Data Analytics for Materials Science

Lecture 4 – Linear Regression in R

Amit K Verma

Revised: 10th Feb. 2021

Carnegie Mellon University

Objectives

- Math behind Linear Regression
    - Analytical Solution
    - Via Matrix Multiplication

- Train a Model within RStudio
    - Understanding the Model Summary
    - Evaluate the Model / Goodness of Fit

- Residuals – What they can tell us?
    - Switching Y and X
    - Class Exercise (at the end)

- A good reference book: http://databookuw.com/databook.pdf (chapter 4)

  Data Driven Science & Engineering (avail. in the CMU library)

  Machine Learning, Dynamical Systems, and Control

  By **Steven L. Brunton, J. Nathan Kutz**, Univ. of Washington

Carnegie Mellon University

Data

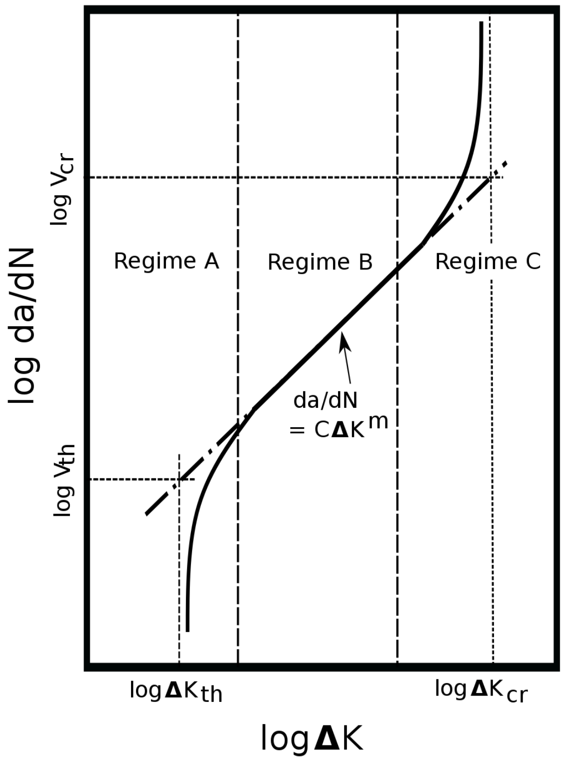
- Bayesian Neural Network Analysis of Fatigue Crack Growth Rate Nickel Base Superalloys – Hidetoshi FUJl; D. J. C. MACKAY; H. K. D. H. BHADESHIA **(1996)**

  - Modeling fatigue crack growth rate using a Neural Net within a Bayesian framework
  - 51 variables - Next Slide
  - Cause: Variations in both thermal and mechanical stress during the flight
    - Typical loading cycle comprises **starting up, takeoff and climb, cruising, landing and shut-down**
    - Highest stresses are experienced in the bore of the disc early in the flight cycle, generally while it is in the lower temperature range 200 – 300 ^{o}C
    - Stress in the rim region is lower, but at a higher temperature, 500 – 600 ^{o}C

  - However, the fatigue propagation is affected by many factors including chemical composition, grain size, heat treatment, temperature, atmosphere, R-ratio, frequency, …

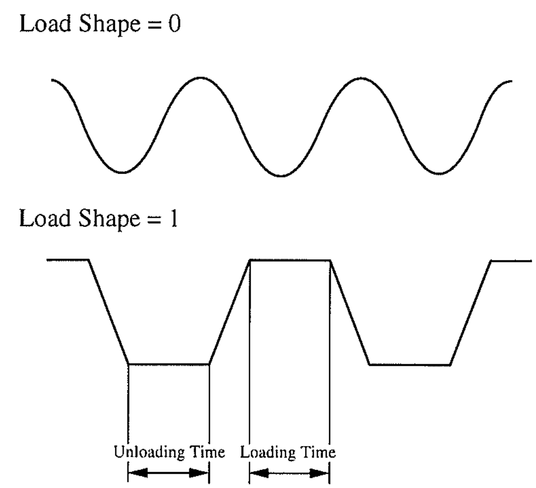
Carnegie Mellon University

Paris Law



Regime A – crack nucleation and initiation
Regime B – crack growth or propagation
Regime C – sudden fracture

Load Shape



- **Feature Engineering**
  - Is it a good representation?

Carnegie Mellon University

Data

	Variable	Range	Mean	Standard Deviation
Output	da/dN, µm	1.0×10^{-8} - 0.646	...	...
	log da/dN, µm	-8 - -0.1898	...	...
Stress Intensity Factor	ΔK, MPa $m^{-1/2}$	4.03 -246	27.16	22.47
	log ΔK, MPa $m^{-1/2}$	0.605 - 2.39	1.316	0.3167
	Temperature, K	293 - 1123	660.3	304.5
Microstructure	Minimum grain size, µm	7 - 5000	295.8	1024
	Maximum grain size, µm	7 - 5000	313.2	1022
	Difference in grain size between major phase and minor phase	-35 - 0	-0.8936	5.432
Heat Treatment	1st step Heat Treatment, Temperature, K	1116 - 1578	1321	103.0
	Duration, hour	0.5 - 7	2.602	1.685
	Cooling rate, K/sec	-15 - 5	-5.629	3.134
	2nd step Heat Treatment, Temperature, K	0 - 1413	955.8	292.1
	Duration, hour	0 - 24	12.22	9.065
	Cooling rate, K/sec	-5 - 0	-3.036	2.291
	3rd step Heat Treatment, Temperature, K	0 - 1143	869.3	312.0
	Duration, hour	0 - 24	10.96	6.179
	Cooling rate, K/sec	-5 - 0	-4.459	1.554
Load Waveform	Frequency, Hz	0.01 - 100	21.47	29.31
	Loading Time, s	0 - 600	15.27	71.96
	Unloading Time, s	0 - 500	7.439	55.14
	Load Shape	0 or 1	0.7355	0.4412
	Atmosphere	1×10^{-6} - 760	691.4	217.9
	R-ratio	0.05 - 0.8	0.171	0.2175
	Short or long crack growth	0 or 1	0.9161	0.2774
Properties	Sample thickness, mm	4.4 -25	11.39	4.063
	Yield Stress, MPa	324 - 1690	911.9	242.3

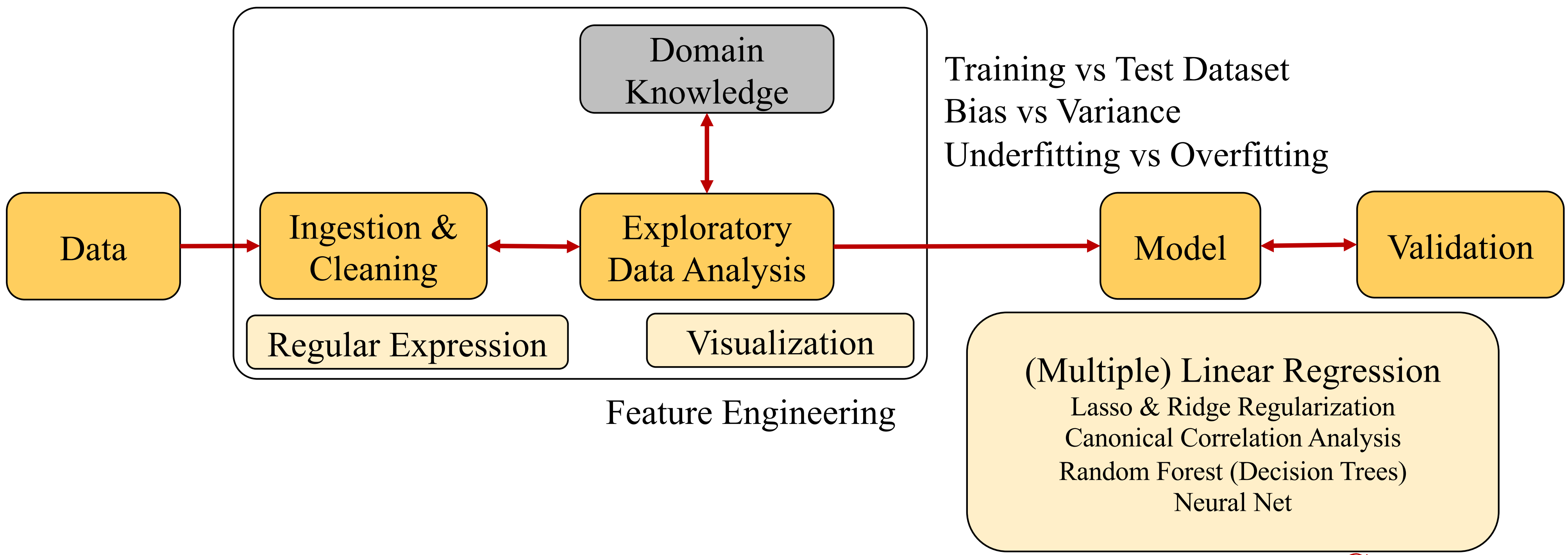
Environment

Composition

Variable	Range	Mean	Standard Deviation
Ni, wt%	40- 73	55.34	8.234
Cr	0.03 - 19.5	14.89	5.127
Co	0 - 17	5.982	7.552
Mo	0 - 6	3.094	1.991
Al	0.3 - 5.5	1.926	1.732
Ti	0.8 - 3.52	2.107	1.098
Fe	0 - 35.56	12.07	12.12
C	0.007 - 0.06	0.03865	0.01176
B	0 - 0.1	0.01418	0.02604
Zr	0 - 0.35	0.01907	0.04495
Si	0 - 0.31	0.05634	0.08841
Nb	0 - 5.35	1.968	2.402
Mn	0 - 0.28	0.03728	0.07740
Cu	0 - 0.06	4.525×10^{-3}	0.01401
P	0 - 0.011	8.551×10^{-4}	2.438×10^{-3}
Ca	0 - 0.006	2.598×10^{-4}	1.221×10^{-3}
Mg	0 - 0.002	8.659×10^{-5}	4.071×10^{-4}
S	0 - 0.005	3.350×10^{-4}	1.031×10^{-3}
Sn	0 - 0.0027	1.169×10^{-4}	5.496×10^{-4}
Pb	0 - 0.00004	1.732×10^{-6}	8.143×10^{-6}
Bi	0 - 0.0000125	5.412×10^{-7}	2.545×10^{-6}
Ag	0 - 0.00001	4.329×10^{-7}	2.036×10^{-6}
W	0 - 6.5	0.4628	1.539
Ta	0 - 6.5	0.3935	1.385
Hf	0 - 0.1	4.488×10^{-3}	0.02071
Re	0 - 3	0.1346	0.6213
Y_2O_3	0 - 1.1	0.04704	0.2226

https://www.jstage.jst.go.jp/article/isijinternational1989/36/11/36_11_1373/_article/-char/en

Exploratory Analysis Pipeline



Carnegie Mellon University

Linear Regression

$$y = \beta_0 + \beta_1 x + \varepsilon$$

$\log_{10}(da/dN) = \beta_0 + \beta_1 x + \varepsilon$;   which x?

Paris Law

$$y = \beta_0 + \beta_1 \text{ x} + \varepsilon$$

$$\log_{10}(da/dN) = \beta_0 + \beta_1 \text{ x} + \varepsilon;$$ which x?

$$\log_{10}(da/dN) = \beta_0 + \beta_1 \log_{10}(\Delta K) + \varepsilon;$$

ON!

Carnegie Mellon University

Error Metrics

$$y = \beta_0 + \beta_1 x + \varepsilon$$

$$\log_{10} (da/dN) = \beta_0 + \beta_1 x + \varepsilon; \qquad \text{which x?}$$

$$\log_{10} (da/dN) = \beta_0 + \beta_1 \log_{10} (\Delta K) + \varepsilon;$$

$$y_k = f(x_k) + \varepsilon_k;$$

$$E_\infty(f) = \max_{1<k<n} |f(x_k) - y_k| \quad \text{Maximum Error } (\ell_\infty)$$

$$E_1(f) = \frac{1}{n} \sum_{k=1}^{n} |f(x_k) - y_k| \quad \text{Mean Absolute Error } (\ell_1)$$

$$\boxed{E_2(f) = \left(\frac{1}{n} \sum_{k=1}^{n} |f(x_k) - y_k|^2 \right)^{1/2} \quad \text{Least-squares Error } (\ell_2)}$$

ℓ_p-norm

$$E_p(f) = \left(\frac{1}{n} \sum_{k=1}^{n} |f(x_k) - y_k|^p \right)^{1/p}$$

Carnegie Mellon University

Solving for Coefficients: Minimize(Least Squares Error)

$$
\begin{aligned}
\frac{\partial E_2}{\partial \beta_1} = 0: &\quad \sum_{k=1}^{n} 2(\beta_1 x_k + \beta_2 - y_k)x_k = 0 \\
\frac{\partial E_2}{\partial \beta_2} = 0: &\quad \sum_{k=1}^{n} 2(\beta_1 x_k + \beta_2 - y_k) = 0\,.
\end{aligned}
$$

(upon rearranging) in matrix form

$$
\begin{pmatrix} \sum_{k=1}^{n} x_k^2 & \sum_{k=1}^{n} x_k \\ \sum_{k=1}^{n} x_k & n \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix} = \begin{pmatrix} \sum_{k=1}^{n} x_k y_k \\ \sum_{k=1}^{n} y_k \end{pmatrix} \longrightarrow \mathbf{Ax} = \mathbf{b}
$$

Analytical solution

Using Matrix Multiplication

$$\begin{aligned} Y_1 &= \beta_0 + \beta_1 X_1 + \epsilon_1 \\ Y_2 &= \beta_0 + \beta_1 X_2 + \epsilon_2 \\ \vdots & \quad \vdots \quad \vdots \\ Y_n &= \beta_0 + \beta_1 X_n + \epsilon_n \end{aligned} \Rightarrow \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} \beta_0 + \beta_1 X_1 \\ \beta_0 + \beta_1 X_2 \\ \vdots \\ \beta_0 + \beta_1 X_n \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix} \Rightarrow \begin{aligned} \mathbf{Y} &= \mathbf{X}\beta + \epsilon \\ \mathbf{Y}_{n\times 1} &= \mathbf{X}_{n\times 2}\beta_{2\times 1} + \epsilon_{n\times 1} \end{aligned}$$

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} 1 & X_1 \\ 1 & X_2 \\ \vdots & \vdots \\ 1 & X_n \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

$$\mathbf{Y}_{n\times 1} \qquad \mathbf{X}_{n\times 2} \qquad \beta_{2\times 1} \qquad \epsilon_{n\times 1}$$

least squares error

$$\sum \epsilon_i^2 = [\epsilon_1 \, \epsilon_2 \, \cdots \epsilon_n] \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix} = \epsilon' \epsilon = (\mathbf{Y} - \mathbf{X}\beta)'(\mathbf{Y} - \mathbf{X}\beta)$$

derivative

$$\frac{d}{d\beta}\left((\mathbf{Y} - \mathbf{X}\beta)'(\mathbf{Y} - \mathbf{X}\beta)\right) = -2\mathbf{X}'(\mathbf{Y} - \mathbf{X}\beta)$$

set derivative to 0

$$-2\mathbf{X}'(\mathbf{Y} - \mathbf{X}\beta) = \mathbf{0}$$

$$\boxed{\mathbf{X}'\mathbf{Y} = \mathbf{X}'\mathbf{X}\beta}$$

Carnegie Mellon University

"prime" – denotes the transpose of the matrix (exchange the rows and columns). Same as superscript "T"

We will have an equation that looks like $\hat{\boldsymbol{\beta}}^T \mathbf{X}^T \mathbf{X} = \mathbf{Y}^T \mathbf{X}$ and will want $\hat{\boldsymbol{\beta}}$, knowing $\mathbf{X}$ and $\mathbf{Y}$.

How will we solve it?

Taking the transpose of both sides of the equation, we have

$$\left(\hat{\boldsymbol{\beta}}^T \mathbf{X}^T \mathbf{X}\right)^T = \mathbf{X}^T \mathbf{X} \hat{\boldsymbol{\beta}} \text{ and } (\mathbf{Y}^T \mathbf{X})^T = \mathbf{X}^T \mathbf{Y}, \text{ or } \mathbf{X}^T \mathbf{X} \hat{\boldsymbol{\beta}} = \mathbf{X}^T \mathbf{Y}$$

Finally, we have: $\hat{\boldsymbol{\beta}} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y}$.

Knowing $\mathbf{X}$ and $\mathbf{Y}$, we can easily find $\hat{\boldsymbol{\beta}}$.

From L3B

Coefficients & Null Hypothesis

```
Call:
lm(formula = log10.da.dN. ~ log10.delta.K., data = temp)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.16209 -0.02413  0.01367  0.03109  0.07302 

Coefficients:
               Estimate Std. Error t value Pr(>|t|)    
(Intercept)   -10.40212    0.13685  -76.01   <2e-16 ***
log10.delta.K.  4.40488    0.09176   48.01   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.05509 on 30 degrees of freedom
Multiple R-squared:  0.9871,	Adjusted R-squared:  0.9867 
F-statistic:  2305 on 1 and 30 DF,  p-value: < 2.2e-16
```

$$\log_{10}(da/dN) = \beta_0 + \beta_1 \log_{10}(\Delta K) + \varepsilon$$

$$\log_{10}(da/dN) = -10.40 + 4.40 \log_{10}(\Delta K) + \varepsilon$$

p-value comes from Null Hypothesis
(there is no relationship between X and Y)

p-value indicates whether you can reject or accept a hypothesis

General Rule: $p < 0.05$
to reject the null hypothesis

Carnegie Mellon University

Residuals

$$\log_{10}(da/dN) = -10.40 + 4.40 \log_{10}(\Delta K) + \varepsilon$$

$$y_k = f(x_k) + \varepsilon_k;\ \text{res} = \mathrm{y_k} - f(x_k)$$

ON!

```
Call:
lm(formula = log10.da.dN. ~ log10.delta.K., data = temp)

Residuals:
     Min       1Q   Median       3Q      Max
-0.16209 -0.02413  0.01367  0.03109  0.07302

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)    -10.40212    0.13685  -76.01   <2e-16 ***
log10.delta.K.   4.40488    0.09176   48.01   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.05509 on 30 degrees of freedom
Multiple R-squared:  0.9871,    Adjusted R-squared:  0.9867
F-statistic:  2305 on 1 and 30 DF,  p-value: < 2.2e-16
```

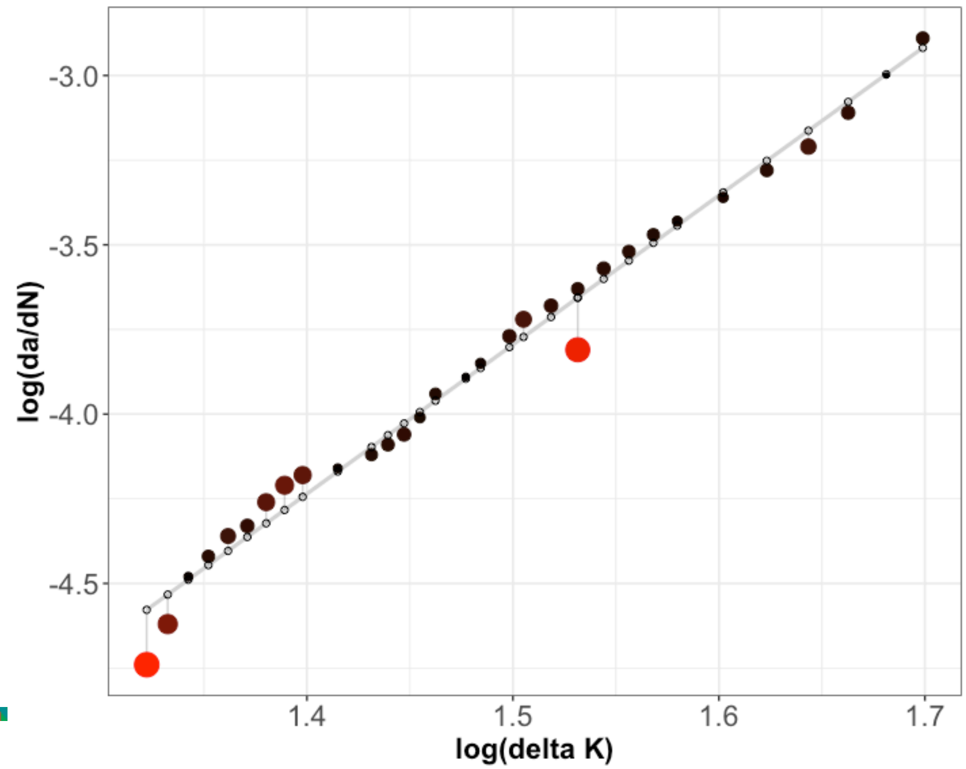
Carnegie Mellon University

Residuals

```
Call:
lm(formula = log10.da.dN. ~ log10.delta.K., data = temp)

Residuals:
     Min       1Q   Median       3Q      Max
-0.16209 -0.02413  0.01367  0.03109  0.07302

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)    -10.40212    0.13685  -76.01   <2e-16 ***
log10.delta.K.   4.40488    0.09176   48.01   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.05509 on 30 degrees of freedom
Multiple R-squared:  0.9871,    Adjusted R-squared:  0.9867
F-statistic:  2305 on 1 and 30 DF,  p-value: < 2.2e-16
```

$$\log_{10}(da/dN) = -10.40 + 4.40 \log_{10}(\Delta K) + \varepsilon$$

$$y_k = f(x_k) + \varepsilon_k; \text{ res} = y_k - f(x_k)$$



Residual Sum of Squares (RSS) = $(\text{res}_1)^2 + (\text{res}_2)^2 + \ldots\ldots (\text{res}_n)^2$

Fitting = Minimize (RSS)

Measure (/Goodness) of Fit

```
Call:
lm(formula = log10.da.dN. ~ log10.delta.K., data = temp)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.16209 -0.02413  0.01367  0.03109  0.07302 

Coefficients:
                Estimate Std. Error t value Pr(>|t|)    
(Intercept)    -10.40212    0.13685  -76.01   <2e-16 ***
log10.delta.K.   4.40488    0.09176   48.01   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.05509 on 30 degrees of freedom
Multiple R-squared:  0.9871,	Adjusted R-squared:  0.9867 
F-statistic:  2305 on 1 and 30 DF,  p-value: < 2.2e-16
```

Residual Sum of Squares (RSS) = $(\text{res}_1)^2 + (\text{res}_2)^2 + \ldots\ldots (\text{res}_n)^2$

Fitting = Minimize (RSS)

$$\log_{10}(da/dN) = -10.40 + 4.40 \log_{10}(\Delta K) + \varepsilon$$

$$y_k = f(x_k) + \varepsilon_k;\ \text{res} = \text{y}_\text{k} - f(x_k)$$

$$\text{RSE} = \sqrt{\frac{1}{n-2}\text{RSS}}$$

Avg. amount that y deviates from regression line; Absolute measure of lack of fit

$$R^2 = \frac{\text{TSS} - \text{RSS}}{\text{TSS}}; \quad \text{TSS} = \sum (y_i - \bar{y})^2$$

Proportion of variability in $\log_{10}(da/dN)$ that can be explained using $\log_{10}(\Delta K)$

ON!

Carnegie Mellon University

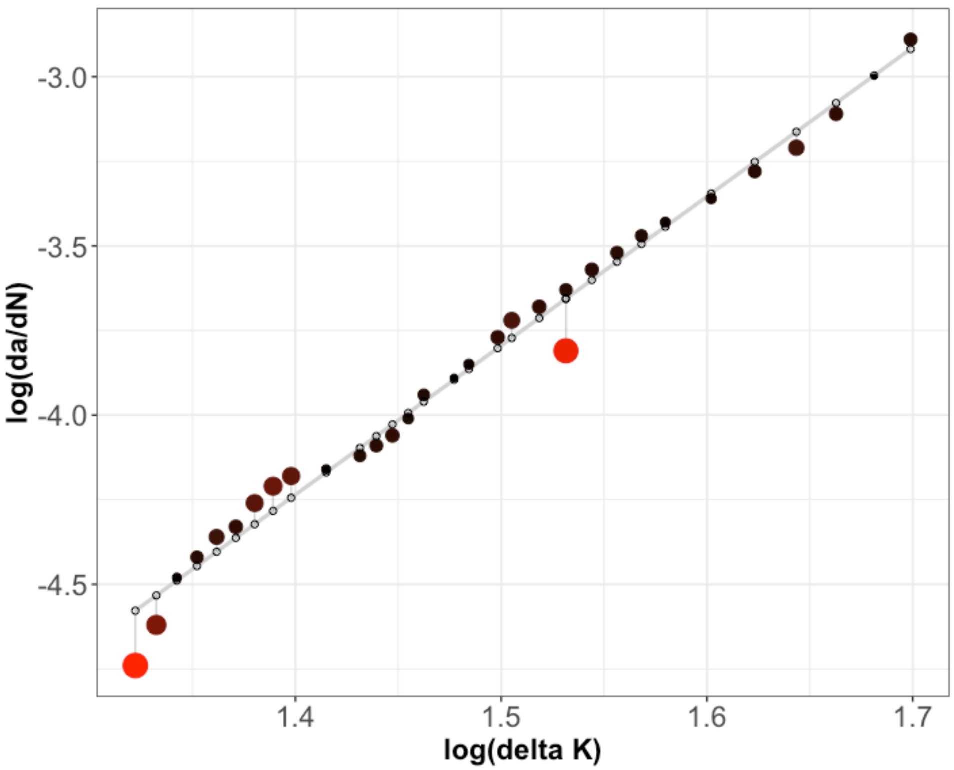
Switching Y and X

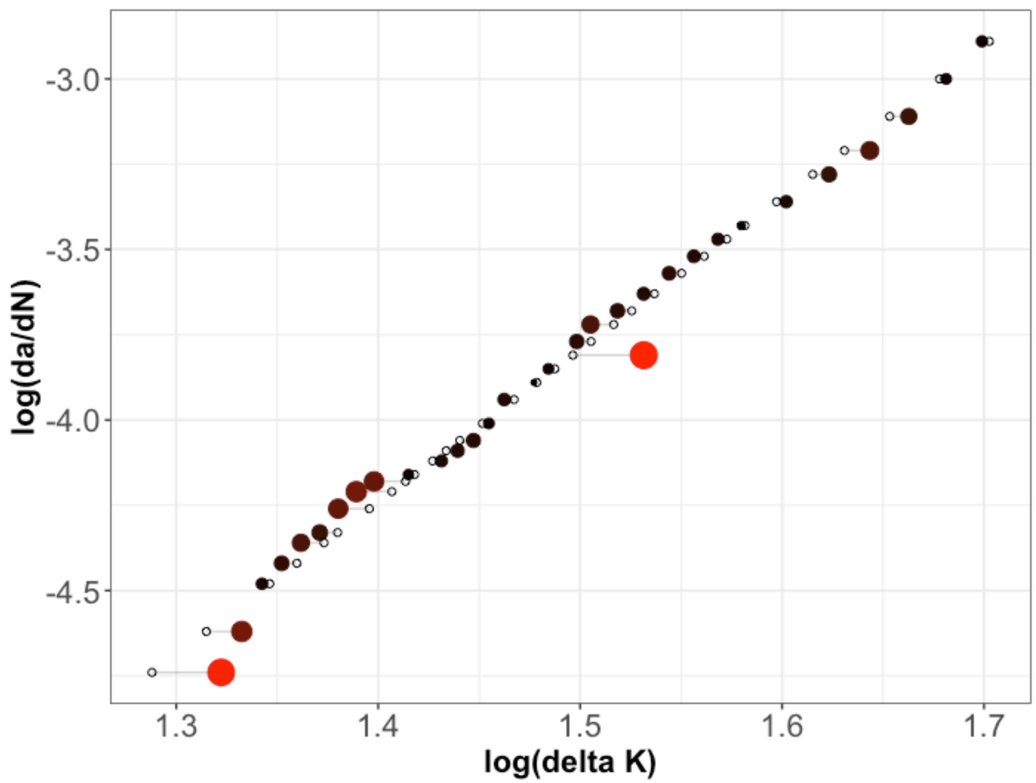
```
Call:
lm(formula = log10.da.dN. ~ log10.delta.K., data = temp)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.16209 -0.02413  0.01367  0.03109  0.07302 

Coefficients:
                Estimate Std. Error t value Pr(>|t|)    
(Intercept)    -10.40212    0.13685  -76.01   <2e-16 ***
log10.delta.K.   4.40488    0.09176   48.01   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.05509 on 30 degrees of freedom
Multiple R-squared:  0.9871,	Adjusted R-squared:  0.9867 
F-statistic:  2305 on 1 and 30 DF,  p-value: < 2.2e-16
```

```
Call:
lm(formula = log10.delta.K. ~ log10.da.dN., data = temp)

Residuals:
      Min        1Q    Median        3Q       Max 
-0.017630 -0.007066 -0.003409  0.004992  0.035041 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)  2.350274   0.018102  129.84   <2e-16 ***
log10.da.dN. 0.224104   0.004668   48.01   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.01243 on 30 degrees of freedom
Multiple R-squared:  0.9871,	Adjusted R-squared:  0.9867 
F-statistic:  2305 on 1 and 30 DF,  p-value: < 2.2e-16
```

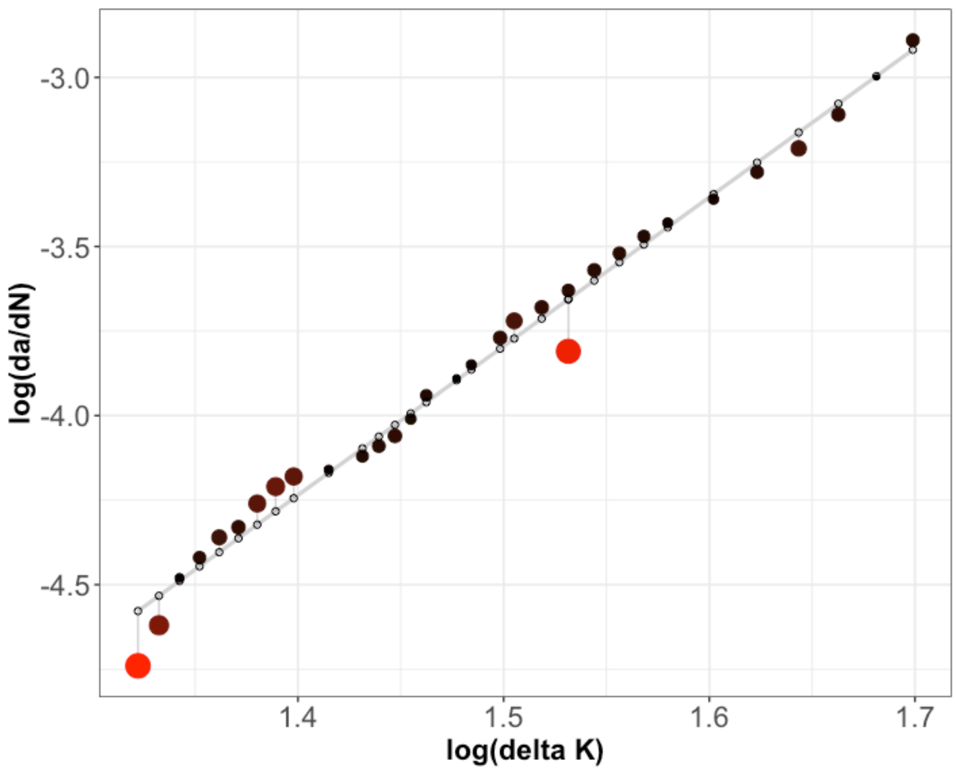
Carnegie Mellon University

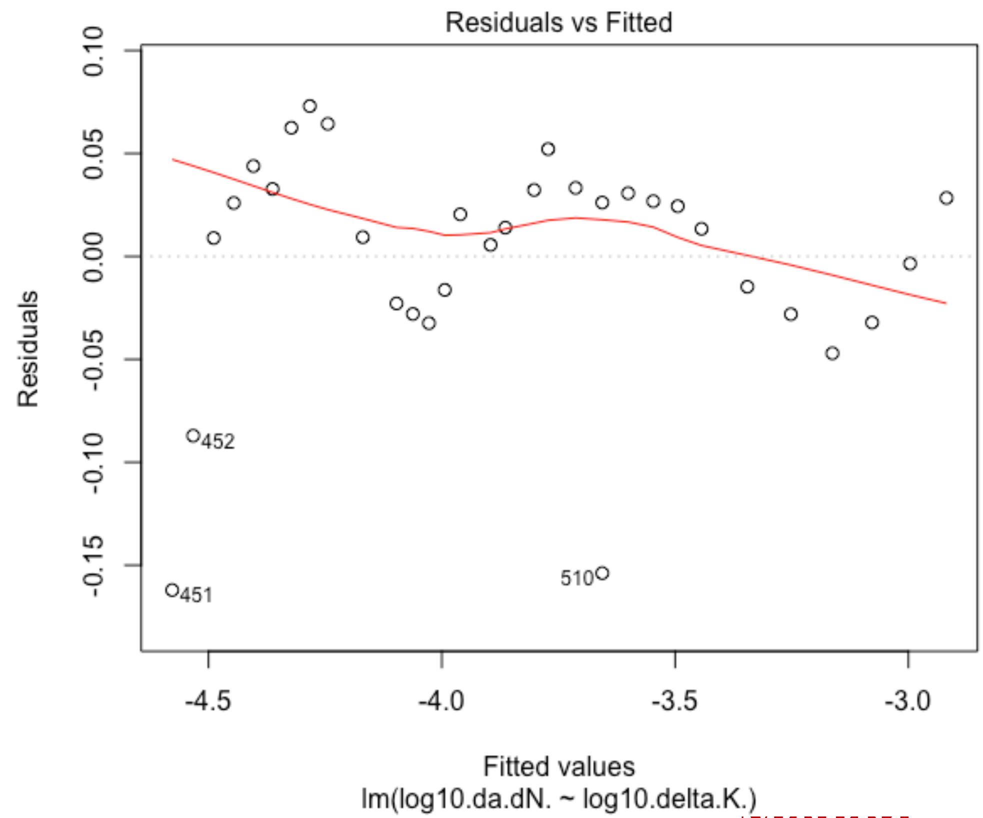
Switching Y and X





More on Residuals





Underlying pattern highlights non-linearity in data

Multiple Linear Regression

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 \,\ldots\ldots + \beta_n x_n + \varepsilon$$

$$\log_{10} (da/dN) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 \,\ldots\ldots + \beta_n x_n + \varepsilon;$$

all vs subset? Composition; Heat Treatment; Grain Size; Temperature

Next Week!

Questions