Lecture 11 ANALYSIS OF VARIANCE

Objective:

- ANOVA technique used for Regression
- One way ANOVA for categorical explanatory variables
- Two way ANOVA/Multiway for two/more categorical variable.
- ANCOVA used for dummy variable model.

Introduction

- The technique of ANOVA was used in conjunction with Least square Regression to interpret the total variation of the data from its mean.
- Traditionally ANOVA technique has also been used for fitting Linear models which are composed of only categorical explanatory variables.
- If the model is composed of only one factor then it is named "One way ANOVA" wheras a model composed of two factors is named "Two way ANOVA".
- An alternative model ANCOVA is composed of quantitative as well as qualitative variables. The acronym "ANCOVA" expands to Analysis of Covariance. The dummy variables model was an interpretation and implementation of this model.

One Way Analysis of Variance

This is the model that is composed of one categorical/qualitative regressor (factor) . The factor could have more than one category.

This classification (one factor with 3 categories) can be represented as:

$Y_i = \alpha + \gamma_1 D_{i1} + \gamma_2 D_{i2} + \varepsilon_i$ Remember if there are 3 categories we need 2 dummy variables.

Group	D1	D2
1	1	0
2	0	1
3	0	0

The expectation of each response variable is the population mean. Recollect the expectation of the error term is zero. The three equation can be reformulated as follows:

Group 1 $\mu_1 = \alpha + \gamma_1(1) + \gamma_2(0)$

$\mu_1 = \alpha + \gamma_1$

Group 2 $\mu_2 = \alpha + \gamma_1(0) + \gamma_2(1)$

$\mu_2 = \alpha + \gamma_2$

Group 3 $\mu_3 = \alpha + \gamma_1(0) + \gamma_2(0)$

$\mu_3 = \alpha$

Solving the equations gives us:

$\alpha = \mu_3$ Alpha captures the mean of the baseline group.

$\gamma_1 = \mu_1 - \mu_3$

$\gamma_2 = \mu_2 - \mu_3$

γ_1 and γ_2 are differences of means of other groups.

The One way Analysis of Variance tests the difference of means amongst the groups of the factor.

The omnibus F statistics tests

Null Hypothesis H_0:μ_1=μ_2=μ_3 No difference of the population means for the three groups of the factor

This hypothesis corresponds to H_0: $\gamma_1 = \gamma_2 = \gamma_3$

ALTERNATE MODEL

The ANOVA model can be formulated as follows:

Analyzing data using a single factor randomized experiment

One Way ANOVA

Treatment *Factor* *GROUPS*	***Observations***						***Totals***	***Average***
1	**y11**	**y21**	**y31**	……	……	**yj1**	$y_{\bullet 1}$	$\overline{y_1}$
2	**y12**	**y22**	**y32**	……	……	**yj2**	$y_{\bullet 2}$	$\overline{y_2}$
….	…….	……	…..	……..	……	…….	……	…….
m	**y1m**	**y2m**	**y3m**	……..	…….	**Yjm**	$y_{\bullet m}$	$\overline{y}_m$
							$y_{\bullet\bullet}$	$\overline{y_{\bullet\bullet}}$

Suppose we have m levels (treatments/groups) of one factor .

Y_{ij} Is the ith observation of the jth group where total groups are m. This response Y_{ij} treatment is a random variable.

n_j is the total number of observations in the jth group.

$n = \sum_{i=1}^{m} n_j$ Overall Total number of observations .

$\mu_j = E(Y_{ij})$

The observation in the above table can be represented by the following linear statistical model:

$$Y_{ij} = \mu + \alpha_j + \varepsilon_{ij} \quad \begin{matrix} i = 1........m \\ j = 1........J \end{matrix}$$

Y_{ij} is a response variable denoting (ij th) observation.

μ is the population mean

α_j is the effect of the jth group

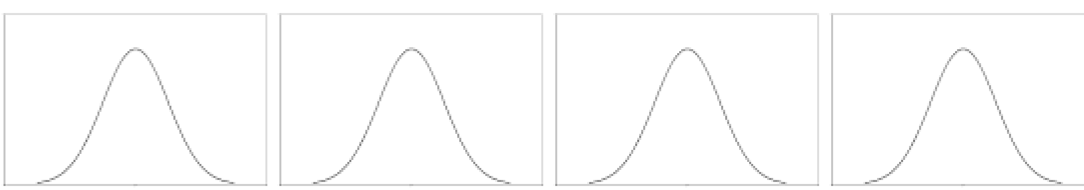
$\in_{ij}$ is the random error component

$Y_{ij} = \mu_j + \varepsilon_{ij}$

Where $\mu_j = \mu + \alpha_j$ the mean of the jth treatment equals the overall mean plus the treatment effect for j th treatment.

We assume by the linear model assumptions $\in_{ij}$ (random error component) is normally and independently distributed with mean zero and variance σ^2. Each treatment can assumed to be normally distributed with mean μ_i and variance σ^2.

σ^2 σ^2 σ^2 σ^2



μ_1 μ_2 μ_3 μ_4

$\mu + \alpha_1$ $\mu + \alpha_2$ $\mu + \alpha_3$ $\mu + \alpha_4$

Here the treatment have been chosen therefore it is called a fixed effect model (explanatory variable is non random) . If the treatment is chosen at random then it is a non fixed effect model whose results can be generalized to treatments in the population.

Here we check the variability of α_j

We are creating an analysis of variance for a fixed effect model.

For a fixed effect model treatment effects α_j are the deviations from the overall mean.

Therefore the sigma constraint or sum to zero constraint(sum will be above the overall eamn some will b below the overall mean):

$$\sum_{i=1}^{m} \alpha_j = 0$$

Hypothesis testing: Testing the equality of a treatment means.

Null Hypothesis $H_0: \mu_1 = \mu_2 = \mu_3$

This can be equivalently written as

$H_0: \alpha_1 = \alpha_2 = \alpha_3 = \ldots\ldots\ldots\alpha_I = 0$

H_1: Atleast one of the $\alpha_i \neq 0$

m=number of categories in a factor n=number of

Sources of Variation	**Sum of Squares**	**Degree of freedom**	**Mean Squared**	**F**	**H_0**
Treatment Groups	$\sum n_j (\overline{Y_j} - \overline{Y})^2$	**m-1**	$\frac{\operatorname{Re} gSS}{m-1} = \operatorname{Re} gMS$	$\frac{\operatorname{Re} gMS}{RMS}$	$\alpha_1 = \alpha_2 = \ldots. = \alpha_m = 0$ $\mu_1 = \mu_2 = \ldots. = \mu_m = 0$
Within groups Error Residuals	$\sum\sum (Y_{ij} - \overline{Y_j})^2$	**n-m**	$\frac{RSS}{n-m} = RMS$		
Total	$\sum\sum (Y_{ij} - \overline{Y})^2$	**n-1**			

Example

Ho: Mean of tensile strengths of hardwoods are not significantly different.

Hardwood Concentration %(Treatments)	Observations 1	2	3	4	5	6	Totals	Average
5	7 y_{11}	8 y_{12}	15	11	9	10	60	10 $\overline{y_1}$
10	12	17	13	18	19	15	94	15.67
15	14	18	19	17	16	18	102	17
20	19	25	22	23	18	20	127	21.17
							383	15.96 $\overline{y}_{..}$

Total $\sum\sum(Y_{ij}-\overline{Y})^2$ **=(7-15.96)²+(8-15.96)²+……………………(20-15.96)² =512.96**

Between $\sum n_j(\overline{Y_j}-\overline{Y})^2$ **= (10-15.96)²+(15.67-15.96)²+………………(21.17-15.96)²=382.79**

Within $\sum\sum(Y_{ij}-\overline{Y_j})^2$ **= (7-10)²+….(10-10)²+(12-15.67)²+…..(20-21.17)²=130.17**

Sources of Variation	Sum of Squares	Degree of freedom	Mean Squared	F $\frac{\mathrm{Re}\,gMS}{RMS}$ 19.60
Treatment Group	SS_B 382.79 Sum of Square Deviation Between groups	m-1 4-1 =3	$\frac{\mathrm{Re}\,gSS}{m-1}=\mathrm{Re}\,gMS$ =382.79/3 =127.60	
Within groups Error	SS_W 130.17 Sum of Square Deviation within factor	n-m=24-4=20	$\frac{RSS}{n-m}=RMS$ =130.17/20 =6.51	
Total	SS_T 512.96	24-1=23		

Taking α=.01 $P(F_{3,20}>19.60)=3.59*10^{-6}$ which is considerably smaller than .01.

We have strong evidence to reject H_0. Therefore our test is significant and we can conclude that Hardwood concentration has an effect on tensile strength.

TWO WAY ANALYSIS CLASSIFICATION

For a two way ANOVA we have two factors. The classification table can be represented as follows:

The means are the means for the response variables for the level of categorical variable C and that of categorical variable R.

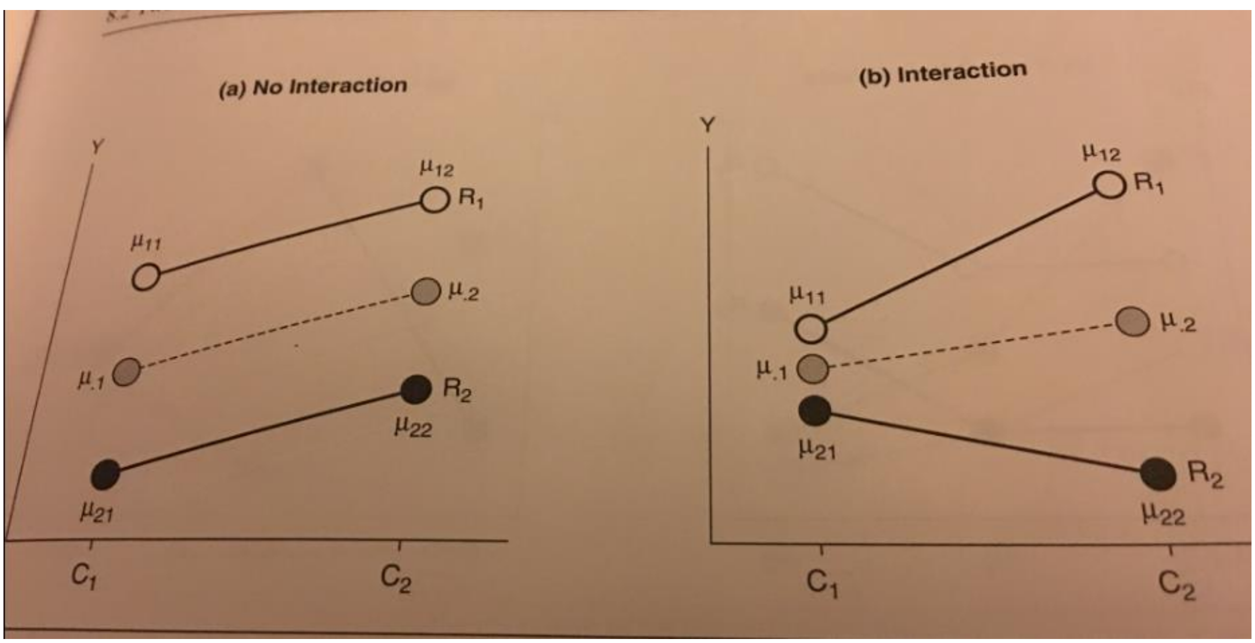
	C_1	C_2	………………………	C_c	
R_1	μ_{11}	μ_{12}		μ_{1c}	$\mu_{1.}$ Marginal mean R1
R_2	μ_{21}	μ_{22}		μ_{2c}	$\mu_{2.}$
..					
..					
R_r	μ_{r1}	μ_{r2}		μ_{rc}	$\mu_{r.}$
	$\mu_{.1}$ Marginal mean C1	$\mu_{.2}$	……………………………	$\mu_{.c}$	$\mu_{..}$

Interactions: Informs the change of response variable by the interaction of R and C. If the lines are parallel then R and C are interacting to affect the response variable Y.

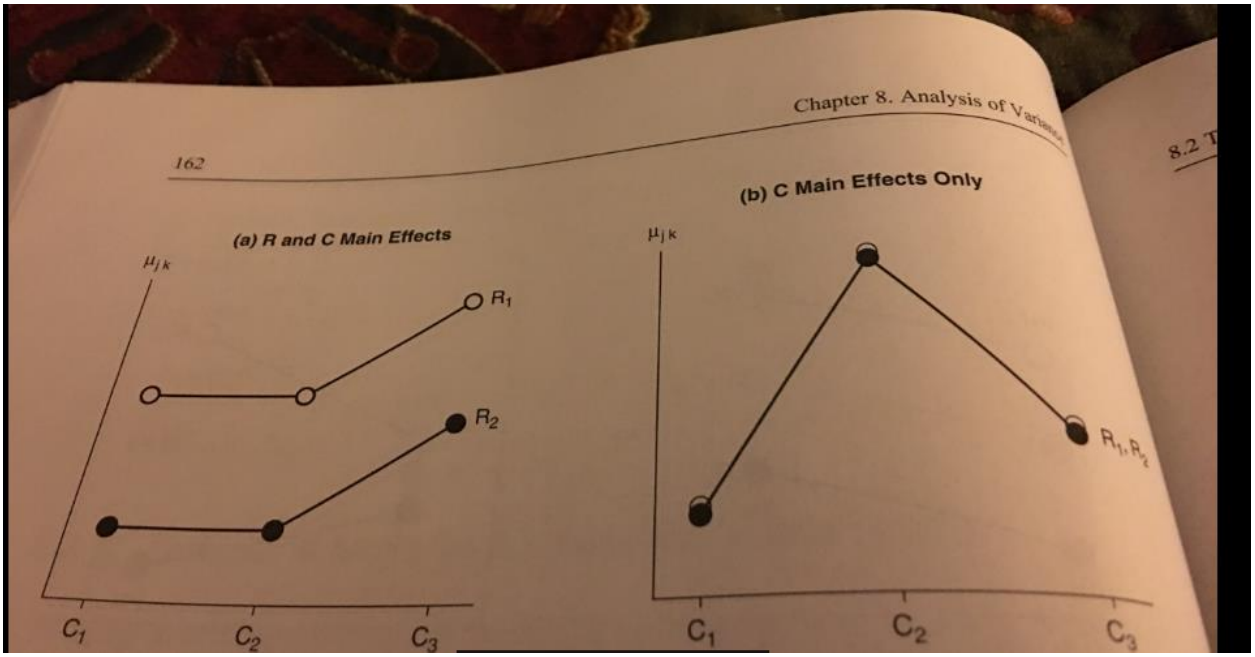
Main effects: Are the partial effects of R or C or both. These are the difference of the marginal means and the grand mean.

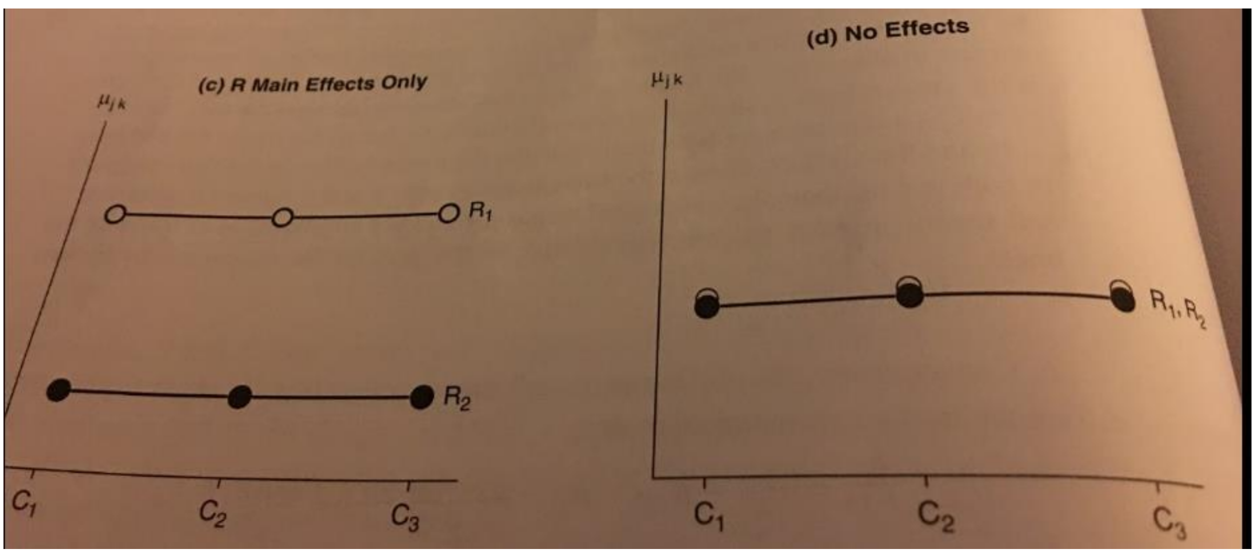
Visualizations

Interactions:



Main Effects: **Ignore one variable and check the means across the other.**





Two way Model and Hypothesis Testing

Interaction Hypothesis

$H_0 : \mu_{jk} - \mu_{jk*} = \mu_{j*k} - \mu_{j*k*}$ **No Interactions**

Columns effects are invariant across rows

Main Effects Hypothesis. Equality of Marginal Means

Row Classification:

$$H_0 : \mu_{1*} = \mu_{2*} \ldots\ldots\ldots\ldots\ldots = \mu_{r*}$$

Column Classification**:**

$$H_0 : \mu_{*1} = \mu_{*2} \ldots\ldots\ldots\ldots\ldots = \mu_{*r}$$

Generally the main effects hypothesis is of consequence if there is no interactions.

The model can be represented as follows:

$$Y_{ijk} = \mu + \alpha_j + \beta_k + \gamma_{jk} + \varepsilon_{ijk}$$

Y_{ijk} is the ith observation of the jth row and kth column of the RC table.

μ is the general mean of Y.

α_j, β_k Are the main effects parameters for row effects and column effects.

γ_{jk} Is the interaction term

ε_{ijk} Is the random error.

Three way ANOVA Classification as well as Multiway ANOVA classification can be conducted in a similar manner.

Three Way Hypothesis Formula for factor A (Factor B and C will have similar Hypothesis):

Interaction and Main effects Hypothesis

A Main effects

$$H_0 : \alpha_A = 0$$

$$H_0 : \alpha_A = 0 \mid \alpha_{AB} = \alpha_{AC} = \alpha_{ABC}$$

Interactions

AB

$$H_0 : \alpha_{AB} = 0$$

$$H_0 : \alpha_{AB} = 0 \mid \alpha_{AB} = \alpha_{AC} = \alpha_{ABC}$$

ABC

$$H_0 : \alpha_{ABC} = 0$$

Three Way classification has the highest order interaction is equal to the number of factors in the model. It is not mandatory that the model is composed of all higher level factors.

ANCOVA

This model is Analysis of Covariance. This model is identical to the Dummy variable Model. The formulation of the model is different though the interpretation is the same. R implements it by the same command anova().