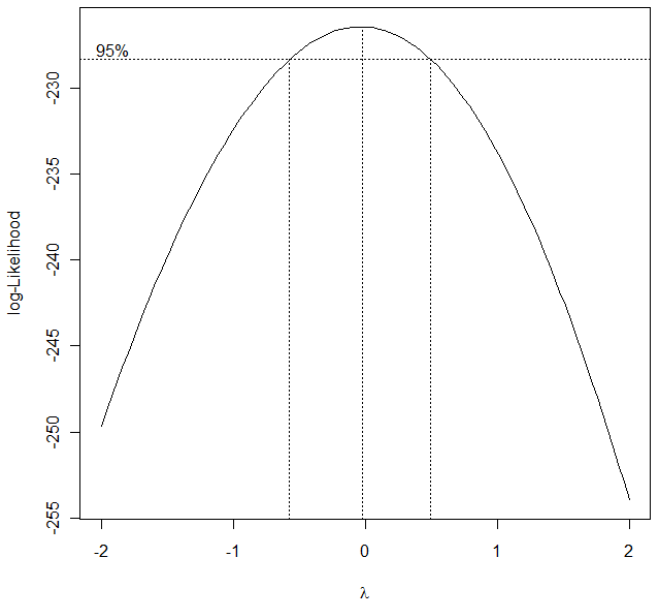
STAT 425 - Homework #6

PROBLEM 1

```
> ##########################
> # STAT 425
> # Homework 6
> # Date: 11/18/11
> # Author: Luis Lin
> ##########################
>
> library(faraway)
> library(MASS)
>
> ############################################################
> ###PROBLEM 1
>
> # Data :
> # A data frame with 54 observations on 3 variables.
> #[,1]  breaks  numeric  The number of breaks
> #[,2]  wool  factor  The type of wool (A or B)
> #[,3]  tension  factor  The level of tension (L, M, H)
```

Part a

```
> ### Part a: Use the Box-Cox method to determine an appropriate
transformation on the re-sponse.
>
> attach(warpbreaks)
The following object(s) are masked from 'warpbreaks (position 5)':

    breaks, tension, wool
> names(warpbreaks)
[1] "breaks"  "wool"    "tension"
> g=lm(breaks ~ wool*tension)
> breaks.trans=boxcox(g, lambda=seq(-2, 2, length=400))
> round(breaks.trans$x[breaks.trans$y == max(breaks.trans$y)],3)
[1] -0.035
> tmp=breaks.trans$x[breaks.trans$y > max(breaks.trans$y) - qchisq(0.95,
1)/2];
> CI=range(tmp) # 95% CI.
> round(CI,3)
[1] -0.566  0.496
> 1>CI[1] & 1<CI[2] # Check contains 1
[1] FALSE
> 0>CI[1] & 0<CI[2] # Check contains 0
[1] TRUE
>
> # Since 1 is not contained in the CI, transformation needed
> # Since 0 is contained in the CI, choose natural log transformation
```



Part b

```
> ### Part b: Determine which factors (including interactions) are
signicant.
>
> newbreaks = log(breaks)
> g=lm(newbreaks ~ wool*tension)
> anova(g)
Analysis of Variance Table

Response: newbreaks
             Df Sum Sq Mean Sq F value   Pr(>F)
wool          1 0.3125 0.31253  2.2344 0.141511
tension       2 2.1762 1.08808  7.7792 0.001185 **
wool:tension  2 0.9131 0.45657  3.2642 0.046863 *
Residuals    48 6.7138 0.13987
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
>
> # Significant factors: p-value < 0.05
> # The Main effect tension and interaction wool-tension are significant.
> # The main effect wool is not significant.
```

Part c

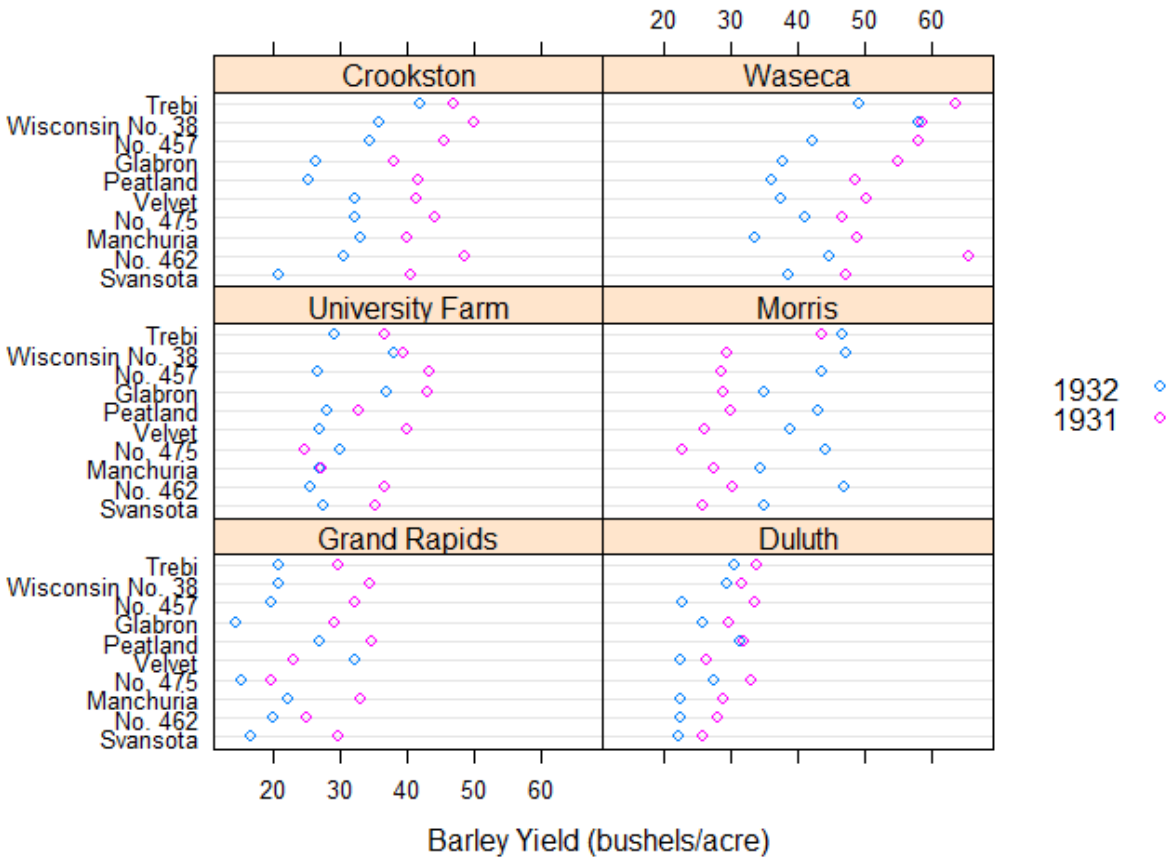
```
> ### Part c: Now form a six-level factor from all combinations of the
wool and tension factors.
> ### Which combinations|there are totally 15 pairs are signicantly
different?
>
> tky = TukeyHSD(aov(g)); tky
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = g)

$wool
          diff        lwr        upr     p adj
B-A -0.1521536 -0.3568127 0.05250558 0.1415114

$tension
          diff        lwr         upr     p adj
M-L -0.2871237 -0.5886239  0.01437636 0.0649432
H-L -0.4892747 -0.7907748 -0.18777463 0.0007948
H-M -0.2021510 -0.5036511  0.09934912 0.2465032

$`wool:tension`
                  diff        lwr         upr     p adj
B:L-A:L -0.4355668365 -0.9588143  0.08768059 0.1534713
A:M-A:L -0.6011957092 -1.1244431 -0.07794828 0.0157632
B:M-A:L -0.4086186238 -0.9318661  0.11462881 0.2071300
A:H-A:L -0.6003226799 -1.1235701 -0.07707525 0.0159794
B:H-A:L -0.8137936293 -1.3370411 -0.29054620 0.0004035
A:M-B:L -0.1656288727 -0.6888763  0.35761856 0.9341161
B:M-B:L  0.0269482127 -0.4962992  0.55019564 0.9999876
A:H-B:L -0.1647558434 -0.6880033  0.35849159 0.9354994
B:H-B:L -0.3782267929 -0.9014742  0.14502064 0.2822984
B:M-A:M  0.1925770854 -0.3306703  0.71582452 0.8820529
A:H-A:M  0.0008730293 -0.5223744  0.52412046 1.0000000
B:H-A:M -0.2125979201 -0.7358454  0.31064951 0.8318529
A:H-B:M -0.1917040562 -0.7149515  0.33154337 0.8840239
B:H-B:M -0.4051750056 -0.9284224  0.11807242 0.2148594
B:H-A:H -0.2134709494 -0.7367184  0.30977648 0.8294547

> names(tky)
[1] "wool"         "tension"      "wool:tension"
> comb.names = row.names(tky$`wool:tension`);
> sig=tky$`wool:tension`[,4]<0.05
> comb.names[sig]
[1] "A:M-A:L" "A:H-A:L" "B:H-A:L"
>
> # Significant different pairs: Look at p-value <0.05 or CI that doesn't
include zero
> # Combinations of wool and tension "A:M-A:L" , "A:H-A:L",  "B:H-A:L" are
significantly different
```

PROBLEM 2

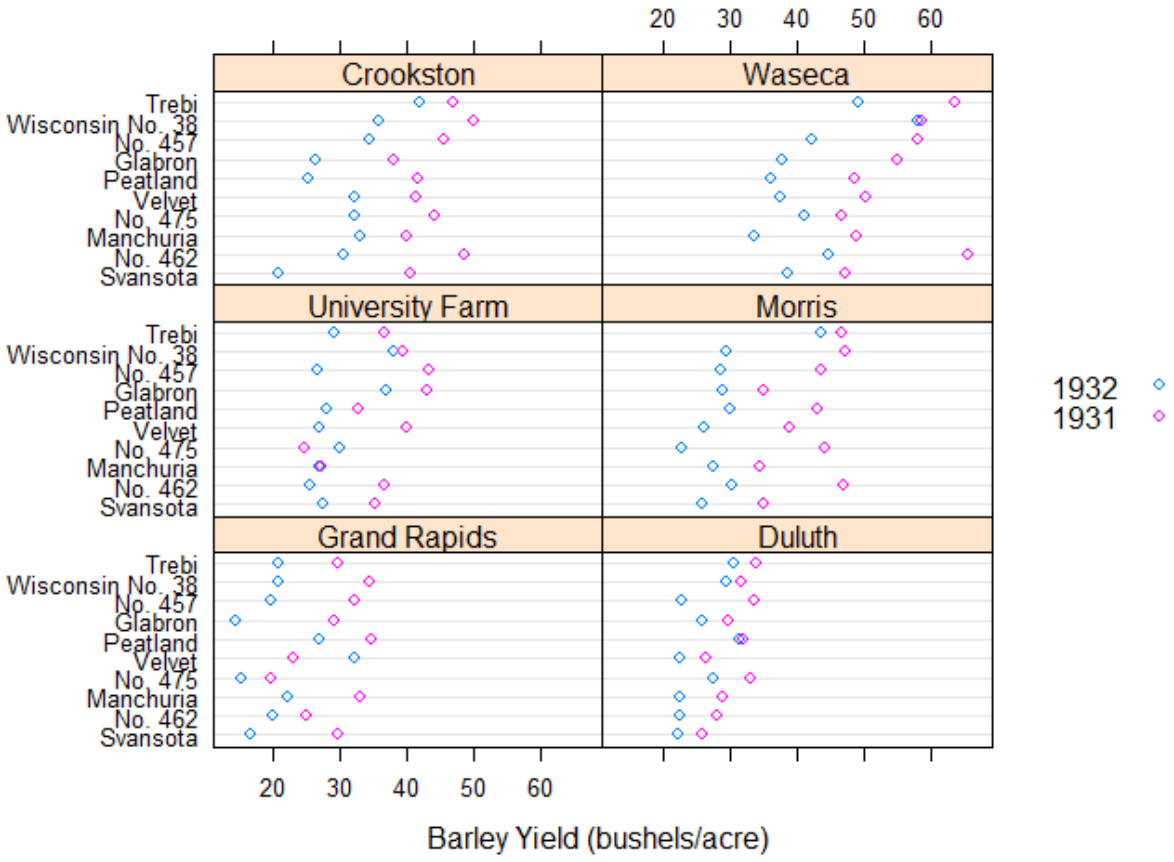
```
> ################################################################
> ###PROBLEM 2
>
> library(lattice)
> names(barley)
[1] "yield"   "variety" "year"    "site"
> ?barley
>
```

Part a

```
> ### Part a: Provide a graphical display of the data
>
> dotplot(variety ~ yield | site, data = barley, groups = year,  key =
simpleKey(levels(barley$year), space = "right"),
+ xlab = "Barley Yield (bushels/acre) ", aspect=0.5, layout = c(2,3),
ylab=NULL)
>
```



```
> newbarley=barley
>
newbarley$year[newbarley$site=="Morris"]=ifelse(newbarley$year[newbarley$s
ite=="Morris"]==1931, 1932, 1931)
>
> dotplot(variety ~ yield | site, data = newbarley, groups = year, key =
simpleKey(levels(newbarley$year), space = "right"),
+ xlab = "Barley Yield (bushels/acre) ", aspect=0.5, layout = c(2,3),
ylab=NULL, font=0.5)
>
```



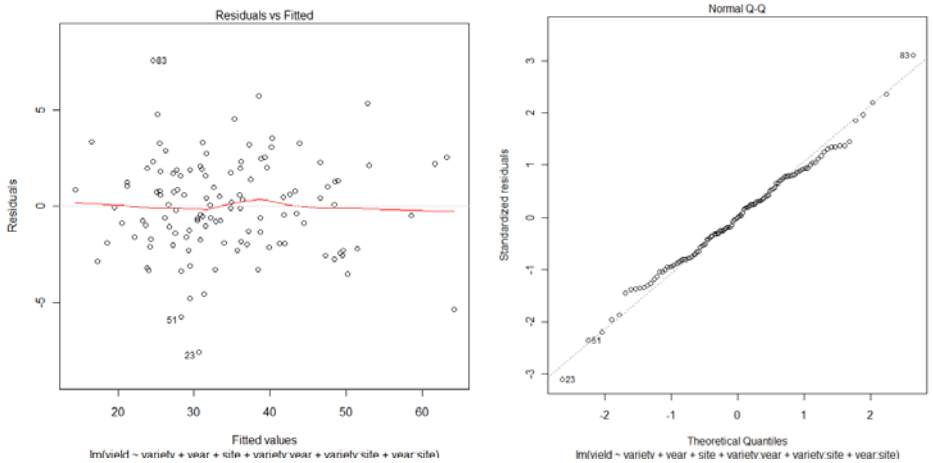
Part b

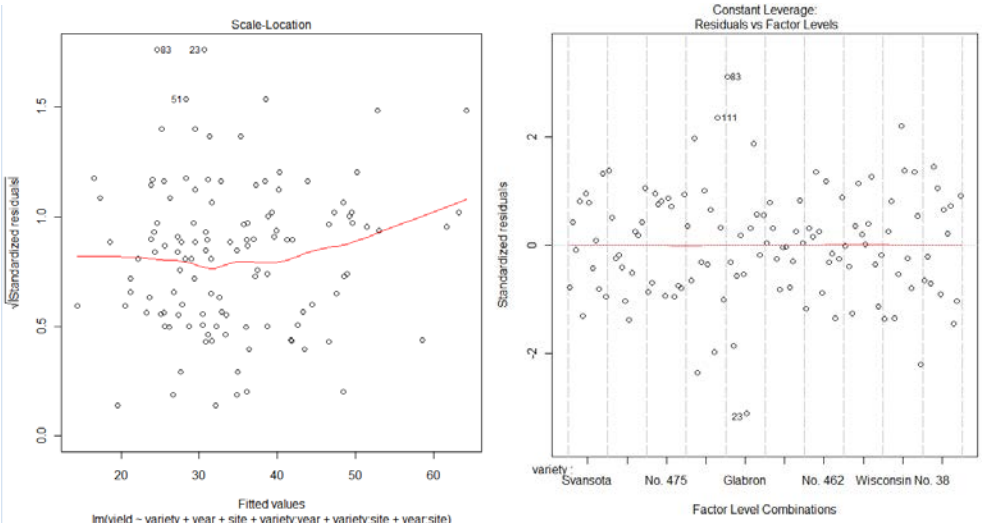
```
> ### Part b: Perform a three-way ANOVA with yield as the response.
Include all the two-way
> ### interactions, but no three-way interactions since we have only one
observation in
> ### each three-way combination. Provide the ANOVA table, and comment on
your result.
>
> g=lm(yield~variety+year+site+variety:year+variety:site+year:site,
data=newbarley)
> #g=lm(yield~(variety+year+site)^2, data=newbarley) # alternative command
>
```

```
> anova(g)
Analysis of Variance Table

Response: yield
             Df Sum Sq Mean Sq  F value    Pr(>F)
variety       9 1052.6  116.95   7.3737 1.637e-06 ***
year          1 2645.2 2645.16 166.7733 < 2.2e-16 ***
site          5 6633.9 1326.77  83.6508 < 2.2e-16 ***
variety:year  9  154.5   17.17   1.0823  0.394275
variety:site 45 1205.8   26.79   1.6894  0.041002 *
year:site     5  304.4   60.87   3.8378  0.005569 **
Residuals    45  713.7   15.86
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> round(summary(g)$r.sq,3)
[1] 0.944
>
> # Significant factors: p-value < 0.05
> # Significant factors: main effects "variety", "year", "site", and
> # intercations "variety:site", "year:site"
>
```

Part c

```
> ### Part c: Check the diagnostics | you will find that two points, the
23rd and the 83rd
> ### samples, stick out. They are the two with the highest residuals (in
absolute value).
> ### Check these two points. Which site or sites are they from? Which
year or years are they from?
>
> plot(g)
Waiting to confirm page change...
```






```
>
> # Samples 23 and 83 have the highest residuals (in absolute value)
> # No evidence of non-constant variance
> # QQ shows normality of residuals, excluding samples 23 and 83
>
> newbarley[c(23, 83),]
      yield variety year         site
23 23.03333  Velvet 1931 Grand Rapids
83 32.23333  Velvet 1932 Grand Rapids
>
> # Sample 23: Variety= velvet, Site= Grand Rapids, Year=1931
> # Sample 83: Variety= vevvet, Site= Grand Rapids, Year=1932
>
```

Part d

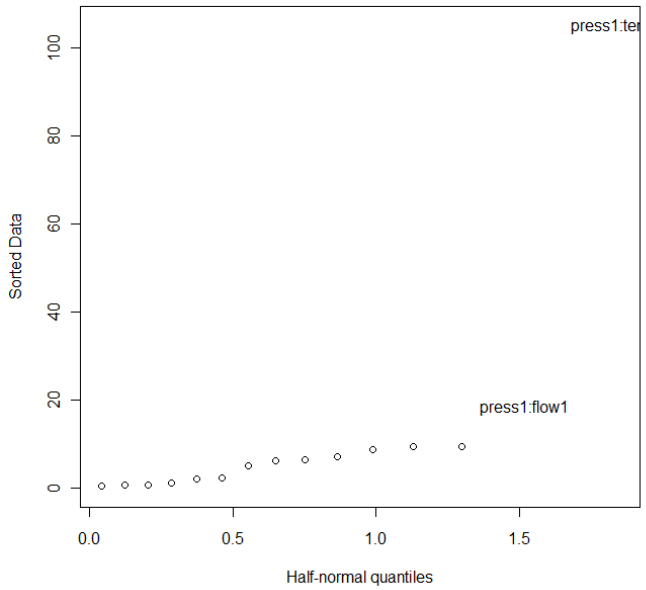
```
> ### Part d: We suspect these two cases, the 23rd and the 83rd samples,
were also switched.
> ### Can you find evidence from the the graphical display produced in
(a)? Switch the
> ### two cases, repeat the analysis, and comment on your results.
>
> # Look at second plot of part a.
> # Yields from 1931 are higher than that of 1932 for all variety across
all sites,
> # with the exception of variety No. 475 in University Farm, variety
velvet in Grand Rapids.
> # For velvet variety, yields from 1931 are higher than that of 1932
across all sites
> # with the exception of Grand Rapids.Thus, this is evidence that the
samples
> # from 1931 and 1932 were possibly switched.
>
```

```
> newbarley[23,3]=1932; newbarley[83,3]=1931;
>
> g=lm(yield~variety+year+site+variety:year+variety:site+year:site,
data=newbarley)
> anova(g)
Analysis of Variance Table

Response: yield
             Df Sum Sq Mean Sq  F value    Pr(>F)
variety       9 1052.6  116.95   9.3632 7.656e-08 ***
year          1 2820.8 2820.76 225.8292 < 2.2e-16 ***
site          5 6633.9 1326.77 106.2209 < 2.2e-16 ***
variety:year  9  137.5   15.28   1.2235  0.304992
variety:site 45 1205.8   26.79   2.1452  0.005935 **
year:site     5  297.4   59.47   4.7613  0.001413 **
Residuals    45  562.1   12.49
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> round(summary(g)$r.sq,3)
[1] 0.956
>
> # Significant factors: p-value < 0.05
> # Significant factors: main effects "variety", "year", "site", and
> # intercations "variety:site", "year:site"
> # The significant factors stay the same
> # The p-value of variety:site seems to have a large decrease
> # (from 0.041 to 0.0059, meaning now significant at a 0.01 level)
> # R-square increased from 0.944 to 0.956.
>
>
```

PROBLEM 3

```
############################################################
> ###PROBLEM 3
>
> ### The peanut data come from a factional factorial experiment to
> ### investigate factors that affect an industrial process using carbon
dioxide to extract oil
> ### from peanuts.
> ### Fit an ANOVA model including all the two-way interactions.
> ### Recommend a couple of important factors, on which further
experiments will be conducted.
> ### You will find that you cannot carry any hypothesis test, since the
residual is zero; explain how
> ### you reach your decision.
>
> ### Fit ANOVA
> ?peanut
> g=lm(solubility ~ (press+temp+moist+flow+size)^2., data=peanut)
> summary(g)

Call:
lm(formula = solubility ~ (press + temp + moist + flow + size)^2,
    data = peanut)

Residuals:
ALL 16 residuals are 0: no residual degrees of freedom!

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)    29.2000         NA      NA       NA
press1         -1.1125         NA      NA       NA
temp1           6.3875         NA      NA       NA
moist1          0.8375         NA      NA       NA
flow1          -9.4125         NA      NA       NA
size1           2.1375         NA      NA       NA
press1:temp1  105.2250         NA      NA       NA
press1:moist1   9.3750         NA      NA       NA
press1:flow1   18.5250         NA      NA       NA
press1:size1   -7.2250         NA      NA       NA
temp1:moist1    6.1750         NA      NA       NA
temp1:flow1     5.1250         NA      NA       NA
temp1:size1    -0.7250         NA      NA       NA
moist1:flow1    2.2750         NA      NA       NA
moist1:size1   -8.8750         NA      NA       NA
flow1:size1     0.4750         NA      NA       NA

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

>
```

```
> # Can't make F-tests because there are as many parameters as cases.
> # Coding: Low=0, High=1
>
> # If no significant effects and errors are normally distributed,
> # then the estimated effects would then just be linear combinations of
the errors and hence normal.
> # Look at normal quantile plot of the main effects
> # Outliers represent significant effects
>
> halfnorm (coef (g) [-1] ,labs=names(coef(g) [-1] ) )
```




```
>
> # From halfnorm, press1:flow1 and press1:temp1 are extremes (ouliers),
representing significant effects
>
> ### Recommend a couple of important factors, on which further
experiments will be conducted.
>
> # Since interactions of press-flow and press-temp were found to be
signficiant
> # conduct another experiment focusing on the factors: press, flow and
temp to determine their effects and significance
>
>
```

PROBLEM 4

```
############################################################
> ###PROBLEM 4
>
> ### Data: The eggprod comes from a randomized block experiment to
> ### determine factors affecting egg production.
>
> ### Data:
> ### The composite data frame has 12 rows and 3 columns. Six pullets were
placed into each of 12 pens.
> ### Four blocks were formed from groups of 3 pens based on location.
> ### Three treatments were applied. The number of eggs produced was
recorded

    eggs
> eggprod
   treat block eggs
1      O     1  330
2      O     2  288
3      O     3  295
4      O     4  313
5      E     1  372
6      E     2  340
7      E     3  343
8      E     4  341
9      F     1  359
10     F     2  337
11     F     3  373
12     F     4  302
> ?eggprod
```

Part a

```
> ### Part a:  What is the blocking variable?
>
> # A blocking variable is a variable that groups similar observations
> # In this case we have 3 treatments, 12 pens available.
> # Divide the pens in 4 blocks of 3 pens each wehre the pens
> # in each block are grouped by location
> # Assign treatment for each block
> # Blocking variable has 4 levels
> # Thus, pens grouped into 3 based on location
> # The blocking variable is the location of pens
> # Each level of the blocking variable represents a location
>
>
```

Part b

```
# Part b: Is there a difference in the treatments?
>
> g=lm(eggs~treat + block)
> anova(g)
Analysis of Variance Table

Response: eggs
          Df Sum Sq Mean Sq F value  Pr(>F)
treat      2 4212.5 2106.25  5.4437 0.04485 *
block      3 2330.2  776.75  2.0075 0.21446
Residuals  6 2321.5  386.92
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
>
> # Significant effect: p-value <0.05
> # Treatment effect significant: there is a difference in treatments
> # (null hypothesis that there are no differences among the three
treatment means is rejected at 0.005 level)
> # Blocking effect not significant
>

> # Note this is a sequential testing of the models:
> # y ~ 1
> # y ~ treat
> # y ~ treat + block
>
> # Orthogonal so same result if change order
> # g=lm(eggs~block+treat)
> # anova(g)
```

Part c

```
> ### Part c: What efficiency was gained by the blocked design?
>
> # Efficiency : Relative advantage of RCBD over CRD
(sigmaCRD/sigmaRCBD)^2
>
>
> gcrd = lm (eggs ~ treat)
> sigmaCRD = summary(gcrd)$sig; sigmaCRD
[1] 22.73458
> sigmaRCBD = summary(g)$sig; sigmaRCBD
[1] 19.6702
>
> efficiency = (sigmaCRD/sigmaRCBD)^2; efficiency
[1] 1.335846
>
> # The relative efficiency of RCBD over CRDis 1.336
> # The interpretation is that a CRD would require 33.6% more observations
to  # obtain the same level of precision as a RCBD.
```

PROBLEM 5

```
> ############################################################
> ###PROBLEM 5

> ### Data: The alfalfa data arise from a Latin square design where the
treatment factor is inoculum and the blocking factors are shade and
irrigation.
```

Part a

```
> ### Part a: Test the signicance of the treatment effects.
>
> g=lm(yield ~ ., data=alfalfa)
> anova(g)
Analysis of Variance Table

Response: yield
           Df  Sum Sq Mean Sq F value   Pr(>F)
shade       4  87.402  21.851  7.1254 0.003533 **
irrigation  4  16.562   4.141  1.3502 0.307872
inoculum    4 155.894  38.974 12.7091 0.000284 ***
Residuals  12  36.799   3.067
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
>
> # Significant effect: p-value <0.05
> # Treatment effect (Inoculum) is significant
> # Blocking effect shade significant
>
```

Part b

```
> ### Part b: Determine which levels of the treatment factor are
signicantly different.
>
>
> tky = TukeyHSD(aov(yield ~ ., data=alfalfa), 'inoculum' ) ; tky
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = yield ~ ., data = alfalfa)

$inoculum
     diff        lwr       upr     p adj
B-A -0.72  -4.250202  2.810202 0.9633433
C-A -0.08  -3.610202  3.450202 0.9999928
D-A -0.86  -4.390202  2.670202 0.9326392
E-A -6.60 -10.130202 -3.069798 0.0005166
C-B  0.64  -2.890202  4.170202 0.9759059
D-B -0.14  -3.670202  3.390202 0.9999332
```

```
E-B -5.88  -9.410202 -2.349798 0.0014163
D-C -0.78  -4.310202  2.750202 0.9515868
E-C -6.52 -10.050202 -2.989798 0.0005764
E-D -5.74  -9.270202 -2.209798 0.0017334

> names(tky)
[1] "inoculum"
> comb.names = row.names(tky$inoculum);
> sig=tky$inoculum[,4]<0.05
> comb.names[sig]
[1] "E-A" "E-B" "E-C" "E-D"
>
> # Signficant different pairs: Look at p-value <0.05 or CI that doesn't
include zero
> # Treatment paris "E-A" , "E-B" , "E-C" and "E-D" are signficantly
different
```

Part c

```
> ### Part c: Compare the effciency (i.e., compare sigmahat^2) of the
Latin square, the completely ran-
> ### domized design, and the blocked designs (with shade only, or with
irrigation only).
>
>
> # efficiency
>
> myanova=anova(g)
> gr=lm(yield ~ inoculum, data=alfalfa)
> (summary(gr)$sig / summary(g)$sig)^2
[1] 2.295115
>
> # Relative efficiency of Latin Squares over CRD is 2.295
>
> gr=lm(yield ~ inoculum + irrigation, data=alfalfa)
> (summary(gr)$sig / summary(g)$sig)^2
[1] 2.531338
>
> # Relative efficiency of Latin Squares over block design with irrigation
only is 2.531
>
> gr=lm(yield ~ inoculum + shade, data=alfalfa)
> (summary(gr)$sig / summary(g)$sig)^2
[1] 1.087556
>
> # Relative efficiency of Latin Squares over block design with  shade
only is 1.088
>
>
```