FISH 604
Module 4:
Statistical estimation I

Instructor: Franz Mueter

Lena Point, Rm 315

796-5448

fmueter@alaska.edu

Review of Module 3: Exploratory data analysis

- Visualizing data
- Assessing distributions
- Outlier detection
- Standardization
- Transformations
- Correlations

→ Complete 'transformations' script

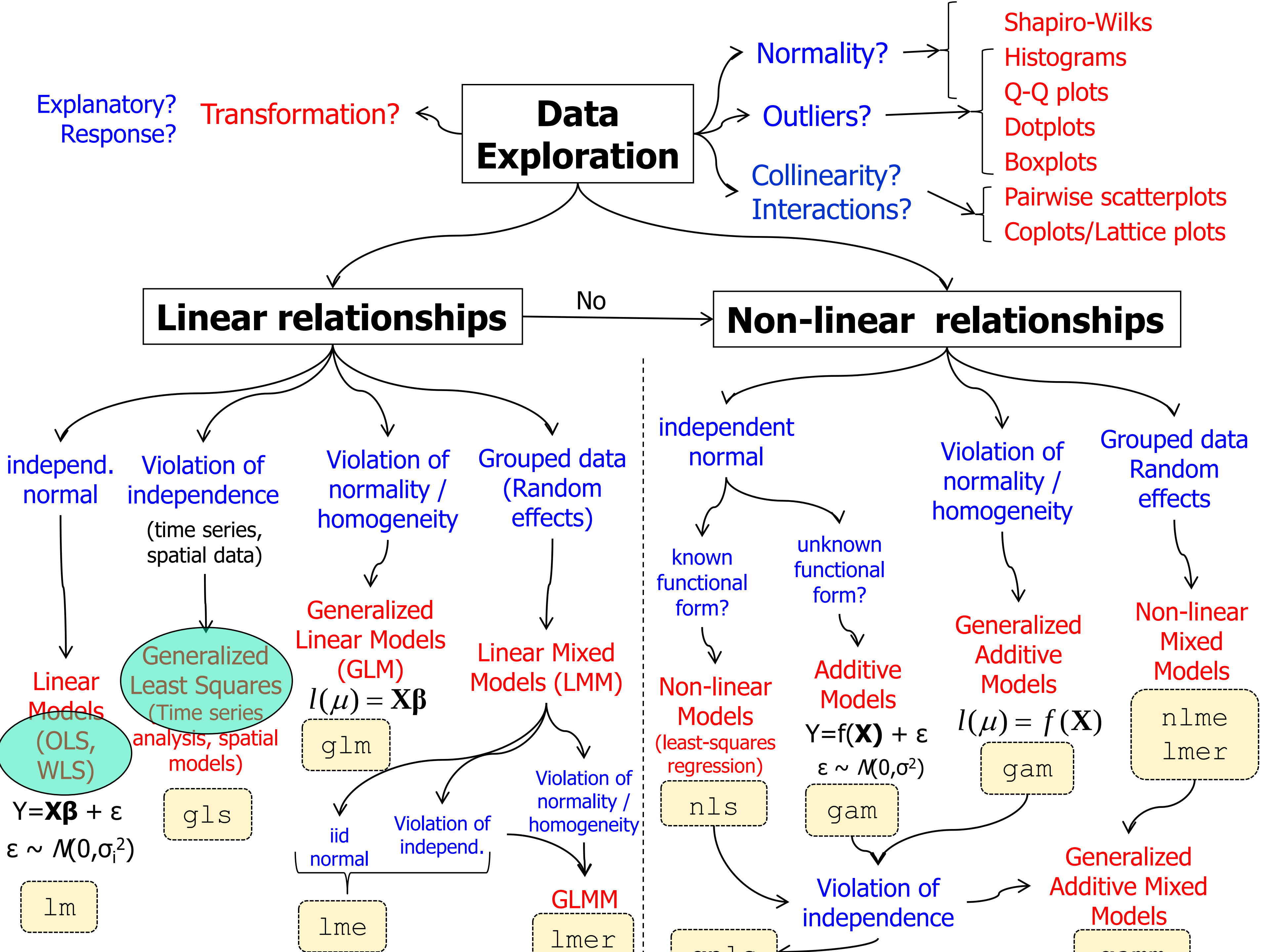
Preview - Student Project

MSL FISH 604

- How to format your data?
  - Observations / samples (rows) x variables (columns)
  - Include any potentially relevant information
    - Aspects of sampling design (incl. spatial / temporal info)
    - Experimental design aspects (Treatment, Replicate)
    - Environmental covariates (measured, derived)

Case.ID / Sample	**Chl. a**	*Station*	*Lat*	*Long*	*Location*	*Stratum*	*Year*	*Month*	*Season*	*DOY*	…
1	12.6	FB1	58.25	-134.9	FunterB	surface	2018	3	Spring	86	…
2	9.2	FB1	58.19	-134.9	FunterB	surface	2018	3	Spring	86	
3	23.2	FB2			FunterB						

…	*Case.ID / Sample*	*Depth*	*Temperature*	*Salinity*	*PAR*	*Weather*	*Fishery*	*Precip (lagged)*
…	1	7	2.3	26.3	12.6	rainy	open	23
	2							
	3							



Data Exploration
Explanatory? Response?
Transformation?
Normality?
Outliers?
Collinearity? Interactions?
Shapiro-Wilks
Histograms
Q-Q plots
Dotplots
Boxplots
Pairwise scatterplots
Coplots/Lattice plots
Linear relationships
No
Non-linear relationships
independ. normal
Linear Models (OLS, WLS)
Y=Xβ + ε
ε ~ N(0,σi2)
lm
Violation of independence
(time series, spatial data)
Generalized Least Squares (Time series analysis, spatial models)
gls
Violation of normality / homogeneity
Generalized Linear Models (GLM)
l(μ) = Xβ
glm
Grouped data (Random effects)
Linear Mixed Models (LMM)
iid normal
Violation of independ.
lme
Violation of normality / homogeneity
GLMM
lmer
independent normal
known functional form?
Non-linear Models (least-squares regression)
nls
unknown functional form?
Additive Models
Y=f(X) + ε
ε ~ N(0,σ2)
gam
Violation of normality / homogeneity
Generalized Additive Models
l(μ) = f(X)
gam
Grouped data Random effects
Non-linear Mixed Models
nlme
lmer
Violation of independence
Generalized Additive Mixed Models

MSL FISH 604

Estimation methods

Today

- Ordinary least-squares (OLS)
- Weighted least-squares (WLS)
- Generalized least-squares (GLS)

- Maximum likelihood estimation (MLE)
- (Bayesian estimation)
- (Bootstrap estimation)

MSL FISH 604

Objectives & Outcomes

- You should understand …
  - … the principle of least-squares regression
  - … the difference between ordinary least-square regression, weighted LS regression, and generalized LS regression
  - … potential pitfalls of non-linear LS regression
- You should be able to…
  - …choose appropriate weights and do a weighted LS regression
  - … fit a generalized LS regression model

MSL FISH 604

Statistical estimation

- Methods for estimating model parameters
  - Ordinary least-squares (OLS)
  - Weighted least-squares (WLS)
  - Generalized least-squares (GLS)
  - Maximum likelihood estimation (MLE)
  - (Bayesian estimation)

Least-squares estimation

- Method of fitting a curve (or surface) to data points so as to minimize the sum of the squares of the distances of the points from the curve (or surface)
- <u>Some notation</u>

  Data: $y_i, \mathbf{x}_i \quad \text{where: } \mathbf{x_i} = \{x_1, x_2, \ldots, x_p\}$

  Parameters: $\alpha, \boldsymbol{\beta} \quad \text{where: } \boldsymbol{\beta} = \{\beta_1, \beta_2, \ldots, \beta_p\}$

  Predicted values: $\hat{\alpha}, \hat{\boldsymbol{\beta}}, \hat{\beta}_i, \hat{y}_i$

MSL FISH 604

Ordinary least-squares (OLS)

- Assumptions
  - Independent, identically distributed errors (iid)!!
  - Normality NOT required for fitting!
  - Normality assumed for drawing inferences (testing coefficients & constructing confidence intervals)

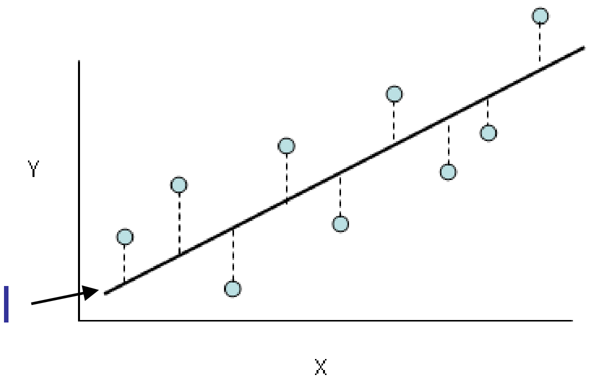
Ordinary least-squares (OLS)

- Minimize squared differences (residuals) between observed and expected values:

$$RSS = \sum_{i=1}^{n} (y_i - f(x_i))^2.$$

Data

Model



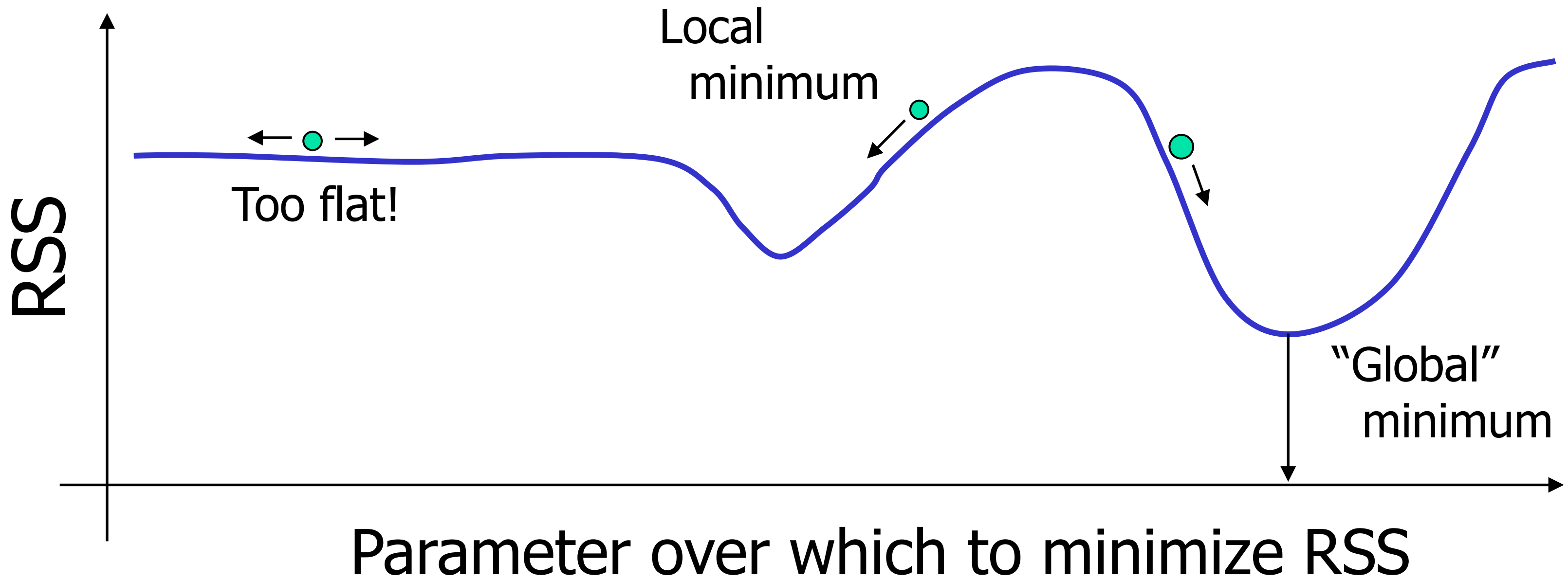
- If the model $f(.)$ is linear in the parameters, e.g.:

$$f(x) = a + b\,x + c\,x^2$$

  the problem simplifies considerably (to a system of linear equations) and can be solved analytically
- If $f(.)$ is not linear in the parameters, an algorithm for general optimization is used, such as Newton's method or gradient descent (iterative methods)

Ordinary least-squares (OLS)

- Non-linear minimization
  - Beware of local minima!
  - Use several different starting values!



Least-squares fitting in R

MSL FISH 604

R functions for OLS (& WLS) estimation:

Function	Purpose
`lsfit`	Find least-squares fit for linear models
`lm`	**Fits linear models via least-squares**
`nls`	Fits <u>non-linear</u> models via least-squares
`nlm`	Function minimization via Newton-type algorithm (If *f* is RSS = least-squares)
`optim` `optimize`	General-purpose optimization (find min or max of <u>any</u> function using one of several algorithms).

MSL FISH 604

Weighted least-squares (WLS)

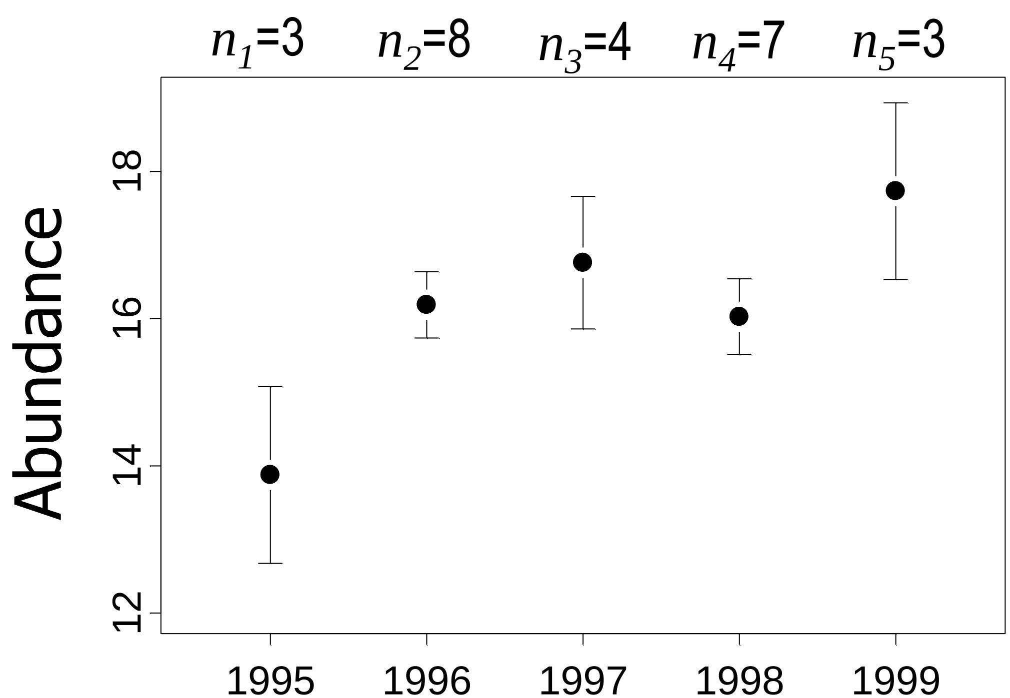
- Method of regression similar to least squares in that it uses the same minimization of the sum of the residuals, but instead of weighting all points equally, they are weighted such that points with a greater weight contribute more to the fit:

$$RSS = \sum_{i=1}^{n} w_i (y_i - f(x_i))^2.$$

- How to choose appropriate weights?
→ Optimal weights are those that weigh each observation by the inverse of its variance, giving points with a lower variance (higher precision) a greater statistical weight: $w_i = 1/\sigma_i^2.$

Weighted least-squares (WLS)

- Main purpose: Deal with responses that have unequal variances!
- Example: What is the trend in abundances of scallops over time?



Each annual estimate based on surveys at n_i randomly selected sites

If estimates of variance ($\hat{\sigma}_i^2$) are available, use:

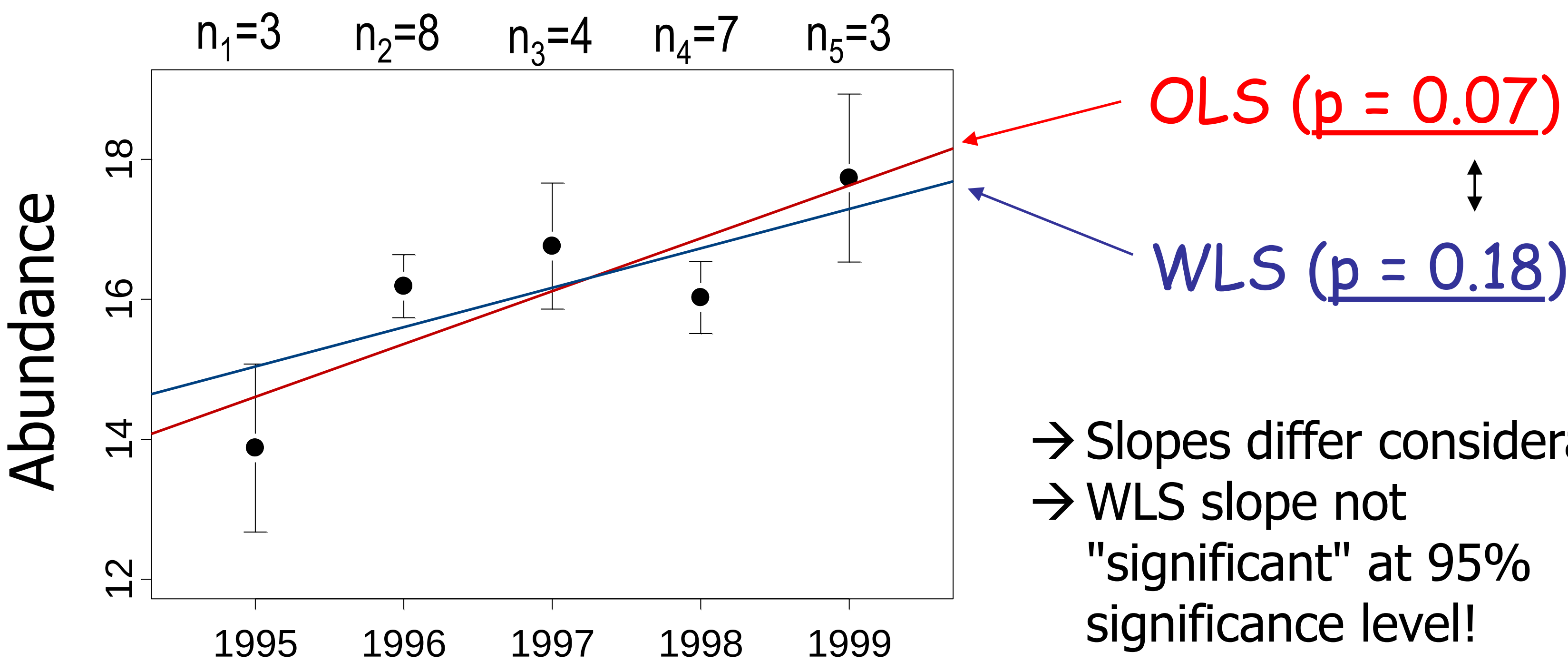
$$w_i = 1/\hat{\sigma}_i^2$$

Otherwise: $w_i = n_i$

(because $\hat{\sigma}_i^2$ is proportional to $1/n_i$)

Weighted least-squares (WLS)

- Unequal variances can have strong influence on regression results!!

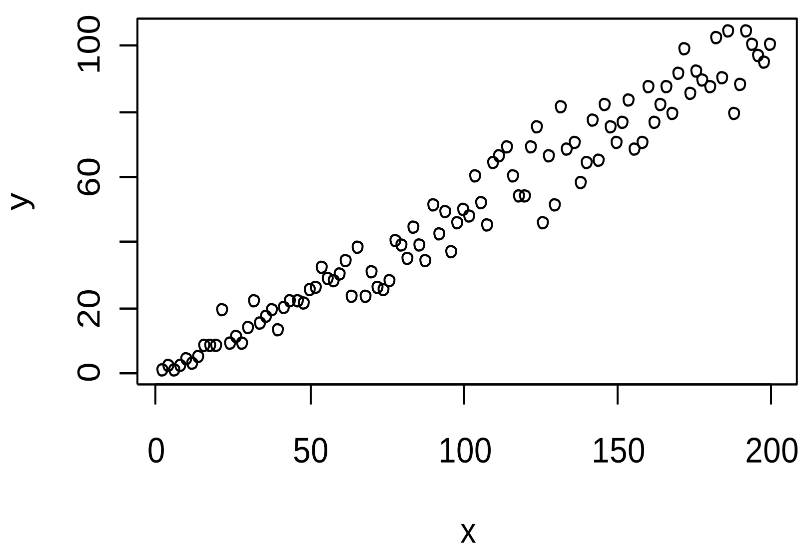


→ Slopes differ considerably
→ WLS slope not "significant" at 95% significance level!

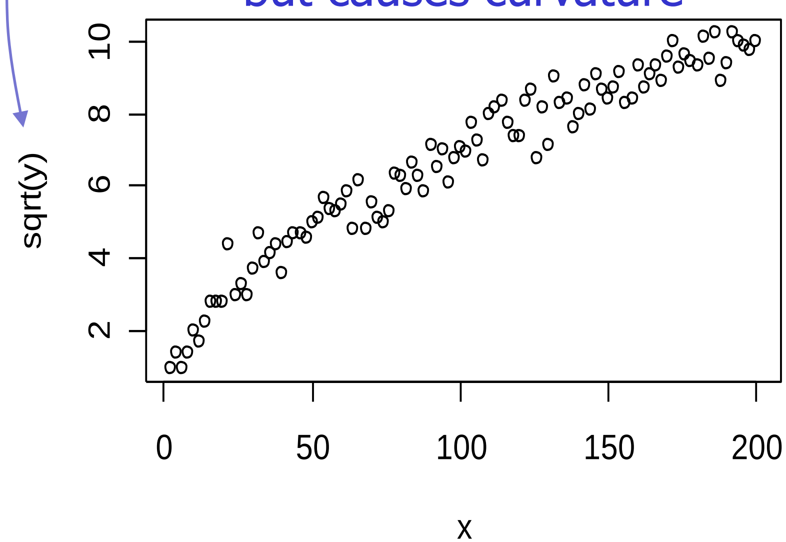
Weighting vs. transformation

- Both weighting and transformation of response variable deal with heteroscedasticity (unequal variances)
- Transformations may <u>also</u> change nature of relationship
- It may be difficult to find transformations (for x and/or y) that result in both homoscedasticity <u>and</u> linearity

Variance of *y* proportional to mean of *y*



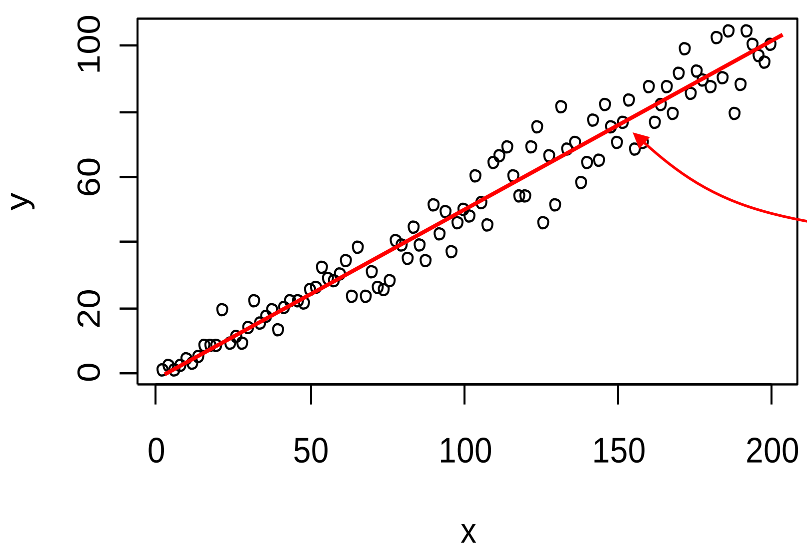
Square-root transformation results in equal variances but causes curvature



Weighting vs. transformation

- Both weighting and transformation of response variable deal with heteroscedasticity (unequal variances)
- Transformations may <u>also</u> change nature of relationship
- It may be difficult to find transformations (for x and/or y) that result in both homoscedasticity <u>and</u> linearity

Variance of y proportional to mean of y



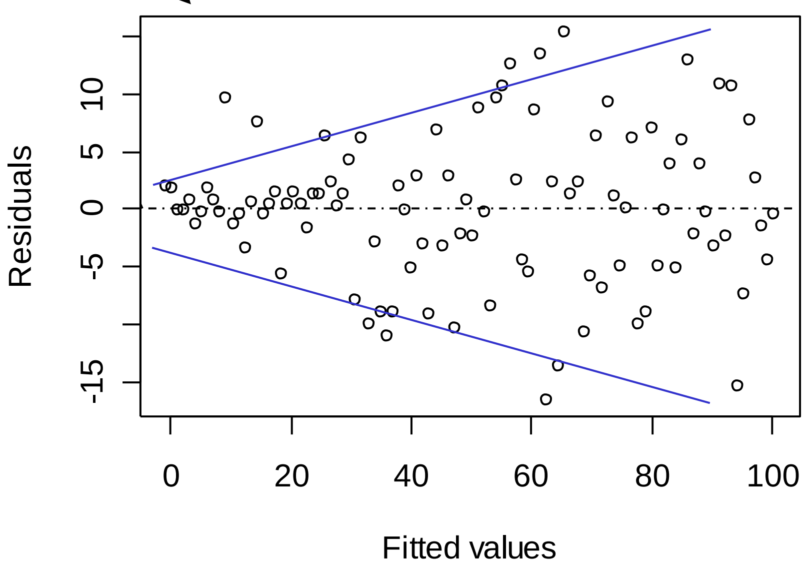
→Use weighting approach

1. Start by fitting OLS regression and taking residuals

Weighted least-squares (WLS)

<u>Weighting approach</u>

1. Fit OLS regression (see previous slide)
2. Plot residual vs. fitted values ($\hat{y}$)
3. Determine relationship between $\hat{\sigma}^2$ and $\hat{y}$: $\boxed{\hat{\sigma} = f(\hat{y})}$??
4. Choose appropriate weights
5. Fit WLS regression



Some common weighting schemes:

If <u>standard deviation</u> is proportional to fitted values:

$$\hat{\sigma} \sim \hat{y} \quad \Rightarrow \quad w_i = 1/\hat{y}_i^2$$

If <u>variance</u> is proportional to fitted values:

$$\hat{\sigma}^2 \sim \hat{y} \quad \Rightarrow \quad w_i = 1/\hat{y}_i$$

Weighted least-squares (WLS)

- Weights w_i essentially transform the y values such that they have equal variances (without distorting nature of relationship!):

Or, equivalently:

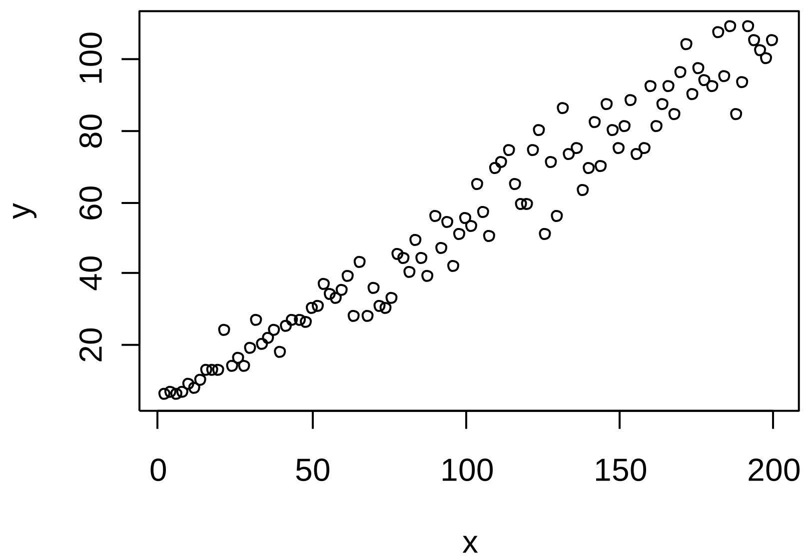
$$\tilde{y}_i = \sqrt{w_i} \cdot y_i$$

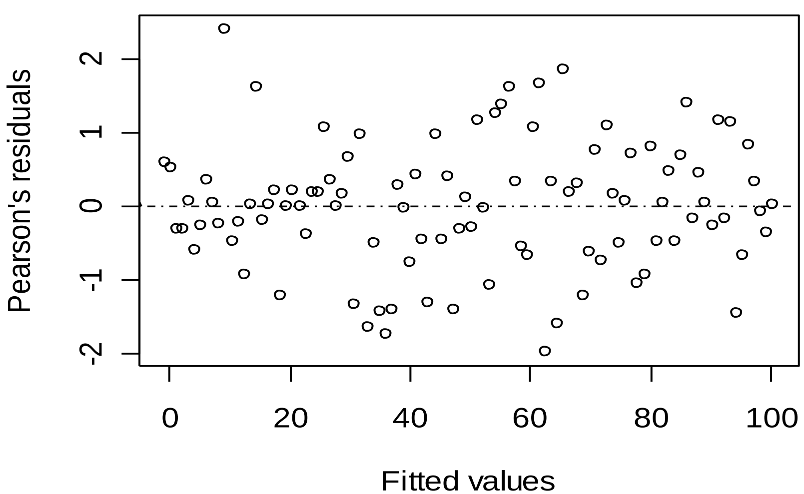
$$\mathrm{var}(\tilde{y}_i) = \mathrm{constant}$$

$$\tilde{\varepsilon}_i = \sqrt{w_i} \cdot \varepsilon_i$$

$$\mathrm{var}(\tilde{\varepsilon}_i) = \mathrm{constant}$$

Weighted least-squares (WLS)





R code for the case where σ^2 is proportional to $\hat{y}$:

```
plot(x,y)
fit.ols <- lm(y ~ x)
y.hat <- fitted(fit.ols)
fit.wls <- lm(y ~ x,
              weights = 1/y.hat)
```

($\hat{y}$ must be positive!!)

```
r <- resid(fit.wls, type =
                    "pearson")
plot(fitted(fit.wls), r)
```

$$\sqrt{w_i} \cdot \hat{\varepsilon}_i$$

MSL
FISH 604

WLS fitting in R

- The main function for linear least-squares regressions (`lm`) has a "weights = *values* " argument (as do most other model fitting functions)
- The main non-linear regression function (`nls`) also allows weights!
Or: use `nls(~ sqrt(W)*(y-f(x)))`

MSL FISH 604

Generalized least-squares (GLS)

- OLS assumes independent, identically distributed errors
- WLS assumes independent errors, but allows unequal variances
- GLS is a further generalization and allows both dependent errors and unequal variances
  - Dependent (correlated) errors produce standard errors that are too small and have fewer d.f. than expected (but have no effect on bias!)
  - Solution: specify correlation structure of errors and account for dependence AND unequal variances

Generalized least-squares (GLS)

- Variance-covariance structure of errors is critical in least-squares fitting and is summarized by var-cov matrix:

$$\operatorname{var}(e_i) = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \sigma_{13} & \sigma_{14} & \cdots & \sigma_{1n} \\ \sigma_{21} & \sigma_{22} & \sigma_{23} & \sigma_{24} & & \sigma_{2n} \\ \sigma_{31} & \sigma_{32} & \sigma_{33} & \sigma_{34} & & \sigma_{3n} \\ \sigma_{41} & \sigma_{42} & \sigma_{43} & \sigma_{44} & & \sigma_{4n} \\ \vdots & & & \vdots & \ddots & \vdots \\ \sigma_{n1} & \sigma_{n2} & \sigma_{n3} & \sigma_{n4} & \cdots & \sigma_{nn} \end{bmatrix}$$

MSL FISH 604

Generalized least-squares (GLS)

Independent, identically distributed errors (equal variances):

$$\mathrm{var}(e_i) = \begin{bmatrix} \sigma^2 & 0 & 0 & 0 \\ 0 & \sigma^2 & 0 & 0 \\ 0 & 0 & \sigma^2 & 0 \\ 0 & 0 & 0 & \sigma^2 \end{bmatrix} = \sigma^2 \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

Independent, unequal variances:

$$\mathrm{var}(e_i) = \begin{bmatrix} \sigma_1^2 & 0 & 0 & 0 \\ 0 & \sigma_2^2 & 0 & 0 \\ 0 & 0 & \sigma_3^2 & 0 \\ 0 & 0 & 0 & \sigma_4^2 \end{bmatrix}$$

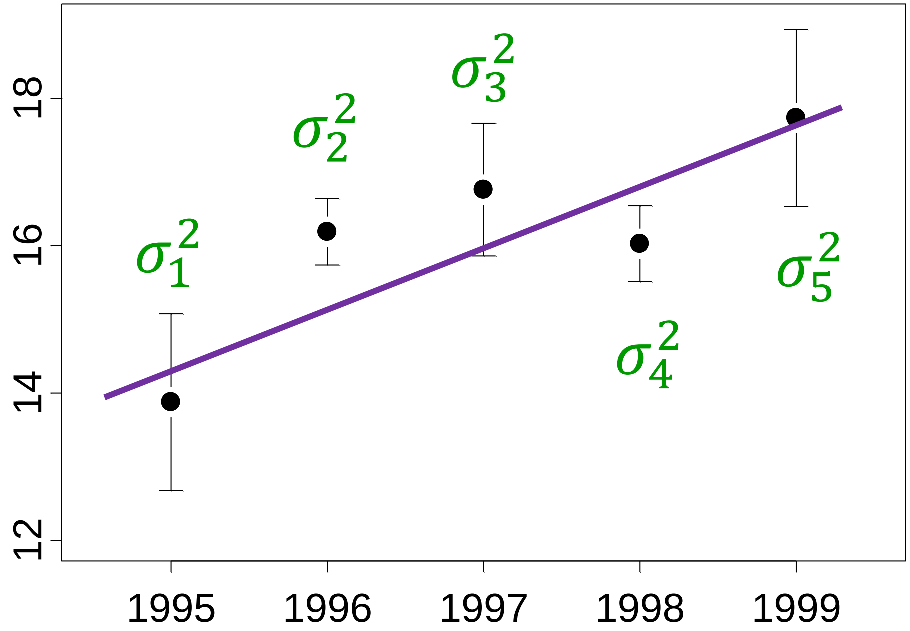
Find weights, such that:

$$\mathrm{var}\left(\sqrt{w_i}\; e_i\right) = \begin{bmatrix} \sigma^2 & 0 & 0 & 0 \\ 0 & \sigma^2 & 0 & 0 \\ 0 & 0 & \sigma^2 & 0 \\ 0 & 0 & 0 & \sigma^2 \end{bmatrix}$$

→ Weighted regression: minimize $RSS_W = \sum_{i=1}^{n} w_i (y_i - \mu_i)^2$

Example: Unequal residual variances

- Recall time series from above: Each residual has a different (but in this case) know variance.
- Often, we make assumptions about the variance of residual *i* (e.g. its relationship with the mean) and then estimate it! If variance is constant, we simply estimate a constant residual variance σ^2 (OLS)



MSL
FISH 604

Generalized least-squares (GLS)

Dependent errors, unequal variances:

$$\mathrm{var}(e_i) = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \sigma_{13} & \sigma_{14} \\ \sigma_{21} & \sigma_{22} & \sigma_{23} & \sigma_{24} \\ \sigma_{31} & \sigma_{32} & \sigma_{33} & \sigma_{34} \\ \sigma_{41} & \sigma_{42} & \sigma_{43} & \sigma_{44} \end{bmatrix}$$

(most general case)

Typically, covariances are not all different (to simplify further, we will look at correlation matrix instead):

Example 1: "adjacent" values have fixed correlation ϕ, which induces correlations at larger d'istances'

$$\rho_e = \begin{bmatrix} 1 & \phi & \phi^2 & \phi^3 \\ \phi & 1 & \phi & \phi^2 \\ \phi^2 & \phi & 1 & \phi \\ \phi^3 & \phi^2 & \phi & 1 \end{bmatrix}$$

This is known as first-order autocorrelation (first-order autoregressive process) if observations are equidistant in time or space:

$$\varepsilon_t = \phi\varepsilon_{t-1} + \nu_t \quad \text{where: } \nu_t \sim N(0,\sigma^2)$$

Generalized least-squares (GLS)

MSL FISH 604

<u>Example 2</u>: Groups of observations are correlated (fixed within-group correlations r_i), no correlation ($=0$) between groups (2 groups with 3 and 2 observations in this example)

$$\rho_e = \left[\begin{array}{ccc:cc} 1 & r_1 & r_1 & 0 & 0 \\ r_1 & 1 & r_1 & 0 & 0 \\ r_1 & r_1 & 1 & 0 & 0 \\ \hdashline 0 & 0 & 0 & 1 & r_2 \\ 0 & 0 & 0 & r_2 & 1 \end{array}\right]$$

<u>Example 3</u>: Correlations decrease with "distance" (e.g. spatial data) This matrix is for a case where observations are equally spaced along a line. In general, correlation is a function of the distance between each pair of observations.

$$\rho_e = \begin{bmatrix} 1 & 0.8 & 0.5 & 0.1 & 0 \\ 0.8 & 1 & 0.8 & 0.5 & 0.1 \\ 0.5 & 0.8 & 1 & 0.8 & 0.5 \\ 0.1 & 0.5 & 0.8 & 1 & 0.8 \\ 0 & 0.1 & 0.5 & 0.8 & 1 \end{bmatrix}$$

MSL FISH 604

Generalized least-squares (GLS)

- As long as the variance-covariance matrix has a certain structure (positive-definite), GLS will result in "best linear unbiased estimates"
  - Generalized LS estimate minimizes the **generalized residual sum of squares**:

$$RSS_G = (\mathbf{y} - \mu)'\mathbf{W}(\mathbf{y} - \mu)$$

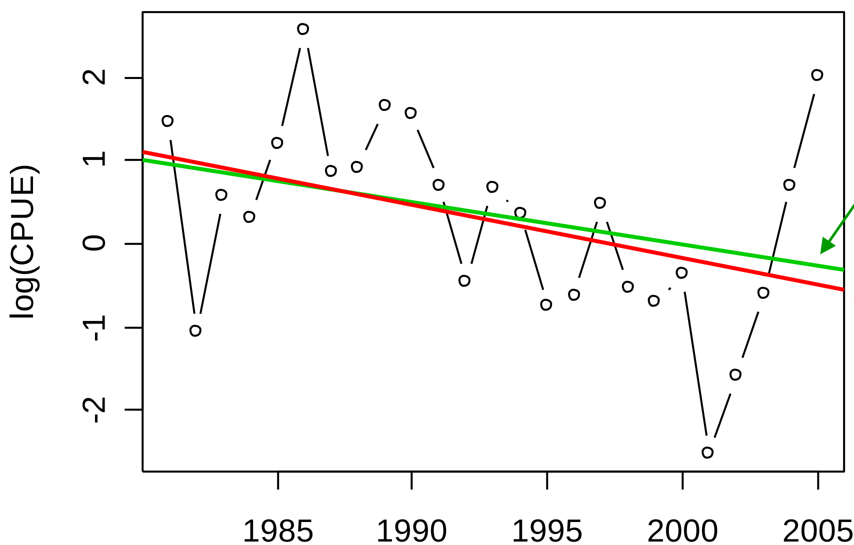
- Variance estimates can be obtained using generalized var-cov matrix $\mathbf{W}$
- F-tests / t-tests are valid (using appropriate generalized RSS_G)

GLS fitting in R

MSL FISH 604

Package: *nlme*

- Functions: `gls` (linear), `gnls` (non-linear)
- Example: GLS regression of rock sole catch-per-unit-effort on year (test for linear trend)
  - minor impact on value of coefficients (both unbiased)
  - strong impact on variances (higher uncertainty!)



GLS fit:

```
gls(y~x, correl = corAR1(),
         method="ML")
```

Slope	se	p-value
-0.051	0.045	0.261

OLS fit:

`lm(y~x)` OR `gls(y~x)`

Slope	se	p-value
-0.064	0.031	0.049

compare

Assigned reading

- Jennrich (1995). An introduction to computational statistics (p. 67-71 and p. 188-190)
- Faraway (2005). Linear models with R (p. 89-94)

→ pdf files in 'Readings'

MSL FISH 604

Further reading

Least-squares regression (& more):

- Jennrich, R.I., 1995. An introduction to computational statistics: Regression analysis. Prentice Hall, Englewood Cliffs, NJ.

- Hilborn, R., Mangel, M., 1997. *The ecological detective: Confronting models with data*. Princeton University Press, Princeton, NJ. (Chpt. 5, 7)