y}$. It turns out that degrees-of-freedom has a nice closed-form expression for linear estimators:

$$\begin{aligned}\mathbb{E}\left[(\mathbf{Y}-\mathbf{m})'\hat{\mathbf{m}}_J\right] &= \mathbb{E}\left[(\mathbf{Y}-\mathbf{m})'\mathbf{H}_J\mathbf{Y}\right] \\ &= \mathrm{Tr}\left(\mathbb{E}\left[(\mathbf{Y}-\mathbf{m})'\mathbf{H}_J\mathbf{Y}\right]\right) \\ &= \mathrm{Tr}\left(\mathbb{E}\left[\mathbf{H}_J\mathbf{Y}(\mathbf{Y}-\mathbf{m})'\right]\right) \\ &= \mathrm{Tr}\left(\mathbf{H}_J\mathbb{E}\left[\mathbf{Y}(\mathbf{Y}-\mathbf{m})'\right]\right) \\ &= \mathrm{Tr}\left(\mathbf{H}_J\sigma^2 I_N\right) \\ &= \sigma^2\mathrm{Tr}\left(\mathbf{H}_J\right) \\ &= \sigma^2\mathrm{Tr}\left(\mathbf{W}_J\left(\mathbf{W}_J'\mathbf{W}_J\right)^{-1}\mathbf{W}_J\right) \\ &= \sigma^2\mathrm{Tr}\left(\mathbf{W}_J'\mathbf{W}_J\left(\mathbf{W}_J'\mathbf{W}_J\right)^{-1}\right) \\ &= \sigma^2\mathrm{Tr}\left(I_J\right) \\ &= \sigma^2 J.\end{aligned}$$

So we have $\mathrm{df}\left(\hat{\mathbf{m}}_J\right) = J$.

The risk of the estimator which, (i) approximates $m(x_i)$ by a linear combination of the vector of basis functions $\lambda(x_i)$ and (ii) choose the linear combination via a least squares fit, is therefore

$$\sum_{i=1}^{N}\mathbb{E}\left[\left(\hat{m}_J(x_i) - m(x_i)\right)^2\right] = -N\sigma^2 + \sum_{i=1}^{N}\mathbb{E}\left[\left(Y_i - \hat{m}_J(x_i)\right)^2\right] + 2\sigma^2 J.$$

Let $\tilde{\sigma}^2$ be an “unbiased” estimate of σ^2; one way of constructing such an estimate is described below.

Mallow's C_J criterion

$$C_J = -N\tilde{\sigma}^2 + \sum_{i=1}^{N} \left(Y_i - \hat{m}_J\left(x_i\right)\right)^2 + 2\tilde{\sigma}^2 J,$$

is an unbiased risk estimate.

Let L be some integer equal to the length of the largest approximating vector of basis functions being considered (we do require that $L \ll N$). Let $\lambda_L\left(x_i\right)$ be the $L \times 1$ vector of basis functions in x_i associated with this "large" model. Let $\hat{m}_L\left(x_i\right) = \lambda_L\left(x_i\right)' \hat{\pi}_L$ and define $\tilde{\sigma}^2 = \frac{1}{N-L}\sum_{i=1}^{N}\left(Y_i - \hat{m}_L\left(x_i\right)\right)^2$. With this feasible estimate of σ^2 we can choose J, the number of basis functions in our approximation of $m\left(x_i\right)$, to minimize C_J.

The calculations above indicate that, for general linear estimators, we have $\mathrm{df}\left(\hat{\mathbf{m}}\right) = \sum_{i=1}^{N} h_{ii}$, where h_{ii} is the i^{th} diagonal matrix of some general smoothing matrix $\hat{\mathbf{m}} = \mathbf{HY}$.

Stein's Unbiased Risk Estimate (SURE)

Consider a, possibly nonlinear and very complicated, estimator with

$$g\left(\mathbf{Y}\right) = \hat{\mathbf{m}} - \mathbf{Y}. \tag{3}$$

Note for an estimator we can always write it as $\hat{\mathbf{m}} = g\left(\mathbf{Y}\right) + \mathbf{Y}$ for some $g\left(\mathbf{Y}\right)$.

Stein (1981), using an elegant integration-by-parts argument, shows that under normality of $\mathbf{Y}$, the equality

$$\mathbb{E}\left[g_i\left(\mathbf{Y}\right)\left(Y_i - m\left(x_i\right)\right)\right] = \sigma^2 \mathbb{E}\left[\frac{\partial g_i\left(\mathbf{Y}\right)}{\partial Y_i}\right]. \tag{4}$$

Next observe that we can write, for $\hat{\mathbf{m}}$as in (3),

$$\begin{aligned}\mathbb{C}\left(\hat{m}\left(x_i\right), Y_i\right) &= \mathbb{E}\left[\hat{m}\left(x_i\right)\left(Y_i - m\left(x_i\right)\right)\right] \\ &= \mathbb{E}\left[\left(\hat{m}\left(x_i\right) - Y_i\right)\left(Y_i - m\left(x_i\right)\right)\right] + \mathbb{E}\left[Y_i\left(Y_i - m\left(x_i\right)\right)\right] \\ &= \mathbb{E}\left[g_i\left(\mathbf{Y}\right)\left(Y_i - m\left(x_i\right)\right)\right] + \sigma^2,\end{aligned}$$

which in conjunction with (4) gives

$$\mathbb{C}\left(\hat{m}\left(x_i\right), Y_i\right) = \sigma^2 \mathbb{E}\left[\frac{\partial g_i\left(\mathbf{Y}\right)}{\partial Y_i}\right] + \sigma^2.$$

However since

$$\frac{\partial g_i\left(\mathbf{Y}\right)}{\partial Y_i} = \frac{\partial \hat{m}\left(x_i\right)}{\partial Y_i} - 1$$

we have the following alternative definition of model degrees-of-freedom

$$\mathrm{df}\left(\hat{\mathbf{m}}\right)=\sum_{i=1}^{N}\mathbb{E}\left[\frac{\partial\hat{m}\left(x_i\right)}{\partial Y_i}\right].$$

Therefore we can re-write IMSE, Equation (2) above, as

$$\sum_{i=1}^{N}\mathbb{E}\left[\left(\hat{m}\left(x_i\right)-m\left(x_i\right)\right)^2\right]=-N\sigma^2+\sum_{i=1}^{N}\mathbb{E}\left[\left(Y_i-\hat{m}\left(x_i\right)\right)^2\right]+2\sigma^2\sum_{i=1}^{N}\mathbb{E}\left[\frac{\partial\hat{m}\left(x_i\right)}{\partial Y_i}\right]. \quad (5)$$

Let $\hat{\mathbf{m}}_\lambda$ we some estimate indexed by the turning parameter λ (e.g., the depth of a tree as discussed in lecture). *Stein's Unbiased Risk Estimate* (SURE) is:

$$\mathrm{SURE}\left(\hat{\mathbf{m}}_\lambda,\mathbf{m}\right)=-N\sigma^2+\sum_{i=1}^{N}\left(Y_i-\hat{m}_\lambda\left(x_i\right)\right)^2+2\sigma^2\sum_{i=1}^{N}\frac{\partial\hat{m}_\lambda\left(x_i\right)}{\partial Y_i}. \quad (6)$$

Taking expectations of (6) returns (5): on average (6) equals actual risk; hence it is an unbiased risk estimate. Equation (6) can be used to re-derive Mallow's C_J criterion in a different way. Let $\mathbf{H}_J=\left[h_{ij}\right]_{N\times N}=\mathbf{W}_J\left(\mathbf{W}_J'\mathbf{W}_J\right)^{-1}\mathbf{W}_J'$. Observe that $J=\mathrm{Tr}\left(\mathbf{H}_J\right)=\sum_{i=1}^{N}h_{ii}=\sum_{i=1}^{N}\frac{\partial\hat{m}_J(x_i)}{\partial Y_i}$.

Ye (1998) proposes a quite general approach to constructing an estimate of $\mathrm{SURE}\left(\hat{\mathbf{m}}_\lambda,\mathbf{m}\right)$. Their procedure can be applied to regression Trees and other complex/nonlinear estimators. Let $\hat{\mathbf{m}}_\lambda$ be some estimation procedure indexed by the tuning parameter λ. Ye (1998) proposes the following algorithm.

1. For $s=1,\ldots,S$
   - (a) Draw $D_i^{(s)}\sim N\left(0,h^2\right)$ for $i=1,\ldots,N$;
   - (b) Compute $\hat{\mathbf{m}}_\lambda^{(s)}$ using the perturbed dataset $\mathbf{Y}+\mathbf{D}^{(s)}$ for $\mathbf{D}^{(s)}=\left(D_1^{(s)},\ldots,D_N^{(s)}\right)$;
2. For each $i=1,\ldots,N$
   - (a) Compute the OLS fit of $\hat{m}_\lambda^{(s)}\left(x_i\right)$ onto 1 and $D_i^{(s)}$ using $s=1,\ldots,S$;
   - (b) Let $\hat{\delta}_i$ be the coefficient on $D_i^{(s)}$.
3. Construct the degrees-of-freedom estimate $\hat{\mathrm{df}}\left(\hat{\mathbf{m}}_\lambda\right)=\sum_{i=1}^{N}\hat{\delta}_i$ and hence the corresponding risk estimate

Ye (1998) recommends setting $h=0.6\tilde{\sigma}$. Intuitively this algorithm is based on the claim that $\hat{\delta}_i\approx\frac{\partial\hat{m}(x_i)}{\partial Y_i}$.

Cross-fitting

Recall that out-of-sample prediction risk equals:

$$\mathbb{E}\left[\left\|\mathbf{Y}^{*}-\hat{\mathbf{m}}\right\|^{2}\right]=N\sigma^{2}+\mathbb{E}\left[\left\|\left(\hat{\mathbf{m}}-\mathbf{m}\right)\right\|^{2}\right].$$

Imagine you had access to two samples: the *training sample*, $\mathbf{Y}$ and the *validation sample*, $\mathbf{Y}^{*}$ (sometimes $\mathbf{Y}$ and $\mathbf{Y}^{*}$ are called *folds*). Let $\hat{\mathbf{m}}_{\lambda}$ be some estimate of $\mathbf{m}$ indexed by tuning parameter λ. We compute $\hat{\mathbf{m}}_{\lambda}$ using the training sample and then construct an unbiased estimate of risk using the validation sample. The sample splitting risk estimate is

$$\mathrm{SS}\left(\hat{\mathbf{m}}_{\lambda},\mathbf{m}\right)\propto\sum_{i=1}^{N}\left(Y_{i}^{*}-\hat{m}_{\lambda}\left(x_{i}\right)\right)^{2}.$$

Cross-fitting: Do above, but the switch the roles of $\mathbf{Y}^{*}$ and $\mathbf{Y}$. Estimate $\hat{\mathbf{m}}_{\lambda}^{*}$ using $\mathbf{Y}^{*}$. We can estimate risk using the average

$$\mathrm{CF}\left(\hat{\mathbf{m}}_{\lambda},\mathbf{m}\right)\propto\frac{1}{2}\sum_{i=1}^{N}\left(Y_{i}^{*}-\hat{m}_{\lambda}\left(x_{i}\right)\right)^{2}+\frac{1}{2}\sum_{i=1}^{N}\left(Y_{i}-\hat{m}_{\lambda}^{*}\left(x_{i}\right)\right)^{2}.$$

This approach to risk estimation is also quite flexible.

References

Efron, B. (2004). The estimation of prediction error. *Journal of the American Statistical Association*, 99(467), 619 – 632.

Stein, C. M. (1981). Estimation of the mean of a multivariate normal distribution. *Annals of Statistics*, 9(6), 1135 – 1151.

Ye, J. (1998). On measuring and correcting the effects of data mining and model selection. *Journal of the American Statistical Association*, 93(441), 120 – 131.