7.1 One-Way ANOVA

ANalysis Of VAriance models are a basic type of model describing the *statistical relation* between one or more independent variables and a dependent variable. Analysis of Variance models differ from ordinary regression models in 2 ways:

(i) Independent variables may be qualitative;

(ii) If independent variables *are* quantitative, **no** assumption is made about the *nature* of the statistical relation between them and the dependent variable.

Therefore, ANOVA models are typically preferred when:

- all independent variables are qualitative, in which case you obtain the same results with regression.
- independent variables are quantitative and there is *no specification of the nature* of the statistical relation, which means that *it is not necessary to assume that* Y *and* X *are* ***linearly*** *dependent*.

So, we may start with ANOVA to study the *effects* of the independent variable(s) on the dependent and then move to regression in order to better describe of the underlying "*statistical relation*".

Single Factor experimental and observational studies are the most basic form of *comparative studies* used in practice. In a single-factor experimental study the **treatments** correspond to the *levels* of the factor, and randomization is used to assign the treatments to the experimental units.

7.1.1 Experimental Design Jargon

A **comparative experiment** is intended to answer research questions regarding the differences between the effects of imposing two or more different conditions. The conditions are the *treatments*, and they are imposed on the *experimental units*. The effects are measured using the *responses* (usually values of a single response variable).

The way treatments are assigned to experimental units is called the **design** of the experiment. Some form of *randomization* is usually used. In that case, it is a *randomized experiment* (or sometimes randomized study).

> **Example: Paper Towel Absorbency**
>
> In an experiment to investigate absorptive properties of **4** different formulations of a paper towel, *five sheets of paper towel* were randomly selected from each of the 4 types (formulation 1–4) of paper towel. **20** 6-ounce beakers of water were prepared, and the 20 paper towel sheets were *randomly* assigned to the beakers. Then, they were fully submerged in the water for 10 seconds, withdrawn, and *the amount of water absorbed be each paper towel sheet* was determined.
>
> *This is an example of a **completely randomized design**, based on a **four-level** qualitative factor.*
>
> In this experiment:
>
> ```
> * "the amount of water absorbed be each paper towel sheet" is the _response_.
> ```
>
> - "paper towel formulation" is a *factor* with 4 levels (4 different formulations $\rightarrow$ **treatments**).
> - "paper towel sheet" is the *experimental unit*.

Another example is the following:

Blood Coagulation Experiment

24 animals were randomly assigned to **4 different diets** with goal to study blood coagulation times. The samples were taken in a *random order.* This data set can be found in the `faraway` library:

```
##    coag diet
## 1    62    A
## 2    60    A
## 3    63    A
## 4    59    A
## 5    63    B
## 6    67    B
## 7    71    B
## 8    64    B
## 9    65    B
## 10   66    B
## 11   68    C
## 12   66    C
## 13   71    C
## 14   67    C
## 15   68    C
## 16   68    C
## 17   56    D
## 18   62    D
## 19   60    D
## 20   61    D
## 21   63    D
## 22   64    D
## 23   63    D
## 24   59    D
```

In this experiment:

- "coagulation time" is the *response*.
- "diet" is a *factor* with 4 levels (4 different diets A, B, C, D $\rightarrow$ 4 treatments).
- "animal" is the *"experimental unit"*.

ANOVA Terminology

In an one-way ANOVA study, we use the following terms:

- Factor: an Independent variable. They can be experimental or observational.
- Level: A particular form of the factor.
- Treatments: Factor levels or factor level combinations (if the study contains more than one factors). They provide insights into mechanisms causing the variation being studied.

  > **Control Treatments**
  > In an experiment, a control treatment is used as a baseline for comparison purposes. For example, if we want to compare diets A, B, C, D with the current diet of the animals, then the current diet can serve as the *control group diet*.

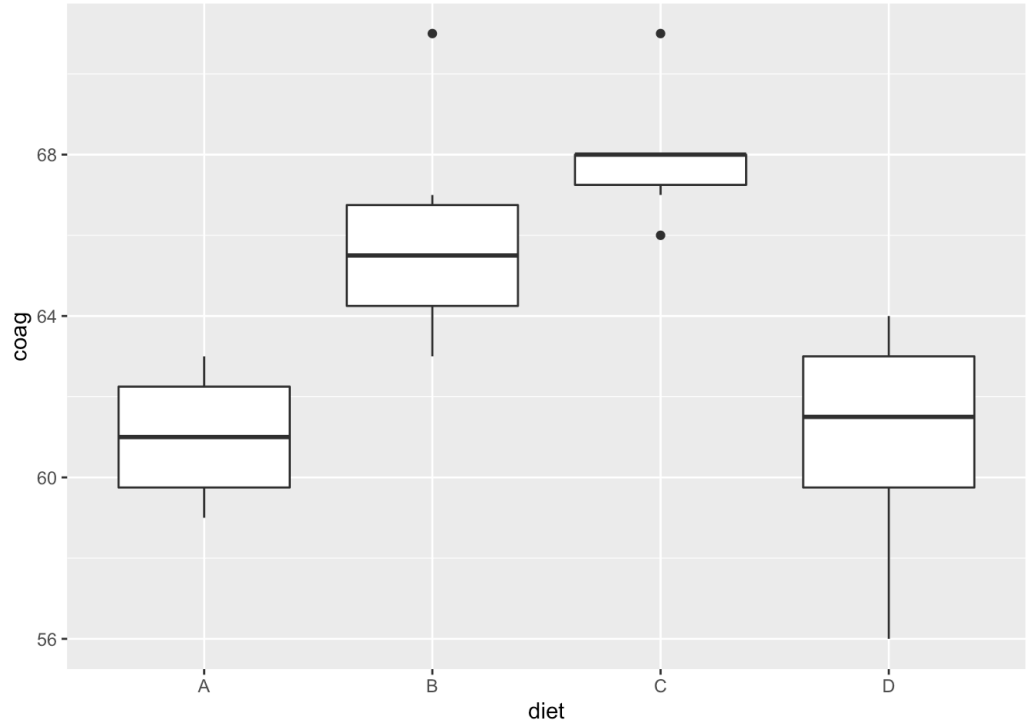
- Complete Randomized Design: Experimental units are randomly split into $\mathbf{r}$ groups, and $\mathbf{r}$ treatments are assigned, one per group.

  Blood Coagulation Experiment

  In this experiment, we do not know whether one of the Diets is the current diet of the animals, so it is uncertain whether they experimenters included a control group in their study.

  To illustrate the data in which the response is continuous and the predictor categorical, we typically use a **side-by-side boxplot**:

```
library(ggplot2)
ggplot(coagulation, aes(diet, coag)) +  geom_boxplot()
```



Based on the plot, we can only draw qualitative conclusions. In this case, we observe that diets A and D have similar coagulation times that are lower than those of diets B and C. Diet C seems to be associated with higher coagulation times compared to the rest of the diets.

7.1.2 One-Way ANOVA Model Formulation(s)

As we can see from the example above, the data from a single-factor ANOVA for a factor with r levels typically has the following structure:

group 1	y_{11}	y_{12}	y_{1n_1}
group 2	y_{21}	y_{22}	y_{2n_2}
...	...	...	...
group r	y_{r1}	y_{r2}	y_{rn_r}

Here we denote by

- r the number of groups

- n_i the number of observations in the ith group
- $n = \sum_{i=1}^{r} n_i$ the total sample size
- y_{ij} the observation j for the ith factor level.

Although we can still use a regression format with indicator variables to define an one-way ANOVA model, we usually prefer a different formulation that focuses on the means of the r groups/levels of the factor. Therefore, we have the following model definition:

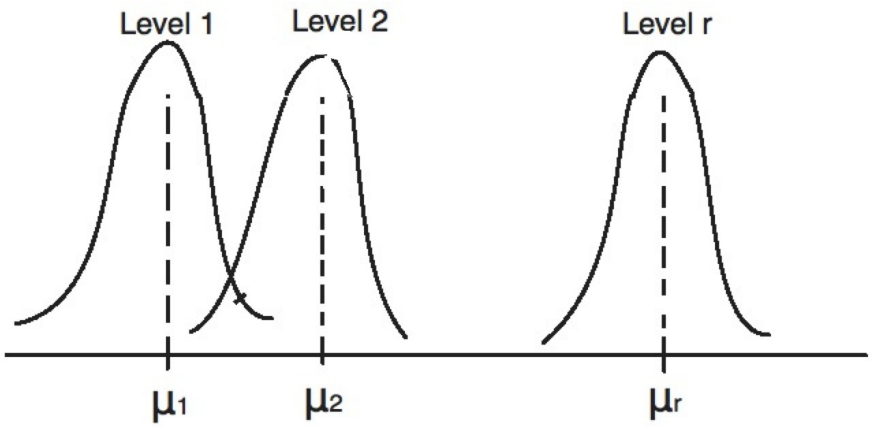
ANOVA (Cell) Means Model

Let y_{ij} be the jth observation for the ith factor level, and μ_i the *population mean* of the ith factor level (treatment), then the *Cell Means* model is defined as

$$y_{ij} = \mu_i + \varepsilon_{ij}, \ i = 1, \ldots, r; \quad j = 1, \ldots, n_i$$

where $\varepsilon_{ij} \sim^{iid} \mathcal{N}(0, \sigma^2)$.

ANOVA Model Representation



This is the analogue of the regression illustration is Section 1.4.1. In an ANOVA model, there is **no** assumption about the relationship between the means of the groups (linear or not), but the only goal is to understand the differences/distances of the means, and

ultimately assess whether those distances are statistically significant.

When we analyze data coming from an one-factor study, sometimes we want to interpret the results differently. Instead of comparing means, we prefer to choose a baseline, denoted by μ, and then *compare the* ***effect*** *of each group* to the baseline. This leads us to the following ANOVA model definition:

ANOVA Factor Effects Model

Let y_{ij} be the jth observation for the ith factor level, and α_i the *treatment effect* of the ith factor level (treatment) on the response defined as

$$\alpha_i = \mu_i - \mu$$

Then, the ANOVA model becomes

$$y_{ij} = \mu + \alpha_i + \varepsilon_{ij},\ i = 1, \ldots, r;\ j = 1, \ldots, n_i$$

where $\varepsilon_{ij} \sim^{iid} \mathcal{N}(0, \sigma^2)$.

The Cell Means model and the Factor Effects model both describe the same study and should be equivalent. However, based on the current definitions the Cell Means model has r (plus the variance) parameters to estimate, i.e. the μ_is, while the Factor Effects model has $r + 1$ (plus the variance) parameters to estimate, i.e. the μ and the α_is. This implies that the Factor Effects model in its current format is **not identifiable**. This is the reason why we need to impose *restrictions* on the parametrization of the Factor Effects models.

7.1.2.1 Different Constraints on the Parameters

Depending on what we choose the *baseline* to be, we have the following parametrizations:

- Reference Cell: $\mu = \mu_1 \Rightarrow \alpha_1 = 0$
- Sum-to-Zero: $\mu = \frac{1}{r}\sum_i \mu_i \Rightarrow \sum_i \alpha_i = 0$

- Weighted Sum-to-Zero: $\mu = \frac{1}{n}\sum_i n_i\mu_i \Rightarrow \sum_i n_i\alpha_i = 0$

Remark: The default in `R` is the Reference Cell parametrization.

How about a Regression Representation of the One-Way ANOVA Model ?

In order to write an equivalent regression model, we need to define appropriate dummy/indicator variables. In fact, the definition/coding of the dummy variables is directly linked to the ANOVA model formulation. So, the indicator variables are defined differently when we write the regression formulation of the Cell Means model compared to the Factor effects with reference cell etc.

So, for the different ANOVA model formulations, we have the following definitions of indicator variables and the corresponding regression models:

- Cell Means model:

  In this case, we need one dummy variable for each level, since there is no baseline (i.e. intercept). So, for the `coagulation example` we have:

$$X_A = \begin{cases} 1, \text{ if case from diet A} \\ 0, \text{ otherwise} \end{cases}, \; X_B = \begin{cases} 1, \text{ if case from diet B} \\ 0, \text{ otherwise} \end{cases}$$

$$X_C = \begin{cases} 1, \text{ if case from diet C} \\ 0, \text{ otherwise} \end{cases}, \; X_D = \begin{cases} 1, \text{ if case from diet D} \\ 0, \text{ otherwise} \end{cases}$$

The corresponding regression model is

$$Y_{ij} = \mu_A X_A + \mu_B X_B + \mu_C X_C + \mu_D X + \varepsilon_{ij}$$

The *partial slope coefficients* here correspond to the *means of each group*, and that is why the no intercept regression model corresponds to the cell means ANOVA model.

- Factor Effects model with Reference-Cell Constraint:

  In this case, we need $r-1$ dummy variables, with the group "missing" from the indicators' definition being the *reference level*. So, for the `coagulation example` we have:

$$X_B = \begin{cases} 1, \text{ if case from diet B} \\ 0, \text{ otherwise} \end{cases}, \; X_C = \begin{cases} 1, \text{ if case from diet C} \\ 0, \text{ otherwise} \end{cases}$$

$$X_D = \begin{cases} 1, \text{ if case from diet D} \\ 0, \text{ otherwise} \end{cases}$$

The corresponding regression model is

$$Y_{ij} = \beta_0 + \beta_1 X_B + \beta_2 X_C + \beta_3 X_D + \varepsilon_{ij}$$

Diet A is the reference level and is "obtained" when all the indicator variables are "0" (the intercept term!). Of course, by slightly changing the indicator variables, you can change the *reference level*.

- Factor Effects model with Sum-to-Zero Constraint:
  In this case, we alos need $r-1$ dummy variables that are now defined as follows:

$$X_A = \begin{cases} 1, \text{ if case from diet A} \\ -1, \text{ if case from diet D} \\ 0, \text{ otherwise} \end{cases}, \quad X_B = \begin{cases} 1, \text{ if case from diet B} \\ -1, \text{ if case from diet D} \\ 0, \text{ otherwise} \end{cases}$$

$$X_C = \begin{cases} 1, \text{ if case from diet C} \\ -1, \text{ if case from diet D} \\ 0, \text{ otherwise} \end{cases}$$

The corresponding regression model is

$$Y_{ij} = \beta_0 + \beta_1 X_A + \beta_2 X_B + \beta_3 X_C + \varepsilon_{ij}$$

In this case, the β_0 corresponds to the "grand" mean μ.

So, as you can see different parametrizations of the ANOVA model lead to different interpretation of the parameters, different regression formulations and different definitions of the corresponding design matrices. However, this **only** changes the *interpretation* and **not** the *statistical significance* of the factor levels.

We illustrate how all these are coded in `R` :

Blood Coagulation Experiment

1. *Cell Means model parametrization*:

```
cell.means = lm(coag~diet-1, data=coagulation)
```

The design matrix that corresponds to the Cell Means model and the corresponding definition of the indicator variables is:

```
model.matrix(cell.means)
```

```
##    dietA dietB dietC dietD
## 1      1     0     0     0
## 2      1     0     0     0
## 3      1     0     0     0
## 4      1     0     0     0
## 5      0     1     0     0
## 6      0     1     0     0
## 7      0     1     0     0
## 8      0     1     0     0
## 9      0     1     0     0
## 10     0     1     0     0
## 11     0     0     1     0
## 12     0     0     1     0
## 13     0     0     1     0
## 14     0     0     1     0
## 15     0     0     1     0
## 16     0     0     1     0
## 17     0     0     0     1
## 18     0     0     0     1
## 19     0     0     0     1
## 20     0     0     0     1
## 21     0     0     0     1
## 22     0     0     0     1
## 23     0     0     0     1
## 24     0     0     0     1
## attr(,"assign")
## [1] 1 1 1 1
## attr(,"contrasts")
## attr(,"contrasts")$diet
## [1] "contr.treatment"
```

2. *Factor Effects model with Reference Cell (default) parametrization*:

```
factor.effects.reference = lm(coag~diet, data=coagulation)
```

The design matrix that corresponds to the Factor Effects model with Reference Cell parametrization and the corresponding definition of the indicator variables is:

```
model.matrix(factor.effects.reference)
```

```
##    (Intercept) dietB dietC dietD
## 1            1     0     0     0
## 2            1     0     0     0
## 3            1     0     0     0
## 4            1     0     0     0
## 5            1     1     0     0
## 6            1     1     0     0
## 7            1     1     0     0
## 8            1     1     0     0
## 9            1     1     0     0
## 10           1     1     0     0
## 11           1     0     1     0
## 12           1     0     1     0
## 13           1     0     1     0
## 14           1     0     1     0
## 15           1     0     1     0
## 16           1     0     1     0
## 17           1     0     0     1
## 18           1     0     0     1
## 19           1     0     0     1
## 20           1     0     0     1
## 21           1     0     0     1
## 22           1     0     0     1
## 23           1     0     0     1
## 24           1     0     0     1
## attr(,"assign")
## [1] 0 1 1 1
## attr(,"contrasts")
## attr(,"contrasts")$diet
## [1] "contr.treatment"
```

Here, the reference level is `diet A`.

3. *Factor Effects model with Sum Constrain parametrization*:

```r
# Set R to use Sum Constraint:
contrasts(coagulation$diet) = contr.sum(4)
factor.effects.sum = lm(coag~diet, data=coagulation)
```

The corresponding design matrix is:

```r
model.matrix(factor.effects.sum)
```

```
##    (Intercept) diet1 diet2 diet3
## 1            1     1     0     0
## 2            1     1     0     0
## 3            1     1     0     0
## 4            1     1     0     0
## 5            1     0     1     0
## 6            1     0     1     0
## 7            1     0     1     0
## 8            1     0     1     0
## 9            1     0     1     0
## 10           1     0     1     0
## 11           1     0     0     1
## 12           1     0     0     1
## 13           1     0     0     1
## 14           1     0     0     1
## 15           1     0     0     1
## 16           1     0     0     1
## 17           1    -1    -1    -1
## 18           1    -1    -1    -1
## 19           1    -1    -1    -1
## 20           1    -1    -1    -1
## 21           1    -1    -1    -1
## 22           1    -1    -1    -1
## 23           1    -1    -1    -1
## 24           1    -1    -1    -1
## attr(,"assign")
## [1] 0 1 1 1
## attr(,"contrasts")
## attr(,"contrasts")$diet
##   [,1] [,2] [,3]
## A    1    0    0
## B    0    1    0
## C    0    0    1
## D   -1   -1   -1
```

The intercept in this case corresponds to the overall (grand) mean μ.

7.1.3 Fitting the ANOVA model

No matter which model we choose to work with, we need to fit the model to the data and estimate its parameters. Because all the definitions are now compatible with each other, fitting one model can easily give us the parameter estimators for the other model. To derive the parameter estimators, we start by looking at the model properties.

ANOVA Model Properties

1. $E(y_{ij}) = \mu_i$
2. $Var(y_{ij}) = Var(\varepsilon_{ij}) = \sigma^2$. Thus, all observations have the same variance, regardless of factor level.
3. $\varepsilon_{ij} \sim \mathcal{N}(0, \sigma^2)$ and independent, therefore $y_{ij} \sim \mathcal{N}(\mu_i, \sigma^2)$ and independent.

So, we can re-state the model as

$$y_{ij} \text{ are independent } \mathcal{N}(\mu_i, \sigma^2)$$

Fitting of the ANOVA Model

As before, we want to **minimize** the sum of squared deviations of the observations around their expected values with respect to the parameters:

$$RSS = \sum_{i=1}^{r} \sum_{j=1}^{n_i} (y_{ij} - \mathbb{E}(y_{ij}))^2$$

If we re-write RSS we have

$$RSS = \sum_j (y_{1j} - \mu_1)^2 + \sum_j (y_{2j} - \mu_2)^2 + \ldots + \sum_j (y_{rj} - \mu_r)^2$$

So the *least squares estimator* of μ_i, denoted by $\hat{\mu}_i$ is

$$\hat{\mu}_i = \bar{y}_{i\cdot} = \frac{1}{n_i} \sum_{j=1}^{n_i} y_{ij}$$

Using the appropriate constraints, we can easily extract the estimators for μ and α_i.

Blood Coagulation Experiment

1. *Cell Means model parametrization*:

```
cell.means = lm(coag~diet-1, data=coagulation)
summary(cell.means)
```

```
##
## Call:
## lm(formula = coag ~ diet - 1, data = coagulation)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -5.00  -1.25   0.00   1.25   5.00
##
## Coefficients:
##       Estimate Std. Error t value Pr(>|t|)
## dietA  61.0000     1.1832   51.55   <2e-16 ***
## dietB  66.0000     0.9661   68.32   <2e-16 ***
## dietC  68.0000     0.9661   70.39   <2e-16 ***
## dietD  61.0000     0.8367   72.91   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.366 on 20 degrees of freedom
## Multiple R-squared:  0.9989, Adjusted R-squared:  0.9986
## F-statistic:  4399 on 4 and 20 DF,  p-value: < 2.2e-16
```

So, we have that

$$\begin{aligned}\hat{\mu}_A &= \bar{Y}_{1\cdot} = 61\\ \hat{\mu}_B &= \bar{Y}_{2\cdot} = 66\\ \hat{\mu}_C &= \bar{Y}_{3\cdot} = 68\\ \hat{\mu}_D &= \bar{Y}_{4\cdot} = 61\end{aligned}$$

2. *Factor Effects model with Reference Cell (default) parametrization*:

```
factor.effects.reference = lm(coag~diet, data=coagulation)
summary(factor.effects.reference)
```

```
##
## Call:
## lm(formula = coag ~ diet, data = coagulation)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -5.00  -1.25   0.00   1.25   5.00
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  64.0000     0.4979 128.537  < 2e-16 ***
## diet1        -3.0000     0.9736  -3.081 0.005889 **
## diet2         2.0000     0.8453   2.366 0.028195 *
## diet3         4.0000     0.8453   4.732 0.000128 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.366 on 20 degrees of freedom
## Multiple R-squared:  0.6706, Adjusted R-squared:  0.6212
## F-statistic: 13.57 on 3 and 20 DF,  p-value: 4.658e-05
```

In the *Reference Cell* parametrization, we assume that α_A is equal to 0 and the remaining **factor effects** are obtained directly using the summary output. Specifically, $\alpha_B = 5.000e+00$, $\alpha_C = 7.000e+00$, $\alpha_D = 2.991e-1$, and of course $\hat{mu} = 6.100e+01$. We can recover the means $\hat{\mu}_A$ etc. as follows:

$$\hat{\mu}_A = \hat{\mu} + \hat{\alpha}_A = 61 + 0 = 61$$

$$\hat{\mu}_B = \hat{\mu} + \hat{\alpha}_B = 61 + 5 = 66$$

$$\hat{\mu}_C = \hat{\mu} + \hat{\alpha}_C = 61 + 7 = 68$$

$$\hat{\mu}_D = \hat{\mu} + \hat{\alpha}_D = 61 + 10^{-15} \approx 61$$

So, we recovered all the means and are exactly the same (up to rounding errors) to the means we obtained with the cell means model.

3. *Factor Effects model with Sum Constrain parametrization*:

```
factor.effects.sum = lm(coag~diet, data=coagulation)
summary(factor.effects.sum)
```

```
##
## Call:
## lm(formula = coag ~ diet, data = coagulation)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -5.00  -1.25   0.00   1.25   5.00
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  64.0000     0.4979 128.537  < 2e-16 ***
## diet1        -3.0000     0.9736  -3.081 0.005889 **
## diet2         2.0000     0.8453   2.366 0.028195 *
## diet3         4.0000     0.8453   4.732 0.000128 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.366 on 20 degrees of freedom
## Multiple R-squared:  0.6706, Adjusted R-squared:  0.6212
## F-statistic: 13.57 on 3 and 20 DF,  p-value: 4.658e-05
```

In the *Sum Constraint* parametrization, we assume that $\sum_i \alpha_i$ is equal to 0 and the **factor effects** are obtained directly using the summary output as follows: $\alpha_A = -3$, $\alpha_B = 1$, $\alpha_C = 4$, $\alpha_D = -(-3 + 2 + 4) = -3$. We can recover the means $\hat{\mu}_A$ etc. as follows:

$$\hat{\mu}_A = \hat{\mu} + \hat{\alpha}_A = 64 - 3 = 61$$

$$\hat{\mu}_B = \hat{\mu} + \hat{\alpha}_B = 64 + 2 = 66$$

$$\hat{\mu}_C = \hat{\mu} + \hat{\alpha}_C = 64 + 4 = 68$$

$$\hat{\mu}_D = \hat{\mu} + \hat{\alpha}_D = 64 - 3 \approx 61$$

So, we recovered all the means and are exactly the same (up to rounding errors) to the means we obtained with the cell means model and the factor effects model with the reference cell parametrization.

Fitted Values & Residuals

The LS fit for y_{ij} is the corresponding group mean

$$\hat{y}_{ij} = \bar{y}_{i\cdot}$$

and the residuals are defined as usual

$$r_{ij} = y_{ij} - \hat{y}_{ij} = y_{ij} - \bar{y}_{i\cdot}$$

with RSS being

$$\sum_{i=1}^{r}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_{i\cdot})^2,$$

i.e. the within-group variation.

We can also verify in our example that the fitted values are the same, independent of the model parametrization (of course up to rounding errors):

Blood Coagulation Experiment

Compute the fitted values obtained with the three different fitted models before:

```
fitted.cell.means = cell.means$fitted
fitted.reference = factor.effects.reference$fitted
fitted.sum = factor.effects.sum$fitted
```

and compare

```
cbind(fitted.cell.means, fitted.reference, fitted.sum)
```

```
##    fitted.cell.means fitted.reference fitted.sum
## 1                 61                61         61
## 2                 61                61         61
## 3                 61                61         61
## 4                 61                61         61
## 5                 66                66         66
## 6                 66                66         66
## 7                 66                66         66
## 8                 66                66         66
## 9                 66                66         66
## 10                66                66         66
## 11                68                68         68
## 12                68                68         68
## 13                68                68         68
## 14                68                68         68
## 15                68                68         68
## 16                68                68         68
## 17                61                61         61
## 18                61                61         61
## 19                61                61         61
## 20                61                61         61
## 21                61                61         61
## 22                61                61         61
## 23                61                61         61
## 24                61                61         61
```

As you can see (up to rounding errors) the fitted values are the same no matter what the parametrization is.

As before we can write down the **ANOVA Table** that corresponds to a single factor study as follows:

Source of Variation	SS	df	MS
Between Groups	$FSS = \sum n_i(\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot})^2$	$r-1$	$\frac{FSS}{r-1}$
Error (within groups)	$RSS = \sum\sum(y_{ij} - \bar{y}_{i\cdot})^2$	$n-r$	$\frac{RSS}{n-r}$
Total	$TSS = \sum\sum(y_{ij} - \bar{y}_{\cdot\cdot})^2$	$n-1$	

Blood Coagulation Experiment We can obtain this table in `R` using the `aov` function for any model or using the `anova` function for any of the factor effects fitted models:

```
aov(factor.effects.sum)
```

```
## Call:
##    aov(formula = factor.effects.sum)
##
## Terms:
##                 diet Residuals
## Sum of Squares   228       112
## Deg. of Freedom    3        20
##
## Residual standard error: 2.366432
## Estimated effects may be unbalanced
```

```
anova(factor.effects.sum)
```

```
## Analysis of Variance Table
##
## Response: coag
##           Df Sum Sq Mean Sq F value    Pr(>F)
## diet       3    228    76.0  13.571 4.658e-05 ***
## Residuals 20    112     5.6
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Testing for equality of Means

All this discussion leads us to the main ANOVA hypothesis test a test for the equality of all the means:

$$\begin{cases} H_0 : \mu_1 = \mu_2 = \ldots = \mu_r \\ H_\alpha : \text{ not all } \mu_i, \ i = 1, \ldots, r \text{ are equal} \end{cases}$$

or in terms of models

$$\begin{cases} H_0 : y_{ij} = \mu + \varepsilon_{ij} \\ H_\alpha : y_{ij} = \mu + \alpha_i + \varepsilon_{ij} \end{cases}$$

or in terms of factor effects

$$\begin{cases} H_0 : \alpha_1 = \alpha_2 = \ldots = \alpha_{r-1} = 0 \\ H_\alpha : \text{ not all } \alpha_i, \ i = 1, \ldots, r-1 \text{ are equal to } 0 \end{cases}$$

The test statistic is defined as a **partial F test** between two nested models

$$\frac{(RSS_0 - RSS_\alpha)/(r-1)}{RSS_\alpha/(n-r)} \sim F_{r-1,n-r}, \text{ under the } H_0$$

This test statistic can also (equivalently) be expressed as

$$\frac{FSS/(r-1)}{RSS/(n-r)} = \frac{\text{Between-group Variation}/(r-1)}{\text{Within-group Variation}/(n-r)},$$

where FSS, RSS are defined in the ANOVA table.

Blood Coagulation Experiment

The partial F test that corresponds to the aforementioned hypothesis in `R` is computed as follows:

```
null.model = lm(coag ~ 1, data=coagulation)
anova(null.model, cell.means)
```

```
## Analysis of Variance Table
##
## Model 1: coag ~ 1
## Model 2: coag ~ diet - 1
##   Res.Df RSS Df Sum of Sq      F    Pr(>F)
## 1     23 340
## 2     20 112  3       228 13.571 4.658e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

It does not matter which coding is used for the mean/effects; the results would be the *same*. Indeed,

```
anova(null.model, factor.effects.reference)
```

```
## Analysis of Variance Table
##
## Model 1: coag ~ 1
## Model 2: coag ~ diet
##   Res.Df RSS Df Sum of Sq      F    Pr(>F)
## 1     23 340
## 2     20 112  3       228 13.571 4.658e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We can also obtain the *same* results by using the ANOVA table F test for any of the factor effects parametrizations:

```
anova(factor.effects.reference)
```

```
## Analysis of Variance Table
##
## Response: coag
##           Df Sum Sq Mean Sq F value    Pr(>F)
## diet       3    228    76.0  13.571 4.658e-05 ***
## Residuals 20    112     5.6
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

In both cases, the p-value is much less than $\alpha = 5\%$, so we reject the null and conclude that there are differences among the different types of diet.

7.1.4 Diagnostics for ANOVA Models

The model assumptions in one-way ANOVA models are exactly the same (minus the linearity assumption) as for a regression model. Therefore, the diagnostics (and remedial measures when needed are the same). Indeed,

- We check for unusual observations: high leverage points, outliers, highly influential observations.
- We check the constancy of variance assumption using residual plots or statistical tests.
- We check the normality assumption using plots or tests.

If any of these assumptions fails, then we take appropriate actions to remedy the issue.

We will only mention here an additional test for equality of variances tailored to the needs of an ANOVA model:

Levene's Test for Equality of Variances

$$\begin{cases} H_0: \text{ All group variances are equal.} \\ H_\alpha: \text{At least one group variances is different.} \end{cases}$$

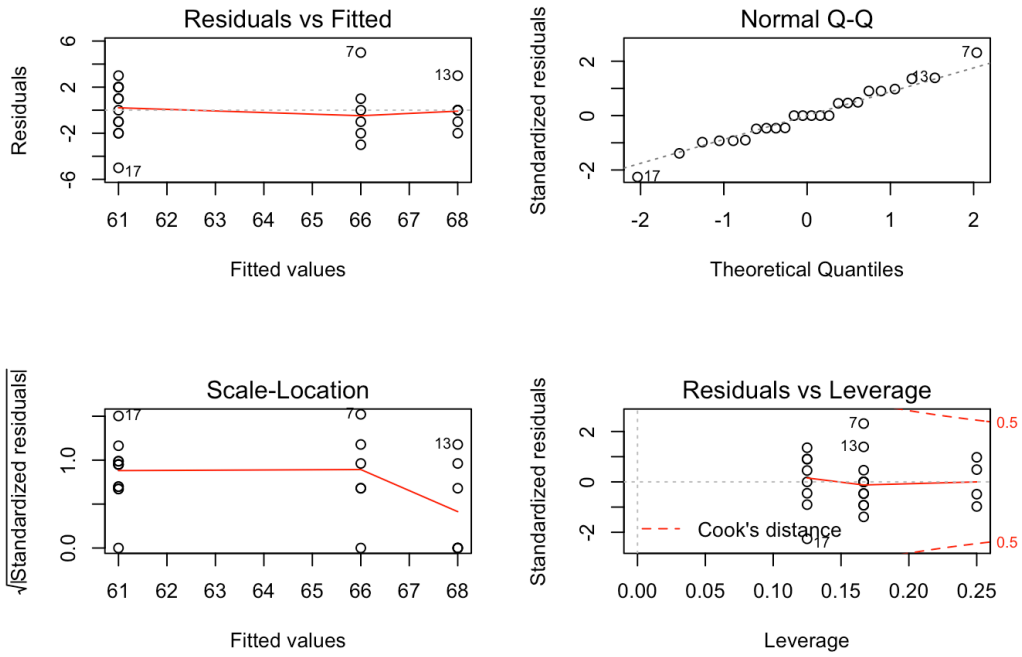
1. Run Regression abs(residuals)$\sim$ X}, i.e. use abs(residuals) as the response in a new one-way ANOVA.
2. If the p-value for the F-test **is greater than 1% level**, then we conclude that there is no evidence of a non-constant variance.

Let's check the diagnostics in our example:

Blood Coagulation Experiment

We start by looking at the residual plots:

```
par(mfrow=c(2,2))
plot(factor.effects.sum)
```



The normality QQ plot seems to be ok, specially given the sample size, but there are some concerns about the constant variance assumption, so we need to investigate this further using Levene's test:

```
levene.model = lm(abs(factor.effects.sum$res) ~ diet, coagulation)
summary(levene.model)
```

```
##
## Call:
## lm(formula = abs(factor.effects.sum$res) ~ diet, data = coagulation)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -2.000 -1.000  0.000  0.625  3.000
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   1.6250     0.3013   5.394  2.8e-05 ***
## diet1        -0.1250     0.5891  -0.212    0.834
## diet2         0.3750     0.5115   0.733    0.472
## diet3        -0.6250     0.5115  -1.222    0.236
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.432 on 20 degrees of freedom
## Multiple R-squared:  0.09559,    Adjusted R-squared:  -0.04007
## F-statistic: 0.7046 on 3 and 20 DF,  p-value: 0.5604
```

The p-value of the overall F test is $0.5604 > 0.01$ which means that there is no evidence that the variances across different groups are not constant.