Properties of the LS estimates

$$V(\hat{\beta}) = \sigma^2 (X'X)^{-1} = \sigma^2 \sum d_k^{-2} \tilde{v}_k \tilde{v}_k'$$

(spectral decomposition)

$$E\|\hat{\beta} - \beta\|^2 = E\left((\hat{\beta}-\beta)'(\hat{\beta}-\beta)\right) = \mathrm{Tr}\{V(\hat{\beta})\} = \sigma^2 \mathrm{Tr}\{(X'X)^{-1}\}$$

(Trace)

$$= \sigma^2 \sum d_k^{-2} \quad (*)$$

So, total MSE of slope estimates depend on eigenvalues $d_1^2, \ldots, d_p^2$ of $(X'X)$.

Now, if (any) d_k^2 is small, $(*)$ can get very large!

- That's what's meant by an unstable regression fit — high variance.
- On the other hand, $E(\hat{\beta} - \beta) = 0$, so LS estimators are unbiased

So, can't have both — trade off some bias to reduce the variance

$\Rightarrow$ Regularized fits!

How can we achieve this?
Stabilize the estimate — the matrix inversion

$$\hat{\beta}_R = \underbrace{(X'X + rI)^{-1}}_{G_r} X'y$$

- $E(\hat{\beta}_R) = G_r X'X\beta = \beta - \underbrace{r G_r \beta}_{\text{bias}}$

- $V(\hat{\beta}_R) = \sigma^2 G_r (X'X) G_r$

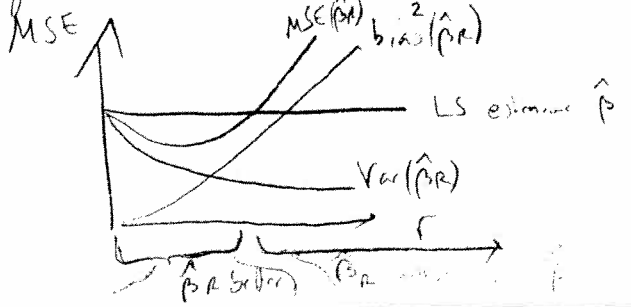
- MSE $= Var + bias^2$
  $= G_r (\sigma^2 X'X + r^2 \beta\beta') G_r$

- $E\|\hat{\beta}_R - \beta\|^2 = Tr\{MSE\} = \sum \frac{\sigma^2 d_k^2 + r^2 \beta_k^2}{(d_k^2 + r)^2}$ $\quad \left(r = 0 \quad \sum \sigma^2 d_k^{-2}\right)$

Now, for some values of r

$$\sum \frac{\sigma^2 d_k^2 + r^2 \beta_k^2}{(d_k^2 + r)^2} < \sum \sigma^2 d_k^{-2}$$

so we get a smaller MSE with $\hat{\beta}_R$ than $\hat{\beta}$



Alternative view : shrinkage estimator

$$\hat{\beta}_S = \frac{\hat{\beta}}{1+c}, \quad c \geq 0$$

$\Rightarrow$ bias $\frac{c\beta}{1+c}$, $V(\hat{\beta}_S) = \sigma^2 (1+c)^{-2} (X'X)^{-1}$

$$MSE = (1+c)^{-2} \left(\sigma^2 (X'X)^{-1} + c^2 \beta\beta' \right)$$

- As before — a trade-off between bias & variance.
- for some c, $MSE(\hat{\beta}_S) < MSE(\hat{\beta})$

If $(X'X)$ diagonal — same thing as $\hat{\beta}_R = (X'X + rI)^{-1} X'y$
— a shrinkage factor applied to all $\hat{\beta}$'s

We can arrive at the $\hat{\beta}_R$ estimate through another argument

- LS : find β to minimize $(y - X\beta)'(y - X\beta)$
- Penalized LS : find β to minimize $(Y - X\beta)'(Y - X\beta)$
  where $\|\beta\|^2 \leq T$

  ie. we place a constraint on the total magnitude of the coefficients.

(Another Bayesian interpretation $p(\beta) \sim N(0, \sigma^2 V)$)

OLS - so we want to minimize

$$(y-X\beta)'(y-X\beta)$$

but we also want to make sure

$\beta'\beta = \|\beta\|^2 = \sum \beta_j^2 \leq T$

These types of constrained optimization problems are solved through Lagrange methods

(*) Minimize $\boxed{(y-X\beta)'(y-X\beta) + r\beta'\beta}$

where r is such that $\beta'\beta \leq T$

(can try different values for r until $\beta'\beta \leq T$ (line search)

(*) $\Rightarrow \frac{\partial (*)}{\partial \beta} = -2X'(y-X\beta) + 2r\beta = 0$

$\Rightarrow (X'X + rI)\beta = X'y$

$\Rightarrow \boxed{\hat{\beta}_R = (X'X + rI)^{-1} X'y}$

This is also called $\boxed{\text{ridge regression}}$

We can also use other penalties on β - maybe not constrain sum of squares $\sum \beta_j^2$ but only sum of absolute values $\sum |\beta_j|$

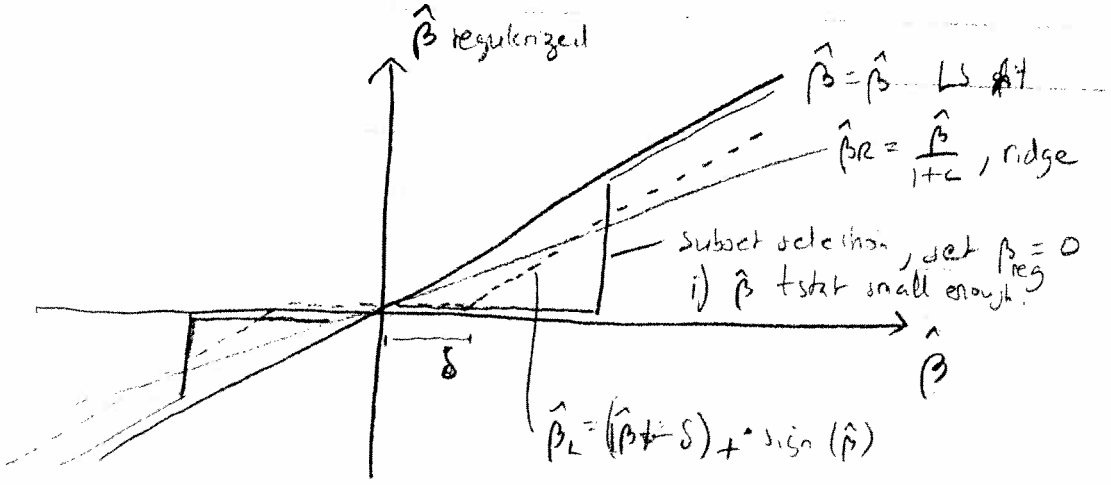
This is called the LASSO

$$\min (y - X\beta)'(y - X\beta) + \tau \frac{\beta'\beta}{\|\beta\|}$$

- If $(X'X)$ diagonal $\hat{\beta}_{Lasso,j} = \text{sign}(\hat{\beta}_j)(|\hat{\beta}_j| - \delta)_+$

  where δ s.t. $\sum |\hat{\beta}_{Lasso,j}| \leq T$

Compare, $\hat{\beta}_R = \frac{\hat{\beta}}{1+c}$, where c s.t. $\sum \hat{\beta}_R^2 \leq T$



- $\hat{\beta}_L$ also called soft shrinkage

Why is it called regularized fit?

LS: $H = X(X'X)^{-1}X'$, $X = UDV'$

$$\hat{y} = X\hat{\beta} = X(X'X)^{-1}X'y = \underset{n\times p}{UU'}y = \left(\sum_{k=1}^{p} u_k u_k'\right) y$$

'Complexity' of projection (regression) H

- Look at its trace

$$Tr\{H\} = Tr\{X(X'X)^{-1}X'\} = Tr\{(X'X)(X'X)^{-1}\}$$
$$= Tr\{I_p\} = p$$

- p is the degrees of freedom of the LS regression line fit

Ridge: $H_R = X(X'X + rL)^{-1}X'$

$$\hat{y}_R = X\hat{\beta}_R = X(X'X+rI)^{-1}X'y = X(VD^2V'+rL)^{-1}X'y$$
$$= X(V(D^2+rI)V')^{-1}X'y$$
$$= UDV'V(D^2+rI)^{-1}VV'DU'y$$
$$= UD(D^2+rL)^{-1}DU'y$$
$$= \sum_{k=1}^{p}\left(\frac{d_k^2}{d_k^2+r}\right) u_k u_k' y$$

$$Tr\{H_R\} = \sum_{k=1}^{p}\frac{d_k^2}{d_k^2+r} < Tr\{H\} \quad \text{if } r>0$$

- so [illegible] (df) of ridge model < [illegible] of LS fit $\Rightarrow$ regularized fit

* So ridge regression — computes coeffint in coordinate system defined by eigenvectors, but then also shrinks them by a factor $\frac{d_k^2}{d_k^2 + r}$.

* The shrinkage will have most effect in the directions of small variance d_k^2 - alt to PCReg!! (like we drop small eigenvalue vectors)

Recap

* Collinearities in the data (correlated x_j) can be problematic → either choose a subset of x_j not highly correlated (A)
→ use a regularized fitting technique

* Reminder, other things to consider
- omitted variables
- transformations ($\sqrt{\ }$, log, $\frac{1}{(\ )}$, ...)

(A) Subset selection? — either by knowledge/understanding of the data
\ selection methods, model search