Applied Analysis of Variance Spring 2019
Final Case Studies

Austin Gregory
Demetrios Papakostas
STP 531

April 30th, 2019

16.51 Refer to the Ischemic heart disease data set in Appendix C.9. Carry out a one-way analysis of variance of this data set, where the response of interest is total cost (variable 2) and the single factor is total number of interventions (variable 5). Recode the number of interventions into six categories: 0, 1, 2, 3-4, 5-7, and greater than or equal to 8. The analysis should consider transformations of the response variable. Document steps taken in your analysis, and justify your conclusions.

Answer: We began the study by recoding the number of interventions into six categories: 0, 1,2,3-4, 5-7, and greater than or equal to 8.

We use the ANOVA model

$$Y_{ij} = \mu_i + \varepsilon_{ij} \tag{1}$$

Where Y_{ij} is the value of the response variable in the jth trial for the ith factor level or treatment, μ_i are the parameters, and ε_{ij} are independent $N(0, \sigma^2)$.

The fitted value for observation Y_{ij}, denoted by $\hat{Y}_{ij}$ for regression models, is simply the corresponding factor level sample mean here:

$$\hat{Y}_{ij} = \overline{Y}_{i\cdot}.$$

In our case,

$$\begin{aligned}
0 &= \hat{Y}_{1j} = \overline{Y}_{1\cdot} = 257.37 \\
1 &= \hat{Y}_{2j} = \overline{Y}_{2\cdot} = 309 \\
2 &= \hat{Y}_{3j} = \overline{Y}_{3\cdot} = 434 \\
\text{3-4} &= \hat{Y}_{4j} = \overline{Y}_{4\cdot} = 1111 \\
\text{5-7} &= \hat{Y}_{5j} = \overline{Y}_{5\cdot} = 2678 \\
\text{8+} &= \hat{Y}_{6j} = \overline{Y}_{6\cdot} = 10153
\end{aligned}$$

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
DS$‘(Recoded) Interventions‘	5	10984631240.29	2196926248.06	70.87	0.0000
Residuals	782	24241162222.15	30998928.67		

Our null and alternative hypothesis are as follows

$$\begin{aligned}
H_0 &= \mu_1 = \mu_2 = \mu_3 \\
H_a &= \text{not all } \mu_i \text{ are equal}
\end{aligned}$$

ie we are testing whether all the means should be the same. We control at $\alpha = 0.05$ risk. Since we know that F^* is distributed as $F(r-1, n_T - r)$ when H_0 holds and that large values of F^* lead to conclusion H_a, the appropriate decision rule to control the level of significance at α is

$$\begin{aligned}
&\text{If} \quad F^* \le F(1-\alpha; r-1, n_T - r), \quad \text{conclude } H_0 \\
&\text{If} \quad F^* \ge F(1-\alpha; r-1, n_T - r), \quad \text{conclude } H_a
\end{aligned} \tag{2}$$

In our case, because $r = 6$ and $n_{\text{total}} = 788$, we find $F(0.95, 5, 782) = 2.22$. Our F^* statistic is given by

$$F^* = \frac{\text{MSTR}}{\text{MSE}} \tag{3}$$

From our ANOVA table

$$F^* = \frac{10.06}{0.6401} = 70.87 > 2.22$$

Therefore, we conclude the alternative hypothesis.

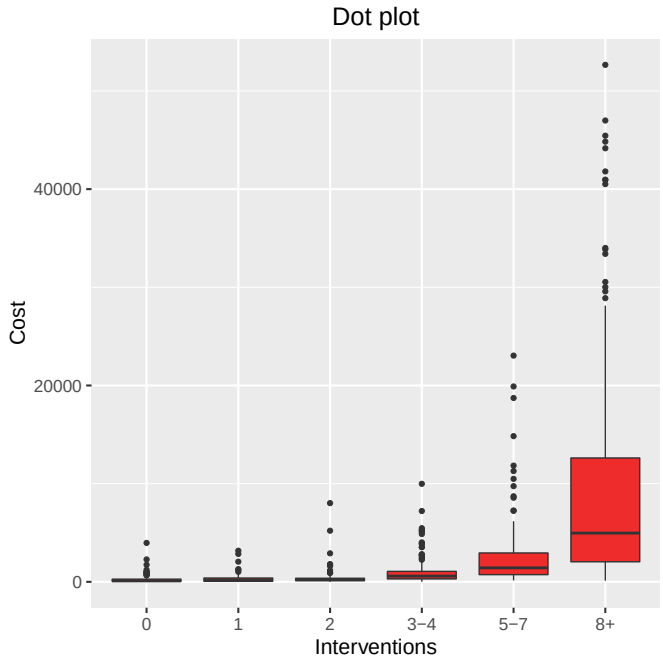

Figure 1: Box Plot across factor levels of number of interventions. Plot indicates transformation of response variable may be appropriate because of apparant breach of constancy of error variance. More formally, a Brown-Forysthe test or Breusch Pagan test would give the same result.

From our Box-Cox transformation, the ideal solution is $Y^{0.02}$, but looking at the plot in figure 3, we see the log transformation is in the neighborhood for maximizing log likelihood. Therefore, we transform using the log transformation on the response variable. Since the log(0) is undefined, we manually encoded cost of 0 dollars to transform to 0 dollars after the log transformation.

The means after the transformation are

$$\begin{aligned}
\ln(\text{intervention=0}) &= \hat{Y}_{1j} = \overline{Y}_{1\cdot} = 4.69 \\
\ln(\text{intervention=1}) &= \hat{Y}_{2j} = \overline{Y}_{2\cdot} = 4.89 \\
\ln(\text{intervention=2}) &= \hat{Y}_{3j} = \overline{Y}_{3\cdot} = 5.33 \\
\ln(\text{intervention=3-4}) &= \hat{Y}_{4j} = \overline{Y}_{4\cdot} = 6.37 \\
\ln(\text{intervention=5-7}) &= \hat{Y}_{5j} = \overline{Y}_{5\cdot} = 7.29 \\
\ln(\text{intervention=8+}) &= \hat{Y}_{6j} = \overline{Y}_{6\cdot} = 8.51
\end{aligned}$$

The ANOVA table for this question is given in table 1. The relevant R-code for this question is pasted below.

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
DS$‘(Recoded) Interventions‘	5	1588.46	317.69	195.36	0.0000
Residuals	782	1271.70	1.63		

Table 1: ANOVA table for 16.51

```
DS<-read.table("STP 531 Final Data Set.txt")
names(DS)<-c("ID", "Total Cost","Age","Gender","Interventions(# of)","Drugs (# of)","ER visits","
    Complications (# of)","Comorbidities (# of","Duration (days)")

#This function codes the 'Intervention' variable according to the problem specifications.
```

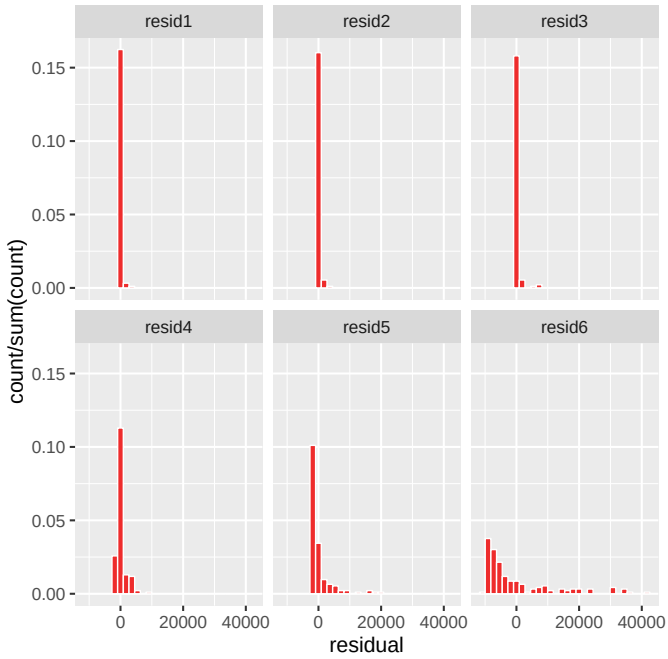

Figure 2: Residuals across different factors. Normality appears to be violated especially at levels corresponding to more intervention

```
RecodedInterventions<-function(x) {
if (x %in% c(0,1,2)) {
result<-as.character(x)
}

if (x %in% c(3,4)) {
result<-"3-4"
}

if (x %in% c(5,6,7)) {
result<-"5-7"
}

if (x >=8 ) {
result<-"8+"
}

return(result)
}
NewCol<-sapply(DS$`Interventions(# of)`,RecodedInterventions)
nDS<-as.data.frame(cbind(DS,NewCol))
names(nDS)<-c(names(DS),"(Recoded) Interventions")
nDS<-nDS[,-c(5)]
DS<-nDS
DS$Gender<-as.factor(DS$Gender)
summary(lm(DS$`Total Cost`~DS$`ER visits`))
mean1=mean(DS$`Total Cost`[DS$`(Recoded) Interventions`=='0'])
mean2=mean(DS$`Total Cost`[DS$`(Recoded) Interventions`=='1'])
mean3=mean(DS$`Total Cost`[DS$`(Recoded) Interventions`=='2'])
mean4=mean(DS$`Total Cost`[DS$`(Recoded) Interventions`=='3-4'])
mean5=mean(DS$`Total Cost`[DS$`(Recoded) Interventions`=='5-7'])
mean6=mean(DS$`Total Cost`[DS$`(Recoded) Interventions`=='8+'])

resid1=DS$`Total Cost`[DS$`(Recoded) Interventions`=='0']-mean1
resid2=DS$`Total Cost`[DS$`(Recoded) Interventions`=='1']-mean2
resid3=DS$`Total Cost`[DS$`(Recoded) Interventions`=='2']-mean3
```

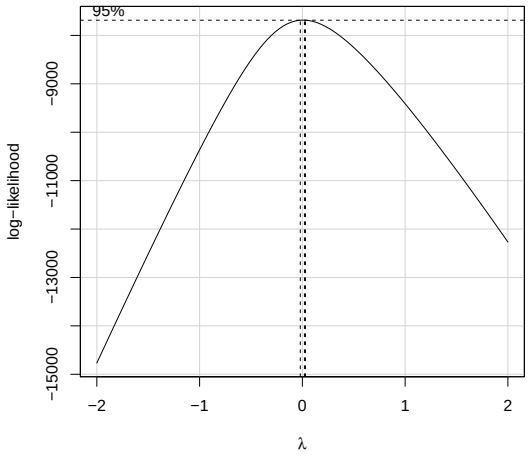

Figure 3: Box-Cox transformations meant to maximize the log-likelihood.

```
resid4=DS$`Total Cost`[DS$`(Recoded) Interventions`=='3-4']-mean4
resid5=DS$`Total Cost`[DS$`(Recoded) Interventions`=='5-7']-mean5
resid6=DS$`Total Cost`[DS$`(Recoded) Interventions`=='8+']-mean6
sum(resid1)+sum(resid2)+sum(resid3)+sum(resid4)+sum(resid5)+sum(resid6)
#ggsave('1651plot1.pdf',
#       DS%>%ggplot(aes(x=`(Recoded) Interventions`, y=`Total Cost`))+geom_boxplot(fill='
    firebrick2')+#+geom_dotplot(binaxis="y", stackdir="center")+
#ggtitle('Dot plot')+xlab('Interventions')+ylab('Cost')+
#theme(plot.title=element_text(hjust=0.5))+theme(legend.title=element_blank()))
newstuff=log(DS$`Total Cost`+1)

stuff=aov(newstuff~DS$`(Recoded) Interventions`)
anova(stuff)
```

17.41 Refer to the Ischemic heart disease data set in Appendix e.9 and Case Study 16.1. Obtain confidence intervals for all pairwise comparisons among the six number-of-intervention categories: use the Tukey procedure and a 90 percent family confidence coefficient. Interpret your results and state your findings. Prepare a line plot of the estimated factor level means, underscoring all non-significant comparisons

Answer: We do the comparison after the log transformation from question 16.51. For each pairwise comparison, we multiply by $Ts(\hat{D})$, where

$$T = \frac{1}{\sqrt{2}} q(1-\alpha, r, n_T - r)$$

and

$$s^2(\hat{D}_i) = \text{MSE}\left(\frac{1}{n_i} + \frac{1}{n_{i'}}\right)$$

This leads to, after calculating each pairwise comparison as the different in mean of total cost across each intervention level:

Comparison	$\hat{D}$	lower	Upper	p-value
1:0	0.197	-0.215	0.609	0.817
2:0	0.642	0.216	1.07	0.001
3-4:0	1.68	1.27	2.08	≈ 0
5-7:-0	2.60	2.19	3.00	≈ 0
8+:0	3.82	3.42	4.20	≈ 0
2:1	0.44	0.017	0.874	0.0771
3-4:1	1.48	1.07	1.88	≈ 0
5-7:1	2.40	1.99	2.81	≈ 0
8+:1	3.62	3.22	4.01	≈ 0
3-4:2	1.03	0.61	1.46	≈ 0
5-7:2	1.96	1.54	2.38	≈ 0
8+:2	3.17	2.76	3.58	≈ 0
5-7:3-4	0.93	0.53	1.32	≈ 0
8+:3-4	2.14	1.753	2.53	≈ 0
8+:5-7	1.21	0.83	1.60	≈ 0

Table 2: Results of the pairwise Tukey family procedure.

We conducted a family of tests of the form

$$H_0 : \mu_i - \mu_{i'} = 0$$
$$H_a : \mu_i - \mu_{i'} \neq 0$$

If zero is contained in the interval, the conclusion H_0 is reached, otherwise H_a is concluded. The results of the Tukey procedure tell us that at a 90% family confidence coefficient, we reject the null for all comparisons except $1:0$. Our intervals not containing 0 are intervals where we reject the Null at confidence level $\alpha = 0.10$. The p-values given in 2 support the conclusion.

If this Tukey procedure were to be performed without transforming the response, 7 of the comparisons would be non-significant. This is because the cost for more interventions dramatically increases, so comparing the means for the levels of low number of interventions will appear to be non-significant, because the difference in their mean cost will be miniscule compared to the mean cost of the high number of interventions level.

The line plot also provides visual confirmation:

The relevant R-code for the Tukey comparisons:

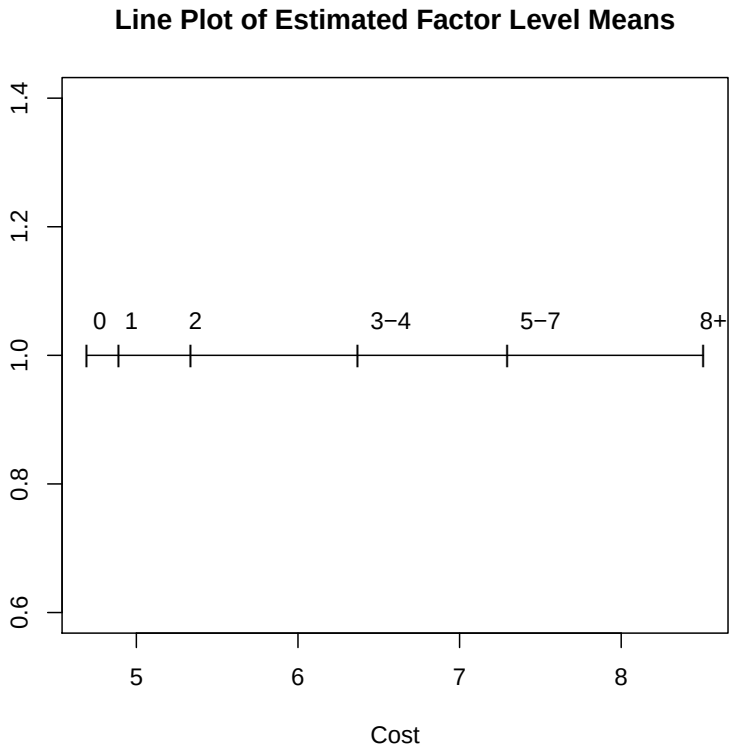

Figure 4: Line plot indicates significance between all pairs except 0 and 1

```
DS<-read.table("STP 531 Final Data Set.txt")
names(DS)<-c("ID", "Total Cost","Age","Gender","Interventions(# of)","Drugs (# of)","ER visits","
    Complicaitons (# of)","Comorbidities (# of","Duration (days)")
DS$Gender<-as.factor(DS$Gender)


#This function codes the 'Intervention' variable according to the problem specifications.
RecodedInterventions<-function(x) {
if (x %in% c(0,1,2)) {
result<-as.character(x)
}

if (x %in% c(3,4)) {
result<-"3-4"
}

if (x %in% c(5,6,7)) {
result<-"5-7"
}

if (x >=8 ) {
result<-"8+"
}

return(result)
}

NewCol<-sapply(DS$`Interventions(# of)`,RecodedInterventions)
nDS<-as.data.frame(cbind(DS,NewCol))
names(nDS)<-c(names(DS),"(Recoded) Interventions")
```

```
nDS<-nDS[,-c(5)]
DS<-nDS
DS$Gender<-as.factor(DS$Gender)

newstuff=log(DS$`Total Cost`+1)

stuff=aov(newstuff~DS$`(Recoded) Interventions`)
anova(stuff)

TukeyHSD(stuff, conf.level=0.9)
```

And for the line plot:

```
meanlog1=mean(log(DS$`Total Cost`+1)[DS$`(Recoded) Interventions`=='0'])
meanlog2=mean(log(DS$`Total Cost`+1)[DS$`(Recoded) Interventions`=='1'])
meanlog3=mean(log(DS$`Total Cost`+1)[DS$`(Recoded) Interventions`=='2'])
meanlog4=mean(log(DS$`Total Cost`+1)[DS$`(Recoded) Interventions`=='3-4'])
meanlog5=mean(log(DS$`Total Cost`+1)[DS$`(Recoded) Interventions`=='5-7'])
meanlog6=mean(log(DS$`Total Cost`+1)[DS$`(Recoded) Interventions`=='8+'])

means2=c(meanlog1,meanlog2,meanlog3, meanlog4, meanlog5, meanlog6)
plot(y=c(1,1,1,1,1,1),x=means2, type = 'o', pch = '|', ylab = '', xlab="productivity improvement",
    main="Line Plot of Estimated Factor Level Means")+text(x=c(meanlog1-.05,meanlog2-.05,meanlog3
    -.1, meanlog4-.01, meanlog5-.01, meanlog6-.11),y=rep(1.05,3), c("0", "1", "2", "3-4", "5-7", "
    8+"), cex=1, pos=4, col="black")
```

23.43 Refer to the Ischemic heart disease data set in Appendix C.9 and Case Study 23.42. Assume that the sample sizes reflect the importance of the treatment means. Carry out an unbalanced two-way analysis of variance of this data set, where the response of interest is total cost (variable 2), the two crossed factors are number of interventions (variable 5) and number of comorbidities (variable 9). Recode the number of interventions into six categories: 0, 1,2,3-4, 5-7, and greater than or equal to 8. Recode the number of comorbidities into two categories: 0-1, and greater than or equal to 2. The analysis should consider transformations of the response variable. Document the steps taken in your analysis and justify your conclusions.

Answer: We saw in problem 22.36 that the log transform of the response meets the model assumptions so we continue with it. The two-way ANOVA with unequal sample size model can be expressed as follows:

$$Y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \varepsilon_{ijk} \tag{4}$$

Where ε_{ijk} are iid $N(0,\sigma^2)$ and

$$\sum_i \alpha_i = \sum_j \beta_j = \sum_i (\alpha\beta)_{ij} = \sum_j (\alpha\beta)_{ij} = 0$$

Where Y_{ijk} is the kth observation of the ijth cell, where $k = 1,\ldots,n_{ij}$ noting that the n_{ij}'s are not all the same, and $i = 1,\ldots,a = 2, j = 1,\ldots,b = 6$. α_i is the unknown parameter describing the main effect of factor A, in this case the number of comorbidities. β_j is the unknown parameter describing the main effect of factor B, the interventions. And, as always, $(\alpha\beta)_{ij}$ is the interaction effects term.

To develop our regression model, we note we need $a-1=1$ indicator variables for factor A main effects and $b-1=5$ main effects for factor B. Therefore, we have

$$\begin{aligned} Y_{ijk} = {} & \mu + \alpha_1 X_{ijk1} + \beta_1 X_{ijk2} + \beta_2 X_{ijk3} + \beta_3 X_{ijk4} + \beta_4 X_{ijk5} + \beta_5 X_{ijk6} + \\ & (\alpha\beta)_{11} X_{ijk1} X_{ijk2} + (\alpha\beta)_{12} X_{ijk1} X_{ijk3} + (\alpha\beta)_{13} X_{ijk1} X_{ijk4} + (\alpha\beta)_{12} X_{ijk1} X_{ijk5} + (\alpha\beta)_{15} X_{ijk1} X_{ijk6} + \varepsilon_{ijk} \end{aligned}$$

Where

$$X_1 = \begin{cases} 1 & \text{if case from level 1 for factor A} \\ -1 & \text{if case from level 2 for factor A} \end{cases} \qquad X_4 = \begin{cases} 1 & \text{if case from level 3 for factor B} \\ -1 & \text{if case from level 6 for factor B} \\ 0 & \text{otherwise} \end{cases}$$

$$X_2 = \begin{cases} 1 & \text{if case from level 1 for factor B} \\ -1 & \text{if case from level 6 for factor B} \\ 0 & \text{otherwise} \end{cases} \qquad X_5 = \begin{cases} 1 & \text{if case from level 4 for factor B} \\ -1 & \text{if case from level 6 for factor B} \\ 0 & \text{otherwise} \end{cases}$$

$$X_3 = \begin{cases} 1 & \text{if case from level 2 for factor B} \\ -1 & \text{if case from level 6 for factor B} \\ 0 & \text{otherwise} \end{cases} \qquad X_6 = \begin{cases} 1 & \text{if case from level 5 for factor B} \\ -1 & \text{if case from level 6 for factor B} \\ 0 & \text{otherwise} \end{cases}$$

The α_1 represents the A main effect, the β_1 and β_2 terms are for the B main effect, and the next two represent the AB interaction effect.

The ANOVA table for the question:

	Sum Sq	Df	Mean Sq	F value	Pr(>F)
(Intercept)	29375.86	1	29375.86	22873.45	0.0000
‘(Recoded) Interventions‘	1366.04	5	273.2	212.73	0.0000
‘(Recoded) Comorbidities‘	217.17	1	217.17	169.10	0.0000
‘(Recoded) Interventions‘:‘(Recoded) Comorbidities‘	38.62	5	7.72	6.01	0.0000
Residuals	996.60	776	1.28		

Table 3: Looking at the p-values, for $\alpha = 0.05$, we conclude factor A and factor B main effects, and AB interaction effects are present.

Because we know we have interaction effects, we look at confidence intervals on difference between interaction level differences.

We held factor intervention fixed at each factor level and compared the difference in means between the two factor levels of comorbidities. We calculate the standard error from $s(\hat{D}) = \sqrt{\text{MSE}\left(\frac{1}{n_i} + \frac{1}{n_{i'}}\right)}$. Using the Bonferroni family comparison method with $\alpha = 0.05$ and $g = 6$ because we have 6 comparisons, we find

	Diff	Lower	Upper
$\mu_{i1} - \mu_{i'1}$	-1.64	-1.94	-1.34
$\mu_{i2} - \mu_{i'2}$	-1.72	-2.01	-1.42
$\mu_{i3} - \mu_{i'3}$	-0.872	-1.18	-0.559
$\mu_{i4} - \mu_{i'4}$	-0.892	-1.18	-0.602
$\mu_{i5} - \mu_{i'5}$	-0.494	-0.781	-.206
$\mu_{i6} - \mu_{i'6}$	-0.811	-1.08	-0.538

Table 4: Family confidence intervals for interaction effects. Because 0 is not contained in any interval, we conclude the interaction effects are significant.

We first give the code that redefines our comorbidities into levels as we did earlier:

```
#This function is to code the comorbidities according to the problem specifications.
RecodeComorbidities<-function(x) {
if (x %in% c(0,1)) {
result<-"0-1"
}
```

```
if (x >= 2) {
result<-"2+"
}
return(result)
}
#apply the function
NewCol2<-sapply(DS$`Comorbidities (# of`,RecodeComorbidities)
nDS<-as.data.frame(cbind(DS,NewCol2))
names(nDS)<-c(names(DS),"(Recoded) Comorbidities")

DS<-nDS[,-c(8)]
```

The following R-code allows us to get the ANOVA table for the full model:

```
library(car)
aov2346=lm(`Total Cost`+1~`(Recoded) Interventions`*`(Recoded) Comorbidities`,data=DS)

xtable(Anova(aov2346, type=3))
```

Finally, the code for the confidence intervals

```
mecom1=mean(log(DS$`Total Cost`+1)[DS$`(Recoded) Interventions`=='0'&DS$`(Recoded) Comorbidities`==
    '0-1'])
mecom2=mean(log(DS$`Total Cost`+1)[DS$`(Recoded) Interventions`=='1'&DS$`(Recoded) Comorbidities`==
    '0-1'])
mecom3=mean(log(DS$`Total Cost`+1)[DS$`(Recoded) Interventions`=='2'&DS$`(Recoded) Comorbidities`==
    '0-1'])
mecom4=mean(log(DS$`Total Cost`+1)[DS$`(Recoded) Interventions`=='3-4'&DS$`(Recoded) Comorbidities
    `=='0-1'])
mecom5=mean(log(DS$`Total Cost`+1)[DS$`(Recoded) Interventions`=='5-7'&DS$`(Recoded) Comorbidities
    `=='0-1'])
mecom6=mean(log(DS$`Total Cost`+1)[DS$`(Recoded) Interventions`=='8+'&DS$`(Recoded) Comorbidities
    `=='0-1'])

MECOM1=mean(log(DS$`Total Cost`+1)[DS$`(Recoded) Interventions`=='0'&DS$`(Recoded) Comorbidities`==
    '2+'])
MECOM2=mean(log(DS$`Total Cost`+1)[DS$`(Recoded) Interventions`=='1'&DS$`(Recoded) Comorbidities`==
    '2+'])
MECOM3=mean(log(DS$`Total Cost`+1)[DS$`(Recoded) Interventions`=='2'&DS$`(Recoded) Comorbidities`==
    '2+'])
MECOM4=mean(log(DS$`Total Cost`+1)[DS$`(Recoded) Interventions`=='3-4'&DS$`(Recoded) Comorbidities
    `=='2+'])
MECOM5=mean(log(DS$`Total Cost`+1)[DS$`(Recoded) Interventions`=='5-7'&DS$`(Recoded) Comorbidities
    `=='2+'])
MECOM6=mean(log(DS$`Total Cost`+1)[DS$`(Recoded) Interventions`=='8+'&DS$`(Recoded) Comorbidities
    `=='2+'])

mecomlist=c(mecom1-MECOM1,mecom2-MECOM2, mecom3-MECOM3,mecom4-MECOM4,mecom5-MECOM5,mecom6-MECOM6)

mecomlist

plot(mecomlist)

mse2343=1.28
sd_D1=sqrt(1.28*(1/length(DS$`Total Cost`[DS$`(Recoded) Interventions`=='0']))+(1/length(DS$`Total
    Cost`[DS$`(Recoded) Comorbidities`=='0-1'])))
sd_D2=sqrt(1.28*(1/length(DS$`Total Cost`[DS$`(Recoded) Interventions`=='1']))+(1/length(DS$`Total
    Cost`[DS$`(Recoded) Comorbidities`=='0-1'])))
sd_D3=sqrt(1.28*(1/length(DS$`Total Cost`[DS$`(Recoded) Interventions`=='2']))+(1/length(DS$`Total
    Cost`[DS$`(Recoded) Comorbidities`=='0-1'])))
sd_D4=sqrt(1.28*(1/length(DS$`Total Cost`[DS$`(Recoded) Interventions`=='3-4']))+(1/length(DS$`
    Total Cost`[DS$`(Recoded) Comorbidities`=='0-1'])))
```

```
sd_D5=sqrt(1.28*(1/length(DS$`Total Cost`[DS$`(Recoded) Interventions`=='5-7']))+(1/length(DS$`
    Total Cost`[DS$`(Recoded) Comorbidities`=='0-1'])))
sd_D6=sqrt(1.28*(1/length(DS$`Total Cost`[DS$`(Recoded) Interventions`=='8+']))+(1/length(DS$`Total
     Cost`[DS$`(Recoded) Comorbidities`=='0-1'])))
sdlist=c(sd_D1,sd_D2, sd_D3, sd_D4, sd_D5, sd_D6)
n=length(DS$`Total Cost`[DS$`(Recoded) Comorbidities`=='0-1'])

bonval=qt(1-.05/(2*6), (408-1)*12 )
bonval*sd_D1
upperrange=c()
lowerrange=c()
for (i in 1:6){
upperrange[[i]]=mecomlist[i]+bonval*sdlist[i]
lowerrange[[i]]=mecomlist[i]-bonval*sdlist[i]
}
upperrange
lowerrange
```

Extra Q An experiment was designed to study the association between the wheat variety and its growth (Y). There are four wheat varieties of interest, and five greenhouse benches are set up as blocks. Within each block, the four varieties of wheat were planted; all the other conditions were kept the same as possible. The data on measurements of plant heights (in inches) after a period of time are given below:

Table 5: Table for the extra problem

Block ↓ Variety →	1	2	3	4
1	9.7	11.8	6.3	4.6
2	6.6	9.7	5.3	3.4
3	7.6	10.9	4.7	2.3
4	8.1	11.3	5.5	3.6
5	6.4	10.7	4.5	2.8

(a) State an appropriate ANOVA model for this study. State and check the model assumptions.

Answer: We use ANOVA model (21.1):

$$Y_{ij} = \mu.. + \rho_i + \tau_j + \varepsilon_{ij} \tag{5}$$

where $\mu..$ is a constant, ρ_i are constants for the block (row) effects, subject to the restriction $\sum \rho_i = 0$, τ_j are constants for the treatment effects, subject to the restriction $\sum \tau_j = 0$, and ε_{ij} are independent $N(0,\sigma^2)$. We note $i = 1,\ldots,n_n$ and $j = 1,\ldots,r$, where r is the number of levels of the treatment factor, in this example the number of varieties, 4.

We check normality of errors, the blocks do not interact, and constant variance.

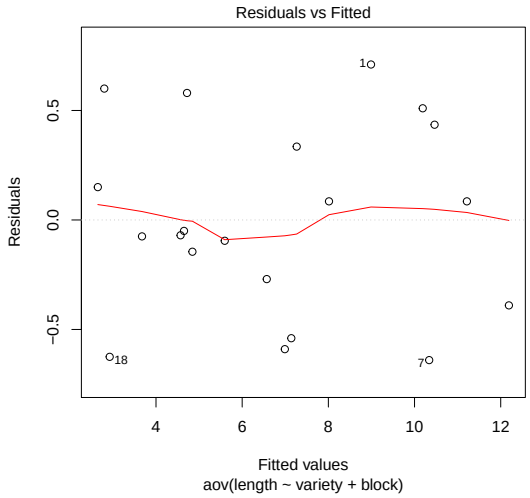

Figure 5: Plot of residuals vs fitted value. Assumptions do not appear to be violated.

From the Tukey test of additivity

Assume

$$(\alpha\beta)_{ij} = D\alpha_i\beta_j$$

Where D is some constant. A regular two-factor ANOVA model with interactions for the case $n = 1$ is

$$Y_{ij} = \mu.. + \alpha_i + \beta_j + D\alpha_i\beta_j + \varepsilon_{ij}$$

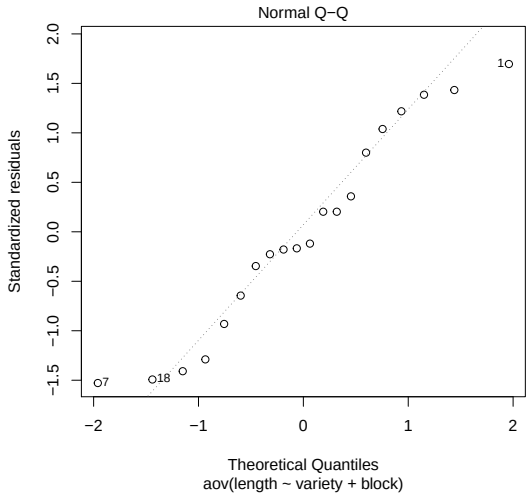

Figure 6: QQ plot. Normality of error assumption does not appear to be violated. Our Null hypothesis is that the The correlation value is 0.979. At significance level $\alpha = 0.05$, with $n = 20$, we compare to correlation value 0.951, from table B.6. Since the observed coefficient exceeds this level, we have support for the conclusion that the distribution of the error terms does not depart substantially from a normal distribution.

where each term has usual meaning. No third subscript because $n = 1$. Least squares and MLE of of D turns out to be

$$\hat{D} = \frac{\sum_i \sum_j \alpha_i \beta_j Y_{ij}}{\sum_i \alpha_i^2 \sum_j \beta_j^2}$$

The usual estimator of α_i is $\overline{Y}_{i\cdot} - \overline{Y}_{\cdot\cdot}$ and that of β_j is $\overline{Y}_{\cdot j} - \overline{Y}_{\cdot\cdot}$. Replacing the parameters in $\hat{D}$ with these estimators, we obtain:

$$\hat{D} = \frac{\sum_i \sum_j (\overline{Y}_{i\cdot} - \overline{Y}_{\cdot\cdot})(\overline{Y}_{\cdot j} - \overline{Y}_{\cdot\cdot}) Y_{ij}}{\sum_i (\overline{Y}_{i\cdot} - \overline{Y}_{\cdot\cdot})^2 \sum_j (\overline{Y}_{\cdot j} - \overline{Y}_{\cdot\cdot})^2}$$

If we substitute the sample estimates into $\sum\sum D^2 \alpha_i^2 \beta_j^2$, we find

$$\begin{aligned} \text{SSAB}^* &= \sum_i \sum_j \hat{D}^2 (\overline{Y}_{i\cdot} - \overline{Y}_{\cdot\cdot})^2 (\overline{Y}_{\cdot j} - \overline{Y}_{\cdot\cdot})^2 \\ &= \frac{\left[\sum_i \sum_j (\overline{Y}_{i\cdot} - \overline{Y}_{\cdot\cdot})(\overline{Y}_{\cdot j} - \overline{Y}_{\cdot\cdot})\right]^2}{\sum_i (\overline{Y}_{i\cdot} - \overline{Y}_{\cdot\cdot})^2 \sum_j (\overline{Y}_{\cdot j} - \overline{Y}_{\cdot\cdot})^2} \end{aligned}$$

The analysis of variance decomposition for the special interaction model is

$$\text{SSTO} = \text{SSA} + \text{SSB} + \text{SSAB}^* + \text{SSRem}^*$$

where SSRem^* is the *remainder sum of squares*:

$$\text{SSRem}^* = \text{SSTO} - \text{SSA} - \text{SSB} - \text{SSAB}^*$$

It can be shown that if $D = 0$, i.e. no interactions of type $D\alpha_i\beta_j$ exist, then SSAB^* and SSRem^* are independently distributed as chi-square random variables with 1 and $ab - a - b$ dof, respectively. Hence, if $D = 0$, the test statistic

$$F^* = \frac{\text{SSAB}^*}{1} \setminus \frac{\text{SSRem}^*}{ab - a - b}$$

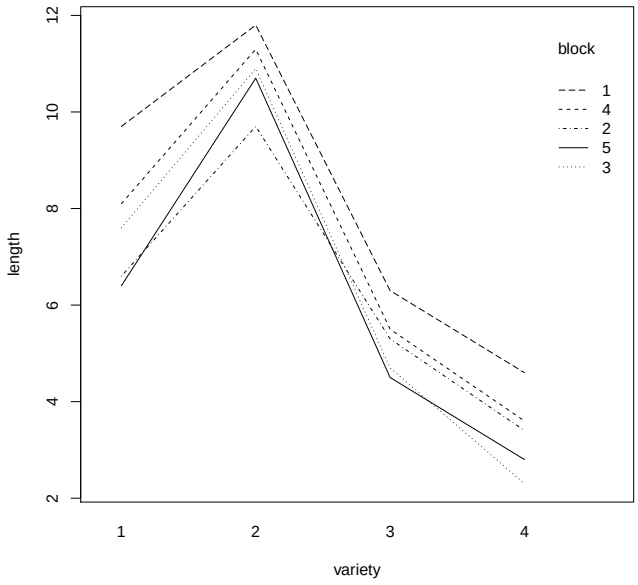

Figure 7: Interaction Plot. There appear to be main effects for block and definitely main effects by variety appear to be the case. No obvious interactions appear to be present.

is distributed as $F(1, ab-a-b)$, so for testing:

$$H_0 : D = 0 \quad \text{no interations present}$$
$$H_a : D \neq 0 \quad \text{interactions } D\alpha_i\beta_j \text{ present}$$

where we control type I error at $\alpha = 0.05$ with the following decision rule, if $F^* \leq F(1-\alpha; 1, ab-a-b) = 4.84$, we conclude H_0, otherwise conclude H_a.

$F^0.00599 < 30.8$, so we conclude H_0, that $D = 0$, and the p-value was 0.937. Therefore, we conclude no interactions.

We assume no interaction between the blocks and the varieties.

(b) Test whether or not the main effect of variety is present. State the hypotheses, test statistic, p-value, and conclusion. Use $\alpha = 0.05$.

Answer: We first give an interaction plot to help visualize: Recall a and b are the number of levels per factor (5 and 4 respectively) Our hypothesis for the test of block effects is:

$$H_0 : \alpha_1 = \alpha_2 = \alpha_3 = \alpha_4 = \alpha_5$$
$$H_a : \text{not all } \alpha_i \text{ equal zero}$$

For the test of block effects:

$$F^* = \frac{\text{MSA}}{\text{MSBLTR}} = 2.77/.292 = 9.47$$

and, with $a = 5$ and $b = 4$, If $F^* \leq F[1-\alpha; a-1, (a-1)(b-1)] = F(.95; 4, 12) = 3.25$, conclude H_a, that block effects are present.

$$H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4$$
$$H_a : \text{not all } \beta_i \text{ equal zero}$$

For variety effects:

$$F^* = \frac{\text{MSBL}}{\text{MSBLTR}} = 52.94/.292 = 181.14$$

and $F^* \le F[1-\alpha; b-1, (a-l)(b-1)] = F(.95; 3, 12) = 3.49$, so conclude H_a, the variety main effects are present.

Our test for location (α factor) is (noting we just swap α and β to do the test for week effects)

(c) Obtain confidence intervals for all pairwise comparisons between the treatment (variety) means; use the most efficicent multiple comparison procedure with a 90% family confidence coefficient. Interpret your results.

Answer:

Bonferroni multiple is (where g=6) $t\big(1-\alpha/2g, (a-1)(b-1)\big) = 2.77$. The Tukey multiple is $T = \frac{1}{\sqrt{2}} q(1-\alpha; r, (a-1)\cdot(b-1)) = 2.56$ and for Scheffe $S = \sqrt{(r-1)F(1-\alpha, r-1, (a-1)\cdot(b-1)} = 2.796$. The Tukey value is the most efficient.

We have 4 varieties and want to compare them, which we means we have $\binom{4}{2} = 6$ combinations.

$$\hat{D}_1 = \overline{Y}_{1\cdot} - \overline{Y}_{\cdot 2} = 7.68 - 10.88 = -3.20$$
$$\hat{D}_2 = \overline{Y}_{1\cdot} - \overline{Y}_{\cdot 3} = 7.68 - 5.26 = 2.42$$
$$\hat{D}_3 = \overline{Y}_{1\cdot} - \overline{Y}_{\cdot 4} = 7.68 - 3.34 = 4.34$$
$$\hat{D}_4 = \overline{Y}_{2\cdot} - \overline{Y}_{\cdot 3} = 10.88 - 5.26 = 5.62$$
$$\hat{D}_5 = \overline{Y}_{2\cdot} - \overline{Y}_{\cdot 4} = 10.88 - 3.34 = 7.54$$
$$\hat{D}_6 = \overline{Y}_{3\cdot} - \overline{Y}_{\cdot 4} = 5.26 - 3.34 = 1.92$$

We need to find $s(\hat{D}_i) = \text{MSE}\frac{(a+b-1)}{ab}$, where $a = 5$ and $b = 4$, the number of levels of the block and variety respectively. Therefore, we find for the first 6 cases, (noting all $c_i = 1$ or -1), since

$$s(\hat{D}_i) = \sqrt{2 \cdot \text{MSE}/5} = \sqrt{.4 * \text{MSE}} = 0.3417$$

because we have 2 levels of week for each location.

The Tukey value is 2.56.

$$-3.20 \pm (2.56)\cdot 0.34 \longrightarrow -4.07 \le \mu_{1\cdot} \le -2.33$$
$$2.42 \pm (2.56)\cdot 0.34 \longrightarrow 1.55 \le \mu_{2\cdot} \le 3.29$$
$$4.34 \pm (2.56)\cdot 0.34 \longrightarrow 3.47 \le \mu_{3\cdot} \le 5.21$$
$$5.62 \pm (2.56)\cdot 0.34 \longrightarrow 4.75 \le \mu_{4\cdot} \le 6.49$$
$$7.54 \pm (2.56)\cdot 0.34 \longrightarrow 6.67 \le \mu_{5\cdot} \le 8.41$$
$$-5.62 \pm (2.56)\cdot 0.34 \longrightarrow 1.05 \le \mu_{6\cdot} \le 2.79$$

(d) Estimate the difference in mean plant height between the first two groups of variety, i.e. $(\mu_1 - \mu_2)$ with a 95% confidence interval. Interpret your findings.

Answer: We found the estimate in part (c):

$$\hat{D} = \overline{Y}_{1\cdot} - \overline{Y}_{\cdot 2} = 7.68 - 10.88 = -3.20$$

Further,

$$s(\hat{D}) = \sqrt{\sum c_i^2 \cdot \text{MSE}} = 0.34$$

As usual, the confidence interval is

$$\hat{D} \pm s(\hat{D}) t\big(1-\alpha/2; (a-1)(b-1)\big) \longrightarrow -3.94 \le D \le -2.45$$

The interpretation is the usual. With confidence coefficent $\alpha = 0.95$, the mean difference in plant height (in inches) between variety 1 and 2 is -3.22 to -5.58 inches, i.e. plants in variety 2 are smaller than variety 1.

(e) Test for $H_0 : \mu_1 = \mu_2$ vs $H_a : \mu_1 \neq \mu_2$ with $\alpha = 0.05$. State the test-statistic, p-value, and your conclusion. Does your conclusion agree with the result in (d)? Explain.

Answer: Test statistic is

$$\frac{\hat{D}}{s(\hat{D})} = \frac{-3.2}{0.34} = -9.364$$

We compare to $t(0.975, 12) = 2.18$. This corresponds to a 2-sided p-value of 7.24e-07, well below the 0.05 risk we are controlling at. We conclude the alternative. This makes sense, as in part (d) our 95% confidence did not contain 0, which the null hypothesis we test here stipulates it should if we want to conclude H_0.

(f) Test for $H_0 : \mu_{\cdot 1} \geq \mu_{\cdot 2}$ vs $H_a : \mu_{\cdot 1} < \mu_{\cdot 2}$.

Answer: This is the one-tail equivalent. We reach the same result, but the p-value is smaller, 1/2 in fact.

(g) Comment on the efficiency of the blocking variable.

Answer: Our ANOVA table is: We estimate E as follows: (using our ANOVA table)

Table 6: ANOVA table for the test

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
variety	3	158.82	52.94	181.14	0.0000
block	4	11.07	2.77	9.47	0.0011
Residuals	12	3.51	0.29		

$$\hat{E} = \frac{s_r^2}{\text{MSBLTR}} = \frac{(n_b - 1) \cdot \text{MSBL} + n_b(r-1)\text{MSBLTR}}{(n_b r - 1) \cdot \text{MSBLTR}} = \frac{4 \cdot 2.77 + 8 \cdot 0.29}{(12 - 1) \cdot 0.29} = 2.80$$

We would require almost three times as many replications per treatment with a completely randomized block design to achieve the same variance of any estimated contrast as is obtained by blocking by pain score.

R-code for this question.

```
extraprob=data.frame(cbind(c(1,2,3,4,5),c(9.7,6.6,7.6,8.1,6.4),
 c(11.8,9.7, 10.9,11.3,10.7),
 c(6.3, 5.3, 4.7, 5.5, 4.5),
 c(4.6, 3.4, 2.3, 3.6, 2.8)))


colnames(extraprob)=c('block', '1','2', '3', '4')

extraprob2=extraprob%>%
gather(variety, length,'1','2','3','4')
extraprob2$block=as.factor(extraprob2$block)
block=extraprob2$block

tukeys.add.test <- function(y,A,B){
## Y is the response vector
## A and B are factors used to predict the mean of y
## Note the ORDER of arguments: Y first, then A and B
dname <- paste(deparse(substitute(A)), "and", deparse(substitute(B)),
"on",deparse(substitute(y)) )
A <- factor(A); B <- factor(B)
ybar.. <- mean(y)
ybari. <- tapply(y,A,mean)
ybar.j <- tapply(y,B,mean)
len.means <- c(length(levels(A)), length(levels(B)))
SSAB <- sum( rep(ybari. - ybar.., len.means[2]) *
rep(ybar.j - ybar.., rep(len.means[1], len.means[2])) *
tapply(y, interaction(A,B), mean))^2 /
( sum((ybari. - ybar..)^2) * sum((ybar.j - ybar..)^2))
```

```
aovm <- anova(lm(y ~ A+B))
SSrem <- aovm[3,2] - SSAB
dfdenom <- aovm[3,1] - 1
STATISTIC <- SSAB/SSrem*dfdenom
names(STATISTIC) <- "F"
PARAMETER <- c(1, dfdenom)
names(PARAMETER) <- c("num df", "denom df")
D <- sqrt(SSAB/  ( sum((ybari. - ybar..)^2) * sum((ybar.j - ybar..)^2)))
names(D) <- "D estimate"
RVAL <- list(statistic = STATISTIC, parameter = PARAMETER,
p.value = 1 - pf(STATISTIC, 1,dfdenom), estimate = D,
method = "Tukey's one df F test for Additivity",
data.name = dname)
attr(RVAL, "class") <- "htest"
return(RVAL)
}

tukeys.add.test(extraprob2$length, extraprob2$variety,extraprob2$block)

#ggsave('intplotfinal.pdf',interaction.plot( extraprob2$variety,block,extraprob2$length, fun=
    mean , xlab='variety',ylab='length', legend=T))

extraprob2%>%
summarize(mean(length[variety==2]))
meanvariety1=mean(extraprob2$length[extraprob2$variety==1])
meanvariety2=mean(extraprob2$length[extraprob2$variety==2])
meanvariety3=mean(extraprob2$length[extraprob2$variety==3])
meanvariety4=mean(extraprob2$length[extraprob2$variety==4])
7.68-5.26
aovextra1=aov(length~variety+block, data=extraprob2)
plot(residuals(aovextra1))
dfextra=data.frame(residuals(aovextra1))

dfextra$index=seq(from=1, to=20)
colnames(dfextra)[1]='resid'

#ggsave('extraprobplot1.pdf',dfextra%>%
#  ggplot(aes(x=index,y=resid))+geom_point()+xlab('index')+ylab('residuals')+geom_hline(
    yintercept = 0))
(TukeyHSD(aovextra1, conf.level=.90)$variety)
library(xtable)
xtable(anova(aovextra1))
qf(0.95,3,12)
sqrt(0.292)
-3.2-0.34*qt(.975, 12)
-4.4/0.54
qt(0.975,12)

qf(.95,1,11)

T=1/sqrt(2)*qtukey(.9, 4,12)
S=sqrt(3*qf(.9,3,12))
B=qt(1-.1/12,12)
T
S
B
sqrt(.4*.292)
1.92+2.56*.34
2*pt(-3.2/.3417, 12)

(4*2.77+15*.29)/(19*.29)
plot(aovextra1)
qq=qqnorm(residuals(aovextra1))
cor(qq$x, qq$y)
```