Two-Factor Full Factorial Design without Replications

Dr. John Mellor-Crummey

Department of Computer Science
Rice University

johnmc@cs.rice.edu

Goals for Today

Understand

- **Two-factor Full Factorial Design Without Replications**
  - —motivation
  - —properties
  - —estimating model parameters to compute effects
  - —estimating experimental errors
  - —allocating variation
  - —analyzing significance of factors
  - —confidence intervals
  - —analysis with multiplicative models (revisited)
  - —analysis with missing data

Two-factor Full Factorial Design

- **Why use them?**
  - —carefully control and vary two parameters
  - —study their impact on performance
- **Example: compare several processors using several workloads**
  - —factor A: processor type
  - —factor B: workload type
  - —full factorial design → study all workloads x processors
- **No limits on number of levels each factor can take**
- **Model:** $y_{ij} = \mu + \alpha_j + \beta_i + e_{ij}$
  - —y_{ij} - response with factor A at level j and factor B at level i
  - —μ - mean response
  - —α_j - effect of level j for factor A
  - —β_j - effect of level i for factor B
  - —e_{ij} - error term
- **All effects are computed so that their sum is 0:**

$$\sum_j \alpha_j = 0 \qquad \sum_j \beta_j = 0$$

Estimating Model Parameters

- **Organize measured data for two-factor full factorial design as**
  - **b x a matrix: (i,j) = factor B at level i and factor A at level j**

    columns = levels of factor A **rows = levels of factor B**

- **Compute parameters so sum of errors in row and column = 0**
- **Compute a column mean**

$$
\begin{aligned}
\bar{y}_{.j} &= \frac{1}{b}\sum_{i=1}^{b}\mu + \frac{1}{b}\sum_{i=1}^{b}\alpha_j + \frac{1}{b}\sum_{i=1}^{b}\beta_i + \frac{1}{b}\sum_{i=1}^{b}e_{ij} \\
&= \mu + \alpha_j + 0 + 0
\end{aligned}
$$

- **Compute a row mean**

$$
\begin{aligned}
\bar{y}_{i.} &= \frac{1}{a}\sum_{j=1}^{a}\mu + \frac{1}{a}\sum_{j=1}^{a}\alpha_j + \frac{1}{a}\sum_{j=1}^{a}\beta_i + \frac{1}{a}\sum_{j=1}^{a}e_{ij} \\
&= \mu + 0 + \beta_i + 0
\end{aligned}
$$

- **Estimate model parameters**

$$
\mu = \bar{y}_{..} \quad \alpha_j = \bar{y}_{.j} - \bar{y}_{..} \quad \beta_i = \bar{y}_{i.} - \bar{y}_{..}
$$

Example: Pgms vs. Cache Configuration

Workload	Two Caches	One Cache	No Cache	Row Sum	Row Mean	Row Effect
ASM	54.0	55.0	106.0	215.0	71.7	-0.5
TECO	60.0	60.0	123.0	243.0	81.0	8.8
SIEVE	43.0	43.0	120.0	206.0	68.7	-3.5
DHRYSTONE	49.0	52.0	111.0	212.0	70.7	-1.5
SORT	49.0	50.0	108.0	207.0	69.0	-3.2
Column Sum	255.0	260.0	568.0	1083.0		
Column Mean	51.0	52.0	113.6		72.2	
Column Effect	-21.2	-20.2	41.4			

- **Computing effects**
  - —column effect = col mean - mean; row effect = row mean - mean
- **Interpretation**
  - —average workload on average processor 72.2ms
  - —caches affect time
    - – two caches: 21.2ms < avg processor
    - – one cache: 20.2ms < avg processor
    - – no cache: 41.4ms > avg processor
    - – mean difference between one cache and no cache is 61.6ms
  - —workloads affect time: average pgm 72.2ms
    - – ASM .5ms < avg pgm; TECO takes 8.8ms > avg pgm, etc

Estimating Experimental Errors

- **Predicted response of (i,j)th experiment**

$$\hat{y}_{ij} = \mu + \alpha_j + \beta_i$$

- **Prediction error** $e_{ij} = y_{ij} - \hat{y}_{ij} = y_{ij} - \mu - \alpha_j - \beta_i$

$$e_{ij} = y_{ij} - \mu - \alpha_j - \beta_i$$

Workload	Two Caches	One Cache	No Cache
ASM	3.5	3.5	-7.1
TECO	0.2	-0.8	0.6
SIEVE	-4.5	-5.5	9.9
DHRYSTONE	-0.5	1.5	-1.1
SORT	1.2	1.2	-2.4

=

Two Caches	One Cache	No Cache
54.0	55.0	106.0
60.0	60.0	123.0
43.0	43.0	120.0
49.0	52.0	111.0
49.0	50.0	108.0

- 72.2 -

Row Effect	Row Effect	Row Effect
-0.5	-0.5	-0.5
8.8	8.8	8.8
-3.5	-3.5	-3.5
-1.5	-1.5	-1.5
-3.2	-3.2	-3.2

-

column effect	column effect	column effect
-21.2	-20.2	41.4
-21.2	-20.2	41.4
-21.2	-20.2	41.4
-21.2	-20.2	41.4
-21.2	-20.2	41.4

- **Estimate variance of errors from sum of squared errors**

$$SSE = \sum_{i=1}^{b} \sum_{j=1}^{a} e_{ij}^2$$

$$SSE = (3.5)^2 + (0.2)^2 + \ldots + (-2.4)^2 = 236.8$$

Allocating Variation

- **Total variation of y can be allocated to the factors and errors**
  - —like one factor designs and 2^2r designs
- **Square both sides of model equation**

$$\sum_{i,j} y_{ij}^2 = \sum_{i,j} \mu^2 + \sum_{i,j} \alpha_j^2 + \sum_{i,j} \beta_i^2 + \sum_{i,j} e_{ij}^2 + \boxed{\text{cross product terms that are all 0}}$$

SSY = SS0 + SSA + SSB + SSE

- **Total variation**

SST = SSY - SS0 = SSA + SSB + SSE

$$SST = \sum_{i,j} (y_{ij} - \bar{y}_{..})^2 = \sum_{i,j} y_{ij}^2 - ab\bar{y}_{..}^2 = SSY - SS0 = SSA + SSB + SSE$$

- **Percent variation for each factor and errors**
  - —factor A: 100(SSA/SST)
  - —factor B: 100(SSB/SST)
  - —error: 100(SSE/SST)

Allocating Variation for Pgm vs. Cache Study

Workload	Two Caches	One Cache	No Cache	Row Sum	Row Mean	Row Effect
ASM	54.0	55.0	106.0	215.0	71.7	-0.5
TECO	60.0	60.0	123.0	243.0	81.0	8.8
SIEVE	43.0	43.0	120.0	206.0	68.7	-3.5
DHRYSTONE	49.0	52.0	111.0	212.0	70.7	-1.5
SORT	49.0	50.0	108.0	207.0	69.0	-3.2
Column Sum	255.0	260.0	568.0	1083.0		
Column Mean	51.0	52.0	113.6		72.2	
Column Effect	-21.2	-20.2	41.4			

$$SSY = \sum_{i,j} y_{ij}^2 = 54^2 + 60^2 + \ldots + 108^2 = 91{,}595$$

$$SS0 = ab\mu^2 = (3)(5)(72.2)^2 = 78{,}192.60$$

$$SSA = b\sum_j \alpha_j^2 = 5((-21.2)^2 + (-20.2)^2 + (41.4)^2) = 12{,}857.20$$

$$SSB = a\sum_j \beta_j^2 = 3((-0.5)^2 + (8.8)^2 + (-3.5)^2 + (-1.5)^2 + (3.2)^2) = 308.40$$

$$SST = SSY - SS0 = 91{,}595 - 78192.6 = 13{,}402.4$$

% variation caches = 100*SSA/SST = 100*12,857.2/13,402.4 = 95.9%
% variation pgms = 100*SSB/SST = 100*308.4/13,402.4 = 2.3%

Computing Mean Squares from Variation

- **Degrees of freedom**

$$\begin{array}{ccccccccccc} SSY & = & SS0 & + & SSA & + & SSB & + & SSE \\ ab & = & 1 & + & (a-1) & + & (b-1) & + & (a-1)(b-1) \end{array}$$

sum of ab obs.
all independent

single term μ^2:
repeated ab times

sum of $(\alpha_j)^2$:
a-1 independent
($\sum\alpha_j = 0$)

sum of $(\beta_i)^2$:
b-1 independent
($\sum\beta_i = 0$)

sum of $(e_{ij})^2$,
(a-1)(b-1)
independent
0 across row,
down column

- **Divide the sum of squares by their DOF to get mean squares**
  - MSA = SSA/(a-1)
  - MSB = SSB/(b-1)
  - MSE = SSE/((a-1)(b-1))

Analyzing Significance of Factors

- **Mean square values**
  - —MSA = SSA/(a-1)
  - —MSB = SSB/(b-1)
  - —MSE = SSE/((a-1)(b-1))

$$s_e = \sqrt{MSE}$$

- **Are effects of the factors significant?**
  - —if errors normally distributed →

    SSE, SSA, SSB have chi-square distributions
  - — MSA/MSE = [SSA/(a-1)]/[SSE/((a-1)(b-1))] has F distribution
    - numerator has ν_A = (a-1) degrees of freedom
    - denominator has ν_E = (a-1)(b-1) degrees of freedom
    - significant if MSA/MSE > $F_{[1-\alpha;\ \nu A;\ \nu E]} = F_{[1-\alpha;\ a-1;\ (a-1)(b-1)]}$
  - —MSB/MSE = [SSB/(b-1)]/[SSE/((a-1)(b-1))] has F distribution
    - numerator has ν_B = (b-1) degrees of freedom
    - denominator has ν_E = (a-1)(b-1) degrees of freedom
    - significant if MSA/MSE > $F_{[1-\alpha;\ \nu B;\ \nu E]} = F_{[1-\alpha;\ b-1;\ (a-1)(b-1)]}$

Analyzing Significance in Pgm / Cache Study

- **Test if factors significant at .10 level (90% confidence level)**

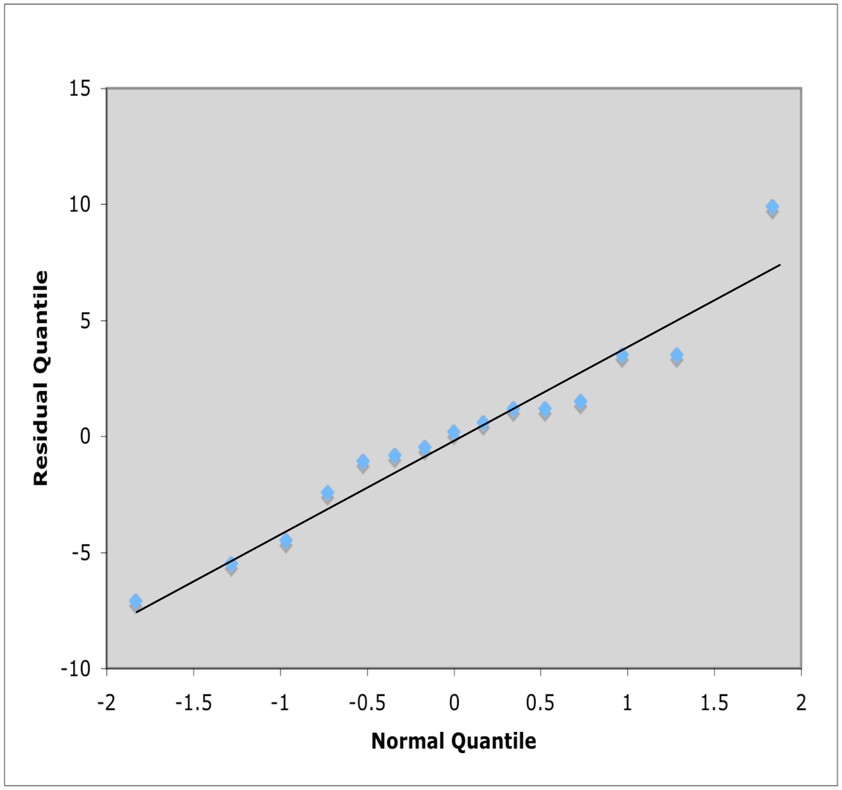
Component	Sum of Squares	% Variation	DOF	Mean Square	F-Computed	F-Table	
SSY	91595						
SS0	78192.60						
SSY-SS0	13402.40	100.0%	14				
Caches	12857.20	95.9%	2	6428.6	217.2	3.11	F[.9,2,8]
Pgms	308.4	2.4%	4	77.1	2.6	2.81	F[.9,4,8]
Errors	236.8	76.8%	8	29.6			

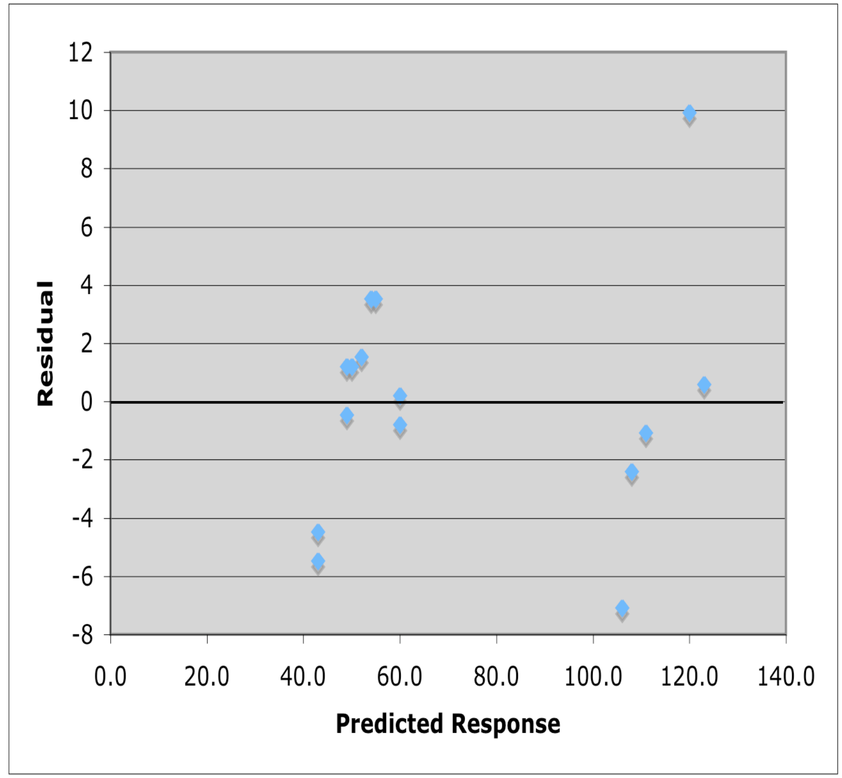
- **Interpretation**
  - —effect of cache makes a *significant* impact on performance
  - —effect of pgms not *significant* impact on performance
- **Notes**
  - —Jain reports this is a real case study
  - —chose pgm workloads that had comparable runtimes
    - – avoid overshadowing cache effect
    - – important point often missed by inexperienced analysts

Assumptions of Two-factor Experiments

- **Effects of various factors are additive**
- **Errors are additive**
- **Errors are independent of factor levels**
- **Errors are normally distributed**
- **Errors have the same variance for all factor levels**

Visual Tests of Pgm / Cache Study





- **Interpretation**
  - —quantile-quantile plot approximately linear → normality OK
  - —scatter plot of residuals shows slightly increasing spread
    - – not ideal

Parameter Estimation for Two Factors

Parameter	Estimate	Variance
μ	$\bar{y}_{..}$	s_e^2 / ab
α_j	$\bar{y}_{.j} - \bar{y}_{..}$	$s_e^2 (a-1)/ab$
$\mu + \alpha_j$	$\bar{y}_{.j}$	s_e^2 / b
β_i	$\bar{y}_{i.} - \bar{y}_{..}$	$s_e^2 (b-1)/ab$
$\mu + \alpha_j + \beta_i$	$\bar{y}_{.j} + \bar{y}_{i.} - \bar{y}_{..}$	$s_e^2 (a+b-1)/ab$
$\sum_{j=1}^{a} h_j \alpha_j, \sum_{j=1}^{a} h_j = 0$	$\sum_{j=1}^{a} h_j \bar{y}_{.j}$	$s_e^2 \sum_{j=1}^{a} h_j^2 / b$
$\sum_{i=1}^{b} h_i \beta_i, \sum_{i=1}^{b} h_i = 0$	$\sum_{i=1}^{b} h_i \bar{y}_{i.}$	$s_e^2 \sum_{i=1}^{b} h_i^2 / a$
s_e^2	$\left(\sum_{i,j} e_{ij}^2 \right) / [(a-1)(b-1)]$	

DOF of errors = (a-1)(b-1)

Variance for μ

- **Expression for μ in terms of random variables y_{ij}'s**

$$\mu = \frac{1}{ab}\sum_{i=1}^{b}\sum_{j=1}^{a} y_{ij}$$

- **What is the coefficient a_{ij} for each y_{ij}?**

$$a_{ij} = \left\{\frac{1}{ab}\right.$$

$$\mathrm{Var}(\mu) = Var\left(\frac{1}{ab}\sum_{i=1}^{a}\sum_{j=1}^{b} y_{ij}\right)$$

Assuming
errors normally distributed
zero mean, variance $(\sigma_e)^2$
What is the variance for μ?

$$\sigma_\mu^2 = \left(ab\left(\frac{1}{ab}\right)^2\right)\sigma_e^2$$

$$= \frac{1}{ab}\sigma_e^2$$

Variance for α_j

- **Expression for α_j in terms of random variables y_{ij}'s**

$$\alpha_j = \bar{y}_{.j} - \bar{y}_{..} = \frac{1}{b}\sum_{i=1}^{b} y_{ij} - \frac{1}{ab}\sum_{i=1}^{b}\sum_{k=1}^{a} y_{ik}$$

- **What is the coefficient a_{ikj} for each y_{ik} for α_j?**

$$a_{ik} = \begin{cases} \dfrac{1}{b} - \dfrac{1}{ab} & k = j \\ -\dfrac{1}{ab} & \text{otherwise} \end{cases}$$

Assuming
errors normally distributed
zero mean, variance $(\sigma_e)^2$
What is the variance for α_j?

$$\boxed{\begin{aligned} \sigma_{\alpha_j}^2 &= \left(b\left(\frac{1}{b} - \frac{1}{ab}\right)^2 + (ab - b)\left(\frac{1}{ab}\right)^2 \right)\sigma_e^2 \\ &= \frac{(a-1)}{ab}\sigma_e^2 \end{aligned}}$$

Variance for $\mu+\alpha_j+\beta_i$

- **Expression for α_j in terms of random variables y_{ij}'s**

$$\hat{y}_{ij} = \mu + \alpha_j + \beta_i = \bar{y}_{.j} + \bar{y}_{i.} - \bar{y}_{..} = \frac{1}{b}\sum_{k=1}^{b} y_{kj} + \frac{1}{a}\sum_{l=1}^{a} y_{il} - \frac{1}{ab}\sum_{k=1}^{b}\sum_{l=1}^{a} y_{kl}$$

- **What is the coefficient a_{lk} for each y_{lk}?**

$$a_{kl} = \begin{cases} \frac{1}{b} - \frac{1}{ab} & l = j \\ \frac{1}{a} - \frac{1}{ab} & k = i \\ \frac{1}{a} + \frac{1}{b} - \frac{1}{ab} & k = i, l = j \\ -\frac{1}{ab} & \text{otherwise} \end{cases}$$

Assuming

errors normally distributed

zero mean, variance $(\sigma_e)^2$

What is the variance for $\mu+\alpha_j+\beta_i$?

$$\sigma^2_{\hat{y}_{ij}} = \left((b-1)\left(\frac{1}{b} - \frac{1}{ab}\right)^2 + (a-1)\left(\frac{1}{a} - \frac{1}{ab}\right)^2 + \left(\frac{1}{a} + \frac{1}{b} - \frac{1}{ab}\right)^2 + (ab - a - b + 1)\left(\frac{1}{ab}\right)^2 \right)\sigma_e^2$$

$$= \frac{(a+b-1)}{ab}\sigma_e^2$$

Confidence Intervals

- **Compute confidence intervals with t values read out of table at (a-1)(b-1) degrees of freedom (DOF of errors)**
- **Examples**

$$90\% \text{ confidence interval for } \mu = \bar{y}_{..} \pm t_{[.95,(a-1)(b-1)]} \times s_{\mu}$$

$$= \bar{y}_{..} \pm t_{[.95,(a-1)(b-1)]} \times s_e / \sqrt{ab}$$

$$90\% \text{ confidence interval for } \alpha_j = \bar{y}_{.j} - \bar{y}_{..} \pm t_{[.95,(a-1)(b-1)]} \times s_{\alpha_j}$$

$$= \bar{y}_{.j} - \bar{y}_{..} \pm t_{[.95,(a-1)(b-1)]} \times s_e \sqrt{(a-1)/ab}$$

- **Parameter not significant if confidence interval includes 0**

Multiplicative Models for Two-factors

- **Why a multiplicative model?**
  - —large range of responses: y_{max}/y_{min} large
  - —whenever physical considerations require it
    - – workloads and processors often multiplicative
  - —spread in residuals increases with mean response
- **How to use one?**
  - —take log of response data
  - —use log transformation of response in an additive model
- **Perform table-based analysis as before on transformed data**

Note: multiplicative and additive models are numerically different only if range covered by observations is large

Handling Missing Observations

- If a few of the a x b observations are missing, the analysis can still be used
  - —if more than one or two values missing in a row/column, better to exclude it
- One method
  - —compute effects as before from row and column means
    - – adjust divisors to use number of observations *available*
    - – adjust degrees of freedoms for sums of squares
      - if a value is missing, it does not contribute a DOF
    - – adjust formulas for standard deviations
      - adjust so DOF reflect observations available
- Alternate methods (some controversy here)
  - —replacement method
    - – replace missing value with prediction so its residual is 0
  - —minimum residual variance method
    - – place symbol y in cell
    - – write SSE as function of y
    - – pick y to minimize SSE
- Problems: different estimates for SSA & SSB

RISC Execution Study: Measured Time

Workload	RISC-I	68000	Z8002	VAX 11/780	PDP 11/70	C/70
String Srch	0.46	1.29	0.74	0.60	0.41	1.01
F-Bit Tst	0.06	0.29	0.43	0.29	0.37	0.55
K-Bit Mtrx	0.43	1.72	2.24	1.29	1.72	4.00
Linked Lst	0.10	0.16	0.24	0.12	0.19	0.25
I-Quick Srt	50.40	206.64	262.08	151.20	181.44	292.32
Ackerman	3200.00		8960.00	5120.00	5120.00	
Rec Qsort	800.00		4720.00	1840.00	2560.00	1040.00
Puzzle Sub	4700.00		19740.00	9400.00	7520.00	15980.00
Puzzle Ptr	3200.00	13440.00	7360.00	4160.00	6400.00	6720.00
SED	5100.00		22440.00	5610.00	5610.00	13260.00
Hanoi	6800.00		28560.00	12240.00	15640.00	10880.00

RISC Execution Study: Multiplicative Model

Workload	RISC-I	68000	Z8002	VAX 11/780	PDP 11/70	C/70	Row Sum	Row Mean	Row Effect
String Srch	-0.34	0.11	-0.13	-0.22	-0.39	0.00	-0.96	-0.16	-2.16
F-Bit Tst	-1.22	-0.54	-0.37	-0.54	-0.43	-0.26	-3.36	-0.56	-2.55
K-Bit Mtrx	-0.37	0.24	0.35	0.11	0.24	0.60	1.17	0.19	-1.80
Linked Lst	-1.00	-0.80	-0.62	-0.92	-0.72	-0.60	-4.66	-0.78	-2.77
I-Quick Srt	1.70	2.32	2.42	2.18	2.26	2.47	13.34	2.22	0.23
Ackerman	3.51		3.95	3.71	3.71		14.88	3.72	1.72
Rec Qsort	2.90		3.67	3.26	3.41	3.02	16.27	3.25	1.26
Puzzle Sub	3.67		4.30	3.97	3.88	4.20	20.02	4.00	2.01
Puzzle Ptr	3.51	4.13	3.87	3.62	3.81	3.83	22.75	3.79	1.80
SED	3.71		4.35	3.75	3.75	4.12	19.68	3.94	1.94
Hanoi	3.83		4.46	4.09	4.19	4.04	20.61	4.12	2.13
Col Sum	19.90	5.46	26.25	23.01	23.70	21.42	119.73		
Col Mean	1.81	0.91	2.39	2.09	2.15	2.14		2.00	
Col Effect	-0.19	-1.09	0.39	0.10	0.16	0.15			

- **Data in table is $\log_{10}$(execution time)**

RISC Execution Study: Modeling Error

Workload	RISC-I	68000	Z8002	VAX 11/780	PDP 11/70	C/70
String Srch	0.01	1.36	-0.36	-0.16	-0.39	0.02
F-Bit Tst	-0.48	1.11	-0.20	-0.07	-0.03	0.15
K-Bit Mtrx	-0.37	1.13	-0.23	-0.18	-0.12	0.26
Linked Lst	-0.04	1.07	-0.23	-0.24	-0.10	0.03
I-Quick Srt	-0.33	1.18	-0.20	-0.14	-0.12	0.10
Ackerman	-0.03		-0.16	-0.11	-0.17	
Rec Qsort	-0.16		0.03	-0.09	0.00	-0.38
Puzzle Sub	-0.15		-0.10	-0.13	-0.29	0.05
Puzzle Ptr	-0.10	1.42	-0.32	-0.27	-0.14	-0.11
SED	-0.04		0.02	-0.28	-0.35	0.04
Hanoi	-0.10		-0.06	-0.13	-0.09	-0.23

$$SSE = \sum_{i=1}^{b}\sum_{j=1}^{a} e_{ij}^2 = (.01)^2 + (-.48)^2 + \ldots + (-.23)^2 = 11.01$$

$$MSE = \frac{SSE}{\nu_E} = \frac{11.01}{\text{DOF}}, \qquad \text{DOF} = (60-1-5-10) = 38, \quad \text{MSE} = .25$$

$$\mu \quad \alpha_j \; \beta_i \qquad\qquad s_e = \sqrt{MSE}$$

Variance for α_j with Missing Data

- **Assuming**
  - —errors are normally distributed with zero mean
  - —assume variance $(\sigma_e)^2$
- **Expression for $\mathrm{Var}(\alpha_j)$ in terms of random variables $\mathrm{Var}(y_{ij})$'s**

$$\text{Let } m_{ij} = \begin{cases} 1 & \text{if } y_{ij} \text{ present} \\ 0 & \text{if } y_{ij} \text{ missing} \end{cases}$$

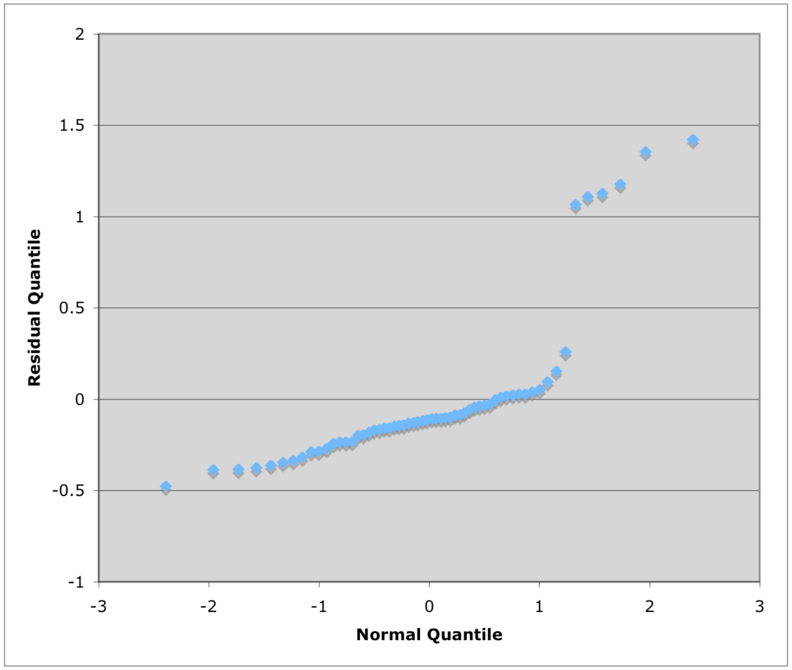
$$\text{Let } c_j = \sum_j m_{ij}, \quad N = \sum_{i,j} m_{ij}$$

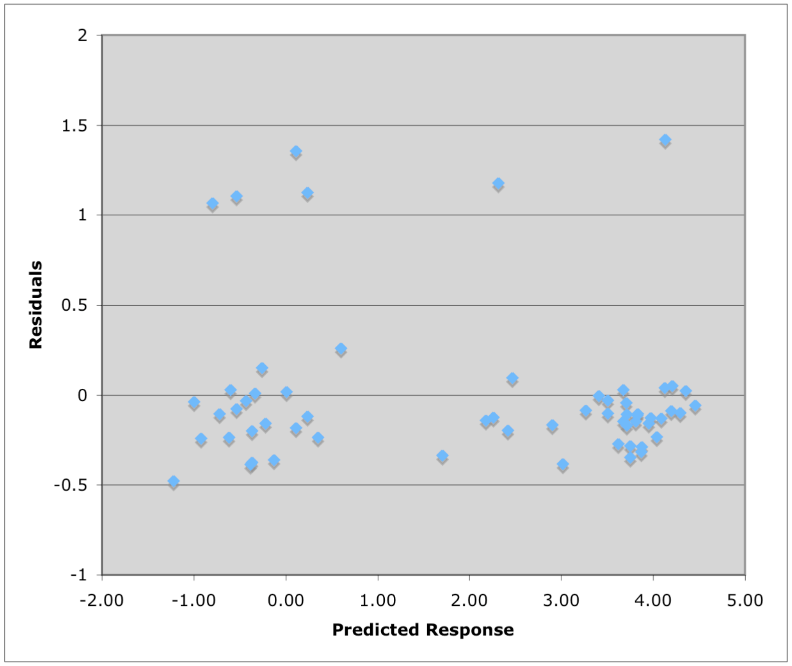
$$\mathrm{Var}(\alpha_j) = \mathrm{Var}(\bar{y}_{.j} - \bar{y}_{..})$$

$$= \mathrm{Var}\left(\frac{1}{c_j} \sum_{i=1}^{a} m_{ij} y_{ij} - \frac{1}{N} \sum_{i=1}^{a} \sum_{j=1}^{b} m_{ij} y_{ij} \right)$$

$$\boxed{\begin{aligned} \sigma_{\alpha_j}^2 &= \left(c_j \left(\frac{1}{c_j} - \frac{1}{N} \right)^2 + (N - c_j) \left(\frac{1}{N} \right)^2 \right) \sigma_e^2 \\ &= \frac{(N - c_j)}{N c_j} \sigma_e^2 \\ s_{\alpha_j} &= s_e \sqrt{\frac{(N - c_j)}{N c_j}} \end{aligned}}$$

Visual Tests





- **Interpretation**
  - —**quantile-quantile plot is not very linear**
  - —**some residuals much higher than others**
    - – **all residuals for 68000 were > 1**

Confidence Intervals