Global Validation of Linear Model Assumptions

Edsel A. Peña*, Elizabeth H. Slate

Department of Statistics, University of South Carolina

Department of Biometry and Epidemiology, Medical University of South Carolina

Research support from NIH, NSF, NIH-COBRE, and USC/MUSC Collaborative Grant.

pena@stat.sc.edu, SlateEH@musc.edu

Linear Model and Assumptions

- Linear Model (LM):

$$\mathbf{Y} = \mathbf{X}\beta + \sigma\epsilon$$

- $\mathbf{Y}$ = observable $n \times 1$ response vector;
- $\mathbf{X}$ = observable $n \times p$ design matrix;
- $\epsilon =$ unobservable error vector;
- β and σ are the parameters.

Linear Model and Assumptions

- Linear Model (LM):

$$\mathbf{Y} = \mathbf{X}\beta + \sigma\epsilon$$

- $\mathbf{Y}$ = observable $n \times 1$ response vector;
- $\mathbf{X}$ = observable $n \times p$ design matrix;
- $\epsilon =$ unobservable error vector;
- β and σ are the parameters.

(A1) *Linearity*:

$$\mathbf{E}\{Y_i|\mathbf{X}\} = \mathbf{x}_i\beta$$

(A2) *Homoscedasticity*:

$$\mathbf{Var}\{Y_i|\mathbf{X}\} = \sigma^2$$

(A3) *Uncorrelatedness*:

$$\mathbf{Cov}\{Y_i, Y_j|\mathbf{X}\} = 0$$

(A4) *Normality*:

$$Y_i|\mathbf{X} \sim \text{Normal}.$$

Estimators

- Estimator of β:

$$\mathbf{b} = \hat{\beta} = (\mathbf{X}^{\mathrm{t}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{t}}\mathbf{Y};$$

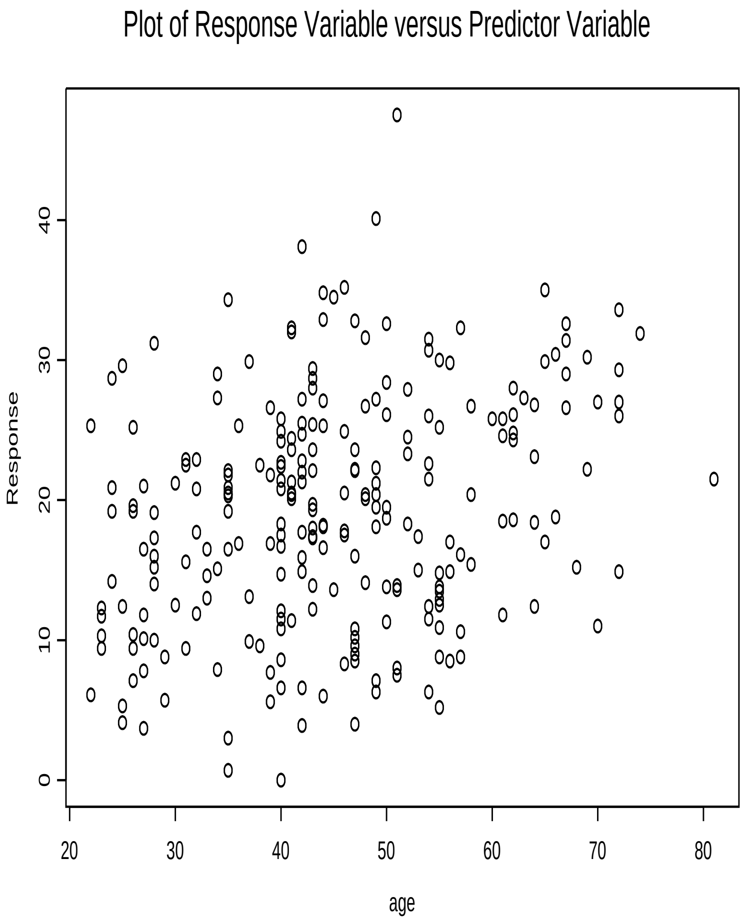
- Estimator of σ^2:

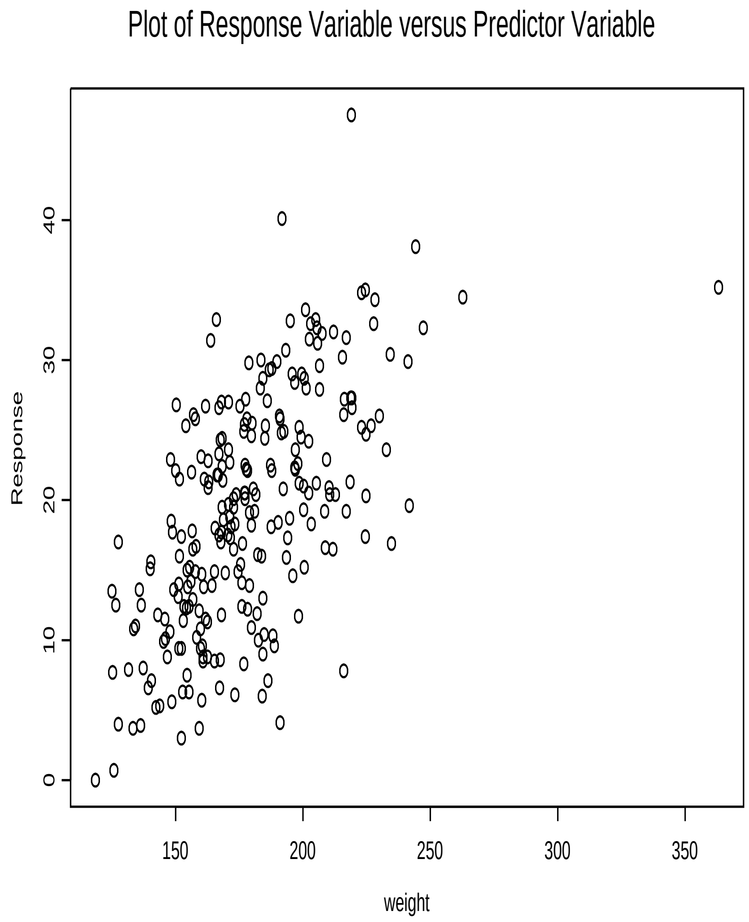
$$s^2 = \hat{\sigma}^2 = \frac{1}{n}\mathbf{Y}^{\mathrm{t}}(\mathbf{I} - \mathbf{P}_{\mathbf{X}})\mathbf{Y},$$

- Projection operator on the linear subspace generated by the columns of $\mathbf{X}$, also denoted by $\mathbf{H}$:

$$\mathbf{P}_{\mathbf{X}} = \mathbf{X}(\mathbf{X}^{\mathrm{t}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{t}}$$

Example: Body Fat Data Set





Example: Fitting LM

- Response: $Y =$ Body fat content.
- Predictors: $X_1 =$ Age; $X_2 =$ Weight.
- Model: $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \sigma\epsilon_i$
- Results of Fitting Model (Using `lm` in `S-Plus`):
- Coefficients: $b_0 = -21.16(se = 2.77, p = 0)$, $b_1 = .20(se = .03, p = 0)$, $b_2 = .18(se = .01, p = .01)$.
- Residual SE: 6.148 on 249 DF. Multiple R^2: 0.4646. F-statistic: 108 on (2, 249) DF. p-value = 0.

Validating LM Assumptions

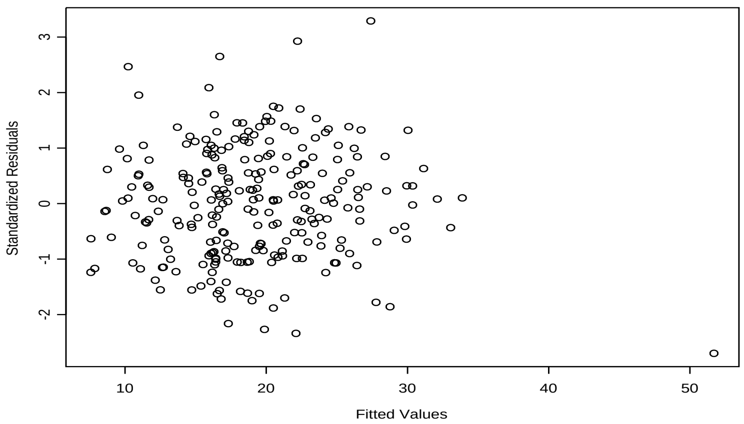
- *Standardized* Residuals:

$$\mathbf{R} = \frac{\mathbf{Y} - \mathbf{Xb}}{s} = \frac{(\mathbf{I} - \mathbf{P_X})\mathbf{Y}}{s}$$

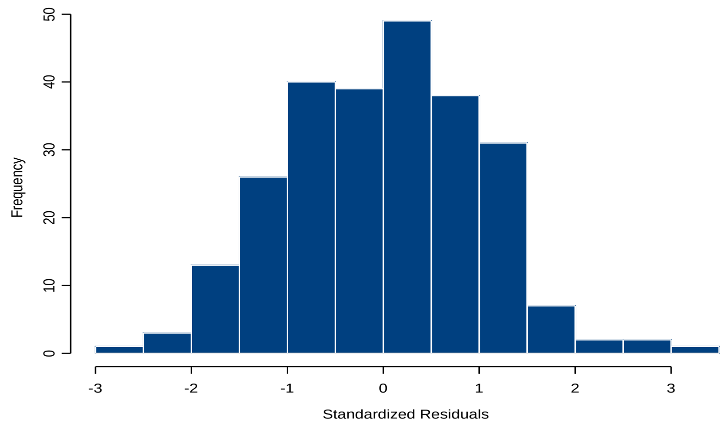
- Graphical Methods.
- Diagnostic plots based on $\mathbf{R}$. Discussed in many (elementary) textbooks!
- Formal tests.
- Such formal hypothesis tests are based on $\mathbf{R}$.

Example: Body Fat Diagnostics

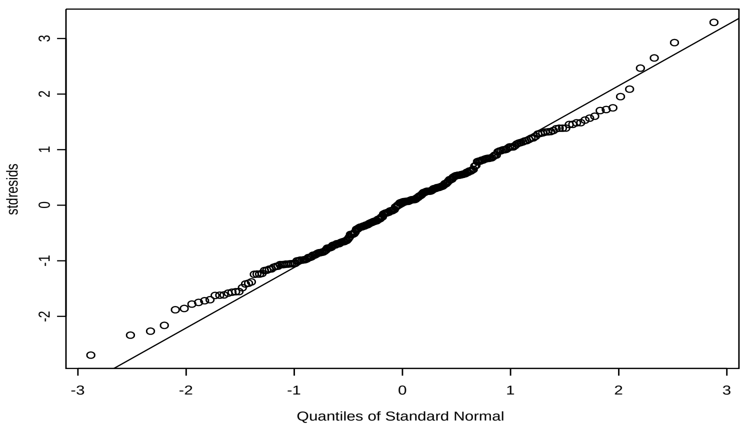
Plot of the Fitted Values versus the Standardized Residuals



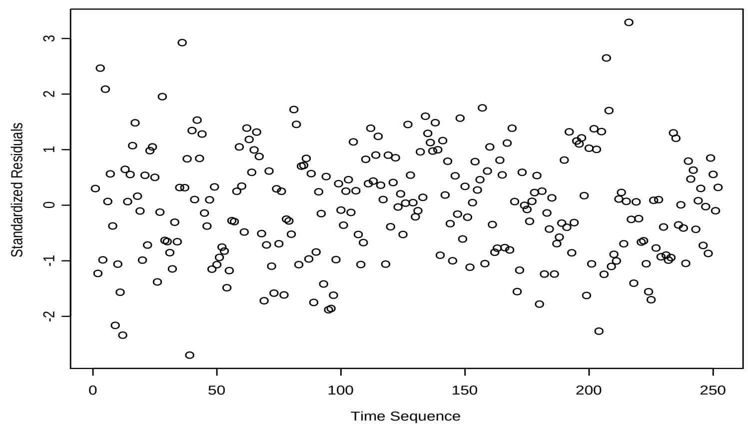
Histogram of the Standardized Residuals



Normal Probability Plot of the Standardized Residuals (with line)



Plot of the Standardized Residuals versus Time Sequence



Issues to Consider

Issues to Consider

- Varied plots to detect varied assumptions. Made easy by statistical packages.

Issues to Consider

- Varied plots to detect varied assumptions. Made easy by statistical packages.
- *“A picture is worth a thousand words, but beauty is always in the eye of the beholder!”*

Issues to Consider

- Varied plots to detect varied assumptions. Made easy by statistical packages.
- *"A picture is worth a thousand words, but beauty is always in the eye of the beholder!"*
- Re-use of data. Parameter estimates are substituted for unknown parameters to obtain $\mathbf{R}$.

Issues to Consider

- Varied plots to detect varied assumptions. Made easy by statistical packages.
- *"A picture is worth a thousand words, but beauty is always in the eye of the beholder!"*
- Re-use of data. Parameter estimates are substituted for unknown parameters to obtain $\mathbf{R}$.
- Formal tests are usually specific to type of departure from assumptions (e.g., Tukey's test for additivity; Durbin and Watson's test for serial correlation; test for normality; tests for heterogeneity of variances).

Problem and Goals

- Based on $(\mathbf{Y}, \mathbf{X})$, to test formally and globally the hypotheses

$$\begin{aligned} H_0 &: \text{ Assumptions (A1)-(A4) all hold;} \\ H_1 &: \text{ At least one of (A1)-(A4) does not hold.} \end{aligned}$$

Problem and Goals

- Based on $(\mathbf{Y}, \mathbf{X})$, to test formally and globally the hypotheses

$$H_0: \text{ Assumptions (A1)-(A4) all hold;}$$
$$H_1: \text{ At least one of (A1)-(A4) does not hold.}$$

- To detect formally the type of departure from the assumptions if the global test decides that a violation has occurred.

Problem and Goals

- Based on $(\mathbf{Y}, \mathbf{X})$, to test formally and globally the hypotheses

$$
\begin{aligned}
H_0 &: \text{ Assumptions (A1)-(A4) all hold;} \\
H_1 &: \text{ At least one of (A1)-(A4) does not hold.}
\end{aligned}
$$

- To detect formally the type of departure from the assumptions if the global test decides that a violation has occurred.
- Objectivity of conclusions and control of probability of error desired.

1st and 2nd Component Statistics

Recalling the standardized residuals

$$R_i = \frac{Y_i - \hat{Y}_i}{s}, \; i = 1, 2, \ldots, n,$$

where $\hat{Y}_i = \mathbf{x}_i \mathbf{b}$ is the ith fitted or predicted value.

$$\hat{S}_1^2 = \left\{ \frac{1}{\sqrt{6n}} \sum_{i=1}^{n} R_i^3 \right\}^2; \quad \hat{S}_2^2 = \left\{ \frac{1}{\sqrt{24n}} \sum_{i=1}^{n} [R_i^4 - 3] \right\}^2;$$

3rd Component Statistic

$$\hat{S}_3^2 = \frac{\left\{\frac{1}{\sqrt{n}}\sum_{i=1}^n(\hat{Y}_i - \bar{Y})^2 R_i\right\}^2}{(\hat{\Omega} - \mathbf{b}^{\mathrm{t}}\hat{\boldsymbol{\Sigma}}_X\mathbf{b} - \hat{\Gamma}\hat{\Sigma}_X^{-1}\hat{\Gamma}^{\mathrm{t}})},$$

$$\hat{\Omega} = \frac{1}{n}\sum_{i=1}^n(\hat{Y}_i - \bar{Y})^4; \quad \hat{\Sigma}_X = \frac{1}{n}\sum_{i=1}^n(\mathbf{x}_i - \bar{\mathbf{x}})^{\mathrm{t}}(\mathbf{x}_i - \bar{\mathbf{x}})$$

$$\hat{\Gamma} = \frac{1}{n}\sum_{i=1}^n(\hat{Y}_i - \bar{Y})^2(\mathbf{x}_i - \bar{\mathbf{x}}).$$

4th Component Statistic

The fourth component statistic requires a user-supplied $n \times 1$ vector $\mathbf{V}$, which by default is set to be the time sequence $\mathbf{V} = (1, 2, \ldots, n)^{\mathrm{t}}$. It is defined via

$$\hat{S}_4^2 = \left\{ \frac{1}{\sqrt{2\hat{\sigma}_V^2 n}} \sum_{i=1}^{n} (V_i - \bar{V})(R_i^2 - 1) \right\}^2,$$

with

$$\hat{\sigma}_V^2 = \frac{1}{n} \sum_{i=1}^{n} (V_i - \bar{V})^2.$$

Global Statistic and Test

- The global test statistic is

$$\hat{G}_4^2 = \hat{S}_1^2 + \hat{S}_2^2 + \hat{S}_3^2 + \hat{S}_4^2.$$

- For large n, a global test of H_0 versus H_1 at asymptotic level α is:

$$\text{Reject } H_0 \text{ if } \hat{G}_4^2 > \chi^2_{4;\alpha},$$

where $\chi^2_{k;\alpha}$ is the $100(1-\alpha)$th percentile of a central chi-squared distribution with degrees-of-freedom k.

Directional Tests

If the global test rejects H_0, type of violation could be detected via:

Directional Tests

If the global test rejects H_0, type of violation could be detected via:

- Skewed error distributions indicated by $\hat{S}_1^2$;

Directional Tests

If the global test rejects H_0, type of violation could be detected via:

- Skewed error distributions indicated by $\hat{S}_1^2$;
- Deviations from the normal distribution kurtosis of the true error distribution generally revealed by $\hat{S}_2^2$;

Directional Tests

If the global test rejects H_0, type of violation could be detected via:

- Skewed error distributions indicated by $\hat{S}_1^2$;
- Deviations from the normal distribution kurtosis of the true error distribution generally revealed by $\hat{S}_2^2$;
- Misspecified link function or the absence of other predictor variables in the model detected by $\hat{S}_3^2$;

Directional Tests

If the global test rejects H_0, type of violation could be detected via:

- Skewed error distributions indicated by $\hat{S}_1^2$;
- Deviations from the normal distribution kurtosis of the true error distribution generally revealed by $\hat{S}_2^2$;
- Misspecified link function or the absence of other predictor variables in the model detected by $\hat{S}_3^2$;
- Presence of heteroscedastic errors and/or dependent errors manifested by $\hat{S}_4^2$; and

Directional Tests

If the global test rejects H_0, type of violation could be detected via:

- Skewed error distributions indicated by $\hat{S}_1^2$;
- Deviations from the normal distribution kurtosis of the true error distribution generally revealed by $\hat{S}_2^2$;
- Misspecified link function or the absence of other predictor variables in the model detected by $\hat{S}_3^2$;
- Presence of heteroscedastic errors and/or dependent errors manifested by $\hat{S}_4^2$; and
- Simultaneous violations revealed by large values of several component statistics.

Global Deletion Statistic

$$\Delta \hat{G}_4^2[i] = \left[\frac{\hat{G}_4^2[i] - \hat{G}_4^2}{\hat{G}_4^2} \right] \times 100, \quad i = 1, 2, \ldots, n.$$

- Percent relative change in value of global statistic $\hat{G}_4^2$ after deletion of ith observation.
- Idea: observation with a large absolute value of $\Delta \hat{G}_4^2[i]$ is either an outlier or has large influence.
- Values of $\Delta \hat{G}_4^2[i]$ can be plotted with respect to time sequence to assess their relative values.

Example: For the Body Fat Data

- Global Test: $\hat{G}_4^2 = 10.15 \; (p = 0.037)$; Decision: Assumptions NOT satisfied!

Example: For the Body Fat Data

- Global Test: $\hat{G}_4^2 = 10.15 \ (p = 0.037)$; Decision: Assumptions NOT satisfied!

- Component Statistics (with p-Value and Decision)
  - $\hat{S}_1 = 0.91 \ (p = 0.33)$; Decision: OK.
  - $\hat{S}_2 = 0.00 \ (p = 0.98)$; Decision: OK.
  - $\hat{S}_3 = 6.89 \ (p = 0.01)$; Decision: Violation!
  - $\hat{S}_4 = 2.33 \ (p = 0.12)$; Decision: OK.

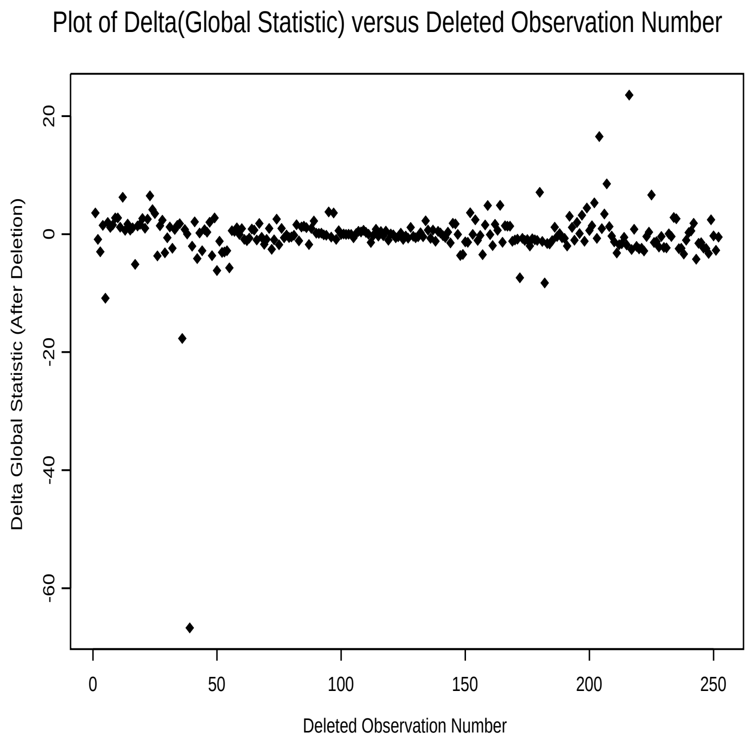
Example: For the Body Fat Data

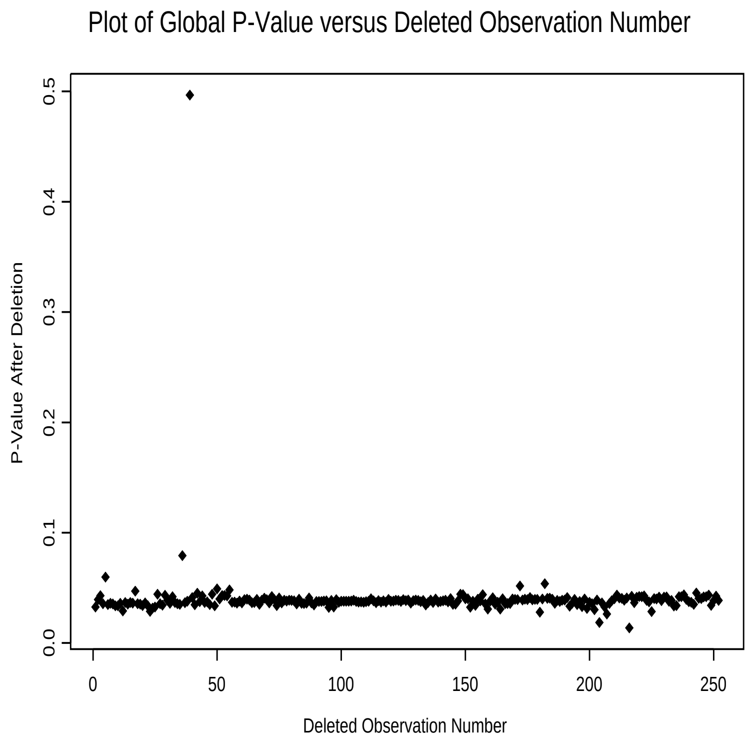
- Global Test: $\hat{G}_4^2 = 10.15 \ (p = 0.037)$; Decision: Assumptions NOT satisfied!

- Component Statistics (with p-Value and Decision)
  - $\hat{S}_1 = 0.91 \ (p = 0.33)$; Decision: OK.
  - $\hat{S}_2 = 0.00 \ (p = 0.98)$; Decision: OK.
  - $\hat{S}_3 = 6.89 \ (p = 0.01)$; Decision: Violation!
  - $\hat{S}_4 = 2.33 \ (p = 0.12)$; Decision: OK.

- Based on the directional tests, the violation appears to be in the link function.

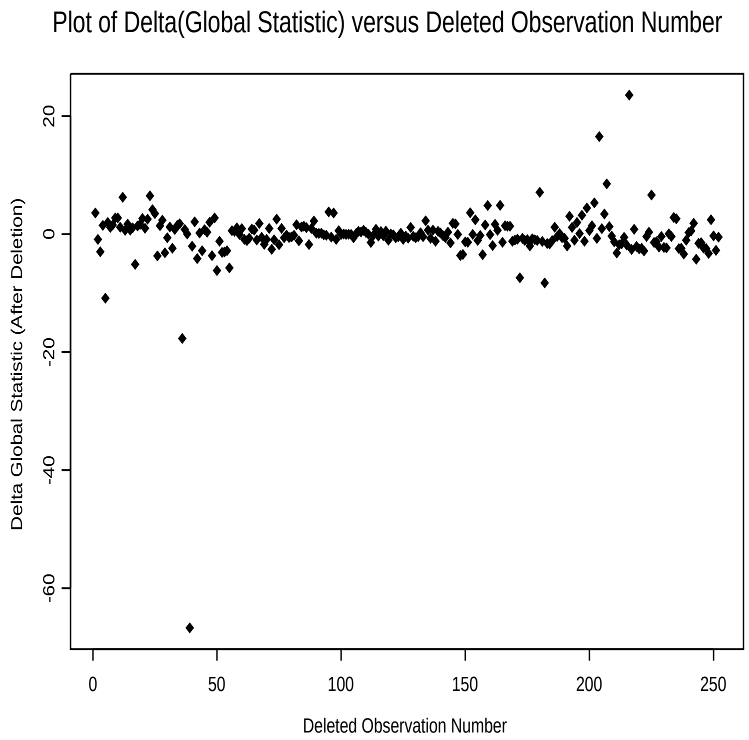
Example: Deletion Statistics

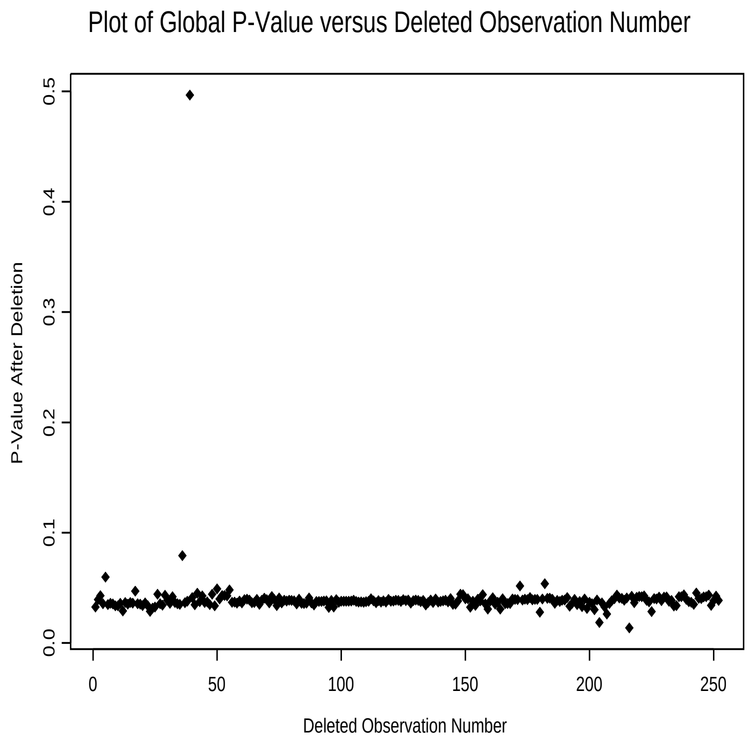




Result: The 39th obs. suspect. Has $\Delta\hat{G}_4^2[39] = -66.73$.

Example: Deletion Statistics





Result: The 39th obs. suspect. Has $\Delta \hat{G}_4^2[39] = -66.73$.

Remark: After deleting the 39th obs: $\hat{G}_4^2 = 3.37 (P = 0.49)$. LM assumptions now acceptable.

Theoretical Interludes

- True Residuals:

$$\mathbf{R}^0 \equiv \mathbf{R}^0(\sigma^2, \beta) = \frac{\mathbf{Y} - \mathbf{X}\beta}{\sigma}$$

- $\mathbf{R}^0$ are iid std normals.
- Density under H_0 of $\mathbf{R}^0$:

$$f_{\mathbf{R}^0}(\mathbf{r}^0) = \prod_{i=1}^{n} \phi(r_i^0)$$

- $\phi(\cdot) =$ std normal pdf.

Theoretical Interludes

- True Residuals:

$$\mathbf{R}^0 \equiv \mathbf{R}^0(\sigma^2, \beta) = \frac{\mathbf{Y} - \mathbf{X}\beta}{\sigma}$$

- $\mathbf{R}^0$ are iid std normals.
- Density under H_0 of $\mathbf{R}^0$:

$$f_{\mathbf{R}^0}(\mathbf{r}^0) = \prod_{i=1}^{n} \phi(r_i^0)$$

- $\phi(\cdot) =$ std normal pdf.

- Embedding Class:

$$f_{\mathbf{R}^0}(\mathbf{r}^0|\theta) = C(\theta) f_{\mathbf{R}^0}(\mathbf{r}^0) \exp\{\theta^{\mathrm{t}} \mathbf{Q}(\mathbf{r}^0)\}$$

$$\mathbf{Q}(\mathbf{r}^0) = \sum_{i=1}^{n} \begin{bmatrix} r_i^0 \\ (r_i^0)^2 - 1 \\ (r_i^0)^3 \\ (r_i^0)^4 - 3 \\ \{(\mathbf{x}_i - \bar{\mathbf{x}})\beta\}^2 r_i^0 \\ (v_i - \bar{v})[(r_i^0)^2 - 1] \end{bmatrix}$$

Score Test Statistic

- The score test statistic within this embedding class for $H_0 : \theta = \mathbf{0}$ versus $H_1 : \theta \neq \mathbf{0}$ when β and σ are known is:

$$\mathbf{U}(\theta = \mathbf{0}, \sigma^2, \beta) = \mathbf{Q}(\mathbf{R}^0; \sigma^2, \beta).$$

Score Test Statistic

- The score test statistic within this embedding class for $H_0 : \theta = \mathbf{0}$ versus $H_1 : \theta \neq \mathbf{0}$ when β and σ are known is:

$$\mathbf{U}(\theta = \mathbf{0}, \sigma^2, \beta) = \mathbf{Q}(\mathbf{R}^0; \sigma^2, \beta).$$

- When the parameters are not known, then the score statistic is:

$$\mathbf{U}(\theta = \mathbf{0}, s^2, \mathbf{b}) = \mathbf{Q}(\mathbf{R}; s^2, \mathbf{b}).$$

Score Test Statistic

- The score test statistic within this embedding class for $H_0 : \theta = \mathbf{0}$ versus $H_1 : \theta \neq \mathbf{0}$ when β and σ are known is:

$$\mathbf{U}(\theta = \mathbf{0}, \sigma^2, \beta) = \mathbf{Q}(\mathbf{R}^0; \sigma^2, \beta).$$

- When the parameters are not known, then the score statistic is:

$$\mathbf{U}(\theta = \mathbf{0}, s^2, \mathbf{b}) = \mathbf{Q}(\mathbf{R}; s^2, \mathbf{b}).$$

- Needed: null asymptotic distribution of

$$\mathbf{Q}(\mathbf{R}; s^2, \mathbf{b}).$$

Asymptotics: Parameters Known

Under H_0 : $\frac{1}{\sqrt{n}}\mathbf{Q}(\mathbf{R}^0;\sigma^2,\beta) \xrightarrow{\mathrm{d}} N\left(\mathbf{0},\mathbf{\Sigma}_{11}(\sigma^2,\beta)\right)$

$$\mathbf{\Sigma}_{11}(\sigma^2,\beta) = \begin{bmatrix} 1 & 0 & 3 & 0 & \beta^{\mathrm{t}}\mathbf{\Sigma}_X\beta & 0 \\ 0 & 2 & 0 & 12 & 0 & 0 \\ 3 & 0 & 15 & 0 & 3\beta^{\mathrm{t}}\mathbf{\Sigma}_X\beta & 0 \\ 0 & 12 & 0 & 96 & 0 & 0 \\ \beta^{\mathrm{t}}\mathbf{\Sigma}_X\beta & 0 & 3\beta^{\mathrm{t}}\mathbf{\Sigma}_X\beta & 0 & \Omega(\beta) & 0 \\ 0 & 0 & 0 & 0 & 0 & 2\sigma_V^2 \end{bmatrix}$$

Asymptotics: Parameters Estimated

Under H_0: $\frac{1}{\sqrt{n}}\mathbf{Q}(\mathbf{R}; s^2, \mathbf{b}) \xrightarrow{\mathrm{d}} N\left(\mathbf{0}, \mathbf{\Xi}_{11.2}(\sigma^2, \beta)\right)$

$$\mathbf{\Xi}_{11.2}(\sigma^2, \beta) = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 6 & 0 & 0 & 0 \\ 0 & 0 & 0 & 24 & 0 & 0 \\ 0 & 0 & 0 & 0 & \xi(\sigma^2, \beta) & 0 \\ 0 & 0 & 0 & 0 & 0 & 2\sigma_V^2 \end{bmatrix}$$

$$\xi(\sigma^2, \beta) = \Omega(\beta) - (\beta^{\mathrm{t}}\mathbf{\Sigma}_X\beta)^2 - \mathbf{\Gamma}(\beta)\mathbf{\Sigma}_X^{-1}\mathbf{\Gamma}(\beta)^{\mathrm{t}}$$

Global Test Statistic

- The test statistic

$$\frac{1}{n}\mathbf{Q}(\mathbf{R}; s^2, \mathbf{b})^{\mathrm{t}}\hat{\boldsymbol{\Xi}}_{11.2}^{-}\mathbf{Q}(\mathbf{R}; s^2, \mathbf{b}) = \hat{S}_1^2 + \hat{S}_2^2 + \hat{S}_3^2 + \hat{S}_4^2 = \hat{G}_4^2$$

  converges in distribution, under H_0, to a four degrees-of-freedom chi-squared random variable.
- This is the justification for the global test procedure, and this test is a score test within the embedding class!
- The estimators of the variances are their natural consistent estimators.

Monte Carlo Adventures

- Goals: to ascertain level and powers of the test procedure for testing the four LM assumptions.
- $n \in \{30, 100, 200\}$
- 2000 replications
- $x_1, x_2, \ldots, x_n$ standard uniform
- Fitted Model: $Y_i = \beta_0 + \beta_1 x_i + \sigma\epsilon_i$
- User-supplied $V = (1, 2, \ldots, n)$
- Level of significance: 5%
- Program implementing the procedure were in `S-Plus` code

Achieved Levels

Model	n	Component Statistics				Global
		$\hat{S}_1^2$	$\hat{S}_2^2$	$\hat{S}_3^2$	$\hat{S}_4^2$	$\hat{G}_4^2$
True	30	4.00	4.00	5.05	5.75	5.10
	100	5.50	4.20	4.35	4.70	5.95
	200	5.70	4.60	4.40	4.05	5.75

Conclusion: The global and directional tests achieve the desired level for the sample size examined in the simulation.

Errors (ϵ_i's): Non-Normal but Symmetric

Error Dist.	n	Component Statistics				Global
		$\hat{S}_1^2$	$\hat{S}_2^2$	$\hat{S}_3^2$	$\hat{S}_4^2$	$\hat{G}_4^2$
t_5	30	18.45	20.35	4.70	9.60	22.40
	100	34.30	57.00	4.40	13.80	54.55
	200	42.25	83.10	4.55	15.15	80.50
Logistic	30	11.80	12.60	5.90	7.20	15.05
	100	17.45	30.30	5.50	8.20	29.35
	200	20.10	52.35	4.25	9.00	47.10
Double Exp.	30	19.50	24.75	5.60	10.35	27.20
	100	35.05	73.55	5.60	14.60	70.65
	200	39.45	95.95	6.35	14.05	92.90

Errors (ϵ_i's): Non-Normal and Skewed

Error Dist	n	Component Statistics				Global
		$\hat{S}_1^2$	$\hat{S}_2^2$	$\hat{S}_3^2$	$\hat{S}_4^2$	$\hat{G}_4^2$
$\chi_1^2 - 1$	30	91.30	59.70	5.25	21.30	80.20
	100	100	98.35	5.05	31.35	100
	200	100	99.95	4.85	33.60	100
$\chi_5^2 - 5$	30	37.15	18.05	4.45	8.65	29.15
	100	96.90	54.25	4.60	11.80	87.70
	200	100	79.40	4.40	13.15	99.90
$\chi_{10}^2 - 10$	30	22.40	12.90	4.75	6.90	18.75
	100	79.70	31.50	4.95	8.60	61.00
	200	98.90	50.00	4.60	8.80	94.70

Model: $Y_i = x_i + x_i^{\gamma}\epsilon_i$ (Heteroscedastic)

Value of γ	Sample Size (n)	Component Statistics				Global
		$\hat{S}_1^2$	$\hat{S}_2^2$	$\hat{S}_3^2$	$\hat{S}_4^2$	$\hat{G}_4^2$
$.5$	30	8.50	10.85	5.65	6.10	14.15
	100	12.00	37.40	4.70	5.55	31.55
	200	10.65	48.85	4.80	8.35	39.50
1	30	11.10	16.65	3.40	6.20	16.15
	100	21.35	78.15	5.15	21.05	72.30
	200	24.10	96.95	5.40	13.35	93.30
2	30	21.40	52.35	6.15	12.60	46.75
	100	34.10	98.95	7.00	10.05	96.45
	200	47.50	100	7.55	31.60	100

Model: $Y_i = x_i + \beta_2 x_i^{\gamma} + \epsilon_i$

Value of (β_2, γ)	Sample Size (n)	Component Statistics $\hat{S}_1^2$	$\hat{S}_2^2$	$\hat{S}_3^2$	$\hat{S}_4^2$	Global $\hat{G}_4^2$
$(3, .5)$	30	5.45	4.15	8.25	4.85	5.45
	100	5.90	4.00	17.45	4.90	8.70
	200	4.00	4.60	32.95	5.00	16.55
$(3, 2)$	30	4.95	3.55	12.70	5.05	5.55
	100	4.95	4.95	43.65	4.55	22.80
	200	4.25	5.50	83.35	5.45	59.90
$(5, .5)$	30	3.70	3.70	14.45	4.20	5.70
	100	5.10	4.25	51.60	4.65	27.05
	200	5.20	5.00	84.25	5.40	62.00

Dependent Errors

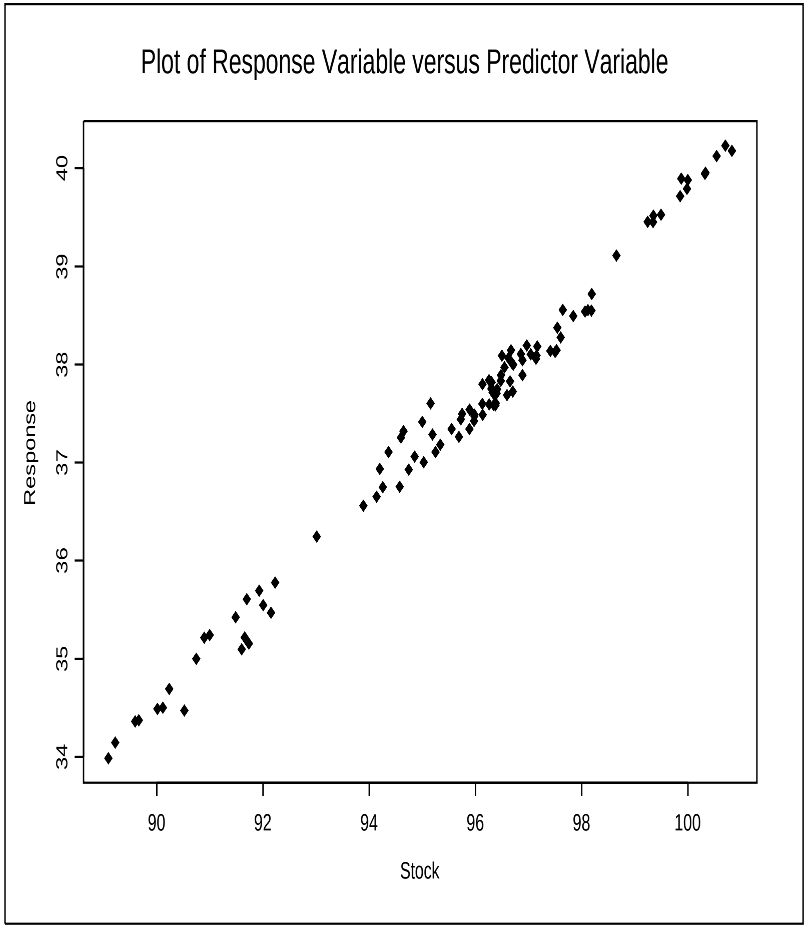
Error Type	Sample Size (n)	Component Statistics				Global
		$\hat{S}_1^2$	$\hat{S}_2^2$	$\hat{S}_3^2$	$\hat{S}_4^2$	$\hat{G}_4^2$
Mart.	30	20.85	10.25	3.80	35.45	27.85
	100	50.70	33.50	3.50	70.25	72.20
	200	63.90	50.35	5.00	79.40	87.25
Markov	30	5.20	3.05	3.50	8.20	5.30
	100	8.90	4.70	6.25	15.55	12.15
	200	11.40	5.85	5.15	18.35	15.70
Markov	30	5.45	2.90	1.85	13.30	6.85
	100	19.60	10.60	5.45	36.45	34.85
	200	29.50	21.40	3.60	47.45	54.45

(Martingale): $\epsilon_i = \frac{1}{\sqrt{i}} \sum_{j=1}^{i} \epsilon_j^*$; (Markov type): $\epsilon_i = \frac{1}{\sqrt{2}}(\epsilon_{i-1} + \epsilon_i^*)$.

Application: CREF Data Set

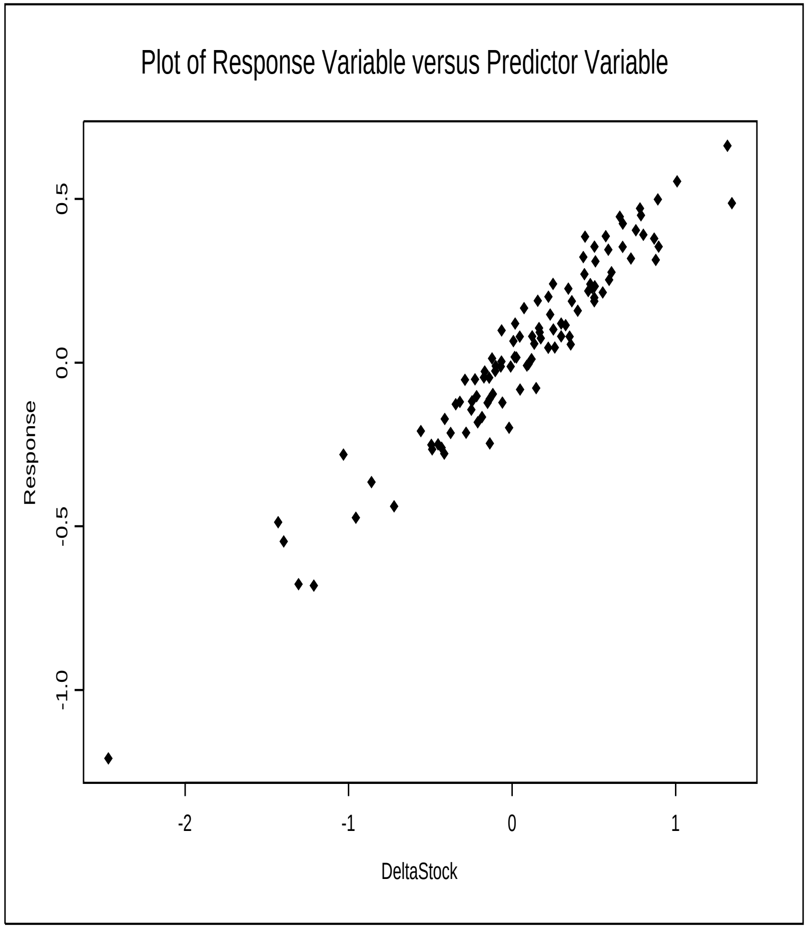
- Source: Data downloaded from TIAA-CREF website.
- Variables: Stock (X) and Growth (Y) Accounts end-of-trading day (EOTD) values
- Period: January 2, 1996 to May 31, 1996
- Size of Data Set: $n = 106$
- Goal: To relate the two accounts EOTD values.
- Question: Is it better to create a model based on the first-order differences: $\Delta Y_i = Y_i - Y_{i-1}$ and $\Delta X_i = X_i - X_{i-1}$?

First Model: Y vs X



- $\hat{Y} = -12 + .52X$
- $R^2 = 98.8\%$
- $\hat{G}_4^2 = 7.87 \; (p = .0965)$
- $\hat{S}_1^2 = 2.32 \; (p = .13)$
- $\hat{S}_2^2 = .55 \; (p = .46)$
- $\hat{S}_3^2 = 4.65 \; (p = .03)$
- $\hat{S}_4^2 = .35 \; (p = .55)$

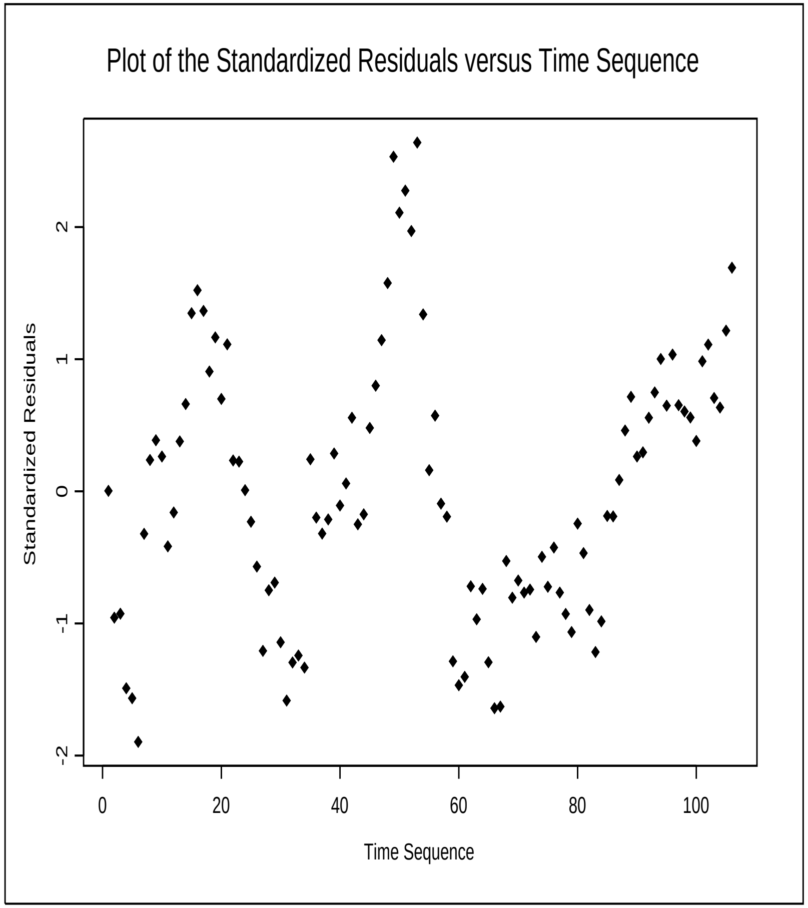
Second Model: ΔY vs ΔX



- $\widehat{\Delta Y} = .0057 + .4760(\Delta X)$
- $R^2 = 92.86\%$
- $\hat{G}_4^2 = 2.81 \; (p = .59)$
- $\hat{S}_1^2 = .11 \; (p = .73)$
- $\hat{S}_2^2 = .0041 \; (p = .95)$
- $\hat{S}_3^2 = .17 \; (p = .68)$
- $\hat{S}_4^2 = 2.51 \; (p = .11)$

Plots: Residuals vs Time

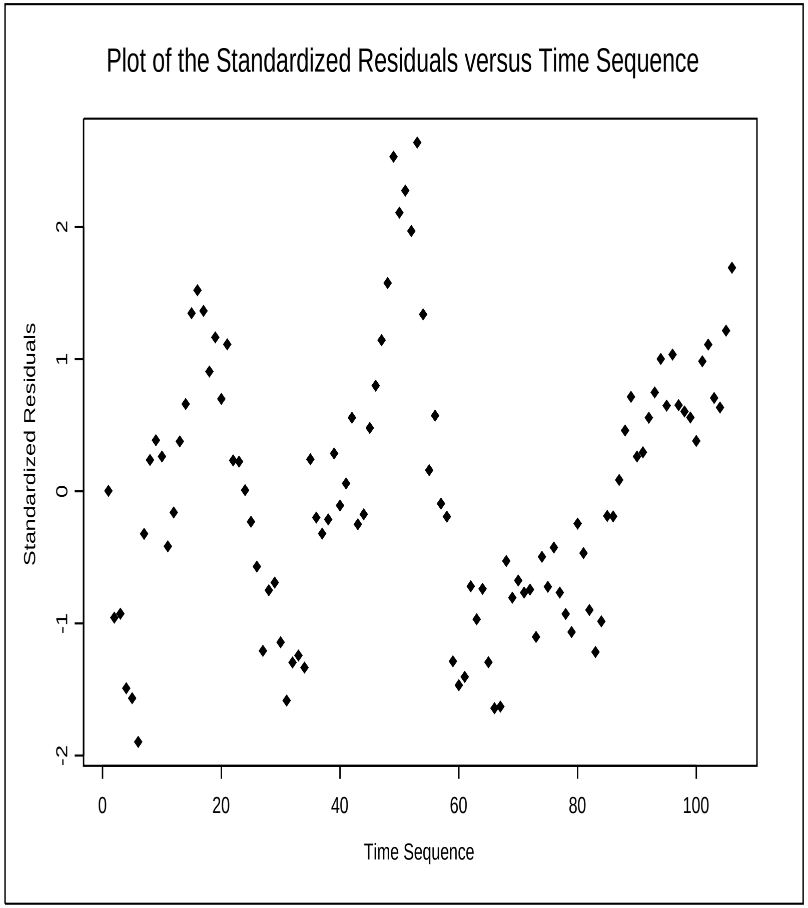
First Model



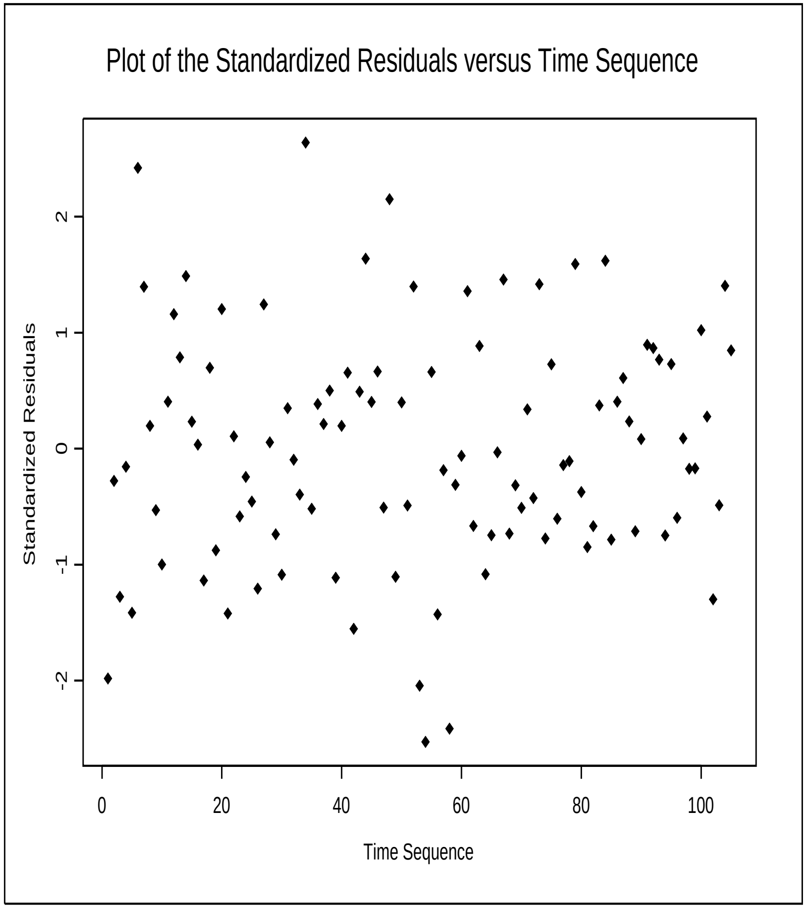
Second Model

Plots: Residuals vs Time

First Model



Second Model



Concluding Remarks

- Considered the problem of validating LM assumptions simultaneously.
- A global procedure making diagnostics formal and objective.
- Easy-to-implement and simple, even doable by undergraduate students!
- Appears to achieve what it purports to do as demonstrated by simulations.
- Will make the procedure ‘adaptive,’ that is, will choose the component statistics for the global statistics on the basis of the data!