Appetizers

Drowning in Information, but Starving for Knowledge. — Rutherford Roger.

I have deeply regretted that I did not proceed far enough at least to understand something of the great leading principles of mathematics; for men thus endowed seem to have extra sense! — Charles Darwin.

Global Validation of Linear Model Assumptions

Edsel A. Peña (Stat, USC)

Joint work with Elizabeth H. Slate (DB^{2}E, MUSC)

Research support from NIH/COBRE, NSF, and USC/MUSC Grants

pena@stat.sc.edu

Outline of Talk

- Motivation and Some Data
- LM Model and Diagnostics
- Specific Problem and Goals
- Proposed Procedure
- Theoretical Interludes
- Monte Carlo Adventures
- Application to Real Data
- Concluding Remarks

Motivation

- Regression analysis

$$Y_i = \beta_0 + \beta_1 X_{1i} + \ldots + \beta_p X_{pi} + \epsilon_i$$

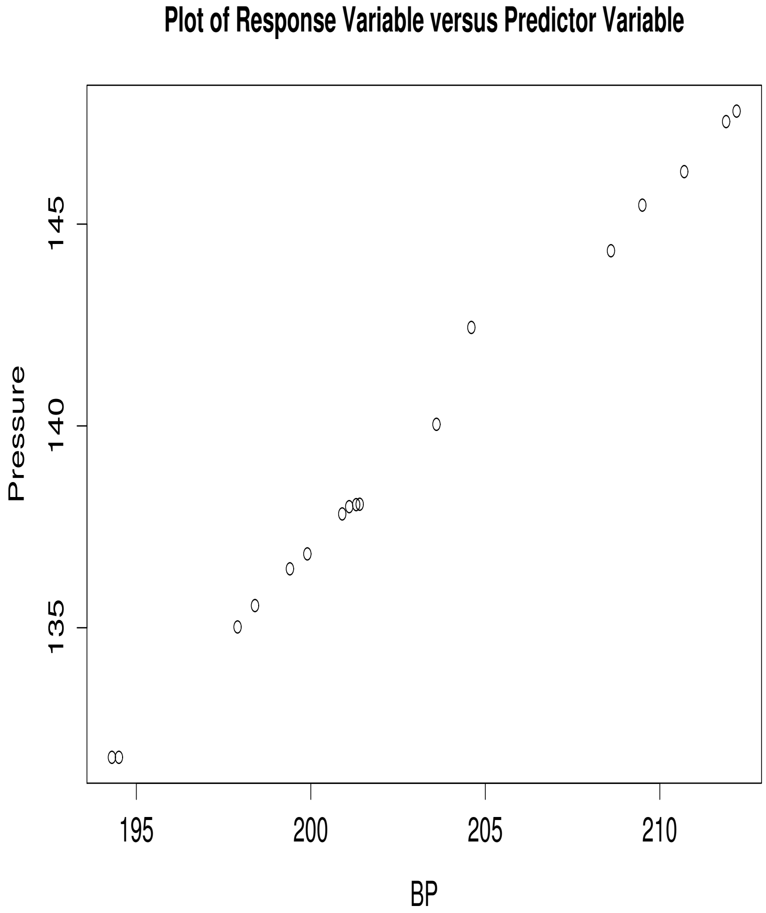
- Analysis of Variance Models

$$Y_{ij} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \epsilon_i$$

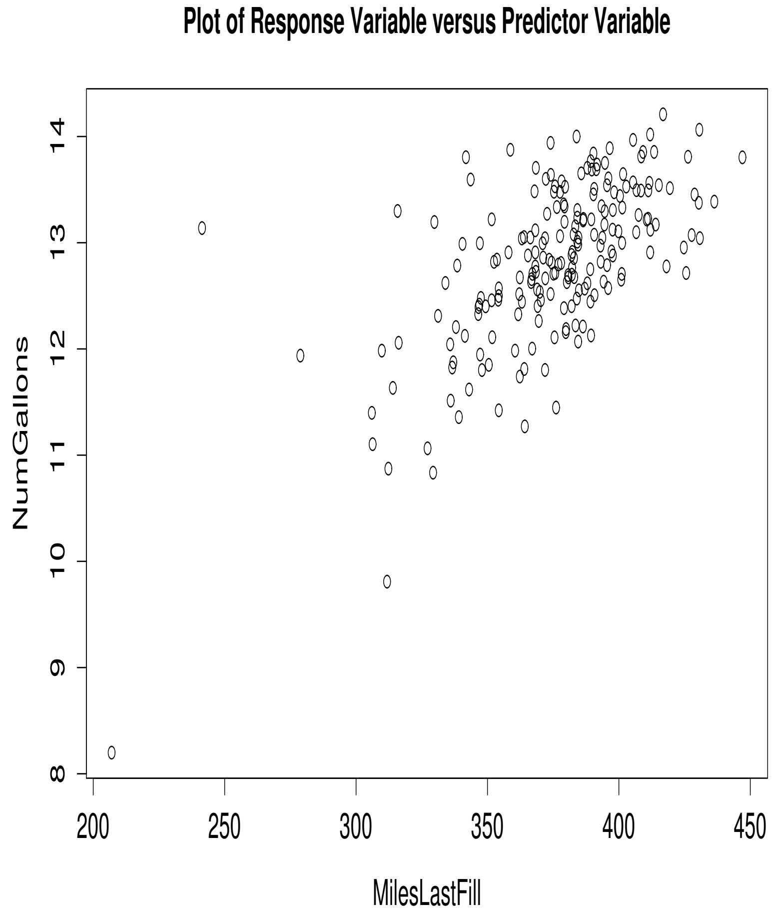
- Impetus from teaching Stat 700-701: Experience with giving a final examination using a real data set from TIAA-CREF.

Two Motivating Data Sets

(i) Forbes BP Data



(ii) Car Efficiency



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*:

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

(A2) *Homoscedasticity*:

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

(A3) *Uncorrelatedness*:

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

(A4) *Normality*:

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

Estimators

- ML Estimator of β:

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

- ML Estimator of σ^2:

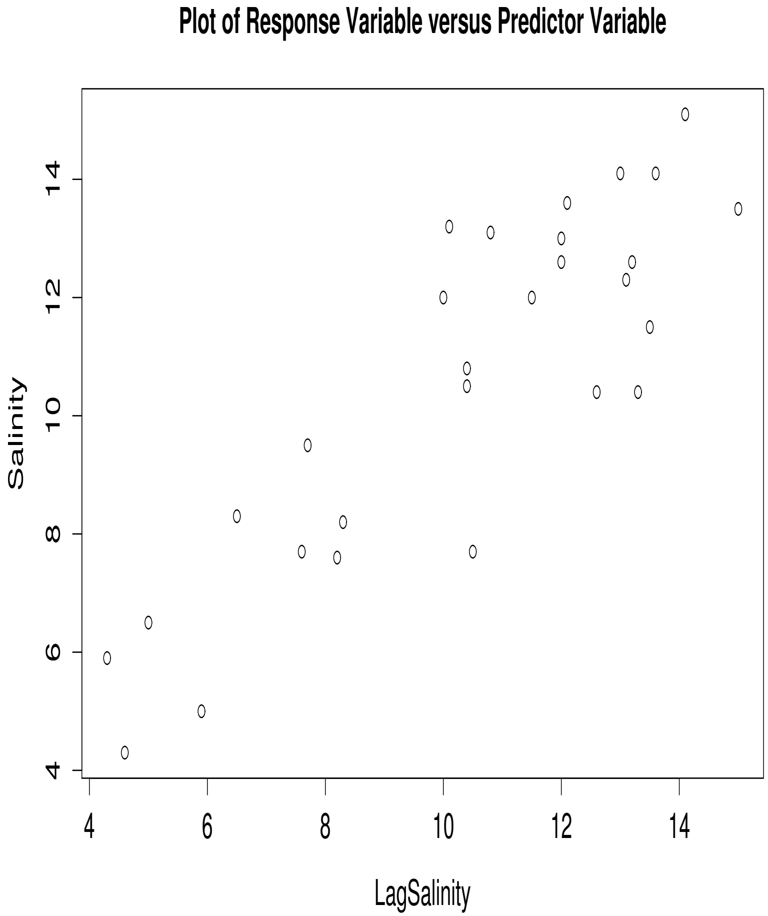
$$s^2 = \hat{\sigma}^2 = \frac{1}{n}\mathbf{Y}^{\mathrm{t}}(\mathbf{I} - \mathbf{P}_{\mathbf{X}})\mathbf{Y},$$

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

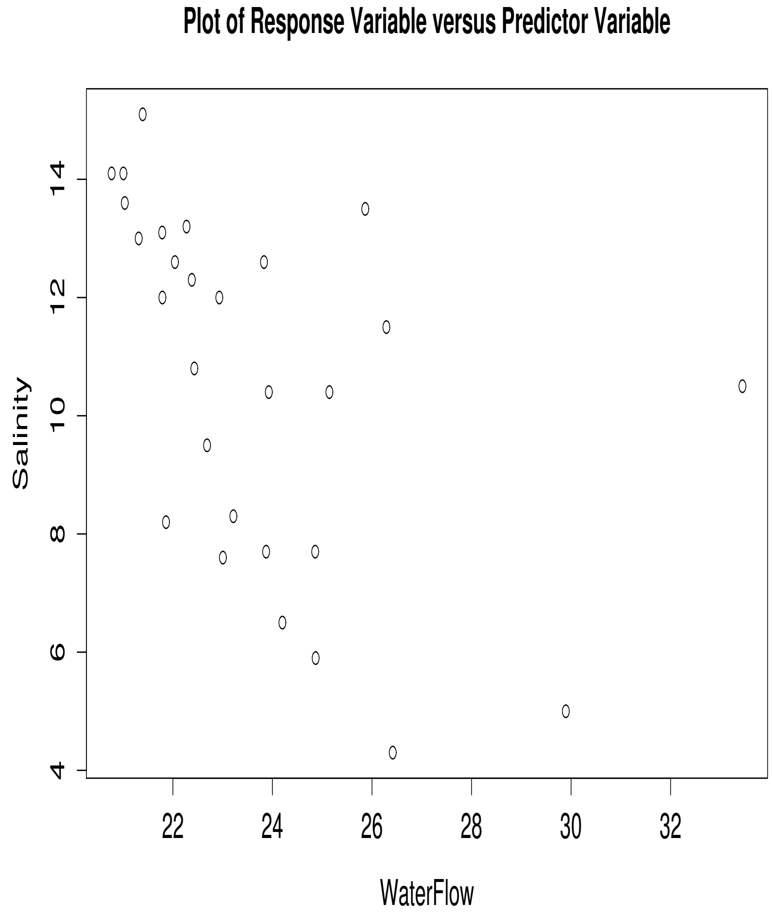
$$\mathbf{P}_{\mathbf{X}} = \mathbf{X}(\mathbf{X}^{\mathrm{t}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{t}}$$

Example: Water Salinity (Carroll & Ruppert; Atkinson)

Predictor = LagSalinity



Predictor = Waterflow



Example: Fitting the LM

- Response: $Y =$ Water Salinity.
- Predictors: $X_1 =$ LagSalinity; $X_2 =$ Trend; $X_3 =$ WaterFlow.
- Model: $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \sigma \epsilon_i$
- Results of Fitting Model (Using `lm` in `R`):
- Coefficients: $b_0 = 9.59$, $b_1 = .78$, $b_2 = -.02$, $b_3 = -.29$.
- If *all* assumptions are satisfied, tests of significance show that all coefficients (β_is) are significantly different from zero.
- Multiple Coefficient of Determination = $R^2 = 83\%$.

Validating LM Assumptions

- *Standardized* Residuals:

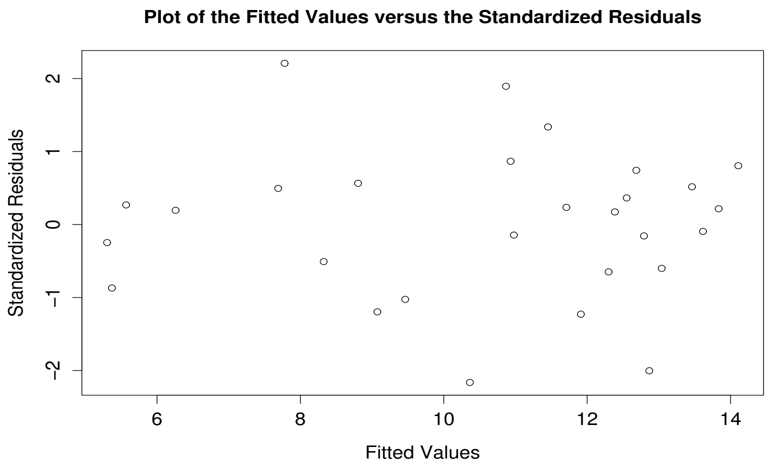
$$\mathbf{R} = \frac{\mathbf{Y} - \mathbf{X}\mathbf{b}}{s} = \frac{(\mathbf{I} - \mathbf{P}_{\mathbf{X}})\mathbf{Y}}{s}$$

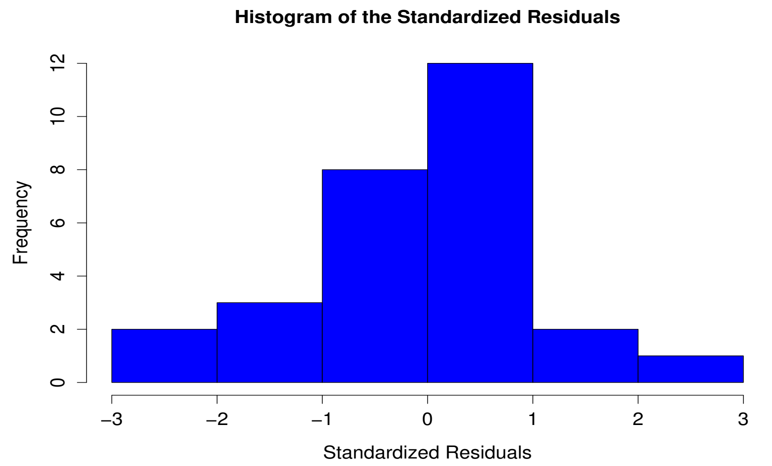
  or, in long form,

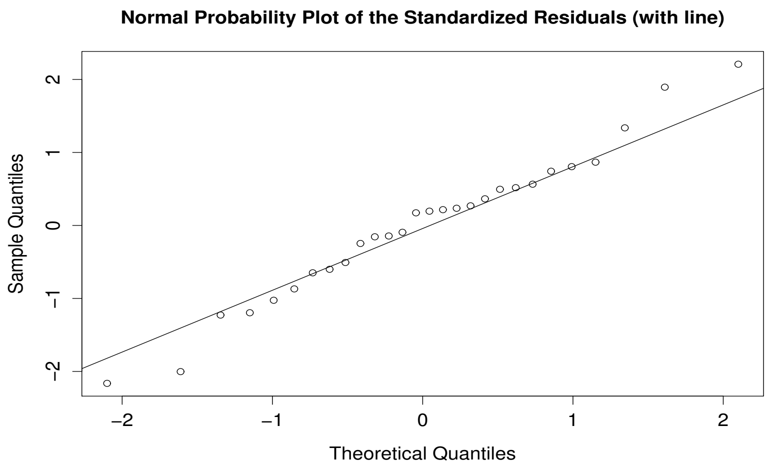
$$R_i = \frac{Y_i - \hat{Y}_i}{s}, \quad i = 1, 2, \ldots, n$$

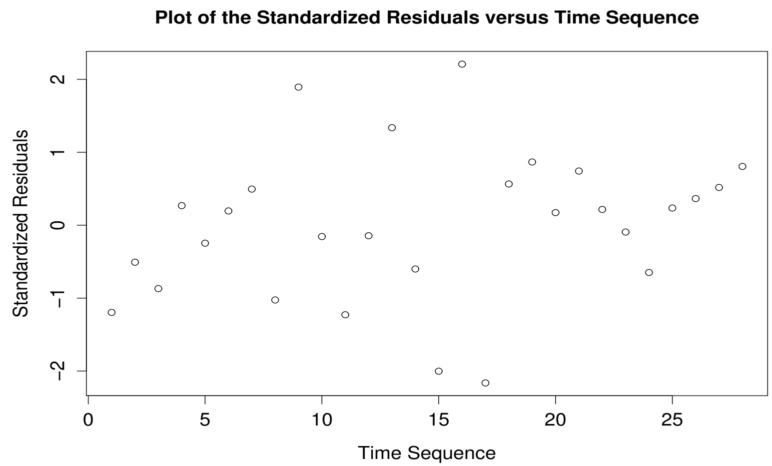
- Graphical or Diagnostic plots based on $\mathbf{R}$. Discussed in many (elementary) textbooks!
- Formal significance tests.
- Such formal hypothesis tests are based on $\mathbf{R}$.

Example: Salinity Data









Question: Are the assumptions OK?

Issues to Consider

Issues to Consider

- Varied plots to detect varied assumptions. Made truly easy by packages: Minitab, SAS, SPSS, Excel, S-Plus, R, etc.

Issues to Consider

- Varied plots to detect varied assumptions. Made truly easy by packages: Minitab, SAS, SPSS, Excel, S-Plus, R, etc.
- *“A picture is worth a thousand words,*

Issues to Consider

- Varied plots to detect varied assumptions. Made truly easy by packages: Minitab, SAS, SPSS, Excel, S-Plus, R, etc.
- *“A picture is worth a thousand words, but beauty is in the eye of the beholder!”*

Issues to Consider

- Varied plots to detect varied assumptions. Made truly easy by packages: Minitab, SAS, SPSS, Excel, S-Plus, R, etc.
- *"A picture is worth a thousand words, but beauty is 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 truly easy by packages: Minitab, SAS, SPSS, Excel, S-Plus, R, etc.
- *"A picture is worth a thousand words, but beauty is 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.

Issues to Consider

- Varied plots to detect varied assumptions. Made truly easy by packages: Minitab, SAS, SPSS, Excel, S-Plus, R, etc.
- *"A picture is worth a thousand words, but beauty is 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.
- Need to be aware of possible synergy among different violations.

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

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

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

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

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

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

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 $\mathbf{V} = (1, 2, \ldots, n)^{\mathrm{t}}/n$. 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 usually be detected via:

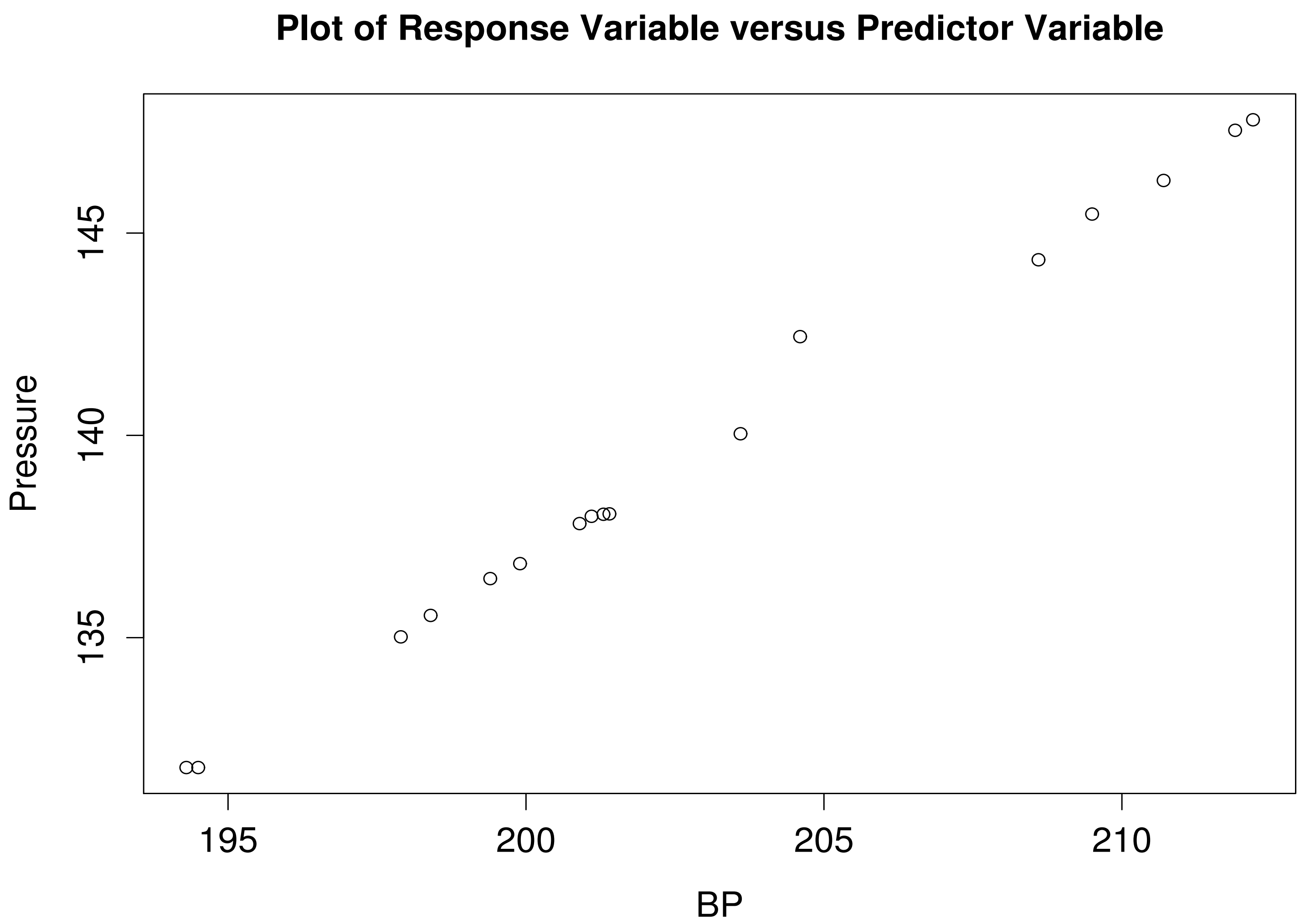
- Skewed error distributions indicated by $\hat{S}_1^2$;
- Deviations from the normal distribution kurtosis of the true error distribution generally revealed by $\hat{S}_2^2$;
- Misspecified link function or the absence of other predictor variables in the model detected by $\hat{S}_3^2$;
- Presence of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

Deletion Statistics

$$\Delta\hat{G}_4^2[i] = \left[\frac{\hat{G}_4^2[i] - \hat{G}_4^2}{\hat{G}_4^2}\right] \times 100; \quad p[i] = \mathbf{P}\left\{\chi_4^2 > \hat{g}_4^2[i]\right\}.$$

- $\Delta\hat{G}_4^2[i] =$ Percent relative change in value of global statistic $\hat{G}_4^2$ after deletion of ith observation.
- $p[i] = p$-value after deletion of ith observation.
- Idea: observation with a large absolute value of $\Delta\hat{G}_4^2[i]$ or a big change in $p[i]$ is either an outlier or has large influence.
- Values of $(\Delta\hat{G}_4^2[i], p[i]), i = 1, 2, \ldots, n$, could be plotted to see outlying or influential observations.

Example: Forbes Data



Example: Model Validation

- Global Test: $\hat{G}_4^2 = 98.4 \; (p = 0)$.
- Decision: Assumptions NOT satisfied!

Example: Model Validation

- Global Test: $\hat{G}_4^2 = 98.4$ $(p = 0)$.
- Decision: Assumptions NOT satisfied!

- Component Statistics (with p-Value and Decision)
  - $\hat{S}_1 = 28.7$ $(p = 0)$; Decision: Violation!
  - $\hat{S}_2 = 65.1$ $(p = 0)$; Decision: Violation!
  - $\hat{S}_3 = 1.9$ $(p = 0.17)$; Decision: OK.
  - $\hat{S}_4 = 2.8$ $(p = 0.10)$; Decision: OK.

Example: Model Validation

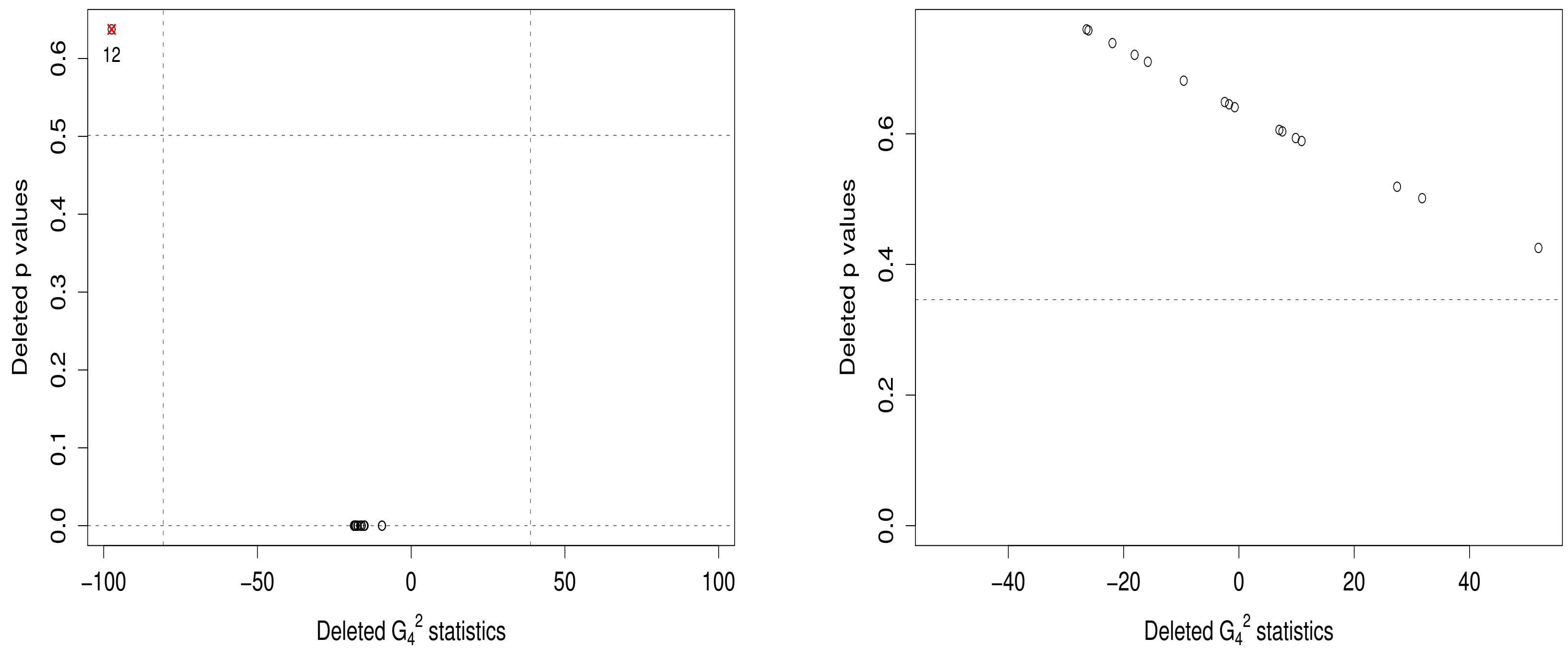
- Global Test: $\hat{G}_4^2 = 98.4 \ (p = 0)$.
- Decision: Assumptions NOT satisfied!

- Component Statistics (with p-Value and Decision)
  - $\hat{S}_1 = 28.7 \ (p = 0)$; Decision: Violation!
  - $\hat{S}_2 = 65.1 \ (p = 0)$; Decision: Violation!
  - $\hat{S}_3 = 1.9 \ (p = 0.17)$; Decision: OK.
  - $\hat{S}_4 = 2.8 \ (p = 0.10)$; Decision: OK.

- Based on the directional tests, there seems to be violations in the normality assumption, or there could be outliers or influential observations.

Example: Deletion Statistics

All Observations | After 12th is Deleted



Result: The 12th obs. is quite different: outlier or too influential. Upon its deletion, $\hat{G}_4^2[12] = 2.54(P = 0.64)$.

Theoretical Interludes: Why it Works!

- 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: Why it Works!

- 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:

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

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

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:

$$\begin{aligned} \mathbf{U}(\theta = \mathbf{0}, \sigma^2, \beta) &\equiv \frac{\partial}{\partial \theta} \log f_{\mathbf{R}^0}(\mathbf{r}^0 | \theta, \beta, \sigma^2) \, |_{\theta = \mathbf{0}} \\ &= \mathbf{Q}(\mathbf{r}^0; \sigma^2, \beta). \end{aligned}$$

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:

$$\begin{aligned} \mathbf{U}(\theta = \mathbf{0}, \sigma^2, \beta) &\equiv \frac{\partial}{\partial \theta} \log f_{\mathbf{R}^0}(\mathbf{r}^0 | \theta, \beta, \sigma^2) \mid_{\theta = \mathbf{0}} \\ &= \mathbf{Q}(\mathbf{r}^0; \sigma^2, \beta). \end{aligned}$$

- 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:

$$\begin{aligned} \mathbf{U}(\theta = \mathbf{0}, \sigma^2, \beta) &\equiv \frac{\partial}{\partial \theta} \log f_{\mathbf{R}^0}(\mathbf{r}^0 | \theta, \beta, \sigma^2) \,|_{\theta = \mathbf{0}} \\ &= \mathbf{Q}(\mathbf{r}^0; \sigma^2, \beta). \end{aligned}$$

- 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

An application of the multivariate CLT yields:

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

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

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

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

  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.
- Sample Size: $n \in \{30, 100, 200\}$
- Replications: 20,000 for levels; 5,000 for powers.
- Covariate: $x_1, x_2, \ldots, x_n$ standard uniform
- Fitted Model: $Y_i = \beta_0 + \beta_1 x_i + \sigma\epsilon_i$
- User-supplied $\mathbf{V}$: $\mathbf{V} = (1, 2, \ldots, n)/n$
- Level of significance: 5%
- Programs implementing the procedure were coded in the `R` language.

Plots of Achieved Levels

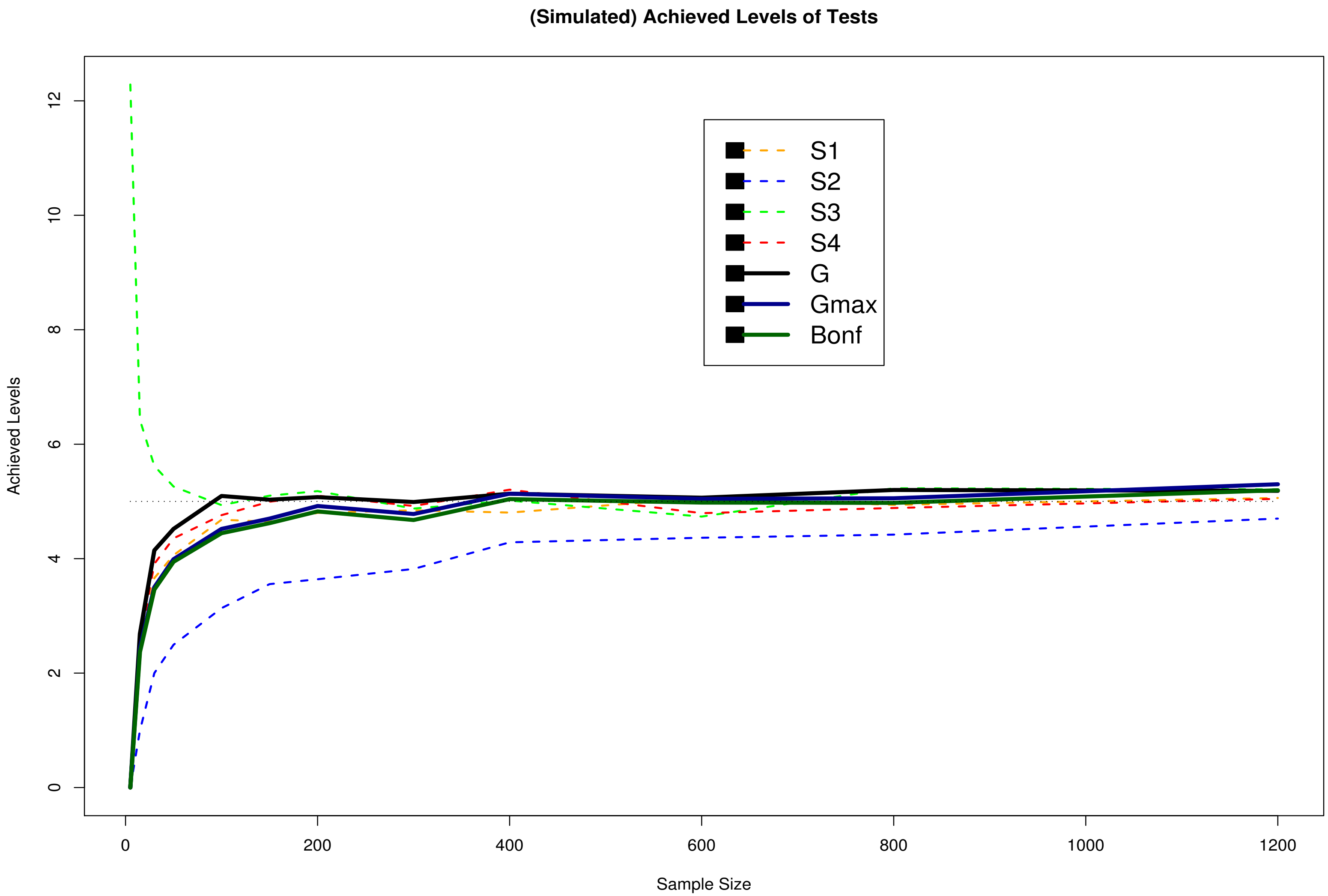


Table of Achieved Levels

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	3.66	2.00	5.62	3.91	4.15
	100	4.68	3.14	4.94	4.76	5.10
	200	4.80	3.64	5.18	5.10	5.08

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

Power Studies

- Generic Data Generation Model for Alternatives.

$$Y_i = x_i + \beta_2 x_i^\gamma + \sigma_i^* x_i^\alpha \epsilon_i, \quad i = 1, \ldots, n$$

- β_2 and γ = misspecified link function parameters.
- α = heteroscedastic parameter.
- $\sigma_i^* = 1$ if $i \leq n/2$; $\sigma_i^* = \sigma_2$ if $i > n/2$.
- With $\epsilon_i^*, i = 1, \ldots, n$ IID $N(0, 1)$:
- Martingale Errors: $\epsilon_i = \frac{1}{\sqrt{i}} \sum_{j=1}^{i} \epsilon_j^*$
- Markov Errors: $\epsilon_i = (\rho\epsilon_{i-1} + \epsilon_i^*)/(\sqrt{1+\rho^2})$

Some Simulated Powers

Model: Errors (ϵ_i's) are Non-Normal

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	21.6	21.1	6.0	10.6	23.9
	100	38.9	61.9	5.1	17.0	59.8
LG	30	11.80	12.60	5.90	7.20	15.05
	100	17.45	30.30	5.50	8.20	29.35
DE	30	19.50	24.75	5.60	10.35	27.20
	100	35.05	73.55	5.60	14.60	70.65
$\chi_5^2 - 5$	30	48.7	19.7	6.0	10.3	34.2
	100	98.7	57.8	5.8	14.5	92.5

Model: Heteroscedastic Variances

$$Y_i = x_i + x_i^{\alpha} \sigma_i^* \epsilon_i$$

Value of (α, σ_2)	Sample Size (n)	Component Statistics				Global
		$\hat{S}_1^2$	$\hat{S}_2^2$	$\hat{S}_3^2$	$\hat{S}_4^2$	$\hat{G}_4^2$
$(2, 1)$	30	40	85	29	30	86
	100	49	100	15	28	99
$(1, 2)$	30	13	12	5	40	27
	100	19	38	7	97	90

Model: Misspecified Link Function

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

Value of (β_2, γ)	Sample Size (n)	Component Statistics				Global
		$\hat{S}_1^2$	$\hat{S}_2^2$	$\hat{S}_3^2$	$\hat{S}_4^2$	$\hat{G}_4^2$
$(3, 2)$	30	3	1.7	19	4	8
	100	5	2.7	55	5	31
$(5, 2)$	30	4	2	41	3	17
	100	4	3	94	4	79

Model: 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	23	10	3	42	32
	100	55	38	4	72	75
Markov	30	8	.7	1.2	22	13
	100	26	24	.7	48	48

Martingale Type: $\epsilon_i = \frac{1}{\sqrt{i}} \sum_{j=1}^{i} \epsilon_j^*$

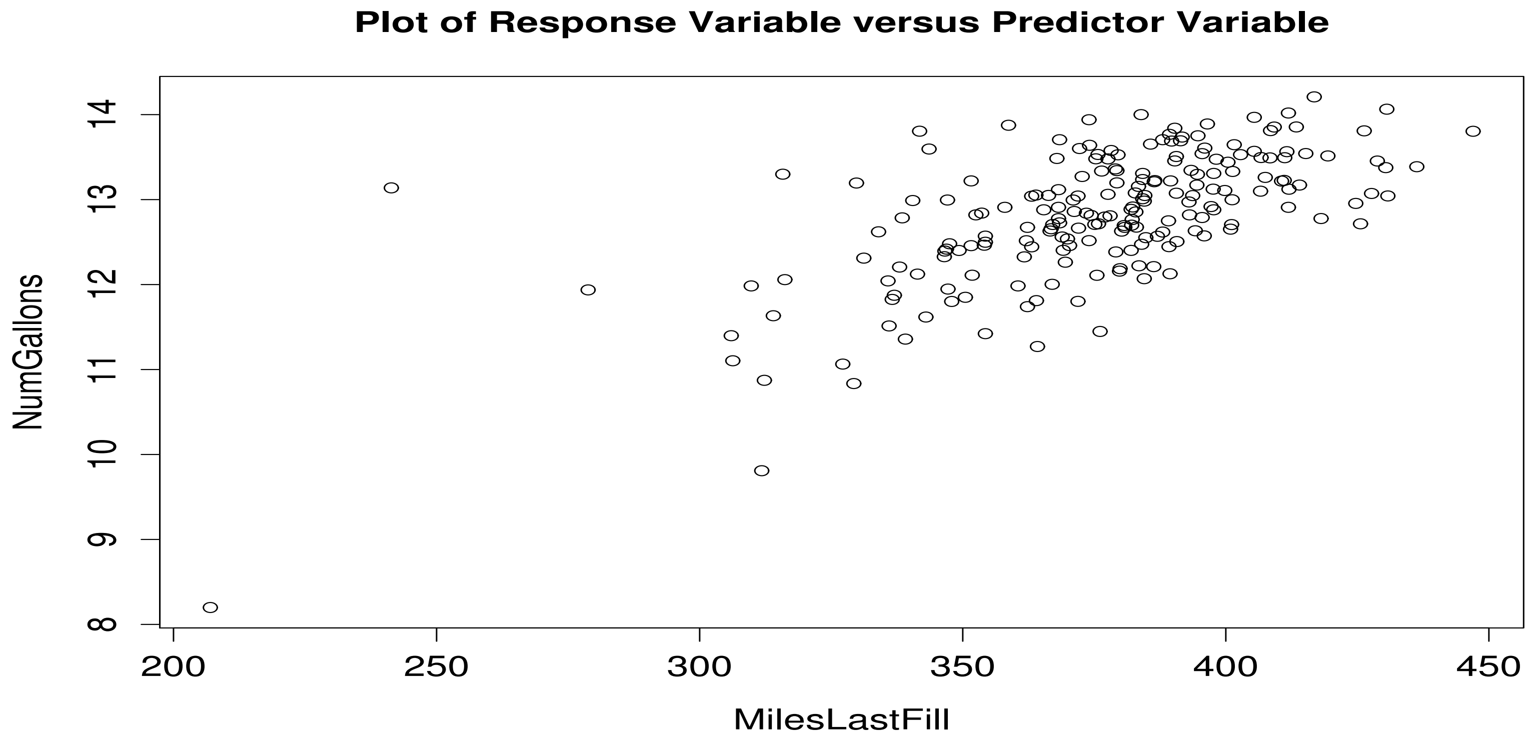
Markov Type: $\epsilon_i = \frac{1}{\sqrt{6}}(5\epsilon_{i-1} + \epsilon_i^*)$

Model: Multiple Violations

Violated Assumptions	Sample Size (n)	Component Statistics				Global
		$\hat{S}_1^2$	$\hat{S}_2^2$	$\hat{S}_3^2$	$\hat{S}_4^2$	$\hat{G}_4^2$
(All)	30	27	47	17	53	63
	100	47	96	51	56	96
(All)	30	42	69	12	10	67
	100	77	99	12	75	99.8
(All)	30	46	72	5.2	41	72
	100	62	99.9	11	93	99.9
(All)	30	42	62	10	73	77
	100	66	98.9	20	71	99.4

Example: Car Efficiency

- Data gathered for 3 years. Mileage recorded every gas fill-up. There were $n = 205$ observations.
- To create regression model with **NumGallons** as response and **MilesLastFill** as predictor.



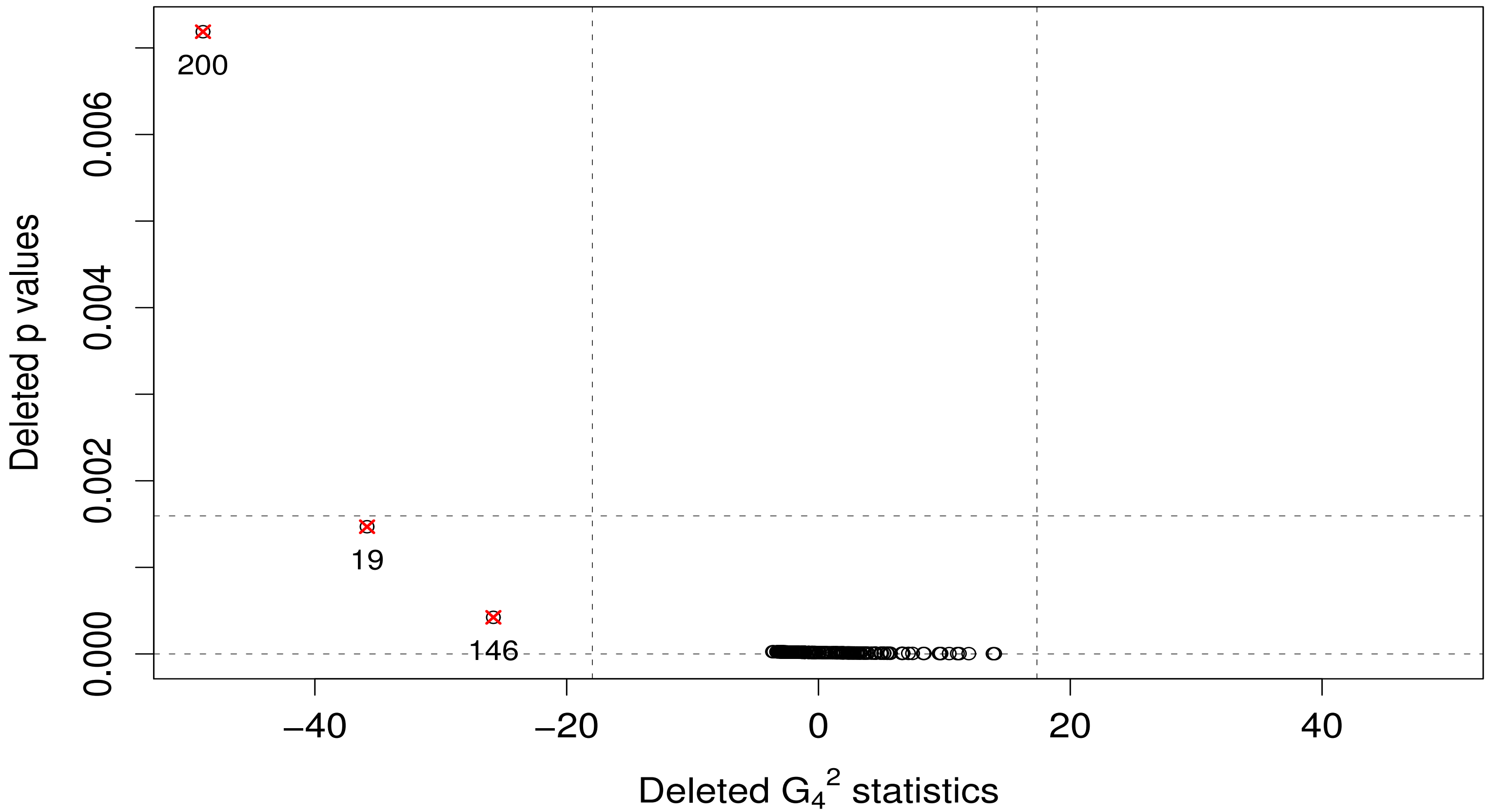
Fitted Model

- $\hat{Y} = 6.808 + 0.016X$
- $\hat{\sigma} = .691$
- F-value = 151 ($p = 0$)
- Coefficient of Determination = 42.6%

Question: Are the model assumptions valid??

- $\hat{G}_4^2 = 27.5(p = 0)$.
- Some assumptions violated!
- $\hat{S}_1^2 = .23(p = .63)$; $\hat{S}_2^2 = 25.1(p = 0)$; $\hat{S}_3^2 = 1.6(p = .20)$; $\hat{S}_4^2 = .48(p = .48)$.

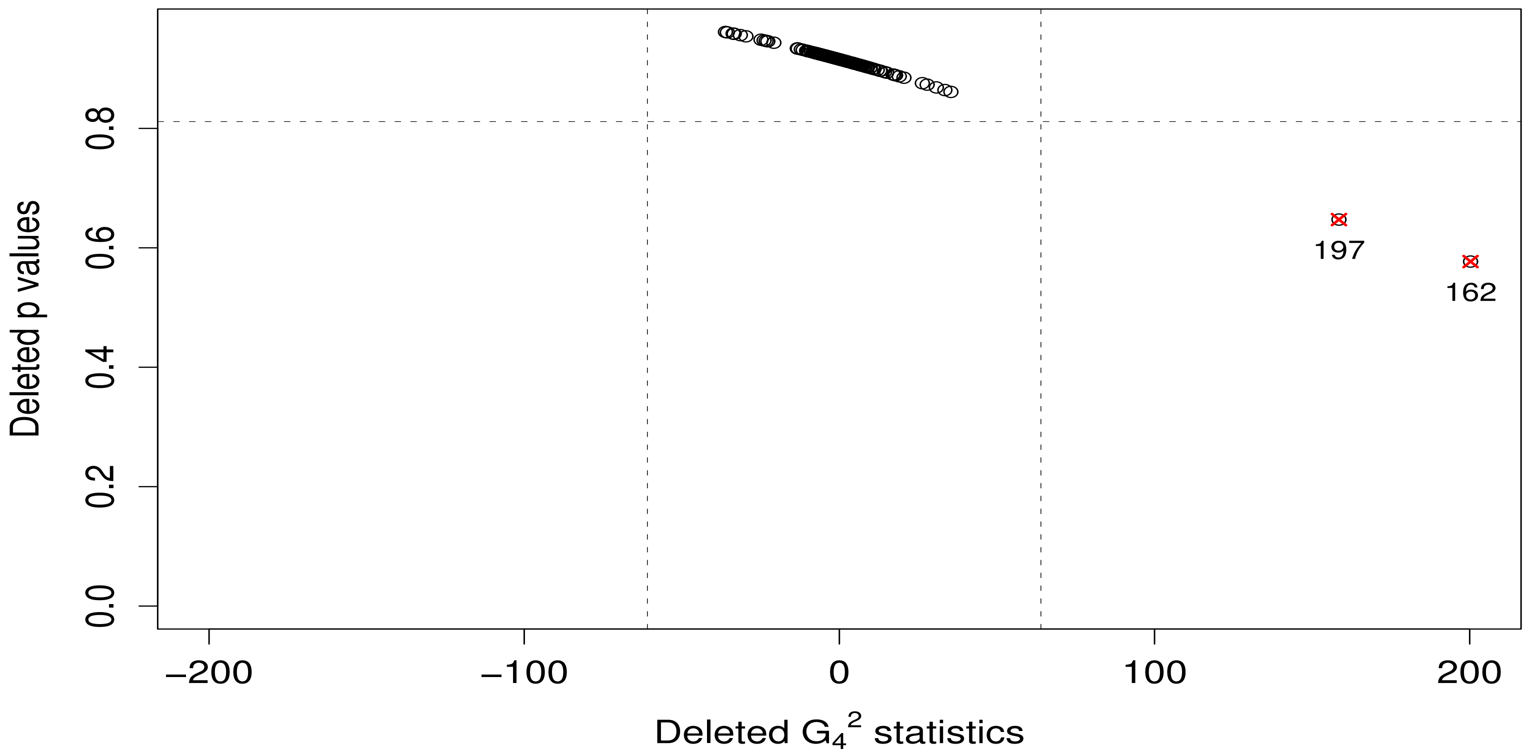
Plot of Deletion Statistics



Question: Why are the 19th, 146th, and 200th observations outliers?

After Exclusion of 19th, 146th, 200th Obs

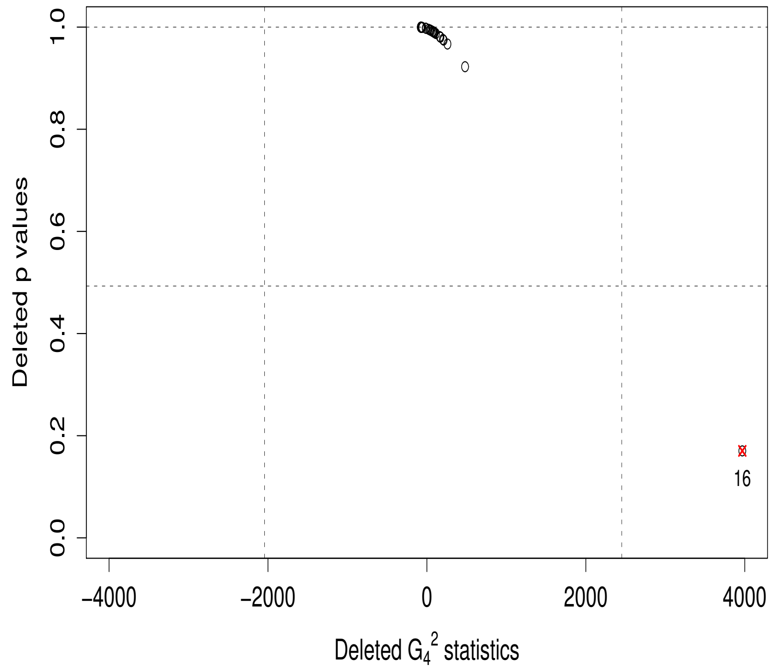
- Global: $\hat{G}_4^2 = .96(p = .92)$
- Components: $\hat{S}_1^2 = .04(p = .84)$; $\hat{S}_2^2 = .002(p = .96)$; $\hat{S}_3^2 = .71(p = .40)$; and $\hat{S}_4^2 = .21(p = .65)$.

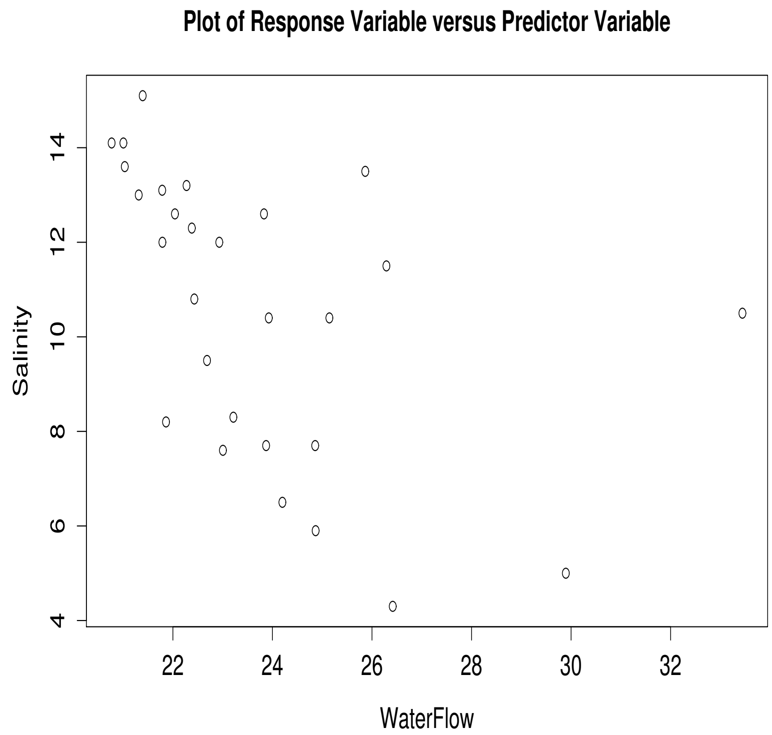


Back to Salinity Data

- Variables: $Y =$ Salinity; $X_1 =$ LagSalinity; $X_2 =$ Trend; $X_3 =$ WaterFlow.
- Model: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \sigma\epsilon$
- Estimates: $b_0 = 9.6$, $b_1 = .78$, $b_2 = -.03$, $b_3 = -.30$. All significant!
- Coefficient of Determination: 83%.
- Global: $\hat{G}_4^2 = .16(p = .99)$.
- Components: $\hat{S}_1^2 = .02(p = .87)$; $\hat{S}_2^2 = .01(p = .95)$; $\hat{S}_3^2 = 0(p = 1.0)$; $\hat{S}_4^2 = .13(p = .72)$.
- Question: Does this mean that model assumptions are satisfied?

Deletion Statistics and Outlier





- Plot 1: 16th observation unusual.
- Waterflow$_{16}$ = 33.443 supposed to be 23.443 (Atkinson).
- Re-analysis: $\hat{G}_4^2 = 6.7(p = .15)$; $\hat{S}_3^2 = 4.2(p = .04)$.

Concluding Remarks

- Presented a simple, but formal, method of validating LM assumptions.
- Lessen subjectivity in model validation.
- Comparisons: Bonferroni-type, Sidak-type, and Box-Cox transformations.
- Adaptive procedure: choose components based on data. Effect of data double-dipping.
- Variety: Use different basis functions: Wavelets?
- What should be done if model assumptions are not satisfied? Issue of two-step process.
- `R` package, or in `DoStat?`!