Assessing Goodness of Fit for Linear Models with Correlated Outcomes*

Andy Houseman

Brent A. Coull

Louise M. Ryan

* Based on work to appear in JASA, in revision at Biometrika

Outline

- Assessing goodness of fit for ordinary linear regression. Impact of estimating model parameters
- Correlated data models
  – Motivating examples
  – Proposed GOF methods
  – Back to examples
  – Some simulations
  – Discussion

Ordinary Linear Model

$$y = X\beta + \varepsilon, \quad \varepsilon \sim \boldsymbol{N}_n\left(\mathbf{0}, \gamma^2 \boldsymbol{I}\right)$$

Many approaches to assessing goodness of fit.
We focus here on residual-based methods, using .

$$z_i = \frac{y_i - X_i\beta}{\gamma}$$

and examining Q-Q plots, or calculating functionals of the residuals (Kolmogorov Smirnov or Cramer von Mises tests)

Empirical CDF

- **Empirical CDF:**

$$\hat{F}_n(x) = n^{-1}\sum_{i=1}^{n} I\{z_i \le x\}$$

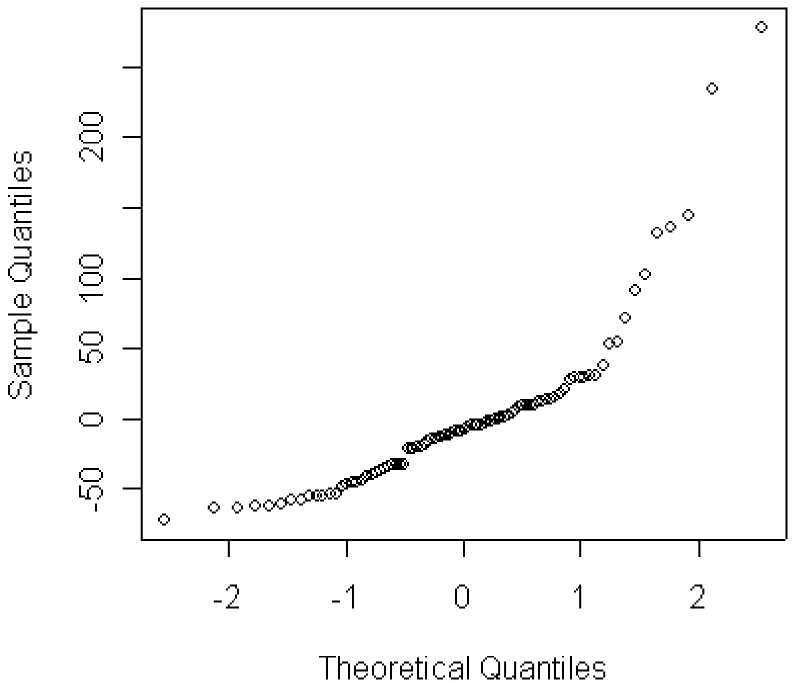
- If $z_i \overset{iid}{\sim} N(0,1)$

$$n^{1/2}\left(\hat{F}_n(x) - \Phi(x)\right) \Rightarrow G\left(0, \sigma^2(x,y)\right)$$

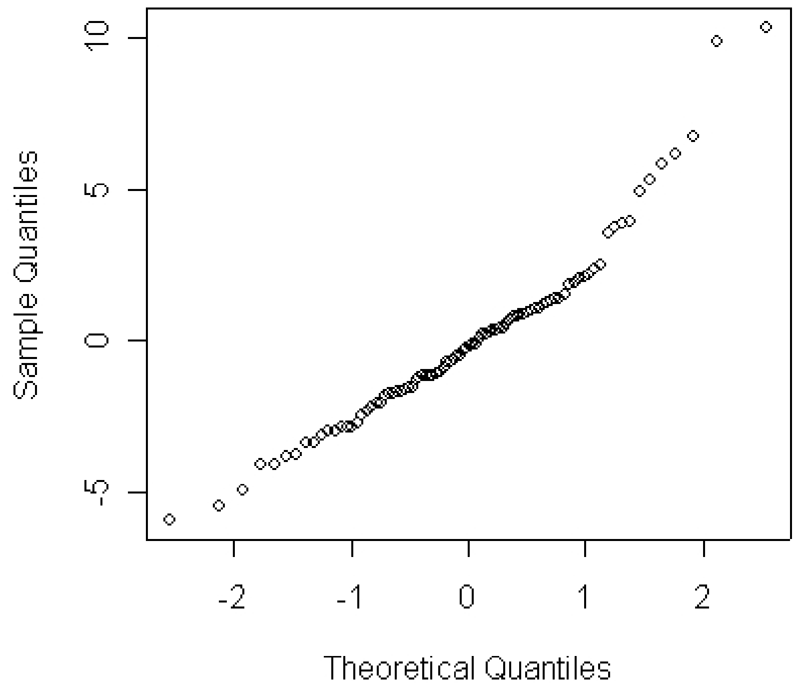
$$\sigma^2(x,y) = \Phi(x)\left(1-\Phi(y)\right)$$

Quantile-Quantile Plot

Definitely Not Normal



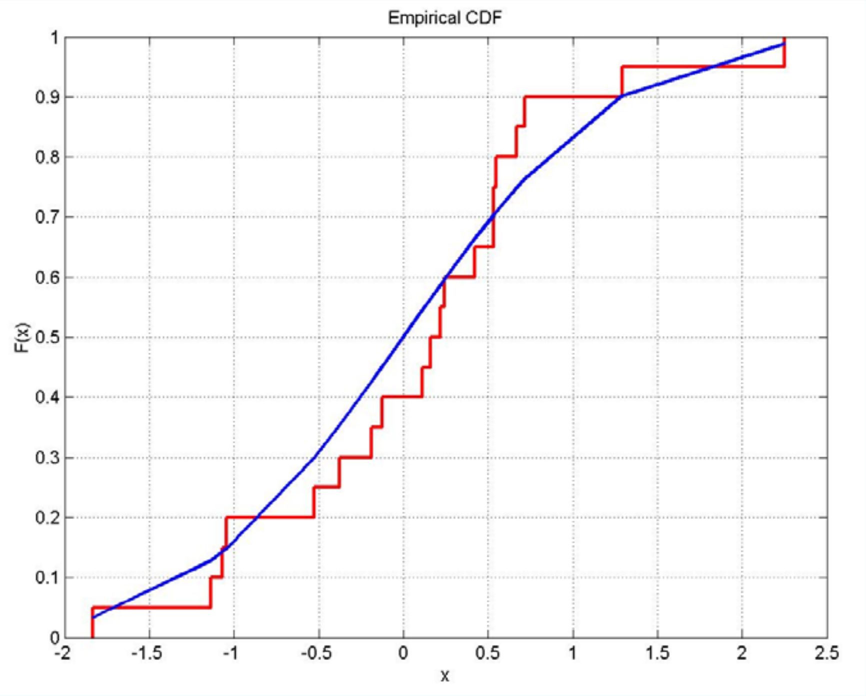
Possibly Normal?



Adding pointwise bands helps with visualization, but global bands needed for formal inference

Tests based on the ECDF

- KS test – maximum departure from expected normal cdf
- Cramer-von Mises – average squared departure



Properties of ECDF and tests based on it affected by the estimation of parameters

Handling unknown model parameters

- Estimated residual $\hat{z}_i = (y_i - X_i\hat{\beta}) / \hat{\gamma}$
- Empirical CDF:

$$\hat{F}_n\left(x; \hat{\beta}, \hat{\gamma}\right) = n^{-1} \sum_{i=1}^{n} I\left\{z_i\left(\hat{\beta}, \hat{\gamma}\right) \le x\right\}$$

- Still asymptotically Gaussian,

$$n^{1/2} \hat{F}_n\left(x; \hat{\beta}, \hat{\gamma}\right) \Rightarrow G\left(\Phi(x), \tilde{\sigma}^2(x, y)\right)$$

- Changed variance (Ron Randles and others)

General Linear Model

$$y = X\beta + \varepsilon, \qquad \varepsilon \sim N_n\left(0, V\left(\gamma\right)\right)$$

- Growth curves
- Linear Mixed Effects Models
- Time Series Regression
- Spatial models
- Crossed random effects Models

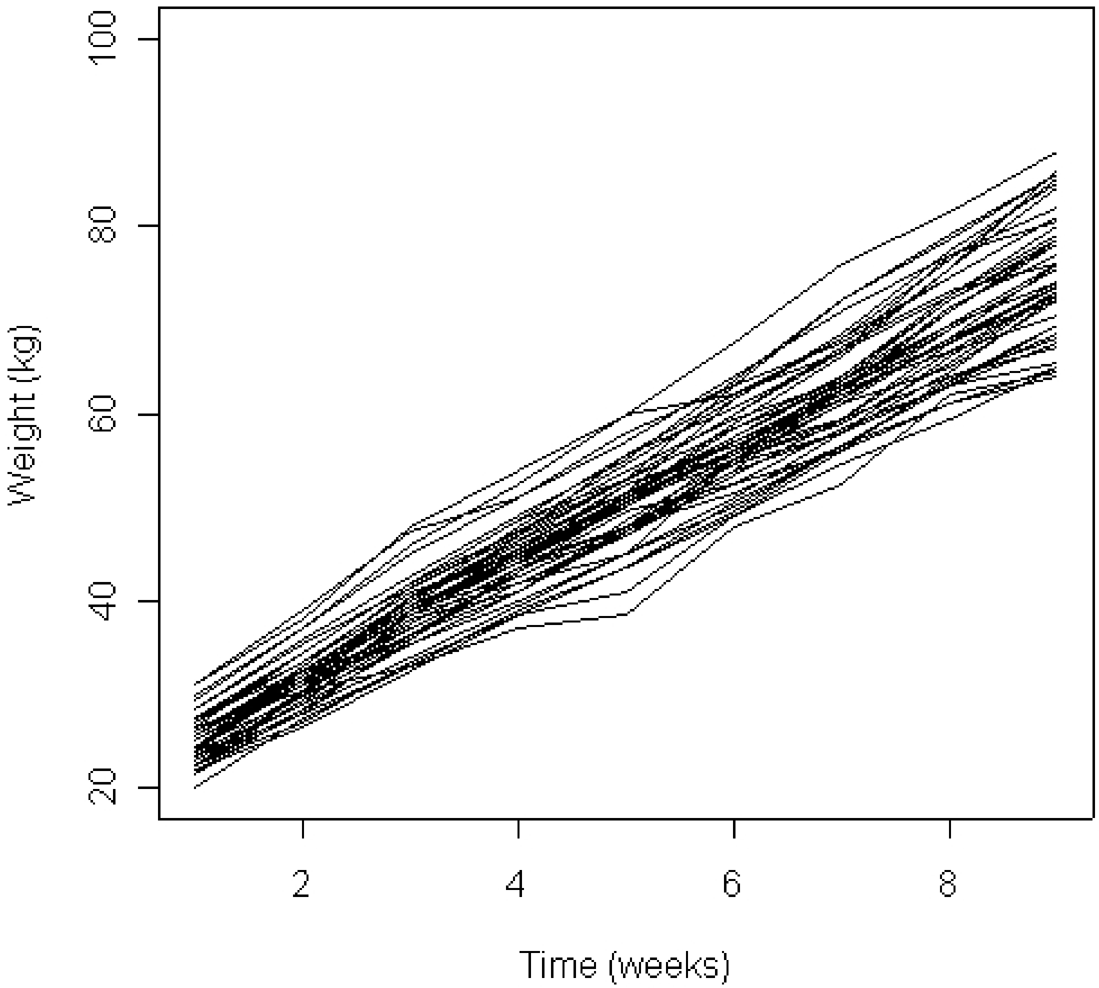
Goodness of fit?

- What does assessing model fit mean?
- How to generalize residual-based methods

Let’s look at a couple of examples….

Pig Weights

Diggle, Heagerty et al.



Mixed Effects Models

$$y_i = X_i \beta + Z_i b_i + e_i$$

$$b_i \sim N_d\left(0, \Delta\right) \qquad e_i \sim N_{k_i}\left(0, \sigma^2 I_{k_i}\right)$$

$$y_i = X_i \beta + \varepsilon_i$$

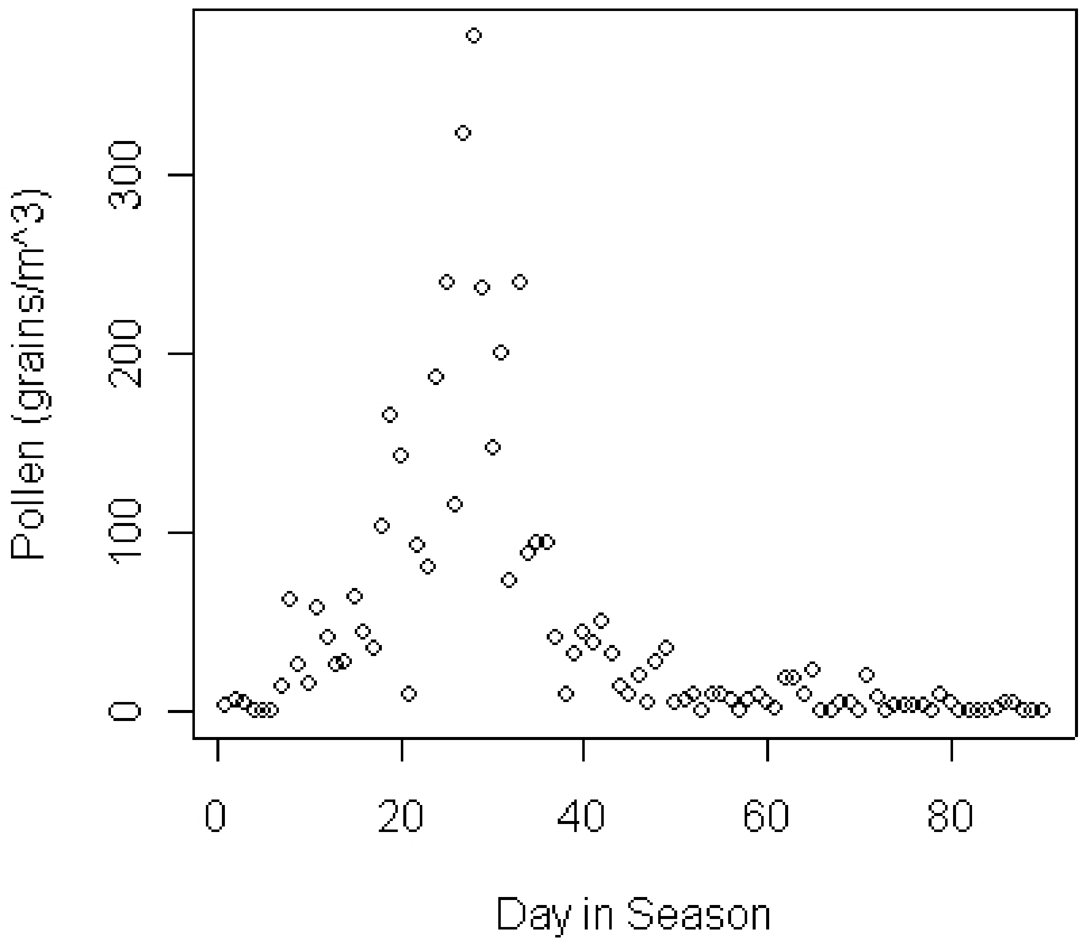
$$\varepsilon_i \sim N_{k_i}\left(0, Z_i \Delta Z_i^T + \sigma^2 I_{k_i}\right)$$

What does it mean to assess fit?

- Are the error terms normal with mean zero?
- Are the random effects normal with mean zero?
- Why important?
  - Is the mean modeled properly?
  - Good properties of random effects BLUPs depend on normality of random effects
  - Fixed effects estimates can depend on normality assumptions (certainly efficiency, maybe even bias – Ray’s talk?)

Pollen Counts

Stark et al. (1997) and Brumback et al. (2000)



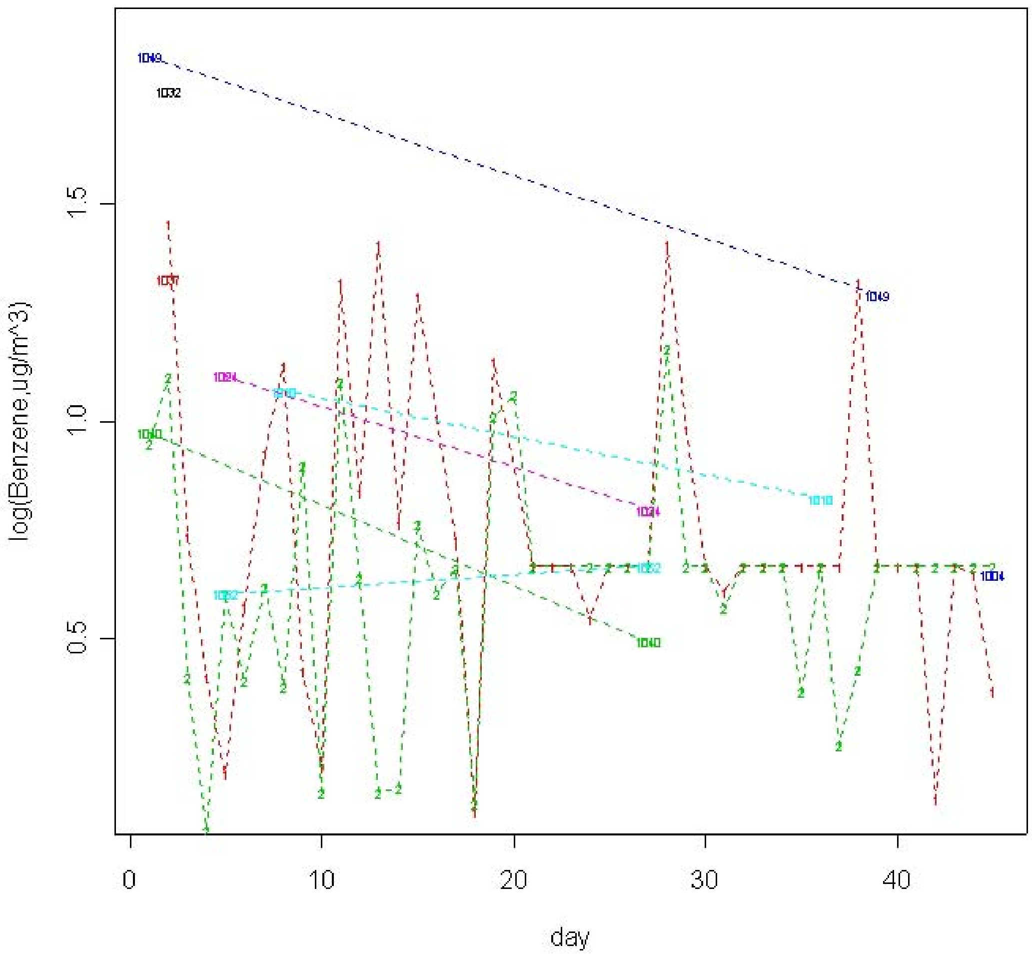
Time Series Regression

$$y_t = X_t\beta + \sum_{s=1}^{k} \rho_s \left(y_{t-s} - X_{t-s}\beta\right) + e_t$$

$$y = X\beta + \varepsilon, \qquad \varepsilon \sim N_n\left(0, \sigma^2 R\left(\rho_1, \ldots, \rho_k\right)\right)$$

Do we have the right error structure?
Important for prediction.

Volatile Organic Compounds



Benzene by time, two central monitors and multiple homes in Mexico City

Toxics **E**xposure **A**ssessment:
a **C**olumbia **H**arvard Project

Crossed Effects Models

$$y_{ij} = X_{ij}\beta + a_i + b_j + e_{ij}$$

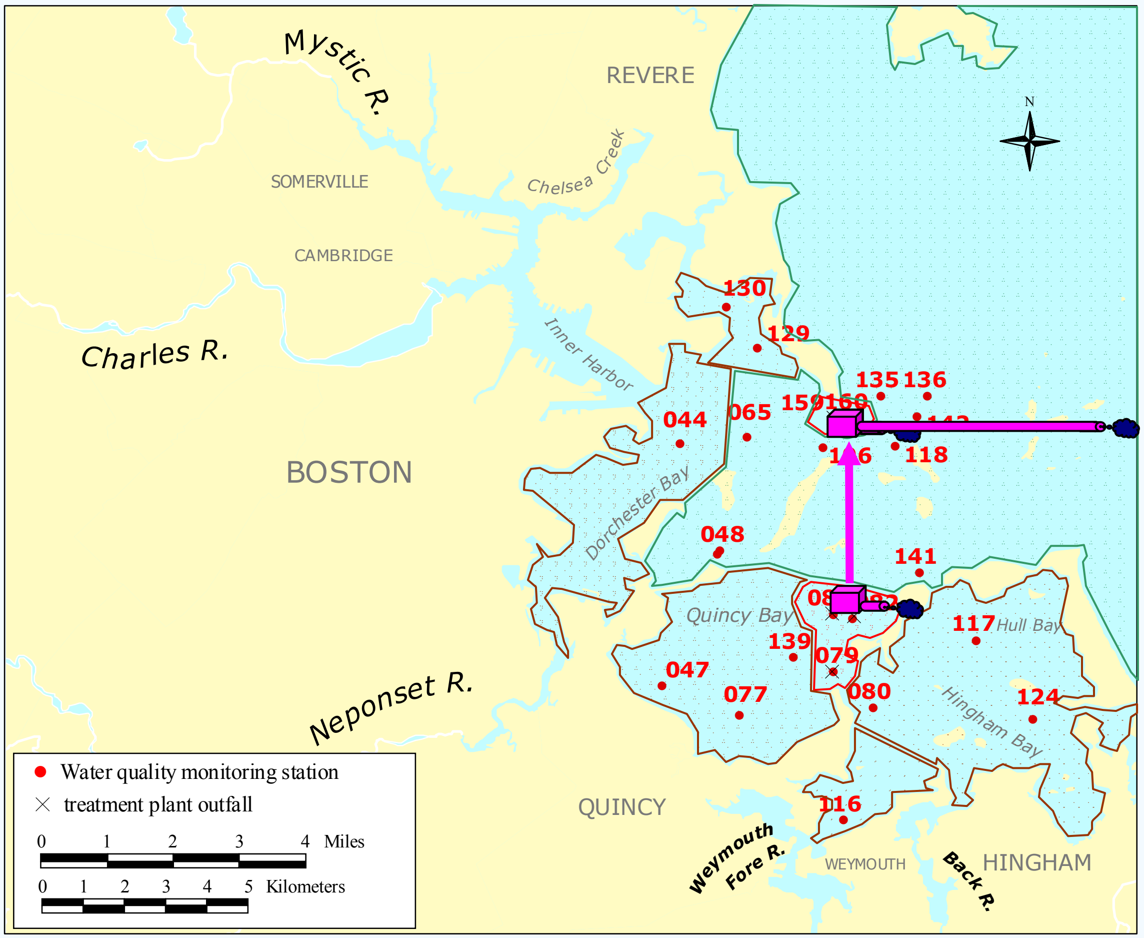
$$a_i \sim N\left(0, \sigma_A^2\right) \qquad b_j \sim N\left(0, \sigma_B^2\right)$$

$$e_{ij} \sim N\left(0, \sigma_0^2\right)$$

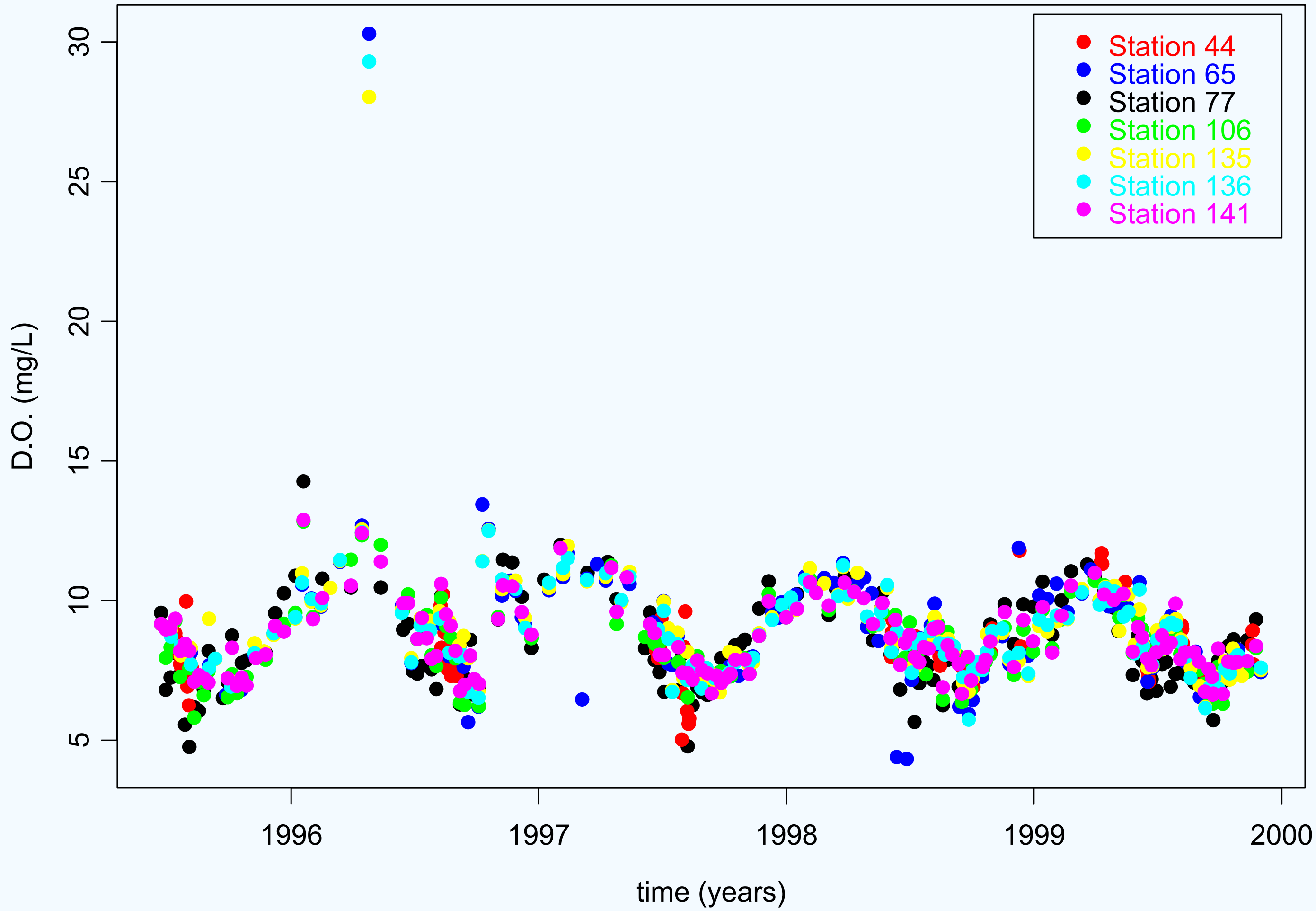
$$y = X\beta + \varepsilon$$

$$\varepsilon \sim N_n\left(0, \sigma_A^2 R_A + \sigma_B^2 R_B + \sigma_0^2 I_n\right)$$

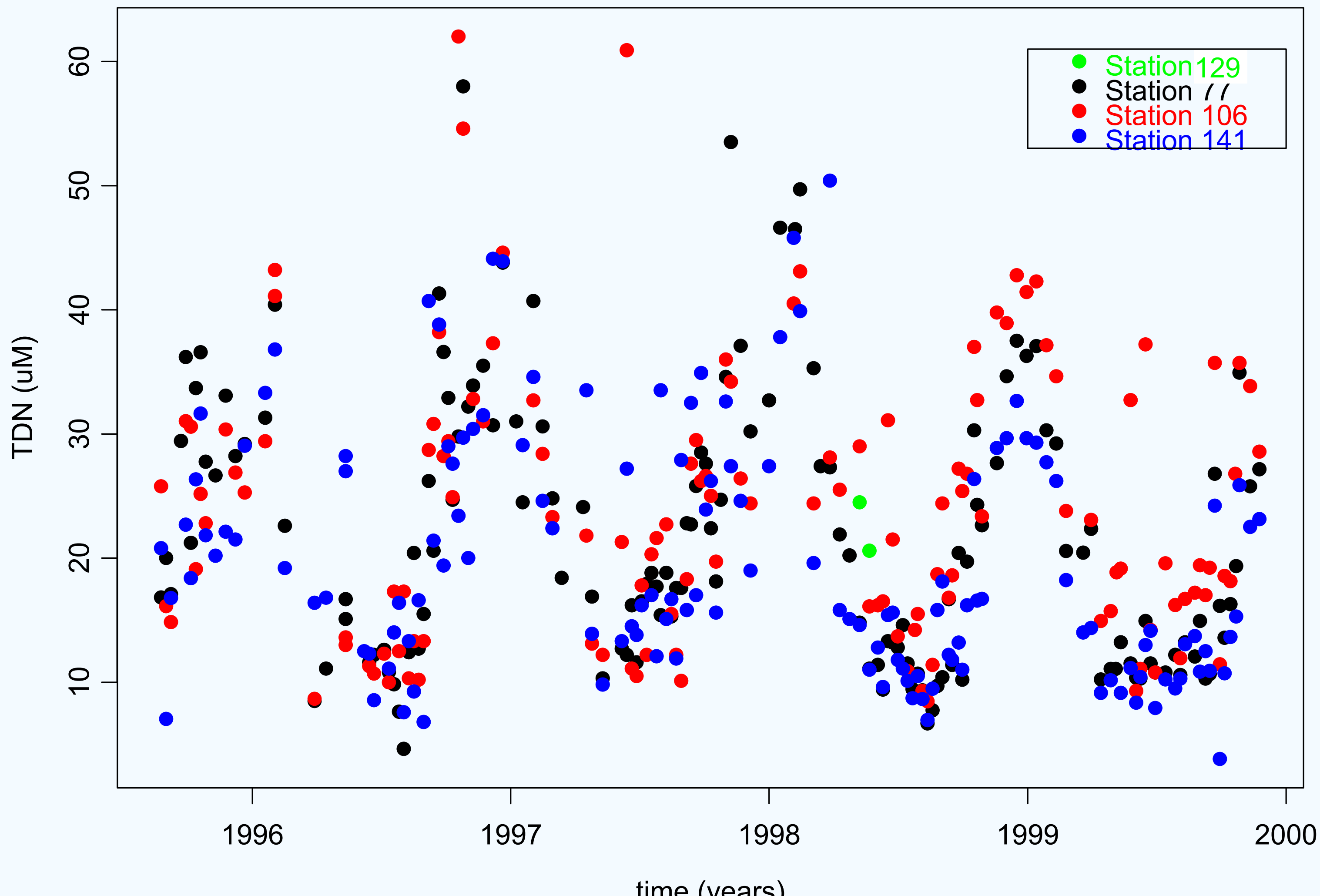
Boston Harbor Data



Dissolved Oxygen (Bottom,1995-2000)



Total Dissolved Nitrogen (Surface, 1995-2002)



Space/Time model?

$$y = X\beta + \varepsilon,$$

$$\varepsilon \sim N_n\left(0, \sigma^2 R\left(\rho_1, \ldots, \rho_k\right)\right)$$

Houseman treated location and time as random effects. Data difficult (lots of zeros etc)

How to assess GOF?

- Pinheiro and Bates (2001) – residuals based on subject-specific means
- Hodges (1998) - complicated
- Jiang (2000) – complicated
- Fraccaro et al. (2000) – residuals for a time series setting; heuristic approach
- Lange & Ryan (1989) –Q-Q plots of standardized random effect estimates (BLUPS)

We generalize the Lange/Ryan approach.

Lange/Ryan standardized BLUPS

- Let

$$\hat{b}_i^{(j)} = \pi_j \hat{\Delta} Z_i \left(Z_i \hat{\Delta} Z_i^T + \hat{\sigma}^2 I_{k_i} \right)^{-1} \left(y_i - X_i \hat{\beta} \right)$$

$$\boldsymbol{Z_i^{(j)} = \hat{b}_i^{(j)} / sd\left(\hat{b}_i^{(j)}\right)}$$

- Pointwise Asymptotics of ECDF of the Z's:

$$n^{1/2}\left(\hat{F}_n\left(x; \hat{\beta}, \hat{\Delta}, \hat{\sigma}^2 \right) - \Phi(x) \right) \Rightarrow N\left(0, \tilde{\sigma}_x^2\right)$$

$$\tilde{\sigma}_x^2 = \Phi(x)\left(1 - \Phi(x)\right) - \delta^T W \delta$$

W = Covariance of estimated parameters, δ a gradient vector (more presently)

Lots of gaps….

- Global Asymptotics?
- Non-clustered data?
- Other diagnostics?
- General residuals?

Cholesky Rotated Residuals

Recall:

$$y = X\beta + \varepsilon, \qquad \varepsilon \sim N_n\left(0, V(\gamma)\right)$$

Define:

$$V(\gamma)^{-1} = L(\gamma)L(\gamma)^T$$

$$z_i(\beta,\gamma) = L(\gamma)^T (y - X\beta)$$

At true parameter values, $z_i(\beta_0,\gamma_0) \overset{iid}{\sim} N(0,1)$

Using the rotated residuals

- Do probability plots of the residuals
- Construct tests such as Kolmogorov or Cramer Von Mises
- Construct functionals of the residuals to target particular departures

Functionals of Rotated Residuals

Choose an appropriate projection:

$$\left(P_1,\ldots,P_N\right), \text{ where } P_i P_j^T = I(i=j)$$

$$z_i\left(\beta,\gamma\right) = P_i L\left(\gamma\right)^T \left(y - X\beta\right)$$

$$z_i\left(\beta_0,\gamma_0\right) \overset{iid}{\sim} N\left(0,1\right)$$

Lange/Ryan standardized BLUPs a special case

Pointwise Asymptotics

$$N^{1/2}\left(\hat{F}_N\left(x;\hat{\theta}\right)-\Phi(x)\right)\Rightarrow N\left(0,\tilde{\sigma}_x^2\right)$$

where

$$\theta=\left(\beta,\gamma\right)^T$$

$$\tilde{\sigma}_x^2=\Phi(x)\left(1-\Phi(x)\right)-\delta_x^T W\delta_x$$

$$W=Var\left(\hat{\theta}\right)$$

Pointwise Asymptotics (2)

$$\delta_x \approx \frac{\partial}{\partial\theta} \sum_{i=1}^{N} \Phi\left(\frac{x - m_i(\theta, \hat{\theta})}{s_i(\theta, \hat{\theta})} \right) \Bigg|_{\theta=\hat{\theta}}$$

$$m_i(\theta, \theta_0) = P_i L(\gamma)^T X(\beta_0 - \beta)$$

$$s_i(\theta, \theta_0) = \left(P_i L(\gamma)^T V(\gamma_0) L(\gamma) P_i^T \right)^{1/2}$$

Closed form exists for δ_x estimate

Global Asymptotics

$$N^{1/2}\left(\hat{F}_N\left(x;\hat{\theta}\right)-\Phi(x)\right)\Rightarrow G\left(0,\tilde{\sigma}^2(x,y)\right)$$

Resampling technique similar to Lin et. al (2002):

$$F^*(x)=\sum_{i=1}^{N} I\left\{P_i z^* \le x\right\}+\hat{\delta}_x^T J^{-1} U\left(\hat{\theta};z^*\right)$$

where $z^* \sim N_n\left(0,I_n\right)$

$$U\left(\hat{\theta};L(\hat{\gamma})^T\left(y-X\hat{\beta}\right)\right)=0 \qquad J=\left.\frac{\partial U}{\partial\theta}\right|_{\theta=\hat{\theta}}$$

Resampling

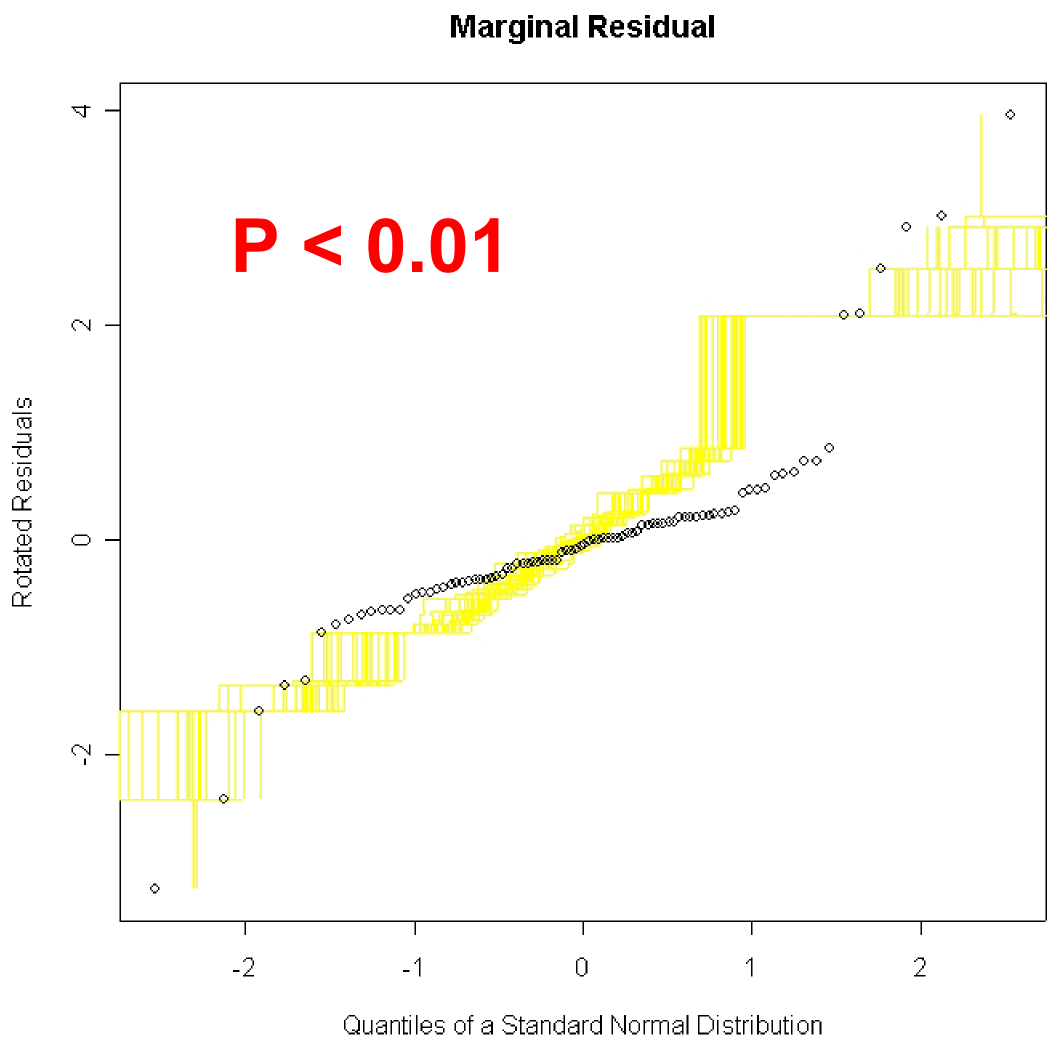
Under H_0: normal $\quad F^*(x \mid y) \cong \hat{F}_n(x;\hat{\theta})$

$$\boldsymbol{\Gamma}\left(F^*\right) \cong \boldsymbol{\Gamma}\left(\hat{F}_n\right)$$

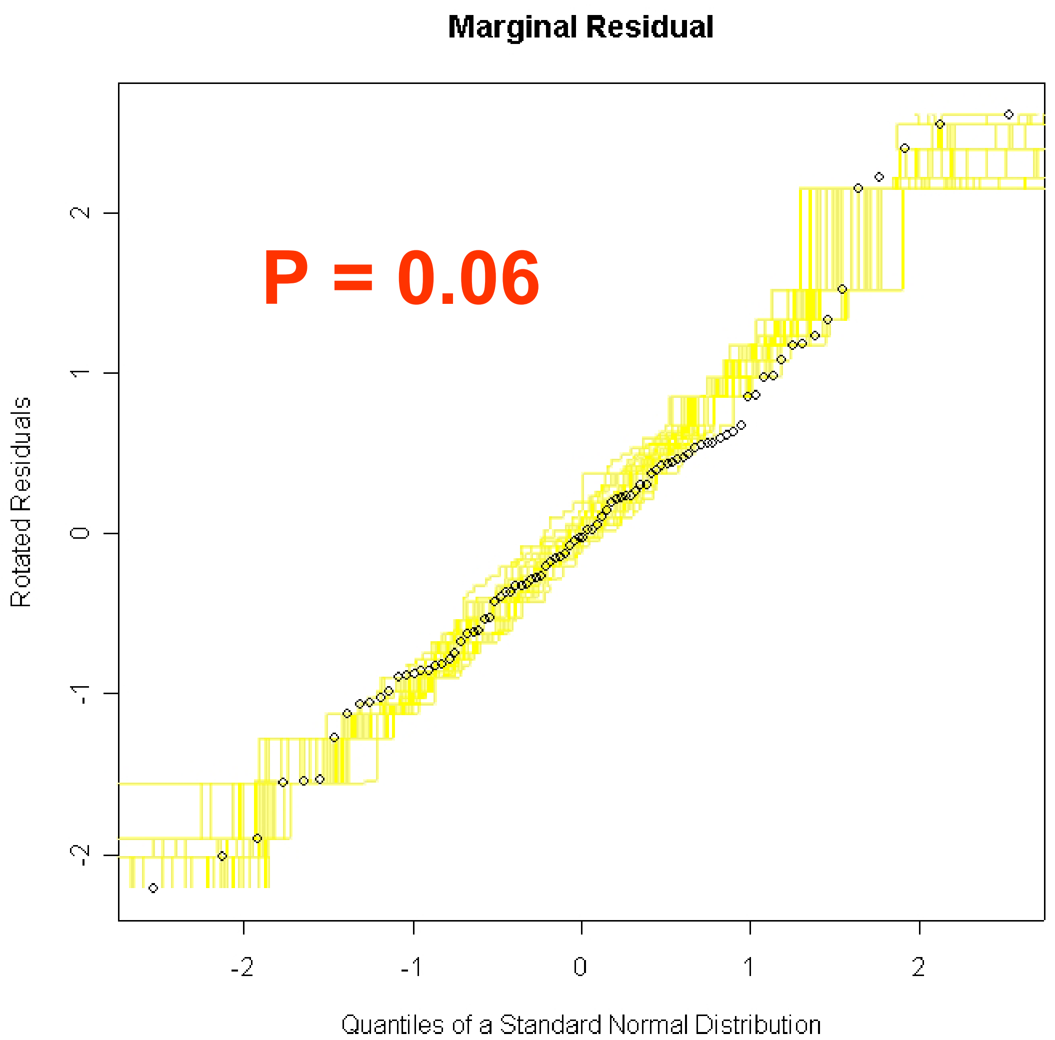
$$\boldsymbol{\Gamma}\left(F\right) = \sup_{x \in \Omega}\left|F(x) - \Phi(x)\right|$$

$$\mathrm{P-value} \approx M^{-1}\sum_{u=1}^{M} I\left(\boldsymbol{\Gamma}\left(F_{(u)}^*\right) > \boldsymbol{\Gamma}\left(\hat{F}_n\right)\right)$$

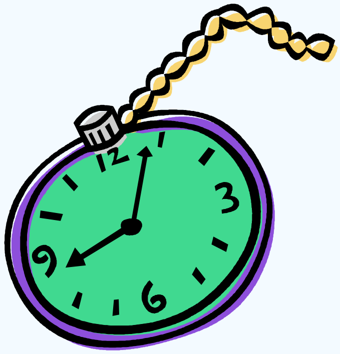
Untransformed Pollen Counts



√ -Transformed Pollen Counts



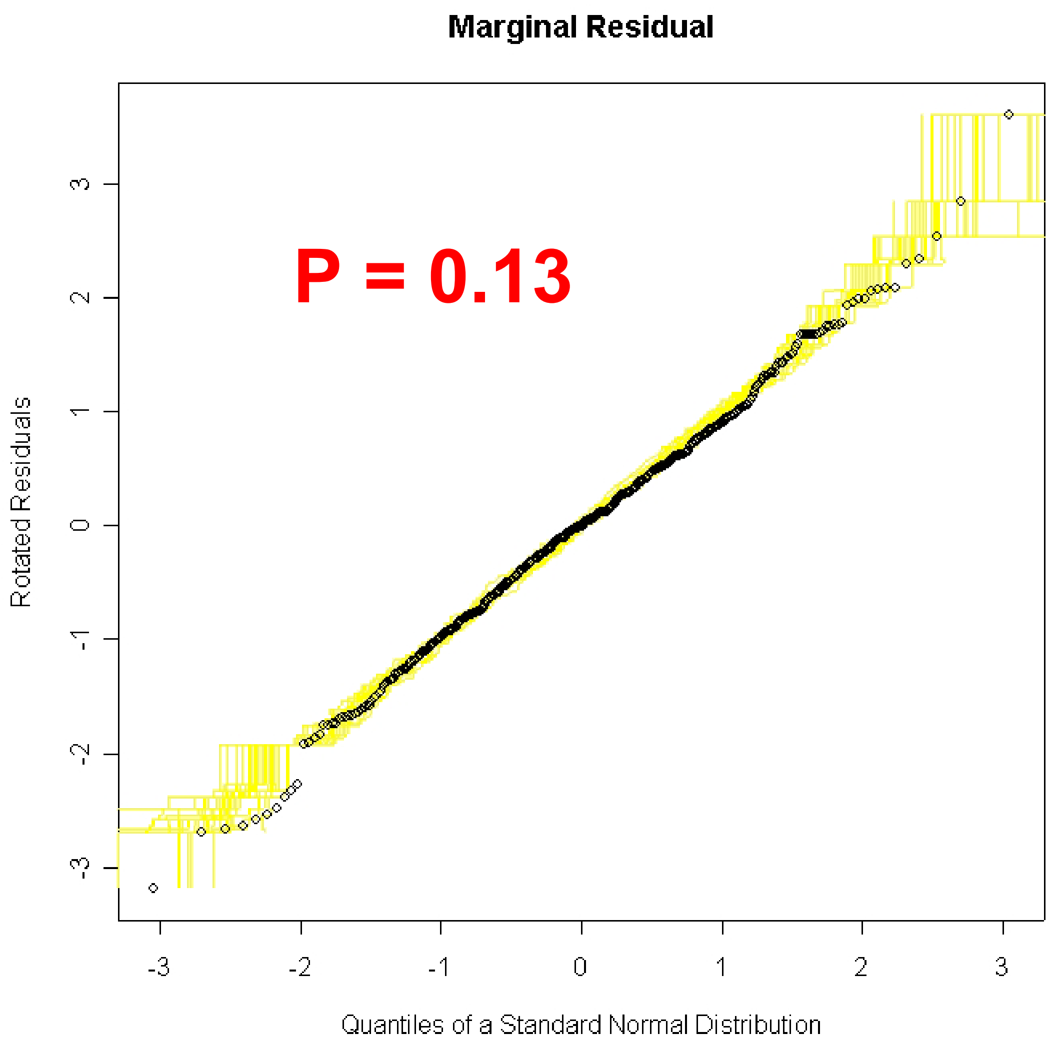
Simulation – time series



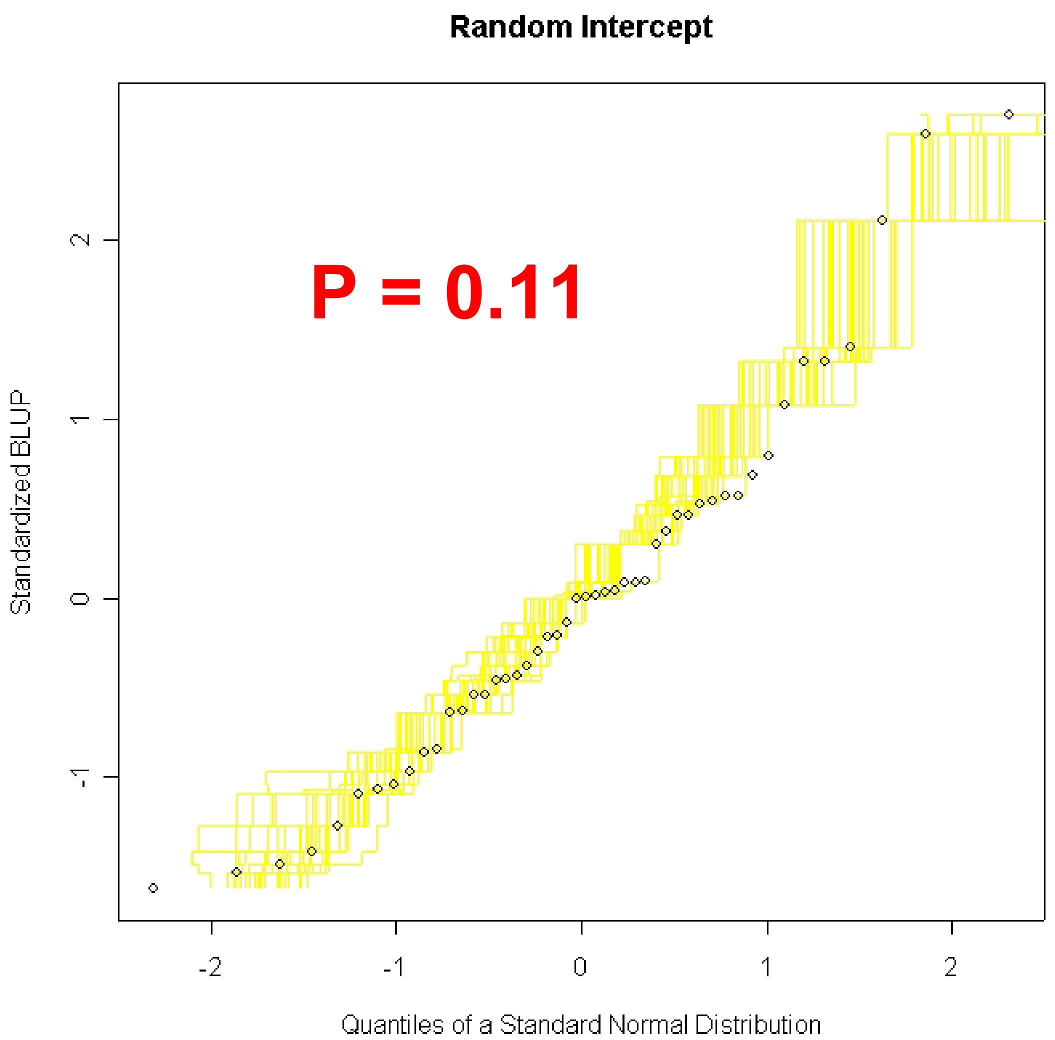
- 1000 simulations of AR(1) time series
- Each series with n=250
- Computed rejection rates (Nominal rate of 5%)

Error distributions	KS	CVM
Normal	.05	.04
Skewed (chisq, 3 df)	.82	.90
Heavy tailed (t, 3df)	.68	.79

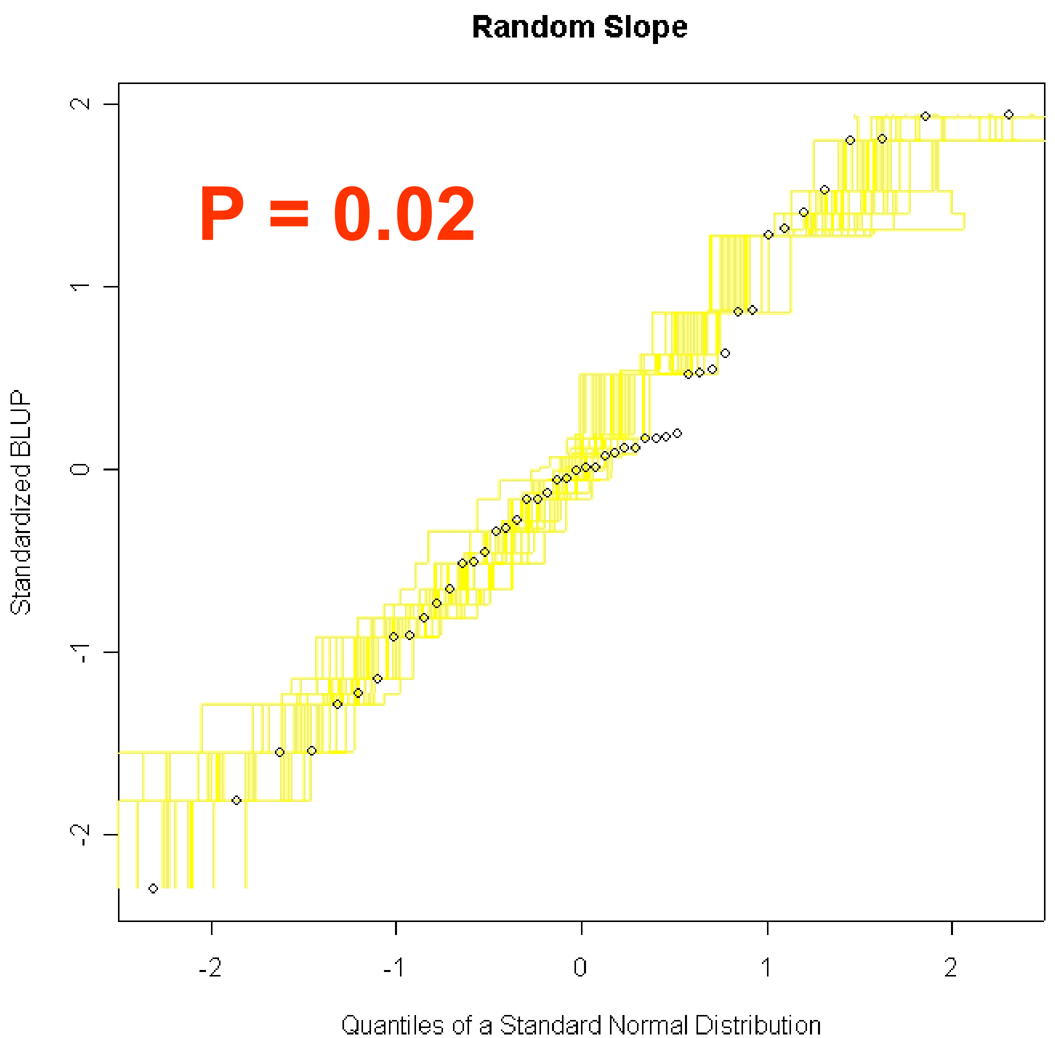
Pig Weights: Marginal Errors



Pig Weights: Random Intercept



Pig Weights: Random Slope



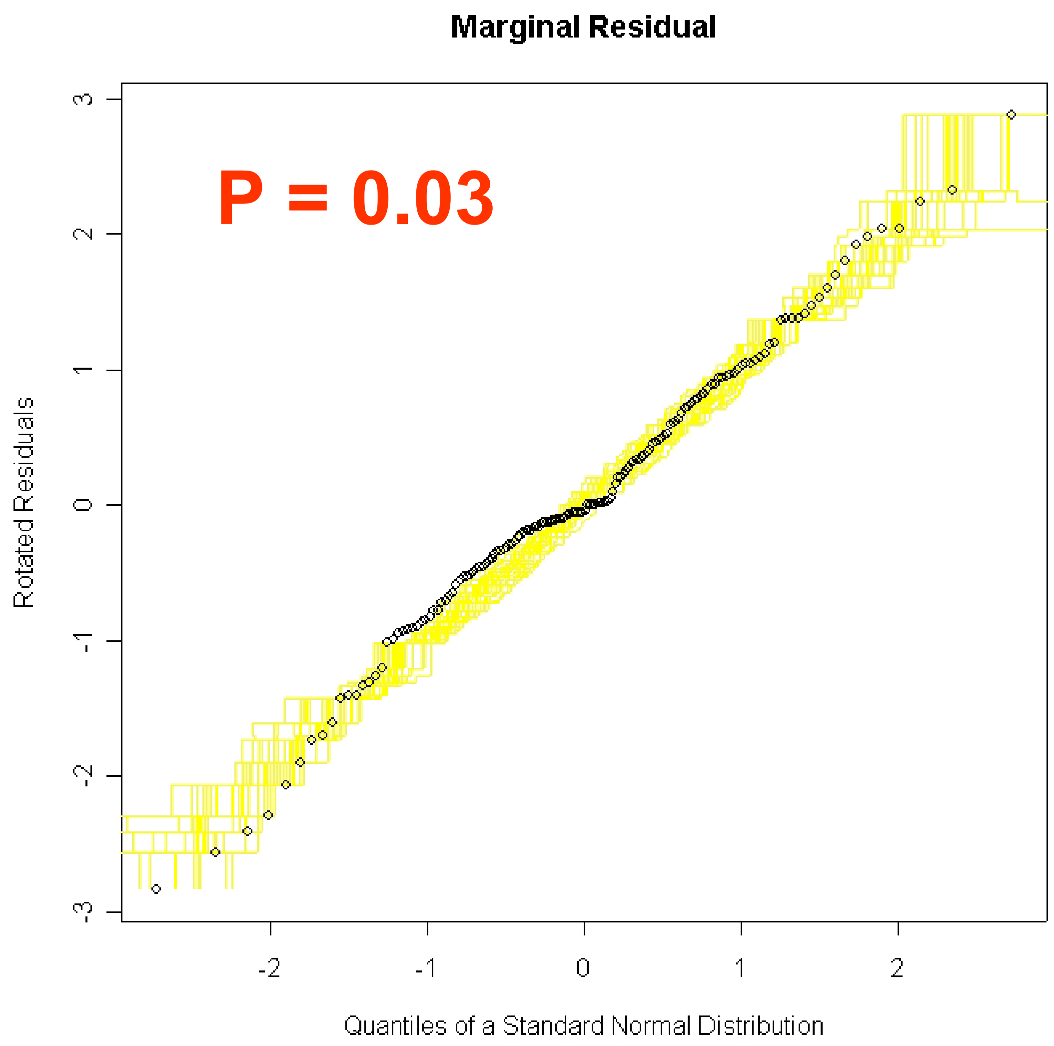
Simulations – random intercept & slope

- 500 simulations of random intercept and slope model, 50 subjects with 5 repeats
- Computed KS and CVM test for Cholesky residuals, random intercept and random slopes
- Rejection rates under following models
  1. Null (random effect and error terms normal)
  2. Skewed random effects
  3. Heavy tailed random effects
  4. Binary random effects
  5. Skewed errors
  6. Heavy tailed errors

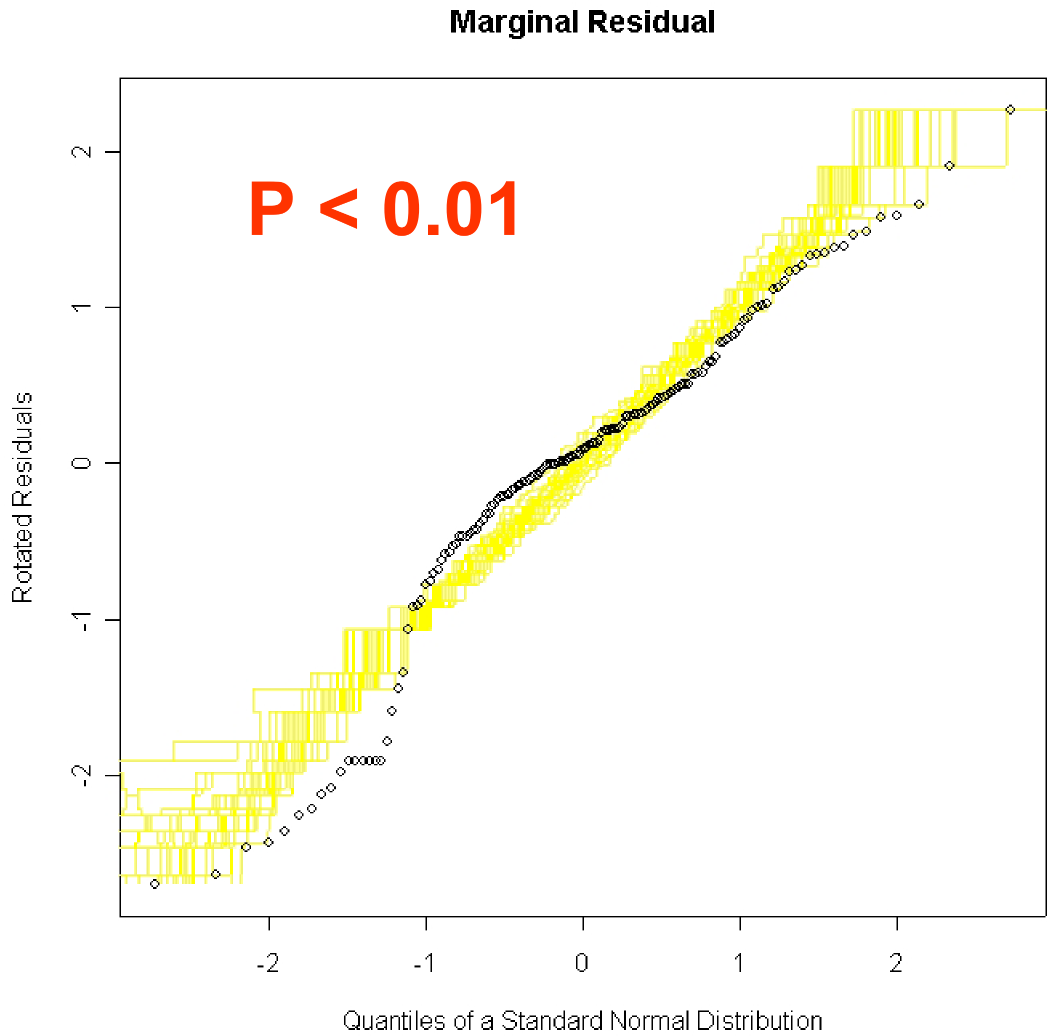
Results

- CVM better than KS for type I error
- CVM had better power than KS to detect non-normality of random effects
- Targeted tests more powerful for detecting skewed and heavy-tailed re distributions
- Easier to detect skewed rather than heavy tailed distributions
- Global test best for detecting non-normality of error terms

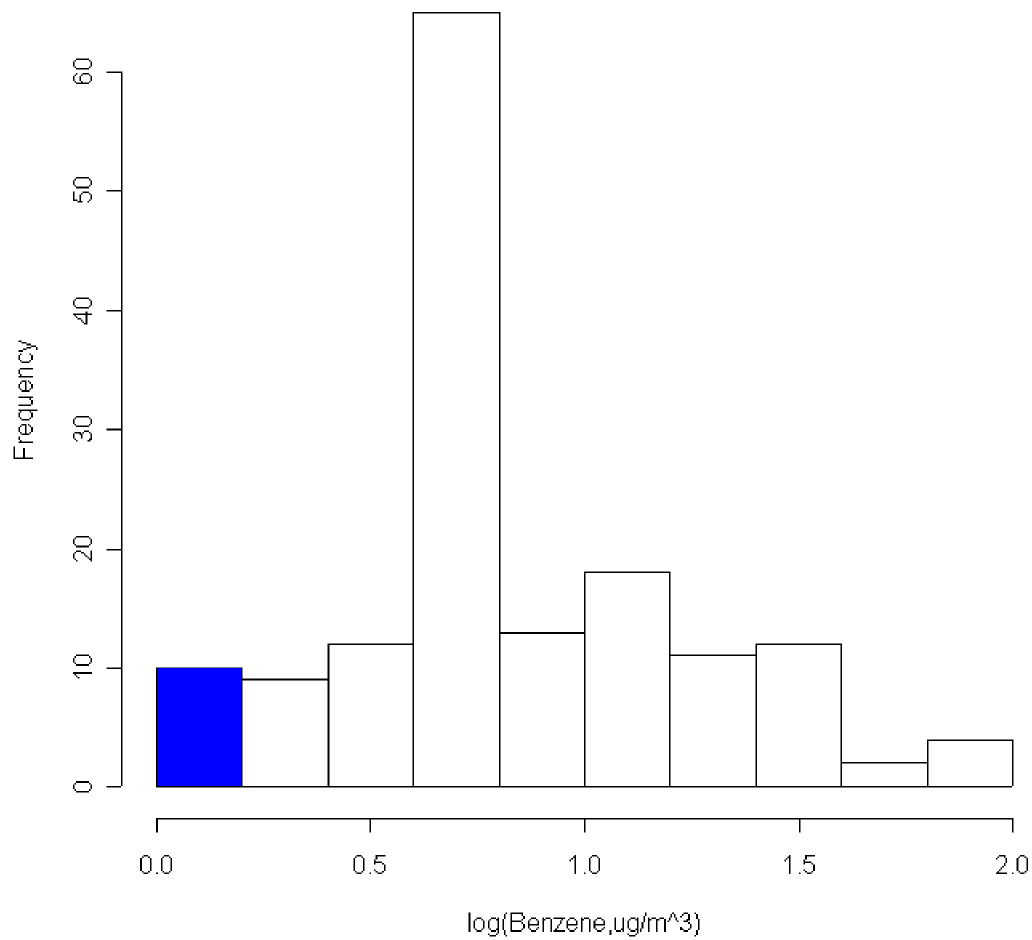
Log-Benzene


Marginal Residual
P = 0.03
Rotated Residuals
Quantiles of a Standard Normal Distribution
3
2
1
0
-1
-2
-3
-2
-1
0
1
2

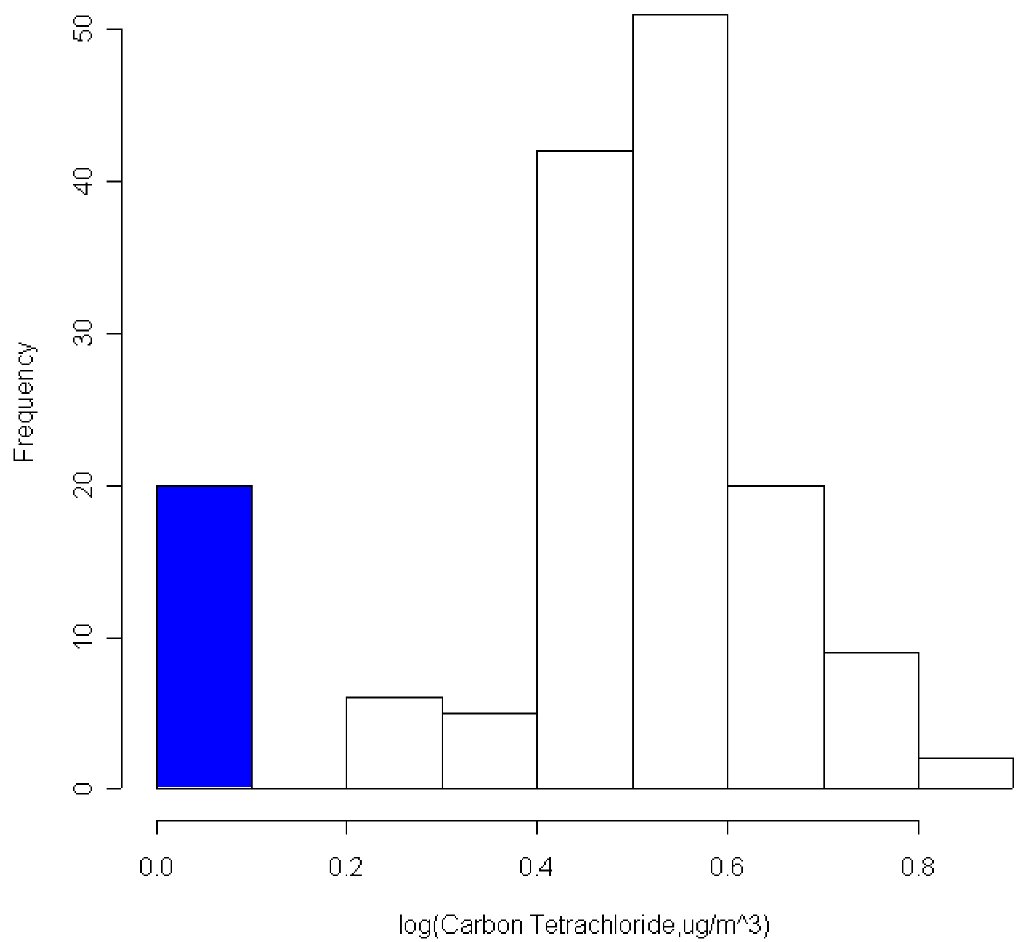
Log-Carbon Tetrachloride



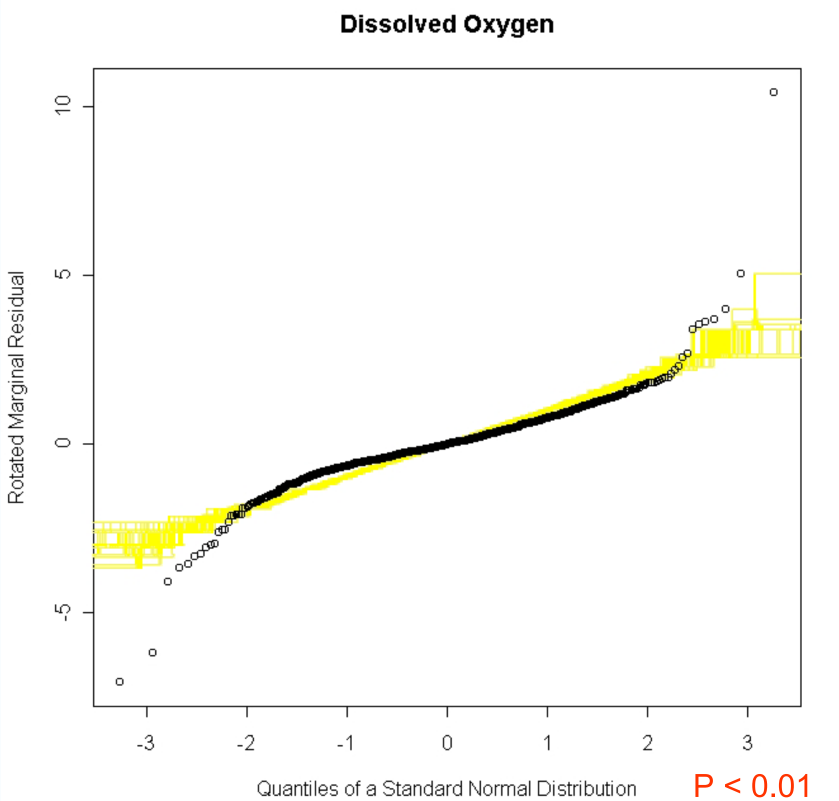
Histogram of Log-Benzene



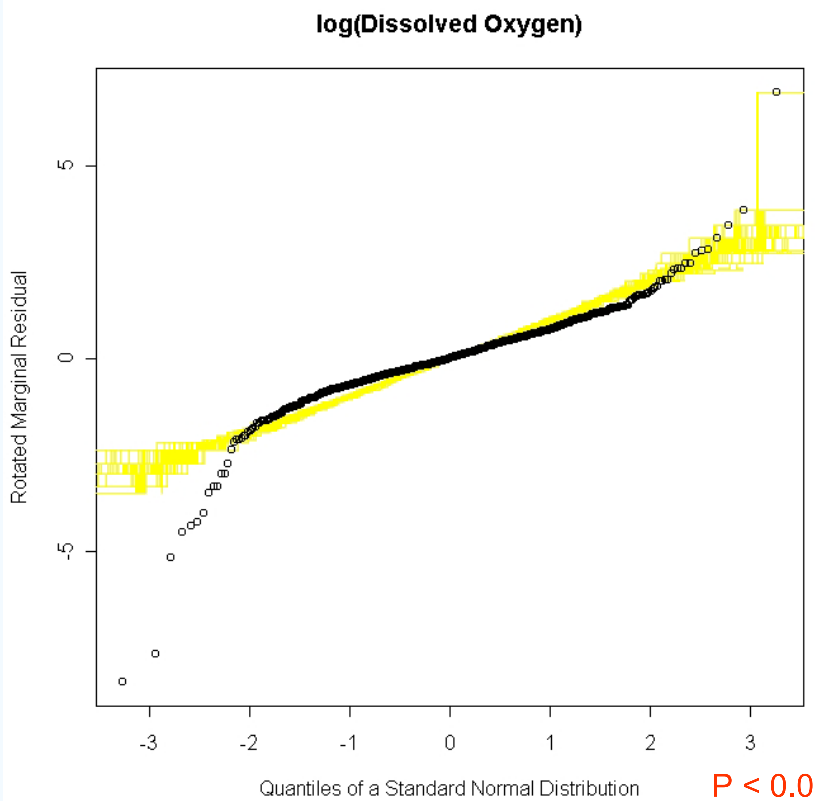
Histogram of Log-Carbon Tetrachloride



QQ Plots for DO Marginal

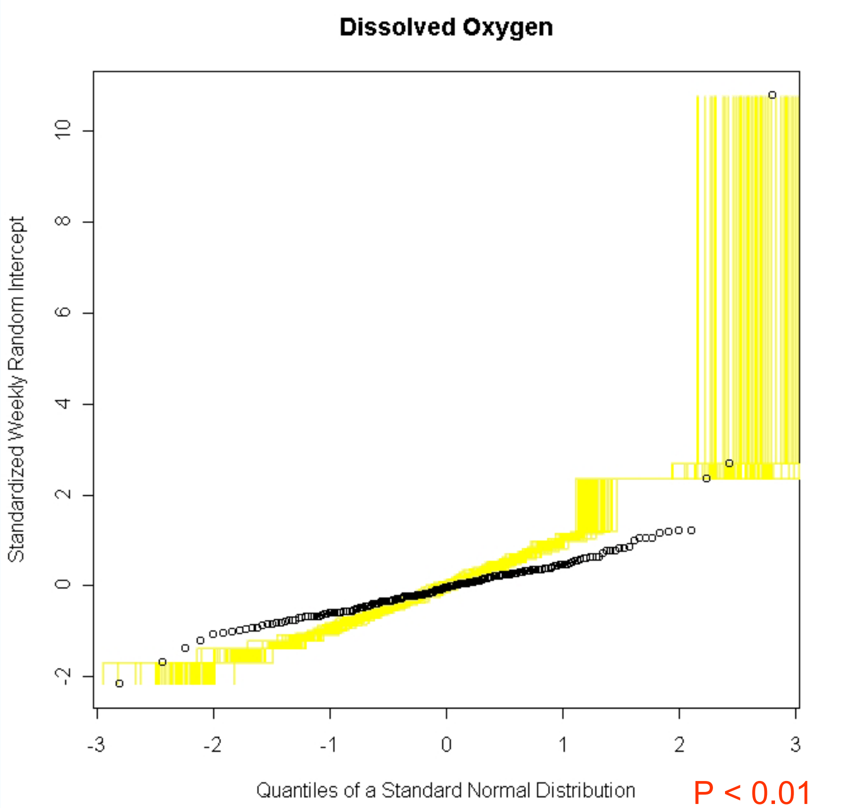


P < 0.01

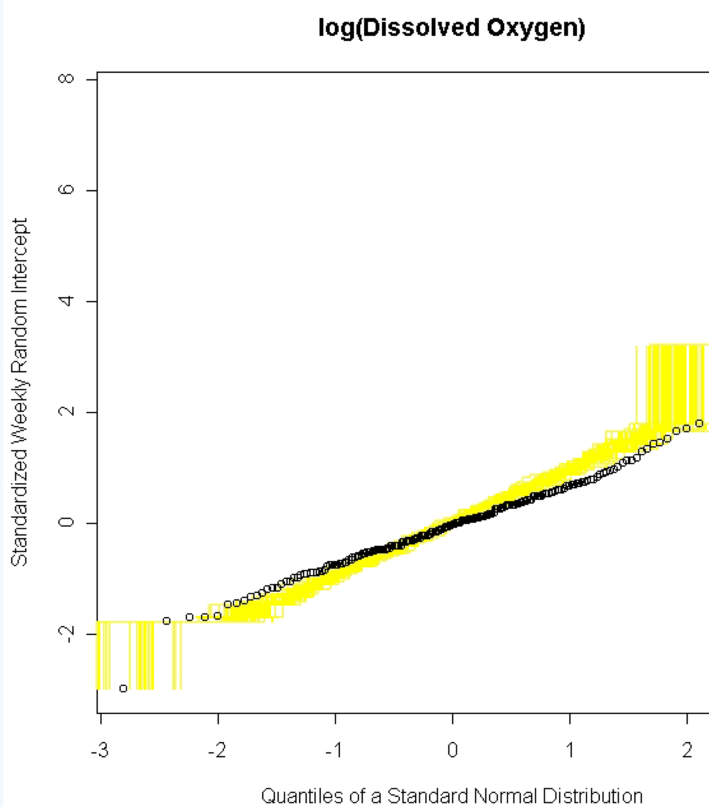


P < 0.01

QQ Plots for DO Random Intercept

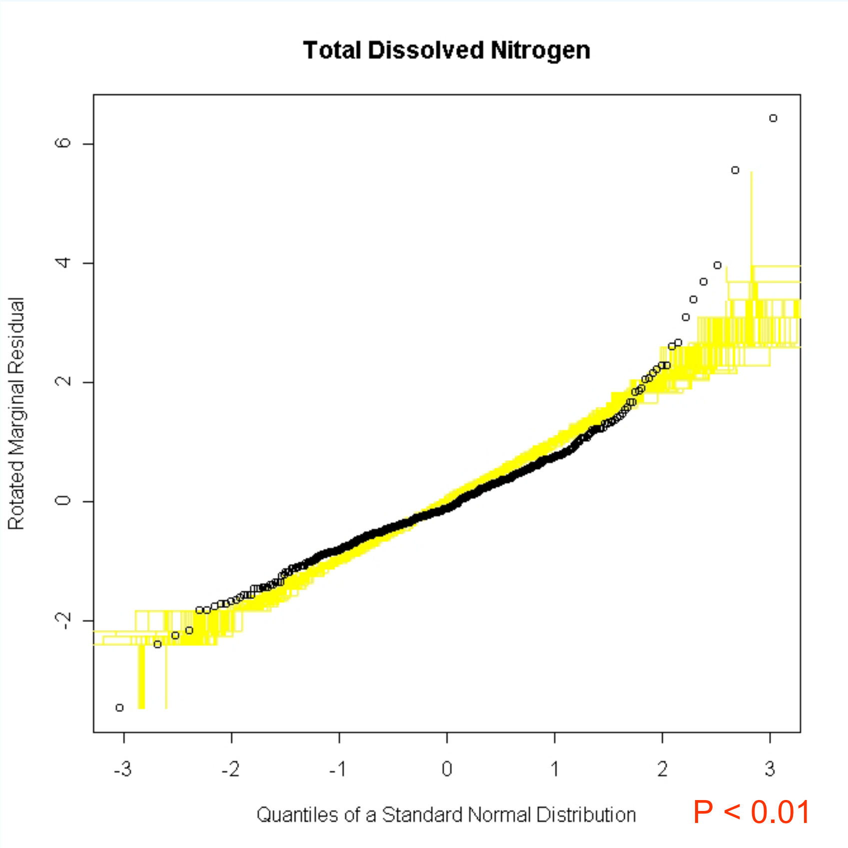


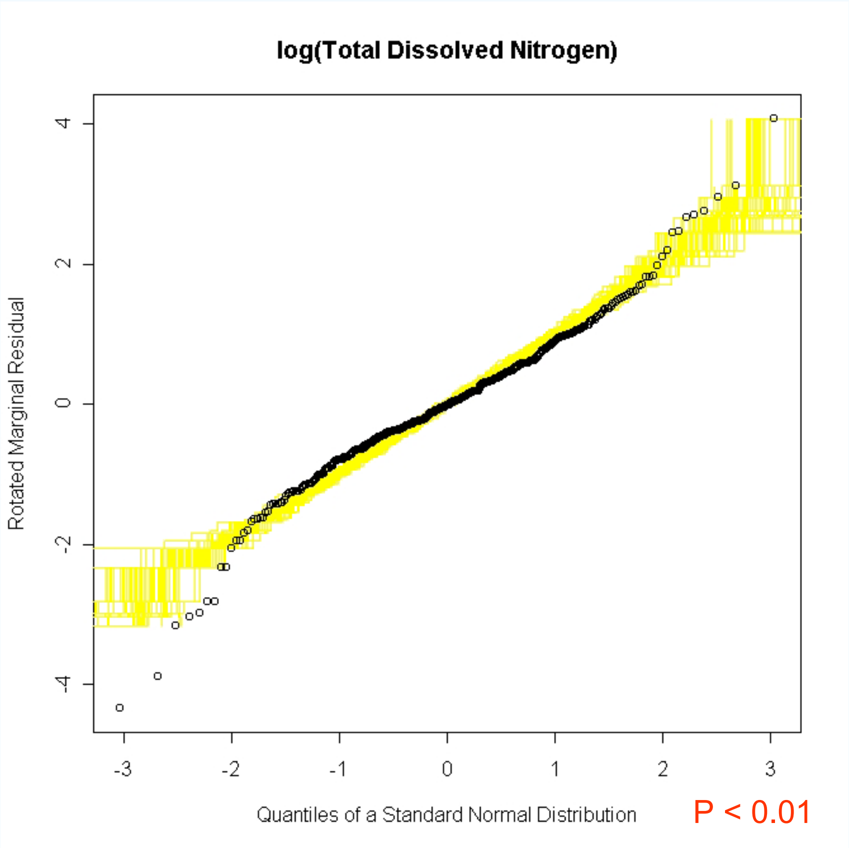
P < 0.01



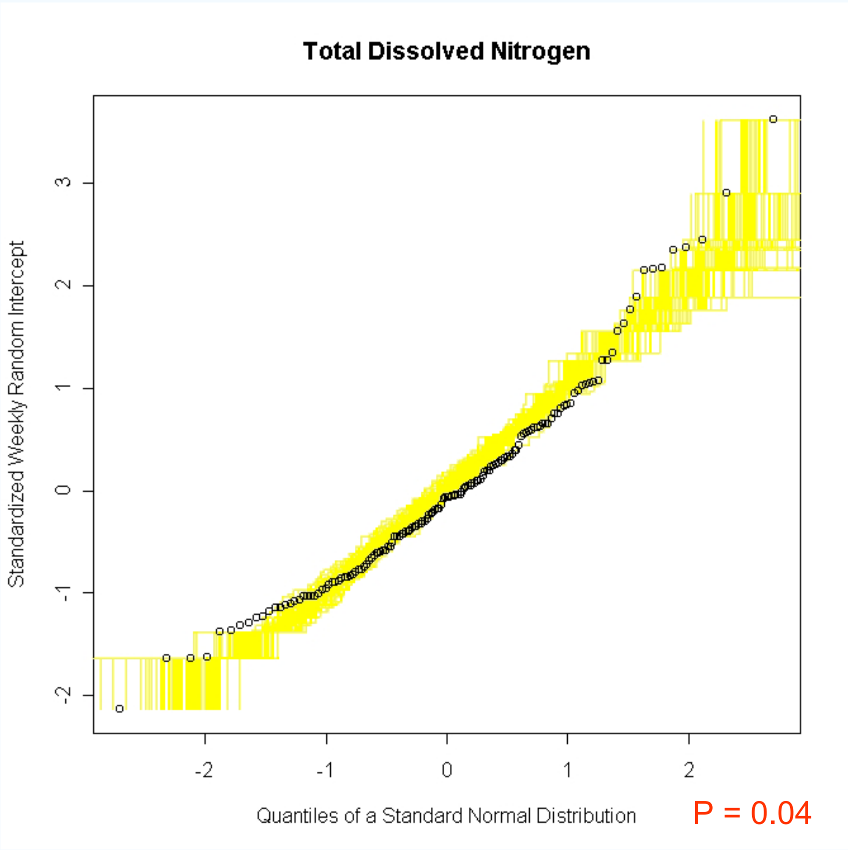
P < 0.01

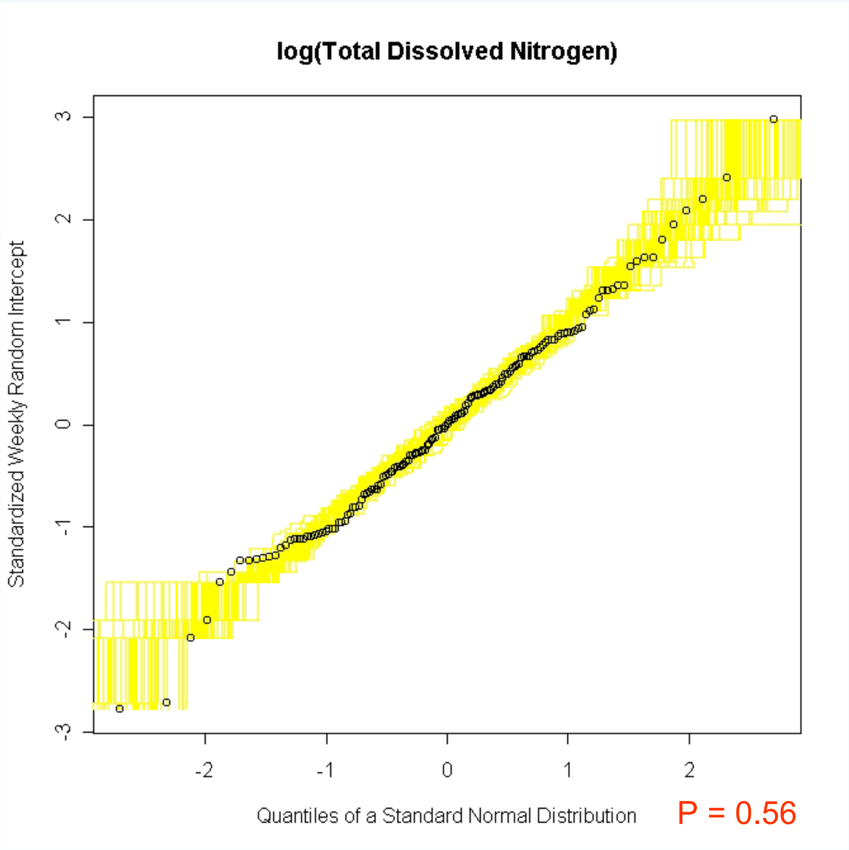
QQ Plots for TDN Marginal





QQ Plots for TDN Random Intercept





Discussion

- Quantifying power to detect specific departures. E.g. how many repeats per subject needed to reliably assess normality of random effects?
- Extensions to GLMMs – use working residuals? Standarized BLUPS?
- Tests targeting particular types of model departures?

And George said …..

The last speaker shall be first to get a glass of wine at dinner….