Discussion 12

Review

Two-way ANOVA

Without interaction

Source	df	SS	MS	F	p-value
factor A	a-1	SS_A	$\frac{SS_A}{(a-1)}$	$F = \frac{MS_A}{MS_{error}}$	$p = P(F_{a-1,df_{error}} > F)$
factor B	b-1	SS_B	$\frac{SS_B}{(b-1)}$	$F = \frac{MS_B}{MS_{error}}$	$p = P(F_{b-1,df_{error}} > F)$
error	n-a-b+1	SS_{error}	$\frac{SS_{error}}{(n-k)}$		
Total	n-1	SS_{total}			

With interaction

Source	df	SS	MS	F	p-value
factor A	a-1	SS_A	$\frac{SS_A}{(a-1)}$	$F = \frac{MS_A}{MS_{error}}$	$p = P(F_{a-1,df_{error}} > F)$
factor B	b-1	SS_B	$\frac{SS_B}{(b-1)}$	$F = \frac{MS_B}{MS_{error}}$	$p = P(F_{b-1,df_{error}} > F)$
A×B	(a-1)(b-1)	SS_{AB}	$\frac{SS_{AB}}{(a-1)(b-1)}$	$F = \frac{MS_{AB}}{MS_{error}}$	$p = P(F_{(a-1)(b-1),df_{error}} > F)$
error	n-ab	SS_{error}	$\frac{SS_{error}}{(n-k)}$		
Total	n-1	SS_{total}			

Linear Regression

ANOVA is an example of a linear model.
If the first group is used as reference in a one-way ANOVA with k groups,

$$Y_i = \beta_0 + \beta_2 I_{\{group=2\}} + \beta_3 I_{\{group=3\}} + \cdots + \beta_k I_{\{group=k\}} + \epsilon_i,$$

where $\epsilon_i \sim$i.i.d.$N(0,1)$.
β_0 is the expected mean of the first group.
β_2 is the expected mean difference between the second and the first group.
...
β_k is the expected mean difference between the kth group and the first group.

Correlation Coefficient

The correlation coefficient r is measure of the strength of the linear relationship between two variables.

$$r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}}$$

Correlation is not affected by linear transformation of the data.

Practice Problems

1. The data are from a statement by Texaco, Inc. to the Air and Water Pollution Subcommittee of the Senate Public Works Committee on June 26, 1973. Mr. John McKinley, President of Texaco, cited an automobile filter developed by Associated Octel Company as effective in reducing pollution. However, questions had been raised about the effects of filters on vehicle performance, fuel consumption, exhaust gas back pressure, and silencing. On the last question, he referred to the data included here as evidence that the silencing properties of the Octel filter were at least equal to those of standard silencers.

   NOISE = Noise level reading (decibels)

   SIZE = Vehicle size: 1 small 2 medium 3 large

   TYPE = 1 standard silencer 2 Octel filter

   SIDE = 1 right side 2 left side of car

NOISE	SIZE	TYPE	SIDE
810	1	1	1
820	1	1	1
820	1	1	1
840	2	1	1
840	2	1	1
845	2	1	1
785	3	1	1
790	3	1	1
785	3	1	1
835	1	1	2
835	1	1	2
835	1	1	2
845	2	1	2
855	2	1	2
850	2	1	2
760	3	1	2
760	3	1	2
770	3	1	2

NOISE	SIZE	TYPE	SIDE
820	1	2	1
820	1	2	1
820	1	2	1
820	2	2	1
820	2	2	1
825	2	2	1
775	3	2	1
775	3	2	1
775	3	2	1
825	1	2	2
825	1	2	2
825	1	2	2
815	2	2	2
825	2	2	2
825	2	2	2
770	3	2	2
760	3	2	2
765	3	2	2

   (a) Plot the noise level versus the vehicle size, type of the filter and filter location, in order to get a visual estimate of which one(s) seem to be most affecting the noise level.

   (b) How many replicates are there in each experimental condition(combination of SIZE, TYPE and SIDE)? is it balanced?

   (c) Fit a linear model to predict noise level using both SIZE and TYPE as predictors, without interaction. Then write the model with the estimated coefficients in two different but equivalent ways:

$$\text{mean noise level } = \mu + \alpha_2 \mathbf{1}_{\text{SIZE}=2} + \alpha_3 \mathbf{1}_{\text{SIZE}=3} + \beta_2 \mathbf{1}_{\text{TYPE}=2}$$

$$\text{mean noise level } = \tilde{\mu} + \tilde{\alpha}_1 \mathbf{1}_{\text{SIZE}=1} + \tilde{\alpha}_3 \mathbf{1}_{\text{SIZE}=3} + \tilde{\beta}_1 \mathbf{1}_{\text{TYPE}=1}$$

   (d) What is the relationship between $beta_4$ and $\tilde{\beta}_4$, and why? What is the relationship between β_2 and $\tilde{\beta}_2$, and why?

   (e) Complete the following table with the mean noise level predicted by the model in 3.

	SIZE=1	SIZE=2	SIZE=3
TYPE=1			
TYPE=2			

Then compare the values in the table with the mean noise level observed in the data.

(f) Fit a linear model to predict noise level using both SIZE and TYPE, including interaction. Write down this model in the form:

$$\begin{aligned}\text{mean noise level } =&\mu + \alpha_1 \mathbf{1}_{\text{SIZE=1}} + \alpha_2 \mathbf{1}_{\text{SIZE=2}} + \beta_2 \mathbf{1}_{\text{TYPE=2}} \\ &+ \gamma_{1,2} \mathbf{1}_{\text{SIZE=1,TYPE=2}} + \gamma_{2,2} \mathbf{1}_{\text{SIZE=2,TYPE=2}}\end{aligned}$$

then write down the same model with a separate equation for each TYPE.

(g) Redo question 5, using the model in question 6.

2. Suppose $cor(X, Y) = 0.2$, calculate the following:

(a) $cor(3X, 5Y)$

(b) $cor(2X + 9, 6Y + 5)$

Solution

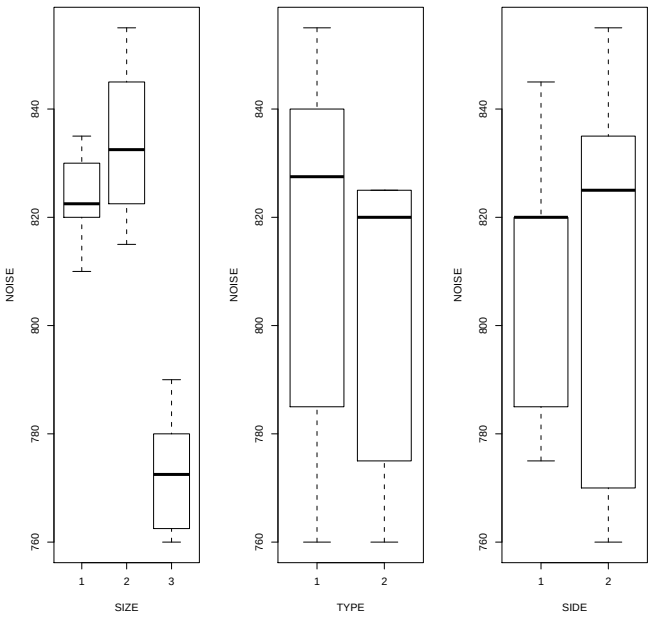
1. ```
   > method1=c(10,9,9,8)
   > method2=c(19,20,21,20)
   > method3=c(11,13,15,13)
   > alldata=c(method1,method2,method3)
   > trt=c(rep(1,4), rep(2,4), rep(3,4))
   > trt=factor(trt); trt
    [1] 1 1 1 1 2 2 2 2 3 3 3 3
   Levels: 1 2 3
   > library(lattice)
   > dotplot(alldata~trt)
   > fit1=lm(alldata~trt)
   > summary(fit1)

   Call:
   lm(formula = alldata ~ trt)

   Residuals:
          Min          1Q      Median          3Q         Max
   -2.000e+00 -2.500e-01 -1.333e-16  2.500e-01  2.000e+00

   Coefficients:
               Estimate Std. Error t value Pr(>|t|)
   (Intercept)   9.0000     0.5774  15.588 8.07e-08 ***
   trt2         11.0000     0.8165  13.472 2.86e-07 ***
   trt3          4.0000     0.8165   4.899 0.000849 ***
   ---
   Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

   Residual standard error: 1.155 on 9 degrees of freedom
   Multiple R-squared: 0.9538,     Adjusted R-squared: 0.9436
   F-statistic:    93 on 2 and 9 DF,  p-value: 9.748e-07
   ```

2. ```
   > noise = read.table("noise.txt",header=T)
   > str(noise)
   'data.frame':   36 obs. of  4 variables:
    $ NOISE: int  810 820 820 840 840 845 785 790 785 835 ...
    $ SIZE : int  1 1 1 2 2 2 3 3 3 1 ...
    $ TYPE : int  1 1 1 1 1 1 1 1 1 1 ...
    $ SIDE : int  1 1 1 1 1 1 1 1 1 2 ...
   > # transform those categorical variables into factors.
   > noise$SIZE = factor(noise$SIZE)
   > noise$TYPE = factor(noise$TYPE)
   > noise$SIDE = factor(noise$SIDE)
   > #1
   > plot(NOISE~SIZE,noise)
   > plot(NOISE~TYPE,noise)
   > plot(NOISE~SIDE,noise)
   > #2
   ```

```
> xtabs(~SIZE+TYPE+SIDE, noise)
, , SIDE = 1

    TYPE
SIZE 1 2
   1 3 3
   2 3 3
   3 3 3

, , SIDE = 2

    TYPE
SIZE 1 2
   1 3 3
   2 3 3
   3 3 3

> ftable(xtabs(~SIZE+TYPE+SIDE, noise))
          SIDE 1 2
SIZE TYPE
1    1         3 3
     2         3 3
2    1         3 3
     2         3 3
3    1         3 3
     2         3 3
> #3
> fit1 = lm(NOISE~SIZE+TYPE, noise)
> summary(fit1)

Call:
lm(formula = NOISE ~ SIZE + TYPE, data = noise)

Residuals:
    Min      1Q  Median      3Q     Max
-19.583  -7.292   1.250   6.250  15.833

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  829.583      3.099 267.657  < 2e-16 ***
SIZE2          9.583      3.796   2.525  0.01674 *
SIZE3        -51.667      3.796 -13.611  7.4e-15 ***
TYPE2        -10.833      3.099  -3.495  0.00141 **
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 9.298 on 32 degrees of freedom
Multiple R-squared: 0.9074,     Adjusted R-squared: 0.8987
F-statistic: 104.5 on 3 and 32 DF,  p-value: < 2.2e-16
```

```
> # change the reference levels to obtain another formula for the same model.
> noise_m = noise
> noise_m$SIZE = relevel(noise_m$SIZE,"2")
> noise_m$TYPE = relevel(noise_m$TYPE,"2")
> fit1_m = lm(NOISE~SIZE+TYPE, noise_m)
> summary(fit1_m)

Call:
lm(formula = NOISE ~ SIZE + TYPE, data = noise_m)

Residuals:
    Min      1Q  Median      3Q     Max
-19.583  -7.292   1.250   6.250  15.833

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  828.333      3.099 267.253  < 2e-16 ***
SIZE1         -9.583      3.796  -2.525  0.01674 *
SIZE3        -61.250      3.796 -16.135  < 2e-16 ***
TYPE1         10.833      3.099   3.495  0.00141 **
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 9.298 on 32 degrees of freedom
Multiple R-squared: 0.9074,     Adjusted R-squared: 0.8987
F-statistic: 104.5 on 3 and 32 DF,  p-value: < 2.2e-16
> #5
> newC = data.frame(
+    SIZE = rep( c("1","2","3"), 2),
+    TYPE = rep( c("1", "2"), each=3)
+ )
> newC
  SIZE TYPE
1    1    1
2    2    1
3    3    1
4    1    2
5    2    2
6    3    2
> predict(fit1, newC)
       1        2        3        4        5        6
829.5833 839.1667 777.9167 818.7500 828.3333 767.0833
> # verify that different formulation doesnt change the predicted values.
> predict(fit1_m, newC)
       1        2        3        4        5        6
829.5833 839.1667 777.9167 818.7500 828.3333 767.0833
> # compute the group means of the noise level.
> with(noise, tapply(NOISE, list(SIZE,TYPE), mean))
         1        2
1 825.8333 822.5000
```

```
2 845.8333 821.6667
3 775.0000 770.0000
> #6
> noise2 = noise
> noise2$SIZE = relevel(noise$SIZE,"3")
> fit2 = lm(NOISE~SIZE*TYPE, noise2)
> summary(fit2)

Call:
lm(formula = NOISE ~ SIZE * TYPE, data = noise2)

Residuals:
     Min       1Q   Median       3Q      Max
-15.8333  -5.2083  -0.4167   5.0000  15.0000

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  775.000      3.302 234.711  < 2e-16 ***
SIZE1         50.833      4.670  10.886 6.11e-12 ***
SIZE2         70.833      4.670  15.169 1.30e-15 ***
TYPE2         -5.000      4.670  -1.071  0.29282
SIZE1:TYPE2    1.667      6.604   0.252  0.80247
SIZE2:TYPE2  -19.167      6.604  -2.902  0.00688 **
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 8.088 on 30 degrees of freedom
Multiple R-squared: 0.9343,     Adjusted R-squared: 0.9234
F-statistic: 85.34 on 5 and 30 DF,  p-value: < 2.2e-16

> #7
> predict(fit2, newC)
       1        2        3        4        5        6
825.8333 845.8333 775.0000 822.5000 821.6667 770.0000
```


NOISE
760
780
800
820
840
1
2
3
SIZE
NOISE
760
780
800
820
840
1
2
TYPE
NOISE
760
780
800
820
840
1
2
SIDE