This morning: Choice of model, and presentation of results

- Example: Cardiac output under increasing workload for 12 CHF patients and 12 healthy volunteers (exercise 7)
- Choice of repeated measurement model
- Model validation
- Presenting the results from the statistical analysis

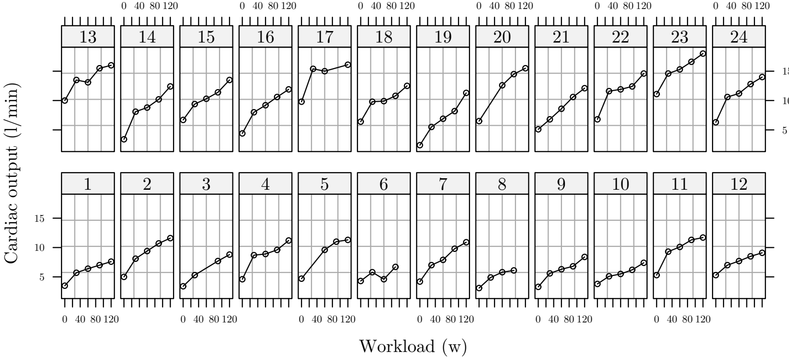
Exercises this morning

Exercises: Summarize and present the analysis of the crp- and/or smoke data.

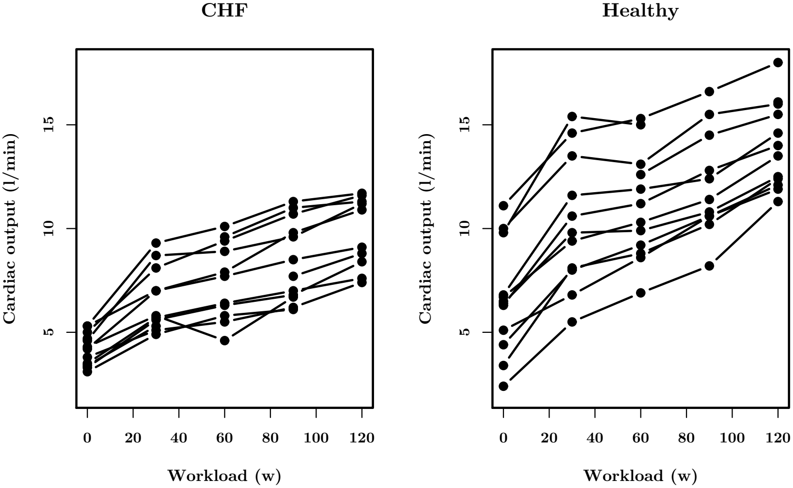
- **Methods**
  - Which statistical analysis was used?
  - Which assumptions were needed?
  - Which hypotheses were tested?
- **Results**
  - Are the assumptions fulfilled?
  - Choose one or two figures to illustrate the major points
  - What are the conclusions of the statistical tests?
  - What are the main results (relevant estimates with CI's and maybe some sd's)?

Example: Cardiac output for CHF patients and healthy

Data: In exercise 7 we considered data on cardiac output (l/min) measured for 12 CHF patients (1 – 12) and 12 healthy volunteers (13 – 24). The measurements are obtained using the inert gas re-breathing technique under increasing workload (0, 30, 60, 90, and 120 watt).

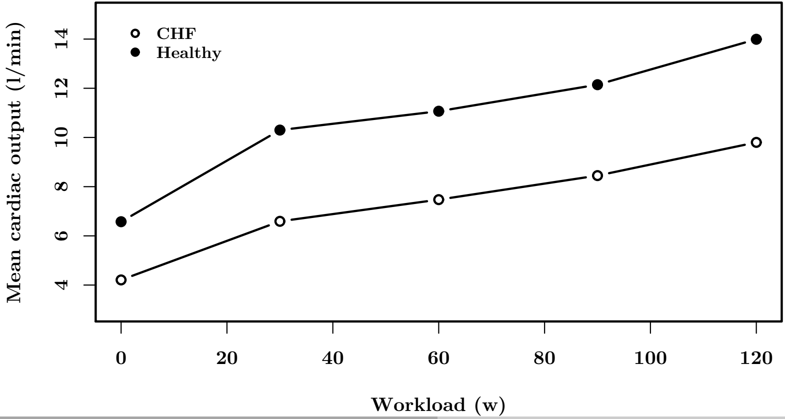


The variation between subjects in the two groups



Cardiac output for CHF patients and healthy: Mean curves

We want to compare the two groups. Is there any difference between the cardiac output for CHF patients compared to healthy individuals, and if so does it depend on the workload?



How should we analyze the data?

So far we have considered several ways of analyzing **repeated measurements** (listed in decreasing order of complexity regarding the correlation between measurements on the same subject):

- Multivariate repeated measurements analysis
- Random coefficient models (RCM)
- Univariate repeated measurements analysis

With **fewer than 6-7 time-points** but several **missing observations** it seems appropriate to consider the **multivariate repeated measurement analysis** for example using `mixed` in **Stata**.
This produces an **approximate** test (likelihood ratio test) for the hypothesis of parallel mean curves in the two groups,

$$\textbf{Test 1:} \quad \text{LR} = 19.45 \sim \chi^2(4), \quad p = 0.0006$$

Model diagnostics in the multivariate analysis

One way to evaluate the model is to compare the observed and expected standard deviations and correlations. Observed standard deviations and correlations in the two groups:

Healthy

2.65				
0.94	3.21			
0.94	0.99	2.60		
0.92	0.96	0.98	2.52	
0.92	0.94	0.97	0.97	2.08

CHF

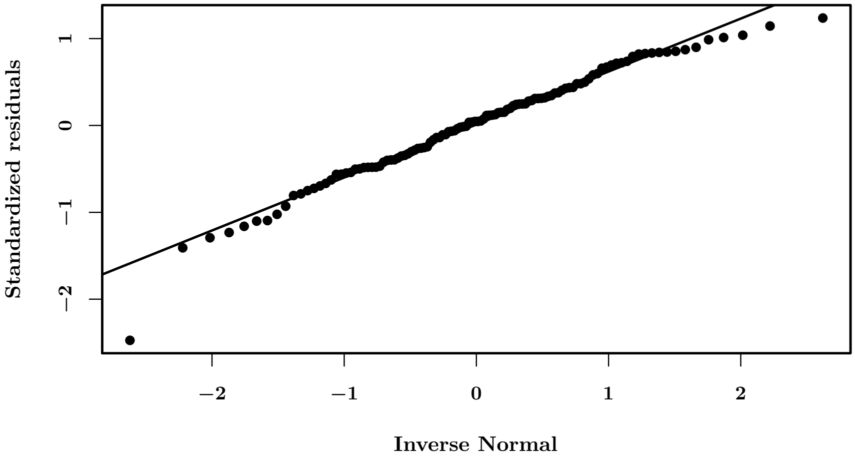
0.79				
0.84	1.53			
0.72	0.93	1.87		
0.78	0.93	0.96	1.95	
0.72	0.91	0.95	0.97	1.71

These are to be compared to each other (they should be similar) and to what is implied by the model ($p = 0.0035$ for the hypothesis of equal standard deviations and correlations in the two groups):

1.95				
0.91	2.46			
0.86	0.96	2.24		
0.85	0.94	0.97	2.32	
0.79	0.90	0.96	0.96	1.95

Except for the correlations with the cardiac output at workload 0 in the CHF-group the correlations look similar and are well described by the model. The sd's, however, are not that similar.

Cardiac output: Model diagnostics (QQ-plot)



Cardiac output: Excluding the data at workload 0

In the multivariate analysis we saw, particularly in the CHF-patients group, that the measurements corresponding to workload 0 were special compared to the rest of the measurements.

If we exclude the measurements corresponding to workload 0 (and only consider cardiac output during exercise), we get

$$p = 0.07$$

for the hypothesis of equal standard errors and correlations in the two groups.

In the multivariate repeated measurements model there is some evidence against the hypothesis of **parallel curves** ($p = 0.0397$).

Of course you should not exclude data from your analysis just like that, but it might be **instructive** to present the **results of both analyses**.

Cardiac output: Analyses with unequal standard deviations and correlations

Back to considering all the data:

In **Stata** it is possible to test the hypothesis of parallel mean curves in the two groups without assuming that the standard deviations and correlations are the same in the two groups.

```
mixed co bn.workload#bn.group || subj:  , nocons
residual(un, t(workload) by(group))
```

This results in an approximate test,

$$\textbf{Test 1:} \quad \mathrm{LR} = 16.75 \sim \chi^2(4), \quad p = 0.0022$$

Conclusion: There is still clear evidence against the hypothesis that the difference between the two groups is the same for all workloads.

Cardiac output: Observed and expected standard deviations and correlations

Observed standard deviations and correlations in the two groups:

Healthy

2.65				
0.94	3.21			
0.94	0.99	2.60		
0.92	0.96	0.98	2.52	
0.92	0.94	0.97	0.97	2.08

CHF

0.79				
0.84	1.53			
0.72	0.93	1.87		
0.78	0.93	0.96	1.95	
0.72	0.91	0.95	0.97	1.71

Expected standard deviations and correlations in the two groups:

Healthy

2.65				
0.93	3.10			
0.94	0.99	2.60		
0.93	0.96	0.98	2.65	
0.92	0.94	0.97	0.97	2.08

CHF

0.79				
0.84	1.53			
0.72	0.92	1.81		
0.78	0.93	0.96	1.95	
0.75	0.93	0.91	0.97	1.74

Conclusion: The observed standard deviations and correlations are well described by the model.

Presenting the results from the statistical analysis

Once the data have been analyzed by an appropriate statistical model the focus should be on how to present the results. Usually this involves:

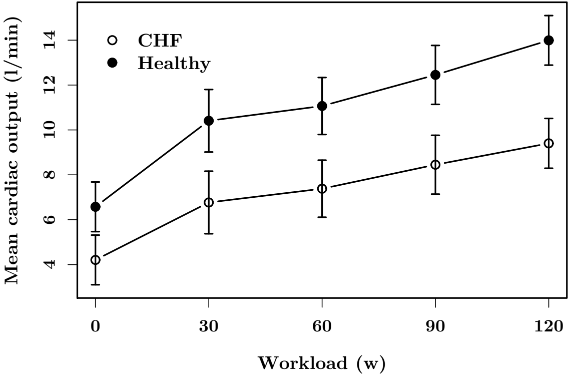
- Describing how the **data were analyzed**
- Presenting **estimated means** with 95%-confidence intervals
- Presenting **estimated effect sizes** (group differences) with confidence intervals (and possibly *p*-values)
- Describing how **model validation** was performed

In repeated measurements analyzes it is most common to focus on potential group differences at the different time-points, but other aspects may be of interest:

- Is the difference between group A and B more pronounced at time 2 than at time 1?

Cardiac output: Presenting the estimated means

In this example we get the following estimated means with 95%-confidence intervals:



A word of caution: The confidence intervals can be **very wide if the variation between subjects is considerable** (and therefore not very informative regarding changes over time).

Cardiac output: Presenting relevant estimated differences

In the cardiac output data it looks as if the difference between CHF-patients and healthy gets more pronounced with increasing workload.

Workload (w)	Healthy−CHF (l/min)	95%-CI	*p*-value
0	2.4	[0.8 , 3.9]	0.003
30	3.7	[1.8 , 5.7]	0.000
60	3.7	[1.9 , 5.5]	0.000
90	4.0	[2.1 , 5.9]	0.000
120	4.5	[3.0 , 6.0]	0.000

To see whether the difference is significantly bigger at workload 30 compared to workload 0, we can in **Stata** do the following:

```
lincom (30.workload#1.group - 30.workload#2.group) -
       (0.workload#1.group - 0.workload#2.group)
```

Cardiac output: Comparing group differences

Comparing successive group differences:

Workload (w)	Healthy−CHF (l/min)	95%-CI	*p*-value
30 − 0	1.36	[0.50 , 2.22]	0.002
60 − 30	−0.05	[−0.62 , 0.52]	0.870
90 − 60	0.33	[−0.12 , 0.79]	0.153
120 − 90	0.50	[−0.06 , 1.05]	0.082

We can summarize the difference between CHF-patients and healthy, when it comes to cardiac output, in the following way:

- The biggest increase in the group difference is seen when going from no workload to a moderate workload (0 − 30 w)
- At moderate workloads (30 − 90 w) the difference between CHF-patients and healthy is more or less constant
- When going from a moderate to a high workload (90 − 120 w) the difference between CHF-patients and healthy is once again increased (though not significantly)

Cardiac output: Comparing workload differences

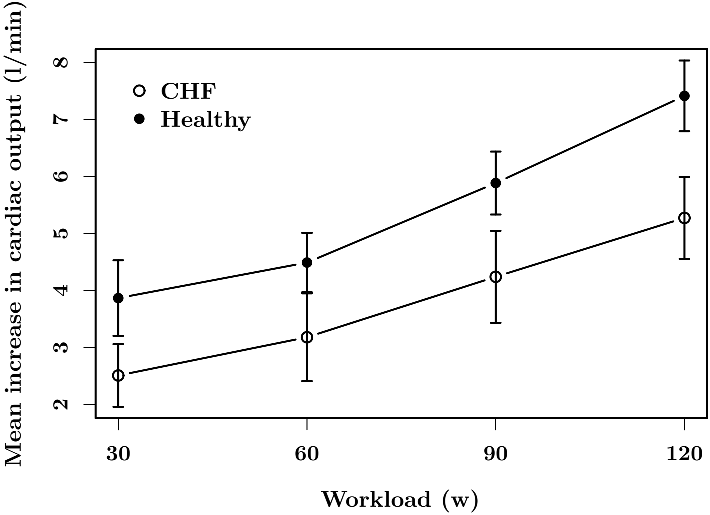
As for the group differences we can give the estimated effect of increasing the workload in the two groups.

Workload (w)	CHF	95%-CI	*p*-value	Healthy	95%-CI	*p*-value
30 − 0	2.5	[2.0 , 3.1]	0.000	3.9	[3.2 , 4.5]	0.000
60 − 30	0.7	[0.3 , 1.1]	0.001	0.6	[0.2 , 1.0]	0.002
90 − 60	1.1	[0.7 , 1.4]	0.000	1.4	[1.1 , 1.7]	0.000
120 − 90	1.0	[0.7 , 1.3]	0.000	1.5	[1.1 , 2.0]	0.000

We can summarize the effect of increasing the workload in the following way:

- The biggest increase in cardiac output is seen from 0 to 30 w in both groups
- With each increase in workload we expect to see an increase in cardiac output
- The increase in cardiac output (with increasing workload) is biggest in the healthy volunteer group (except from 30 to 60 w).

Increase (from rest) in mean cardiac output for each group



Cardiac output: Describing how the data were analyzed

A description of the data analysis should contain enough information to enable the reader to achieve the same results if he/she had access to the data and the statistical software used. In this example we might use the formulation:

- The cardiac output data were analyzed using a multivariate repeated measurements ANOVA with workload and group (and the interaction between them) as factors.
- The unequal standard errors and correlations in the two groups was taken into account in the analysis.
- The data were analyzed using **Stata** version 13.0.

Comments:

- There is no reason to mention the other simpler models that were considered but did not adequately describe the data.
- Different statistical packages can do different analyzes and can produce slightly different results when performing the same statistical analysis.
- Don't just write: "$p < 0.05$ was considered statistically significant" and then say significant / not significant. **Give the p-values**.

Cardiac output: Describing the test results

The results of the statistical tests can be formulated in the following way:

- There was clear evidence against the hypothesis of parallel curves in the two groups (identical to the hypothesis of no interaction between workload and group), $p = 0.0022$.
- If the data corresponding to workload 0 were excluded, the p-value was 0.0397.
- At all workloads there was a clear difference between the two groups (all p-values less than 0.003).
- The hypotheses of equal differences between CHF-patients and healthy volunteers at workload 30, 60, and 90 w were accepted (p-values over 0.15). At workload 0 w the difference was significantly smaller ($p = 0.002$), and at workload 120 w larger (though not significantly, $p = 0.082$).
- There were clearly significant increases in cardiac output with increasing workload in both groups (all p-values less than 0.003).

It may not be relevant to report all these test results.

Cardiac output: Describing the model validation

Of course we have to assume that people have checked the validity of the statistical analysis even though they do not report what they have done.

Often, though, it is fairly easy to reassure the readers that the results can be trusted by shortly describing what was done to validate the model.

In this example we could simply write the following:

- Model validation was performed by comparing observed and expected within subject standard deviations and correlations and by inspecting QQ-plots.
- There was some evidence against the hypothesis of equal standard deviations and correlations in the two groups ($p = 0.0035$), but this was confined to the measurements corresponding to workload 0 w in the CHF-patients group. If the data corresponding to workload 0 w were excluded, the p-value was 0.07 and all conclusions were unchanged.