39) Ridge regression, lasso, and elastic net can be used if $n \geq p$ and if $p > n$. Lasso and elastic net do variable selection: $\hat{\beta}_L$ and $\hat{\beta}_E$ have 0's. Fit OLS to the predictors, including a constant, with $\hat{\beta}_i \neq 0$.

ex) $\hat{\beta}_L = (\hat{\beta}_1, 0, \hat{\beta}_3, 0)^T \rightarrow$ fit OLS regression of Y on (x_1, x_3).

40) For $\underline{Y} = X\underline{\beta} + \underline{e}$ and $\underline{Z} = W\underline{\eta} + \underline{e}$

the OLS estimator $\hat{\underline{\eta}} = (W^T W)^{-1} W^T \underline{Z}$.

§7.8

41) The ridge regression estimator $\hat{\underline{\eta}}_R$

minimizes $Q_R(\underline{\eta}) = \frac{1}{a}(\underline{Z} - W\underline{\eta})^T(\underline{Z} - W\underline{\eta}) + \frac{\lambda_n}{a}\underbrace{\sum_{i=1}^{p-1} \eta_i^2}_{\underline{\eta}^T\underline{\eta}}$

$= \frac{1}{a} RSS_W(\underline{\eta}) + \frac{\lambda_n}{a} \|\underline{\eta}\|_2^2$ with $a = 1, 2, n, 2n$ common.

If $\lambda_n = 0$, $\hat{\underline{\eta}}_R = \hat{\underline{\eta}}_{OLS}$. As $\lambda_n \to \infty$, $\hat{\underline{\eta}}_R \to 0$ and $\hat{\underline{Y}} \to \bar{\underline{Y}}$. So RR is a shrinkage method.

$\hat{\underline{\eta}}_R = \underbrace{(W^T W + \lambda_n I_{p-1})^{-1}}_{\text{inverse exists if } \lambda_n > 0} W^T \underline{Z}$.

warning: The literature typically uses $\lambda = \frac{\lambda_n}{n}$.

42) Usually a grid of m λ_n values is used

$0 \le \lambda_1 < \lambda_2 \le \cdots \le \lambda_m$ where $\lambda_i = \lambda_{ni}$.

10 fold CV is used to select $\lambda_S = \hat{\lambda}_n$.

43) Suppose $n > p$ and $(W^TW)^{-1}$ exists.

Then $\hat{\beta}_R = \underbrace{(W^TW + \lambda_n I_{p-1})^{-1}}_{A_n} \underbrace{W^TW (W^TW)^{-1}}_{I_{p-1}} W^T Z$ $\quad \hat{\beta}_{OLS}$

So $\hat{\beta}_R = A_n \hat{\beta}_{OLS}$. HW 10 shows

$A_n = B_n = I_{p-1} - \lambda_n (W^TW + \lambda_n I_{p-1})^{-1}$.

44) Suppose $\frac{W^TW}{n} \xrightarrow{P} V^{-1}$ and the OLS CLT holds.

$\sqrt{n}(\hat{\beta}_{OLS} - \beta) \xrightarrow{D} N_{p-1}(0, \sigma^2 V)$.

If $\frac{\lambda_n}{n} \to 0$, then $\frac{W^TW + \lambda_n I_{p-1}}{n} \xrightarrow{P} V^{-1}$ and

$n (W^TW + \lambda_n I_{p-1})^{-1} \xrightarrow{P} V$. Also $A_n = \frac{n}{n} A_n =$

$\left(\frac{W^TW + \lambda_n I_{p-1}}{n}\right)^{-1} \frac{W^TW}{n} \xrightarrow{P} V V^{-1} = I_{p-1}$.

45} RR CLT Assume p is fixed and the OLS CLT holds for $\underline{Z} = W\underline{\eta} + \underline{e}$.

a) If $\frac{\hat{\lambda}_n}{\sqrt{n}} \xrightarrow{P} 0$, $\sqrt{n}(\hat{\underline{\eta}}_R - \underline{\eta}) \xrightarrow{D} N_{p-1}(\underline{0}, \sigma^2 V)$.

So OLS full model and RR are asymptotically equivalent.

b) If $\frac{\hat{\lambda}_n}{\sqrt{n}} \xrightarrow{P} \gamma > 0$, then

$\sqrt{n}(\hat{\underline{\eta}}_R - \underline{\eta}) \xrightarrow{D} N_{p-1}(-\gamma V \underline{\eta}, \sigma^2 V)$

interior to full model OLS unless this term $= 0$

c) $\hat{\underline{\eta}}_R$ is a consistent estimator of $\underline{\eta}$ if $\frac{\hat{\lambda}_n}{n} \xrightarrow{P} 0$ as $n \to \infty$.

---

§7.9

46} $\underline{Y} = X\underline{\beta} + \underline{e}$, $\underline{Z} = W\underline{\eta} + \underline{e}$. The lasso estimator

$\hat{\underline{\eta}}_L$ minimizes the criterion

$$Q_n(\underline{\eta}) = \frac{1}{a}(\underline{Z} - W\underline{\eta})^T(\underline{Z} - W\underline{\eta}) + \frac{\lambda_n}{a}\sum_{i=1}^{p-1}|\eta_i| =$$

$\frac{1}{a} RSS_W(\underline{\eta}) + \frac{\lambda_n}{a} \|\underline{\eta}\|_1$. $\leftarrow L_1$ norm.

Lasso is a shrinkage method and a variable selection method! often some $\hat{\eta}_i = 0$.

is used. λ_m is the smallest value of λ such that $\hat{\underset{\sim}{\eta}}_{\lambda_m} = \underset{\sim}{0}$. Hence $\hat{\underset{\sim}{\eta}}_{\lambda_i} \neq 0$ for $i < M$.

48] By the Karush Kuhn Tucker (KKT) conditions for convex optimality (see Math 471)

$$-W^T(\underset{\sim}{Z} - W\hat{\underset{\sim}{\eta}}_L) + \frac{\lambda_n}{2}\underset{\sim}{s}_\eta = \underset{\sim}{0} \quad \text{where } s_i \in [-1, 1].$$

Thus $\hat{\underset{\sim}{\eta}}_L = \underbrace{(W^TW)^{-1}W^T\underset{\sim}{Z}}_{\hat{\underset{\sim}{\eta}}_{OLS}} - n(W^TW)^{-1}\frac{\lambda_n}{n}\underset{\sim}{s}_\eta$

49] Lasso CLT: Assume p is fixed and the OLS CLT holds for $\underset{\sim}{Z} = W\underset{\sim}{\eta} + \underset{\sim}{e}$.

a) If $\frac{\hat{\lambda}_n}{\sqrt{n}} \xrightarrow{P} 0$, then $\sqrt{n}(\hat{\underset{\sim}{\eta}}_L - \underset{\sim}{\eta}) \xrightarrow{D} N_{p-1}(\underset{\sim}{0}, \sigma^2 V)$,

b) If $\frac{\hat{\lambda}_n}{\sqrt{n}} \xrightarrow{P} \tau > 0$, and $\underset{\sim}{s}_n \xrightarrow{P} \underset{\sim}{s}$, then $\sqrt{n}(\hat{\underset{\sim}{\eta}}_L - \underset{\sim}{\eta}) \xrightarrow{D} N_{p-1}(-\frac{\tau}{2}V\underset{\sim}{s}, \sigma^2 V)$,

c) If $\frac{\hat{\lambda}_n}{n} \xrightarrow{P} 0$, then $\hat{\underset{\sim}{\eta}}_L$ is a consistent estimator of $\underset{\sim}{\eta}$.

§7.11 50] $\underset{\sim}{Y} = X\underset{\sim}{\beta} + \underset{\sim}{e}$, $\underset{\sim}{Z} = W\underset{\sim}{\eta} + \underset{\sim}{e}$

The elastic net estimator $\hat{\underline{m}}_{EN}$ minimizes the criterion $Q_{EN}(\underline{m}) = RSS_w(\underline{m}) + \lambda_1 \|\underline{m}\|_2^2 + \lambda_2 \|\underline{m}\|_1$.

where $\lambda_1 = (1-\alpha)\lambda_n$, $\lambda_2 = 2\alpha\lambda_n$ and $0 \le \alpha \le 1$, So $\alpha = 1$ corresponds to lasso using $2\lambda_n$ and $\alpha = 0$ corresponds to RR. The EN estimator has a CLT similar to that of lasso, and EN is a shrinkage estimator that does variable selection.

51) The VS CLT showed

$\hat{\beta}_{I_{min}}$ is $\sqrt{n}$ consistent when lasso or EN are consistent where I_{min} corresponds to the lasso or EN $\hat{\beta}_i \neq 0$.

52) know Consider choosing $\hat{\underline{m}}$ to minimize

$Q(\underline{m}) = \frac{1}{a}(\underline{z} - w\underline{m})^T(\underline{z} - w\underline{m}) + \frac{\lambda_n}{a}\sum_{i=1}^{p-1} |m_i|^j$ where $\lambda_n \ge 0$, $a > 0$ and $j > 0$ are known constants.

Then $j=2$ corresponds to RR $\hat{\underline{m}}_R$, $j=1$ to lasso $\hat{\underline{m}}_L$, and $\lambda_n = 0$ to OLS $\hat{\underline{m}}_{OLS}$. Do a similar problem with EN criterion from 50].

5.3) Get outlier resistant estimators by replacing $\hat{\beta}_{OLS}$ by the alternative estimator $\hat{\beta}_A$.

i) RMVN, RFCH, COVMB2 Set D applied to nontrivial predictors, then fit $\hat{\beta}_A$ on cases corresponding to D.

ii) tvreg using $\hat{\beta}_A$ instead of $\hat{\beta}_{OLS}$

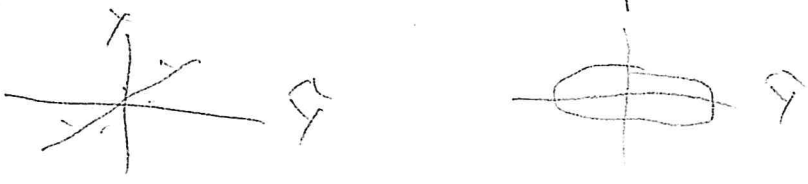
iii) RMVN ~~Set~~ U from $(Y_i, \underline{x}_i)$

Fit $\hat{\beta}_A$ to cases in U.

Ch 8 §8.1

1) The additive error regression (AER) model is $Y = m(\underline{x}) + e$ where $m(\underline{x})$ is a real valued function, e_i iid $E(e)=0$, $V(e)=\sigma^2$, $\hat{Y} = \hat{m}(\underline{x})$, $r = Y - \hat{Y}$. The AER model is a 1D reg model with $SP = h(\underline{x}) = m(\underline{x})$, $ESP = \hat{m}(\underline{x})$

2) If the error distribution is not highly skewed, the response and residual plots look exactly like those for MLR

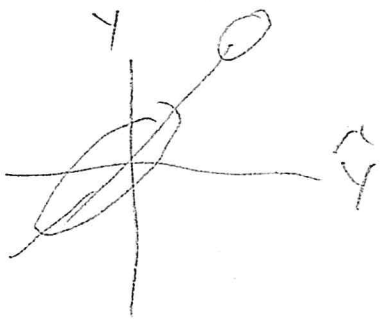


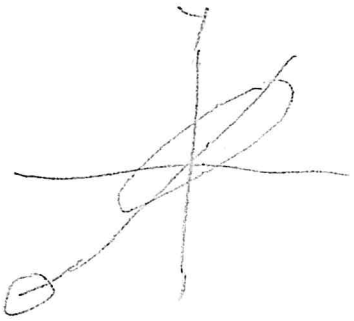
3) MLR has $m(\underline{x}) = \underline{x}^T\underline{\beta}$. Nonlinear regression, nonparametric regression and the AE Gam

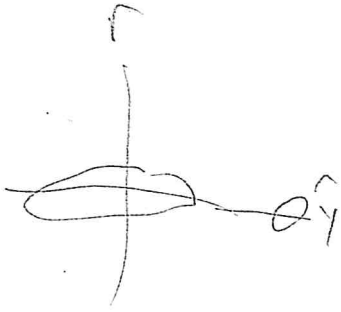
$Y = \alpha + \sum_{i=2}^{p} S_i(x_i) + e = AP + \bar{e}$, are also special cases of AER.

(could start at $i=1$)

4) Compared to the inflexible hyperplane $\underline{x}^T\underline{\beta}$, several methods for finding $\hat{m}(\underline{x})$ are flexible. Hence $\hat{Y}_i \approx Y_i$ often occurs even if Y_i is an outlier. Thus outlier detection is more difficult.







Look for gaps in the response and residual plots.

5) PIs are similar. Let $[r_L, r_U]$ be a PI for r_ϵ. Then a PI for Y_ϵ is $[\hat{Y}_\epsilon + r_L, \hat{Y}_\epsilon + r_U]$.

Ch 9) 1) Good references 585 59

Cook and Weisberg 1999a and Olive 2010a, 2017a

2) Know p Regression is the study of the conditional distribution $Y|\underline{x}$ of the response Y given the $(p-1)\times 1$ vector of nontrivial predictors $\underline{x}$ (Now $\underline{x}$ does not contain a constant 1).

Want to know how Y changes as the predictors $\underline{x}$ change. ($Y = Y|\underline{x}$ suppressed)

NO 595

3) Know: In a 1D regression model

Y is conditionally independent of $\underline{x}$ given $h(\underline{x})$ (often $h(\underline{x}) = \underline{x}^T\underline{\beta}$ a single linear combination of the predictors), written $Y \perp\!\!\!\perp \underline{x} \mid \underline{\beta}^T\underline{x}$.

4) Know: If $Y \perp\!\!\!\perp \underline{x} \mid \underline{\beta}^T\underline{x}$, then $Y \perp\!\!\!\perp \underline{x} \mid (a + c\underline{\beta}^T\underline{x})$ for any constants a and $c \neq 0$. The quantity $a + c\underline{\beta}^T\underline{x}$ is called a sufficient predictor (SP).

A sufficient summary plot (SSP) is a plot of SP vs Y. An estimated sufficient predictor (ESP) is $\tilde{\alpha} + \tilde{\underline{\beta}}^T\underline{x}$ where $\tilde{\underline{\beta}}$ is an estimator of $c\underline{\beta}$ for some $c \neq 0$. An estimated sufficient summary plot (ESSP) or response plot is a plot of ESP vs Y.

5) * The most used statistical regression models are 1D models.

ex] $y = g(\alpha + \underline{\beta}^T\underline{x}, e)$

2.3 Single Index Models

$$y = m(\alpha + \underline{\beta}^T \underline{x}) + e$$

(often the function m is unknown)

ex) $y = \alpha + \underline{\beta}^T \underline{x} + e$ (MLR)

ex) $y = t^{-1}(\alpha + \underline{\beta}^T \underline{x} + e)$ (transformation model)

ex) logistic regression in ch 10

ex) poisson regression in ch 10

6) Any model with a single (nontrivial) predictor x ($p-1 = 1$) is a 1D model. In this case x is both a SP and ESP and a plot of x vs y is both a SSP and an ESSP.

7) If $p-1 > 1$ the SP is unknown and an SSP can not be made (except for simulated data). For the MLR model $Y = \alpha + \underline{\beta}^T \underline{x} + e$ the ESP $= \hat{y} = \hat{\alpha} + \hat{\underline{\beta}}^T \underline{x}$, and the [illegible] ESSP = [illegible] response plot [illegible].