4.2: Variations on OLS (Ordinary Least Squares)

Dr. Bean - Stat 5100

1 Why alternatives?

Remember this: when standard model assumptions are met, OLS is the "best" linear modeling approach.
No matter how good we are at performing variable transformations, there are some situations where we simply cannot satisfy linear model assumptions of constant variance, normality, or independence.

Fortunately, there are several OLS alternatives that address one or more of these issues.

The cost:

- Lose our ability to conduct inference on the coefficients.
- The models become harder to fit/harder to explain.

2 Weighted Least Squares (textbook §11.1)

Recall regression model $Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_{p-1} X_{p-1} + \varepsilon$ in matrix form:
(Ch. 5, Handout #12 p. 2)

$$Y = X\beta + \varepsilon$$

Model assumption: $\varepsilon \sim N(0, \sigma^2 I)$

- If constant variance, (i.e., $Cov(\varepsilon) = \sigma^2 I$), then use OLS:

$$b = (X'X)^{-1}X'Y$$

- If non-constant variance, then can estimate and account for it (WLS):

$$V = Cov(\varepsilon) \qquad \text{(typically assumed diagonal)}$$

$$W = V^{-1} \qquad \text{(i.e. the weights)}$$

$$b_w = (X'WX)^{-1}X'WY$$

Why give *smaller* weight to observations with *larger* variance when calculating the model coefficients?

Smaller variance is equivalent to greater certainty. Certain information should have greater "value" than uncertain information.

Typically, $Cov(\varepsilon)$ must be estimated

- can often relate variance of residuals (or squared residuals) to predictors or $\hat{Y}$ values
- example (as in Ex. 1 of Handout 4.2.1): residual vs. X_1 is megaphone-shaped (linear relationship between SD of residual and X_1)
  - regress absolute residuals on X_1 and get predicted values s (as function of X_1)
  - define weights $w = 1/s^2$
- see p. 425 for other examples
  – key is how to estimate w for given scenario, as a function of X's

Some things to remember:

- The *pattern* of the residuals against the other variables determines how we should estimate the weights.
- Its OK to see non-constant variance in weighted model.
- In **Spatial Statistics, weights are calculated using geographic similarity.**

3 Robust Regression (textbook §11.3)

Rather than remove influential observations and outliers, we may choose to reduce their influence by changing the way we measure "error".

3.1 IRLS (iteratively reweighted least squares)

1. Obtain (maybe from OLS) b, then calculate $\hat{Y} = Xb$ and $e = Y - \hat{Y}$
2. Calculate weights W, based on e (lots of weight functions available)
3. Calculate (WLS) $b_w = (X'WX)^{-1}X'WY$ and resulting $e = Y - Xb_w$
4. Iterate steps 2 & 3 to convergence of b_w

How to calculate weights?
- usually chosen to optimize some criterion
- the choice of criterion determines the method of weight calculation

3.2 M-estimation

- If $u_1, \ldots, u_n$ are *iid* from some distribution with parameter θ, then the type-M estimate of θ is of the form

$$\hat{\theta} \;=\; \arg\min \sum \rho(u_i; \theta)$$

  where ρ is some "scalar objective function"

- Example: $\rho(u;\theta) = -\frac{1}{n}\log f(u;\theta)$, f is pdf of distribution of $u_1, \ldots, u_n$. Then

$$\begin{aligned} \hat{\theta} &= \arg\max \sum \frac{1}{n} \log f(u_i;\theta) \\ &= \arg\max (\text{likelihood}) \\ &= (\text{what is this called?}) \end{aligned}$$

- W-estimation approach in IRLS:
  1. Calculate robust estimate of σ, such as $s = \frac{MAD(e)}{0.6745}$
  2. Let $u_i = \frac{e_i}{s}$ be "scaled" (or standardized) residual
  3. Calculate (diagonal) weights $w_i = \frac{\psi(u_i)}{u_i}$
     – where $\psi(u) = \rho'(u)$ for some scalar objective function ρ

Example – Tukey Bisquare (sometimes called Tukey's Biweight):

$$\rho(u) = \begin{cases} \frac{c^2}{3}\left(1 - [1 - \left(\frac{u}{c}\right)^2]^3\right) & \text{if } |u| \le c\text{; default } c = 4.685 \\ \frac{c^2}{3} & \text{otherwise} \end{cases}$$

Bisquare weight function: $w(u) = \left(1 - \left(\frac{u}{c}\right)^2\right)^2$ for $|u| \le c$, 0 otherwise

Note: M-estimation works well for outliers; for leverage points, use MM-estimation (see SAS help)

3. Nonlinear Regression (textbook §13.1 – 13.2)

What if Y vs $X_1, \ldots X_{p-1}$ not linear (in β's)?
– Usually need mechanistic theory

Mechanistic Theory: the assumption that a natural phenomenon can be understood through the use of an equation.
Example: Population Growth

$$\frac{dN}{dt} = rN(1 - N/K)$$

`proc nlin` fits these nonlinear models

- Parameters estimated by an iterative process to reduce the SSE at each iteration, until convergence
- Keys to [useful] convergence:
  - form of nonlinear equation
  - initial parameter estimates

If you were dropped randomly on the side of a mountain with dense fog, how would you find your way down? How would you know when you have made it to the bottom (assuming the fog persists at the bottom)?

You would most likely take each step in a direction that would cause you to be lower than you were before. You would know that you (hopefully) arrived at the bottom of the mountain when you can no longer find a direction to take a step in which you could decrease your altitude. This approach is often called **gradient descent**.

Example 3.1: $Y = \beta_0 + \beta_1 X_1^{\beta_2} - \beta_3 \exp(\beta_4 X_2) \quad (+\epsilon)$
(with simulated data)

Example 3.2: a nonlinear curve to describe sand compression, from Lagioia et al. (1996) Computers and Geotechnics 19(3):171-191

$$f = \frac{p}{p_c} - \frac{\left(1 + \frac{q}{p \cdot M \cdot k_2}\right)^{\frac{k_2}{(1-\mu)(k_1-k_2)}}}{\left(1 + \frac{q}{p \cdot M \cdot k_1}\right)^{\frac{k_1}{(1-\mu)(k_1-k_2)}}},$$

where

f = yield surface (response)
q = deviatoric stress (predictor)
p = mean effective stress (predictor)
p_c = hardening / softening constant defining current size of surface (known)
η = stress ratio p/q
M = parameter defining value of η with no strain increment
μ = parameter defining general slope of d vs. η curve
α = parameter defining how close to $\eta = 0$ axis curve bends towards $d = \infty$
d = dilatancy, $2\mu M(1 - \alpha)$

Goal: find μ, α, and M to make $f \approx 0$, and look at the relationship between these three parameters

`proc model` estimates such nonlinear systems (can do multiple equations)

From playing with this in SAS, it appears that to achieve convergence of estimates in `proc model`, the most important thing is that at least one of the tails of the $q * p$ curve to be fit has data along most of it. To make the convergent estimates "good", it appears necessary to have data along both tails. It is also crucial that the initial starting estimates be good, especially for M (maybe within .2 or so).