Administration

- Homework 1 available Thursday
- Discussion of project requirements on Thursday
- NSERC summer undergraduate awards
- Fields-MITACS undergraduate summer research

  `http://www.fields.utoronto.ca/programs/scientific/10-11/summer-research/`

Geometric view of least squares fitting

- $\hat{\beta} = (X^T X)^{-1} X^T y$
- $\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p)$
- $\hat{\beta}_p$ can be obtained by a series of regressions (projections) as outlined in algorithm 3.1 on p.54

  regress $\mathbf{x}_1$ on 1, get coefficient $\hat{\gamma}_{01}$, form residual $z_1 = \mathbf{x}_1 - \hat{\mathbf{x}}_1$

  regress $\mathbf{x}_2$ on 1, z_1, get coefs $\hat{\gamma}_{02}, \hat{\gamma}_{12}$, form residual $z_2 = \mathbf{x}_2 - \hat{\gamma}_{02} 1 - \hat{\gamma}_{12} z_1$

  $\vdots$

  regress $\mathbf{x}_p$ on $z_{p-1}, z_{p-2}, \ldots, z_1, 1$ to get $z_p = \mathbf{x}_p - \hat{\mathbf{x}}_p$

  regress y on z_p to get $\hat{\beta}_p$
- illustration on prostate training data – see `prostateRsession.txt`

QR Decomposition $X = Z\Gamma$

- matrix representation $X = Z\Gamma$
- Z has columns z_j
- $$\Gamma = \begin{pmatrix} 0 & \hat\gamma_{01} & \hat\gamma_{02} & \cdots & \hat\gamma_{0p} \\ 0 & 0 & \hat\gamma_{12} & \cdots & \hat\gamma_{1p} \\ & & \ddots & & \\ & & & & \hat\gamma_{p-1,p} \\ 0 & \dots & & & 0 \end{pmatrix}$$
- Let $D_{jj} = (z_j^T z_j)^{1/2} = ||z_j||$ and $D = \mathrm{diag}\,(D_{jj})$ dimension?
- $X = ZD^{-1}D\Gamma = QR$
- $Q^TQ = I, \quad R$ is upper triangular
- $\hat\beta = R^{-1}Q^Ty$
- $\hat y = QQ^Ty$ check `pr.lm$qr`
- Gram-Schmidt

Singular value decomposition, $X = UDV$

- X assumed to be centered, so all columns add to zero (p.64, l.-4)
- U is $N \times p$, V is $p \times p$
- D is diagonal, $d_1 \geq d_2 \geq \cdots \geq d_p \geq 0$ not the same D
- $$X^T X = V D^2 V^T$$

  eigendecomposition of $X^T X$ and of $NS = XX^T$
- $\hat{y} = X\hat{\beta} = UU^T y$

  $= QQ^T y$ different orthogonal bases for ...
- define

  $$\mathrm{z}_1 = X v_1 = u_1 d_1$$
- note that $\mathrm{var}(\mathrm{z}_1) = d_1^2 / N$
- z_1 is the derived variable with the largest variance: the first principal component of X
- z_2 has the largest variance among linear combinations orthogonal to z_1

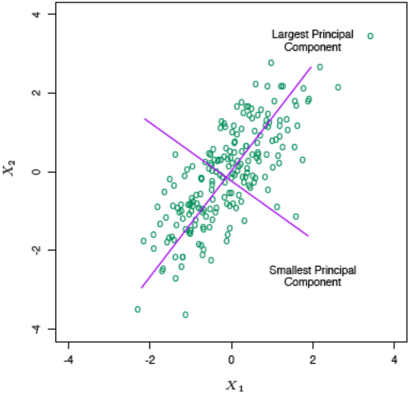


FIGURE 3.9. *Principal components of some input data points. The largest principal component is the direction that maximizes the variance of the projected data, and the smallest principal component minimizes that variance. Ridge regression projects* $\mathbf{y}$ *onto these components, and then shrinks the coefficients of the low–variance components more than the high-variance components.*

... singular value decomposition

- $$\begin{aligned} X_{N\times p} &= U_{N\times p} D_{p\times p} V^T_{p\times p}, \quad U^T U = I, \quad V^T V = I, \\ & \qquad D = \mathrm{diag}(d_1, \dots d_p) \end{aligned}$$

$$\begin{aligned} X\hat{\beta}_{LS} &= X(X^T X)^{-1} X^T y \\ &= UDV^T (VDU^T UDV^T)^{-1} VDU^T y \\ &= UDV^T V^{T^{-1}} D^{-2} V^{-1} VDU^T y \\ &= UU^T y = \Sigma_{j=1}^p u_j u_j^T y \end{aligned}$$

`svd(model.matrix(pr.lm))`, for example

Model selection: subsets (§3.3.1)

- linear regression: forward, backward, stepwise, all possible subsets regression
- $RSS(\hat\beta) = (y - X\hat\beta)^T(y - X\hat\beta) \qquad = \Sigma(y_i - \hat y_i)^2$
  Figure 3.5 this is called SSE in the 302 text
- if we add a regressor, say from X_{p-1} to X_p, $RSS(\hat\beta)$ necessarily decreases
- forward (stepwise) selection starts with one predictor (usually the constant term) and stops when no additional predictor is statistically significant `step(pr.lm, direction = "forward", ...)`
- backward (stepwise) selection starts with all predictors and deletes least significant . . . `direction = "backward"...`
- stepwise selection checks at each stage whether or not to add variables back in `direction = "both"`

... subset selection

- forward stagewise: a “slow” version of forward stepwise, in which coefficients are not re-computed
- all possible subsets regression considers all 2^p models.
- for $p < 30$, feasible with the “leaps and bounds” algorithm, implemented in package `leaps` (See Figure 3.6), also `regsubsets`
- Figure 3.7
- huh?
- 10-fold cross-validation
- model selection related to expected prediction error: theory to come in Ch. 7

"Mallows' C_p"

- a common adjustment to measure benefit of adding further parameters:

$$C_p = \frac{RSS_p}{\sigma^2} + 2p - N \quad \text{if } \sigma^2 \text{ is known;}$$

- or an estimate of this, if σ^2 is unknown
- rule of thumb: choose p so that C_p is small and $C_p \simeq p$
- can be shown to be a good choice for prediction (details deferred until Chapter 7)
- a closely related, more general, criterion *AIC* (Akaike's Information Criterion)
- for our linear model

$$AIC \approx N \log(RSS_p/N) + 2p + constant$$

- `stepAIC` in `MASS` library; `step` in `base`

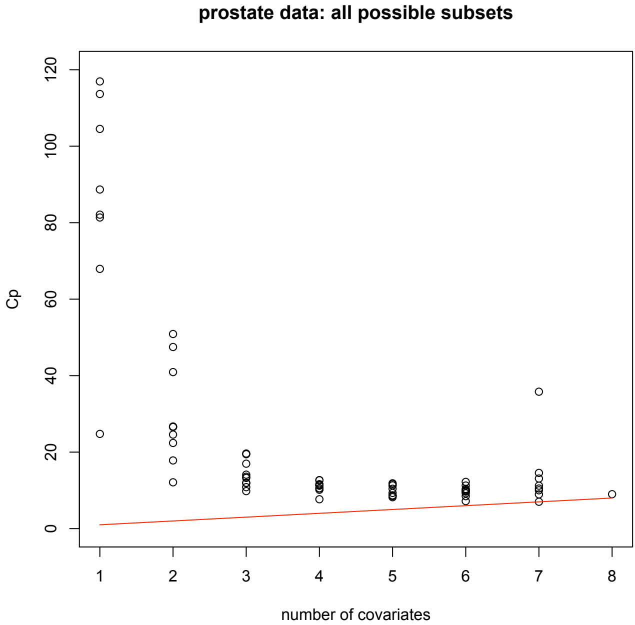

prostate data: all possible subsets
Cp
0
20
40
60
80
100
120
1
2
3
4
5
6
7
8
number of covariates

Figure 3.5

step

```
> step(pr.lm)
Start:  AIC=-37.13
lpsa ˜ lcavol + lweight + age + lbph + svi + lcp + gleason +
    pgg45

          Df Sum of Sq    RSS     AIC
- gleason  1     0.011 29.437 -39.103
<none>                 29.426 -37.128
- age      1     0.989 30.415 -36.914
- pgg45    1     1.532 30.959 -35.727
- lcp      1     1.768 31.195 -35.218
- lbph     1     2.144 31.571 -34.415
- svi      1     3.093 32.520 -32.430
- lweight  1     3.839 33.265 -30.912
- lcavol   1    14.610 44.037 -12.118

Step:  AIC=-39.1
lpsa ˜ lcavol + lweight + age + lbph + svi + lcp + pgg45

          Df Sum of Sq    RSS     AIC
<none>                 29.437 -39.103
- age      1     1.102 30.540 -38.639
- lcp      1     1.758 31.196 -37.216
- lbph     1     2.135 31.573 -36.411
- pgg45    1     2.376 31.813 -35.903
- svi      1     3.166 32.604 -34.258
- lweight  1     4.005 33.442 -32.557
- lcavol   1    14.887 44.325 -13.681

Call:
lm(formula = lpsa ˜ lcavol + lweight + age + lbph + svi + lcp +     pgg45, data = train)

Coefficients:
```

Shrinkage Methods: (§3.4)

- Ridge regression
-

$$\begin{aligned}\hat{\beta}_{LS} &= (X^TX)^{-1}X^Ty \\ \hat{\beta}_{ridge} &= (X^TX+\lambda I)^{-1}X^Ty\end{aligned}$$

- can show that $\hat{\beta}_{ridge}$ satisfies

$$\min_{\beta}\left(\Sigma\{y_i-\beta_0-\Sigma_{j=1}^{p}x_{ij}\beta_j\}^2+\lambda\Sigma_{j=1}^{p}\beta_j^2\right)$$

$$\min_{\beta}\Sigma\{y_i-\beta_0-\Sigma_{j=1}^{p}x_{ij}\beta_j\}^2 \quad \text{s.t. } \Sigma\beta_j^2\le t$$

- Assume x_j's are centered and put these in matrix X (with no column of 1's:

$$\min_{\beta}(y-X\beta)^T(y-X\beta) \qquad \text{s.t. } ||\beta||^2\le t$$

... ridge regression

- $\min_\beta\{(y - X\beta)^T(y - X\beta) + \lambda||\beta||^2\}$
- λ is a tuning parameter: $\lambda = 0$ gives $\hat{\beta}_{LS}$, $\lambda \to \infty$

Figure 3.8

- in R the `library MASS` `library(MASS )` has a ridge regression version of `lm` called `lm.ridge`
- if columns of X are nearly linearly dependent (multicollinearity), $\hat{\beta}$'s for these columns should be shrunk towards 0.
- essential that the predictors are all scaled to the same units
- this is difficult for interpretation of the coefficients

$$\begin{aligned}
X\hat{\beta}_{ridge} &= X(X^TX+\lambda I)^{-1}X^Ty \\
&= UDV^T(VD^2V^T+\lambda I)^{-1}VDU^Ty \\
&= UDV^T(VD^2V^T+\lambda VV^T)^{-1}VDU^Ty \\
&= UD(D^2+\lambda I)^{-1}DU^Ty \\
&= \Sigma_{j=1}^p u_j(\frac{d_j^2}{d_j^2+\lambda})u_j^Ty
\end{aligned}$$

$$df(\lambda) = \mathrm{tr}[X(X^TX+\lambda I)^{-1}X^T] = \Sigma_{j=1}^p \frac{d_j^2}{d_j^2+\lambda}$$

Figure 3.7

$df(\lambda)$ called effective number of parameters in Ch. 7

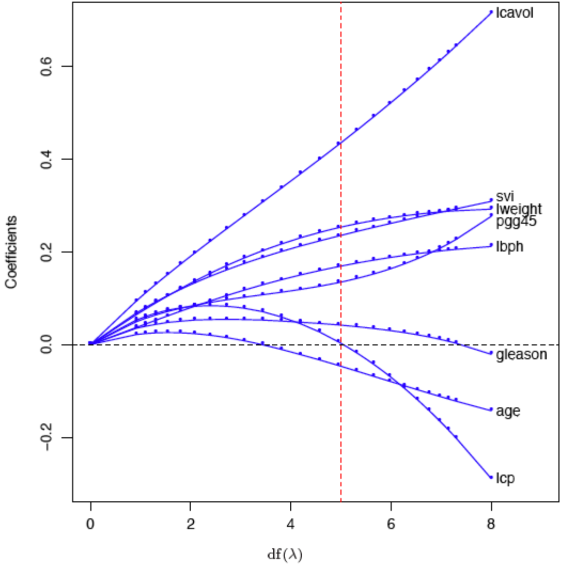


FIGURE 3.8. *Profiles of ridge coefficients for the prostate cancer example, as the tuning parameter* λ *is varied. Coefficients are plotted versus* $\mathrm{df}(\lambda)$*, the effective degrees of freedom. A vertical line is drawn at* $\mathrm{df} = 5.0$*, the value chosen by cross-validation.*

Lasso

- $$\min_{\beta}\left(\Sigma\{y_i - \beta_0 - \Sigma_{j=1}^{p} x_{ij}\beta_j\}^2 + \lambda\Sigma_{j=1}^{p}|\beta_j|\right)$$
- $$\min_{\beta}\Sigma\{y_i - \beta_0 - \Sigma_{j=1}^{p} x_{ij}\beta_j\}^2 \quad \text{s.t. } \Sigma|\beta_j| \le t$$
- quadratic programming problem
- $\hat{\beta}^{lasso}$ is nonlinear function of y
- Figure 3.10
- Table 3.3

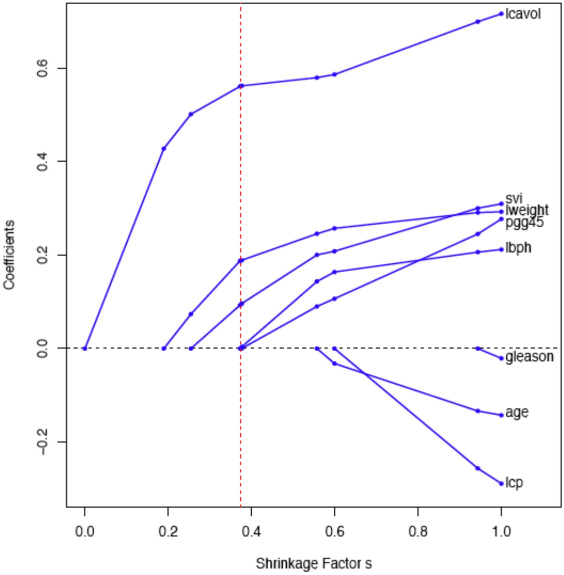


FIGURE 3.10. *Profiles of lasso coefficients, as the tuning parameter t is varied. Coefficients are plotted versus $s = t/\sum_1^p |\hat{\beta}_j|$. A vertical line is drawn at $s = 0.36$, the value chosen by cross-validation. Compare Figure 3.8 on page 65; the lasso*

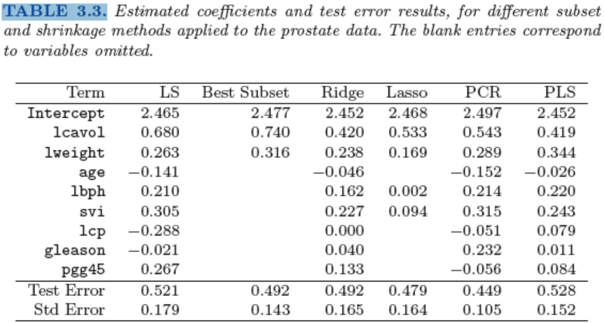


TABLE 3.3. *Estimated coefficients and test error results, for different subset and shrinkage methods applied to the prostate data. The blank entries correspond to variables omitted.*

Term	LS	Best Subset	Ridge	Lasso	PCR	PLS
`Intercept`	2.465	2.477	2.452	2.468	2.497	2.452
`lcavol`	0.680	0.740	0.420	0.533	0.543	0.419
`lweight`	0.263	0.316	0.238	0.169	0.289	0.344
`age`	−0.141		−0.046		−0.152	−0.026
`lbph`	0.210		0.162	0.002	0.214	0.220
`svi`	0.305		0.227	0.094	0.315	0.243
`lcp`	−0.288		0.000		−0.051	0.079
`gleason`	−0.021		0.040		0.232	0.011
`pgg45`	0.267		0.133		−0.056	0.084
Test Error	0.521	0.492	0.492	0.479	0.449	0.528
Std Error	0.179	0.143	0.165	0.164	0.105	0.152

...Lasso

- in Table 3.3 each method had a tuning parameter to choose; they used cross-validation within the training data
- in `lm.ridge` you can extract a component called `$GCV`
- the quantity $\Sigma d_j^2/(d_j^2 + \lambda)$ has an interpretation as the number of 'degrees of freedom' or number of 'parameters' used by the ridge regression fit
- book says that the best value is 4.16, which corresponds to quite a large λ (39); the GCV criterion chooses $\lambda = 5$
- analysis of lasso more difficult; note Figure 3.10 plotted against $t/\Sigma|\hat{\beta}_j|$

... smoothing

- ridge regression gives “proportional shrinkage”
- subset selection gives “hard thresholding” (some $\beta_j \rightarrow 0$)
- lasso gives “soft thresholding”: blend of shrinkage and zeroing (Figure 3.10 and Figure 3.11)
- Least Angle Regression (LAR): combine forward stagewise regression with the lasso
- related to the Dantzig selector (Candes and Tao, AS 2007)

Mean squared error of prediction in linear models

Let $\tilde{\beta} = \tilde{\beta}(y)$ be a competing estimator of β (not $\hat{\beta}$, the LS estimator). Using $\tilde{\beta}$ for prediction would give $\tilde{y}_0 = x_0^T \tilde{\beta}$, where $x_0^T = (1, x_{01}, \ldots, x_{0p})$ is the new value of the inputs. The expected prediction error is

$$
\begin{aligned}
E(\tilde{y}_0 - y_0)^2 &= E(x_0^T \tilde{\beta} - y_0)^2 \quad \text{(by definition)} \\
&= E(y_0 - x_0^T \beta + x_0^T \beta - x_0^T \tilde{\beta})^2 \\
&= \mathrm{var}(y_0) + E(x_0^T \tilde{\beta} - x_0^T \beta)^2 \quad \text{(why is cross prod 0?)} \\
&= \sigma^2 + E\{x_0^T (\tilde{\beta} - \beta)(\tilde{\beta} - \beta)^T x_0\} \\
&= \sigma^2 + x_0^T E\{(\tilde{\beta} - \beta)(\tilde{\beta} - \beta)^T\} x_0 \\
&= \sigma^2 + x_0^T E\{(\tilde{\beta} - E\tilde{\beta} + E\tilde{\beta} - \beta)(\tilde{\beta} - E\tilde{\beta} + E\tilde{\beta} - \beta) x_0 \\
&= \sigma^2 + x_0^T [E\{(\tilde{\beta} - E\tilde{\beta})(\tilde{\beta} - E\tilde{\beta})^T\} + (E\tilde{\beta} - \beta)(E\tilde{\beta} - \beta)^T] x_0 \\
&= \sigma^2 + x_0^T \{\mathrm{cov}(\tilde{\beta}) + \mathrm{bias}^2(\tilde{\beta})\} x_0
\end{aligned}
$$

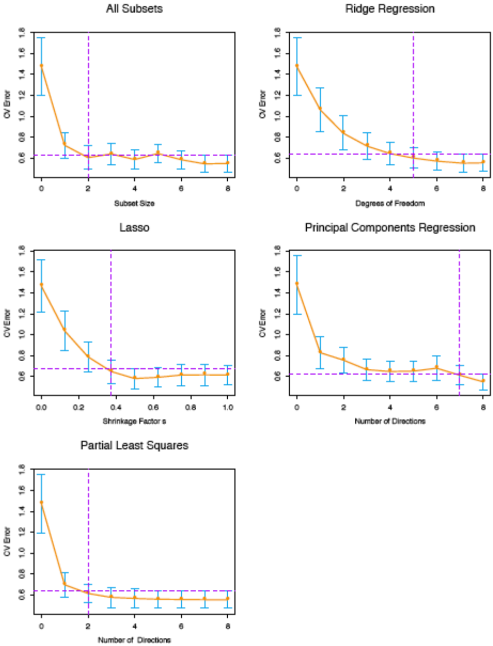
The first term, σ^2, is unavoidable. The next two terms together are the Mean Squared Error (MSE) of the prediction $\tilde{y}_0$, and are shown here to be a function of x_0 and the MSE of $\tilde{\beta}$. If $\tilde{\beta}$ is unbiased, i.e. $E\tilde{\beta} = \beta$, then we only need to worry about the covariance terms. Estimates of β obtained by ridge regression, Lasso, and LARS are all biased. This could be useful if the variance is decreased enough to give smaller MSE.

Derived features §3.5

- replace $\mathbf{x}_1, \ldots \mathbf{x}_p$ with linear combinations of columns
- principal components from SVD are natural candidates
- $X = UDV$
- $\mathrm{z}_m = Xv_m, \quad m = 1, \ldots, M < p$
- z_m are orthogonal by construction
- $$\hat{y}^{pcr}_{(M)} = \bar{y}\mathbf{1} + \sum_{m=1}^{M} \hat{\theta}_m \mathrm{z}_m$$
- $$\hat{\theta}_m = \frac{\langle \mathrm{z}_m, y \rangle}{\langle \mathrm{z}_m, \mathrm{z}_m \rangle}$$
- inputs should be scaled first (mean 0, variance 1)
- Figure 3.17

... derived features

- closely related method Partial least squares
- also constructs derived variables
- widely used in chemometrics, where often $p > N$
- see §3.6 for discussion


All Subsets
CV Error
0.6
0.8
1.0
1.2
1.4
1.6
1.8
0
2
4
6
8
Subset Size
Ridge Regression
CV Error
Degrees of Freedom
Lasso
CV Error
0.0
0.2
0.4
0.6
0.8
1.0
Shrinkage Factor s
Principal Components Regression
CV Error
Number of Directions
Partial Least Squares
CV Error
Number of Directions

Thursday and next week

- more on ridge regression and lasso in R
- construction of Table 3.3 in R test set error
- discussion of project
- HW 1 will be available
- Next week Chapter 4: §4.1 to §4.4