Assumptions in OLS Regression

1. $\epsilon$ 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. $E[\epsilon] = 0$ (the population mean of the true residual is zero); this will always be true for a model with an intercept
3. All $\epsilon$ are independent of each other (uncorrelated for the population, though not for a sample)
4. All $\epsilon$ have the same probability density function (pdf), and thus the same variance (called homoscedasticity)
5. $\epsilon \sim N(0, \sigma^2)$ (the residuals, and thus the $y_i$, are normally distributed)
6. The values of each $x_i$ are known exactly

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

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

Run Sequence Plots

- NBS measurements for a standard weight (Data Set 1)
- Intensity measurements for a chaotic laser (Data Set 1)
- Here, run number is the patient in sequence of measurement (Data Set 2). (The small number of data points makes any conclusions tentative.)
Lag Plots

• A plot of $e_i$ versus $e_{i-1}$ (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 $e_i$ versus $e_{i-n}$

NBS measurements for a standard weight (Data Set 1)

Michelson speed of light measurements (Data Set 1)

Correlated Residuals – Model Error

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

Linear Model

Parabolic Model

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

Runs Test

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

Note: $R \sim \binom{n}{\frac{n}{2}}$
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?