MATH 60604A Statistical Modelling

Chapter 2 Part 3: t-tests and ANOVA

Juliana Schulz

---

1) t-test

Motivating Example

Does paying by credit vs. cash effect the amount consumers are willing to pay? Here we consider the amount consumers are willing to pay for two tickets for the last regular season game of the Boston Celtics (NBA basketball team). There were a total of 64 subjects who were *randomly* placed into one of two groups:

- group 1: must pay by cash
- group 2: must pay by credit card

Subjects were asked to fill out a questionnaire asking how much they were willing to pay for the tickets.

The data are saved in the `comp` file:

```
comp<-read.csv("Data/comp.csv")
head(comp)
```

```
##   off grp
## 1  62   0
## 2  44   0
## 3  46   0
## 4  48   0
## 5  50   0
## 6  58   0
```

The variable `off` gives the amount willing to offer (in $) and `grp` is the group indicator (grp=0 for cash, and grp=1 for credit). Quick overview of the data:

```
summary(comp)
```

```
##       off              grp        
##  Min.   : 34.00   Min.   :0.0000  
##  1st Qu.: 52.00   1st Qu.:0.0000  
##  Median : 61.00   Median :0.0000  
##  Mean   : 63.88   Mean   :0.4844  
##  3rd Qu.: 70.50   3rd Qu.:1.0000  
##  Max.   :144.00   Max.   :1.0000
```

```
table(comp$grp)
```

```
## 
##  0  1 
## 33 31
```

Classic t-test

We're interested in comparing the mean (average amount individuals are willing to spend) for the two groups (cash vs. credit). Thus, we're interested in testing

$$H_0: \mu_1 = \mu_2 \quad \text{vs.} \quad H_1: \mu_1 \neq \mu_2$$

or equivalently,

$$H_0: \mu_1 - \mu_2 = 0 \quad \text{vs.} \quad H_1: \mu_1 - \mu_2 \neq 0$$

where μ_1 represents the mean for group 1 (cash, i.e. $grp = 0$) and μ_2 represents the mean for group 2 (credit, i.e. $grp = 1$).

We can do this using a classic t-test:

```r
# t-test (two-sided, bilateral)
t.test(off~grp,alternative="two.sided",var.equal=TRUE,conf.level=0.95,data=comp)
```

```
##
##  Two Sample t-test
##
## data:  off by grp
## t = -3.4775, df = 62, p-value = 0.0009308
## alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0
## 95 percent confidence interval:
##  -23.633106  -6.380579
## sample estimates:
## mean in group 0 mean in group 1
##        56.60606        71.61290
```

We obtain a test statistic value of $t = -3.4775$, and a p-value of 0.0009308. Since $p < \alpha$ for any reasonable α (eg. 5%), we can reject H_0 and conclude that the μ_1 and μ_2 are significantly different, that is, there is a significant difference in the amount people are willing to pay for the tickets. We also obtain a 95% confidence interval for the difference $\mu_1 - \mu_2$: $(-23.633106, -6.380579)$.

We can also formulate this as a simple linear regression model. Let Y denote the offer amount and X be the indicator variable for credit (i.e., $X = 1$ for credit, i.e. group=1, and $X = 0$ for cash, i.e. group 0). We can write the model as

$$Y = \beta_0 + \beta_1 X + \epsilon$$

The hypotheses can then be expressed in terms of the linear regression model as

$$H_0: \beta_1 = 0 \quad \text{vs.} \quad H_1: \beta_1 \neq 0$$

```r
# En utilisant un modele de regression
# Using linear regression
mod1<-lm(off~grp,data=comp)
summary(mod1)
```

```
##
## Call:
## lm(formula = off ~ grp, data = comp)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -37.613  -8.858  -0.606   5.394  72.387
##
## Coefficients:
```

```
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   56.606      3.003  18.848  < 2e-16 ***
## grp           15.007      4.315   3.478 0.000931 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 17.25 on 62 degrees of freedom
## Multiple R-squared:  0.1632, Adjusted R-squared:  0.1497
## F-statistic: 12.09 on 1 and 62 DF,  p-value: 0.0009308
```

```r
# alternative: as.factor(grp)
summary(lm(off~as.factor(grp),data=comp))
```

```
##
## Call:
## lm(formula = off ~ as.factor(grp), data = comp)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -37.613  -8.858  -0.606   5.394  72.387
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)       56.606      3.003  18.848  < 2e-16 ***
## as.factor(grp)1   15.007      4.315   3.478 0.000931 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 17.25 on 62 degrees of freedom
## Multiple R-squared:  0.1632, Adjusted R-squared:  0.1497
## F-statistic: 12.09 on 1 and 62 DF,  p-value: 0.0009308
```

We obtain the same results as before! The estimated mean difference between the credit group and cash group is 15.007. This difference is significantly different from zero, with the test statistic value of $t = 3.478$ and a p-value of 0.000931, exactly as before! We can also obtain a 95% confidence interval:

```r
confint(mod1)
```

```
##                 2.5 %   97.5 %
## (Intercept) 50.602436 62.60969
## grp          6.380579 23.63311
```

Note that now the confidence interval is $(6.380579, 23.63311)$, which is *reversed* to what we saw before as now we are considering the difference $\mu_2 - \mu_1$ rather than $\mu_1 - \mu_2$. But the results are indeed exactly what we had before.

Side note: here we did a two-sided test of the for $H_0 : \mu_1 = \mu_2$ vs. $H_1 : \mu_1 \neq \mu_2$, so we are testing whether there is a difference. If we were rather interested in testing whether those paying credit pay more, then our test would be $H_1 : \mu_1 < \mu_2$. To do this in R:

```r
# one sided / unilateral
t.test(off~grp,alternative="less",var.equal=TRUE,conf.level=0.95,data=comp)
```

```
##
##  Two Sample t-test
##
## data:  off by grp
## t = -3.4775, df = 62, p-value = 0.0004654
```

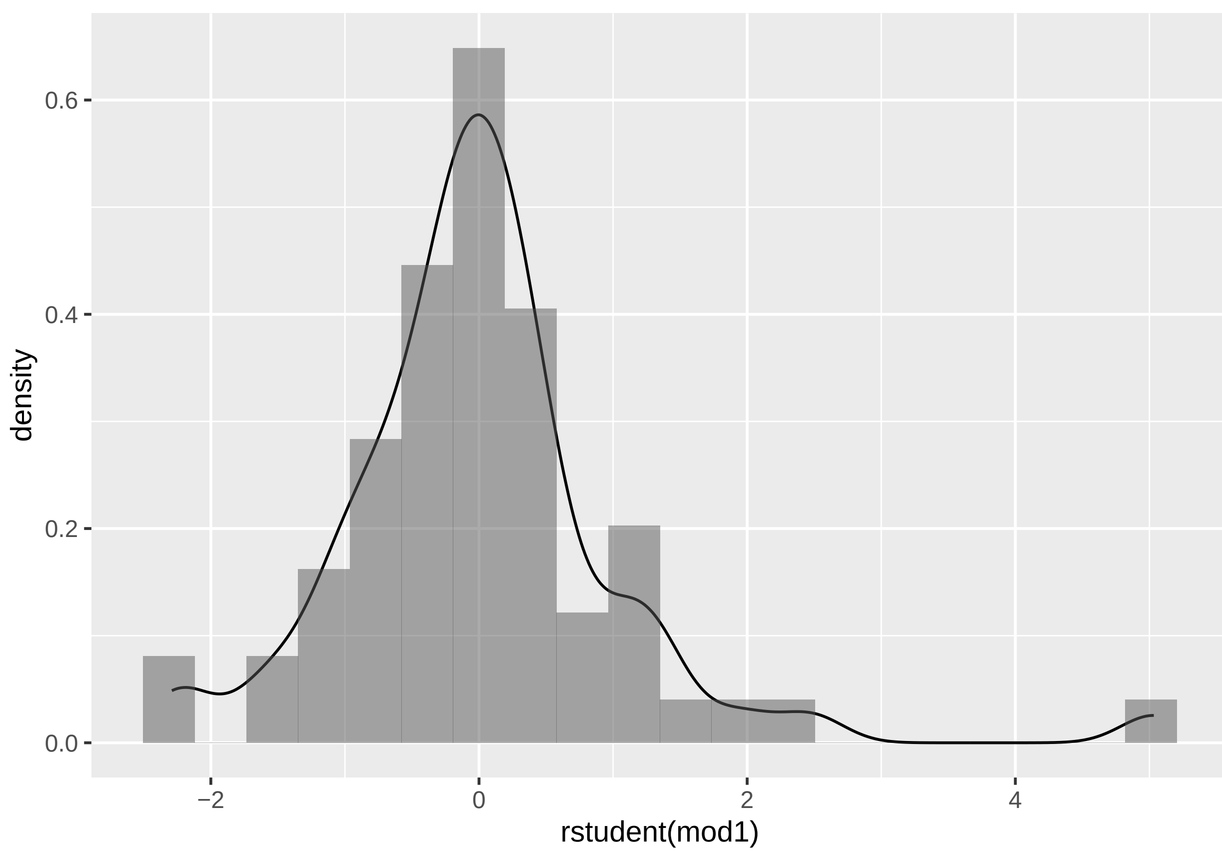
```
## alternative hypothesis: true difference in means between group 0 and group 1 is less than 0
## 95 percent confidence interval:
##       -Inf -7.801052
## sample estimates:
## mean in group 0 mean in group 1
##        56.60606        71.61290
```

We see that the value of the test statistic is the same, but the p-value is different since now we have a direction in the test. The p-value is 0.0004654*, which is less than* $\alpha = 5\%$ *and so we can reject* H_0 *and conclude that the mean amount offered for those paying by credit is significantly larger than the amount offered for those paying by cash.*

Assumptions

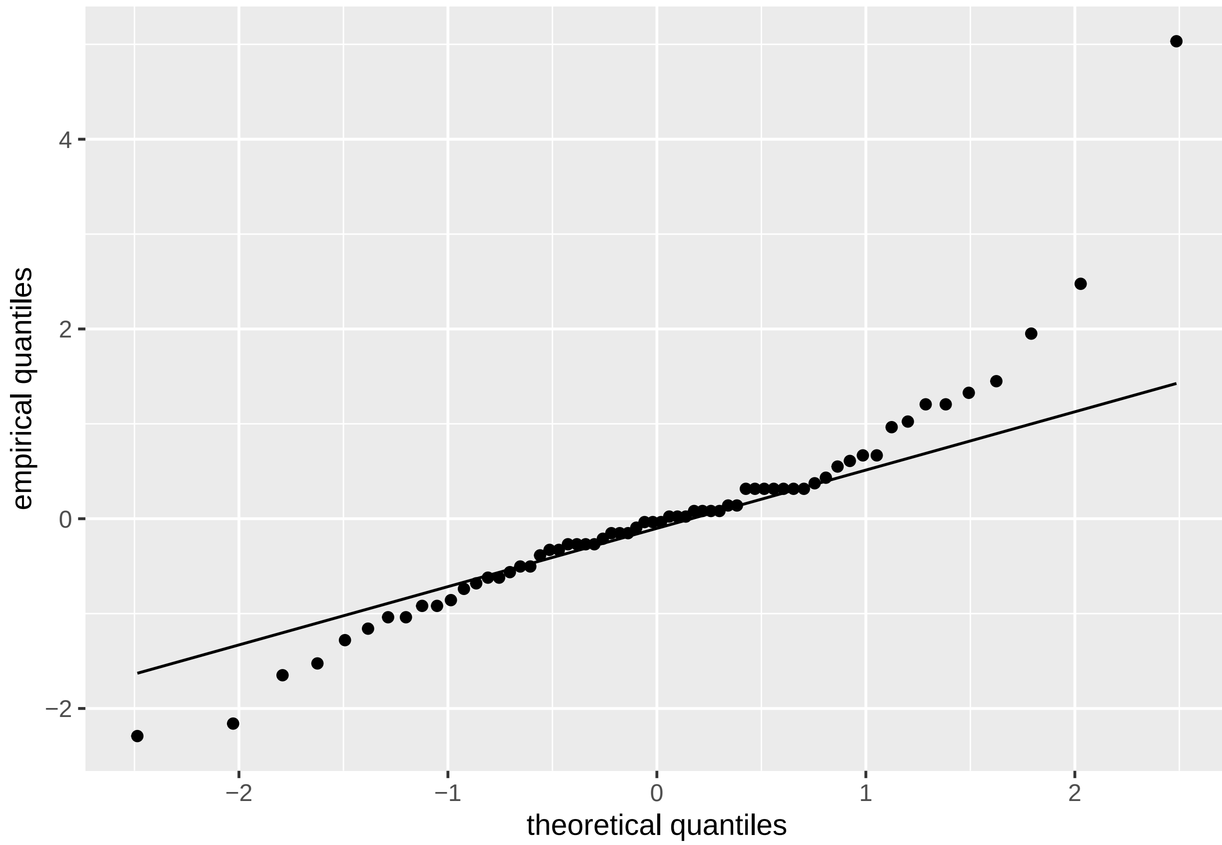
An important underlying assumption in the t-test (and linear regression model) is normality. We can assess normality by looking at the residuals, as before.

```r
# residuals / residus
# histogram
library(ggplot2)
ggplot(comp, aes(x = rstudent(mod1))) +
  geom_density() +
  geom_histogram(aes(y = after_stat(density)), bins = 20, alpha = 0.5)
```



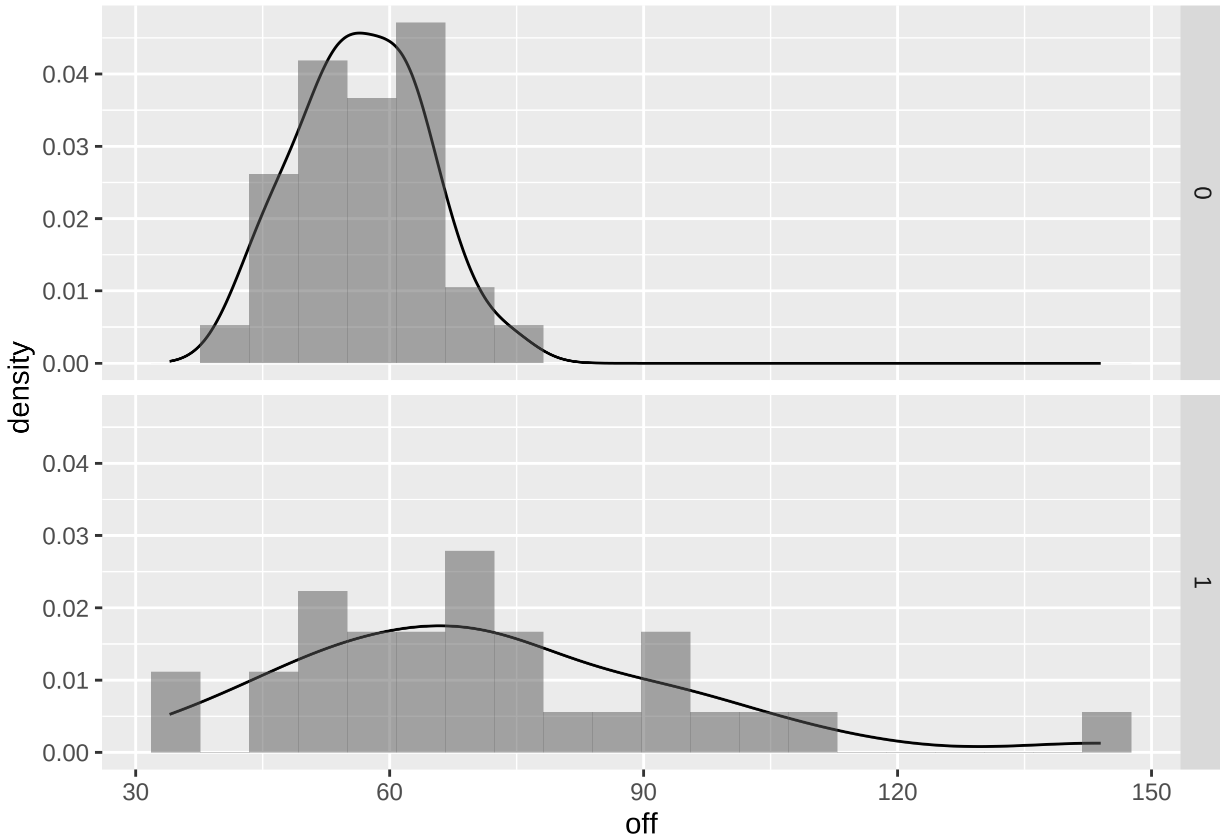

```r
# qqplot
ggplot(comp, mapping = aes(sample = rstudent(mod1))) +
  stat_qq(distribution = qt, dparams = mod1$df.residual) +
  stat_qq_line(distribution = qt, dparams = mod1$df.residual) +
```

```r
  labs(x = "theoretical quantiles",
       y = "empirical quantiles")
```

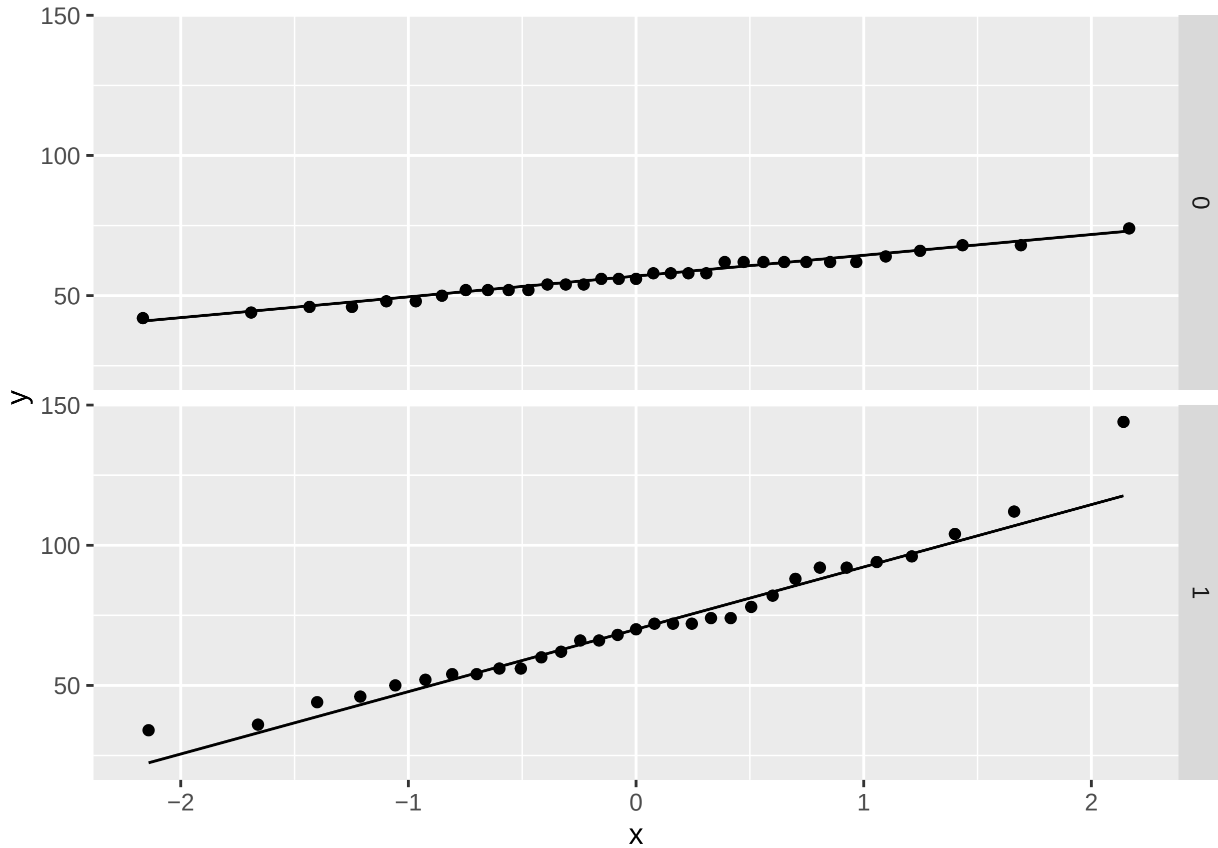


We can also assess normality by inspecting the observations themselves in each group... (Why do you think it's important to look at the observations within each group?) This is done as part of default output in SAS when carrying out a t-test

```r
# Y
# histogram
library(ggplot2)
ggplot(comp, aes(x = off)) +
  geom_density() +
  geom_histogram(aes(y = after_stat(density)), bins = 20, alpha = 0.5) +
  facet_grid(grp ~ .)
```




```r
# qqplot
ggplot(data = comp, aes(sample = off)) +
  geom_qq( ) +
  geom_qq_line( ) +
  facet_grid(grp ~ .)
```



The assumption of normality seems reasonable here.

Another important assumption for the t-test (as well as linear regression) is constant variance, that is, equality of variance between the two groups. In the context of a t-test, we can actually formally test the assumption that $\sigma_1^2 = \sigma_2^2$, where σ_j^2 denotes the variance for group $j =, 1, 2$. We can test

$$H_0 : \sigma_1^2 = \sigma_2^2 \quad \text{vs.} \quad H_1 : \sigma_1^2 \neq \sigma_2^2$$

using an F-test. To do this in R:

```
# test: egalite des variances / equality of variances
var.test(off~grp,data=comp,alternative="two.sided")
```

```
##
##  F test to compare two variances
##
## data:  off by grp
## F = 0.10206, num df = 32, denom df = 30, p-value = 5.595e-09
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.04959248 0.20828966
## sample estimates:
## ratio of variances
##          0.1020609
```

We obtain a test statistic of $F = 0.10206$ (with corresponding df1=32, df2=30) and a p-value of 5.595×10^{-09}, which is less than any reasonable α. Thus we can reject H_0 and conclude that the variances in the two groups are significantly different. As a result, the analyses we just did (t-test / linear regression) are not quite valid...

2) Welch's test

Welch's test is a modified version of the t-test which allows for $\sigma_1^2 \neq \sigma_2^2$. This can be done in R using the `t.test` function, specifying `var.equal=FALSE` in the function input:

```
# Welch test
t.test(off~grp,alternative="two.sided",var.equal=FALSE,conf.level=0.95,data=comp)
```

```
##
##  Welch Two Sample t-test
##
## data:  off by grp
## t = -3.3887, df = 35.72, p-value = 0.001725
## alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0
## 95 percent confidence interval:
##  -23.990758  -6.022927
## sample estimates:
## mean in group 0 mean in group 1
##        56.60606        71.61290
```

We obtain a test statistic of -3.3887, with $df \approx 35.72$ (Satterthwaite approximation), yielding a p-value of 0.001725. The p-value is smaller than $\alpha = 5\%$ and thus we can reject H_0 and conclude that the mean amount offered in the two groups differs significantly. In the output, we also have a 95% confidence interval for the difference: $(-23.990758, -6.022927)$ (recall: this is showing the difference between $grp = 0$ and $grp = 1$, i.e., the difference between the cash vs. credit group). Finally, estimated means for both groups is given: $\hat{\mu}_1 = 56.60606$ and $\hat{\mu}_2 = 71.61290$. While the estimated means are the same as those obtained in the classic t-test framework, the test results and confidence intervals are indeed different.

Side note

Side note: obtaining similar results using linear regression is a bit tricky... We can allow for the variance to differ by group using the `gls` function in R (GLS stands for generalized least squares). The function input is similar to `lm`, but requires specifying the variance structure to allow for the variance to vary by group. This is done by specifying weights as `varIdent(form = ~ 1 | grp)`.

```
# linear regression: generalized least squares
# allows for non-constant variance (and correlation)
library(nlme)
mod2 <- gls(off ~ grp, data = comp,
          weights=varIdent(form = ~ 1 | grp))
summary(mod2)
```

```
## Generalized least squares fit by REML
##   Model: off ~ grp
##   Data: comp
##        AIC      BIC    logLik
##   509.6002 518.1088 -250.8001
##
## Variance function:
##  Structure: Different standard deviations per stratum
##  Formula: ~1 | grp
##  Parameter estimates:
##        0        1
## 1.000000 3.130186
##
## Coefficients:
```

```
##                Value Std.Error  t-value p-value
## (Intercept) 56.60606  1.309883 43.21458  0.0000
## grp         15.00684  4.428530  3.38867  0.0012
##
##  Correlation:
##     (Intr)
## grp -0.296
##
## Standardized residuals:
##         Min          Q1         Med          Q3         Max
## -1.94108033 -0.66286314 -0.07451017  0.71683049  3.07327455
##
## Residual standard error: 7.524707
## Degrees of freedom: 64 total; 62 residual
```

The results displayed are presented in a slightly different manner than traditional least squares regression via `lm`*... We still obtain an estimated difference of* 15.007*. Note that the p-value presented here is not quite what we obtained from Welch's test - this is because the default does not use the Satterthwaite approximation for the df, but rather the default within linear regression (n-2=62 here). To obtain the results for Welch's test, we must use* `contrast` *from the* `emmeans` *library:*

```r
library(emmeans)
```

```
## Welcome to emmeans.
## Caution: You lose important information if you filter this package's results.
## See '? untidy'
```

```r
m2<-emmeans(mod2, specs="grp")
m2
```

```
##  grp emmean   SE df lower.CL upper.CL
##    0   56.6 1.31 32     53.9     59.3
##    1   71.6 4.23 30     63.0     80.3
##
## Degrees-of-freedom method: satterthwaite
## Confidence level used: 0.95
```

```r
m2.cont<-contrast(m2,method="pairwise")
m2.cont
```

```
##  contrast    estimate   SE   df t.ratio p.value
##  grp0 - grp1      -15 4.43 35.7  -3.389  0.0017
##
## Degrees-of-freedom method: satterthwaite
```

```r
confint(m2.cont)
```

```
##  contrast    estimate   SE   df lower.CL upper.CL
##  grp0 - grp1      -15 4.43 35.7      -24    -6.02
##
## Degrees-of-freedom method: satterthwaite
## Confidence level used: 0.95
```

4) ANOVA

Motivating Examples

The study objective was to compare the quality of service at 3 banks. The data are saved in the `comp3.csv` file:

```
comp3<-read.csv("Data/comp3.csv")
head(comp3)
```

```
##   ban score
## 1   1 1.599
## 2   1 2.364
## 3   1 2.106
## 4   1 1.683
## 5   1 1.959
## 6   1 1.909
```

The `score` gives the reliability score (response variable of interest) and the `ban` identifies the bank (1=United Bank of India, 2=State Bank of India, 3=Allahabad Bank). A quick overview of the data:

```
summary(comp3)
```

```
##       ban            score      
##  Min.   :1.000   Min.   :1.365  
##  1st Qu.:1.000   1st Qu.:1.819  
##  Median :2.000   Median :2.044  
##  Mean   :2.033   Mean   :2.061  
##  3rd Qu.:3.000   3rd Qu.:2.305  
##  Max.   :3.000   Max.   :3.018  
```

```
table(comp3$ban)
```

```
## 
##  1  2  3 
## 27 33 30
```

And some further descriptive statistics...

```
stats<-data.frame(as.matrix(aggregate(score~ban, data=comp3,
                                      function(x)cbind(mean(x),sd(x),min(x),max(x),length(x)))))
names(stats)<-c("ban","mean","sd","min","max","n")
stats
```

```
##   ban     mean        sd   min   max  n
## 1   1 1.897519 0.2566102 1.395 2.374 27
## 2   2 2.072515 0.3304285 1.365 2.791 33
## 3   3 2.195000 0.3292763 1.574 3.018 30
```

```
library(dplyr)
```

```
## 
## Attaching package: 'dplyr'

## The following object is masked from 'package:nlme':
## 
##     collapse

## The following objects are masked from 'package:stats':
## 
```

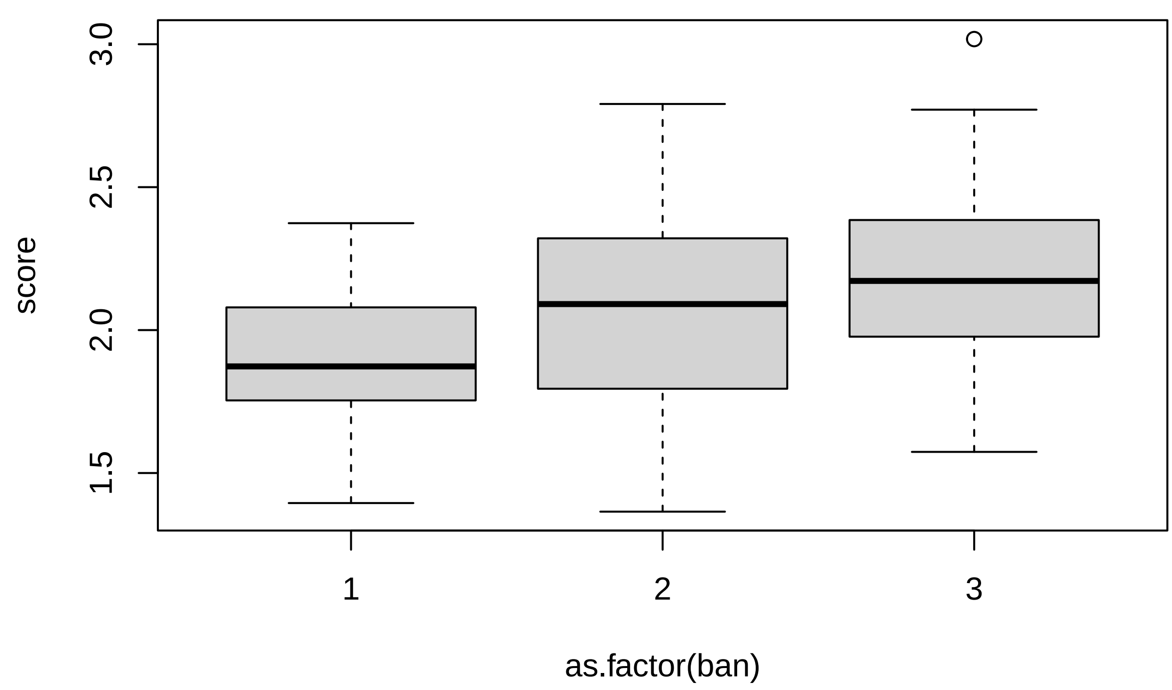
```
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
group_by(comp3, ban) %>%
  summarise(
    mean = mean(score),
    sd = sd(score),
    min = min(score),
    max = max(score),
    count = n(),
  )
```

```
## # A tibble: 3 x 6
##     ban  mean    sd   min   max count
##   <int> <dbl> <dbl> <dbl> <dbl> <int>
## 1     1  1.90 0.257  1.40  2.37    27
## 2     2  2.07 0.330  1.36  2.79    33
## 3     3  2.19 0.329  1.57  3.02    30
```

```
# visualization
boxplot(score~as.factor(ban),data=comp3)
```



So we see that the sample means are indeed different, but is there a *significant* difference?

5) ANOVA (one-way)

We're interested in testing

$$H_0 : \mu_1 = \mu_2 = \mu_3$$
$$H_1 : \text{ at least two means differ}$$

The traditional (one-way) ANOVA can be done using `oneway.test`:

```
oneway.test(score ~ ban, data = comp3, var.equal = TRUE)
```

```
##
##  One-way analysis of means
##
## data:  score and ban
## F = 6.5882, num df = 2, denom df = 87, p-value = 0.002167
```

We obtain a test statistic $F = 6.5882$ (with df1=2, df2=87) and corresponding p-value 0.002167. The p-value is smaller than $\alpha = 5\%$ and thus we can reject H_0 and conclude that the mean scores differ significantly for at least two banks.

We can also approach this using a linear regression model of the form

$$Y = \beta_0 + \beta_1 X_{bank1} + \beta_2 X_{bank2} + \epsilon$$

where Y denotes the reliability score and $X_{bankj} = 1\{bank = j\}$ is the indicator for bank j, $j = 1, 2$. The test of interest is then a global F-test for the above model, i.e.

$$H_0 : \beta_1 = \beta_2 = 0$$
$$H_1 : \text{ at least one } \beta_j \neq 0$$

We can carry out this test in the linear regression modelling framework:

```
mod3<-lm(score~as.factor(ban),data=comp3)
summary(mod3)
```

```
##
## Call:
## lm(formula = score ~ as.factor(ban), data = comp3)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.70752 -0.21713 -0.02126  0.22623  0.82300
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)      1.89752    0.05962  31.826  < 2e-16 ***
## as.factor(ban)2  0.17500    0.08039   2.177 0.032210 *
## as.factor(ban)3  0.29748    0.08218   3.620 0.000495 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3098 on 87 degrees of freedom
## Multiple R-squared:  0.1315, Adjusted R-squared:  0.1116
## F-statistic: 6.588 on 2 and 87 DF,  p-value: 0.002167
```

```r
library(car)
```

```
## Loading required package: carData

##
## Attaching package: 'car'

## The following object is masked from 'package:dplyr':
##
##     recode
```

```r
Anova(mod3)
```

```
## Anova Table (Type II tests)
##
## Response: score
##                Sum Sq Df F value   Pr(>F)
## as.factor(ban) 1.2647  2  6.5882 0.002167 **
## Residuals      8.3502 87
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

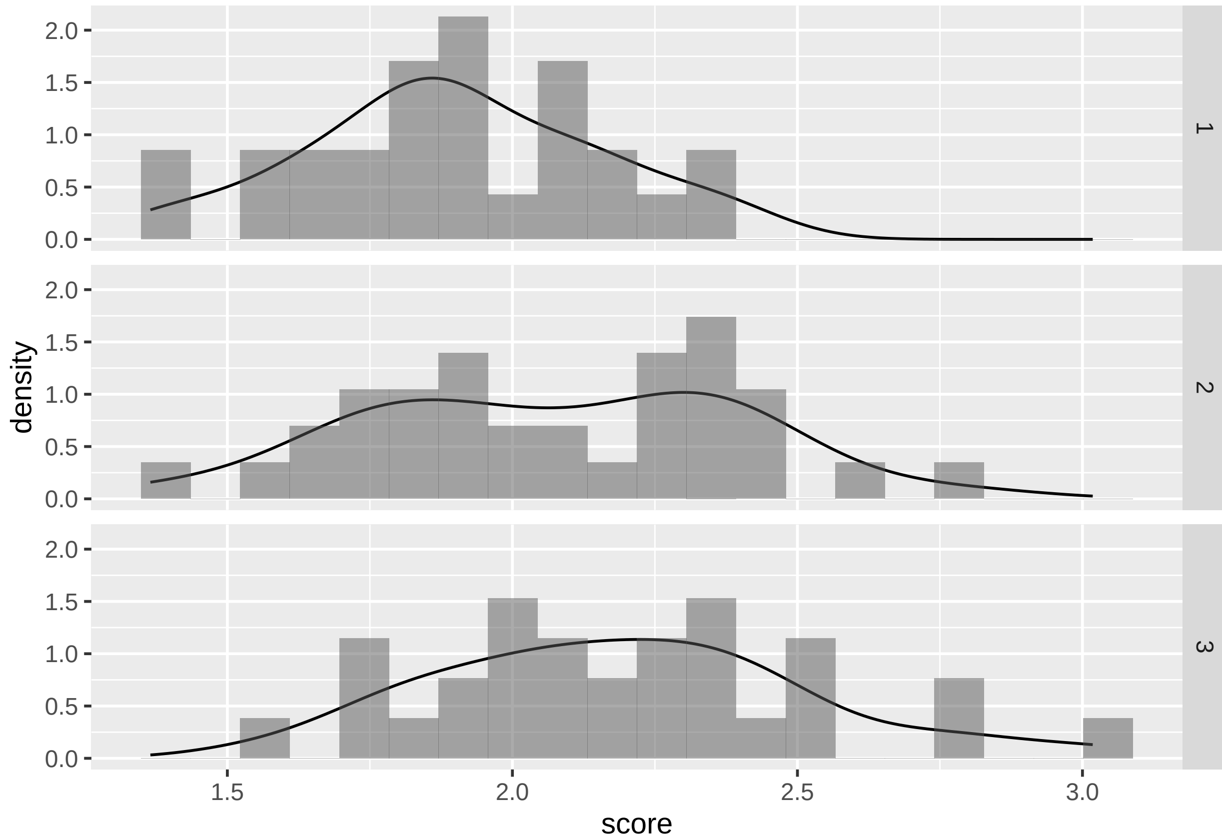
The global F-test yields a test statistic of 6.588 (with df1=2 and df2=87) and corresponding p-value 0.002167; exactly as before.

What do the different p-values in the summary results for the model represent?

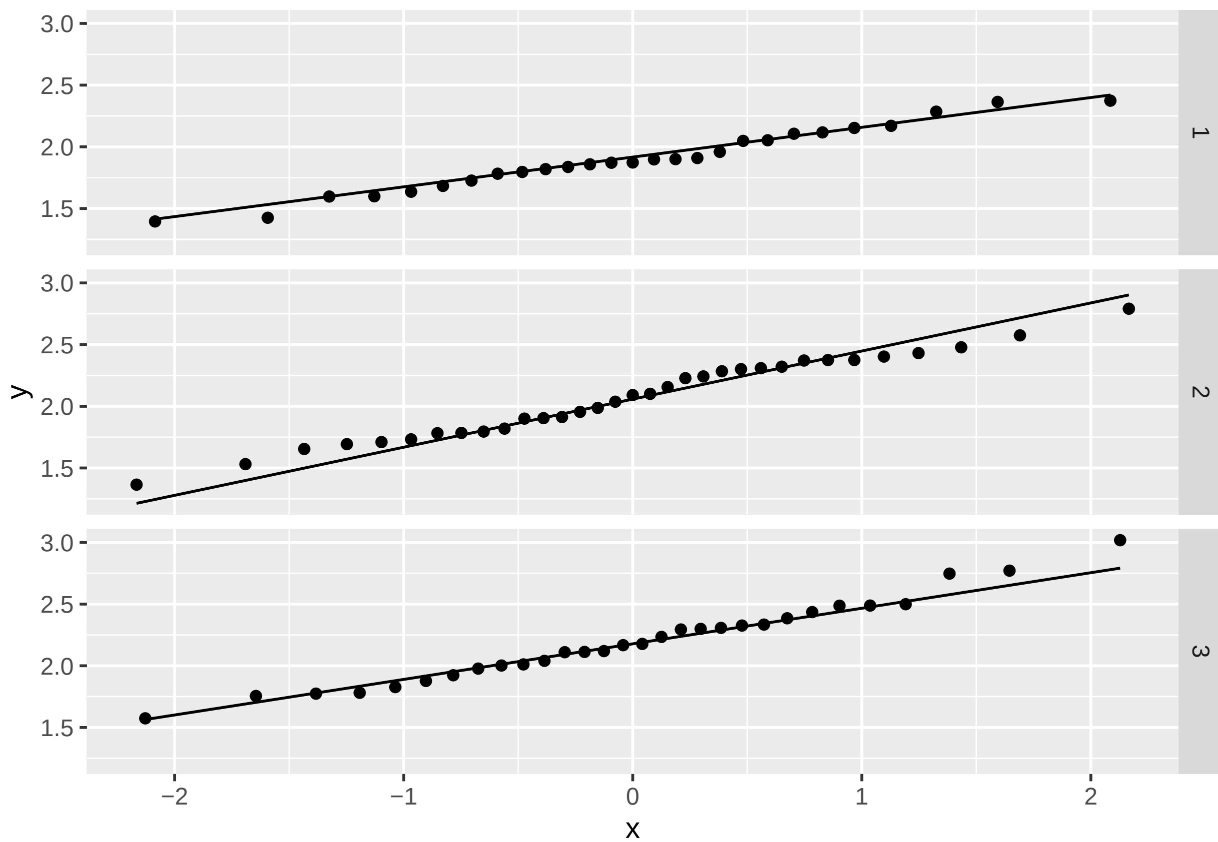
Assumptions

As was the case for the classic t-test, here we rely on a few assumptions (which parallel those in the linear regression model). We assume normality, which can be checked using the model residuals or directly on the observations themselves:

```r
# histograms
library(ggplot2)
ggplot(comp3, aes(x = score)) +
  geom_density() +
  geom_histogram(aes(y = after_stat(density)), bins = 20, alpha = 0.5) +
  facet_grid(ban ~ .)
```




```r
# qqplot
ggplot(data = comp3, aes(sample = score)) +
  geom_qq( ) +
  geom_qq_line( ) +
  facet_grid(ban ~ .)
```



We also rely on the assumption of constant variance. Again, we can visually assess this through the linear regression model residual plots. However, here, we can formally test for equality of variances across the groups. That is, we can test

$$H_0 : \sigma_1^2 = \sigma_2^2 = \sigma_3^2$$
$$H_1 : \text{ at least two variances differ}$$

There are various statistical tests that allows to test these hypotheses, two are presented here:

```
# test: equality of variance / egalites des variances
bartlett.test(score~as.factor(ban),data=comp3)
```

```
##
##  Bartlett test of homogeneity of variances
##
## data:  score by as.factor(ban)
## Bartlett's K-squared = 2.1036, df = 2, p-value = 0.3493
```

```
library(car)
leveneTest(score~as.factor(ban),data=comp3)
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##       Df F value Pr(>F)
## group  2  1.6234 0.2031
##       87
```

```
leveneTest(score~as.factor(ban),data=comp3,center=mean)
```

```
## Levene's Test for Homogeneity of Variance (center = mean)
##       Df F value Pr(>F)
## group  2  1.6644 0.1953
##       87
```

While each yield different results, all lead to the same conclusion: we fail to reject H_0. As such, the analysis we previously carried out are valid.

Comments: Levene's test is less sensitive to the assumption of normality in comparison to Bartlett's test. There are different versions of Levene's test depending on what is used to center the observations: the mean, the median, the trimmed mean.

If the variances were found to be significantly different amongst the three groups, we could use Welch's test:

```
oneway.test(score ~ ban, data = comp3, var.equal = FALSE)
```

```
##
##  One-way analysis of means (not assuming equal variances)
##
## data:  score and ban
## F = 7.5421, num df = 2.000, denom df = 57.802, p-value = 0.001229
```

Note: we could also do this within a linear regression model, as before, allowing for heteroscedasticity by specifying a variance which differs by group.

Pairwise comparisons

Recall that in the one-way ANOVA, we are testing

$$
\begin{aligned}
&H_0 : \mu_1 = \mu_2 = \cdots = \mu_K \\
&H_1 : \text{ at least two means differ}
\end{aligned}
$$

If we reject H_0 with the F-test (or Welch's test), we can only conclude that at least two means differ - we do not know which particular groups have different means.

In a linear regression model (which includes only a categorical variable with K levels) the global F-test has the form

$$
\begin{aligned}
&H_0 : \beta_1 = \beta_2 = \cdots = \beta_{K-1} = 0 \\
&H_1 : \text{ at least one } \beta_j \neq 0
\end{aligned}
$$

That is, the global F-test allows to test whether at least two means differ. The individual t-tests

$$
\begin{aligned}
&H_0 : \beta_j = 0 \\
&H_1 : \beta_j \neq 0
\end{aligned}
$$

allow to assess whether there is a significant difference between the given level (e.g., j) and the reference level.

Often, we're interested in all pairwise comparisons of the form

$$H_0 : \mu_i = \mu_j \text{ vs. } H_1 : \mu_i \neq \mu_j$$

for all possible comginations (i, j). This can be done in the classic set-up (assuming constant variance):

```
# Pairwise comparisons / comparaisons paires
# linear regression / regression lineaire
library(nlme)
library(emmeans)
# constant variance / variance constante
contrast(emmeans(mod3, ~ban),method="pairwise",adjust="none")
```

```
##   contrast    estimate     SE df t.ratio p.value
##   ban1 - ban2   -0.175 0.0804 87  -2.177  0.0322
##   ban1 - ban3   -0.297 0.0822 87  -3.620  0.0005
##   ban2 - ban3   -0.122 0.0782 87  -1.567  0.1207
```

We see that there is a significant difference in the mean reliability score for State Bank of India (ban=2) and United Bank of India (ban=1) (p-value of 0.0322). There is also a significant difference in the mean score for the Allahabad Bank (ban=3) and United Bank of India (ban=1) (p-value 0.0005). However, there is not a significant difference in the mean scores for State Bank of India (ban=2) and Allahabad Bank (bank=3) (p-value=0.1207).

Note: We can do the same in the context of Welch's test where the variance is allowed to differ by group:

```
# different variance by bank
# differente variance par banque
mod4 <- gls(score ~ as.factor(ban), data = comp3,
            weights=varIdent(form = ~ 1 | as.factor(ban)))
summary(mod4)
```

```
## Generalized least squares fit by REML
##   Model: score ~ as.factor(ban)
##   Data: comp3
##        AIC      BIC    logLik
##   63.05748 77.85293 -25.52874
##
## Variance function:
##  Structure: Different standard deviations per stratum
##  Formula: ~1 | as.factor(ban)
##  Parameter estimates:
##        1        2        3
## 1.000000 1.287667 1.283177
##
## Coefficients:
##                     Value  Std.Error  t-value p-value
## (Intercept)     1.8975185 0.04938465 38.42324  0.0000
## as.factor(ban)2 0.1749966 0.07581174  2.30831  0.0234
## as.factor(ban)3 0.2974815 0.07780065  3.82364  0.0002
##
##  Correlation:
##                 (Intr) as.()2
## as.factor(ban)2 -0.651
## as.factor(ban)3 -0.635  0.413
##
## Standardized residuals:
##         Min          Q1         Med          Q3         Max
## -2.14120480 -0.74252383 -0.06985015  0.72557616  2.49942042
##
## Residual standard error: 0.2566102
## Degrees of freedom: 90 total; 87 residual
```

```
contrast(emmeans(mod4, ~ban),method="pairwise",adjust="none")
```

```
##   contrast    estimate     SE   df t.ratio p.value
##   ban1 - ban2   -0.175 0.0758 57.9  -2.308  0.0246
##   ban1 - ban3   -0.297 0.0778 53.9  -3.824  0.0003
##   ban2 - ban3   -0.122 0.0832 60.5  -1.472  0.1462
```

```
##
## Degrees-of-freedom method: satterthwaite
```

While the numerical results are different here (i.e., different p-values), we obtain the same conclusions as before.

6) ANOVA (two-way)

Motivating example

The study assesses the effect of the type of delay (`del`) and the stage of advancement (`sta`) on the evaluation (`eval`) of the service and the perceived waiting time (`t`):

- `t`: time measured in minutes
- `eval`: standardized score of evaluation of service
- `sta`: 1=close to end, 2=far from end
- `del`: 1=procedural, 2=correctional, 3=unknown

Here is a quick overview of the data:

```
#  Exemple: UX
#  Example: UX
comp4<-read.csv("Data/comp4.csv")
nrow(comp4)
```

```
## [1] 109
```

```
head(comp4)
```

```
##   del sta  t  eval
## 1   1   1 11 -1.02
## 2   1   1 11 -0.69
## 3   1   1 15 -0.18
## 4   1   1 20 -0.21
## 5   1   1 16  0.10
## 6   1   1 12 -0.60
```

```
summary(comp4)
```

```
##       del             sta               t              eval
##  Min.   :1.000   Min.   :1.000   Min.   : 3.0   Min.   :-1.62000
##  1st Qu.:1.000   1st Qu.:1.000   1st Qu.: 9.0   1st Qu.:-0.41000
##  Median :2.000   Median :1.000   Median :11.0   Median :-0.03000
##  Mean   :1.982   Mean   :1.477   Mean   :11.2   Mean   :-0.04229
##  3rd Qu.:3.000   3rd Qu.:2.000   3rd Qu.:14.0   3rd Qu.: 0.43000
##  Max.   :3.000   Max.   :2.000   Max.   :20.0   Max.   : 1.28000
```

We're ultimately interested in evaluating the effect of the factors delay and stage on the evaluation of service.

Main effects model

We can start with the main effects model:

```
# modeles a effets principaux
# main effects model
mod5<-lm(eval~as.factor(sta)+as.factor(del),data=comp4)
summary(mod5)
```

```
##
```

```
## Call:
## lm(formula = eval ~ as.factor(sta) + as.factor(del), data = comp4)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.40788 -0.42049  0.02264  0.44999  1.33165
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)     -0.03951    0.11779  -0.335    0.738
## as.factor(sta)2 -0.10048    0.12124  -0.829    0.409
## as.factor(del)2 -0.07214    0.14693  -0.491    0.624
## as.factor(del)3  0.21687    0.14887   1.457    0.148
##
## Residual standard error: 0.6313 on 105 degrees of freedom
## Multiple R-squared:  0.04432,    Adjusted R-squared:  0.01702
## F-statistic: 1.623 on 3 and 105 DF,  p-value: 0.1884
```

How can we interpret these parameters? What conclusions can we make based on the results? What can we conclude based on the global F-test?

```r
#
library(car)
Anova(mod5)
```

```
## Anova Table (Type II tests)
##
## Response: eval
##                Sum Sq  Df F value Pr(>F)
## as.factor(sta)  0.274   1  0.6869 0.4091
## as.factor(del)  1.614   2  2.0249 0.1371
## Residuals      41.852 105
```

Here we see that both `sta` and `del` are not significant in the model. We should check, however, if there is an interaction between the two variables `sta` and `del` on the response variable `eval`.

Interaction model

We can fit the interaction model:

```r
# modele avec interaction
# interaction model
mod6<-lm(eval~as.factor(sta)*as.factor(del),data=comp4)
summary(mod6)
```

```
##
## Call:
## lm(formula = eval ~ as.factor(sta) * as.factor(del), data = comp4)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.25263 -0.29063  0.04937  0.29737  1.01937
##
## Coefficients:
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                     -0.4350     0.1017  -4.277 4.25e-05 ***
## as.factor(sta)2                  0.7603     0.1501   5.067 1.79e-06 ***
```

```
## as.factor(del)2                    0.9261     0.1478   6.266 8.71e-09 ***
## as.factor(del)3                    0.4576     0.1457   3.140   0.0022 **
## as.factor(sta)2:as.factor(del)2  -2.0346     0.2119  -9.601 5.71e-16 ***
## as.factor(sta)2:as.factor(del)3  -0.5223     0.2153  -2.426   0.0170 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4549 on 103 degrees of freedom
## Multiple R-squared:  0.5133, Adjusted R-squared:  0.4897
## F-statistic: 21.73 on 5 and 103 DF,  p-value: 8.514e-15
```

How can we interpret these parameters? What conclusions can we make based on the results? What can we conclude based on the global F-test?

```r
#
library(car)
Anova(mod6)
```

```
## Anova Table (Type II tests)
##
## Response: eval
##                                  Sum Sq  Df F value    Pr(>F)
## as.factor(sta)                   0.2738   1  1.3231   0.25271
## as.factor(del)                   1.6142   2  3.9005   0.02329 *
## as.factor(sta):as.factor(del) 20.5388   2 49.6284 8.074e-16 ***
## Residuals                       21.3134 103
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We must be careful here to make sure we understand what the output is really giving us.

The interaction term:

```r
Anova(mod6)
```

```
## Anova Table (Type II tests)
##
## Response: eval
##                                  Sum Sq  Df F value    Pr(>F)
## as.factor(sta)                   0.2738   1  1.3231   0.25271
## as.factor(del)                   1.6142   2  3.9005   0.02329 *
## as.factor(sta):as.factor(del) 20.5388   2 49.6284 8.074e-16 ***
## Residuals                       21.3134 103
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
anova(mod6,mod5)
```

```
## Analysis of Variance Table
##
## Model 1: eval ~ as.factor(sta) * as.factor(del)
## Model 2: eval ~ as.factor(sta) + as.factor(del)
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1    103 21.313
## 2    105 41.852 -2   -20.539 49.628 8.074e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The global F-test for the interaction term is testing $H_0 : \beta_4 = \beta_5 = 0$ vs. H_1 : at least one β_4 or $\beta_5 \neq 0$. This leads to a test statistic value of 49.628 and a p-value of 8.074×10^{-16}. Thus, we can reject H_0 and conclude that the interaction is significant - that is, the effect of the stage depends on the type of delay, and vice versa (the effect of the type of delay depends on the stage).

Note: since the interaction is significant, it does not matter if the main effects are not significant! Recall that here the "main effects" terms are in fact the effects for the reference levels! We must be careful with interpreting the results for the tests for `sta` and `del`, in particular, as they are involved in interactions.

What are these tests really showing? Let's take a closer look with what we get by comparing models with `anova`:

```
Anova(mod6,error=mod5)
```

```
## Anova Table (Type II tests)
##
## Response: eval
##                                Sum Sq  Df F value    Pr(>F)
## as.factor(sta)                  0.274   1  0.6869    0.4091
## as.factor(del)                  1.614   2  2.0249    0.1371
## as.factor(sta):as.factor(del) 20.539   2 25.7642 7.876e-10 ***
## Residuals                     41.852 105
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Anova(mod6,error=mod6)
```

```
## Anova Table (Type II tests)
##
## Response: eval
##                                 Sum Sq  Df F value    Pr(>F)
## as.factor(sta)                  0.2738   1  1.3231   0.25271
## as.factor(del)                  1.6142   2  3.9005   0.02329 *
## as.factor(sta):as.factor(del) 20.5388   2 49.6284 8.074e-16 ***
## Residuals                     21.3134 103
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Anova(mod6)
```

```
## Anova Table (Type II tests)
##
## Response: eval
##                                 Sum Sq  Df F value    Pr(>F)
## as.factor(sta)                  0.2738   1  1.3231   0.25271
## as.factor(del)                  1.6142   2  3.9005   0.02329 *
## as.factor(sta):as.factor(del) 20.5388   2 49.6284 8.074e-16 ***
## Residuals                     21.3134 103
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
anova(mod6,lm(eval~as.factor(del),data=comp4))
```

```
## Analysis of Variance Table
##
## Model 1: eval ~ as.factor(sta) * as.factor(del)
## Model 2: eval ~ as.factor(del)
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
```

```
## 1    103 21.313
## 2    106 42.126 -3   -20.813 33.527 3.341e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
anova(mod5,lm(eval~as.factor(del),data=comp4))
```

```
## Analysis of Variance Table
##
## Model 1: eval ~ as.factor(sta) + as.factor(del)
## Model 2: eval ~ as.factor(del)
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1    105 41.852
## 2    106 42.126 -1  -0.27377 0.6869 0.4091
```

```r
anova(mod5,lm(eval~as.factor(del),data=comp4),error=mod6)
```

```
## Analysis of Variance Table
##
## Model 1: eval ~ as.factor(sta) + as.factor(del)
## Model 2: eval ~ as.factor(del)
## Model 3: eval ~ as.factor(sta) * as.factor(del)
##   Res.Df    RSS Df Sum of Sq       F    Pr(>F)
## 1    105 41.852
## 2    106 42.126 -1   -0.2738  1.3231    0.2527
## 3    103 21.313  3   20.8126 33.5266 3.341e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
anova(mod6,lm(eval~as.factor(sta),data=comp4))
```

```
## Analysis of Variance Table
##
## Model 1: eval ~ as.factor(sta) * as.factor(del)
## Model 2: eval ~ as.factor(sta)
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1    103 21.313
## 2    107 43.466 -4   -22.153 26.764 3.133e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
anova(mod5,lm(eval~as.factor(sta),data=comp4))
```

```
## Analysis of Variance Table
##
## Model 1: eval ~ as.factor(sta) + as.factor(del)
## Model 2: eval ~ as.factor(sta)
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1    105 41.852
## 2    107 43.466 -2   -1.6142 2.0249 0.1371
```

```r
anova(mod5,lm(eval~as.factor(sta),data=comp4),error=mod6)
```

```
## Analysis of Variance Table
##
## Model 1: eval ~ as.factor(sta) + as.factor(del)
## Model 2: eval ~ as.factor(sta)
## Model 3: eval ~ as.factor(sta) * as.factor(del)
```

```
##   Res.Df    RSS Df Sum of Sq       F    Pr(>F)
## 1    105 41.852
## 2    107 43.466 -2   -1.6142  3.9005   0.02329 *
## 3    103 21.313  4   22.1530 26.7644 3.133e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

From the `R` documentation on `Anova`:

Type-II tests are calculated according to the principle of marginality, testing each term after all others, except ignoring the term's higher-order relatives; so-called type-III tests violate marginality, testing each term in the model after all of the others. This definition of Type-II tests corresponds to the tests produced by SAS for analysis-of-variance models, where all of the predictors are factors, but not more generally (i.e., when there are quantitative predictors). Be very careful in formulating the model for type-III tests, or the hypotheses tested will not make sense.

To illustrate what the type III tests are showing exactly, we will manually create indicator variables to represent the various levels of the categorical variables as well as the interaction:

```r
# manually / manuellement
data<-comp4
data$sta2<-as.numeric(comp4$sta==2)
data$del2<-as.numeric(comp4$del==2)
data$del3<-as.numeric(comp4$del==3)
data$int22<-as.numeric(comp4$sta==2 & comp4$del==2)
data$int23<-as.numeric(comp4$sta==2 & comp4$del==3)
head(data)
```

```
##   del sta  t  eval sta2 del2 del3 int22 int23
## 1   1   1 11 -1.02    0    0    0     0     0
## 2   1   1 11 -0.69    0    0    0     0     0
## 3   1   1 15 -0.18    0    0    0     0     0
## 4   1   1 20 -0.21    0    0    0     0     0
## 5   1   1 16  0.10    0    0    0     0     0
## 6   1   1 12 -0.60    0    0    0     0     0
```

We can then refit the same model previously considered by working with these indicator variables:

```r
#
mod7<-lm(eval~sta2+del2+del3+int22+int23,data=data)
summary(mod7)
```

```
##
## Call:
## lm(formula = eval ~ sta2 + del2 + del3 + int22 + int23, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.25263 -0.29063  0.04937  0.29737  1.01937
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -0.4350     0.1017  -4.277 4.25e-05 ***
## sta2          0.7603     0.1501   5.067 1.79e-06 ***
## del2          0.9261     0.1478   6.266 8.71e-09 ***
## del3          0.4576     0.1457   3.140   0.0022 **
## int22        -2.0346     0.2119  -9.601 5.71e-16 ***
## int23        -0.5223     0.2153  -2.426   0.0170 *
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4549 on 103 degrees of freedom
## Multiple R-squared:  0.5133, Adjusted R-squared:  0.4897
## F-statistic: 21.73 on 5 and 103 DF,  p-value: 8.514e-15
```

```
summary(mod6)
```

```
##
## Call:
## lm(formula = eval ~ as.factor(sta) * as.factor(del), data = comp4)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.25263 -0.29063  0.04937  0.29737  1.01937
##
## Coefficients:
##                                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)                        -0.4350     0.1017  -4.277 4.25e-05 ***
## as.factor(sta)2                     0.7603     0.1501   5.067 1.79e-06 ***
## as.factor(del)2                     0.9261     0.1478   6.266 8.71e-09 ***
## as.factor(del)3                     0.4576     0.1457   3.140   0.0022 **
## as.factor(sta)2:as.factor(del)2  -2.0346     0.2119  -9.601 5.71e-16 ***
## as.factor(sta)2:as.factor(del)3  -0.5223     0.2153  -2.426   0.0170 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4549 on 103 degrees of freedom
## Multiple R-squared:  0.5133, Adjusted R-squared:  0.4897
## F-statistic: 21.73 on 5 and 103 DF,  p-value: 8.514e-15
```

Indeed, this model (`mod7`) is identical to the one we previously fit (`mod6`).

With this model, it's easier to understand what is being showed with `Anova` type III:

```
Anova(mod6,type=3)
```

```
## Anova Table (Type III tests)
##
## Response: eval
##                                Sum Sq  Df F value    Pr(>F)
## (Intercept)                    3.7845   1  18.289 4.254e-05 ***
## as.factor(sta)                 5.3118   1  25.670 1.788e-06 ***
## as.factor(del)                 8.1300   2  19.645 5.926e-08 ***
## as.factor(sta):as.factor(del) 20.5388   2  49.628 8.074e-16 ***
## Residuals                     21.3134 103
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
anova(mod7,lm(eval~del2+del3+int22+int23,data=data))
```

```
## Analysis of Variance Table
##
## Model 1: eval ~ sta2 + del2 + del3 + int22 + int23
## Model 2: eval ~ del2 + del3 + int22 + int23
##   Res.Df    RSS Df Sum of Sq     F    Pr(>F)
```

```
## 1    103 21.313
## 2    104 26.625 -1   -5.3118 25.67 1.788e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
anova(mod7,lm(eval~sta2+int22+int23,data=data))
```

```
## Analysis of Variance Table
##
## Model 1: eval ~ sta2 + del2 + del3 + int22 + int23
## Model 2: eval ~ sta2 + int22 + int23
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1    103 21.313
## 2    105 29.443 -2     -8.13 19.645 5.926e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

From this we see that `Anova` type III is yielding tests which just remove the "main effects" terms for `sta` and for `del`, respectively. Here (and in general) these tests are not quite of interest. *Why?*

Pairwise comparisons

We can also consider the simple effects of the factors within the interaction model. First,

```r
# comparisons
library(emmeans)
emmeans(mod6, ~del*sta)
```

```
##  del sta  emmean    SE  df lower.CL upper.CL
##    1   1 -0.4350 0.102 103  -0.6367   -0.233
##    2   1  0.4911 0.107 103   0.2785    0.704
##    3   1  0.0226 0.104 103  -0.1843    0.230
##    1   2  0.3253 0.110 103   0.1065    0.544
##    2   2 -0.7832 0.104 103  -0.9901   -0.576
##    3   2  0.2606 0.114 103   0.0351    0.486
##
## Confidence level used: 0.95
```

We obtain the estimated mean, SE and C.I. for every group resulting from crossing `sta` and `del`. (Note that these estimated means are the mean of the observations in the corresponding group.)

We can also consider contrasts between the various groups; in total there are $\binom{6}{2} = 15$ distinct pairwise comparisons:

```r
contrast(emmeans(mod6, ~del*sta),method="pairwise",adjust="none")
```

```
##  contrast              estimate    SE  df t.ratio p.value
##  del1 sta1 - del2 sta1  -0.9261 0.148 103  -6.266  <.0001
##  del1 sta1 - del3 sta1  -0.4576 0.146 103  -3.140  0.0022
##  del1 sta1 - del1 sta2  -0.7603 0.150 103  -5.067  <.0001
##  del1 sta1 - del2 sta2   0.3482 0.146 103   2.389  0.0187
##  del1 sta1 - del3 sta2  -0.6956 0.153 103  -4.559  <.0001
##  del2 sta1 - del3 sta1   0.4685 0.150 103   3.131  0.0023
##  del2 sta1 - del1 sta2   0.1658 0.154 103   1.078  0.2836
##  del2 sta1 - del2 sta2   1.2743 0.150 103   8.517  <.0001
##  del2 sta1 - del3 sta2   0.2305 0.156 103   1.475  0.1434
##  del3 sta1 - del1 sta2  -0.3027 0.152 103  -1.993  0.0489
##  del3 sta1 - del2 sta2   0.8058 0.148 103   5.460  <.0001
```

```
##  del3 sta1 - del3 sta2  -0.2380 0.154 103  -1.542  0.1262
##  del1 sta2 - del2 sta2   1.1085 0.152 103   7.299  <.0001
##  del1 sta2 - del3 sta2   0.0647 0.158 103   0.408  0.6840
##  del2 sta2 - del3 sta2  -1.0438 0.154 103  -6.762  <.0001
```

For the delay factor `del`, there are 6 possible comparisons:

- When `sta=1` (close to end):
  - `del=1` (procedural) vs `del=2` (correctional): p-value $< .0001$
  - `del=1` (procedural) vs `del=3` (unknown): p-value 0.0022
  - `del=2` (correctional) vs `del=3` (unknown): p-value 0.0023
- When `sta=2` (far from end):
  - `del=1` (procedural) vs `del=2` (correctional): p-value $< .0001$
  - `del=1` (procedural) vs `del=3` (unknown): p-value 0.6840
  - `del=2` (correctional) vs `del=3` (unknown): p-value $< .0001$

For the stage factor `sta`, there are 3 possible comparisons:

- When `del=1` (procedural)
  - `sta=1` (close to end) vs. `sta=2` (far from end): p-value $< .0001$
- When `del=2` (correctional)
  - `sta=1` (close to end) vs. `sta=2` (far from end): p-value $< .0001$
- When `del=3` (unknown)
  - `sta=1` (close to end) vs. `sta=2` (far from end): p-value 0.1262

What conclusions can we make here?

- e.g., when the type of delay is procedural, the mean evaluation score is significantly different when the stage is far from the end vs. when the stage is close to the end. (p-vaue < 0.0001).

Non-constant variance

An underlying assumption in the 2-way ANOVA (and linear regression model) is homoscedasticity, that is, constant variance. As we saw before, we can formally test this assumption.

Here, we first need to create the groups representing the crossing of the two variables `sta` and `del`:

```
# testing equality of variance
data2<-comp4
data2$grp<-1*as.numeric(comp4$del==1 & comp4$sta==1) +
  2*as.numeric(comp4$del==2 & comp4$sta==1)+
  3*as.numeric(comp4$del==3 & comp4$sta==1)+
  4*as.numeric(comp4$del==1 & comp4$sta==2)+
  5*as.numeric(comp4$del==2 & comp4$sta==2)+
  6*as.numeric(comp4$del==3 & comp4$sta==2)
head(data2)
```

```
##   del sta  t  eval grp
## 1   1   1 11 -1.02   1
## 2   1   1 11 -0.69   1
## 3   1   1 15 -0.18   1
## 4   1   1 20 -0.21   1
## 5   1   1 16  0.10   1
## 6   1   1 12 -0.60   1
```

Now we can test the assumption of constant variance across the 6 groups, that is, we can test

$$H_0 : \sigma_1^2 = \sigma_2^2 = \cdots = \sigma_6^2$$
$$H_1 : \text{ at least two } \sigma_j^2 \text{ differ}$$

As we saw before, there are different tests that can be used to evaluate these hypotheses:

```r
# test: equality of variance / egalites des variances
bartlett.test(eval~as.factor(grp),data=data2)
```

```
##
##  Bartlett test of homogeneity of variances
##
## data:  eval by as.factor(grp)
## Bartlett's K-squared = 3.8364, df = 5, p-value = 0.5732
```

```r
library(car)
leveneTest(eval~as.factor(grp),data=data2)
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##        Df F value Pr(>F)
## group   5  1.1588 0.3346
##       103
```

```r
leveneTest(eval~as.factor(grp),data=data2,center=mean)
```

```
## Levene's Test for Homogeneity of Variance (center = mean)
##        Df F value Pr(>F)
## group   5  1.2736  0.281
##       103
```

In all cases, the p-values are large and thus we fail to reject H_0, and thus the assumption of constant variance seems reasonable here.

Note: as we saw before, if it was found that there was heteroscedasticity, we could adapt the approach to allow for the variance to differ across the groups. This can be done in the GLS framework, exactly as we saw before:

```r
library(nlme)
mod8 <- gls(eval ~ as.factor(del)*as.factor(sta), data = data2,
            weights=varIdent(form = ~ 1 | grp))
summary(mod8)
```

```
## Generalized least squares fit by REML
##   Model: eval ~ as.factor(del) * as.factor(sta)
##   Data: data2
##        AIC      BIC   logLik
##   167.4926 199.1094 -71.7463
##
## Variance function:
##  Structure: Different standard deviations per stratum
##  Formula: ~1 | grp
##  Parameter estimates:
##         1         4         2         5         3         6
## 1.0000000 0.7485762 0.7369905 0.9400609 1.0703174 0.8662404
##
## Coefficients:
##                                    Value Std.Error    t-value p-value
## (Intercept)                   -0.4350000 0.1119816  -3.884566  0.0002
## as.factor(del)2                0.9261111 0.1418019   6.531020  0.0000
## as.factor(del)3                0.4576316 0.1663172   2.751560  0.0070
## as.factor(sta)2                0.7602941 0.1442458   5.270822  0.0000
## as.factor(del)2:as.factor(sta)2 -2.0345631 0.2000992 -10.167773  0.0000
## as.factor(del)3:as.factor(sta)2 -0.5223007 0.2183812  -2.391692  0.0186
```

```
##
##  Correlation:
##                                  (Intr) as.fctr(d)2 as.()3 as.fctr(s)2 a.()2:
## as.factor(del)2                  -0.790
## as.factor(del)3                  -0.673  0.532
## as.factor(sta)2                  -0.776  0.613       0.523
## as.factor(del)2:as.factor(sta)2   0.560 -0.709      -0.377 -0.721
## as.factor(del)3:as.factor(sta)2   0.513 -0.405      -0.762 -0.661       0.476
##
## Standardized residuals:
##        Min         Q1        Med         Q3        Max
## -2.3369475 -0.6699351  0.1070283  0.6439484  2.3498155
##
## Residual standard error: 0.5007971
## Degrees of freedom: 109 total; 103 residual
```