

Review Problem Set 4 Solutions, Fall 2021

Problem i (from Kirk's book)

```
Kirk1 <- scan("c:/temp/KirkTable1431.txt",  
what = list(a=0,b=0,c=0,d=0,y=0),multi.line=F)
```

```
Kirk1 <- data.frame(Kirk1)
```

```
Kirk1$a <- as.factor(Kirk1$a)
```

```
Kirk1$b <- as.factor(Kirk1$b)
```

```
Kirk1$c <- as.factor(Kirk1$c)
```

```
Kirk1$d <- as.factor(Kirk1$d)
```

```
#head(Kirk1)
```

```
Kirk1.lm <- lm(y ~ a*b*c, data=Kirk1)  
anova(Kirk1.lm)
```

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

```
##
```

```
## Response: y
```

##	Df	Sum Sq	Mean Sq	F value	Pr(>F)
## a	1	1.00	1.000	0.8889	0.37337
## b	1	72.25	72.250	64.2222	4.313e-05 ***
## c	1	6.25	6.250	5.5556	0.04617 *
## a:b	1	6.25	6.250	5.5556	0.04617 *
## a:c	1	6.25	6.250	5.5556	0.04617 *
## b:c	1	9.00	9.000	8.0000	0.02220 *
## a:b:c	1	1.00	1.000	0.8889	0.37337
## Residuals	8	9.00	1.125		

```
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

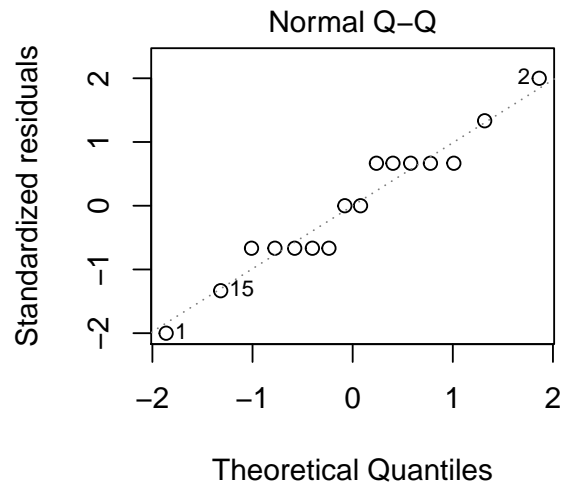
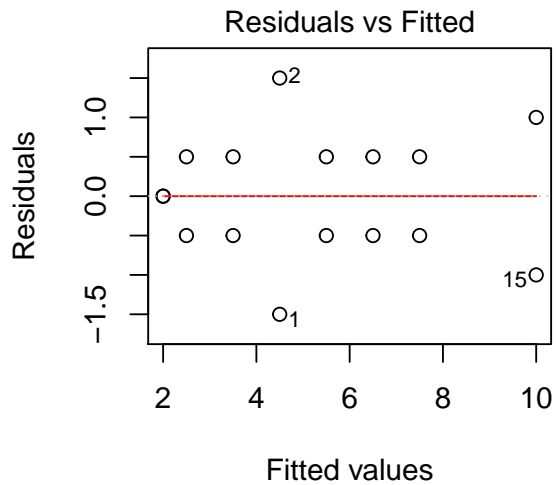
```
# Alias structure (through 2-way interactions) is:
```

```
# A = BCD, B = ACD, C = ABD, D = ABC
```

```
# AB = CD, AC = BD, AD = BC
```

```
par(mfrow=c(1,2))
```

```
plot(Kirk1.lm, which = c(1,2))
```



```
par(mfrow=c(1,1))
```

This is a 2^{4-1} fractional factorial design using the defining relation $I = ABCD$, with two replicates per combination. Since there are significant results for B and C, we might conclude that BC is significant instead of AD, but the significant effects of AB (or CD) and AC (or BD) are difficult to interpret. Given the small data set, the residual plots look reasonable.

Problem ii

```
#Problem ii: I = ABD = -CDE so we can use the embedded 2^3 ABC design
```

#	A	B	C	D = AB	E = -CD	combination
#(1)	-	-	-	+	+	de
# a	+	-	-	-	-	a
# b	-	+	-	-	-	b
#ab	+	+	-	+	+	abde
# c	-	-	+	+	-	cd
#ac	+	-	+	-	+	ace
#bc	-	+	+	-	+	bce
#abc	+	+	+	+	-	abcd

```
# I = ABD = -CDE = -ABCE, so the resolution is 3 (III)
```

```
# alias structure (excluding signs) :
```

```
# A = BD = ACDE = BCE
# B = AD = BCDE = ACE
# C = ABCD = DE = ABE
# D = AB = CE = ABCDE
# E = ABDE = CD = ABC
```

```
# AC = BCD = ADE = BE
# AE = BDE = ACD = BC
```

Problem iii (Problem 18.6)

```
Prb18.6 <- read.table("http://www.stat.umn.edu/~gary/book/fcdae.data/pr18.6",header=TRUE)

Prb18.6$A <- as.factor(Prb18.6$A)
Prb18.6$B <- as.factor(Prb18.6$B)
Prb18.6$C <- as.factor(Prb18.6$C)
Prb18.6$D <- as.factor(Prb18.6$D)
Prb18.6$E <- as.factor(Prb18.6$E)
Prb18.6$F <- as.factor(Prb18.6$F)
Prb18.6$G <- as.factor(Prb18.6$G)
Prb18.6$H <- as.factor(Prb18.6$H)

# pick four letters not involved in the entire defining relation, which is
# I = ABCD = BCEF = ACEG = BCEH = ADEF = BDEG = ADEH = ABFG = FH = ABGH =
#     CDGF = ABCDFH = CDGH = ACEFGH = BDEFGH

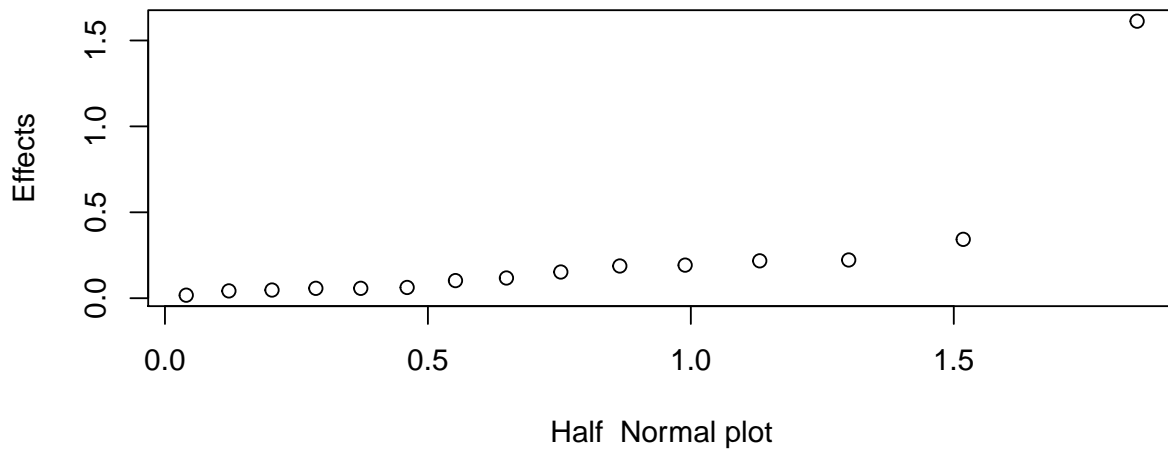
Prb18.6.aov <- aov(lm(thickness ~ A*B*C*E, data = Prb18.6))

library(gplots)

##
## Attaching package: 'gplots'

## The following object is masked from 'package:stats':
##
##     lowess

qqnorm(Prb18.6.aov)
```



```
anova(lm(thickness ~ A*B*C*E, data = Prb18.6))
```

```
## Warning in anova.lm(lm(thickness ~ A * B * C * E, data = Prb18.6)): ANOVA F-
## tests on an essentially perfect fit are unreliable
```

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

```
##
```

```
## Response: thickness
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
## A	1	0.02326	0.02326		
## B	1	0.00331	0.00331		
## C	1	0.04951	0.04951		
## E	1	0.00226	0.00226		
## A:B	1	0.00391	0.00391		
## A:C	1	0.03516	0.03516		
## B:C	1	0.01051	0.01051		
## A:E	1	0.00331	0.00331		
## B:E	1	0.01381	0.01381		
## C:E	1	0.00181	0.00181		
## A:B:C	1	2.60016	2.60016		
## A:B:E	1	0.03706	0.03706		
## A:C:E	1	0.04731	0.04731		
## B:C:E	1	0.11731	0.11731		
## A:B:C:E	1	0.00031	0.00031		
## Residuals	0	0.00000			

```
# with this command, you can click on and label points
# to get out of this mode, you can select the "R Console" window and hit the Escape button
# (this command does not work if not interactive)
#if (dev.interactive()) qqnorm(Prb18.6.aov, label=TRUE)
```

```

# now fitting the model with only "active-appearing" effects

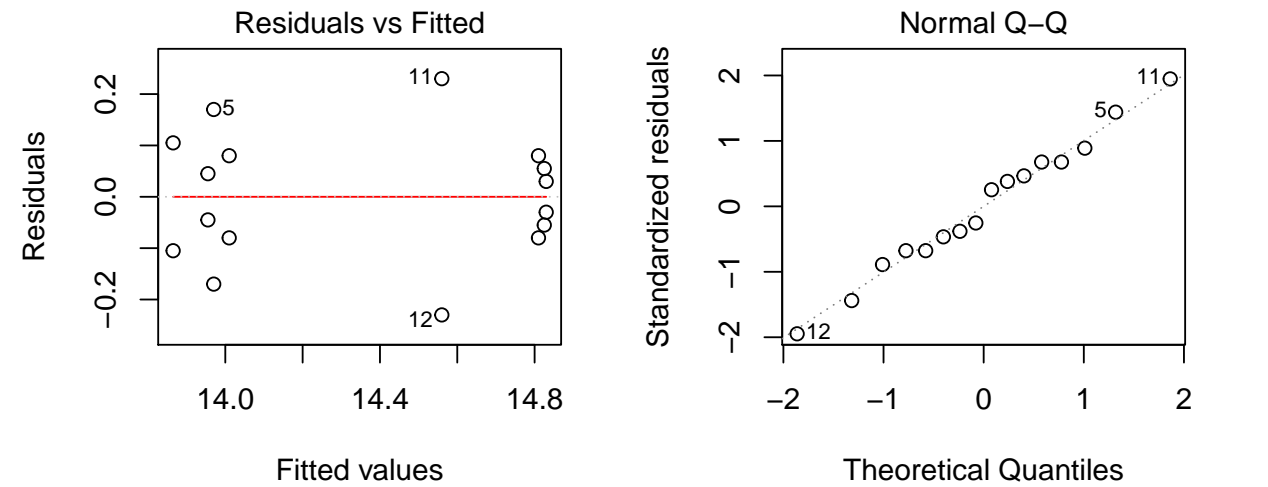
Prb18.6.lm2 <- lm(thickness ~ A*B*C, data = Prb18.6)

anova(Prb18.6.lm2)

## Analysis of Variance Table
##
## Response: thickness
##          Df Sum Sq Mean Sq F value    Pr(>F)
## A           1  0.02326  0.02326   0.8337   0.3879
## B           1  0.00331  0.00331   0.1185   0.7395
## C           1  0.04951  0.04951   1.7748   0.2195
## A:B         1  0.00391  0.00391   0.1400   0.7180
## A:C         1  0.03516  0.03516   1.2604   0.2941
## B:C         1  0.01051  0.01051   0.3767   0.5564
## A:B:C       1  2.60016  2.60016  93.2164 1.103e-05 ***
## Residuals   8  0.22315  0.02789
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

par(mfrow=c(1,2))
plot(Prb18.6.lm2, which = c(1,2))

```



```

par(mfrow=c(1,1))

```

First we can list the entire defining relation with the result shown above. Although the author claims that this design has resolution IV, since FH appears in the defining relation it is actually only resolution II. A half-normal plot shows one large effect. One line of code above (commented out) can be used to identify the effect interactively, but otherwise we can fit a saturated model (with 0 error df) to the embedded design

ABCE, and we can identify the large effect as ABC. We can also fit a reduced ABC model to get F tests and residual plots. The ABC effect is aliased with the D main effect, so we conclude that it is an important effect.

Problem iv (Sales and Assets data)

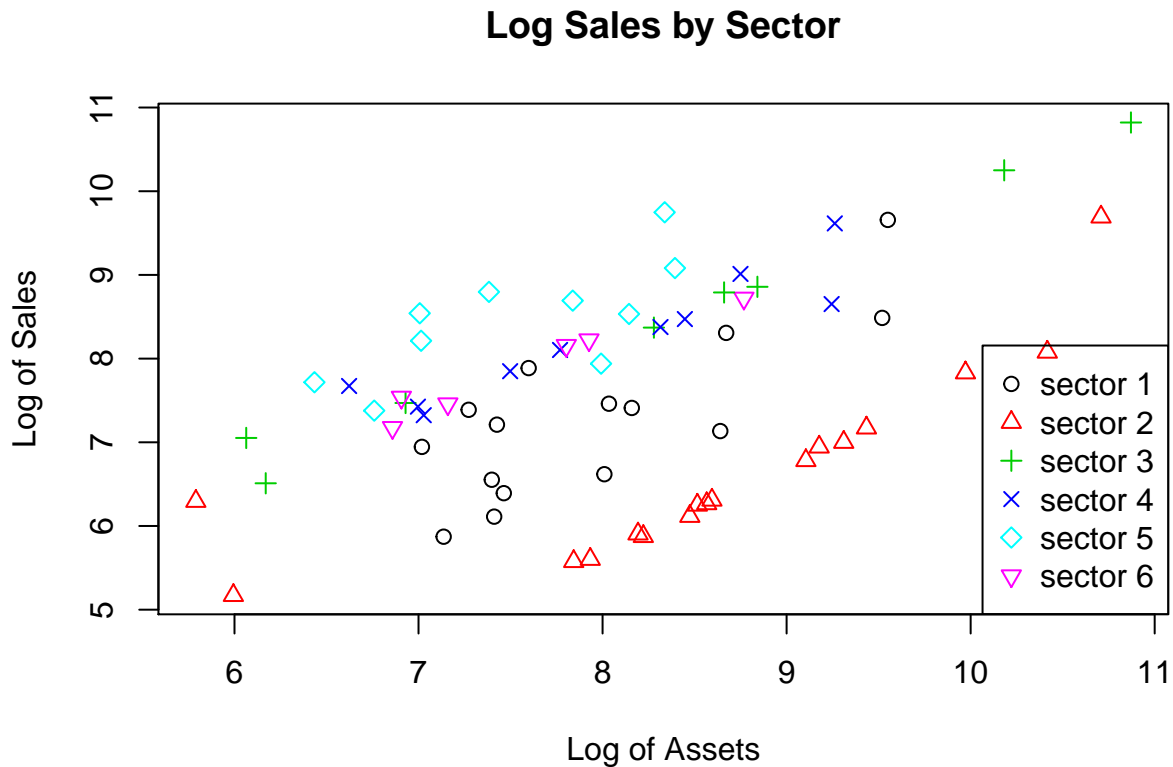
```
# first read the data directly from my website

Forbes <- read.table("http://webpages.uidaho.edu/~chrisw/stat507live/companies.txt",header=TRUE)

# head(Forbes)

# look at the data

plot(Forbes$logassets,Forbes$logsales,pch=as.numeric(Forbes$sector),
     col=as.numeric(Forbes$sector),ylab="Log of Sales",xlab="Log of Assets")
legend("bottomright",legend=c("sector 1","sector 2","sector 3","sector 4","sector 5","sector 6"),
      levels(Forbes$sector),pch=1:6,col=1:6)
title("Log Sales by Sector")
```



The scatterplot shows a clear linear relationship between logSales and logAssets, with likely different intercepts in the different sectors.

```
# standard ANCOVA assuming equal slopes
Forbes.lm1 <- lm(lsales ~ lassets +sector , data=Forbes)
anova(Forbes.lm1)
```

```
## Analysis of Variance Table
##
## Response: lsales
##           Df Sum Sq Mean Sq F value    Pr(>F)
## lassets    1 23.291  23.2912  85.719 4.249e-13 ***
## sector     5 56.400  11.2800  41.514 < 2.2e-16 ***
## Residuals 59 16.031   0.2717
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
summary(Forbes.lm1)
```

```
##
## Call:
## lm(formula = lsales ~ lassets + sector, data = Forbes)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.88028 -0.32250 -0.04517  0.20966  1.82195
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    1.17104    0.50027   2.341 0.022641 *
## lassets         0.77001    0.06058  12.711 < 2e-16 ***
## sectorFinance  -1.15605    0.18879  -6.123 8.06e-08 ***
## sectorHiTech    0.99185    0.22891   4.333 5.80e-05 ***
## sectorManufacturing 0.92444    0.21282   4.344 5.59e-05 ***
## sectorRetail    1.49490    0.21435   6.974 2.99e-09 ***
## sectorTransportation 0.87832    0.25286   3.473 0.000969 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5213 on 59 degrees of freedom
## Multiple R-squared:  0.8325, Adjusted R-squared:  0.8155
## F-statistic: 48.88 on 6 and 59 DF,  p-value: < 2.2e-16
```

```
# calculate adjustedMeans and standard errors from the effects package
library(effects)
```

```
## Loading required package: carData
## lattice theme set by effectsTheme()
## See ?effectsTheme for details.
```

```
adjustedMeans <- effect("sector",Forbes.lm1)
adjustedMeans
```

```
##
## sector effect
## sector
##      Energy      Finance      HiTech  Manufacturing      Retail
##      7.380207    6.224159    8.372060    8.304651    8.875104
## Transportation
##      8.258530
```

```
adjustedMeans$se
```

```
## [1] 0.1347544 0.1305734 0.1846384 0.1648937 0.1679773 0.2148943
```

```
# information for calculating adjusted means
```

```
summary(Forbes)
```

```
##          company      lsales      lassets
## AmericanElectricPower: 1  Min.   : 5.170  Min.   : 5.790
## AmericanSavgsBankFSB  : 1  1st Qu.: 6.823  1st Qu.: 7.188
## AMR                    : 1  Median : 7.504  Median : 8.022
## AppleComputer         : 1  Mean   : 7.649  Mean   : 8.064
## ArmstrongWorldInd     : 1  3rd Qu.: 8.483  3rd Qu.: 8.669
## BankSouth             : 1  Max.   :10.821  Max.   :10.871
## (Other)               :60
##          sector
## Energy           :15
## Finance          :17
## HiTech           : 8
## Manufacturing    :10
## Retail           :10
## Transportation: 6
##
```

```
tapply(Forbes$lsales,Forbes$sector,summary)
```

```
## $Energy
##   Min. 1st Qu. Median   Mean 3rd Qu.  Max.
##   5.872 6.586  7.211  7.296  7.674  9.657
##
## $Finance
##   Min. 1st Qu. Median   Mean 3rd Qu.  Max.
##   5.170 5.905  6.295  6.639  7.000  9.693
##
## $HiTech
##   Min. 1st Qu. Median   Mean 3rd Qu.  Max.
##   6.510 7.365  8.580  8.515  9.206 10.821
##
## $Manufacturing
```



```
##   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   7.324  7.717  8.240  8.250  8.607  9.615
##
## $Retail
##   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   7.378  8.008  8.537  8.464  8.771  9.748
##
## $Transportation
##   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   7.175  7.478  7.847  7.879  8.206  8.721
```

```
tapply(Forbes$assets,Forbes$sector,summary)
```

```
## $Energy
##   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   7.018  7.405  7.598  7.954  8.399  9.550
##
## $Finance
##   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   5.790  8.193  8.566  8.603  9.310 10.709
##
## $HiTech
##   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   6.064  6.740  8.470  8.250  9.177 10.871
##
## $Manufacturing
##   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   6.623  7.146  8.042  7.993  8.674  9.262
##
## $Retail
##   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   6.434  7.009  7.610  7.530  8.106  8.393
##
## $Transportation
##   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   6.859  6.970  7.481  7.570  7.895  8.768
```

```
# (for the HiTech sector)  8.52 - .77 * (8.25 -8.06) =
#      8.52 - .77 * 0.19 = 8.52 - .15 = 8.37
```

```
# multiple comparisons using the multcomp package
```

```
library(multcomp)
```

```
## Warning: package 'multcomp' was built under R version 3.6.3
```

```
## Loading required package: mvtnorm
```

```
## Loading required package: survival
```

```
## Loading required package: TH.data
```

```
## Warning: package 'TH.data' was built under R version 3.6.3
```

```
## Loading required package: MASS
```

```
##
```

```
## Attaching package: 'TH.data'
```

```
## The following object is masked from 'package:MASS':
```

```
##
```

```
## geyser
```

```
summary(glht(Forbes.lm1, linfct=mcp(sector="Tukey")))
```

```
##
```

```
## Simultaneous Tests for General Linear Hypotheses
```

```
##
```

```
## Multiple Comparisons of Means: Tukey Contrasts
```

```
##
```

```
##
```

```
## Fit: lm(formula = lsales ~ lassets + sector, data = Forbes)
```

```
##
```

```
## Linear Hypotheses:
```

	Estimate	Std. Error	t value	Pr(> t)	
## Finance - Energy == 0	-1.15605	0.18879	-6.123	<0.001	***
## HiTech - Energy == 0	0.99185	0.22891	4.333	<0.001	***
## Manufacturing - Energy == 0	0.92444	0.21282	4.344	<0.001	***
## Retail - Energy == 0	1.49490	0.21435	6.974	<0.001	***
## Transportation - Energy == 0	0.87832	0.25286	3.473	0.0117	*
## HiTech - Finance == 0	2.14790	0.22451	9.567	<0.001	***
## Manufacturing - Finance == 0	2.08049	0.21099	9.860	<0.001	***
## Retail - Finance == 0	2.65094	0.21766	12.179	<0.001	***
## Transportation - Finance == 0	2.03437	0.25531	7.968	<0.001	***
## Manufacturing - HiTech == 0	-0.06741	0.24774	-0.272	0.9998	
## Retail - HiTech == 0	0.50304	0.25107	2.004	0.3481	
## Transportation - HiTech == 0	-0.11353	0.28451	-0.399	0.9986	
## Retail - Manufacturing == 0	0.57045	0.23480	2.430	0.1601	
## Transportation - Manufacturing == 0	-0.04612	0.27040	-0.171	1.0000	
## Transportation - Retail == 0	-0.61657	0.26919	-2.290	0.2108	

```
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
## (Adjusted p values reported -- single-step method)
```

```
# check equal slopes assumption
```

```
Forbes.lm2 <- lm(lsales ~ lassets + sector + lassets:sector, data=Forbes)
```

```
anova(Forbes.lm2)
```

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

```
##
```

```
## Response: lsales
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
## lassets	1	23.291	23.2912	82.4047	1.873e-12 ***

```
## sector          5 56.400 11.2800 39.9089 < 2.2e-16 ***
## lassets:sector  5  0.768  0.1537  0.5438    0.7423
## Residuals      54 15.263  0.2826
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# how do the logAssets means compare?
anova(lm(lassets ~ sector, data=Forbes))
```

```
## Analysis of Variance Table
##
## Response: lassets
##          Df Sum Sq Mean Sq F value Pr(>F)
## sector    5  9.754  1.9507  1.5807 0.1793
## Residuals 60 74.044  1.2341
```

The ANCOVA shows a significant effect of sector with $F = 41.5$ on 5 and 59 df and $P < .0001$. The logAssets covariate was also significant and accounted for a large SS relative to its 1 df. The adjusted means show an ordering of: Finance < Energy < Transp. < Manufac. < HiTech < Retail

For the HiTech adjusted mean we use:

$$\bar{y}_{HiTech(adj)} = \bar{y}_{HiTech} - \hat{\beta}(\bar{x}_{HiTech.} - \bar{x}_{..}),$$

giving

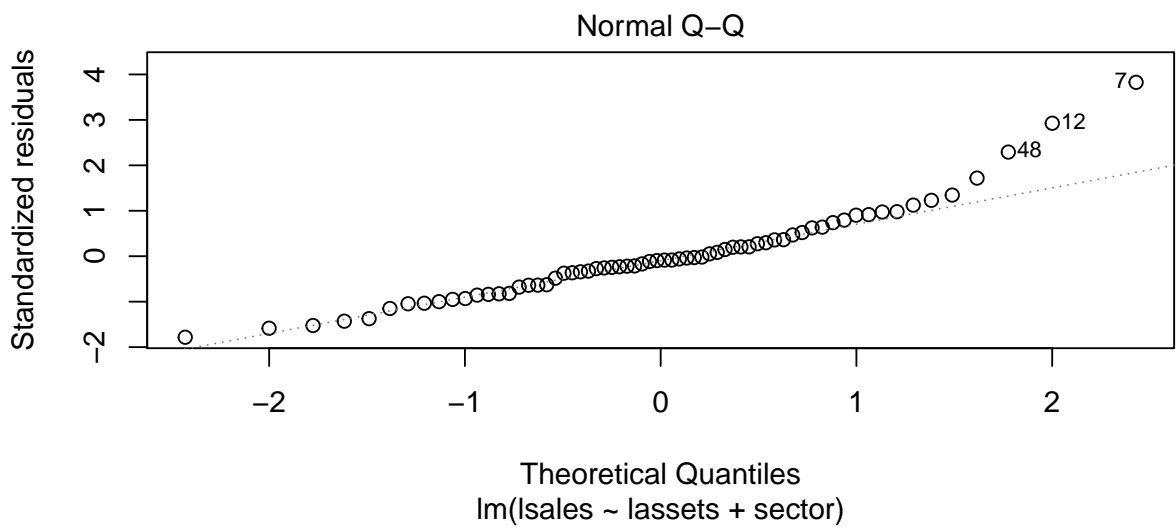
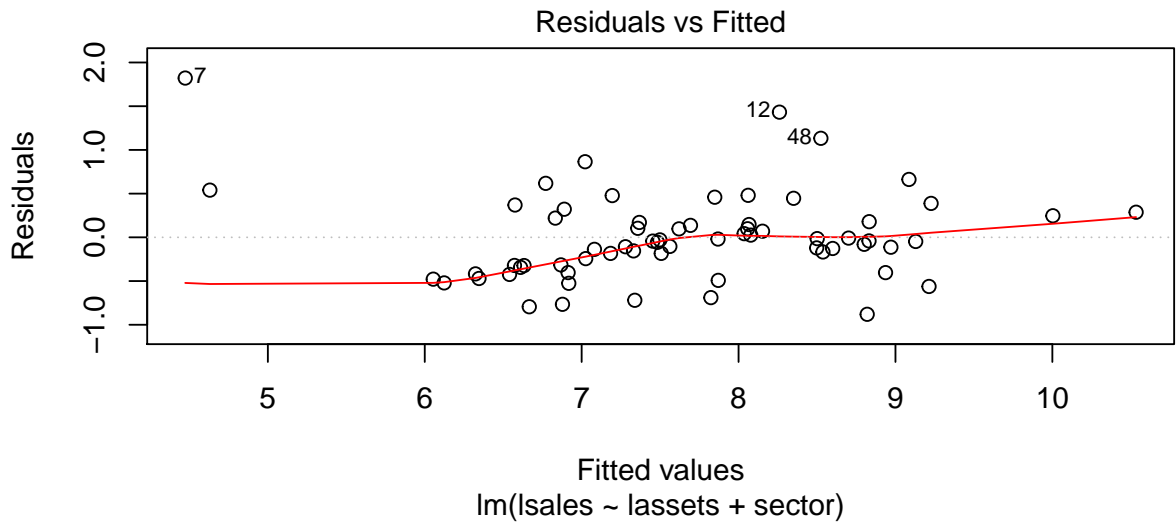
$$\bar{y}_{HiTech(adj)} = 8.515 - 0.77(8.25 - 8.064) = 8.37.$$

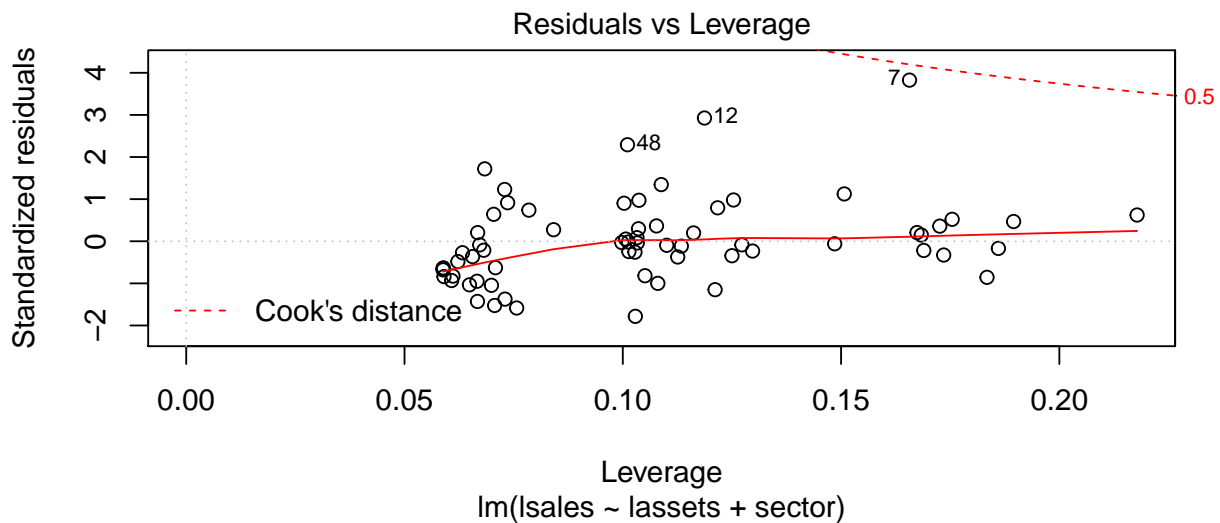
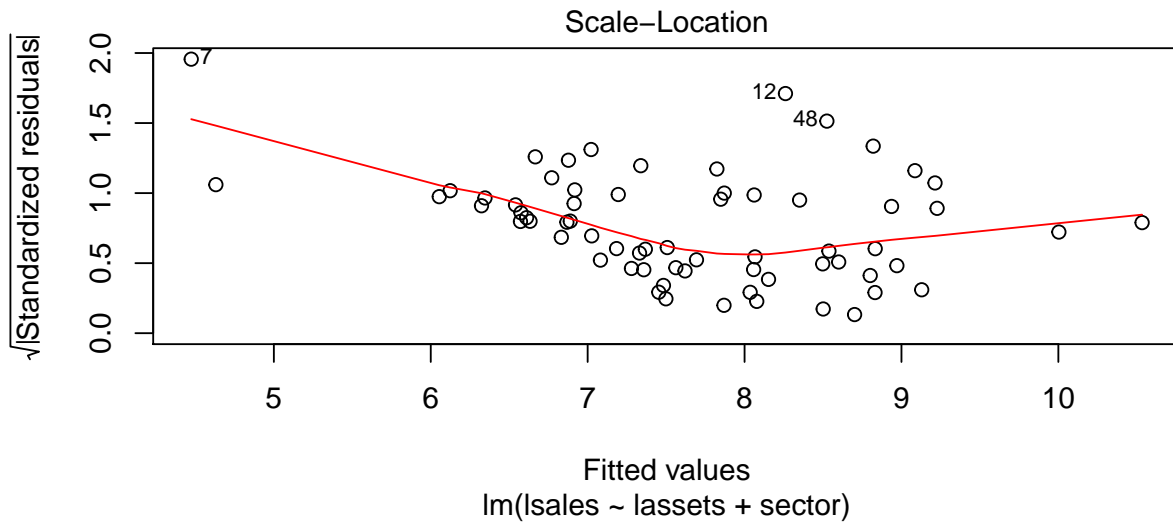
The Tukey multiple comparison tests show that the adjusted Energy and Finance means differ from each other sector, but the other four sectors do not differ from each other.

We then test for parallelism, and find that the null hypothesis of equal slopes among sectors cannot be rejected, with $F = .54$ on 5 and 54 df, $P = .74$. So we can be comfortable with the equal slopes model.

Finally we do an ANOVA of the covariate to see if its' mean differs across sectors. If we see a significant difference it may focus our attention on the question of whether the “treatment affects the covariate” but this is observational data so the question is less relevant. In this case the means of logAssets do not differ significantly between sectors.

```
# check assumptions
#par(mfrow=c(2,2))
plot(Forbes.lm1)
```





```
#par(mfrow=c(1,1))
```

Checking distributional assumptions yields some problems. The residual x predicted plot shows one outlier (obs #7) with a low predicted value. This observation is seen with some others as falling off of the normal line, and it has large leverage and nearly crosses into the region of high Cook's distance values. Further study of these data would include an examination of that observation and possibly a Box-Cox estimate of a power transformation for Sales on its' original scale.

Problem v (Problems 18.1a, 18.2d, and 18.2h)

Prob. 18.1a: There is a dual randomization, first to the stress treatment, then for each student the order of group vs. individual activity is randomized, so this is a split plot design. The whole plot treatment is completely randomized, as children are randomized to a stress level. Children are the whole plot units and the factors are stress level and activity. The skeleton looks like:

#	Source	df	
#	Stress	1	
#	Child(Stress)	30	<- whole plot error term
#	Activity	1	
#	Stress*Activity	1	
#	Error	30	<- split plot error term
#	Total	63	

Prob. 18.2d: Here rooms are blocks that receive treatments (different texts). But each room can only see three of the four treatment levels. With four rooms we can construct a balanced incomplete block design (BIBD) with $g = 4$, $k = 3$, $b = 4$, and $r = 3$.

Prob. 18.2h: The treatment factor is pavement type with 5 levels. Also there is variability due to weather pattern (5 levels) and traffic volume (5 levels). With a budget allowing 25 observations, a Latin-square design can be used with rows and columns consisting of weather and traffic levels, and the letters in the square are the pavement types.