“Statistics on Statistics”

(On validating linear model assumptions)

Edsel A. Peña

Department of Statistics, Univ. of South Carolina

Support from NIH and NSF

Joint work with Prof. Elizabeth Slate

pena@stat.sc.edu

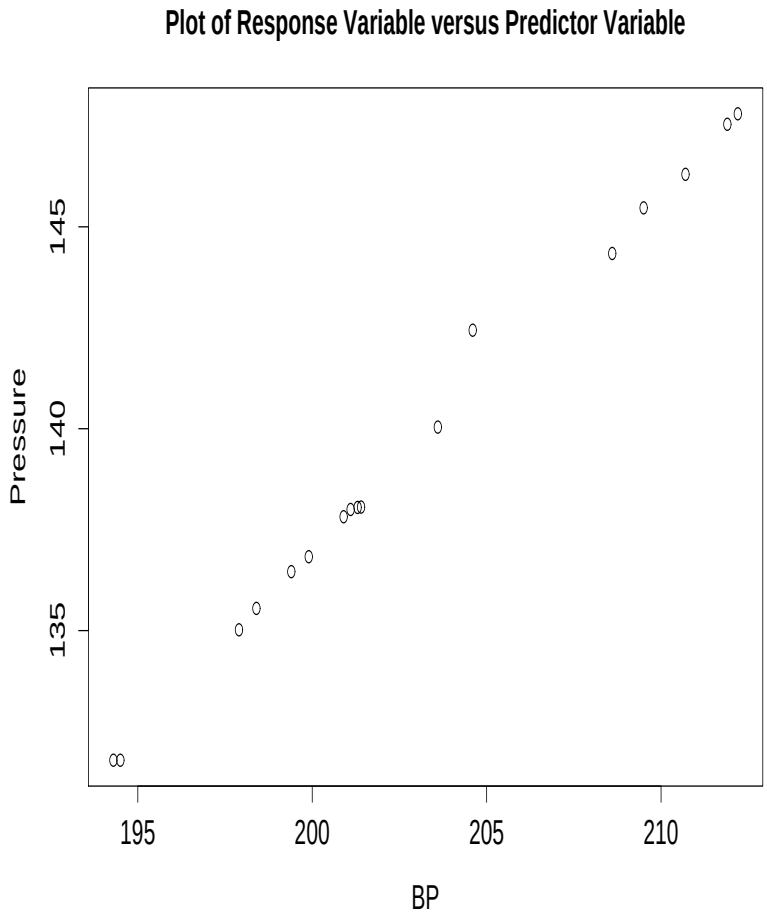
Goal and Outline of Talk

Main Goal: Interplay among statistics, mathematics, and computing in developing new methods.

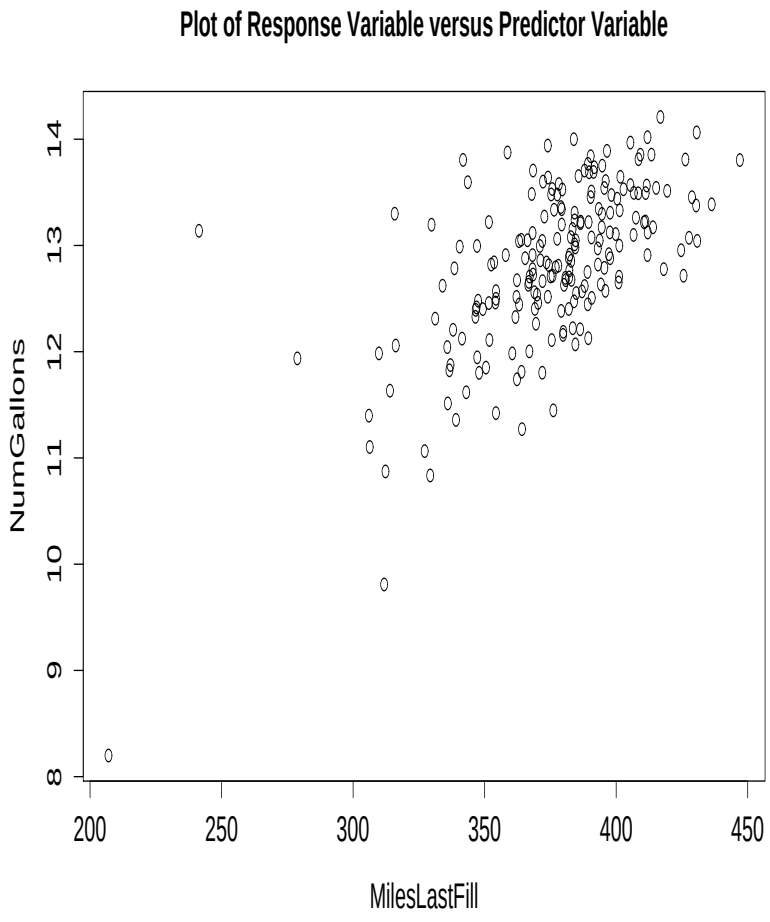
- Illustrative Examples and Review
- Linear Model and Diagnostics
- Problem and Goals
- Proposed Procedure
- Theoretical Interludes
- Monte Carlo Adventures
- Application to Illustrative Data
- Concluding Remarks

Some Illustrative Data

Boiling Point Vs Pressure



Gas Mileage Data



Simple Linear Regression Model

- Model: $Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$
- Assumption: ϵ_is are IID $N(0, \sigma^2)$
- Least-Squares Method: The best fitting line is $\hat{Y} = b_0 + b_1 X$, with b_0 and b_1 minimizing

$$Q(b_0, b_1) = \sum_{i=1}^{n} (Y_i - (b_0 + b_1 X_i))^2$$

- $b_1 = \frac{\sum (Y_i - \bar{Y})(X_i - \bar{X})}{\sum (X_i - \bar{X})^2}$
- $b_0 = \bar{Y} - b_1 \bar{X}$
- $\hat{\sigma}^2 = \frac{1}{n-2} \sum (Y_i - (b_0 + b_1 X_i))^2$

Inferences: Tests and Predictions

- Is the predictor variable X (e.g. Boiling Point) significant for the response variable Y (e.g. Pressure)?
- Declare significant predictor if $|T_c| > t_{n-2;\alpha/2}$ where

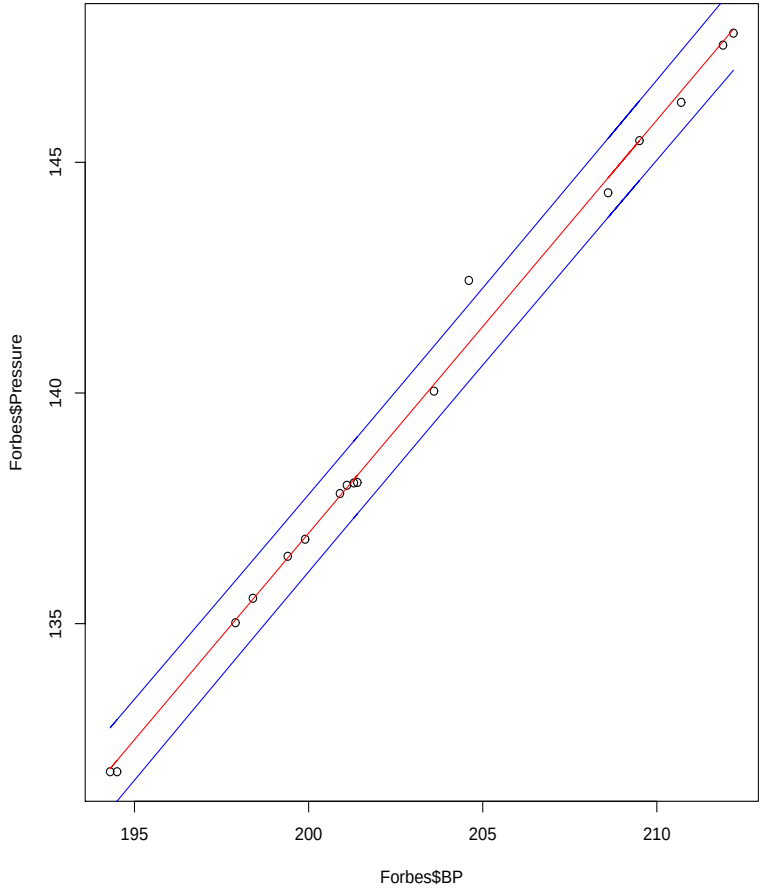
$$T_c = \frac{b_1}{\hat{\sigma}/\sqrt{\sum(X_i - \bar{X})^2}}$$

- To predict the value of Y at $X = x_0$, one constructs the confidence interval:

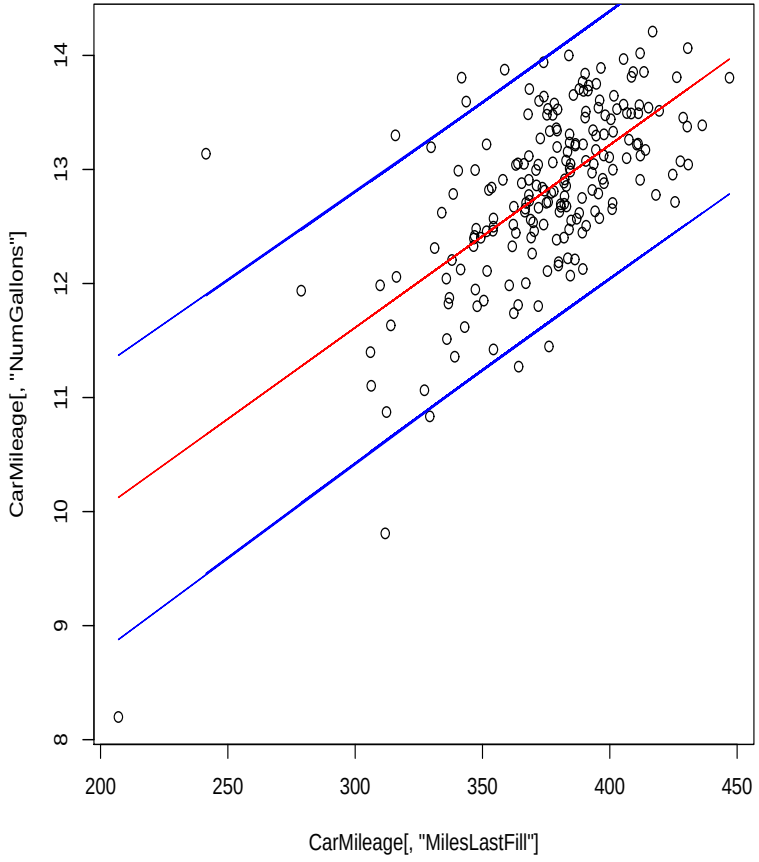
$$(b_0 + b_1 x_0) \pm t_{n-2;\alpha/2}\hat{\sigma}\sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{X})^2}{\sum(X_i - \bar{X})^2}}$$

Fitted Line and Prediction Interval

BP vs Pressure



Car Mileage Data



Summary of Regression Fits

Estimator/Quantity	BP vs Pressure	Car Mileage
b_0	$-42.1 (p = 0)$	$6.81 (p = 0)$
b_1	$.89 (p = 0)$	$.016 (p = 0)$
$\hat{\sigma}$	$.38$	0.591
R^2	$.995$	0.426
F	$2965 (p = 0)$	$150.7 (p = 0)$

- **The validity of these results, however, especially those pertaining to testing, confidence intervals, and prediction, are highly dependent on the model assumptions being true.**
- **It is imperative that model assumptions be validated!**

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 $\mathbf{H}$:

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

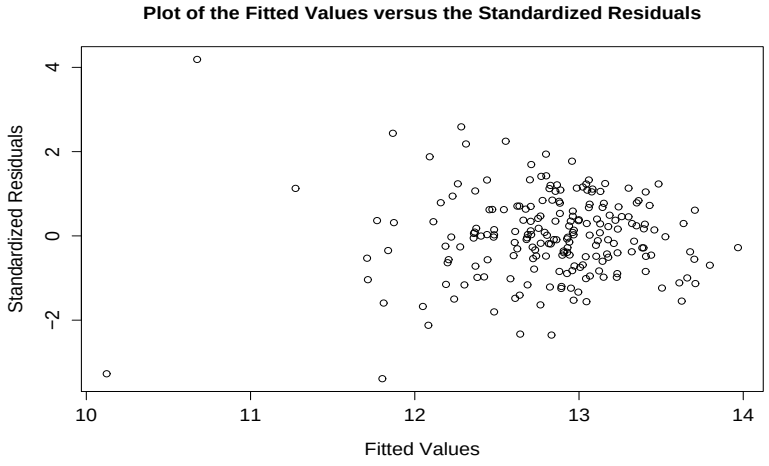
Validating LM Assumptions

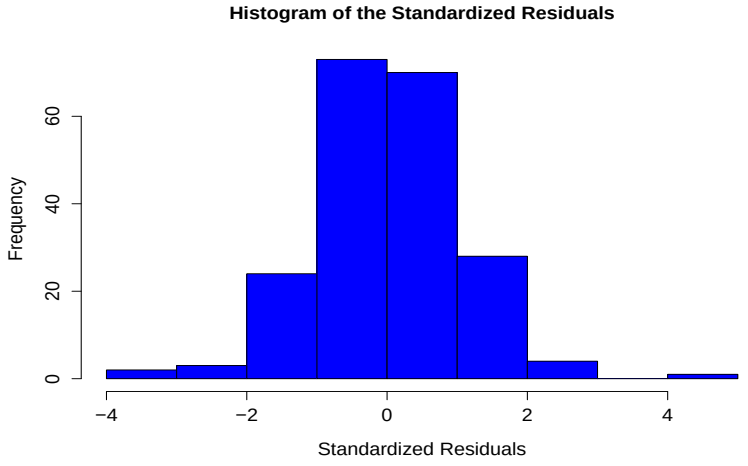
- *Standardized* Residuals:

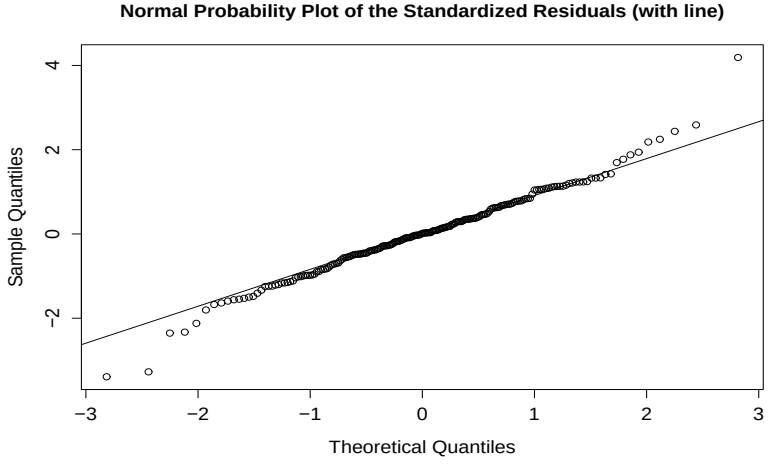
$$\mathbf{R} = \frac{\mathbf{Y} - \mathbf{X}\mathbf{b}}{s} = \frac{(\mathbf{I} - \mathbf{P}_{\mathbf{X}})\mathbf{Y}}{s}$$

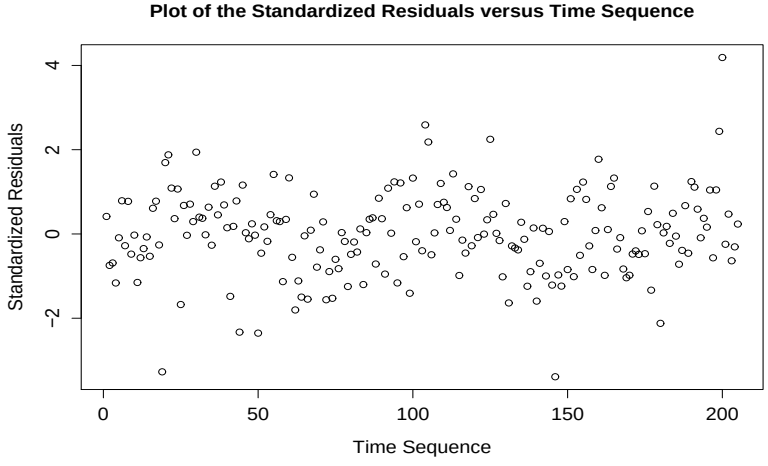
- Graphical Methods.
- Diagnostic plots based on $\mathbf{R}$. Discussed in many (elementary) textbooks!
- Formal tests.
- Such formal hypothesis tests are based on $\mathbf{R}$.

Example: Car Mileage Diagnostics









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,*

Issues to Consider

- Varied plots to detect varied assumptions. Made easy by statistical packages.
- *“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 easy by statistical packages.
- *"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 easy by statistical packages.
- *"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 (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

  H_0 : Assumptions (A1)-(A4) all hold;

  H_1 : At least one of (A1)-(A4) does not hold.

Problem and Goals

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

  H_0 : Assumptions (A1)-(A4) all hold;

  H_1 : 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 : Assumptions (A1)-(A4) all hold;

  H_1 : 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

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

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

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

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 and p-value Deletion Statistic

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

$p[i] =$ associated p-value after the ith observation is excluded from the analysis.

- 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 $p[i]$ to assess their relative values.

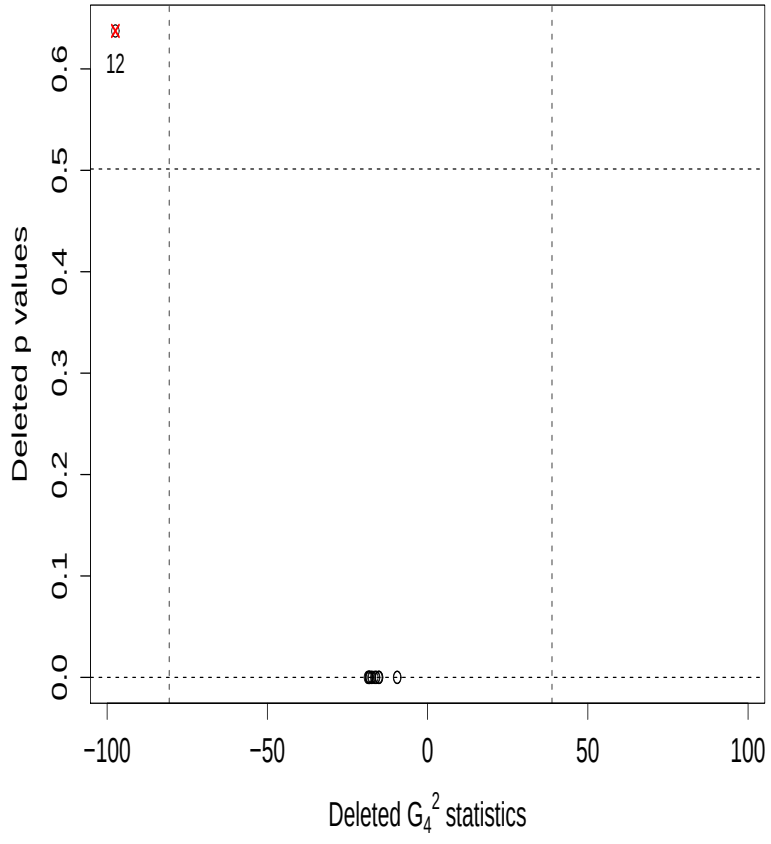
Applications to the Data Sets

Statistic	BP vs Pressure	Car Mileage
Global ($\hat{G}_4^2$)	$98.4(p = 0)$	$27.5(p \approx 0)$
$\hat{S}_1^2$	$28.7(p \approx 0)$	$.235(p = .63)$
$\hat{S}_2^2$	$65.1(p \approx 0)$	$0.1(p \approx 0)$
$\hat{S}_3^2$	$1.9(p = .17)$	$1.63(p = .20)$
$\hat{S}_4^2$	$2.8(p \approx .1)$	$.48(p = .48)$

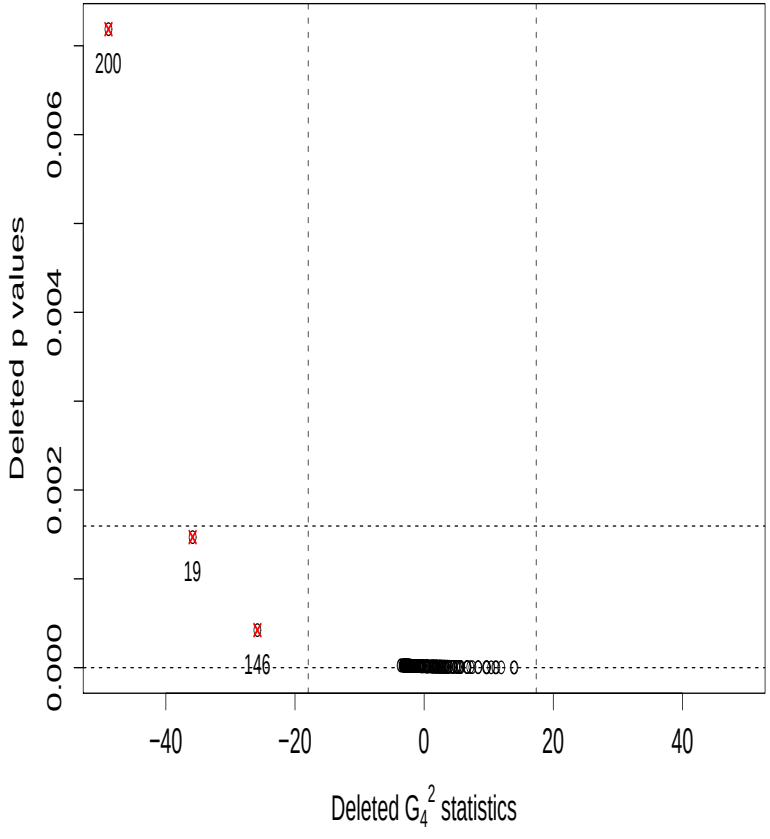
- Violations of model assumptions!
- For the BP vs Pressure data, problems with the normality assumption.
- For the car mileage data, second statistic indicates problems with normality.

Plots: Deletion Statistics

BP vs Pressure



Car Mileage



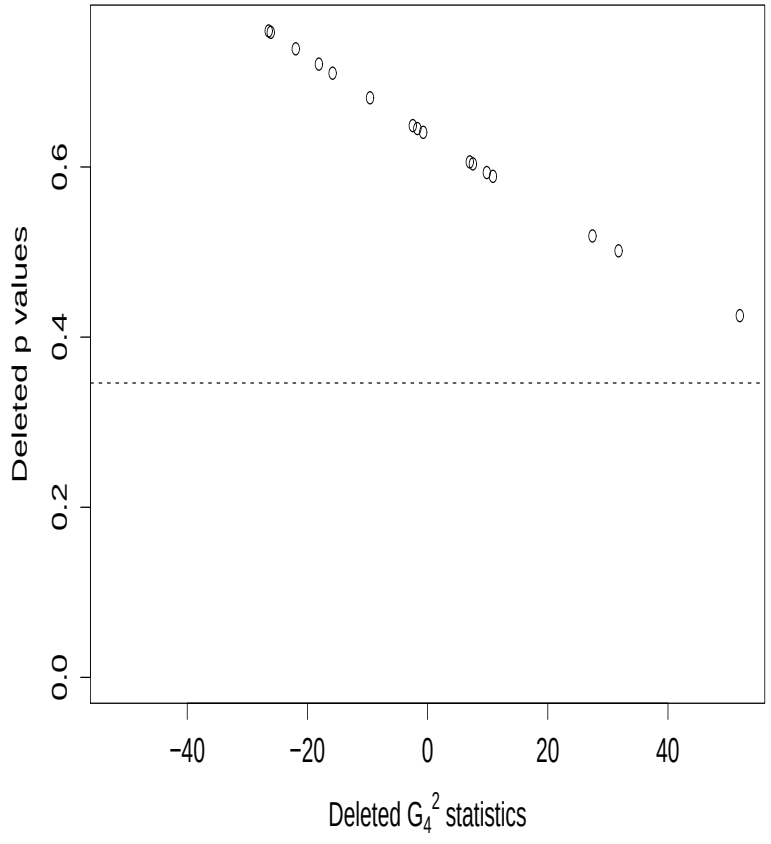
After Excluding Unusual Obs.

Statistic	BP vs Pressure	Car Mileage
Global ($\hat{G}_4^2$)	$2.54(p = .64)$	$.96(p = .92)$
$\hat{S}_1^2$	$1.06(p \approx .3)$	$.04(p = .92)$
$\hat{S}_2^2$	$.26(p \approx .61)$	$.002(p = .96)$
$\hat{S}_3^2$	$1.21(p = .27)$	$.71(p = .40)$
$\hat{S}_4^2$	$.01(p \approx .92)$	$.21(p = .65)$

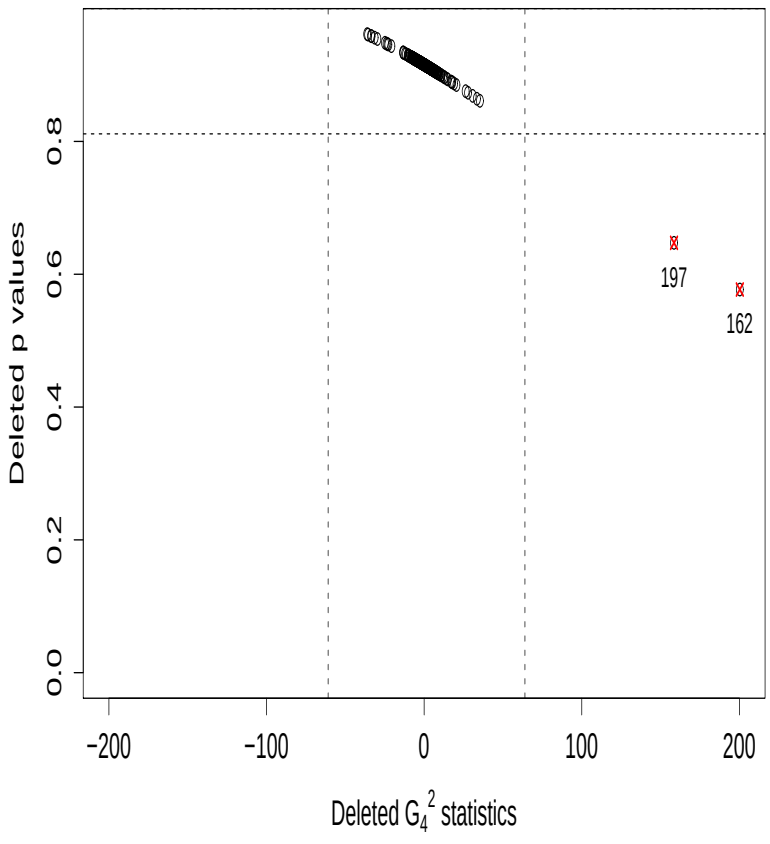
- For both data sets, when the unusual observations revealed by the deletion statistics are excluded, the global validation statistic does not reject the model assumptions.

Deletion Plots: After Exclusions!

BP vs Pressure



Car Mileage



Why It Works: 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.

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

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

Asymptotics: Parameters Estimated

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

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

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

Global Test Statistic

- The test statistic

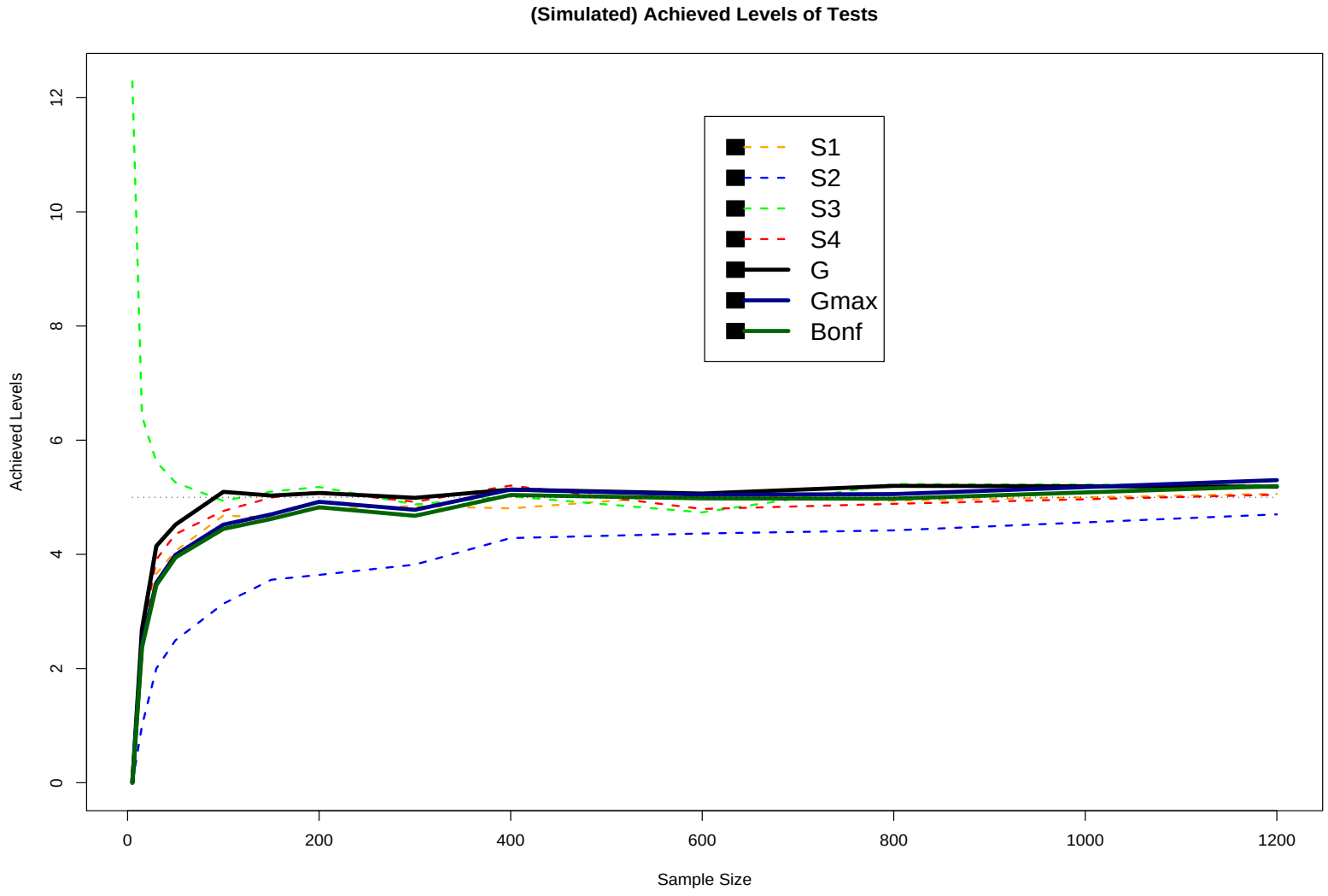
$$\frac{1}{n}\mathbf{Q}(\mathbf{R};s^2,\mathbf{b})^{\mathrm{t}}\hat{\boldsymbol{\Xi}}_{11.2}^{-}\mathbf{Q}(\mathbf{R};s^2,\mathbf{b}) = \hat{S}_1^2+\hat{S}_2^2+\hat{S}_3^2+\hat{S}_4^2 = \hat{G}_4^2$$

  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.
- $n \in \{30, 100\}$
- 20000 replications for level simulations; 5000 for power simulations
- $x_1, x_2, \ldots, x_n$ 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%
- Programs implementing the procedure were in an R code.

Achieved Levels



A Sampler of Simulated Powers

Viol	Para	n	$\hat{S}_1^2$	$\hat{S}_2^2$	$\hat{S}_3^2$	$\hat{S}_4^2$	$\hat{G}_4^2$
A4	t_5	30	21.6	21.1	6.0	10.6	23.9
		100	38.9	61.9	5.1	17.0	59.8
	χ_5^2	30	48.7	19.7	6.0	10.3	34.2
		100	98.7	57.8	5.8	14.5	92.5
A2	$\alpha = 2$	30	40	85	29	30	86
		100	49	100	15	28	99
	$\sigma_2 = 2$	30	13	12	5	40	27
		100	19	38	7	97	90
A1	$\beta_2 = 3$	30	3	1.7	19	4	8
	$\gamma = 2$	100	5	2.7	55	5	31
A3	MA	30	23.1	9.7	2.6	41.9	31.5
		100	55.0	38.4	4.0	72.1	75.2

A Never Ending Process!

- Research leads to further research problems: ‘peeling an onion!’. Some new problems arising from this research are:
- How to improve the asymptotic approximation?
- How to improve the power of the procedure?
- The use of a different ‘basis’ such as using wavelets in the density embedding?!
- Use the data to determine the components to use, hence an adaptive procedure. But, beware of the data double-dipping!
- Further studies on how to use deletion statistics.

A Final Word

- Research begets research!
- For those who are planning to pursue further studies, a good knowledge of mathematics, computers, probability, and statistics, as well as some applied science (e.g., biology), is a wonderful combination.