12 Sept 01.

Can now compute

$$\widehat{\operatorname{var}\, \underset{\sim}{p}' \hat{\underset{\sim}{\beta}}} = \underset{\sim}{p}'(\underset{\sim}{X}'\underset{\sim}{X})^{-}\underset{\sim}{p}\, \hat{\sigma}^2$$

and so get c.i.'s

 12 Sept 2001

Influential observations, regression diagnostics, outliers (

(∃ variants for nonlinear regression, glm, ...)

Influential observation: crucial to inferences drawn from the data

Examine its value carefully

Full rank case (for simplicity)

Hat matrix

$$\hat{\underset{\sim}{Y}} = \underset{\sim}{H}\,\underset{\sim}{Y}$$

$$\underset{\sim}{H} = \underset{\sim}{X}(\underset{\sim}{X}'\underset{\sim}{X})^{-}\underset{\sim}{X}'$$

$$\boxed{\hat{Y}_i = H_{ii} Y_i + \sum_{j \neq i} H_{ij} Y_j}$$

H_{ii}: the leverage or influence of Y_i or $\hat{Y}_i$

12 Sept 2001

Properties of $\underset{\sim}{H}$.

a) $0 \le H_{ii} \le 1$

Proof $\text{var } \hat{\underset{\sim}{Y}} = \sigma^2 \underset{\sim}{H}$ and $\text{var } \hat{\underset{\sim}{\epsilon}} = \sigma^2 (\underset{\sim}{I} - \underset{\sim}{H})$

b) i) $H_{ii} = 1 \Rightarrow \hat{Y}_i = Y_i$ A high value, model requires a parameter to fit that Y_i

$(\simeq 1 \Rightarrow \hat{Y}_i \simeq Y_i)$

ii) $H_{ii} = 0 \Rightarrow \hat{Y}_i = 0$ (fixed by design)

Proof

$$\underset{\sim}{H}^2 = \underset{\sim}{H}$$

$$H_{ii} = H_{ii}^2 + \sum_{j \ne i} H_{ij}^2$$

$$H_{ii} = 1, 0 \Rightarrow H_{ij} = 0 \quad j \ne i$$

$$\hat{Y}_i = H_{ii} Y_i + \sum_{j \ne i} H_{ij} Y_j$$

c) $\bar{H} = r/n \qquad r = r(\underset{\sim}{X})$

$$\sum_i H_{ii} = \text{trace}(\underset{\sim}{H}) = \text{trace}(\underset{\sim}{X}(\underset{\sim}{X}'\underset{\sim}{X})^{-}\underset{\sim}{X}')$$

$$= r$$

A definition of leverage points:

$$H_{ii} > 2r/n$$

Such points have a lot of influence

All have the same influence if $H_{ii} = r/n$
(unusual)

Eg. $r = 2$

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

$$H_{ii} = \frac{1}{n} + \frac{(X_i - \bar{X})^2}{\sum (X_i - \bar{X})^2} \qquad \sum (X_i - \bar{X})^2 \neq 0$$

influence relates directly to how near X_i is to $\bar{X}$

{Why condition on X's?}

Turning to <u>residuals</u>:

$$\hat{\underset{\sim}{\epsilon}} = (\underset{\sim}{I} - \underset{\sim}{H})\underset{\sim}{\epsilon}$$

$$\hat{\epsilon}_i = \epsilon_i - \sum_{j=1}^{n} H_{ij}\,\epsilon_j$$

Shows how $\hat{\epsilon}_i$ gets at ϵ_i

$$\text{var}\,\hat{\underset{\sim}{\epsilon}} = \sigma^2(\underset{\sim}{I} - \underset{\sim}{H})$$

$$\text{corr}\{\epsilon_i, \epsilon_j\} = -H_{ij} \Big/ \sqrt{(1-H_{ii})(1-H_{jj})}$$

$H_{ii}, H_{jj} \neq 0$

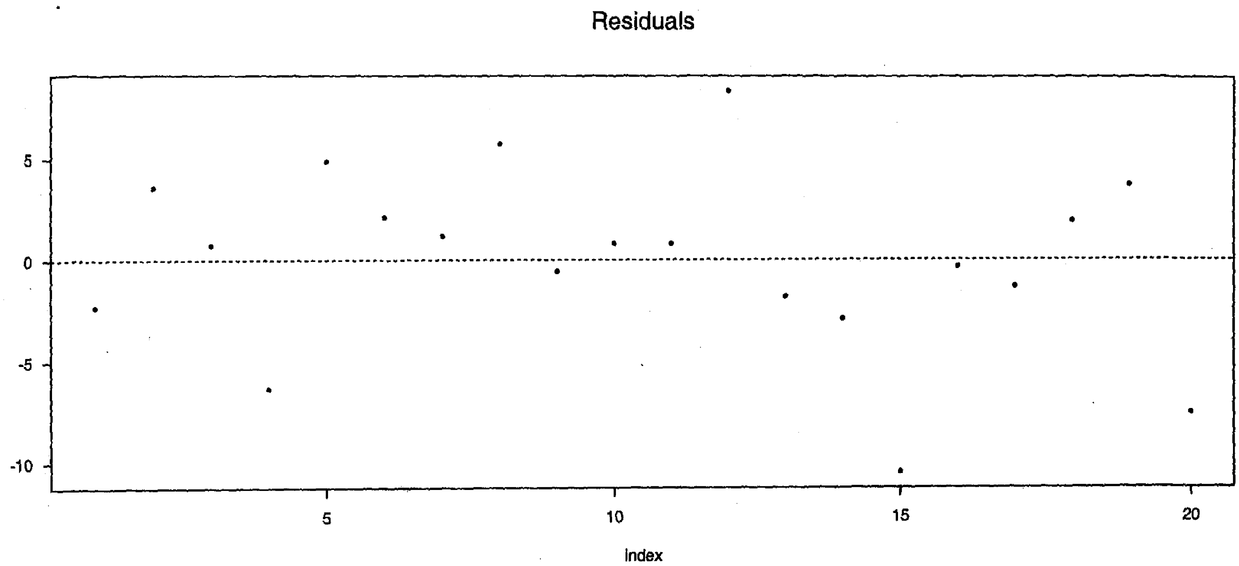
$$\left(\text{var}\,\hat{Y}_i = \sigma^2 H_{ii}\right)$$

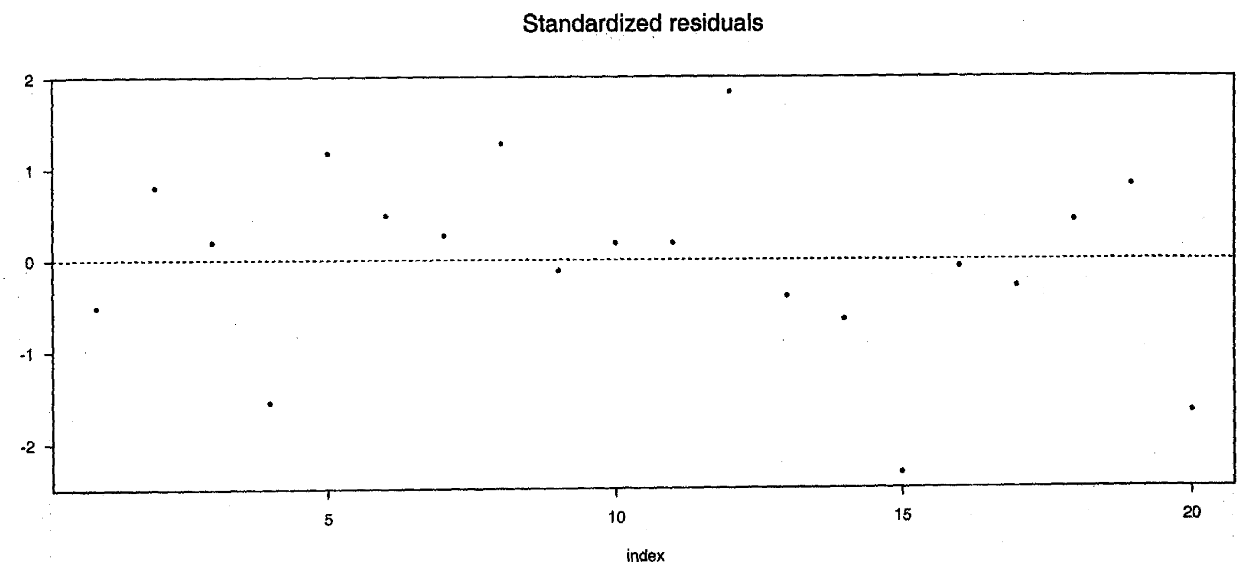
way to get H_{ii}

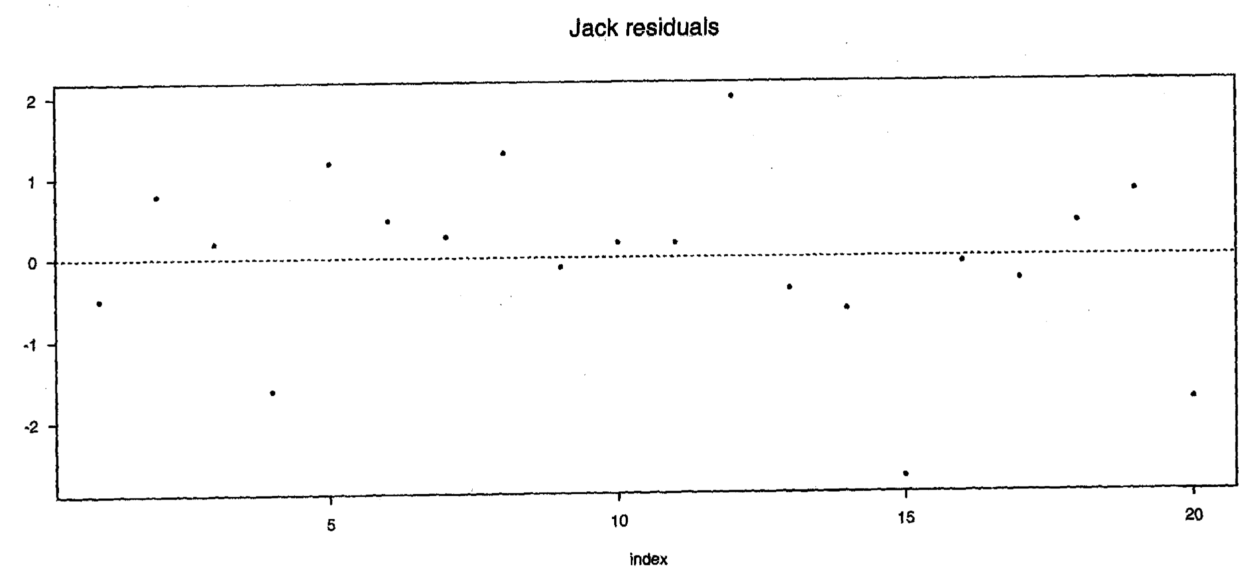
<u>Standardized residual</u>

$$\hat{\epsilon}_i \Big/ \sqrt{1 - H_{ii}}$$

$H_{ii} \neq 1$


Residuals
5
0
-5
-10
5
10
15
20
index


Standardized residuals
2
1
0
-1
-2
5
10
15
20
index


Jack residuals
2
1
0
-1
-2
5
10
15
20
index

12 Sept 2001

Jack-knife / cross-validatory residuals

$$\hat{\epsilon}_i^* = \hat{\epsilon}_i \Big/ s_{-i}\sqrt{1 - H_{ii}} \qquad H_{ii} \neq 1$$

$$= \frac{Y_i - \underset{\sim}{x}_i' \hat{\underset{\sim}{\beta}}_{-i}}{s_{-i}\sqrt{1 + \underset{\sim}{x}_i' (\underset{\sim}{X}_{-i}' \underset{\sim}{X}_{-i})^{-1} \underset{\sim}{x}_i}} \qquad \text{need inverse}$$

Need lemma (Algebra of deletion)

$\hat{\underset{\sim}{\beta}}_{-i}$, s_{-i}^2 are estimates of $\underset{\sim}{\beta}$, σ^2 based on $n-1$ observations excluding ith

$$(n-1-r)s_{-i}^2 = \sum_{l \neq i} [Y_l - \underset{\sim}{x}_l' \hat{\beta}_{-i}]^2$$

1. A standardized measure of the distance between the i-th case and the model estimated on the remaining cases.
2. A "test statistic" to see if the i-th point belongs in the model

12 Sept. 2001

Cook's Distance

$$\frac{(\hat{\beta}_{-i} - \hat{\beta})' X'X (\hat{\beta}_{-i} - \hat{\beta})}{r s^2}$$

$$= \frac{\hat{\epsilon}_i^2 H_{ii}}{s^2 (1 - H_{ii})^2 r}$$

$$= \sup_{\alpha} \frac{|\alpha' (\hat{\beta} - \hat{\beta}_{-i})|^2}{s^2 \alpha' (X'X)^{-1} \alpha}$$

1. i.e. distance $\hat{\beta}$ would move if ith point omitted
2. large values indicate observations which are influential on joint inferences re β.

Compare with $4/(n-r-1)$.

$\exists$ modified Cook

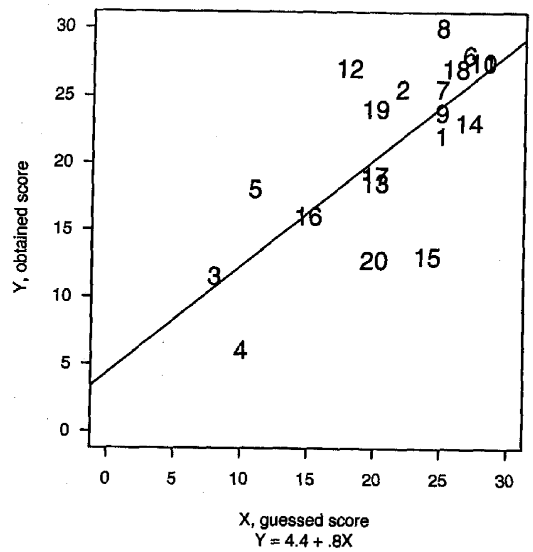
12 Sept 2001

Some routines:

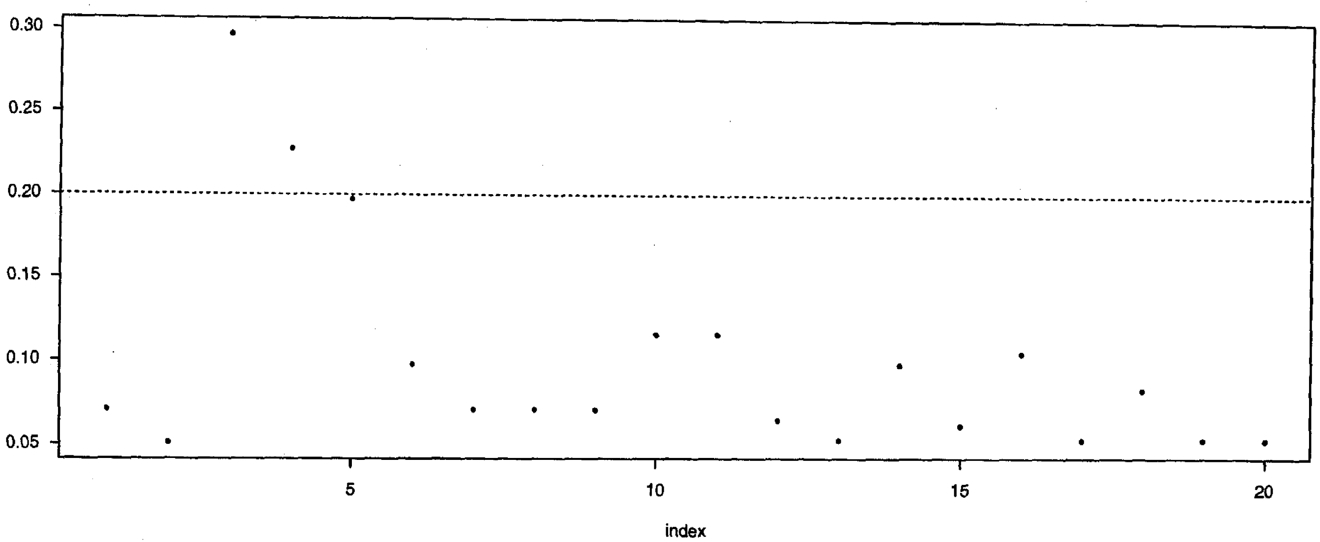
standardized resid stdres()
{studentized / jackknifed} resid studres()

lm.influence() $ hat

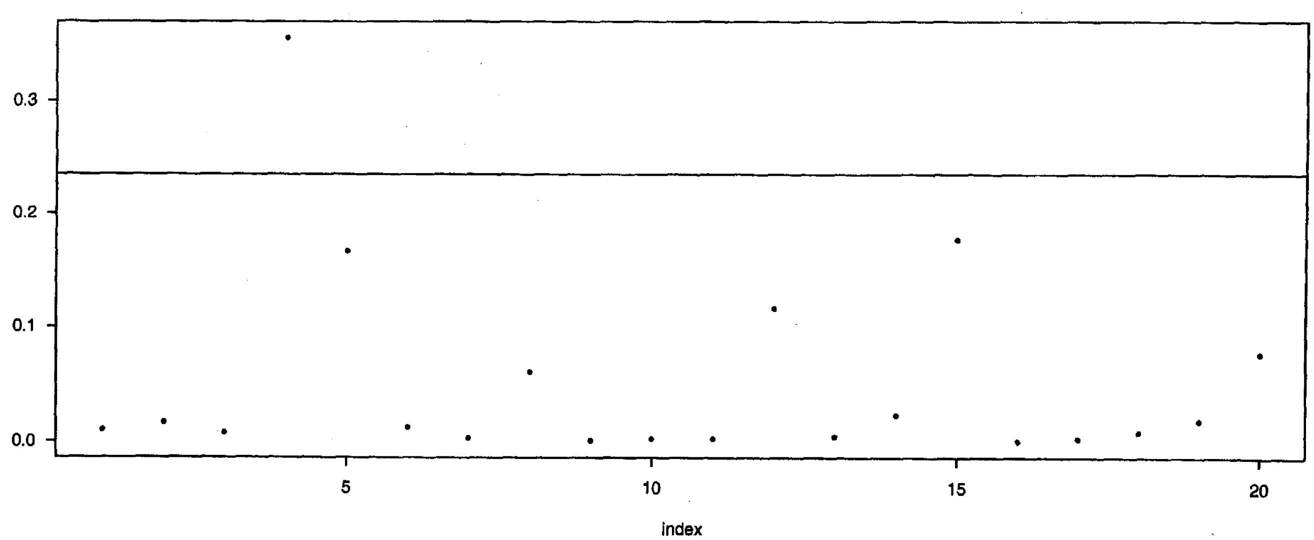
Stat 131a : Midterm



leverage



Cook's distance



```
> summary(junkr)

Call: lm(formula = y ~ x)
Residuals:
    Min     1Q Median     3Q   Max
 -10.51 -1.94 0.7755 2.474 8.275

Coefficients:
              Value Std. Error t value Pr(>|t|)
(Intercept) 4.3783 3.8509     1.1370  0.2705
          x 0.7970 0.1749     4.5571  0.0002

Residual standard error: 4.651 on 18 degrees of freedom
Multiple R-Squared: 0.5357
F-statistic: 20.77 on 1 and 18 degrees of freedom, the p-value is 0.0002442

Correlation of Coefficients:
  (Intercept)
x -0.9628
```

Interpretation of Regression Coefts.

15 Sept. 06

Consider the ~~estimation~~

$$\underset{\sim}{Y} = \underset{\sim}{X}\underset{\sim}{\beta} + \underset{\sim}{Z}\underset{\sim}{\gamma} + \underset{\sim}{\epsilon}$$

$\underset{\sim}{Z}$ possibly unmeasured

Normal equations are:

$$X'X\hat{\beta} + X'Z\hat{\gamma} = X'Y \qquad (*)$$

$$Z'X\hat{\beta} + Z'Z\hat{\gamma} = Z'Y$$

$$\hat{\beta} = (X'X)^{-1}X'Y - (X'X)^{-1}X'Z\hat{\gamma}$$

from (*) if inverse exists

$$= \hat{\beta}_* - \hat{\delta}_*\hat{\gamma}$$

where

$$\hat{\delta}_* = (X'X)^{-1}X'Z$$

$$\boxed{\hat{\beta}_* = \hat{\beta} + \hat{\delta}_*\hat{\gamma}}$$

13 Sept 01

<u>Consider fits</u>

Would like

$$\hat{Y} = X\hat{\beta} + Z\hat{\gamma} \qquad (**)$$

Have

$$\hat{Y}_* = X\hat{\beta}_*$$

$$= X(\hat{\beta} + \hat{\delta}_*\hat{\gamma})$$

$$= X\hat{\beta} + X(X'X)^{-1}X'Z\hat{\gamma}$$

Suppose regress Z on X

$$\hat{Z} = HZ = X(X'X)^{-1}X'Z$$

get $\hat{Y}_* = X\hat{\beta} + \hat{Z}\hat{\gamma}$

Have replaced Z in (**) by fitted value based on X

The use of X alone, replaces (unknown) Z by prediction based on X

13 Sept 01

Not including a variable in a regression is equivalent to replacing it by its best linear predictor based on the variables included.

(Shows predictions OK)

Variables included are acting as <u>proxies</u> for those left out. X drawn in, not because should be, but because related to Z that should be.

Example. During World War II, in investigating aiming errors made during bomber flights over Europe, one of the research organizations developed a regression equation with several carriers. Among its nine or so carriers were altitude, type of aircraft, speed of the bombing group, size of group, and the amount of fighter opposition. On physical grounds, one might expect higher altitudes and higher speeds to produce larger aiming errors. It would not be surprising if different aircraft differed in performance. What the effect of size of group might be can be argued either way. But few people will believe that additional fighter opposition would help a pilot and bombardier do a better job. Nevertheless, amount of fighter opposition appeared as a strong term in the regression equation—the more opposition, the smaller the aiming error. The effect is generally regarded as a proxy phenomenon, arising because the equation had no variable for amount of cloud cover. If clouds obscured the target, the fighters usually did not come up and the aiming errors were ordinarily very large.

13 Sept. 04

Can get apparent effect when none exists (β, $\hat{\beta} \approx 0$)

Can get no apparent effect, when effect exists ($\beta \neq 0$)

Impossible to draw valid conclusions as to how interference with an X will affect the system.

13 Sept 01

In a designed experiment, levels of the regressor variables are chosen in some deliberately random manner — impossible for their values to be affected by the z's.

$$\hat{y}_* = X\hat{\beta} + \hat{z}\hat{\gamma}$$

$$\hat{z} \simeq 0$$

$$\hat{\delta}_* = (X'X)^{-1}X'z$$

$$\simeq 0$$

$$\hat{\beta}_* \simeq \hat{\beta} \quad \text{then}$$

Can see effect of changing X

13 Sept 01

<u>Lurking variable</u> – a variable that has an important effect and yet is not included amongst the predictor variables under consideration (perhaps existence unknown or effect unsuspected)

13 Sept 01

Questions

1. Is a variable serving well, but only a proxy?
2. Is a proxy's coefficient taking part of the coeff we would like for the proper variable!

Difficulties arise because of unplanned/observational data vs. experimental

13 Sept. 91

Observational study

John Snow (1854) Cholera in London

Suspected it related to 2 separate water companies

Company	1851 pop^n	Deaths (7/8/54-8/26/54)	Rate /1000
1	167,654	844	5.0
2	19,133	18	.9
Both	300,149	652	2.2

Makes Company 1 look bad, but could be division by social class.

Snow focused on 3rd group, i.e. area where both companies served. Homes served seemed completely intermixed.

1	98,862	419	4.2
2	154,615	80	.5

Company 1 appears at fault. Company 2 had moved water source away from the Thames

MacMahon + Pugh (1970) Epidemiology

14 Sept 01

__Example__. One-way classification
(one factor, one treatment)

$$Y_{jk} = \mu_j + \epsilon_{jk}$$

$$j = 1, \ldots, J \; ; \quad k = 1, \ldots, K$$

J categories, K replicates (Could have K_j)

$H: \quad \mu_1 = \ldots = \mu_J$

$\mu_1 - \mu_J = \mu_2 - \mu_J = \ldots$ (Present $J-1$)

$$\underset{\sim}{Y} = \underset{\sim}{X} \, \underset{\sim}{\beta} + \underset{\sim}{\epsilon}$$

$$\begin{bmatrix} Y_{11} \\ Y_{12} \\ \vdots \\ Y_{1K_1} \\ \hline Y_{21} \\ \vdots \\ Y_{2K_2} \\ \hline \vdots \\ \hline Y_{J1} \\ \vdots \\ Y_{JK_J} \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & & 0 \\ 1 & 0 & 0 & & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & 0 & 0 & & 0 \\ \hline 0 & 1 & 0 & & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & 1 & 0 & & 0 \\ \hline \vdots & \vdots & \vdots & & \vdots \\ \hline 0 & 0 & 0 & & 1 \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & 0 & & 1 \end{bmatrix} \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_J \end{bmatrix} + \begin{bmatrix} \epsilon_{11} \\ \epsilon_{12} \\ \vdots \\ \epsilon_{1K_1} \\ \hline \epsilon_{21} \\ \vdots \\ \epsilon_{2K_2} \\ \hline \vdots \\ \epsilon_{J1} \\ \vdots \\ \epsilon_{JK_J} \end{bmatrix}$$

14 Sept. 01

$$\underset{\sim}{Y} = \underset{\sim}{X}\,\underset{\sim}{\beta} + \underset{\sim}{\epsilon}\,, \qquad \underset{\sim}{\beta} = \underset{\sim}{\mu}$$

$$\begin{bmatrix} 1 & 0 & 0 & & 0 & -1 \\ 0 & 1 & 0 & & 0 & -1 \\ \vdots & \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & 0 & & 1 & -1 \end{bmatrix} \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_J \end{bmatrix} = \underset{\sim}{0}$$

$$\underset{\sim}{P}^{T}\underset{\sim}{\beta} = \underset{\sim}{0} = \underset{\sim}{P}^{T}\underset{\sim}{\mu}$$

ANOVA Table

Source	SS	df
Mean	$\sum\sum \bar{Y}_{..}^2$	1
Between categories	$\sum\sum(\bar{Y}_{j.} - \bar{Y}_{..})^2$	$J-1$
Error	$\sum\sum(Y_{jk} - \bar{Y}_{j.})^2$	~~n~~ $-J$
Total	$\sum\sum Y_{jk}^2$	Jn

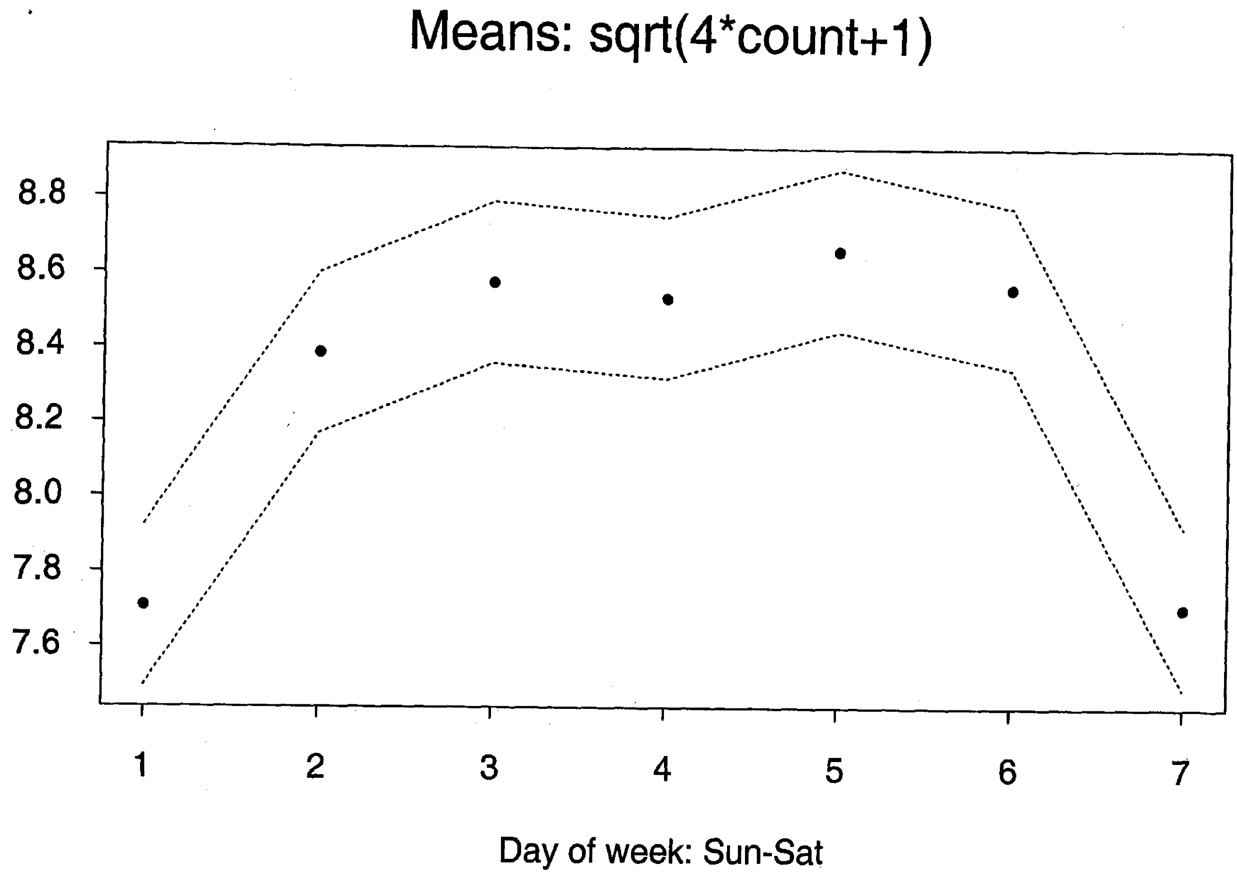
Alternate parametrizations:

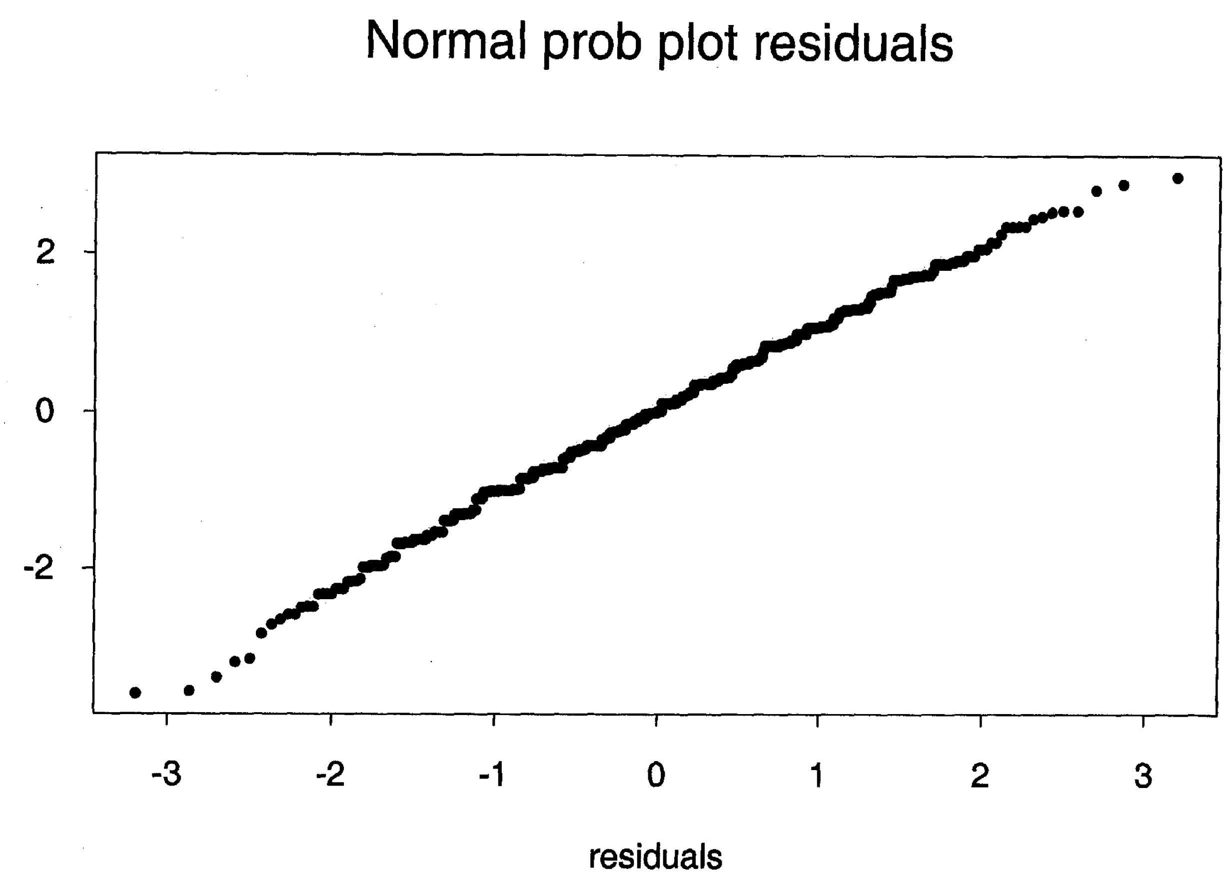
$$Y_{jk} = \mu + \alpha_j + \epsilon_{jk}\,, \qquad \sum_j \alpha_j = 0$$

$$\sum_j \sum_k \alpha_j = 0$$

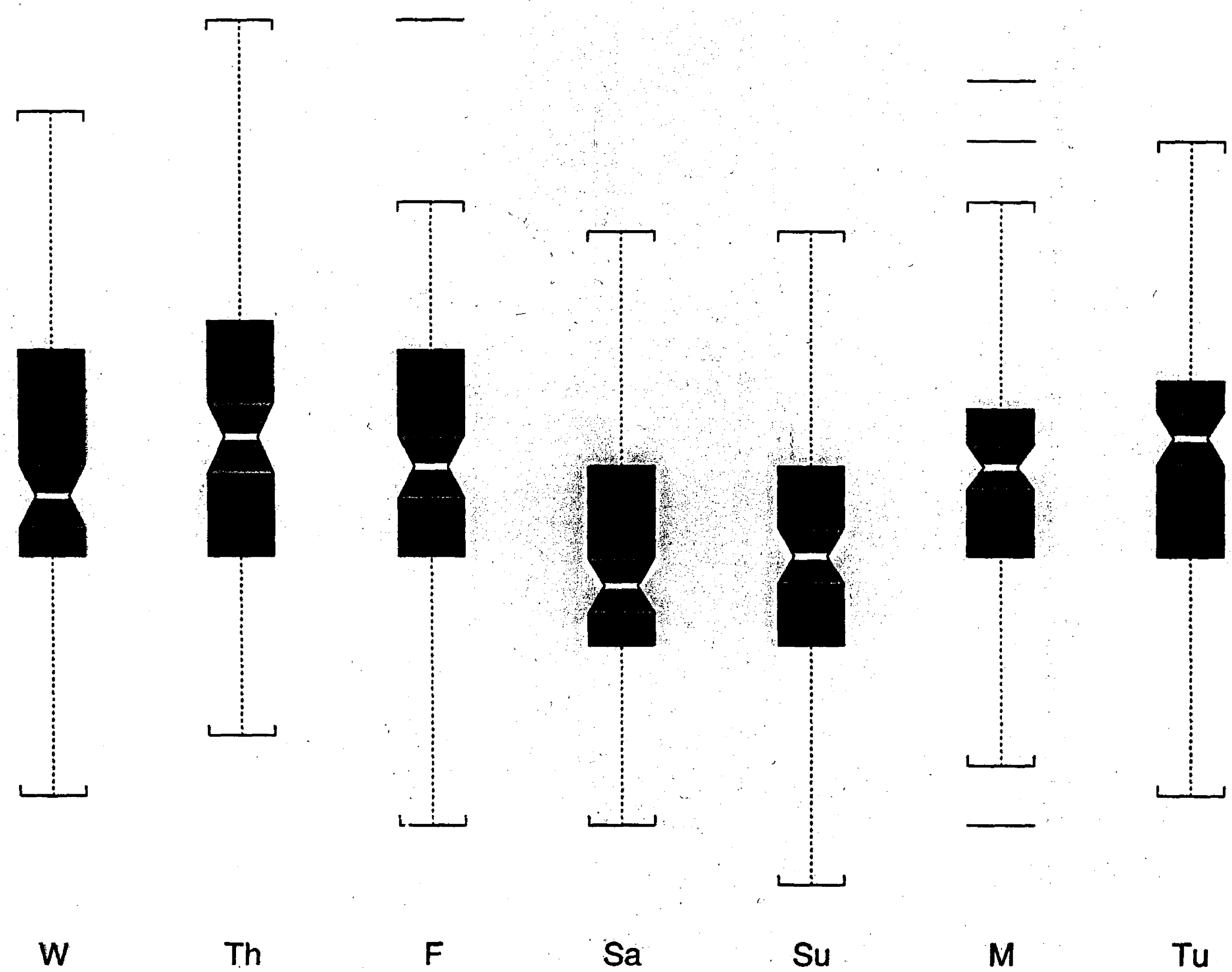
$$\alpha_1 = 0$$

```
Call: lm(formula = sqrt(4 * y + 1) ~ -1 + A)
Residuals:
    Min      1Q   Median     3Q  Max
 -3.587 -0.7245 0.009242 0.8339 2.97

Coefficients:
     Value Std. Error t value Pr(>|t|)
A1  7.7101  0.1080    71.3960  0.0000
A2  8.3854  0.1080    77.6491  0.0000
A3  8.5749  0.1080    79.4037  0.0000
A4  8.5348  0.1080    79.0320  0.0000
A5  8.6633  0.1080    80.2220  0.0000
A6  8.5622  0.1080    79.2861  0.0000
A7  7.7084  0.1080    71.3797  0.0000

Residual standard error: 1.096 on 714 degrees of freedom
Multiple R-Squared: 0.9831
F-statistic: 5928 on 7 and 714 degrees of freedom, the p-value is 0

Correlation of Coefficients:
   A1 A2 A3 A4 A5 A6
A2 0
A3 0  0
A4 0  0  0
A5 0  0  0  0
A6 0  0  0  0  0
A7 0  0  0  0  0  0
Analysis of Variance Table

Response: sqrt(4 * y + 1)

Terms added sequentially (first to last)
           Df Sum of Sq  Mean Sq  F Value Pr(F)
A           7  49843.34 7120.478 5927.813     0
Residuals 714    857.66    1.201
```


Means: sqrt(4*count+1)
8.8
8.6
8.4
8.2
8.0
7.8
7.6
1
2
3
4
5
6
7
Day of week: Sun-Sat


Normal prob plot residuals
2
0
-2
-3
-2
-1
0
1
2
3
residuals

Daily births in Saskatchewan, 1986-1987



D.R. Brillinger notes

14 Sept. 01

$$\underset{n\times 1}{Y} = \underset{n\times m}{X}\ \underset{m\times 1}{\beta} + \underset{n\times 1}{\epsilon}$$

$$\underset{m\times k}{P} = [P_1, \ldots, P_k] \qquad r(P) = q$$

$P_j'\beta$ estimable $\qquad P' = QX \qquad \underset{m\times k}{P} = X'\underset{n\times k}{Q}$

H_0: $\quad P'\beta = \delta$

$$X'X\hat{\beta} = X'Y$$

$$(P'\hat{\beta} - \delta)'[P'(X'X)^- P]^-(P'\hat{\beta} - \delta) : \sigma^2\chi^2_q$$

$$\| Y - X\hat{\beta}\|^2 : \sigma^2\chi^2_{n-r}$$

$$\frac{(P'\hat{\beta} - \delta)'[P'(X'X)^- P]^-(P'\hat{\beta} - \delta)/q}{\|Y - X\hat{\beta}\|^2/(n-r)} : F_{q, n-r}$$

$$r = r(X)$$

$$= \frac{SSH/q}{SSE/(n-r)}$$

$$SSE = \min_{\beta} \|Y - X\beta\|^2$$

$$RSS_H = \min_{P'\beta = \theta} \|Y - X\beta\|^2 = SSE + SSH$$

14 Sept. 01

Lets consider the restricted least squares problem

$$\min_{P'\beta = \delta} (Y - X\beta)'(Y - X\beta)$$

Use a Lagrange multiplier to handle restriction

Criterion $(Y - X\beta)'(Y - X\beta) + 2\lambda'(P'\beta - \delta)$

Diff

$$X'X\hat{\beta}_* + P\lambda = X'Y \quad (*)$$

$$P'\hat{\beta}_* = \delta \quad (**)$$

To "solve"

$$\begin{bmatrix} X'X & P \\ P' & 0 \end{bmatrix} \begin{bmatrix} \hat{\beta}_* \\ \lambda \end{bmatrix} = \begin{bmatrix} X'Y \\ \delta \end{bmatrix}$$

Rohde <u>SIAM J Appl Math</u> 13, 1033-1035

$\begin{bmatrix} A & C \\ C' & B \end{bmatrix}$ nonneg.d. has g-inverse

$$\begin{bmatrix} A^- + A^-CD^-C'A^- & -A^-CD^- \\ -D^-C'A^- & D^- \end{bmatrix}$$

where $D = B - C'A^-C$

14 Sept. 01

Suppose $\begin{bmatrix} X'X & P \\ P' & 0 \end{bmatrix}$ nonsingular

From (*) $\hat{\beta}_* = (X'X)^-(X'Y - P\lambda)$ — unique value if consistent eq^n

$= (X'X)^- X'Y - (X'X)^- P\lambda$ (***)

So $P'\hat{\beta}_* = P'(X'X)^- X'Y - P'(X'X)^- P\lambda = \delta$ from (**)

and $\lambda = [P'(X'X)^- P]^{-1}[P'(X'X)^- X'Y - \delta]$ a solution

so $\hat{\beta}_* = (X'X)^- X'Y - (X'X)^- P[P'(X'X)^- P]^- [P'(X'X)^- X'Y - \delta]$

What is restricted minimum?

$(Y - X\hat{\beta}_*)'(Y - X\hat{\beta}_*) = [Y - X\hat{\beta} + X(\hat{\beta} - \hat{\beta}_*)]'[\qquad]$

$= (Y - X\hat{\beta})'(Y - X\hat{\beta}) + (\hat{\beta} - \hat{\beta}_*)'X'X(\hat{\beta} - \hat{\beta}_*)$

since $X'(Y - X\hat{\beta}) = 0$

Note that $(\hat{\beta} - \hat{\beta}_*)'X'X(\hat{\beta} - \hat{\beta}_*)$ the numerator of the

$= [P'\hat{\beta} - \delta]'[P'(X'X)^{-1}P]^- P'(X'X)^{-1}X'X(X'X)^{-1}P[P'(X'X)^- P]^-$
$[P'\hat{\beta} - \delta]$

$=$ SSH numerator of F statistic

14 Sept 01

ANOVA Table

Source	SS	df
Deviation from $H, \underset{\sim}{P}'\underset{\sim}{\beta} = \underset{\sim}{\delta}$	$(\hat{\underset{\sim}{\beta}} - \hat{\underset{\sim}{\beta}}_*)' \underset{\sim}{X}' \underset{\sim}{X} (\hat{\underset{\sim}{\beta}} - \hat{\underset{\sim}{\beta}}_*)$ $\equiv SSH$	q
Residual (unrestricted β)	SSE	$n - r$
Total (restricted β)	$(\underset{\sim}{Y} - \underset{\sim}{X}\hat{\underset{\sim}{\beta}}_*)'(\underset{\sim}{Y} - \underset{\sim}{X}\hat{\underset{\sim}{\beta}}_*)$	$n - r + q$

$r = r(\underset{\sim}{X})$, $q = r(\underset{\sim}{P})$

If $\delta = 0$ 14 Sept. 01

$$Y'Y = (Y - X\hat{\beta}_* + X\hat{\beta}_*)'(Y - X\hat{\beta}_* + X\hat{\beta}_*)$$

$$= (Y - X\hat{\beta}_*)'(Y - X\hat{\beta}_*) + \hat{\beta}_*' X' X \hat{\beta}_*$$

For

$$\hat{\beta}_*' X'(Y - X\hat{\beta}_*) = \hat{\beta}_*' P \lambda \quad \text{from } (*)$$

$$= 0 \quad \text{from } (**)$$

When $\delta = 0$

ANOVA Table

Source	SS	df
Reduced model	$\lVert X\hat{\beta}_* \rVert^2$	$r - q$
Deviation from H	$SSH = \lVert X(\hat{\beta} - \hat{\beta}_*) \rVert^2$	q
Error	$SSE = \lVert Y - X\hat{\beta} \rVert^2$	$n - r$
Total	$SST = \lVert Y \rVert^2$	n

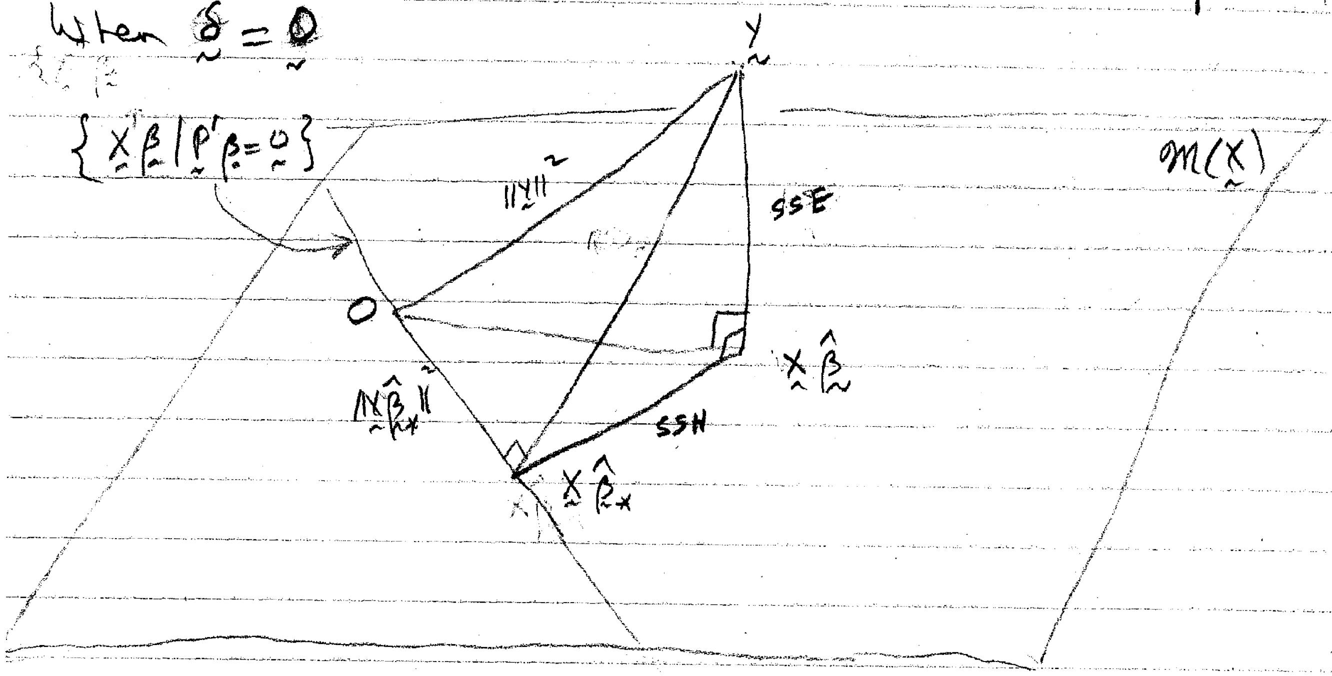
"correct" for β_0 H: $P'\beta = P'\beta_0 = \delta_0$

	SS	df
To	$\lVert X(\hat{\beta}_* - \beta_0) \rVert^2$	$r - q$
	$\lVert X(\hat{\beta} - \hat{\beta}_*) \rVert^2$	q
	$\lVert Y - X\hat{\beta} \rVert^2$	$n - r$
	$\lVert Y - X\beta_0 \rVert^2$	n

Can use Fisher - Cochran

14 Sept. 01

When $\underset{\sim}{\delta} = \underset{\sim}{0}$

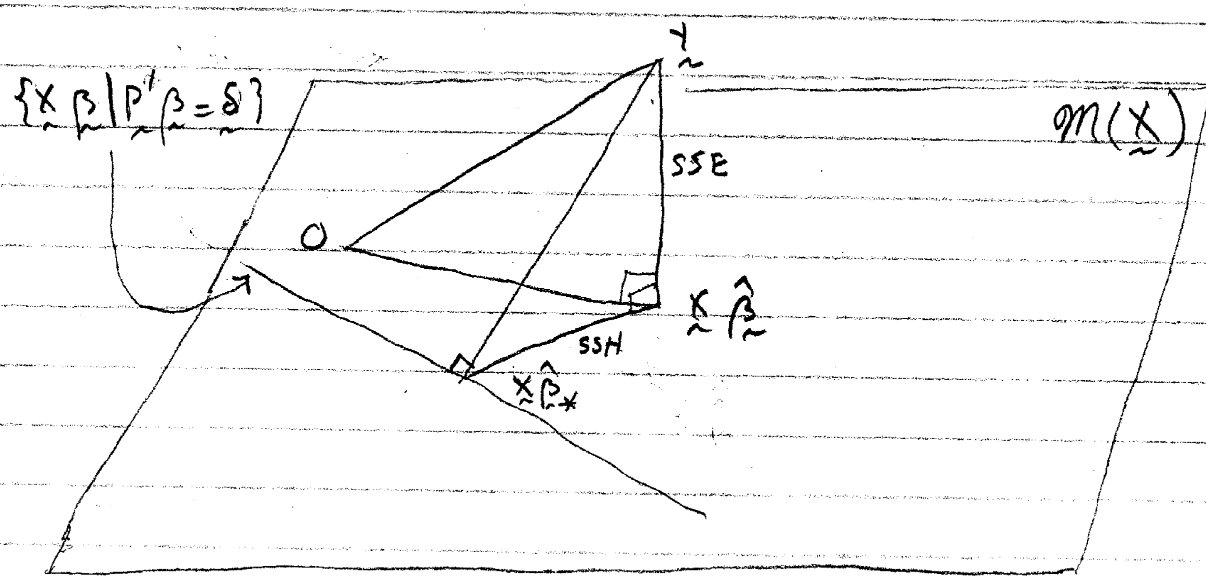


$$\|\underset{\sim}{Y}\|^2 = \|\underset{\sim}{X}\hat{\underset{\sim}{\beta}}_*\|^2 + \|\underset{\sim}{X}(\hat{\underset{\sim}{\beta}} - \hat{\underset{\sim}{\beta}}_*)\|^2 + \|\underset{\sim}{Y} - \underset{\sim}{X}\hat{\underset{\sim}{\beta}}\|^2$$

If line doesn't go through O don't have $\perp$

14 Sept 01

Case of Non-zero Constraint



No longer have

$$\|\underset{\sim}{Y}\|^2 = \|\underset{\sim}{X}\hat{\underset{\sim}{\beta}}_*\|^2 + \|\underset{\sim}{X}(\hat{\underset{\sim}{\beta}} - \hat{\underset{\sim}{\beta}}_*)\|^2 + \|\underset{\sim}{Y} - \underset{\sim}{X}\hat{\underset{\sim}{\beta}}\|^2$$

14 Sept. 01

Anova table for succession of (nested) hypotheses

Can make SS smaller + smaller

Get different results if not orthogonal

Entry of table = additional reduction

19 Sept 01

One-way ANOVA

$$Y_{jk} = \mu_j + \epsilon_{jk} \qquad k = 1, \dots, K_j$$
$$j = 1, \dots, J$$

OR

$$Y_{jk} = \mu + \alpha_j + \epsilon_{jk}$$
$$\sum_1^J \alpha_j = 0$$

ANOVA TABLE

Source	SS	DF
Between treatments	$\sum_j \sum_k (\bar{Y}_{j.} - \bar{Y})^2$	$J-1$
Residual	$\sum_j \sum_k (Y_{jk} - \bar{Y}_{j.})^2$	$n-J$
Total	$\sum_j \sum_k (Y_{jk} - \bar{Y}_{..})^2$	n

Joint strengths for wood (and glue)

Tensile strength were measured for 9 treatments. There were 16 units.

Data

Treatment	Stress at failure (psi)	$\hat{\mu}_j$
1	829, 596	712.5 (120.1)
2	1169	1169.0 (169.9)
3	1263, 1029	1146.0 (120.1)
⋮		
9	1489	1489.0 (169.9)

Model I. $Y_{jk} = \mu_j + \epsilon_{jk}$

Model II. $Y_{jk} = \mu + \alpha_j + \epsilon_{jk}$ $\quad \sum_{i}^{I} \alpha_j = 0$

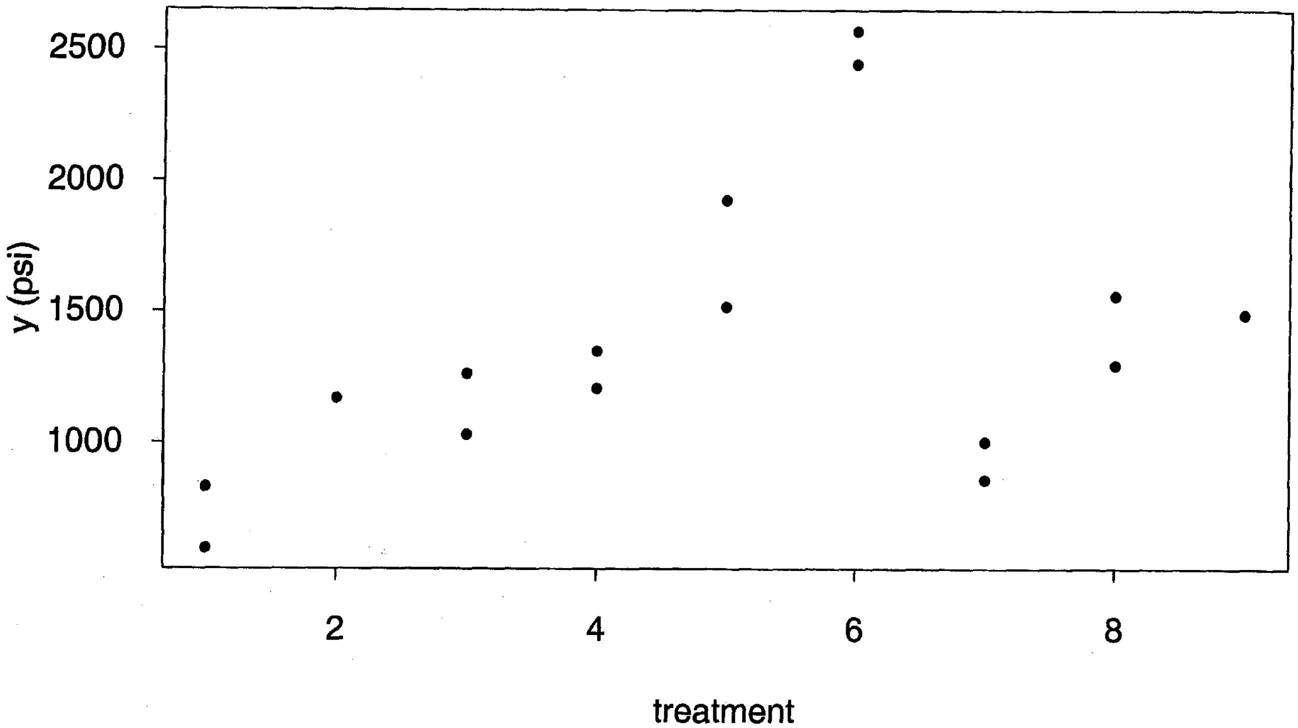
$\hat{\mu} = 1375.8$
$\hat{\alpha}_1 = -663.3$
$\hat{\alpha}_2 = -206.8$
⋮

Question: Do the treatments differ?

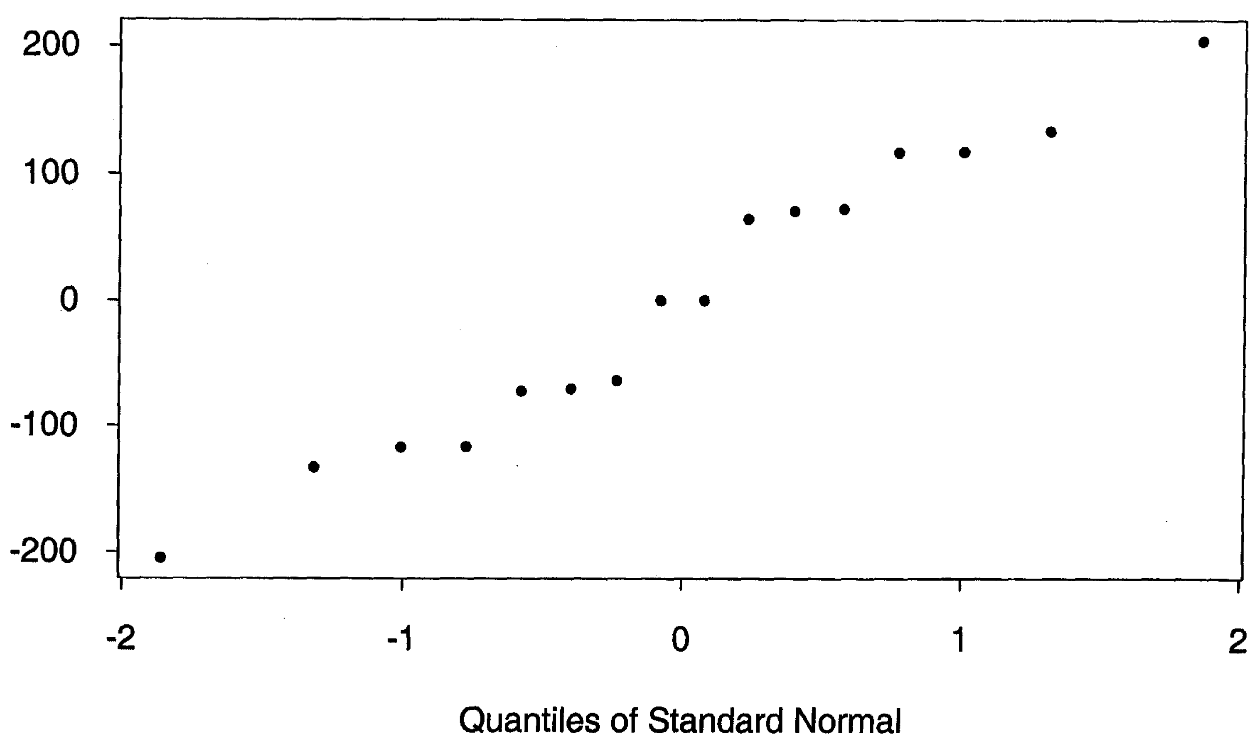
ANOVA TABLE

Source	SS	DF	MS
Between treatments	4260674	8	532584.1
Residual	202042	7	28863.1
Total	4462715	15	

Stress at failure



Normal prob plot of residuals fitting single factor



Actually the treatments had a factorial structure

	A	B
1 ≡	(1,	1)
2 ≡	(1,	2)
3 ≡	(1,	3)
⋮		

There were 3 different types of joints (butt, lap, beveled) and 3 types of wood (pine, oak, walnut)

One can reparametrize to reflect this

Data Y_{mno} $m = 1, 2, 3$
$n = 1, 2, 3$
$o = 1, O_{mn}$ replicate

$$E Y_{mno} = E Y_{mn}$$

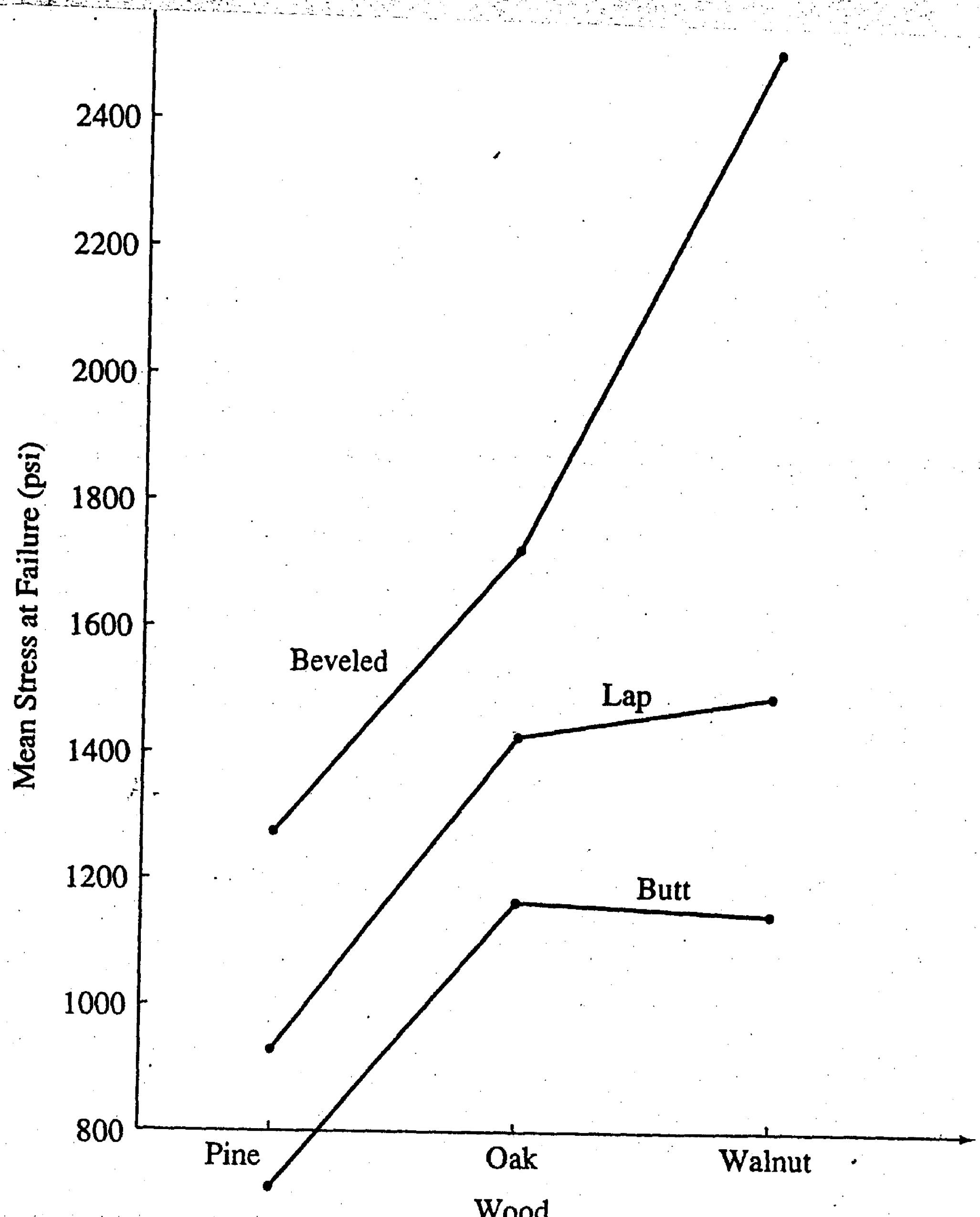
$$= \mu + \alpha_m + \beta_n + \gamma_{mn}$$

with $\sum_m \alpha_m = 0 = \sum_n \beta_n$

$$\sum_m \gamma_{mn} = 0 = \sum_n \gamma_{mn}$$

γ_{mn} : <u>interaction</u>

Specimen	Joint	Wood	y Stress at Failure (psi)
1	beveled	oak	1518
2	butt	pine	829
3	beveled	walnut	2571
4	butt	oak	1169
5	beveled	oak	1927
6	beveled	pine	1348
7	lap	walnut	1489
8	beveled	walnut	2443
9	butt	walnut	1263
10	lap	oak	1295
11	lap	oak	1561
12	lap	pine	1000
13	butt	pine	596
14	lap	pine	859
15	butt	walnut	1029
16	beveled	pine	1207



2400
2200
2000
1800
1600
1400
1200
1000
800
Mean Stress at Failure (psi)
Beveled
Lap
Butt
Pine
Oak
Walnut
Wood

The ANOVA Table will look the same, but perhaps things simplify.

Perhaps $\gamma_{mn} \equiv 0 \quad \forall m, n$

i.e. the model is additive

Fit $E Y_{mno} = \mu + \alpha_m + \beta_n$

ANOVA TABLE

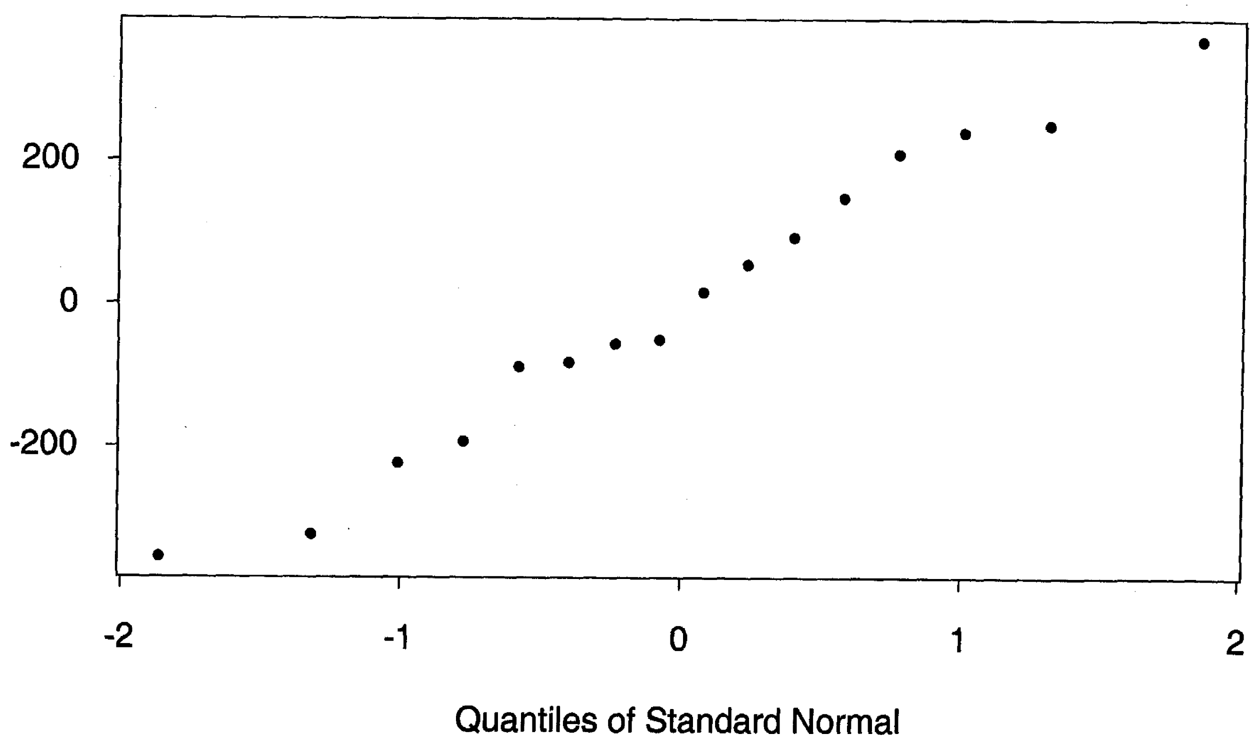
Source	SS	DF	MS
A+B	3792143	4	948035.8
A+B+A:B	468530	4	117132.5
Residual	202041	7	28863
Total	4462715	15	

Generally

ANOVA Table if balanced $O_{mn} = O$

Source	SS	DF
A+B	$\sum\sum\sum(\bar{Y}_{m..} - \bar{Y}_{...})^2 + \sum\sum\sum(\bar{Y}_{.n.} - \bar{Y}_{...})^2$	M+N-2
A+B+AB	$\sum\sum\sum(\bar{Y}_{mn.} - \bar{Y}_{m..} - \bar{Y}_{.n.} + \bar{Y}_{...})^2$	(M-1)(N-1)
Residual	$\sum\sum\sum(Y_{mno} - \bar{Y}_{mn.})^2$	MN(O-1)
Total	$\sum\sum\sum(Y_{mno} - \bar{Y}_{...})^2$	MNO-1

Residuals fitting additive model



19 Sept. 2001

Experiment.
designed study
often comparative

Object to compare a number of treatments

Treatments. things being compared

Experimental unit. smallest division of the experimental material such that any two units may receive different treatments in the actual experiment

Factor. A basic treatment
The possible forms of a factor are its levels

Factorial experiment. All factor combinations are of interest.