MSDS Assignemnt Two

Etana Disasa

11/02/2018

ANOVA

Analysis of Variance (ANOVA) is a systematic approach to test the hypothesis that the means of two or more variables/population equal. It compares the variable means at different factor levels. The null hypothesis claims that all population means (factor level means) are equal while the alternative hypothesis states that at least one is different.

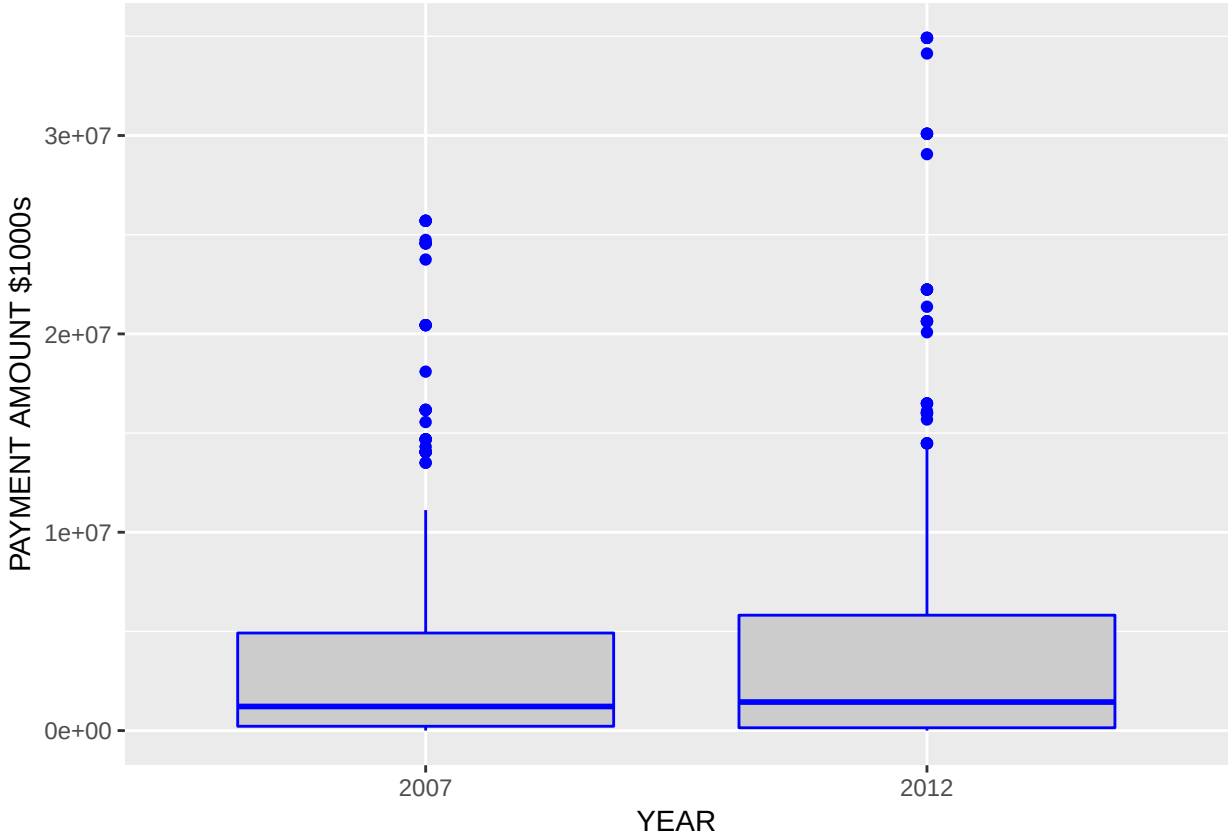
What does an Analysis of Variance tell you? What types of questions does it answer?

Annova uses variance to determine if means are different in a particular variables of a dataset. It runs these tests on fixed factor variables that are observed in the data. It attepts to answer if the mean of certain variables (columns) in the data set are different.

```
library(ggplot2)
library(data.table)
library(e1071)
library(stringr)
library(dplyr)
library(ggpubr)
```

ANOVA (ONE-WAY) #1

```
eco.dt <- read.csv("ECN_FULL_US_55A1_with_ann.csv")

eco.dt$YEAR = factor(eco.dt$YEAR, labels = c("2007", "2012"))


require(ggplot2)

ggplot(eco.dt, aes(x = YEAR, y = PAYANN)) +
  geom_boxplot(fill = "grey80", colour = "blue") +
  scale_x_discrete() + xlab("YEAR") +
  ylab("PAYMENT AMOUNT $1000s")
```




```
eco.mod1 = lm(PAYANN ~ YEAR, data = eco.dt)

summary(eco.mod1)
```

```
##
## Call:
## lm(formula = PAYANN ~ YEAR, data = eco.dt)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -4634811 -3752645 -2646770  1155646 30286523
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3764967     335165  11.233   <2e-16 ***
## YEAR2012      869844     473994   1.835   0.0669 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6324000 on 710 degrees of freedom
## Multiple R-squared:  0.004721,   Adjusted R-squared:  0.003319
## F-statistic: 3.368 on 1 and 710 DF,  p-value: 0.0669
```

```
anova(eco.mod1)
```

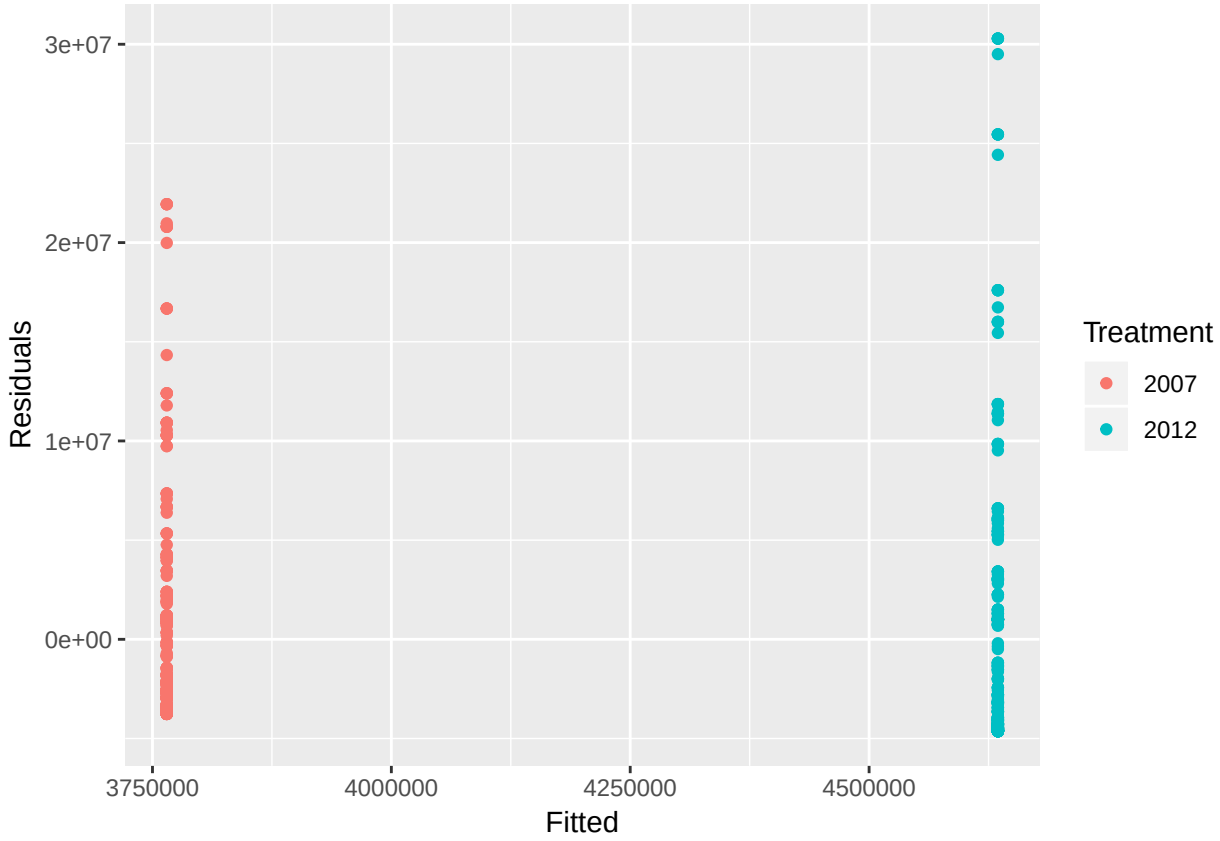
```
## Analysis of Variance Table
##
## Response: PAYANN
```

```
##             Df     Sum Sq    Mean Sq F value Pr(>F)
## YEAR         1 1.3468e+14 1.3468e+14  3.3677 0.0669 .
## Residuals  710 2.8394e+16 3.9991e+13
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
confint(eco.mod1)
```

```
##                  2.5 %  97.5 %
## (Intercept) 3106934.7 4423000
## YEAR2012     -60754.3 1800443
```

```
eco.mod = data.frame(Fitted = fitted(eco.mod1),
                     Residuals = resid(eco.mod1), Treatment = eco.dt$YEAR)
ggplot(eco.mod, aes(Fitted, Residuals, colour = Treatment)) + geom_point()
```

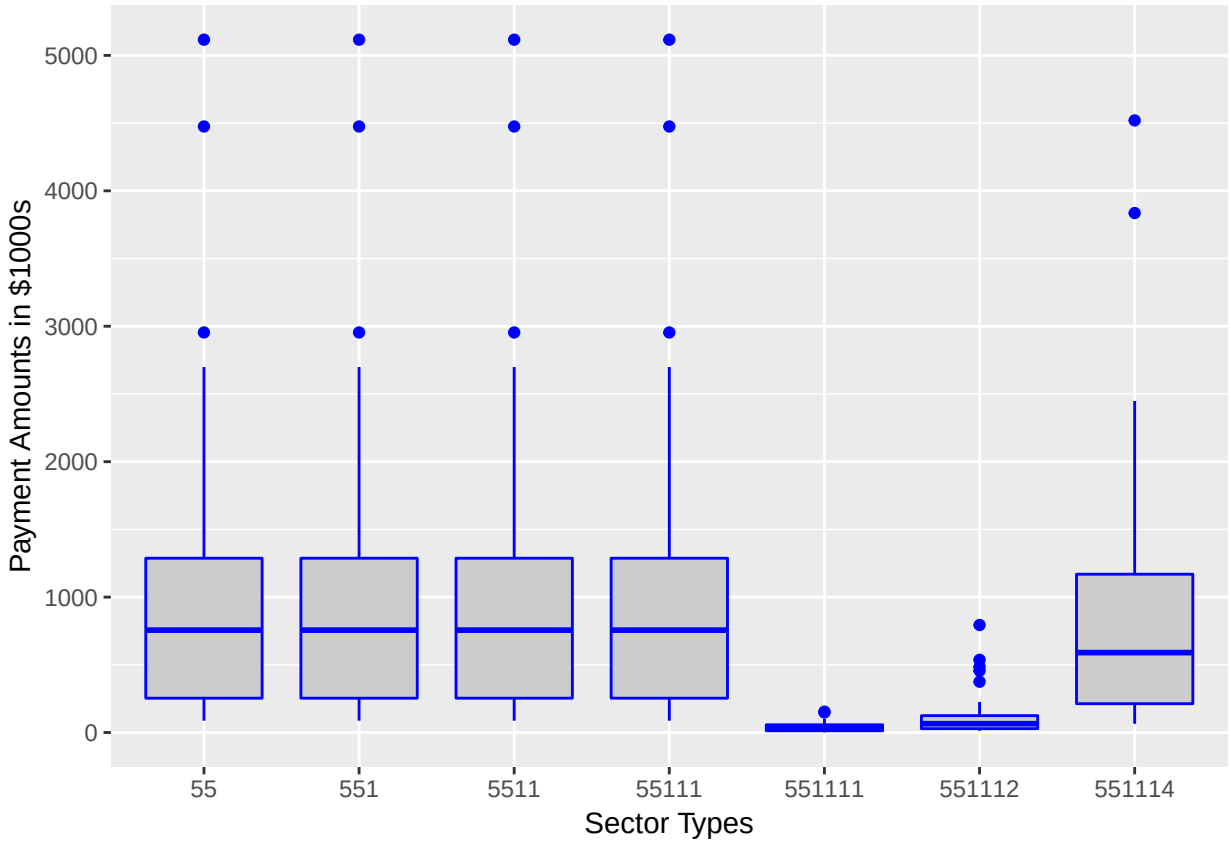


This analysis shows that the tested variable (PAYANN) does not show any significant difference between YEAR 2007 and 2012.

AVNOVA (ONE-WAY) #2

```
eco.dt <- read.csv("ECN_2012_US_55A1/ECN_2012_US_55A1_with_ann.csv")

eco.dt$SECTORID = factor(eco.dt$SECTORID,
                         labels = c("55",
                                    "551",
                                    "5511",
                                    "55111",
```

```
                                    "551111",
                                    "551112",
                                    "551114"))


require(ggplot2)

ggplot(eco.dt, aes(x = SECTORID, y = ESTAB)) +
  geom_boxplot(fill = "grey80", colour = "blue") +
  scale_x_discrete() + xlab("Sector Types") +
  ylab("Payment Amounts in $1000s")
```




```
eco.mod1 = lm(ESTAB ~ SECTORID, data = eco.dt)

summary(eco.mod1)
```

```
##
## Call:
## lm(formula = ESTAB ~ SECTORID, data = eco.dt)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -967.2 -520.2  -45.2  164.3 4061.8
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)     1.054e+03  1.214e+02   8.686  < 2e-16 ***
```

```
## SECTORID551    -1.337e-13  1.716e+02   0.000    1.000
## SECTORID5511   -1.467e-13  1.716e+02   0.000    1.000
## SECTORID55111  -3.561e-14  1.716e+02   0.000    1.000
## SECTORID551111 -1.015e+03  1.725e+02  -5.886 9.29e-09 ***
## SECTORID551112 -9.324e+02  1.716e+02  -5.432 1.04e-07 ***
## SECTORID551114 -1.600e+02  1.716e+02  -0.932    0.352
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 866.7 on 349 degrees of freedom
## Multiple R-squared:  0.1995, Adjusted R-squared:  0.1857
## F-statistic:  14.5 on 6 and 349 DF,  p-value: 8.834e-15
```

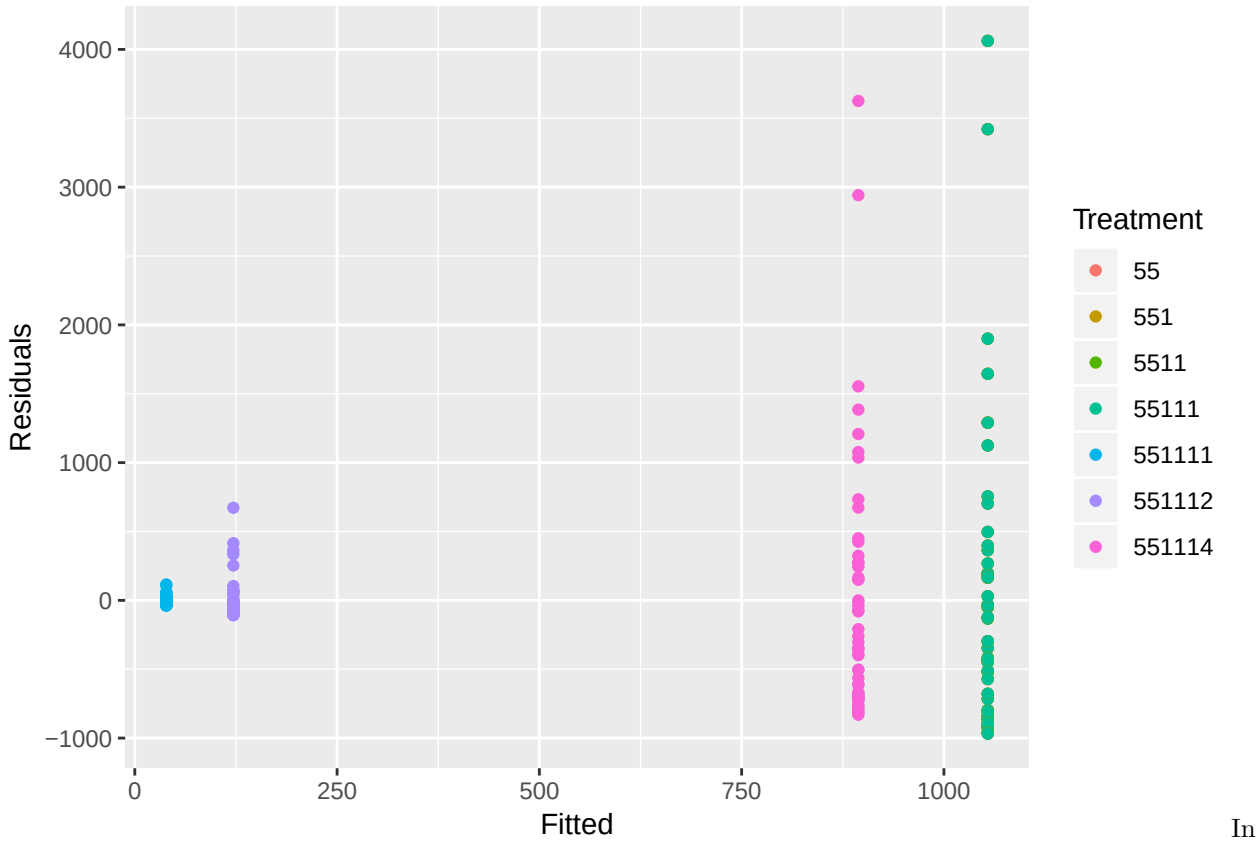
```
anova(eco.mod1)
```

```
## Analysis of Variance Table
##
## Response: ESTAB
##            Df    Sum Sq  Mean Sq F value    Pr(>F)
## SECTORID    6  65339437 10889906  14.496 8.834e-15 ***
## Residuals 349 262174362   751216
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
confint(eco.mod1)
```

```
##                     2.5 %    97.5 %
## (Intercept)      815.5148 1292.9166
## SECTORID551     -337.5741  337.5741
## SECTORID5511    -337.5741  337.5741
## SECTORID55111   -337.5741  337.5741
## SECTORID551111 -1354.5534 -676.0379
## SECTORID551112 -1269.9858 -594.8377
## SECTORID551114  -497.5349  177.6133
```

```
eco.mod = data.frame(Fitted = fitted(eco.mod1),
                     Residuals = resid(eco.mod1), Treatment = eco.dt$SECTORID)
ggplot(eco.mod, aes(Fitted, Residuals, colour = Treatment)) + geom_point()
```



In this analysis, the number of establishments(ESTAB) in differernt sector types (SECTORID) have displayed variance in their means. Furthermore, the P-Value is <0.05. Furthermore, the plot displays that sectors identified by SECTORID 551111 and 551112 display signficantly lower mean.

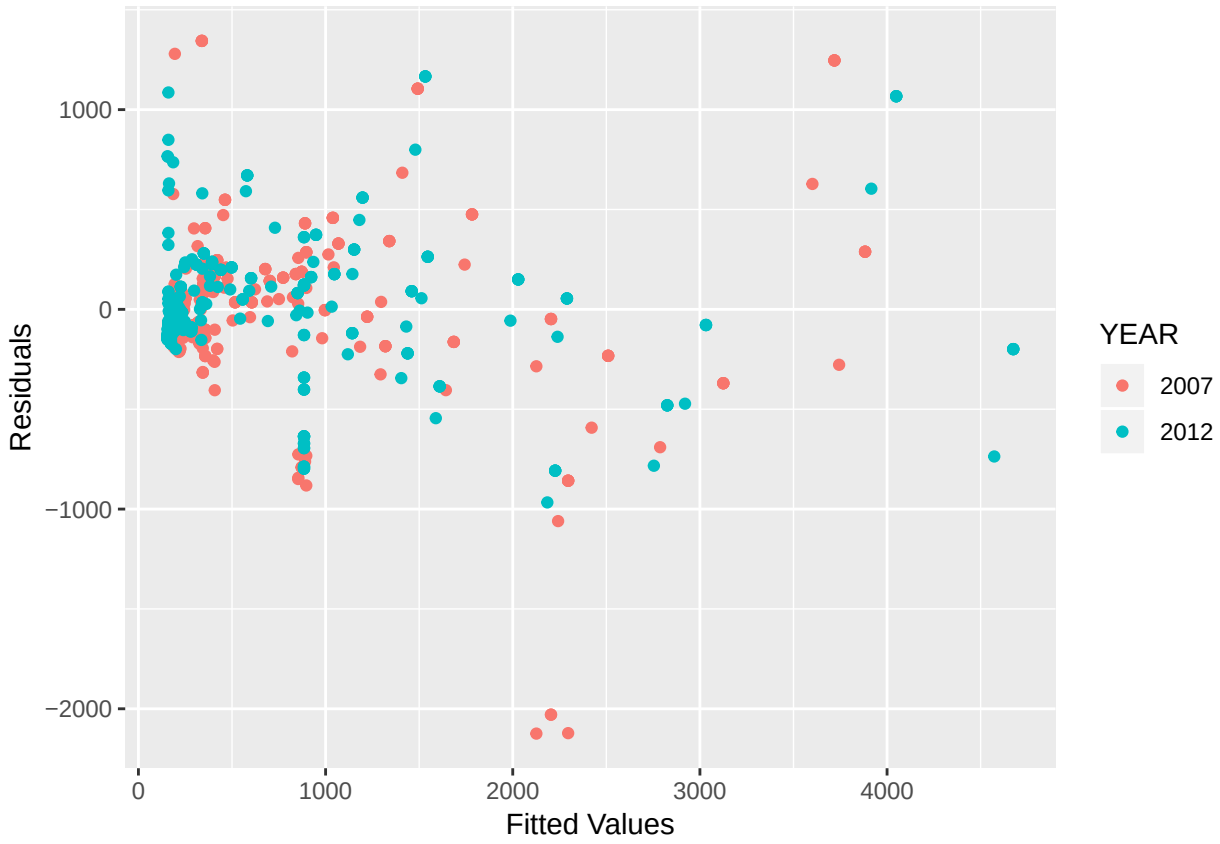
ANOVA (TWO-WAY)

```
eco.dt <- read.csv("ECN_FULL_US_55A1_with_ann.csv")

eco.dt$YEAR = factor(eco.dt$YEAR,
                         labels = c("2007","2012"))


require(ggplot2)

ggplot(eco.dt, aes(ESTAB, PAYANN, color = YEAR)) +
  geom_point()
```

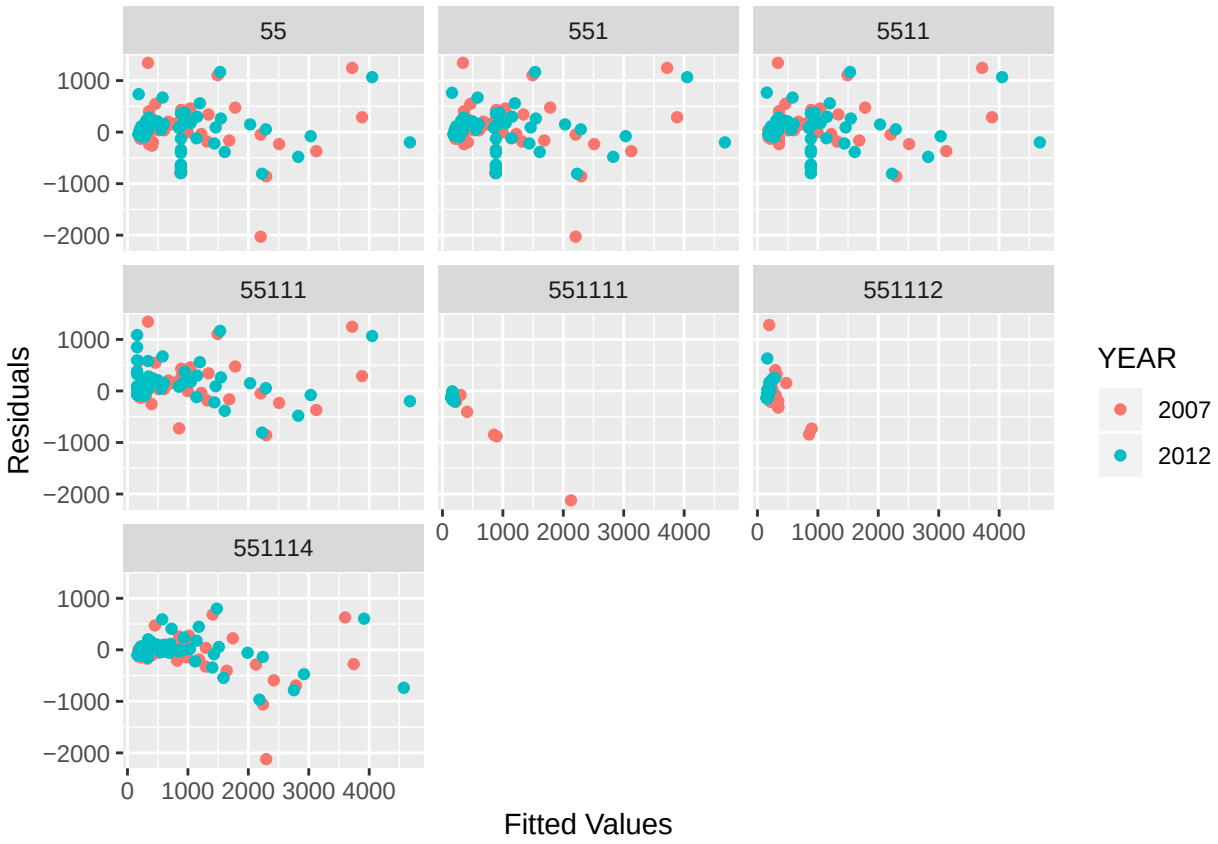

```
eco.mod1 = aov(ESTAB ~ PAYANN*YEAR, data=eco.dt)
summary(eco.mod1)
```

```
##               Df    Sum Sq   Mean Sq  F value  Pr(>F)
## PAYANN         1 518200100 518200100 3527.728 < 2e-16 ***
## YEAR           1   1320226   1320226    8.988 0.00281 **
## PAYANN:YEAR    1   1451319   1451319    9.880 0.00174 **
## Residuals    708 104000550    146893
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
eco.res = eco.dt
eco.res$M1.Fit = fitted(eco.mod1)
eco.res$M1.Resid = resid(eco.mod1)

ggplot(eco.res, aes(M1.Fit, M1.Resid, colour = YEAR)) + geom_point() +
  xlab("Fitted Values") + ylab("Residuals")
```

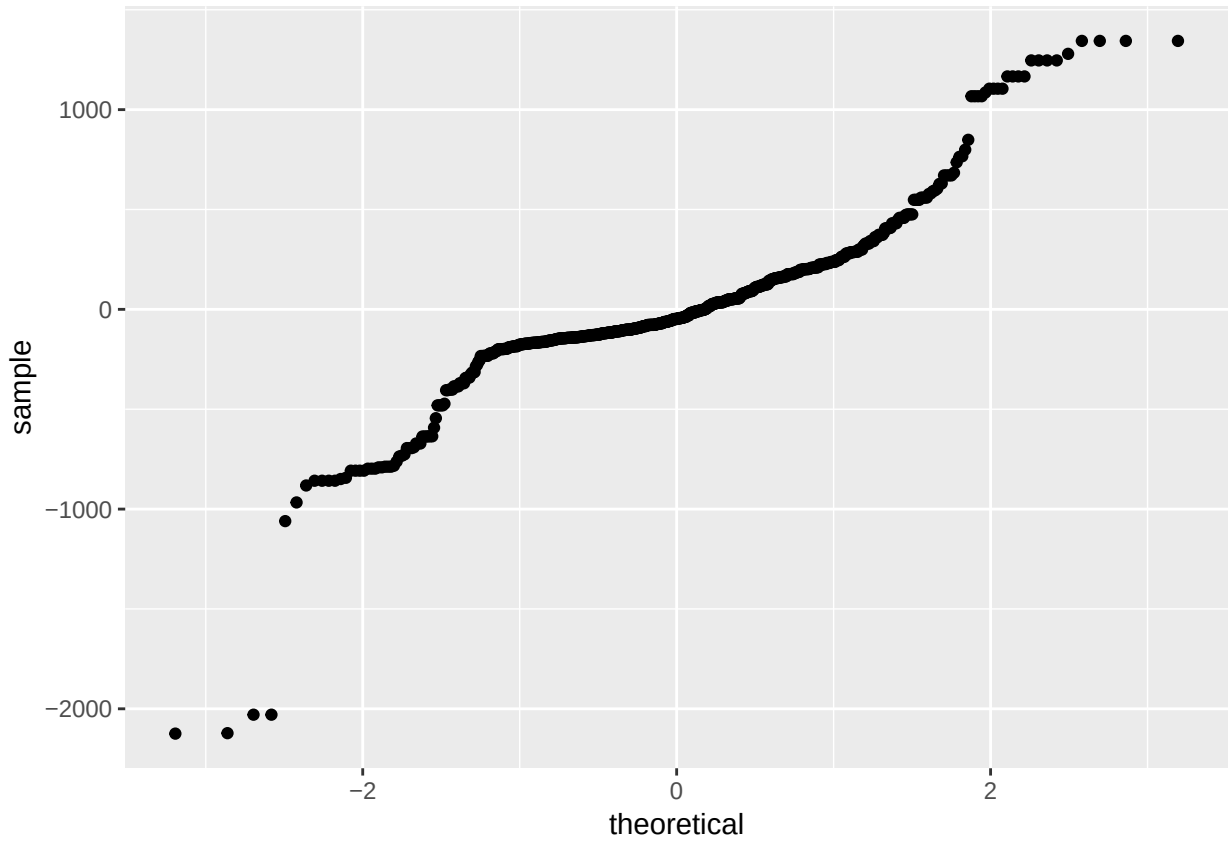



```
ggplot(eco.res, aes(M1.Fit, M1.Resid, colour = YEAR)) + geom_point() +
  xlab("Fitted Values") + ylab("Residuals") + facet_wrap( ~ SECTORID)
```




```
ggplot(eco.res, aes(M1.Fit, M1.Resid, colour = YEAR)) + geom_point() +
  xlab("Fitted Values") + ylab("Residuals") + facet_wrap( ~ YEAR)
```


```
ggplot(eco.res, aes(sample = M1.Resid)) + stat_qq()
```


1000
0
−1000
−2000
sample
−2
0
2
theoretical