Design of experiments: Lab 3

1.1. The patients are a blocking variable and the gas is the treatment. We only have $r = 1$, so we'll fit an additive model:

$$y_{ijk} = \mu + \beta_i + \tau_j + \varepsilon_{ijk}, \qquad \varepsilon_{ijk} \overset{\text{ind}}{\sim} N(0, \sigma^2)$$

with sum-to-zero restrictions

$$\sum_{i=1}^{7} \beta_i = \sum_{j=1}^{4} \tau_j = 0.$$

Here's the ANOVA table:

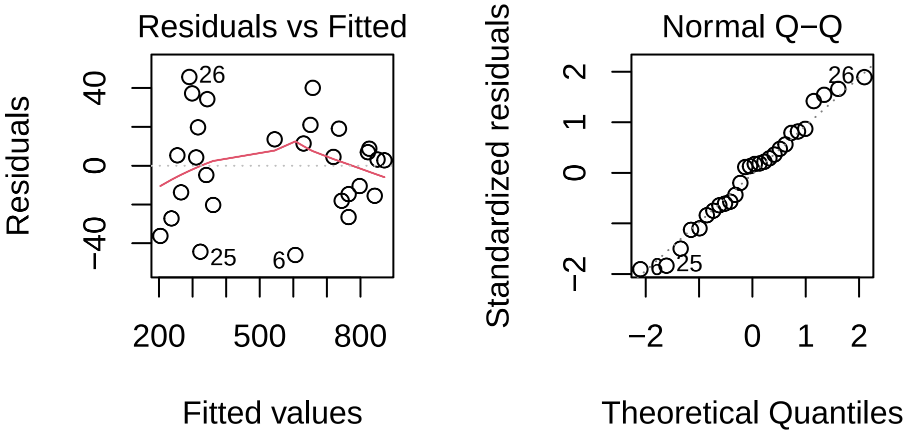
```r
library(tidyverse)
gasos = read.csv2("http://vicpena.github.io/doe/lab3/Gases.csv")
gasos$Gas = factor(gasos$Gas)
gasos$Sujeto = factor(gasos$Sujeto)
options(contrasts = c("contr.sum", "contr.poly"))
mod_add = aov(Valor ~ Sujeto+Gas, data = gasos)
summary(mod_add)
```

```
##             Df  Sum Sq Mean Sq F value   Pr(>F)    
## Sujeto       6 1471772  245295  270.61  < 2e-16 ***
## Gas          3   44827   14942   16.48 2.11e-05 ***
## Residuals   18   16316     906                     
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Both the block and the treatment are significant.

Let's take a look at the residual plots:

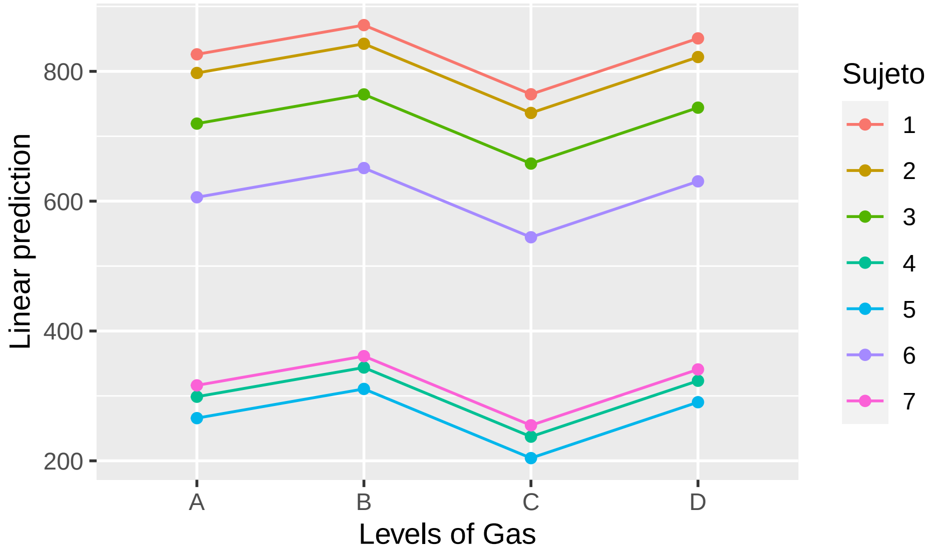
```r
par(mfrow = c(1,2))
plot(mod_add, 1:2)
```



The residual plots look fine.

Let's take a look at the effect plots

```
library(emmeans)
emmip(mod_add, Sujeto ~ Gas)
```



There are obvious differences between subjects. It seems that gas C might be significantly worse than the others (the response is distance walked in 12 minutes). We can compare the gases with `TukeyHSD`:

```
TukeyHSD(mod_add, which = "Gas")
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = Valor ~ Sujeto + Gas, data = gasos)
##
## $Gas
##           diff          lwr       upr     p adj
## B-A   45.00000   -0.4840367  90.48404 0.0530703
## C-A  -61.57143 -107.0554653 -16.08739 0.0061872
## D-A   24.57143  -20.9126082  70.05547 0.4429649
## C-B -106.57143 -152.0554653 -61.08739 0.0000178
## D-B  -20.42857  -65.9126082  25.05547 0.5929753
## D-C   86.14286   40.6588204 131.62689 0.0002338
```

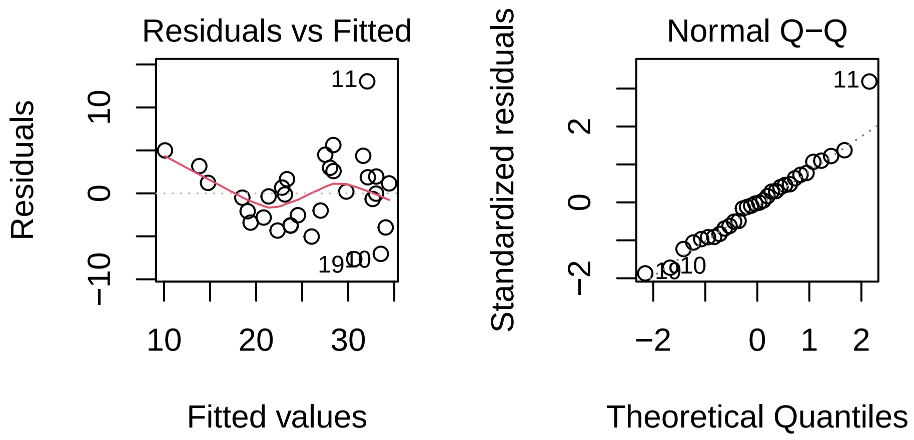
The p-values for the tests comparing gas C to the others are significant, confirming our initial intuition.

1.2. The rats are blocks and the zones are treatments. We only have one replicate ($r = 1$), so we fit an additive model with both variables. The p-values are all significant. The residuals look fine, with the exception of observation 11, which seems to be badly predicted by the model (large residual).

```
rates = read.csv2("http://vicpena.github.io/doe/lab3/Rates.csv")
rates$Sujeto = factor(rates$Sujeto)
rates$Zona = factor(rates$Zona)
mod_add = aov(Cobre ~ Sujeto+Zona, data = rates)
summary(mod_add)
```

```
##             Df Sum Sq Mean Sq F value  Pr(>F)
## Sujeto       7  703.3  100.48   3.938 0.00678 **
## Zona         3  565.9  188.63   7.393 0.00146 **
## Residuals   21  535.8   25.51
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
par(mfrow = c(1, 2))
plot(mod_add, 1:2)
```



Since we're interested in comparing zones, let's run `TukeyHSD`:

```
TukeyHSD(mod_add, which = "Zona")
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = Cobre ~ Sujeto + Zona, data = rates)
```

```
##
## $Zona
##          diff        lwr       upr     p adj
## Z1-N  10.7125   3.672793 17.752207 0.0019145
## Z2-N   9.3000   2.260293 16.339707 0.0069970
## Z3-N   4.6750  -2.364707 11.714707 0.2786427
## Z2-Z1 -1.4125  -8.452207  5.627207 0.9429572
## Z3-Z1 -6.0375 -13.077207  1.002207 0.1098481
## Z3-Z2 -4.6250 -11.664707  2.414707 0.2872316
```

There are significant differences between Z1 and N and also between Z2 and N.

1.3. We'd put the loaves randomly to avoid systematic biases (for example, some parts of the oven might be hotter than other). The batches are a block effect and the recipes are the treatment. This is another complete block design with $r = 1$ – we'll fit an additive model. There are significant batch and recipe effects. The residuals look fine (not including them for concreteness).

```
pa = read.csv2("http://vicpena.github.io/doe/lab3/Pan.csv")
pa$Receta = factor(pa$Receta)
pa$Hornada = factor(pa$Hornada)
mod_add = aov(Densidad ~ Receta + Hornada, data = pa)
summary(mod_add)
```

```
##             Df  Sum Sq Mean Sq F value Pr(>F)
## Receta       2 0.08657 0.04329   8.137 0.0118 *
## Hornada      4 0.09884 0.02471   4.645 0.0312 *
## Residuals    8 0.04256 0.00532
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Let's compare the recipes with `TukeyHSD`:

```
TukeyHSD(mod_add, which = "Receta")
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = Densidad ~ Receta + Hornada, data = pa)
##
## $Receta
##       diff        lwr         upr     p adj
## B-A -0.088 -0.2198146  0.04381462 0.1983163
## C-A -0.186 -0.3178146 -0.05418538 0.0093756
## C-B -0.098 -0.2298146  0.03381462 0.1460142
```

There are significant differences between recipes A and C.

1.4. Now we have $r = 2$, so we can fit a model with an interaction and see if we need it

```
options(contrasts = c("contr.sum", "contr.poly"))
aigua = read.csv2("http://vicpena.github.io/doe/lab3/Aigua.csv")
mod_inter = aov(Reduccio ~ Accio*Densitat, data = aigua)
summary(mod_inter)
```

```
##                Df Sum Sq Mean Sq F value   Pr(>F)
## Accio           2 182.36   91.18  32.242 1.49e-05 ***
## Densitat        3 260.32   86.77  30.685 6.55e-06 ***
## Accio:Densitat  6  30.39    5.07   1.791    0.184
## Residuals      12  33.94    2.83
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

It turns out that the interaction isn't significant, so we go ahead and fit an additive model:

```
mod_add = aov(Reduccio ~ Accio + Densitat, data = aigua)
summary(mod_add)
```

```
##             Df Sum Sq Mean Sq F value   Pr(>F)    
## Accio        2 182.36   91.18   25.51 5.58e-06 ***
## Densitat     3 260.32   86.77   24.28 1.50e-06 ***
## Residuals   18  64.33    3.57                     
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Our final model is

$$y_{ijk} = \mu + \alpha_i + \tau_j + \varepsilon_{ijk}, \qquad \varepsilon_{ijk} \overset{\text{iid}}{\sim} N(0, \sigma^2),$$

where τ_i represents the effect of the "action" and τ_j the population density. The model has sum-to-zero restrictions on the effects, as usual. The hypothesis tests are

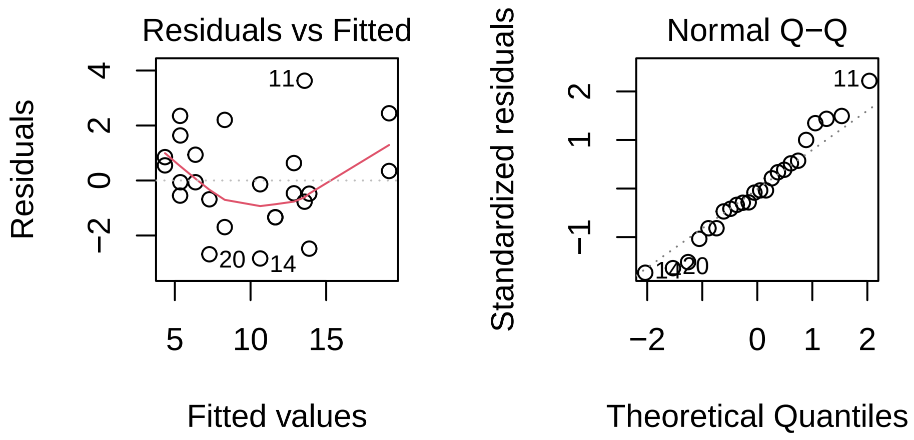
$$H_{0,\alpha} : \alpha_i = 0 \text{ for all } i, \qquad H_{1,\alpha} : \alpha_i \neq 0 \text{ for at least one } i,$$

and

$$H_{0,\tau} : \tau_j = 0 \text{ for all } j, \qquad H_{1,\tau} : \tau_j \neq 0 \text{ for at least one } j.$$

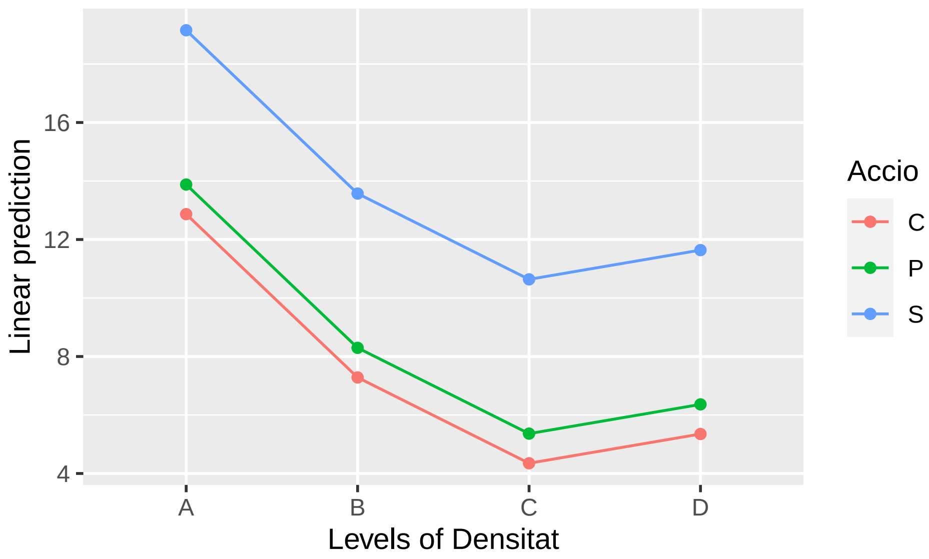
Here are the residual plots, which look fine:

```
par(mfrow = c(1, 2))
plot(mod_add, which = c(1,2))
```



And here's the effects plot:

```
emmip(mod_add, Accio ~ Densitat)
```



It seems that action S (a subsidy for changing old equipment) works best. This can be confirmed with `TukeyHSD`. If we want to find intervals for the actions, we can use `emmeans`:

```
emmeans(mod_add, ~ Accio)
```

```
##  Accio emmean    SE df lower.CL upper.CL
##  C       7.46 0.668 18     6.06     8.87
##  P       8.47 0.668 18     7.07     9.88
##  S      13.75 0.668 18    12.35    15.15
##
## Results are averaged over the levels of: Densitat
## Confidence level used: 0.95
```

The 95% confidence interval for S goes from 12.35 to 15.15.

Finally, we can find the estimated mean for action S and density A with `emmeans`, which is 19.15:

```
emmeans(mod_add, ~ Accio + Densitat)
```

```
##  Accio Densitat emmean    SE df lower.CL upper.CL
##  C     A         12.87 0.945 18    10.88    14.85
##  P     A         13.88 0.945 18    11.89    15.87
##  S     A         19.15 0.945 18    17.17    21.14
##  C     B          7.28 0.945 18     5.30     9.27
##  P     B          8.30 0.945 18     6.31    10.28
##  S     B         13.57 0.945 18    11.58    15.56
##  C     C          4.35 0.945 18     2.36     6.34
##  P     C          5.36 0.945 18     3.38     7.35
##  S     C         10.64 0.945 18     8.65    12.62
##  C     D          5.35 0.945 18     3.36     7.34
##  P     D          6.36 0.945 18     4.38     8.35
##  S     D         11.64 0.945 18     9.65    13.62
##
## Confidence level used: 0.95
```

2.1. We read in the data and convert the factors into `factor` type. We also edit the levels of `Estacion` because there was a formatting error.

```
farmac = read.csv2("http://vicpena.github.io/doe/lab3/Farmaco.csv")
farmac$Farmaco = factor(farmac$Farmaco); farmac$Estacion = factor(farmac$Estacion)
levels(farmac$Estacion)[2] = "Otoño"
```

Let's fit a model with an interaction

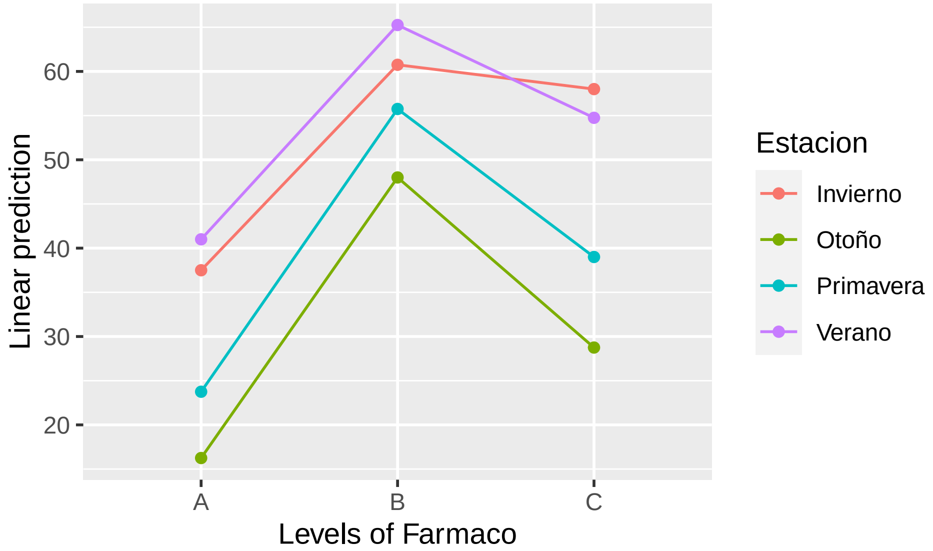
```
mod_inter = aov(Escala ~ Estacion*Farmaco, data = farmac)
summary(mod_inter)
```

```
##                   Df Sum Sq Mean Sq F value   Pr(>F)    
## Estacion           3   4176  1392.1  56.932 9.63e-14 ***
## Farmaco            2   6215  3107.7 127.097  < 2e-16 ***
## Estacion:Farmaco   6    355    59.2   2.419   0.0456 *  
## Residuals         36    880    24.5                     
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The interaction is significant. If we take a look at the residuals, the assumptions of normality and equality of variances seem to be satisfied.

Since there is an interaction, it doesn't make much sense to report an "average" treatment effect over the seasons. We can report an interaction plot like the one below

```
emmip(mod_inter, Estacion ~ Farmaco)
```



The treatment effect looks mostly the same across seasons except Winter, where treatments B and C seem to be equally effective.

If we want to find intervals, we can find them with `emmeans`

```
emmeans(mod_inter, ~ Estacion*Farmaco)
```

```
##  Estacion  Farmaco emmean   SE df lower.CL upper.CL
##  Invierno  A         37.5 2.47 36     32.5     42.5
##  Otoño     A         16.2 2.47 36     11.2     21.3
##  Primavera A         23.8 2.47 36     18.7     28.8
##  Verano    A         41.0 2.47 36     36.0     46.0
##  Invierno  B         60.8 2.47 36     55.7     65.8
##  Otoño     B         48.0 2.47 36     43.0     53.0
##  Primavera B          55.8 2.47 36      50.7     60.8
##  Verano    B          65.2 2.47 36      60.2     70.3
##  Invierno  C          58.0 2.47 36      53.0     63.0
##  Otoño     C          28.8 2.47 36      23.7     33.8
##  Primavera C          39.0 2.47 36      34.0     44.0
##  Verano    C          54.8 2.47 36      49.7     59.8
##
## Confidence level used: 0.95
```

2.2 The goal is to investigate how ozone pollution depends on climate and the size of the car population. We'll start fitting a model with an interaction.

$$y_{ijkl} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \varepsilon_{ijkl}, \qquad \varepsilon_{ijkl} \overset{\text{ind}}{\sim} N(0, \sigma^2),$$

where μ is the grand mean, α_i represents the main effect of the climate, β_j the main effect of the size of the car population, and $(\alpha\beta)_{ij}$ is the interaction term. The model has to sum-to-zero constraints

$$\sum_{i=1}^{2} \alpha_i = \sum_{j=1}^{3} \beta_j = \sum_{i=1}^{2} (\alpha\beta)_{ij} = \sum_{j=1}^{3} (\alpha\beta)_{ij} = 0,$$

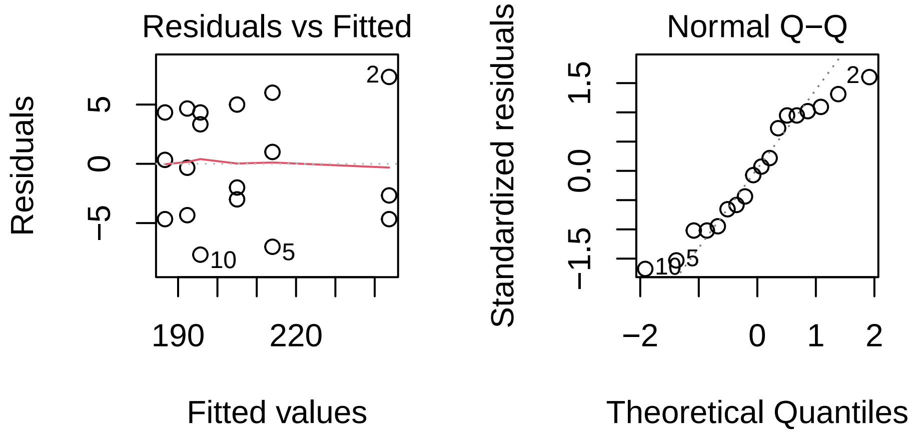
Let's fit the model with interaction

```
cotxes = read.csv2("http://vicpena.github.io/doe/lab3/Polucion.csv")
cotxes$Cotxes = factor(cotxes$Cotxes); cotxes$Clima = factor(cotxes$Clima)
options(contrasts = c("contr.sum", "contr.poly"))
mod_inter = aov(Contaminacio ~ Cotxes*Clima, data = cotxes)
summary(mod_inter)
```

```
##              Df Sum Sq Mean Sq F value   Pr(>F)    
## Cotxes        2   4953  2476.7  78.904 1.25e-07 ***
## Clima         1    998   997.6  31.781 0.000109 ***
## Cotxes:Clima  2    501   250.7   7.988 0.006229 ** 
## Residuals    12    377    31.4                     
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The interaction is significant at $\alpha = 0.05$, so we don't drop it from the model. Time now to check the residuals:

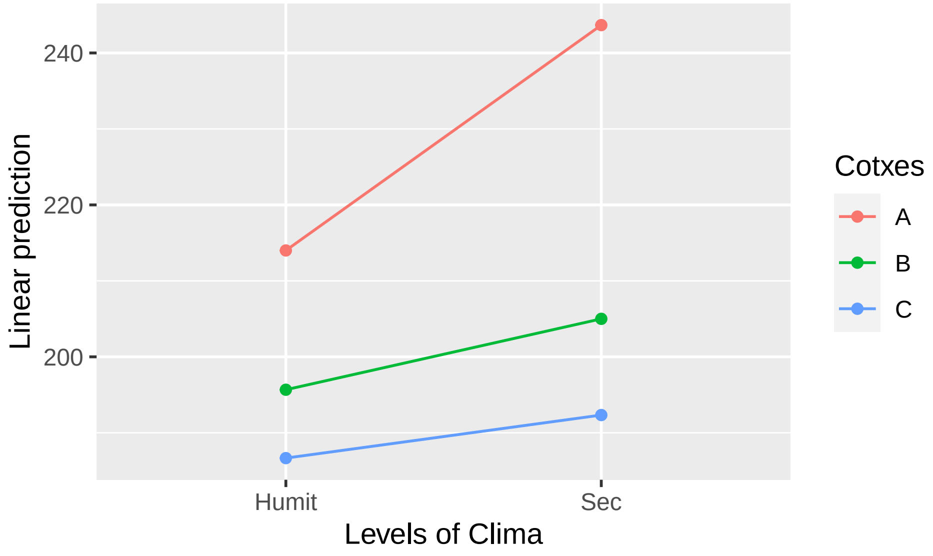
```
par(mfrow = c(1, 2))
plot(mod_inter, which = 1:2)
```



The residuals look fine.

We can take a look at the interaction plot

```
library(emmeans)
emmip(mod_inter, Cotxes ~ Clima)
```



The effect of climate depends on the car population size. The effect of climate is most noticeable when the car population is big (i.e. `Cotxes == A`). In all cases, the model predicts more ozone contamination when the climate is dry than when it is humid.

2.3 The goal is to investigate how hydrogen sulfide depends on three treatments and the type of soil. We'll start fitting a model with an interaction.

$$y_{ijkl} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \varepsilon_{ijkl}, \qquad \varepsilon_{ijkl} \overset{\text{ind}}{\sim} N(0, \sigma^2),$$

where μ is the grand mean, α_i represents the main effect of the soil, β_j the main effect of the treatment, and $(\alpha\beta)_{ij}$ is the interaction term. The model has to sum-to-zero constraints

$$\sum_{i=1}^{2} \alpha_i = \sum_{j=1}^{3} \beta_j = \sum_{i=1}^{2} (\alpha\beta)_{ij} = \sum_{j=1}^{3} (\alpha\beta)_{ij} = 0,$$

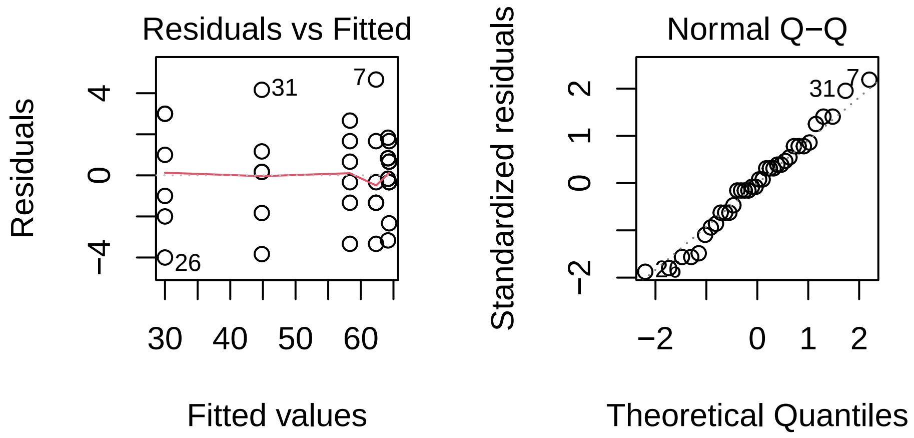
Let's read in the data and fit the model we described above

```
sulfur = read.csv2("http://vicpena.github.io/doe/lab3/Sulfuro.csv")
sulfur$Metodo = factor(sulfur$Metodo); sulfur$Sol = factor(sulfur$Sol)
mod_inter = aov(Reduccio ~ Metodo*Sol, data = sulfur)
summary(mod_inter)
```

```
##             Df Sum Sq Mean Sq F value   Pr(>F)    
## Metodo       2   5042  2521.1  462.11  < 2e-16 ***
## Sol          1    361   361.0   66.17 4.43e-09 ***
## Metodo:Sol   2    347   173.6   31.82 3.85e-08 ***
## Residuals   30    164     5.5                     
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

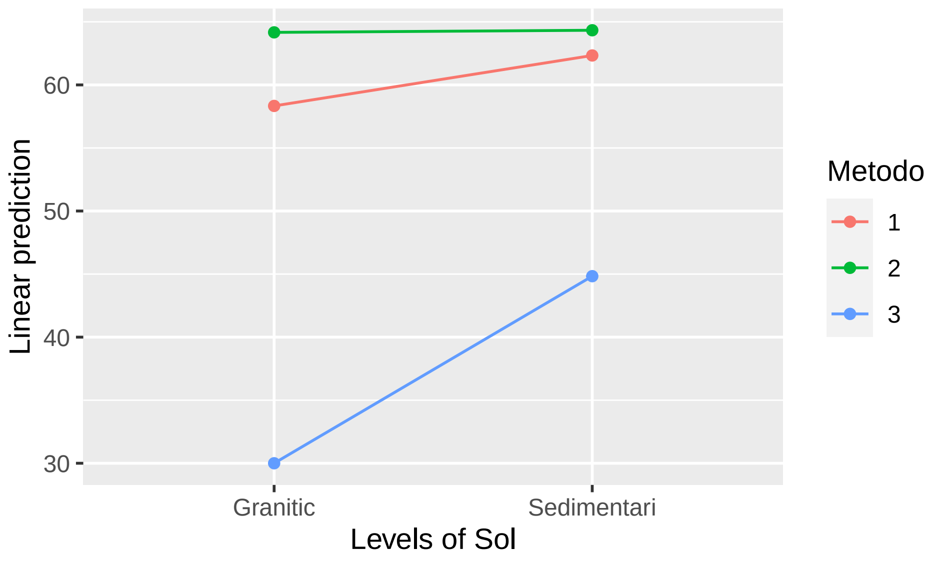
Everything is significant at $\alpha = 0.05$. Time to check residuals:

```
par(mfrow = c(1,2))
plot(mod_inter, 1:2)
```



The residuals look good. Now, let's try to understand what the model is telling us with `emmip`:

```
emmip(mod_inter, Metodo ~ Sol)
```



There is a strong interaction effect between the soil type and the treatment. Since the goal is reducing hydrogen sulfide, we recommend treatment 3: it's the one with the smallest predicted values for the outcome for both types of soil.

2.4 The model, potentially, can include 3 two-way interactions and a three-way interaction. It can be written as

$$Y_{ijkl} = \mu + \alpha_i + \beta_j + \gamma_k + (\alpha\beta)_{ij} + (\alpha\gamma)_{ik} + (\beta\gamma)_{jk} + (\alpha\beta\gamma)_{ijk} + \varepsilon_{ijkl}, \qquad \varepsilon_{ijkl} \overset{\text{ind}}{\sim} N(0, \sigma^2)$$

where μ is the grand mean, α_i is the main effect of the species, β_j is the main effect of the sex, γ_k is the main effect of the area, and then the rest of terms are interactions that relate to the main effects in an obvious manner. There are the usual sum-to-zero constraints that I'm not including here for concreteness.

This is the biggest model we'll consider. However, we'll see that most effects are not important, and we'll end up proposing a (much) reduced model.

Let's fit the model we described:

```r
mosques = read.csv2("http://vicpena.github.io/doe/lab3/Mosques.csv")
mosques$Especie = factor(mosques$Especie)
mosques$Area = factor(mosques$Area)
mosques$Genero = factor(mosques$Genero)
mod_inter = aov(Longitud ~ Especie*Area*Genero, data = mosques)
summary(mod_inter)
```

```
##                     Df Sum Sq Mean Sq F value   Pr(>F)    
## Especie              1 0.0004  0.0004   0.017 0.898040    
## Area                 1 0.0038  0.0038   0.153 0.701269    
## Genero               1 0.6337  0.6337  25.780 0.000112 ***
## Especie:Area         1 0.0038  0.0038   0.153 0.701269    
## Especie:Genero       1 0.0338  0.0338   1.373 0.258470    
## Area:Genero          1 0.0004  0.0004   0.017 0.898040    
## Especie:Area:Genero  1 0.0504  0.0504   2.051 0.171368    
## Residuals           16 0.3933  0.0246                     
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

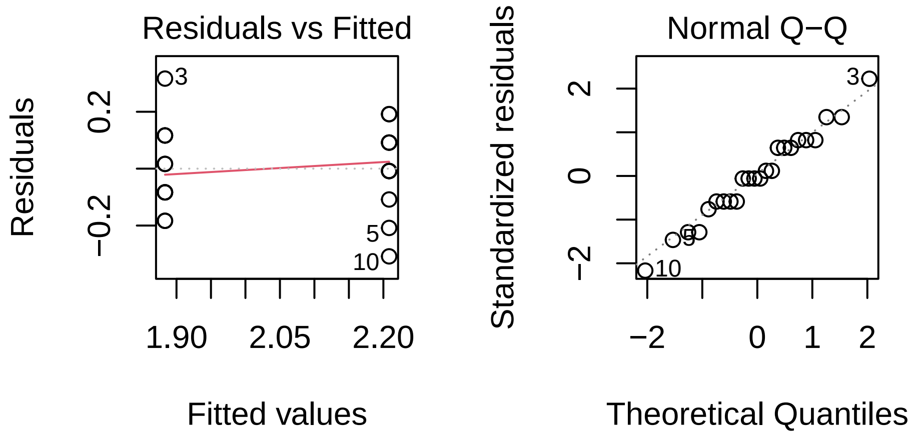
Most terms aren't significant. If, for example, we apply the backward algorithm described in the course slides, we select the following model:

```
mod_final = aov(Longitud ~  Genero, data = mosques)
summary(mod_final)
```

```
##             Df Sum Sq Mean Sq F value   Pr(>F)    
## Genero       1 0.6337  0.6337    28.7 2.23e-05 ***
## Residuals   22 0.4858  0.0221                     
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Let's check the residuals of this final model

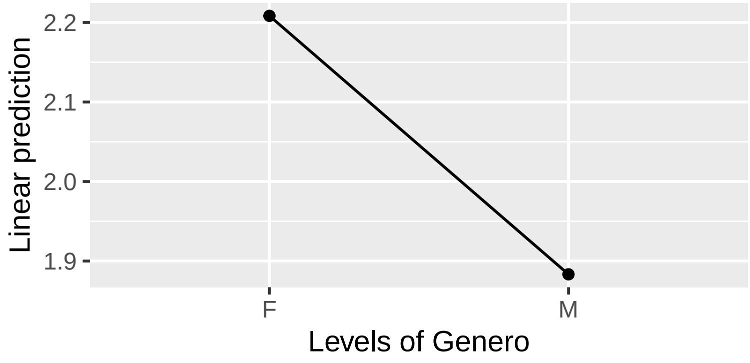
```
par(mfrow = c(1,2))
plot(mod_final, which = 1:2)
```


Residuals vs Fitted

Normal Q−Q

And finally check the effects plot

```
emmip(mod_final, ~ Genero)
```



The conclusion is that male fruit flies have shorter wings than female fruit flies.