ANOVA and Design of Experiments

Problem #1 - DoE

Use the following data to conduct a Design of Experiments analysis by hand (i.e. not using R). This is a 2^3 design, which is about the limit of complexity you'd want to handle by hand. You should produce a table similar to those in page 52 the DOE slides. I've provided a generic 2^3 table below; you would augment it to include the data below.

In [158]:
```
temporaryFile <- tempfile()
download.file("https://www.dropbox.com/s/avt401loc9f8qk6/3-factor-sign-table.csv?dl=0",
                  destfile=temporaryFile, method="curl", extra="-L")
signTable = read.csv(temporaryFile)
```

In [159]:
```
signTable$Y = c(100, 15, 40, 30, 120, 110, 20, 50)
signTable
```

Out[159]:

	I	A	B	C	AB	AC	BC	ABC	Y
1	1	-1	-1	-1	1	1	1	-1	100
2	1	-1	-1	1	1	-1	-1	1	15
3	1	-1	1	-1	-1	1	-1	1	40
4	1	-1	1	1	-1	-1	1	-1	30
5	1	1	-1	-1	-1	-1	1	1	120
6	1	1	-1	1	-1	1	-1	-1	110
7	1	1	1	-1	1	-1	-1	-1	20
8	1	1	1	1	1	1	1	1	50

You should compute terms qA, qB, .., qABC, or the effects size of each term.

In [160]:
```
qA = (-100-15-40-30+120+110+20+50)/8

qB = (-100-15+40+30-120-110+20+50)/8

qC = (-100+15-40+30-120+110-20+50)/8

qAB = (+100+15-40-30-120-110+20+50)/8

qAC = (+100-15+40-30-120+110-20+50)/8

qBC = (+100-15-40+30+120-110-20+50)/8

qABC = (-100+15+40-30+120-110-20+50)/8

effects = c(qA, qB, qC, qAB, qAC, qBC, qABC)
effects
```

Out[160]: 14.375 -25.625 -9.375 -14.375 14.375 14.375 -4.375

Now, compute the percentage variation due to each

In [161]:
```
#Find sst
eff = c(qA,qB,qC,qAB,qAC,qBC,qABC)
eff2 = eff^2
(SST = sum(8*eff2))

#Find percentage variation due to each
(8*eff2/SST)
(8*eff2)
```

Out[161]: 12721.875

Out[161]: 0.129943502824859 0.41292065831491 0.0552689756816507 0.129943502824859 0.129943502824859 0.129943502824859 0.0120363547040039

Out[161]: 1653.125 5253.125 703.125 1653.125 1653.125 1653.125 153.125

You can now compare it to the results using the **lm** and **anova** functions

In [162]:
```
#linear model
model = lm(Y ~ A + B + C + AB + AC + BC + ABC , data=signTable)
(model)

#anova using aov
fit <- aov(Y ~ A*B*C, data=signTable)
(fit)

#anova using anova
fit2 <- anova(model)
(fit2)
```

Out[162]:
```
Call:
lm(formula = Y ~ A + B + C + AB + AC + BC + ABC, data = signTable)

Coefficients:
(Intercept)            A            B            C           AB           AC
     60.625       14.375      -25.625       -9.375      -14.375       14.375
         BC          ABC
     14.375       -4.375
```

Out[162]:
```
Call:
   aov(formula = Y ~ A * B * C, data = signTable)

Terms:
                       A        B        C      A:B      A:C      B:C    A:B:C
Sum of Squares  1653.125 5253.125  703.125 1653.125 1653.125 1653.125  153.125
Deg. of Freedom        1        1        1        1        1        1        1

Estimated effects are balanced


Warning message:
In anova.lm(model): ANOVA F-tests on an essentially perfect fit are unreliable
```

Out[162]:

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
A	1	1653.125	1653.125	NaN	NaN
B	1	5253.125	5253.125	NaN	NaN
C	1	703.125	703.125	NaN	NaN
AB	1	1653.125	1653.125	NaN	NaN
AC	1	1653.125	1653.125	NaN	NaN
BC	1	1653.125	1653.125	NaN	NaN
ABC	1	153.125	153.125	NaN	NaN
Residuals	0	0	NaN	NA	NA

Problem #2 - 2-Level Dhrystone

This file (dhry-2level.csv) contains the following table

In [163]:

```
temporaryFile <- tempfile()
download.file("https://www.dropbox.com/s/1eabtdr5tsrtlaf/dhry-2level.csv?dl=0",
                  destfile=temporaryFile, method="curl", extra="-L")
dhry2lvl = read.csv(temporaryFile)
#dhry2lvl # - print this out and it will have 24 rows
```

This is the result of running a $3 * 2^3$ experimental design for evaluating the importance of certain compiler optimizations -- you can ignore their meaning for this problem (this is a subset of data from an experiment described below). Use R to conduct an analysis of variance for this data.

a) Set up a linear model. Treat the a, f and g level as categorical factors. You can do this using e.g.

```
data$bits = factor(data$bits).
```

In [164]:

```
#make them catigorical factors
dhry2lvl$a = factor(dhry2lvl$a)
dhry2lvl$f = factor(dhry2lvl$f)
dhry2lvl$g = factor(dhry2lvl$g)

#linear model
m = lm(mips ~ a*f*g, data=dhry2lvl)
```

b) Report the linear model.

In [165]:

```
#report linear model
(m)
summary(m)
```

Out[165]:

```
Call:
lm(formula = mips ~ a * f * g, data = dhry2lvl)

Coefficients:
(Intercept)           a1           f1           g1        a1:f1        a1:g1
  1440.5070       4.1047    -114.7363       4.8220       0.6523      17.5640
      f1:g1     a1:f1:g1
     8.5200     -34.2030
```

Out[165]:

```
Call:
lm(formula = mips ~ a * f * g, data = dhry2lvl)

Residuals:
    Min      1Q  Median      3Q     Max
-15.319  -6.148   1.662   5.069  17.420

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 1440.5070     6.3922 225.355  < 2e-16 ***
a1             4.1047     9.0399   0.454   0.6559
f1          -114.7363     9.0399 -12.692 9.09e-10 ***
g1             4.8220     9.0399   0.533   0.6011
a1:f1          0.6523    12.7843   0.051   0.9599
a1:g1         17.5640    12.7843   1.374   0.1884
f1:g1          8.5200    12.7843   0.666   0.5146
a1:f1:g1     -34.2030    18.0798  -1.892   0.0768 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 11.07 on 16 degrees of freedom
Multiple R-squared:  0.9777,    Adjusted R-squared:  0.968
F-statistic: 100.4 on 7 and 16 DF,  p-value: 5.093e-12
```

c) Note that R reports this differently for categorical data. The report should be missing terms for "a", "f" and so on. Why is that? What does the intercept represent? Note that categorical values must take one some value and the intercept needs to include the results of some value(s).

In [166]:

```
print("The a, f terms are split into a linear and constant part.")
print("The constants of a, f... are absorbed in the intercept.")
print("The intercept of the lm is the the every factors constant term combined.")
```

```
[1] "The a, f terms are split into a linear and constant part."
[1] "The constants of a, f... are absorbed in the intercept."
[1] "The intercept of the lm is the the every factors constant term combined."
```

d) Determine the percentage of variation attributable to each factor. You can do this using the anova table as described in the lecture notes.

In [167]:

```
#anova
anova(m)
anova_m1 <- anova(m)

#pct variation
sst = sum(anova(m)$Sum)
pct_var = anova(m)$Sum/sst
#rename columns
anova_m1 <- setNames(cbind(rownames(anova_m1), anova_m1, row.names = NULL),
         c("factor", "Df", "SumSq", "MeanSq", "FValue", "PR(>F)"))

#put results in table
ff <- data.frame(matrix(pct_var, nrow=1, ncol=8))
colnames(ff) <- anova_m1$factor
attr(ff,"title") <- "Percentage Variation"

#display table
print("Percentage Variation of residuals")
(ff)
```

Out[167]:

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
a	1	130.4101	130.4101	1.063884	0.3176712
f	1	84539.45	84539.45	689.6719	1.385664e-14
g	1	520.4198	520.4198	4.245579	0.05599911
a:f	1	405.8626	405.8626	3.311023	0.08757807
a:g	1	0.3208594	0.3208594	0.002617567	0.9598296
f:g	1	110.4632	110.4632	0.9011577	0.356591
a:f:g	1	438.692	438.692	3.578844	0.07676194
Residuals	16	1961.268	122.5792	NA	NA

```
[1] "Percentage Variation of residuals"
```

Out[167]:

	a	f	g	a:f	a:g	f:g	a:f:g	Residuals
1	0.001480136	0.9595101	0.005906687	0.00460648	3.641706e-06	0.001253741	0.004979088	0.0222601

e) Compute the 95% confidence interval for the "opt" factor. The "anova" function doesn't do this, but you can see something similar using

```
confint( aov (l ) )
```

where 'l' is your linear model. For models involving factors, the confidence interval is expressed for the different levels in each categorical factor.

In [168]:

```
#find confidence interval
confint(aov(m))
print("There is no opt factor so here is the confidence interval for all of them.")
```

Out[168]:

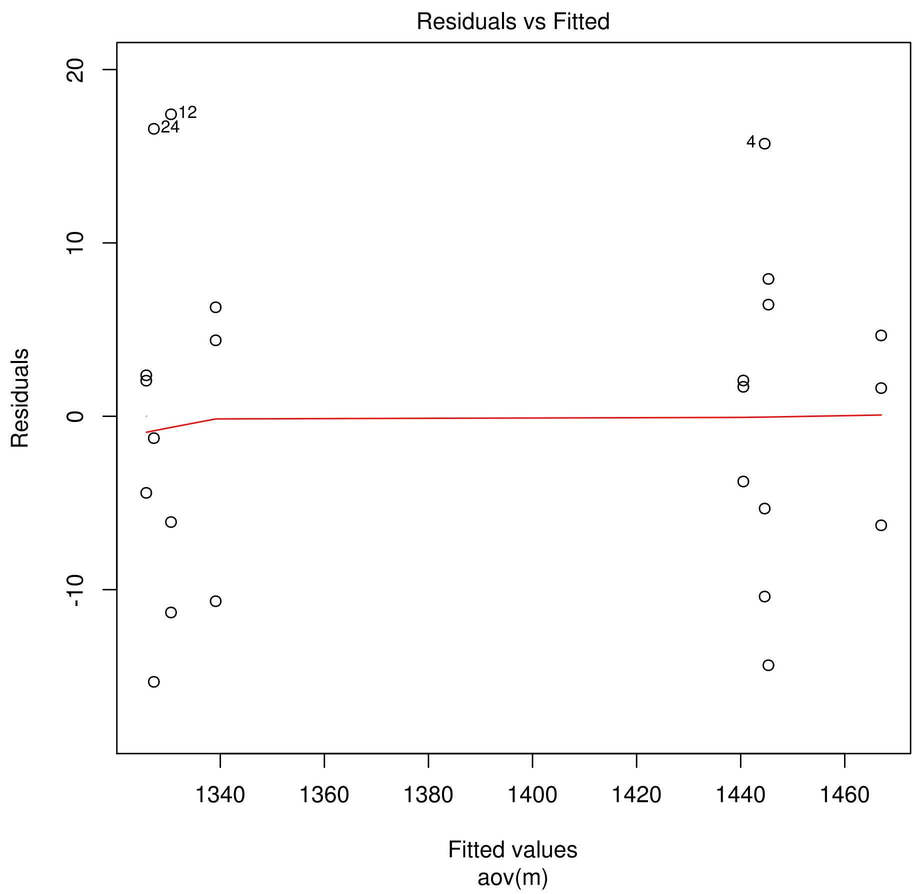
	2.5 %	97.5 %
(Intercept)	1426.956	1454.058
a1	-15.05903	23.26836
f1	-133.90003	-95.57264
g1	-14.3417	23.9857
a1:f1	-26.44923	27.75389
a1:g1	-9.537558	44.665558
f1:g1	-18.58156	35.62156
a1:f1:g1	-72.530392	4.124392

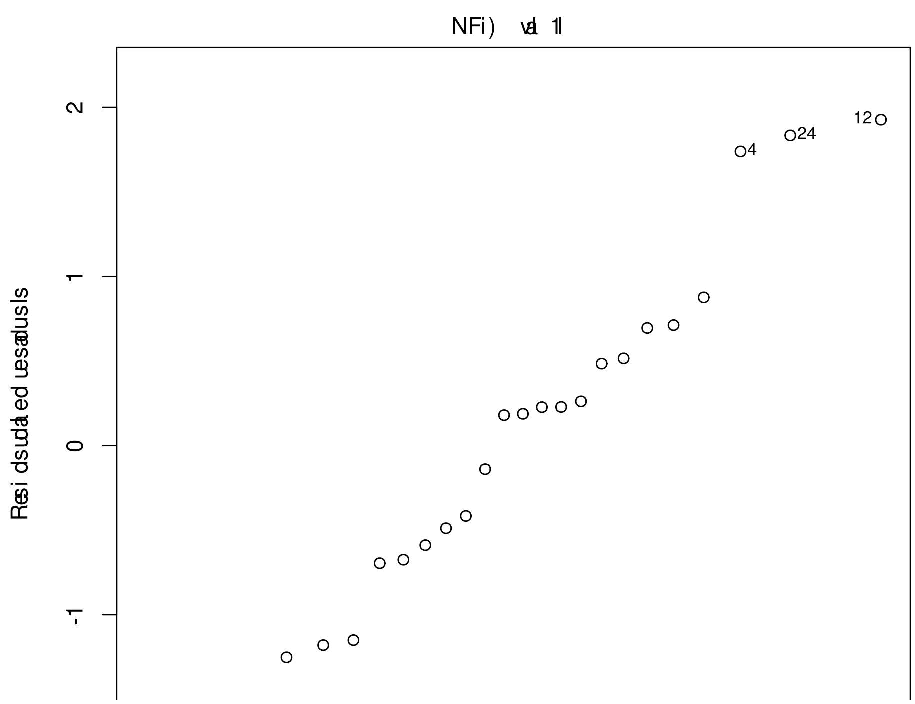
```
[1] "There is no opt factor so here is the confidence interval for all of them."
```

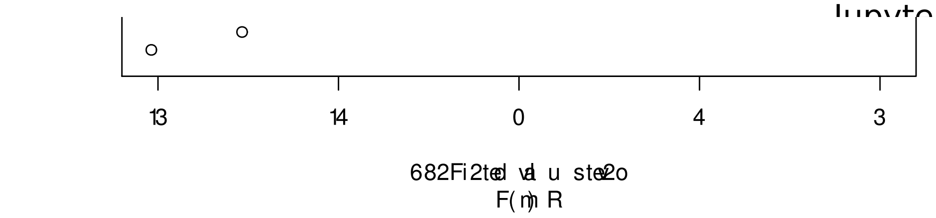
f) Plot out the diagnostic plots for the model. You can do this using "plot(aov(l))" where "l" is your model. Can you justify that your model is reasonable?

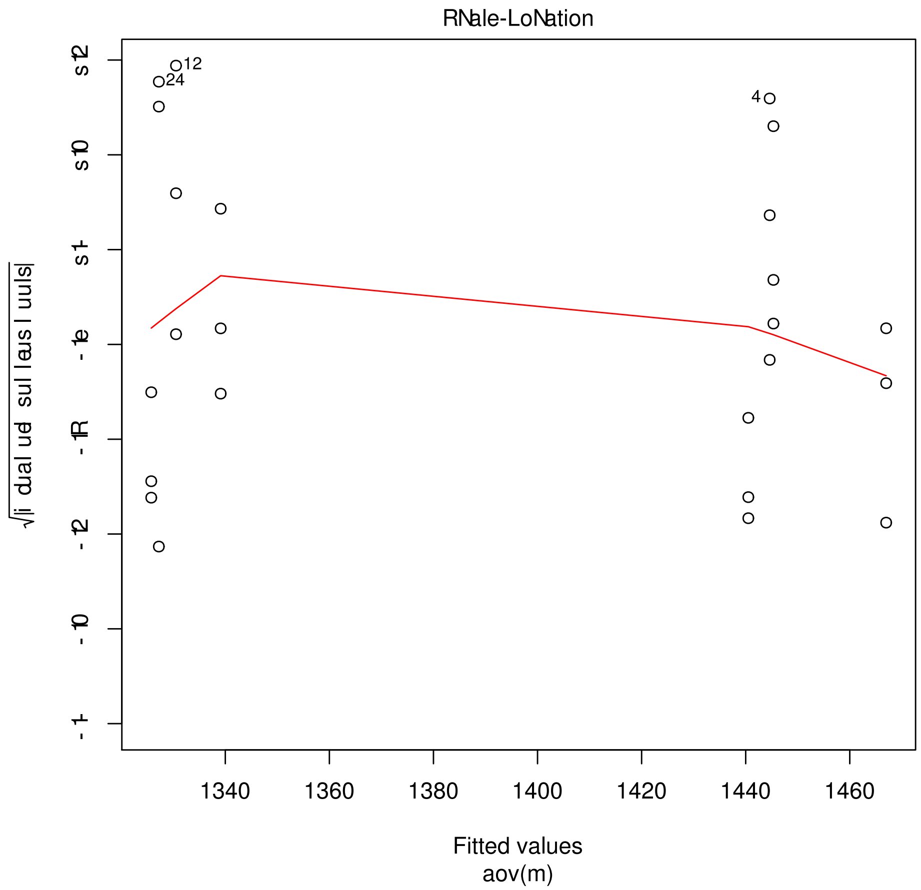
In [169]:

```
#plot anova stuff
plot(aov(m))
print("The residuals are fairly small and don't seen to have a pattern associated with them.")
print("Also they are evenly distributed in the + and - side.")
print("Furthermore the QQ plot is pretty linear.")
print("These all point to a reasonable model.")
print("Only worrysome point is the lack of data.")
```

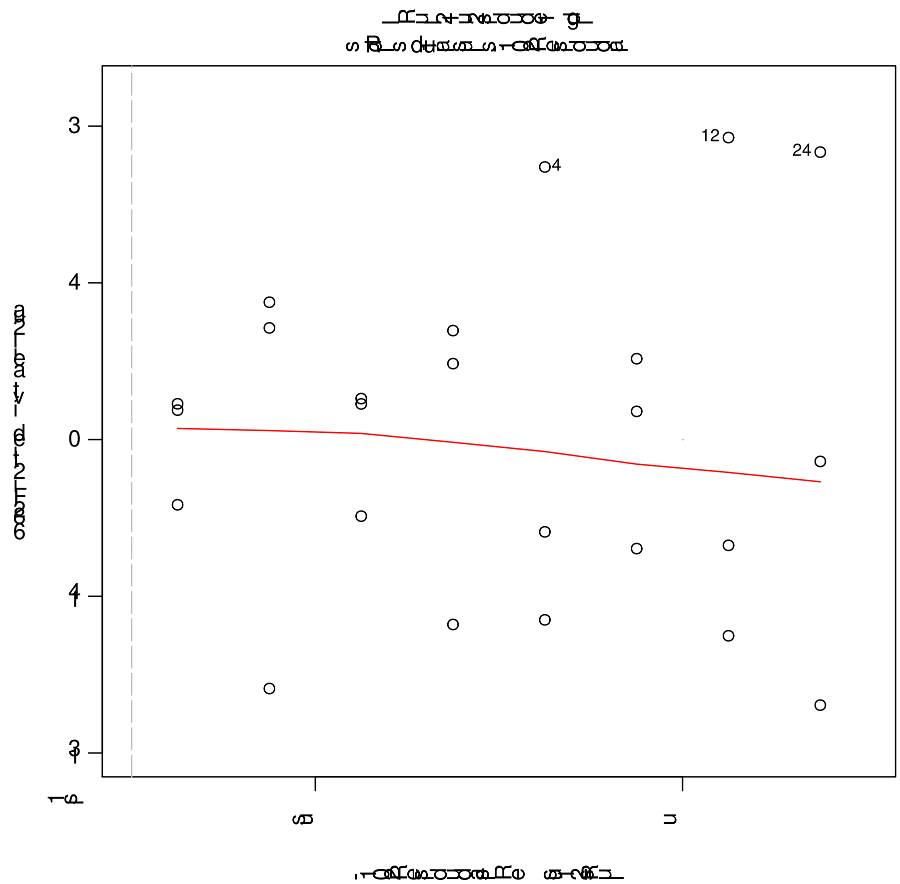









```
[1] "The residuals are fairly small and don't seen to have a pattern associated with them."
[1] "Also they are evenly distributed in the + and - side."
[1] "Furthermore the QQ plot is pretty linear."
[1] "These all point to a reasonable model."
[1] "Only worrysome point is the lack of data."
```



Problem #3 -

The file OUTPUT-optlevel (see below) contains data from a 422 study with 2 replicates. The first column is the trial number (you shouldn't include this in you model, or you won't get any "residuals"). The first factor ('opt') has four levels (-O0, -O1, -O2, -O3). The second factor ('bits') has two levels (specifying -m32 or -m64 for compilation as a 32-bit or 64-bit application) and the third ('benchmark') has two levels (dhry11 or dhry21). In each case, the dependent variable is the performance in VAX MIPS (this is the output of the Dhrystone program). Use R to conduct an analysis of variance for this data.

In [170]:

```
temporaryFile <- tempfile()
download.file("https://www.dropbox.com/s/l0lenc8prwfo7e3/OUTPUT-optlevels.orig?dl=0",
                  destfile=temporaryFile, method="curl", extra="-L")
optlevels = read.csv(temporaryFile)
#optlevels # - this will have 32 rows
```

a) Set up a linear model. Treat the bits, progs and optimization level as catagorical factors.

In [171]:

```
#make them catigorical
optlevels$opt = factor(optlevels$opt)
optlevels$bits = factor(optlevels$bits)
optlevels$prog = factor(optlevels$prog)

#linear model
m2 = lm(mips ~ opt*bits*prog, data=optlevels)
```

b) Report the linear model. Note that R reports this differently for catagorical data. The report should be missing terms for "opt0", "bits0" and so on. Why is that? What does the intercept represent?

In [172]:

```
#diplay lm
(m2)
summary(m2)
print("The opt0 and bits0 terms are absorbed into the intercept.")
print("The intercept of the lm is the the every factors constant term combined.")
```

Out[172]:

```
Call:
lm(formula = mips ~ opt * bits * prog, data = optlevels)

Coefficients:
     (Intercept)               opt1               opt2               opt3  
          3077.7              504.3             4911.3             4486.1  
           bits1              prog1         opt1:bits1         opt2:bits1  
          2329.9              462.3             -634.6             -808.9  
      opt3:bits1         opt1:prog1         opt2:prog1         opt3:prog1  
          -187.5             -375.8             -848.0             -194.8  
     bits1:prog1   opt1:bits1:prog1   opt2:bits1:prog1   opt3:bits1:prog1  
          -472.5              383.9             1111.2              204.2  
```

Out[172]:

```
Call:
lm(formula = mips ~ opt * bits * prog, data = optlevels)

Residuals:
    Min      1Q  Median      3Q     Max 
-278.14  -71.26    0.00   71.26  278.14 

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)    
(Intercept)        3077.7      134.3  22.921 1.16e-13 ***
opt1                504.3      189.9   2.656  0.01726 *  
opt2               4911.3      189.9  25.864 1.76e-14 ***
opt3               4486.1      189.9  23.624 7.24e-14 ***
bits1              2329.9      189.9  12.270 1.49e-09 ***
prog1               462.3      189.9   2.435  0.02700 *  
opt1:bits1         -634.6      268.5  -2.363  0.03113 *  
opt2:bits1         -808.9      268.5  -3.012  0.00827 ** 
opt3:bits1         -187.5      268.5  -0.698  0.49517    
opt1:prog1         -375.8      268.5  -1.400  0.18074    
opt2:prog1         -848.0      268.5  -3.158  0.00609 ** 
opt3:prog1         -194.8      268.5  -0.725  0.47867    
bits1:prog1        -472.5      268.5  -1.760  0.09759 .  
opt1:bits1:prog1    383.9      379.8   1.011  0.32720    
opt2:bits1:prog1   1111.2      379.8   2.926  0.00989 ** 
opt3:bits1:prog1    204.2      379.8   0.538  0.59822    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 189.9 on 16 degrees of freedom
Multiple R-squared:  0.9968,    Adjusted R-squared:  0.9937 
F-statistic: 328.9 on 15 and 16 DF,  p-value: < 2.2e-16

[1] "The opt0 and bits0 terms are absorbed into the intercept."
[1] "The intercept of the lm is the the every factors constant term combined."
```

c) Determine the percentage of variation attributable to each factor. You can do this using the anova table.

In [173]:

```
#anova
anova(m2)
anova_m2 <- anova(m2)

#pct variation
sst2 = sum(anova(m2)$Sum)
pct_var2 = anova(m2)$Sum/sst2
#(pct_var2)

#change col names
anova_m2 <- setNames(cbind(rownames(anova_m2), anova_m2, row.names = NULL),
         c("factor", "Df", "SumSq", "MeanSq", "FValue", "PR(>F)"))

#put results in table
ff <- data.frame(matrix(pct_var2, nrow=1, ncol=8))
colnames(ff) <- anova_m2$factor
attr(ff,"title") <- "Percentage Variation"

#display table
print("Percentage Variation of residuals")
(ff)
```

Out[173]:

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
opt	3	148353389	49451130	1371.402	1.690382e-19
bits	1	28829066	28829066	799.5012	4.349935e-15
prog	1	56143.58	56143.58	1.557	0.2300588
opt:bits	3	229529.3	76509.76	2.121804	0.1376009
opt:prog	3	93947.14	31315.71	0.868462	0.4777577
bits:prog	1	4552.321	4552.321	0.1262471	0.7269961
opt:bits:prog	3	350977.3	116992.4	3.244489	0.04975912
Residuals	16	576941.1	36058.82	NA	NA

```
[1] "Percentage Variation of residuals"
```

Out[173]:

	opt	bits	prog	opt:bits	opt:prog	bits:prog	opt:bits:prog	Residuals
1	0.8311368	0.1615123	0.0003145394	0.001285918	0.0005263306	2.550398e-05	0.001966319	0.003232262

d) Compute the 95% confidence interval for the "opt" factor. The "anova" function doesn't do this, but you can see something similar using

```
confint( aov (l ) )
```

where 'l' is your linear model. For models involving factors, the confidence interval is expressed for the different levels in each categorical factor.

In [174]:

```
#display confidence intervals
confint(aov(m2))
```

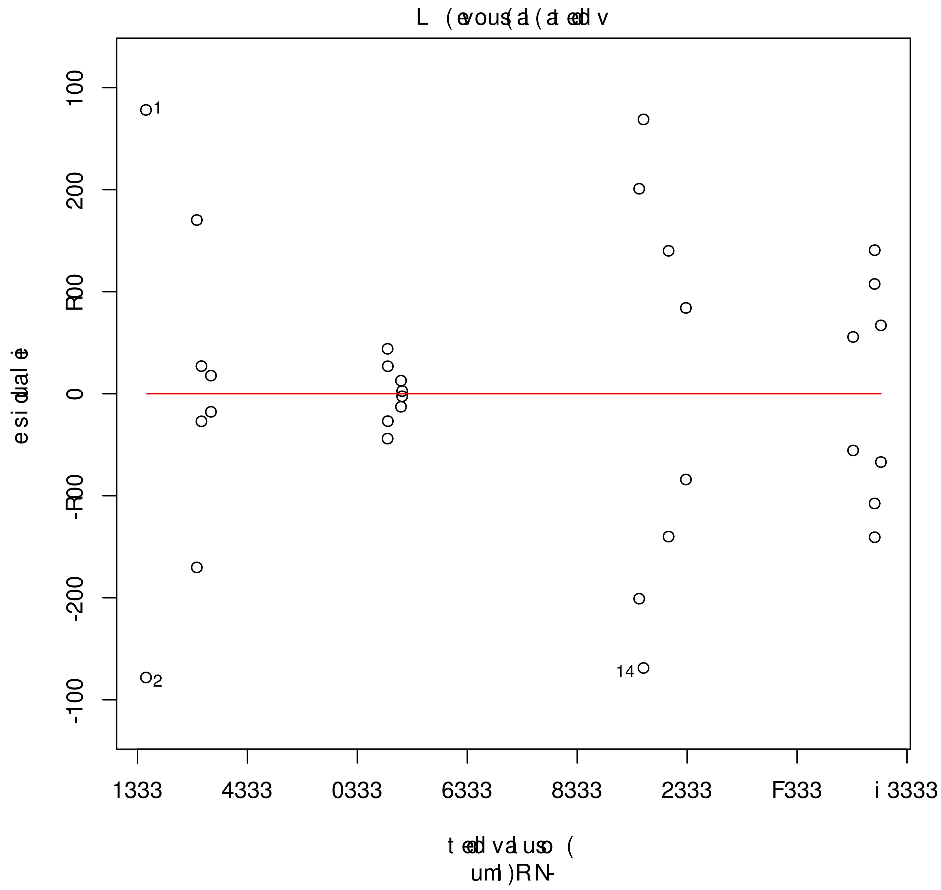
Out[174]:

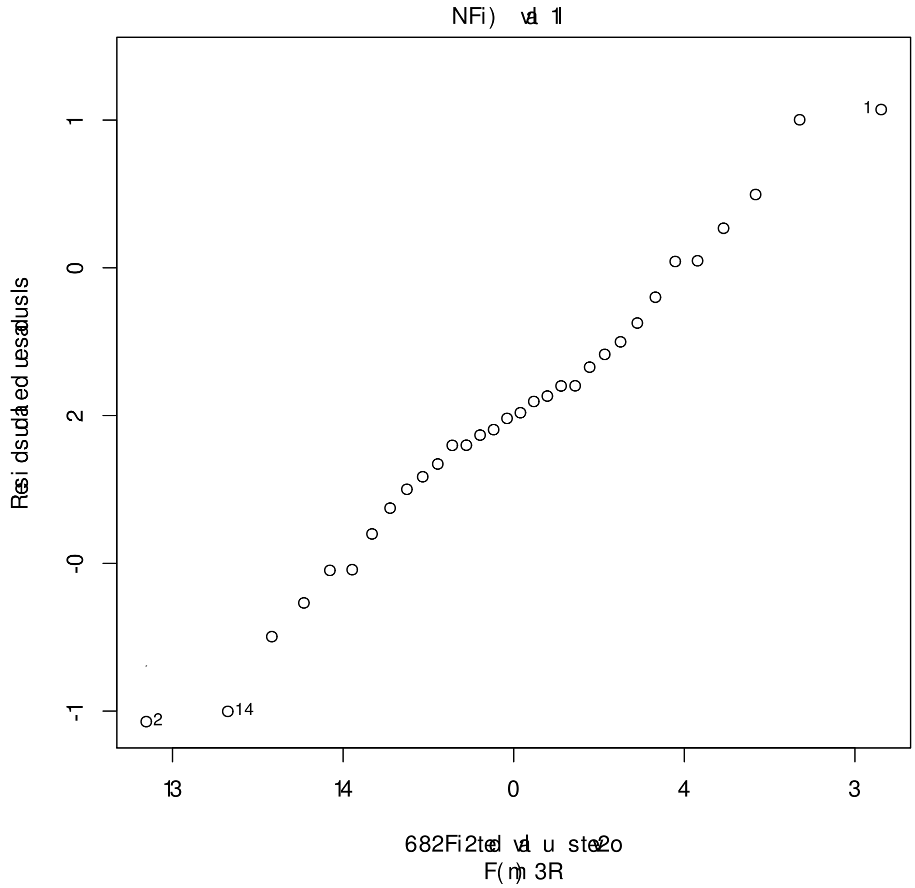
	2.5 %	97.5 %
(Intercept)	2793.061	3362.355
opt1	101.7688	906.8732
opt2	4508.722	5313.827
opt3	4083.535	4888.639
bits1	1927.361	2732.465
prog1	59.74181	864.84619
opt1:bits1	-1203.84477	-65.25523
opt2:bits1	-1378.2248	-239.6352
opt3:bits1	-756.7533	381.8363
opt1:prog1	-945.1343	193.4553
opt2:prog1	-1417.3108	-278.7212
opt3:prog1	-764.1043	374.4853
bits1:prog1	-1041.81577	96.77377
opt1:bits1:prog1	-421.2534	1188.9554
opt2:bits1:prog1	306.1021	1916.3109
opt3:bits1:prog1	-600.9144	1009.2944

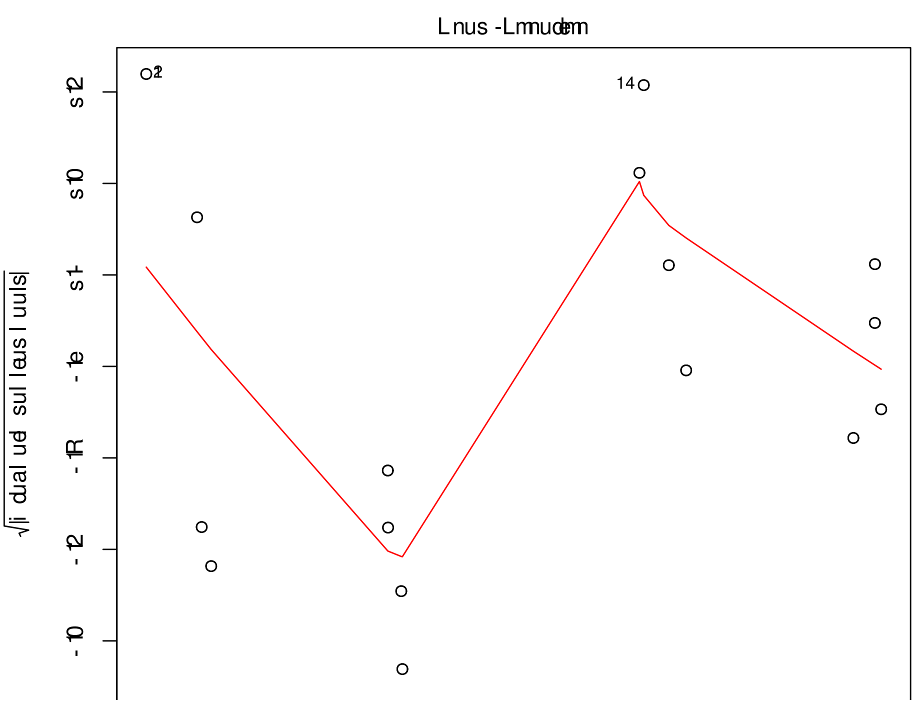
e) Plot out the diagnostic plots for the model. You can do this using "plot(aov(l))" where "l" is your model. Can you justify that your model is reasonable

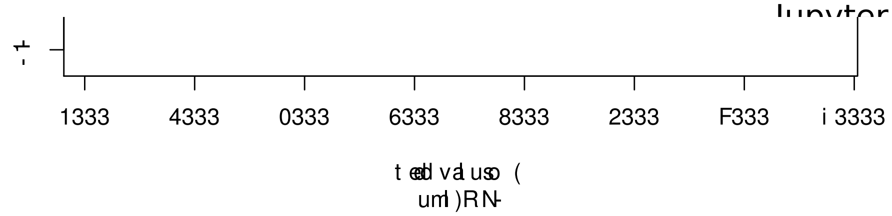
In [175]:

```
#plot anova stuff
plot(aov(m2))
print("The residuals are fairly small and don't seen to have a pattern associated with them.")
print("Also they are evenly distributed in the + and - side.")
print("Furthermore the QQ plot is pretty linear.")
print("Worrysome point is the lack of data.")
print("Also different points have very different residuals associated with them.")
print("still seems to be a reasonable model.")
```

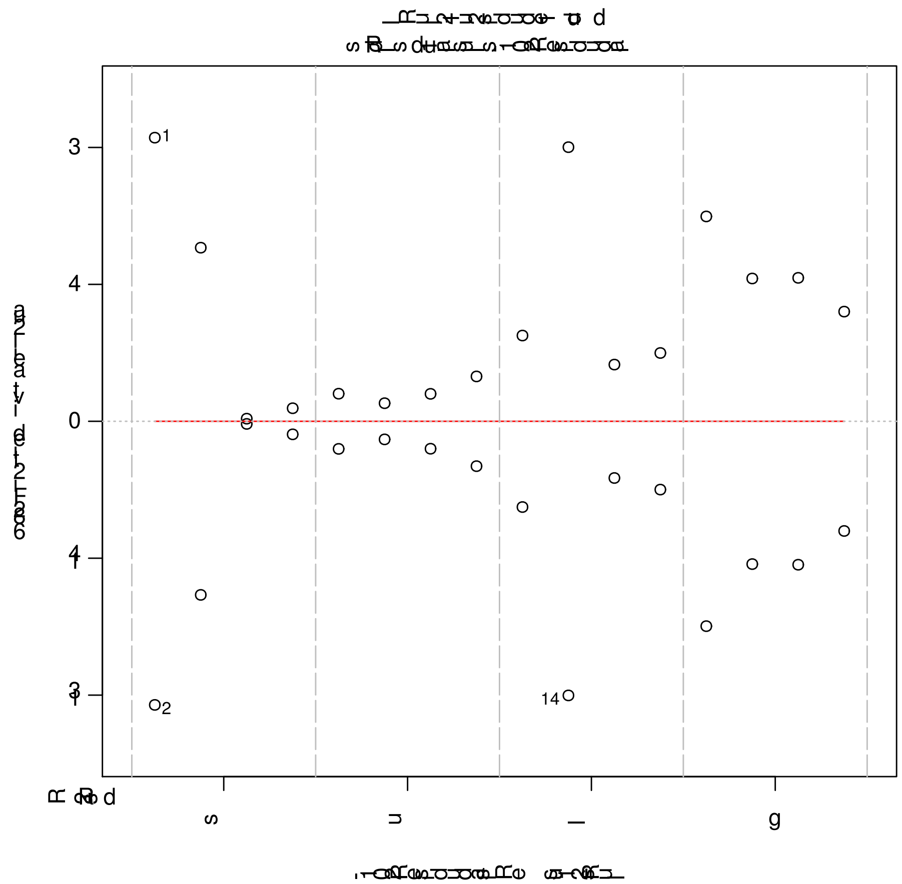








```
[1] "The residuals are fairly small and don't seen to have a pattern associated with them."
[1] "Also they are evenly distributed in the + and - side."
[1] "Furthermore the QQ plot is pretty linear."
[1] "Worrysome point is the lack of data."
[1] "Also different points have very different residuals associated with them."
[1] "still seems to be a reasonable model."
```



Problem #4 -

This file (OUTPUT-21.csv) contains a $3 * 2^7$ experiment. The data was generated from the dhrystone 2.1 benchmark on a specific machine. The factors in the experiment correspond to different optimizations as indicated in the RUN script:

```
DATA = (
('-fstrength-reduce', '-fno-strength-reduce'), # a
('-fgcse', '-fno-gcse'), # b
('-floop-optimize', '-fno-loop-optimize'), # c
('-fpeephole', '-fno-peephole'), # d
('-finline-functions', '-fno-inline-functions'), # e
('-fomit-frame-pointer', '-fno-omit-frame-pointer'), # f
('-fwhole-program', '-fno-whole-program') # g
)
```

each factor has two levels - either the optimization is enabled or not. In my model, I called the factors o.a, o.b, o.c, etc for the different possible optimizations.

In [176]:

```
temporaryFile <- tempfile()
download.file("https://www.dropbox.com/s/0hq1r63ym1h904q/OUTPUT-21.csv?dl=0",
                  destfile=temporaryFile, method="curl", extra="-L")
fullOpt = read.csv(temporaryFile)
#fullOpt # - if you print this out, it will have 384 rows
```

a) There are several terms that contribute more variation to the results than the error/residual term, but most of the terms contribute little variation to the overall performance. Determine which single factor contributes in some meaningful way to variation in performance and report the percentage of variation explained by that single term. Explain how you did this.

In [177]:

```
#Turn into categorical
fullOpt$o.a = factor(fullOpt$o.a)
fullOpt$o.b = factor(fullOpt$o.b)
fullOpt$o.c = factor(fullOpt$o.c)
fullOpt$o.d = factor(fullOpt$o.d)
fullOpt$o.e = factor(fullOpt$o.e)
fullOpt$o.f = factor(fullOpt$o.f)
fullOpt$o.g = factor(fullOpt$o.g)

#linear model
m4 = lm(mips ~ o.a*o.b*o.c*o.d*o.e*o.f*o.g, data=fullOpt)

#anova
anova_m4 = anova(m4)

#percentage variation
sst4 = sum(anova(m4)$Sum)
pct_var4 = anova(m4)$Sum/sst4

#find highest single factor and associated row
max(pct_var4)
max_row <- which.max(pct_var4)

#rename column names
anova_m4 <- setNames(cbind(rownames(anova_m4), anova_m4, row.names = NULL),
         c("factor", "Df", "SumSq", "MeanSq", "FValue", "PR(>F)"))

#print row
(anova_m4[max_row,])

print('The most important factor is o.f with a 64.99 percentage variation')
print("I found this by turning the factors into categorical variables,")
print("creating a linear model, performing anova analysis.")
print("Finally i calculated the percentage variation of all factors,")
print("and picked the highest one.")
```

Out[177]: 0.649967839195917

Out[177]:

	factor	Df	SumSq	MeanSq	FValue	PR(>F)
6	o.f	1	2077012	2077012	726.4939	9.973427e-77

```
[1] "The most important factor is o.f with a 64.99 percentage variation"
[1] "I found this by turning the factors into categorical variables,"
[1] "creating a linear model, performing anova analysis."
[1] "Finally i calculated the percentage variation of all factors,"
[1] "and picked the highest one."
```

b) What 2-factor interaction of factors contributes in some meaningful way to variation in performance. Explain how you determine which factors you're including.

In [178]:

```
#Pick out 2-factor interactions
#anova_m4$factor
start_2s <- which(grepl("o.a:o.b", anova_m4$factor))[1]
end_2s <- which(grepl("o.f:o.g", anova_m4$factor))[1]

#Find highest
pct_var_2s <- pct_var4[start_2s:end_2s]
print("pct variation of most influential")
max(pct_var_2s)
print("row number of that factor")
which.max(pct_var_2s)

#get row using value, print row
max_row <- which(pct_var4 == max(pct_var_2s))
print("factors row from anova")
(anova_m4[max_row,])

#Find second highest, first zero out previous highest
pct_var_2s[13] <- 0
print("pct variation of second most influential")
max(pct_var_2s)
print("row number of that factor")
which.max(pct_var_2s)

#get row using value, print row
max_row <- which(pct_var4 == max(pct_var_2s))
print("factors row from anova")
(anova_m4[max_row,])

print("2-Factor interactions that have highest percentage variation are o.c:o.f at 0.56%,")
print("and o.a:o.b at 0.26%.")
print("I included all factors and then picked the ones with highest percentage variation")
print("In the 2-factor category.")
```

```
[1] "pct variation of most influential"
```

Out[178]: 0.00569151029703818

```
[1] "row number of that factor"
```

Out[178]: 13

```
[1] "factors row from anova"
```

Out[178]:

	factor	Df	SumSq	MeanSq	FValue	PR(>F)
20	o.c:o.f	1	18187.57	18187.57	6.361619	0.01226866

```
[1] "pct variation of second most influential"
```

Out[178]: 0.00265805714650366

```
[1] "row number of that factor"
```

Out[178]: 1

```
[1] "factors row from anova"
```

Out[178]:

	factor	Df	SumSq	MeanSq	FValue	PR(>F)
8	o.a:o.b	1	8493.985	8493.985	2.971012	0.08597612

```
[1] "2-Factor interactions that have highest percentage variation are o.c:o.f at 0.56%,"
[1] "and o.a:o.b at 0.26%."
[1] "I included all factors and then picked the ones with highest percentage variation"
[1] "In the 2-factor category."
```

c) There any two significant intercations that involve more than 2 factors. List those & explain how you did this.

In [179]:
```
#Pick out 3-factor interactions
#anova_m4$factor
start_3s <- which(grepl("o.a:o.b:o.c", anova_m4$factor))[1]
end_3s <- which(grepl("o.e:o.f:o.g", anova_m4$factor))[1]

#Find highest
pct_var_3s <- pct_var4[start_3s:end_3s]
print("pct variation of most influential")
max(pct_var_3s)
print("row number of that factor")
which.max(pct_var_3s)

#get row using value, print row
max_row <- which(pct_var4 == max(pct_var_3s))
print("factors row from anova")
(anova_m4[max_row,])

#Find second highest, first zero out previous highest
pct_var_3s[5] <- 0
print("pct variation of second most influential")
max(pct_var_3s)
print("row number of that factor")
which.max(pct_var_3s)

#get row using value, print row
max_row <- which(pct_var4 == max(pct_var_3s))
print("factors row from anova")
(anova_m4[max_row,])

print("3-Factor interactions that have highest percentage variation are: ")
print("and o.a:o.b:o.e at 0.45% and o.a:o.b:o.c at 0.37%.")
print("I included all factors and then picked the ones with highest percentage variation")
print("in the 3-factor category.")
```

```
[1] "pct variation of most influential"
```

Out[179]: 0.00453043169162491

```
[1] "row number of that factor"
```

Out[179]: 5

```
[1] "factors row from anova"
```

Out[179]:

	factor	Df	SumSq	MeanSq	FValue	PR(>F)
33	o.a:o.b:o.e	1	14477.27	14477.27	5.063837	0.02527951

```
[1] "pct variation of second most influential"
```

Out[179]: 0.00373568245534494

```
[1] "row number of that factor"
```

Out[179]: 1

```
[1] "factors row from anova"
```

Out[179]:

	factor	Df	SumSq	MeanSq	FValue	PR(>F)
29	o.a:o.b:o.c	1	11937.6	11937.6	4.175515	0.04203591

```
[1] "3-Factor interactions that have highest percentage variation are: "
[1] "and o.a:o.b:o.e at 0.45% and o.a:o.b:o.c at 0.37%."
[1] "I included all factors and then picked the ones with highest percentage variation"
[1] "in the 3-factor category."
```

Problem #4 - reduced

I've taken the data from the previous question and prepared a $3 * 2^{7-4}$ design table. You're left with a 24 row table (8 factors, repeated 3 times). This is in the following table.

In [183]:
```
temporaryFile <- tempfile()
download.file("https://www.dropbox.com/s/n9urefzm9jzggh6/OUTPUT-21-reduced.csv?dl=0",
                  destfile=temporaryFile, method="curl", extra="-L")
reducedOpt = read.csv(temporaryFile)
#reducedOpt # if you print this out, it will have 24 rows
```

a) Again, there should be a single factor that explains the most variation, but the percentage of variation explained by that factor should now be different. Compute the percentage for that single factor.

In [184]:

```
#Turn into categorical
reducedOpt$o.a = factor(reducedOpt$o.a)
reducedOpt$o.b = factor(reducedOpt$o.b)
reducedOpt$o.c = factor(reducedOpt$o.c)
reducedOpt$o.d = factor(reducedOpt$o.d)
reducedOpt$o.e = factor(reducedOpt$o.e)
reducedOpt$o.f = factor(reducedOpt$o.f)
reducedOpt$o.g = factor(reducedOpt$o.g)

#linear model
m5 = lm(mips ~ o.a*o.b*o.c*o.d*o.e*o.f*o.g, data=reducedOpt)
#m5 = lm(mips ~ o.a+o.b+o.c+o.d+o.e+o.f+o.g, data=reducedOpt)

#anova
anova_m5 = anova(m5)

#percentage variation
sst5 = sum(anova(m5)$Sum)
pct_var5 = anova(m5)$Sum/sst5

#find highest single factor, print row
max(pct_var5)
max_row <- which.max(pct_var5)
anova_m5 <- setNames(cbind(rownames(anova_m5), anova_m5, row.names = NULL),
         c("factor", "Df", "SumSq", "MeanSq", "FValue", "PR(>F)"))
(anova_m5[max_row,])

print("Factor o.f has the highest percentage of variation with 76.27%.")
```

Out[184]: 0.762742739498329

Out[184]:

	factor	Df	SumSq	MeanSq	FValue	PR(>F)
6	o.f	1	126683.4	126683.4	59.34043	9.023391e-07

```
[1] "Factor o.f has the highest percentage of variation with 76.27%."
```

b) Now, determine why that variation is different. Show that the "F" factor is confounded with the ABCDEG factor and explain why that changes the variation attributed to F.

In [182]:

```
library(car)
vif(m5)
print("Find Alias")
(alias(m5))

print("vif() couldn't be used to find variable inflation because there are aliased coefficients in the")
print("model. Thus alias() was used to look for correlated variables. Many variables were correlated. F")
print("and ABCDEG had some collinearity.")
```

```
Error in vif.default(m5): there are aliased coefficients in the model


[1] "Find Alias"
```

Out[182]:

```
Model :
mips ~ o.a * o.b * o.c * o.d * o.e * o.f * o.g

Complete :
                                  (Intercept) o.a1 o.b1 o.c1 o.d1 o.e1 o.f1
o.a1:o.b1                         -1/2         1/2  1/2    0  1/2    0    0
o.a1:o.c1                         -1/2         1/2    0  1/2    0  1/2    0
o.b1:o.c1                         -1/2           0  1/2  1/2    0    0  1/2
o.a1:o.d1                         -1/2         1/2  1/2    0  1/2    0    0
o.b1:o.d1                         -1/2         1/2  1/2    0  1/2    0    0
o.c1:o.d1                         -1/2           0    0  1/2  1/2    0    0
o.a1:o.e1                         -1/2         1/2    0  1/2    0  1/2    0
o.b1:o.e1                         -1/2           0  1/2    0    0  1/2    0
o.c1:o.e1                         -1/2         1/2    0  1/2    0  1/2    0
o.d1:o.e1                         -1/2           0    0    0  1/2  1/2  1/2
o.a1:o.f1                         -1/2         1/2    0    0    0    0  1/2
o.b1:o.f1                         -1/2           0  1/2  1/2    0    0  1/2
o.c1:o.f1                         -1/2           0  1/2  1/2    0    0  1/2
o.d1:o.f1                         -1/2           0    0    0  1/2  1/2  1/2
o.e1:o.f1                         -1/2           0    0    0  1/2  1/2  1/2
o.a1:o.g1                         -1/2         1/2    0    0    0    0  1/2
o.b1:o.g1                         -1/2           0  1/2    0    0  1/2    0
o.c1:o.g1                         -1/2           0    0  1/2  1/2    0    0
o.d1:o.g1                         -1/2           0    0  1/2  1/2    0    0
o.e1:o.g1                         -1/2           0  1/2    0    0  1/2    0
o.f1:o.g1                         -1/2         1/2    0    0    0    0  1/2
o.a1:o.b1:o.c1                    -3/4         1/4  1/4  1/4  1/4  1/4  1/4
o.a1:o.b1:o.d1                    -1/2         1/2  1/2    0  1/2    0    0
o.a1:o.c1:o.d1                    -3/4         1/4  1/4  1/4  1/4  1/4  1/4
o.b1:o.c1:o.d1                    -3/4         1/4  1/4  1/4  1/4  1/4  1/4
o.a1:o.b1:o.e1                    -3/4         1/4  1/4  1/4  1/4  1/4  1/4
o.a1:o.c1:o.e1                    -1/2         1/2    0  1/2    0  1/2    0
o.b1:o.c1:o.e1                    -3/4         1/4  1/4  1/4  1/4  1/4  1/4
o.a1:o.d1:o.e1                    -3/4         1/4  1/4  1/4  1/4  1/4  1/4
o.b1:o.d1:o.e1                    -3/4         1/4  1/4  1/4  1/4  1/4  1/4
o.c1:o.d1:o.e1                    -3/4         1/4  1/4  1/4  1/4  1/4  1/4
o.a1:o.b1:o.f1                    -3/4         1/4  1/4  1/4  1/4  1/4  1/4
o.a1:o.c1:o.f1                    -3/4         1/4  1/4  1/4  1/4  1/4  1/4
o.b1:o.c1:o.f1                    -1/2           0  1/2  1/2    0    0  1/2
o.a1:o.d1:o.f1                    -3/4         1/4  1/4  1/4  1/4  1/4  1/4
o.b1:o.d1:o.f1                    -3/4         1/4  1/4  1/4  1/4  1/4  1/4
o.c1:o.d1:o.f1                    -3/4         1/4  1/4  1/4  1/4  1/4  1/4
o.a1:o.e1:o.f1                    -3/4         1/4  1/4  1/4  1/4  1/4  1/4
o.b1:o.e1:o.f1                    -3/4         1/4  1/4  1/4  1/4  1/4  1/4
o.c1:o.e1:o.f1                    -3/4         1/4  1/4  1/4  1/4  1/4  1/4
o.d1:o.e1:o.f1                    -1/2           0    0    0  1/2  1/2  1/2
o.a1:o.b1:o.g1                    -3/4         1/4  1/4  1/4  1/4  1/4  1/4
o.a1:o.c1:o.g1                    -3/4         1/4  1/4  1/4  1/4  1/4  1/4
o.b1:o.c1:o.g1                    -3/4         1/4  1/4  1/4  1/4  1/4  1/4
o.a1:o.d1:o.g1                    -3/4         1/4  1/4  1/4  1/4  1/4  1/4
o.b1:o.d1:o.g1                    -3/4         1/4  1/4  1/4  1/4  1/4  1/4
o.c1:o.d1:o.g1                    -1/2           0    0  1/2  1/2    0    0
o.a1:o.e1:o.g1                    -3/4         1/4  1/4  1/4  1/4  1/4  1/4
o.b1:o.e1:o.g1                    -1/2           0  1/2    0    0  1/2    0
o.c1:o.e1:o.g1                    -3/4         1/4  1/4  1/4  1/4  1/4  1/4
o.d1:o.e1:o.g1                    -3/4         1/4  1/4  1/4  1/4  1/4  1/4
o.a1:o.f1:o.g1                    -1/2         1/2    0    0    0    0  1/2
o.b1:o.f1:o.g1                    -3/4         1/4  1/4  1/4  1/4  1/4  1/4
o.c1:o.f1:o.g1                    -3/4         1/4  1/4  1/4  1/4  1/4  1/4
o.d1:o.f1:o.g1                    -3/4         1/4  1/4  1/4  1/4  1/4  1/4
o.e1:o.f1:o.g1                    -3/4         1/4  1/4  1/4  1/4  1/4  1/4
o.a1:o.b1:o.c1:o.d1               -3/4         1/4  1/4  1/4  1/4  1/4  1/4
o.a1:o.b1:o.c1:o.e1               -3/4         1/4  1/4  1/4  1/4  1/4  1/4
o.a1:o.b1:o.d1:o.e1               -3/4         1/4  1/4  1/4  1/4  1/4  1/4
o.a1:o.c1:o.d1:o.e1               -3/4         1/4  1/4  1/4  1/4  1/4  1/4
o.b1:o.c1:o.d1:o.e1               -3/4         1/4  1/4  1/4  1/4  1/4  1/4
o.a1:o.b1:o.c1:o.f1               -3/4         1/4  1/4  1/4  1/4  1/4  1/4
o.a1:o.b1:o.d1:o.f1               -3/4         1/4  1/4  1/4  1/4  1/4  1/4
o.a1:o.c1:o.d1:o.f1               -3/4         1/4  1/4  1/4  1/4  1/4  1/4
o.b1:o.c1:o.d1:o.f1               -3/4         1/4  1/4  1/4  1/4  1/4  1/4
o.a1:o.b1:o.e1:o.f1               -3/4         1/4  1/4  1/4  1/4  1/4  1/4
o.a1:o.c1:o.e1:o.f1               -3/4         1/4  1/4  1/4  1/4  1/4  1/4
o.b1:o.c1:o.e1:o.f1               -3/4         1/4  1/4  1/4  1/4  1/4  1/4
o.a1:o.d1:o.e1:o.f1               -3/4         1/4  1/4  1/4  1/4  1/4  1/4
o.b1:o.d1:o.e1:o.f1               -3/4         1/4  1/4  1/4  1/4  1/4  1/4
o.c1:o.d1:o.e1:o.f1               -3/4         1/4  1/4  1/4  1/4  1/4  1/4
o.a1:o.b1:o.c1:o.g1               -3/4         1/4  1/4  1/4  1/4  1/4  1/4
o.a1:o.b1:o.d1:o.g1               -3/4         1/4  1/4  1/4  1/4  1/4  1/4
o.a1:o.c1:o.d1:o.g1               -3/4         1/4  1/4  1/4  1/4  1/4  1/4
o.b1:o.c1:o.d1:o.g1               -3/4         1/4  1/4  1/4  1/4  1/4  1/4
```

```
o.a1:o.b1:o.e1:o.g1              -3/4        1/4  1/4  1/4  1/4  1/4  1/4
o.a1:o.c1:o.e1:o.g1              -3/4        1/4  1/4  1/4  1/4  1/4  1/4
o.b1:o.c1:o.e1:o.g1              -3/4        1/4  1/4  1/4  1/4  1/4  1/4
o.a1:o.d1:o.e1:o.g1              -3/4        1/4  1/4  1/4  1/4  1/4  1/4
o.b1:o.d1:o.e1:o.g1              -3/4        1/4  1/4  1/4  1/4  1/4  1/4
o.c1:o.d1:o.e1:o.g1              -3/4        1/4  1/4  1/4  1/4  1/4  1/4
o.a1:o.b1:o.f1:o.g1              -3/4        1/4  1/4  1/4  1/4  1/4  1/4
o.a1:o.c1:o.f1:o.g1              -3/4        1/4  1/4  1/4  1/4  1/4  1/4
o.b1:o.c1:o.f1:o.g1              -3/4        1/4  1/4  1/4  1/4  1/4  1/4
o.a1:o.d1:o.f1:o.g1              -3/4        1/4  1/4  1/4  1/4  1/4  1/4
o.b1:o.d1:o.f1:o.g1              -3/4        1/4  1/4  1/4  1/4  1/4  1/4
o.c1:o.d1:o.f1:o.g1              -3/4        1/4  1/4  1/4  1/4  1/4  1/4
o.a1:o.e1:o.f1:o.g1              -3/4        1/4  1/4  1/4  1/4  1/4  1/4
o.b1:o.e1:o.f1:o.g1              -3/4        1/4  1/4  1/4  1/4  1/4  1/4
o.c1:o.e1:o.f1:o.g1              -3/4        1/4  1/4  1/4  1/4  1/4  1/4
o.d1:o.e1:o.f1:o.g1              -3/4        1/4  1/4  1/4  1/4  1/4  1/4
o.a1:o.b1:o.c1:o.d1:o.e1         -3/4        1/4  1/4  1/4  1/4  1/4  1/4
o.a1:o.b1:o.c1:o.d1:o.f1         -3/4        1/4  1/4  1/4  1/4  1/4  1/4
o.a1:o.b1:o.c1:o.e1:o.f1         -3/4        1/4  1/4  1/4  1/4  1/4  1/4
o.a1:o.b1:o.d1:o.e1:o.f1         -3/4        1/4  1/4  1/4  1/4  1/4  1/4
o.a1:o.c1:o.d1:o.e1:o.f1         -3/4        1/4  1/4  1/4  1/4  1/4  1/4
o.b1:o.c1:o.d1:o.e1:o.f1         -3/4        1/4  1/4  1/4  1/4  1/4  1/4
o.a1:o.b1:o.c1:o.d1:o.g1         -3/4        1/4  1/4  1/4  1/4  1/4  1/4
o.a1:o.b1:o.c1:o.e1:o.g1         -3/4        1/4  1/4  1/4  1/4  1/4  1/4
o.a1:o.b1:o.d1:o.e1:o.g1         -3/4        1/4  1/4  1/4  1/4  1/4  1/4
o.a1:o.c1:o.d1:o.e1:o.g1         -3/4        1/4  1/4  1/4  1/4  1/4  1/4
o.b1:o.c1:o.d1:o.e1:o.g1         -3/4        1/4  1/4  1/4  1/4  1/4  1/4
o.a1:o.b1:o.c1:o.f1:o.g1         -3/4        1/4  1/4  1/4  1/4  1/4  1/4
o.a1:o.b1:o.d1:o.f1:o.g1         -3/4        1/4  1/4  1/4  1/4  1/4  1/4
o.a1:o.c1:o.d1:o.f1:o.g1         -3/4        1/4  1/4  1/4  1/4  1/4  1/4
o.b1:o.c1:o.d1:o.f1:o.g1         -3/4        1/4  1/4  1/4  1/4  1/4  1/4
o.a1:o.b1:o.e1:o.f1:o.g1         -3/4        1/4  1/4  1/4  1/4  1/4  1/4
o.a1:o.c1:o.e1:o.f1:o.g1         -3/4        1/4  1/4  1/4  1/4  1/4  1/4
o.b1:o.c1:o.e1:o.f1:o.g1         -3/4        1/4  1/4  1/4  1/4  1/4  1/4
o.a1:o.d1:o.e1:o.f1:o.g1         -3/4        1/4  1/4  1/4  1/4  1/4  1/4
o.b1:o.d1:o.e1:o.f1:o.g1         -3/4        1/4  1/4  1/4  1/4  1/4  1/4
o.c1:o.d1:o.e1:o.f1:o.g1         -3/4        1/4  1/4  1/4  1/4  1/4  1/4
o.a1:o.b1:o.c1:o.d1:o.e1:o.f1    -3/4        1/4  1/4  1/4  1/4  1/4  1/4
o.a1:o.b1:o.c1:o.d1:o.e1:o.g1    -3/4        1/4  1/4  1/4  1/4  1/4  1/4
o.a1:o.b1:o.c1:o.d1:o.f1:o.g1    -3/4        1/4  1/4  1/4  1/4  1/4  1/4
o.a1:o.b1:o.c1:o.e1:o.f1:o.g1    -3/4        1/4  1/4  1/4  1/4  1/4  1/4
o.a1:o.b1:o.d1:o.e1:o.f1:o.g1    -3/4        1/4  1/4  1/4  1/4  1/4  1/4
o.a1:o.c1:o.d1:o.e1:o.f1:o.g1    -3/4        1/4  1/4  1/4  1/4  1/4  1/4
o.b1:o.c1:o.d1:o.e1:o.f1:o.g1    -3/4        1/4  1/4  1/4  1/4  1/4  1/4
o.a1:o.b1:o.c1:o.d1:o.e1:o.f1:o.g1 -3/4      1/4  1/4  1/4  1/4  1/4  1/4
                                 o.g1
o.a1:o.b1                           0
o.a1:o.c1                           0
o.b1:o.c1                           0
o.a1:o.d1                           0
o.b1:o.d1                           0
o.c1:o.d1                         1/2
o.a1:o.e1                           0
o.b1:o.e1                         1/2
o.c1:o.e1                           0
o.d1:o.e1                           0
o.a1:o.f1                         1/2
o.b1:o.f1                           0
o.c1:o.f1                           0
o.d1:o.f1                           0
o.e1:o.f1                           0
o.a1:o.g1                         1/2
o.b1:o.g1                         1/2
o.c1:o.g1                         1/2
o.d1:o.g1                         1/2
o.e1:o.g1                         1/2
o.f1:o.g1                         1/2
o.a1:o.b1:o.c1                    1/4
o.a1:o.b1:o.d1                      0
o.a1:o.c1:o.d1                    1/4
o.b1:o.c1:o.d1                    1/4
o.a1:o.b1:o.e1                    1/4
o.a1:o.c1:o.e1                      0
o.b1:o.c1:o.e1                    1/4
o.a1:o.d1:o.e1                    1/4
o.b1:o.d1:o.e1                    1/4
o.c1:o.d1:o.e1                    1/4
o.a1:o.b1:o.f1                    1/4
o.a1:o.c1:o.f1                    1/4
o.b1:o.c1:o.f1                      0
```

```
o.a1:o.d1:o.f1                1/4
o.b1:o.d1:o.f1                1/4
o.c1:o.d1:o.f1                1/4
o.a1:o.e1:o.f1                1/4
o.b1:o.e1:o.f1                1/4
o.c1:o.e1:o.f1                1/4
o.d1:o.e1:o.f1                  0
o.a1:o.b1:o.g1                1/4
o.a1:o.c1:o.g1                1/4
o.b1:o.c1:o.g1                1/4
o.a1:o.d1:o.g1                1/4
o.b1:o.d1:o.g1                1/4
o.c1:o.d1:o.g1                1/2
o.a1:o.e1:o.g1                1/4
o.b1:o.e1:o.g1                1/2
o.c1:o.e1:o.g1                1/4
o.d1:o.e1:o.g1                1/4
o.a1:o.f1:o.g1                1/2
o.b1:o.f1:o.g1                1/4
o.c1:o.f1:o.g1                1/4
o.d1:o.f1:o.g1                1/4
o.e1:o.f1:o.g1                1/4
o.a1:o.b1:o.c1:o.d1           1/4
o.a1:o.b1:o.c1:o.e1           1/4
o.a1:o.b1:o.d1:o.e1           1/4
o.a1:o.c1:o.d1:o.e1           1/4
o.b1:o.c1:o.d1:o.e1           1/4
o.a1:o.b1:o.c1:o.f1           1/4
o.a1:o.b1:o.d1:o.f1           1/4
o.a1:o.c1:o.d1:o.f1           1/4
o.b1:o.c1:o.d1:o.f1           1/4
o.a1:o.b1:o.e1:o.f1           1/4
o.a1:o.c1:o.e1:o.f1           1/4
o.b1:o.c1:o.e1:o.f1           1/4
o.a1:o.d1:o.e1:o.f1           1/4
o.b1:o.d1:o.e1:o.f1           1/4
o.c1:o.d1:o.e1:o.f1           1/4
o.a1:o.b1:o.c1:o.g1           1/4
o.a1:o.b1:o.d1:o.g1           1/4
o.a1:o.c1:o.d1:o.g1           1/4
o.b1:o.c1:o.d1:o.g1           1/4
o.a1:o.b1:o.e1:o.g1           1/4
o.a1:o.c1:o.e1:o.g1           1/4
o.b1:o.c1:o.e1:o.g1           1/4
o.a1:o.d1:o.e1:o.g1           1/4
o.b1:o.d1:o.e1:o.g1           1/4
o.c1:o.d1:o.e1:o.g1           1/4
o.a1:o.b1:o.f1:o.g1           1/4
o.a1:o.c1:o.f1:o.g1           1/4
o.b1:o.c1:o.f1:o.g1           1/4
o.a1:o.d1:o.f1:o.g1           1/4
o.b1:o.d1:o.f1:o.g1           1/4
o.c1:o.d1:o.f1:o.g1           1/4
o.a1:o.e1:o.f1:o.g1           1/4
o.b1:o.e1:o.f1:o.g1           1/4
o.c1:o.e1:o.f1:o.g1           1/4
o.d1:o.e1:o.f1:o.g1           1/4
o.a1:o.b1:o.c1:o.d1:o.e1      1/4
o.a1:o.b1:o.c1:o.d1:o.f1      1/4
o.a1:o.b1:o.c1:o.e1:o.f1      1/4
o.a1:o.b1:o.d1:o.e1:o.f1      1/4
o.a1:o.c1:o.d1:o.e1:o.f1      1/4
o.b1:o.c1:o.d1:o.e1:o.f1      1/4
o.a1:o.b1:o.c1:o.d1:o.g1      1/4
o.a1:o.b1:o.c1:o.e1:o.g1      1/4
o.a1:o.b1:o.d1:o.e1:o.g1      1/4
o.a1:o.c1:o.d1:o.e1:o.g1      1/4
o.b1:o.c1:o.d1:o.e1:o.g1      1/4
o.a1:o.b1:o.c1:o.f1:o.g1      1/4
o.a1:o.b1:o.d1:o.f1:o.g1      1/4
o.a1:o.c1:o.d1:o.f1:o.g1      1/4
o.b1:o.c1:o.d1:o.f1:o.g1      1/4
o.a1:o.b1:o.e1:o.f1:o.g1      1/4
o.a1:o.c1:o.e1:o.f1:o.g1      1/4
o.b1:o.c1:o.e1:o.f1:o.g1      1/4
o.a1:o.d1:o.e1:o.f1:o.g1      1/4
o.b1:o.d1:o.e1:o.f1:o.g1      1/4
o.c1:o.d1:o.e1:o.f1:o.g1      1/4
o.a1:o.b1:o.c1:o.d1:o.e1:o.f1 1/4
o.a1:o.b1:o.c1:o.d1:o.e1:o.g1 1/4
```

```
o.a1:o.b1:o.c1:o.d1:o.f1:o.g1       1/4
o.a1:o.b1:o.c1:o.e1:o.f1:o.g1       1/4
o.a1:o.b1:o.d1:o.e1:o.f1:o.g1       1/4
o.a1:o.c1:o.d1:o.e1:o.f1:o.g1       1/4
o.b1:o.c1:o.d1:o.e1:o.f1:o.g1       1/4
o.a1:o.b1:o.c1:o.d1:o.e1:o.f1:o.g1  1/4

[1] "vif() couldn't be used to find variable inflation because there are aliased coefficients in the
[1] "model. Thus alias() was used to look for correlated variables. Many variables were correlated.
[1] "and ABCDEG had some collinearity."
```

In [ ]: