Biostatistics & Epidemiological Data Analysis using R

11

Analysis of variance

Stefan Konigorski

Health Intervention Analytics Group, HPI

January 27, 2022

HPI Hasso Plattner Institut
Digital Engineering · Universität Potsdam

Content

Block	Class	Content	Date
R, Data manipulation, Descriptives	1	Overview & Introduction to R and data analysis	2021.10.28
	2	First steps in data analysis using R	2021.11.04
	3	Second steps in data analysis using R	2021.11.11
Epidemiology & Statistics: concepts	4	Epidemiological study designs	2021.11.18
	5	Estimation	2021.11.25
	6	Hypothesis testing & study planning	2021.12.02
	7	Missing data	2021.12.09
Data analysis w/ regression models	8	Linear regression I	2021.12.16
	9	Linear regression II	2022.01.13
	10	Regression models for binary and count data	2022.01.20
	11	Analysis of variance & Linear mixed models I	2022.01.27
	12	Linear mixed models II & Meta analysis	2022.02.03
	13	Survival analysis	2022.02.10
	14	Causal inference & Data analysis challenge	2022.02.17

(see full schedule online)

Introduction

Aim

- Predict a normally-distributed variable Y by a factor (i.e. categorical variable X).
- Test, if Y differs between the levels (=categories) of X.

Introduction

Examples

- Does the birth weight (as a continuous variable) differ between children, whose mothers have different BMI (categories)?
- Do the blood glucose levels differ between people who have had their last meal (i) 1 hour ago, (ii) 2 hours ago, (iii) 5 hours ago, (iv) 10 hours ago?
- Does the systolic blood pressure differ between age groups of children (1-way)?
- Does the systolic blood pressure differ between age groups of children and boys/girls (2-way)?

- 1-way ANOVA: with 1 factor.
- 2-way ANOVA: with 2 factors.

Introduction

How can we generally investigate and test the association between a normally-distributed variable Y and a categorical X?

Introduction

How can we generally investigate and test the association between a normally-distributed variable Y and a categorical X?

X has 2 levels

- 2-sample t-test or
- Test of the regression coefficients (Wald test with t-distribution) in a linear regression model - this allows to consider covariates as well.

Introduction

How can we generally investigate and test the association between a normally-distributed variable Y and a categorical X?

X has $k > 2$ levels

Introduction

How can we generally investigate and test the association between a normally-distributed variable Y and a categorical X?

X has $k > 2$ levels

- Pairwise t-tests of each level against each other.
- Test of the regression coefficients in a linear regression model, where X is included as a numerical (ordinal) variable (`lm(Y ~ as.numeric(X))`).
- Test of the regression coefficients in a linear regression model, where X is included as $k - 1$ dummy variables (`lm(Y ~ as.factor(X))`).

Introduction

How can we generally investigate and test the association between a normally-distributed variable Y and a categorical X?

X has $k > 2$ levels

- Pairwise t-tests of each level against each other.
- Test of the regression coefficients in a linear regression model, where X is included as a numerical (ordinal) variable (`lm(Y ~ as.numeric(X))`).
- Test of the regression coefficients in a linear regression model, where X is included as $k - 1$ dummy variables (`lm(Y ~ as.factor(X))`).

$\longrightarrow$ Disadvantages of these approaches?

Introduction

How can we generally investigate and test the association between a normally-distributed variable Y and a categorical X?

X has $k > 2$ levels

- Pairwise t-tests of each level against each other.
- Test of the regression coefficients in a linear regression model, where X is included as a numerical (ordinal) variable (`lm(Y ~ as.numeric(X))`).
- Test of the regression coefficients in a linear regression model, where X is included as $k - 1$ dummy variables (`lm(Y ~ as.factor(X))`).

$\longrightarrow$ Disadvantages of these approaches?

$\longrightarrow$ Construct new test within GLM framework: ANOVA
(has therefore same assumptions as linear regression)

Introduction

Perspectives on ANOVA

1. Generalization of t-test with > 2 groups.
2. Test within linear regression model, i.e. in GLM with normally-distributed Y with identity as link function.

Approach

- Null hypothesis: the means of Y are the same across all levels of the factor X.
- Look at variation (sum of squares) of Y between the levels of X and within the levels of X.
- This comparison allows to test mean differences in Y between the levels of X.

Decomposition of sum of squares

Decomposition of sum of squares in 1-way ANOVA

$$SS_{total} = SS_{\text{between groups}} + SS_{\text{within groups}}$$

$$\sum_{i=1}^{n} (y_i - \overline{y})^2 = \sum_{i=1}^{n} (\hat{y}_i - \overline{y})^2 + \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

Decomposition of sum of squares

Decomposition of sum of squares in 1-way ANOVA

$$SS_{total} = SS_{\text{between groups}} + SS_{\text{within groups}}$$

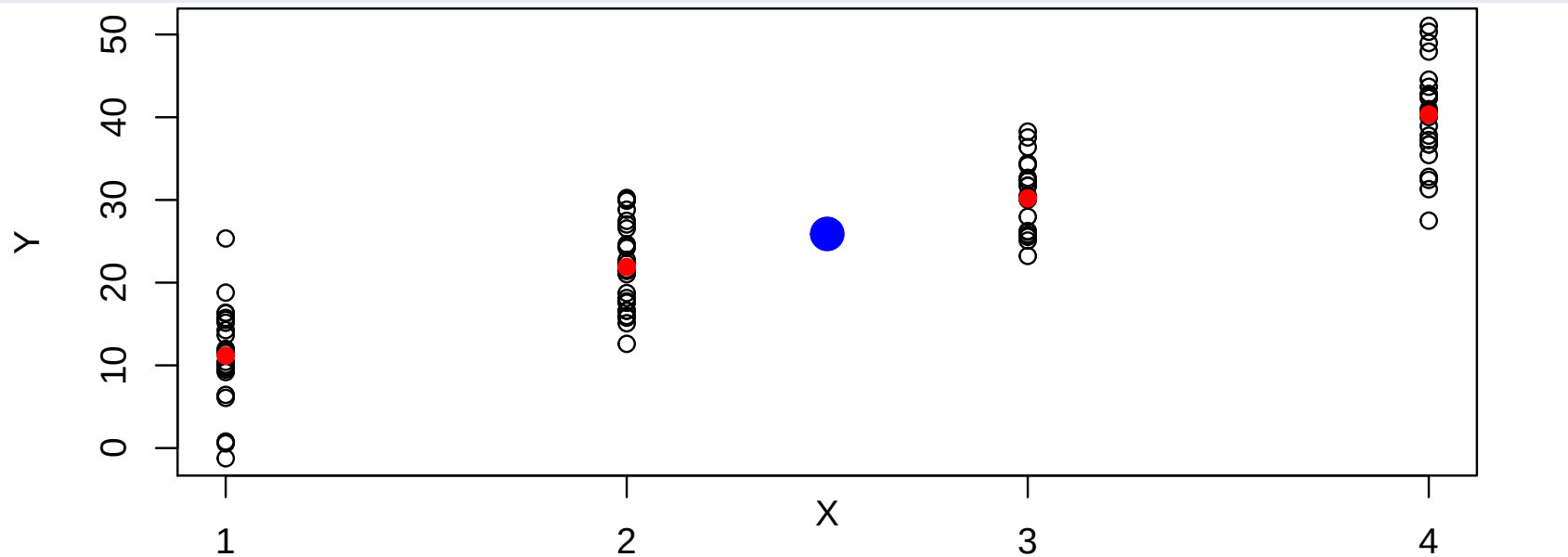
$$\sum_{i=1}^{n} (y_i - \overline{y})^2 = \sum_{i=1}^{n} (\hat{y}_i - \overline{y})^2 + \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

- SS_{total}: total variation of Y
- $SS_{\text{between groups}}$: Variation between the predicted Y values from the model (=mean of Y in factor levels) around the mean of Y. $\longrightarrow$ Variation of Y values which can be explained by the factor levels.
- $SS_{\text{within groups}}$: Variation between the actual and predicted values of Y. $\longrightarrow$ Variation of Y values around the respective group mean (hence "within group").

Decomposition of sum of squares

$$\sum_{i=1}^{n} (y_i - \overline{y})^2 = \sum_{i=1}^{n} (\hat{y}_i - \overline{y})^2 + \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

Example



See `R_11a_ANOVA_SS.Rmd`

Decomposition of sum of squares

Decomposition of sum of squares in 1-way ANOVA

$$SS_{total} = SS_{\text{between groups}} + SS_{\text{within groups}}$$

$$\sum_{i=1}^{n} (y_i - \overline{y})^2 = \sum_{i=1}^{n} (\hat{y}_i - \overline{y})^2 + \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

Using SS to test mean differences

- Evaluate the proportion of the variation between and within the groups to test the null hypothesis that the mean of Y is the same in all groups (vs. at least two are different).
- Test statistic: $F = \frac{SS_{\text{between groups}}}{df_{\text{between groups}}} \Big/ \frac{SS_{\text{within groups}}}{df_{\text{within groups}}}$

Decomposition of sum of squares

1-way ANOVA

$$SS_{total} = SS_{\text{between groups}} + SS_{\text{within groups}}$$
$$= SS_X + SS_{resid}$$

Model	Sum of squares	Degrees of freedom	Mean sum of squares	F value	p value
Factor X	SS_X	$k-1$	$MS_X = SS_X/(k-1)$	$F = MS_X/MS_{resid}$	$p = P(F_{df_1=k-1, df_2=n-k} > F)$
Residual	SS_{resid}	$n-k$	$MS_{resid} = SS_{resid}/(n-k)$		
Total	SS_{total}	$n-1$			

k = Number of levels of factor X

Decomposition of sum of squares

2-way ANOVA

$$\begin{aligned} SS_{total} &= SS_{\text{between groups}} + SS_{\text{within groups}} \\ &= (SS_{X_1} + SS_{X_2}) + SS_{\text{resid}} \end{aligned}$$

Decomposition of sum of squares

2-way ANOVA

$$SS_{total} = SS_{\text{between groups}} + SS_{\text{within groups}}$$
$$= (SS_{X_1} + SS_{X_2}) + SS_{\text{resid}}$$

Model	Sum of squares	Degrees of freedom	Mean sum of squares	F value	p value
Factor X_1	SS_{X_1}	$k_1 - 1$	$MS_{X_1} = SS_{X_1}/(k_1 - 1)$	$F_1 = MS_{X_1}/MS_{resid}$	$p = P(F_{df_1=k_1-1, df_2=n-k_1-k_2+1} > F_1)$
Factor X_2	SS_{X_2}	$k_2 - 1$	$MS_{X_2} = SS_{X_2}/(k_2 - 1)$	$F_2 = MS_{X_2}/MS_{resid}$	$p = P(F_{df_1=k_2-1, df_2=n-k_1-k_2+1} > F_2)$
Residual	SS_{Resid}	$n - k_1 - k_2 + 1$	$MS_{resid} = SS_{resid}/(n - k_1 - k_2 + 1)$		
Total	SS_{Gesamt}	$n - 1$			

k_1 = Number of levels of X_1, k_2 = Number of levels of X_2

Model equation

Model can be written like a linear regression equation, or also as:

1-way ANOVA

$$Y_{ij} = \mu + \alpha_j + \varepsilon_{ij}$$

- for person $i = 1 \ldots n$, factor levels $j = 1 \ldots k$,
- $Y \sim N(\mu, \sigma^2)$,
- μ is overall mean, α_j is difference between μ and mean of level j of the factor.

Model equation

Model can be written like a linear regression equation, or also as:

1-way ANOVA

$$Y_{ij} = \mu + \alpha_j + \varepsilon_{ij}$$

- for person $i = 1 \ldots n$, factor levels $j = 1 \ldots k$,
- $Y \sim N(\mu, \sigma^2)$,
- μ is overall mean, α_j is difference between μ and mean of level j of the factor.

2-way ANOVA

$$Y_{ijl} = \mu + \alpha_j + \gamma_l + \varepsilon_{ijl}$$

- for person $i = 1 \ldots n$, levels $j = 1 \ldots k_1$ of factor 1, levels $l = 1 \ldots k_2$ of factor 2, $Y \sim N(\mu, \sigma^2)$.

ANOVA in R

ANOVA in R

- `aov(Y ~ X1 + X2)` or `anova(lm(Y ~ X1 + X2))`

ANOVA in R

ANOVA in R

- `aov(Y ~ X1 + X2)` or `anova(lm(Y ~ X1 + X2))`

Exercise 1

In KiGGS dataset:

- Compute a 1-way ANOVA to investigate if systolic blood pressure (sys12) differs between the age groups of the children (age2).
- Compute a 2-way ANOVA to investigate if systolic blood pressure (sys12) differs between the age groups of the children (age2) and between boys and girls (sex).
- See exercise 1 in `R_11a_exercises.Rmd`.

Additional material

- Interactions
- Posthoc tests
- Contrasts
- Extensions to ANOVA
- SS types

Interactions - overview

Technical

Add another factor into the model, which is the product of two factors.

Interpretation

Interaction contains whether the effect of factor X_1 on Y differs between the levels of factor X_2 (or vice versa).

Interactions - overview

Technical

Add another factor into the model, which is the product of two factors.

Interpretation

Interaction contains whether the effect of factor X_1 on Y differs between the levels of factor X_2 (or vice versa).

Visualization

Using profile plots = line/scatter plots of Y on the y-axis, one factor on the x-axis, and different lines for the levels of the other factor.

Interaction - theory

2-way ANOVA with interaction - SS decomposition

$$\begin{aligned} SS_{total} &= SS_{\text{between groups}} + SS_{\text{within groups}} \\ &= (SS_{X_1} + SS_{X_2} + SS_{X_1X_2}) + SS_{resid} \end{aligned}$$

Table analogue to slide 12; the degrees of freedom of the interaction term are $(k_1 - 1) \cdot (k_2 - 1)$.

2-way ANOVA with interaction - model equation

$$Y_{ijl} = \mu + \alpha_j + \gamma_l + (\alpha\gamma)_{jl} + \varepsilon_{ijl}$$

for person $i = 1 \ldots n$, level $j = 1 \ldots k_1$ of factor 1, level $l = 1 \ldots k_2$ of factor 2, $Y \sim N(\mu, \sigma^2)$.

Interactions in ANOVA in R

Interactions

`aov(Y ~ X1 + X2 + X1:X2)` or short `aov(Y ~ X1*X2)`

Line (profile) plots of interactions

Eg with `interactions::cat_plot(aov(Y ~ X1*X2), pred = X1, modx = X2, geom = "line")`

(Adjusted) means

- Profile plots show the predicted means of Y by the model for the different factor levels. They can be extracted with `predict(aov(), Xvaluesforprediction)`.
- Non-adjusted means of Y for the different factor levels can be computed with: `tapply(as.numeric(Y), list(X1, X2), mean, na.rm = TRUE)`
- See exercise 2 in `R_11a_exercises.Rmd`.

Exercise

Exercise 2

Follow-up to exercise 1:

- Compute an ANOVA to investigate if systolic blood pressure (sys12) differs between the age groups of the children (age2), between boys and girls (sex), and between the levels of their interaction term.
- Visualize the interaction in a profile plot.
- See exercise 2 in `R_11a_exercises.Rmd`

Analysis of covariance

Analysis of covariance (ANCOVA)

ANCOVA = ANOVA with additional (quantitative) covariates.

Analysis of covariance

Analysis of covariance (ANCOVA)

ANCOVA = ANOVA with additional (quantitative) covariates.

Exercise 3

- Compute an ANOVA to investigate if systolic blood pressure (sys12) differs between the age groups of the children (age2), between boys and girls (sex), and between the levels of their interaction term, adjusting for BMI.
- Look at the predicted adjusted systolic blood pressure means. Interpret the results.
- See exercise 3 in `R_11a_exercises.Rmd`.

Posthoc tests vs. contrasts

Posthoc tests

- If a factor with 5 levels was tested, and has a significant effect on Y, what does it tell us? Do all levels differ or only some?

Posthoc tests vs. contrasts

Posthoc tests

- If a factor with 5 levels was tested, and has a significant effect on Y, what does it tell us? Do all levels differ or only some?

  $\longrightarrow$ Posthoc tests (=pairwise t-tests) with correction for multiple testing

Posthoc tests vs. contrasts

Posthoc tests

- If a factor with 5 levels was tested, and has a significant effect on Y, what does it tell us? Do all levels differ or only some?

  $\longrightarrow$ Posthoc tests (=pairwise t-tests) with correction for multiple testing

- In R: eg with `pairwise.t.test` or `TukeyHSD` function.

Posthoc tests vs. contrasts

Posthoc tests

- If a factor with 5 levels was tested, and has a significant effect on Y, what does it tell us? Do all levels differ or only some?

  $\longrightarrow$ Posthoc tests (=pairwise t-tests) with correction for multiple testing
- In R: eg with `pairwise.t.test` or `TukeyHSD` function.

Contrasts

- Is the research question a priori to test differences between specific levels or combinations of levels? Then this should be done with "a priori t-tests" (contrasts).
- In R: directly in `aov` function or eg with functions in packages `lsmeans` or `multcomp`.

Extensions of ANOVA

If the assumptions of ANOVA are violated (check assumptions of linear regression), there are several possibilities:

Continous Y/distribution/homoscedasticity not satisfied

- Friedman test: ANOVA based on ranks for >2 dependent samples (in R: `friedman.test` function).
- Kruskal-Wallis test: ANOVA based on ranks for >2 independent samples (in R: `kruskal.test` function).
- But with the restriction to 1-way ANOVA and without covariates.
- Alternative: eg ordinal/multinomial regression.

Independent observations

- Generalization of t-test for dependent samples: Between-within ANOVA, GLMMs.

SS types

SS types

- There are different types of SS in ANOVA, which are relevant when there are 2 or more factors in the model.
- They differ in how the variation of Y that can be explained by different factors is partitioned over the factors.

SS types

SS types

- There are different types of SS in ANOVA, which are relevant when there are 2 or more factors in the model.
- They differ in how the variation of Y that can be explained by different factors is partitioned over the factors.
- If the data is balanced (i.e. no missing factor level combinations, all factor levels have same number of observations), then all SS types yield same results.
- There is no consensus which SS type is best for unbalanced data.
- You can just use the default (R: type I; e.g. SAS: type III), but might have to consider it in the interpretation.

SS types

Type I (sequential)

- Explain total SS first by X_1, the rest by X_2 etc.
- Advantage: balanced, good partition if there is an intrinsic order in factors.
- Disadvantage: order makes a difference.

SS types

Type I (sequential)

- Explain total SS first by X_1, the rest by X_2 etc.
- Advantage: balanced, good partition if there is an intrinsic order in factors.
- Disadvantage: order makes a difference.

Type II (hierarchichal/partially sequential)

- SS of a factor = reduction of SS_{resid} when the factor is included in the model after all other factors are in the model, except for those that contain him.
- Advantage: Good for model selection, has highest power without interaction, is not dependent on order.
- Disadvantage: Doesn't consider interaction terms, not optimal if factors levels don't occur equally often.

SS types

Type III (marginal/orthogonal)

- SS of a factor = reduction of SS_{resid} when the factor is included as last factor into the model.
- Advantage: Balanced, are equal to the LS means (=best estimates of the marginal means), not dependent on whether factors levels occur equally often.
- Disadvantage: Not appropriate for missing cells (=factor level combinations), main effects are tested adjusted for interaction effects.

Type IV (balanced)

- Modification of SS type III to handle missing cells.

SS types

In R

- `aov()` computes SS type I $\longrightarrow$ order of factors plays a role!
- To compute SS type II or III, you can use the `car::aov` or `car::Anova` functions. Before, you have to change the default setting in the factor coding by running `options(contrasts = c("contr.sum", "contr.poly"))`.

References

- Menard (2000) Coefficients of determination for multiple logistic regression analysis. The American Statistician 54: 17-24.
- Mittlbock, Schemper (1996) Explained variation in logistic regression. Statistics in Medicine 15: 1987-1997.
- Agresti (2012) Categorical data analysis. Wiley.