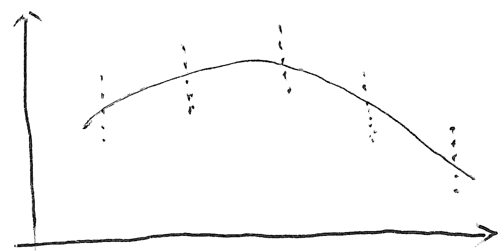
Analysis of factor effects if Factor is quantitative

Then one can try to see if there is a "response function" relating predictor x and response y.

Ex: x = price y = sales.



(1). plot and see (dot plot)

try different curves

(2) Then test for lack of fit

$$H_0: Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \varepsilon_i$$

	Regression		ANOVA	
SSR	$p-1$		SSTR	$r-1$
SSE	$n_T - p$		SSPE	$n_T - r$
SSTO	$n_T - 1$		SSTO	$n_T - 1$

$$SSLF = SSTR - SSR = SSE - SSPE$$

$$F_{LOF} = \frac{MSLF}{MSPE} = \frac{SSLF/(r-p)}{SSPE/(n_T-r)}$$

Ex: Page 771

Chapter 18 ANOVA Diagnostics

(1) Examine if the model is appropriate (check assumptions)
indep. constant variance
Normal

(2) If not, consider remedial measures.

(3) after that, repeat inference

---

① Analysis of residuals.

Assumption: $\varepsilon_{ij} \sim$ iid $N(0, \sigma^2)$

$$e_{ij} := y_{ij} - \bar{y}_{i\cdot}$$

- standardized residual (divided by an est. of σ)

$$e_{ij}^* = \frac{e_{ij}}{\sqrt{MSE}}$$

e_{ij}^* approximately $N(0,1)$.

$|e_{ij}^*| > 3$ are potential outliers.

- studentized residuals

$$Var(e_{ij}) = Var(y_{ij}) + Var(\bar{y}_{i\cdot}) - 2\,Cov(y_{ij}, \bar{y}_{i\cdot})$$
$$= \sigma^2 + \frac{\sigma^2}{n_i} - 2\frac{\sigma^2}{n_i} = \sigma^2\left(1 - \frac{1}{n_i}\right)$$

$$S(e_{ij}) = s\sqrt{1 - \frac{1}{n_i}}$$

$$r_{ij} = \frac{e_{ij}}{s\sqrt{1 - \frac{1}{n_i}}}$$

- studentized deleted residuals

(delete case ij, and refit the model compute the predicted value and residual for the case ij)

$$t_{ij} = \frac{e_{ij}}{\sqrt{\dfrac{SSE(1 - \frac{1}{n_i}) - e_{ij}^2}{n_T - r - 1}}}$$

- both $\bar{y}_{i\cdot}$ and $s^2_{(-ij)}$ excludes y_{ij},

The new $\bar{y}_{i\cdot(-ij)} = \dfrac{n_i \bar{y}_{i\cdot} - y_{ij}}{n_i - 1}$, $e_{ij'} = y_{ij'} - \bar{y}_{i\cdot(-ij)}$

Outliers test

- We are selecting the further outlier, it is not legitimate to use a simple t-test.
- Bonferroni adjustment

Ex. if 5 observations, in 2 factor levels)

By the Bonferroni method, compare with

$$t_{\frac{.05}{2\times 5}} = t_{.005} = 9.925 \quad (\text{for } 5-2-1 = 2 \text{ df})$$

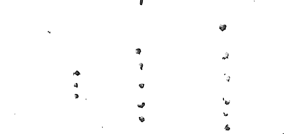
- Tools of diagnotics

Residual plots

r_{ij} vs $\bar{y}_{i\cdot}$

r_{ij} aligned dot plots , Normal prob. plot

They show:
- non constant variance



- non independent errors (in a time plot)
- outliers (Bonferroni , check 2-3 points)
- nonnormality

· Test for constant variance

$H_0: \sigma_1^2 = \sigma_2^2 = \cdots = \sigma_r^2$ vs H_A: H_0 not hold.

① Hartley test.

$$H_{obs} = \frac{\max S_i^2}{\min S_i^2},$$ one-sided test

it has H statistic distr. under H_0 with parameters (r, n): r: # of levels

n: size of each group (need same n_i)

$n_i \approx n \Rightarrow$ still ok.

Use $H(1-\alpha, r, n-1)$ right-sided.

② Modified Levene test (Brown-Forsythe test)

$\tilde{y}_i = i^{th}$ median $=$ median of $(y_{i1}, \cdots y_{in_i})$

$d_{ij} = |y_{ij} - \tilde{y}_i|$

Idea: Under H_0, $E(d_{ij}) \equiv$ const

$\Rightarrow$ do ANOVA for $\{d_{ij}\}$

$$MSTR = \frac{\sum n_i (\bar{d}_{i\cdot} - \bar{d}_{\cdot\cdot})^2}{r-1}$$

$$MSE = \frac{\sum \sum (d_{ij} - \bar{d}_{i\cdot})^2}{n_T - r}$$

$$F_L^* = \frac{MSTR}{MSE}$$

Under H_0, $F_L^* \approx F_{(r-1,\, n_T - r)}$

Levene's – Use $\bar{y}_{i\cdot}$ mean

R command leveneTest( )

Ex: 6 patients with 3 treatment, the responses were recorded

A	30	40
B	20	50
C	10	60

Use the modified Levene test to test a const. variance

~~Tre~~Treatment	y_{ij}		$\tilde{y}_i$	$d_{ij} = \lvert y_{ij} - \tilde{y}_i \rvert$		$\bar{d}_{i\cdot}$	$d_{ij} - \bar{d}_{i\cdot}$		$\bar{d}_{i\cdot} - \bar{d}_{\cdot\cdot}$
A	30	40	35	5	5	5	0	0	−10
B	20	50	35	15	15	15	0	0	0
C	10	60	35	25	25	25	0	0	10

Test $H_0: \sigma_1 = \sigma_2 = \sigma_3$. H_A: H_0 doesn't hold.

Carry out an ANOVA F based on d_{ij}.

Get $SSTR = \sum_i n_i (\bar{d}_{i.} - \bar{d}_{..})^2 > 0$, $r-1 = 2$

$SSE = \sum \sum (d_{ij} - \bar{d}_{i.})^2 = 0$ $n_T - r = 3$

$\Rightarrow F = \frac{MSTR}{MSE} = \infty$, H_0 is rejected at any sig-level

Comment.

If $n_i \leq 2$ for all i, then $d_{ij} - \bar{d}_{i.} = 0$

So $SSE = 0$ $\Rightarrow F = +\infty$ always.

We hesitate to apply it.

③. Bartlett's test

doesn't require $n_1 = n_2 = \cdots n_r$.

$$\chi^2 = \frac{\sum_{i=1}^{r} \log \left(\frac{S^2_{pooled}}{S_i^2} \right)^{n_i - 1}}{1 + \frac{1}{3(r-1)} \left(\sum \frac{1}{n_i - 1} - \frac{1}{\sum (n_i - 1)} \right)} \approx \chi^2_{r-1}$$

one-sided

Rule of thumb

ANOVA F test can tolerate non-constant variance to some extent.

It is usually fine

$$\frac{\max\{S_i\}}{\min\{S_i\}} \leq 2 \text{ or even } 3, \text{ especially}$$

when n_i are roughly equal.

Remedies

$\begin{cases} \text{Normal} \\ \sigma_i^2 \neq \text{constant} \end{cases}$ $\Rightarrow$ weighted least squares

$\begin{cases} \text{non-Normal} \\ \sigma_i^2 \neq \text{constant} \end{cases}$ $\Rightarrow$ transform y_{ij}, box-cox

Don't help
or other departures
from assumption $\Rightarrow$ Nonparametric test

1. Weighted least squares

$$y_{ij} = \mu_i + \varepsilon_{ij}, \qquad \varepsilon_{ij} \sim N(0, \sigma_i^2)$$

estimate σ_i^2 by S_i^2 and use weights $w_i = \frac{1}{S_i^2}$

$$SSTR(W) = \sum_i w_i n_i (\bar{y}_{i.} - \bar{y}_{..})^2 = \sum_i n_i \left(\frac{\bar{y}_{i.} - \bar{y}_{..}}{S_i} \right)^2$$

$$SSE(w) = \sum_i w_i \sum_j (y_{ij} - \bar{y}_{i.})^2 = \sum_i \frac{1}{S_i^2} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{i.})^2$$

$$= \sum_i (n_i - 1) = n_T - r$$

Then $$F_W = \frac{MSTR(w)}{MSE(w)} \sim F_{(r-1,\, n_T - r)}$$

2. Transformations

variance - stabilizing

$\sigma^2 \sim \mu$ (poisson) $\Rightarrow y' = \sqrt{y}$ or $y' = \sqrt{y} + \sqrt{y+1}$

$\sigma^2 \sim \mu^2 \Rightarrow y' = \log y$

$\sigma^2 \sim \mu^4 \Rightarrow y' = \frac{1}{y}$

Generally, $\sigma \sim \mu^\alpha$

$$y \longrightarrow \begin{cases} y^{1-\alpha} & \alpha \neq 1 \\ \log y & \alpha = 1 \end{cases}$$

Box-cox

$y' = y^{\lambda}$, search for best λ numerically to minimize SSE.

Drawback:

- Except for a few special transformations ($\log$, $\sqrt{\ }$, $\frac{1}{y}$) the transformed response lacks natural interpretation.
- Transformation doesn't work (helps little) for symmetric but heavy-tailed distr. (many outliers)

- ANOVA F test is robust to non-normality, but it is not risistent to outliers.
- If outliers ~~are~~ cannot be removed, try non-parametric test.

③ Nonparametric Rank F-test.

Def: R_{ij} = rank of y_{ij} among all n_T obs.

($R_{ij} = r \Leftrightarrow y_{ij}$ is the rth smallest

If $y_{ij} = y_{i'j'} = y_{i''j''}$ etc $\Rightarrow$ average ranks)

Then do ANOVA of ranks

$$R_{ij} = \mu_i^{(R)} + \varepsilon_{ij}^{(R)}$$

$$F_R = \frac{MSTR(R)}{MSE(R)} \sim F_{r-1,\, n_T - r} \text{ if } n_i$$

are not very small.

Ex:

	Observations		
Treatment 1	-5	1	9
2	-4	3	11
3	0	6	17

$\Rightarrow$

	Ranks			
	1	4	7	4
	2	5	8	5
	3	6	9	6

$$SSTR = \sum_i n_i (\bar{R}_{i\cdot} - \bar{R}_{\cdot\cdot})^2 \qquad \bar{R}_{\cdot\cdot} = \frac{n_T + 1}{2} = 5$$

$$= 3(4-5)^2 + 3(5-5)^2 + 3(6-5)^2 = 6 \text{ with df } 2$$

$$SSE = \sum_i \sum_j (R_{ij} - \bar{R}_{i\cdot})^2$$

$$= (1-4)^2 + \cancel{(4-4)^2} + (7-4)^2 + (2-5)^2 + \cancel{(5-5)^2} + (8-5)^2 + (3-6)^2 + \cancel{(6-6)^2} + (9-6)^2$$

$$= 54 \text{ with df. } 6$$

$$F_R = \frac{SSTR/2}{SSE/6} = \frac{3}{9} = \frac{1}{3}$$

· Kruskal-Wallis Test

$H_0: \mu_1 = \mu_2 = \cdots \mu_r$

Test statistic.

$$\chi^2_{KW} = \frac{SS_{TR} \text{ (based on ranks)}}{SS_{Total} \text{ (based on ranks)} / (n_T - 1)}$$

$$\approx \chi^2_{r-1} \text{ under } H_0 \text{ for large } n.$$

$$SS_{Total} = \sum_i \sum_j (R_{ij} - \bar{R}_{..})^2$$

$$= \sum_{k=1}^{n_T} \left(k - \frac{n_T+1}{2}\right)^2$$

$$= \frac{n_T(n_T^2 - 1)}{12}$$

$$\chi^2_{KW} = \frac{SS_{TR}(R)}{n_T(n_T+1)/12}$$

Two way ANOVA

Consider two factors, $\underset{a}{A}$ and $\underset{b}{B}$

Treatment = each combination of a level of A and a level of B

Ex ; $a = 2$ levels of A
$b = 3$ levels of B 6 treatment

Say we have $n_T = 36$ obs, 6 for each treatment.

This is a complete design.

If not all, but only a fraction of treatments is used in study $\Rightarrow$ fractional design

2-way ANOVA model.

cell mean $\quad \mu_{ij} = E(Y \mid \text{ith level of A}, \text{jth level of B})$

$i = 1, \cdots a$

$j = 1, \cdots b$

$$\mu_{i\cdot} = \frac{1}{b}\sum_{j=1}^{b}\mu_{ij}$$

$$\mu_{\cdot j} = \frac{1}{a}\sum_{i=1}^{a}\mu_{ij}$$

$$\mu_{\cdot\cdot} = \frac{1}{ab}\sum_{i}\sum_{j}\mu_{ij}$$

main effects $\quad \alpha_i = \mu_{i\cdot} - \mu_{\cdot\cdot} \Rightarrow \sum_{i=1}^{a}\alpha_i = 0$

$\qquad\qquad\qquad \beta_j = \mu_{\cdot j} - \mu_{\cdot\cdot} \Rightarrow \sum_{j}\beta_j = 0$

Additive factor effects

$$\Leftrightarrow \mu_{ij} = \mu_{\cdot\cdot} + \alpha_i + \beta_j$$

$$= \mu_{\cdot\cdot} + \mu_{i\cdot} - \mu_{\cdot\cdot} + \mu_{\cdot j} - \mu_{\cdot\cdot} = \mu_{i\cdot} + \mu_{\cdot j} - \mu_{\cdot\cdot}$$

Then for any i, j, k, l

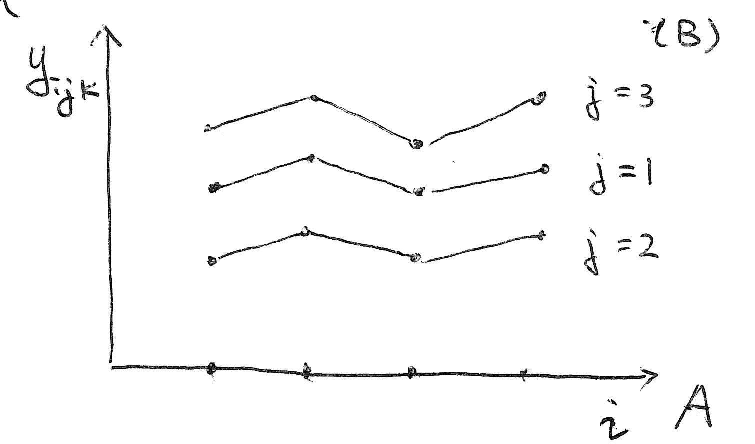
$$\mu_{ij} + \mu_{kl} = 2\mu_{\cdot\cdot} + \alpha_i + \beta_j + \alpha_k + \beta_l$$

$$\mu_{il} + \mu_{kj} = 2\mu_{\cdot\cdot} + \alpha_i + \beta_l + \alpha_k + \beta_j$$

$$\Rightarrow \mu_{ij} - \mu_{il} = \mu_{kj} - \mu_{kl} \qquad \text{no interaction}$$

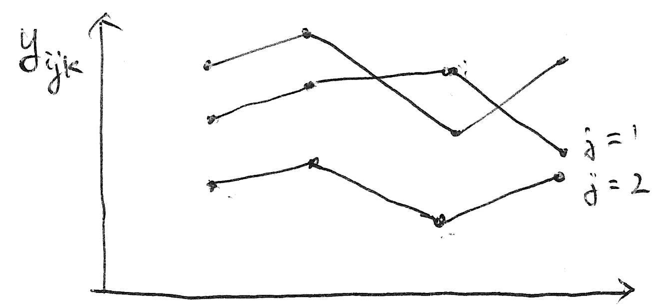
(Doesn't matter if (i) paired with (j) or (l))
effect of A doesn't depend on the level of B)

Illustration



← treatment means plot

Non-additive (interacting) factor effects



interaction effect is

$$(\alpha\beta)_{ij} := \mu_{ij} - (\mu_{..} + \alpha_i + \beta_j)$$
$$= \mu_{ij} - \alpha_i - \beta_j - \mu_{..}$$
$$= \mu_{ij} - \mu_{i.} - \mu_{.j} + \mu_{..}$$

$\exists$ interaction if

(1) $\mu_{ij} \neq \mu_{\cdot\cdot} + \alpha_i + \beta_j$ for some i, j

(2) $\mu_{ij} - \mu_{il} \neq \mu_{kj} - \mu_{kl}$ for some i, j, k, l

(3) treatment means curves are not parallel.

Some $(\alpha\beta)_{ij} = 0$ is possible

always $\sum_i (\alpha\beta)_{ij} = \sum_j (\alpha\beta)_{ij} = 0$

$$\sum_i (\alpha\beta)_{ij} = \sum_i (\mu_{ij} - \alpha_i - \beta_j - \mu_{\cdot\cdot})$$

$$= a\mu_{\cdot j} - 0 - a\beta_j - a\mu_{\cdot\cdot}$$

· Transformable interactions.

Say $\mu_{ij} = \mu_{\cdot\cdot}\alpha_i\beta_j$

$$\log \mu_{ij} = \log \mu_{\cdot\cdot} + \log \alpha_i + \log \beta_j$$

or $\mu_{ij} = \alpha_i + \beta_j + 2\sqrt{\alpha_i\beta_j}$

$$\sqrt{\mu_{ij}} = \sqrt{\alpha_i} + \sqrt{\beta_j}$$