Statistics & Explanatory Data Analysis

ANOVA & related (more than 2 samples tests)

dr Marcin Chlebus, dr Ewa Cukrowska - Torzewska

Uniwersytet Warszawski
Wydział Nauk Ekonomicznych

One-way ANOVA (analysis of variance)

- One-way data. (1 DEPVAR in 2 or more groups)
- Dependent variable is interval/ratio, and is continuous
- Independent variable is a factor with two or more levels.
- Residuals are normally distributed
- Groups have the same variance (homoscedasticity)
- Observations among groups are independent.
- Moderate deviation from normally-distributed residuals is permissible

HYPOTHESIS:

- H0: All means in all groups are equal.
- H1 (2-sided): Exist at least one mean which is different then the rest of means

- **One-way (one factor) analysis of variance is a hypothesis-testing technique that is used to compare the means more than 2 samples.**

$$\begin{cases} H_0 : \mu_1 = \mu_2 = \mu_3 = \ldots = \mu_k \\ H_1 : \exists \mu_i \neq \mu_j \end{cases}$$

$$MSR_B = \frac{SS_B = \sum_{i=1}^{N} \left(\overline{y}_j - \overline{y}\right)^2}{df_B = k - 1}$$

$$MSR_W = \frac{SS_W = \sum_{i=1}^{N} \left(y_{i,j} - \overline{y}_j\right)^2}{df_W = N - k}$$

$$F = \frac{MSR_B}{MSR_W} \sim F(k-1, N-k)$$

One-way ANOVA assumptions testing

- **Residual normality:**
  - Shapiro-Wilk
  - Jarque-Bera etc.
- **Variance equality:**
  - Bartlett’s test: Compare the variances of k samples, where k can be more than two samples. The data must be normally distributed.
  - Levene’s test: Compare the variances of k samples, where k can be more than two samples. It’s an alternative to the Bartlett’s test that is less sensitive to departures from normality.
  - Fligner-Killeen test: a non-parametric test which is very robust against departures from normality.

HYPOTHESIS:
H0: all populations variances are equal;
H1: at least two of them differ

Source: http://www.sthda.com/english/

One-way ANOVA follow-up analysis

- **Pairwise comparison of means between different independent variable levels**
  - Least Square Means comparison - means for groups that are adjusted for means of other factors in the model (example: average salary for 2 different groups of people, where in first group much less have higher education)
  - Problem **unadjusted p-values & confidence intervals**
    - Goal is to adjust p-values in a way that rejecting any of single test at α level would be consistent in a way with α significance level for joint hypothesis

"**tukey**" Uses the Studentized range distribution with the number of means in the family. (Available for two-sided cases only.)
"**scheffe**" Computes p values from the F distribution, according to the Scheffe critical value of p kF(k, d), where d is the error degrees of freedom and k is (family size minus 1) for contrasts, and (number of estimates) otherwise. (Available for two-sided cases only.)
"**sidak**" Makes adjustments as if the estimates were independent (a conservative adjustment in many cases).
"**bonferroni**" Multiplies p values, or divides significance levels by the number of estimates. This is a conservative adjustment.
"**dunnettx**" Uses an approximation to the Dunnett distribution for a family of estimates having pairwise correlations of 0.5 (as is true when comparing treatments with a control with equal sample sizes). The accuracy of the approximation improves with the number of simultaneous estimates, and is much faster than "mvt". (Available for two-sided cases only.)
"**mvt**" Uses the multivariate t distribution to assess the probability or critical value for the maximum of k estimates. This method produces the same p values and intervals as the default summary or confint methods to the results of as.glht. In the context of pairwise comparisons or comparisons with a control, this produces "exact" Tukey or Dunnett adjustments, respectively. However, the algorithm (from the mvtnorm package) uses a Monte Carlo method, so results are not exactly repeatable unless the random-number seed is used (see set.seed). As the family size increases, the required computation time will become noticeable or even intolerable, making the "tukey", "dunnettx", or others more attractive.
"**holm**" the FWER controlled using Holm's (1979) progressive step-up procedure to relax control on subsequent tests. pvalues are ordered from smallest to largest, and adjusted p-values = max[1, p(m+1-i)], where i indexes the ordering. All tests after and including the first test to not be rejected at the alpha/2 level are not rejected.
"**hochberg**" the FWER is controlled using Hochberg's (1988) progressive step-down procedure to increase control on successive tests. p values are ordered from largest smallest, and adjusted p-values = max[1, p*i], where i indexes the ordering. All tests after and including the first to be rejected at the alpha/2 level are rejected.
"**none**" Makes no adjustments to the p values.

Multiple comparisons

- Standard statistical procedures can be misleading when researchers conduct a large group of hypothesis (finding significance even if there is none)
- Why multiple comparisons?
  - Many factors to be analysed (n groups of employees)
  - Heterogeneity of influence (different results for subgroups – ex. female vs. male)
  - Different tests
  - Different measures to prove effects (training vs total pasess, shots on target,.. etc.)
- Type I error is not α, but $1-(1-\alpha)^K$ under all H0 are true
- Two measures included in adjustments
  - Controlling **Family-Wise Error Rate** (FWER) at α level
    - α is a probability of at least one test result is false positive under all true H0
  - Controlling **False Discovery Rate** (FDR) at q level
    - q is a threshold for ratio (False Positive/All Positives)

http://egap.org/methods-guides/10-things-you-need-know-about-multiple-comparisons

Controlling FWER & FDR examples

- **Bonferroni correction - FWER**
  - Rejection on $\alpha_{corr} = \frac{\alpha}{k}$ level guarantee FWER at α level
  - Very conservative approach, especially for many tests
  - This correction works in the worst-case scenario that all tests are independent
  - II type error may be sufficient
- **Holm correction - FWER**
  - Procedure:
    - Order all k p-values in ascending order
    - Find the smallest p-value that $p > \frac{\alpha}{k+1-i}$, where *i* is the p-value index
    - The p-value and all higher are insignificant, all smaller are significant
- **Benjamini – Hochberg correction - FDR**
    - Order all k p-values in ascending order
    - Find the largest p-value that $p \leq \frac{i}{k}\alpha$, where *i* is the p-value index
    - The p-value and all lower are significant

Kruskall – Wallis test

DATA TYPE:

- One-way data
- Dependent variable is ordinal, interval, or ratio
- Independent variable is a factor with two or more levels.
- Observations between groups are independent.
- In order to be a test of medians, the distributions of values for each group need to be of **similar shape and spread.** Otherwise the test is a test of distributions.

HYPOTHESIS:

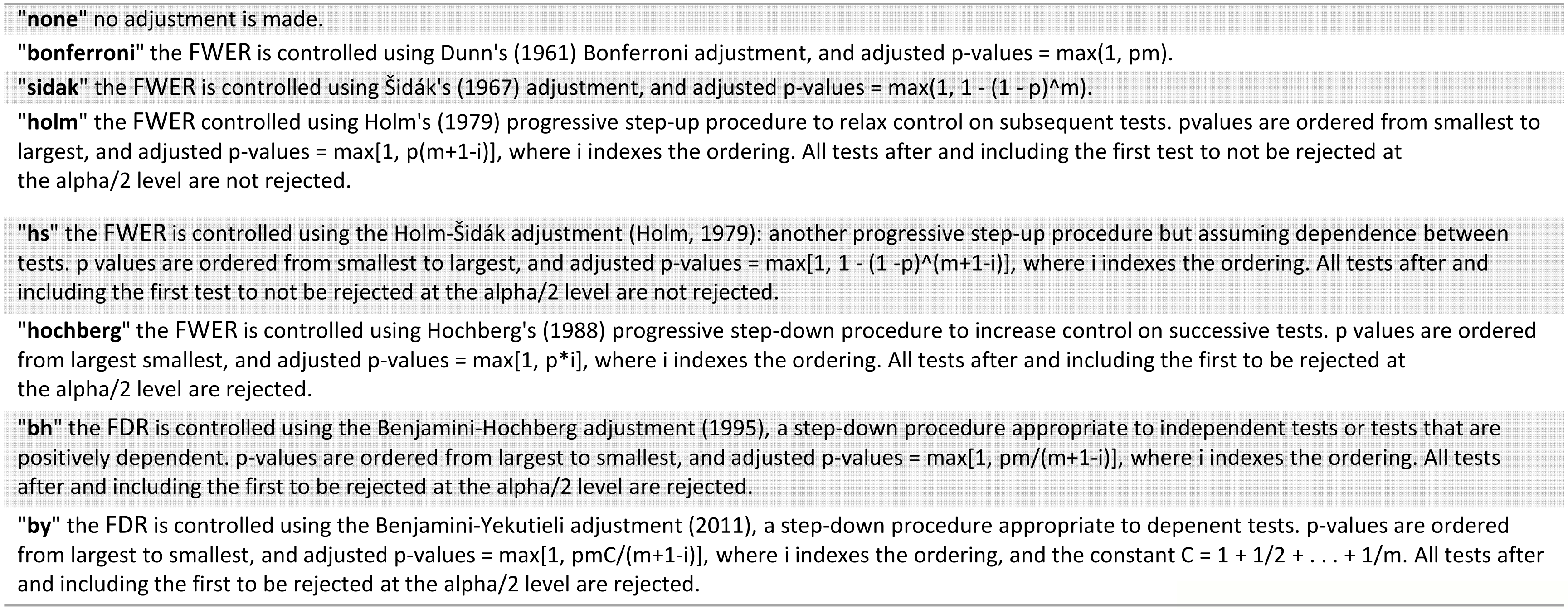
- H0: The medians of values for each group are equal (*distributions are similar in shape and spread)*/The distribution of values for each group are equal (otherwise)/there is no evidence of stochastic dominance between the samples.
- H1 (2-sided): The medians of values for each group are not equal/there is systematic difference in the distribution of values for the groups/ at least one sample stochastically dominates another sample

- **When groups are not similar in spread → Median Mood's test recommended**
- **Procedure:**
  1. Order all samples together in ascending order
  2. Assign Ranks to each observation (if ties assign average of ranks)

$$H = (N-1)\frac{\sum_{i=1}^{g} n_i(\overline{r}_i - \overline{r})^2}{\sum_{i=1}^{g}\sum_{j=1}^{n_i}\left(r_{ij} - \overline{r}\right)^2} \sim \chi^2(g-1)$$

Kruskal – Wallis follow-up analysis

- **Pairwise comparison of means between different independent variable levels**
  - Dunn test
  - Conover-Iman test - more powerful (works only if Kruskall – Wallis test is rejected)
  - Mann–Whitney tests on each pair of groups.
- **H0: The null hypothesis for each pairwise comparison is that the probability of observing a randomly selected value from the first group that is larger than a randomly selected value from the second group equals one half**



"**none**" no adjustment is made.

"**bonferroni**" the FWER is controlled using Dunn's (1961) Bonferroni adjustment, and adjusted p-values = max(1, pm).

"**sidak**" the FWER is controlled using Šidák's (1967) adjustment, and adjusted p-values = max(1, 1 - (1 - p)^m).

"**holm**" the FWER controlled using Holm's (1979) progressive step-up procedure to relax control on subsequent tests. pvalues are ordered from smallest to largest, and adjusted p-values = max[1, p(m+1-i)], where i indexes the ordering. All tests after and including the first test to not be rejected at the alpha/2 level are not rejected.

"**hs**" the FWER is controlled using the Holm-Šidák adjustment (Holm, 1979): another progressive step-up procedure but assuming dependence between tests. p values are ordered from smallest to largest, and adjusted p-values = max[1, 1 - (1 -p)^(m+1-i)], where i indexes the ordering. All tests after and including the first test to not be rejected at the alpha/2 level are not rejected.

"**hochberg**" the FWER is controlled using Hochberg's (1988) progressive step-down procedure to increase control on successive tests. p values are ordered from largest smallest, and adjusted p-values = max[1, p*i], where i indexes the ordering. All tests after and including the first to be rejected at the alpha/2 level are rejected.

"**bh**" the FDR is controlled using the Benjamini-Hochberg adjustment (1995), a step-down procedure appropriate to independent tests or tests that are positively dependent. p-values are ordered from largest to smallest, and adjusted p-values = max[1, pm/(m+1-i)], where i indexes the ordering. All tests after and including the first to be rejected at the alpha/2 level are rejected.

"**by**" the FDR is controlled using the Benjamini-Yekutieli adjustment (2011), a step-down procedure appropriate to depenent tests. p-values are ordered from largest to smallest, and adjusted p-values = max[1, pmC/(m+1-i)], where i indexes the ordering, and the constant C = 1 + 1/2 + . . . + 1/m. All tests after and including the first to be rejected at the alpha/2 level are rejected.

Mood’s median test

DATA TYPE:

- One-way data
- Dependent variable is ordinal, interval, or ratio
- Independent variable is a factor with two or more levels.
- Observations between groups are independent.
- Distributions of values for each group are similar in shape; however, the test is not sensitive to outliers

HYPOTHESIS:

- H0: The medians of values for each group are equal.
- H1 (2-sided): The medians of values for each group are not equal

- **When groups are not similar in spread → Median Mood’s test recommended**
- **Procedure**
  1. Order all samples together in ascending order
  2. Calculate median for joint sample
  3. Prepare contingency table (Below/Above median v
  4. Perform Fischer Exact or Pearson Chi2 test.
- **Post-hoc analysis**
  - Pairwise Mood’s Median Test

Low power in comparison to K-W test, but do not require approximately equal variances (scale/spread)

Only option for data with serious outliers

One-way ANOVA with blocks

DATA:

- One-way data, with blocks. That is, one measurement variable in two or more groups, where each group is also distributed among at least two blocks
- Dependent variable is interval/ratio, and is continuous
- Independent variable is a factor with two or more levels.
- A second independent variable is a blocking factor variable with two or more levels
- Residuals are normally distributed
- Groups have the same variance
- Observations among groups are independent.
- Moderate deviation from normally-distributed residuals is permissible

HYPOTHESIS:

- H0: The means of the measurement variable for each group are equal
- H1 (two-sided): The means of the measurement variable for among groups are not equal
- additional H0 (for the effect of blocks): The means of the measurement variable for each block are equal

- Blocks are used in ANOVA in order to account for suspected variation from factors other than the treatments or main independent variables being investigated (example: Earnings in Cities (fraction of Females).
- It helps to sort out a problem with non-independency of data
- When using blocks, the experimenter isn’t concerned necessarily with the effect of the blocks or even the factors behind assigning those blocks.

One-way ANOVA – types of SS

- When data is unbalanced, there are different ways to calculate the sums of squares for ANOVA.
  - For balanced data all gives the same results
- Consider a model that includes two factors **A and B** and its interaction **AB**.
  - The full model is represented by SS(A, B, AB).
  - Other models are represented similarly:
    - SS(A, B) indicates the model with no interaction,
    - SS(B, AB) indicates the model that does not account for effects from factor A, and so on.
- The influence of particular factors (including interactions) can be tested by examining the differences between models. For example, to determine the presence of an interaction effect, an F-test of the models SS(A, B, AB) and the no-interaction model SS(A, B) would be carried out.
- It is convenient to define incremental sums of squares to represent these differences:
  - $SS(AB|A,B) = SS(A,B,AB) – SS(A,B), SS(A|B,AB) = SS(A,B,AB) – SS(B,AB),$ $SS(B|A,AB) = SS(A,B,AB) – SS(A,AB),\ SS(A|B) = SS(A,B) – SS(B),\ SS(B|A) = SS(A,B) – SS(A)$
- Different types of sums of squares:
  - Type I: $SS(A)$ for A, $SS(B|A)$ for B & $SS(AB|B,A)$ for AB (different results for unbalanced data depending on order of factors) – it is testing A factor not controlling B factor
  - Type II: : $SS(A|B)$ for A, $SS(B|A)$ for B (powerful for cases with no interactions)
  - Type III: $SS(A|B,AB)$ for factor $A, SS(B|A,AB)$ for factor B (valid with interactions)
    - Note: if interactions exist interpretation of main effect alone is not so interesting

https://www.r-bloggers.com/anova-%E2%80%93-type-iiiiii-ss-explained/

One-way ANOVA with random blocks

DATA:

- One-way data, with blocks. That is, one measurement variable in two or more groups, where each group is also distributed among at least two blocks
- Dependent variable is interval/ratio, and is continuous
- Independent variable is a factor with two or more levels.
- A second independent variable is a blocking factor variable with two or more levels
- Residuals are normally distributed
- Groups have the same variance
- Observations among groups are independent.
- Moderate deviation from normally-distributed residuals is permissible

HYPOTHESIS:

- H0: The means of the measurement variable for each group are equal
- H1 (two-sided): The means of the measurement variable for among groups are not equal

- In analysis of variance, blocking variables are often treated as random variables. This is because the blocking variable represents a random selection of levels of that variable. The analyst wants to take the effects of the blocking variable into account, but the identity of the specific levels of the blocks are not of interest.
- Example: (example: Earnings in respect to Gender (from 3 Cities).

Fixed effects vs. Random Effects vs. Mixed Effects

- **Fixed effects model -** parameters are fixed or non-random quantities (regression models).
  - Fixed effects on the other hand are expected to have a systematic and predictable influence on your data.
- **Random effects models -** parameters are random variables.
  - So, a random effect is generally something that can be expected to have a non-systematic, idiosyncratic, unpredictable, or “random” influence on your data.
- **Mixed effects models -** some parameters are fixed and some are considered as random variables**.**
  - Fixed effect to estimate a parameter if all possible levels are included (e.g., both males and females
  - Use a random effect to account for a variable if the levels included are just a random sample from a population
- **Why to use mixed effects models?**
  - Multiple measures for a subject (multiple scenarios, reapeted measures)
    - Radom effects should help with violate indepednce assumption
  - NOTE: Panel data – two main approaches to estimate – different properties

http://www.bodowinter.com/tutorial/bw_LME_tutorial2.pdf

Different R^2

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{N}\left(y_i - \bar{y}_i\right)^2}$$

$$R^2_{McFadden} = 1 - \frac{ln\left(\hat{L}(M_{FULL})\right)}{ln\left(\hat{L}(M_{NULL})\right)}$$

McFadden

- R-squared as improvement from null model to fitted model
- Values from [0;1]
- The higher the better

$$R^2_{Cox\;\&\;Snell} = 1 - \left\{\frac{\hat{L}(M_{NULL})}{\hat{L}(M_{FULL})}\right\}^{2/N}$$

Cox - Snell

- R-squared as improvement from null model to fitted model
- Values from [0; $1 - \hat{L}(M_{NULL})^{2/N}$]
- The higher the better

$$R^2_{Negelkerke} = \frac{1 - \left\{\frac{\hat{L}(M_{NULL})}{\hat{L}(M_{FULL})}\right\}^{2/N}}{1 - \hat{L}(M_{NULL})^{2/N}}$$

Nagelkerke

- R-squared as improvement from null model to fitted model
- Values from [0; 1]
- The higher the better

https://stats.idre.ucla.edu/other/mult-pkg/faq/general/faq-what-are-pseudo-r-squareds/

Friedman test

DATA:

- Two-way data structured in *unreplicated complete block design (each block has one and only one observation of each treatment)*
- Dependent variable is ordinal, interval, or ratio
- Treatment or group independent variable is a factor with two or more levels.
- Blocking variable is a factor with two or more levels
- Blocks are independent of each other and have no interaction with treatments
- In order to be a test of medians, the distribution of the differences between each pair of groups are all symmetrical, or the distributions of values for each group have similar shape and spread. Otherwise the test is a test of distributions.

HYPOTHESIS:

- H0: The medians of values for each group are equal (*distribution of the differences between each pair of groups are all symmetrical, or the distributions of values for each group have similar shape and spread)*/The distribution of values for each group are equal (otherwise).
- H1 (2-sided): The medians of values for each group are not equal/there is systematic difference in the distribution of values for the groups

- The Friedman test is a generalization of the **paired sign test**
- The Friedman test may be preferable when there are a larger number of groups (five or more), while the Quade is preferable for fewer groups.
- For tests with ordinal dependent variables, cumulative link models or permutation tests may be alternative
- **Post-hoc testing:** pairwise Sign Test

Quade test

DATA:

- Two-way data structured in *unreplicated complete block design (each block has one and only one observation of each treatment)*
- Dependent variable is ordinal, interval, or ratio, although some authors say data must be interval or ratio only
- Treatment or group independent variable is a factor with two or more levels.
- Blocking variable is a factor with two or more levels
- Blocks are independent of each other and have no interaction with treatments
- In order to be a test of medians, the distributions of values for each group should have similar shape and spread. Otherwise the test is a test of distributions.

HYPOTHESIS:

- H0: The medians of values for each group are equal (*the distributions of values for each group have similar shape and spread)*/The distribution of values for each group are equal (otherwise).
- H1 (2-sided): The medians of values for each group are not equal/there is systematic difference in the distribution of values for the groups

- The Quade test is a generalization of the **two-sample Wilcoxon signed-rank test**
- The Friedman test may be preferable when there are a larger number of groups (five or more), while the Quade is preferable for fewer groups.
- Another alternative is to use cumulative link models for ordinal data
- **Post-hoc testing:** pairwise Wilcoxon signed-rank test

Two-way ANOVA (factorial ANOVA)

DATA:

- Two-way data. That is, one dependent variable measured across two independent factor variables
- Dependent variable is interval/ratio, and is continuous
- Independent variables are a factor with two or more levels.
- Residuals are normally distributed
- Groups have the same variance. That is, homoscedasticity
- Observations among groups are independent. That is, not paired or repeated measures data
- Moderate deviation from normally-distributed residuals is permissible

HYPOTHESIS:

- H0: The means of the first/second/interaction variable for each group are equal
- H1 (two-sided): The means of the first/second/interaction variable among groups are not equal

- A two-way (multiple-way) anova can investigate the *main* ***effects*** of each independent **factor** variables, as well as the effect of the ***interactions***.
- Post-hoc analysis:
  - *Neither the main effects nor the interaction effect is statistically significant* - **no post-hoc mean-separation testing**
  - *Only the* ***main effects*** *are* ***statistically significant*** - **post-hoc mean-separation testing for significant main effects only**.
  - ***Interaction effect is statistically significant*** - post-hoc mean-separation testing **interactions effect only**.

Scheirer–Ray–Hare test

DATA:

- Two-way data arranged in a factorial design
- Dependent variable is interval/ratio, and is continuous
- There are two treatment or group independent variables. Each is a factor with two or more levels.
- Observations among groups are independent. That is, not paired or repeated measures data
- It has been suggested that the observations should be balanced and that each cell in the interaction should have at least five observations.

HYPOTHESIS:

- H0: The distributions of values for each group in first/second variable or interaction are equal
- H1 (2-sided): There is systematic difference in the distributions of values for the groups in first/second variable or interaction

- The Scheirer–Ray–Hare test is a nonparametric test used for a two-way factorial design.
  - **It is ANOVA made on ranks (with some amendments)**
- The Scheirer–Ray–Hare test is less likely to find the interaction effect significant than would an ordinary least squares analysis of variance (less powerful)
- Appropriate post-hoc tests might be Dunn test or pairwise Mann–Whitney tests for each significant factor or interaction

Repeated measures ANOVA

DATA:

- One-way data
- Dependent variable is interval/ratio, and is continuous
- Independent variable is a factor with two or more levels.
- Unit ID – the same unit
- Time ID – moment of measurement
- Residuals are normally distributed
- Groups have the same variance
- Moderate deviation from normally-distributed residuals is permissible

HYPOTHESIS:

- H0: The means of the first/second/interaction variable for each group are equal
- H1 (two-sided): The means of the first/second/interaction variable among groups are not equal

- Idea:
  - experimental design takes measurements on the same unit over time, t
  - analysis must take into account possible (auto)correlation
- In ANOVA WITH (RANDOM) BLOCKS to deal with non-independent observations a blocking variable is used.
- In REPEATED MEASURES ANOVA autocorrelation structure is incorporated (we have time)
- **DIAGNOSTICS:** Loy, Adam Madison Montgomery, "Diagnostics for mixed/hierarchical linear models" (2013). Graduate Theses and Dissertations. 13277. http://lib.dr.iastate.edu/etd/13277

Correlation structures

$$\begin{bmatrix} \sigma_1^2 & \rho\sigma_1^2 & \rho^2\sigma_1^2 & & & \\ \rho\sigma_1^2 & \sigma_1^2 & \rho\sigma_1^2 & & 0 & \\ \rho^2\sigma_1^2 & \rho\sigma_1^2 & \sigma_1^2 & & & \\ & & & \sigma_2^2 & \rho\sigma_2^2 & \rho^2\sigma_2^2 \\ & 0 & & \rho\sigma_2^2 & \sigma_2^2 & \rho\sigma_2^2 \\ & & & \rho^2\sigma_2^2 & \rho\sigma_2^2 & \sigma_2^2 \end{bmatrix}$$

AR(1) – declinig exponentially with time

$$\begin{bmatrix} \sigma^2+\sigma_1^2 & \sigma_1^2 & \sigma_1^2 & & & \\ \sigma_1^2 & \sigma^2+\sigma_1^2 & \sigma_1^2 & & 0 & \\ \sigma_1^2 & \sigma_1^2 & \sigma^2+\sigma_1^2 & & & \\ & & & \sigma^2+\sigma_2^2 & \sigma_2^2 & \sigma_2^2 \\ & 0 & & \sigma_2^2 & \sigma^2+\sigma_2^2 & \sigma_2^2 \\ & & & \sigma_2^2 & \sigma_2^2 & \sigma^2+\sigma_2^2 \end{bmatrix}$$

Compound symmetry – constant correlation

$$\begin{bmatrix} \sigma_1^2 & \sigma_{12} & \sigma_{13} & & & \\ \sigma_{21} & \sigma_2^2 & \sigma_{23} & & 0 & \\ \sigma_{31} & \sigma_{32} & \sigma_3^2 & & & \\ & & & \sigma_4^2 & \sigma_{45} & \sigma_{46} \\ & 0 & & \sigma_{54} & \sigma_5^2 & \sigma_{56} \\ & & & \sigma_{64} & \sigma_{65} & \sigma_6^2 \end{bmatrix}$$

Unstructured– all values not related

Which structure to choose?

- **BY PARSIMONY**
  - Just as in traditional regression we want to **have as few parameters in the model as possible**.
  - The more data you have the more parameters you can fit, but they do not always add to our knowledge and often take away (**overfitting**).
  - From the fixed effects perspective, selecting a structure that is **too simple increases the fixed effects Type I error** rate, and selecting a structure that is **too complex sacrifices power and efficiency**.
- **BY MEANING**
  - Using understanding of the design and data structures and the meaning of the covariance structures will usually give you a few candidate structures to work with.
- **BY INFORMATION CRITERIA**
  - AIC = Akaike's Information Criteria,
  - AICC = AIC Corrected, and
  - BIC = Bayesian Information Criteria.
  - These statistics are functions of the log likelihood and number of parameters (to compare across models the fixed effects part of the model should be constant).
  - NOTE: Keselman, Algina, Kowalchuk, and Wolfinger 1998:
    - AIC selected the correct structure only 47 percent of the time
    - BIC only 35 percent of the time.

http://www2.sas.com/proceedings/sugi30/198-30.pdf

Guidelines for model selection based on IC

1. Carefully construct your candidate model set. Each model should represent a specific (interesting) hypothesis to test.
2. Keep your candidate model set short.
3. Check model fit. Use your global model (most complex model) or subglobal models to determine if the assumptions are valid. If none of your models fit the data well, information criteria will only indicate the most parsimonious of the poor models.
4. Avoid data dredging - data fishing, data snooping or p-hacking (looking for patterns after an initial round of analysis).
5. Avoid overfitting models. You should not estimate too many parameters for the number of observations available in the sample.
6. Be careful of missing values. Remember that values that are missing only for certain variables change the data set and sample size, depending on which variable is included in any given model.
7. Use the same response variable for all models of the candidate model set.
8. Determining the ranking of the models is just the first step. Akaike weights sum to 1 for the entire model set and can be interpreted as the weight of evidence in favor of a given model being the best one given the candidate model set considered and the data at hand.
    1. In cases where the top ranking model has an Akaike weight > 0.9, one can base inference on this single most parsimonious model.
    2. When many models rank highly (i.e., delta (Q)AIC(c) < 4), one should model-average effect sizes for the parameters with most support across the entire set of models.
    3. Model averaging consists in making inference based on the whole set of candidate models, instead of basing conclusions on a single 'best' model. It is an elegant way of making inference based on the information contained in the entire model set.

https://artax.karlin.mff.cuni.cz/r-help/library/AICcmodavg/html/aictab.html