Salmon Project

Team Data Driven

Deisharrah Allen-Benso, Jonathan Acevedo, Reuben Lopez, Rachel Tekchandani

Introduction.

In this project we investigate contamination levels in farm raised salmon. The obtained data set consists of 153 instances of farm raised salmon. It includes salmons' farms origin and 11 different contaminants found in salmon. We can split our data into three regions: North America, South America and Europe. In this research we would like to focus on which region of farm salmon has a higher mean Mirex contaminant. Additionally we are interested to see if some contaminants are independent from other contaminants in their respective levels. To address these questions, we will use statistical methods like ANOVA(analysis of variance), permutation based ANOVA and permutation test of independence of two variables. There are papers that also focused salmon contamination levels. For example, 'Global Assessment of Organic Contaminants in Farmed Salmon. (2003)'. However, above mentioned paper and others were focusing on mean contamination locations in particular locations rather than by regions. Additionally, none of them focused on any form of correlation between contaminants.

Results and discussion.

Our data analysis starts with data exploration. Results of data analysis can be seen in the Appendix section of this paper. We can observe that Mirex column contains has three NaNs values. Hence, we omit them from further analysis. This makes our data set to contain 150 rows going forward. We can also observe that there are 8 unique locations in our dataset.

```
## [1] "Chile"          "Eastern Canada" "Faroe Islands"  "Maine"
## [5] "Norway"         "Scotland"       "Washington"     "Western Canada"
```

We will define the three regions using the following guidelines:

North America - Eastern Canada, Western Canada,Washington, Maine.

South America - Chile.

Europe - Norway, Scotland and Faroe Islands

We will encode these three regions as 1,2, 3 respectively.

Let's see number of examples available by Region.

```
##          Region Instances
## 1 North America        57
## 2 South America        30
## 3        Europe        63
```

Let's see assess correlation coefficients between Mirex contaminant and all other contaminants.

```
##           Contaminant Correlation
## 1               Mirex   1.0000000
## 2   Hexachlorobenzene   0.8810309
## 3           HCH_gamma  -0.0287125
## 4  Heptachlor.Epoxide   0.6663536
## 5            Dieldrin   0.8530482
## 6              Endrin   0.8875020
## 7     Total.Chlordane   0.8517805
## 8           Total.DDT   0.5510607
## 9              Dioxin   0.6915263
## 10  Total.Pesticides   0.7262804
## 11         Total.PCBs   0.7242992
```

We can observe that all contaminants but HCH_gamma are highly correlated to Mirex. HCH_gamma contaminant has negative correlation of -0.028. We will investigate whether we have a small negative correlation or a correlation of 0 cannot be ruled out. Lets check whether we can reject the hypothesis that correlation between HCH_gamma and Mirex is statistically 0. Due to sample sizes we will employ the permutation test of independence of two variables.

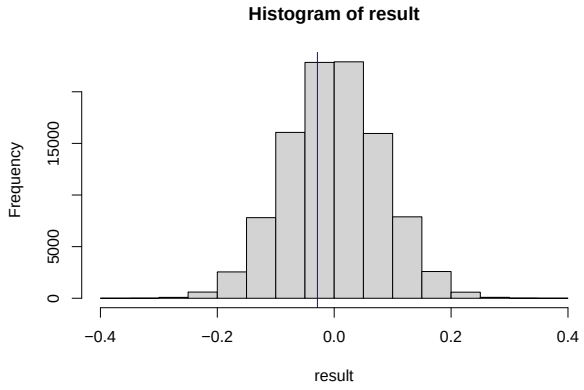


Figure 1: Bootstrap correlation distribution.

```
## [1] "The p-value is"
```

```
## [1] 0.6375636
```

Figure 1 contains bootstrap distribution of correlation values.The observed p-value is ~ 0.6. Since p-value is greater that 5% alpha, we fail to reject the null hypothesis that contaminants Mirex and HCH_gamma are independent.

Lets check average Mirex levels for each region.

```
##          Region       Mean
## 1 North America 0.05415789
## 2 South America 0.05993333
## 3        Europe 0.13993651
```

Although average values seem to differ, we see from box-plots in Figure 2 that overall distributions are similar across 3 regions. Hence, to assess whether we can reject the null hypothesis that mean values between regions are different we employ the ANOVA test. Let's define μ_1, μ_2, μ_3 as means of Mirex contaminant values for three regions. Therefore, ANOVA test hypotheses look like this:

$H_0 : \mu_1 = \mu_2 = \mu_3$
H_A : means are not equal

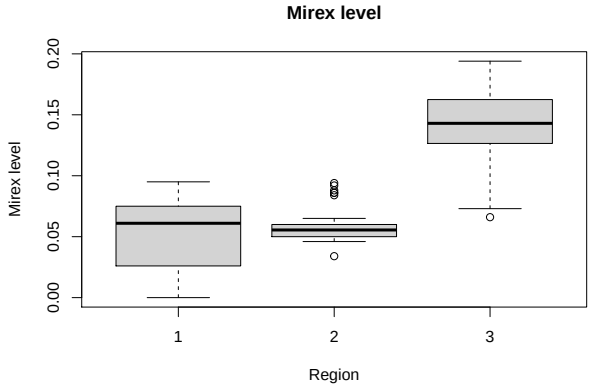


Figure 2: Box-plots of Mirex level per region

```
## Analysis of Variance Table
##
## Response: data$Mirex
##            Df  Sum Sq  Mean Sq F value    Pr(>F)
## Regions     2 0.25718 0.128588   174.6 < 2.2e-16 ***
## Residuals 147 0.10826 0.000736
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = data$Mirex ~ Regions)
##
## $Regions
##            diff          lwr        upr     p adj
## 2-1 0.005775439 -0.008717911 0.02026879 0.6137033
## 3-1 0.085778613  0.074032610 0.09752462 0.0000000
## 3-2 0.080003175  0.065749798 0.09425655 0.0000000
```

We can observe that at 5% significance level there is a difference between Regions 2 and 1 , 3 and 1, 3 and 2. Our overall p-value is 2.2e-16. Hence, we have sufficient evidence to reject the null hypothesis that means of Mirex between 3 regions are the same.

Based on quantile plot of residuals of ANOVA table, we can see that residuals are not normally distributed. Hence, standard ANOVA assumption of residuals being normally distributed is violated. Therefore, it is better to proceed with permutation based version.

```
## Analysis of Variance Table
##
## Response: data$Mirex
##            Df  Sum Sq  Mean Sq F value    Pr(>F)
## Regions     2 0.25718 0.128588   174.6 < 2.2e-16 ***
## Residuals 147 0.10826 0.000736
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
## [1] "Our p-value is"
```

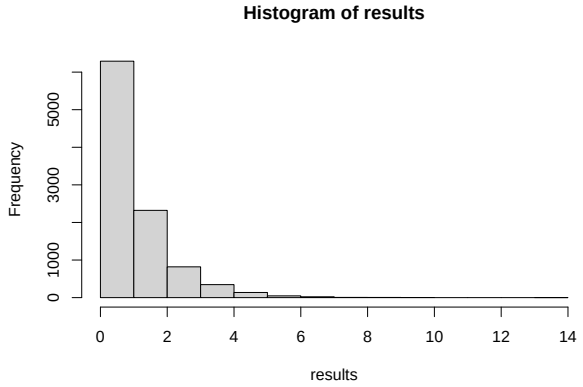


Figure 3: Bootstrap correlation distribution.

```
## [1] 1e-04
```

The permutation based p-value is eˆ-4. Hence, we reject the null hypothesis that average Mirex contaminant values are the same across 3 regions.

Conclusion.

In our findings we have seen that Mirex contamination levels in Europe are higher than in other two regions. We have also observed that Mirex and HCH contaminants do not seem to have any linear association between them.

References.

1. Chihara, L., & Hesterberg, T. (2019). Mathematical statistics with resampling and R. Hoboken, NJ: Wiley.
2. Global Assessment of Organic Contaminants in Farmed Salmon. (2003). Ronald A. Hites, Jeffery A. Foran, David O. Carpenter, M. Coreen Hamilton, Barbara A. Knuth, Steven J. Schwager
3. B. Charron, An IntraFish.com Industry Report on Salmon Product Development–The Fish of the Future? (1999)

Appendix.

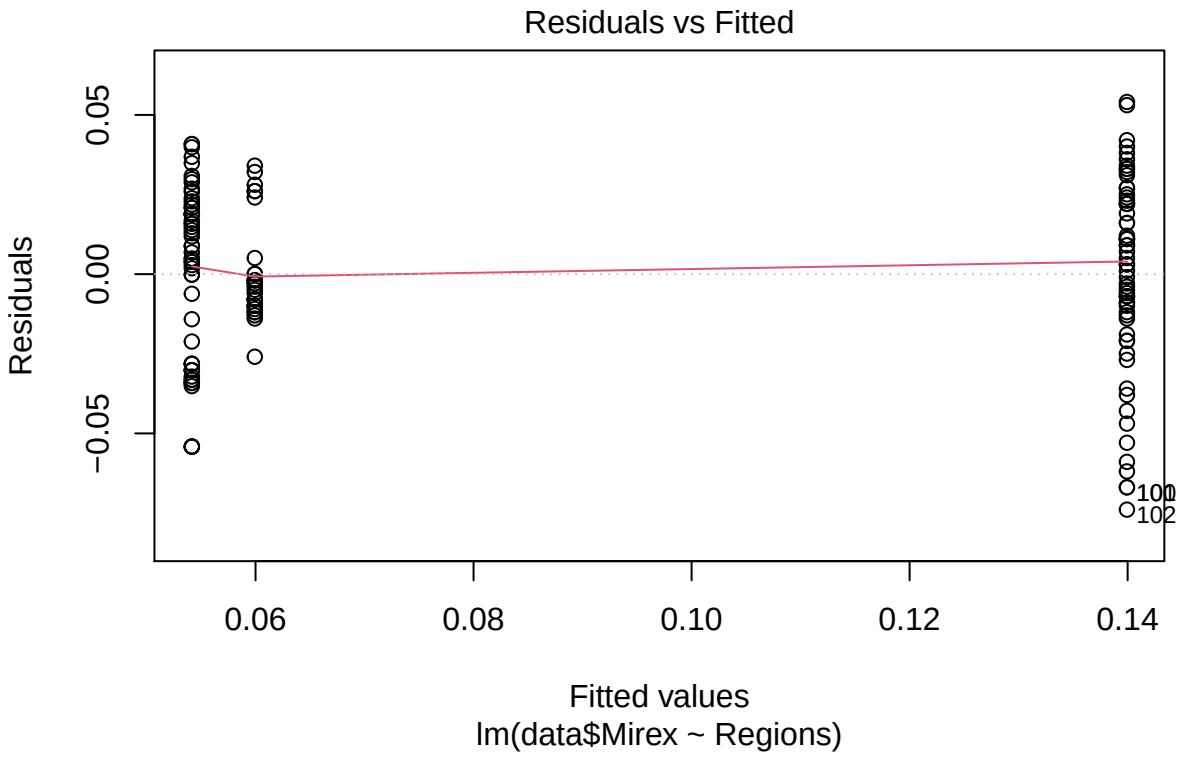
Summary of Salmon data.

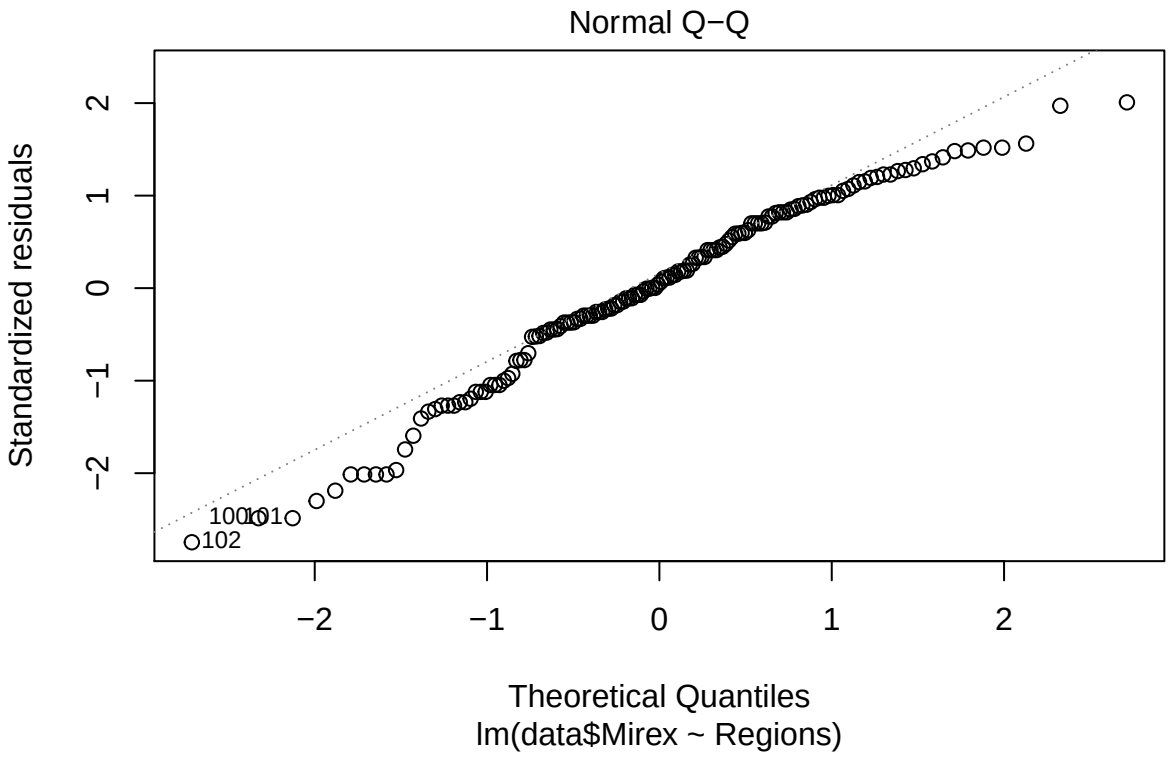
```
## data
##
##  14  Variables      150  Observations
## ---------------------------------------------------------------------------
## Kind
##        n  missing distinct    value
##      150        0        1   Farmed
##
## Value      Farmed
## Frequency     150
## Proportion      1
## ---------------------------------------------------------------------------
## Location
##        n  missing distinct
##      150        0        8
##
## lowest : Chile          Eastern Canada Faroe Islands  Maine          Norway
## highest: Maine          Norway         Scotland       Washington     Western Canada
##
## Value               Chile Eastern Canada  Faroe Islands          Maine
## Frequency              30             24             21              6
## Proportion           0.20           0.16           0.14           0.04
##
## Value              Norway       Scotland     Washington Western Canada
## Frequency              12             30              9             18
## Proportion           0.08           0.20           0.06           0.12
## ---------------------------------------------------------------------------
## Mirex
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##      150        0       93        1  0.09134   0.0565  0.02045  0.02600
##      .25      .50      .75      .90      .95
##  0.05600  0.07900  0.13475  0.16310  0.17300
##
## lowest : 0.000 0.019 0.020 0.021 0.022, highest: 0.178 0.180 0.182 0.193 0.194
## ---------------------------------------------------------------------------
## Hexachlorobenzene
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##      150        0      133        1     2.45    1.692   0.5549   0.5909
##      .25      .50      .75      .90      .95
##   1.0225   2.2850   3.8450   4.5210   4.8175
##
## lowest : 0.421 0.528 0.529 0.537 0.541, highest: 4.910 4.920 5.050 5.080 5.240
## ---------------------------------------------------------------------------
## HCH_gamma
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##      150        0      126        1   0.3049   0.2037  0.01145  0.01400
##      .25      .50      .75      .90      .95
##  0.20200  0.31350  0.41925  0.48040  0.62705
##
## lowest : 0.000 0.008 0.009 0.010 0.011, highest: 0.651 0.691 0.700 0.701 0.789
## ---------------------------------------------------------------------------
```

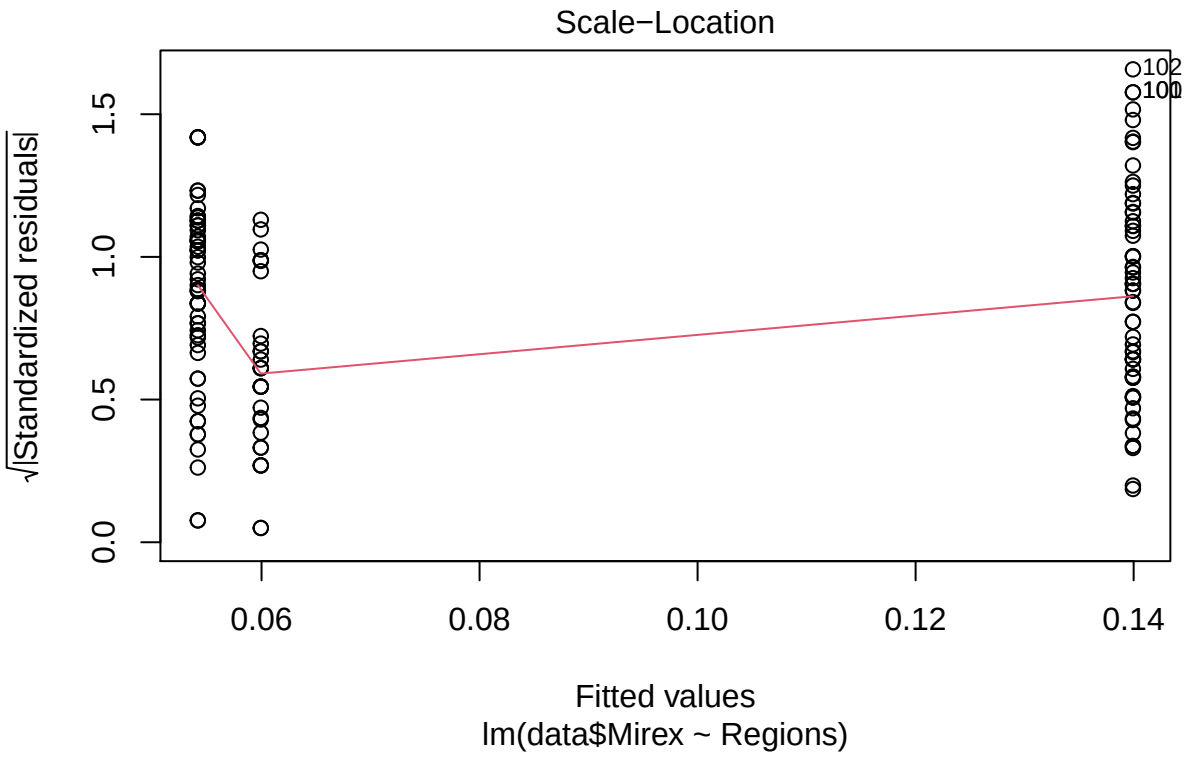
```
## Heptachlor.Epoxide
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##      150        0      132        1   0.5965   0.3629  0.07845  0.08690
##      .25      .50      .75      .90      .95
##  0.26975  0.69350  0.84025  0.91100  0.97295
##
## lowest : 0.060 0.064 0.071 0.072 0.078, highest: 1.020 1.120 1.390 1.430 1.520
## ------------------------------------------------------------------------------
## Dieldrin
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##      150        0      131        1    3.305    2.172   0.5263   0.5719
##      .25      .50      .75      .90      .95
##   0.9465   3.4900   5.1500   5.7120   5.9265
##
## lowest : 0.000 0.374 0.481 0.500 0.507, highest: 5.990 6.010 6.140 6.360 6.370
## ------------------------------------------------------------------------------
## Endrin
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##      150        0      135        1   0.3298   0.2663   0.0339   0.0419
##      .25      .50      .75      .90      .95
##   0.1265   0.2790   0.5172   0.6911   0.7193
##
## lowest : 0.000 0.030 0.033 0.035 0.037, highest: 0.769 0.771 0.772 0.784 0.814
## ------------------------------------------------------------------------------
## Total.Chlordane
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##      150        0      146        1    6.268    4.704   0.8243   0.8853
##      .25      .50      .75      .90      .95
##   1.2035   6.3820   9.9790  11.6001  12.4446
##
## lowest :  0.337  0.351  0.365  0.702  0.762, highest: 12.592 12.721 12.986 13.217 13.251
## ------------------------------------------------------------------------------
## Total.DDT
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##      150        0      149        1    18.72    11.69    4.543    5.125
##      .25      .50      .75      .90      .95
##    9.988   18.597   24.164   32.042   37.843
##
## lowest :  3.026  3.309  3.314  4.451  4.496, highest: 42.564 42.793 47.560 48.546 49.439
## ------------------------------------------------------------------------------
## Dioxin
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##      150        0      139        1    0.445   0.3234   0.0949   0.1358
##      .25      .50      .75      .90      .95
##   0.1965   0.3905   0.6520   0.8367   0.9260
##
## lowest : 0.036 0.037 0.038 0.044 0.075, highest: 1.140 1.150 1.180 1.200 1.310
## ------------------------------------------------------------------------------
## Total.Pesticides
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##      150        0      150        1    34.34    19.04    7.006    7.626
##      .25      .50      .75      .90      .95
##   21.566   36.898   47.441   52.864   58.086
##
```

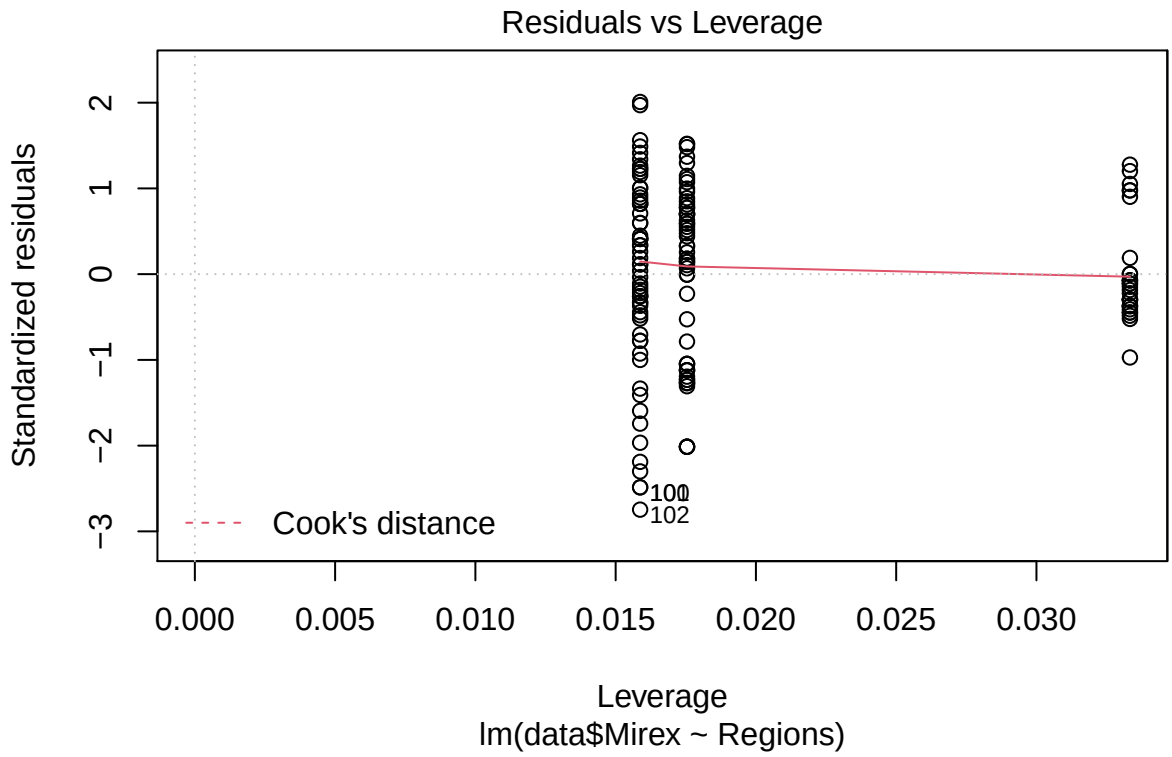
```
## lowest :  4.870  5.437  5.524  6.548  6.611, highest: 61.035 62.490 62.723 64.166 64.570
## ---------------------------------------------------------------------------
## Total.PCBs
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##      150        0      123        1    36422    17916    13600    15090
##      .25      .50      .75      .90      .95
##    20350    40800    46900    56010    59620
##
## lowest :  4940  5070  7440 11100 12500, highest: 64100 65900 70300 72000 76200
## ---------------------------------------------------------------------------
## Region
##        n  missing distinct     Info     Mean      Gmd
##      150        0        3    0.863     2.04   0.9648
##
## Value         1    2    3
## Frequency    57   30   63
## Proportion 0.38 0.20 0.42
## ---------------------------------------------------------------------------
```

```
## Analysis of Variance Table
##
## Response: data$Mirex
##            Df  Sum Sq  Mean Sq F value    Pr(>F)
## Regions     2 0.25718 0.128588   174.6 < 2.2e-16 ***
## Residuals 147 0.10826 0.000736
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```


Residuals vs Fitted
Residuals
0.05
0.00
−0.05
0.06
0.08
0.10
0.12
0.14
Fitted values
lm(data$Mirex ~ Regions)


Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(data$Mirex ~ Regions)
100
101
102
−2
−1
0
1
2


Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
0.06
0.08
0.10
0.12
0.14
Fitted values
lm(data$Mirex ~ Regions)


Residuals vs Leverage
Standardized residuals
2
1
0
−1
−2
−3
Cook's distance
100
102
0.000
0.005
0.010
0.015
0.020
0.025
0.030
Leverage
lm(data$Mirex ~ Regions)

```
## NULL
```