2 April 08

Sums of squares.

Normal linear model

$$\underset{n\times 1}{y} = \underset{1\times 1}{1_n \beta_0} + \underset{n\times(r_1-1-\nu_1)}{X_1 \beta_1} + \underset{n\times(\nu_2-\nu_1)}{X_2 \beta_2} + \dots + \underset{n\times(\nu_m-\nu_{m-1})}{X_m \beta_m} + \underset{n\times 1}{\varepsilon}$$

simplest model $y = 1_n \beta_0 + \varepsilon$

$\hat\beta = \bar y \quad \hat y_0 = 1_n \bar y \quad SS_0 = \sum (y_j - \bar y)^2 \quad df = \nu_0 = n-1$

suppose bring in X's successively

$$y - \hat y_0 = (y - \hat y_m) + (\hat y_m - \hat y_{m-1}) + \dots + (\hat y_1 - \hat y_0)$$

terms orthogonal

$$SS_0 = SS_m + (SS_{m-1} - SS_m) + \dots + (SS_0 - SS_1)$$

$SS_{r-1} - SS_r$: reduction in residual SS due to adding X_r

$$\|y - \hat y_0\|^2 = \|y - \hat y_m\|^2 + \|\hat y_m - \hat y_{m-1}\|^2 + \dots + \|\hat y_1 - \hat y_0\|^2 \quad (*)$$

y normal $\Rightarrow$ $\hat y_r - \hat y_{r-1}$, $y - \hat y_m$ normal

SS_m & $SS_{r-1} - SS_r$ independent

2 April 08

ANOVA Table based on (x)

Terms added	df	Reductions is SS	MS = SS/df
X_1	$n-1-\nu_1$	$SS_0 - SS_1$	
X_2	$\nu_1 - \nu_2$	$SS_1 - SS_2$	
X_3	$\nu_2 - \nu_3$	$SS_2 - SS_3$	
$\vdots$			
X_m	$\nu_{m-1} - \nu_m$	$SS_{m-1} - SS_m$	
Residual	ν_m	SS_m	
Total	$n-1$	SS_0	

F ratios $$\frac{(SS_{r-1} - SS_r)/(\nu_{r-1} - \nu_r)}{SS_m / \nu_m}$$

Cement data. p. 381

$$y_j = \beta_0 + \beta_1 x_{1j} + \beta_2 x_{2j} + \beta_3 x_{3j} + \beta_4 x_{4j} + \varepsilon_j$$

$n = 13$ β 5×1

[illegible]le 8.1 Cement data ([illegible]oods *et al.*, 1932): y is [illegible] evolved in calories [illegible]gram of cement, and [illegible] x_2, x_3, and x_4 are [illegible]entage weight of [illegible]kers, with x_1, [illegible]$aO.Al_2O_3$, x_2, [illegible]$O.SiO_2$, x_3, [illegible]$O.Al_2O_3.Fe_2O_3$, [illegible] x_4, $2CaO.SiO_2$.

Case	x_1	x_2	x_3	x_4	y
1	7	26	6	60	78.5
2	1	29	15	52	74.3
3	11	56	8	20	104.3
4	11	31	8	47	87.6
5	7	52	6	33	95.9
6	11	55	9	22	109.2
7	3	71	17	6	102.7
8	1	31	22	44	72.5
9	2	54	18	22	93.1
10	21	47	4	26	115.9
11	1	40	23	34	83.8
12	11	66	9	12	113.3
13	10	68	8	12	109.4

Bring in x's successively

```
junk<-matrix(scan("cement"),byrow=T,ncol=5)
x1<-junk[,1];x2<-junk[,2];x3<-junk[,3];x4<-junk[,4];y<-junk[,5]
m0<-lm(y~1)
m1<-lm(y~x1)
m2<-lm(y~x1+x2)
m3<-lm(y~x1+x2+x3)
m4<-lm(y~x1+x2+x3+x4)
anova(m0,m1,m2,m3,m4)
```

```
Model 1: y ~ 1
Model 2: y ~ x1
Model 3: y ~ x1 + x2
Model 4: y ~ x1 + x2 + x3
Model 5: y ~ x1 + x2 + x3 + x4
  Res.Df     RSS Df Sum of Sq        F    Pr(>F)
1     12 2715.76
2     11 1265.69  1   1450.08 242.3679 2.888e-07 ***
3     10   57.90  1   1207.78 201.8705 5.863e-07 ***
4      9   48.11  1      9.79   1.6370    0.2366
5      8   47.86  1      0.25   0.0413    0.8441
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Orthogonality

The order of inserting terms affects their reduction in sum of squares, generally

Suppose there are X_1 and X_2 and

$$y = 1_n \beta_0 + X_1 \beta_1 + X_2 \beta_2 + \varepsilon$$

The normal equations are

$$\begin{bmatrix} 1^T 1 & 1^T X_1 & 1' X_2 \\ X_1^T 1 & X_1^T X_1 & X_1^T X_2 \\ X_2^T 1 & X_2^T X_1 & X_2^T X_2 \end{bmatrix} \begin{bmatrix} \hat\beta_0 \\ \hat\beta_1 \\ \hat\beta_2 \end{bmatrix} = \begin{bmatrix} 1^T y \\ X_1^T y \\ X_2^T y \end{bmatrix} \quad (**)$$

Suppose $X_1^T 1, X_2^T 1, X_1^T X_2 = 0$ orthogonal

Then (**)

$$\begin{bmatrix} 1^T 1 & 0 & 0 \\ 0 & X_1^T X_1 & 0 \\ 0 & 0 & X_2^T X_2 \end{bmatrix} \begin{bmatrix} \hat\beta_0 \\ \hat\beta_1 \\ \hat\beta_2 \end{bmatrix} = \begin{bmatrix} 1^T y \\ X_1^T y \\ X_2^T y \end{bmatrix}$$

If inverses exist

$$\hat\beta_0 = \bar{y}, \quad \hat\beta_r = (X_r^T X_r)^{-1} X_r^T y \qquad r = 1, 2$$

Residual sum of squares

$$(y - X\hat{\beta})^T (y - X\hat{\beta}) = y^T y - n\bar{y}^2 - \hat{\beta}_1 X_1^T X_1 \hat{\beta}_1 - \hat{\beta}_2 X_2^T X_2 \hat{\beta}_2$$

Order of fitting does not matter

If ε's are $IN(0, \sigma^2)$, then

$$\hat{\beta}_0 \sim N(\beta_0, \frac{\sigma^2}{n})$$

$$\perp$$

$$\hat{\beta}_1 \sim N(\beta_1, (X_1^T X_1)^{-1} \sigma^2)$$

$$\perp$$

$$\hat{\beta}_2 \sim N(\beta_2, (X_2^T X_2)^{-1} \sigma^2)$$

cycling data

7

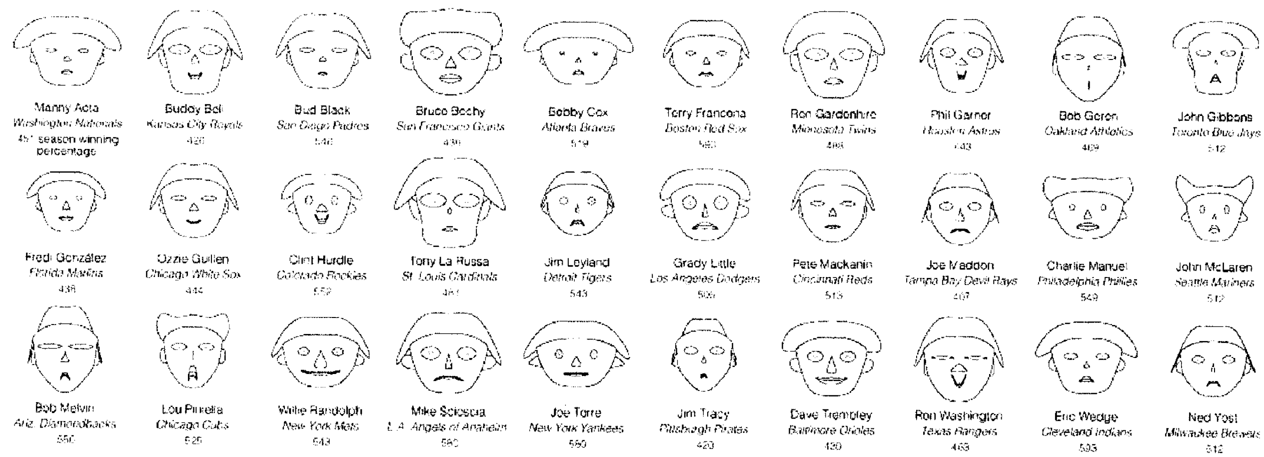
Table 8.2 Data and experimental setup for bicycle experiment (Box et al., 1978, pp. 368–372). The lower part of the table shows the average times for each of the eight combinations of settings of seat height, tyre pressure, and dynamo, and the average times for the eight observations at each setting, considered separately.

Setup	Day	Run	Seat height (inches)	Dynamo	Tyre pressure (psi)	Time (secs)
1	3	2	−	−	−	51
2	4	1	−	−	−	54
3	2	2	+	−	−	41
4	2	3	+	−	−	43
5	3	3	−	+	−	54
6	2	1	−	+	−	60
7	3	1	+	+	−	44
8	4	3	+	+	−	43
9	1	1	−	−	+	50
10	4	4	−	−	+	48
11	3	5	+	−	+	39
12	4	2	+	−	+	39
13	3	4	−	+	+	53
14	1	3	−	+	+	51
15	1	2	+	+	+	41
16	2	4	+	+	+	44

$$\begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ y_5 \\ y_6 \\ y_7 \\ y_8 \\ y_9 \\ y_{10} \\ y_{11} \\ y_{12} \\ y_{13} \\ y_{14} \\ y_{15} \\ y_{16} \end{pmatrix} = \begin{pmatrix} 1 & -1 & -1 & -1 \\ 1 & -1 & -1 & -1 \\ 1 & 1 & -1 & -1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & 1 & -1 \\ 1 & -1 & 1 & -1 \\ 1 & 1 & 1 & -1 \\ 1 & 1 & 1 & -1 \\ 1 & -1 & -1 & 1 \\ 1 & -1 & -1 & 1 \\ 1 & 1 & -1 & 1 \\ 1 & 1 & -1 & 1 \\ 1 & -1 & 1 & 1 \\ 1 & -1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \varepsilon_4 \\ \varepsilon_5 \\ \varepsilon_6 \\ \varepsilon_7 \\ \varepsilon_8 \\ \varepsilon_9 \\ \varepsilon_{10} \\ \varepsilon_{11} \\ \varepsilon_{12} \\ \varepsilon_{13} \\ \varepsilon_{14} \\ \varepsilon_{15} \\ \varepsilon_{16} \end{pmatrix}. \qquad (8.2)$$

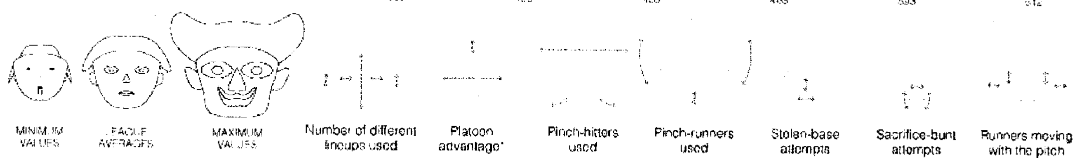
The New York Times

April 1, 200



SMILE IF YOU BUNT

Steve C. Wang, an associate professor of statistics at Swarthmore College, charted baseball managers from the 2007 season as Chernoff faces, a method of using the heights, widths and angles of facial features to represent different sets of numbers



*Percentage of players who had the advantage of batting against an opposite-handed pitcher at the start of the game.

Note: Because different rules cause National League managers to use more pinch-hitters, for example, each manager's rates are compared with his league's average

Close Window

Print This Image »

6 April 08

Model checking

What are the assumptions of the normal/Gaussian regression model?

$Y = X\beta + \epsilon \qquad \epsilon\text{'s } IN(0, \sigma^2)$

1. linearity in β, X
2. constant variance σ^2
3. independence ε's (uncorrelated)
4. normality

What to check for?
isolated discrepancy, outliers
systematic dependency
transform of y needed
transformation of covariate
omitted variable
correlated errors

How?
residuals

Raw residuals

$$\hat{\varepsilon} = e = y - \hat{y} = y - X(X^TX)^{-1}X^Ty$$
$$= (I - H)y \qquad H = X(X^TX)^{-1}X^T$$
$$= (I - H)\varepsilon$$

Properties

$$E(e) = 0$$
$$\mathrm{var}(e) = \sigma^2(I_n - H)$$

$$\mathrm{var}\, e_j = \sigma^2(1 - h_{jj})$$

$$\mathrm{cov}(e_j, e_k) = -\sigma^2 h_{jk} \qquad j \neq k$$

 6 April 08

<u>Standardized residuals</u>

$$r_j = e_j \Big/ s\sqrt{(1-h_{jj})}$$
$$= (y_j - x_j^T\hat{\beta}) \Big/ s\sqrt{1-h_{jj}}$$

$E r_j = 0$ Why?

$\text{var}\, r_j \approx 1$

Check on <u>linearity</u>

Plot $\underset{n\times 1}{y}$ on $\underset{n\times 1}{X_\ell}$ $\ell = 1, \dots, p$

Plot r on X_ℓ $\ell = 1, \dots, p$

$\underset{n\times 1}{r}$ on $\underset{n\times 1}{V}$, omitted variable

Look for pattern

(4)

Cycling data

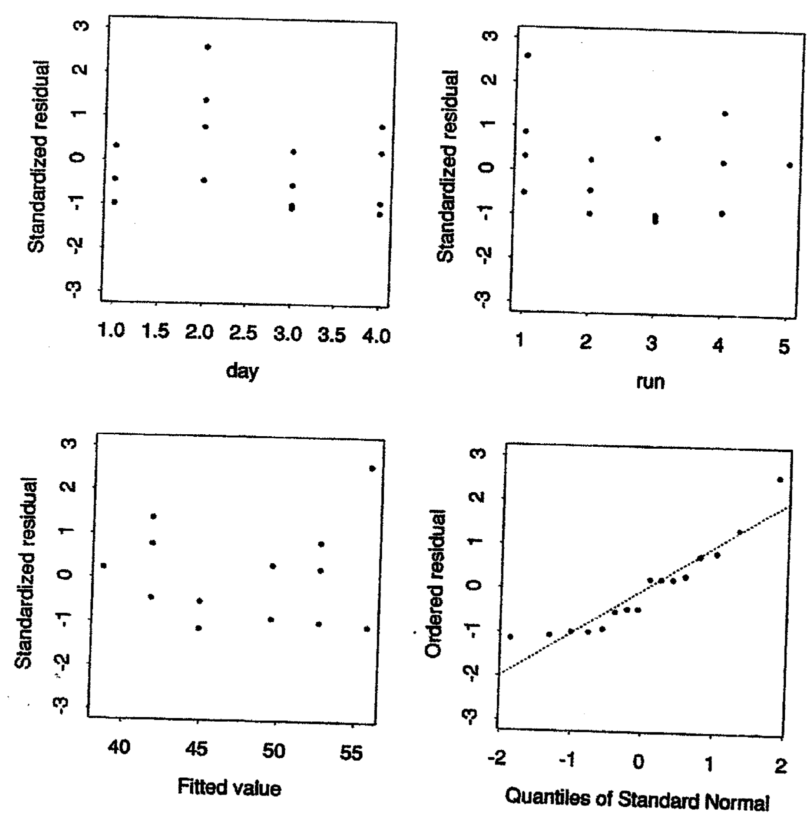


Figure 8.4 Residual plots for data on cycling up a hill. The panels showing residuals plotte against levels of day an run, and against fitted values, would show random variation if the model is adequate, as seems to be the case. Th normal scores plot show that the errors appear close to normal.

<u>Constancy of variance</u> r_j (or $|r_j|$) vs $\hat{y}$

wedging

<u>Independence</u> r_j vs. t_j or run number

<u>Distribution of errors</u>

normal prob plot of r_j

Nonlinearity

$$a(y) = x^T \beta + \varepsilon$$

$$y = a^{-1}(x^T \beta + \varepsilon)$$

$$y = b(x^T \beta) + \varepsilon$$

Box-Cox transform

$$u^{(\lambda)} = a(u) = \begin{cases} (u^\lambda - 1)/\lambda & \lambda \neq 0 \\ \log u & \lambda = 0 \end{cases}$$

$$y^{(\lambda)} = X\beta + \varepsilon$$

Jacobian $\frac{dy^{(\lambda)}}{dy} = \lambda y^{\lambda-1}$

$$g(y^{(\lambda)})\,dy^{(\lambda)} = g(y^{(\lambda)})\,\lambda y^{\lambda-1}\,dy$$

λ fixed

log likelihood

$$-\frac{1}{2}\left\{ n \log \sigma^2 + \frac{1}{\sigma^2} \sum_{j=1}^{n} \left(y_j^{(\lambda)} - x_j^T \beta\right)^2 \right.$$

$$+ (\lambda - 1) \sum \log y_j$$

$$\hat{\beta}_\lambda = (X^T X)^{-1} X^T y^{(\lambda)}$$

$$SS(\hat{\beta}_\lambda)/n = \hat{\sigma}^2_\lambda$$

profile log likelihood

$$\ell_p(\lambda) = -\frac{n}{2}\left\{ \log SS(\hat{\beta}_\lambda) - \log g^{2(\lambda-1)} \right\}$$

$$g = \left(\prod y_j\right)^{1/n}$$

Generalized additive model

$$y = b(x^T\beta) + \varepsilon$$

b: smooth
spline

lowess( ), loess( ), gam( )

<u>leverage</u> h_{jj} of j-th case

$$\hat{y} = X\hat{\beta}$$
$$= X(X^TX)^{-1}X^Ty$$
$$= Hy$$

$$\hat{y}_j = h_{jj}y_j + \sum_{j \neq i} h_{ij}y_i$$

$\hat{y}_j$ will be dominated by y_j if

y_j is an outlier

$$0 \leq h_{jj} \leq 1$$

If h_{jj} is large, changing y_j will move $\hat{y}_j$ a lot

$$tr(H) = \sum_i h_{ii} = p$$

Cases with $h_{jj} > 2p/n$ deserve close inspection

Table 8.3 Data from bicycle experiment, together with fitted values $\widehat{y}$, raw residuals e, standardized residuals, r, deletion residuals r', leverages h and Cook distances C.

Setup	Seat height	Dynamo	Tyre pressure	Time y	$\widehat{y}$	e	r	r'	h	C
1	−1	−1	−1	51	52.62	−1.625	−0.99	−0.99	0.25	0.08
2	−1	−1	−1	54	52.62	1.375	−0.84	0.83	0.25	0.06
3	1	−1	−1	41	41.75	−0.750	−0.46	−0.44	0.25	0.02
4	1	−1	−1	43	41.75	1.250	0.76	0.75	0.25	0.05
5	−1	1	−1	54	55.75	−1.750	−1.06	−1.07	0.25	0.09
6	−1	1	−1	60	55.75	4.250	2.59	3.72	0.25	0.56
7	1	1	−1	44	44.87	−0.875	−0.53	−0.52	0.25	0.02
8	1	1	−1	43	44.87	−1.875	−1.14	−1.16	0.25	0.11
9	−1	−1	1	50	49.50	0.500	0.30	0.29	0.25	0.01
10	−1	−1	1	48	49.50	−1.500	−0.91	−0.91	0.25	0.07
11	1	−1	1	39	38.62	0.375	0.23	0.22	0.25	0.00
12	1	−1	1	39	38.62	0.375	0.23	0.22	0.25	0.00
13	−1	1	1	53	52.62	0.375	0.23	0.22	0.25	0.00
14	−1	1	1	51	52.62	−1.625	−0.99	−0.99	0.25	0.08
15	1	1	1	41	41.75	−0.750	−0.46	−0.44	0.25	0.02
16	1	1	1	44	41.75	2.250	1.37	1.43	0.25	0.16

Sometimes e_j is called a *raw residual*.

we obtain $\widehat{\beta} = (X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}y$. The fitted value $\widehat{y} = X\widehat{\beta} = X(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}y = Hy$ is the orthogonal projection of y onto the plane spanned by the columns of X, and the matrix representing that projection is H. Notice that $\widehat{y}$ is unique whether or not $X^{\mathrm{T}}X$ is invertible.

Figure 8.2 shows that the vector of residuals, $e = y - \widehat{y} = (I_n - H)y$, and the vector of fitted values, $\widehat{y} = Hy$, are orthogonal. To see this algebraically, note that

$$\widehat{y}^{\mathrm{T}}e = y^{\mathrm{T}}H^{\mathrm{T}}(I_n - H)y = y^{\mathrm{T}}(H - H)y = 0, \tag{8.6}$$

because $H^{\mathrm{T}} = H$ and $HH = H$, that is, the projection matrix H is symmetric and idempotent (Exercise 8.2.5). The close link between orthogonality and independence for normally distributed vectors means that (8.6) has important consequences, as we shall see in Section 8.3. For now, notice that (8.6) implies that

$$y^{\mathrm{T}}y = (y - \widehat{y} + \widehat{y})^{\mathrm{T}}(y - \widehat{y} + \widehat{y}) = (e + \widehat{y})^{\mathrm{T}}(e + \widehat{y}) = e^{\mathrm{T}}e + \widehat{y}^{\mathrm{T}}\widehat{y}, \tag{8.7}$$

as is clear from Figure 8.2 by Pythagoras' theorem. That is, the overall sum of squares of the data, $\sum y_j^2 = y^{\mathrm{T}}y$, equals the sum of the residual sum of squares, $SS(\widehat{\beta}) = \sum(y_j - \widehat{y}_j)^2 = e^{\mathrm{T}}e$, and the sum of squares for the fitted model, $\sum \widehat{y}_j^2 = \widehat{y}^{\mathrm{T}}\widehat{y}$. Such decompositions are central to analysis of variance, discussed below.

8.2.3 Likelihood quantities

Chapter 4 shows how the observed and expected information matrices play a central role in likelihood inference, by providing approximate variances for maximum likelihood estimates. To obtain these matrices for the normal linear model, note that the

Simple linear regression

$$y_j = \gamma_0 + (x_j - \bar{x})\gamma_1 + \varepsilon_j$$

$$\begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_1 - \bar{x} \\ \vdots & \vdots \\ 1 & x_n - \bar{x} \end{pmatrix} \begin{pmatrix} \gamma_0 \\ \gamma_1 \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

$$\hat{\beta} = \begin{pmatrix} \hat{\gamma}_0 \\ \hat{\gamma}_1 \end{pmatrix} = \begin{pmatrix} n & \sum(x_j - \bar{x}) \\ \sum(x_j - \bar{x}) & \sum(x_j - \bar{x})^2 \end{pmatrix}^{-1} \begin{pmatrix} \sum y_j \\ \sum(x_j - \bar{x}) y_j \end{pmatrix}$$

$$= \underbrace{\begin{pmatrix} n^{-1} & 0 \\ 0 & 1/\sum(x_j - \bar{x})^2 \end{pmatrix}}_{(X^T X)^{-1}} \begin{pmatrix} \sum y_j \\ \sum(x_j - \bar{x}) y_j \end{pmatrix}$$

$$\overset{n\times 2}{X} \overset{2\times 2}{(X^T X)^{-1}} \overset{2\times n}{X^T}$$

$$= \left(\frac{1}{n} + \frac{(x_j - \bar{x})^2}{\sum (x_k - \bar{x})^2} \right)$$