THE UNIVERSITY OF TEXAS AT AUSTIN — WHAT STARTS HERE CHANGES THE WORLD

CHE384, From Data to Decisions: Measurement, Uncertainty, Analysis, and Modeling

Lecture 37
Independence of Residuals

Chris A. Mack
Adjunct Associate Professor

http://www.lithoguru.com/scientist/statistics/

THE UNIVERSITY OF TEXAS AT AUSTIN — WHAT STARTS HERE CHANGES THE WORLD

Assumptions in OLS Regression

1. ε is a random variable that does not depend on x (i.e., the model is perfect, it properly accounts for the role of x in predicting y)
2. $\mathrm{E}[\varepsilon_i] = 0$ (the population mean of the true residual is zero); this will always be true for a model with an intercept
3. All ε_i are independent of each other (uncorrelated for the population, though not for a sample)
4. All ε_i have the same probability density function (pdf), and thus the same variance (called homoscedasticity)
5. $\varepsilon \sim \mathrm{N}(0,\sigma_\varepsilon)$ (the residuals, and thus the y_i, are normally distributed)
6. The values of each x_i are known exactly

THE UNIVERSITY OF TEXAS AT AUSTIN — WHAT STARTS HERE CHANGES THE WORLD

Residuals Are Not Independent

- Residuals are variation unexplained by the fitted model
- Assumption 3 (all residuals are independent of each other) is never true for a sample
  - Residuals depend on the fitted regression function, which depends on the same data that the residuals come from
  - With p fitting parameters, n residuals only have $n - p$ degrees of freedom
- For n sufficiently large compared to p, we can ignore this dependence

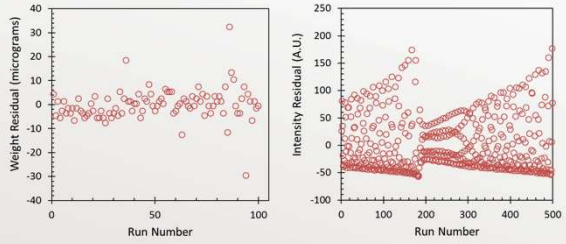
THE UNIVERSITY OF TEXAS AT AUSTIN — WHAT STARTS HERE CHANGES THE WORLD

Residuals in Sequence

- Other factors can prevent residual independence
  - Model error (usually a missing predictor variable)
  - Time dependence: sample aging, measurement drift
  - Spatial dependence: where the measurement was taken
- A run sequence plot shows the residuals in time sequence or other natural order to look for systematic variation
  - If time order corresponds to changing predictor values, the drift/aging may be hidden in the functional relationship
  - Randomization in experimental design prevents this
- A lag plot can make systematic variation more visible

THE UNIVERSITY OF TEXAS AT AUSTIN — WHAT STARTS HERE CHANGES THE WORLD

Run Sequence Plots

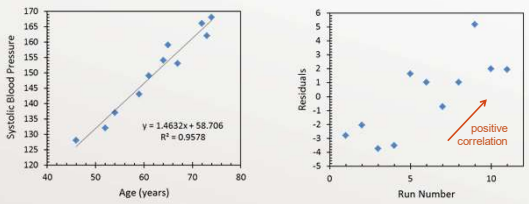


NBS measurements for a standard weight (Data Set 1)

Intensity measurements for a chaotic laser (Data Set 1)

THE UNIVERSITY OF TEXAS AT AUSTIN — WHAT STARTS HERE CHANGES THE WORLD

Run Sequence Plots

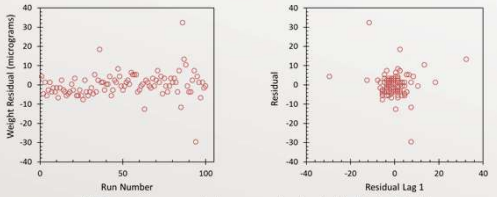


Here, run number is the patient in sequence of measurement (Data Set 2). (The small number of data points makes any conclusions tentative.)

THE UNIVERSITY OF TEXAS AT AUSTIN — WHAT STARTS HERE CHANGES THE WORLD

Lag Plots

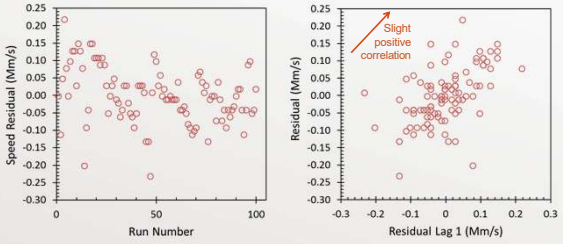
- A plot of ε_i versus ε_{i-1} (when residuals are ordered in time or other natural sequence) helps to discover correlations between a residual and its preceding residual
  - A lag of n plots ε_i versus ε_{i-n}



NBS measurements for a standard weight (Data Set 1)

THE UNIVERSITY OF TEXAS AT AUSTIN — WHAT STARTS HERE CHANGES THE WORLD

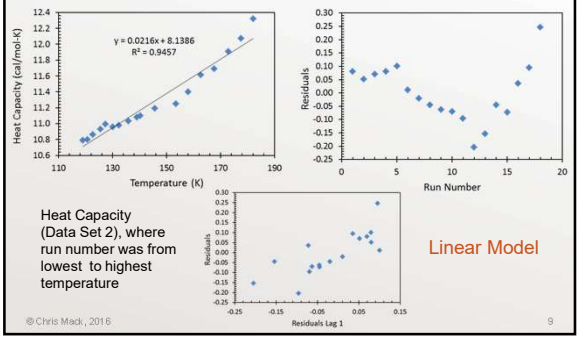
Lag Plots



Michelson speed of light measurements (Data Set 1)

THE UNIVERSITY OF TEXAS AT AUSTIN — WHAT STARTS HERE CHANGES THE WORLD

Correlated Residuals – Model Error

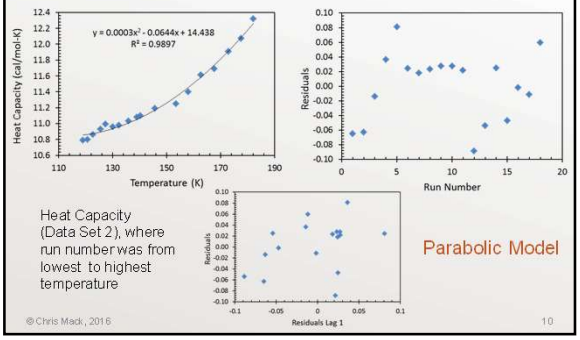


Heat Capacity (Data Set 2), where run number was from lowest to highest temperature

Linear Model

THE UNIVERSITY OF TEXAS AT AUSTIN — WHAT STARTS HERE CHANGES THE WORLD

Correlated Residuals – Model Error



Heat Capacity (Data Set 2), where run number was from lowest to highest temperature

Parabolic Model

THE UNIVERSITY OF TEXAS AT AUSTIN — WHAT STARTS HERE CHANGES THE WORLD

Runs Test – Is the Sequence Random?

- A random sequence of residuals will bounce back and forth between + and – according to a binomial distribution
  - Too many or two few bounces means the sequence is likely not random
- We define a "run" as any sequence (one or more in length) on the same side of 0
  - Sequence: ++-+-+++---++--++++-
  - This series has n_+ = 12, n_- = 8, and R = 10 runs

THE UNIVERSITY OF TEXAS AT AUSTIN — WHAT STARTS HERE CHANGES THE WORLD

Runs Test

- For about thirty or more residuals, we can perform the following test (usually one-tailed to test for positive correlation):
  - R = number of runs in the data
  - $\bar{R} = \frac{2n_+n_-}{n} + 1, \quad n = n_+ + n_-$
  - $s_R^2 = \frac{2n_+n_-(2n_+n_- - n)}{n^2(n-1)}$ Note: $\bar{R} \sim \frac{n}{2}$, $s_R \sim \frac{\sqrt{n}}{2}$
  - $Z_R = \frac{R-\bar{R}}{s_R}$ is about normally distributed

THE UNIVERSITY OF TEXAS AT AUSTIN — WHAT STARTS HERE CHANGES THE WORLD

Lecture 37: What have we learned?

- What can cause correlated (non-independent) residuals?
- Be able to generate and interpret a lag plot if the data sequence (order) is known
- Why is randomization of data order important in experimental design?
- What is a runs test and how is it performed?

