Math 455 Class 28 April 7

Midterm 2: After April 15 but before the last week of classes. Covers at least through Ch. 9.

Emphasizes material covered after MT 1.

New HW (from Ch. 8) will be posted tonight.

Notes (from last week) are now on course website.

Last time: Generalized Least Squares (+ special case, Weighted Least Squares)

Summary: if our error $\varepsilon = \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{bmatrix}$ does NOT satisfy $V(\varepsilon) = \sigma^2 I$

but instead satisfies $V(\varepsilon) = \sigma^2 \Sigma$, where Σ is some other matrix, what can we do?

Answer: linear alg. reduces to the standard case after we make some adjustments. This is GLS.

This can be done automatically for you, with R.

(Discussed French Provincial Election Example Ch. 8)

Other topics in Ch. 8:

"Goodness-Of-Fit" tests

How well does the model fit the data?

Crude example: is the R^2 high?

Another example: are the data plausibly from a normal distribution? Shapiro-Wilk Test. for normality asks "how well does a normal dist fit our data?"

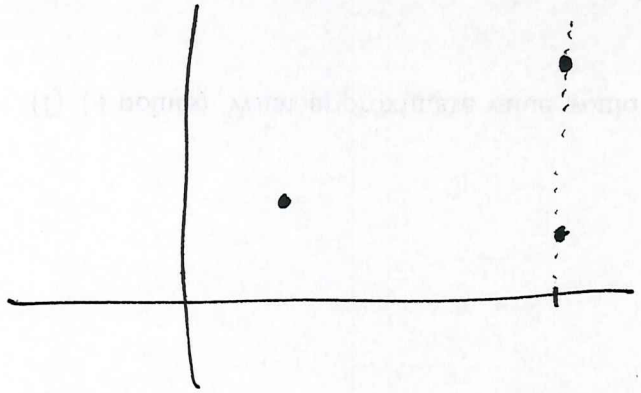
Example: χ^2 test (more about this later.)

Saturated Models: this means a model with as many parameters as we have rows of data.

Q: Wait! this seems silly. We can fit a degree $d-1$ polynomial (d parameters) to any d points. Why use a saturated model?

A: First, even if # parameters = # rows, we might NOT be able to have a perfect fit, because of the functional form of our model. What if, for example, there are two points with

Exactly the same X-values.



We can't fit a quadratic polyn. perfectly to these 3 points.

Q: Ok, funny exception acknowledged.
Why a saturated model?

A: When you're working on any statistical problem it's good to have two "benchmarks" for your model. One, a "baseline" which is very weak. The second should be "too good", more than you can hope to achieve.

Example: We're trying to predict the winner of football games. A baseline: every team has a 50% chance of winning every game. Too good: a model that uses the score at halftime. (Uses information not available.)

What if, in the context of our problem, we can't find a natural model that is to good? We can use a "saturated model".

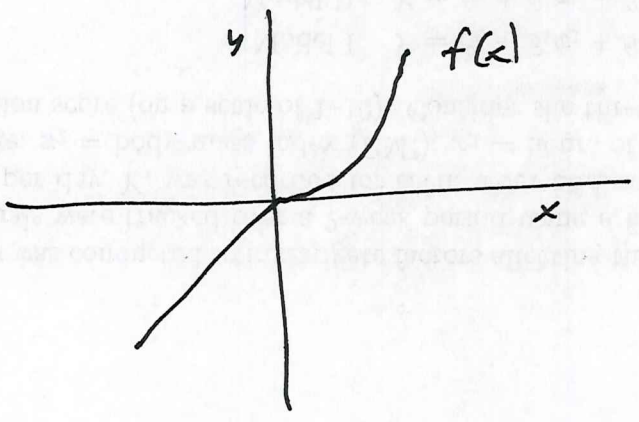
Other reasons to consider a saturated model: we are using regression with penalty, so the extra parameters matter less.

There is a lot of detail on splines.

A "spline" is a function which is "piecewise polynomial". Example:

$$f(x) = \begin{cases} x & x < 0 \\ x^2 & x \geq 0 . \end{cases}$$

This is a continuous function:



But $f'(x)$ is NOT continuous.

There are various requirements that could be placed on splines, e.g. $f'(x)$ continuous.

These splines are used when we want to fit a continuous, nonlinear fcn to our data, but aren't sure about the functional form.

8.4. M-Estimation:

What if we wanted to minimize something other than $RSS = \sum_{i=1}^{n} e_i^2$.

We can minimize $\sum_{i=1}^{n} |e_i|$.

"Least Absolute Deviation" (LAD).

We could compromise between LAD and OLS:

Choose a c and define.

$$\rho(x) = \begin{cases} x^2/2, & |x| \le c \\ c|x| - c^2/2, & \text{if } |x| > c. \end{cases}$$

Then minimize $\sum_{i=1}^{n} \rho(e_i)$.

This is called "Huber's Method":

"OLS up to c and LAD after that."

This can be done using Weighted Least Squares.

(There is a package in R.)