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

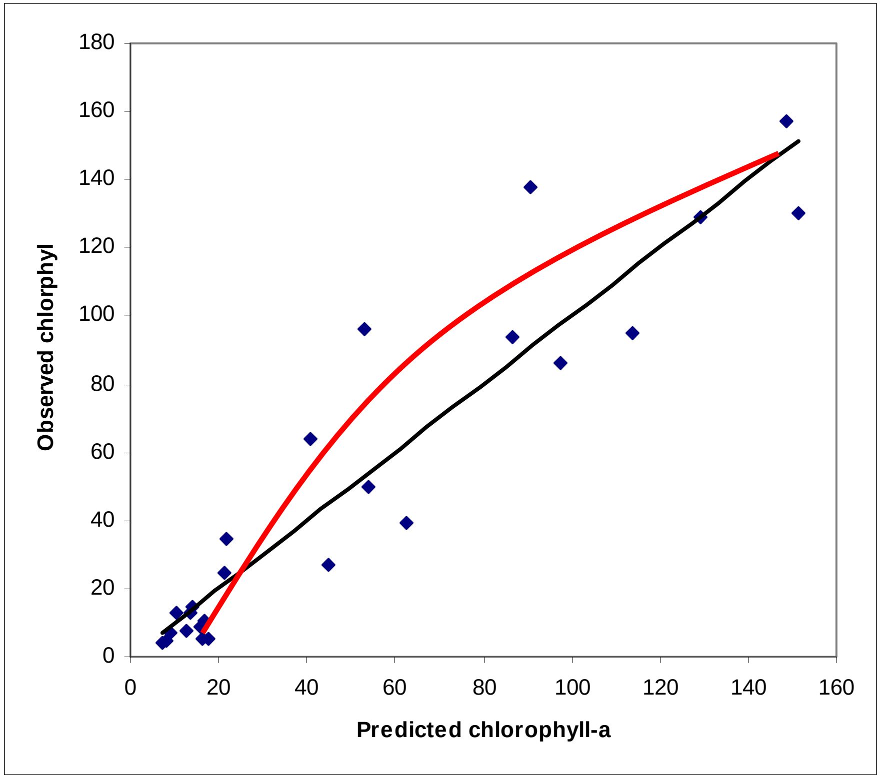
- **Problem:** Can’t fit the model!
- **Diagnosis:** Look at the model
- **Solutions:**
  1. Re-frame the model
  2. Use *nonlinear least squares (NLS) regression*

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)*

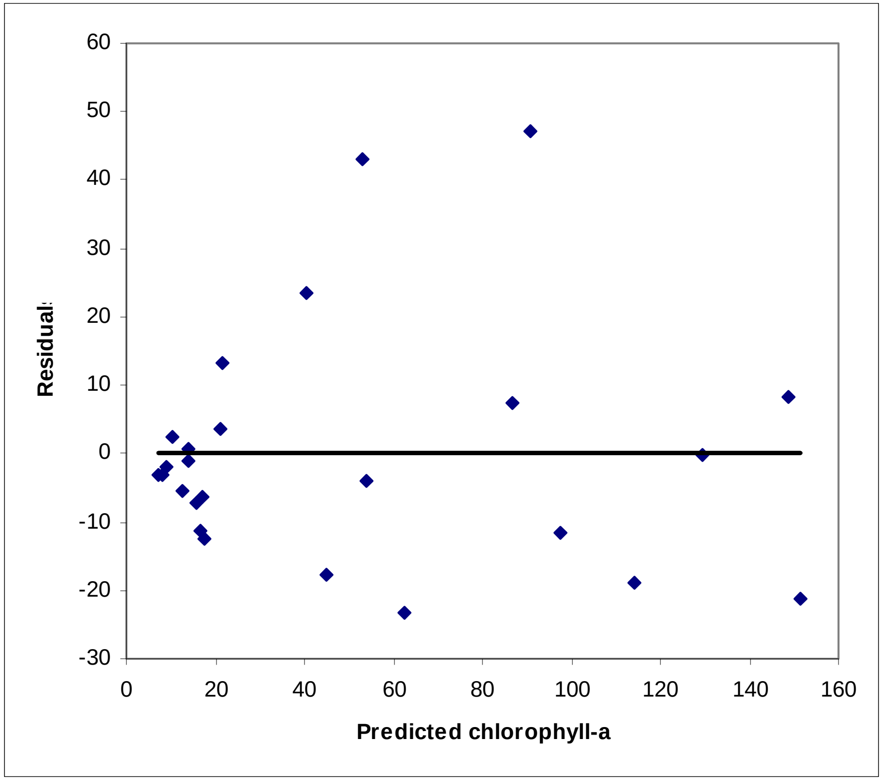
3. Average error not everywhere zero ("nonlinearity")

- **Problem:** indicates that *model is wrong*
- **Diagnosis:**
  - Look for curvature in plot of observed vs. predicted Y



3. Average error not everywhere zero (“nonlinearity”)

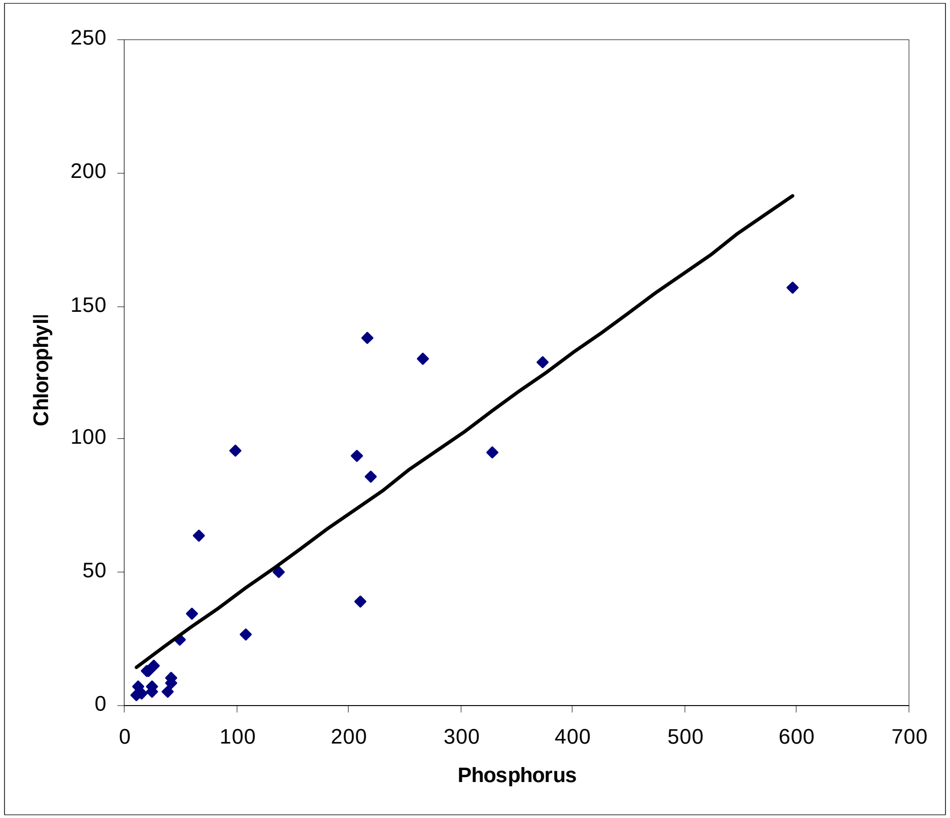
- **Problem:** indicates that *model is wrong*
- **Diagnosis:**
  - Look for curvature in plot of observed vs. predicted Y
  - Look for curvature in plot of residuals vs. predicted Y

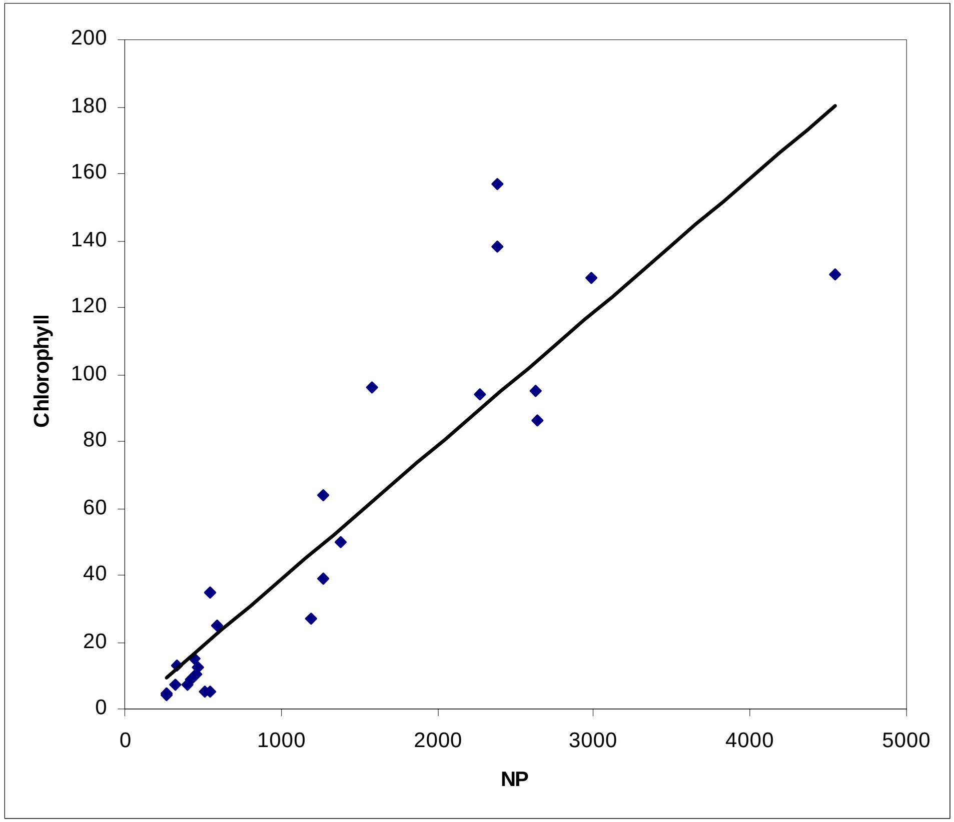


3. Average error not everywhere zero ("nonlinearity")

- **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 a 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:*

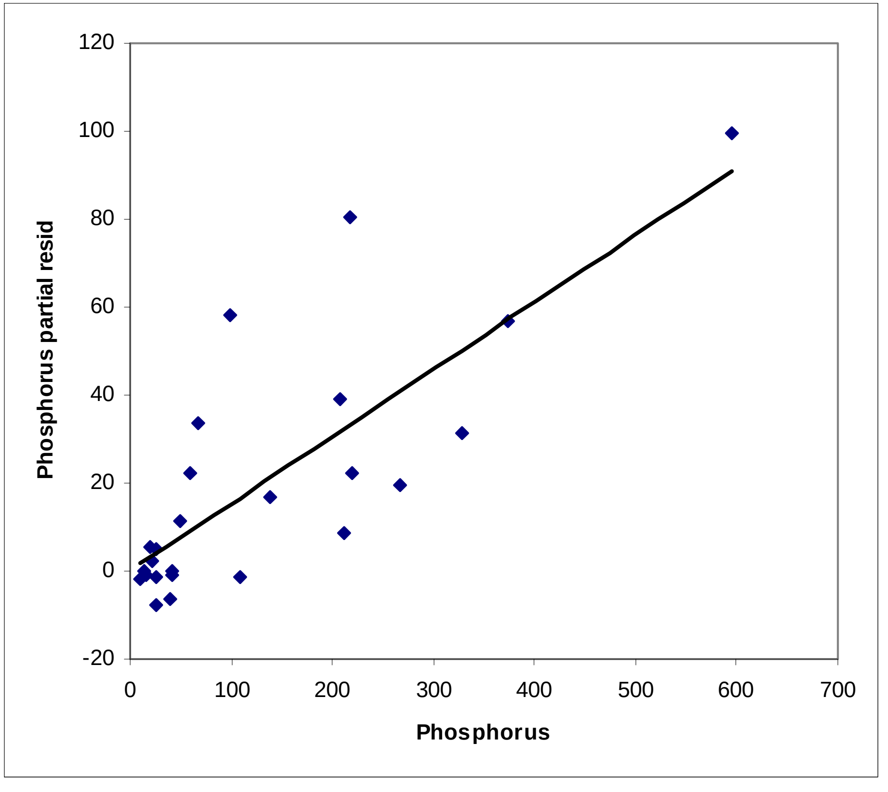
Phosphorus

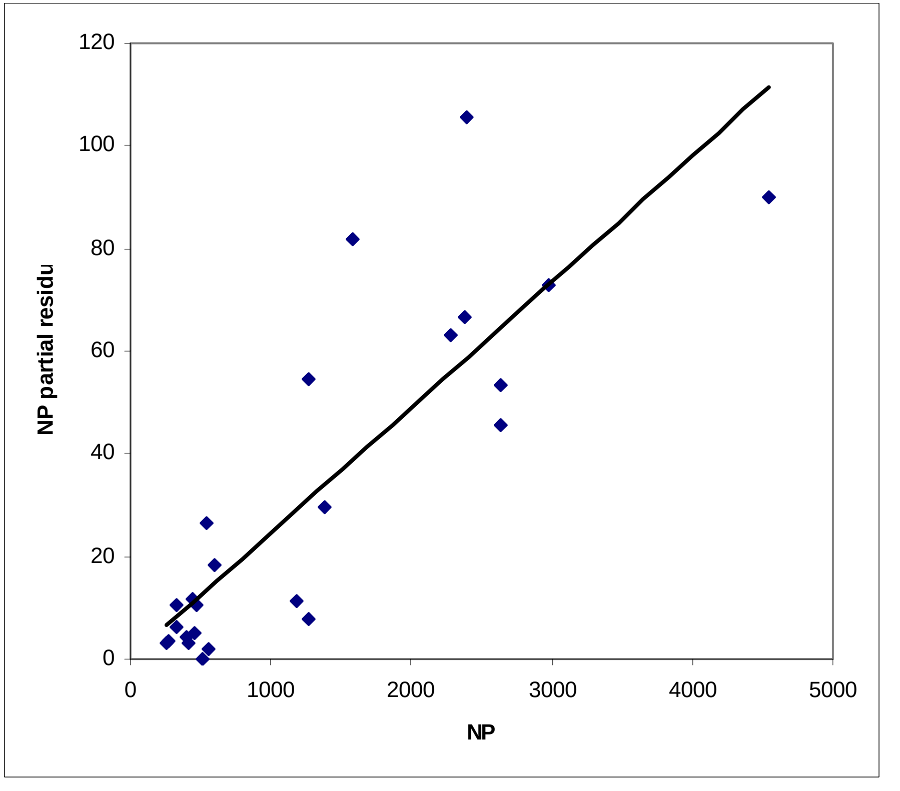
$$PR_i = b_1 P_i + e_i$$

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 = b0 + b1.x1 + b2.x2
R2	0.534192			
R2(adj)	0.499688			
F	15.48191			
Prob(>F)	3.32E-05			

$$\beta_0 = 0; \beta_1 = 1; \beta_2 = 0.5; \rho_{XZ} = 0.95$$

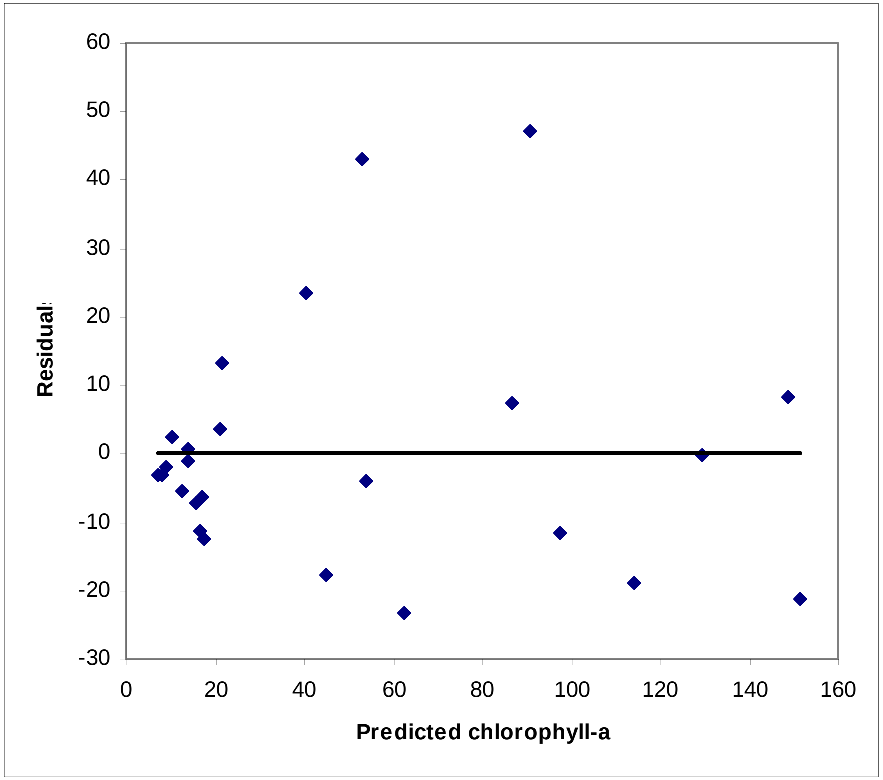
5. Independent variables not precise (“measurement error”)

- **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



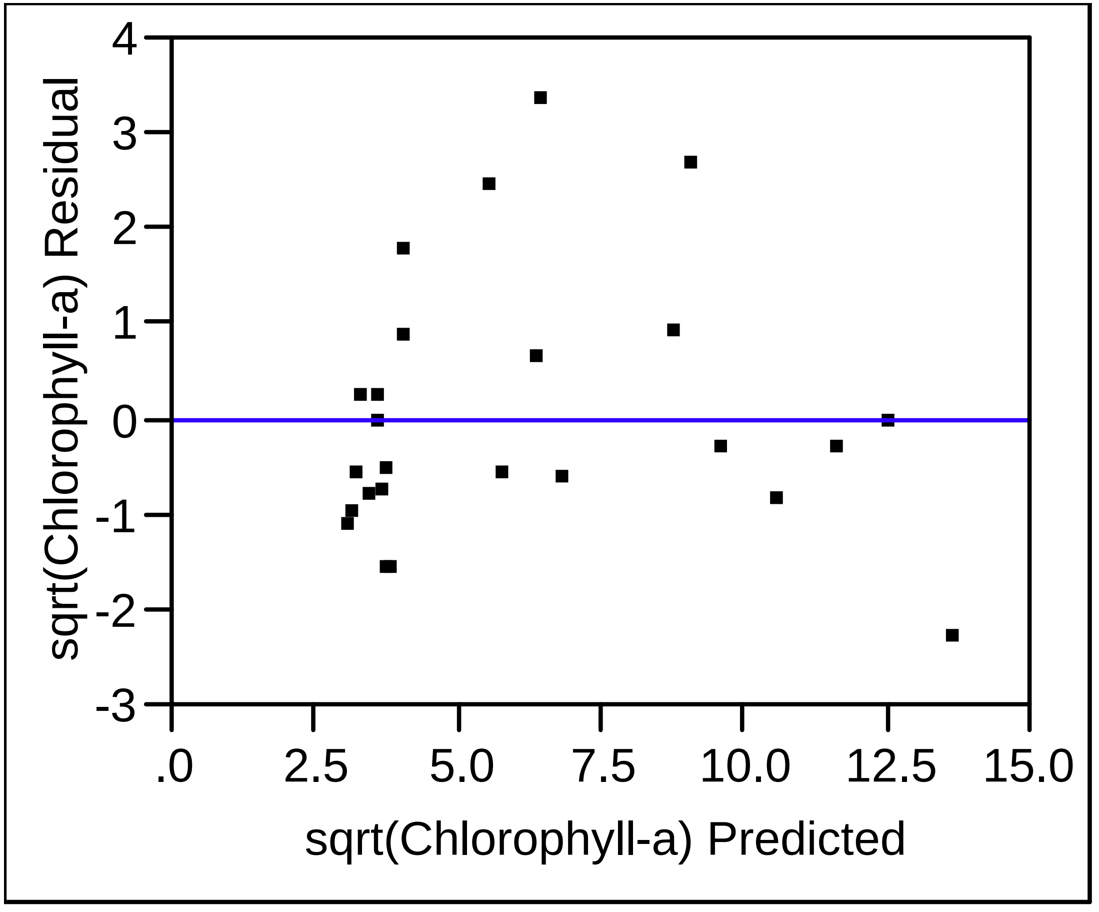
60
50
40
30
20
10
0
-10
-20
-30
Residuals
0
20
40
60
80
100
120
140
160
Predicted chlorophyll-a

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

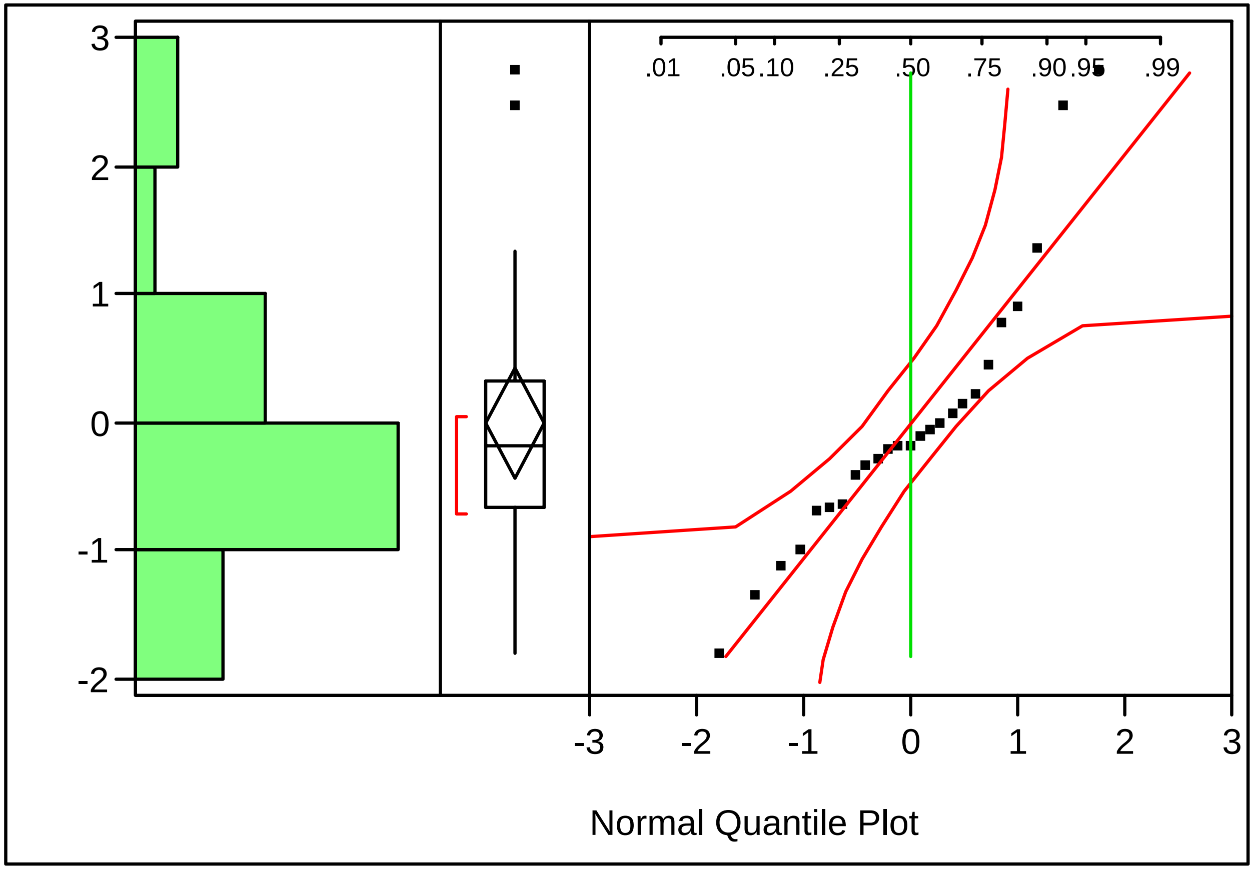
- **Solutions:**
  - Transform the dependent variable
    - May create nonlinearity in the model
  - Fit a *generalized linear model (GLM)*
    - For some distributions, the variance changes with the mean in predictable ways
  - Fit a *generalized least squares* model (GLS)
    - Specifies how variance 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*

Distributions

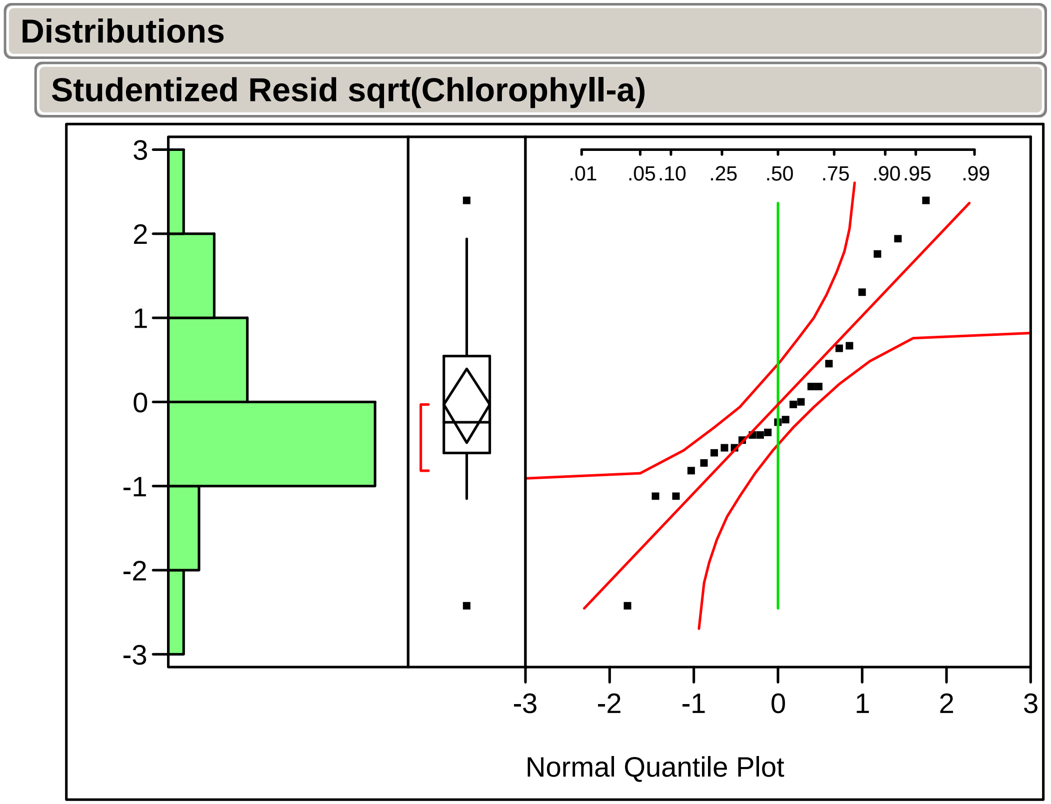
Studentized Resid Chlorophyll-a



Errors not normally distributed

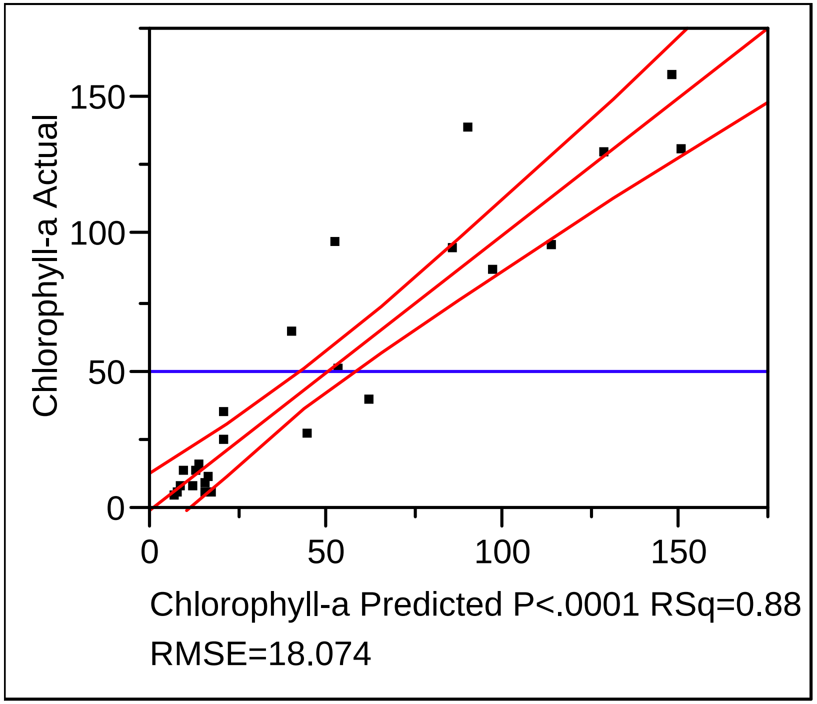
- **Problem:**
  - Parameter estimates are *unbiased*
  - P-values are *unreliable*
  - Regression fits the mean; with skewed residuals the mean is not a good measure of central tendency
- **Diagnosis:** examine QQ plot of *Studentized residuals*
  - Corrects for bias in estimates of residual variance
- **Solutions:**
  - Transform the *dependent* variable
    - May create nonlinearity in the model

Try transforming the response variable

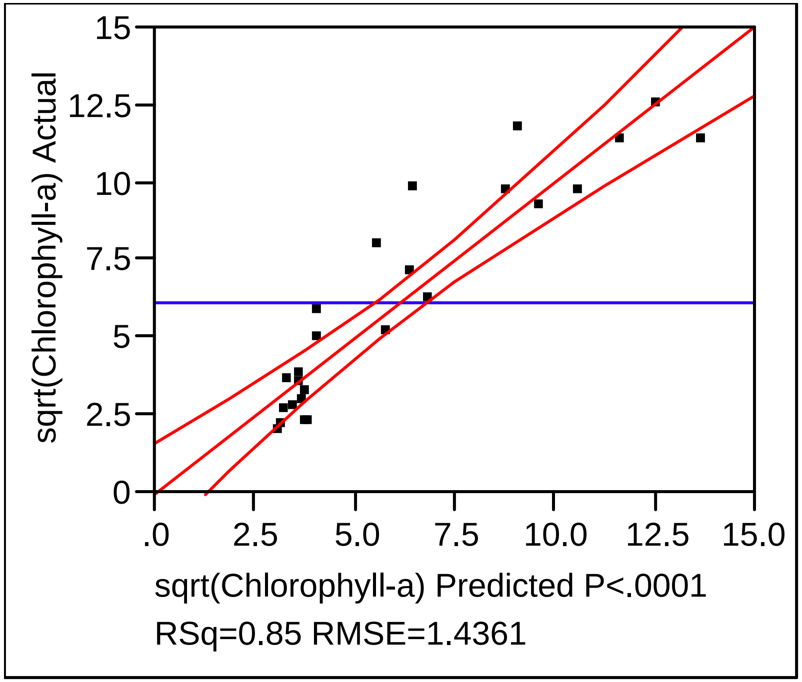


But we’ve introduced nonlinearity…

Actual by Predicted Plot (Chlorophyll)



Actual by Predicted Plot (sqrt[Chlorophyll])



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

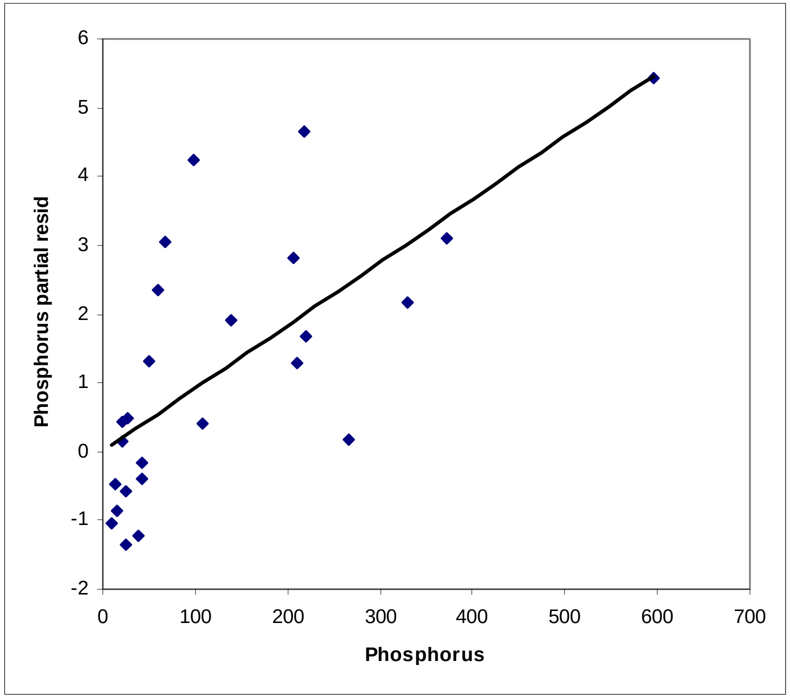
Violation	Problem	Solution
Nonlinear in parameters	Can't fit model	NLS
Non-normal errors	Bad P-values	Transform Y; GLM
Heteroskedasticity	Bad P-values	Transform Y; GLM
Nonlinearity	Wrong model	Transform X; add terms
Nonindependence	Biased parameter estimates	GLS
Measurement error	Biased parameter estimates	Hard!!!
Collinearity	Individual P-values inflated	Remove X terms?

Fixing assumptions via data transformations is an iterative process

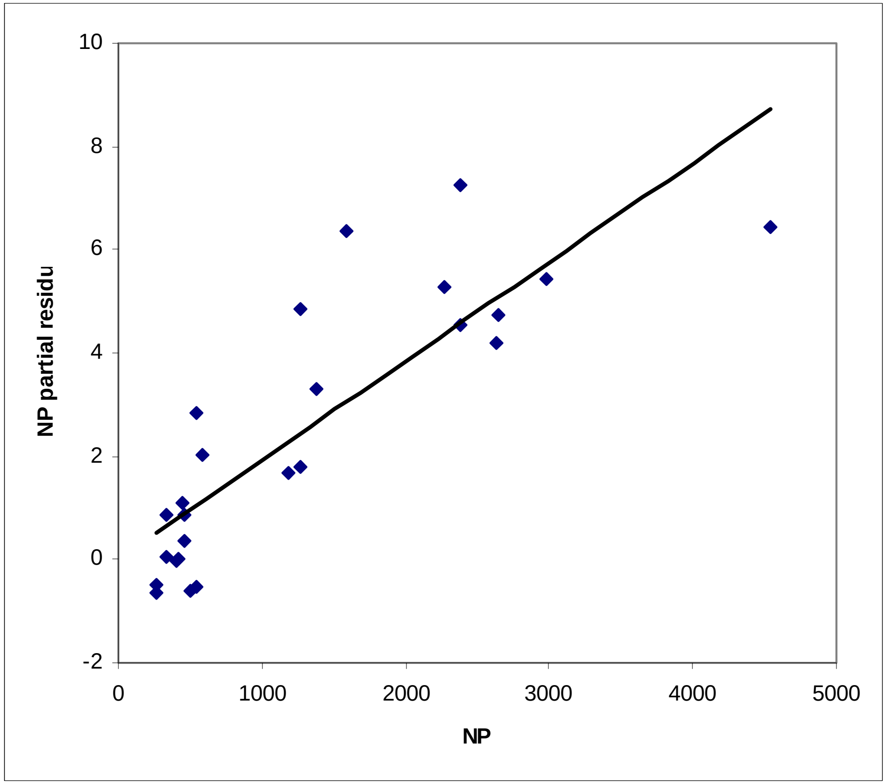
- After each modification, fit the new model and look at all the assumptions again

What can we do about chlorophyll regression?

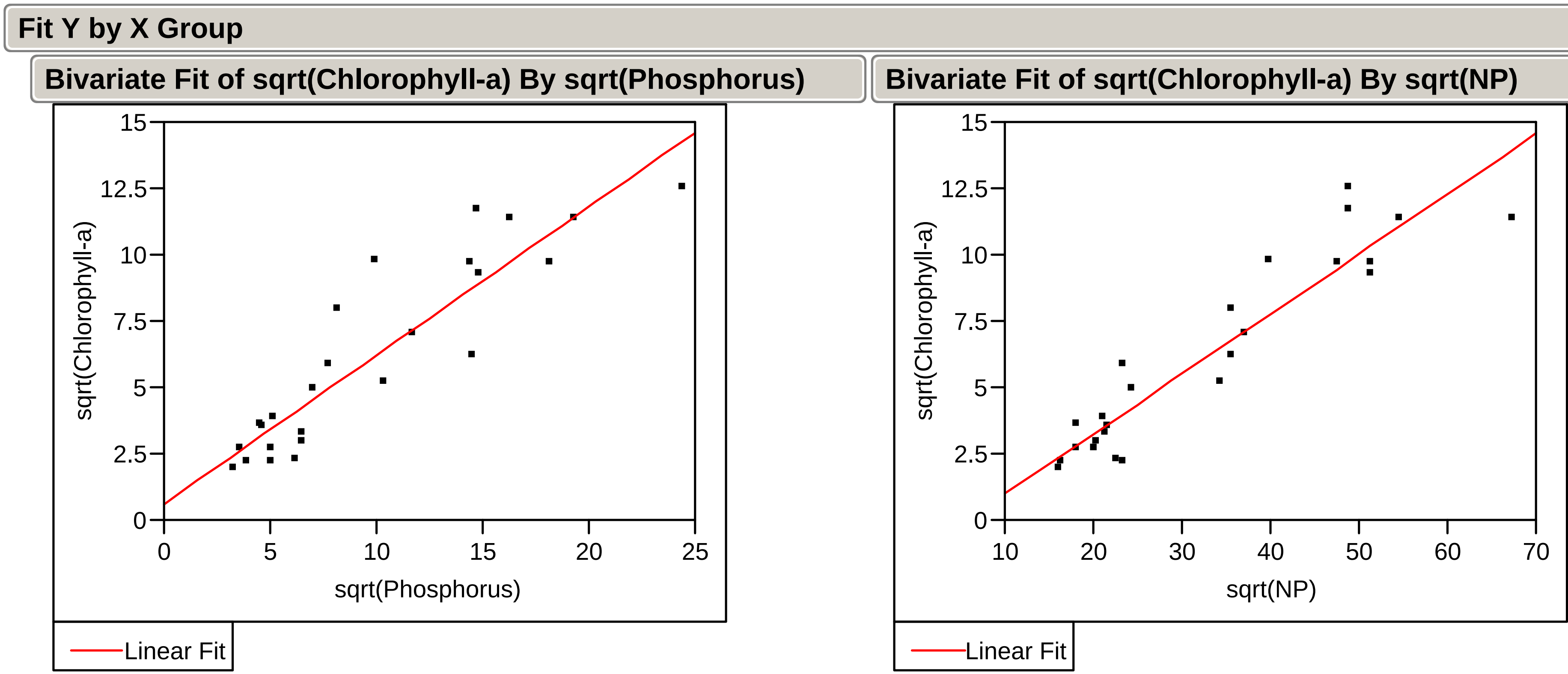
- Square root transform helps a little with non-normality and a lot with heteroskedasticity



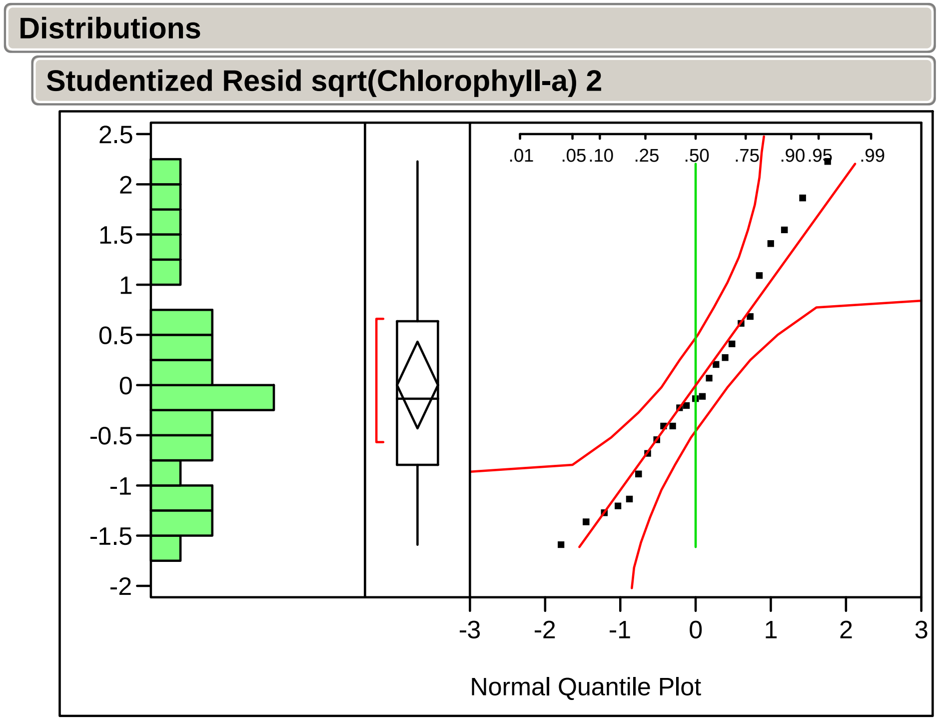
- But it creates nonlinearity

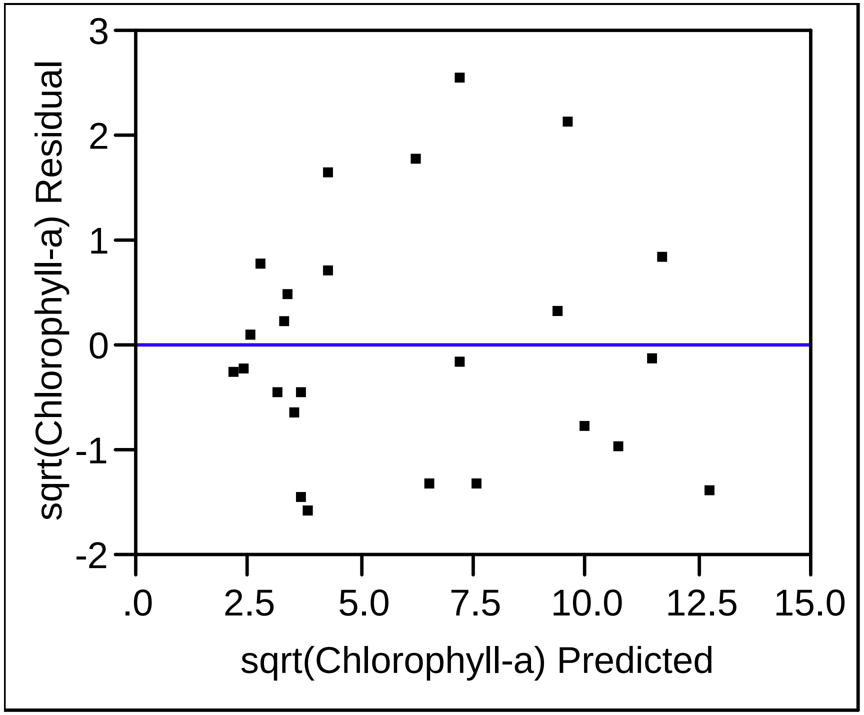


A new model … it's linear …



… it’s normal (sort of) and homoskedastic …





… and it fits well!

Response sqrt(Chlorophyll-a)

Whole Model

Summary of Fit

RSquare	0.896972
RSquare Adj	0.887606
Root Mean Square Error	1.198382
Mean of Response	6.167458
Observations (or Sum Wgts)	25

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	2	275.06699	137.533	95.7674
Error	22	31.59463	1.436	Prob > F
C. Total	24	306.66163		<.0001

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

Source	Nparm	DF	Sum of Squares	F Ratio	Prob > F
sqrt(Phosphorus)	1	1	7.220755	5.0280	0.0353
sqrt(NP)	1	1	23.488665	16.3556	0.0005