MAST30025 Week 11 Lab R Code

#Question 1: #Part a: Express this as a two factor model with no interaction in matrix form.

y = c(18.8,21.2,16.7,19.8,23.9,22.3,15.9,19.2,21.8)  
#Need a shortcut version for this  
X1 = matrix(c(rep(1,9),1,1,1,1,1,0,0,0,0,0,0,0,0,0,1,1,1,1,1,1,0,0,0,1,0,0,0,0,0,1,0,0,0,1,1,0,0,0,0,1,1,0,0,0,1),9,6)  
X1

## [,1] [,2] [,3] [,4] [,5] [,6]  
## [1,] 1 1 0 1 0 0  
## [2,] 1 1 0 1 0 0  
## [3,] 1 1 0 0 1 0  
## [4,] 1 1 0 0 0 1  
## [5,] 1 1 0 0 0 1  
## [6,] 1 0 1 1 0 0  
## [7,] 1 0 1 0 1 0  
## [8,] 1 0 1 0 1 0  
## [9,] 1 0 1 0 0 1

#Where error terms are obvious!

#Part b: Express this as a two factor model with interaction in matrix form.

y = c(18.8,21.2,16.7,19.8,23.9,22.3,15.9,19.2,21.8)  
#Need a shortcut version for this  
X = matrix(c(rep(1,9),1,1,1,1,1,0,0,0,0,0,0,0,0,0,1,1,1,1,1,1,0,0,0,1,0,0,0,0,0,1,0,0,0,1,1,0,0,0,0,1,1,0,0,0,1,1,1,0,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,1,1,0,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,1,1,0,0,0,0,0,0,0,0,0,1),9,12)  
X

## [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10] [,11] [,12]  
## [1,] 1 1 0 1 0 0 1 0 0 0 0 0  
## [2,] 1 1 0 1 0 0 1 0 0 0 0 0  
## [3,] 1 1 0 0 1 0 0 1 0 0 0 0  
## [4,] 1 1 0 0 0 1 0 0 1 0 0 0  
## [5,] 1 1 0 0 0 1 0 0 1 0 0 0  
## [6,] 1 0 1 1 0 0 0 0 0 1 0 0  
## [7,] 1 0 1 0 1 0 0 0 0 0 1 0  
## [8,] 1 0 1 0 1 0 0 0 0 0 1 0  
## [9,] 1 0 1 0 0 1 0 0 0 0 0 1

## Part c: Express the hypothesis that there is no interaction in terms of your parmeters. Eliminate any redundancies.

#Using Theorem 7.5 from the less than full rank model page 60!

#Part d: Input this data into R. Plot an interaction plot between breed and diet.

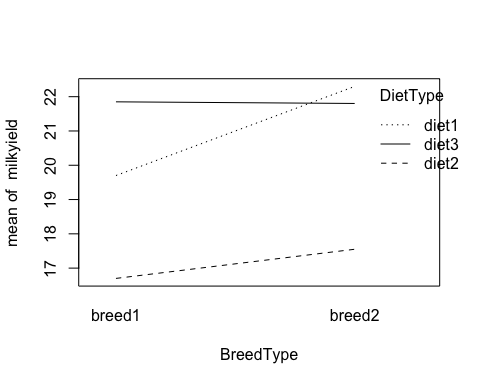
#Attempt 1: Create Dataframe  
Breed = c('breed1','breed1','breed1','breed1','breed1','breed2','breed2','breed2','breed2')  
Diet = c('diet1','diet1','diet2','diet3','diet3','diet1','diet2','diet2','diet3')  
milkyield = c(18.2,21.2,16.7,19.8,23.9,22.3,15.9,19.2,21.8)  
cow = data.frame(milkyield,Breed,Diet)  
str(cow)

## 'data.frame': 9 obs. of 3 variables:  
## $ milkyield: num 18.2 21.2 16.7 19.8 23.9 22.3 15.9 19.2 21.8  
## $ Breed : chr "breed1" "breed1" "breed1" "breed1" ...  
## $ Diet : chr "diet1" "diet1" "diet2" "diet3" ...

# However we want Breed and Diet to become factors!  
BreedType = factor(cow$Breed)  
DietType= factor(cow$Diet)  
newcow = data.frame(milkyield,BreedType,DietType) #This we want to recreate our dataframe!  
str(newcow)

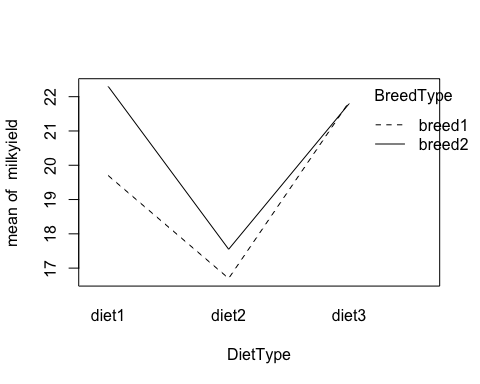
## 'data.frame': 9 obs. of 3 variables:  
## $ milkyield: num 18.2 21.2 16.7 19.8 23.9 22.3 15.9 19.2 21.8  
## $ BreedType: Factor w/ 2 levels "breed1","breed2": 1 1 1 1 1 2 2 2 2  
## $ DietType : Factor w/ 3 levels "diet1","diet2",..: 1 1 2 3 3 1 2 2 3

with(newcow, interaction.plot(BreedType,DietType,milkyield))



#Actual solution

with(newcow, interaction.plot(DietType,BreedType,milkyield)) #Switch tao with beta



#Part e: Test for the presense of interaction

library(Matrix)  
library(MASS)  
#Use Matrix package!  
r = rankMatrix(X)[1]  
r

## [1] 6

library(Matrix)  
library(MASS)  
XtXc = ginv(t(X)%\*%X)  
b = XtXc%\*%t(X)%\*%y  
b

## [,1]  
## [1,] 10.016667  
## [2,] 4.620833  
## [3,] 5.395833  
## [4,] 4.083333  
## [5,] 1.400000  
## [6,] 4.533333  
## [7,] 1.279167  
## [8,] 0.662500  
## [9,] 2.679167  
## [10,] 2.804167  
## [11,] 0.737500  
## [12,] 1.854167

s2 = sum((y-X%\*%b)^2)/(length(y)-r)  
s2

## [1] 5.576667

#To compute our C matrix!  
C = matrix(0,2,12)  
C[1,c(7,8,10,11)] = c(1,-1,-1,1)  
C[2,c(7,9,10,12)] = c(1,-1,-1,1)  
C[,-(1:6)]

## [,1] [,2] [,3] [,4] [,5] [,6]  
## [1,] 1 -1 0 -1 1 0  
## [2,] 1 0 -1 -1 0 1

Fstat = (t(C%\*%b)%\*%solve(C%\*%XtXc%\*%t(C))%\*%C%\*%b/2/s2)  
Fstat

## [,1]  
## [1,] 0.1680614

pf(Fstat,2,9-6,lower = F)

## [,1]  
## [1,] 0.8527444

#There is no interaction (without the lm function) #Now with the lm

model = lm(milkyield~BreedType\*DietType, data = newcow)  
anova(model)

## Analysis of Variance Table  
##   
## Response: milkyield  
## Df Sum Sq Mean Sq F value Pr(>F)  
## BreedType 1 0.057 0.0569 0.0093 0.9293  
## DietType 2 35.861 17.9304 2.9314 0.1969  
## BreedType:DietType 2 2.421 1.2106 0.1979 0.8304  
## Residuals 3 18.350 6.1167

#There is no interaction, values are sightly different :(.

#Part f: What is the degrees of freedom used for the interaction test? #We use 2 and 3 degrees of freedom! from the F distribution

#Part g: From the interaction model, what is the estimated amount of milk produced from breed 2 and diet 3

#Actual Method (#Slight value difference compared to the solutions!)  
model$coefficients

## (Intercept) BreedTypebreed2   
## 19.70 2.60   
## DietTypediet2 DietTypediet3   
## -3.00 2.15   
## BreedTypebreed2:DietTypediet2 BreedTypebreed2:DietTypediet3   
## -1.75 -2.65

#Actual Method  
c(1,1,0,1,0,1)%\*%model$coefficients

## [,1]  
## [1,] 21.8

#Part i: Test the hypothesis (under the additive model) #Using Matrix theory Method 1

#Use Matrix package!  
r1 = rankMatrix(X1)[1]  
r1

## [1] 4

XtXc2 = ginv(t(X1)%\*%X1)  
b = XtXc2%\*%t(X1)%\*%y  
b

## [,1]  
## [1,] 10.916162  
## [2,] 4.941414  
## [3,] 5.974747  
## [4,] 4.564646  
## [5,] 0.720202  
## [6,] 5.631313

s2a = sum((y-X1%\*%b)^2)/(length(y)-r1)  
s2a

## [1] 3.720889

#To compute our C matrix! #Attempt 1   
C1 = matrix(0,1,6)  
C1[1,c(1,5,6)] = c(0,1,-1)  
C1

## [,1] [,2] [,3] [,4] [,5] [,6]  
## [1,] 0 0 0 0 1 -1

Fstat = (t(C1%\*%b)%\*%solve(C1%\*%XtXc2%\*%t(C1))%\*%C1%\*%b/1/s2a)  
Fstat

## [,1]  
## [1,] 8.975155

pf(Fstat,1,9-length(r1),lower = F)

## [,1]  
## [1,] 0.01717985

#2nd Method: Use the car package (Correct)

model2 = lm(milkyield~BreedType+DietType, data = newcow)  
library(car)

## Loading required package: carData

linearHypothesis(model2,c(0,0,1,-1),0)

## Linear hypothesis test  
##   
## Hypothesis:  
## DietTypediet2 - DietTypediet3 = 0  
##   
## Model 1: restricted model  
## Model 2: milkyield ~ BreedType + DietType  
##   
## Res.Df RSS Df Sum of Sq F Pr(>F)   
## 1 6 54.622   
## 2 5 20.771 1 33.85 8.1484 0.03563 \*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

#We reject the null hypothesis at 5% signifiance level. Conclude the 2nd and 3rd diets are equivalent in terms of milk produced.

#Part h: Fit the additive model. What is the estimated amount of milk produced from breed 2 and diet 3 now?

#Additive model  
model2 = lm(milkyield~1+BreedType+DietType, data = newcow)  
anova(model2)

## Analysis of Variance Table  
##   
## Response: milkyield  
## Df Sum Sq Mean Sq F value Pr(>F)   
## BreedType 1 0.057 0.0569 0.0137 0.91140   
## DietType 2 35.861 17.9304 4.3162 0.08147 .  
## Residuals 5 20.771 4.1542   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

model2$coefficients

## (Intercept) BreedTypebreed2 DietTypediet2 DietTypediet3   
## 20.188889 1.133333 -3.677778 1.266667

c(1,1,0,1)%\*%model2$coefficients #We select the intercept, Breed Type 2 and Diet Type 3

## [,1]  
## [1,] 22.58889

#Part j: Find a 95% Confidence Interval, under the additive model, for the amount of milk produced from breed 2 and diet 3. Use both matrix calculations and the estimable function from the gmodels package.

#Attempt 1  
library(gmodels)  
ci =estimable(model2,c(1,1,0,1),conf.int = 0.95) #NOTE from c(1,1,0,1)   
#We select the intercept, Breed Type 2 and Diet Type 3  
c(ci$Lower,ci$Upper)

## [1] 18.68372 26.49406

#Part k: Find the same confidence interval under the interaction model

#Attempt 1  
library(gmodels)  
ci =estimable(model,c(1,1,0,1,0,1),conf.int = 0.95) #NOTE: More to add  
#We select the intercept, Breed Type 2 and Diet Type 3  
c(ci$Lower,ci$Upper)

## [1] 13.92921 29.67079

#Part l: Why is the second interval wider than the first? #Demonstrator Answer: The second interval is wider than the first because we are attributing some degrees of freedom to the interaction term(s). The resulting loss in degrees of freedom for the residuals leads to greater error in our estimations

#Question 2: #Part a: Fit an additive model with all variables: Then repeat without the block variables. Does the fitted model change? Are the block variables significant?

str(npk)

## 'data.frame': 24 obs. of 5 variables:  
## $ block: Factor w/ 6 levels "1","2","3","4",..: 1 1 1 1 2 2 2 2 3 3 ...  
## $ N : Factor w/ 2 levels "0","1": 1 2 1 2 2 2 1 1 1 2 ...  
## $ P : Factor w/ 2 levels "0","1": 2 2 1 1 1 2 1 2 2 2 ...  
## $ K : Factor w/ 2 levels "0","1": 2 1 1 2 1 2 2 1 1 2 ...  
## $ yield: num 49.5 62.8 46.8 57 59.8 58.5 55.5 56 62.8 55.8 ...

#Attempt 1  
#Additive Model with blocks  
adb = lm(yield~1+N+P+K+block, data = npk)  
summary(adb)

##   
## Call:  
## lm(formula = yield ~ 1 + N + P + K + block, data = npk)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -7.0000 -1.7083 -0.0833 2.2458 6.4833   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 53.800 2.450 21.955 8.13e-13 \*\*\*  
## N1 5.617 1.634 3.438 0.00366 \*\*   
## P1 -1.183 1.634 -0.724 0.47999   
## K1 -3.983 1.634 -2.438 0.02767 \*   
## block2 3.425 2.830 1.210 0.24483   
## block3 6.750 2.830 2.386 0.03068 \*   
## block4 -3.900 2.830 -1.378 0.18831   
## block5 -3.500 2.830 -1.237 0.23512   
## block6 2.325 2.830 0.822 0.42412   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 4.002 on 15 degrees of freedom  
## Multiple R-squared: 0.7259, Adjusted R-squared: 0.5798   
## F-statistic: 4.966 on 8 and 15 DF, p-value: 0.003761

#Attempt 1  
#Additive Model without blocks  
adnb = lm(yield~1+N+P+K, data = npk)  
summary(adnb)

##   
## Call:  
## lm(formula = yield ~ 1 + N + P + K, data = npk)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -9.2667 -3.6542 0.7083 3.4792 9.3333   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 54.650 2.205 24.784 <2e-16 \*\*\*  
## N1 5.617 2.205 2.547 0.0192 \*   
## P1 -1.183 2.205 -0.537 0.5974   
## K1 -3.983 2.205 -1.806 0.0859 .   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 5.401 on 20 degrees of freedom  
## Multiple R-squared: 0.3342, Adjusted R-squared: 0.2343   
## F-statistic: 3.346 on 3 and 20 DF, p-value: 0.0397

#MORE TO ADD!

anova(adnb,adb)

## Analysis of Variance Table  
##   
## Model 1: yield ~ 1 + N + P + K  
## Model 2: yield ~ 1 + N + P + K + block  
## Res.Df RSS Df Sum of Sq F Pr(>F)   
## 1 20 583.48   
## 2 15 240.19 5 343.29 4.2879 0.01272 \*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

#We reject the null for both models conclue that the block variables are significant.

#Demonstrator comment: The fitted model does not change, in the sense that the parameters corresponding to N, P and K are the same for the two models. This is because the design is balanced and complete: the overall effect of the predictors of interest are observed in each individual block. However, the blocks themselves are significant: there is a difference in the yield of each block. Because the blocks have been carefully designed, this does not affect the fitted model itself when the blocks are removed from consideration.

#Part b: Fit a model with the fertilizer variables and all pairwise interaction terms. Are the interaction terms significant?

#Attempt 1  
imodel = lm(yield~1+N+P+K+N\*P+N\*K+P\*K,data = npk)  
summary(imodel)

##   
## Call:  
## lm(formula = yield ~ 1 + N + P + K + N \* P + N \* K + P \* K, data = npk)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -8.8917 -3.2875 0.4583 3.4000 9.7083   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 52.6750 3.0114 17.492 2.64e-12 \*\*\*  
## N1 9.8500 3.9429 2.498 0.023 \*   
## P1 0.4167 3.9429 0.106 0.917   
## K1 -1.9167 3.9429 -0.486 0.633   
## N1:P1 -3.7667 4.5529 -0.827 0.420   
## N1:K1 -4.7000 4.5529 -1.032 0.316   
## P1:K1 0.5667 4.5529 0.124 0.902   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 5.576 on 17 degrees of freedom  
## Multiple R-squared: 0.3968, Adjusted R-squared: 0.184   
## F-statistic: 1.864 on 6 and 17 DF, p-value: 0.146

#Attempt 1  
anova(adnb,imodel) #no blocks

## Analysis of Variance Table  
##   
## Model 1: yield ~ 1 + N + P + K  
## Model 2: yield ~ 1 + N + P + K + N \* P + N \* K + P \* K  
## Res.Df RSS Df Sum of Sq F Pr(>F)  
## 1 20 583.48   
## 2 17 528.58 3 54.898 0.5885 0.6308

#We do not reject the null. The interaction terms are not significant!

#Part c: Perform variable selection using stepwise selection with AIC, starting from the model with no interaction terms (but considering them for inclusion). What do you find?

#Attempt 1  
noimodel = step(adnb,scope =~.+N+P+K)

## Start: AIC=84.58  
## yield ~ 1 + N + P + K  
##   
## Df Sum of Sq RSS AIC  
## - P 1 8.402 591.88 82.926  
## <none> 583.48 84.583  
## - K 1 95.202 678.68 86.210  
## - N 1 189.282 772.76 89.326  
##   
## Step: AIC=82.93  
## yield ~ N + K  
##   
## Df Sum of Sq RSS AIC  
## <none> 591.88 82.926  
## - K 1 95.202 687.08 84.506  
## + P 1 8.402 583.48 84.583  
## - N 1 189.282 781.16 87.586

#We find that our final model includes the variables corresponding to nitrogen and potassium, but not phosphate.

#Attempt 1  
withimodel = step(imodel,scope =~.+N+P+K+N\*P+N\*K+P\*K)

## Start: AIC=88.21  
## yield ~ 1 + N + P + K + N \* P + N \* K + P \* K  
##   
## Df Sum of Sq RSS AIC  
## - P:K 1 0.482 529.06 86.233  
## - N:P 1 21.282 549.86 87.159  
## - N:K 1 33.135 561.72 87.671  
## <none> 528.58 88.211  
##   
## Step: AIC=86.23  
## yield ~ N + P + K + N:P + N:K  
##   
## Df Sum of Sq RSS AIC  
## - N:P 1 21.282 550.34 85.180  
## - N:K 1 33.135 562.20 85.691  
## <none> 529.06 86.233  
## + P:K 1 0.482 528.58 88.211  
##   
## Step: AIC=85.18  
## yield ~ N + P + K + N:K  
##   
## Df Sum of Sq RSS AIC  
## - P 1 8.402 558.75 83.543  
## - N:K 1 33.135 583.48 84.583  
## <none> 550.34 85.180  
## + N:P 1 21.282 529.06 86.233  
## + P:K 1 0.482 549.86 87.159  
##   
## Step: AIC=83.54  
## yield ~ N + K + N:K  
##   
## Df Sum of Sq RSS AIC  
## - N:K 1 33.135 591.88 82.926  
## <none> 558.75 83.543  
## + P 1 8.402 550.34 85.180  
##   
## Step: AIC=82.93  
## yield ~ N + K  
##   
## Df Sum of Sq RSS AIC  
## <none> 591.88 82.926  
## + N:K 1 33.135 558.75 83.543  
## - K 1 95.202 687.08 84.506  
## + P 1 8.402 583.48 84.583  
## - N 1 189.282 781.16 87.586

#Similar with the interaction model

#Part d: What is the best treatment for peas, according to your final model? Find a 95% confidence interval for the yield of this treatment.

#Attempt 1 No interaction  
finalmodel = lm(yield~N+K, data = npk)  
finalmodel$coefficients

## (Intercept) N1 K1   
## 54.058333 5.616667 -3.983333

#Actual solution: ASK FOR HELP!  
ib = estimable(finalmodel, c(1,1,0),conf.int = 0.95)  
c(ib$Lower,ib$Upper)

## [1] 55.77158 63.57842