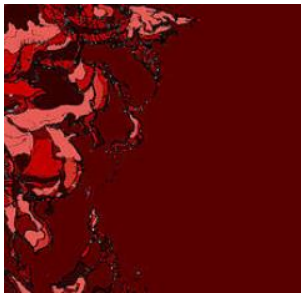
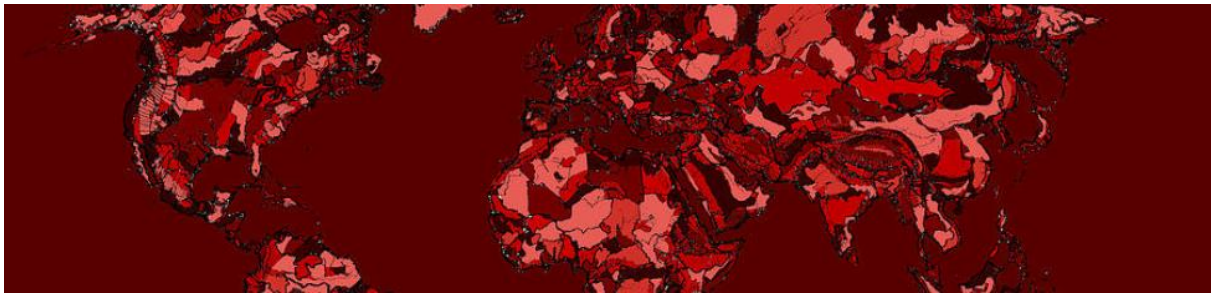
STA501: Data-based Decision Making

Lecture 4: ANOVA to multi-variate regressions

Swiss Institute of
Artificial Intelligence





Nature of Chi-square test

- Z-test for multivariable cases to test 1.Independence, 2.Homogeneity, (and goodness-of-fit)

Chi-square test for independence

$$Q = \sum_{i=1}^{k} X_i^2 \sim \chi_k^2$$

- A test based on chi-squared distribution
- X are standard normal distributions
  - Why squared normal? / why k(>1) variables?
- Test for independence
  - Null: Two factors are independent
  - Alter: Two factors are NOT independent
  - *Expected: (Row value x Column value) / Total

		Heart Rate		
		Low	**High**	**Row Total**
Gen-der	**Girl**	*O = 11 (E = 12.6)*	*O = 7 (E = 5.4)*	*R = 18*
	Boy	*O = 17 (E = 15.4)*	*O = 5 (E = 6.6)*	*R = 22*
Column Total		*C = 28*	*C = 12*	*n = 40*

$$\chi^2 = \sum \frac{(\text{Observed} - \text{Expected})^2}{\text{Expected}}$$

$$\Sigma \frac{(O-E)^2}{E} = \frac{(11-12.6)^2}{12.6} + \frac{(7-5.4)^2}{5.4} + \frac{(17-15.4)^2}{15.4} + \frac{(5-6.6)^2}{6.6} = 1.231$$

Chi-square test of homogeneity

- A test for two categorical variables are homogeneous

Observed Counts (Expected Counts)				
	Music			
Wine	**None**	**French**	**Italian**	**Total**
French	30 (34.22)	39 (30.56)	30 (34.22)	**99**
Italian	11 (10.72)	1 (9.57)	19 (10.72)	**31**
Other	43 (39.06)	35 (34.88)	35 (39.06)	**113**
Total	**84**	**75**	**84**	**243**

  - Are Music and Wine preferences driven by the same hidden characteristics? (In vector space terms?)
- Ex. Doctors vs. Patients
  - For hospital advertisements, doctors and patients responded a common questionnaire with a scale from 1 to 10, and compared the scoring to test whether two groups are homogeneous in responses
  - If they are, then no reason to block hospital advertisements
    - How would the responses change if ads change?

Nature of F test

- A ratio of two chi-squared distribution (A ratio of two squared normal distribution)

Definition of F test

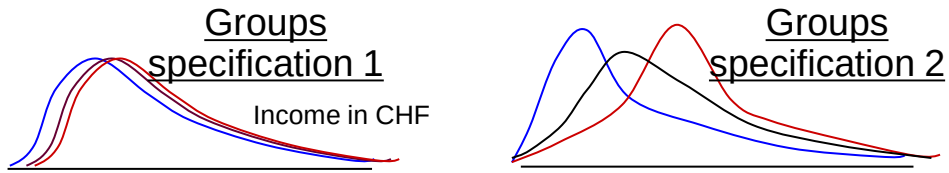
- F-distribution for F-test

$$\frac{X/d_1}{Y/d_2} \sim \mathrm{F}(d_1, d_2) \qquad X \sim \chi^2_{d_1} \text{ and } Y \sim \chi^2_{d_2}$$

  - Where X, Y are independent chi-square distribution
  - To test how X, Y are different (by variance or S.S.)
  - Test of variance btwn two groups of data
  - What happens if the underlying is t distribution?
- Why F test?
  1. 3 t-tests combined vs. 1 F-test
     - At 5% significance level, 3 t-tests will give us total 15% of Type I error – more tests, more errors
  2. Mean is not the right criterion for comparison
     - In finite sample and finite variance for distinct groups, variance contains significant information
- Two use cases
  - Btwn group variance / within-group (ANOVA)
  - Explained variance / Unexplained variance (Chow test)

Group comparison (ANOVA)

- How different are two groups (in variance terms)
  - Btwn: deviation of group from the population (systematic)
  - Within: deviation of individual within group (error)
  - Ex. Income level difference btwn BBA vs. MBA
    - One variable in control to affect the other variable



- Could be a sampling error or group specific hidden issue
  - Similar group – lower btwn variance
  - Distinct group – lower within variance

$$F = \frac{BSS/g-1}{WSS/n-g}$$

H_o: $\mu_1 = \mu_2 = \mu_3 = \mu_4$
H_a: At least one group's mean does not equal the others'

- Post-hoc test to identify which group rejects the null
  - A series of independent t-tests comparing each group
    - Significance level has to be adjusted for the aggregate test = sum of individual tests

Dummy variable regressions and Chow test

- Applying F test for structural (time-dependent) changes

Dummification in regression

- Simple dummy variable
  - Any categorical variable can be turned into a set of dummy variables
  - If n categories, either n-1 dummies with intercept or n dummy variables
  - But if there are too many, better to group them
  - The same logic applies to time variables
    - N periods -> Before/After the critical event
- Interaction variables
  - Like sub-dividing the group/time
    - Ex. Male/Before & After MBA

$$y_i = \beta_1 + \beta_2 D_i + \beta_3 X_i + \beta_4 D_i X_i + u_i$$

  - When D= 0 / 1, the regression will have different intercept and slope

$$y_i = \beta_1 + \beta_2 D_i + \beta_3 X_i + \beta_4 D_i X_i + \beta_5 T_i X_i + u_i$$

  - With T, X's effect can be analyzed with respect to time variation from 0 to 1 (Before/After the event)
- If y is binary, it is Linear Probability Model, but logit, probit perform better

Time comparison (Chow test)

- For a restricted vs. unrestricted, by the number of test conditions (q), we have below F-test

$$F = \frac{(RRSS - URSS)/q}{URSS/(n-k-1)}$$

- What if the restriction is time variable? – Chow test
  - If there were no structural changes, then before/after (U)RSS will be identical to (R)RSS

$$F = \frac{\left[RSS_C - (RSS_1 + RSS_2)\right]/(k)}{(RSS_1 + RSS_2)/(n-2k)}$$

$$y = \begin{pmatrix} X_I & 0 \\ 0 & X_{II} \end{pmatrix} \begin{pmatrix} \beta_I \\ \beta_{II} \end{pmatrix} + \epsilon$$

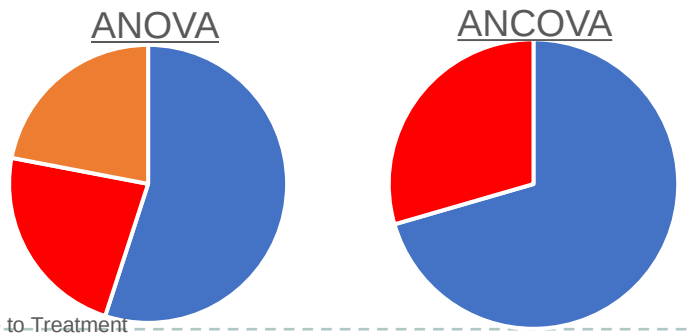
- If identical, both betas will be the same
- What if the restriction is time variable?
  - If there were no structural changes, then RSS_c = RSS_1 + RSS_2
- Chow test is, in fact, a time version of ANOVA

Extensions – ANCOVA

- Where Co-variance matters and why we stick to multivariate regressions

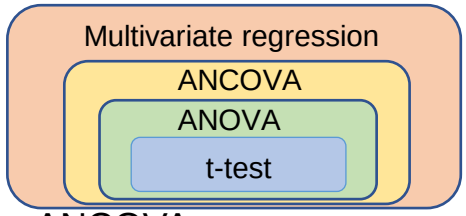
Analysis of CoVariance (ANCOVA)

- What if there is a confounding variable (CV)?
  - Ex. For Group 1, do exercise A, and Group 2, exercise B, both of which raises blood pressure. Then, if Group 1 is elder, would it be the key factor for higher blood pressure?
  - Adjust dependent variable (DV) by DV, then do ANOVA for remaining variables
    - Find common slope by CV for two groups ("Co-Variance"), and adjust the DV
- Types of variance
  - In ANOVA, the covariance generated by a 3rd factor is shadowed, which often results in misleading argument



ANCOVA, a subset of Multivariate regression

- ANCOVA uses "Co-Variance" to adjust the DV, which eliminates the effect of CV



- Conditions for ANCOVA
  1. Normality of residuals
  2. Homogeneity of variances
  3. Homogeneity of regression slopes (by CV)
  4. Linearity of regression
  5. Independence of error terms
  - Condition 4 is GM's A2, and the other conditions are the same as ANOVA
  - ANOVA controlling other variables is ANCOVA
- AN(C)OVA is a subset of multivariate regression
- This is why we only focus on multivariate (w/ or w/o dummy) regressions for ANOVA like tests