Model selection, unbiased estimators, and the Gauss-Markov theorem

MATH 697 AM:ST

Tuesday, September 19th

Warmup

Positive matrices

Let us recall a fundamental linear algebra fact

$$M \text{ a self-adjoint linear operator } (M^t = M)$$
$$\Rightarrow M \text{ is diagonal in some orthonormal basis}$$

In other words, symmetric matrices M are equivalent to a diagonal matrix under an orthogonal change of basis.

Warmup

Positive matrices

A symmetric matrix M is said to be non-negative if all of its eigenvalues are non-negative.

Equivalently, M is non-negative if

$$(Mv, v) \geq 0 \text{ for all } v \in \mathbb{R}^n.$$

Likewise, we say that $M \geq N$ if $M - N$ is non-negative.

In summary

$$M \geq N \Leftrightarrow (Mv, v) \geq (Nv, v) \text{ for all } v \in \mathbb{R}^n.$$

Warmup

Covariance matrices

An important example: Covariance matrices of random vectors.

Let X be a random vector in $\mathbb{R}^n$. Then, for $v, w \in \mathbb{R}^n$ define

$$\mathrm{Cov}_X(v, w) = \mathbb{E}[(X - \mathbb{E}[X], v)(X - \mathbb{E}[X], w)]$$

This is a bilinear form that can be written as $(\Sigma_X v, w)$, where

$$(\Sigma_X)_{ij} = \mathbb{E}[(X_i - \mathbb{E}[X_i])(X_j - \mathbb{E}[X_j])]$$

(note that $\Sigma_X^t = \Sigma_X$ follows immediately from this formula)

Warmup

Covariance matrices

Observe that given $v \in \mathbb{R}^d$, then the real random variable (X, v) has mean $(\mathbb{E}[(X], v)$ and variance $\mathrm{Cov}_X(v, v)$, since

$$\mathrm{Cov}_X(v, v) = \mathbb{E}[((X, v) - (\mathbb{E}[X], v))^2]$$

In particular, given two random vectors X and X' in $\mathbb{R}^n$, when we say X has smaller variance than X', written as

$$\mathrm{Cov}_X \leq \mathrm{Cov}_{X'}$$

this simply means that for any v we have the inequality

$$\mathrm{Var}((X, v)) \leq \mathrm{Var}((X', v))$$

Warmup

Covariance matrices

Example: If $X = (X_1, X_2)$ where $X_i \sim N(0, \sigma^2)$ and the X_i are uncorrelated, then

$$\Sigma_X = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

If $X' := (2X_1, 3X_2)$, then clearly $\mathrm{Cov}(X') \geq \mathrm{Cov}(X)$.

Model Selection

Problem: All of the structures we have seen in the past two weels come with some kind of complexity parameter. Once we decide on one of these structures, we face the following question: What is a judicious choice for

... the penalty parameter λ?

... the number and nature of basis functions?

... the size of the support for the kernel?

... the number of nearest neighbors considered?

In other words, what is the right amount of **smoothing**?

Model Selection

Training error versus Test error

Higher Model Complexity $\Rightarrow$ Higher variance, lower bias.

For least squares the larger p, the more complexity.
For k-NN the smaller the k, the more complexity.

Complexity is intrinsically connected with smoothness: their relationship is a tense one where one pulls against the other.

Model Selection

Training error versus Test error

Fix N points $x_1, \ldots, x_N$, which are fixed and known.
(so for simplicity we are assuming no randomness in the x_i).

Along with this, we have a training data set modeled by random observations

$$Y_1, \ldots, Y_N$$

yielding training data $\{(x_i, Y_i)\}_{i=1}^N$. A particular realization of this data $\{(x_i, y_i)\}_{i=1}^N$, we denote by $\mathcal{T}$.

Alternatively, we think of $\mathcal{T}$ as the event given by

$$\{Y_1 = y_1, \ldots, Y_N = y_n\}$$

Model Selection

Training error versus Test error

We also fix a **loss function** $L(\hat{y}, y)$, typically it's just the squared error

$$L(\hat{y}, y) = |\hat{y} - y|^2$$

Model Selection

Training error versus Test error

Let $\hat{f}$ be a function which intends to estimate an unknown f.

The **training error** for $\hat{f}$ is defined as the quantity

$$\frac{1}{N}\sum_{i=1}^{N} L(\hat{f}(x_i), y_i)$$

Most commonly, in the case $L(\hat{y}, y) = |\hat{y} - y|^2$, it is given by

$$\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{f}(x_i))^2$$

Model Selection

Training error versus Test error

What is important to remember is the training error quanitifies how closely $\hat{f}$ matches the training data it was used in its selection.

One can imagine a situation where the $\hat{f}$ we choose may have a very small training error, but tends to give large errors when tested against predictions outside the training data set.

This issue leads us to the notion of **test error**.

Model Selection

Training error versus Test error

The **test error** for $\hat{f}$, also known as the **generalization error** is a more subtle notion which is also more useful in assessing the accuracy of a predictive model.

It pressuposes a statistical model for the data, and it is given by

$$\mathbb{E}[L(Y, \hat{f}(X)) \mid \mathcal{T}]$$

Where $\mathcal{T}$ corresponds to the training data (x_i, y_i).

The quantity above corresponds to the average error we make when using $\hat{f}(X)$ to estimate Y, having sampled the training data $\mathcal{T}$ and used it to select the function $\hat{f}$.

Model Selection

Training error versus Test error

…of course, if we take $L(\hat{y}, y) = |\hat{y} - y|^2$, then we obtain the expected squared prediction error, conditioned on $\mathcal{T}$,

$$\mathbb{E}[|Y - \hat{f}(X)|^2 \mid \mathcal{T}]$$

Model Selection

Training error versus Test error

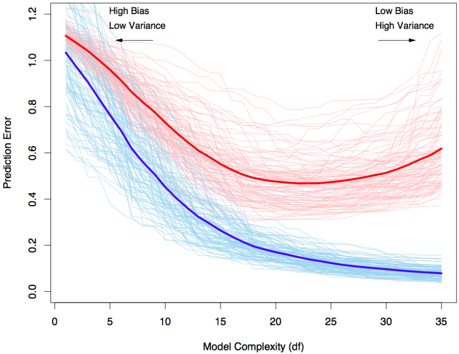
By the total expectation formula, we have in any casee

$$\mathbb{E}[L(Y, \hat{f}(X))] = \mathbb{E}[\mathbb{E}[L(Y, \hat{f}(X)) \mid \mathcal{T}]]$$

Thus: the Expected Prediction Error ought to be understood as the expected value for the **generalization error** for $\mathcal{T}$, this being a random variable computed from the model (X, Y) we have for our data, as well as the predictive model which generates $\hat{f}$ from a training set.

Model Selection

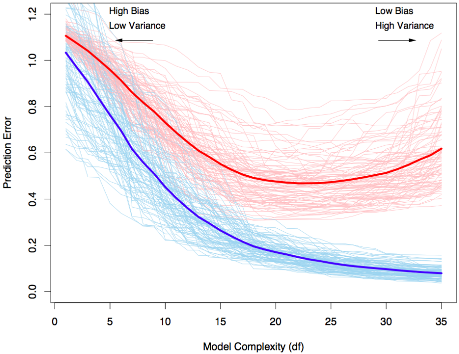
From Tibsharini, Hastie, and Friedman (p. 220-221)



As the model becomes more and more complex, it uses the training data more and is able to adapt to more complicated underlying structures...

Model Selection

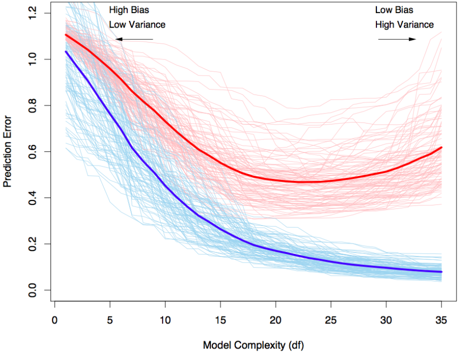
From Tibsharini, Hastie, and Friedman (p. 220-221):



...Hence there is a decrease in bias but an increase in variance. ***There is some intermediate model complexity that gives minimum expected test error.***

Model Selection

From Tibsharini, Hastie, and Friedman (p. 220-221):



Unfortunately training error is not a good estimate of the test error, as seen in Figure 7.1.

Model Selection

They conclude, thusly:

Training error consistently decreases with model complexity, typically dropping to zero if we increase the model complexity enough. However, a model with zero training error is overfit to the training data and will typically generalize poorly.

Now, a more in-depth look at **linear regression**

Least Squares' Statistics

Distribution of $\hat{\beta}$

Let us consider the additive model

$$Y = f(X) + \varepsilon$$

and approximate f among affine functions

$$\hat{f}(x) = \beta_0 + x \cdot \beta, \;\; \beta_0 \in \mathbb{R}, \beta \in \mathbb{R}^p$$

(the "real" f may or may not be affine).

By data, we mean N samples (X_i, Y_i) $i = 1, \ldots, N$.

Least Squares' Statistics

Distribution of $\hat{\beta}$

That is, there are random variables Y_i, X_i and ε_i satisfying

$$Y_i = \beta_0 + X_i \cdot \beta + \varepsilon_i$$

Assumptions: Assume that the X_i are known and fixed (not random) and that the ε_i all have mean zero, common variance σ^2 and are uncorrelated.

Least Squares' Statistics

Distribution of $\hat{\beta}$

From the data (X_i, Y_i) we recall the $N \times p$ matrix $\mathbf{X}$

$$\mathbf{X}v = (X_1 \cdot v, \ldots, X_N \cdot v)$$

and the random vectors

$$\begin{aligned}\mathbf{y} &= (Y_1, \ldots, Y_N)\\ \bar{\varepsilon} &= (\varepsilon_1, \ldots, \varepsilon_N)\end{aligned}$$

Under these assumptions, we have the equation

$$\mathbf{y} = \mathbf{X}\beta + \bar{\varepsilon}$$

Least Squares' Statistics

Distribution of $\hat{\beta}$

The least squares estimator for β, is the random vector

$$\hat{\beta} = (\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\mathbf{y}$$

A number of facts can be gleaned from the formula for $\hat{\beta}$, depending on what model assumptions we make about the data $\{(x_1, y_1), \ldots, (x_N, y_N)\}$

Least Squares' Statistics

Distribution of $\hat{\beta}$

First, under the current assumptions (namely the x_i being fixed and known and the y_i being uncorrelated and with common variance σ^2) one can show that

$$\mathbb{E}[\hat{\beta}] = \beta, \ \ \mathrm{Var}(\hat{\beta}) = (\mathbf{X}^t\mathbf{X})^{-1}\sigma^2$$

Least Squares' Statistics

Distribution of $\hat{\beta}$

Second, if one further assumes that $\varepsilon \sim \mathcal{N}(0, \sigma^2)$ and the ε_i are independent, then

$$\hat{\beta} \sim \mathcal{N}(\beta, (\mathbf{X}^t\mathbf{X})^{-1}\sigma^2)$$

The Gauss-Markov Theorem

Setup

The celebrated Gauss-Markov Theorem
This theorem says that the least squares estimator has the least variance among all **linear estimators**.

Let us explain what we mean by this.

The Gauss-Markov Theorem

Setup

First, let us repeat the assumptions.

For some N, we have $x_1, \ldots, x_N \in \mathbb{R}^p$, fixed and known vectors. Then, we have N random variables

$$Y_i = x_i \cdot \beta + \varepsilon_i$$

The ε_i are of mean zero and are pairwise uncorrelated. Then, our goal is to infer β from the Y_i. Concretely, we are looking at **estimators** for β.

The Gauss-Markov Theorem

Setup

Let us repeat the assumptions.

For some N, we have $x_1, \ldots, x_N \in \mathbb{R}^p$, fixed and known vectors. Then, we have the vector random variable $\mathbf{y}$

$$\mathbf{y} = \mathbf{X}\beta + \bar{\varepsilon}$$

The $\bar{\varepsilon}$ is a zero-mean vector with covariance matrix $\sigma^2\mathbf{I}$
Then, our goal is to infer β from $\mathbf{y}$. Concretely, we are looking at **estimators** for β.

The Gauss-Markov Theorem

Linear & unbiased estimators

An estimator $\tilde{\beta}$ is said to be **linear** if it has the form

$$\tilde{\beta} = \mathbf{C}\mathbf{y}, \quad \mathbf{y} = (Y_1, \ldots, Y_N) \in \mathbb{R}^N$$

where $\mathbf{C}$ is some $p \times N$ matrix.

Recall that an estimator for a parameter is said to be **unbiased** if its expectation returns the parameter for all values of the parameter. In this context, this means that we always have

$$\mathbb{E}[\mathbf{C}\mathbf{y}] = \beta$$

The Gauss-Markov Theorem

Linear & unbiased estimators

Going forward, let us use the notation

$$\mathbf{L} := (\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t$$

(the **L** stands for Least Squares). In this notation then, we have

$$\hat{\beta} := \mathbf{L}\mathbf{y}$$

The Gauss-Markov Theorem

Linear & unbiased estimators

As mentioned earlier, the least squares estimator is unbiased

Proposition

If the x_i are fixed and $\hat{\beta}$ is as above then for any β we have

$$\mathbb{E}[\hat{\beta}] = \beta$$

The Gauss-Markov Theorem

Linear & unbiased estimators

Proof. Observe that $\mathbf{LX} = \mathbf{I}$, so

$$\begin{aligned}\hat{\beta} &= \mathbf{Ly} = \mathbf{L}(\mathbf{X}\beta + \bar{\varepsilon}) \\ &= \beta + \mathbf{L}\bar{\varepsilon}\end{aligned}$$

Since the entries of $\mathbf{L}$ are non-random, we have

$$\mathbb{E}[\mathbf{L}\bar{\varepsilon}] = \mathbf{L}\mathbb{E}[\bar{\varepsilon}] = 0$$

Therefore, we always have $\mathbb{E}[\hat{\beta}] = \beta$.

The Gauss-Markov Theorem

Theorem

Among all unbiased linear estimators for β, the least squares estimator has the least variance.

The Gauss-Markov Theorem

Theorem

... In other words: if $\boldsymbol{C}$ *is a* $p \times N$ *matrix such that*

$$\mathbb{E}[\boldsymbol{C}\boldsymbol{y}] = \beta$$

Then,

$$\operatorname{Cov}(\boldsymbol{C}\boldsymbol{y}) \geq \operatorname{Cov}(\hat{\beta}).$$

The Gauss-Markov Theorem

Equivalently, we have that for any vector $v \in \mathbb{R}^p$,

$$\mathrm{Var}((\tilde{\beta}, v)) \geq \mathrm{Var}((\hat{\beta}, v)).$$

where we are writing $\tilde{\beta} = \mathbf{C}\mathbf{y}$ and $\hat{\beta} = \mathbf{L}\mathbf{y}$

Let us prove the theorem by showing the above inequality.

The Gauss-Markov Theorem

Fix $v \in \mathbb{R}^p$, then

$$\mathbb{E}[(\tilde{\beta}, v)] = \mathbb{E}[(\hat{\beta}, v)] = (\beta, v)$$

Therefore,

$$\mathrm{Var}((\tilde{\beta}, v)) = \mathbb{E}[(\tilde{\beta} - \beta, v)^2]$$
$$\mathrm{Var}((\hat{\beta}, v)) = \mathbb{E}[(\hat{\beta} - \beta, v)^2]$$

The Gauss-Markov Theorem

Recall that $\mathbf{y} = \mathbf{X}\beta + \bar{\varepsilon}$, where $\bar{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_N)$.

Then,

$$\mathbf{Cy} = \mathbf{CX}\beta + \mathbf{C}\bar{\varepsilon} \Rightarrow \mathbb{E}[\mathbf{Cy}] = \mathbf{CX}\beta$$

It follows then that $\mathbf{C}$ yields an unbiased estimator if and only if

$$\mathbf{CX} = \mathrm{I}$$

(Note that, as it should be, we have $\mathbf{LX} = (\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\mathbf{X} = \mathbf{I}$)

The Gauss-Markov Theorem

In any case, going forward, we keep in mind this:

$$\mathbf{C}\mathbf{X} = \mathrm{I}$$

We have also established that

$$\hat{\beta} = \beta + \mathbf{L}\bar{\varepsilon},\ \tilde{\beta} = \beta + \mathbf{C}\bar{\varepsilon}$$

Let us use this to compute the variances of $(\hat{\beta}, v)$ and $(\tilde{\beta}, v)$

The Gauss-Markov Theorem

We have that

$$\mathrm{Var}((\hat{\beta}, v)) = \mathbb{E}[(\hat{\beta} - \beta, v)^2] = \mathbb{E}[(\mathbf{L}\bar{\varepsilon}, v)^2]$$
$$\mathrm{Var}((\tilde{\beta}, v)) = \mathbb{E}[(\tilde{\beta} - \beta, v)^2] = \mathbb{E}[(\mathbf{C}\bar{\varepsilon}, v)^2]$$

what we need to prove is that

$$\mathbb{E}[(\mathbf{L}\bar{\varepsilon}, v)^2] \leq \mathbb{E}[(\mathbf{C}\bar{\varepsilon}, v)^2]$$

The Gauss-Markov Theorem

Let us write one expectation in terms of the other:

$$\begin{aligned}\mathbb{E}[(\mathbf{C}\bar{\varepsilon}, v)^2] &= \mathbb{E}[(\mathbf{L}\bar{\varepsilon} + (\mathbf{C} - \mathbf{L})\bar{\varepsilon}, v)^2] \\ &= \mathbb{E}[(\mathbf{L}\bar{\varepsilon}, v)^2] + \mathbb{E}[((\mathbf{C} - \mathbf{L})\bar{\varepsilon}, v)^2] \\ &\quad + 2\mathbb{E}[(\mathbf{L}\bar{\varepsilon}, v)((\mathbf{C} - \mathbf{L})\bar{\varepsilon}, v)]\end{aligned}$$

Claim: In these circumstances, we have that

$$\mathbb{E}[(\mathbf{L}\bar{\varepsilon}, v)((\mathbf{C} - \mathbf{L})\bar{\varepsilon}, v)] = 0$$

and this immediately proves the theorem, since $\mathbb{E}[(\mathbf{C}\bar{\varepsilon}, v)^2]$ is then equal to $\mathbb{E}[(\mathbf{L}\bar{\varepsilon}, v)^2]$ plus a non-negative term!

The Gauss-Markov Theorem

It remains to prove the claim –here is where the assumption that the ε_i are uncorrelated and have the same variance is used!

First, note that

$$\mathbb{E}[(\mathbf{L}\bar{\varepsilon}, v)((\mathbf{C}-\mathbf{L})\bar{\varepsilon}, v)] = \mathbb{E}[(\bar{\varepsilon}, \mathbf{L}^t v)(\bar{\varepsilon}, (\mathbf{C}-\mathbf{L})^t v)]$$

Therefore, from the definition of the Covariance matrix

$$\begin{aligned} \mathbb{E}[(\mathbf{L}\bar{\varepsilon}, v)((\mathbf{C}-\mathbf{L})\bar{\varepsilon}, v)] &= \mathrm{Cov}_{\bar{\varepsilon}}(\mathbf{L}^t v, (\mathbf{C}-\mathbf{L})^t v) \\ &= \sigma^2(\mathbf{L}^t v, (\mathbf{C}-\mathbf{L})^t v) \end{aligned}$$

The latter beign simply because $\Sigma_{\bar{\varepsilon}} = \sigma^2\mathbf{I}$.

The Gauss-Markov Theorem

Now,

$$(\mathbf{L}^t v, (\mathbf{C} - \mathbf{L})^t v) = (v, \mathbf{L}(\mathbf{C} - \mathbf{L})^t v)$$

Finally, we make use of two facts $\mathbf{CX} = \mathbf{I}$ and the form of $\mathbf{L}$

$$\begin{aligned} \mathbf{LL}^t &= (\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\mathbf{X}(\mathbf{X}^t\mathbf{X})^{-1} = (\mathbf{X}^t\mathbf{X})^{-1} \\ \mathbf{LC}^t &= (\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\mathbf{C}^t = (\mathbf{X}^t\mathbf{X})^{-1} \end{aligned}$$

This means that $\mathbf{L}(\mathbf{C} - \mathbf{L})^t = 0$, and therefore

$$(\mathbf{L}^t v, (\mathbf{C} - \mathbf{L})^t v) = 0 \;\; \forall \, v$$

as we wanted!.

The biased estimators: ridge regression

MATH 697 AM:ST

Tuesday, September 19th

Warm up

Polar factorization

Let $\mathbf{M}$ be a $n \times n$ **invertible** matrix. Then the matrix

$$\mathbf{M}\mathbf{M}^t$$

is symmetric, non-negative, and invertible.

Thusly, there is a unique symmetric, positive $\mathbf{P}$ such that

$$\mathbf{P}^2 = \mathbf{M}\mathbf{M}^t$$

Warm up

Polar factorization

In this case, moreover, the matrix

$$\mathbf{V} := \mathbf{M}^{-1}\mathbf{P}$$

is orthogonal and moreover, one has that

$$\mathbf{M} = \mathbf{P}\mathbf{V}^t$$

This expression for $\mathbf{M}$ is known as its **polar factorization**.

Warm up

Singular Value Decomposition of a matrix

What happens for general matrices? What we have is the **Singular Value Decomposition (SVD)**:

It says that given a $n \times m$ matrix $\mathbf{M}$, there exists

$\mathbf{U}$ a $n \times m$ matrix whose columns yield
an orthonormal basis for the image of $\mathbf{M}$,

$\mathbf{D}$ a $n \times n$ diagonal non-negative matrix,

$\mathbf{V}$ a $n \times n$ orthogonal matrix,

all such that $\mathbf{M}$ is decomposed as

$$\mathbf{M} = \mathbf{U}\mathbf{D}\mathbf{V}^t$$

Warm up

The SVD and Least Squares

Let $\mathbf{X}$ be a $N \times p$ matrix (as considered in least squares), and let us consider it's SVD

$$\mathbf{X} = \mathbf{U}\mathbf{D}\mathbf{V}^t$$

Then, a brief computation yields the identity

$$(\mathbf{X}^t\mathbf{X})^{-1} = \mathbf{V}^{-t}\mathbf{D}^{-2}\mathbf{V}^{-1}$$

as well as the important identity,

$$\begin{aligned} (\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t &= \mathbf{V}^{-t}\mathbf{D}^{-1}\mathbf{U}^t \\ \mathbf{X}(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t &= \mathbf{U}\mathbf{U}^t \end{aligned}$$

Warm up

The SVD and Least Squares

Let us denote by $u_1, \ldots, u_p$ the non-zero columns of $\mathbf{U}$.

Now, given $y \in \mathbb{R}^N$, the entries of

$$\mathbf{U}^t y$$

are either zero or of the form (u_i, y), thus, they represent the components of $\hat{y}$ in the basis $u_1, \ldots, u_r$. Accordingly,

$$\hat{y} = \sum_{i=1}^{p} (u_i, y) u_i$$

Univariate v. Multivariate regression

When doing a regression where the input space has dimension $p = 1$ we say the regression is **univariate**.

Then, suppose we are given data $(x_1, y_1), \ldots, (x_N, y_N)$ with $x_i, y_i \in \mathbb{R}$, and write

$$\mathbf{x} = (x_1, \ldots, x_N), \;\; \mathbf{y} = (y_1, \ldots, y_N)$$

Univariate v. Multivariate regression

The least square regression reduces to finding β_0 minimizing

$$|\mathbf{x}\beta - \mathbf{y}|^2$$

It is immediate that the solution is given by

$$\hat{\beta} = \frac{\sum\limits_{i=1}^{N} x_i y_i}{\sum\limits_{i=1}^{N} x_i^2} = \frac{\mathbf{x} \cdot \mathbf{y}}{|\mathbf{x}|^2}$$

Univariate v. Multivariate regression

It is illustrative to compare the univariate formula

$$\hat{\beta} = |\mathbf{x}|^{-2}\mathbf{x} \cdot \mathbf{y}$$

with the multivariate one

$$\hat{\beta} = (\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\mathbf{y}$$

The multivariate regression **can** be decomposed into several independent univariate regressions, in the right system of coordinates.

Univariate v. Multivariate regression

Suppose for a moment that the matrix $\mathbf{X}^t\mathbf{X}$ is diagonal

$$(\mathbf{X}^t\mathbf{X})_{ij} = d_i^2 \delta_{ij}$$

where $d_i^2 = x_{1i}^2 + \ldots + x_{Ni}^2$

$$\hat{\beta} = (\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\mathbf{y}$$

becomes

$$\hat{\beta}_j = d_i^{-2} \sum_{i=1}^{N} x_{ji} y_j$$

Univariate v. Multivariate regression

When the matrix $\mathbf{X}^t\mathbf{X}$ is not diagonal, since it is in any case symmetric and positive, we may use the **Gram-Schmidt process** to select a new orthonormal system of coordinates where $\mathbf{X}^t\mathbf{X}$ is diagonal.

Therefore, up to “just” a change of orthonormal basis, multivariate regression is the same as p separate univariate regressions.

Univariate v. Multivariate regression

Centering and the intercept

Let us add a word about ways of dealing with β_0.

First, the way we have dealt with so far more or less implicitly: we think of the numbers

$$(x_i, \beta) + \beta_0$$

as the dot product between $(1, x_i)$ and (β_0, β), vectors in $\mathbb{R}^{p+1}$.

Then, by adding an extra column to $\mathbf{X}$, we make β_0 part of the β parameter, increasing its dimension by 1.

Univariate v. Multivariate regression

Centering and the intercept

Another option is to **center the data** before applying the least squares algorithm

$$\sum_{i=1}^{N} |y_i - (x_i, \beta) - \beta_0|^2$$

Meaning the following: thinking first of β as **fixed**, minimize the above as a function of β_0.

Univariate v. Multivariate regression

Centering and the intercept

This is a parabola with respect to β_0, so finding the minimum is simple…

Differentiating, we have

$$\frac{d}{d\beta_0}\sum_{i=1}^{N}|y_i-(x_i,\beta)-\beta_0|^2 = 2\sum_{i=1}^{N}(y_i-(x_i,\beta)-\beta_0)$$

Therefore, the minimizer $\hat{\beta}_0$ is given by

$$\sum_{i=1}^{N}(y_i-(x_i,\beta)-\hat{\beta}_0)=0$$

Univariate v. Multivariate regression

Centering and the intercept

This, would yield the solution

$$\hat{\beta}_0 = \frac{1}{N}\sum_{i=1}^{N} y_i - (x_i, \beta) = \frac{1}{N}\sum_{i=1}^{N} y_i - \left(\frac{1}{N}\sum_{i=1}^{N} x_i, \beta\right)$$

However, this does not mean we choose this $\hat{\beta}_0$.

Instead we change the y_i and x_i to make $\hat{\beta}_0 = 0$

$$y_i' = y_i - \bar{y}, \;\; x_i' = x_i - \bar{x}$$

Univariate v. Multivariate regression

Centering and the intercept

where $\bar{y}$ and $\bar{x}$ are

$$\bar{y} = \frac{1}{N}\sum_{i=1}^{N} y_i, \;\; \bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$$

This means that $y'_1 + \ldots + y'_N = 0$ and $x'_1 + \ldots + x'_N = 0$, and thus, $\hat{\beta}_0 = 0$ **regardless** of β.

The process of changing the variables y_i and x_i for $y_i - \bar{y}$ is what is known as **centering the data**.

Variable subset selection

We have already seen that the EPE is given by the addition of the **biase squared**, the **variance** underlying to the data, and the **variance inherent to the estimator**.

What if a great deal of the variance appearing above is due to a component $\hat{\beta}_j$ of $\hat{\beta}$ with small mean ($= \beta_j$) and large variance?

Well, if we replaced $\hat{\beta}_j$ with 0, then the bias of $\hat{\beta}$ would change from zero to something small, while at the same time the variance would decrease by a lot. Yielding a net improvement on the EPE. This means that **unbiased estimators are preferable** under the right circumstances.

Variable subset selection

1. Best-subset selection via *leaps and bounds* algorithm.
2. Forward and backward stepwise selection.
3. Forward-stagewise regression.

The above involve a discrete selection process. What if instead of altogether eliminating a variable, we diminished, in a continuous manner, its influence on our prediction model?

This is the idea behind shrinking methods, which are by now more widely used than selection methods. We start with Ridge regression and the Lasso.

Ridge regression

So, in least squares, we have an affine structure

$$y = (x, \beta) + \beta_0$$

and data $\mathbf{X}$ and $\mathbf{y}$, leading to

$$\hat{\beta} = \left(\mathbf{X}^t \mathbf{X}\right)^{-1} \mathbf{X}^t \mathbf{y}$$

For random data, what may cause $\hat{\beta}$ to have high variance?

Ridge regression

Ridge regression

Hoerl and Kennard (1970):
Add $\lambda\mathbf{I}$ to $\mathbf{X}^t\mathbf{X}$ before taking the inverse, then set

$$\hat{\beta}^{\text{ridge}} = \left(\mathbf{X}^t\mathbf{X} + \lambda\mathbf{I}\right)^{-1}\mathbf{X}^t\mathbf{y}$$

For $\lambda > 0$, $\mathbf{X}^t\mathbf{X} + \lambda$ is always invertible, even if $\mathbf{X}^t\mathbf{X}$ is not.

Ridge regression

On the other hand, we have

$$\mathbf{H}_\lambda = \mathbf{X}\left(\mathbf{X}^t\mathbf{X} + \lambda\mathbf{I}\right)^{-1}\mathbf{X}^t$$

Let us use the SVD for $\mathbf{X}$, $\mathbf{X} = \mathbf{U}\mathbf{D}\mathbf{V}^t$, then

$$\mathbf{X}^t\mathbf{X} + \lambda\mathbf{I} = \mathbf{V}\mathbf{D}^2\mathbf{V}^t + \lambda\mathbf{I}$$

Now, since $\mathbf{V}(\lambda\mathbf{I})\mathbf{V}^t = \lambda\mathbf{I}$, we have

$$\begin{aligned} &\mathbf{X}^t\mathbf{X} + \lambda\mathbf{I} = \mathbf{V}(\mathbf{D}^2 + \lambda\mathbf{I})\mathbf{V}^t \\ &\Rightarrow (\mathbf{X}^t\mathbf{X} + \lambda\mathbf{I})^{-1} = \mathbf{V}(\mathbf{D}^2 + \lambda\mathbf{I})^{-1}\mathbf{V}^t \end{aligned}$$

Ridge regression

Therefore,

$$\begin{aligned}\mathbf{H}_\lambda &= (\mathbf{U}\mathbf{D}\mathbf{V}^t)\mathbf{V}(\mathbf{D}^2 + \lambda\mathbf{I})^{-1}\mathbf{V}^t(\mathbf{V}\mathbf{D}\mathbf{U}^t) \\ &= \mathbf{U}(\mathbf{D}(\mathbf{D}^2 + \lambda\mathbf{I})^{-1}\mathbf{D})\mathbf{U}^t\end{aligned}$$

In terms of the column vectors of $\mathbf{U}$, this is

$$\hat{y}^{\text{ridge}} = \sum_{i=1}^{p} (u_i, y)\frac{d_i^2}{d_i^2 + \lambda} u_i$$

Ridge regression

What about $\hat{\beta}^{\text{ridge}}$?

Well, we saw from the SVD for $\mathbf{X}$, $\mathbf{UDV}^t$ that

$$(\mathbf{X}^t\mathbf{X} + \lambda\mathbf{I})^{-1} = \mathbf{V}(\mathbf{D}^2 + \lambda\mathbf{I})^{-1}\mathbf{V}^t$$

from where it follows that

$$\begin{aligned}(\mathbf{X}^t\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^t &= \mathbf{V}(\mathbf{D}^2 + \lambda\mathbf{I})^{-1}\mathbf{V}^t(\mathbf{VDU}^t)\\ &= \mathbf{V}(\mathbf{D}^2 + \lambda\mathbf{I})^{-1}\mathbf{DU}^t\end{aligned}$$

Ridge regression

What about $\hat{\beta}^{\text{ridge}}$?

We conclude that

$$\hat{\beta}^{\text{ridge}} = \mathbf{V}(\mathbf{D}^2 + \lambda\mathbf{I})^{-1}\mathbf{D}\mathbf{U}^t\mathbf{y}$$

which, it's convenient to write as

$$\begin{aligned}\hat{\beta}^{\text{ridge}} &= \sum_{i=1}^{p} \frac{d_i}{d_i^2 + \lambda}(\mathbf{u}_i, y)\mathbf{v}_i \\ &= \sum_{i=1}^{p} \frac{d_i^2}{d_i^2 + \lambda}\frac{1}{d_i}(\mathbf{u}_i, y)\mathbf{v}_i\end{aligned}$$

Ridge regression

…or, in the more suggestive form

$$\hat{\beta}^{\text{ridge}} = \sum_{i=1}^{p} \frac{d_i^2}{d_i^2 + \lambda} \frac{1}{d_i} (\mathbf{u}_i, y) \mathbf{v}_i$$

Compare this with the least squares solution $\hat{\beta}$, which is

$$\hat{\beta}^{\text{ridge}} = \sum_{i=1}^{p} \frac{1}{d_i} (\mathbf{u}_i, y) \mathbf{v}_i$$

The numbers $d_i^2/(d_i^2 + \lambda)$ are all < 1 when $\lambda > 0$ and $d_i > 0$, therefore, ridge regression **shrinks the components of** $\hat{\beta}$ in the basis $\mathbf{v}_1, \ldots, \mathbf{v}_p$.

Ridge regression

More so,

$$\hat{\beta}^{\text{ridge}} = \sum_{i=1}^{p} \frac{d_i^2}{d_i^2 + \lambda} \frac{1}{d_i} (\mathbf{u}_i, y) \mathbf{v}_i$$

Those components corresponding to smaller d_i are shrunk way more than the components corresponding to larger d_i.

It is helpful to think of ridge regression as diminishing the effects corresponding to the directions along which the sample points x_i exhibit the smaller variance.

Ridge regression

Suppose for a second that the sample point data $\mathbf{X}$ is such that

$$\mathbf{D} = d\mathbf{I}$$

Then, ridge regression yields smaller multiples of the least squares solution

$$\hat{\beta}^{\text{ridge}} = \frac{d^2}{d^2 + \lambda}\hat{\beta}, \quad \hat{\mathbf{y}}^{\text{ridge}} = \frac{d^2}{d^2 + \lambda}\hat{\mathbf{y}}$$

Ridge regression

The variational side of ridge regression

Problem:

Given $\{(x_i, y_i)\}_{i=1}^N$ find $\beta_0 \in \mathbb{R}, \beta \in \mathbb{R}^p$ minimizing

$$\sum_{i=1}^{N} (y_i - \beta_0 - (x_i, \beta))^2 + \lambda|\beta|^2$$

Note the “penalization” term does not involve β_0.
We do the usual centering,

$$y_i \to y_i - \bar{y}, \;\; x_i \to x_i - \bar{x}$$

Where $\bar{y}$ and $\bar{x}$ are the respective means.

Ridge regression

The variational side of ridge regression

Having done this centering, we aim to minimize

$$\mathrm{RSS}_\lambda(\beta) = \sum_{i=1}^{N} (y_i - (x_i, \beta))^2 + \lambda |\beta|^2$$

As in standard least squares, we may rewrite this in vector form

$$|\mathbf{X}\beta - \mathbf{y}|^2 + \lambda |\beta|^2$$

Ridge regression

The variational side of ridge regression

Computing the gradient of RSS_λ

$$\begin{aligned}\nabla(|\mathbf{X}\beta-\mathbf{y}|^2+\lambda|\beta|^2) &= 2\mathbf{X}^t\,(\mathbf{X}\beta-\mathbf{y})+2\lambda\beta\\ &= 2(\mathbf{X}^t\mathbf{X}+\lambda\mathbf{I})\beta-2\mathbf{X}^t\mathbf{y}\end{aligned}$$

Therefore, the minimizer is given by

$$(\mathbf{X}^t\mathbf{X}+\lambda\mathbf{I})^{-1}\mathbf{X}^t\mathbf{y}$$

That is, by $\hat{\beta}^{\mathrm{ridge}}$! This shows ridge regression is simply least squares plus a quadratic penalization term for β having coefficient λ.

The Lasso

ℓ^2 and ℓ^1

Let us make a brief comment about the norms

$$|x|_{\ell^2} = \left(\sum_{j=1}^{p} |x_j|^2 \right)^{\frac{1}{2}}, \quad |x|_{\ell^1} = \sum_{j=1}^{p} |x_j|$$

The “unit ball” for the ℓ^2 is the standard Euclidean ball in $\mathbb{R}^p$, meanwhile, the “unit ball” for ℓ^1 is a diamond shape with flat faces and corners.

The Lasso

ℓ^2 and ℓ^1

This fact about "flat sides" and "corners" for $B_1^{\ell^1}(0)$ has a very important consequence, which we state as a lemma

Lemma

Let $f : \mathbb{R}^p \mapsto \mathbb{R}$ *be a linear functional, then, the maximum of* f *over* $B_1^{\ell^1}(0)$ *is always attained at least at a vector* $x = (x_1, \ldots, x_p) \in B_1^{\ell^1}(0)$ *having at least one zero component, i.e.* $x_j = 0$ *for some* j*.*

The Lasso

The **L**east **A**bsolute **S**hrinkage and **S**election **O**perator

Given $\{(x_i, y_i)\}_{i=1}^N$ find $\beta_0 \in \mathbb{R}, \beta \in \mathbb{R}^p$ minimizing

$$\frac{1}{2}\sum_{i=1}^{N}(y_i - \beta_0 - x_i \cdot \beta)^2 + \lambda \sum_{j=1}^{p} |\beta_j|$$

or, equivalently

$$\sum_{i=1}^{N}(y_i - \beta_0 - x_i \cdot \beta)^2 + \lambda|\beta|_{\ell^1}$$

The Lasso

The **L**east **A**bsolute **S**hrinkage and **S**election **O**perator

After the proper centering, we find ourselves with the problem of minimizing

$$\begin{aligned} J(\beta) &= \frac{1}{2}\sum_{i=1}^{N}(y_i - x_i \cdot \beta)^2 + \lambda|\beta|_{\ell^1} \\ &= \frac{1}{2}|\mathbf{X}\beta - \mathbf{y}|^2 + \lambda|\beta|_{\ell^1} \end{aligned}$$

This functional J is **convex**, and second order differentiable away from $\beta = 0$, at which point it fails to be differentiable of first order.

The Lasso

The **L**east **A**bsolute **S**hrinkage and **S**election **O**perator

We observe that for β's such that $\beta_j \neq 0 \; \forall \; j$

$$\nabla|\beta|_{\ell^1} = \mathrm{sign}(\beta) := (\mathrm{sign}(\beta_1), \ldots, \mathrm{sign}(\beta_p))$$

Then, for such β, we have

$$\nabla J(\beta) = \mathbf{X}^t(\mathbf{X}\beta - \mathbf{y}) + \lambda \mathrm{sign}(\beta)$$

Trying to solve $\nabla J(\beta) = 0$ is not as straightforward now as in least squares! The resulting equation is not linear.

The Lasso

The **L**east **A**bsolute **S**hrinkage and **S**election **O**perator

Example:
Consider the case $p = 1$ and with data $x_1, \ldots, x_N$ such that

$$x_1^2 + \ldots + x_N^2 = 1$$

Then, if we consider the function of the real variable β

$$J(\beta) = \frac{1}{2} \sum_{i=1}^{N} |x_i \beta - y_i|^2 + \lambda |\beta|$$

it's minimizer is given by

$$\operatorname{sign}(\hat{\beta})(|\hat{\beta}| - \lambda)_+$$

where $\hat{\beta}$ is the least squares solution.

The Lasso

The **L**east **A**bsolute **S**hrinkage and **S**election **O**perator

Example: (continued)
This means that the lasso estimator takes the least square solution $\hat{\beta}$ and returns **zero** if $|\hat{\beta}|$ is no larger than λ, and otherwise returns the $\beta - \lambda$ or $\beta + \lambda$ according to the sign of β. The function

$$\beta \mapsto \operatorname{sign}(\beta)(|\beta| - \lambda)_+$$

is known as a **shrinkage operator**.

The third week, in one slide

1. When selecting a predictive model, one may judge it by looking at the **training error** and the **test error**. Of these, the training error is by the better evaluation metric.
2. A biased algorithm is often preferable to an unbiased one: the tradeoff may lead to a reduced mean squared error.
3. The Gauss-Markov Theorem says that among **linear, unbiased** estimators, least squares has the least variance.
4. Ridge regression is an alternative linear estimator with bias. It diminishes the components β_j according to the degeneracy associated to each direction.
5. The Lasso takes advantage of the geometry of the ℓ^1 metric to shrink the components of an estimation for β in a more dramatic manner than ridge regression.