400138679_4qz3_a1

Jeff Suitor

10/1/2021

Import libraries.

```
library("lm.beta")
library("carData")
library("MASS")
library("agricolae")
```

Question 1

Import the data and drop the patient column.

```
setwd("~/code/4qz3-modelling/a1")
q1_data = read.delim("Q1.txt", header=TRUE, "\t")[-c(1)]
```

To calculate which variables are most a linear regression using standardized values should be used. The variables with the largest absolute standardized coefficients are the most impactful on the regression.

```
fit = lm(q1_data$mg.Urinase ~ q1_data$RBC + q1_data$Protein + q1_data$Glucose + q1_data$Specific.Gravity
summary(fit)
```

```
##
## Call:
## lm(formula = q1_data$mg.Urinase ~ q1_data$RBC + q1_data$Protein +
##     q1_data$Glucose + q1_data$Specific.Gravity + q1_data$Bilirubin,
##     data = q1_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -38.386 -16.984  -5.134  14.409  58.262
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)              -399.9123    33.0269 -12.109 2.76e-15 ***
## q1_data$RBC                 3.2571     4.6231   0.705    0.485
## q1_data$Protein            23.4102     3.3856   6.915 1.93e-08 ***
## q1_data$Glucose             2.9242     0.2321  12.598 7.42e-16 ***
## q1_data$Specific.Gravity   10.2713     6.2342   1.648    0.107
## q1_data$Bilirubin          -0.6876    13.2309  -0.052    0.959
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 25.11 on 42 degrees of freedom
```

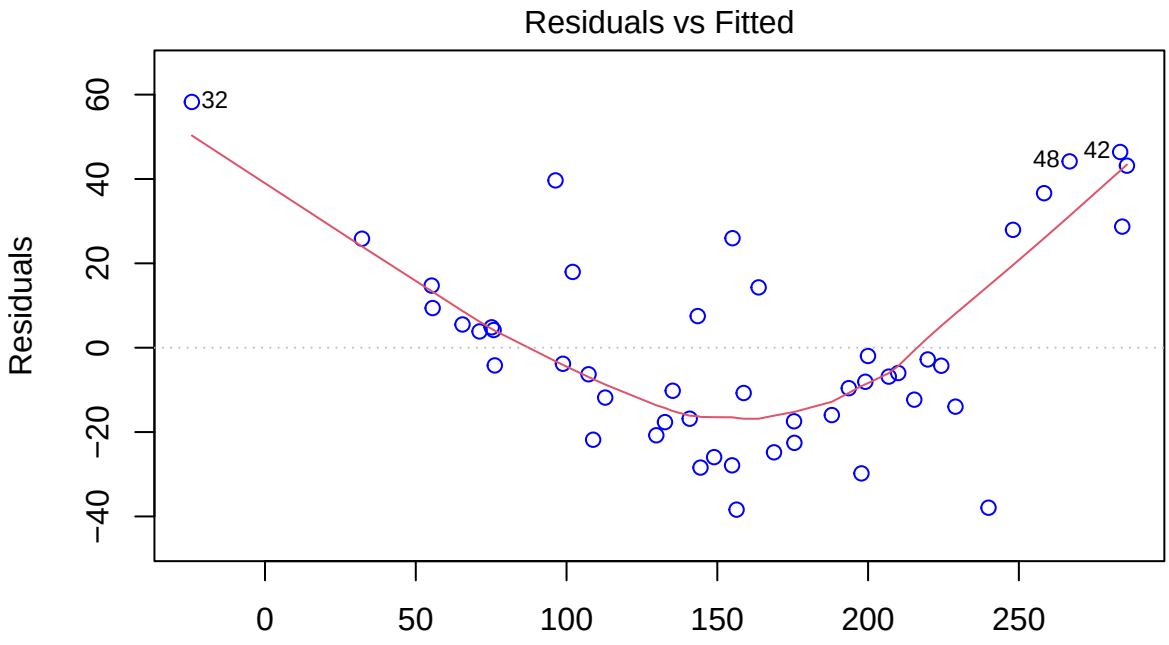
```
## Multiple R-squared:  0.901,  Adjusted R-squared:  0.8892
## F-statistic: 76.44 on 5 and 42 DF,  p-value: < 2.2e-16
```

```
lm.beta(fit)
```

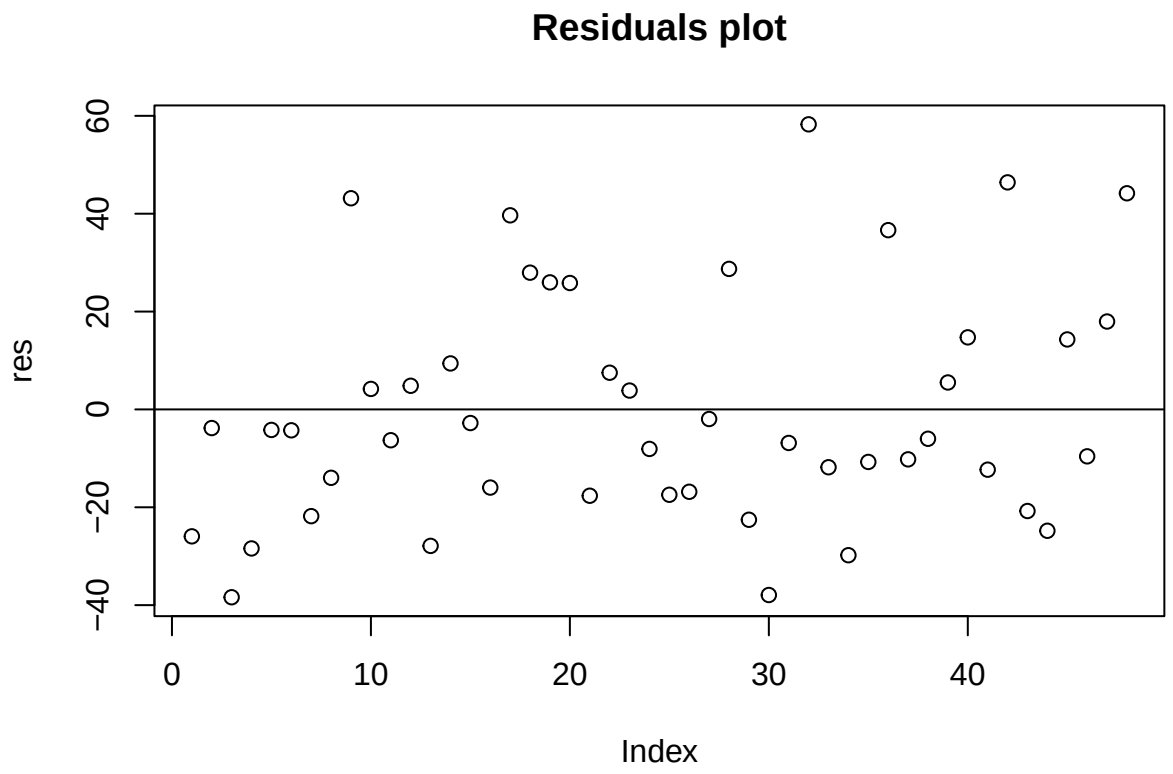
```
##
## Call:
## lm(formula = q1_data$mg.Urinase ~ q1_data$RBC + q1_data$Protein +
##     q1_data$Glucose + q1_data$Specific.Gravity + q1_data$Bilirubin,
##     data = q1_data)
##
## Standardized Coefficients::
##              (Intercept)              q1_data$RBC          q1_data$Protein
##               0.00000000               0.69707775               0.42448209
##          q1_data$Glucose q1_data$Specific.Gravity        q1_data$Bilirubin
##               0.78253853               0.10331061              -0.05151029
```

Therefore, the variables in order of importance from most important to least important is glucose, RBC, urine protein, urine specific gravity, and lastly bilirubin levels.
The final regression model is: $Y = -399.9123 + 3.2571 * RBC + 23.4102 * *Protein + 2.9242 * X_3 * Glucose + 10.2713 * Specific.gravity - 0.6876 * Bilirubin$ Interestingly the RBC, specific gravity and bilirubin have non significant P values indicating that they can likely be eliminated from the model. When examining the standardized coefficients it can be seen that the bilirubin has the least impact on the model and can likely be removed. Due to the P value being below 0.05 the model is significant although it can be further optimized by removing unnecessary parameters.
To analyze the residuals start by first plotting the linear model with residuals. When plotting the residuals there appear to be no appear making a linear regression an appropriate regression time. To get a better sense of the residuals create a Q-Q plot and plot the density of the residuals. The residuals are left shifted showing that there is data which is altering the normal distribution of the data.

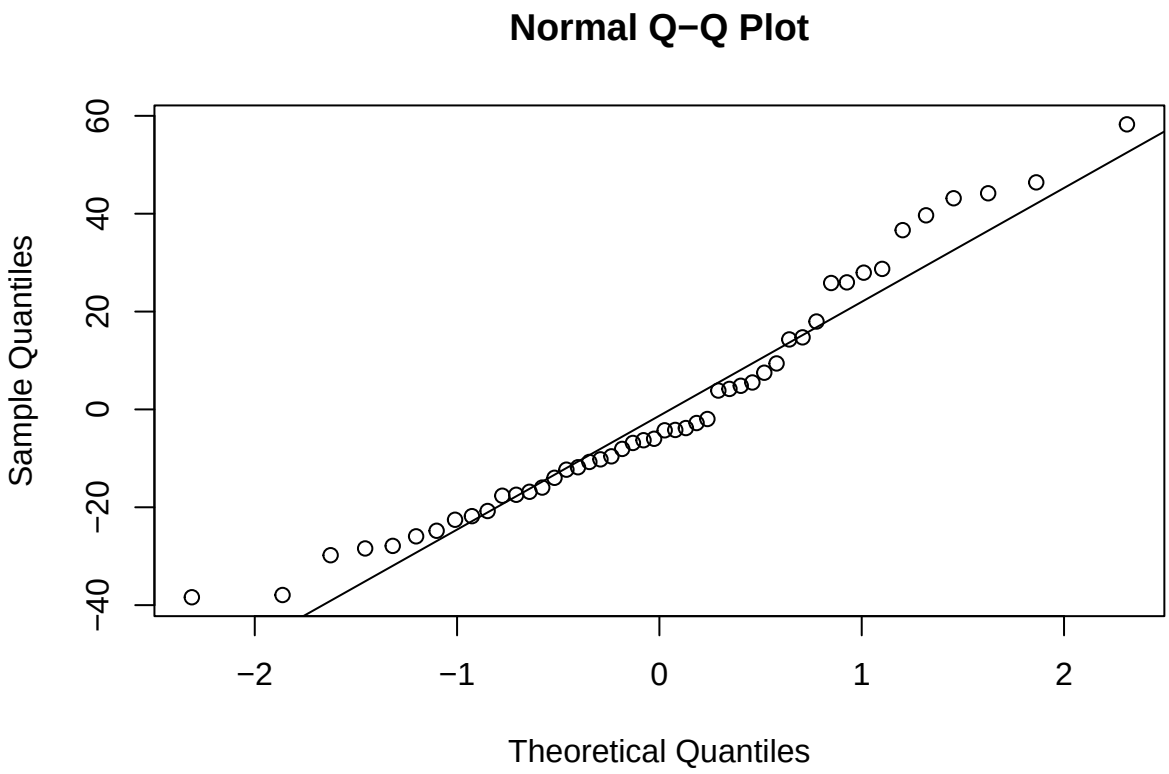
```
plot(fit, which=1, col=c("blue")) # Residuals vs Fitted Plot
```



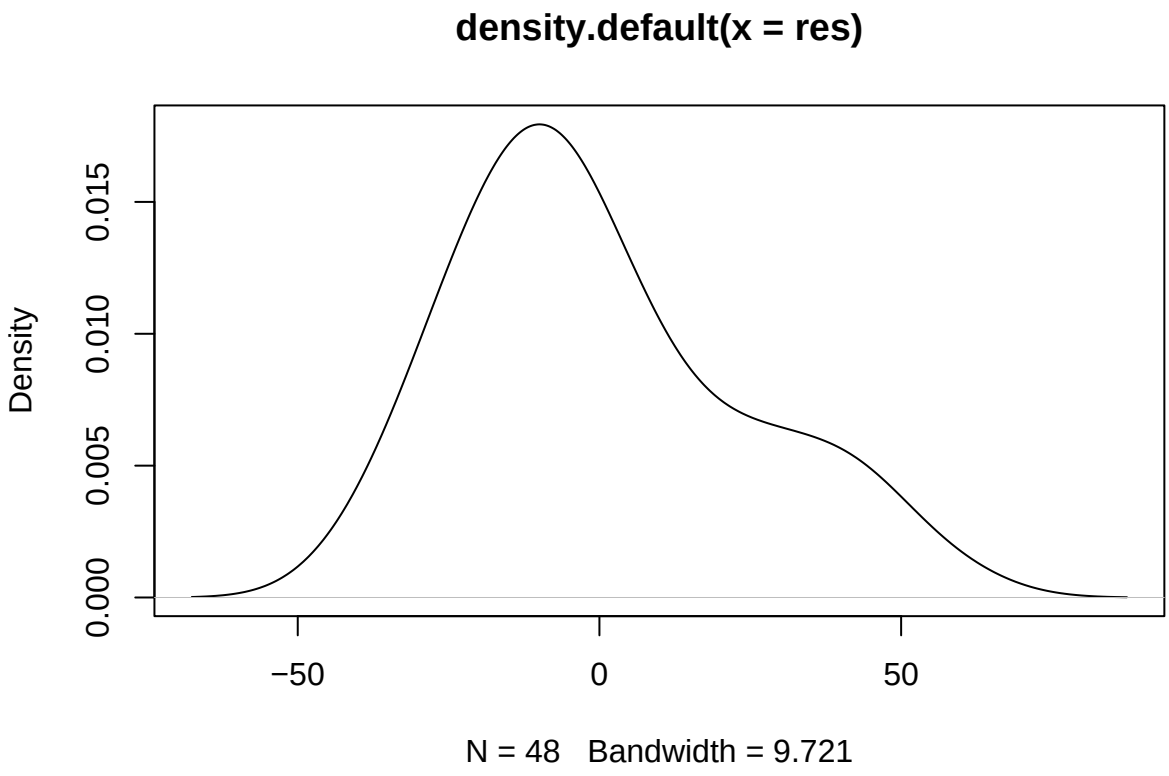

```
res = resid(fit)

plot(res)
qqline(0)
title("Residuals plot")
```




```
qqnorm(res)
qqline(res)
```




```
plot(density(res))
```



Analyzing the results of the ANOVA test it can be seen that both the specific gravity and the bilirubin parameters are not significant for the model due to their P values being greater than 0.05.

```
anova(fit)
```

```
## Analysis of Variance Table
##
## Response: q1_data$mg.Urinase
##                           Df Sum Sq Mean Sq  F value    Pr(>F)
## q1_data$RBC                1  65638   65638 104.0676 6.167e-13 ***
## q1_data$Protein            1  13708   13708  21.7341 3.165e-05 ***
## q1_data$Glucose            1 160013  160013 253.6973 < 2.2e-16 ***
## q1_data$Specific.Gravity   1   1713    1713   2.7156    0.1068
## q1_data$Bilirubin          1      2       2   0.0027    0.9588
## Residuals                 42  26490     631
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Using this information an optimal model was created and plotted.

```
optimal_fit = lm(q1_data$mg.Urinase ~ q1_data$RBC + q1_data$Protein + q1_data$Glucose, data=q1_data)
plot(optimal_fit, which=1, col=c("blue")) # Residuals vs Fitted Plot
```


```
summary(optimal_fit)
```

```
##
## Call:
## lm(formula = q1_data$mg.Urinase ~ q1_data$RBC + q1_data$Protein +
##     q1_data$Glucose, data = q1_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -43.519 -15.095  -9.113  13.803  62.181
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)     -414.7980    30.8755 -13.435  < 2e-16 ***
## q1_data$RBC        3.1496     0.2352  13.390  < 2e-16 ***
## q1_data$Protein   26.4612     2.8415   9.312 5.74e-12 ***
## q1_data$Glucose    3.1240     0.1977  15.799  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 25.32 on 44 degrees of freedom
## Multiple R-squared:  0.8946, Adjusted R-squared:  0.8874
## F-statistic: 124.5 on 3 and 44 DF,  p-value: < 2.2e-16
```

```
anova(optimal_fit)
```

```
## Analysis of Variance Table
```

```
##
## Response: q1_data$mg.Urinase
##                 Df Sum Sq Mean Sq F value    Pr(>F)
## q1_data$RBC      1  65638   65638 102.396 4.632e-13 ***
## q1_data$Protein  1  13708   13708  21.385 3.298e-05 ***
## q1_data$Glucose  1 160013  160013 249.622 < 2.2e-16 ***
## Residuals       44  28205     641
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The model remains significant as it's P value is less than 0.05 and is the best fit as an ANOVA of the optimized model shows that all variables are significant.
The optimal regression is: $Y = -414.7980 + 3.1496 * RBC + 26.4612 * Protein + 3.1240 * Glucose$

Question 2

Import the data

```
q2_data = read.delim("Q2.txt", header=FALSE, "\t")
treatments = c("3d-cast", "plaster", "fiberglass", "fixator", "splint", "semi-rigid", "spongy", "airboo
hospitals = c("H1", "H2", "H3", "H4", "H5", "H6", "H7", "H8")
colnames(q2_data) = treatments
```

A

Create the data frame

```
vect = c(t(as.matrix(q2_data))) # vector of all data
treatment_num = 9 # number of treatment levels
patient_num = 8  # number of patients in each hospital (block size)
hospital_num = 8 # number of hospitals
RCBD_df = data.frame(vect)
colnames(RCBD_df) = c("BMD")
RCBD_df$treatments = gl(treatment_num, 1, treatment_num * patient_num * hospital_num, factor(treatments)
RCBD_df$hospital = gl(hospital_num, patient_num*treatment_num, treatment_num * patient_num * hospital_nu
```

Run the anova.

Hospitals:
N̂ull hypothesis: There is no variation between different hospitals.
Â̂lternative hypothesis: There is a variation between hospitals.

Treatments:
N̂ull hypothesis: There is no variation between different treatments.
Â̂lternative hypothesis: There is a variation between treatments.

```
anova_res = aov(RCBD_df$BMD ~RCBD_df$hospital + RCBD_df$treatments)
summary(anova_res)
```

```
##                      Df Sum Sq Mean Sq F value   Pr(>F)
## RCBD_df$hospital      7   8611  1230.1   1.716 0.102669
## RCBD_df$treatments    8  23223  2902.8   4.049 0.000107 ***
## Residuals           560 401476   716.9
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The anova indicates that the differences between hospitals is not significant because P $>$ 0.05 so we accept the null hypothesis but P is less than 0.05 so we reject the null hypothesis.

```
duncan.test(anova_res, "RCBD_df$hospital", console=TRUE)
```

```
##
## Study: anova_res ~ "RCBD_df$hospital"
##
## Duncan's new multiple range test
## for RCBD_df$BMD
##
## Mean Square Error:  716.9214
##
## RCBD_df$hospital,  means
```

```
##
##    RCBD_df.BMD      std  r      Min      Max
## H1    64.41379 29.36056 72 10.23500 200.3460
## H2    58.72463 26.52487 72 10.69400 154.6570
## H3    66.73124 26.20743 72 10.25360 132.5752
## H4    66.52889 25.22167 72 10.23600 125.6637
## H5    65.35612 27.04276 72 11.25600 125.6637
## H6    69.43830 28.65513 72 18.84956 154.8690
## H7    63.84355 28.47022 72 11.93805 111.2124
## H8    72.81514 27.02068 72 22.36400 121.3580
##
## Alpha: 0.05 ; DF Error: 560
##
## Critical Range
##         2         3         4         5         6         7         8
##  8.765411  9.228255  9.537876  9.766529  9.945542 10.091236 10.213145
##
## Means with the same letter are not significantly different.
##
##    RCBD_df$BMD groups
## H8    72.81514      a
## H6    69.43830      a
## H3    66.73124     ab
## H4    66.52889     ab
## H5    65.35612     ab
## H1    64.41379     ab
## H7    63.84355     ab
## H2    58.72463      b
```

A DMRT confirms that only hospitals 6 and 8 differed from hospital 2. However, there was a larger variance due to treatments as can be seen when performing a DMRT on the treatments.

```
duncan.test(anova_res, "RCBD_df$treatments", console=TRUE)
```

```
##
## Study: anova_res ~ "RCBD_df$treatments"
##
## Duncan's new multiple range test
## for RCBD_df$BMD
##
## Mean Square Error:  716.9214
##
## RCBD_df$treatments,  means
##
##            RCBD_df.BMD      std  r      Min      Max
## 3d-cast       67.88388 31.60247 64 11.93805 154.6570
## 3d-splint     60.33121 21.50167 64 21.56900 111.2124
## airboot       68.89927 27.82098 64 21.35700 109.3274
## fiberglass    66.77035 25.89123 64 23.24779 134.7743
## fixator       51.60940 24.91496 64 10.23500 125.6390
## plaster       70.21071 25.53377 64 22.56900 109.6416
## semi-rigid    70.29001 25.75067 64 22.36570 154.8690
## splint        74.30003 26.06295 64 10.23600 116.8672
## spongy        63.53825 31.44530 64 11.25600 200.3460
```

```
##
## Alpha: 0.05 ; DF Error: 560
##
## Critical Range
##         2         3         4         5         6         7         8         9
##  9.297122  9.788043 10.116445 10.358968 10.548840 10.703373 10.832676 10.943148
##
## Means with the same letter are not significantly different.
##
##            RCBD_df$BMD groups
## splint        74.30003      a
## semi-rigid    70.29001     ab
## plaster       70.21071     ab
## airboot       68.89927     ab
## 3d-cast       67.88388     ab
## fiberglass    66.77035     ab
## spongy        63.53825      b
## 3d-splint     60.33121     bc
## fixator       51.60940      c
```

From the DMRT we can see that following groups of treatments have no significant differences:

- external fixator, 3D printed splint
- 3D printed splint, spongy cast, fiberglass cast, 3D printed case, Airboot, plaster cast, semi-rigid external fixator
- fiberglass cast, 3D printed case, Airboot, plaster cast, semi-rigid external fixator, simple splinting

The variance in this design is coming from the difference in treatments with minimal impact from which hospitals performed the procedures.

B

Setup the data frame

```
scanners = c("D1", "D2", "D3", "D4", "D5", "D6", "D7", "D8")
RCBD_df$scanners = gl(hospital_num, treatment_num, treatment_num * patient_num * hospital_num, factor(s
```

Run the anova. Scanners:
N̂ull hypothesis: There is no variation between DEXA scanners.
Âlternative hypothesis: There is a variation between DEXA scanners.

```
anova_dexa_res = aov(RCBD_df$BMD ~ RCBD_df$scanners + RCBD_df$treatments)
summary(anova_dexa_res)
```

```
##                      Df Sum Sq Mean Sq F value  Pr(>F)
## RCBD_df$scanners      7   2341   334.5   0.459 0.86387
## RCBD_df$treatments    8  23223  2902.8   3.987 0.00013 ***
## Residuals           560 407745   728.1
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Run a DMRT to examine the variance in scanners.

```
duncan.test(anova_dexa_res, "RCBD_df$scanners", console=TRUE)
```

```
##
## Study: anova_dexa_res ~ "RCBD_df$scanners"
##
## Duncan's new multiple range test
## for RCBD_df$BMD
##
## Mean Square Error:  728.1166
##
## RCBD_df$scanners,  means
##
##    RCBD_df.BMD      std  r      Min      Max
## D1    64.48069 23.90844 72 10.23500 109.3274
## D2    62.94695 26.21584 72 16.33628 132.5752
## D3    63.69626 25.36210 72 10.69400 120.6372
## D4    68.94173 27.29291 72 11.93805 125.6637
## D5    65.47290 28.25873 72 18.84956 119.6947
## D6    67.21539 29.01403 72 10.25360 154.6570
## D7    67.98342 34.38304 72 10.23600 200.3460
## D8    67.11431 24.50279 72 22.63500 109.3274
##
## Alpha: 0.05 ; DF Error: 560
##
## Critical Range
##         2         3         4         5         6         7         8
##  8.833584  9.300029  9.612058  9.842488 10.022894 10.169722 10.292578
##
## Means with the same letter are not significantly different.
##
##    RCBD_df$BMD groups
## D4    68.94173      a
## D7    67.98342      a
## D6    67.21539      a
```

```
## D8    67.11431      a
## D5    65.47290      a
## D1    64.48069      a
## D3    63.69626      a
## D2    62.94695      a
```

The DMRT of the dexa scanners confirms that there is no significant variance between them. Run a DMRT to see if there is any change in the differences between treatments.

```
duncan.test(anova_dexa_res, "RCBD_df$treatments", console=TRUE)
```

```
##
## Study: anova_dexa_res ~ "RCBD_df$treatments"
##
## Duncan's new multiple range test
## for RCBD_df$BMD
##
## Mean Square Error:  728.1166
##
## RCBD_df$treatments,  means
##
##            RCBD_df.BMD      std  r      Min      Max
## 3d-cast       67.88388 31.60247 64 11.93805 154.6570
## 3d-splint     60.33121 21.50167 64 21.56900 111.2124
## airboot       68.89927 27.82098 64 21.35700 109.3274
## fiberglass    66.77035 25.89123 64 23.24779 134.7743
## fixator       51.60940 24.91496 64 10.23500 125.6390
## plaster       70.21071 25.53377 64 22.56900 109.6416
## semi-rigid    70.29001 25.75067 64 22.36570 154.8690
## splint        74.30003 26.06295 64 10.23600 116.8672
## spongy        63.53825 31.44530 64 11.25600 200.3460
##
## Alpha: 0.05 ; DF Error: 560
##
## Critical Range
##        2        3        4        5        6        7        8        9
##  9.369431  9.864170 10.195127 10.439535 10.630885 10.786619 10.916928 11.028259
##
## Means with the same letter are not significantly different.
##
##            RCBD_df$BMD groups
## splint        74.30003      a
## semi-rigid    70.29001     ab
## plaster       70.21071     ab
## airboot       68.89927     ab
## 3d-cast       67.88388     ab
## fiberglass    66.77035     ab
## spongy        63.53825     ab
## 3d-splint     60.33121     bc
## fixator       51.60940      c
```

From the DMRT we can see that following groups of treatments have no significant differences:

- external fixator, 3D printed splint
- 3D printed splint, spongy cast, fiberglass cast, 3D printed case, Airboot, plaster cast, semi-rigid external fixator

- spongy cast, fiberglass cast, 3D printed case, Airboot, plaster cast, semi-rigid external fixator, simple splinting

The spongy cast is now no longer considered different from the fiberglass cast, 3D printed case, Airboot, plaster cast, semi-rigid external fixator, or simple splinting.

The the key difference between the approach in A and the approach in B is that A used a blocking approaching which reduces experimental error due to the blocking methodology while B used a CRD with subsamples.

All code shown throughout this assignment is the only code that was used.