STAT 40001/STAT 50001 Statistical Computing

Lecture 20

Department of Mathematics and Statistics
Purdue University Northwest

Two-way ANOVA

In a two way analysis of variance analyzing the effects of two factors A and B it is possible and is often of interest to study the interaction effect. Interaction means examining the simultaneous effect of factor A and B on the variable Y. A two-way ANOVA setup is not the same as a two-way contingency table.

- In two-way ANOVA, we classify subjects according to their values for two categorical variables. Then, for each of these subjects, we record the value of some quantitative response variable.
- In a two-way contingency table, we also classify subjects according to their values for two categorical variables. However, that is all we do — we simply count the number of subjects in each cell (group) and put that number in the table.

The two-way ANOVA model is written as

$$y_{ijk} = \mu + \tau_i + \beta_j + \gamma_{ij} + \epsilon_{ijk} \begin{cases} i = 1, 2, \cdots, a \\ j = 1, 2, \cdots, b \end{cases}$$

where
μ is the overall (grand) mean,
τ_i is the i^{th} facto A effect
β_j is the j^{th} factor B effect
γ_{ij} is the interaction effect
ϵ_{ijk} is the error and we assume that $\epsilon_{ij} \sim N(0, \sigma^2)$.

Example

An automobile web site wishes to test the mile-per-gallon rating of a car. It has three drivers and two cars of the same type. Each driver is asked to drive each car three times and record the miles per gallon. Below is the data

		Driver				Driver	
Car	a	b	c	Car	a	b	c
A	33.3	34.5	37.4	B	32.6	33.4	36.6
	33.4	34.8	36.8		32.5	33.7	37.0
	32.9	33.8	37.6		33.0	33.9	36.7

Ideally, there should be little variation. But is this the case with this data? We can use R code bellow to perform the analysis.

Example- Two way ANOVA

```
> y<-c(33.3,33.4,32.9,32.6,32.5,33.0,34.5,34.8,33.8,33.4,
   33.7,33.9,37.4,36.9,37.6,36.6,37.0,36.7)
> car=factor(rep(rep(1:2,c(3,3)),3))
> levels(car)=c("A","B")
> driver<-factor(rep(1:3,c(6,6,6)))
> levels(driver)=c("a","b","c")
> anova(lm(y~car+driver))
> anova(lm(y~car+driver), lm(y~driver))
Analysis of Variance Table

Model 1: y ~ car + driver
Model 2: y ~ driver
  Res.Df    RSS Df Sum of Sq  F   Pr(>F)
1     14 1.3144
2     15 2.8167 -1   -1.5022 16 0.001316 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Decision: The p-value indicates that there is a significance difference in the models

Example

The extra factor in two way ANOVA (interaction term) can be analyzed using R code below.

```
> y<-c(33.3,33.4,32.9,32.6,32.5,33.0,34.5,34.8,33.8,33.4,
   33.7,33.9,37.4,36.9,37.6,36.6,37.0,36.7)
> car=factor(rep(rep(1:2,c(3,3)),3))
> levels(car)=c("A","B")
> driver<-factor(rep(1:3,c(6,6,6)))
> levels(driver)=c("a","b","c")
> interaction.plot(driver, car, y) # Interaction plot
> summary(lm(y~car*driver))
> anova(lm(y~car+driver), lm(y~car*driver))
Analysis of Variance Table

Model 1: y ~ car + driver
Model 2: y ~ car * driver
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1     14 1.3144
2     12 1.2800  2  0.034444 0.1615 0.8527
```

Example-Rats

An experiment was conducted as part of an investigation to combat the effects of certain toxic agents. The data set `rats` is available in the `faraway` package. The data frame includes 48 observations on the following 3 variables. `time` - survival time in tens of hours, `poison` the poison type - a factor with levels I II III, `treat` the treatment - a factor with levels A B C D.

```
> library(faraway)
> data(rats)
> attach(rats)
> par(mfrow=c(1,2))
> plot(time~treat+poison, data=rats)
> interaction.plot(treat, poison, time)
> model=lm(time~poison*treat, data=rats)
> anova(model)
Analysis of Variance Table
Response: time
             Df  Sum Sq Mean Sq F value    Pr(>F)
poison        2 1.03301 0.51651 23.2217 3.331e-07 ***
treat         3 0.92121 0.30707 13.8056 3.777e-06 ***
poison:treat  6 0.25014 0.04169  1.8743    0.1123
Residuals    36 0.80073 0.02224
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Note that the interaction effect is not significant but the main effects are. We can perform the diagnostic test using the `checking.plot` function available in `PASWR` package or simply using: qqnorm(model$res)

```
> library(PASWR)
> model=lm(time~poison*treat, data=rats)
> checking.plots(model)
```

Example

The proprietors of a movie house want to maximize their customers's movie going experience. They are interested to know whether either eating popcorn or sitting in more comfortable seats makes a difference in customer enjoyment. They randomly assign 16 people equally to the four possible combinations and ask them to rate the same movie on a 0-100 scale.

seat type		good				bad			
popcorn	yes	92	80	80	78	60	59	57	51
	no	63	65	65	69	60	58	52	65

```
> y<-c(92,80,80,78,63,65,65,69,60,59,57,51,60,58,52,65)
> Seat<-factor(rep(c("Good","Bad"), c(8,8)))
> Popcorn<-factor(rep(rep(c("Y","N"), c(4,4)),2))
> interaction.plot(Seat,Popcorn,y)
> anova(lm(y~Seat*Popcorn),lm(y~Seat+Popcorn))
Analysis of Variance Table
Model 1: y ~ Seat * Popcorn
Model 2: y ~ Seat + Popcorn
  Res.Df   RSS Df Sum of Sq      F   Pr(>F)
1     12 277.5
2     13 638.5 -1      -361 15.611 0.001924 **
Signif. codes:  Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.
```

Example

A study was performed to test the efficiency of a new drug developed to increase high-density lipoprotein (HDL) cholesterol levels in patients. 18 volunteers were split into 3 groups (Placebo/5 mg/10 mg) and the difference in HDL levels before and after the treatment was measured. The 18 volunteers were also categorized into two age groups young(18-39) and old ($\geq$ 40)

Drug	young (18-39)	old(40+)
Placebo	4,3,-1	3,2,0
5mg	9,5,6	3,6,7
10mg	14,12,10	10,8,7

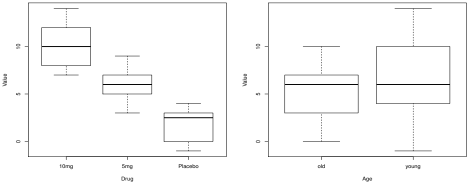
We are interested in determining the answers to the following questions:

- Does the amount of drug have an effect on the HDL level?
- Does the age of the patient have an effect on the HDL level?
- Is there an interaction between age and amount of drug?

Example-HDL data

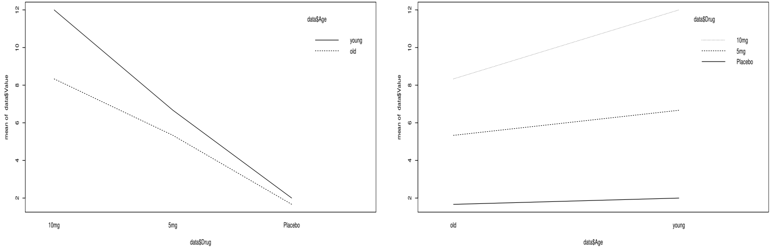
We will read the data and plot the side by side box plots using R code below

```
> data<-read.table("C://Aryal//STAT 40001//HDL.txt", header=T)
> data
      Drug   Age Value
1  Placebo young     4
2  Placebo young     3
3  Placebo young    -1
4  Placebo   old     3
5  Placebo   old     2
6  Placebo   old     0
7      5mg young     9
8      5mg young     5
9      5mg young     6
10     5mg   old     3
11     5mg   old     6
12     5mg   old     7
13    10mg young    14
14    10mg young    12
15    10mg young    10
16    10mg   old    10
17    10mg   old     8
18    10mg   old     7
> par(mfrow=c(1,2))
> plot(Value ~ Drug + Age, data=data)
```



Example-HDL data

Observing the boxplots there appears to be a difference in HDL level for the different drug levels. However, the difference is less pronounced between the two age groups. An interaction plot displays the levels of one factor on the x-axis and the mean response for each treatment on the y-axis.



The above graph are obtained by using the code below. Please pay attention to the order of the factors.

```
> interaction.plot(data$Drug, data$Age, data$Value)
> interaction.plot(data$Age, data$Drug, data$Value)
```

When no interaction is present the lines should be roughly parallel. These plots helps to determine whether an interaction term should be included in ANOVA

Example-HDL data

We can use the R code below to create the ANOVA table:

```
> model=lm(Value˜Drug + Age + Drug*Age, data=data)
> anova(model)
Analysis of Variance Table

Response: Value
          Df  Sum Sq Mean Sq F value    Pr(>F)
Drug       2 208.333 104.167 25.6849 4.611e-05 ***
Age        1  14.222  14.222  3.5068   0.08568 .
Drug:Age   2   8.778   4.389  1.0822   0.36974
Residuals 12  48.667   4.056
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Observe that there is no evidence of a significant interaction effect (F=1.08, p=0.370). We therefore cannot conclude that there is an interaction between age and the amount of drug taken. The test for the main effect of treatment (F=25.68, $p < 0.0001$) shows a significant drug effect on the HDL level. Finally, the test for the main effect of age (F=3.51, p=0.0857) tells us there is not enough evidence to conclude that there is a significant age effect.

MANOVA

MANOVA stands for Multivariate ANOVA or Multivariate Analysis Of Variance. It's an extension of regular ANOVA. The general idea is the same, but the MANOVA test has to include at least two dependent variables to analyze differences between multiple groups (factors) of the independent variable.

Assumptions of MANOVA

- The dependent variables should be normally distribute within groups. The R function mshapiro.test( )[in the mvnormtest package] can be used to perform the Shapiro-Wilk test for multivariate normality. This is useful in the case of MANOVA, which assumes multivariate normality.
- Homogeneity of variances across the range of predictors.
- Linearity between all pairs of dependent variables, all pairs of covariates, and all dependent variable-covariate pairs in each cell
- Multicollinearity: There should be no multicollinearity (very high correlations i.e., > 0.9) among dependent variables
- There should be no outliers in the dependent variables. dependent variables should be continuous

Test Procedure

When performing a MANOVA, most statistical software will actually produce four test statistics:

- Pillai's Trace
$$V = \text{trace}\left[\mathbf{H}(\mathbf{H}{+}\mathbf{E})^{-1}\right]$$
- Wilks' Lambda The Wilks' Lambda can be interpreted as the proportion of the variance in the outcomes that is not explained by an effect. To calculate Wilks' Lambda, for each eigenvalue, calculate $1/(1 +$ the eigenvalue), then find the product of these ratios.
$$\Lambda = \frac{|\boldsymbol{E}|}{|\boldsymbol{E} + \boldsymbol{H}|}$$
- Lawley-Hotelling Trace The Lawley-Hotelling trace is calculated as
$$T_0^2 = trace[\boldsymbol{E}^{-1}\boldsymbol{H}]$$
- Roy's Largest Root Roy's greatest root is the largest eigenvalue of
$$R = \boldsymbol{E}^{-1}\boldsymbol{H}$$

Example

Example: Suppose we have a dataset of various plant varieties and their associated phenotypic measurements for plant heights (height) and canopy volume. We want to see if plant heights and canopy volume are associated with different plant varieties using MANOVA. The dataset is provided in the Brightspace.

```
> df=read.csv( "C:\\Users\\aryalg\\Desktop\\manova_data.csv")
> head(df,3)
  plant_var height canopy_vol
1         A     20       0.70
2         A     22       0.80
3         A     24       0.95
> attach(df)
> dep_vars <- cbind(height, canopy_vol)
> fit <- manova(dep_vars ~ plant_var, data = df)
> summary(fit)
          Df Pillai approx F num Df den Df    Pr(>F)
plant_var  3 1.0365   12.909      6     72 7.575e-10 ***
Residuals 36
```

Analysis of covariance(ANCOVA)

The analysis of covariance aims to model the relationship between a quantitative response variable Y and quantitative and qualitative explanatory variables. So Analysis of Covariance refers to regression problems where there is a mixture of quantitative and qualitative predictors. It augments the ANOVA model with one or more additional quantitative variables, called covariates, which are related to the response variable. The covariates are included to reduce the variance in the error terms and provide more precise measurement of the treatment effects. ANCOVA is used to test the main and interaction effects of the factors, while controlling for the effects of the covariate.

Example

The sample data in the present example come from a study of the effects of childhood sexual abuse on adult females. 45 women being treated at a clinic, who reported childhood sexual abuse (`csa`), were measured for post-traumatic stress disorder (`ptsd`) and childhood physical abuse (`cpa`), both on standardized scales. Also measured were 31 women treated at the same clinic, who did not report childhood sexual abuse, were also measured. All women were treated for problems in their committed relationships with male living partners. The full study was more complex than reported here. You can learn more in the original article (Rodriguez et al., 1997).
The complete data set `sexab` available in the `faraway` package.
`cpa` - Childhood physical abuse on a standard scale
`ptsd` - Post-traumatic stress disorder on a standard scale
`csa` - Childhood sexual abuse:(Abused or NotAbused)
We want to know

- Is there a difference in the population means of ptsd for the Abused and the NotAbused
- Is there a treatment effect of `csa` on `ptsd`

Example

We cane use the R code below to perform the analysis

```
> library(faraway)
> data(sexab)
> attach(sexab)
> head(sexab,5)
       cpa     ptsd    csa
1  2.04786  9.71365 Abused
2  0.83895  6.16933 Abused
3 -0.24139 15.15926 Abused
4 -1.11461 11.31277 Abused
5  2.01468  9.95384 Abused
> by(sexab,sexab$csa,summary)
> plot(ptsd˜csa,sexab)
> plot(ptsd˜cpa, pch=as.character(csa),sexab)
> model=lm(ptsd˜cpa*csa, sexab)
> model.matrix(model)
```

Note that “Absued” is coded as zero and “NotAbsued” is coded as one. This defult choise is made in the alphabetical order. So absued is the refreence level.

Example

Since the interaction term is not significant we have

```
> model=lm(lm(ptsd~cpa+csa, sexab))
> summary(model)
Call:
lm(formula = lm(ptsd ~ cpa + csa, sexab))
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   10.2480     0.7187  14.260  < 2e-16 ***
cpa            0.5506     0.1716   3.209  0.00198 **
csaNotAbused  -6.2728     0.8219  -7.632 6.91e-11 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

> plot(ptsd~cpa, pch=as.character(csa),sexab)
> abline(10.248,0.551)
> abline(10.248-6.273,0.551,lty=2)# slope of both lines is 0.551
> fit=fitted(model)
> resid=residuals(model)
> plot(fit,resid, pch=as.character(csa))
```

Example

We would like to test whether there is a significance difference in the PTSD for sexual Abused and NotAbsued groups and the effect of childhood physical abuse

```
> library(faraway)
> data(sexab)
> attach(sexab)
> t.test(sexab$ptsd[1:45],sexab$ptsd[46:76])
        Welch Two Sample t-test
data:  sexab$ptsd[1:45] and sexab$ptsd[46:76]
t = 8.9006, df = 63.675, p-value = 8.803e-13
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 5.618873 8.871565
sample estimates:
mean of x mean of y
11.941093  4.695874
```

Observe that there is a significant difference in the PTSD. The estimated average of childhood sexual abuse is 11.9411-4.6959=7.2452. Note that after adjusting for the effect of childhood physical abuse, our estimate of the effect of childhood sexual abuse on PTSD is mildly reduced.