EDA Oct 24

Ex1

Recall the hearing data from our first lecture. The data came from Cuthbert Daniel, a gentle and very smart British statistician (1904-1997). The data are prevalence rates of hearing loss in males aged 55-64 with hearing levels at least 16 dB above audible zero, at 500, 1000, 2000, 3000, 4000, 6000 Hz (cycles per second) and also "normal speech". Daniel suggested, from the pattern of LS residuals, that observations in cells [3,2], [4,3], [5,3], [6,3], [3,1] might be suspicious, as well as any whose LS residual exceeded 3 times the "approximate sigma" (10.2) in magnitude.

a. Conduct both the means analysis and the median polish for this table.

```
hearing.meaned
```

```
##              profl        farm      sales     crafts       oper       serv
## 500      2.157143  -1.2714286   8.685714 -5.6857143  0.5714286 -2.4000000
## 1000     2.642857   0.9142857   9.571429 -4.8000000 -1.1428571 -5.7142857
## 2000    -6.314286 -14.0428571   6.514286  3.0428571  1.7000000  5.3285714
## 3000     5.257143   2.1285714 -14.514286  4.0142857 -0.7285714  3.1000000
## 4000    -2.142857   5.2285714 -14.814286  5.2142857 -2.7285714  2.1000000
## 6000    -2.885714   7.7857143  -3.557143  2.6714286  1.1285714 -0.6428571
## normal   1.285714  -0.7428571   8.114286 -4.4571429  1.2000000 -1.7714286
## colmean -6.628571   1.5000000  -6.857143  0.5142857  7.4571429  3.7285714
##              labor   rowmean
## 500     -2.0571429 -31.64286
## 1000    -1.4714286 -32.52857
## 2000     3.7714286 -10.87143
## 3000     0.7428571  20.55714
## 4000     7.1428571  36.75714
## 6000    -4.5000000  46.50000
## normal  -3.6285714 -28.77143
## colmean  0.2857143  38.21429
```

```
med$residuals
```

```
##        profl  farm sales crafts oper serv labor
## 500      1.4   0.0   5.5   -5.4  2.2  0.0   0.0
## 1000     1.3   1.6   5.8   -5.1 -0.1 -3.9   0.0
## 2000   -10.4 -16.1   0.0    0.0  0.0  4.4   2.5
## 3000     1.1   0.0 -21.1    0.9 -2.5  2.1  -0.6
## 4000    -7.4   2.0 -22.5    1.0 -5.6  0.0   4.7
## 6000    -5.4   7.3  -8.5    1.2  1.0  0.0  -4.2
## normal   0.0   0.0   4.4   -4.7  2.3  0.1  -2.1
```

b. Calculate the matrix of residuals and stem-and-leaf them (back-to-back).

```
## ________________________________________________________________
##   1 | 2: represents 12, leaf unit: 1
##                med$residuals       hearing.meaned[-8, -8]
## LO: -22.5
## ________________________________________________________________
##     2                       1| -2* |
##                              | -1. |
##     3                       6|   s |
```

```
##                              |   f |444                    3
##                              |   t |
##     4                       0| -1* |
##     5                       8| -0. |
##     6                       7|   s |6                      4
##    12                  445555|   f |55444                  9
##    15                     223|   t |3322222               16
##    17                      00| -0* |1111000               23
##   (21)  111111100000000000000|  0* |0001111              (7)
##    11                   22222|   t |22222333              19
##     6                   55444|   f |45555                 11
##     1                       7|   s |677                    6
##                              |  0. |889                    3
##                              |  1* |
## ----------------------------------------------------------------
## n:                         49      49
## ----------------------------------------------------------------
```

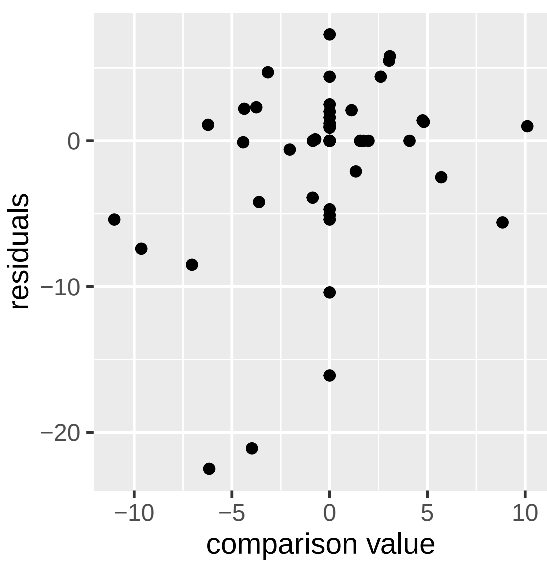
c. What is the effect of rows (Hz) and of columns (professionals)? Calculate a robust measure of the percent of the variation in the data explained by your fit (pseudo-R2).

```
##    500   1000   2000   3000   4000   6000 normal
##  -24.1  -24.4    0.0   31.5   48.8   55.8  -20.7
```

```
##  profl   farm  sales crafts   oper   serv  labor
##   -6.1    0.0   -3.9    0.0    5.6    1.1   -2.0
```

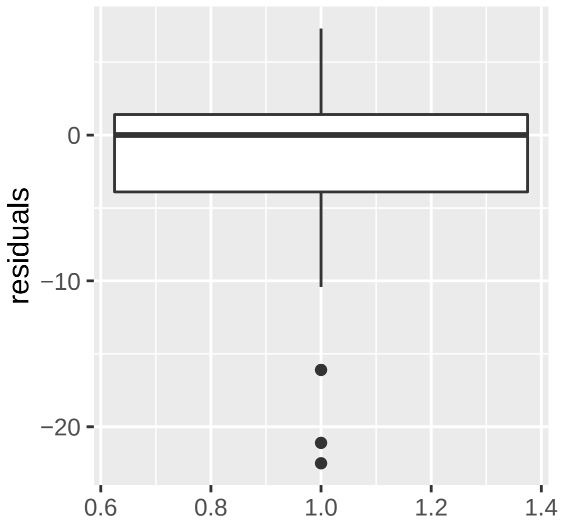
pseudo-R^2 = 0.8748421

d. Construct the diagnostic plot. Does it suggest the need for re-expression?



The diagnostic does not show a systematic position of the points, re-expression not needed.

e. Does the median polish table of residuals suggest any outliers (based on the boxplot outlier labeling rule)? If so, which ones?



There are three outliers according to boxplot.

(2000Hz, farm)=-16.1

(3000Hz,sales)=-21.1

(4000Hz, sales)=-22.5.

Ex2

a. Use median polish to construct a fit to the two sets of values in Table 7-4, p.267 (see data below): corn1.mat = yield in pounds filed weight of ear corn, corn2.mat = number of plants.

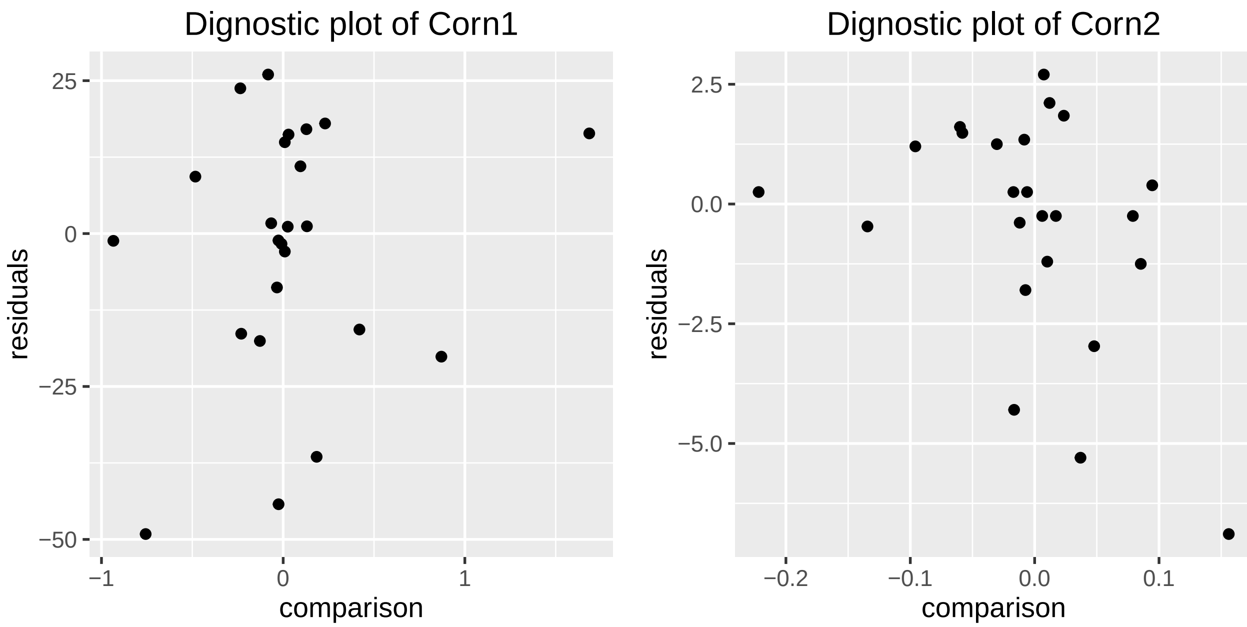
```
medcorn1$residuals
```

```
##         1        2       3       4
## A  26.000 -15.6875 16.1875 -49.125
## B -44.250  17.0625 14.9375 -16.375
## C  -1.125   1.1875 -2.9375  23.750
## D -36.500  -1.1875  1.6875  16.375
## E   1.125 -17.5625 -1.6875  18.000
## F  11.000   9.3125 -8.8125 -20.125
```

```
medcorn2$residuals
```

```
##           1        2        3         4
## A  1.609375 -1.25000  1.25000 -5.296875
## B -4.296875  1.84375  1.34375 -1.203125
## C -0.390625 -0.25000  0.25000  2.703125
## D -6.890625  0.25000 -0.25000  1.203125
## E  0.390625 -0.46875 -2.96875  1.484375
## F  2.109375  0.25000 -0.25000 -1.796875
```

b. Construct diagnostic plots for both median polish fits. Is re-expression indicated?

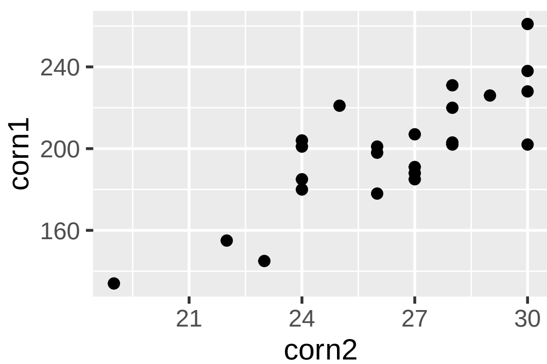


No need for re-expression.

c. Calculate pseudo-R2. How good are the median polish fits?

pseudo-R^2 for corn1 = 0.2589641. pseudo-R^2 for corn2 = 0.2307692.

d. Plot "corn1.mat" (y-axis) vs "corn2.mat" (x-axis). Fit RR line (you may round intercept and slope to nearest integer). How good is the "fit" ("pseudo R-squared")?




```
##            a       b    |res|
## 1 -13.27273 8.09091 321.0909
## 2 -18.59504 0.74380 313.6529
## 3  -3.38092 0.13524 312.3005
## 4  -0.61471 0.02459 312.0546
## 5  -0.11177 0.00447 312.0099
##   -35.97516 8.99901 312.0099
```

The pseudo-R^2 is 0.9245624.

e. Calculate residuals from RRline you fitted above. Place the residuals back into the matrix and perform median polish.

```
## 1: 146.997
## 2: 142.004
## Final: 142.004
```

```
##           1          2          3         4
## A 11.066194  -3.842524  2.001987 -2.347988
## B -3.065387   4.034836  2.876366 -3.469635
## C  1.748261   3.842524 -8.314953 -1.655986
## D 22.055266  -5.843517 -2.001987  1.655986
## E -3.250745 -13.157476 21.682067  3.343020
## F -1.748261  14.344014 -2.814456  1.840537
```

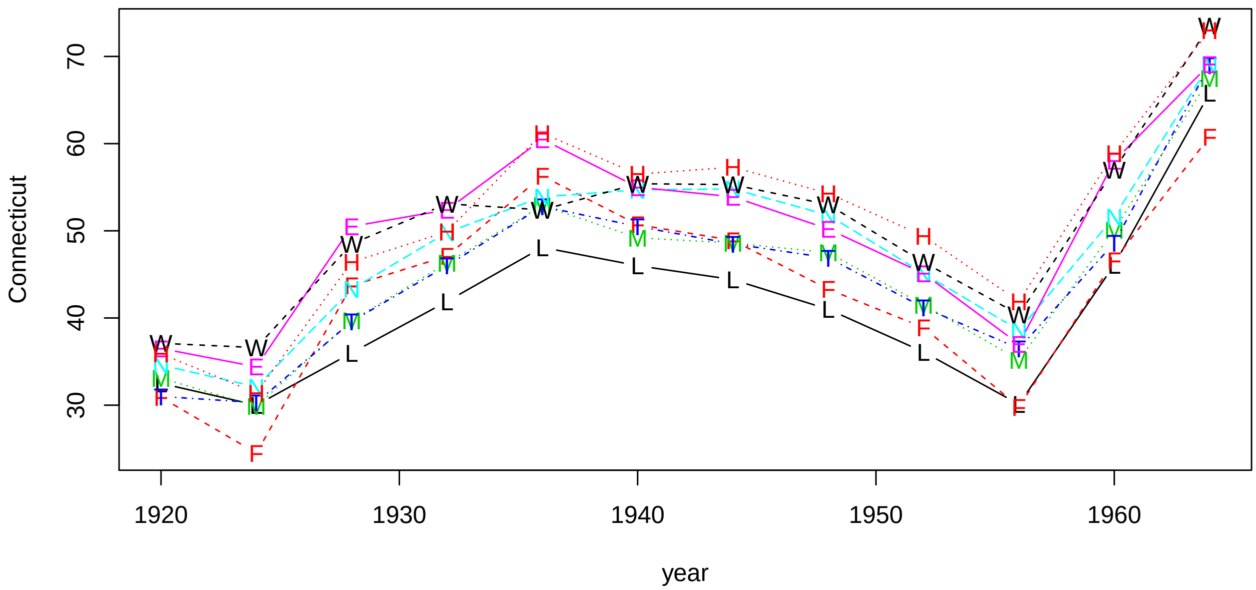
f. What do you learn from the analysis? Calculate your final pseudo-R2 .

The final pseudo-R^2 is 0.5880897. I learned that when two sets of data have significant correlation, we should first perform regression between the two then use median polish method.

EX3

3. *Connecticut elections*: Below are the data for the percent Democratic vote in the 8 counties in Connecticut in each election year from 1920 to 1964. The order of the columns is: Litchfield, Fairfield, Middlesex, Tolland, New London, New Haven, Windham, Hartford.

a. Plot the results for the 8 counties as a function of year, all on one graph. (If using R, use `matplot(year,CTelections,type="b")`). You can add `pch="LFMTNEWH"` as a way to distinguish the lines so you can identify the counties by a letter instead of by a number.)



b. Median polish the table. You may round values to the nearest tenth of a percentage point.

```r
med3b<-medpolish(Connecticut)
```

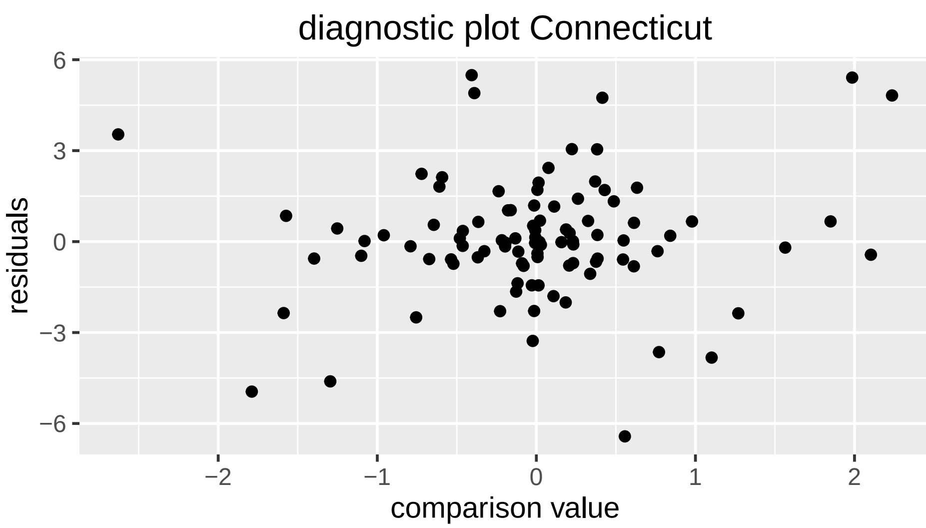
```
## 1: 143.4
## 2: 132.25
## Final: 132.025
```

```r
med3b$res
```

```
##      Litchfield  Fairfield Middlesex   Tolland New_London New_Haven
## 1920  5.4109375  0.6640625  1.328125 -0.590625  -0.140625  0.209375
## 1924  4.8203125 -3.8265625  0.037500  0.618750  -0.731250  0.018750
## 1928  0.1921875  4.7453125 -0.790625 -0.709375  -0.159375  5.490625
## 1932 -0.1078125  1.9453125 -0.390625 -0.509375   0.140625  1.190625
```

```
## 1936 -0.1546875  4.8984375 -0.037500  0.043750  -2.006250  3.043750
## 1940  0.1078125  1.6609375 -1.375000  0.106250   1.156250 -0.093750
## 1944 -0.3171875  1.0359375 -0.800000 -0.718750   2.431250 -0.018750
## 1948 -1.4421875 -2.2890625  0.375000 -0.043750   1.706250 -1.443750
## 1952 -0.3171875 -0.6640625  0.400000  0.281250   1.031250 -0.518750
## 1956 -0.1984375 -3.6453125  0.218750  1.700000   0.650000 -2.500000
## 1960  0.3515625 -2.2953125 -0.331250 -1.650000  -1.800000  3.050000
## 1964  3.5359375 -4.6109375  0.553125  2.234375  -0.815625 -2.365625
##        Windham   Hartford
## 1920 -0.559375 -2.3578125
## 1924  0.850000 -4.9484375
## 1928  2.121875 -0.5765625
## 1932  0.521875 -3.2765625
## 1936 -6.425000  1.7765625
## 1940 -1.062500 -0.5609375
## 1944  0.012500  1.4140625
## 1948 -0.012500  0.6890625
## 1952 -0.587500  1.8140625
## 1956 -0.468750  0.4328125
## 1960  0.681250  1.9828125
## 1964  0.665625 -0.4328125
```

c. Construct the diagnostic plot. What does it tell you?



The residuals seem to have a linear relationship with comparison values, indicating a re-expression.

d. If a re-expression of the data are needed, re-express the data and re-fit by median polish.

```
RRfit3d<-run.rrline(as.vector(comparison3c),as.vector(med3b$residuals))
```

```
##          a       b    |res|
## 1 -0.01559 0.10824 131.8309
## 2 -0.00001 0.01652 131.8223
## 3 -0.00073 0.00585 131.8222
## 4 -0.00026 0.00207 131.8221
## 5 -0.00009 0.00073 131.8221
##   -0.01668 0.13341 131.8221
```

```r
RRfit3d$b
```

```
## [1] 0.1334137
```

```r
lm(as.vector(med3b$residuals)~as.vector(comparison3c))
```

```
##
## Call:
## lm(formula = as.vector(med3b$residuals) ~ as.vector(comparison3c))
##
## Coefficients:
##             (Intercept)  as.vector(comparison3c)
##                 0.09211                  0.31005
```

RRfit comparison value to medpolish residuals, the slope is 0.1334137, 1-slope is 0.8665863,close to 0.5, square root-transformation.

```r
med3d<-medpolish(sqrt(Connecticut))
```

```
## 1: 10.77229
## 2: 9.72916
## Final: 9.698271
```

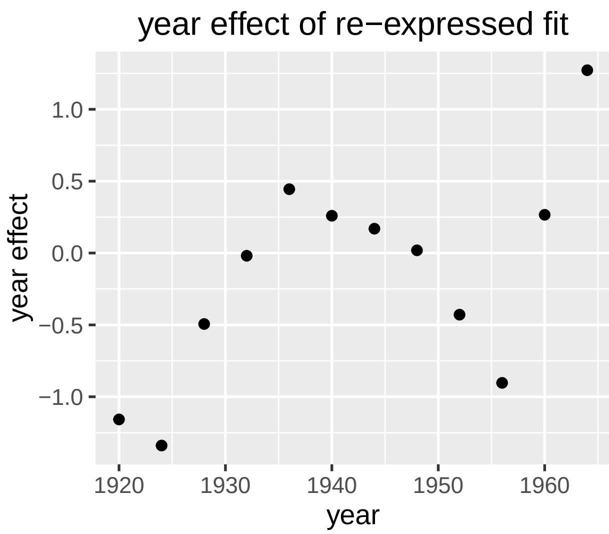
```r
med3d$residuals
```

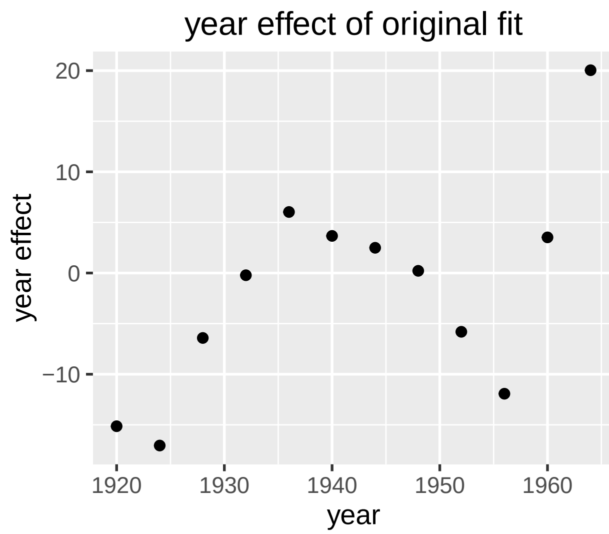
```
##         Litchfield  Fairfield    Middlesex      Tolland    New_London
## 1920  0.3647201589  0.02533033  0.084781510 -0.100301240 -0.0350372320
## 1924  0.3226533984 -0.40211426 -0.018804399  0.018062765 -0.0699427204
## 1928 -0.0001210727  0.41318693 -0.031648940 -0.039176620 -0.0009042555
## 1932 -0.0019295837  0.19070391 -0.002861651 -0.024529170  0.0009042555
## 1936 -0.0026391452  0.36798755  0.002861651 -0.003603286 -0.1774502952
## 1940  0.0289441393  0.16971786 -0.071436146  0.021040721  0.0615022576
## 1944  0.0001210727  0.13235252 -0.024162872 -0.030925925  0.1584340490
## 1948 -0.1097213075 -0.12986154  0.046857363  0.003603286  0.1033194078
## 1952 -0.0566996789 -0.02533033  0.044700662  0.021786745  0.0695693895
## 1956 -0.1042261936 -0.32892481  0.010064438  0.119041467  0.0412602101
## 1960  0.0296694151 -0.13089895 -0.014214784 -0.120567277 -0.1577713985
## 1964  0.3457548956 -0.16581909  0.117551873  0.208750206 -0.0343807778
##        New_Haven      Windham    Hartford
## 1920  0.01345775 -0.015404981 -0.17831674
## 1924  0.01867088  0.124996261 -0.38482254
## 1928  0.41430586  0.193848119 -0.02218718
## 1932  0.07192565  0.041799217 -0.24477796
## 1936  0.14821083 -0.469496018  0.05116674
## 1940 -0.02909562 -0.080493165 -0.07055830
## 1944 -0.01345775  0.002960603  0.07264427
## 1948 -0.11955798 -0.002960603  0.02218718
## 1952 -0.04128214 -0.023497634  0.12966547
## 1956 -0.19971862 -0.004698398  0.04862740
## 1960  0.16383618  0.019580824  0.08078423
## 1964 -0.14523231  0.037024132 -0.05578193
```

e. Approximately how much of the variation in the original table does your final median polish fit explain? (i.e., calculate a robust measure of the traditional R^2.)

The pseudo-R^2 is 0.8375255 in the re-expressed medpolish.

f. Plot the year effects. Do you observe a pattern?

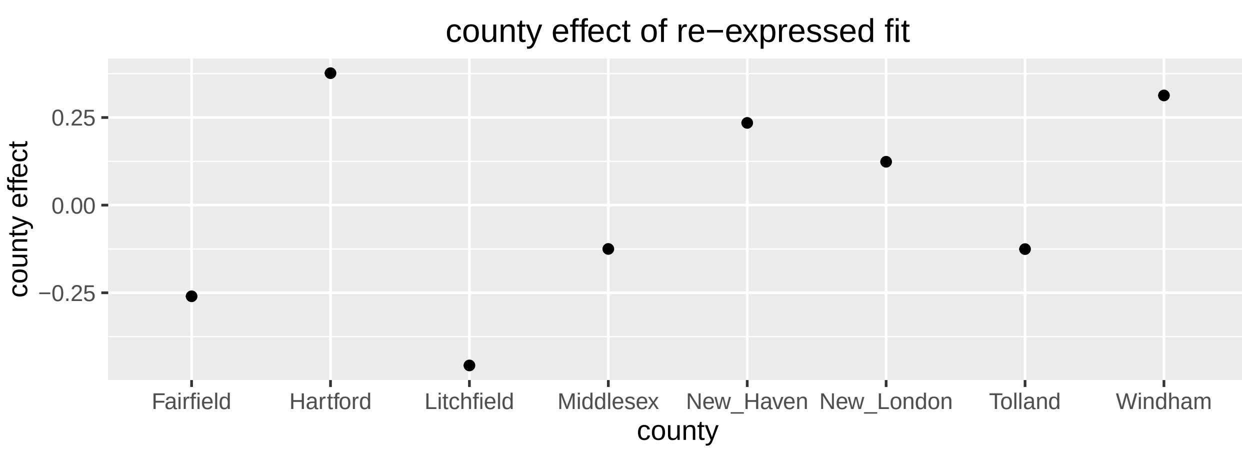


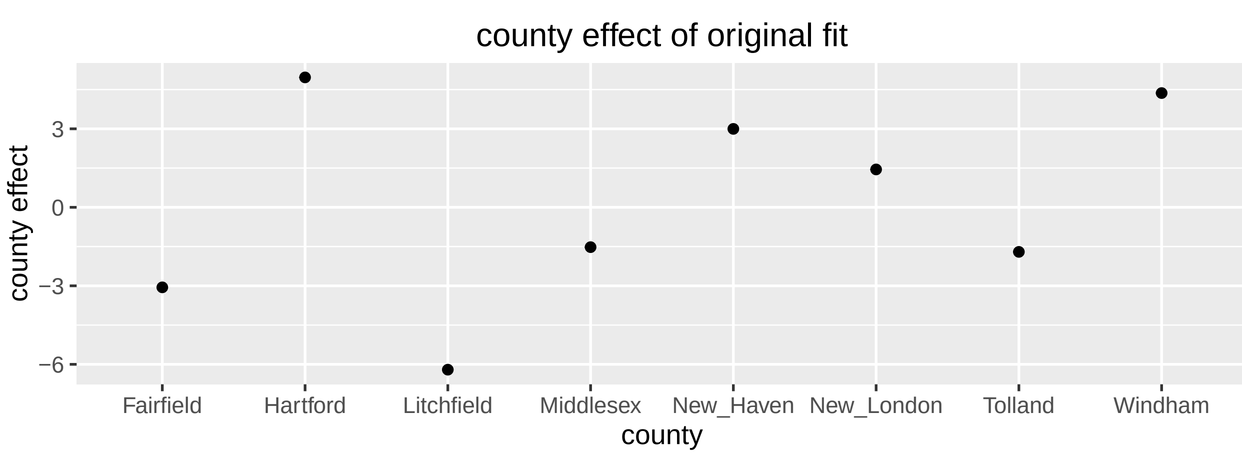


The year effect is very similar to the plot in a), which indicates our fit is suitable.

g. Plot the county effects. What do you notice?

The effects between counties have a significant difference.



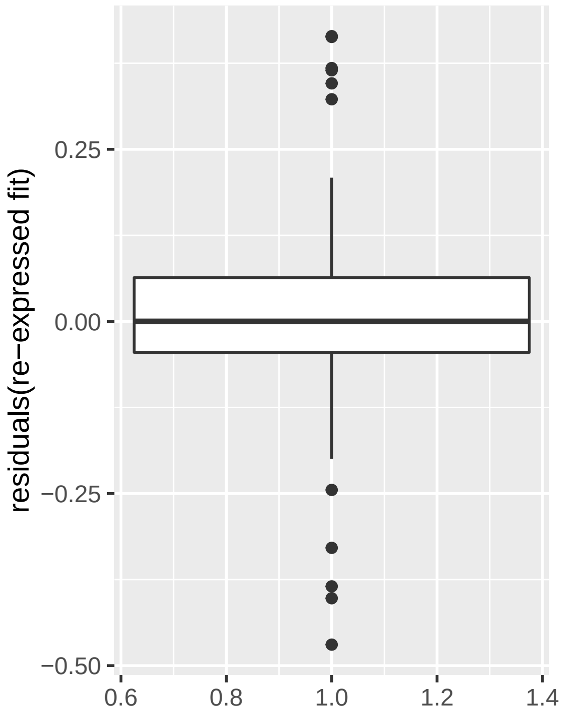


h. Stem and leaf the residuals, and plot the residuals as a function of year. From a five-number summary of all 96 residuals, which (county,year) residuals are "out" or "far-out" (based on the boxplot labeling rules)?

```r
stem.leaf(med3d$residuals)
```

```
## 1 | 2: represents 0.12
##  leaf unit: 0.01
##            n: 96
## LO: -0.469496017904904 -0.402114256335361 -0.384822536024916 -0.328924805702369
##     5    -2* | 4
##    10    -1. | 97765
##    18    -1* | 43221000
##    24    -0. | 877655
##    48    -0* | 433333222222111100000000
##   (22)    0* | 0000011111222222344444
##    26     0. | 5667788
##    19     1* | 0112234
##    12     1. | 56699
##     7     2* | 0
##           2. |
##     6     3* | 24
## HI: 0.364720158925149 0.367987545972848 0.413186931801386 0.414305863847269
```

```r
qplot(1,as.vector(med3d$residuals),geom = "boxplot")+labs(x="",y="residuals(re-expressed fit)")
```



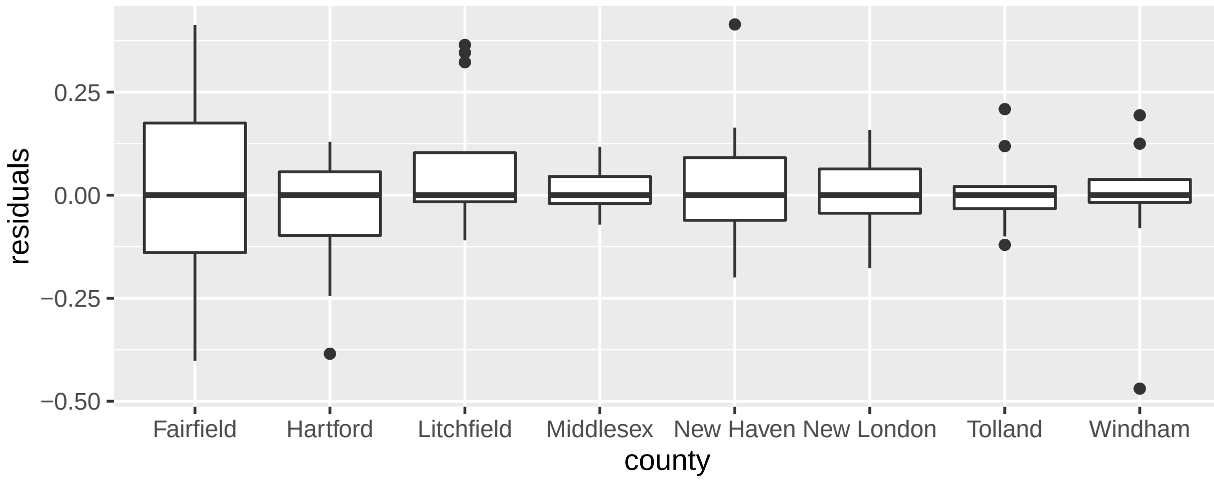
There are 10 outliers according to the boxplot. They are:

(1920,Litchfield),(1964,Litchfield),(1928,Fairfield),(1936,Fairfield),(1928,New Haven),

(1924,Fairfield),(1956,Fairfield),(1936,Windham),(1924,Hartford),(1932,Hartford).

i. Construct side-by-side boxplots of the 12 residuals in each state. Which counties are most variable?

```r
county<-c(rep("Litchfield",12),rep("Fairfield",12),rep("Middlesex",12),rep("Tolland",12),rep("New London
residuals3i<-as.vector(med3d$residuals)
box3i<-data.frame(county,residuals3i);colnames(box3i)<-c("county","residuals")
ggplot(box3i,aes(x=factor(county),y=residuals))+geom_boxplot()+labs(x="county")
```



From the aligned boxplot, Fairfield is the most variable county.

Ex5

a) Median polish starting with rows.

```
##                                 dataex5.rowmedian
##                     9  0 0 0  0                -2
##                     0  0 0 0  0                -1
##                     0  0 0 0  0                 0
##                     0  0 0 0  0                 1
##                     0  0 0 0 11                 2
## dataex5.colmedian -2 -1 0 1  2                 5
```

b) Median polish starting with columns.

```
##                                 dataex5.rowmedian
##                     9  0 0 0  0                -2
##                     0  0 0 0  0                -1
##                     0  0 0 0  0                 0
##                     0  0 0 0  0                 1
##                     0  0 0 0 11                 2
## dataex5.colmedian -2 -1 0 1  2                 5
```

c) Analysis by means.

```
##                                          dataex5.rowmean
##                   6.2 -1.0 -1.0 -1.0 -3.2            -1.0
##                  -1.0  0.8  0.8  0.8 -1.4            -1.8
##                  -1.0  0.8  0.8  0.8 -1.4            -0.8
##                  -1.0  0.8  0.8  0.8 -1.4             0.2
##                  -3.2 -1.4 -1.4 -1.4  7.4             3.4
## dataex5.colmean -1.0 -1.8 -0.8  0.2  3.4             5.8
```

d) 20% trimmed mean.

```
##                                                dataex5.rowmean
##                   7.27 -0.73 -0.73 -0.73 -0.73           -1.27
##                  -0.73  0.27  0.27  0.27  0.27           -1.27
##                  -0.73  0.27  0.27  0.27  0.27           -0.27
##                  -0.73  0.27  0.27  0.27  0.27            0.73
```

```
##                  -0.73  0.27  0.27  0.27 11.27            1.73
## dataex5.colmean -1.27 -1.27 -0.27  0.73  1.73            5.27
```