“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

Plot of Response Variable versus Predictor Variable

Plot of Response Variable versus Predictor Variable
**Simple Linear Regression Model**

- **Model:** \( Y_i = \beta_0 + \beta_1 X_i + \epsilon_i \)
- **Assumption:** \( \epsilon_i \)'s 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

Idaho State University Talk – p.
Summary of Regression Fits

<table>
<thead>
<tr>
<th>Estimator/Quantity</th>
<th>BP vs Pressure</th>
<th>Car Mileage</th>
</tr>
</thead>
<tbody>
<tr>
<td>$b_0$</td>
<td>$-42.1 (p = 0)$</td>
<td>$6.81 (p = 0)$</td>
</tr>
<tr>
<td>$b_1$</td>
<td>$0.89 (p = 0)$</td>
<td>$0.016 (p = 0)$</td>
</tr>
<tr>
<td>$\hat{\sigma}$</td>
<td>$0.38$</td>
<td>$0.591$</td>
</tr>
<tr>
<td>$R^2$</td>
<td>$0.995$</td>
<td>$0.426$</td>
</tr>
<tr>
<td>$F$</td>
<td>$2965 (p = 0)$</td>
<td>$150.7 (p = 0)$</td>
</tr>
</tbody>
</table>

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):

  \[ Y = X\beta + \sigma \epsilon \]

- \( Y \) = observable \( n \times 1 \) response vector;

- \( X \) = observable \( n \times p \) design matrix;

- \( \epsilon \) = unobservable error vector;

- \( \beta \) and \( \sigma \) are the parameters.
Linear Model and Assumptions

- **Linear Model (LM):**
  \[ Y = X\beta + \sigma \epsilon \]
- \( Y \) = observable \( n \times 1 \) response vector;
- \( X \) = observable \( n \times p \) design matrix;
- \( \epsilon \) = unobservable error vector;
- \( \beta \) and \( \sigma \) are the parameters.

(A1) **Linearity:**
\[ E\{Y_i|X\} = x_i\beta \]

(A2) **Homoscedasticity:**
\[ \text{Var}\{Y_i|X\} = \sigma^2 \]

(A3) **Uncorrelatedness:**
\[ \text{Cov}\{Y_i, Y_j|X\} = 0 \]

(A4) **Normality:**
\[ Y_i|X \sim \text{Normal}. \]
Estimators

- Estimator of $\beta$:

$$b = \hat{\beta} = (X^tX)^{-1}X^tY;$$

- Estimator of $\sigma^2$:

$$s^2 = \hat{\sigma}^2 = \frac{1}{n}Y^t(I - P_X)Y,$$

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

$$P_X = X(X^tX)^{-1}X^t$$
Validating LM Assumptions

- **Standardized Residuals:**

  \[ R = \frac{Y - Xb}{s} = \frac{(I - PX)Y}{s} \]

- **Graphical Methods.**

- **Diagnostic plots** based on \( R \). Discussed in many (elementary) textbooks!

- **Formal tests.**

- Such formal hypothesis tests are based on \( R \).
Example: Car Mileage 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,”
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 $R$. 
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 $R$.

- **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 \((Y, X)\), to test \textit{formally} and \textit{globally} the hypotheses

\[
H_0 : \text{Assumptions (A1)-(A4) all hold;}
\]
\[
H_1 : \text{At least one of (A1)-(A4) does not hold.}
\]
Problem and Goals

Based on \((Y, X)\), 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 \((Y, X)\), 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.
Recalling the standardized residuals

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

where \( \hat{Y}_i = x_i \beta \) is the \( i \)th fitted or predicted value.

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

\[ \hat{S}_3^2 = \left\{ \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2 R_i \right\}^2 \]

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

\[ \hat{\Gamma} = \frac{1}{n} \sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2 (x_i - \bar{x}) \]
4th Component Statistic

The fourth component statistic requires a user-supplied $n \times 1$ vector $V$, which by default is set to be the time sequence $V = (1, 2, \ldots, n)^t$. 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 \( \alpha \) is:

Reject \( H_0 \) 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

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

\( p[i] \) = associated \( p \)-value after the \( i \)th observation is excluded from the analysis.

- Percent relative change in value of global statistic \( \hat{G}^2_4 \) after deletion of \( i \)th observation.
- **Idea:** observation with a large absolute value of \( \Delta \hat{G}^2_4[i] \) is either an outlier or has large influence.
- Values of \( \Delta \hat{G}^2_4[i] \) can be plotted with respect to \( p[i] \) to assess their relative values.
Applications to the Data Sets

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

- **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

Deleted $G^2$ statistics

Deleted p values

Deleted $G^2$ statistics

Deleted p values
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.

<table>
<thead>
<tr>
<th>Statistic</th>
<th>BP vs Pressure</th>
<th>Car Mileage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Global $\hat{G}_4^2$</td>
<td>$2.54(p = .64)$</td>
<td>$.96(p = .92)$</td>
</tr>
<tr>
<td>$S_1^2$</td>
<td>$1.06(p \approx .3)$</td>
<td>$.04(p = .92)$</td>
</tr>
<tr>
<td>$S_2^2$</td>
<td>$.26(p \approx .61)$</td>
<td>$.002(p = .96)$</td>
</tr>
<tr>
<td>$S_3^2$</td>
<td>$1.21(p = .27)$</td>
<td>$.71(p = .40)$</td>
</tr>
<tr>
<td>$S_4^2$</td>
<td>$.01(p \approx .92)$</td>
<td>$.21(p = .65)$</td>
</tr>
</tbody>
</table>
Deletion Plots: After Exclusions!

BP vs Pressure

Car Mileage

 Deleted $G^2$ statistics

 Deleted p values

Deleted $G^2$ statistics

197

162
Why It Works: Theoretical Interludes

- True Residuals:

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

- \( R^0 \) are iid std normals.

- Density under \( H_0 \) of \( R^0 \):

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

- \( \phi(\cdot) = \) std normal pdf.
Why It Works: Theoretical Interludes

- True Residuals:

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

- True Residuals:

\[ R^0 \text{ are iid std normals.} \]

- Density under \( H_0 \) of \( R^0 \):

\[ f_{R^0}(r^0) = \prod_{i=1}^{n} \phi(r^0_i) \]

- \( \phi(\cdot) = \text{std normal pdf.} \)

Embedding Class:

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

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

The score test statistic within this embedding class for $H_0 : \theta = 0$ versus $H_1 : \theta \neq 0$ when $\beta$ and $\sigma$ are known is:

$$U(\theta = 0, \sigma^2, \beta) = Q(\mathbf{R}^0; \sigma^2, \beta).$$
The score test statistic within this embedding class for $H_0 : \theta = 0$ versus $H_1 : \theta \neq 0$ when $\beta$ and $\sigma$ are known is:

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

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

$$U(\theta = 0, s^2, b) = Q(R; s^2, b).$$
Score Test Statistic

- The score test statistic within this embedding class for $H_0 : \theta = 0$ versus $H_1 : \theta \neq 0$ when $\beta$ and $\sigma$ are known is:

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

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

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

- Needed: null asymptotic distribution of

$$Q(R; s^2, b).$$
Asymptotics: Parameters Known

Under $H_0$:

$$\frac{1}{\sqrt{n}} Q(R^0; \sigma^2, \beta) \xrightarrow{d} N(0, \Sigma_{11}(\sigma^2, \beta))$$

$$\Sigma_{11}(\sigma^2, \beta) = \begin{bmatrix}
1 & 0 & 3 & 0 & \beta^t \Sigma_X \beta & 0 \\
0 & 2 & 0 & 12 & 0 & 0 \\
3 & 0 & 15 & 0 & 3 \beta^t \Sigma_X \beta & 0 \\
0 & 12 & 0 & 96 & 0 & 0 \\
\beta^t \Sigma_X \beta & 0 & 3 \beta^t \Sigma_X \beta & 0 & \Omega(\beta) & 0 \\
0 & 0 & 0 & 0 & 0 & 2\sigma^2_V
\end{bmatrix}$$
Asymptotics: Parameters Estimated

Under $H_0$: \[ \frac{1}{\sqrt{n}} Q(R; s^2, b) \xrightarrow{d} N(0, \Xi_{11.2}(\sigma^2, \beta)) \]

\[ \Xi_{11.2}(\sigma^2, \beta) = \begin{bmatrix}
0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 6 & 0 & 0 & 0 \\
0 & 0 & 0 & 24 & 0 & 0 \\
0 & 0 & 0 & 0 & \xi(\sigma^2, \beta) & 0 \\
0 & 0 & 0 & 0 & 0 & 2\sigma^2_V
\end{bmatrix} \]

\[ \xi(\sigma^2, \beta) = \Omega(\beta) - (\beta^t \Sigma_X \beta)^2 - \Gamma(\beta) \Sigma_X^{-1} \Gamma(\beta)^t \]
Global Test Statistic

The test statistic

\[
\frac{1}{n} Q(R; s^2, b)^t \Xi_1 - 1.2 Q(R; s^2, b) = \hat{S}_1^2 + \hat{S}_2^2 + \hat{S}_3^2 + \hat{S}_4^2 = \hat{G}_4^2
\]

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 \( \mathbb{R} \) code.
### A Sampler of Simulated Powers

<table>
<thead>
<tr>
<th>Viol Para</th>
<th>$n$</th>
<th>$\hat{S}_1^2$</th>
<th>$\hat{S}_2^2$</th>
<th>$\hat{S}_3^2$</th>
<th>$\hat{S}_4^2$</th>
<th>$\hat{G}_4^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>A4 $t_5$</td>
<td>30</td>
<td>21.6</td>
<td>21.1</td>
<td>6.0</td>
<td>10.6</td>
<td>23.9</td>
</tr>
<tr>
<td></td>
<td>100</td>
<td>38.9</td>
<td>61.9</td>
<td>5.1</td>
<td>17.0</td>
<td>59.8</td>
</tr>
<tr>
<td>$\chi^2_5$</td>
<td>30</td>
<td>48.7</td>
<td>19.7</td>
<td>6.0</td>
<td>10.3</td>
<td>34.2</td>
</tr>
<tr>
<td></td>
<td>100</td>
<td>98.7</td>
<td>57.8</td>
<td>5.8</td>
<td>14.5</td>
<td>92.5</td>
</tr>
<tr>
<td>A2 $\alpha = 2$</td>
<td>30</td>
<td>40</td>
<td>85</td>
<td>29</td>
<td>30</td>
<td>86</td>
</tr>
<tr>
<td></td>
<td>100</td>
<td>49</td>
<td>100</td>
<td>15</td>
<td>28</td>
<td>99</td>
</tr>
<tr>
<td>$\sigma_2 = 2$</td>
<td>30</td>
<td>13</td>
<td>12</td>
<td>5</td>
<td>40</td>
<td>27</td>
</tr>
<tr>
<td></td>
<td>100</td>
<td>19</td>
<td>38</td>
<td>7</td>
<td>97</td>
<td>90</td>
</tr>
<tr>
<td>A1 $\beta_2 = 3$</td>
<td>30</td>
<td>3</td>
<td>1.7</td>
<td>19</td>
<td>4</td>
<td>8</td>
</tr>
<tr>
<td>$\gamma = 2$</td>
<td>100</td>
<td>5</td>
<td>2.7</td>
<td>55</td>
<td>5</td>
<td>31</td>
</tr>
<tr>
<td>A3 MA</td>
<td>30</td>
<td>23.1</td>
<td>9.7</td>
<td>2.6</td>
<td>41.9</td>
<td>31.5</td>
</tr>
<tr>
<td></td>
<td>100</td>
<td>55.0</td>
<td>38.4</td>
<td>4.0</td>
<td>72.1</td>
<td>75.2</td>
</tr>
</tbody>
</table>
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.