Math 455 Class 29 April 8

"Robust Regression" Section 8.4

Robust = not affected much by outliers / errors in data.

Big theme of this book: alternatives to ordinary regression are often equivalent to ordinary regression after some modification.

Example: We decide to use MLE (Max. Likelihood) instead of OLS. This is equivalent: the ML estimators are the OLS estimators, assuming the standard hypotheses for linear regression.

Example: We change the hypotheses of regression from $\varepsilon \sim N(0, \sigma^2 I)$ to $\varepsilon \sim N(0, \sigma^2 \Sigma)$.

This is equivalent to OLS after a transformation of the variables (See 8.1)

Example: If we want to minimize (instead of $SSR = \sum_{i=1}^{n} e_i^2$), $\sum_{i=1}^{n} \rho(e_i)$; this is equivalent to ordinary regression with weights.

We discussed two functions which are common choices for ρ (apart from $\rho(x) = x^2$) $\leftarrow$ 1.

2. Least absolute deviation (LAD)

$$\rho(x) = |x|.$$

3. Huber:

$$\rho(x) = \begin{cases} x^2/2 & : |x| \leq c \\ c|x| - c^2/2 & : |x| > c \end{cases}$$

Notice that these agree if $|x| = c$:

$$c|c| - \frac{c^2}{2} = c^2 - \frac{c^2}{2} = \frac{c^2}{2} \text{ if } x = +c$$

and if $x = -c$ $\quad c|-c| - \dfrac{(-c)^2}{2} = \dfrac{c^2}{2}.$

So the ρ is continuous.

What is c?

c is usually chosen to be an estimate of σ.

Faraway suggests a value "proportional to the ~~mid~~ median value of $e_1, \ldots, e_n$."

In practice: this choice will be made for you by the "robust regression" package selected.

Example in text: rlm function in MASS package. [Also quantreg package]

When we use these "robust" methods, we lose some information: R^2 is not given, because it doesn't make much sense in this context.

We said that the robust methods are equivalent to ~~the~~ weighted least squares: these packages allow you to extract the weights.

Why do we care what the weights are?

If points are significantly down- or up-weighted, we should try to figure out <u>why</u>.

Extreme case of weighted least squares:

"Least Trimmed Squares". (LTS)

Some weights are <u>zero</u>.

Instead of minimizing $\sum_{i=1}^{n} e_i$ we write

$e_{(1)} \leq e_{(2)} \leq \cdots \leq e_{(n)}$ for the <u>sorted</u>

residuals, and minimize $\sum_{i=1}^{q} e_{(i)}^2$.

For a choice of q.

That is, we weight residuals $e_{(q+1)}, \ldots, e_{(n)}$

at ZERO.

⚠ Ignoring certain errors or data points should be cause for concern.

This may make sense if we know that there are data entry errors or exceptional cases which we can't find in advance ~~etc~~, or want to ignore.

Some of these methods sacrifice things we get with OLS, e.g. LTS doesn't give standard errors for coeff. estimates.

Often we can recover these things using approximate numerical techniques, e.g. Bootstrap.

LTS involves a choice of q.

Default: $q = \lfloor \frac{n}{2} \rfloor + \lfloor \frac{(p+1)}{2} \rfloor$