STAT 350: Lecture 31

Heteroscedastic Errors

If plots and/or tests show that the error variances $\sigma_i^2 = Var(\epsilon_i)$ depend on i there are several standard approaches to fixing the problem, depending on the nature of the dependence.

- If the variances are known except for a constant factor then we may write $\sigma_i^2 = \sigma^2/w_i$ and use a technique called **weighted least squares**. (See Chapter 10 in the text.)

  This usually arises realistically in the following situations:

  - Y_i is an average of n_i measurements where you know n_i. Then $w_i = n_i$.
  - Plots suggest that σ_i^2 might be proportional to some power of some covariate: $\sigma_i^2 = kx_i^\gamma$. Then $w_i = x_i^{-\gamma}$.

- If the variances are thought to depend on the mean of Y_i then two standard approaches are available:

  - The older approach is **transformation**. We compute $Y_i^* = g(Y_i)$ for some function g like the logarithm or square root. Then we regress Y_i^* on the covariates. This approach sometimes works for skewed response variables like income; after transformation we occasionally find the errors are more nearly normal, more homoscedastic and that the model is simpler. See page 130ff and check under transformations and Box-Cox in the index.
  - The newer approach is to use a **generalized linear model**; see STAT 402.
  - Transformation uses the model

    $$E(g(Y_i)) = x_i^T\beta$$

    while generalized linear models use

    $$g(E(Y_i)) = x_i^T\beta$$

Generally the latter approach offers more flexibility since it is then possible to model the variance as a general function of the mean while for transformation followed by ordinary least squares the transformed data must follow a *homoscedastic* linear model.

- It is also possible to use a hybrid approach in which the parameters are *estimated* by least squares but inference (estimation of standard errors, testing, confidence intervals) are based on a model in which the errors may be heteroscedastic.

Weighted Least Squares

If

$$E(Y_i) = x_i^T \beta$$

and

$$Var(Y_i) = \sigma^2 / w_i$$

and *the errors are independent with normal distributions* then the likelihood is

$$\prod_{i=1}^{n} \frac{\sqrt{w_i}}{\sqrt{2\pi}\sigma} \exp\left[-\frac{w_i}{2\sigma^2}(Y_i - x_i^T\beta)^2\right]$$

To choose β to maximize this likelihood we minimize the quantity

$$\sum_{i=1}^{n} w_i (Y_i - x_i^T \beta)^2 \,.$$

The process is called weighted least squares.

Algebraically it is easy to see how to do the minimization. Rewrite the quantity to be minimized as

$$\sum_{i=1}^{n} \left[w_i^{1/2} Y_i - (w_i^{1/2} x_i)^T \beta\right]^2 \,.$$

This is just an ordinary least squares problem with the response variable being

$$Y_i^* = w_i^{1/2} Y_i$$

and the covariates being

$$x_i^* = w_i^{1/2} x_i \,.$$

The calculation can be written in matrix form. If $W^{1/2}$ is a diagonal matrix with $w_i^{1/2}$ in the ith diagonal position then put $Y^* = W^{1/2}Y$ and $X^* = W^{1/2}X$. Then

$$Y = X\beta + \epsilon$$

becomes

$$Y^* = X^*\beta = W^{1/2}\epsilon$$

If ϵ had mean 0, independent entries and $Var(\epsilon_i) = \sigma^2/w_i$ then $\epsilon^* = W^{1/2}\epsilon$ has mean 0, independent entries $\epsilon_i^* = w_i^{1/2}\epsilon$ and $Var(\epsilon_i^*) = \sigma^2$ so that ordinary multiple regression theory applies. The estimate of β is

$$\hat{\beta}_w = \left[(X^*)^T X^*\right]^{-1} (X^*)^T Y^* = (X^T W X)^{-1} X^T W Y$$

where now $W = W^{1/2}W^{1/2}$ is a diagonal matrix with w_i on the diagonal. This estimate is unbiased and has variance covariance matrix

$$\sigma^2 \left[(X^*)^T X^*\right]^{-1} = \sigma^2 (X^T W X)^{-1}\,.$$

Example

It is possible to do weighted least squares in SAS fairly easily. As an example we consider using the SENIC data set taking the variance of RISK to be proportional to 1/CENSUS. (Motivation: RISK is an estimated proportion; variance of a Binomial proportion is inversely proportional to the sample size. This makes the weight just CENSUS.

```
proc reg  data=scenic;
  model Risk = Culture Stay Nratio Chest Facil;
  weight Census;
run ;
```

EDITED OUTPUT