Transformation Approach

- Suppose $Y = X\beta + \varepsilon$ where $\sigma^2(\varepsilon) = W^{-1}$
- Have linear model but potentially correlated errors and unequal variances
- Consider a transformation based on $W$
  $$w^{1/2}Y = w^{1/2}X\beta + w^{1/2}\varepsilon$$
  $$\downarrow$$
  $$Y_w = X_w\beta + \varepsilon_w$$
- Can show $E(\varepsilon_w) = 0$ and $\sigma^2(\varepsilon_w) = I$
- Weighted least squares special case of generalized least squares where only variances may differ ($W$ is a diagonal matrix)

Maximum Likelihood

- Consider
  $$Y_i \sim N(X_i\beta, \sigma_i^2) \quad (\sigma_i's \text{ known})$$
  $$\downarrow$$
  $$f_i = \frac{1}{\sqrt{2\pi\sigma_i^2}} \exp\left\{-\frac{1}{2\sigma_i^2}(Y_i - X_i\beta)^2\right\}$$
- Likelihood function $L = f_1 \times f_2 \times \cdots \times f_n$
- Find $\beta$ which maximizes $L$
- Similar to minimizing
  $$Q_w = \sum_{i=1}^{n} \frac{1}{\sigma_i^2}(Y_i - X_i\beta)^2$$

Weighted Least Squares

- Expressed in matrix form
  $$Q_w = (Y - X\beta)' W (Y - X\beta)$$
- Where
  $$W = \begin{bmatrix}
  1/\sigma_1^2 & 0 & \cdots & \cdots & 0 \\
  0 & 1/\sigma_2^2 & \ddots & \vdots & \\
  \vdots & \ddots & \ddots & \ddots & \vdots \\
  \vdots & \cdots & 1/\sigma_{n-1}^2 & 0 & \\
  0 & \cdots & \cdots & 0 & 1/\sigma_n^2
  \end{bmatrix}$$
- Normal equations: $(X'WX)b_w = X'WY$
- Solution: $b_w = (X'WX)^{-1}X'WY$
Weighted Least Squares

- Can be implemented in SAS using the `weight` option
- Must determine optimal weights
- Optimal weights \( \propto 1/\text{variance} \)
- Methods to determine weights
  - Find relationship between the absolute residual and another variable and use this as a model for the standard deviation
  - Instead of the absolute residual, use the squared residual and find function for the variance
  - Use grouped data or approximately grouped data to estimate the variance

Example Page 427

- Interested in the relationship between diastolic blood pressure and age
- Have measurements on 54 adult women
- Age range is 20 to 60 years old
- Issue:
  - Variability increases as the mean increases
  - Appears to be nice linear relationship
  - Don’t want to transform \( X \) or \( Y \) and lose this

SAS Commands I

```sas
data a1;
  infile 'U:\.www\datasets525\Ch11ta01.txt';
  input age diast;
  symbol1 v=circle i=sm70;
  proc gplot data=a1;
    plot diast*age/frame;
  proc reg data=a1;
    model diast=age;
    output out=a2 r=resid;
  proc gplot data=a2;
    plot resid*age;
run;
```
Output

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Squares</th>
<th>Square</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>1</td>
<td>2374.96833</td>
<td>2374.96833</td>
<td>35.79</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Error</td>
<td>52</td>
<td>3450.36501</td>
<td>66.35317</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Corrected Total</td>
<td>53</td>
<td>5825.33333</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Root MSE 8.14575 R-Square 0.4077 Dependent Mean 79.11111 Adj R-Sq 0.3963 Coeff Var 10.29659

Parameter Estimates

| Parameter | DF | Estimate | Error | t Value | Pr > |t| |
|-----------|----|----------|-------|---------|------|--|
| Intercept | 1  | 56.15693 | 3.99367| 14.06  | <.0001|
| age       | 1  | 0.58003  | 0.09695| 5.98   | <.0001|

SAS Commands II

data a2; set a2;
  absr=abs(resid); sqrr=resid*resid;
proc gplot data=a2;
  plot (resid absr sqrr)*age;
proc reg data=a2;
  model absr=age;
  output out=a3 p=shat;
data a3; set a3;
  wt=1/(shat*shat);
proc reg data=a3;
  model diast=age / clb;
  weight wt;
run;
Construction of Weights

- Will assume abs(res) is linearly related to age
- Fit least squares model to predict $SD_i$

```plaintext
proc reg data=a2;
model absr=age;
output out=a3 p=shat;
```

- Weight is $= 1/SD_i^2$

```plaintext
data a3; set a3;
wt=1/(shat*shat);
```

```plaintext
proc reg data=a3;
model diast=age / clb;weight wt;
```
Summary

- Not much change in the parameter estimates
- Slight reduction in the parameters' standard errors
- Be wary:
  - $R^2$ does not have usual meaning
  - Interpretation of residual plots
  - Construction of confidence and prediction intervals
- Since weights based on residuals, can take iterative approach and re-estimate weights based on new residuals and repeat.

Iterative Approach

- Usually converges quite quickly
- For this example:

<table>
<thead>
<tr>
<th>Iteration</th>
<th>$b_0$</th>
<th>SE($b_0$)</th>
<th>$b_1$</th>
<th>SE($b_1$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>55.56577</td>
<td>2.52092</td>
<td>0.59634</td>
<td>0.07924</td>
</tr>
<tr>
<td>2</td>
<td>55.56264</td>
<td>2.51851</td>
<td>0.59643</td>
<td>0.07922</td>
</tr>
<tr>
<td>3</td>
<td>55.56261</td>
<td>2.51849</td>
<td>0.59643</td>
<td>0.07922</td>
</tr>
<tr>
<td>4</td>
<td>55.56261</td>
<td>2.51849</td>
<td>0.59643</td>
<td>0.07922</td>
</tr>
</tbody>
</table>

- Usually changes within level of accuracy so run only once

Mixed model approach(?)

- We’re assuming the variance/covariance matrix is a diagonal matrix whose values along the main diagonal (the variances) are either a
  - Linear function of age
  - Quadratic function of age
- This relationship along with the estimation of parameters can be done simultaneously using the lin(q) covariance structure
- Need to create appropriate diagonal matrices and specify reasonable starting values...sample code provided
### Output - Linear Relationship

<table>
<thead>
<tr>
<th>Cov Parm Subject</th>
<th>Estimate</th>
</tr>
</thead>
<tbody>
<tr>
<td>LIN(1) Intercept</td>
<td>-51.039</td>
</tr>
<tr>
<td>LIN(2) Intercept</td>
<td>2.870</td>
</tr>
</tbody>
</table>

**Fit Statistics**
- $-2$ Res Log Likelihood: 365.0
- AIC (smaller is better): 369.0
- AICC (smaller is better): 369.3
- BIC (smaller is better): 372.9

**Solution for Fixed Effects**

| Effect   | Estimate | Standard Error | DF  | tValue | Pr > |t|     |
|----------|----------|----------------|-----|--------|-------|-------|
| Intercept| 55.3831  | 2.5720         | 17.8| 21.53  | <.0001|
| age      | 0.5996   | 0.07855        | 39.2| 7.63   | <.0001|

### Output - Quadratic Relationship

<table>
<thead>
<tr>
<th>Cov Parm Subject</th>
<th>Estimate</th>
</tr>
</thead>
<tbody>
<tr>
<td>LIN(1) Intercept</td>
<td>9.4647</td>
</tr>
<tr>
<td>LIN(2) Intercept</td>
<td>-1.2853</td>
</tr>
<tr>
<td>LIN(3) Intercept</td>
<td>0.06331</td>
</tr>
</tbody>
</table>

**Fit Statistics**
- $-2$ Res Log Likelihood: 364.4
- AIC (smaller is better): 370.4
- AICC (smaller is better): 370.9
- BIC (smaller is better): 376.2

**Solution for Fixed Effects**

| Effect   | Estimate | Standard Error | DF  | tValue | Pr > |t|     |
|----------|----------|----------------|-----|--------|-------|-------|
| Intercept| 55.6087  | 2.8634         | 14.8| 19.42  | <.0001|
| age      | 0.5954   | 0.08666        | 25.4| 6.87   | <.0001|

### Background Reading

- KNNL Section 11.1
- knnl427.sas
- KNNL Sections 11.2-11.6