HW3 Solution, ECE7252, February 13, 2008

1. Let $\{(x_i, y_i), i = 1, ..., N\}$ be N pairs of training data for linear regression (no need to do the exercise for part on nearest neighbor regressions). Let's also assume without a loss of generosity that the intercept is zero, i.e. x_i has zero mean as well as y_i. Let $\mathbf{X} = \{x_1, ..., x_i, ..., x_N\}$ and $\mathbf{Y} = \{y_1, ..., y_i, ..., y_N\}$. Then it is clear that for linear regression $y = f(x) + \epsilon = x\beta + \epsilon$, then the LS estimate for β can be expressed as $\hat{\beta} = \frac{\sum_{i=1}^N x_i y_i}{\sum_{j=1}^N x_j^2}$.

   (a) Now for a new x_0, the predicted $\hat{y}_0 = x_0\hat{\beta}$, or [refer to Eq. (3.20)],

$$\hat{y}_0 = \hat{f}(x_0) = \sum_{i=1}^N [\frac{x_i x_0}{\sum_{j=1}^N x_j^2}] y_i = \sum_{i=1}^N [l_i(x_0, \mathbf{X})] y_i, \tag{1}$$

   with the weights $l_i(x_0, \mathbf{X})$ independent of $\mathbf{Y}$.

   (b) Given $\mathbf{X}$ the conditional density of $y_i = f(x_i) = x_i\beta$ is zero mean with variance σ^2. So

$$E_{\mathbf{Y}|\mathbf{X}}[f(x_0) - \hat{f}(x_0)]^2 = E_{\mathbf{Y}|\mathbf{X}}[\hat{f}(x_0) - E_{\mathbf{Y}|\mathbf{X}}(\hat{f}(x_0))]^2 + E_{\mathbf{Y}|\mathbf{X}}[E_{\mathbf{Y}|\mathbf{X}}(\hat{f}(x_0)) - f(x_0)]^2, \tag{2}$$

   or in the case of linear regression as explained above, $E_{\mathbf{Y}|\mathbf{X}}(y_i) = x_i\beta$ and $E_{\mathbf{Y}|\mathbf{X}}(\hat{f}(x_0)) = x_0\beta$. Now let $q_i = \frac{x_i}{\sum_{j=1}^N x_j^2}$

$$E_{\mathbf{Y}|\mathbf{X}}[f(x_0) - \hat{f}(x_0)]^2 = E_{\mathbf{Y}}[x_0(\hat{\beta} - \beta)]^2 + 0, \tag{3}$$

   or since x_is are i.i.d. that $E[y_i y_j] = E[y_j]E[y_j] = \beta^2 x_i x_j$ if $i \neq j$, we have

$$E_{\mathbf{Y}|\mathbf{X}}[f(x_0) - \hat{f}(x_0)]^2 = x_0^2[\sum_{i=1}^N E_{\mathbf{Y}|\mathbf{X}}\{q_i^2(y_i - \beta x_i)^2\}] = \frac{x_0^2\sigma^2}{\sum_{j=1}^N x_j^2}. \tag{4}$$

   (c) Now for the join expectation, we have

$$E_{\mathbf{Y},\mathbf{X}}[f(x_0) - \hat{f}(x_0)]^2 = E_{\mathbf{X}}[E_{\mathbf{Y}|\mathbf{X}}[f(x_0) - \hat{f}(x_0)]^2 = \sigma^2 E_{\mathbf{X}}[\frac{x_0^2}{\sum_{j=1}^N x_j^2}]. \tag{5}$$

   (d) In the situation of linear regression, the bias is zero in both cases in (c) and (d), i.e. $E_{\mathbf{Y}|\mathbf{X}}(\hat{\beta}) = \beta$, while the variance in (c) depends on the values of the regressors but the variance of (d) depends on the density of x_i, $h(x)$. Sometimes the variance in (d) may be difficult to compute. However in some special cases we may still get a close-form expression for the expectation.

2. If β_is are random variables the posterior density of $\beta^T = [\beta_0, \beta_1, ..., \beta_p]$ after observing $\mathbf{Y} = \{y_1, ..., y_i, ..., y_n\}$ is $f(\beta|\mathbf{Y}) = [\prod_{i=1}^{N} f(y_i|\beta)]f(\beta)/f(\mathbf{Y})$. Since $f(\mathbf{Y})$ is a constant after observing $\mathbf{Y}$, and the likelihood $f(\mathbf{Y}|\beta)$ i.i.d. y_i and the prior density $f(\beta)$ for independent β_i are simply:

$$f(\mathbf{Y}|\beta) = \prod_{i=1}^{N}[\frac{1}{\sqrt{2\pi\sigma^2}}\exp[-\frac{1}{2\sigma^2}(y_i - \beta_0 - \sum_{k=1}^{p} x_{ik}\beta_k)^2]. \tag{6}$$

and

$$f(\beta) = \prod_{j=0}^{p}[\frac{1}{\sqrt{2\pi\tau^2}}\exp(-\frac{1}{2\tau^2}\beta_j^2). \tag{7}$$

Therefore the log posterior density can be expressed as

$$\log f(\beta|\mathbf{Y}) = K - \sum_{i=1}^{N}[\frac{1}{2\sigma^2}(y_i - \beta_0\sum_{k=1}^{p} x_{ik}\beta_k)^2] - \sum_{j=0}^{p}[\frac{1}{2\tau^2}\beta_j^2], \tag{8}$$

where K is a constant. Let $\lambda = \sigma^2/\tau^2$, then we finally have

$$\log f(\beta|\mathbf{Y}) = K - \frac{1}{2\sigma^2}\{\sum_{i=1}^{N}[(y_i - \beta_0\sum_{k=1}^{p} x_{ik}\beta_k)^2] + \lambda[\sum_{j=0}^{p}\beta_j^2]\}. \tag{9}$$

3. Again let us just assume $\beta_0 = 0$ without a loss of generosity. Let $A = [X^T|\sqrt{\lambda}\mathbf{I}_p]^T$, and $\mathbf{q} = [\mathbf{y}^T|\mathbf{0}_p^T]^T$ with I_p a $p * p$ identity matrix, and $\mathbf{0}_p$ a $p * 1$ zero vector. Then A is a $(N + p) * p$ matrix and $\mathbf{q}$ is a $(N + p) * 1$ vector. By regressing $\mathbf{q}$ on A, we have a modified LS solution β^* such that

$$\hat{\beta}^* = (A^T A)^{-1}A^T\mathbf{q} \tag{10}$$

Since

$$A^T\mathbf{q} = X^T\mathbf{y}, \tag{11}$$

and

$$A^T A = X^T X + \lambda\mathbf{I}_p, \tag{12}$$

we arrive at

$$\hat{\beta}^* = (X^T X + \lambda\mathbf{I}_p)^{-1}X^T\mathbf{y} = \hat{\beta}^{ridge}. \tag{13}$$

4. For principle component regression, we derive a new set of regressors $z_m = Xv_m$, with v_m being orthonornal, and then regress $\mathbf{y}$ on $z_i, ..., z_M$, $M \leq p$, so we have Eq. (3.52) in HTF assuming $\bar{y} = 0$, we have Eq. (3.53) clearly shown in the following expression:

$$\hat{\mathbf{y}}^{pcr} = \sum_{m=1}^{M} \hat{\theta}_m z_m = X \sum_{m=1}^{M} \hat{\theta}_m v_m = X\hat{\beta}^{pcr}. \tag{14}$$

Now in the case of $M = p$, we have

$$\hat{\beta}^{pcr} = \sum_{m=1}^{p} \hat{\theta}_m v_m = V\hat{\theta}, \tag{15}$$

where the orthonormal matrix $V = [v_1|...|v_p]$, and $\hat{\theta}^T = [\hat{\theta}_1, ..., \hat{\theta}_p]$. Now since $< z_m, z_l >= 0$ if $l \neq m$, if we let $Z = [z_1|...|z_p]$, then we have

$$\hat{\theta} = (Z^T Z)^{-1} Z^T \mathbf{y}. \tag{16}$$

Now with singular value decomposition we have $X = UDV^T$. With $Z = XV = UD$, because $V^{-1} = V^T$, so we have $Z^T = V^T X^T$. Similarly $Z^T Z = V^T (X^T X)V$, so we have

$$(Z^T Z)^{-1} = V^T (X^T X)^{-1} V, \tag{17}$$

Now we are ready to derive

$$\hat{\beta}^{pcr} = V[V^T (X^T X)^{-1} V][V^T X^T]\mathbf{y} = (X^T X)^{-1} X^T \mathbf{y} = \hat{\beta}_{LS}. \tag{18}$$

5.

$$h(n) = \sum_{k=1}^{p} \alpha_k h(n-k) + G\delta(n)$$

$$\tilde{R}(m) = \sum_{n=0}^{\infty} h(n)h(n+m)$$

a) $$\tilde{R}(-m) = \sum_{n=0}^{\infty} h(n)h(n-m)$$

let $m' = n - m$

$$\tilde{R}(-m) = \sum_{m'=-n}^{\infty} h(m+m')h(m')$$

since $h(m')$ is causal, $h(m') = 0,\ m' < 0 \Rightarrow$

$$\tilde{R}(-m) = \sum_{m'=0}^{\infty} h(m+m')h(m') = \tilde{R}(m)$$

b) $$\tilde{R}(-m) = \sum_{n=0}^{\infty} h(n)h(n-m) = \tilde{R}(m)$$

$$= \sum_{n=0}^{\infty} \left[\sum_{k=1}^{p} \alpha_k h(n-k) + G\delta(n) \right] \left[\sum_{k=1}^{p} \alpha_l h(n-m-l) + G\delta(n-m) \right]$$

$$= \sum_{n=0}^{\infty} \left\{ \sum_{k=1}^{p} \sum_{l=1}^{p} \alpha_k \alpha_l h(n-k)h(n-m-l) + G\sum_{k=1}^{p} \alpha_k h(n-k)\delta(n-m) + G\sum_{l=1}^{p} \alpha_l h(n-m-l)\delta(n) + G^2\delta(n)\delta(n-m) \right\}$$

assume $m \geq 0$

$$\tilde{R}(m) = \sum_{n=0}^{\infty} \sum_{k=1}^{p} \sum_{l=1}^{p} \alpha_k \alpha_l h(n-k)h(n-m-l) + G\sum_{k=1}^{p} \alpha_k h(m-k) + G\sum_{l=1}^{p} \alpha_l h(-m-l)$$

since for $m \geq 0,\ h(-m-l) = 0$ for $l = 1, 2, ..., p$ then

$$\tilde{R}(m) = \sum_{n=0}^{\infty} \sum_{k=1}^{p} \alpha_k h(n-k) \left[\sum_{l=1}^{p} \alpha_l h(n-m-l) + G\delta(n-m) \right]$$

but $$\left[\sum_{l=1}^{p} \alpha_l h(n-m-l) + G\delta(n-m) \right] = h(n-m)$$

thus $$\tilde{R}(m) = \sum_{k=1}^{p} \alpha_k \sum_{n=0}^{\infty} h(n-k)h(n-m)$$

6.

a) $d(n) = x(n) - \tilde{x}(n)$

$$Var\left[d(n)\right] = E\left\{\left[d(n) - E\left[d(n)\right]\right]^2\right\} = E\left[d^2(n)\right] - E^2\left[d(n)\right]$$

where

$$E\left[d^2(n)\right] = E\left\{x^2(n) - 2\alpha x(n)x(n-1) + \alpha^2 x^2(n-1)\right\}$$

$$= E\left[x^2(n)\right] - 2\alpha E\left[x(n)x(n-1)\right] + \alpha^2 E\left[x^2(n-1)\right]$$

$$= \sigma_x^2 - 2\alpha\phi_x(1) + \alpha^2\sigma_x^2$$

$$\therefore \sigma_d^2 = \sigma_x^2\left[1 + \alpha^2 - 2\alpha\frac{\phi_x(1)}{\sigma_x^2}\right]$$

b) To find the minimum, differentiate σ_d^2 with respect to α and equate to zero

$$\frac{\partial\sigma_d^2}{\partial\alpha} = \sigma_x^2\left[2\alpha - 2\frac{\phi(1)}{\sigma_x^2}\right] = 0 \Rightarrow \hat{\alpha} = \frac{\phi_x(1)}{\sigma_x^2}$$

Note that this represents a minimum since the second derivative is positive.

c) $\sigma_{d_{\min}}^2$ is obtained by substituting the value of $\hat{\alpha}$ from part (b)

$$\sigma_{d_{\min}}^2 = \sigma_x^2\left[1 + \frac{\phi_x{}^2(1)}{\sigma_x^4} - 2\frac{\phi_x{}^2(1)}{\sigma_x^4}\right]$$

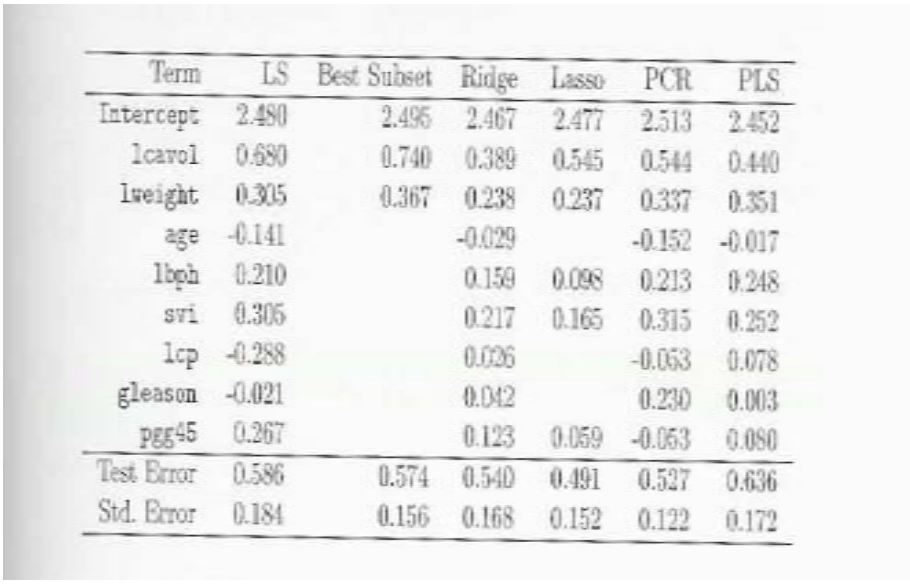
$$= \sigma_x^2\left[1 - \frac{\phi_x{}^2(1)}{\sigma_x^4}\right] = \sigma_x^2\left(1 - \hat{\alpha}^2\right)$$

d) $\sigma_d^2 \le \sigma_x^2 \Rightarrow \sigma_x^2\left[1 + \hat{\alpha}^2 - 2\hat{\alpha}\frac{\phi_x(1)}{\sigma_x^2}\right] \le \sigma_x^2$

$$1 + \hat{\alpha}^2 - 2\hat{\alpha}\frac{\phi_x(1)}{\sigma_x^2} \le 1$$

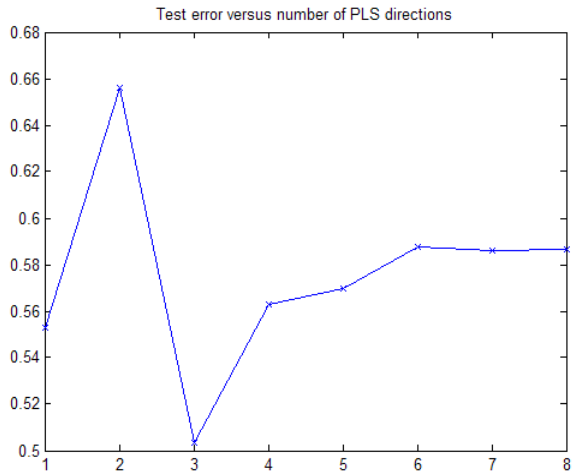
$$\hat{\alpha}^2 - 2\hat{\alpha}\frac{\phi_x(1)}{\sigma_x^2} \le 0,$$

$$\left|\hat{\alpha}\right| \le 2\frac{\phi_x(1)}{\sigma_x^2}.$$



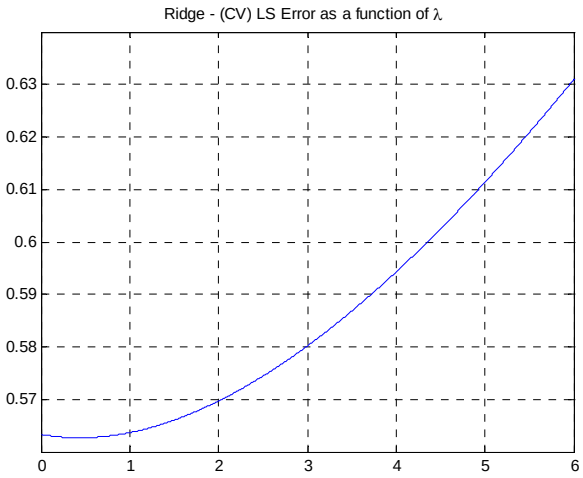
Term	LS	Best Subset	Ridge	Lasso	PCR	PLS
Intercept	2.480	2.496	2.467	2.477	2.513	2.452
lcavol	0.680	0.740	0.389	0.545	0.544	0.440
lweight	0.305	0.367	0.238	0.237	0.337	0.351
age	-0.141		-0.029		-0.152	-0.017
lbph	0.210		0.159	0.098	0.213	0.248
svi	0.305		0.217	0.165	0.315	0.252
lcp	-0.288		0.026		-0.053	0.078
gleason	-0.021		0.042		0.230	0.003
pgg45	0.267		0.123	0.059	-0.053	0.080
Test Error	0.586	0.574	0.540	0.491	0.527	0.636
Std. Error	0.184	0.156	0.168	0.152	0.122	0.172

	LS	Best Subset (lowest error on the test set)	Best Subset (lowest error on CV set)	Verification (when 1st,2nd and 3rd parameters are nonzero)	Ridge (lowest error on the CV test) (λ=0.444)	Ridge (lowest error on the test set)	PC	PLS (best error on test data) (3 PLS directions)	PLS Verification (2 PLS directions)
Intercept	2.4523	2.4523	2.4523	2.4523	2.4362	2.3644	2.4523	2.4523	2.4523
lcavol	0.7110	0.7341	0.7078	0.7740	0.6993	0.6521	0.5663	0.5932	0.4331
lweight	0.2905		0.2929	0.3493	0.2901	0.2879	0.3209	0.3110	0.3578
age	-0.1415		-0.1450		-0.1382	-0.1244	-0.1526	-0.1822	-0.0213
lbph	0.2104		0.2098		0.2095	0.2054	0.2144	0.2033	0.2415
svi	0.3073	0.2234	0.3092		0.3047	0.2940	0.3197	0.3100	0.2574
lcp	-0.2868		-0.2856		-0.2719	-0.2139	-0.0500	-0.0367	0.0852
gleason	-0.0208	0.0289			-0.0163	-0.0003	0.2269	0.0063	0.0061
Pgg45	0.2753		0.2601		0.2663	0.2330	-0.0631	0.1221	0.0837
Std error	0.1839	0.0980	0.1805	0.1563	0.1840	0.1868	0.1211	0.1275	0.1800
Test error	0.5863	0.3828	0.5820	0.5737	0.5823	0.5744	0.5269	0.5036	0.6562



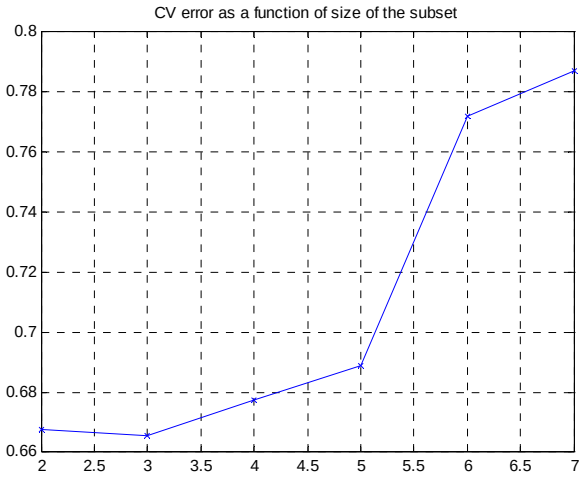
PLS Regression:

The test error was at its minimum when 3 PLS directions were used.



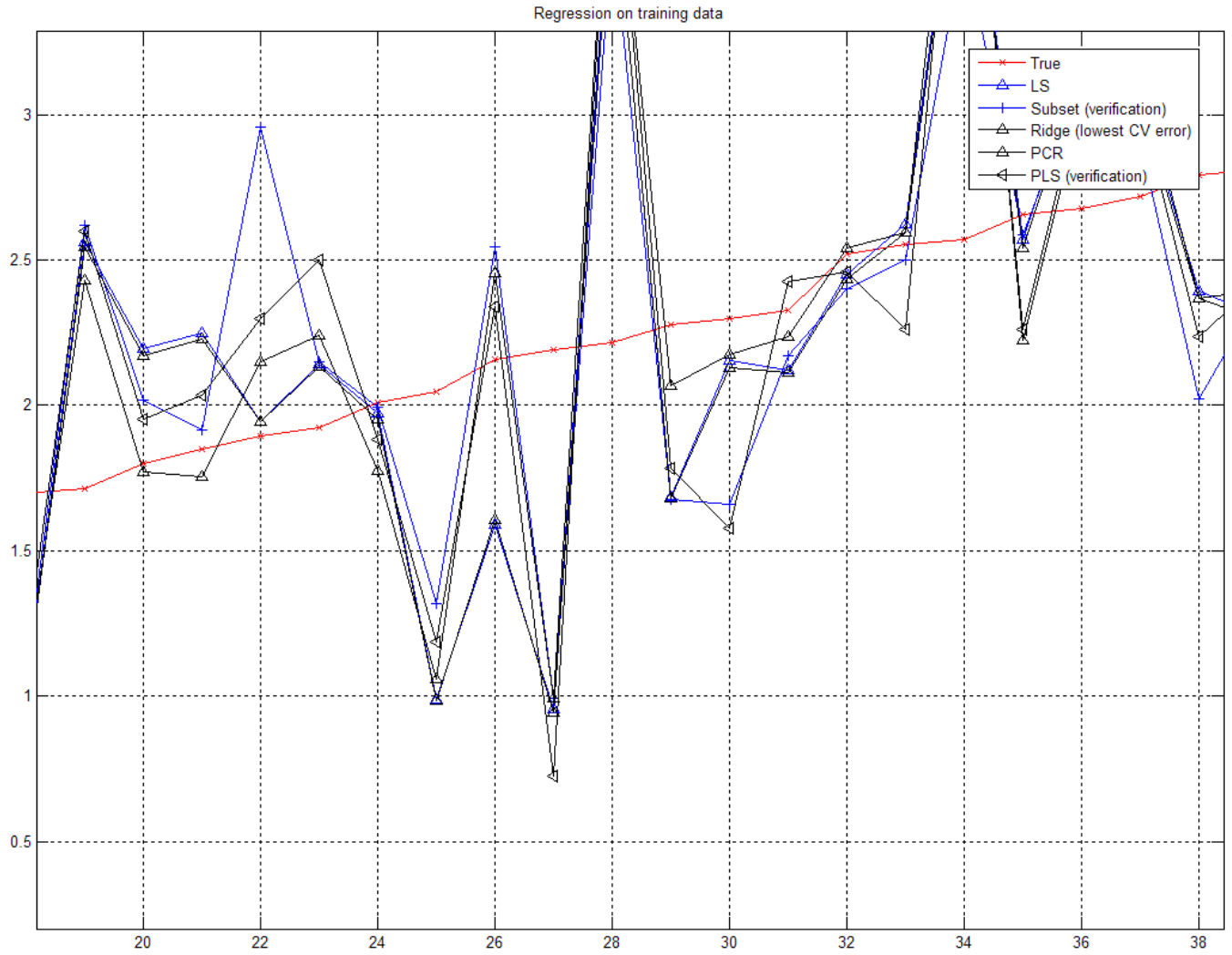
Ridge Regression:

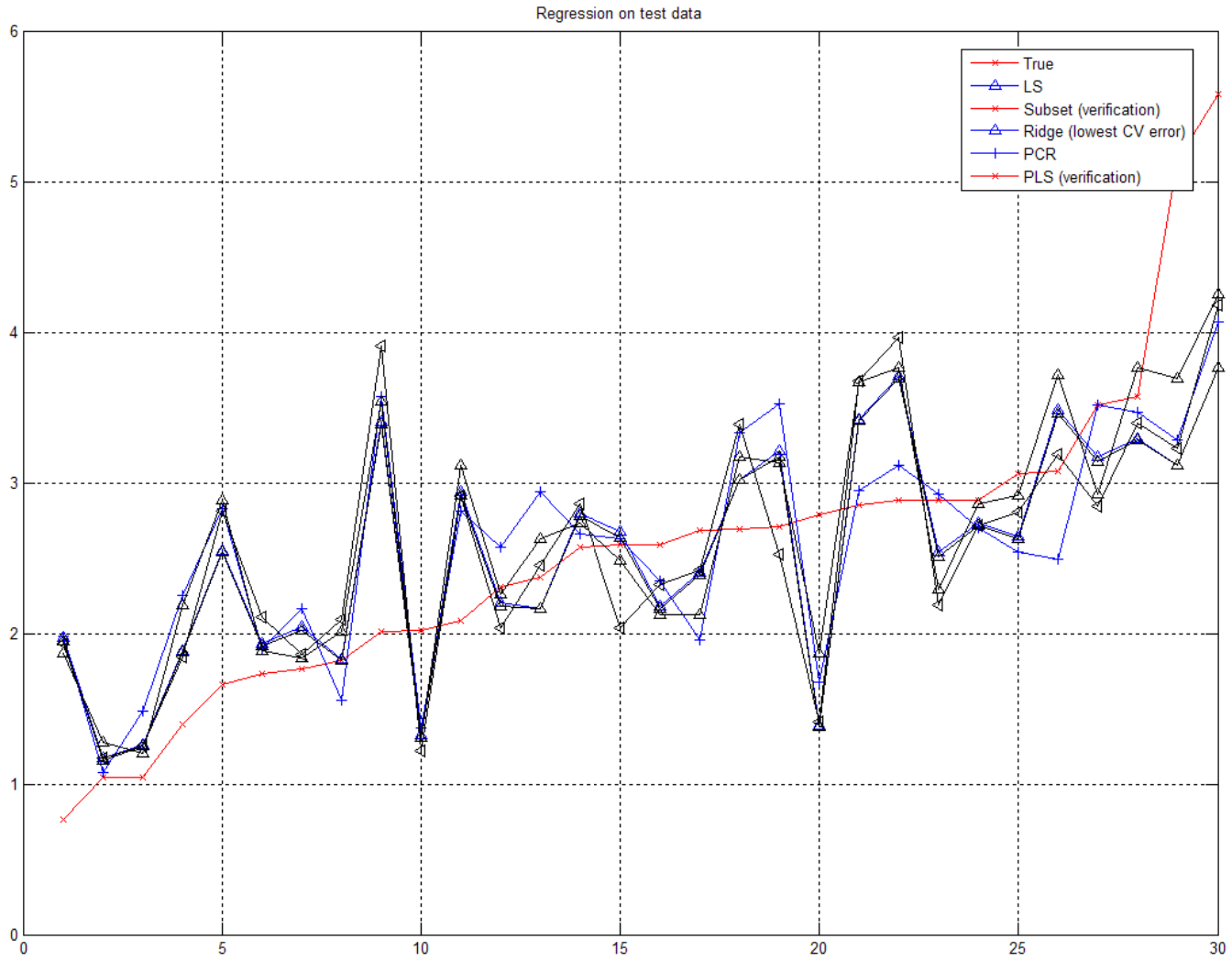
The CV error first reduced then increased with increasing λ.

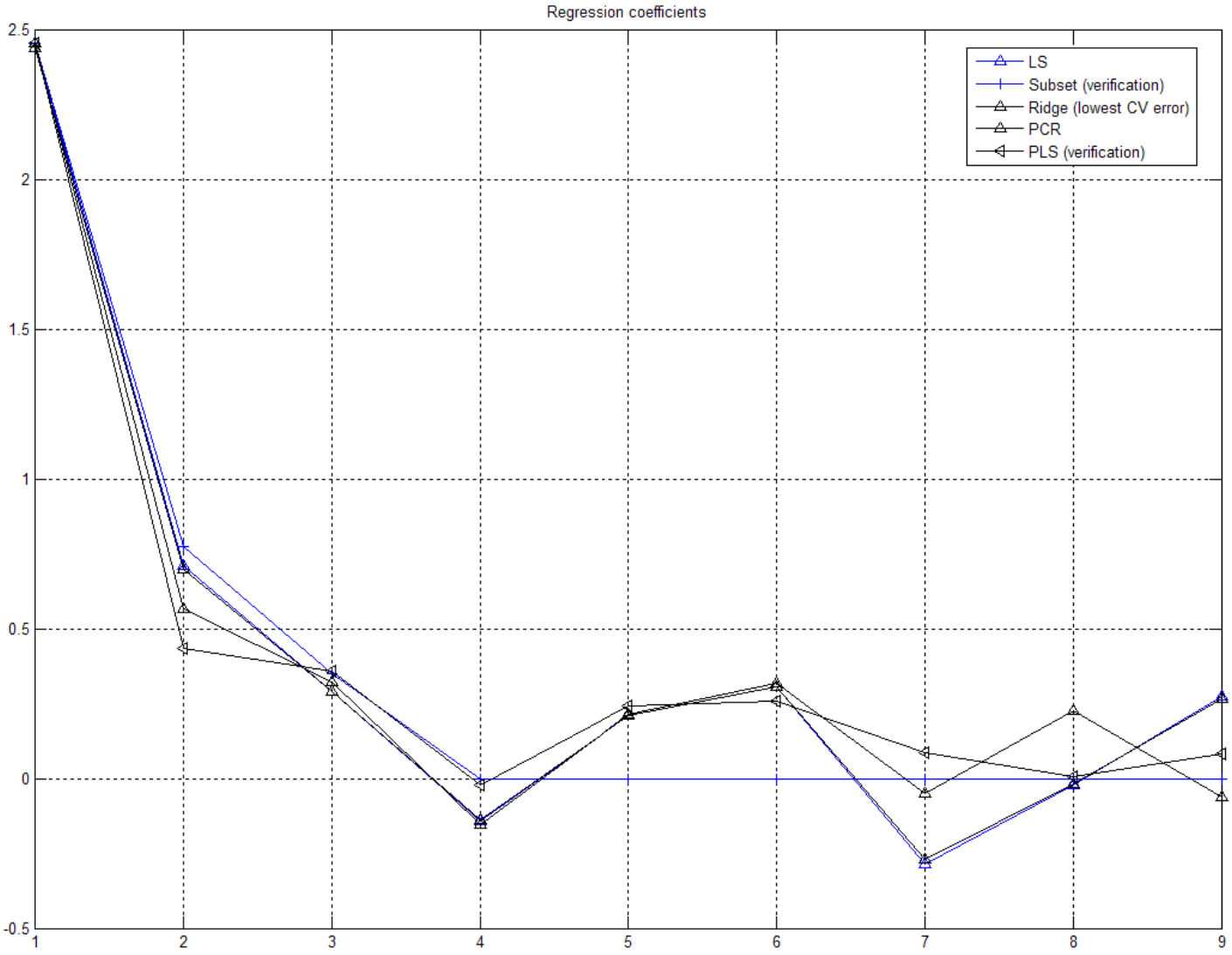


Regression with Subset Selection

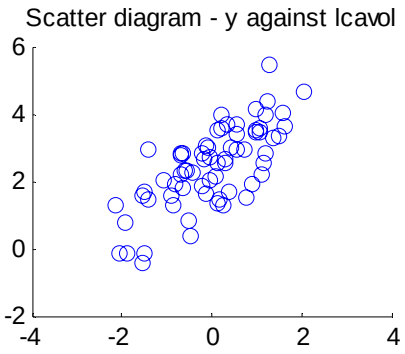
Best subset size was 3. This meant all the coefficients other than the ones for 'lcoval', 'lweight' and 'gleason' were 0.

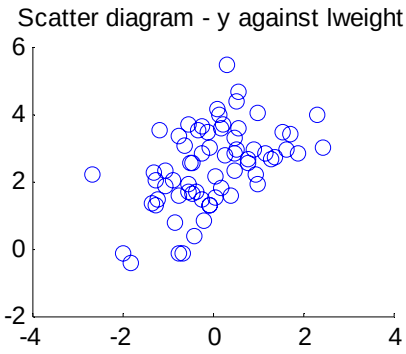

Regression on training data
True
LS
Subset (verification)
Ridge (lowest CV error)
PCR
PLS (verification)
3
2.5
2
1.5
1
0.5
20
22
24
26
28
30
32
34
36
38

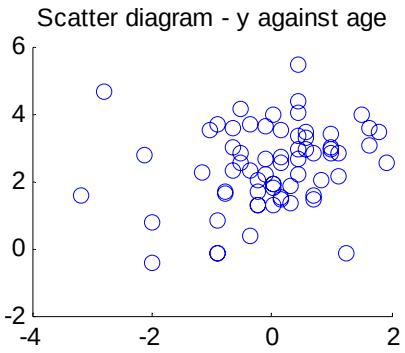

Regression on test data
True
LS
Subset (verification)
Ridge (lowest CV error)
PCR
PLS (verification)
0
1
2
3
4
5
6
0
5
10
15
20
25
30

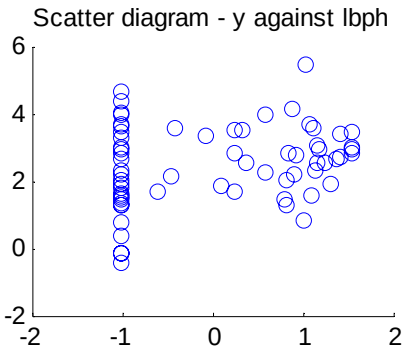

Regression coefficients
LS
Subset (verification)
Ridge (lowest CV error)
PCR
PLS (verification)
2.5
2
1.5
1
0.5
0
-0.5
1
2
3
4
5
6
7
8
9

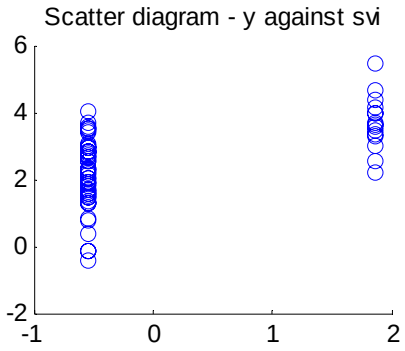
ScatterDiagrams

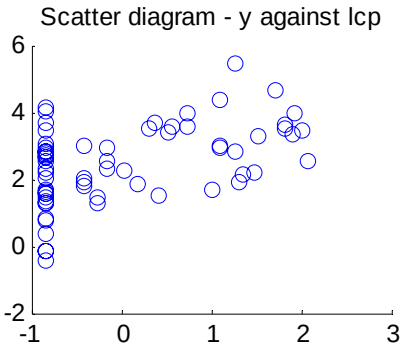


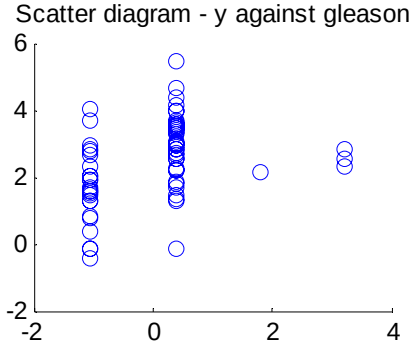


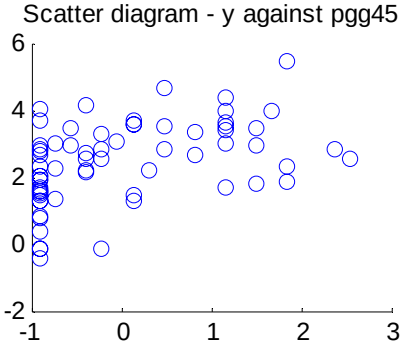












Correlation Coefficients

Corr('lcavol',y) = 0.7332 (highest correlation)
Corr('lweight',y) = 0.4852
Corr('age',y) = 0.2276
Corr('lbph',y) = 0.2629
Corr('svi',y) = 0.5569 (second highest correlation)
Corr('lcp',y) = 0.4892
Corr('gleason',y) = 0.3424
Corr('pgg45',y) = 0.4480

Regression with Outliers

I assumed there were errors with the 20th data point: 'lweight' and 'lcp' were 20. The regression coefficients changed as in the table below. Six of the nine regression coefficients were less severely affected when Ridge regression was performed.

	LS	LS with outliers	Ridge (λ=0.444)	Ridge with outliers (λ=3.15)
Intercept	2.4523	2.4523	2.4362	2.3424
lcavol	0.7110	0.7799	0.6993	0.6824
lweight	0.2905	0.7358	0.2901	0.3775
age	-0.1415	-0.1225	-0.1382	-0.0829
lbph	0.2104	0.2480	0.2095	0.2782
svi	0.3073	0.3508	0.3047	0.3178
lcp	-0.2868	-0.8269	-0.2719	-0.4140
gleason	-0.0208	-0.0592	-0.0163	-0.0346
Pgg45	0.2753	0.3407	0.2663	0.2311
Std error	0.1839	0.2026	0.1840	0.1799
Test error	0.5863	0.5996	0.5823	0.5464