Two-way ANOVA (with replications) and MLR

- We look at the case where the two factors *jointly* affects the response variable
- E.g. testing effectiveness of detergents at removing marks on different types if fabric, but some detergents may work better on certain types of fabric
- To use the interaction model, we must have more than one measurements for each combination of the interactions

Interaction model

response variable Y_{ij} has k replications

		column factor		
	1	2		J
1	$y_{111}, y_{112}, \ldots, y_{11k}$	$y_{121}, y_{122}, \ldots, y_{12k}$	$\cdots$	$y_{1J1}, y_{1J2}, \ldots, y_{1Jk}$
row factor 2	$y_{211}, y_{212}, \ldots, y_{21k}$	$y_{221}, y_{222}, \ldots, y_{22k}$	$\cdots$	$y_{2J1}, y_{2J2}, \ldots, y_{2Jk}$
$\vdots$	$\vdots$	$\vdots$		$\vdots$
I	$y_{I11}, y_{I12}, \ldots, y_{I1k}$	$y_{I21}, y_{I22}, \ldots, y_{I2k}$	$\cdots$	$y_{IJ1}, y_{IJ2}, \ldots, y_{IJk}$

total number of observations is IJK.

$$Y_{ijk} \sim N(\mu_{ij}, \sigma^2)$$

$$\mu_{ij} = \mu + \underbrace{\alpha_i}_{\text{main effect of the row factor}} + \underbrace{\beta_j}_{\text{main effect of the column factor}} + \underbrace{\gamma_{ij}}_{\text{interaction term}}$$

Two-way layout with interaction (replication)

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

with

$$\epsilon_{ijk} \sim i.i.d.\, N(0, \sigma^2)$$

for $i = 1, 2, \ldots, I; j = 1, 2, \ldots, J, k = 1,2, \ldots, K$.

Parameters are $\mu, \alpha_i, \beta_j, \gamma_{ij}$

There are $1 + I + J + IJ$ parameters

But μ_{ij} just need IJ parameters

We need $1 + I + J$ constraints.

$$\cancel{\mu} = \frac{1}{IJ} \sum_{i=1}^{I} \sum_{j=1}^{J} \mu_{ij} = \frac{1}{IJ} \sum_{i=1}^{I} \sum_{j=1}^{J} (\mu + \alpha_i + \beta_j + \gamma_{ij})$$

$$= \cancel{\mu} + \frac{1}{I} \sum_{i=1}^{I} \alpha_i + \frac{1}{J} \sum_{j=1}^{J} \beta_j + \frac{1}{IJ} \sum_{i=1}^{I} \sum_{j=1}^{J} \gamma_{ij}$$

$$0 = \frac{1}{I} \sum_{i=1}^{I} \alpha_i + \frac{1}{J} \sum_{j=1}^{J} \beta_j + \frac{1}{IJ} \sum_{i=1}^{I} \sum_{j=1}^{J} \gamma_{ij}$$

We can use the zero-sum constraints

or the reference category constraints

in this case as well.

Constraints

- Zero Sum Constraints:

$$\sum_{i=1}^{I} \alpha_i = \sum_{j=1}^{J} \beta_j = \sum_{i=1}^{I} \gamma_{ij} = \sum_{j=1}^{J} \gamma_{ij} = 0.$$

1 + 1 + J + I-1 = 1 + I + J constraints

- Reference Category Constraints:

1 + 1 + J + I-1 = 1 + I + J constraints

$$\alpha_1 = \beta_1 = \boxed{\gamma_{1j} = \gamma_{i1} = 0,}$$

for $i = 1,2, \ldots, I, j = 1, 2, \ldots, J.$

ANOVA table

Source	SS	df	MSE	F	
Tx A	$JK\sum_i(\bar{y}_{i\bullet\bullet} - \bar{y}_{\bullet\bullet\bullet})^2$	I-1	MSA	MSA/MSE	H_1
Tx B	$IK\sum_j(\bar{y}_{\bullet j\bullet} - \bar{y}_{\bullet\bullet\bullet})^2$	J-1	MSB	MSB/MSE	H_2
Interaction	$K\sum_{ij}(y_{ij\bullet} - \bar{y}_{i\bullet\bullet} - \bar{y}_{\bullet j\bullet} + \bar{y}_{\bullet\bullet\bullet})^2$	(I-1)(J-1)	MSI	MSI / MSE	H_0
Residual	$\sum_{ijk}(y_{ijk} - \bar{y}_{ij\bullet})^2$	IJ(K-1)	MSE		
Total	$\sum_{ij}(y_{ij} - \bar{y}_{\bullet\bullet})^2$	IJ-1	MST		

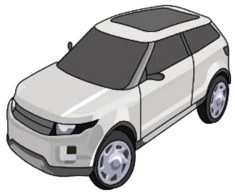
H_1: no difference in row means : $\alpha_1 = \alpha_2 = \dots = \alpha_I = 0$

H_2: no difference in column means : $\beta_1 = \beta_2 = \dots = \beta_J = 0$

H_0: no interaction between column and row factors: $\gamma_{ij} = 0 \ \forall i,j$

(similar to saying the row and column factors are independent)

Example 4.10



Consider the mpg data used in the Practicals. We wish to investigate the relationship between the response variable city miles per gallon (cty) and the predictors Drive (drv) and Transmission (trans).

The first two observations for each Drive are given in the table on the right:

cty	drv	trans
18	4	manual
16	4	auto
18	f	auto
21	f	manual
14	r	auto
11	r	auto

a) Write down the design matrix for the above observations if we were to perform a two-way ANOVA without interactions.
b) Write down the design matrix for the above observations if we were to perform a two-way ANOVA with interactions.

Example 4.10

c) The additive and interaction models were fitted using R. From the given output, which model is better?

d) Use the chosen model in part (c) to predict the mean City Miles per Gallon for an automatic with front-wheel drive.

e) If you change your car from automatic to manual, but keeping the same Drive, what is the expected change in City Miles per Gallon?

Example 4.10 Solution

a) Our model is $Y_{ij} = \mu + \alpha_i + \beta_j + \epsilon_{ij}$ where $\epsilon_{ij} \sim N(0, \sigma^2)$.
Index $i = 1,2,3$ corresponds to 4, f, and r respectively (for the predictor drv).
Index $j = 1, 2$ corresponds to auto and manual, respectively.

Using the reference category constraint, with drv=4 and trans=auto as the reference categories, we have

$$X = \begin{bmatrix} 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 1 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \end{bmatrix} \text{ and } \beta = \begin{bmatrix} \mu \\ \alpha_2 \\ \alpha_3 \\ \beta_2 \end{bmatrix}$$

b) The model is $Y_{ij} = \mu + \alpha_i + \beta_j + \boxed{\gamma_{ij}} + \epsilon_{ij}$ where $\epsilon_{ij} \sim N(0, \sigma^2)$.

$$X = \begin{bmatrix} 1 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 1 & 1 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 \end{bmatrix} \text{ and } \beta = \begin{bmatrix} \mu \\ \alpha_2 \\ \alpha_3 \\ \beta_2 \\ \gamma_{22} \\ \gamma_{32} \end{bmatrix}$$

Example 4.10 Solution

```
library(tidyverse)
data(mpg)
mpg
```

```
# A tibble: 234 x 11
   manufacturer model       displ  year   cyl trans      drv     cty   hwy fl    class
   <chr>        <chr>       <dbl> <int> <int> <chr>      <chr> <int> <int> <chr> <chr>
 1 audi         a4            1.8  1999     4 auto(l5)   f        18    29 p     compact
 2 audi         a4            1.8  1999     4 manual(m5) f        21    29 p     compact
 3 audi         a4            2    2008     4 manual(m6) f        20    31 p     compact
 4 audi         a4            2    2008     4 auto(av)   f        21    30 p     compact
 5 audi         a4            2.8  1999     6 auto(l5)   f        16    26 p     compact
 6 audi         a4            2.8  1999     6 manual(m5) f        18    26 p     compact
 7 audi         a4            3.1  2008     6 auto(av)   f        18    27 p     compact
 8 audi         a4 quattro    1.8  1999     4 manual(m5) 4        18    26 p     compact
 9 audi         a4 quattro    1.8  1999     4 auto(l5)   4        16    25 p     compact
10 audi         a4 quattro    2    2008     4 manual(m6) 4        20    28 p     compact
# ... with 224 more rows
```

Notice the variable `trans` needs recoding so that there are only two levels: "auto" and "manual".

Example 4.10 Solution

```
mpg <- mpg %>%
  mutate(trans = ifelse(
     str_detect(trans, “auto”), “auto”, “manual”
))
mpg
```

```
# A tibble: 234 x 11
   manufacturer model      displ  year   cyl trans  drv     cty   hwy fl    class
   <chr>        <chr>      <dbl> <int> <int> <chr>  <chr> <int> <int> <chr> <chr>
 1 audi         a4           1.8  1999     4 auto   f        18    29 p     compact
 2 audi         a4           1.8  1999     4 manual f        21    29 p     compact
 3 audi         a4           2    2008     4 manual f        20    31 p     compact
 4 audi         a4           2    2008     4 auto   f        21    30 p     compact
 5 audi         a4           2.8  1999     6 auto   f        16    26 p     compact
 6 audi         a4           2.8  1999     6 manual f        18    26 p     compact
 7 audi         a4           3.1  2008     6 auto   f        18    27 p     compact
 8 audi         a4 quattro   1.8  1999     4 manual 4        18    26 p     compact
 9 audi         a4 quattro   1.8  1999     4 auto   4        16    25 p     compact
10 audi         a4 quattro   2    2008     4 manual 4        20    28 p     compact
# ... with 224 more rows
```

Example 4.10 Solution

a)

```
library(modelr)
mpg %>%
  select(cty, drv, trans) %>%
  group_by(drv) %>%
  slice(1:2) %>%
  model_matrix(cty ~ drv + trans)
```

```
## # A tibble: 6 x 4
##   `(Intercept)`  drvf  drvr transmanual
##           <dbl> <dbl> <dbl>       <dbl>
## 1             1     0     0           1
## 2             1     0     0           0
## 3             1     1     0           0
## 4             1     1     0           1
## 5             1     0     1           0
## 6             1     0     1           0
```

X =

The command `model_matrix` gives us the design matrix for your regression.

Example 4.10 Solution

b)

```
mpg %>%
  select(cty, drv, trans) %>%
  group_by(drv) %>%
  slice(1:2) %>%
  model_matrix(cty ~ drv * trans)
```

```
## # A tibble: 6 x 6
##   `(Intercept)`  drvf  drvr transmanual `drvf:transmanua~ `drvr:transmanu~
##           <dbl> <dbl> <dbl>       <dbl>             <dbl>            <dbl>
## 1             1     0     0           1                 0                0
## 2             1     0     0           0                 0                0
## 3             1     1     0           0                 0                0
## 4             1     1     0           1                 1                0
## 5             1     0     1           0                 0                0
## 6             1     0     1           0                 0                0
```

There are now two additional columns (compared to part a) due to the interaction terms.

Example 4.10 Solution

c)

```r
two_way_interaction <- lm(cty ~ drv * trans, data = mpg)
two_way_main <- lm(cty ~ drv + trans, data = mpg)
anova(two_way_interaction)
```

```
## Analysis of Variance Table
##
## Response: cty
##            Df  Sum Sq Mean Sq  F value    Pr(>F)
## drv         2 1878.81  939.41 101.0039 < 2.2e-16 ***
## trans       1  217.45  217.45  23.3799 2.445e-06 ***
## drv:trans   2    3.52    1.76   0.1893    0.8276
## Residuals 228 2120.56    9.30
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Example 4.10 Solution

```
anova(two_way_main)
```

```
## Analysis of Variance Table
##
## Response: cty
##            Df  Sum Sq Mean Sq F value    Pr(>F)
## drv         2 1878.81  939.41 101.721 < 2.2e-16 ***
## trans       1  217.45  217.45  23.546 2.25e-06 ***
## Residuals 230 2124.08    9.24
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

both variables are significant.

final model is additive model

```
anova(two_way_interaction, two_way_main)
```

```
## Analysis of Variance Table
##
## Model 1: cty ~ drv * trans
## Model 2: cty ~ drv + trans
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1    228 2120.6
## 2    230 2124.1 -2   -3.5217 0.1893 0.8276
```

Example 4.10 Solution

d)
```r
summary(two_way_main)
```

```
##
## Call:
## lm(formula = cty ~ drv + trans, data = mpg)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -8.1728 -2.0881 -0.3287  1.5810 13.7617
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  13.7686     0.3210  42.891  < 2e-16 ***
## drvf          5.4042     0.4233  12.767  < 2e-16 ***
## drvr         -0.3496     0.6779  -0.516    0.607
## transmanual   2.0655     0.4257   4.852 2.25e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.039 on 230 degrees of freedom
## Multiple R-squared:  0.4967, Adjusted R-squared:  0.4901
## F-statistic: 75.66 on 3 and 230 DF,  p-value: < 2.2e-16
```

e)

front, auto

cty = 13.7686 + 5.4042

= 19.1728

mean increase of 2.0655 miles per gallon if we switch from an auto to a manual

Example 4.10 Solution

```r
summary(two_way_interaction)
```

```
##
## Call:
## lm(formula = cty ~ drv * trans, data = mpg)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.1077 -2.0441 -0.3415  1.6940 13.6585
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)        13.8533     0.3521  39.339   <2e-16 ***
## drvf                5.2544     0.5168  10.167   <2e-16 ***
## drvr               -0.5592     0.8192  -0.683    0.496
## transmanual         1.7538     0.6754   2.597    0.010 *
## drvf:transmanual    0.4800     0.9089   0.528    0.598
## drvr:transmanual    0.7021     1.4717   0.477    0.634
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.05 on 228 degrees of freedom
## Multiple R-squared:  0.4975, Adjusted R-squared:  0.4865
## F-statistic: 45.15 on 5 and 228 DF,  p-value: < 2.2e-16
```

Example 4.10 Solution

c) One approach is to look at the output of `anova(two_way_interaction, two_way_main)`. The ANOVA table gives a P-value for 0.8276 when testing Model 2 against Model 1, suggesting there is not enough evidence to reject Model 2. So we would choose Model 2, which is the additive model.

Alternatively, we can also look at the ANOVA table for the main effects. It shows that both main effects are significant and so we keep both terms in the model.
The final model is therefore

$$cty \sim drv + trans$$

d) Prediction for an automatic with front-wheel drive is $13.7686 + 5.4042 = 19.1728$.

e) The coefficient for `transmanual` is 2.0655. Hence if we keep drive the same, but change from an automatic to a manual, the mean increase in fuel efficiency is 2.0655 miles per gallon.