

Statistics: part 2

Regression Analysis and SPSS

Analysis of Variance I

(syllabus 6.1 – 6.2)

Bjorn Winkens
Methodology and Statistics
University of Maastricht
Bjorn.Winkens@stat.unimaas.nl

6 June 2008

Content

Analysis of variance:

- Example
- Comparison with linear regression
- Pair-wise comparisons + contrasts
- Multiple-testing: corrections
- Assumptions + check

ANOVA (1)

Application:

- Effect of one or more categorical variables (factors) on a continuous outcome variable
- ➔ One categorical variable: **one-way ANOVA**
 - **Example:** effect of different treatments on blood pressure
- ➔ Two categorical variables: **two-way ANOVA**
 - **Example:** effect of sex and different treatments on blood pressure

ANOVA (2)

= **linear regression**, where all independent variables/predictors are *categorical*

→ same results as linear regression with *dummy variables*

→ old-fashioned

→ GLM: combination of regression and ANOVA

Analysis of covariance (ANCOVA):

→ correction for a continuous covariate

= **linear regression** with *categorical and continuous* variables

Example: obstruction time (1)

Design:

- randomized clinical trial
- 189 rats
- 9 different diets:
 - sunflower-seed oil (50SO)
 - coconut fat with sunflower-seed oil (5SO)
 - crude palm oil (CPO)
 - 6 refined palm oils (POR, PORO, PORS, PON, PONO, PONS)

Example: obstruction time (2)

Research question:

Effect of diet on obstruction time (= time between insertion and complete obstruction of aorta loop)

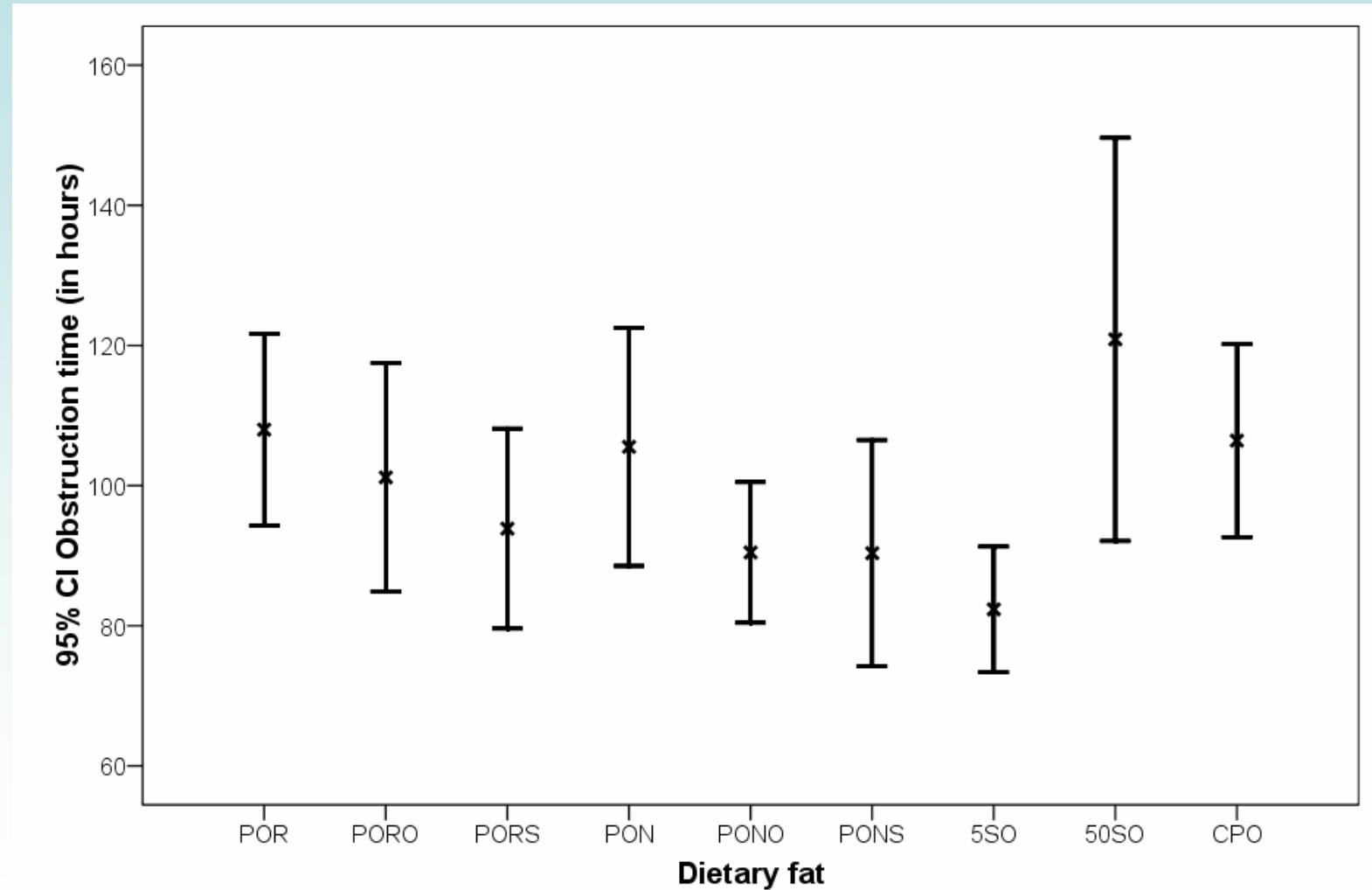
→ The longer the obstruction time, the lower the arterial thrombosis tendency

Example: obstruction time (3)

Obstruction time (in hours)

Dietary fat	N	Mean	SD	Min	Max
POR	22	107.99	30.86	54.75	186.75
PORO	20	101.19	34.81	46.00	174.75
PORS	21	93.87	31.36	51.25	162.25
PON	22	105.53	38.32	48.75	186.50
PONO	23	90.47	23.20	42.50	142.25
PONS	18	90.36	32.46	46.00	190.75
5SO	23	82.35	20.73	42.00	120.00
50SO	21	120.87	63.17	45.75	296.25
CPO	19	106.42	28.58	48.00	164.00

Example: obstruction time (4)



Example: obstruction time (5)

ANOVA (SPSS - GLM):

Tests of Between-Subjects Effects

Dependent Variable: Obstruction time (in hours)

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	23742.529 ^a	8	2967.816	2.351	.020
Intercept	1874102.341	1	1874102.341	1484.370	.000
Diet	23742.529	8	2967.816	2.351	.020
Error	227260.401	180	1262.558		
Total	2131814.125	189			
Corrected Total	251002.929	188			

a. R Squared = .095 (Adjusted R Squared = .054)

Remark - Type III Sum of squares:

correct for all other variables included in model

(Type I Sum of Squares: only correction for variables that precede the corresponding one)

Example: obstruction time (6)

Linear regression (ANOVA table):

ANOVA^b

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	23742.529	8	2967.816	2.351	.020 ^a
	Residual	227260.4	180	1262.558		
	Total	251002.9	188			

a. Predictors: (Constant), diet1, diet2, diet3, diet4, diet5, diet6, diet7, diet8

b. Dependent Variable: Obstruction time (in hours)

Model: $OT = \beta_0 + \beta_1 \text{Diet1} + \dots + \beta_8 \text{Diet8} + \varepsilon$

→ same results as ANOVA

→ what is tested?

Example: obstruction time (7)

F-test (ANOVA):

- $H_0: \mu_1 = \dots = \mu_9$
- H_1 : not all means are equal
- **$F = MS(\text{corrected model}) / MS(\text{Error})$** ←

F-test (Linear regression):

- $H_0: \beta_1 = \dots = \beta_8 = 0$
- H_1 : not all beta's are equal to 0
- **$F = MS(\text{regression}) / MS(\text{Residual})$** ←

equivalent



Example: obstruction time (8)

- **F-test significant ($p = 0.020$)**
 - conclusion?
 - finished?

**ANOVA:
(GLM)**

Parameter Estimates						
Dependent Variable: Obstruction time (in hours)						
Parameter	B	Std. Error	t	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
Intercept	106.421	8.152	13.055	.000	90.336	122.506
[Diet=1.00]	1.568	11.128	.141	.888	-20.391	23.526
[Diet=2.00]	-5.234	11.383	-.460	.646	-27.695	17.228
[Diet=3.00]	-12.552	11.250	-1.116	.266	-34.752	9.648
[Diet=4.00]	-.887	11.128	-.080	.937	-22.846	21.072
[Diet=5.00]	-15.954	11.016	-1.448	.149	-37.690	5.783
[Diet=6.00]	-16.060	11.687	-1.374	.171	-39.122	7.002
[Diet=7.00]	-24.073	11.016	-2.185	.030	-45.810	-2.337
[Diet=8.00]	14.448	11.250	1.284	.201	-7.752	36.648
[Diet=9.00]	0 ^a

a. This parameter is set to zero because it is redundant.

[link7](#)

Example: obstruction time (9)

Linear regression:

Coefficients^a

Model	Unstandardized Coefficients		Standardized Coefficients	t	Sig.	
	B	Std. Error	Beta			
1	(Constant)	106.421	8.152		13.055	.000
	diet1	1.568	11.128	.014	.141	.888
	diet2	-5.234	11.383	-.044	-.460	.646
	diet3	-12.552	11.250	-.108	-1.116	.266
	diet4	-.887	11.128	-.008	-.080	.937
	diet5	-15.954	11.016	-.143	-1.448	.149
	diet6	-16.060	11.687	-.129	-1.374	.171
	diet7	-24.073	11.016	-.216	-2.185	.030
	diet8	14.448	11.250	.125	1.284	.201

a. Dependent Variable: Obstruction time (in hours)

→ same results as ANOVA

Pair-wise comparisons (1)

- Compare means of group i and j
- Number of pair-wise tests for k groups $k*(k - 1)/2$

- Test:

$$t = \frac{\text{difference in means}}{\text{se(difference)}} = \frac{M_i - M_j}{s \sqrt{\frac{1}{n_i} + \frac{1}{n_j}}}$$

[link16](#)

- s = standard deviation, computed over **all** groups
- $s = \sqrt{\text{residual mean square}}$
- $df = n - k$, where n = total sample size
- Alternative for *unequal variances* available

Pair-wise comparisons (2)

Multiple-testing problem:

- $P(\text{type I error}) = P(H_0 \text{ is rejected} | H_0 \text{ is true}) = \alpha$,
for each test
- ➔ overall type I error $> \alpha$
- Need a method to correct for multiple testing
- Choice of method depends on number of comparisons (tests)

Pair-wise comparisons (3)

Fisher's LSD:

- *t*-test with significance level α
- Least significant difference = $t_{1-\alpha/2, n-k} * \text{se}(\text{diff})$
- **no correction for multiple testing**

[link14](#)

Bonferroni correction:

- *t*-test with $\alpha^+ = \alpha / (\text{number of tests})$
- **SPSS:**
 - $p^+ = (\text{number of tests}) * p \rightarrow$ compare with original α
 - assumes number of tests = $k*(k - 1)/2$
- Very conservative
- Improvement: **Hochberg & Rom (1995)**

Pair-wise comparisons (4)

Tukey's range test:

- Significant if $t > q_{(k, n-k, \alpha)} / \sqrt{2}$
- $q_{(k, n-k, \alpha)} / \sqrt{2}$ in Appendix E1 & E2 (syllabus)
- **Honestly significant difference (HSD):**
- if $n_1 = \dots = n_k = n/k$, **HSD = $s * q_{(k, n-k, \alpha)} / \sqrt{(n/k)}$**

Scheffé's test:

- Significant if $t > \sqrt{[(k-1) * F_{1-\alpha, k-1, n-k}]}$
- **More conservative than Tukey's range test**

Pair-wise comparisons (5)

Dunnett's test:

- Compare multiple treatment groups with the same (a priori chosen) control group
- Often used in dose-finding studies

Duncan's test:

- No control of overall type I error (can be larger than α)

... and many more

Pair-wise comparisons (6)

Recommendations:

- Small number of tests (< 5): **Bonferroni**
- All groups with one control group: **Dunnett**
- All pair-wise comparisons: **Tukey**

Linear contrasts:

- Small number (< 5): **Bonferroni**
- Large number: **Scheffé**

Linear contrasts

Definition:

- $L = \sum c_j \mu_j$, where $\sum c_j = 0$

Example (obstruction time):

- $\mu_{\text{POR}} - \mu_{\text{PORO}}$
- $(\mu_{\text{POR}} + \mu_{\text{PORO}} + \mu_{\text{PORS}})/3 - (\mu_{\text{PON}} + \mu_{\text{PONO}} + \mu_{\text{PONS}})/3$

Test:

- $H_0: \sum c_j \mu_j = 0; H_1: \sum c_j \mu_j \neq 0$
- $t = \sum c_j M_j / [s * \sqrt{(\sum c_j^2 / n_j)}]$
- **SPSS:** complicated contrasts only by **syntax**

Example: obstruction time (10)

- Overall *F*-test significant ($p = 0.020$)
→ pair-wise comparisons ($\alpha = 0.05$)
 - LSD:
5SO - POR (0.02), PON (0.03), **50SO (0.00)**, CPO (0.03)
50SO - PORS (0.02), PONO (0.01), PONS (0.01)
 - Bonferroni: **5SO with 50SO ($p = 0.015$)**
 - Tukey: **5SO with 50SO ($p = 0.012$)**
 - Scheffé: **no significant differences**
(5SO with 50SO → $p = 0.124$)
- } **CONCLUSION?**

Example: obstruction time (11)

POR, PORO, PORS, PON, PONO, PONS, 5SO, 50SO, CPO

Interesting contrasts:

L1) Refined oils (POR, ..., PONS) vs crude oil (CPO)

$$(\mu_{\text{POR}} + \mu_{\text{PORO}} + \mu_{\text{PORS}} + \mu_{\text{PON}} + \mu_{\text{PONO}} + \mu_{\text{PONS}})/6 - \mu_{\text{CPO}}$$

$$- c = (1/6, 1/6, 1/6, 1/6, 1/6, 1/6, 0, 0, -1)$$

$$- \sum c_j = 0$$

– **SPSS:**

/CONTRAST (Diet)= special(1, 1, 1, 1, 1, 1, 0, 0, -6)

$$- 6 * L1 = -49.11 \rightarrow L1 = -49.11/6 = -8.19$$

$$- p = 0.351$$

Example: obstruction time (12)

L2) Physically refined oils (POR, PORO, PORS) vs
Chemically refined oils (PON, PONO, PONS)

$$(\mu_{\text{POR}} + \mu_{\text{PORO}} + \mu_{\text{PORS}})/3 - (\mu_{\text{PON}} + \mu_{\text{PONO}} + \mu_{\text{PONS}})/3$$

– $c = (1/3, 1/3, 1/3, -1/3, -1/3, -1/3, 0, 0, 0)$

– $\sum c_j = 0$

– **SPSS:**

 /CONTRAST (Diet)= special(1, 1, 1, -1, -1, -1, 0, 0, 0)

– $3*L2 = 16.68 \rightarrow L1 = 16.68/3 = 5.56$

– $p = 0.382$

Note: SPSS gives uncorrected p -values for contrasts

Assumptions (1)

1. Independence

- **check:** by design

2. Homoscedasticity

- homogeneity of variances
- **check:** several options (next sheet)

3. Normality

- outcome variable is normally distributed *within each group*
- **check:** normal pp-plot (qq-plot), histogram

Assumptions (2)

Check for homoscedasticity:

1. Levene's test (H_0 : equal variances)

- SPSS
- Not recommended

2. Ratio of variances (largest/smallest):

- If largest/smallest ≤ 2 , then OK
- Rough indication

3. Plot $\log(\text{SD})$ vs $\log(\text{mean})$

- Compute slope b
- Indication which data transformation may be useful (syllabus 6.1.5)

Assumptions (3)

Homoscedasticity and/or normality violated:

1. Data transformation (e.g. Y^{-1} , $\log(Y)$, \sqrt{Y}):

- Which one? Indication by plot $\log(\text{SD})$ vs $\log(\text{mean})$

2. Welch test:

- does not assume homoscedasticity
- **SPSS:** Analyze > Compare Means > One-way ANOVA

3. Non-parametric:

- Two groups: **Mann-Whitney**
- More than two groups: **Kruskal-Wallis**

Example: obstruction time (13)

- **Independence:** OK
- **Homoscedasticity:**
 - Levene's test: $p = 0.002$
 - Ratio of variances: $(63.17/20.73)^2 = 9.29$
 - Log(SD) vs Log(mean): $b = 2.2$
- **Normality:**
 - Normal qq-plots: some show small deviation of normality

Conclusion: homoscedasticity assumption violated
→ data transformation (**syllabus** & **SPSS practicum**)

Summary

- **ANOVA/ANCOVA** = linear regression with dummy variables
- **Goal:** study effect of categorical variables on one continuous outcome variable
 - practice: comparison of more than two means
- **Multiple-testing problem:** several solutions
 - choice depends on number of tests
 - see *recommendations (sheet 19)*

Next week: repeated measurements

Which analysis method?

- **ANOVA?**
 - Repeated measurements ANOVA?
 - ANOVA with subject as fixed factor?
 - ANOVA with subject as random factor?
- **Summary Measures?**
- **Linear Mixed Models?**

LAST LECTURE: which topic?

Option 1: overview

- Correlation, linear, logistic regression, ANOVA
 - Applications, assumptions, brief summary
 - Examples: which method is preferred?
 - Modeling strategies

Option 2: own research

- Questions, practical problems, ...

Option 3: ???