Section 3*

Econ 240A - Second Half

Ingrid Haegele†
University of California, Berkeley

October 26, 2018

*These section notes rely on the notes prepared and revised by Markus Pelger and Raffaele Saggio.

†E-mail: inha@berkeley.edu

1 Prediction when the population is known

Prediction is gaining more and more attention from economists. There are many examples in economics that show the importance of prediction. As motivated in lecture, prediction can be an important tool for policy decision. The rising interest in machine learning is also contributing to the popularity of prediction. In this section we start by formally defining the prediction problem.

In a prediction problem, we have an outcome variable, Y, that we would like to predict using a set of variables, X. X is often defined as the set of covariates. Our first, and absolute key, assumption is that the researcher knows the joint distribution of (Y, X), denoted as $\Pr(Y \leq y, X \leq x) \equiv F_{Y,X}(y, x)$. We will relax this assumption in the next part that talks about prediction in cases when the population is unknown.

The researcher would like to minimize the cost associated with wrong predictions. Let g denote our prediction, then the loss function is defined as $L(y - g)$. We are interested in conditional predictions, which means that the decision function g is a function of x, $g(x)$. After observing $X = x$, the researcher computes the prediction rule $g(x)$. 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.

$$Risk(x) = \mathrm{E}[L(Y - g(x))|X = x] = \int L(y - g(x)) f_{Y|X}(y|x) dy \tag{1}$$

Notice that the objective above is computable given our initial assumption that the researcher knows the conditional distribution.

In addition, if we assume loss is proportional to the square of our prediction error (a convenient and historically important loss function), then, for $\mathrm{E}[Y^2] < \infty$, risk is equal to

mean square error (MSE):

$$\begin{aligned}
\mathrm{E}[L(Y-g(x))|X=x] &= \mathrm{E}[(Y-g(x))^2|X=x] \\
&= \mathrm{E}[(Y-E[Y|X=x]-g(x)-\mathrm{E}[Y|X=x])^2|X=x] \\
&= \mathrm{E}[(Y-E[Y|X=x])^2|X=x] + \mathrm{E}[(g(x)-\mathrm{E}[Y|X=x])^2|X=x] + \\
&+ 2\,\mathrm{E}[(Y-E[Y|X=x])(g(x)-\mathrm{E}[Y|X=x])|X=x] \\
&= \mathrm{Var}(Y|X=x) + (g(x)-\mathrm{E}[Y|X=x])^2
\end{aligned} \tag{2}$$

One can easily see that the expression above is going to be minimized at the CEF, i.e. $g^*(x) = \mathrm{E}[Y|X=x]$. Thus, under squared error loss the conditional expectation of Y given X=x is the optimal predictor.

This formula for risk enables us to understand the determinants of risk a bit better. As the last line of equation (2) shows, risk has two components: the first is the intrinsic variability of Y given X ($\mathrm{Var}(Y|X=x)$) and the second is the structural error ($(g(x) - \mathrm{E}[Y|X=x])^2|X=x]$).

2 Prediction when the population is unknown

We learned that under mean squared loss, the conditional expectation function (CEF) represents the best predictor for the outcome variable Y given knowledge of some covariates $\mathbf{X}=\mathbf{x}$. However, computation of the CEF requires knowledge of the joint distribution. Above, we assumed that the researcher knows the joint distribution. In practice this is not an realistic assumption. Instead all the researcher (i.e. the decision maker) has available is only a random sample (also called training sample) of N observations $\{X_i, Y_i\}_{i=1}^N$. The researcher's task is to use this training sample to predict the value of a new draw of Y, the response vector.

We will assume that $\mathbf{X} = (X_1^\top, \ldots, X_N^\top)$, i.e. the $N \times P$ vector of covariates (or design matrix), to be non-stochastic. This assumption implies that we are working under the assumption that if we were to obtain a *new* random sample of $\{X_i, Y_i\}_{i=1}^N$ this new random sample will have the same design matrix $\mathbf{X}$ as our original training sample. The only randomness comes from the new information collected in $\{Y_i\}_{i=1}^N$. This situation typically arises with stratified data, see the lecture notes for a specific example.

The researcher's goal is to construct a decision rule that will generate good predictions on average. A good measurement of the quality of predictions is the associated risk, as we discussed in the context of prediction with a known joint distribution.

The decision maker knows that, under squared loss, the (unfeasible) vector of optimal prediction is given by $\mathbf{m} = \mathrm{E}[\mathbf{Y}|\mathbf{X}]$. The question we are after is how to use the training sample is order to construct a good estimate of $\mathbf{m}$, call it $\hat{\mathbf{m}}$, that does well on average. Crucially, this average is computed over the universe of possible training samples the decision maker might have observed. Let $m(X_i)$ be the CEF evaluated at observation i, i.e. the CEF evaluated at the point X_i. Consequently, $\mathbf{m} = (m(X_1), \ldots, m(X_N))$. Also, for simplicity we are going to denote the ith coordinate of $\mathbf{m}$ as m_i.
We start by analyzing the within sample squared prediction error under squared loss[1]

$$\begin{aligned}
\mathrm{E}\,||\mathbf{Y} - \hat{\mathbf{m}}||^2 &= \mathrm{E}\left[\sum_{i=1}^{N}(Y_i - m_i - (\hat{m}_i - m_i))^2\right] \\
&= \mathrm{E}\left[\sum_{i=1}^{N}(Y_i - m_i)^2\right] + \mathrm{E}\left[\sum_{i=1}^{N}(\hat{m}_i - m_i)^2\right] - 2\,\mathrm{E}\left[\sum_{i=1}^{N}(\hat{m}_i - m_i)(Y_i - m_i)\right] \\
&= N\sigma^2 + \mathrm{E}\,||\hat{\mathbf{m}} - \mathbf{m}||^2 - 2\sum_{i=1}^{N}\mathrm{E}(\hat{m}_i - m_i)(Y_i - m_i) \\
&= N\sigma^2 + \mathrm{E}\,||\hat{\mathbf{m}} - \mathbf{m}||^2 - 2\sum_{i=1}^{N}\mathrm{Cov}(\hat{m}_i, Y_i) \\
&= N\sigma^2 + \mathrm{E}\,||\hat{\mathbf{m}} - \mathbf{m}||^2 - 2\sigma^2 df(\hat{\mathbf{m}})
\end{aligned} \tag{3}$$

where $df(\hat{\mathbf{m}}) = \sum_{i=1}^{N}\frac{\mathrm{Cov}(Y_i, \hat{m}_i)}{\sigma^2}$ is the degrees of freedom associated with the rule $\hat{\mathbf{m}}$. We can rearrange this equation to obtain an expression for the risk of our procedure as follows

$$\mathrm{E}\,||\hat{\mathbf{m}} - \mathbf{m}||^2 = \mathrm{E}\,||\mathbf{Y} - \hat{\mathbf{m}}||^2 - N\sigma^2 + 2\sigma^2 df(\hat{\mathbf{m}}). \tag{4}$$

- The risk of our estimator is increasing in expected training error, $\mathrm{E}\,||\mathbf{Y} - \hat{\mathbf{m}}||^2$.
- The risk of our estimator increases with model complexity which is represented by the degrees of freedom. Model complexity is going to increase with the number of regressors P. In this framework we think of P to be large but still not larger than N.

[1]Recall that for any $N \times 1$ vector $\mathbf{W}$, $||\mathbf{W}|| = (\sum_{i=1}^{N} w_i^2)^{1/2}$

- Consequently, we face a trade-off: complex models produce good sample fits, which lowers risk, but they also raise risk due to over-fitting.

3 K Normal Means

In the following part, we develop a canonical formulation of our prediction problem. The new set-up of our prediction problem closely corresponds to the so-called K Normal Means Problem.

3.1 Assumptions

To make the problem more tractable, we introduce the following two assumptions

Assumption 1: We can rewrite the outcome variable as

$$Y = m(X) + \sigma U \tag{5}$$

where $U|X \sim \mathcal{N}(0;1)$ and σ is known. Notice that decomposing a random variable in its CEF plus a CEF error is not an assumption, it can always be accomplished provided first moments exists. The assumption here is the parametric form of the CEF error which is assumed to be normal with a known variance.

Assumption 2: The CEF can be expressed as

$$m(x) = \sum_{k=1}^{K} \alpha_k g_k(x) \tag{6}$$

Notice that if X is discrete than this is not really an assumption as we can always write the CEF of Y given X as a linear expression of a set of indicator functions for each support point of X. If X is continuously distributed then we can think of the basis functions $\{g_k\}_{k=1}^{K}$ as polynomials in x of order K.

3.2 Normalization

Let

$$W(X_i) = \begin{pmatrix} g_1(X_i) \\ \vdots \\ g_K(X_i) \end{pmatrix} \tag{7}$$

$$\alpha = \begin{pmatrix} \alpha_1 \\ \vdots \\ \alpha_K \end{pmatrix} \tag{8}$$

that is $W(X_i)$ stacks the basis functions of a given observation into a $K \times 1$ vector[2]. This allows us to write $m(X_i)$ as a *linear* function $W(X_i)^\top \alpha$. This is great because we know how to estimate parameters from a linear model[3]. However remember that in our setting K is relatively large. Hence computing an inverse[4] can be extremely tedious, as each polynomial inside $W(X_i)$ is going to to be correlated with each other. Ideally, letting $\mathbf{W} = (W(X_1)^\top, \ldots, W(X_N)^\top)$, we want $\mathbf{W}^\top \mathbf{W}$ to be diagonal. This can be achieved using the Gram-Schmidt orthogonalization technique.

In order to define this technique, we need to properly define a vector space and embedded it with an inner product. Let $\mathbf{X} = (X_1, \ldots, X_N)$ be an $N \times 1$ vector of design points. Let $F_N(x) = \frac{1}{N} \sum_{i=1}^N \mathbf{1}(X \leq x)$ be the empirical cdf. For two $N \times 1$ vector of functions of x, $(\mathbf{f}, \mathbf{g})$ define the inner product as

$$\langle \mathbf{f}, \mathbf{g} \rangle = \int f(x) g(x) dF_N(x) = \frac{1}{N} \sum_{i=1}^N f(X_i) g(X_i) \tag{9}$$

You should check that such an inner product satisfies the typical three axioms for inner products (symmetry, linearity and positive definiteness). Armed with this definition, we define also the norm as follows

$$||\mathbf{f}|| = \sqrt{\langle \mathbf{f}, \mathbf{f} \rangle} \tag{10}$$

Recall that two vector are orthogonal if their inner product is zero. Now, define the orthogonal projection of $\mathbf{g}$ into $\mathbf{f}$ as follows

$$proj_{\mathbf{f}}(\mathbf{g}) = \frac{\langle \mathbf{f}, \mathbf{g} \rangle}{\langle \mathbf{f}, \mathbf{f} \rangle} \mathbf{f} \tag{11}$$

[2]To keep it simple, from now on I will consider the case where $P = 1$, it is straightforward to extend this to the case where $W(X_i)$is of dimension $KP \times 1$.

[3]We are interested in the coefficients because our goal is to find the best possible estimator for $\hat{m}$.

[4]Remember that computing such an inverse is required in order to compute the least squares fit

Then if we define $\mathbf{u} = \mathbf{g} - proj_{\mathbf{f}}(\mathbf{g})$ as the projection error, we can see that

$$\begin{aligned} \langle \mathbf{f}, \mathbf{u} \rangle &= \langle \mathbf{g} - proj_{\mathbf{f}}(\mathbf{g}), \mathbf{f} \rangle \\ &= \langle \mathbf{f}, \mathbf{g} \rangle - \frac{\langle \mathbf{f}, \mathbf{g} \rangle}{\langle \mathbf{f}, \mathbf{f} \rangle} \langle \mathbf{f}, \mathbf{f} \rangle \\ &= 0 \end{aligned} \tag{12}$$

Hence the projection error is orthogonal to $\mathbf{f}$. This idea is central to the Gram-Schmidt orthogonalization, which is an iterative procedure which we are going to use to transform our original basis functions into a new set of basis functions which are going to be orthogonal to each other. In order to do that, start by defining $\mathbf{g}_k = (g_k(X_1), \ldots, g_k(X_N))^\top$ as the $N \times 1$ vector collecting the value of the $k - th$ basis for our sample. The goal is to transform $\mathbf{g} = (\mathbf{g_1}, \ldots, \mathbf{g_K})$ into a new matrix in which each column is going to be orthogonal to each other. To do that, we start by defining

$$f_1(x) = g_1(x) \tag{13}$$

$$f_2(x) = g_2(x) - \frac{\langle \mathbf{g_2}, \mathbf{f_1} \rangle}{\langle \mathbf{f_1}, \mathbf{f_1} \rangle} f_1(x) \tag{14}$$

$$f_3(x) = g_3(x) - \frac{\langle \mathbf{g_3}, \mathbf{f_2} \rangle}{\langle \mathbf{f_2}, \mathbf{f_2} \rangle} f_2(x) - \frac{\langle \mathbf{g_3}, \mathbf{f_1} \rangle}{\langle \mathbf{f_1}, \mathbf{f_1} \rangle} f_1(x) \tag{15}$$

$$\vdots$$

$$f_K(x) = g_K(x) - \sum_{k=1}^{K-1} \frac{\langle \mathbf{g_K}, \mathbf{f_k} \rangle}{\langle \mathbf{f_k}, \mathbf{f_k} \rangle} f_k(x) \tag{16}$$

It is easy to show that $\mathbf{f_2}$ and $\mathbf{f_1}$ are orthogonal. Let us show orthogonality between $\mathbf{f_1}$ and $\mathbf{f_2}$ based on orthogonality between $\mathbf{f_1}$ and $\mathbf{f_2}$.

$$\begin{aligned} \langle \mathbf{f_1}, \mathbf{f_3} \rangle &= \langle \mathbf{f_1}, \mathbf{g_3} - \frac{\langle \mathbf{g_3}, \mathbf{f_2} \rangle}{\langle \mathbf{f_2}, \mathbf{f_2} \rangle} \mathbf{f}_2 - \frac{\langle \mathbf{g_3}, \mathbf{f_1} \rangle}{\langle \mathbf{f_1}, \mathbf{f_1} \rangle} \mathbf{f}_1 \rangle \\ (\mathbf{f_2} \perp \mathbf{f_1}) &= \langle \mathbf{f_1}, \mathbf{g_3} \rangle + 0 - \frac{\langle \mathbf{g_3}, \mathbf{f_1} \rangle}{\langle \mathbf{f_1}, \mathbf{f_1} \rangle} \langle \mathbf{f_1}, \mathbf{f_1} \rangle \\ &= 0 \end{aligned} \tag{17}$$

One can proceed by induction to complete the proof to show orthogonality between all the columns of $\mathbf{f} = (\mathbf{f_1}, \ldots, \mathbf{f_K})$.

A final convenient step is to further transform our newly constructed basis to have length one. This will allow us to obtain an *orthonormal* set of basis functions.

$$\phi_1(x) = \frac{f_1(x)}{||\mathbf{f}_1||} \tag{18}$$

$$\vdots$$

$$\phi_K(x) = \frac{f_K(x)}{||\mathbf{f}_K||} \tag{19}$$

Our new set of functions is therefore going to be orthonormal, meaning that for $j \neq k \in \{1, \ldots, K\}$ given our vector space and associated inner product and norm

$$\int \phi_j(x)\phi_k(x)dF_N(x) = 0 \tag{20}$$

$$\int \phi_k^2(x)dF_N(x) = 1 \tag{21}$$

3.3 Maximum Likelihood

Our newly constructed set of functions remains a linear combination of the original basis $\mathbf{g}$, which implies that we still have the key property that

$$m(x) = \sum_{k=1}^{K} \theta_k \phi_k(x) \tag{22}$$

Therefore using our Assumption 1, we can estimate $\theta = (\theta_1, \ldots, \theta_K)^\top$ as a simple MLE. Redefine the following object

$$W(X_i) \equiv W_i = \begin{pmatrix} \phi_1(X_i) \\ \vdots \\ \phi_K(X_i) \end{pmatrix} \tag{23}$$

then it follows, using what we have learned in the first half of the course, that

$$\begin{aligned} \hat{\theta}_{ML} &= \left[\frac{1}{N}\sum_{i=1}^{N} W_i W_i^\top\right]^{-1} \left[\frac{1}{N}\sum_{i=1}^{N} W_i Y_i\right] \\ \text{(by orthonormality)} &= \left[\frac{1}{N}\sum_{i=1}^{N} W_i Y_i\right] \\ &\equiv \mathbf{Z} \end{aligned} \tag{24}$$

Recall that in our setting $\mathbf{X}$ does not vary across target samples, hence the source of randomness in this estimator is only coming from Y.

Now that we defined $\mathbf{Z}$ as an estimator of θ, the question arises how good the estimate is. To evaluate this, we will examine this estimator based on the usual properties of interest.

First, we are interested in whether $\mathbf{Z}$ is a conditionally unbiased estimate of θ. Conditionally on $\mathbf{X}$, we can show that

$$\begin{aligned} \mathrm{E}[\mathbf{Z}|\mathbf{X}] = \frac{N}{N}\,\mathrm{E}(W_i Y_i|\mathbf{X}) &= \mathrm{E}(W_i m(X_i)) + \mathrm{E}(W_i U|\mathbf{X}) \\ \text{(Using Assumption 1+2)} &= \mathrm{E}(W_i W_i^\top \theta) \\ \text{(by orthonormality)} &= \theta \end{aligned} \tag{25}$$

Hence, the MLE is centered at the truth and represents an unbiased estimator of θ.
Second, the conditional variance of the MLE estimator is

$$\begin{aligned} \mathrm{Var}[\mathbf{Z}|\mathbf{X}] &= \frac{1}{N^2}\sum_{i=1}^{N} W_i\,\mathrm{Var}(Y_i|\mathbf{X})W_i^\top \\ \text{(assumption 1+orthonormality)} &= \frac{\sigma^2}{N} I_K \end{aligned} \tag{26}$$

The formula tells us that its conditional sample variance is inversely proportional to the sample size.

Moreover, given that $\mathbf{Z}$ is a linear combination of normal random variables and since any affine transformation of normal random variables is normally distributed, one can summarize the overall distribution of $\mathbf{Z}$ as

$$\mathbf{Z} \sim \mathcal{N}_K(\theta, \sigma^2 I_k N^{-1}) \tag{27}$$

3.4 How good is our estimator ?

Recall that our intial goal was to find the best estimate of m(x), which itself is the best possible predictor for the outcome variable Y. After introducing basis functions and creating an orthonormal system in order to make the problem more tractable, we established that $\hat{\mathbf{m}} = \mathbf{W}^\top \mathbf{Z}$ where $\hat{\theta} = \mathbf{Z}$.

In order to gage how good our estimate $\hat{\mathbf{m}}$ is, we want to express its corresponding risk function in a convenient form. We start with the loss function and manipulate the expression based on the assumptions we introduced.

$$\begin{aligned} ||\hat{\mathbf{m}} - \mathbf{m}||^2 &= \sum_{i=1}^{N} \left[W_i^\top (\mathbf{Z} - \theta)\right]^2 \\ \text{(by definition)} &= \sum_{i=1}^{N} \left[\sum_{k=1}^{K} \phi_k(X_i)(Z_k - \theta_k)\right]^2 \\ \text{(orthonormality)} &= \sum_{i=1}^{N} \phi_1(X_i)^2(Z_1 - \theta_1)^2 + \ldots + \phi_K(X_i)^2(Z_K - \theta_K)^2 \\ \text{(orthonormality)} &= \sum_{k=1}^{K} (Z_k - \theta_k)^2 \end{aligned} \tag{28}$$

This establishes that squared error loss for $\mathbf{m}$ corresponds to squared error loss for θ. Therefore, we can work with the loss function for $\mathbf{Z}$ instead of $\mathbf{m}$ when assesing the risk of our estimate $\hat{\mathbf{m}}$.

The loss function for $\mathbf{Z}$ is:

$$L(\mathbf{Z}, \theta) = ||\mathbf{Z} - \theta||^2 = \sum_{k=1}^{K} (Z_k - \theta_k)^2 \tag{29}$$

Since risk equals the average loss across repeated samples associated with using $\mathbf{Z}$ as an estimte of θ, we can write risk as:

$$\begin{aligned} R(\mathbf{Z}, \theta) &= \mathrm{E}\left[L(\mathbf{Z}, \theta)\right] \\ &= \mathrm{E} \sum_{k=1}^{K} (Z_k - \theta_k)^2 \\ \text{(properties of MLE)} &= \frac{K}{N}\sigma^2 \end{aligned} \tag{30}$$

This last expression represents the risk associated with our MLE procedure. So, how good is our MLE estimator? In the first half of this course you have learned that the MLE attains the Cramer Rao Lower Bound. This implies that in the class of unbiased estimators no estimator is going to have a variance strictly lower than the MLE. This is a positive result: it implies that given knowledge of the full parametric distribution (which

is needed to compute the MLE) we can compute estimators that are unbiased and have minimal variance.

Still, you may think that by trading off a little bit of bias with smaller variance we can actually achieve a better estimator. This is the main intuition behind the James-Stein estimator and other types of shrinkage estimators. The question that remains is how can we define a better estimator? In the next section we look for an estimator that is uniformly better (in terms of lower risk) than the MLE.

4 James-Stein Type Estimators

In this part, we show that there exist estimators with a risk lower than the one of MLE. We start by deriving the lower bound of risk that these oracle estimators achieve and then introduce Stein's Unbiased Risk Estimate (SURE), which enables us to define feasible estimators.

MLE is a member of the family of linear estimators. Define this class of linear estimators as

$$\mathcal{L} = \{C\mathbf{Z}; C = \text{diag}\{c_1, \ldots, c_K\}, c_k \in [0; 1]\} \tag{31}$$

The associated risk is going to be

$$\begin{aligned}
\mathrm{E}\left[\sum_{k=1}^{K}(c_k Z_k - \theta_k)^2\right] &= \mathrm{E}\left[\sum_{k=1}^{K}\left(c_k(Z_k - \theta_k) - (1-c_k)\theta_k\right)^2\right] \\
\text{(by properties of } \mathbf{Z}) &= \frac{\sigma^2}{N}\sum_{k=1}^{K} c_k^2 + \sum_{k=1}^{K}(1-c_k)^2\theta_k^2 + \sum_{k=1}^{K}(1-c_k)c_k\,\mathrm{E}(Z_k - \theta_k)\theta_k \qquad (32) \\
\text{(by properties of } \mathbf{Z}) &= \frac{\sigma^2}{N}\sum_{k=1}^{K} c_k^2 + \sum_{k=1}^{K}(1-c_k)^2\theta_k^2
\end{aligned}$$

In order to obtain the lowest possible risk for this class of estimators, we minimize this expression with respect to $\{c_1, \ldots, c_K\}$, which yields the following optimal choice

$$c_k^* = \frac{\theta_k^2}{\theta_k^2 + \sigma^2 N^{-1}} \tag{33}$$

Using these weights by plugging back this expression into (32), we obtain the risk associated with our new estimator $C\mathbf{Z}$, which represents the lower bound and is sometimes

called oracle inequality:

$$\begin{aligned} Risk(C\mathbf{Z}) &= \frac{\sigma^2}{N}\sum_{k=1}^{K}\left(\frac{\theta_k^2}{\theta_k^2+\sigma^2N^{-1}}\right)^2 + \sum_{k=1}^{K}\left(1-\frac{\theta_k^2}{\theta_k^2+\sigma^2N^{-1}}\right)^2\theta_k^2 \\ &= \sum_{k=1}^{K}\left(\frac{\theta_k^2}{\theta_k^2+\sigma^2N^{-1}}\right)^2\left[\sigma^2N^{-1}+\theta_k^2\right] + \theta_k^2 - \frac{2\theta_k^4}{\theta_k^2+\sigma^2N^{-1}} \\ &= \sum_{k=1}^{K}\frac{\theta_k^4+\theta_k^4+\theta_k^2\sigma^2N^{-1}-2\theta_k^4}{\theta_k^2+\sigma^2N^{-1}} \\ &= \frac{\sigma^2}{N}\sum_{k=1}^{K}\frac{\theta_k^2}{\theta_k^2+\sigma^2N^{-1}} \end{aligned} \tag{34}$$

You should notice that this risk is *smaller* than the one we have computed in the previous part for the MLE (i.e. σ^2KN^{-1}). This is indeed a great result. However, the problem is that is not operational, meaning that if you open Python and try to compute something like $C^*\mathbf{Z}$ it turns out that you cannot do this because knowledge of C^* requires knowledge of the truth, which is exactly what you are trying to estimate. Consequently, we are interested in whether there exists a *feasible* estimator with lower risk than that of MLE at all possible values of the paramterer θ.

This is where the result of Stein's Theorem becomes useful. It shows that a particular risk measure (defined as SURE in the notes) will be unbiased relative to the true risk of our estimator under our assumption 1 and 2, that is

$$\mathrm{E}[\hat{Risk}_{SURE}(\hat{\theta})] = \mathrm{E}\left[||\hat{\theta}-\theta||^2\right] \tag{35}$$

with

$$\hat{Risk}_{SURE}(\hat{\theta}) = \frac{K}{N}\sigma^2 + 2\frac{\sigma^2}{N}\sum_{k=1}^{K}\frac{\partial g_k}{\partial Z_k} + \sum_{k=1}^{K}(\hat{\theta}_k - Z_k)^2 \tag{36}$$

where $\hat{\theta}$ is any estimator of θ as a function of $\mathbf{Z}$, i.e. $\hat{\theta} = \hat{\theta}(\mathbf{Z})$ and

$$g(\mathbf{Z}) = \hat{\theta} - \mathbf{Z} \tag{37}$$

which has to be assumed to be weakly differentiable for this result to work. An important take-away of (36) is that $\hat{Risk}_{SURE}$ is operational. Thus the theorem can be used to conduct risk comparisons of different estimators. Moreover, it enables researchers to perform model selection.

Since now our estimator will be *operational*, we can repeat the steps we did to define an oracle type of estimator and compute $\hat{Risk}_{SURE}$ for a feasible estimator of the type $C\mathbf{Z}$:

$$\begin{aligned}
\hat{Risk}_{SURE}(C\mathbf{Z}) &\equiv \frac{K}{N}\sigma^2 + 2\frac{\sigma^2}{N}\sum_{k=1}^{K}\frac{\partial g_k}{\partial Z_k} + \sum_{k=1}^{K}(c_k Z_k - Z_k)^2 \\
\text{(recall definition of } g(\mathbf{Z})) &= \frac{K}{N}\sigma^2 - 2\frac{\sigma^2}{N}\sum_{k=1}^{K}(1-c_k) + \sum_{k=1}^{K} Z_k^2(1-c_k)^2 \\
\text{(add and subtract)} &= \frac{K}{N}\sigma^2 - 2\frac{\sigma^2}{N}\sum_{k=1}^{K}(1-c_k) + \frac{\sigma^2}{N}\sum_{k=1}^{K}(1-c_k)^2 + \sum_{k=1}^{K}(Z_k^2 - \frac{\sigma^2}{N})(1-c_k)^2 \\
&= \frac{\sigma^2}{N}\left[K - 2\sum_{k=1}^{K}(1-c_k) + \sum_{k=1}^{K}(1-c_k)^2\right] + \sum_{k=1}^{K}(Z_k^2 - \frac{\sigma^2}{N})(1-c_k)^2 \\
&= \frac{\sigma^2}{N}\sum_{k=1}^{K} c_k^2 + \sum_{k=1}^{K}(Z_k^2 - \frac{\sigma^2}{N})(1-c_k)^2
\end{aligned} \tag{38}$$

Therefore, the FOC w.r.t c_k yields

$$\begin{aligned}
&2\frac{\sigma^2}{N}c_k^* - 2(Z_k - \frac{\sigma^2}{N})(1-c_k^*) = 0 \\
&\rightarrow c_k^* = (1 - \frac{N^{-1}\sigma^2}{Z_k^2})
\end{aligned} \tag{39}$$

This is related to the estimator in Efromovich. Using SURE we chose the weights to minimize an unbiased estimate of risk, which is indeed operational.

5 Why it matters

This section's material involved a lot of tedious transformations. Nevertheless, the content is highly relevant for applied research in economics.

First, remember that a trade-off arises when specifying a prediction model. More comlex models tend to increase the sample fit and therefore reduce risk. However, increasing the model complexity also increases the risk. A common explanation is the problem of over-fitting the training samle, which leads to poor predictions using new data. As mentioned in the lecture notes, this trade-off is a central point in prediction and model selection problems.

This brings us to the second point: Model selection matters. Economic theory often does not provide intution into which variables to include and which ones to exclude. Many estimates are indeed sensitive to model selection and in the context of big data, this question becomes more and more important. This section also provides intuition for why shrinkage estimators, as the LASSO, can be of advantage.