Math 455 Class 19 March 9

Chapter 6 - Diagnostics.

The assumptions of linear regression:

Linearity : an approximate linear relation

Independence : of errors

Homoscedasticity : errors have same variance

Normality : errors Normal.

Q: How do we know whether these are true?

A: This is the wrong question.

They are virtually never true about real data.

Instead we should ask a statistical question:

HOW true are these assumptions?

We answer this question (in part) with diagnostic plots.

Q: What are diagnostic plots?

A: If we have lmod <- lm(Y ~ X)

What should we actually do?

A: at a minimum, look at the 4 plots you get "for free" with plot(lmod).

To do this, we need to know what these plots are.

What is the Q-Q plot and what is it checking?

The Q-Q plot plots the sorted residuals against $\Phi^{-1}\left(\frac{i}{n+1}\right)$ where Φ is the cdf of std. normal.

Why? Q-Q is checking <u>normality</u>.

Residuals are a stand-in for errors.

The idea is that sorted residuals should behave like order statistics. (Wackerly 6.7)

Let's imagine that we have 100 samples from a std. normal and we order them.

quadratic pattern in the residuals, then you could try adding an X^2 term.

If you see that a residual variance increases with fitted values, you could change the model.

See diagrams Faraway Ch. 6 p. 74.

Fundamental Questions:

(1) Are the data we have relevant to the problem we are trying to solve?

If NO, stop!

(2) OK, we go around "checking assumptions" by looking at "diagnostic plots". How do we know when we are done?

A: We do not. This is not pure math. This is more like plumbing, or corporate law. We can always make things better, but they will never be perfect.

And we say plot (lmod)

We get 4 plots:

1. Fitted vs Residuals

2. Q-Q.

3. Scale-Location. (Fitted vs. $\sqrt{\text{Std. Resid.}}$)

4. Leverage vs. Standardized Residuals

or : Cook's Distance

Residuals vs Fitted:

$[e_i]$ $[\hat{Y}_i]$

This tests linearity and (to some extent) homoskedasticity, two assumptions of linear regression.

Big idea: if a pattern is present in this plot, the model can be improved.

Example: if you see that there is a

Start: $\varepsilon_1, \ldots, \varepsilon_{100}$. Ordered $\varepsilon_{[1]} \le \ldots \le \varepsilon_{[100]}$.

The distribution of ε_i is $N(0,1)$.

The distribution of $\varepsilon_{[1]}$ is NOT.

It is very sharply peaked at $\Phi^{-1}\left(\frac{1}{101}\right)$.

The distn of $\varepsilon_{[i]}$ is very sharply peaked at $\Phi^{-1}\left(\frac{i}{n+1}\right)$. This is why the defn of the Q-Q-plot is what it is.

Idea: <u>if</u> the normality assumption is satisfied, the dots in the Q-Q plot will be approximately on the line $y=x$.

If there is a deviation, we can see the way normality fails.