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

\[ y = X\beta + \varepsilon, \quad \varepsilon \sim N_n(0, \gamma^2 I) \]

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:
  \[
  \hat{F}_n (x) = n^{-1} \sum_{i=1}^{n} I\{z_i \leq x\}
  \]

- If \( z_i \sim N(0,1) \)
  \[
  n^{1/2} \left( \hat{F}_n (x) - \Phi(x) \right) \Rightarrow G \left( 0, \sigma^2(x, y) \right)
  \]
  \[
  \sigma^2(x, y) = \Phi(x)(1 - \Phi(y))
  \]
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:
  \[
  \hat{F}_n(x; \hat{\beta}, \hat{\gamma}) = n^{-1} \sum_{i=1}^{n} I\left\{ z_i(\hat{\beta}, \hat{\gamma}) \leq x \right\}
  \]
- Still asymptotically Gaussian,
  \[
  n^{1/2} \hat{F}_n(x; \hat{\beta}, \hat{\gamma}) \Rightarrow G(\Phi(x), \bar{\sigma}^2(x, y))
  \]
- Changed variance (Ron Randles and others)
General Linear Model

\[ y = X\beta + \varepsilon, \quad \varepsilon \sim N_n(0,V(\gamma)) \]

- 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

\[ y_i = X_i \beta + Z_i b_i + e_i \]

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

\[ y_i = X_i \beta + \varepsilon_i \]

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

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

\[ y = X \beta + \epsilon, \quad \epsilon \sim N_n \left(0, \sigma^2 R(\rho_1, \ldots, \rho_k)\right) \]

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 Exposure Assessment: a Columbia Harvard Project
Crossed Effects Models

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

\[ a_i \sim N\left(0, \sigma_A^2\right) \quad b_j \sim N\left(0, \sigma_B^2\right) \]

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

\[ y = X\beta + \varepsilon \]

\[ \varepsilon \sim N_n\left(0, \sigma_A^2 R_A + \sigma_B^2 R_B + \sigma_0^2 I_n\right) \]
Boston Harbor Data

![Map of Boston Harbor with water quality monitoring stations and treatment plant outfalls.]

- Water quality monitoring station
- Treatment plant outfall

Legend:
- 0-1, 1-2, 2-3, 3-4, 4-5 Miles
- 0-1, 1-2, 2-3, 3-4, 4-5 Kilometers

10 January 2005
Longitudinal workshop, U Florida Gainesville
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>44</td>
<td>5</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>65</td>
<td>52</td>
<td>02</td>
<td>53</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>77</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>106</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>135</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>136</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>141</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Graph:**

Dissolved Oxygen (Bottom, 1995-2000)

- **D.O. (mg/L)**
- **time (years)**
  - 1996
  - 1997
  - 1998
  - 1999
  - 2000

Legend:
- Station 44
- Station 65
- Station 77
- Station 106
- Station 135
- Station 136
- Station 141

Graph shows the dissolved oxygen levels at various stations from 1995 to 2000, with fluctuating trends over the years.
Total Dissolved Nitrogen
(Surface, 1995-2002)
Space/Time model?

\[ y = X \beta + \varepsilon, \]

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

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

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

\[ Z_i^{(j)} = \frac{\hat{b}_{(j)}^{(i)}}{sd(\hat{b}_{(j)}^{(i)})} \]

Pointwise Asymptotics of ECDF of the Z’s:

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

\[ \tilde{\sigma}_x^2 = \Phi(x)(1 - \Phi(x)) - \delta^T W \delta \]

\( W = \) Covariance of estimated parameters, \( \delta \) a gradient vector (more presently)
Lots of gaps....

• Global Asymptotics?
• Non-clustered data?
• Other diagnostics?
• General residuals?
Cholesky Rotated Residuals

Recall:

\[ y = X\beta + \varepsilon, \quad \varepsilon \sim N_n(0, V(\gamma)) \]

Define:

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

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

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

\[(P_1, \ldots, P_N), \text{ where } P_i P_j^T = I(i = j)\]

\[z_i(\beta, \gamma) = P_i L(\gamma)^T (\gamma - X\beta)\]

\[
z_i(\beta_0, \gamma_0) \overset{iid}{\sim} N(0,1)
\]

Lange/Ryan standardized BLUPs a special case
Pointwise Asymptotics

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

where

\[ \theta = (\beta, \gamma)^T \]

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

\[ W = \text{Var}(\hat{\theta}) \]
Pointwise Asymptotics (2)

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

\[ m_i(\theta, \theta_0) = P_iL(\gamma)^T X(\beta_0 - \beta) \]

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

Closed form exists for \( \delta_x \) estimate
Global Asymptotics

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

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

\[ F^* (x) = \sum_{i=1}^{N} I \{ P_i z^* \leq x \} + \delta_x^T J^{-1} U \left( \hat{\theta}; z^* \right) \]

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

\[ U \left( \hat{\theta}; L(\hat{\gamma})^T (y - X\hat{\beta}) \right) = 0 \]

\[ J = \frac{\partial U}{\partial \theta} \bigg|_{\theta = \hat{\theta}} \]
Resampling

Under $H_0$: normal

\[ F^*(x \mid y) \equiv \hat{F}_n (x; \hat{\theta}) \]

\[ \Gamma(F^*) \equiv \Gamma(\hat{F}_n) \]

\[ \Gamma(F) = \sup_{x \in \Omega} |F(x) - \Phi(x)| \]

\[ P - \text{value} \approx M^{-1} \sum_{u=1}^{M} I(\Gamma(F^*_u) > \Gamma(\hat{F}_n)) \]
Untransformed Pollen Counts

Marginal Residual

P < 0.01
$\sqrt{\cdot}$-Transformed Pollen Counts

**Marginal Residual**

P = 0.06
Simulation – time series

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

<table>
<thead>
<tr>
<th>Error distributions</th>
<th>KS</th>
<th>CVM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Normal</td>
<td>.05</td>
<td>.04</td>
</tr>
<tr>
<td>Skewed (chisq, 3 df)</td>
<td>.82</td>
<td>.90</td>
</tr>
<tr>
<td>Heavy tailed (t, 3df)</td>
<td>.68</td>
<td>.79</td>
</tr>
</tbody>
</table>
Pig Weights: Marginal Errors

P = 0.13
Pig Weights: Random Intercept

Random Intercept

P = 0.11
Pig Weights: Random Slope

$P = 0.02$
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
Log-Carbon Tetrachloride

Marginal Residual

P < 0.01
Histogram of Log-Benzene
Histogram of Log-Carbon Tetrachloride
QQ Plots for DO
Marginal

Dissolved Oxygen

log(Dissolved Oxygen)

Quantiles of a Standard Normal Distribution

P < 0.01

Quantiles of a Standard Normal Distribution

P < 0.01
QQ Plots for DO Random Intercept

Dissolved Oxygen

log(Dissolved Oxygen)

P < 0.01
QQ Plots for TDN Marginal

Total Dissolved Nitrogen

$P < 0.01$

Log(Total Dissolved Nitrogen)

$P < 0.01$
QQ Plots for TDN
Random Intercept

Total Dissolved Nitrogen

P = 0.04

log(Total Dissolved Nitrogen)

P = 0.56
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....