1 Use of indicator variables. (Chapter 8)

- let $I(A) = 1$ if the event A occurs, and $I(A) = 0$ otherwise.
- $I(A)$ is referred to as the indicator of the event A.
- The notation I_A is often used.

2 One-way Analysis of Variance

Recall Example: A group of 32 rats were randomly assigned to each of 4 diets labelled (A,B,C,and D). The response is the liver weight as a percentage of body weight. Two rats escaped and another died, resulting in the following data

A	B	C	D
3.42	3.17	3.34	3.65
3.96	3.63	3.72	3.93
3.87	3.38	3.81	3.77
4.19	3.47	3.66	4.18
3.58	3.39	3.55	4.21
3.76	3.41	3.51	3.88
3.84	3.55		3.96
	3.44		3.91

R code to enter data and calculate sample means by diet.

```
> weight=c(3.42 ,3.17 ,3.34 ,3.65 ,
+    3.96 ,3.63 ,3.72 ,3.93 ,
+    3.87 ,3.38 ,3.81 ,3.77 ,
+    4.19 ,3.47 ,3.66 ,4.18 ,
+    3.58 ,3.39 ,3.55 ,4.21 ,
+    3.76 ,3.41 ,3.51 ,3.88 ,
+    3.84 ,3.55       ,3.96 ,
+          3.44       ,3.91 )
> diet=as.factor(c(rep(c("A","B","C","D"),6),"A","B","D","B","D"))
> tapply(weight,diet,FUN=mean) # calculate mean weight by diet

       A        B        C        D
3.802857 3.430000 3.598333 3.936250
```

2.1 One way ANOVA using multiple regression and indicator variables

Indicator Variables

Where there are k groups, define $k-1$ indicator variables to identify the k groups. In this example, $k = 4$.
Index the subjects using a single index i, $i = 1, \ldots, n$
Let

$X_{i1} = 1$ if subject i is in group 1, and 0 otherwise. In the example, variable X_1 is the indictor of diet A.
$X_{i2} = 1$ if subject i is in group 2, and 0 otherwise.
$X_{i3} = 1$ if subject i is in group 3, and 0 otherwise.
...
$X_{i,k-1} = 1$ if subject i is in group $k-1$, and 0 otherwise.

Multiple regression Model :

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_{a-1} X_{i,k-1} + \epsilon_i$$

Table of Means

Group	ANOVA mean	Regression mean
1	μ_1	$\beta_0 + \beta_1$
2	μ_2	$\beta_0 + \beta_2$
	...	
$k-1$	μ_{k-1}	$\beta_0 + \beta_{k-1}$
k	μ_k	β_0

```
> IA=ifelse(diet=="A",1,0)
> IB=ifelse(diet=="B",1,0)
> IC=ifelse(diet=="C",1,0)
> ID=ifelse(diet=="D",1,0)
> lmfull=lm(weight~IA+IB+IC)
> summary(lmfull)

Call:
lm(formula = weight ~ IA + IB + IC)

Residuals:
     Min       1Q   Median       3Q      Max
-0.38286 -0.05625 -0.00625  0.12000  0.38714

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.93625    0.06691  58.828  < 2e-16 ***
IA          -0.13339    0.09795  -1.362  0.18538
IB          -0.50625    0.09463  -5.350 1.51e-05 ***
IC          -0.33792    0.10221  -3.306  0.00286 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1893 on 25 degrees of freedom
Multiple R-squared:  0.5654,        Adjusted R-squared:  0.5132
F-statistic: 10.84 on 3 and 25 DF,  p-value: 9.502e-05
```

The F statistic and p-value are identical to those using the partial F test to compare the full model to the reduced model $y = \beta_0 + \epsilon$.

```
> lmred=lm(weight~1)
> anova(lmfull,lmred)

Analysis of Variance Table

Model 1: weight ~ IA + IB + IC
Model 2: weight ~ 1
  Res.Df     RSS Df Sum of Sq      F    Pr(>F)
1     25 0.89541
2     28 2.06032 -3   -1.1649 10.841 9.502e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

3 Two way analysis of variance

Example: Tests were conducted to assess the effects of two factors, enginge type, and propellant type, on propellant burn rate in fired missiles. Three engine types and four propellant types were tested.
Twenty-four missiles were selected from a large production batch. The missiles were randomly split into three groups of size eight. The first group of eight had engine type 1 installed, the second group had engine type 2, and the third group received engine type 3.‘
Each group of eight was randomly divided into four groups of two. The first such group was assigned propellant type 1, the second group was assigned propellant type 2, and so on.
Data on burn rate were collected, as follows:

Engine	Propellant Type			
type	1	2	3	4
1	34.0	30.1	29.8	29.0
	32.7	32.8	26.7	28.9
2	32.0	30.2	28.7	27.6
	33.2	29.8	28.1	27.8
3	28.4	27.3	29.7	28.8
	29.3	28.9	27.3	29.1

The twoway ANOVA model including interaction is

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \epsilon_{ijk}$$

$i = 1, 2, \ldots, I, \qquad j = 1, 2, \ldots, J, \qquad k = 1, 2, \ldots, K.$
In the example, $I = 3$, $J = 4$, and $K = 2$.

- μ is the overall mean
- $\sum_{i=1}^{I} \alpha_i = 0$
- $\sum_{j=1}^{J} \beta_j = 0$
- $\sum_{i=1}^{I} \gamma_{ij} = 0$ for each $j = 1, 2, \ldots, J$
- $\sum_{j=1}^{J} \gamma_{ij} = 0$ for each $i = 1, 2, \ldots, I$
- we assume ϵ_{ijk} are iid $N(0, \sigma^2)$
- The sum constraints ensure that there are the same number of parameters as there are cell means, IJ, which in the example is 3(4)=12.
- The hypothesis of no interaction between propellant type and engine type is

$$H_0 : \gamma_{ij} = 0 \text{ for all } i, j$$

- The hypothesis of no main effect of propellant is

$$H_0 : \alpha_i = 0 \text{ for all } i$$

- The hypothesis of no main effect of engine type is

$$H_0 : \beta_j = 0 \text{ for all } j$$

```
> propellant=as.factor(rep(c(1:4),6))
> engine=as.factor(c(rep(1,8),rep(2,8),rep(3,8)))
> rate=c(34.0 , 30.1 , 29.8 , 29.0, 32.7 , 32.8 , 26.7 , 28.9
+       ,32.0 , 30.2 , 28.7 , 27.6, 33.2 , 29.8 , 28.1 , 27.8
+       ,28.4 , 27.3 , 29.7 , 28.8, 29.3 , 28.9 , 27.3 , 29.1)
```

The following uses the builtin lm command to carry out a twoway analysis of variance including main effects of propellant and engine, and their interaction.

```
> anova(lm(rate~propellant*engine))

Analysis of Variance Table

Response: rate
                  Df Sum Sq Mean Sq F value  Pr(>F)
propellant         3 40.082 13.3606 10.7530 0.00102 **
engine             2 14.523  7.2617  5.8444 0.01690 *
propellant:engine  6 22.163  3.6939  2.9729 0.05117 .
Residuals         12 14.910  1.2425
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

- If testing at level of significance $\alpha = .05$, the conclusion is no significant interaction (p=.05117), but siginficant main effects of engine type (p=.0169) and propellant (p=.00102).

3.1 Multiple regression approach to two way ANOVA

- Let $x_1, x_2, \ldots, x_{I-1}$ be $I-1$ indicator variables coding for a row factor having I levels, where x_l is an indicator for the l'th level of the row factor.
- Let $z_1, z_2, \ldots, z_{J-1}$ be $J-1$ indicator variables coding for a column factor having J levels, where x_j is an indicator for the j'th level of the column factor.

The multiple regression model is

$$\begin{aligned} y &= \beta_0 \\ &+ \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_{I-1} X_{I-1} \\ &+ \beta_I Z_1 + \beta_{I+1} Z_2 + \ldots + \beta_{(I-1)+(J-1)} Z_{J-1} \\ &+ \beta_{I+J-1} X_1 Z_1 + \beta_{I+J} X_1 Z_2 + \ldots + \beta_{IJ-1} X_{I-1} Z_{J-1} + \epsilon \end{aligned}$$

The number of β parameters is 1 (for β_0), plus $I-1$ for the indicator variables associated with the row factor, plus $J-1$ for the indicator variables associated with the column factor, $(I-1)(J-1)$ for the interaction terms, so $IJ = (I-1)(J-1) + (I-1) + (J-1) + 1$, which is the number of means in the two way layout.

In the example.

The following carries out a multiple regression to fit the twoway anova model with two qualitative factors, the first having 3 levels, and second having 4 levels, plus replicate observations. Replicates are necessary in order to accommodate interactions.

- First define two indicator variables which identify the engine type, and three indicator variables to identify the propellant type.
- Then fit a regression model which includes these indicators, and the products of the engine type indicators with the propellant type indicators. The product terms code for interactions of engine type with propellant type.
- The specific multiple regression model is

$$\begin{aligned} y &= \beta_0 \\ &+ \beta_1 X_1 + \beta_2 X_2 \\ &+ \beta_3 Z_1 + \beta_4 Z_2 + \beta_5 Z_3 \\ &+ \beta_6 X_1 Z_1 + \beta_7 X_1 Z_2 + \beta_8 X_1 Z_3 \\ &+ \beta_9 X_2 Z_1 + \beta_{10} X_2 Z_2 + \beta_{11} X_2 Z_3 + \epsilon \end{aligned}$$

- The null hypothesis for the test of no interaction between engine type and propellant type is

$$H_0 : \beta_6 = \beta_7 = \ldots \ \beta_{11} = 0$$

- To carry out the test
  - first fit the full model
    lm(rate∼X1+X2+X3+Z1+Z2+X1*Z1+X2*Z1+X3*Z1+X1*Z2+X2*Z2+X3*Z2),
  - and then fit the reduced model under the null hypothesis
    lm(rate∼X1+X2+X3+Z1+Z2)
  - and calculate the partial F statistic using the error sums of squares for the two models. As usual, the test statistic is $F = \frac{(SSE_{red} - SSE_{full})/r}{MSE_{full}}$, and has an F distribution whose numerator degrees of freedom is the number of parameters set to 0 under the null hypothesis, and denominator degrees of freedom is the error degrees of freedom for the full model.

```
> X1=ifelse(propellant==1,1,0)
> X2=ifelse(propellant==2,1,0)
> X3=ifelse(propellant==3,1,0)
> Z1=ifelse(engine==1,1,0)
> Z2=ifelse(engine==2,1,0)
> #following fits the full model including
> #two indicator variables for engine type
> #three indicator variables for propellant type
> #2(3)=6 products of indicators for interaction of engine and propellant
> lm.full=lm(rate~X1+X2+X3+Z1+Z2+X1*Z1+X2*Z1+X3*Z1+X1*Z2+X2*Z2+X3*Z2)
> #following fits a reduced model without the 6 interaction terms
> lm.noint=lm(rate~X1+X2+X3+Z1+Z2)
> #following carries out the test of no interaction
> anova(lm.full,lm.noint)

Analysis of Variance Table

Model 1: rate ~ X1 + X2 + X3 + Z1 + Z2 + X1 * Z1 + X2 * Z1 + X3 * Z1 +
    X1 * Z2 + X2 * Z2 + X3 * Z2
Model 2: rate ~ X1 + X2 + X3 + Z1 + Z2
  Res.Df    RSS Df Sum of Sq      F  Pr(>F)
1     12 14.910
2     18 37.073 -6   -22.163 2.9729 0.05117 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

- Recall from Stat2080 that if the interaction is significant (ie reject H_0 : no interaction) then it makes no sense to test for the main effects, as we know that there are significant mean diffences between levels of each factor but that the mean differences between levels of one factor will depend on the level of the other factor.
- If the interaction is not significant then proceed to test for the main effects of propellant and engine type. At the 5% level of signficiance conclude that there are no interactions, so proceed to test for main effects of engine type and propellant.
  - to test for effect of engine type, fit a third model without engine type (and of course, without interaction terms), and compare the SSE for this model to that with main effects of engine type and propellant.
  - to test for the effect of propellant, fit a model without propellant, and compare to model with both engine type and propellant.
  - because the design is balanced (same number of observaions at each level of engine type and propellant) the order in which the main effects are entered into the regression model is irrelevant, which is a major advantage of a balanced design. A consequence is that in this case, testing for the effect of propellant can also be accomplished by comparing a model only including propellant type (lm(rate∼X1+X2+X3)), to a model only including the overall mean β_0 (lm(rate∼1)).

```
> lm.prop=lm(rate~X1+X2+X3)
> lm.0=lm(rate~1)
> anova(lm.full,lm.noint,lm.prop,lm.0)

Analysis of Variance Table

Model 1: rate ~ X1 + X2 + X3 + Z1 + Z2 + X1 * Z1 + X2 * Z1 + X3 * Z1 +
    X1 * Z2 + X2 * Z2 + X3 * Z2
Model 2: rate ~ X1 + X2 + X3 + Z1 + Z2
Model 3: rate ~ X1 + X2 + X3
Model 4: rate ~ 1
  Res.Df    RSS Df Sum of Sq       F  Pr(>F)
1     12 14.910
2     18 37.073 -6   -22.163  2.9729 0.05117 .
3     20 51.597 -2   -14.523  5.8444 0.01690 *
4     23 91.678 -3   -40.082 10.7530 0.00102 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

- note that the three p-values from the comparison of the multiple regression models are identical to those using the builtin anova(lm()) procedure with qualitative predictor variables, also known as factor level predictor variables.