Math 455 Class 28 April 4

Big picture: we learned how to do a regression in R and how to interpret it.

We have been studying "what if something goes wrong?"

The diagnostic plots will help us figure this out.

The most important diagnostic plot is "Fitted vs Residuals" (The first one obtained with plot(lmod) in R).

The principle: any pattern we can see suggests an improvement to the model.

But we also have other ways of checking if things have gone wrong. e.g. tests for collinearity.

Q: How do we know when we've got it right?

A: We don't, at least not for most people's definition of right.

Most people's def'n of right means something like "a causal relationship".

(Between the predictor & response.)

Regression, by itself, does not establish the existence of a causal relationship.

When we look at diagnostic plots, test for collinearity, etc, we are trying to make sure that whatever relationship we see didn't happen for a stupid reason.

Q: There are lots and lots of tests: how do we know we are done?

A: We don't. The tests, even all of them together, don't eliminate all problems. You have to know, for domain-specific reasons, which problems are most likely, and test for those.

We'll look at a couple of examples in time-series: the rows of data correspond to increasing times, e.g. row i is day i of n days.

Errors ε / Residuals in time series data

For example, if we can (somehow) predict e_i from e_{i-3} and e_{i-5}, the DW statistic might not "see" this.

- This is usually OK: the prediction usually is mostly about e_{i-1}.

Coding: we looked at the durtest code in Ch. 6. In that example we found that e_i was (somewhat) predictable from e_{i-1}.

What do we do about this?

We can change our model to reflect the predictability of e_i from e_{i-1}.

There are packages to do this for you automatically.

One simple model of somewhat-predictable errors:

$$\varepsilon_i = \phi \varepsilon_{i-1} + \delta_i$$

where $\delta_i \sim N(0, \tau^2)$.

here $|\phi| < 1$ or we'll have problems.

Package : nlme

Command : ~~glsrobd~~ gls

there is a "switch" to the command that specifies the time-dependence of the error.

that is the "correlation = ..." part of the code.

Q: How would we know to do this in advance?

A: Experience and domain-specific knowledge.

Or: collaboration between statisticians and domain experts.

Interpretation of models may change when we make changes, e.g.

$$\log(Y) = \beta_0 + \beta_1 X + \varepsilon \qquad (1)$$

vs. $$Y = \beta_0 + \beta_1 X + \varepsilon. \qquad (2)$$

You learned how to interpret $\hat{\beta}_1$ in (2).

The interpretation of $\hat{\beta}_1$ in (1). is different.

Likewise, if X is a "Categorical Variable"

e.g. X = "Red", "Yellow", "Green"

(no numerical value)

R will still do the regression and give coefficients: we should know how to interpret these!