STAT 535

MORE ON "LINEAR STUFF"

Updating Least-Squares Fit

Solve $(X'X)\beta = X'y$.

Now add a column to X and want new fit.

Have to start from scratch? Not necessarily.

Say $R = Chol(X)$. Then $R^* = Chol([Xv])$ has form

$$R^* = \begin{pmatrix} R & a \\ 0' & b \end{pmatrix},$$

where $R'a = X'v$, and $b^2 = v'v - a'a$.

Generalized Inverse?

If X is full rank, $\hat{\beta} = (X'X)^{-1}X'y$

If $rank(X) < p$, many solutions to $(X'X)\beta = X'y$.

Say B is a generalized inverse of A if $ABA = A$.

Further conditions to get the unique Moore-Penrose inverse A^+ of A.

$\hat{\beta} = (X'X)^+X'y$ is the solution to the normal equations having smallest norm.

Condition Number

Confusing! - multiple definitions.

One definition, condition number of invertible matrix is ratio of largest to smallest eigenvalues.

Larger condition number for $X^{prime}X$ means solution to $(X^{prime}X)\beta = X'y$ is more sensitive to small changes in X and y.

Also view as reflecting closeness to singularity.

Common to hear talk of ill-conditioning!

Penalized Least-Squares?

$$\hat{\beta} \quad = \quad \text{argmin}_\beta \left\{ \|y - X\beta\|^2 + \lambda(\beta-\mu)'B(\beta-\mu) \right\}$$

WHY???

Think of terms as representing fit and (prior) plausibility of possible solutions.

Ridge Regression
Penalized likelihood
MAP Bayesian inference

Computation?

Differentiate to get

$$(X'X + \lambda B)\beta \quad = \quad X'y + \lambda B\mu$$

QR or SVD decomposition of X no longer directly helpful.

But Choleski decomposition of $X'X + \lambda B$ applicable as before (remember two backsolves).

Example: estimating a function from noisy observations

Interested in $f(t)$.

Observe $Y_i = f(t_i) + \epsilon_i$, $\epsilon_i \sim N(0, \sigma^2)$

Say $(t_1, \ldots, t_{101})$ is equal-spaced sequence from 0 to 20

Not sure linear form is appropriate.

Allow flexibility by regarding $f()$ as piecewise linear, parameterized as $f(0) = \beta_0, f(1) = \beta_1, \ldots, f(20) = \beta_{20}$.

In matrix form, $Y \sim N(X\beta, \sigma^2 I)$.

What about penalty term. In many applications don't expect the true $f()$ is very "wiggly".

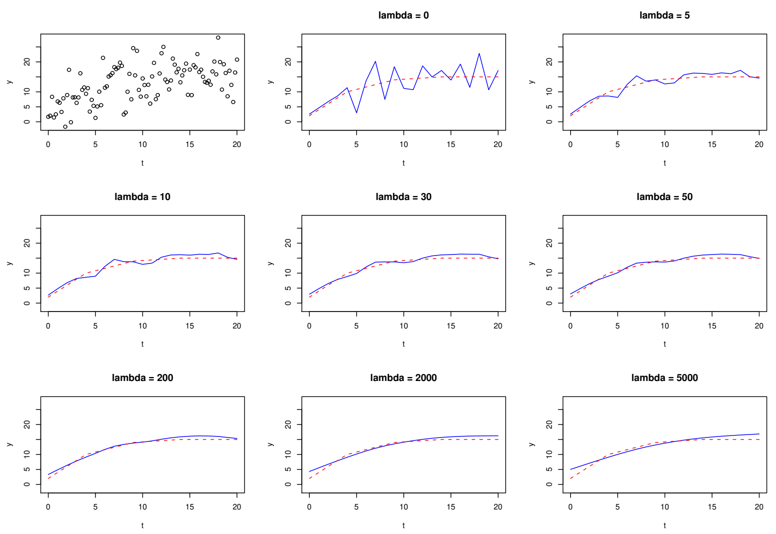
How to measure and penalize "wiggle"?

Choose B (and $\mu = 0$) such that

$$\beta' B \beta \quad = \quad \sum_{i=2}^{20} \left\{ \beta_i - \frac{(\beta_{i-1} + \beta_{i+1})}{2} \right\}^2$$

Form of B?

Plots of data and $\hat{f}()$ (corresponding to $\hat{\beta}$) for different values of λ ...



Numerical check - how much wiggle in solutions?

```
u   <- chol( t(x) %*% x + lambda*b)
tmp <-  backsolve(u, t(x) %*% y + lambda*b%*%mu, trans=T)
ans <-  backsolve(u, tmp, transpose=F)

print(c(lambda,
    sum((ans[2:(p-1)]-.5*ans[1:(p-2)]-.5*ans[3:p])^2)))

[1]    0   893
[1]    5    21.7
[1]   10     8.23
[1]   30     1.66
[1]   50     0.777
[1]  200     0.121
[1] 2000     0.024
[1] 5000     0.013
```


Other applications of penalties?

“ridge regression” type applications

$\beta_1, \ldots, \beta_p$ are the effects of the same treatment at p different hospitals. Penalize in a way that pushes the p estimates toward one another.

$\beta_1, \ldots, \beta_p$ are regression coefficients for p predictors in a regression. Penalize as $\sum_{i=1}^p \beta_i^2$, as alternative to “model selection.” (Or try $\sum_{i=1}^p |\beta_i|$ for Lasso).

Bayesian hierarchical models, shrinkage, ...

Bayesian Interpretation

Model expressed as distribution for $Y|\beta$.

Prior beliefs expressed as prior distribution for β.

Turn Bayes theorem crank to obtain posterior distribution for $\beta|y$.

One point estimate for β would be the mode of the posterior density.

Interpret penalized least-squares estimate this way....