Booklet of Code and Output
for
STAC32 Midterm Exam

December 19, 2016

```
library(tidyverse)
## -- Attaching packages ---------------------------------
tidyverse 1.2.1 --
## v ggplot2 3.0.0     v purrr   0.2.5
## v tibble  1.4.2     v dplyr   0.7.6
## v tidyr   0.8.1     v stringr 1.3.1
## v readr   1.1.1     v forcats 0.3.0
## -- Conflicts ------------------------------------
tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

Figure 1: R packages needed

2016 version of the data:

```
  1   F    35   17   7    2   2
 17   M    50   14   5    5   3
 33   F    45    6   7    2   7
 49   M    24   14   7    5   7
 65   F    52    9   4    7   7
 81   M    44   11   7    7   7
  2   F    34   17   6    5   3
 18   M    40   14   7    5   2
 34   F    47    6   6    5   6
 50   M    35   17   5    7   5
```

2017 version of the data:

```
 1 F 35 17 7 2 2
 17 M 50 14 5 5 3
 33 F 45 6 7 2 7
 49 M 24 14 7 5 7
 65 F 52 9 4 7 7
 81 M 44 11 7 7 7
 2 F 34 17 6 5 3
 18 M 40 14 7 5 2
 34 F 47 6 6 5 6
 50 M 35 17 5 7 5
```

Figure 2: Survey data

```
weightloss=read.table("weightloss.txt",header=T)
weightloss
##    client before after
## 1       1    210   197
## 2       2    205   195
## 3       3    193   191
## 4       4    182   174
## 5       5    259   236
## 6       6    239   226
## 7       7    164   157
## 8       8    197   196
## 9       9    222   201
## 10     10    211   196
## 11     11    187   181
## 12     12    175   164
## 13     13    186   181
## 14     14    243   229
## 15     15    246   231
```

Figure 3: Weight loss data

```
wtloss2=weightloss %>% gather(when,weight,before:after)
```

The actual spaghetti plot is printed in colour at the end of this booklet.

Figure 4: Spaghetti plot preliminaries

```
mark group
4 exam
9 exam
12 exam
8 exam
9 exam
13 exam
12 exam
13 exam
13 exam
7 exam
6 exam
7 threat
8 threat
7 threat
2 threat
6 threat
9 threat
7 threat
10 threat
5 threat
0 threat
10 threat
8 threat
```

Figure 5: Data for stereotype threat experiment

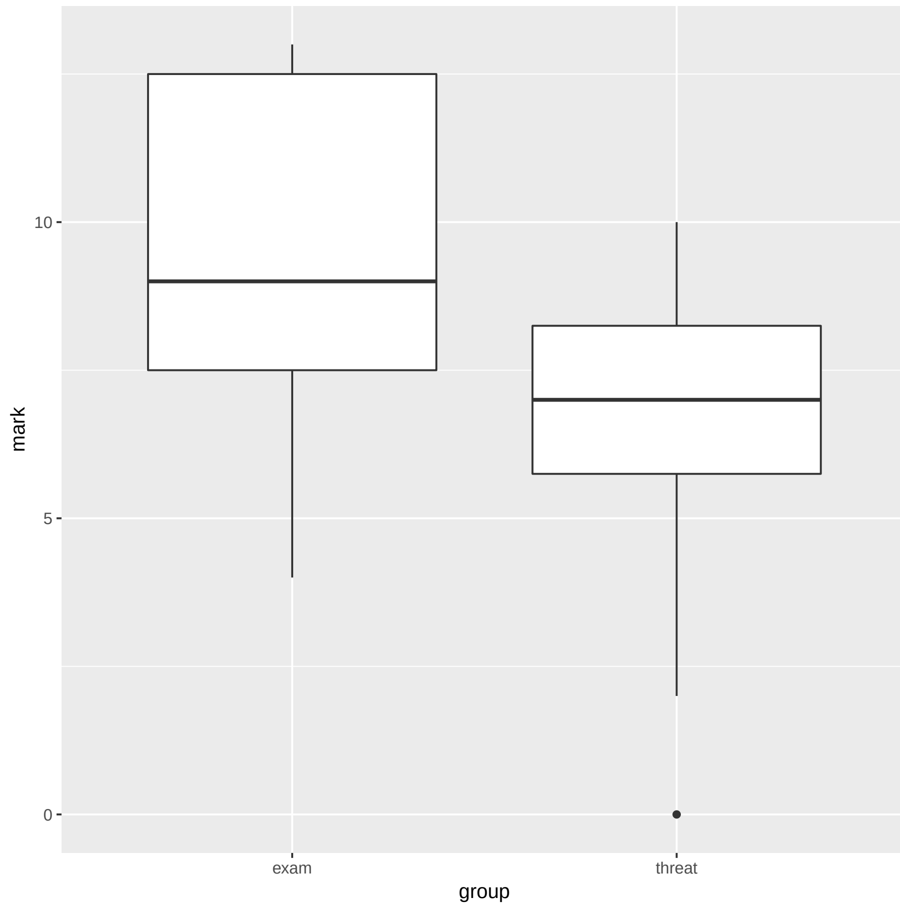

Figure 6: Boxplots of stereotype threat data

```
obs = stereo %>% group_by(group) %>%
    summarize(med=median(mark))
obs
## # A tibble: 2 x 2
##   group    med
##   <fct>  <dbl>
## 1 exam       9
## 2 threat     7
omd=obs$med[2]-obs$med[1]
omd
## [1] -2
```

Figure 7: Computations for stereotype threat data

```
rd=function(x) {
  sh=sample(x$group)
  med=aggregate(mark~sh,x,median)
  return(med$mark[2]-med$mark[1])
}
```

Figure 8: A function

```
randm.dist=replicate(1000,rd(stereo))
table(randm.dist<=omd)
##
## FALSE  TRUE
##   854   146
```

Figure 9: Randomization test

```
power.t.test(delta=10,sd=80,n=100,type="one.sample",alternative="one.sided")
##
##      One-sample t test power calculation
##
##               n = 100
##           delta = 10
##              sd = 80
##       sig.level = 0.05
##           power = 0.3433285
##     alternative = one.sided
```

Figure 10: Power analysis 1 for New England college

```
power.t.test(delta=530,sd=80,n=100,type="one.sample",alternative="one.sided")
##
##      One-sample t test power calculation
##
##               n = 100
##           delta = 530
##              sd = 80
##       sig.level = 0.05
##           power = 1
##     alternative = one.sided
```

Figure 11: Power analysis 2 for New England college

```
power.t.test(delta=10,sd=80,n=100,type="one.sample",alternative="two.sided")
##
##      One-sample t test power calculation
##
##               n = 100
##           delta = 10
##              sd = 80
##       sig.level = 0.05
##           power = 0.2351253
##     alternative = two.sided
```

Figure 12: Power analysis 3 for New England college

```
power.t.test(delta=530,sd=80,n=100,type="one.sample",alternative="two.sided")
##
##      One-sample t test power calculation
##
##               n = 100
##           delta = 530
##              sd = 80
##       sig.level = 0.05
##           power = 1
##     alternative = two.sided
```

Figure 13: Power analysis 4 for New England college

```
safelight=read.table("safelight.txt",header=T)
str(safelight)
## 'data.frame': 40 obs. of  2 variables:
##  $ treatment: Factor w/ 5 levels "AH","AL","BH",..: 5 5 5 5 5 5 5 5 2 2 ...
##  $ height   : num  32.9 36 34.8 32.4 32.8 ...
```

Figure 14: Structure of safelight data

```
ggplot(safelight,aes(x=treatment,y=height))+geom_boxplot()
```

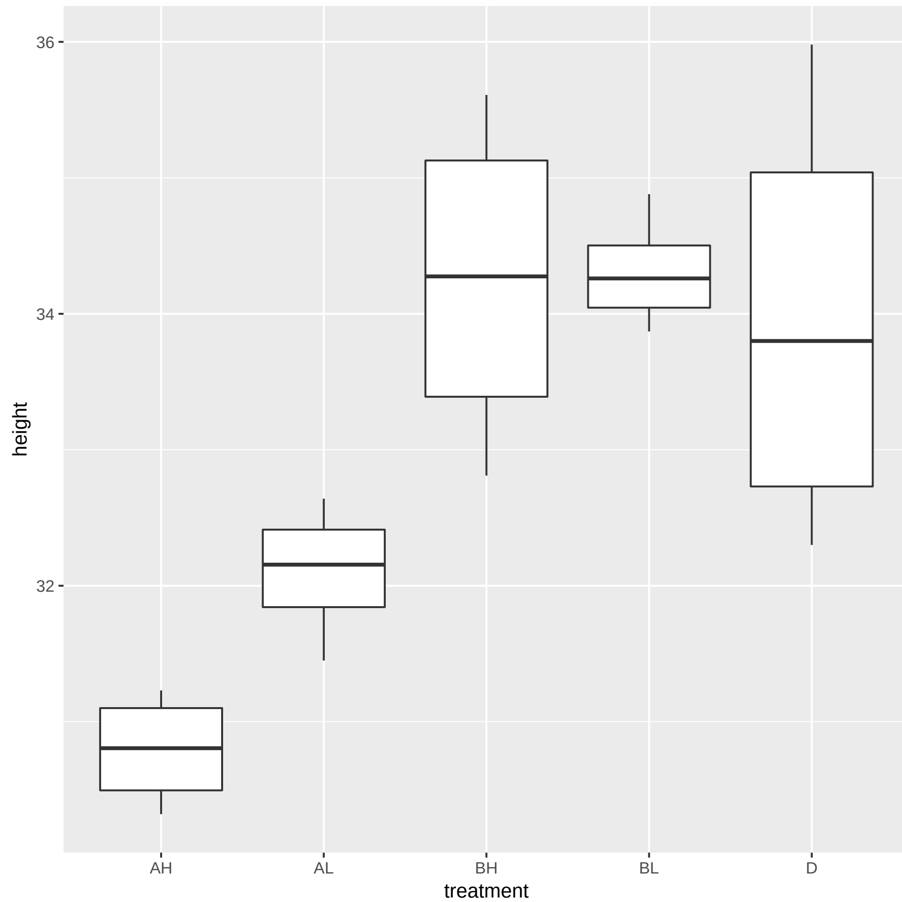

Figure 15: Boxplots of safelight data

```
safelight.1=aov(height~treatment,data=safelight)
summary(safelight.1)
##             Df Sum Sq Mean Sq F value   Pr(>F)
## treatment    4  78.94   19.73   24.07 1.24e-09 ***
## Residuals   35  28.69    0.82
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figure 16: Analysis of variance for safelight data

```
m=median(safelight$height)
tab=with(safelight,table(treatment,height<m))
tab
##
## treatment FALSE TRUE
##        AH     0    8
##        AL     0    8
##        BH     7    1
##        BL     8    0
##        D      5    3
chisq.test(tab)
## Warning in chisq.test(tab):  Chi-squared approximation may be
incorrect
##
##  Pearson's Chi-squared test
##
## data:  tab
## X-squared = 29, df = 4, p-value = 7.817e-06
```

Figure 17: Mood's median test for safelight data

```
data employees;
  infile '/home/ken/salaries.txt' firstobs=2;
  input salary degree experience supervised;

proc print data=employees(obs=20);
```

Obs	salary	degree	experience	supervised
1	58.8	3	4.49	0
2	34.8	1	2.92	0
3	163.7	3	29.54	42
4	70.0	3	9.92	0
5	55.5	3	0.14	0
6	85.0	2	15.96	4
7	34.0	1	2.27	0
8	29.7	1	1.20	0
9	56.1	2	5.33	3
10	70.6	3	15.74	0
11	74.2	1	22.46	2
12	34.1	1	3.16	0
13	31.6	1	2.62	0
14	65.5	1	15.06	5
15	57.2	3	2.92	0
16	60.3	3	2.26	0
17	41.8	1	9.76	1
18	76.5	3	14.71	4
19	122.1	3	21.76	10
20	85.9	3	15.63	8

Figure 18: Employee salaries data (some)

```
proc reg;
  model salary=degree experience supervised;
```

```
                          The REG Procedure
                            Model: MODEL1
                     Dependent Variable: salary

               Number of Observations Read          65
               Number of Observations Used          65
                          Analysis of Variance

                                  Sum of           Mean
Source                  DF        Squares         Square    F Value    Pr > F

Model                    3          39005          13002     128.35    <.0001
Error                   61     6179.05100      101.29592
Corrected Total         64          45184
             Root MSE            10.06459    R-Square     0.8632
             Dependent Mean      60.01846    Adj R-Sq     0.8565
             Coeff Var           16.76915
                          Parameter Estimates

                        Parameter       Standard
    Variable      DF     Estimate          Error    t Value    Pr > |t|

    Intercept      1     19.86899        3.87249       5.13      <.0001
    degree         1     11.34087        1.72365       6.58      <.0001
    experience     1      1.26085        0.22507       5.60      <.0001
    supervised     1      1.85315        0.22580       8.21      <.0001
```

Figure 19: Regression 1 for employee salaries data

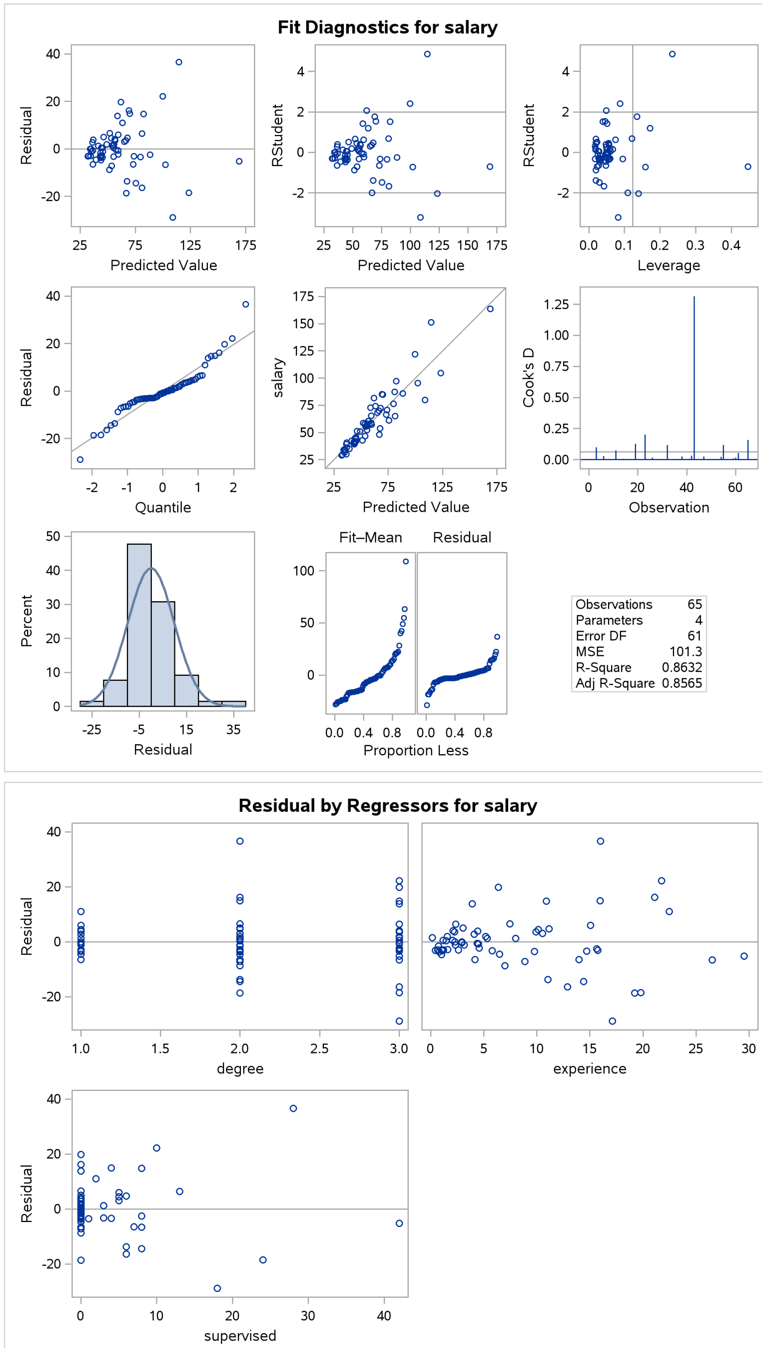

Figure 20: Regression 1 graphical output

```
proc transreg;
  model boxcox(salary)=identity(degree experience supervised);
```

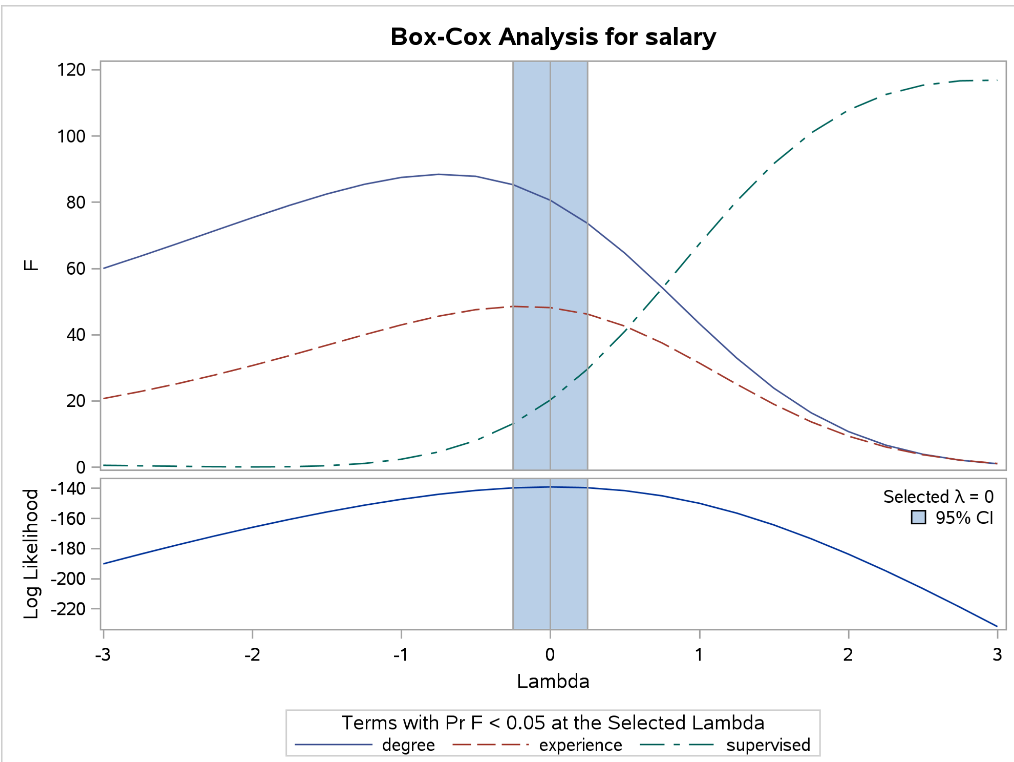

Figure 21: Output from `proc transreg`

```
data employees2;
  set employees;
  logsal=log(salary);

proc reg;
  model logsal=degree experience supervised;
```

```
                          The REG Procedure
                            Model: MODEL1
                     Dependent Variable: logsal

                  Number of Observations Read          65
                  Number of Observations Used          65
                         Analysis of Variance

                                       Sum of           Mean
Source                   DF           Squares         Square    F Value    Pr > F

Model                     3           8.06274        2.68758     114.24    <.0001
Error                    61           1.43513        0.02353
Corrected Total          64           9.49787
             Root MSE               0.15338    R-Square     0.8489
             Dependent Mean         4.01625    Adj R-Sq     0.8415
             Coeff Var              3.81909
                         Parameter Estimates

                         Parameter       Standard
    Variable      DF      Estimate          Error    t Value    Pr > |t|

    Intercept      1       3.28035        0.05902      55.58      <.0001
    degree         1       0.23573        0.02627       8.97      <.0001
    experience     1       0.02379        0.00343       6.94      <.0001
    supervised     1       0.01547        0.00344       4.50      <.0001
```

Figure 22: Regression 2

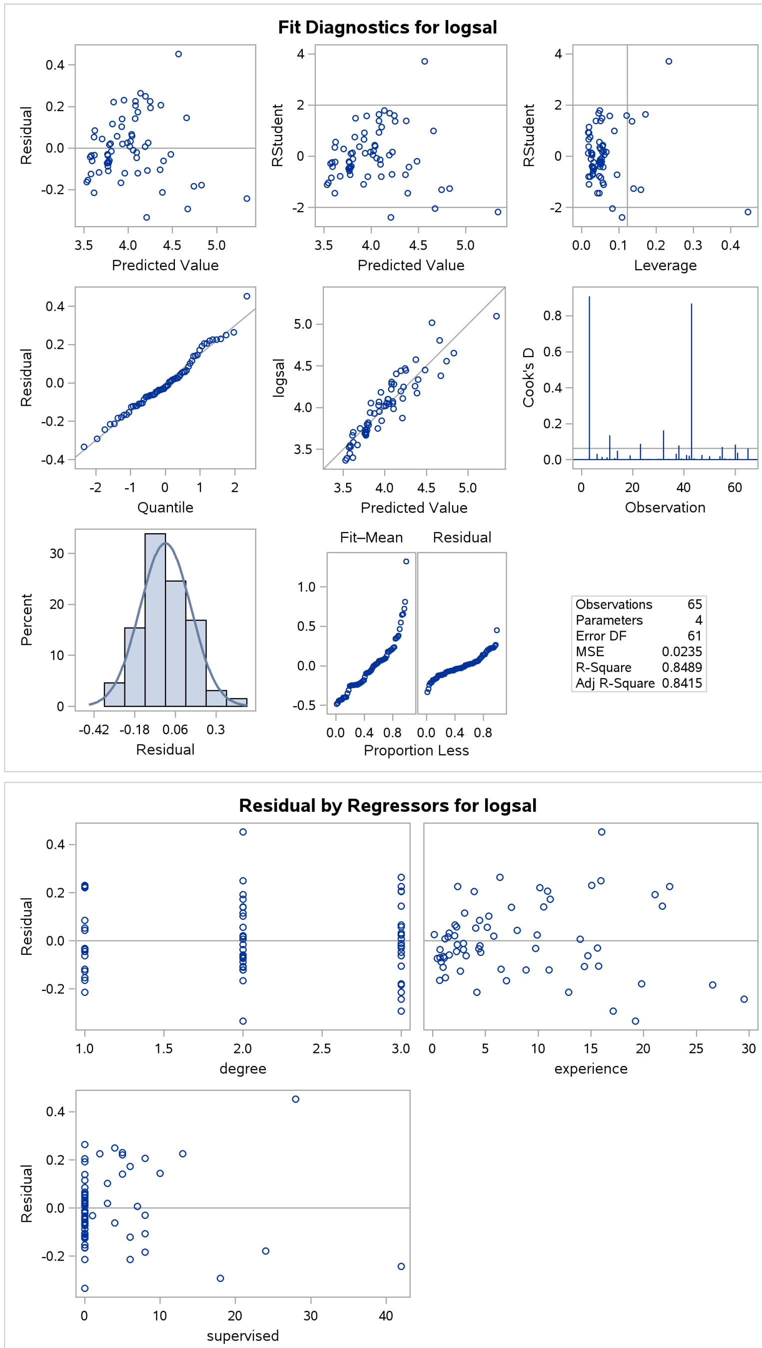

Figure 23: Regression 2 graphical output

```
davis2=read.csv("davis2.csv",header=T)
davis2 %>% select(Height,GPA,Sex,Alchol,momheight,dadheight) %>%
  head(20)
##    Height  GPA    Sex Alchol momheight dadheight
## 1    64.0 2.60 Female     15        64        70
## 2    69.0 2.70   Male     14        67        68
## 3    66.0 3.00 Female     NA        61        70
## 4    63.0 3.11 Female     10        62        68
## 5    72.0 3.40   Male     30        66        69
## 6    67.0 3.43 Female     20        68        69
## 7    69.0 3.70   Male     15        67        69
## 8    74.0 3.70   Male     15        69        76
## 9    72.0 3.77   Male      0        NA        72
## 10   63.0 3.50 Female      0        NA        NA
## 11   68.5 3.00   Male     NA        64        NA
## 12   70.0 3.00   Male      0        61        74
## 13   71.0 3.50   Male      0        NA        73
## 14   68.0 3.25   Male      0        63        73
## 15   60.0 2.83 Female      0        60        68
## 16   71.0 2.62   Male      0        61        67
## 17   68.0 3.15 Female     NA        67        72
## 18   67.0 4.20 Female      1        70        76
## 19   66.0 3.70 Female      0        60        71
## 20   69.0 4.38   Male      2        64        64
```

Figure 24: Cal-Davis data (some)

```
davis2 %>% filter(!is.na(GPA),
                  !is.na(Alchol),
                  !is.na(momheight),
                  !is.na(dadheight)) -> davis3
```

@

Figure 25: Cal-Davis data organization

```
height.1=lm(Height~Sex+GPA+Alchol+momheight+dadheight,data=davis3)
summary(height.1)
##
## Call:
## lm(formula = Height ~ Sex + GPA + Alchol + momheight + dadheight,
##     data = davis3)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.4526 -1.7467 -0.1142  1.5053 12.4837
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 35.35227    3.99558   8.848 1.42e-15 ***
## SexMale      5.30682    0.42538  12.476  < 2e-16 ***
## GPA         -0.31955    0.36611  -0.873   0.3841
## Alchol       0.01340    0.03158   0.424   0.6719
## momheight    0.20001    0.07888   2.536   0.0122 *
## dadheight    0.25674    0.05711   4.495 1.31e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.524 on 163 degrees of freedom
## Multiple R-squared:  0.5803,Adjusted R-squared:  0.5674
## F-statistic: 45.07 on 5 and 163 DF,  p-value: < 2.2e-16
```

Figure 26: Cal-Davis first regression

```
height.2=update(height.1,.~.-GPA-Alchol)
summary(height.2)
##
## Call:
## lm(formula = Height ~ Sex + momheight + dadheight, data = davis3)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.5755 -1.6840 -0.0808  1.4906 12.5341
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 34.43822    3.86730   8.905 9.35e-16 ***
## SexMale      5.38748    0.40459  13.316  < 2e-16 ***
## momheight    0.20372    0.07657   2.661  0.00857 **
## dadheight    0.25263    0.05683   4.446 1.60e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.517 on 165 degrees of freedom
## Multiple R-squared:  0.5774,Adjusted R-squared:  0.5697
## F-statistic: 75.15 on 3 and 165 DF,  p-value: < 2.2e-16
```

Figure 27: Cal-Davis second regression

```
anova(height.2,height.1)
## Analysis of Variance Table
##
## Model 1: Height ~ Sex + momheight + dadheight
## Model 2: Height ~ Sex + GPA + Alchol + momheight + dadheight
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1    165 1045.4
## 2    163 1038.3  2     7.143 0.5607 0.5719
```

Figure 28: Cal-Davis last output

```
ggplot(height.2,aes(x=.fitted,y=.resid))+geom_point()
```

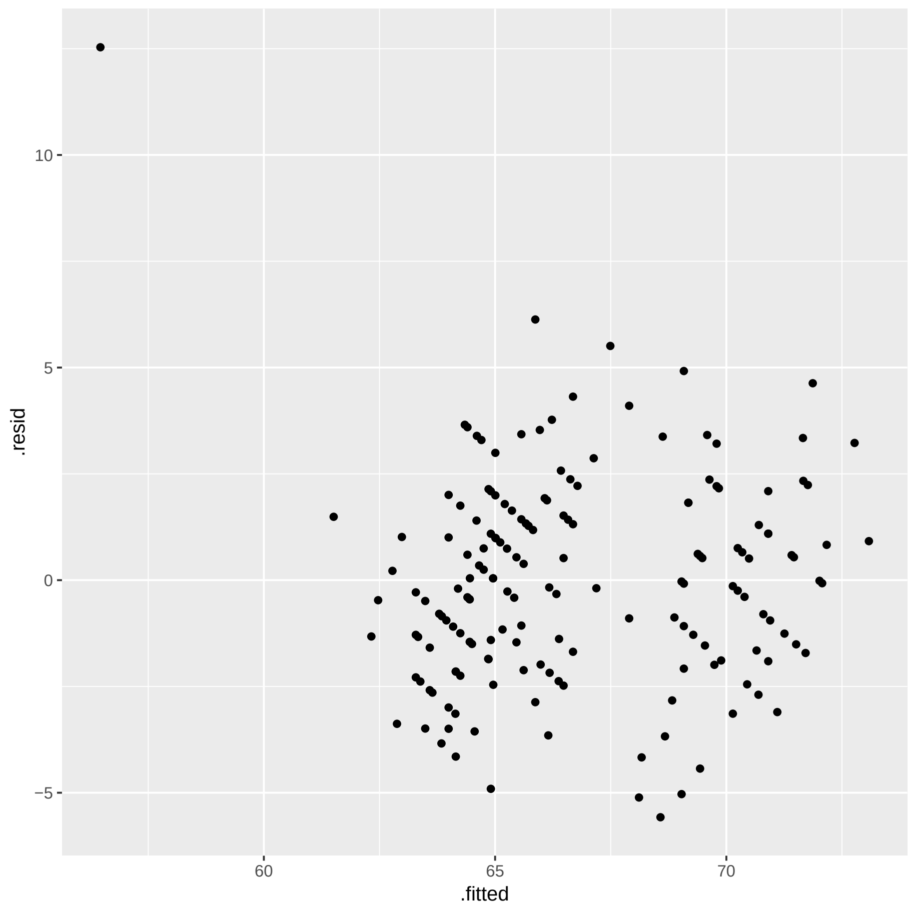

Figure 29: Cal-Davis residual plot

```
shingles=read.table("shingles.txt",header=T)
shingles
##    district sales promotion active competing potential
## 1         1  79.3       5.5     31        10         8
## 2         2 200.1       2.5     55         8         6
## 3         3 163.2       8.0     67        12         9
## 4         4 200.1       3.0     50         7        16
## 5         5 146.0       3.0     38         8        15
## 6         6 177.7       2.9     71        12        17
## 7         7  30.9       8.0     30        12         8
## 8         8 291.9       9.0     56         5         4
## 9         9 160.0       4.0     42         8         4
## 10       10 339.4       6.5     73         5        16
## 11       11 159.6       5.5     60        11         7
## 12       12  86.3       5.0     44        12        12
## 13       13 237.5       6.0     50         6         6
## 14       14 107.2       5.0     39        10         4
## 15       15 155.0       3.5     55        10         4
## 16       16 291.4       8.0     70         6        14
## 17       17 100.2       6.0     40        11         6
## 18       18 135.8       4.0     50        11         8
## 19       19 223.3       7.5     62         9        13
## 20       20 195.0       7.0     59         9        11
## 21       21  73.4       6.7     53        13         5
## 22       22  47.7       6.1     38        13        10
## 23       23 140.7       3.6     43         9        17
## 24       24  93.5       4.2     26         8         3
## 25       25 259.0       4.5     75         8        19
## 26       26 331.2       5.6     71         4         9
```

Figure 30: Roofing shingles sales data

```
apply(shingles[,2:6],2,summary)
##            sales promotion   active competing potential
## Min.     30.9000  2.500000 26.00000  4.000000  3.000000
## 1st Qu. 101.9500  4.000000 40.50000  8.000000  6.000000
## Median  159.8000  5.500000 51.50000  9.000000  8.500000
## Mean    170.2077  5.407692 51.84615  9.115385  9.653846
## 3rd Qu. 217.5000  6.650000 61.50000 11.000000 13.750000
## Max.    339.4000  9.000000 75.00000 13.000000 19.000000
```

Figure 31: Summaries of roofing shingle variables

```
z=rep(1,26)
shingles.1=lm(z~sales+promotion+active+competing+potential,data=shingles)
hatvalues(shingles.1)
##          1          2          3          4          5          6
## 0.13975772 0.23776316 0.24613015 0.21492480 0.24262317 0.37380734
##          7          8          9         10         11         12
## 0.33722509 0.43964316 0.14946533 0.25469374 0.12972647 0.12272428
##         13         14         15         16         17         18
## 0.15469957 0.10428845 0.22923944 0.27410226 0.13013926 0.20248892
##         19         20         21         22         23         24
## 0.20367969 0.08812773 0.40380805 0.18297350 0.22102630 0.30259729
##         25         26
## 0.29639805 0.31794706
2*(5+1)/26
## [1] 0.4615385
```

Figure 32: Roofing shingles regression and "hatvalues"

```
date          ,team1                        ,team2                        ,s1,s2
2016-08-13    ,Southampton                  ,Watford                      ,1 , 1
2016-08-13    ,Middlesbrough                ,Stoke City                   ,1 , 1
2016-08-13    ,Everton                      ,Tottenham Hotspur            ,1 , 1
2016-08-13    ,Manchester City              ,Sunderland                   ,2 , 1
2016-08-13    ,Crystal Palace               ,West Bromwich Albion         ,0 , 1
2016-08-13    ,Burnley                      ,Swansea City                 ,0 , 1
2016-08-13    ,Hull City                    ,Leicester City               ,2 , 1
2016-08-14    ,Arsenal                      ,Liverpool                    ,3 , 4
2016-08-14    ,AFC Bournemouth              ,Manchester United            ,1 , 3
2016-08-15    ,Chelsea                      ,West Ham United              ,2 , 1
2016-08-19    ,Manchester United            ,Southampton                  ,2 , 0
2016-08-20    ,Tottenham Hotspur            ,Crystal Palace               ,1 , 0
2016-08-20    ,West Bromwich Albion         ,Everton                      ,1 , 2
2016-08-20    ,Leicester City               ,Arsenal                      ,0 , 0
...
```

There are more lines of data for a total of 130 lines.

Figure 33: England soccer data (some)

```
proc print;
```

Obs	id	brakepower	fuel	massburnrate
1	a	4	DF-2	13.2
2	b	4	Blended	17.5
3	c	4	AdvancedTiming	17.5
4	d	6	DF-2	26.1
5	e	6	Blended	32.7
6	f	6	AdvancedTiming	43.5
7	g	8	DF-2	25.9
8	h	8	Blended	46.3
9	i	8	AdvancedTiming	45.6
10	j	10	DF-2	30.7
11	k	10	Blended	50.8
12	l	10	AdvancedTiming	68.9
13	m	12	DF-2	32.3
14	n	12	Blended	57.1

Figure 34: Synthetic fuels data

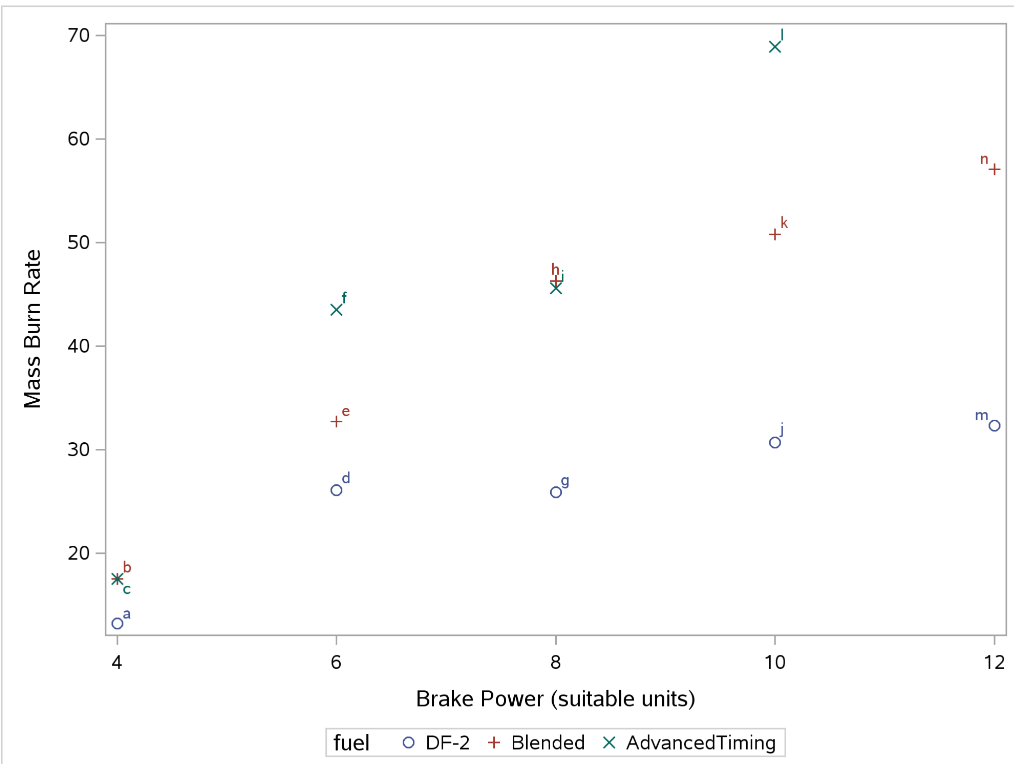

Figure 35: Plot of synthetic fuel data

```
ggplot(wtloss2,aes(x=when,y=weight,colour=factor(client),group=factor(client)))+
  geom_point()+geom_line()+guides(colour=F)
```

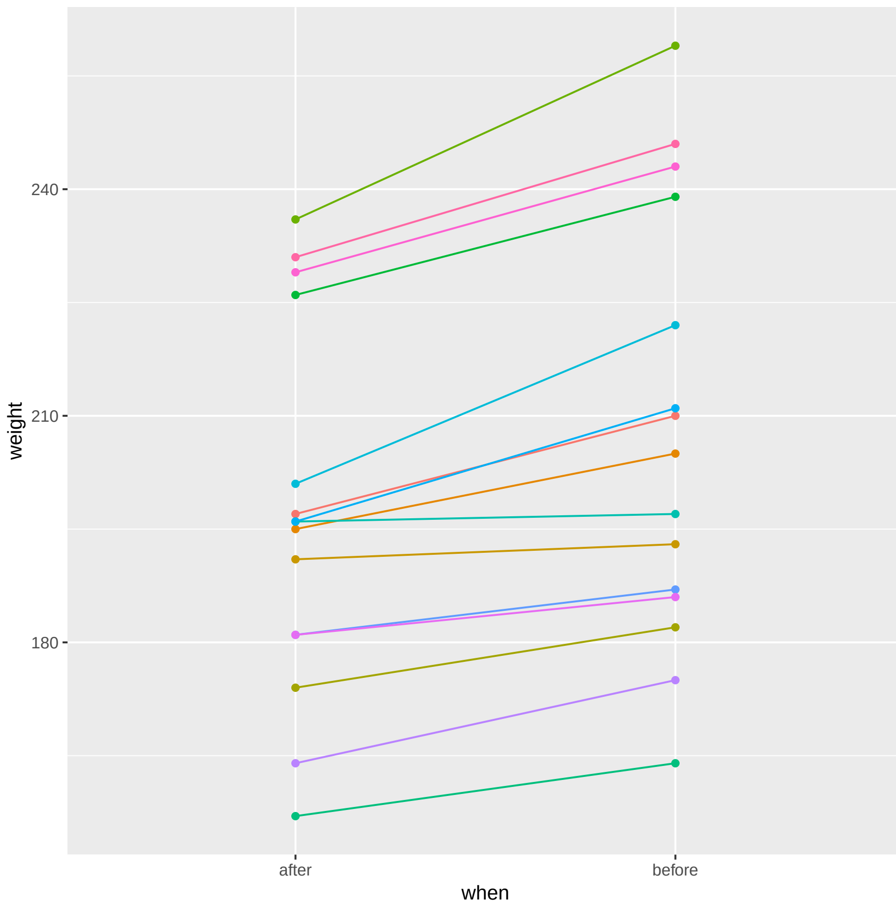

Figure 36: Spaghetti plot