Stat 414 - Day 4
Heterogeneity Corrected Standard Errors

Last Time

- We care about several different standard deviations
  - Standard deviation of response (related to $(y_i - \bar{y})^2 \;\; or \; Y'Y - n\bar{Y}^2$)
  - Variability and Correlation between explanatory variables (related to $X'X$)
  - Standard deviation of regression coefficients $se(\hat{\beta}) = \frac{\hat{\sigma}}{(n-1)s_x} aka \; \hat{\sigma}^2(X'X)^{-1}$
  - Standard deviation of fits $se(\hat{y}) = \hat{\sigma}\sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{((n-1)s_x^2)}}$ aka $\hat{\sigma}^2(X(X'X)^{-1}X')$
- One of the main reasons for dealing with the heteroscedasticity is otherwise our estimates of the standard errors of our slope coefficients may be off, which impacts our p-values and confidence intervals.

Weighted least squares, a special case of *generalized least squares*, minimizes $\Sigma w_i(y_i - XB)^2$
With GLS: $Var(\hat{\beta}) = (X'X)^{-1}X'\Sigma X(X'X)^{-1} where \; \Sigma = VarCov(\epsilon_{ik})$

- $Var(\epsilon_i) = \sigma^2/w_i$
- Hope the weighted residuals have equal variance ($\sigma^2\boldsymbol{I}$) without impacting linearity and normality of residuals

Common choices of weights include

$V(y_i) = \sigma_i^2$	$w_i = \frac{1}{\sigma_i^2}$	Regress $\lvert e_i \rvert$ on x_i or $\lvert e_i \rvert$ on $\hat{y}_i$ to estimate σ_i Regress e_i^2 on x_i or $\hat{y}_i$ to estimate σ_i^2
$V(y_i) \propto 1/n_i$	$w_i = n_i$	
$V(y_i) \propto x_i^2 \; aka \; SD(y_i) \propto x_i$	$w_i = 1/x_i^2$	Equivalent to regressing $\frac{y}{x}$ on $\frac{1}{x}$
$V(y_i) \propto x_i$	$w_i = 1/x_i$	

Unweighted vs. weighted regression (OLS and REML match)

```
> summary(model1REML)
Generalized least squares fit by REML
  Model: Testisweight ~ DML
  Data: Squid
       AIC      BIC    logLik
  4055.094 4069.018 -2024.547

Coefficients:
                Value Std.Error   t-value p-
(Intercept) -6.534226 0.3925936 -16.64374
DML          0.046660 0.0014749  31.63582

 Correlation:
    (Intr)
DML -0.951

Standardized residuals:
       Min         Q1        Med         Q3
-3.4469532 -0.6797156  0.0477543  0.6189041

Residual standard error: 3.352301
Degrees of freedom: 768 total; 766 residual

Multiple R-squared:  0.5665
```

```
> summary(model2REML)
Generalized least squares fit by REML
  Model: Testisweight ~ DML
  Data: Squid
       AIC      BIC    logLik
  3885.837 3899.761 -1939.919

Variance function:
 Structure: fixed weights
 Formula: ~DML

Coefficients:
                Value Std.Error   t-value p-value
(Intercept) -5.623937 0.3382932 -16.62445       0
DML          0.043065 0.0014061  30.62659       0

 Correlation:
    (Intr)
DML -0.95

Standardized residuals:
        Min          Q1         Med          Q3
-2.66818179 -0.75305668  0.01266351  0.71346611  4.

Residual standard error: 0.1935302
Degrees of freedom: 768 total; 766 residual

Multiple R-squared:  0.5505
```

$V(\epsilon_i) = \sigma^2 \, DML_i$

What is the estimated variance for squid with DML = 136 with the weighted regression?

.19353^2 x 136

How do the parameter estimates change between the two models?

They did not

DML SE = .043 to .0014

How do the standard errors of the regression coefficients compare?

smaller with weighted regression (because sigma-hat smaller)

How does the behavior of the confidence and prediction intervals change?

They are now much wider (most noticeably the PIs) for the larger DML values

Sandwich estimators

If we don't have good candidates for weights (or not appropriate), an alternative approach is stick with OLS but use heterogeneity corrected (HC) "sandwich" standard errors. The main idea is to estimate the standard errors of the coefficients with $(X'X)^{-1}(X'\hat{U}X)(X'X)^{-1}$ where $\hat{U}$ is a diagonal matrix of the squared residuals (aka HC0, White 1980).

- These use the squared residuals to tells us about the variance (and covariance) structure of the residuals
- HC1: scales the residuals by the df (Huber-White)
- HC2: scales the residuals by the leverage values $(1 - h_{ii})$
- HC3: scales the residuals by $(1 - h_{ii})^2$

The main idea is you have taken into account the heteroscedasticity without having to know about or model the functional form of the heteroscedasticity or use "arbitrary" transformations.

In R:

```
library(lmtest); library(sandwich)
sqrt(diag(vcovHC(model1, "HC1"))) # HC1 gives us the White-Huber standard errors
coeftest(model1, vcov = vcovHC(model1, type = "HC1"))   #updates the significance tests
```

(e) How do the standard errors of the slope coefficients change? Does the statistical significance of any of the variables change? (If not, then can claim analysis was not being affected by the heterogeneity.)

In this case, things don't change too much, but they could! DML SE is now .0021 but still statistically significant

Reminders:

- When heteroscedasticity is discovered, we should not simply ask "What can I do to make the problem go away?" without also asking "What does heteroscedasticity tell me about the process I am studying?" (Hayes & Cai, 2007).
- Keep in mind that non-constant variance could be due to a misspecified model (e.g., missing key predictors, interactions, or non-linear effects).