Applied Generalized Linear Models (FS 20)

ANOVA and ANCOVA

Viviana Amati

Course structure - schedule

Date	Topic	Assignment
18.02	Introduction to the course	Ass. 1 released on 25.02 due to 19.03
25.02	Introduction to R and review of the linear regression model	
03.03	**The general linear model: ANOVA and ANCOVA**	
10.03	Practical: ANOVA and ANCOVA	
17.03	Binary outcomes: logistic regression and probit models	Ass. 2 released on 17.03 due to 23.04
24.03	Practical: logistic regression and probit models	
31.03	Nominal outcomes: multinomial logistic regression	
07.04	Practical: multinomial logistic regression	
21.04	Ordinal outcomes: ordered logistic regression and probit models	Ass. 3 released on 21.04 due to 21.05
28.04	Practical: ordered logistic regression and probit models	
05.05	Count outcomes: Poisson and negative binomial models	
12.05	Practical: Poisson and negative binomial models	
19.05	Survival models (lecture+practical)	
26.05	Exam	

Today's agenda

- Introduction to ANOVA and ANCOVA

- Material:
  Slides and lecture notes (Chapter 2)

An Example

- The “classic pullover” company has collected data on pullover sales (number of pullovers sold over the last week) in 30 shops which have adopted different marketing strategies
- Three marketing strategies each adopted by 10 shops
  - advertisement in local newspaper
  - presence of sales assistant
  - luxury presentation in shop windows
- Strategy were randomly assigned to the shops
- Is there an association between the pullover sales and the marketing strategy?

Association

	Observations									
Strategy	1	2	3	4	5	6	7	8	9	10
Assistant	8	10	10	11	12	15	12	13	13	11
Newspaper	7	9	10	10	11	9	11	12	13	13
Window	12	14	14	14	15	16	17	17	17	18

Any idea?

Association

Strategy	Observations										Average
	1	2	3	4	5	6	7	8	9	10	
Assistant	8	10	10	11	12	15	12	13	13	11	11.5
Newspaper	7	9	10	10	11	9	11	12	13	13	10.5
Window	12	14	14	14	15	16	17	17	17	18	15.4

Comparing group means

$$H_0 : \mu_A = \mu_N = \mu_W \quad ,$$

with μ_j the population mean in group j,

Association

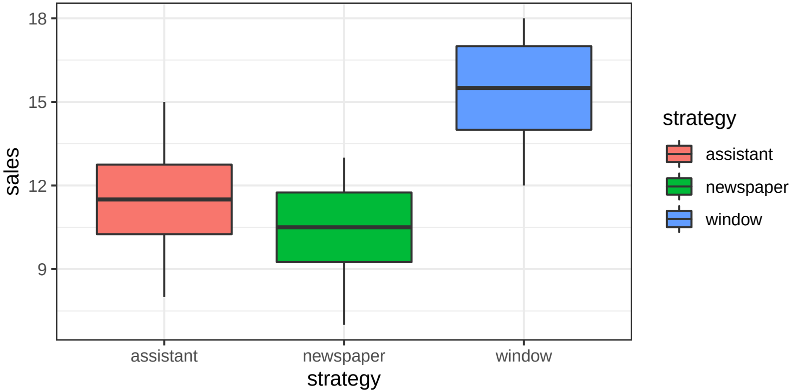
Strategy	Observations 1	2	3	4	5	6	7	8	9	10	Average	s.d.
Assistant	8	10	10	11	12	15	12	13	13	11	11.5	1.958
Newspaper	7	9	10	10	11	9	11	12	13	13	10.5	1.900
Window	12	14	14	14	15	16	17	17	17	18	15.4	1.897

Comparing group means

$$H_0 : \mu_A = \mu_N = \mu_W \quad ,$$

with μ_j the population mean in group j, while accounting for group variability

Comparing group means



1. Linear regression model with categorical explanatory variables
2. Analysis of variance ANOVA

LRM with categorical explanatory variables

- Y continuous and X categorical with M categories $c_1, \ldots, c_M$
- For each category c_j $(j = 1 \ldots, M)$ we create dummy variables D_j

$$D_{ij} = \begin{cases} 1 & \text{if} \quad x_i = c_j \\ 0 & \text{otherwise} \end{cases}$$

- LRM
  - Model without intercept

    $$E[Y|\mathbf{D}] = \gamma_1 D_1 + \ldots + \gamma_M D_M \quad ,$$

    with $\mathbf{D}$ the matrix of dummy variables
  - γ_j: expected value of Y when X takes category c_j

LRM with categorical explanatory variables

- Y continuous and X categorical with M categories $c_1, \ldots, c_M$
- For each category c_j $(j = 1 \ldots, M)$ we create dummy variable D_j

$$D_{ij} = \begin{cases} 1 & \text{if} \quad x_i = c_j \\ 0 & \text{otherwise} \end{cases}$$

- LRM
  - Model with intercept and reference category c_M

    $$E[Y|\mathbf{D}] = \beta_0 + \beta_1 D_1 + \ldots + \beta_{M-1} D_{M-1}$$

    with $\mathbf{D}$ the matrix with a first column of 1s and $M-1$ dummy variables
  - β_0: expected value of Y when X takes the reference category c_M
  - β_j: expected difference in Y for two subjects with categories c_j and c_M, respectively

LRM with categorical explanatory variables

Testing group means

- We would like to test

$$H_0 : \mu_1 = \mu_2 = \ldots = \mu_M$$

$$H_1 : \text{at least two of the group means are different}$$

- This is equivalent to test

$$H_0 : \beta_1 = \ldots = \beta_{M-1} = 0$$

$$H_1 : \text{at least one of } \beta_j \neq 0, \quad \text{j=1,...,M-1}$$

- F-test

$$\frac{\text{SSReg}/(M-1)}{\text{SSR}/(n-M)} \sim F_{M-1,n-M}$$

LRM: back to the example

Dummy variables

Strategy	D_1	D_2	D_3
presence of sales assistant	1	0	0
luxury presentation in shop windows	0	1	0
advertisement in local newspaper	0	0	1

LRM: back to the example

Example: model estimation

	Estimate	Std. Error	t value	Pr(>|t|)
intercept	10.5000	0.6068	17.31	0.0000
assistant	1.0000	0.8581	1.17	0.2541
window	4.9000	0.8581	5.71	0.0000

At significance level $\alpha = 0.05$

- β_0: on average a shop with “advertisement in local newspaper” sells nearly 11 pullovers per week
- β_1: not significantly different from 0. No difference in weekly pullover sales when using advertisement in local newspaper or sales assistant
- β_2: significantly different from 0. On average the weekly pullover sales of a shop with strategy “luxury presentation in shop windows” is of nearly 5 pullovers larger than that of a shop with strategy “advertisement in local newspaper”

LRM: back to the example

Example: model estimation

	Estimate	Std. Error	t value	Pr(>|t|)
intercept	10.5000	0.6068	17.31	0.0000
assistant	1.0000	0.8581	1.17	0.2541
window	4.9000	0.8581	5.71	0.0000

At significance level $\alpha = 0.05$

- β_0: on average a shop with "advertisement in local newspaper" sells nearly 11 pullovers per week
- β_1: not significantly different from 0. No difference in weekly pullover sales when using advertisement in local newspaper or sales assistant
- β_2: significantly different from 0. On average the weekly pullover sales of a shop with strategy "luxury presentation in shop windows" is of nearly 5 pullovers larger than that of a shop with strategy "advertisement in local newspaper"
- F-test: $F_{2,27} = 18.21$, p-value: < 0.001
  There is a difference among the population means of pullover sales for the three marketing strategies

ANOVA

Term used in two different contexts:

- LRMs: the partition of the SST into the SSReg and SSR
  used to compute the coefficient of determination R^2 and the (overall and partial) F-tests to evaluate the fit of a model

- Design of experiments: statistical methods for testing and fitting linear models in which the explanatory variables are categorical
  - categorical variables are referred to as factors and their categories as levels
  - experiments aim to test whether one (or more factors) have an effect on an outcome variable

ANOVA: data

Factor	Observations			
Level 1	y_{11}	y_{21}	$\dots$	y_{n1}
Level 2	y_{12}	y_{22}	$\dots$	y_{n2}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
Level M	y_{1M}	y_{2M}	$\dots$	y_{nM}

- Y be a continuous dependent variable
- observations partitioned into M groups determined by the levels of the factor
- y_{ij}: observation of Y for the i-th unit in the j-th level
- n_j: number of observations in the j-th group is n_j
- $N = \sum_{j=1}^{M} n_j$: total number of observations

ANOVA: model

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

with

- μ: the population mean of Y
- α_j: effect on the dependent variable in the j-th group
- ε_{ij}: error term
  - independent and normally distributed
  - $E[\varepsilon_{ij}] = 0$ and $\mathrm{Var}[\varepsilon_{ij}] = \sigma^2$

Implication:

$$Y_{ij} \sim N(\mu + \alpha_j, \sigma^2)$$

ANOVA: estimation

- The parameters cannot be uniquely estimated

  M categories, $M-1$ equations identifying the parameters
- Sigma constraint on the model parameters

$$\sum_{j=1}^{M} \alpha_j = 0$$

- The estimates of the model parameters are

$$\mu = \frac{1}{M} \sum_{j=1}^{M} \mu_j = \mu.$$

$$\alpha_j = \mu_j - \mu.$$

with

  - $\mu.$: population mean
  - μ_j: mean in the j-th group
  - α_j: difference between the mean of the j-th group and the general mean

ANOVA: deviation regressors

- To estimate the parameters under the sigma constraint, we use $M-1$ deviation regressors S_j with elements

$$S_{ij} = \begin{cases} 1 & \text{for observations in group } j \\ -1 & \text{for observations in group } M \\ 0 & \text{otherwise} \end{cases} .$$

- Deviation-coded model can be expressed as

$$Y_{ij} = \mu + \alpha_1 S_{i1} + \alpha_2 S_{i2} + \ldots + \alpha_{M-1} S_{i(M-1)} + \varepsilon_{ij}$$

- Equations for the group means:

$$\begin{aligned} \mu_1 &= \mu + \alpha_1 \\ &\vdots \\ \mu_{M-1} &= \mu + \alpha_{M-1} \\ \mu_M &= \mu - \alpha_1 - \alpha_2 - \ldots - \alpha_{M-1} \end{aligned}$$

ANOVA: hypothesis testing

Under the sigma constraint,

$$H_0 : \mu_1 = \mu_2 = \ldots = \mu_M$$

is equivalent to

$$H_0 : \alpha_1 = \alpha_2 = \ldots = \alpha_{M-1} = 0$$

H_0 is tested by the overall F-test for the deviation-coded model

ANOVA: F-test

SST, SSReg and SSR of the deviation-code model take simple formulas as illustrated by the ANOVA table

	Sum of squares (SS)	df	Mean Square (MS)	F-test
Factor	$\sum_{j=1}^{M} n_j(\overline{y}_j - \overline{y})^2$	$M-1$	$\frac{SSReg}{M-1}$	$\frac{MSReg}{MSR}$
Residuals	$\sum_{j=1}^{M}\sum_{i=1}^{n_j}(y_{ij} - \overline{y}_j)^2$	$N-M$	$\frac{SSR}{N-M}$	
Total	$\sum_{j=1}^{M}\sum_{i=1}^{n_j}(y_{ij} - \overline{y})^2$	$N-1$		

- ANOVA is usually presented as an F-test for comparing group means
- The table above tells us whether group means are different but not which group means differ $\rightarrow$ post-hoc analysis (Bonferroni test)

ANOVA: back to the example

Deviation-coded model

$$Y_{ij} = \mu + \alpha_1 S_{i1} + \alpha_2 S_{i2} + \varepsilon_{ij} \quad ,$$

with deviation regressors:

Strategy	S_1	S_2
assistant	1	0
window	0	1
newspaper	-1	-1

ANOVA: back to the example

ANOVA table

	Sum Sq	Df	Mean Sq	F value	Pr(>F)
Strategy	134.07	2	67.03	18.21	0.0000
Residuals	99.40	27	3.68		
Total	233.47	29			

We reject

$$H_0 : \mu_A = \mu_N = \mu_W$$

The deviation coding is equivalent to the dummy coding in that they both lead to the same fit to the data and overall F-test

ANOVA: back to the example

Estimates of the deviation-coded model

	Estimate	Std. Error	t value	Pr(>|t|)
intercept	12.467	0.350	35.588	0.000
assistant	-0.967	0.495	-1.951	0.061
window	2.933	0.495	5.921	0.000

- μ: average number of pullover sold by the 30 shops is 12.4
- α_1: not significant. No difference between the average number of pullovers sold in one week by all the shops and the group of shops with “sales assistant” strategy
- α_2: shops with strategy “presentation in shop windows” on average sell nearly 3 pullover more than the average number of pullovers sold in one week by all the shops

ANCOVA

- Linear models that contains both qualitative and quantitative explanatory variables
- ANOVA formulation for the main effects of the categorical explanatory variables (i.e. deviation coding)
- Quantitative explanatory variables are expressed as deviations from their means (i.e. centred explanatory variables)
- ANCOVA provides a more intuitive interpretation of the model parameters