Chapter 2

Conditional prediction

In many countries a variety of social transfer programs exists to ameliorate the negative effects of poverty. Examples include direct cash transfers, in kind transfers (e.g., meals, food stamps), and price subsidies (e.g, for basic foodstuffs, fuel etc.). In some cases such programs are provided universally, to all households, but in many more they are putatively targeted toward poorer households. Program targeting requires a method to determine eligibility; that is, to quickly determine whether a household is poor (i.e., eligible) or not. Unfortunately targeting, as practiced in many countries, is very inefficient. Coady, Grosh and Hoddinott (2004), reviewing a sample of programs across 48 countries, find that while the median program is progressively targeted, meaning that poor households do disproportionately benefit, up to a quarter of programs are, in fact, regressive.

Implementing a good targeting mechanisms is difficult for many reasons. Concerns about eligibility manipulation may loom large in some settings (e.g., Camacho and Conover, 2011). Another challenge is that it may be difficult to determine which households are poor (i.e., eligible) and which are not (even in the absence of any strategic behavior on the part of households). Poverty can be hard to measure. In poorer countries, where many individuals may work informally, engage in subsistence activities, and make in-kind transfers to one another, an income-based measure of poverty may be very noisy. In these settings poverty is typically assessed by measuring a household's total (real) outlay on all goods and services within a certain time period (e.g., the past month). Such a consumption aggregate can be constructed using a detailed expenditure survey, such as a Living Standards Measurement Survey collected under the auspices of the World Bank (cf., Deaton and Zaidi, 2002).

Unfortunately expenditure surveys are costly to field. A typical expenditure survey may contain hundreds of questions about food consumption alone. Furthermore a skilled enumerator is required to ensure useful data are actually collected. Post-collection processing time can also be substantial. These considerations motivate the question of whether cheaper, but relatively accurate, mechanisms for categorizing households as poor are available. For example, a government official might categorize a household as poor in they lack some combination of (i) running water, (ii) electricity and (iii) a permanent floor (an "unmet basic needs" type index).

Are there other, easily observed covariates, which could be used to improve targeting? Colombia, and some other countries in Latin America, use an elaborate proxy means approach to determining program eligibility. In Colombia each household is assigned a SISBEN number, a particular linear combination of many easily measured household characteristics. The SISBEN number is used to determined eligibility for various social safety net programs. For our purposes we can view the SISBEN number as a proxy for true total household outlay. How good is this proxy? Could it be improved? Grosh and Baker (1995) discuss approaches to proxy means testing for social programs in the developing world.

2.1 Prediction when the population distribution is known

Let $F_{X,Y}(x,y) = \Pr(X \leq x, Y \leq y)$ be the joint distribution of household characteristics (X) and consumption (Y). We will assume that X contains household characteristics that are easy to measure/observe. In contrast aggregate household consumption, Y, is difficult and costly to measure. Hence our interest in constructing a "cheap" but "accurate" prediction of Y using knowledge of X. We will call X the **covariates** and Y the **response** or outcome. In Machine Learning covariates are called **features** or attributes.

Initially assume that the policy maker knows the population joint distribution of X and Y. She faces the following prediction problem. She observes a sequence of random draws from the population and observes their value for X. On the basis of these observations she seeks to predict consumption Y. Concretely you can imagine our government official entering a village and predicting Y for each household on the basis of their (easily observed) values of X. Households with low levels of predicted consumption will be deemed eligible for social assistance.

Our policymaker would like her predictions to be accurate. We will assume she pays a penalty or **loss** based on the accuracy of her prediction. Let $L(y-g)$ be the loss associated with making a prediction of g, when actual consumption equals y. We assume that $L(\bullet)$ is weakly increasing in $|y-g|$, however it may be asymmetric. In many situations it may be reasonable to attach different utility costs to under- versus over-prediction. Our choice of loss function can be highly consequential, nevertheless, in practice, considerations of convenience, analytic tractability, and historic practice often drive this choice.

Since our decision-maker observes $X = x$ prior to making her prediction she conditions her action on the observed value of X. Let $g(x)$ equal the prediction she makes when she observes $X = x$. For different values of x she may make different predictions. We may therefore think of $g(x)$ as a prediction function or **decision rule**. If the conditional distribution of Y given $X = x$ varies with x, then we should typically expect $g(x)$, the policy-maker's prediction/decision, to vary with x as well.

Ideally our decision-maker would like to choose $g(x)$ to minimize loss. However this is non-operational since loss is only observed/experienced *after* she makes her prediction. Instead we assume she chooses $g(x)$ to minimize the *average* penalty paid across many (i.e., an infinite number of) replications of her prediction problem; an object we will call **risk**. That is she choose $g(x)$ to minimize

$$\mathbb{E}\left[L(Y - g(x))|\, X = x\right] = \int L(y - g(x))\, f_{Y|X}(y|\, x)\, \mathrm{d}m(y), \tag{2.1.1}$$

where m is the counting measure for discrete y:

$$\int_{y\in\mathbb{Y}} y f_{Y|X}(y|\, x)\, \mathrm{d}m(y) = \sum_{y_k\in\mathbb{Y}} y_k p(y_k|\, x)$$

and the Lebesgue measure for continuous y:

$$\int_{y\in\mathbb{Y}} y f_{Y|X}(y|\, x)\, \mathrm{d}m(y) = \int_{y\in\mathbb{Y}} y f_{Y|X}(y|\, x)\, \mathrm{d}y.$$

Choosing $g(x)$ to minimize (2.1.1) *is* operational. Since our decision-maker knows the joint distribution of X and Y in the population she can numerically evaluate (2.1.1) for any candidate conditional prediction. By varying her candidate prediction she can find the/a risk-minimizing one.

If we assume loss is proportional to the square of our prediction error (a convenient and historically important loss function). (i.e., $L(Y - g(x)) = (Y - g(x))^2$), then, for $\mathbb{E}[Y^2] < \infty$, risk is equal to **mean square error (MSE)**

$$\begin{aligned}
\mathbb{E}\left[L(Y - g(x))|\, X = x\right] &= \mathbb{E}\left[\left.(Y - g(x))^2\right| x\right] \\
&= \mathbb{E}\left[\left.(Y - \mathbb{E}[Y|\, x] - (g(x) - E[Y|\, x]))^2\right| x\right] \\
&= \mathbb{E}\left[\left.(Y - \mathbb{E}[Y|\, x])^2\right| x\right] + 2\mathbb{E}\left[(Y - \mathbb{E}[Y|\, x])|\, x\right](g(x) - E[Y|\, x]) \\
&+ (g(x) - E[Y|\, x])^2 \\
&= \mathbb{V}(Y|\, x) + (g(x) - E[Y|\, x])^2,
\end{aligned}$$

where we use the notation $\mathbb{E}[Y|\, X = x] = \mathbb{E}[Y|\, x]$ and $\mathrm{Var}(Y|\, X = x) = \mathbb{V}(Y|\, x)$. Clearly

MSE is minimized at $g^*(x) = \mathbb{E}[Y|\,x]$. Under squared error loss the conditional expectation of Y given $X = x$ is the optimal predictor.

Observe that risk has two components. The first is the intrinsic variability of Y given X. In our example, the variance of household consumption across households identical in X. In Machine Learning this is called the **noise floor**. The second component is bias or structural error ($|g(x) - E[Y|\,x]|$).

In our problem the decision maker knowns $F_{X,Y}$ and observes X for all households. In such a situation the optimal prediction function *under squared error loss* is the conditional expectation function (CEF) . Below we will explore how the decision maker's behavior changes when she does not observe $F_{X,Y}$. Before doing so we briefly review some basic properties of the conditional expectation function.

Conditional expectation function

Let (X, Y) denote a random draw from the joint distribution $F_{X,Y}$ with $\mathbb{E}[|Y|] < \infty$. The **conditional expectation function** (CEF) of Y given $X = x$ is

$$\mathbb{E}[Y|\,x] = \mu_{Y|X}(x) = \int y f_{Y|X}(y|\,x)\,\mathrm{d}m(y).$$

We assume that the relevant integral or sum exists. The definition of a conditional expectation function for the case where x is continuously-valued involves some technical issues which we will ignore here (Proschan and Presnell, 1988).

The roof distribution

As a simple example of a CEF consider the so-called roof distribution introduced by Goldberger (1991). This is a distribution for the bivariate random variable (X, Y) with joint probability density function $f_{X,Y}(x, y) = x + y$ if $x, y \in [0, 1]$ and zero otherwise. Integrating over y we get a marginal pdf for X of

$$f_X(x) = \int_{-\infty}^{+\infty} f_{X,Y}(x, y)\,\mathrm{d}y = \int_0^1 (x + y)\,\mathrm{d}y = \left[xy + \frac{1}{2}y^2\right]_0^1 = x + \frac{1}{2}.$$

Therefore the conditional pdf of Ygiven $X = x$ is

$$f_{Y|X}(y|\,x) = \frac{f_{Y,X}(x, y)}{f_X(x)} = \frac{x + y}{x + \frac{1}{2}},$$

for all $x, y \in [0, 1]$ and zero otherwise. Computing the CEF we get

$$\mu_{Y|X}(x) = \int_0^1 y\left(\frac{x + y}{x + \frac{1}{2}}\right)\mathrm{d}y = \frac{\left[\frac{1}{2}xy^2 + \frac{1}{3}y^3\right]_0^1}{x + \frac{1}{2}} = \frac{3x + 2}{6x + 3}$$

for $x \in [0, 1]$.

The CEF has a number of useful properties. First consider the **Law of Iterated Expectations**. For any two random variables Y and X with $\mathbb{E}[|Y|] < \infty$ we have

$$\mathbb{E}[Y] = \mathbb{E}[\mathbb{E}[Y|\,X]]. \tag{2.1.2}$$

Equation (2.1.2) follows from the equalities

$$\begin{aligned}
\mathbb{E}[Y] &= \int y f_Y(y)\,\mathrm{d}m(y) \\
&= \int y\left[\int f_{X,Y}(x, y)\,\mathrm{d}m(x)\right]\mathrm{d}m(y) \\
&= \int y\left[\int f_{Y|X}(y|\,x) f_X(x)\,\mathrm{d}m(x)\right]\mathrm{d}m(y) \\
&= \int\left[\int y f_{Y|X}(y|\,x)\,\mathrm{d}m(y)\right] f_X(x)\,\mathrm{d}m(x) \\
&= \int \mathbb{E}[Y|\,x] f_X(x)\,\mathrm{d}m(x) \\
&= \mathbb{E}_X[\mathbb{E}[Y|\,X]].
\end{aligned}$$

Relationship (2.1.2) often facilitates the computation of expected values. We will use it repeatedly.

The conditional expectation also has a **linearity** property

$$\mathbb{E}\left[\alpha Y+\beta Z|\, X\right]=\alpha\mathbb{E}\left[Y|\, X\right]+\beta\mathbb{E}\left[Z|\, X\right].$$

Another property that we will use is the **modulus inequality**

$$\left|\mathbb{E}\left[Y|\, X\right]\right|\leq\mathbb{E}\left[\left|Y\right||\, X\right].$$

Mean regression

The CEF is often called the **mean regression function** (e.g., Hansen, 2011; Manski, 2007). Specifically we say that the mean regression of Y on X is equal to $\mu_{Y|X}\left(X\right).$ Letting $U=Y-\mu_{Y|X}\left(X\right)$ be the regression function error we may write

$$Y=\mu_{Y|X}\left(X\right)+U,$$

so that realizations of Y may be viewed as random deviations about their conditional mean. Using the linearity and iteration properties of the conditional expectation we have immediately:

1. $\mathbb{E}\left[U|\, X\right]=\mathbb{E}\left[Y-\mu_{Y|X}\left(X\right)\middle|\, X\right]=\mathbb{E}\left[Y|\, X\right]-\mu_{Y|X}\left(X\right)=0$; the regression error is conditionally mean zero.

2. $\mathbb{E}\left[U\right]=\mathbb{E}\left[\mathbb{E}\left[U|\, X\right]\right]=\mathbb{E}\left[0\right]=0$; the regression error is unconditionally mean zero.

3. $\mathbb{E}\left[h\left(X\right)U\right]=\mathbb{E}\left[\mathbb{E}\left[h\left(X\right)U|\, X\right]\right]=\mathbb{E}\left[h\left(X\right)\mathbb{E}\left[U|\, X\right]\right]=\mathbb{E}\left[h\left(X\right)0\right]=0$; the regression error is uncorrelated with any function of X.

An intuitive understanding of these properties may be formed by recalling that the conditional expectation is a MSE-optimal predictor. If $\mathbb{E}\left[U\right]>0$, then prediction error is systematically positive (i.e., systematic under prediction). In this case a decision-maker would be able to lower MSE risk by raising their prediction (for all values of X); hence the requirement that prediction error is mean zero is an implication of predictive optimality (in the MSE sense). Likewise if $\mathbb{E}\left[h\left(X\right)U\right]>0$, then large values of $h\left(X\right)$ are systematically associated with under prediction. We can once again lower MSE risk by raising our prediction when we observe large values of $h\left(X\right)$. Interpreting mean independence conditions in terms of predictive optimality is frequently helpful in econometric modeling.

2.2 Prediction when the population distribution is unknown

Returning to the prediction problem introduced above, lets consider the more realistic case where the decision maker lacks complete knowledge of the joint distribution of X and Y. Instead all she has at her disposal is a random **training sample** of size N: $\{(X_i, Y_i)\}_{i=1}^N$.

Let $\mathbf{Y} = (Y_1, \ldots, Y_N)'$ be our response vector and $\mathbf{X} = (X_1, \ldots, X_N)'$ our covariate or **design matrix**. Our decision maker wonders how she can advantageously use the training sample *in hand* to predict the value of a *new* draw of Y. To keep things simple we will assume that the covariate value for any unit drawn in the future will coincide with a value contained in the training sample.

Our decision maker wants to construct a procedure or **decision rule** that will generate good predictions on average. Here the averaging is over the universe of possible training samples she might have observed. We will assume that the configuration of $\mathbf{X}$ is fixed across these different possible training samples, with only the value of $\mathbf{Y}$ changing. Conditioning on $\mathbf{X}$ simplifies her problem, allowing her to behave 'as if' it were non-stochastic. This could literally be true, as would be the case if the sample were a stratified random one. For example, returning to our proxy targeting problem, a government official could partition the universe of households according to X and then select a single household at random from each cell, measuring their total outlay Y. In other cases treating $\mathbf{X}$ 'as if' it were non-stochastic is nothing more than a convenient thought experiment. Even with this simplification, prediction on the basis of a single training sample is a considerably more difficult than prediction when $F_{X,Y}$ is known.

The decision maker knows that the mean squared error minimizing prediction of Y given $X = x$ is its conditional mean. Her goal, therefore, is to use the target sample to construct a good estimate of the $N \times 1$ conditional mean vector $\mathbf{m} = \mathbb{E}\left[\mathbf{Y}|\, \mathbf{X}\right]$. Let $\hat{m}_i = \hat{m}\left(X_i\right)$ denote her prediction of Y when $X = X_i$. We should think of $\hat{m}_i$ as a (decision) rule. The decision maker will use $\hat{m}_i$ as her outcome prediction for any future unit with $X = X_i$. Note that her rule is random; it depends on the realized values of Y in the training sample.

Let $\hat{\mathbf{m}}$ be the stacked $N \times 1$ vector of predictions. Recall that the design matrix $\mathbf{X}$ is fixed throughout, the randomness in $\hat{\mathbf{m}}$ therefore arises from randomness in $\mathbf{Y}$ across different possible training samples. We could write $\hat{\mathbf{m}} = \mathbf{d}\left(\mathbf{Y}\right)$, say, to emphasize this dependence. Risk equals $\mathbb{E}\left[\|(\hat{\mathbf{m}} - \mathbf{m})\|^2\right]$ (here $\|\bullet\|$ denotes the Euclidian norm $\|\mathbf{m}\| = \left[\sum_{i=1}^N m_i^2\right]^{1/2}$). The expected within-sample squared prediction error, sometimes called **apparent error** (e.g., Efron, 2004), is

$$\begin{aligned}
\mathbb{E}\left[\|\mathbf{Y}-\hat{\mathbf{m}}\|^2\right] &= \mathbb{E}\left[\|(\mathbf{Y}-\mathbf{m})+(\mathbf{m}-\hat{\mathbf{m}})\|^2\right] \\
&= \mathbb{E}\left[\|(\mathbf{Y}-\mathbf{m})\|^2\right]+\mathbb{E}\left[\|(\mathbf{m}-\hat{\mathbf{m}})\|^2\right] \\
&\quad +2\mathbb{E}\left[(\mathbf{Y}-\mathbf{m})'(\mathbf{m}-\hat{\mathbf{m}})\right] \\
&= N\sigma^2+\mathbb{E}\left[\|(\hat{\mathbf{m}}-\mathbf{m})\|^2\right]-2\sum_{i=1}^{N}\mathbb{C}\left(Y_i,\hat{\mathbf{m}}_i\right) \\
&= N\sigma^2+\mathbb{E}\left[\|(\hat{\mathbf{m}}-\mathbf{m})\|^2\right]-2\sigma^2\mathrm{df}\left(\hat{\mathbf{m}}\right) \qquad (2.2.1)
\end{aligned}$$

where

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

is the **degrees of freedom** associated with the rule $\hat{\mathbf{m}}$ (Ye, 1998; Efron, 2004). The reasons for this nomenclature will become apparent later on, but for now you may think of $\mathrm{df}\left(\hat{\mathbf{m}}\right)$ as a measure of model complexity.

Rearranging (2.2.1) we see that risk equals

$$\mathbb{E}\left[\|(\hat{\mathbf{m}}-\mathbf{m})\|^2\right]=-N\sigma^2+\mathbb{E}\left[\|\mathbf{Y}-\hat{\mathbf{m}}\|^2\right]+2\sigma^2\mathrm{df}\left(\hat{\mathbf{m}}\right). \qquad (2.2.2)$$

Expression (2.2.2) indicates that risk is increasing in expected training error, $\mathbb{E}\left[\|\mathbf{Y}-\hat{\mathbf{m}}\|^2\right]$. That is, that is all things equal, rules with low risk produce good in sample fits. However risk is also increasing in model degrees of freedom or complexity, $\mathrm{df}\left(\hat{\mathbf{m}}\right)$. There is a tension here: complex models produce good in sample fits, which lowers risk, but complex models also raise risk. They may over-fit the training sample, and hence do poorly in predicting the response value of new units. This trade-off is central to our prediction (model selection) problem.

2.3 K Normal means

It will be convenient to develop a canonical formulation of our prediction problem. Specifically we will show that our problem, after suitable transformation, coincides with the so called K Normal Means Problem. This is a classic problem in frequentist decision theory. Wasserman (2006, Chapter 7) and Lehmann and Casella (1998, Chapter 5) provide textbook treatments. Efron and Morris (1977) provide an account of the problem for the popular press.

We begin by adding a distributional assumption. Specifically we will assume that the conditional distribution of Y given X is Gaussian:

$$Y = m(X) + \sigma U, \;\; U|\, X \sim \mathcal{N}(0,1) \tag{2.3.1}$$

with σ known. As noted above, our interest centers on $m(X)$, which coincides with the mean regression function. Our Gaussian assumption is restrictive.

We will also let $m(x)$ have the finite basis function representation

$$m(x) = \sum_{k=1}^{K} \alpha_k g_k(x) \tag{2.3.2}$$

with $g_1(x), g_2(x), \ldots, g_K(x)$ a sequence of known basis functions. If all the components of X are discretely-valued then these basis functions may consist of a set of indicator functions for each support point of X. If X is a scalar, continuously-valued, variable, then the basis functions might be the first K polynomials in x:

$$g_k(x) = x^{k-1}, \;\; k = 1, \ldots, K.$$

In many applications the basis functions will consist of combinations of dummy variables, interactions and polynomials. Although we will allow K to be large, we will assume its dimension is smaller than the sample size N. In some situations condition (2.3.2) may only hold approximately, although the approximation will be quite good when $m(x)$ is smooth and/or K large enough.

In general the basis functions in (2.3.2) will be correlated with one another (e.g., x^2 is increasing in x). In what follows it will be very convenient to work instead with an orthogonal basis function representation.

Gram-Schmidt orthogonalization

Recall that the shortest route from a point to a line is given by the segment perpendicular or orthogonal to it. This fact can be generalized to solve a wide range of optimization problems within econometrics (as well as other areas of economics). A classic introduction to optimization by vector space methods is given by Luenberger (1969, Chapters 2 - 4). A less technical introduction is provided by Tsiatis (2006, Chapter 2). The key idea is to use **orthogonal projection** within a complete normed linear **vector space**. Here I use this basic insight to construct a set of basis functions that are orthogonal with respect to a set of irregularly spaced design points.

Let $\mathbf{X} = (X_1, \ldots, X_N)'$ be the set of N design points. Let $F_N(x) = \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}(X_i \leq x)$

denote the empirical distribution function of X. Let $\mathbf{f}$ and $\mathbf{g}$ be two $N \times 1$ vectors of functions of x. Define the **inner product**

$$\langle \mathbf{f}, \mathbf{g} \rangle = \int f(x)\, g(x)\, \mathrm{d}F_N(x) = \frac{1}{N} \sum_{i=1}^{N} f(X_i)\, g(X_i)$$

with the associated *norm*

$$\|\mathbf{f}\| = \langle \mathbf{f}, \mathbf{f} \rangle^{1/2} = \left[\int f(x)^2\, \mathrm{d}F_N(x) \right]^{1/2} = \left[\frac{1}{N} \sum_{i=1}^{N} f(X_i)^2 \right]^{1/2}.$$

We say that the vectors $\mathbf{f}$ and $\mathbf{g}$ are orthogonal if their inner product is zero.

Write $\mathbf{g}_k = (g_k(X_1), \ldots, g_k(X_N))'$ as the N vector with i element $g_k(X_i)$ (as it appears in (2.3.2)). Our goal is to transform $(\mathbf{g}_1, \ldots, \mathbf{g}_K)$ such that, after transformation, each column in this matrix is orthogonal to all others. Our transformed vector of functions will consist of linear combinations of our original vectors. This means that any function which is well approximated by some linear combination of our original basis functions will also be well approximated by our transformed basis functions.

Define the orthogonal projection of $\mathbf{g}$ onto $\mathbf{f}$ as

$$\mathrm{proj}_{\mathbf{f}}(\mathbf{g}) = \frac{\langle \mathbf{f}, \mathbf{g} \rangle}{\langle \mathbf{f}, \mathbf{f} \rangle} \mathbf{f}. \tag{2.3.3}$$

Let $\mathbf{u} = \mathbf{g} - \mathrm{proj}_{\mathbf{f}}(\mathbf{g})$ denote the prediction error. Observe that

$$\begin{aligned} \langle \mathbf{u}, \mathbf{f} \rangle &= \langle \mathbf{g} - \mathrm{proj}_{\mathbf{f}}(\mathbf{g}), \mathbf{f} \rangle \\ &= \left\langle \mathbf{g} - \frac{\langle \mathbf{f}, \mathbf{g} \rangle}{\langle \mathbf{f}, \mathbf{f} \rangle} \mathbf{f}, \mathbf{f} \right\rangle \\ &= \langle \mathbf{g}, \mathbf{f} \rangle - \frac{\langle \mathbf{f}, \mathbf{g} \rangle}{\langle \mathbf{f}, \mathbf{f} \rangle} \langle \mathbf{f}, \mathbf{f} \rangle \\ &= 0 \end{aligned}$$

so that prediction error is orthogonal to $\mathbf{f}$.

We will construct our transformed set of basis functions recursively as follows:

$$\begin{aligned}
f_1(x) &= g_1(x) \\
f_2(x) &= g_2(x) - \frac{\langle \mathbf{g}_2, \mathbf{f}_1 \rangle}{\langle \mathbf{f}_1, \mathbf{f}_1 \rangle} f_1(x) \\
f_3(x) &= g_3(x) - \frac{\langle \mathbf{g}_3, \mathbf{f}_2 \rangle}{\langle \mathbf{f}_2, \mathbf{f}_2 \rangle} f_2(x) - \frac{\langle \mathbf{g}_3, \mathbf{f}_1 \rangle}{\langle \mathbf{f}_1, \mathbf{f}_1 \rangle} f_1(x) \\
&\vdots \\
f_K(x) &= g_K(x) - \sum_{k=1}^{K-1} \frac{\langle \mathbf{g}_K, \mathbf{f}_k \rangle}{\langle \mathbf{f}_k, \mathbf{f}_k \rangle} f_k(x).
\end{aligned}$$

It is simple to show that $\mathbf{f}_1$ and $\mathbf{f}_2$ are orthogonal:

$$\begin{aligned}
\langle \mathbf{f}_1, \mathbf{f}_2 \rangle &= \left\langle \mathbf{f}_1, \mathbf{g}_2 - \frac{\langle \mathbf{g}_2, \mathbf{f}_1 \rangle}{\langle \mathbf{f}_1, \mathbf{f}_1 \rangle} \mathbf{f}_1 \right\rangle \\
&= \langle \mathbf{f}_1, \mathbf{g}_2 \rangle - \frac{\langle \mathbf{g}_2, \mathbf{f}_1 \rangle}{\langle \mathbf{f}_1, \mathbf{f}_1 \rangle} \langle \mathbf{f}_1, \mathbf{f}_1 \rangle \\
&= \langle \mathbf{f}_1, \mathbf{g}_2 \rangle - \langle \mathbf{g}_2, \mathbf{f}_1 \rangle = 0.
\end{aligned}$$

Using orthogonality of $\mathbf{f}_1$ and $\mathbf{f}_2$ we can also show orthogonality of $\mathbf{f}_1$ and $\mathbf{f}_3$

$$\begin{aligned}
\langle \mathbf{f}_1, \mathbf{f}_3 \rangle &= \left\langle \mathbf{f}_1, \mathbf{g}_3 - \frac{\langle \mathbf{g}_3, \mathbf{f}_2 \rangle}{\langle \mathbf{f}_2, \mathbf{f}_2 \rangle} \mathbf{f}_2 - \frac{\langle \mathbf{g}_3, \mathbf{f}_1 \rangle}{\langle \mathbf{f}_1, \mathbf{f}_1 \rangle} \mathbf{f}_1 \right\rangle \\
&= \langle \mathbf{f}_1, \mathbf{g}_3 \rangle - \frac{\langle \mathbf{g}_3, \mathbf{f}_2 \rangle}{\langle \mathbf{f}_2, \mathbf{f}_2 \rangle} \langle \mathbf{f}_1, \mathbf{f}_2 \rangle - \frac{\langle \mathbf{g}_3, \mathbf{f}_1 \rangle}{\langle \mathbf{f}_1, \mathbf{f}_1 \rangle} \langle \mathbf{f}_1, \mathbf{f}_1 \rangle \\
&= \langle \mathbf{f}_1, \mathbf{g}_3 \rangle - 0 - \langle \mathbf{g}_3, \mathbf{f}_1 \rangle = 0.
\end{aligned}$$

To show that all elements in the above sequence are orthogonal to one another we can use an inductive argument.

For our purposes it will also be convenient to normalize the basis so that each element has length one. This results in the orthonormal set of basis functions

$$\begin{aligned}
\phi_1(x) &= \frac{f_1(x)}{\|\mathbf{f}_1\|} \\
&\vdots \\
\phi_K(x) &= \frac{f_K(x)}{\|\mathbf{f}_K\|}.
\end{aligned}$$

Our transformed set of functions $\phi_1(x), \ldots, \phi_K(x)$ constitute an **orthonormal system** with respect to $F_N(x)$ with the properties

$$\begin{aligned}\int \phi_j(x)\,\phi_k(x)\,\mathrm{d}F_N(x) &= 0,\ j \neq k \\ \int \phi_k(x)^2\,\mathrm{d}F_N(x) &= 1.\end{aligned}$$

The first condition implies, given our inner product definition, orthogonality of our functions with one another. The second condition implies that they are normalized to have length one. Hence the orthonormal nomenclature.

Maximum likelihood estimate as a normal random K vector

Because each $\phi_k(X)$ is a linear combination of our original set of basis functions we can re-write (2.3.2) as

$$m(x) = \sum_{k=1}^{K} \theta_k \phi_k(x). \tag{2.3.4}$$

If we define

$$w(X) = \begin{pmatrix} \phi_1(X) \\ \vdots \\ \phi_K(X) \end{pmatrix},\ \theta = \begin{pmatrix} \theta_1 \\ \vdots \\ \theta_K \end{pmatrix}$$

we can write $m(x)$ as the linear regression function $m(x) = w(x)'\theta$.

The maximum likelihood estimate (MLE) of θ equals the maximizer of the log-likelihood for $\mathbf{Y} = (Y_1, \ldots, Y_N)'$ given $\mathbf{X}$:

$$l(\mathbf{Y}|\,\mathbf{X};\theta) = -\frac{N}{2}\ln 2\pi - N\ln\sigma - \frac{1}{2\sigma^2}\sum_{i=1}^{N}\left(Y_i - w(X_i)'\theta\right)^2. \tag{2.3.5}$$

The form of the log-likelihood follows from the Gaussian assumption on $U_1, \ldots, U_N$. Maximizing (2.3.5) yields an MLE of

$$\hat{\theta}_{ML} = \left[\frac{1}{N}\sum_{i=1}^{N} W_i W_i'\right]^{-1} \times \left[\frac{1}{N}\sum_{i=1}^{N} W_i Y_i\right]$$

with $W_i = w(X_i)$. This estimate is also the ordinary least squares (OLS) estimate. We can then form an estimate of, invoking our targeting example, the consumption response function as $\hat{m}(x) = w(x)'\hat{\theta}_{ML}$.

One implication of the orthonormal basis is that

$$\int w(x)\,w(x)'\,\mathrm{d}F_N(x) = \frac{1}{N}\sum_{i=1}^{N} W_i W_i' \quad = \quad I_K. \tag{2.3.6}$$

We use this observation to express the MLE of θ in the simplified form

$$\hat{\theta}_{ML} = \mathbf{Z} \overset{def}{=} \frac{1}{N}\sum_{i=1}^{N} W_i Y_i. \tag{2.3.7}$$

Using the notation $\mathbf{Z}$ to represent an estimate of θ may appear odd, but doing so highlights the connection between our concrete prediction problem and the (more abstract) canonical K Normal means problem. In what follows we will compare this benchmark estimate with alternatives.

Observe that $\mathbf{Z}$ is a **conditionally unbiased** estimate of θ:

$$\begin{aligned}
\mathbb{E}\left[\mathbf{Z}\middle|\,\mathbf{X}\right] &= \frac{1}{N}\sum_{i=1}^{N} W_i \mathbb{E}\left[Y_i\middle|\,\mathbf{X}\right] \\
&= \frac{1}{N}\sum_{i=1}^{N} W_i W_i'\theta + \frac{\sigma}{N}\sum_{i=1}^{N} W_i \mathbb{E}\left[U_i\middle|\,\mathbf{X}\right] \\
&= \theta.
\end{aligned}$$

The conditional sampling variance of its k^{th} element, using independence of Y_i and Y_j for $i \neq j$, equals

$$\begin{aligned}
\mathbb{V}\left(Z_k\middle|\,\mathbf{X}\right) &= \mathbb{V}\left(\frac{1}{N}\sum_{i=1}^{N} \phi_k(X_i)\,Y_i\middle|\,\mathbf{X}\right) \\
&= \frac{1}{N^2}\sum_{i=1}^{N} \mathbb{V}\left(\phi_k(X_i)\,Y_i\middle|\,\mathbf{X}\right) \\
&= \frac{\sigma^2}{N^2}\sum_{i=1}^{N} \phi_k(X_i)^2 \\
&= \frac{\sigma^2}{N},
\end{aligned}$$

where the last line follows from the equality $\frac{1}{N}\sum_{i=1}^{N}\phi_k(X_i)^2 = 1$.

Finally each element of $\mathbf{Z}$ is (conditionally) uncorrelated with the others:

$$\begin{aligned}\mathbb{C}\left(Z_j, Z_k\middle|\,\mathbf{X}\right) &= \mathbb{C}\left(\frac{1}{N}\sum_{i=1}^{N}\phi_j\left(X_i\right)Y_i, \frac{1}{N}\sum_{i=1}^{N}\phi_k\left(X_i\right)Y_i\middle|\,\mathbf{X}\right)\\ &= \frac{\sigma^2}{N^2}\sum_{i=1}^{N}\phi_j\left(X_i\right)\phi_k\left(X_i\right)\\ &= 0,\end{aligned}$$

where the last line follows from the fact that $\frac{1}{N}\sum_{i=1}^{N}\phi_j\left(X_i\right)\phi_k\left(X_i\right) = 0$ for $j \neq k$.

Finally, noting that $\mathbf{Z}$ is a linear combination of normal random variables, we get

$$\mathbf{Z}|\,\mathbf{X} \sim \mathcal{N}\left(\theta, \frac{\sigma^2}{N}I_K\right). \tag{2.3.8}$$

The conditional distribution of our MLE takes a convenient form. Our MLE estimate has a number of desirable properties. First, it is centered at the target parameter in the sense that $\mathbb{E}\left[\mathbf{Z}|\,\mathbf{X}\right] = \theta$. We say that $\mathbf{Z}$ is a conditionally unbiased estimate of θ. Second, its conditional sampling variability is inversely proportional to the sample size $\mathbb{V}\left(\mathbf{Z}|\,\mathbf{X}\right) = \sigma^2/N$.

The MLE of $m\left(x\right)$ is give by $\hat{m}\left(x\right) = w\left(x\right)'\mathbf{Z}$ with conditional distribution

$$\hat{m}\left(x\right)|\,\mathbf{X} \sim \mathcal{N}\left(m\left(x\right), \frac{\sigma^2}{N}w\left(x\right)w\left(x\right)'\right).$$

Later on we will develop the large sample properties of $\mathbf{Z}$ and show how they can be used to conduct approximate inference on θ. Initially, however, we will consider its finite sample properties. How good is $\mathbf{Z}$, as an estimate of θ, across repeated finite samples of size N?

2.4 Risk of MLE

Recall that our hypothetical decision maker's goal is estimate $m\left(x\right)$ accurately at each of the N design points in our training sample. Define $\mathbf{W} = \left(W_1, \ldots, W_N\right)'$ so that $\hat{\mathbf{m}} = \mathbf{W}'\mathbf{Z}$ equals our N vector of MLEs of $\mathbf{m} = \left(m\left(X_1\right), \ldots, m\left(X_N\right)\right)'$. We continue to assume that the loss associated with $\hat{\mathbf{m}}$ deviating from $\mathbf{m}$ equals the squared Euclidian distance between them. Risk equals the average loss associated with our procedure across the distribution of possible training samples. Recall that across each of these repeated samples we will keep the configuration of our design points fixed. From hereon I will keep this conditioning on $\mathbf{X}$ implicit, but all expectations, variances and covariances in what immediately follows should be understood as conditional on $\mathbf{X}$.

Manipulating our loss function yields:

$$\begin{aligned} \|\hat{\mathbf{m}} - \mathbf{m}\|^2 &= \frac{1}{N}\sum_{i=1}^{N}\left(W_i'\left(\mathbf{Z}-\theta\right)\right)^2 \\ &= \frac{1}{N}\sum_{i=1}^{N}\left(\sum_{k=1}^{K}\phi_k\left(X_i\right)\left(Z_k-\theta_k\right)\right)^2 \\ &= \sum_{k=1}^{K}\left(Z_k-\theta_k\right)^2. \end{aligned}$$

The last equality follows from the orthonormality of our basis with respect to $F_N(x)$. Squared error loss for our target $N \times 1$ vector $\mathbf{m}$ corresponds to square error loss for θ. We may therefore, without loss of generality, work with the loss function

$$L\left(\mathbf{Z},\theta\right) = \|\mathbf{Z}-\theta\|^2 = \sum_{k=1}^{K}\left(Z_k-\theta_k\right)^2. \tag{2.4.1}$$

Risk equals the average loss (across repeated samples) associated with using $\mathbf{Z}$ as an estimate of θ:

$$R\left(\mathbf{Z},\theta\right) = \mathbb{E}\left[L\left(Z,\theta\right)\right]. \tag{2.4.2}$$

The expectation in (2.2.2) is over $\mathbf{Z}$, the random MLE (θ is held fixed). The MLE is random because $\mathbf{Y}$ is. The risk function associated with the MLE is constant in θ and equals

$$\mathbb{E}\left[\sum_{k=1}^{K}\left(Z_k-\theta_k\right)^2\right] = \frac{K}{N}\sigma^2. \tag{2.4.3}$$

The risk of $\mathbf{Z} = \hat{\theta}_{ML}$ depends on the degree of model complexity, as indexed by the order of our basis function representation, K, relative to our sample size, N.

2.5 James-Stein type estimators

The MLE is a member of the family of linear estimators

$$\mathcal{L} = \left\{C\mathbf{Z} \,:\, C = \operatorname{diag}\left\{c_1,\ldots,c_K\right\},\, c_k \in [0,1],\, k=1,\ldots,K\right\}.$$

The risk associated with a generic member of this family, say $C\mathbf{Z}$, is

$$\begin{aligned}\mathbb{E}\left[\sum_{k=1}^{K}\left(c_k Z_k-\theta_k\right)^2\right] &= \mathbb{E}\left[\sum_{k=1}^{K}\left(c_k\left(Z_k-\theta_k\right)-\left(1-c_k\right)\theta_k\right)^2\right] \\ &= \frac{\sigma^2}{N}\sum_{k=1}^{K}c_k^2+\sum_{k=1}^{K}\left(1-c_k\right)^2\theta_k^2. \qquad (2.5.1)\end{aligned}$$

To find the risk-minimizing member of this class of estimators we minimize (2.5.1) with respect to $c_1,\ldots,c_K$. This yields

$$c_k^* = \frac{\theta_k^2}{\frac{\sigma^2}{N}+\theta_k^2},\ k=1,\ldots,K \qquad (2.5.2)$$

Evaluating the risk of the optimal estimator yields the lower bound or oracle inequality:

$$\begin{aligned}\inf_{\hat{\theta}\in\mathcal{L}} R\left(\hat{\theta},\theta\right) &= \frac{\sigma^2}{N}\sum_{k=1}^{K}\left(\frac{\theta_k^2}{\frac{\sigma^2}{N}+\theta_k^2}\right)^2+\sum_{k=1}^{K}\left(1-\frac{\theta_k^2}{\frac{\sigma^2}{N}+\theta_k^2}\right)^2\theta_k^2 \\ &= \frac{\sigma^2}{N}\left(\sum_{k=1}^{K}\frac{\theta_k^2}{\frac{\sigma^2}{N}+\theta_k^2}\right). \qquad (2.5.3)\end{aligned}$$

The risk of the MLE clearly exceeds (2.5.3) for all $\theta\in\Theta$. This raises the question of whether a feasible estimator which uniformly improves upon the MLE is available. Since constructing c_k^* requires knowledge of θ_k, directly using the "oracle weights" defined in (2.5.2) is non-operational. Nevertheless it may be possible to construct an estimate with risk properties close to that of the oracle estimator.

In a brilliant paper Stein (1981) proved a more general version of the following result.[1]

Theorem 2.5.1. *(Stein's Unbiased Risk Estimate, SURE). Let $\mathbf{Z}\sim\mathcal{N}\left(\theta,\sigma^2 I_K\right)$, $\hat{\theta}=\hat{\theta}\left(\mathbf{Z}\right)$ be an estimate of θ and $g\left(\mathbf{Z}\right)=\hat{\theta}-\mathbf{Z}$. Define Stein's Unbiased Risk Estimate (SURE) as*

$$\hat{R}_{\text{SURE}}\left(\mathbf{Z}\right)=K\sigma^2+2\sigma^2\sum_{k=1}^{K}\frac{\partial g_k\left(\mathbf{Z}\right)}{\partial Z_k}+\sum_{k=1}^{K}\left(\hat{\theta}_k-Z_k\right)^2. \qquad (2.5.4)$$

[1] Charles Stein began his academic career at Berkeley. He was hired by Neyman on the basis of his doctoral dissertation, the topic of which was suggested to him by Ken Arrow when they both worked in the meteorology division of the Pentagon during WW II. When Stein arrived at Berkeley there was shortage of office space and he was forced to share an office with Evelyn Fix, Joseph Hodges and Erich Lehmann (cf., Degroot, 1986; Lehmann, 2008). Theorem 2.5.1 actually predates the 1981 paper by at least a decade. Stein delivered a paper with the result in a symposium in Prague organized by Jaroslav Hájek in 1973 (see Stein (1973)).

If $g(\mathbf{Z})$ is weakly differentiable, then

$$\mathbb{E}\left[\hat{R}_{\text{SURE}}(\mathbf{Z})\right] = E\left[\left\|\hat{\theta}-\theta\right\|^2\right].$$

The proof of Theorem 2.5.1 is left as an exercise, but it is helpful to connect Stein's representation of risk with our earlier apparent error calculation (Equation (2.2.2)). Analogous calculations applied to the K Normal means problem yield as risk expression of

$$\mathbb{E}\left[\left\|\hat{\theta}-\theta\right\|^2\right] = -K\sigma^2 + 2\sum_{k=1}^{K}\mathbb{C}\left(Z_k,\hat{\theta}_k\right) + \mathbb{E}\left[\left\|\hat{\theta}-\mathbf{Z}\right\|^2\right].$$

Expanding the covariance term in the expression above we have that

$$\begin{aligned}\mathbb{C}\left(Z_k,\hat{\theta}_k\right) &= \mathbb{E}\left[\hat{\theta}_k\left(Z_k-\theta_k\right)\right]\\ &= \mathbb{E}\left[\left(\hat{\theta}_k - Z_k\right)\left(Z_k-\theta_k\right)\right] + \mathbb{E}\left[\left(Z_k-\theta_k\right)^2\right]\\ &= \mathbb{E}\left[g_k(\mathbf{Z})\left(Z_k-\theta_k\right)\right] + \sigma^2\end{aligned}$$

Stein's Theorem therefore follows if

$$\mathbb{E}\left[g_k(\mathbf{Z})\left(Z_k-\theta_k\right)\right] = \sigma^2\mathbb{E}\left[\frac{\partial\theta_k(\mathbf{Z})}{\partial Z_k}\right]$$

for $k = 1,\ldots,K$. A simple integration by parts argument can be used to establish this equality (which is known as Stein's Lemma).

Theorem 2.5.1 has many uses, one being it facilitates risk comparisons of different estimators. Another is for model section (see the Exercises). A large literature in theoretical and applied statistics uses Theorem 2.5.1 in these ways.

Here I will use Stein's result to show that the MLE of θ is **inadmissible**. By inadmissible I mean there exist feasible estimators with risk lower than that of the MLE at all possible values of the parameter θ. This is a surprising result, after all the MLE is the usual estimator. The demonstration will proceed by using Theorem 2.5.1 to evaluate the risk of an alternative estimator with risk uniformly lower that $\frac{K}{N}\sigma^2$.

I will consider the following estimator proposed by James and Stein (1961) at the Fourth Berkeley Symposium on Mathematical Statistics*:*

$$\hat{\theta}_{\text{JS}}(\mathbf{Z}) = \left(1 - \frac{(K-2)}{\mathbf{Z}'\mathbf{Z}}\frac{\sigma^2}{N}\right)\mathbf{Z}. \qquad (2.5.5)$$

This estimator shrinks the components of the MLE toward zero.

Note that $g_{\mathrm{JS}}(\mathbf{Z}) = -\left(\frac{(K-2)\frac{\sigma^2}{N}}{\mathbf{Z}'\mathbf{Z}}\right)\mathbf{Z}$ and hence that

$$\begin{aligned}\sum_{k=1}^{K}\frac{\partial g_k(\mathbf{Z})}{\partial Z_k} &= - \frac{(K-2)\frac{\sigma^2}{N}}{\mathbf{Z}'\mathbf{Z}}\sum_{k=1}^{K}\left(1-\frac{2Z_k^2}{\mathbf{Z}'\mathbf{Z}}\right) \\ &= -\frac{(K-2)^2\frac{\sigma^2}{N}}{\mathbf{Z}'\mathbf{Z}}\end{aligned}$$

for $k=1,\ldots,K$. Stein's unbiased estimate of the risk of $\hat{\theta}_{\mathrm{JS}}(\mathbf{Z})$, using Theorem 2.5.1, therefore equals

$$\begin{aligned}\hat{R}_{\mathrm{SURE}}(\mathbf{Z}) &= \frac{K}{N}\sigma^2 + \frac{2}{N}\sigma^2\sum_{k=1}^{K}\frac{\partial g_k(\mathbf{Z})}{\partial Z_k} + \sum_{k=1}^{K}\left(\hat{\theta}_k - Z_k\right)^2 \\ &= \frac{K}{N}\sigma^2 - \frac{2\sigma^2}{N}\frac{(K-2)^2\frac{\sigma^2}{N}}{\mathbf{Z}'\mathbf{Z}} + \frac{(K-2)^2\left(\frac{\sigma^2}{N}\right)^2}{(\mathbf{Z}'\mathbf{Z})^2}\sum_{k=1}^{K}Z_k^2 \\ &= \frac{K}{N}\sigma^2 - 2\left(\frac{\sigma^2}{N}\right)^2\frac{(K-2)^2}{\mathbf{Z}'\mathbf{Z}} + \left(\frac{\sigma^2}{N}\right)^2\frac{(K-2)^2}{\mathbf{Z}'\mathbf{Z}} \\ &= \frac{K}{N}\sigma^2 - \frac{(K-2)^2}{\mathbf{Z}'\mathbf{Z}}\frac{\sigma^4}{N^2}. \end{aligned} \tag{2.5.6}$$

Theorem 2.5.6 implies that we can compute the actual risk of the James-Stein estimator by computing the expectation of (2.5.6):

$$R\left(\hat{\theta}_{\mathrm{JS}},\theta\right) = \mathbb{E}\left[\hat{R}_{\mathrm{SURE}}(\mathbf{Z})\right] = \frac{K}{N}\sigma^2 - (K-2)^2\frac{\sigma^4}{N^2}\mathbb{E}\left[\frac{1}{\mathbf{Z}'\mathbf{Z}}\right].$$

Observe that $Z_k^2 = \frac{\sigma^2}{N}\left(\frac{\theta_k}{\sigma/\sqrt{N}} + U\right)^2$ with $U \sim \mathcal{N}(0,1)$ and hence $\mathbf{Z}'\mathbf{Z} \sim \frac{\sigma^2}{N}V$ where V is a *non-central* χ^2 random variable with K degrees of freedom and non-centrality parameter $\rho = \sum_{k=1}^{K} N\left(\theta_k^2/\sigma^2\right)$. Using a result in Johnson, Kotz and Balakrishnan (1995) we can represent V as a mixture of a *central* χ^2 and Poisson random variable. Specifically

$$V \sim \chi^2_{K+2W},\ W \sim \text{Poisson}\left(\frac{\rho}{2}\right).$$

Also recall that for $K > 2$ and $S \sim \chi^2_K$ we have $\mathbb{E}\left[\frac{1}{S}\right] = \frac{1}{K-2}$. Using this fact, the law-of-iterated expectations, and Jensen's Inequality we get

$$\begin{aligned}\mathbb{E}\left[\frac{1}{\mathbf{Z}'\mathbf{Z}}\right] &= \frac{N}{\sigma^2}\mathbb{E}\left[\frac{1}{V}\right] \\ &= \frac{N}{\sigma^2}\mathbb{E}\left[\mathbb{E}\left[\frac{1}{V}\middle| W\right]\right] \\ &= \frac{N}{\sigma^2}\mathbb{E}\left[\frac{1}{K-2+2W}\right] \\ &\geq \frac{N}{\sigma^2}\frac{1}{K-2+N\sum_{k=1}^{K}\left(\theta_k^2/\sigma^2\right)} \\ &= \frac{1}{(K-2)\frac{\sigma^2}{N}+\|\theta\|^2}. \qquad (2.5.7)\end{aligned}$$

Using this inequality we get a bound on the risk of $\hat{\theta}_{JS}$ equal to

$$\begin{aligned}R\left(\hat{\theta}_{\mathrm{JS}},\theta\right) &\leq \frac{K}{N}\sigma^2 - \frac{(K-2)^2\frac{\sigma^4}{N^2}}{(K-2)\frac{\sigma^2}{N}+\|\theta\|^2} \\ &= \frac{2}{N}\sigma^2 + (K-2)\frac{\sigma^2}{N} - \frac{(K-2)^2\frac{\sigma^4}{N^2}}{(K-2)\frac{\sigma^2}{N}+\|\theta\|^2} \\ &= \frac{2}{N}\sigma^2 + (K-2)\frac{\sigma^2}{N}\left\{1-\frac{(K-2)\frac{\sigma^2}{N}}{(K-2)\frac{\sigma^2}{N}+\|\theta\|^2}\right\} \\ &= \frac{2}{N}\sigma^2 + \frac{(K-2)\frac{\sigma^2}{N}\|\theta\|^2}{(K-2)\frac{\sigma^2}{N}+\|\theta\|^2}.\end{aligned}$$

Note that

$$R\left(\hat{\theta}_{\mathrm{ML}},\theta\right) - R\left(\hat{\theta}_{\mathrm{JS}},\theta\right) = \left(\frac{(K-2)^2\frac{\sigma^4}{N^2}}{(K-2)\frac{\sigma^2}{N}+\|\theta\|^2}\right) > 0$$

for all $\theta \in \Theta$. The econometrician is never worse off, under squared error loss, when she uses $\hat{\theta}_{\mathrm{JS}}$ instead of $\hat{\theta}_{\mathrm{ML}}$. When the risk properties of an estimator are uniformly dominated by those of an alternative across all possible data generating processes we say that the dominated estimator is **inadmissible**. An implication of James and Stein (1961) is that least squares is inadmissible under squared error loss.

The James-Stein estimator is a member of subset of $\mathcal{L}$ where $C = cI_K$ for some $c \in [0,1]$. It can be shown that its risk properties are close to those of the oracle estimator within this restricted class (e.g., Wasserman, 2006; Theorem 7.47), although it turns out that the James-Stein estimator is itself inadmissible.

James and Stein (1961) is landmark paper. Stein's (1981) later paper is also of enduring importance for research on model selection (e.g., Efron, 2004). This is a topic of particular interest in econometrics and statistics today given the growing prevalence of estimation problems with large numbers of covariates/features. At the same time we should be careful about interpreting the implications of the inadmissibility result for empirical practice.

One useful application of SURE is due to Efromovich (1999) (see also Beran (2000)). Efromovich's (1999) focus was on nonparametric curve estimation, but his methods apply to any conditional mean estimation problem with a basis function representation. His estimator attempts to mimic the oracle estimator defined in (2.5.2).

Consider SURE for a generic estimator in the class $\mathcal{L}$:

$$\begin{aligned}
\hat{R}_{\text{SURE}}\left(\mathbf{Z},\mathbf{c}\right) &= \frac{K}{N}\sigma^2 - \frac{2}{N}\sigma^2\sum_{k=1}^{K}\left(1-c_k\right) + \sum_{k=1}^{K} Z_k^2\left(1-c_k\right)^2 \\
&= \frac{K}{N}\sigma^2 - \frac{2}{N}\sigma^2\sum_{k=1}^{K}\left(1-c_k\right) + \frac{\sigma^2}{N}\sum_{k=1}^{K}\left(1-c_k\right)^2 + \sum_{k=1}^{K}\left(Z_k^2 - \frac{\sigma^2}{N}\right)\left(1-c_k\right)^2 \\
&= \frac{\sigma^2}{N}\left[K - 2\sum_{k=1}^{K}\left(1-c_k\right) + \sum_{k=1}^{K}\left(1-c_k\right)^2\right] + \sum_{k=1}^{K}\left(Z_k^2 - \frac{\sigma^2}{N}\right)\left(1-c_k\right)^2 \\
&= \frac{\sigma^2}{N}\sum_{k=1}^{K} c_k^2 + \sum_{k=1}^{K}\left(Z_k^2 - \frac{\sigma^2}{N}\right)\left(1-c_k\right)^2.
\end{aligned}$$

Minimizing this expression with respect to $c_1,\ldots,c_K$ yields

$$\hat{c}_k = 1 - \frac{N^{-1}\sigma^2}{Z_k^2},\ k = 1,\ldots,K. \tag{2.5.8}$$

Let $\hat{C} = \operatorname{diag}\left\{\hat{c}_1,\ldots,\hat{c}_K\right\}$. The estimator $\hat{\theta}_{EF} = \hat{C}\mathbf{Z}$ is closely related to the universal series estimator in Efromovich (1999, p. 125).

Note that to characterize the oracle estimator I chose $c_1,\ldots,c_K$ to minimize actual risk. This is non-operational in the real world since θ in unknown. However using SURE we can choose $c_1,\ldots,c_K$ to minimize an unbiased estimate of risk. This is operational.

2.6 Additional reading

Wasserman (2006, Chapters 7 - 9) provides an accessible introduction to the K Normal Means problem and applications. Lehmann and Casella (1998, Chapter 5) is a more theoretical treatment. Efromovich (1999) is a monograph-length, and highly practical, discussion of nonparametric curve estimation using series methods and empirical risk minimization procedures.

In Machine Learning the K Normal Means problem is an example of so-called **supervised learning**. Prediction and classification problems with large covariate, or feature, vectors arise frequently in Machine Learning. Empirical risk minimization methods may be used to select which features to retain and which to discard and, more generally, to reduce model complexity. Murphy (2012, Chapters 7, 13, 16) provides an elementary textbook treatment. Efron (2004) provides a nice synoptic discussion of model selection with many interesting examples.

Admissibility provides a very weak ranking of estimation procedures. Different procedures may dominate in different parts of the parameter space. One approach to ranking procedures is to consider their minimax risk. Let Θ be a subset of $\mathbb{R}^K$, the **minimax** risk of $\hat{\theta}$ over Θ is

$$R(\Theta) = \inf_{\hat{\theta}} \sup_{\theta \in \Theta} R\left(\hat{\theta}, \theta\right).$$

In words: compute the maximal risk for each estimator and choose one with minimal maximal risk. For connections between minimax theory and the K Normal Means problem see Wasserman (2006) and Lehmann and Casella (1998).

The soft threshold estimator introduced in the exercise 3 is a special case of Tibshirani's (1996) LASSO. Hansen (2015) compares James-Stein and LASSO shrinkage.

2.7 Exercises

1. Let $\mathbf{Z} \sim \mathcal{N}\left(\theta, \frac{\sigma^2}{N} I_K\right)$. Show that $\mathbb{E}\left[g(\mathbf{Z})(\mathbf{Z} - \theta)\right] = \frac{\sigma^2}{N}\mathbb{E}\left[\nabla_{\mathbf{Z}} g(\mathbf{Z})\right]$ (<u>HINT</u>: Use integration by parts).

2. Consider the estimate of the Risk associated with the weakly differentiable estimate $\hat{\theta}$ introduced in Theorem 2.5.1 of the main text:

$$\hat{R}(\mathbf{Z}) = K\sigma^2 + 2\sigma^2 \sum_{k=1}^{K} \frac{\partial g(\mathbf{Z})}{\partial Z_k} + \sum_{k=1}^{K} \left(\hat{\theta}_k - Z_k\right)^2.$$

   Prove that this risk estimate is unbiased under square error loss: $\mathbb{E}_\theta\left[\hat{R}(\mathbf{Z})\right] = R\left(\hat{\theta}, \theta\right)$.

3. (Wasserman, 2006) Let $\mathbf{Z} \sim \mathcal{N}\left(\theta, \frac{\sigma^2}{N} I_K\right)$ and consider the following soft threshold estimate of θ :

$$\hat{\theta}_k = \operatorname{sgn}(Z_k)\left(|Z_k| - \lambda\right)_+, \; k = 1, \ldots, K.$$

   In words this estimator shrinks the MLE of θ_k toward zero when it is large (in absolute value) and shrinks it exactly to zero when it is small (in absolute value). Use

Theorem 2.5.1 to show that

$$\hat{R}_{\text{SURE}}\left(\mathbf{Z},\lambda\right)=\frac{K}{N}\sigma^2-\frac{2\sigma^2}{N}\sum_{k=1}^{K}\mathbf{1}\left(|Z_k|\leq\lambda\right)+\sum_{k=1}^{K}\min\left(Z_k^2,\lambda^2\right).$$

Provide a concrete prediction problem where you would expect the risk properties of the soft threshold estimator to be attractive.

4. (Wasserman, 2006) Let $\mathbf{Z}\sim\mathcal{N}\left(\theta,\frac{\sigma^2}{N}I_K\right)$ and $\mathcal{M}$ be the class of ordered subsets

$$\left\{\emptyset,\left\{1\right\},\left\{1,2\right\},\left\{1,2,3\right\},\ldots,\left\{1,\ldots,K\right\}\right\}$$

and consider the estimator $\hat{\theta}_k\left(M\right)=Z_k\mathbf{1}\left(k\in M\right)$ for $k=1,\ldots,K$. Use Theorem 2.5.1 to show that

$$\hat{R}_{\text{SURE}}\left(\mathbf{Z},M\right)=\frac{\sigma^2}{N}\left|M\right|+\sum_{k\in M^c}\left(Z_k^2-\frac{\sigma^2}{N}\right)$$

with $|M|$ denoting the cardinality of M and M^c the absolute complement of M in the universe $\mathcal{M}$.

5. (Efromovich, 1999) Consider the cosine orthonormal system of $[0,1]$ with elements

$$\phi_1\left(x\right)=1,\ \phi_k\left(x\right)=\sqrt{2}\cos\left(\pi\left(k-1\right)x\right),\ k=2,\ldots.$$

Let $\mathcal{N}\left(x;\mu,\sigma\right)$ denote the normal density with mean μ and standard deviation σ at x. Consider the following "bimodal" and "steps" test functions introduced by Efromovich (1999, p. 18):

$$\begin{aligned} m_{\text{B}}\left(x\right) &= \frac{1}{2}\mathcal{N}\left(x;0.4,0.12\right)+\frac{1}{2}\mathcal{N}\left(x;0.7,0.08\right)\\ m_{\text{S}}\left(x\right) &= 0.6\times\mathbf{1}\left(0\leq x<\frac{1}{3}\right)+0.9\times\mathbf{1}\left(\frac{1}{3}\leq x<\frac{3}{4}\right)+\frac{204}{120}\times\mathbf{1}\left(\frac{3}{4}\leq x\leq 1\right). \end{aligned}$$

   (a) Let $X_i=i/\left(N+1\right)$ for $i=1,\ldots,N$ and $Y_i=m_{\text{j}}\left(x\right)+\sigma U_i$ with U_i a standard normal random variable and $j=\text{B},\text{S}$. For each of the two functions generate 1000 random samples of size $N=1000$ with $\sigma=1/2$. Set $K=20$. For each sample compute $Z_k=\frac{1}{N}\sum_{i=1}^{N}\phi_k\left(X_i\right)Y_i$ for $k=1,\ldots,K$ and construct two estimates of $\mathbf{m}$: (i) the MLE with $\hat{m}_{\text{ML}}\left(x\right)=\sum_{k=1}^{N}\phi_k\left(x\right)Z_k$ and (ii) Efromovich's with $\hat{m}_{\text{EF}}\left(x\right)=\sum_{k=1}^{N}\hat{c}_k\phi_k\left(x\right)Z_k$ with $\hat{c}_k$ as defined in (2.5.8). Compute squared error loss for each sample, estimate and test function.

(b) Compute average loss across your 1000 random samples for each test function and estimator and report the results in a Table. For each test function and estimator find the sample with median loss (across the 1000 simulated samples). For this sample plot the true test function, your estimate and a scatter of the data (four figures in all). Briefly discuss your findings.

(c) Using the same simulated samples from parts (a) and (b) compute the soft thresholding estimate of $\mathbf{m}$. For each sample choose λ to minimize the SURE expression from problem 3 above. Compute average loss across your 1000 random samples for each test function and report the results in a table. For each test function find the sample with median loss (across the 1000 simulated samples). For this sample plot the true test function, your estimate and a scatter of the data (two figures in all). Briefly discuss your findings.