20074027 STAT2003 TASK 2

Bruce Odek Omondi

2025-05-12

# QUESTION B

## 1. Part 1: Analysis for 16-run design only with a 2-level factorial design

Opening the dataset

# Load required packages  
library(tidyverse)

## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --  
## v dplyr 1.1.2 v readr 2.1.4  
## v forcats 1.0.0 v stringr 1.5.1  
## v ggplot2 3.4.4 v tibble 3.2.1  
## v lubridate 1.9.2 v tidyr 1.3.0  
## v purrr 1.0.1   
## -- Conflicts ------------------------------------------ tidyverse\_conflicts() --  
## x dplyr::filter() masks stats::filter()  
## x dplyr::lag() masks stats::lag()  
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors

# Read data  
drugdata <- read.csv("antiviraldrugs.csv")

Start by subsetting the dataset to include only the first 16 runs for 2-level factorial design

# Subset only the first 16 runs  
drugdata\_16 <- drugdata[1:16, ]

Then define the model that will be used to get the effect estimates

# get the average readout and encode replicates as -1 and 1   
drugs16 <- drugdata\_16 %>%  
 mutate(avg\_readout = (Replicate1 + Replicate2) / 2,  
 replicate = rep(c(-1, 1), each = 8))  
  
# model on the first 16 runs  
model16 <- lm(avg\_readout ~ A + B + C + D + E +  
 A:B + A:C + A:D + A:E +  
 B:C + B:D + B:E +  
 C:D + C:E + D:E +  
 replicate,  
 data = drugs16)  
  
summary(model16)

##   
## Call:  
## lm(formula = avg\_readout ~ A + B + C + D + E + A:B + A:C + A:D +   
## A:E + B:C + B:D + B:E + C:D + C:E + D:E + replicate, data = drugs16)  
##   
## Residuals:  
## ALL 16 residuals are 0: no residual degrees of freedom!  
##   
## Coefficients: (1 not defined because of singularities)  
## Estimate Std. Error t value Pr(>|t|)  
## (Intercept) 27.8469 NaN NaN NaN  
## A -2.9906 NaN NaN NaN  
## B -2.5469 NaN NaN NaN  
## C 0.5375 NaN NaN NaN  
## D -19.6250 NaN NaN NaN  
## E -10.9687 NaN NaN NaN  
## replicate -2.7062 NaN NaN NaN  
## A:B NA NA NA NA  
## A:C 0.8094 NaN NaN NaN  
## A:D 1.9969 NaN NaN NaN  
## A:E -1.3219 NaN NaN NaN  
## B:C 2.4781 NaN NaN NaN  
## B:D 0.2781 NaN NaN NaN  
## B:E 1.1719 NaN NaN NaN  
## C:D -0.8906 NaN NaN NaN  
## C:E 0.1656 NaN NaN NaN  
## D:E 9.0156 NaN NaN NaN  
##   
## Residual standard error: NaN on 0 degrees of freedom  
## Multiple R-squared: 1, Adjusted R-squared: NaN   
## F-statistic: NaN on 15 and 0 DF, p-value: NA

The model turns out to be saturated (no degrees of freedom, p-val, etc.) so we cannot get the significance of the estimates. According to the textbook, this situation calls for the use of fold-over for the factorial design.

Fixing the 16-run design using fold-over to allow us to get the significance of the effect estimates

# load libraries  
library(tidyverse)  
  
# getting only the first 16 runs   
original16 <- drugdata[1:16, ] %>%  
 mutate(  
 avg\_readout = (Replicate1 + Replicate2) / 2, # average of Rep 1 and 2  
 replicate = rep(c(-1, 1), each = 8) # encode replicates as -1 and 1  
 )

Fold-over requires us to flip the signs of all the levels for each factor

# create folded-over design by reversing the signs of all factor levels  
foldover16 <- original16 %>%  
 mutate(  
 A = -A, B = -B, C = -C, D = -D, E = -E, # reverse signs as required for fold-over  
 Run = Run + 100 # label the foldover to differentiate from first 16  
 )

Since we don’t have actual data for the fold-over runs, I opted to simulate readout values for the fold-over runs. Ideally, these would be actual data from doing the fold-over during the experiment

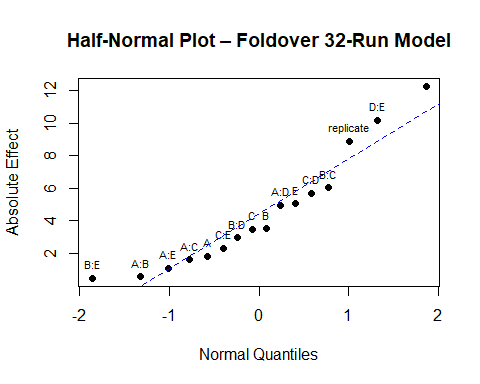
# simulate placeholder readout values for fold-over runs  
set.seed(123) # for reproducibility  
  
foldover16$avg\_readout <- rnorm(16, mean = mean(original16$avg\_readout), sd = sd(original16$avg\_readout))  
foldover16$replicate <- rep(c(-1, 1), each = 8)  
  
# combine original and fold-over data into one dataset which will have 32 runs total  
folded\_data <- bind\_rows(original16, foldover16)  
  
# fit full second-order interaction model using the combined data   
model\_folded <- lm(avg\_readout ~ A + B + C + D + E +  
 A:B + A:C + A:D + A:E +  
 B:C + B:D + B:E +  
 C:D + C:E + D:E +  
 replicate,  
 data = folded\_data)

Then I defined a function that would be used to plot the half-normal plots of effects

# define and generate a half-normal plot to assess effect significance  
custom\_half\_normal <- function(model, model\_name) {  
 effects <- coef(model)[-1] # exclude intercept  
 abs\_effects <- sort(abs(effects), decreasing = FALSE)  
 n <- length(abs\_effects)  
 quantiles <- qnorm((1:n - 0.5) / n)  
   
 plot(quantiles, abs\_effects,  
 main = paste("Half-Normal Plot –", model\_name),  
 xlab = "Normal Quantiles", ylab = "Absolute Effect", pch = 19)  
 text(quantiles, abs\_effects, labels = names(abs\_effects), pos = 3, cex = 0.7)  
 abline(lm(abs\_effects ~ quantiles), col = "blue", lty = 2)  
}  
  
# output the summary and the plots  
summary(model\_folded) # Check model estimates

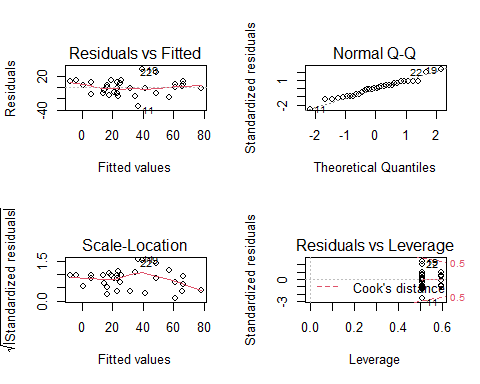
##   
## Call:  
## lm(formula = avg\_readout ~ A + B + C + D + E + A:B + A:C + A:D +   
## A:E + B:C + B:D + B:E + C:D + C:E + D:E + replicate, data = folded\_data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -32.903 -10.049 0.492 9.610 30.991   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 31.3393 3.3339 9.400 1.12e-07 \*\*\*  
## A -1.8256 3.3339 -0.548 0.59203   
## B -3.5074 3.3339 -1.052 0.30943   
## C -3.5060 3.3339 -1.052 0.30962   
## D -12.2452 3.3339 -3.673 0.00226 \*\*   
## E -5.0442 3.3339 -1.513 0.15106   
## replicate -8.8802 3.8497 -2.307 0.03576 \*   
## A:B -0.5664 3.8497 -0.147 0.88498   
## A:C 1.6112 3.3339 0.483 0.63587   
## A:D 4.9725 3.3339 1.492 0.15656   
## A:E 1.0932 3.3339 0.328 0.74752   
## B:C 6.0404 3.3339 1.812 0.09008 .   
## B:D 2.9756 3.3339 0.893 0.38620   
## B:E -0.4793 3.3339 -0.144 0.88759   
## C:D -5.6911 3.3339 -1.707 0.10842   
## C:E -2.3173 3.3339 -0.695 0.49763   
## D:E 10.1721 3.3339 3.051 0.00809 \*\*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 18.86 on 15 degrees of freedom  
## Multiple R-squared: 0.7472, Adjusted R-squared: 0.4775   
## F-statistic: 2.771 on 16 and 15 DF, p-value: 0.02755

custom\_half\_normal(model\_folded, "Foldover 32-Run Model") # Plot effect significance

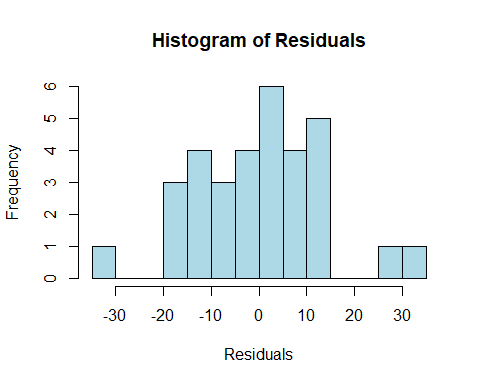


Running some model diagnostic checks

# Plot standard diagnostic plots starting with the basic 4 plots  
par(mfrow = c(2, 2))  
plot(model\_folded)



# and some additional checks for normality  
# Histogram of residuals  
par(mfrow = c(1, 1)) # Reset layout  
hist(residuals(model\_folded), main = "Histogram of Residuals",  
 xlab = "Residuals", col = "lightblue", breaks = 10)



# Shapiro-Wilk test  
shapiro.test(residuals(model\_folded))

##   
## Shapiro-Wilk normality test  
##   
## data: residuals(model\_folded)  
## W = 0.97519, p-value = 0.6529

## 2. Part 2(Bonus marks): Analysis to replicate estimates a,b, and c in Table III using all 34 runs (16-run and 18-run designs combined as they did in the paper)

I. Estimates in Table III (a) with y=readout and all runs included

# we use the average readout, so y= readout  
drugs\_a <- drugdata %>%  
 mutate(avg\_readout = (Replicate1 + Replicate2) / 2,  
 replicate = rep(c(-1, 1), each = 17))  
  
# Fit full second-order model using raw response  
model\_a <- lm(avg\_readout ~ A + B + C + D + E +  
 I(A^2) + I(B^2) + I(C^2) + I(D^2) + I(E^2) +  
 A:B + A:C + A:D + A:E +  
 B:C + B:D + B:E +  
 C:D + C:E + D:E +  
 replicate,  
 data = drugs\_a)  
  
# Summary of model (estimates should match Table III(a))  
summary(model\_a)

##   
## Call:  
## lm(formula = avg\_readout ~ A + B + C + D + E + I(A^2) + I(B^2) +   
## I(C^2) + I(D^2) + I(E^2) + A:B + A:C + A:D + A:E + B:C +   
## B:D + B:E + C:D + C:E + D:E + replicate, data = drugs\_a)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -7.1366 -2.6036 -0.8743 2.4686 8.4523   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 17.87589 6.76783 2.641 0.02152 \*   
## A -1.59155 1.29155 -1.232 0.24144   
## B -2.35092 1.29155 -1.820 0.09374 .   
## C -2.00099 1.29155 -1.549 0.14727   
## D -19.57614 1.29155 -15.157 3.45e-09 \*\*\*  
## E -13.93498 1.29155 -10.789 1.57e-07 \*\*\*  
## I(A^2) 3.52854 3.38347 1.043 0.31756   
## I(B^2) 0.04055 3.38347 0.012 0.99063   
## I(C^2) -1.82319 3.38347 -0.539 0.59984   
## I(D^2) -7.33321 3.38347 -2.167 0.05103 .   
## I(E^2) 15.27194 3.38347 4.514 0.00071 \*\*\*  
## replicate -0.17232 1.83306 -0.094 0.92665   
## A:B 0.71654 1.40509 0.510 0.61933   
## A:C 2.33747 1.40509 1.664 0.12207   
## A:D 1.46124 1.40502 1.040 0.31883   
## A:E -1.32398 1.40502 -0.942 0.36461   
## B:C 1.63732 1.40502 1.165 0.26652   
## B:D -0.27370 1.40509 -0.195 0.84881   
## B:E 1.30919 1.40502 0.932 0.36981   
## C:D -0.65668 1.40502 -0.467 0.64860   
## C:E 0.11626 1.40509 0.083 0.93542   
## D:E 9.56342 1.40509 6.806 1.89e-05 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 6.504 on 12 degrees of freedom  
## Multiple R-squared: 0.9762, Adjusted R-squared: 0.9345   
## F-statistic: 23.42 on 21 and 12 DF, p-value: 9.551e-07

1. Estimates in Table III (b) with y= sqrt(readout) after transformation and all runs included

# use the data but with y= sqrt(readout)  
drugs\_b <- drugdata %>%  
 mutate(avg\_readout = (Replicate1 + Replicate2) / 2,  
 sqrt\_readout = sqrt(avg\_readout),  
 replicate = rep(c(-1, 1), each = 17))  
  
# fit model using sqrt(readout)  
model\_b <- lm(sqrt\_readout ~ A + B + C + D + E +  
 I(A^2) + I(B^2) + I(C^2) + I(D^2) + I(E^2) +  
 A:B + A:C + A:D + A:E +  
 B:C + B:D + B:E +  
 C:D + C:E + D:E +  
 replicate,  
 data = drugs\_b)  
  
# Summary of model (estimates should match Table III(b))  
summary(model\_b)

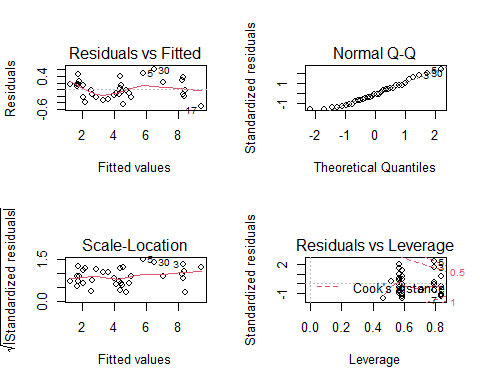
##   
## Call:  
## lm(formula = sqrt\_readout ~ A + B + C + D + E + I(A^2) + I(B^2) +   
## I(C^2) + I(D^2) + I(E^2) + A:B + A:C + A:D + A:E + B:C +   
## B:D + B:E + C:D + C:E + D:E + replicate, data = drugs\_b)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.70275 -0.18779 -0.04771 0.17332 0.77122   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 4.045619 0.568363 7.118 1.22e-05 \*\*\*  
## A -0.133744 0.108464 -1.233 0.241159   
## B -0.230536 0.108464 -2.125 0.054998 .   
## C -0.200151 0.108464 -1.845 0.089798 .   
## D -2.059020 0.108464 -18.983 2.56e-10 \*\*\*  
## E -1.217984 0.108464 -11.229 1.01e-07 \*\*\*  
## I(A^2) 0.244341 0.284144 0.860 0.406686   
## I(B^2) 0.074547 0.284144 0.262 0.797497   
## I(C^2) -0.001232 0.284144 -0.004 0.996612   
## I(D^2) -1.172732 0.284144 -4.127 0.001402 \*\*   
## I(E^2) 1.393339 0.284144 4.904 0.000364 \*\*\*  
## replicate -0.012681 0.153941 -0.082 0.935707   
## A:B 0.124954 0.118000 1.059 0.310481   
## A:C 0.259862 0.118000 2.202 0.047945 \*   
## A:D 0.072360 0.117994 0.613 0.551156   
## A:E -0.130754 0.117994 -1.108 0.289515   
## B:C 0.142028 0.117994 1.204 0.251915   
## B:D -0.093080 0.118000 -0.789 0.445524   
## B:E 0.129679 0.117994 1.099 0.293313   
## C:D -0.105543 0.117994 -0.894 0.388652   
## C:E 0.053797 0.118000 0.456 0.656596   
## D:E 0.535907 0.118000 4.542 0.000676 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.5462 on 12 degrees of freedom  
## Multiple R-squared: 0.9815, Adjusted R-squared: 0.9492   
## F-statistic: 30.37 on 21 and 12 DF, p-value: 2.17e-07

1. Estimates in Table III (c) with y= sqrt(readout) and Replicate 1 of run 14 removed

# Clean the dataset - remove replicate 1 of run 14 and use sqrt(readout)  
drugs\_c <- drugdata %>%  
 mutate(Replicate1 = ifelse(Run == 14, NA, Replicate1)) %>%  
 mutate(avg\_readout = rowMeans(select(., Replicate1, Replicate2), na.rm = TRUE),  
 sqrt\_readout = sqrt(avg\_readout),  
 replicate = rep(c(-1, 1), each = 17))  
  
# fit the full second-order model  
model\_c <- lm(sqrt\_readout ~ A + B + C + D + E +  
 I(A^2) + I(B^2) + I(C^2) + I(D^2) + I(E^2) +  
 A:B + A:C + A:D + A:E +  
 B:C + B:D + B:E +  
 C:D + C:E + D:E +  
 replicate,  
 data = drugs\_c)  
  
# View model summary (same estimates as Table III(c))  
summary(model\_c)

##   
## Call:  
## lm(formula = sqrt\_readout ~ A + B + C + D + E + I(A^2) + I(B^2) +   
## I(C^2) + I(D^2) + I(E^2) + A:B + A:C + A:D + A:E + B:C +   
## B:D + B:E + C:D + C:E + D:E + replicate, data = drugs\_c)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.5020 -0.1979 -0.0292 0.1546 0.5969   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 4.02584 0.47738 8.433 2.18e-06 \*\*\*  
## A -0.10962 0.09110 -1.203 0.252052   
## B -0.26608 0.09110 -2.921 0.012823 \*   
## C -0.22067 0.09110 -2.422 0.032183 \*   
## D -2.03490 0.09110 -22.337 3.83e-11 \*\*\*  
## E -1.23851 0.09110 -13.595 1.19e-08 \*\*\*  
## I(A^2) 0.25507 0.23866 1.069 0.306197   
## I(B^2) 0.06670 0.23866 0.279 0.784649   
## I(C^2) -0.01213 0.23866 -0.051 0.960285   
## I(D^2) -1.16200 0.23866 -4.869 0.000386 \*\*\*  
## I(E^2) 1.38244 0.23866 5.793 8.57e-05 \*\*\*  
## replicate 0.01778 0.12930 0.137 0.892924   
## A:B 0.16523 0.09911 1.667 0.121348   
## A:C 0.28653 0.09911 2.891 0.013552 \*   
## A:D 0.04496 0.09911 0.454 0.658203   
## A:E -0.09556 0.09911 -0.964 0.353968   
## B:C 0.10195 0.09911 1.029 0.323922   
## B:D -0.05280 0.09911 -0.533 0.603937   
## B:E 0.08960 0.09911 0.904 0.383750   
## C:D -0.07035 0.09911 -0.710 0.491375   
## C:E 0.02021 0.09911 0.204 0.841847   
## D:E 0.56257 0.09911 5.676 0.000103 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.4588 on 12 degrees of freedom  
## Multiple R-squared: 0.9869, Adjusted R-squared: 0.9639   
## F-statistic: 42.96 on 21 and 12 DF, p-value: 2.923e-08

# Diagnostic plots to check model assumptions  
par(mfrow = c(2, 2))  
plot(model\_c)



Performing variable selection via stepwise regression to ensure there are no further significant effects before we re-plot the diagnostic plots to check for model assumptions.

library(MASS)

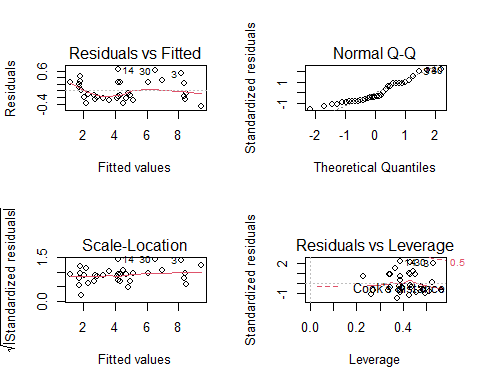
##   
## Attaching package: 'MASS'

## The following object is masked from 'package:dplyr':  
##   
## select

# Stepwise regression to select significant effects  
step\_model <- stepAIC(model\_c, direction = "both", trace = FALSE)  
  
# final model output  
summary(step\_model)

##   
## Call:  
## lm(formula = sqrt\_readout ~ A + B + C + D + E + I(A^2) + I(D^2) +   
## I(E^2) + A:B + A:C + A:E + B:C + D:E, data = drugs\_c)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.4590 -0.2182 -0.1234 0.2554 0.6367   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 4.09019 0.22218 18.409 5.22e-14 \*\*\*  
## A -0.10242 0.07341 -1.395 0.17827   
## B -0.27317 0.07371 -3.706 0.00140 \*\*   
## C -0.21201 0.07360 -2.880 0.00925 \*\*   
## D -2.04714 0.07421 -27.585 < 2e-16 \*\*\*  
## E -1.21788 0.07382 -16.499 4.10e-13 \*\*\*  
## I(A^2) 0.24187 0.17887 1.352 0.19140   
## I(D^2) -1.14291 0.17969 -6.361 3.31e-06 \*\*\*  
## I(E^2) 1.35297 0.17909 7.555 2.79e-07 \*\*\*  
## A:B 0.15727 0.08096 1.943 0.06628 .   
## A:C 0.29607 0.08096 3.657 0.00157 \*\*   
## A:E -0.10175 0.08060 -1.262 0.22133   
## B:C 0.10690 0.08166 1.309 0.20537   
## D:E 0.56397 0.08189 6.887 1.09e-06 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.3821 on 20 degrees of freedom  
## Multiple R-squared: 0.9848, Adjusted R-squared: 0.975   
## F-statistic: 99.81 on 13 and 20 DF, p-value: 2.989e-15

# Diagnostic plots  
par(mfrow = c(2, 2))  
plot(step\_model)



Models a, b, and c outputs are mostly similar to the output in Table 3 of the study, as explained in the MS Word file.