R Markdown Basics

Inference in small and nonnormal samples In chapter 1, we outlined the process for conducting inference for models with repeated measures. When sample size is small, both Kenward-Rogers (KR) and Sattherthwaite approximations have been implemented to reduce Type I error rates.

Kenward-Rogers (1997)(CITE) proposes a Wald statistic in the form of:

$$F = 1/l(\hat{\beta} - \beta)^T L(L^T \hat{\Phi}_A L)^{-1} L^T (\hat{\beta} - \beta)$$

where l represents the number of linear combinations of the elements in β, L is a fixed matrix, and $\hat{\Phi}_A$ is the adjusted estimator for the covariance matrix of $\hat{\beta}$. As mentioned in chapter 1, $\hat{\Phi}$ is a biased estimator of Φ when samples are small, and underestimates. This adjusted estimator is broken down into $\hat{\Phi}_A = \hat{\Phi} + 2\hat{\Lambda}$, where $\hat{\Lambda}$ accounts for the amount of variation that was underestimated by the original estimator of covariance of $\hat{\beta}$. This Wald statistic that uses the adjusted estimator is scaled in the form:

$$F^* = \frac{m}{m + l - 1} \lambda F,$$

where m is the denominator degrees of freedom, and λ is a scale factor. Using the expectation and variance of the Wald statistic, F Both m and λ need to be calculated from the data, such that:

$$m = 4 + \frac{l+2}{l\rho - 1},$$

, where $\rho = \frac{V[F]}{2E[F]^2}$ and $\lambda = \frac{m}{E[F](m-2)}$. This statistic will ultimately follow an exact $F_{l,m}$ distribution.

Sattherthwaite approximation was developed by Fai & Cornelius (1996), with the F statistic following the form:

$$F = \frac{1}{l} \hat{\beta}' L' (L \Phi L')^{-1} L \hat{\beta}.$$

Note in this approximation we use the original Φ as the variance of $\hat{\beta}$. For the denominator degrees of freedom we perform spectral decomposition on $L'\Phi L = P'DP$, where D is a diagonal matrix of eigenvalues and P is an orthogonal matrix of eigenvectors. When r represents the r^{th} row of $P'L$, we have $v_r = \frac{2(d_r)^2}{g_r' W g_r}$, where g_r is a gradient vector, d_r is the r^{th} diagonal element of D, and W is the covariance matrix of $\hat{\sigma}^2$. The denominator degrees of freedom is calculated by:

$$\frac{2E}{E - l}$$

, where $E = \sum_{r=1}^{l} \frac{v_r}{v_r - 2} I(v_r > 2)$ if $E > l$, otherwise $DF = 1$.

When $l =1 $ the KR and Satterthwaite approximation will produce the same denominator degrees of freedom. However, since the statistic used for the two methods are not the same, the results for inference will not be the same.

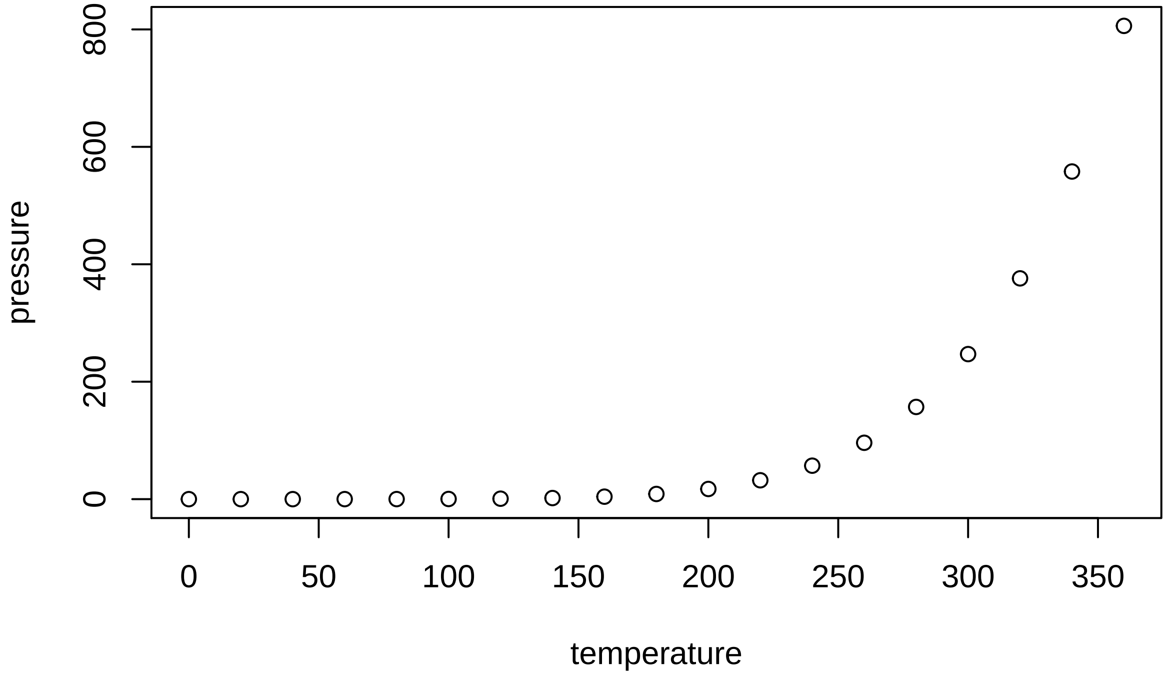
Both methods are frequently used and compared, and its performance is highly dependent on the structure of the data. A majority of studies focusing on DF method comparison in mixed models use split-plot design, as small sample sizes are more common in agricultural and biological fields. Schaalje, et al. (2002) found that in comparison to other degrees of freedom-adjusting methods like Satterthaite, KR was the most suitable for small sample data. Using factors such as imbalance, covariance structure, and sample size, they demonstrated that the KR method produced simulated Type I error rates closest to target values. However, their focus was primarily on complexity of covariance structure, and they found that more complicated structures, such as ante-dependence, produced inflated error rates when coupled with small sample size.

Arnau

As you see with `$2 \pi$` above, mathematics can be added by surrounding the mathematical text with dollar signs. More examples of this are in [Mathematics and Science] if you uncomment the code in [Math].

Including plots

You can also embed plots. For example, here is a way to use the base **R** graphics package to produce a plot using the built-in `pressure` dataset:



Note that the `echo=FALSE` parameter was added to the code chunk to prevent printing of the **R** code that generated the plot. There are plenty of other ways to add chunk options. More information is available at http://yihui.name/knitr/options/.

Another useful chunk option is the setting of `cache=TRUE` as you see here. If document rendering becomes time consuming due to long computations or plots that are expensive to generate you can use knitr caching to improve performance. Later in this file, you'll see a way to reference plots created in **R** or external figures.

Loading and exploring data

Included in this template is a file called `flights.csv`. This file includes a subset of the larger dataset of information about all flights that departed from Seattle and Portland in 2014. More information about this dataset and its **R** package is available at http://github.com/ismayc/pnwflights14. This subset includes only Portland flights and only rows that were complete with no missing values. Merges were also done with the `airports` and `airlines` data sets in the `pnwflights14` package to get more descriptive airport and airline names.

We can load in this data set using the following command:

```
flights <- read.file("data/flights.csv")
```

```
## Reading data with read.csv()
```

The data is now stored in the data frame called `flights` in **R**. To get a better feel for the variables included in this dataset we can use a variety of functions. Here we can see the dimensions (rows by columns) and also the names of the columns.

```r
dim(flights)
```

```
## [1] 52808    16
```

```r
names(flights)
```

```
##  [1] "month"        "day"          "dep_time"     "dep_delay"    "arr_time"    
##  [6] "arr_delay"    "carrier"      "tailnum"      "flight"       "dest"        
## [11] "air_time"     "distance"     "hour"         "minute"       "carrier_name"
## [16] "dest_name"
```

```r
# read long paragraph file
longtext <- readLines("data/paragraphs.txt")
```

```
## Warning in readLines("data/paragraphs.txt"): incomplete final line found on
## 'data/paragraphs.txt'
```

```r
# display text as vector
longtext
```

```
## [1] "Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut lab
## [2] ""
## [3] "Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla par
```

```r
# display text as paragraphs
cat(longtext)
```

```
## Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore
```

```r
# display text without linewidth option specified
longtext
```

```
## [1] "Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut lab
## [2] ""
## [3] "Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla par
```

Another good idea is to take a look at the dataset in table form. With this dataset having more than 50,000 rows, we won't explicitly show the results of the command here. I recommend you enter the command into the Console ***after*** you have run the **R** chunks above to load the data into **R**.

```r
View(flights)
```

While not required, it is highly recommended you use the `dplyr` package to manipulate and summarize your data set as needed. It uses a syntax that is easy to understand using chaining operations. Below I've created a few examples of using `dplyr` to get information about the Portland flights in 2014. You will also see the use of the `ggplot2` package, which produces beautiful, high-quality academic visuals.

The example we show here does the following:

- Selects only the `carrier_name` and `arr_delay` from the `flights` dataset and then assigns this subset to a new variable called `flights2`.
- Using `flights2`, we determine the largest arrival delay for each of the carriers.

```
flights2 <- flights %>%
  select(carrier_name, arr_delay)
max_delays <- flights2 %>%
  group_by(carrier_name) %>%
  summarize(max_arr_delay = max(arr_delay, na.rm = TRUE))
```

A useful function in the `knitr` package for making nice tables in *R Markdown* is called `kable`. It is much easier to use than manually entering values into a table by copying and pasting values into Excel or LaTeX. This again goes to show how nice reproducible documents can be! (Note the use of `results="asis"`, which will produce the table instead of the code to create the table.) The `caption.short` argument is used to include a shorter title to appear in the List of Tables.

```
kable(max_delays,
      col.names = c("Airline", "Max Arrival Delay"),
      caption = "Maximum Delays by Airline",
      caption.short = "Max Delays by Airline",
      longtable = TRUE,
      booktabs = TRUE)
```

Table 1: Maximum Delays by Airline

Airline	Max Arrival Delay
Alaska Airlines Inc.	338
American Airlines Inc.	1539
Delta Air Lines Inc.	651
Frontier Airlines Inc.	575
Hawaiian Airlines Inc.	407
JetBlue Airways	273
SkyWest Airlines Inc.	421
Southwest Airlines Co.	694
United Air Lines Inc.	472
US Airways Inc.	347
Virgin America	366

The last two options make the table a little easier-to-read.

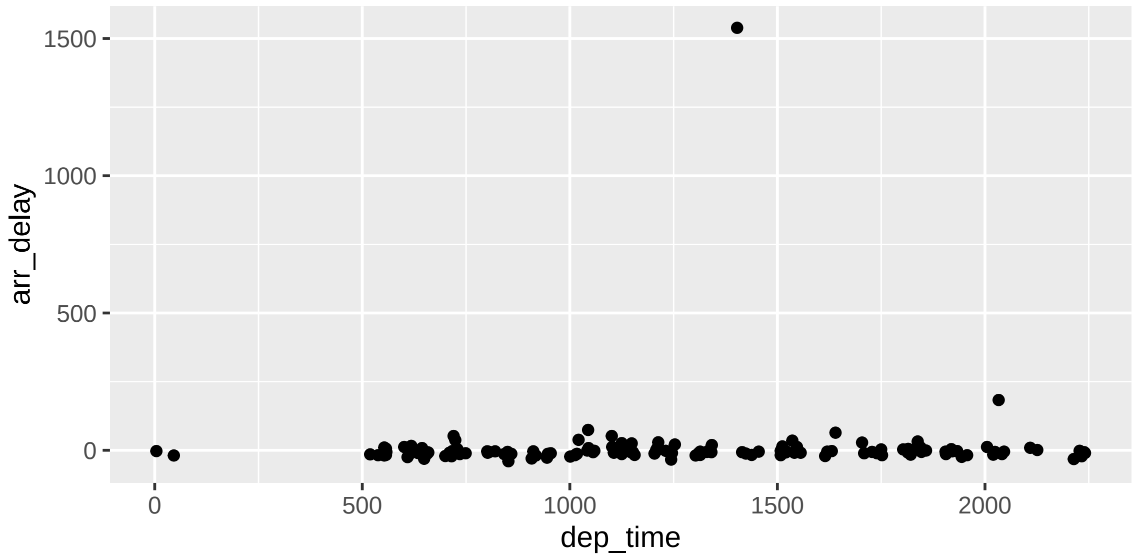
We can further look into the properties of the largest value here for American Airlines Inc. To do so, we can isolate the row corresponding to the arrival delay of 1539 minutes for American in our original `flights` dataset.

```
flights %>% filter(arr_delay == 1539,
                  carrier_name == "American Airlines Inc.") %>%
  select(-c(month, day, carrier, dest_name, hour,
            minute, carrier_name, arr_delay))
```

```
##   dep_time dep_delay arr_time tailnum flight dest air_time distance
## 1     1403      1553     1934  N595AA   1568  DFW      182     1616
```

We see that the flight occurred on March 3rd and departed a little after 2 PM on its way to Dallas/Fort Worth. Lastly, we show how we can visualize the arrival delay of all departing flights from Portland on March 3rd against time of departure.

```r
flights %>% filter(month == 3, day == 3) %>%
  ggplot(aes(x = dep_time, y = arr_delay)) + geom_point()
```



```
There is a proof environment in which you can create equations

$$
\hat{\beta}_0 + \hat{\beta}_1x
$$
```

Notes from lit review

Arnau (2014): Investigated robustness of the KR procedure when groups have different nonnormal distributions.Parameters considered: skewness, kurtosis, ratio of kurtosis differences across groups, unequal vs equal group sizes, pairing of large kurtosis with larger group size. violation of the sphericity assumption does not affect the robustness of the LMM combined with the KR procedure. the violation of skewness appears to have a greater effect on KR robustness than does the violation of kurtosis

KR was less robust when the relationship between the kurtosis values in the groups increased or when the largest group was associated with the largest value of kurtosis. Both of these effects on KR robustness were greater when total sample size was smaller, and especially when total sample sizes were 30.

he pairing of kurtosis with group size and the relationship between the kurtosis values in the groups are shown to be relevant variables to consider when using the LMM with KR

Arnau (2013): Specifically, we sought to examine whether skewness and kurtosis have a differential effect on KR robustness by exploring both independently.that for the repeated measures effect the LMM with KR was robust mainly when data were normal, regardless of whether the sphericity assumption was met. Likewise, for the interaction effect, the procedure was also robust when the total sample size was 45 or larger, but it was liberal when the total sample size was 30.

arnau(2012): that the test is more robust with a normal than with a log-normal distribution, whereas, overall, there are no significant differ- ences in performance between normal and exponential dis- tributions. In addition, when the covariance matrix is spherical, the test tends to become more robust with normal and exponential distributions, especially when the number of observations increases; this is contrary to what occurs with the log-normal distribution. Finally, a comparison of the two exponential distributions shows that the test becomes more robust as kurtosis increases, regardless of whether or not the assumption of sphericity is fulfilled.

s compared to the estimation of the time effect alone, the interaction between time and group leads to a consider- able increase in the test's robustness when the distribution is log-normal.

KR procedure is compromised with log- normal distributions that show moderate skewness, especial- ly as regards the estimation of the time effect. By contrast, when distributions are normal or have slight skewness (1 0 0.8), the test is robust even with extreme kurtosis

vallejo: when data were obtained from a normal distribution, the BF procedure of the main effect provided better control of the Type I error rate than did the Proc Mixed solution based on either the AIC or BIC criterion. For the non-normal distributions, both approaches become conservative or liberal as a function of the covariance structure and forms of the distribution used to generate the data. Nonetheless, the BF test was slightly less liberal than the test based on either the AIC or BIC criterion when both procedures had Type I error rates above the significance level and slightly more conservative when both had Type I error rates below the significance level. With regard to power, no test was uniformly most powerful

Additional resources

- *Markdown* Cheatsheet - https://github.com/adam-p/markdown-here/wiki/Markdown-Cheatsheet
- *R Markdown* Reference Guide - https://www.rstudio.com/wp-content/uploads/2015/03/rmarkdown-reference.pdf
- Introduction to `dplyr` - https://cran.rstudio.com/web/packages/dplyr/vignettes/introduction.html
- `ggplot2` Documentation - http://docs.ggplot2.org/current/