Linear regression

Hastie, Tibshirani, Friedman Ch 6-7

Kevin Murphy Ch. 7

CS 6140
Machine Learning
Professor Olga Vitek

February 2, 2017

Generative vs discriminative models

- Goal: predict Y
  - Bayes rule:

$$p(Y|\mathbf{X}) = \frac{p(Y) \cdot p(\mathbf{X}|Y)}{p(\mathbf{X})}$$

- Generative classifiers
  - Specify prior probability of $p(Y)$
  - Assume conditional distribution $p(\mathbf{X}|Y)$
  - Use Bayes rule to derive the posterior $p(Y|\mathbf{X})$
  - **Example:** Linear discriminant analysis

- Discriminative classifiers
  - Estimate the posterior the posterior $p(Y|\mathbf{X})$
  - Do not assume the distribution on $\mathbf{X}$
  - **Example:** Y continuous: linear regression

Linear regression with two predictors

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \varepsilon_i; \qquad i = 1, \ldots, n$$

- β_0 is the intercept
- β_1 and β_2 are the regression coefficients
- Meaning of regression coefficients
  - β_1 describes change in mean response per unit increase in X_1 when X_2 is held constant
  - β_2 describes change in mean response per unit increase in X_2 when X_1 is held constant
- Variables X_1 and X_2 are **additive**.
- Same change in X_1 for all X_2.
- The response surface is a plane.

Interaction model

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i1} X_{i2} + \varepsilon_i$$

- Meaning of parameters:
  - Change in X_1 when $X_2 = x_2$

$$\begin{aligned} \Delta Y &= (\beta_0 + \beta_1(X_1 + 1) + \beta_2 x_2 + \beta_3(X_1 + 1)x_2) - \\ &\quad (\beta_0 + \beta_1 X_1 + \beta_2 x_2 + \beta_3 X_1 x_2) \\ &= \beta_1 + \beta_3 x_2 \end{aligned}$$

  - Change in X_2 when $X_1 = x_1$

$$\Delta Y = \beta_2 + \beta_3 x_1$$

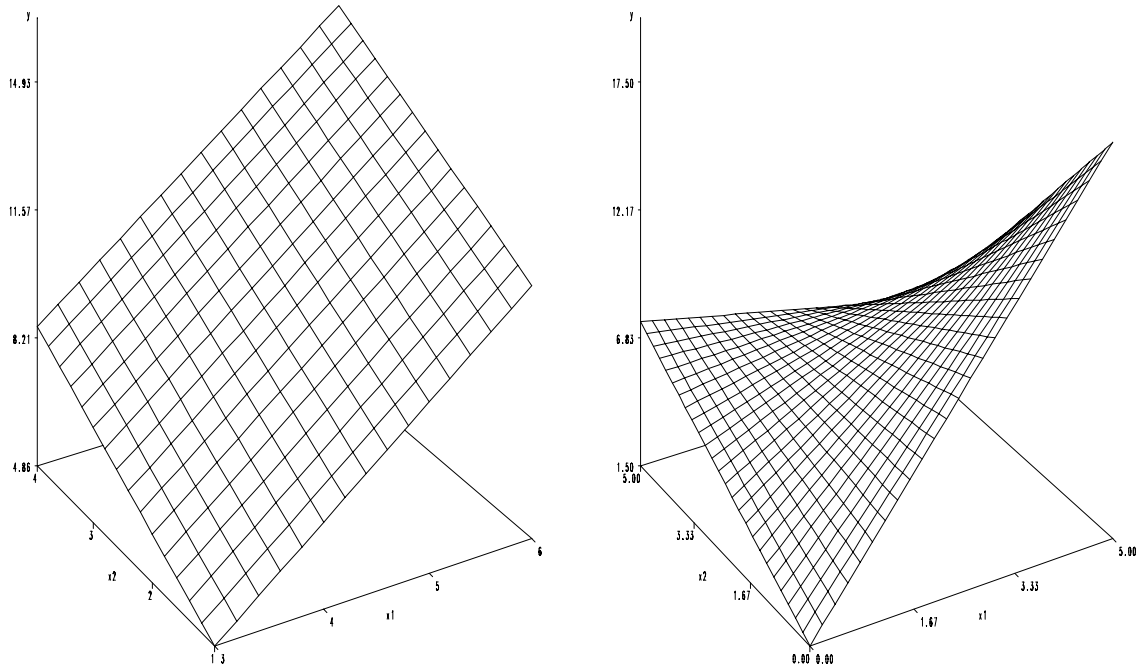
- Rate of change due to one variable affected by the other

Additive vs interaction model

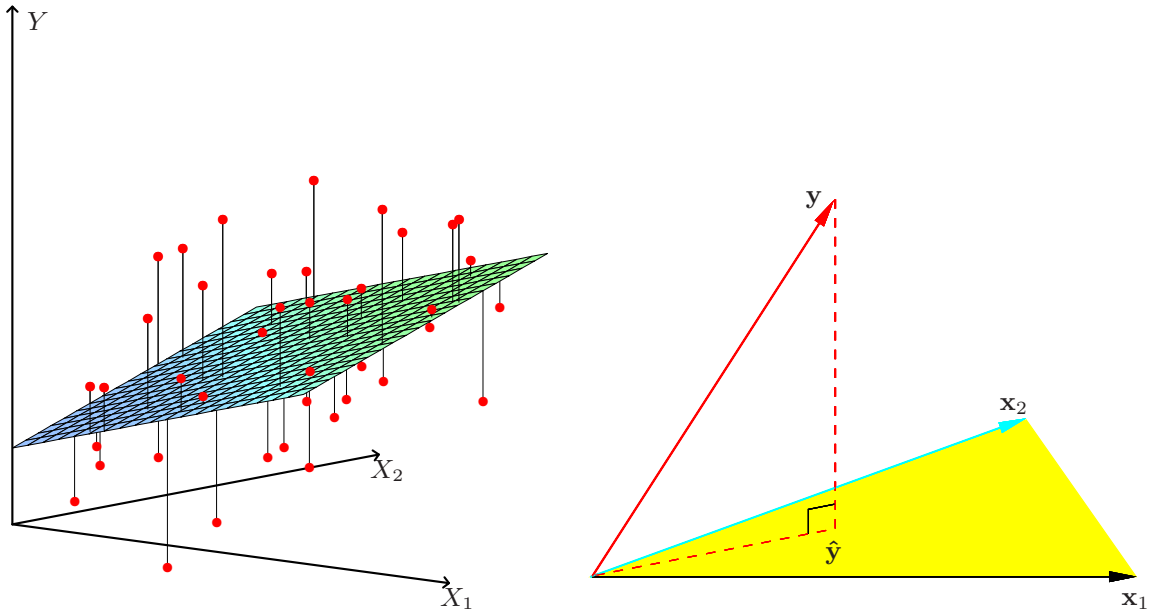
$$\widehat{Y}_i = -2.79 + 2.14X_{i1} + 1.21X_{i2}$$

versus

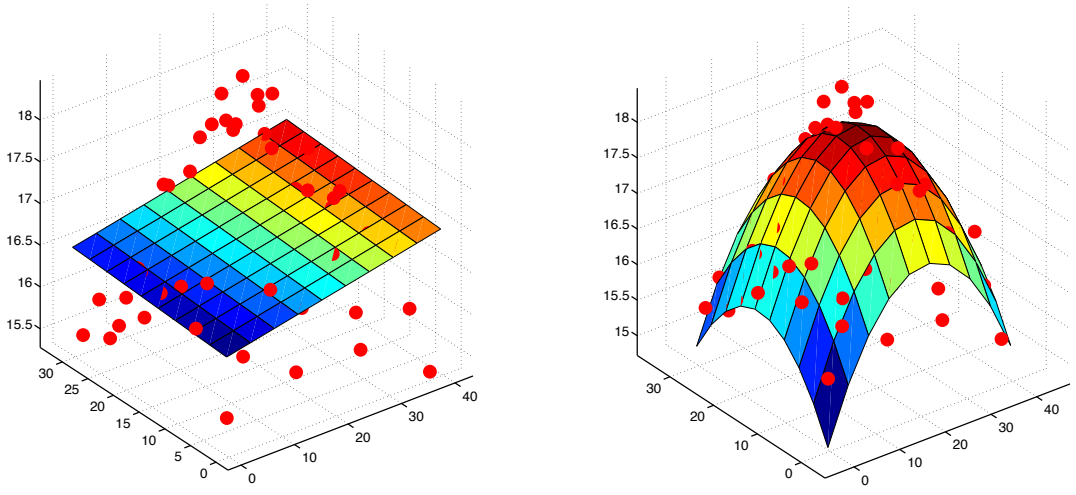
$$\widehat{Y}_i = 1.5 + 3.2X_{i1} + 1.2X_{i2} - .75X_{i1}X_{i2}$$



Linear regression with two predictors



Hastie, Tibshirani, Friedman, Fig 3.1 and 3.2



K. Murphy, Fig 7.1

Polynomial regression and transformations

- Polynomial regression:

$$\begin{aligned} Y_i &= \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \varepsilon_i \\ &= \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \varepsilon_i \end{aligned}$$

where $X_{i2} = X_i^2$.

  - this is a linear model because it is a linear function of parameters β

- Transformations

$$log(Y_i) = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \varepsilon_i$$

$$Y_i = \frac{1}{\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \varepsilon_i}$$

  - this is a linear model on the $log(Y_i)$ scale

General linear regression in matrix terms

- As an equation

$$Y_i = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_{p-1} X_{i,p-1} + \varepsilon_i$$

- As an array

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} 1 & X_{11} & X_{12} & \cdots & X_{1\ p-1} \\ 1 & X_{21} & X_{22} & \cdots & X_{2\ p-1} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & X_{n1} & X_{n2} & \cdots & X_{n\ p-1} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \cdots \\ \beta_{p-1} \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

- In matrix notation

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

Estimation of regression coefficients

- Objective function: least squares
  - find $\hat{\beta}$ to minimize

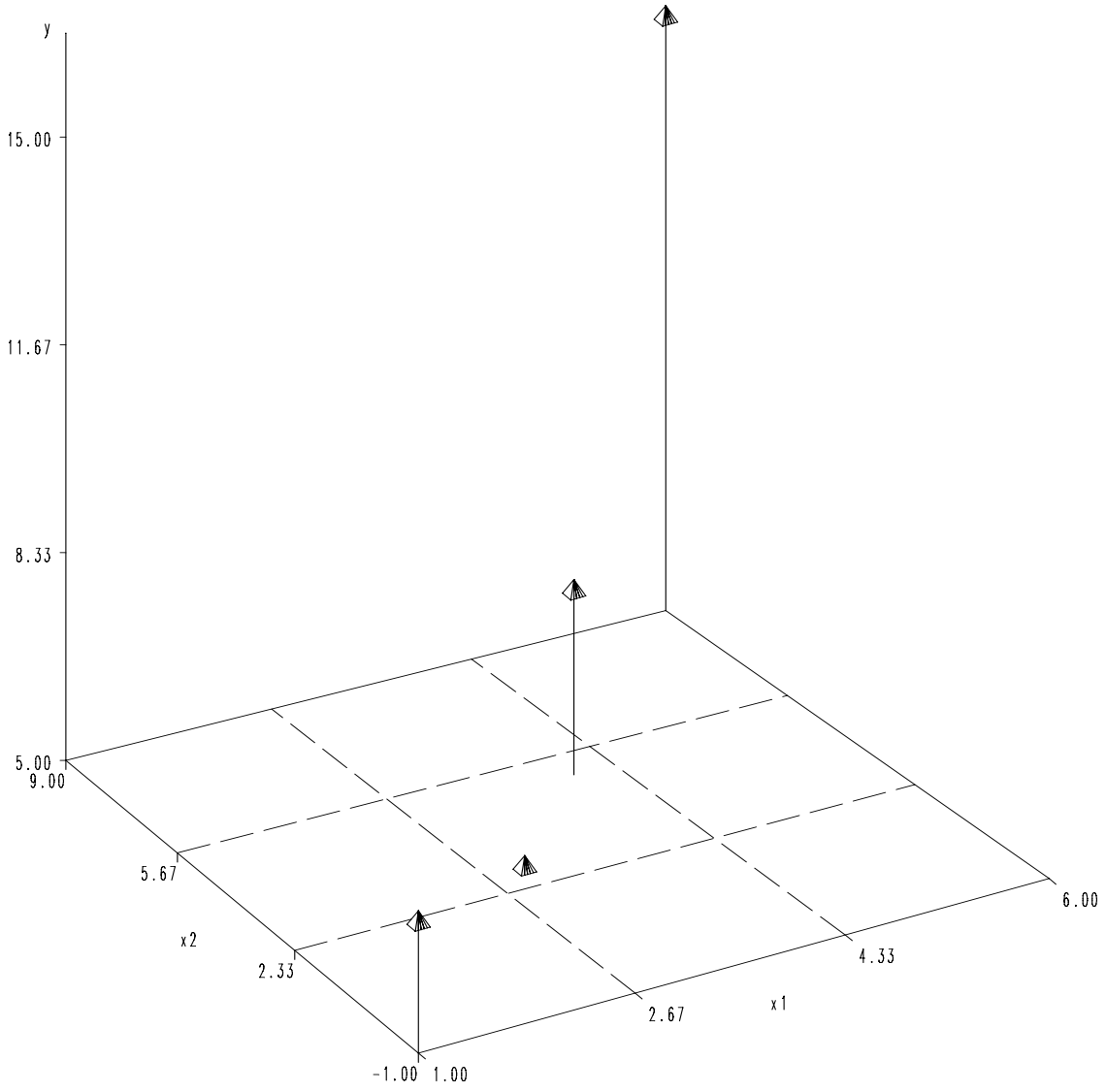
    $$\sum_{i=1}^{N}(y_i - x_i'\beta)^2 = (\mathbf{Y} - \mathbf{X}\hat{\beta})'(\mathbf{Y} - \mathbf{X}\hat{\beta})$$

  - Quadratic objective function $\Rightarrow$
    its minimum always exists, but may not be unique

- Finding estimates
  - Differentiating wrt β:
  - Normal equations $\mathbf{X}'(y - \mathbf{X}\beta) = 0 \Rightarrow \hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$

- Fitted values define a (hyper)plane
  - $\hat{\mathbf{Y}} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} = \mathbf{H}\mathbf{Y}$
  - Residuals: $\mathbf{e} = \mathbf{Y} - \hat{\mathbf{Y}} = (\mathbf{I} - \mathbf{H})\mathbf{Y}$

Multicollinearity



Qualitative predictors

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i1} X_{i2} + \varepsilon_i$$

- Let $X_2 = 1$ if case from Massachusetts

- Meaning of parameters:
  - Case from Massachusetts ($X_2 = 1$):

$$\begin{aligned} Y &= \beta_0 + \beta_1 X_1 + \beta_2 1 + \beta_3 X_1 (1) \\ &= (\beta_0 + \beta_2) + (\beta_1 + \beta_3) X_1 \end{aligned}$$

  - Case from other location ($X_2 = 0$)

$$\begin{aligned} Y &= \beta_0 + \beta_1 X_1 + \beta_2 0 + \beta_3 X_1 (0) \\ &= \beta_0 + \beta_1 X_1 \end{aligned}$$

- Have <u>two</u> regression lines

- β_2 and β_3 quantify the differences

Two groups: Wrong coding

- Assume an additive model with two groups

- Wrong approach: add both indicators

$$X_2 = \begin{cases} 1 \text{ , if stock firm} \\ 0 \text{ , otherwise} \end{cases} \quad X_3 = \begin{cases} 1 \text{ , if mutual fund} \\ 0 \text{ , otherwise} \end{cases}$$

  - the model below is wrong

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \varepsilon_i$$

- The corresponding design matrix

  - 4 data points (first 2 from stock firm, last 2 from mutual fund)

$$\mathbf{X} = \begin{pmatrix} 1 & X_{11} & 1 & 0 \\ 1 & X_{21} & 1 & 0 \\ 1 & X_{31} & 0 & 1 \\ 1 & X_{41} & 0 & 1 \end{pmatrix}$$

  - this model creates fully collinear columns in the design matrix $\mathbf{X}$ (R will drop the first)

Two groups: Correct coding

- Correct approach 1:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \varepsilon_i$$

  - interpretation:

$$\begin{aligned} E\{Y_i\} &= \beta_0 + \beta_1 X_{i1} && \text{if mutual fund} \\ E\{Y_i\} &= (\beta_0 + \beta_2) + \beta_1 X_{i1} && \text{if stock firm} \end{aligned}$$

  - Mutual fund is the reference group
  - β_2: the deviation of the intercept of the stock firm from the reference

- The corresponding design matrix:

  - 4 data points (first 2 from stock firm, last 2 from mutual fund)

$$\mathbf{X} = \begin{pmatrix} 1 & X_{11} & 1 \\ 1 & X_{21} & 1 \\ 1 & X_{31} & 0 \\ 1 & X_{41} & 0 \end{pmatrix}$$

Three groups: Wrong coding

- Extend the indicator

$$X_2 = \begin{cases} 0 \text{ , if mutual fund} \\ 1 \text{ , if stock firm} \\ 2 \text{ , if foreign firm} \end{cases}$$

- The model below is still appropriate

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \varepsilon_i$$

  - interpretation: enforces an equal change in $E\{Y\}$ for each extra indicator

$$\begin{aligned} E\{Y_i\} &= \beta_0 + \beta_1 X_{i1} && \text{if mutual fund} \\ E\{Y_i\} &= (\beta_0 + \beta_2) + \beta_1 X_{i1} && \text{if stock firm} \\ E\{Y_i\} &= (\beta_0 + 2\beta_2) + \beta_1 X_{i1} && \text{if foreign firm} \end{aligned}$$

- The corresponding design matrix:

  - 6 data points (first 2 from mutual fund, 2 from stock, 2 foreign)

$$\mathbf{X} = \begin{pmatrix} 1 & X_{11} & 0 \\ 1 & X_{21} & 0 \\ 1 & X_{31} & 1 \\ 1 & X_{41} & 1 \\ 1 & X_{41} & 2 \\ 1 & X_{41} & 2 \end{pmatrix}$$

Three groups: Correct coding

- First option:

$$X_2 = \begin{cases} 1 \text{ , if stock firm} \\ 0 \text{ , otherwise} \end{cases} \qquad X_3 = \begin{cases} 1 \text{ , if foreign firm} \\ 0 \text{ , otherwise} \end{cases}$$

- The model below contains two indicators

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \varepsilon_i$$

  - interpretation:

$$\begin{aligned} E\{Y_i\} &= \beta_0 + \beta_1 X_{i1} && \text{if mutual fund} \\ E\{Y_i\} &= (\beta_0 + \beta_2) + \beta_1 X_{i1} && \text{if stock firm} \\ E\{Y_i\} &= (\beta_0 + \beta_3) + \beta_1 X_{i1} && \text{if foreign firm} \end{aligned}$$

  - mutual fund is the reference
  - β_2 and β_3 are deviations of the intercepts from the reference
  - also more flexibility in presence of interactions X_1X_2 and X_1X_3
  - the number of indicators is always one less than the number of groups

Normal Error Model

- **The least square estimates of the parameters do not require the assumption of Normality**

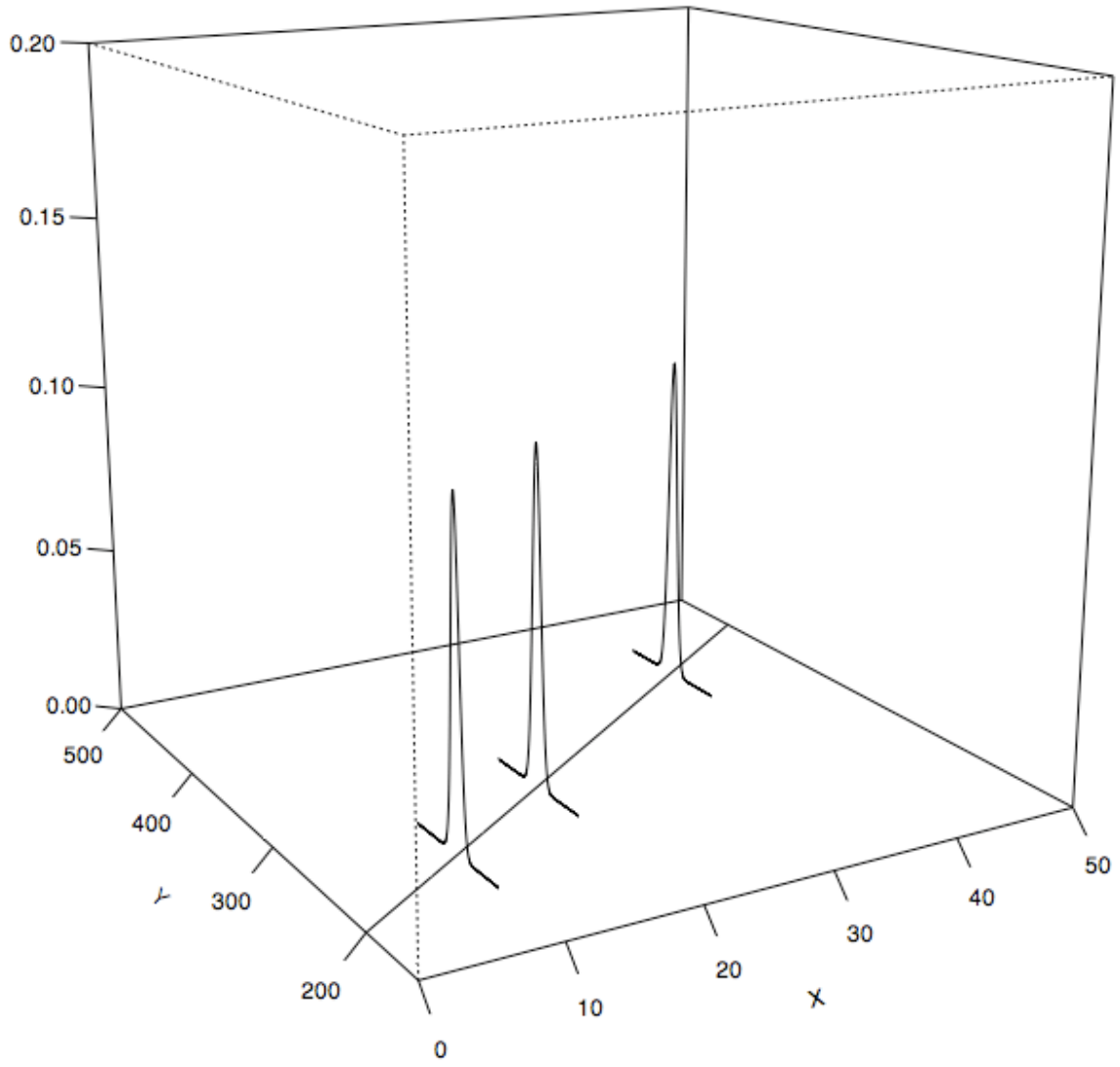
- Normal error assumption greatly simplifies the theory of analysis

- Normality is used to construct confidence intervals / perform hypothesis tests follow known distributions (e.g., t, F)

- While not always true in practice, most inference only sensitive to large departures from normality

Normal Error regression model

- $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$



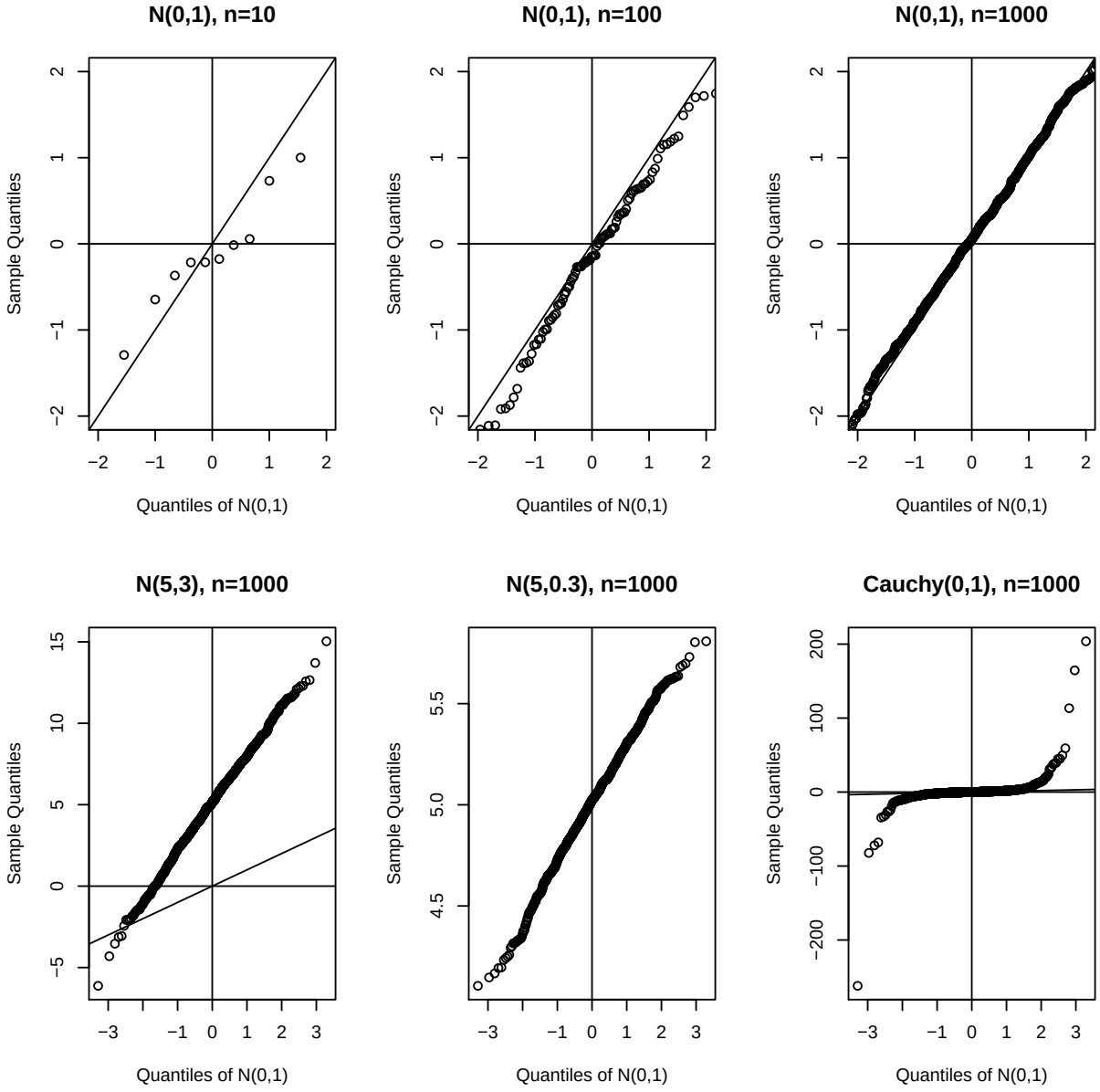
Normal Error regression model

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

- β_0 is the intercept
- β_1 in the slope
- ε_i is the i^{th} random error term
  - $\varepsilon_i \sim \mathsf{N}(0, \sigma^2) \longleftarrow$ **NEW**
  - Uncorrelated $\longrightarrow$ independent error terms
- Defines distribution of Y: $p(Y|\mathbf{X})$

$$Y_i \sim \mathsf{N}(\beta_0 + \beta_1 X_i, \sigma^2)$$

Assessing Normality: Quantile-quantile plot



Can be used with any other distribution

Example

Height of 11 women

i	Observed height	Adj. percentile $100(i-\frac{1}{2})/11)$	z	Sample quantiles
1	61.0	4.55	-1.69	60.6
2	62.5	13.64	-1.10	62.3
3	63.0	22.73	-0.75	63.4
4	64.0	31.82	-0.47	64.1
5	64.5	40.91	-0.23	64.8
6	65.0	50.00	0.00	65.5
7	66.5	59.09	0.23	66.2
8	67.0	68.18	0.47	66.9
9	68. 0	77.27	0.75	67.6
10	68.5	86.36	1.10	68.7
11	70.5	95.45	1.69	70.4

QQplot: plot Observed height vs sample quantiles

Sample quantiles $= x + Z \cdot \hat{\sigma} + \hat{\mu}$

```
> ?qqplot
> ?qqnorm
```

Maximum Likelihood Estimation

- Assumption of Normality gives us more choices of methods for parameter estimation

$$Y_i \sim \mathsf{N}(\beta_0 + \beta_1 X_i, \sigma^2)$$

$$\downarrow$$

$$f_i = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}(Y_i - \beta_0 - \beta_1 X_i)^2\right\}$$

- Likelihood function $L = f_1 \times f_2 \times \cdots \times f_n$ (i.e. the joint probability distribution of the observations, viewed as function of parameters)

- Find β_0, β_1 and σ^2 which maximizes L

- Obtain same estimators $\widehat{\beta}_0$ and $\widehat{\beta}_1$

- A slightly smaller estimate of σ^2
  - See KM 7.3 for derivation in vector notation

Partitioning sums of squares

- Organizes results arithmetically
- Total sums of squares in Y is defined

$$\text{SSTO} = \sum (Y_i - \overline{Y})^2$$

- Can partition sum of squares into
  - Model (explained by regression)
  - Error (unexplained / residual)
- Rewrite the total sum of squares as

$$\sum (Y_i - \overline{Y})^2 = \sum (Y_i - \hat{Y}_i + \hat{Y}_i - \overline{Y})^2$$

$$\begin{aligned} \sum (Y_i - \overline{Y})^2 &= \sum (\hat{Y}_i - \overline{Y})^2 + \sum (Y_i - \hat{Y}_i)^2 \\ &= b_1^2 \sum (X_i - \overline{X})^2 + \sum (Y_i - \hat{Y}_i)^2 \\ \text{SSTO} &= \text{SSR} + \text{SSE} \end{aligned}$$

Coefficient of multiple determination

- Coefficient of Determination R^2 describes proportionate reduction in total variation associated with the **full set** of X variables

$$R^2 = \frac{SSR}{SSTO} = 1 - \frac{SSE}{SSTO},\ 0 \leq R^2 \leq 1$$

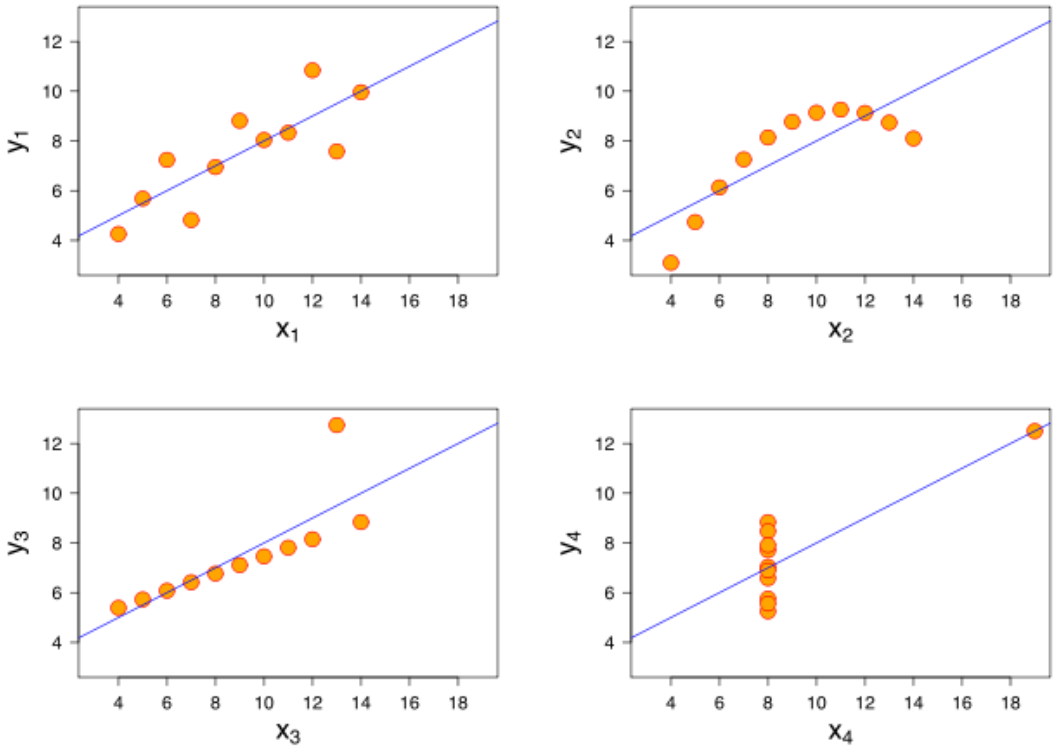
- R^2 usually increases with the increasing p
  - Adjusted R_a^2 attempts to account for p

$$R_a^2 = 1 - \frac{SSE/n-p}{SSTO/n-1},\ 0 \leq R_a^2 \leq 1$$

  - The adjustment is often insufficient

Anscombe's quartet

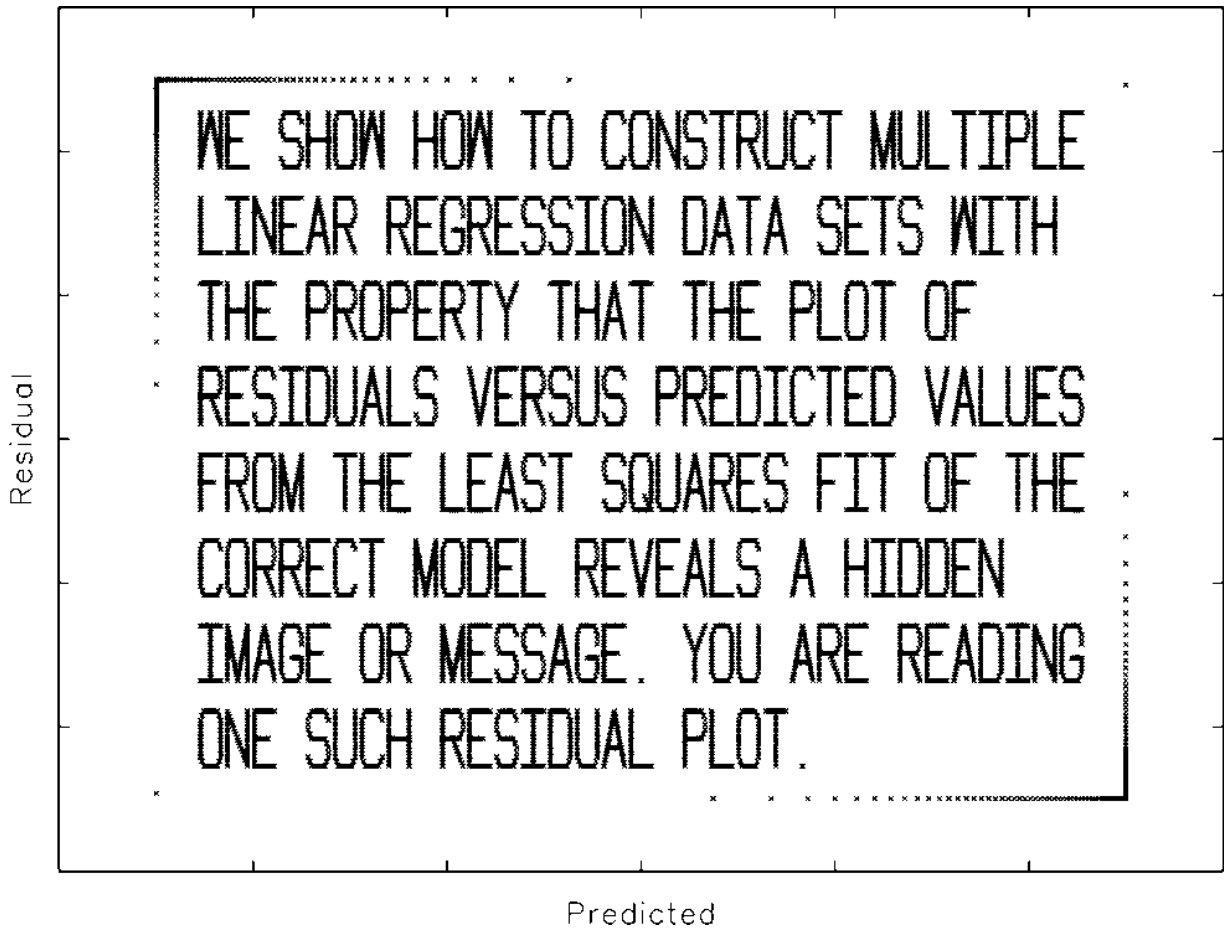
- Constructed in 1973 by Francis Anscombe
- All four datasets share properties
  - $n = 11$, $\bar{x} = 9$, $\bar{y} = 7.50$
  - $Var\{x\} = 11$, $Var\{y\} = 4.125$
  - $Corr(x, y) = 0.816$, $\hat{y} = 0.5x + 3$



https://en.wikipedia.org/wiki/Anscombe's_quartet

Residual (Sur)Realism

- Algorithm for creating multi-variable linear regressions with arbitrary residuals



L. Stefanski, Residual (Sur)Realism. *The American Statistician*, vol. 61, p.163, 2007.

Properties of sampling distribution of $\widehat{\beta}$

- Consistent estimators $\widehat{\beta} \to \beta$ as $n \to \infty$
- Unbiased estimators $E\{\widehat{\beta}\} = \beta$
- Minimum variance estimators: min $Var\{\widehat{\beta}\}$
- Gauss-Markov theorem: $\widehat{\beta}$ least squares
  - Are unbiased
  - Have **minimum variance** among all unbiased linear estimators
  - i.e., are the most precise of any estimators where

    b_l is of the form $\sum k_i Y_i$ and $E(b_l) = \beta_l$

- See KM 6.4 and HTF 3.2.2 for details

Bias-variance decomposition

- Mean Squared Error (MSE):

$$\begin{aligned} MSE &= E\{\widehat{\beta} - \beta\}^2 \\ &= E\left\{\widehat{\beta} - E\{\widehat{\beta}\} + E\{\widehat{\beta}\} - \beta\right\}^2 \\ &= E\left\{\widehat{\beta} - E\{\widehat{\beta}\}\right\}^2 + E\left\{\{\widehat{\beta}\} - \beta\right\}^2 \\ &= \text{Var}\{\widehat{\beta}\} + \text{Bias}\{\widehat{\beta}\}^2 \end{aligned}$$

- For unbiased estimators, MSE=Var$\{\widehat{\beta}\}$

- Biased estimators can $\downarrow$ MSE if Var$\{\widehat{\beta}\}$ $\downarrow$

Empirical risk minimization

- Consider loss function $L\left(Y, f(\mathbf{X})\right)$
  - y is true but unknown response
  - $f(\mathbf{X})$ is function (e.g., linear combination) of observed predictors
- The risk (of making incorrect decision) is
  - $R = E\{L\left(Y, f(\mathbf{X})\right)\} = \sum_x \sum_y L\left(Y, f(\mathbf{X})\right) \cdot p(\mathbf{X}, Y)$
  - When $L\left(Y, f(\mathbf{X})\right)$ is 0-1- function: misclassification rate
  - When $L\left(Y, f(\mathbf{X})\right) = (Y - \delta(\mathbf{X}))^2$: mean squared error
- The estimated (i.e. empirical) risk is
  - $R = \frac{1}{N} \sum_{i=1}^{N}, L\left(Y, f(\mathbf{X})\right)$
  - If validation set is not available: use cross-validation (see KM 6.5.3)

Model selection and bias-variance tradeoff

- Consider polynomial regression, fixed X

  $Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \ldots + \beta_k X^k = f_k(X) + \epsilon$

- When $L\left(Y, f_k(\mathbf{X})\right) = \left(Y - f_k(\mathbf{X})\right)^2$

  - The expected loss for fixed $X = x$ $L\left(Y, f_k(x)\right)$ is

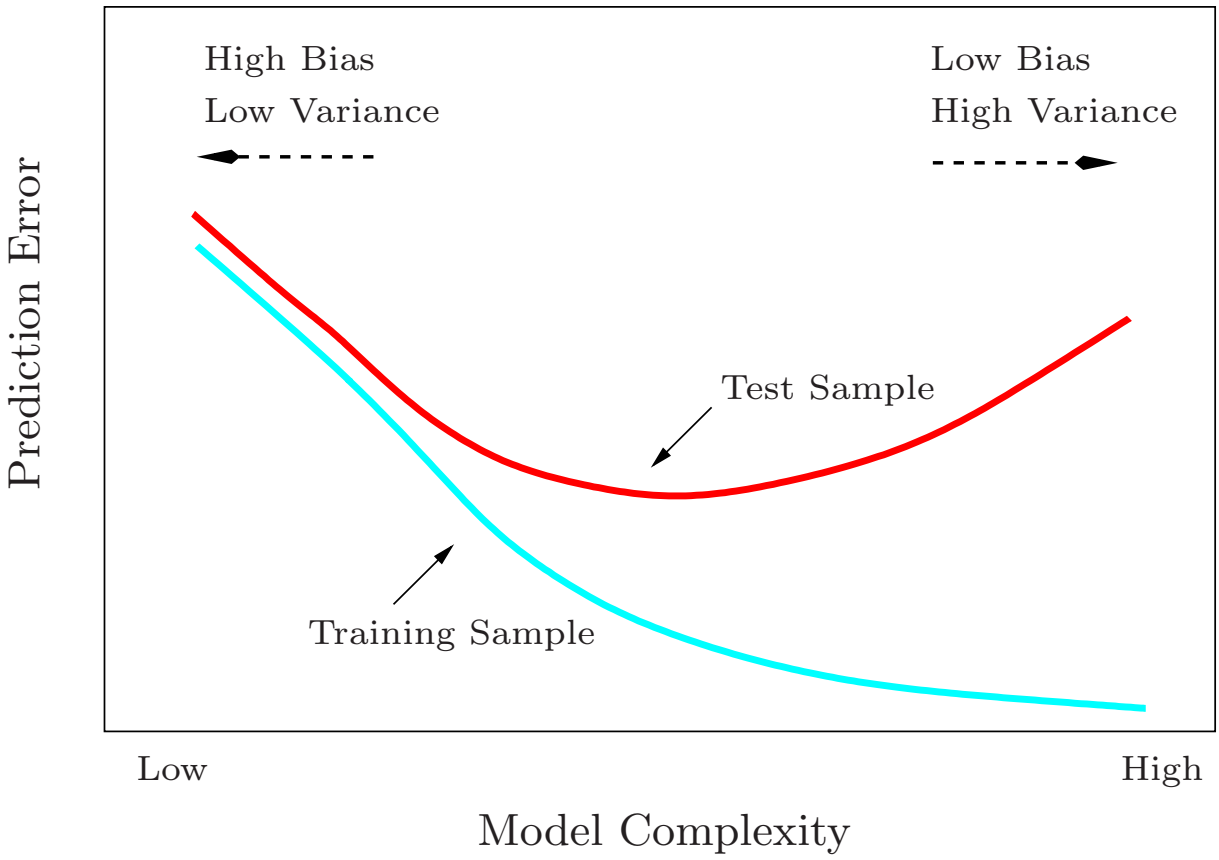
$$
\begin{aligned}
& E\left\{(Y - f_k(x) | X = x)^2\right\} \\
= \; & \sigma^2 + \left[f(x) - \frac{1}{n'} \sum_{l=1}^{n'} f_k(x_l)\right]^2 + \frac{\sigma^2}{n'} \\
= \; & \text{irreducible error} + \text{Bias}^2\{\widehat{f}_k(x)\} + Var\{\widehat{f}_k(x)\}
\end{aligned}
$$

  - n' is the number of nearest points of x

See Hastie, Friedman, Tibshirani Section 2.9 for details

Model selection and bias-variance tradeoff

- Select k that minimizes the loss



Hastie, Friedman, Tibshirani Section 2.9

AIC and BIC

- In Normal linear regression, the likelihood:

$$L_p = \frac{1}{(2\pi\sigma^2)^{\frac{n}{2}}} e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(Y_i-\beta_0-\beta_1 X_{1i}-\ldots-\beta_{p-1}X_{p-1,i})^2}$$

$$-2logL_p \propto \frac{1}{\sigma^2}\sum_{i=1}^{n}(Y_i-\beta_0-\beta_1 X_{1i}-\ldots\beta_{p-1}X_{p-1,i})^2$$

- Min -2log(likelihood), while penalizing p

- AIC - Akaike's information criterion

$$AIC = \frac{\text{SSE}_\text{p}}{MSE_P} + 2p \text{ (proportional to } C_p\text{)}$$

also written as $AIC = n\log\left(\frac{\text{SSE}_p}{n}\right) + 2p$

- SBC - Schwarz Bayesian Criterion

$$BIC = \frac{\text{SSE}_\text{p}}{MSE_P} + p\,log(n) \text{ (heavier penalty for } p\text{)}$$

also written as $BIC = n\log\left(\frac{\text{SSE}_p}{n}\right) + \log(n)p$

- Can use to compare non-nested models

Steps of Model Building (1)

- Data examination
  - outliers? errors? missing data?
  - correct records; complete missings; remove unreliable predictors
- Preliminary model investigation
  - scatterplots; correlations between Xs and between Xs and Y; normality of errors
  - potential transformations of Y
  - remove redundant or uninformative variables
  - identify potentially important predictors that are not part of the dataset
    - in designed experiments, randomization helps avoid the bias due to important unobserved predictors

Steps of Model Building (2)

- Further reduction of potential predictors
  - domain knowledge
  - (semi-)automated subset selection techniques
- Model refinement
  - higher-order terms (curvature, interactions)
  - consider influential or atypical observations
  - a small number of competing models can be kept at this stage
- Model validation
  - stability of estimated coefficients on new dataset
  - predictive ability on new dataset
    - one model can be better at estimation, but another better at prediction

Surgical Unit Example, p. 350

- Random sample of 54 patients undergoing a liver operation

- Response `surv` or `lsurv` post-operation survival (or log-survival) time

- Predictor variables
  - `blood` blood clotting score
  - `prog` prognostic index
  - `enz` enzyme function score
  - `liver` liver function score
  - `age` in years
  - `female` gender, 0=male, 1=female
  - `modAlc` and `heavyAlc` alcool use

Getting to know the data

```
> require(RCurl)
# https://netfiles.umn.edu/users/nacht001/www/nachtsheim/5th
> ch09ta01.file <- getURL("[...]", ssl.verifypeer=FALSE)
>
> X <- read.table(textConnection(ch09ta01.file), sep='')
> dimnames(X)[[2]] <- c('blood', 'prog', 'enz', 'liver',
+  'age', 'female', 'modAlc', 'heavyAlc', 'surv', 'lsurv')

> dim(X)
[1] 54 10


> head(X)
 blood prog enz liver age female modAlc heavyAlc surv lsurv
1  6.7   62  81  2.59  50      0      1        0  695 6.544
2  5.1   59  66  1.70  39      0      0        0  403 5.999
3  7.4   57  83  2.16  55      0      0        0  710 6.565
4  6.5   73  41  2.01  48      0      0        0  349 5.854
5  7.8   65 115  4.30  45      0      0        1 2343 7.759
6  5.8   38  72  1.42  65      1      1        0  348 5.852


> sum(is.na(X))
[1] 0
```

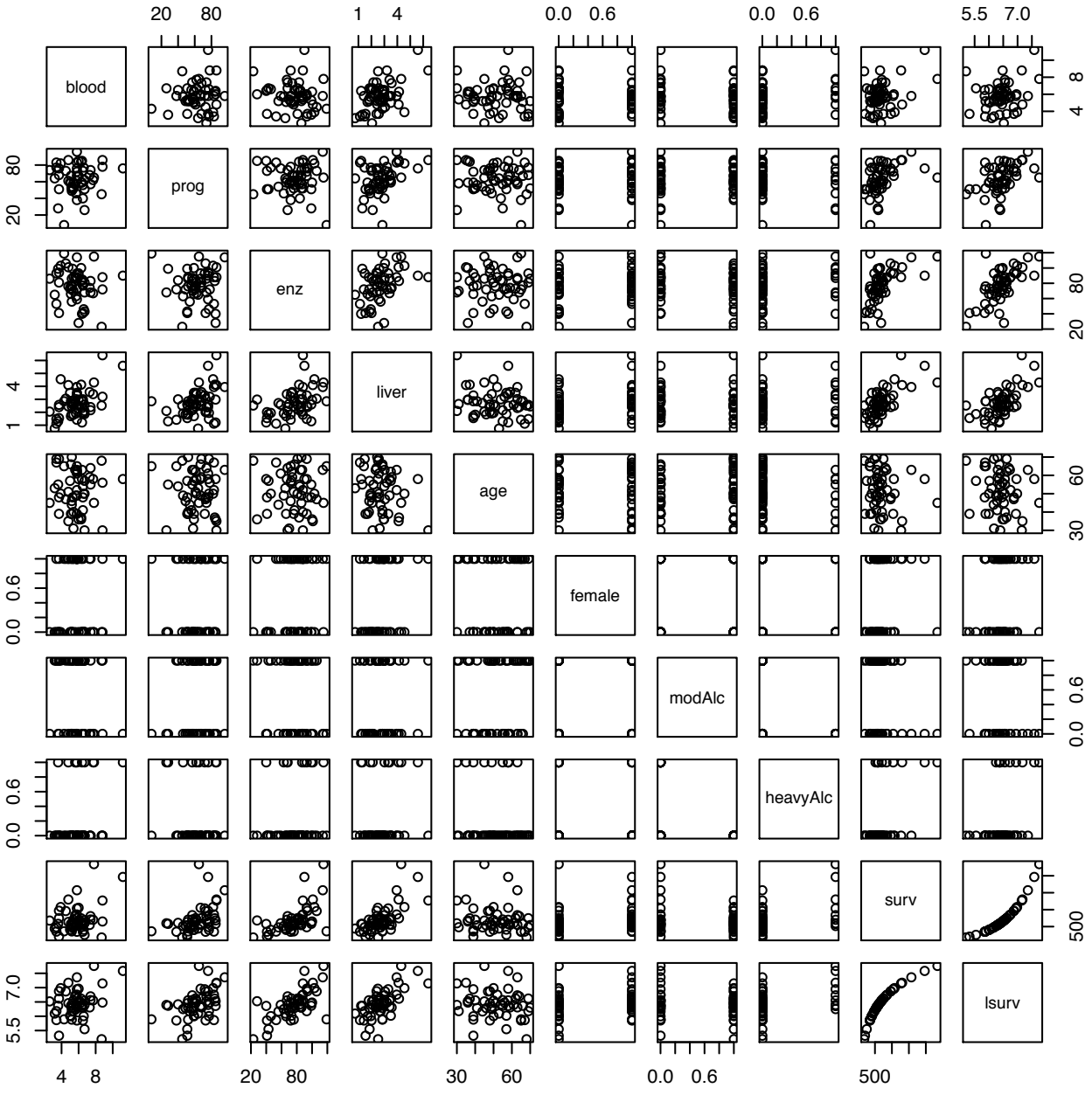
Getting to know the data

Pairwise correlation of predictors:

```
> round(cor(X[,-c(9:10)]), digits=1)
         blood prog  enz liver  age female modAlc heavyAlc
blood      1.0  0.1 -0.1   0.5  0.0    0.0   -0.1      0.2
prog       0.1  1.0  0.0   0.4  0.0    0.1    0.1     -0.1
enz       -0.1  0.0  1.0   0.4  0.0    0.1   -0.1      0.1
liver      0.5  0.4  0.4   1.0 -0.2    0.3    0.0      0.1
age        0.0  0.0  0.0  -0.2  1.0    0.0    0.1     -0.1
female     0.0  0.1  0.1   0.3  0.0    1.0    0.0     -0.1
modAlc    -0.1  0.1 -0.1   0.0  0.1    0.0    1.0     -0.5
heavyAlc   0.2 -0.1  0.1   0.1 -0.1   -0.1   -0.5      1.0
```

```
> pairs(X)
```

Pairwise plots



Exhaustive subset selection

```
> library(leaps)
> # By default - exhaustive search
> regfit.full <- regsubsets(lsurv ~ ., data=X[,-9])
> reg.summary <- summary(regfit.full)
> reg.summary

....

1 subsets of each size up to 8
Selection Algorithm: exhaustive
         blood prog enz liver age female modAlc heavyAlc
1  ( 1 ) " "   " "  "*" " "   " " " "    " "    " "
2  ( 1 ) " "   "*"  "*" " "   " " " "    " "    " "
3  ( 1 ) " "   "*"  "*" " "   " " " "    " "    "*"
4  ( 1 ) "*"   "*"  "*" " "   " " " "    " "    "*"
5  ( 1 ) "*"   "*"  "*" " "   " " "*"    " "    "*"
6  ( 1 ) "*"   "*"  "*" " "   "*" "*"    " "    "*"
7  ( 1 ) "*"   "*"  "*" " "   "*" "*"    "*"    "*"
8  ( 1 ) "*"   "*"  "*" "*"   "*" "*"    "*"    "*"

> names(reg.summary)
[1] "which" "rsq" "rss" "adjr2" "cp" "bic" "outmat" "obj"
```

The `leaps` library uses an efficient branch and bound algorithm

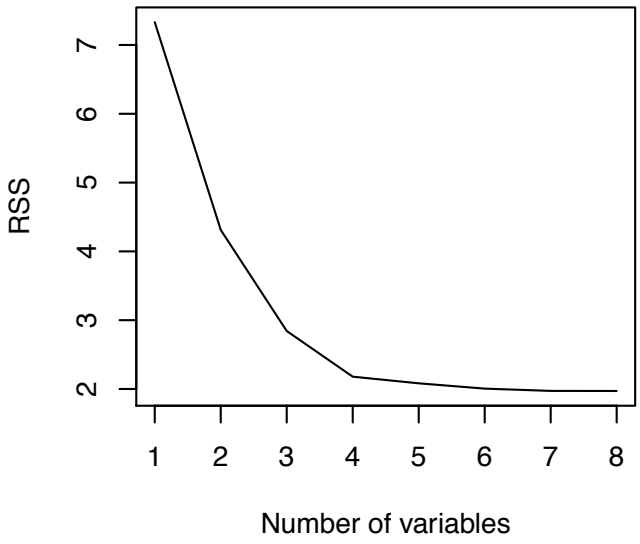
Exhaustive subset selection

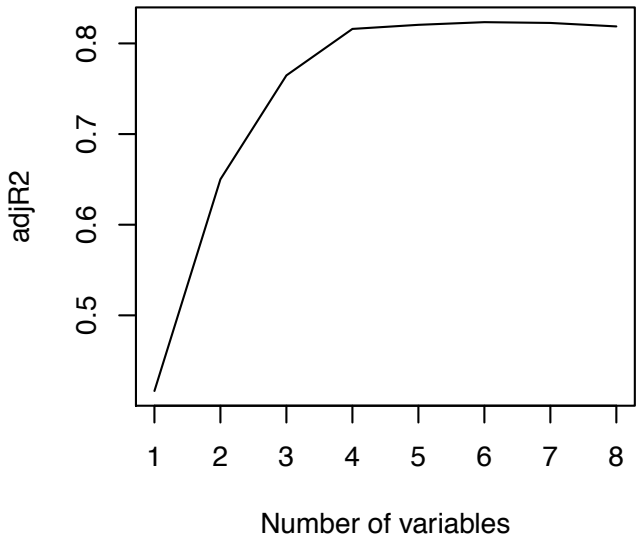
```
> par(mfrow=c(2,2))
> plot(reg.summary$rss, xlab='Number of variables',
+        ylab='RSS', type='l')
> plot(reg.summary$adjr2, xlab='Number of variables',
+        ylab='adjR2', type='l')
> plot(reg.summary$cp, xlab='Number of variables',
+        ylab='Cp', type='l')
> abline(a=0,b=1, lty=3, lwd=2)
> plot(reg.summary$bic, xlab='Number of variables',
+        ylab='BIC', type='l')
```

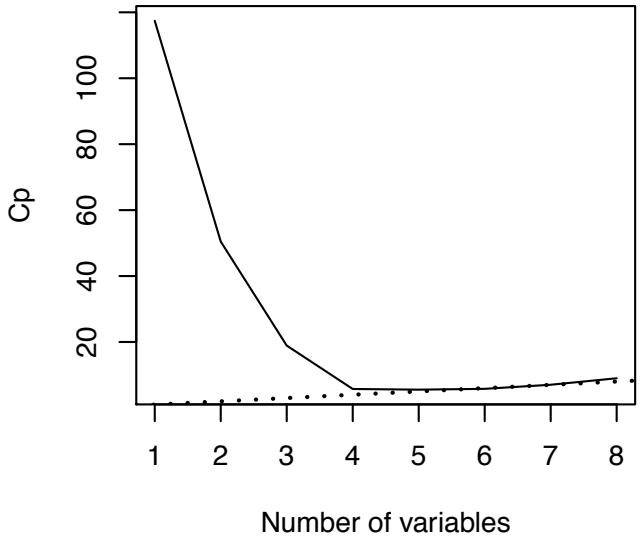
```
> which.min(reg.summary$bic)
[1] 4

> coef(regfit.full, 4)
(Intercept)      blood        prog         enz    heavyAlc
 3.85241856 0.07332263  0.01418507 0.01545270  0.35296762
```

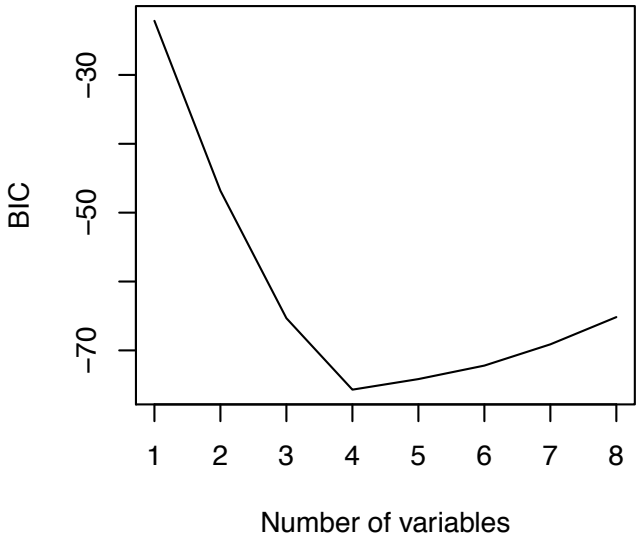
Conclusion: Model with 4 predictors appears best

Best model visualization









Note: BIC has a heavier penalty than C_p

Data-rich situation: independent validation

Randomly partition the dataset into 3 parts

1 Training set
  - predictive ability of any model is too optimistic (model fit caters to the training set)

2 Independent variable selection set
  - select predictors that minimize predictive error on this independent set
  - predictive ability of the "best" model is still too optimistic (variable selection caters to the variable selection set)

- Independent validation set
  - predictive ability of the model on independent data

$$MSPR = \frac{\sum_{i=1}^{n^*} (Y_i - \hat{Y}_i)^2}{n^*}$$

  - n^* = # of observations in validation set

Data-poor situation: cross-validation

- If # of observations is relatively small, but larger than # of variables, randomly partition the dataset into three parts
  - (1) training, (2) var. selection, (3) validation
- Iteratively use each part for training / variable selection / validation
  - each observation will play each role once
  - a value of predictive error for each observation
  - better use of the resources
  - may have a different model at different iteration of cross-validation
- See JWHT Sec. 6.5.3 for R code
  http://www-bcf.usc.edu/ gareth/ISL/
  - Or, use `library(DAAG)`
    Maindonald, J.H. and Braun, W.J. (3rd Ed., 2010) *Data Analysis and Graphics Using R*

Cross-validation: full model

```
> library(DAAG)
> lm.full <- lm(lsurv ~ ., data=X[,-9])
> CVlm(X[,-9], lm.full)

[...]

fold 1
Observations in test set: 18
                1     3    5     12    16     20   [...]
Predicted   6.455 6.387 7.44  5.708 6.503 6.6898
cvpred      6.372 6.227 7.29  5.660 6.541 6.6539
lsurv       6.544 6.565 7.76  5.549 6.695 6.7310
CV residual 0.172 0.338 0.47 -0.111 0.154 0.0771

Sum of squares = 1.25    Mean square = 0.07    n = 18


fold 2
Observations in test set: 18
                  2      7     9     24     25    27  [...]
Predicted    6.0551  6.315 6.625  6.630  6.809 6.220
cvpred       6.0634  6.516 6.600  6.733  7.026 6.166
lsurv        5.9990  6.250 6.962  6.332  6.478 6.302
CV residual -0.0644 -0.266 0.362 -0.401 -0.548 0.136

Sum of squares = 1.3    Mean square = 0.07    n = 18
```

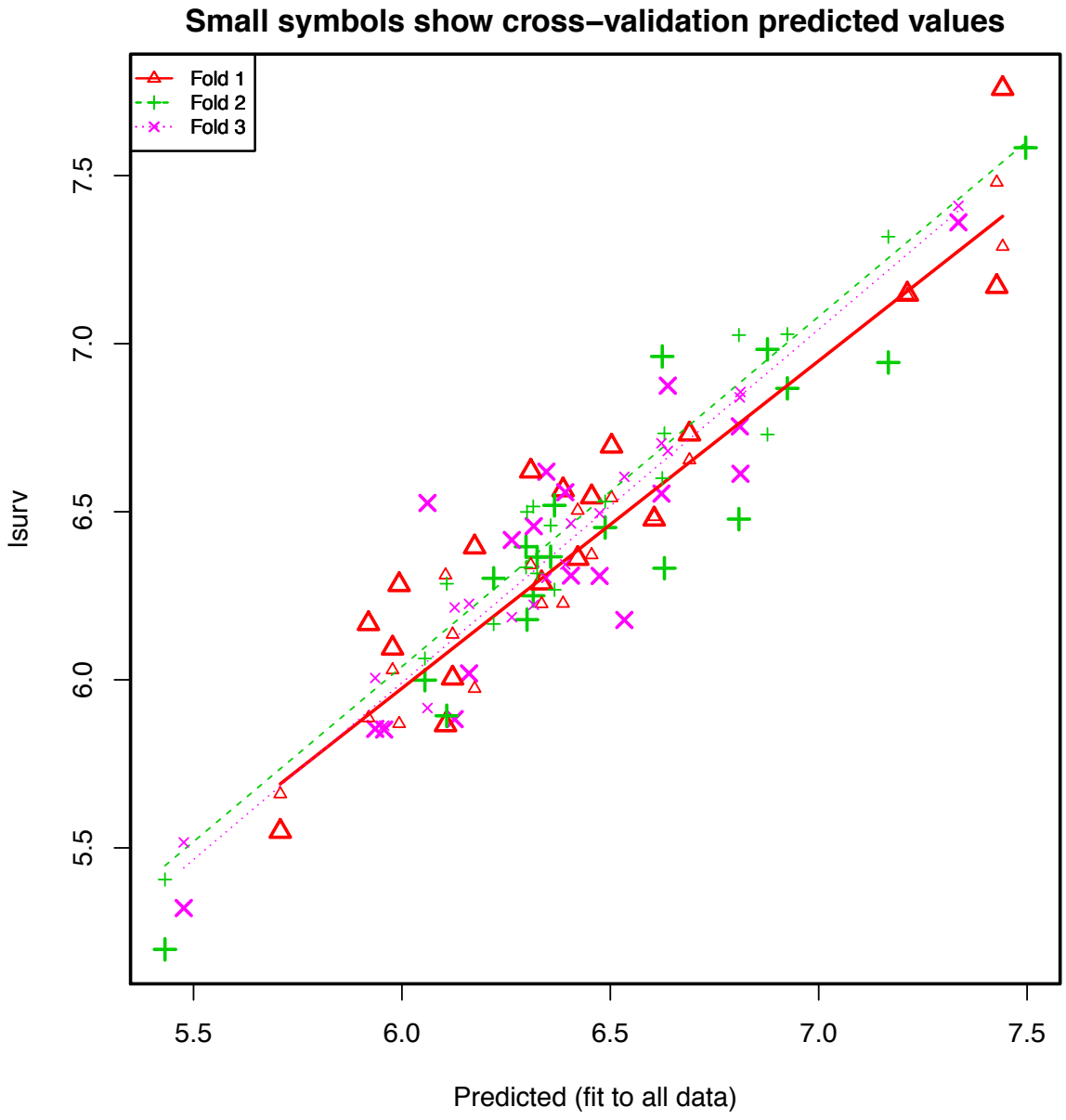
Cross-validation: full model

```
fold 3
Observations in test set: 18
                    4       6     8    10     11      13
Predicted       5.936  5.9574 6.347 6.638  6.812  7.3352
cvpred          6.005  5.8639 6.303 6.681  6.856  7.4101
lsurv           5.854  5.8520 6.619 6.875  6.613  7.3610
CV residual -0.151 -0.0119 0.316 0.194 -0.243

Sum of squares = 1.21    Mean square = 0.07    n = 18

Overall (Sum over all 18 folds)
    ms
0.0696
```

Note: Cross-validation should be done as part of variable selection.

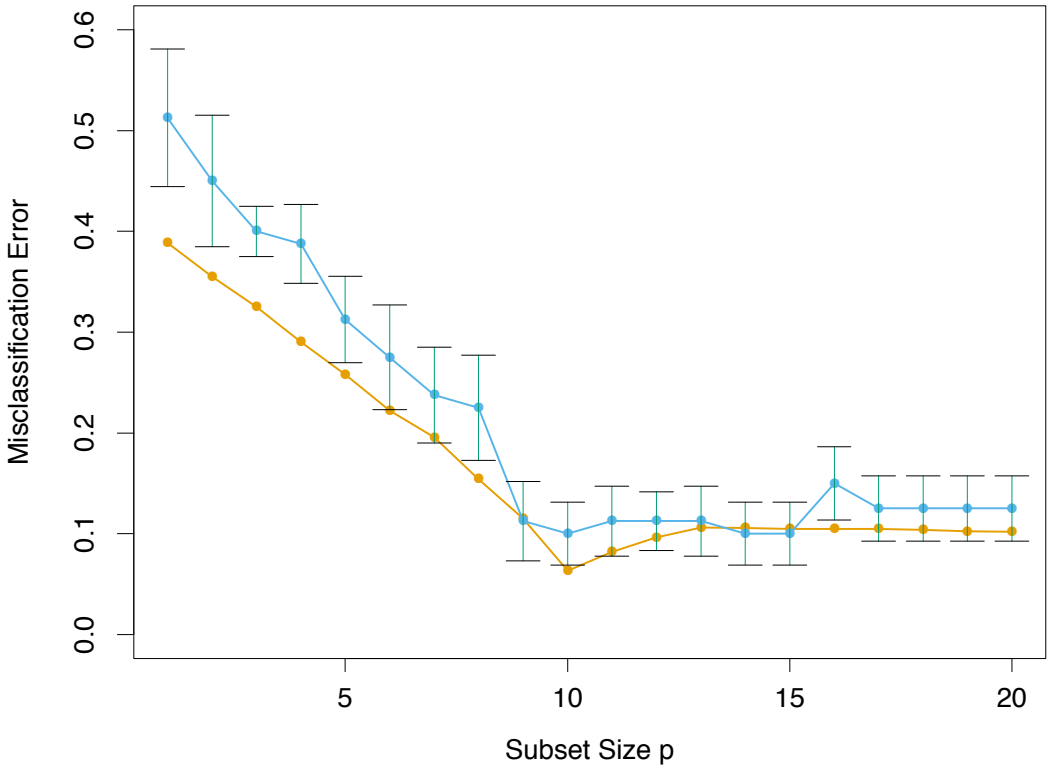
- Selecting the model on the full dataset, and then using cross-validation to evaluate the predictive accuracy, will yield biased results.

Visualization of cross-validation: fit



Cross-validation as part of variable selection

- Orange line: in-sample prediction error
- Blue line: cross-validated prediction error
  - Error bars are obtained over each fold (alternatively, by repeatedly partitioning data into folds)



From Hastie, Tibshirani, Friedman *The elements of Statistical Learning*, 2nd Ed., Springer

Small n large p situation: Variable selection by regularization.

Ridge, Lasso and Elastic Net

Linear Regression: Ridge

- Consider linear model
  $\mathbf{Y} = \mathbf{X}\beta + \varepsilon,\ \varepsilon \sim \mathcal{MVN}(0, \sigma^2\mathbf{I})$

- Properties of least squares estimates
  - $\widehat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$
  - unbiased: $\mathrm{E}\{\widehat{\beta}\} = \beta$
  - minimal $\mathrm{Var}\{\widehat{\beta}\}$ among all unbiased estimates

- In problems with many predictors and few replicates, $\mathrm{Var}\{\widehat{\beta}\} = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$ is large

- Solution by ridge regression: introduce a bias to reduce the variance
  - reduce the overall mean squared error
    $\mathrm{MSE} = \mathrm{E}\{(\widehat{\beta} - \beta)^2\} = \mathrm{Bias}^2\{\widehat{\beta}\} + \mathrm{Var}\{\widehat{\beta}\}$

Parameter Estimation in Ridge Regression

- Ordinary least squares estimates
  - Standardize the predictors $\mathbf{X}^*$
  - All β are comparable after standardization
  - $\widehat{\beta} = (\mathbf{X}^{*\prime}\mathbf{X})^{-1}\mathbf{X}^{*\prime}\mathbf{Y}^*$ (standardized regression)
  - $\text{Var}\{\widehat{\beta}\} = \sigma^2(\mathbf{X}^{*\prime}\mathbf{X}^*)^{-1}$
  - Difficulty inverting $\mathbf{X}^{*\prime}\mathbf{X}^*$ when multicollinearity
- Ridge regression estimates
  - $\widehat{\beta} = (\mathbf{X}^{*\prime}\mathbf{X}^* + \lambda\mathbf{I})^{-1}\mathbf{X}^{*\prime}\mathbf{Y}^*$
  - Biased estimates $\widehat{\beta}$, but stable $(\mathbf{X}^{*\prime}\mathbf{X}^* + \lambda\mathbf{I})^{-1}$
- Difficulty: choice of λ
  - λ varies between datasets, subjective choice
- Difficulty: approximate inference only

Connection to Penalized Least Squares

- Can show that ridge parameter estimates can be obtained by minimizing

$$\sum_{i=1}^{n}\left[Y_i^* - (\beta_0^* + \beta_1^* X_{i1}^* + \ldots + \beta_{p-1}^* X_{ip}^*)\right]^2 + \lambda \cdot \left[\sum_{j=1}^{p-1} \beta_j^{*2}\right]$$
$$= (\mathbf{Y}^* - \mathbf{X}^*\boldsymbol{\beta}^*)'(\mathbf{Y}^* - \mathbf{X}^*\boldsymbol{\beta}^*) + \lambda \cdot \boldsymbol{\beta}^{*\prime}\boldsymbol{\beta}^*$$

  - No intercept if variables are standardized **and** centered

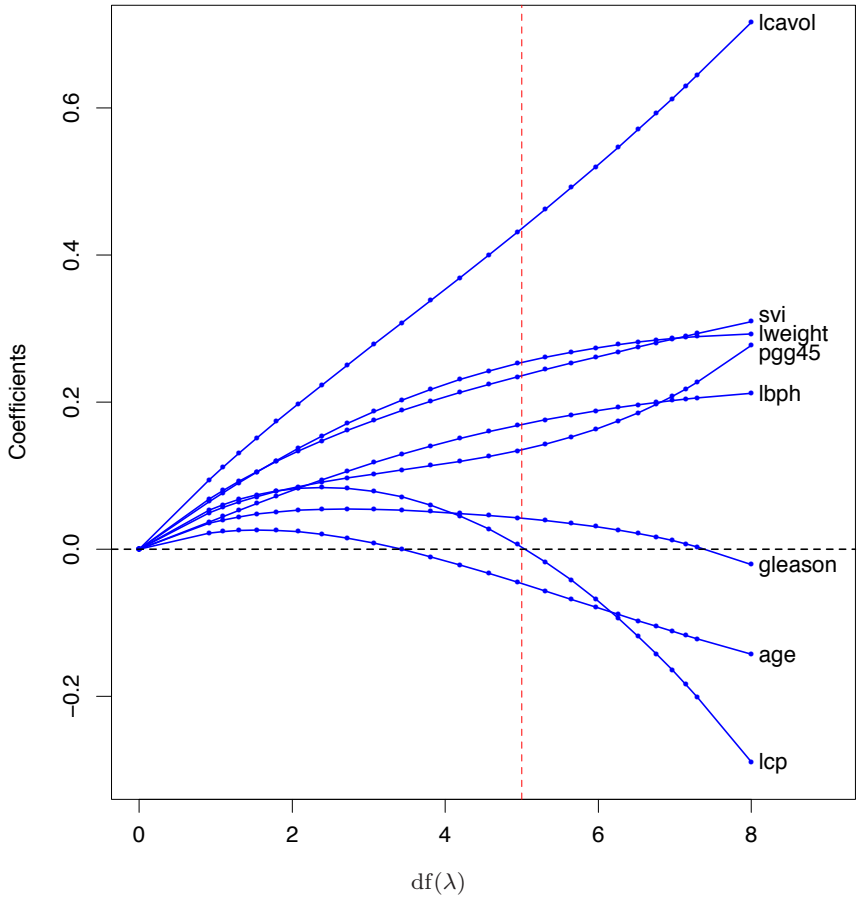
- Or by minimizing

$$\sum_{i=1}^{n}\left[Y_i^* - (\beta_0^* + \beta_1^* X_{i1}^* + \ldots + \beta_{p-1}^* X_{ip}^*)\right]^2 s.t. \sum_{j=1}^{p-1} \beta_j^{*2} \leq t$$

  λ and t have a one-to-one correspondence

- Or as the mean or mode of a posterior distribution, with a suitably chosen prior

Example: Ridge



X axis: 'effective degrees of freedom'

$$df(\lambda) = \text{tr}\left[\mathbf{X}(\mathbf{X}'\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}'\right] = \sum_{j=1}^{p} \frac{d_j^2}{d_j^2 + \lambda}$$

Here d_j are the eigenvalues of $\mathbf{X}$

I.e., variables with smaller d_j have more shrinkage

Hastie, Tibshirani, Friedman
The Elements of Statistical Learning 2008

Adaptive choice of λ: Ridge Trace

- Ridge trace:
  - Simultaneous plot of all parameter estimates for different values of $\lambda \geq 0$.
  - Curves may fluctuate widely when $\lambda \approx 0$
  - Eventually stabilize and converge to $\hat{\beta} = 0$ for large λ.

- Choose λ
  - Where things tend to "stabilize"
  - Better yet, use cross-validation

- λ determines the amount of penalty
  - Large λ is associated with smaller $|\beta_j^*|$
  - Therefore these are *shrinkage* estimators

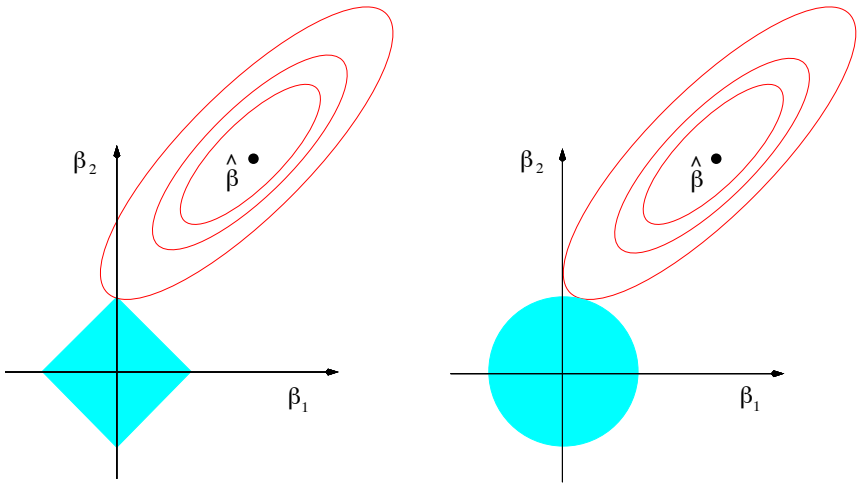
Modifications to Penalized Least Squares: LASSO

- Least absolute shrinkage and selection operator (LASSO): minimize

$$\sum_{i=1}^{n}\left[Y_i^* - (\beta_0^* + \beta_1^* X_{i1}^* + \ldots + \beta_{p-1}^* X_{ip}^*)\right]^2 + \lambda \cdot \left[\sum_{j=1}^{p-1} |\beta^*|_j\right]$$

- Weaker penalty on β than Ridge
  - Solution is non-linear in y_i.
  - Efron et al. (2004) derived an algorithm to obtain a sequence of solutions for discrete λ
  - A subset of parameters in the solutions = 0
  - It can be viewed as a stepwise procedure with a single addition to or deletion from the set of nonzero regression coefficients at any step.
  - Usually estimate $\widehat{\beta}_{LS}$ for the selected predictors

Graphical illustration of regularized estimation



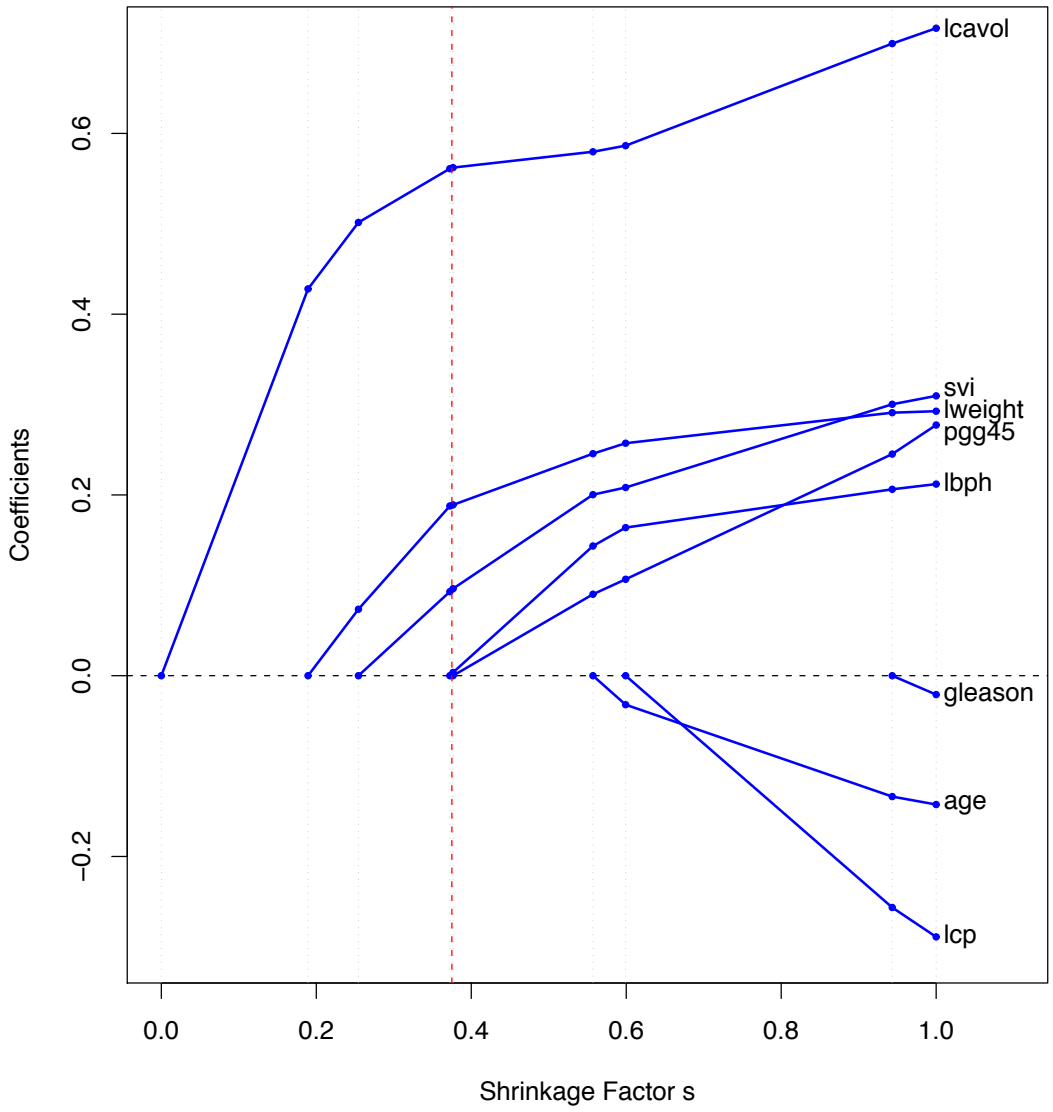
- Iterative least steep coordinate descent
  - Objective function convex, differentiable, convex in each dimension
  - Compute the solution at $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_r$
  - For tuning parameter value λ_k, initialize coordinate descent at the computed solution for λ_{k+1}
  - LASSO optimum is on intersection with axes.

    $\Rightarrow$ LASSO is more effective for feature selection.

    $\Rightarrow$ LASSO is faster for sparse problems.

Hastie, Tibshirani, Friedman
The Elements of Statistical Learning 2008

Example: Lasso

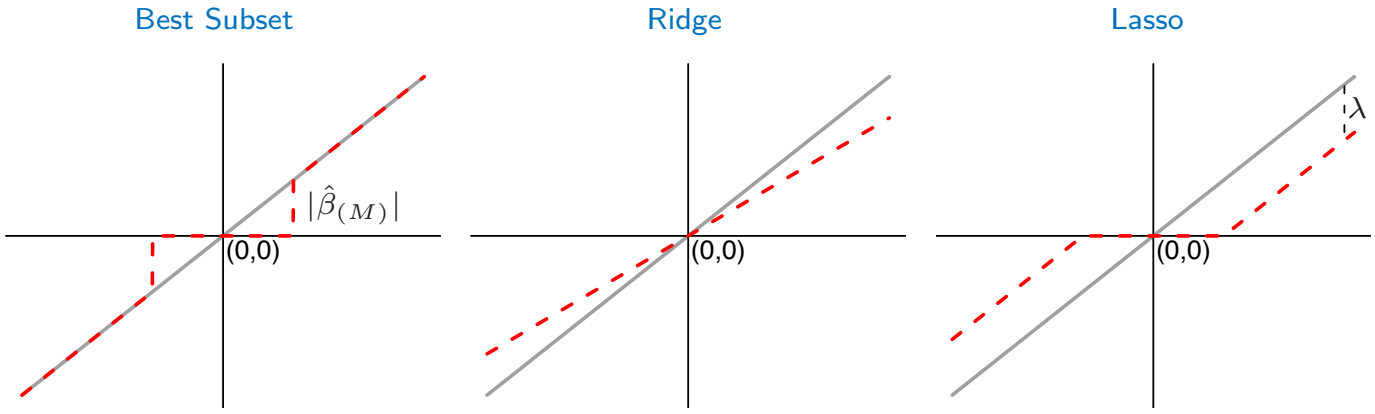


Shrinkage factor $s = t / \sum_{j=1}^{P} |\beta_j^*|$

Hastie, Tibshirani, Friedman
The Elements of Statistical Learning 2008

Explicit estimators in case of orthonormal columns

Estimator	Formula
Best subset (size M)	$\hat{\beta}_j \cdot I(\lvert\hat{\beta}_j\rvert \geq \lvert\hat{\beta}_{(M)}\rvert)$
Ridge	$\hat{\beta}_j/(1+\lambda)$
Lasso	$\text{sign}(\hat{\beta}_j)(\lvert\hat{\beta}_j\rvert - \lambda)_+$

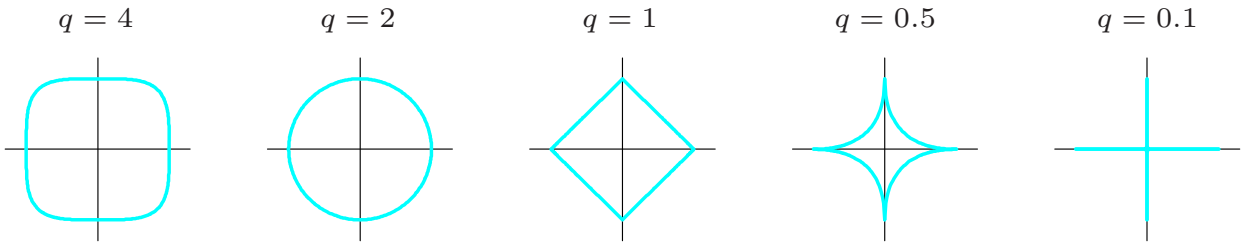


- Best subset: least squares estimates, if selected
- Ridge: Scaled least squares estimates
- Lasso: soft thresholding

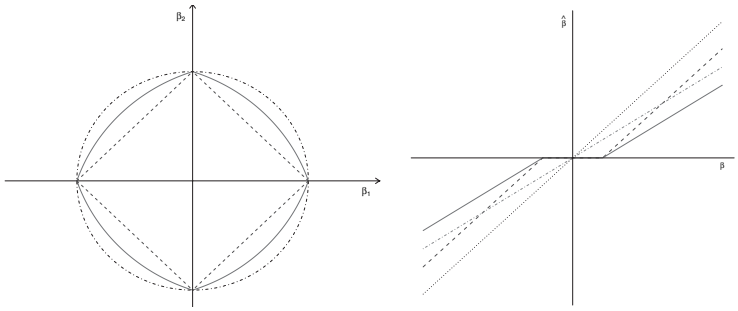
Hastie, Tibshirani, Friedman
The Elements of Statistical Learning 2008

More extensions

Power penalty $\lambda \cdot \sum_{j=1}^{p-1} |\beta_j^*|^q$



Elastic net penalty $\lambda \cdot \sum_{j=1}^{p-1} \left[\alpha \beta_j^{*2} + (1-\alpha)|\beta^*|_j\right]$



Hastie, Tibshirani, Friedman
The Elements of Statistical Learning 2008 Sec. 3.4.3

Ridge selection

```
library(glmnet)
grid=10^seq(10,-2,length=100)
ridge.mod <- glmnet(x=as.matrix(X[,-c(9:10)]), y=X[,10],
+       alpha=0, lambda=grid)
plot(ridge.mod$lambda, ridge.mod$beta)

> ridge.mod$lambda[20]
[1] 49770236
> coef(ridge.mod)[,20]
  (Intercept)        blood         prog          enz
 6.430481e+00 7.387398e-10  1.337399e-10 1.479902e-10 [...]

> ridge.mod$lambda[80]
[1] 2.656088
> coef(ridge.mod)[,80]
  (Intercept)        blood         prog          enz
 5.9968742835 0.0093395444  0.0019819344 0.0021961768 [...]
```

- α is a mixing parameter in elastic net penalty

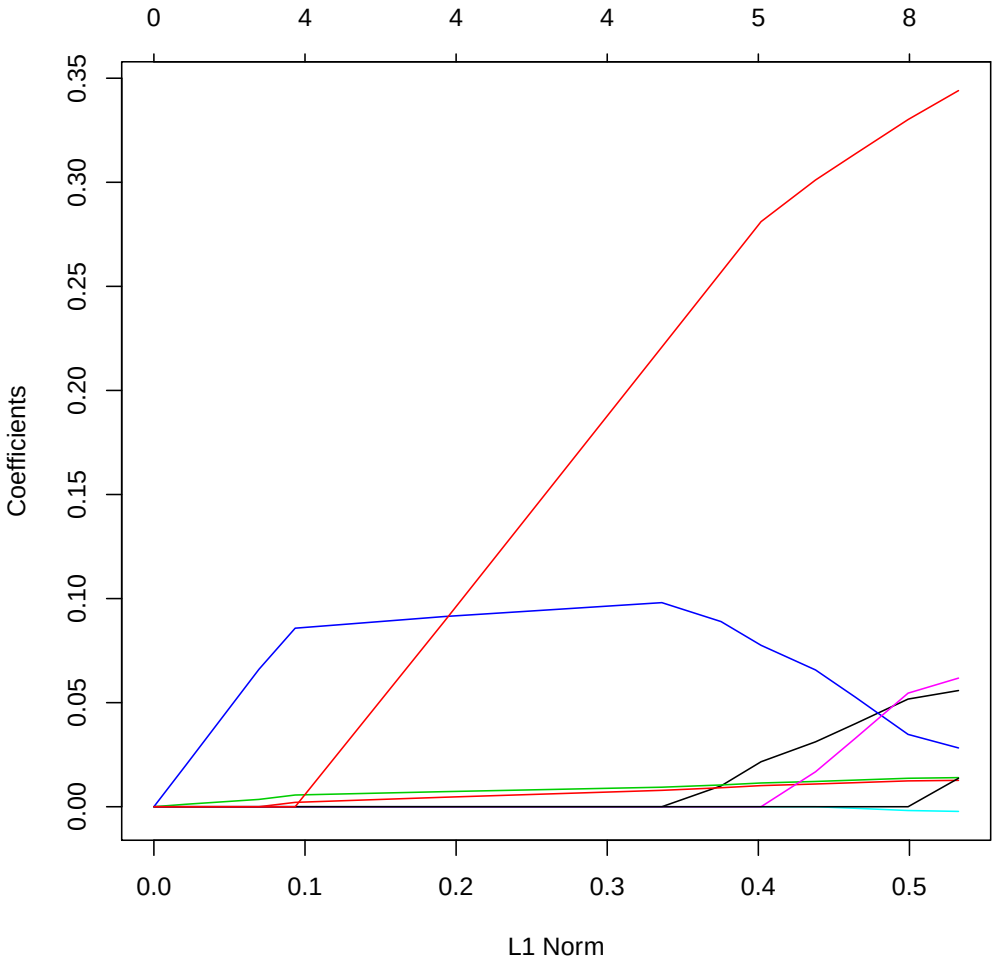
$$(1-\alpha)/2||\beta||_2^2 + \alpha||\beta||_1$$

- By default, $\mathbf{X}$ is standardized but Y is not
- Smaller λ corresponds to larger coefficients

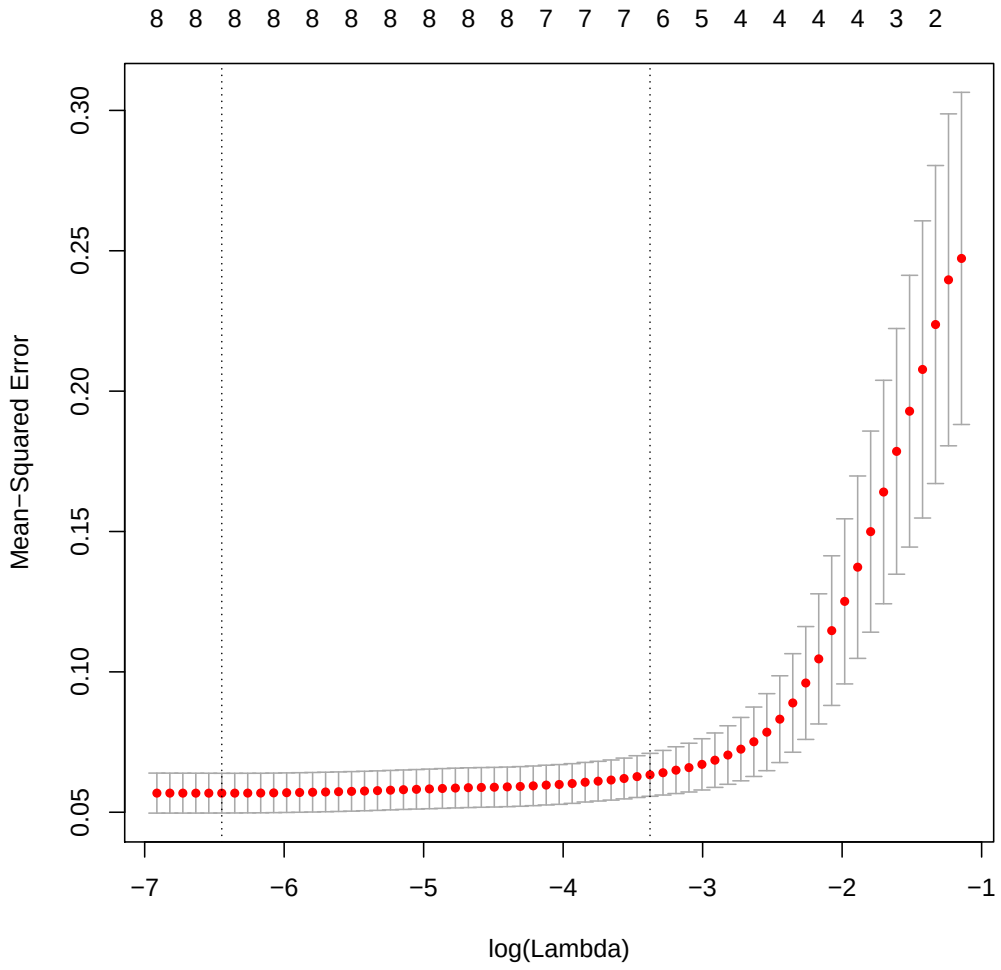
Lasso fit

```
> # Fit lasso for a pre-specified (set of) lambda
> #-----------------------------------------------
> lasso.mod  <- glmnet(x=as.matrix(X[,-c(9:10)]),
+       y=X[,10], alpha=1, lambda=grid)
> plot(lasso.mod)
>

> # Select optimal lambda by cross-validation
> #-----------------------------------------------
> cv.out <- cv.glmnet(x=as.matrix(X[,-c(9:10)]),
+      y=X[,10], alpha=1)
> plot(cv.out)
> bestlam <- cv.out$lambda.min
> bestlam
[1] 0.00144
```

Lasso model fit



Select λ by cross-validation



- Usually refit the model with the selected predictors on the training set
- Evaluate the predictive ability on the validation set