Lecture 3: Just a little more math
Last time

Through simple algebra and some facts about sums of normal random variables, we derived some basic results about orthogonal regression.

We used as our major test case a set of orthogonal polynomials as computed by R.

We derived the mean and variance of OLS estimates and established their distribution; in particular, we found that OLS estimates are unbiased.
Last time

We then considered a model selection rule and demonstrated how the orthogonal model allowed us to compute the “best” fit exactly, no matter how many variables we entered into our model.

We interpreted this rule in terms of a keep-or-kill operator acting on the OLS estimates.
Today

We start with one more property of OLS estimates as formalized by the Gauss-Markov theorem.

We use this result to think about why we might want to use a model selection criterion in the first place.

We then introduce a new kind of estimate that acts by shrinking the OLS coefficients; we see that it also has many desirable properties.

We finish by seeing this new shrinkage estimate as coming from a penalized least squares criterion.
Some rationale

Ok, this class is meant to be an applied tour of regression, and this quarter generalized linear models. But...

Sticking to OLS leaves us somewhere in the late 1800’s and early 1900’s methodologically

Today’s lecture will bring us up to the mid-1970’s and by Monday, we will be discussing techniques developed in your lifetime
Some history

R. A. Fisher is often credited as the single most important figure in 20th century statistics

Before Fisher, statistics was an "ingenious collection of ad hoc devices" (Efron, 1996)

Statistical Methods for Research Workers (1958) looks remarkably like most of the texts we use for Statistics 11, 13, and so on
Some history

He is responsible for creating a mathematical framework for statistics; but equally important was his impact on data analysis.

Fisher once commented that over his calculator, “he had learned all he knew”.

At one time, Fisher believed that model specification was outside the field of mathematical statistics; his work focused on estimation assessing uncertainty after a model has been selected.
Some history

John Tukey is another pioneer of statistical theory and practice.

In his text *Exploratory Data Analysis* (1977), Tukey creates graphical tools for exploring features in data.

Many of the graphical displays in R were developed by Tukey.

His style is iterative, advocating many different analyses.

John Wilder Tukey (1915-2000)

From the Bell Laboratories archives, circa 1965
stat.bell-labs.com/who/tukey
What happened between Fisher’s time (the 1920’s and 30’s) and Tukey’s (the 1960’s and 70’s)?
The computer

In 1947 the transistor was developed

In 1964 the IBM 360 was introduced and “quickly becomes the standard institutional mainframe computer”*

By the end of the 1960’s computer resources were generally available at major institutions

* Taken from a PBS history of the computer, www.pbs.org/nerds/timeline/micro.html
The computer

The computer changed the process of modeling, allowing statisticians to attempt larger, more challenging problems.

In the late 70’s Box and his co-authors articulated a pipeline of analysis, a feedback loop that runs from model specification, to fitting and diagnostics and back.

Once we were free to think about model selection, there is an irresistible urge to formalize the process.

Hence, in the 70’s you see the rise of selection criteria like AIC, BIC and Cp.
And since 1970?
Home computers and the Internet

First personal computer in 1975 (MITS Altair 8800)

In 1977 Apple introduces the Apple II at a price of $1,195; 16K of RAM, no monitor

The first spreadsheet, VisiCalc, ships in 1979 and is designed for the Apple II

The Apple Macintosh appears in 1984

Microsoft Windows 1.0 ships in 1985

* Taken from a PBS history of the computer, www.pbs.org/nerds/timeline/micro.html
Home computers and the Internet

In the 1990’s there is a migration to “ubiquitous computing”: There are small but powerful computers in phones, PDAs, cars, you name it.

The internet (or rather a nationwide fiber optic network) connects us, with wireless access becoming standard.

At the same time, technologies for data collection, and in particular those associated with environmental monitoring are undergoing a small revolution in the form of sensor networks.

All of these developments have had an enormous impact on the practice of statistics.
Upcoming events

An Undergraduate Summer Program
Data visualization and its role in the practice of statistics

http://summer.stat.ucla.edu

June 19-25, 2005
UCLA Campus

Designed for university sophomores and juniors with backgrounds in some quantitative field; hands-on projects related to current “hot topics” in statistics

Bioinformatics
Computer vision
Earth and space sciences
Computer networks
Industrial statistics
... but I digress
Regression revisited

Recall the basic assumptions for the normal linear model

- We have a response $y$ and (possible) predictor variables $x_1, x_2, \ldots, x_p$

- We hypothesize a linear model for the response

$$y = \beta_1^* x_1 + \beta_2^* x_2 + \cdots + \beta_p^* x_p + \epsilon$$

where $\epsilon$ has a normal distribution with mean 0 and variance $\sigma^2$
Regression revisited

Then, we collect data...

- We have a $n$ observations $y_1, \ldots, y_n$ and each response $y_i$ is associated with the predictors $x_{i1}, x_{i2}, \ldots, x_{ip}$

- Then, according to our linear model

$$y_i = \beta_1^* x_{i1} + \beta_2^* x_{i2} + \cdots + \beta_p^* x_{ip} + \epsilon_i$$

and we assume the $\epsilon_i$ are independent, each having a normal distribution with mean 0 and variance $\sigma^2$
Regression revisited

To estimate the coefficients $\beta_1^*, \beta_2^*, \ldots, \beta_p^*$ we turn to OLS, ordinary least squares (this is also maximum likelihood under our normal linear model)

- We want to choose $\beta_1, \beta_2, \ldots, \beta_p$ to minimize the OLS criterion

$$\sum_{i=1}^{n} [y_i - \beta_1 x_{i1} - \beta_2 x_{i2} - \cdots - \beta_p x_{ip}]^2$$

- You recall that this is just the sum of squared errors that we incur if we predict $y_i$ with $\beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip}$
Regression revisited

Computing the OLS estimates was easy; for example, we can differentiate this expression with respect to $\beta_1$ and solve to find

$$\hat{\beta}_1 = \sum_i y_i x_{i1}$$

From here we found that $E \hat{\beta}_1 = \beta_1^*$ and that $\text{var}(\hat{\beta}_1) = \sigma^2$
Regression revisited

OLS estimates have a lot to recommend them

- They are unbiased (repeated sampling)

- Gauss-Markov Theorem: Among all “linear” unbiased estimates, they have the smallest variance (BLUE or best unbiased linear estimates)
Regression revisited

The BLUEs

To see this, look at some other linear estimate of $\beta_1^*$; that is, we can write

$$\tilde{\beta}_1 = \sum d_i y_i$$

Then, if we set $c_i = d_i - x_{1i}$ we can rewrite it as

$$\tilde{\beta}_1 = \hat{\beta}_1 + \sum c_i y_i$$
Regression revisited

The BLUEs

For \( \tilde{\beta}_1 \) to be unbiased, we must have

\[
E\tilde{\beta}_1 = E\hat{\beta}_1 + \sum c_i E\hat{y}_i
\]

\[
= \beta^*_1 + \beta^*_1 \sum c_i x_{i1} + \cdots + \beta^*_p \sum c_i x_{ip}
\]

If this is to hold for all values of \( \beta^*_1, \ldots, \beta^*_p \) we must have

\[
\sum c_i x_{i1} = 0 \quad \sum c_i x_{i2} = 0 \quad \text{and so on}
\]
Regression revisited

The BLUEs

For the variance of $\tilde{\beta}_1$ we just follow our noses

\[
\text{var} \tilde{\beta}_1 = \text{var} \sum d_i y_i \\
= \text{var} \sum (c_i + x_{i1}) y_i \\
= \sum (c_i + x_{i1})^2 \text{var}(y_i) \\
= \sigma^2 \sum x_{i1}^2 + \sigma^2 \sum c_i^2 + 2\sigma^2 \sum c_i x_{i1} \\
= \sigma^2 + \sigma^2 \sum c_i^2
\]

which is larger than $\sigma^2$ unless the $c_i$ are all zero
Mean squared error

OK, so the OLS estimates are looking pretty good; but suppose we give up unbiasedness?

• Define the mean squared error of any estimator $\hat{\beta}$ to be

$$E(\hat{\beta} - \beta^*)^2 = E(\hat{\beta} - \hat{\beta}^* + \hat{\beta}^* - \beta^*)^2 = E(\hat{\beta} - \hat{\beta}^*)^2 + (\hat{\beta}^* - \beta^*)^2 = \text{var} \hat{\beta} + (\hat{\beta}^* - \beta^*)^2$$

• This relationship is often called the bias-variance tradeoff; the first term is the variance of an estimator and the second is its squared bias
Mean squared error

The Gauss-Markov theorem implies that our OLS estimates have the smallest mean squared error among all linear estimators with no bias.

That begs a question: Can there be a biased estimator with a smaller mean squared error?
Shrinkage estimators

Let’s consider something that initially seems crazy

We will replace our OLS estimates $\hat{\beta}_k$ with something slightly smaller

$$\tilde{\beta}_k = \frac{1}{1 + \lambda} \hat{\beta}_k$$

If $\lambda$ is zero, we get our OLS estimates back; if $\lambda$ gets really really big, things crush to zero
Mean squared error

Working out the mean squared error, we again follow our noses to find

\[
\sum_{k=1}^{p} E(\hat{\beta}_k - \beta^*_k)^2 = \sum_{k=1}^{p} E\left(\frac{1}{1 + \lambda} \hat{\beta}_k - \beta^*_k\right)^2
\]

\[
= \left(\frac{1}{1 + \lambda}\right)^2 \sum_{k=1}^{p} E(\hat{\beta}_k - \beta^*_k)^2 + \left(\frac{\lambda}{1 + \lambda}\right)^2 \sum_{k=1}^{p} \beta^*_k
\]

\[
= p\sigma^2 \left(\frac{1}{1 + \lambda}\right)^2 + \left(\frac{\lambda}{1 + \lambda}\right)^2 \sum_{k=1}^{p} \beta^*_k
\]
Mean squared error

Let’s look at these two terms

\[ p\sigma^2 \left( \frac{1}{1 + \lambda} \right)^2 + \left( \frac{\lambda}{1 + \lambda} \right)^2 \sum_{k=1}^{p} \beta_k^* \]

The first is the variance component; it is smallest when \( \lambda \) is zero; the second is the squared variance and it goes to zero as \( \lambda \) gets large (why?)
Shrinkage

In principle, with the right choice of $\lambda$, we can get an estimator with a better MSE.

The estimate is not unbiased, but what we pay for in bias, we make up for in variance.
Mean squared error

We can find the minimum by balancing the two terms

\[ p\sigma^2 \left( \frac{1}{1 + \lambda} \right)^2 + \left( \frac{\lambda}{1 + \lambda} \right)^2 \sum_{k=1}^{p} \beta_k^* \]

It is not hard to show that the minimum must occur at

\[ \lambda = \frac{p\sigma^2}{\sum \beta_k^*} \]
Mean squared error

Therefore, if all of the coefficients are large relative to their variances, we want to set $\lambda$ small.

On the other hand, if we have a lot of small coefficients, we will want to pull them closer to zero.
Comparison

So, if our wee bit of math suggests we can do better in terms of MSE by shrinking the coefficients when we have lots of small ones, what advice does AIC give?

How does it compare?
Next time

OK, so the last two days have been a lot of math; I want you to focus on the concepts rather than the algebra

Think about how you might design a simulation to see how well subset selection is doing versus shrinkage

We will talk about this next time and you will implement it for homework next week