Assumptions of Ordinary Least Squares Regression
Assumptions of OLS regression

1. Model is *linear in parameters*
2. The data are a *random sample* of the population
   1. The errors are *statistically independent* from one another
3. The expected value of the errors is always zero
4. The independent variables are not too strongly *collinear*
5. The independent variables are measured *precisely*
6. The residuals have *constant variance*
7. The errors are normally distributed

• If assumptions 1-5 are satisfied, then 
  OLS estimator is *unbiased*
• If assumption 6 is also satisfied, then 
  OLS estimator has *minimum variance* of all unbiased estimators.
• If assumption 7 is also satisfied, then 
  we can do hypothesis testing using *t* and *F* tests
• How can we test these assumptions?
• If assumptions are violated,
  – what does this do to our conclusions?
  – how do we fix the problem?
1. Model not linear in parameters

- **Problem**: Can’t fit the model!
- **Diagnosis**: Look at the model
- **Solutions**:
  1. Re-frame the model
  2. Use *nonlinear least squares (NLS)*
     *regression*
2. Errors not independent

- **Problem:** parameter estimates are biased
- **Diagnosis (1):** look for correlation between residuals and another variable (not in the model)
  - I.e., residuals are dominated by another variable, Z, which is not random with respect to the other independent variables
- **Solution (1):** add the variable to the model
- **Diagnosis (2):** look at *autocorrelation function* of residuals to find patterns in
  - time
  - Space
  - I.e., observations that are nearby in time or space have residuals that are more similar than average
- **Solution (2):** fit model using *generalized least squares (GLS)*
3. Average error not everywhere zero ("nonlinearity")

- **Problem:** indicates that *model is wrong*
- **Diagnosis:**
  - Look for curvature in plot of observed vs. predicted Y
3. Average error not everywhere zero (“nonlinearity”)

- **Problem:** indicates that *model is wrong*
- **Diagnosis:**
  - Look for curvature in plot of observed vs. predicted Y
  - Look for curvature in plot of residuals vs. predicted Y
3. Average error not everywhere zero (“nonlinearity”)

- **Problem:** indicates that *model is wrong*
- **Diagnosis:**
  - Look for curvature in plot of observed vs. predicted Y
  - Look for curvature in plot of residuals vs. predicted Y
  - Look for curvature in *partial-residual plots* (also *component+residual plots* [CR plots])
    - Most software doesn’t provide these, so instead can take a quick look at plots of Y vs. each of the independent variables
A simple look at nonlinearity: bivariate plots
A better way to look at nonlinearity: partial residual plots

- The previous plots are fitting a *different model*:
  - for phosphorus, we are looking at residuals from the model
    \[
    C_i = a_0 + a_1 P_i + e_i
    \]
  - We want to look at residuals from
    \[
    C_i = b_0 + b_1 P_i + b_2 N_i P_i + e_i
    \]
- Construct *Partial Residuals*:

  Phosphorus \[ PR_i = b_1 P_i + e_i \]

  NP \[ PR_i = b_2 N_i P_i + e_i \]
A better way to look at nonlinearity: partial residual plots

\[ PR_i = b_1 P_i + e_i \]

\[ PR_i = b_2 N_i P_i + e_i \]
Average error not everywhere zero ("nonlinearity")

- **Solutions:**
  - If pattern is monotonic*, try transforming *independent* variable
    - Downward curving: use powers less than one
      - E.g. Square root, log, inverse
    - Upward curving: use powers greater than one
      - E.g. square
  
  * Monotonic: always increasing or always decreasing

- If not, try adding *additional terms* in the independent variable (e.g., quadratic)
4. Independent variables are collinear

- **Problem:** parameter estimates are imprecise
- **Diagnosis:**
  - Look for correlations among independent variables
  - In regression output, none of the individual terms are significant, even though the model as a whole is
- **Solutions:**
  - Live with it
  - Remove statistically redundant variables
| Parameter | Est value | St dev  | t student | Prob(>|t|) |
|-----------|-----------|---------|-----------|-----------|
| b0        | 16.37383  | 41.50584| 0.394495  | 0.696315  |
| b1        | 1.986335  | 1.02642 | 1.935206  | 0.063504  |
| b2        | -1.22964  | 2.131899| -0.57678  | 0.568867  |
| Residual St dev | 31.6315 |         |           |           |

\[ y = b_0 + b_1.x_1 + b_2.x_2 \]

\[ \beta_0 = 0; \beta_1 = 1; \beta_2 = 0.5; \rho_{XZ} = 0.95 \]
5. Independent variables not precise ("measurement error")

- **Problem:** parameter estimates are biased
- **Diagnosis:** know how your data were collected!

- **Solution:** very hard
  - State space models
  - Restricted maximum likelihood (REML)
  - Use simulations to estimate bias
  - Consult a professional!
6. Errors have non-constant variance ("heteroskedasticity")

- **Problem:**
  - Parameter estimates are *unbiased*
  - P-values are *unreliable*

- **Diagnosis:** plot residuals against fitted values
Errors have non-constant variance (‘‘heteroskedasticity’’)

• **Problem:**
  – Parameter estimates are *unbiased*
  – P-values are *unreliable*

• **Diagnosis:** plot studentized residuals against fitted values

• **Solutions:**
  – Transform the *dependent* variable
    • If residual variance increases with predicted value, try transforming with power less than one
Try square root transform
Errors have non-constant variance ("heteroskedasticity")

• **Problem:**
  – Parameter estimates are *unbiased*
  – P-values are *unreliable*

• **Diagnosis:** plot studentized residuals against fitted values

• **Solutions:**
  – Transform the dependent variable
    • May create nonlinearity in the model
  – Fit a *generalized linear model (GLM)*
    • For some distributions, the variance changes with the mean in predictable ways
  – Fit a *generalized least squares model (GLS)*
    • Specifies how variance depends on one or more variables
  – Fit a *weighted least squares regression (WLS)*
    • Also good when data points have differing amount of precision
7. Errors not normally distributed

- **Problem:**
  - Parameter estimates are *unbiased*
  - P-values are *unreliable*
  - Regression fits the mean; with skewed residuals the mean is not a good measure of central tendency

- **Diagnosis:** examine QQ plot of *residuals*
Studentized Resid Chlorophyll-a Distributions

Normal Quantile Plot
Errors not normally distributed

• **Problem:**
  – Parameter estimates are *unbiased*
  – P-values are *unreliable*
  – Regression fits the mean; with skewed residuals the mean is not a good measure of central tendency

• **Diagnosis:** examine QQ plot of *Studentized residuals*
  – Corrects for bias in estimates of residual variance

• **Solutions:**
  – Transform the *dependent* variable
    • May create nonlinearity in the model
Try transforming the response variable

![Normal Quantile Plot](image-url)
But we’ve introduced nonlinearity…

Actual by Predicted Plot (Chlorophyll)

Chlorophyll-a Actual

Chlorophyll-a Predicted P<.0001 RSq=0.88 RMSE=18.074

Actual by Predicted Plot (sqrt[Chlorophyll])

sqrt(Chlorophyll-a) Actual

sqrt(Chlorophyll-a) Predicted P<.0001 RSq=0.85 RMSE=1.4361
Errors not normally distributed

- **Problem:**
  - Parameter estimates are *unbiased*
  - P-values are *unreliable*
  - Regression fits the mean; with skewed residuals the mean is not a good measure of central tendency

- **Diagnosis:** examine QQ plot of *Studentized residuals*
  - Corrects for bias in estimates of residual variance

- **Solutions:**
  - Transform the dependent variable
    - May create nonlinearity in the model
  - Fit a *generalized linear model (GLM)*
    - Allows us to assume the residuals follow a different distribution (binomial, gamma, etc.)
## Summary of OLS assumptions

<table>
<thead>
<tr>
<th>Violation</th>
<th>Problem</th>
<th>Solution</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nonlinear in parameters</td>
<td>Can’t fit model</td>
<td>NLS</td>
</tr>
<tr>
<td>Non-normal errors</td>
<td>Bad P-values</td>
<td>Transform Y; GLM</td>
</tr>
<tr>
<td>Heteroskedasticity</td>
<td>Bad P-values</td>
<td>Transform Y; GLM</td>
</tr>
<tr>
<td>Nonlinearity</td>
<td>Wrong model</td>
<td>Transform X; add terms</td>
</tr>
<tr>
<td>Nonindependence</td>
<td>Biased parameter estimates</td>
<td>GLS</td>
</tr>
<tr>
<td>Measurement error</td>
<td>Biased parameter estimates</td>
<td>Hard!!!</td>
</tr>
<tr>
<td>Collinearity</td>
<td>Individual P-values inflated</td>
<td>Remove X terms?</td>
</tr>
</tbody>
</table>
Fixing assumptions via data transformations is an iterative process

• After each modification, fit the new model and look at all the assumptions again
What can we do about chlorophyll regression?

- Square root transform helps a little with non-normality and a lot with heteroskedasticity
- But it creates nonlinearity
A new model … it’s linear …

**Fit Y by X Group**

**Bivariate Fit of \( \sqrt{\text{Chlorophyll-a}} \) By \( \sqrt{\text{Phosphorus}} \)**

**Bivariate Fit of \( \sqrt{\text{Chlorophyll-a}} \) By \( \sqrt{\text{NP}} \)**
… it’s normal (sort of) and homoskedastic …
... and it fits well!

**Response sqrt(Chlorophyll-a)**

### Whole Model

<table>
<thead>
<tr>
<th>Summary of Fit</th>
</tr>
</thead>
<tbody>
<tr>
<td>RSquare</td>
</tr>
<tr>
<td>RSquare Adj</td>
</tr>
<tr>
<td>Root Mean Square Error</td>
</tr>
<tr>
<td>Mean of Response</td>
</tr>
<tr>
<td>Observations (or Sum Wgts)</td>
</tr>
</tbody>
</table>

### Analysis of Variance

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Sum of Squares</th>
<th>Mean Square</th>
<th>F Ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>2</td>
<td>275.06699</td>
<td>137.533</td>
<td>95.7674</td>
</tr>
<tr>
<td>Error</td>
<td>22</td>
<td>31.59463</td>
<td>1.436</td>
<td>Prob &gt; F</td>
</tr>
<tr>
<td>C. Total</td>
<td>24</td>
<td>306.66163</td>
<td></td>
<td>&lt;.0001</td>
</tr>
</tbody>
</table>

### Parameter Estimates

| Term       | Estimate | Std Error | t Ratio | Prob>|t| |
|------------|----------|-----------|---------|------|
| Intercept  | -0.901414| 0.61584   | -1.46   | 0.1574|
| sqrt(Phosphorus) | 0.214075| 0.095471  | 2.24    | 0.0353|
| sqrt(NP)   | 0.1513313| 0.037419  | 4.04    | 0.0005|

### Effect Tests

<table>
<thead>
<tr>
<th>Source</th>
<th>Nparm</th>
<th>DF</th>
<th>Sum of Squares</th>
<th>F Ratio</th>
<th>Prob &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>sqrt(Phosphorus)</td>
<td>1</td>
<td>1</td>
<td>7.220755</td>
<td>5.0280</td>
<td>0.0353</td>
</tr>
<tr>
<td>sqrt(NP)</td>
<td>1</td>
<td>1</td>
<td>23.488665</td>
<td>16.3556</td>
<td>0.0005</td>
</tr>
</tbody>
</table>

RSq=0.90 RMSE=1.1984