207 Xavier Hall Annex seating arrangement

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Tahani Al-Rajeh | Yit Mui Khoo | Nathan House | Laura Gardner | Diana Ramirez |
| Joe Fetter | X | Malcolm Townes | Darren O’Brien | X |
| James Hillis | X | Ashley | Destiny Brooks | Courtney |
| Renee Jonas | Saeed Asiri | X | X | X |

Recommend text: SPSS Survival Manual

Data for final project

* The data set for the project must include nominal, ordinal, and ratio variables and a minimum sample size of n=100.
* Clean the data before using it.

Article review assignment

* The article for the article review assignment must use regression analysis.
* Pair up for the article review assignment and presentation.
* Decide on pairs in the next couple of weeks.

Nominal and ordinal data are both referred to as categorical data.

Mean, median, and mode

* Three moments in statistics
* Measures of central tendency
* We’re not concerned with outliers in statistical analysis

Confidence interval is the level of confidence that the real mean falls between two points.

It’s a good practice to include descriptive statistics in data analysis.

This course is concerned with Type 1 errors; Type 2 errors concern Bayesian statistics.

Significance levels are 0.05 (\*), 0.01 (\*\*), and 0.001 (\*\*\*)

Degrees of freedom = sample size - number of variables

Chi-Square Test

* Useful for nominal variables
* Requires a minimum count of 5 for each variable

Standard coefficient is equivalent to the Pearson correlation only in simple regression

Data set requirements for final project

* Minimum of 5 variables
* Minimum of 3 ratio variables
* Cross-sectional data is fine (i.e., longitudinal data is not necessary)

Options for identifying a group for the journal article review and presentation assignment

* Work with other students in the Public and Social Policy (PSP) program
* Work with Nathan House and Saeed Asiri \*

Technically ordinal data should not be used for regression analysis but is sometime acceptable.

General Social Survey (GSS) is conducted by the National Opinion Research Center (NORC)

Residual is the difference from the target value (e.g., mean value)

Null hypothesis: The percent of male in the sample is the same as the percent of female in the sample

Alternative hypothesis: The pct of male in the sample is NOT the same as the pct of female in the sample

H0: pM - pF and HA = pM ≠ pF

Null hypothesis: The percent of each race in the sample is the same as the percent of each race in the population

Alternative hypothesis: The percent of each race in the sample is the same as the percent of each race in the population H0: white = pwhite , etc. and HA = white ≠ pwhite , etc.

Degrees of freedom for two variables = (n1-1)( n2-1)

Generally, hypotheses are structured as follows

* H0: There is NOT an association
* HA: There is an association

H0: μM = μF 🡪 The population mean of males is the same as the population mean of females

HA1: μM ≠ μF 🡪 The population mean of males is NOT the same as the pop. mean of females (2-tail test)

HA2: μM > μF 🡪 The population mean of males is greater than the pop. mean of females (1-tail test)

To use T-test, the test variable must be a ratio variable and there must only be a total of two variables.

ANOVA is used for testing 3 or more variables

Paired T-test is used for comparing two sets of observations on the same unit of analysis (i.e., before and after tests, twins, married couples, etc.)

Data set for class

* NBER patent citation data is okay, per Dr. Matsuo
  + Use data from 1994-1999
  + Between 100 and 200,000 observations is good
  + Data must have at least three continuous variables, not necessarily ratio variables

Assignment #1

* Due by 12 noon on Thursday, September 20th.

Journal article review and presentation

* Submit chosen article to Dr. Matsuo for pre-approval

The more objective the data the more likely to get a high R2 value if a strong correlation exists.

In statistics, the term “parsimonious” means economical.

Sample distribution is NOT the same as sampling distribution.

Example…

Population, N=10,000

Sample size, n=100

Number of samples possible (of size n=100), k=100 (i.e., 10,000 / 100)

Sample means are : = V1 … = Vk

Sample distribution is the distribution of x values for each sample (i.e., frequencies as in a histogram)

Sampling distribution is the distribution of the sample means, which should be approximately normally distributed according to the Central Limit Theorem

Standard error = standard deviation of the sampling distribution

Ideally, errors (residuals) should not correlate with each other.

Autocorrelation is when the residuals of a single variable correlate with one another.

Autocorrelation is about how the residual of one data point influences the residual of other data points in a sample for one variable.

Multicollinearity is about whether there is a linear relationship between two independent variables.

Standardized coefficient beta

* Measures the increase in the contribution of a variable that results from a one standard deviation change in the value of the variable.
* |t-value| ≥ 2.0 🡪 likely significant

Types of variables appropriate for regression analysis

* Can’t use nominal variables in regression analysis.
* Can possible use ordinal variables in regression analysis.

ANOVA

* df1 🡪 number of parameters
* df2 🡪 number of

The difference between concepts and constructs

* Concepts are theoretical
* Constructs are measureable

Four types of validity

* Content
* Face
* Predictive/Concurrent
* Construct

Collinearity statistics

* Tolerance = 1 🡪 no multicollinearity
* Tolerance = 0 🡪 high multicollinearity

Hetreroscedasticity

* Regression standardized residuals get wider or narrower as the value of X increases.

Correlations

* Pearson correlation
  + Zero order correlation
  + The influence of other independent variables is NOT removed.
* First order partial correlation
  + The influence of one independent variable is removed.
* Second order partial correlation
  + The influence of two independent variables is removed.
* Higher order partial correlation
  + The influence of three or more independent variables is removed.

Absent from class. No notes taken.

Adjusted R2 is adjusted for the sample size.

Significant regression 🡪 model shows a good fit

Use natural log (ln) rather than log10

Presentation 01

* Group meeting on Friday, October 5, 2018 at 3:00 PM at PIUS XII Library
  + Saeed Asiri to reserve room
* Basic division of presentation
  + Introduction, literature review, theoretical framework 🡪 Malcolm Townes
  + Data and Methods 🡪 Saeed Asiri
  + Results 🡪 Nathan House

Assignment02

* Provide output file to Dr. Hisoku Matsuo by Wednesday, October 10, 2018 at 4:15 PM
* Assignment due by Thursday, October 11, 2018 by 12:00 PM (i.e., noon)

Dr. Matsuo also generally available on Friday mornings and Saturday evenings.

Assignment 02

* Due by Friday, October 12, 2018 at 12:00 PM (noon)
* The write up doesn’t have to be as extensive as Assignment 01

Path models

* Can’t use the enter method
* Must create independent regression models and combine
* Use the standardized coefficient

Unstandardized coefficients

* 1 unit change in the independent variable produces X change in the dependent variable.

Standardized coefficients

* 1 standard deviation change in the independent variable produces X change in the dependent variable

Residuals =

Class on October 24, 2018

* No class on because of instructor schedule conflict
* Office hours after 5 pm for those that have questions (1905 Morrissey)
* Due date for Assignment 03 postponed until Week 10, October 31

Presenting and Interpreting Hierarchical Regression

* Discuss any change in sign
* Discuss any change in significance level
* Show coding of dichotomous variables
* Show correlation table for manuscripts
* Include labels for variable names (include in SPSS output file)
* Use proportions instead of raw values where appropriate (i.e., normalize data when necessary)
  + Using percentages may increase multicollinearity
* Significance level 🡪 chance of making Type I error (i.e., false positive; rejecting null hypothesis when it is in fact true)

Assignment03

* General process for preparing manuscripts
  + Develop table first
  + Convert table data into words
* Recommended process for preparing Assignment03 manuscript
  + Develop model first
  + Test model
  + Revise model
  + Convert model into words (i.e., explain the model)
* Rule of thumb is a minimum of 20 observations per arrow for causal pathway models
* Can’t combine logistic regression and multiple regression in same analysis
  + Dichotomous variables require logistic regression analysis
  + Dichotomous variable can only be used in the first series of variables in causal pathway model (i.e., not in the model of the model)

Class canceled. No notes.

Assignment 03

* Due between Thursday, November 1, 2018, 12:00 PM and Friday, November 2, 2018, 5:00 PM

Solutions for Missing Data

* Use pair-wise analysis rather than list-wise analysis
  + List-wise analysis 🡪 observation is dropped if data is missing in at least one of the variables of interest
  + Pair-wise analysis 🡪 observation with missing data in certain variables may still be used to analyze other variables where the data is not missing
* Imputing data
  + If using another dataset to impute missing data for variables in the primary data one is using, the other data should have similar characteristics to the primary data or it may cause problems.
    - e.g., primary data of interest has 40% African-American respondents but data used to impute missing data only had 10% African-Americans respondents
* Convert the variable of interest into a dummy variable
  + Answered = 1, Did Not Answer = 0

Power Analysis

* Used to determine sample size
* Must consider type of statistical analysis
* Statistical power 🡪 the likelihood of rejecting null hypothesis if the null hypothesis is false
  + The higher the statistical power the better
  + Different from significance level which is the probability of rejecting the null hypothesis if the null hypothesis is true (i.e., Type 1 error)

Solutions for Categorical Variables

* Chi square tests association between only two variables
* Log-linear describes association among three or more variables

Proposed final presentation schedule

* 8 presentations on November 28, 2018
* 4 presentations on December 5, 2018

Binary logistic regression

* Model fit test is calculated using chi-square
* Always comparing something against something

Odds ratio

* If calculation is less than 1, take inverse to interpret odds ratio (OR)

Use of chi-square test

* Test contingency table with two categorical variables
  + e.g., social class and political affiliation
  + e.g., sex and voting status
  + H0: There is NO association between the two variables
  + Degrees of freedom (d.f.) is number of categories for the first variable minus one times the number of categories in the second variable minus one [i.e., (n-1)(k-1)]
  + Requires minimum of 5 observations per cell
* Test data goodness-of-fit
  + Analogous to F-test in ANOVA
  + H0: There is equal proportion of variable 1 and variable 2 in the data
  + If chi-square is small (p-value is large), do NOT reject null hypothesis

Binary logistic regression

* Probability of 0.5 corresponds to logit of zero.
* Probability greater than 0.5 corresponds to a positive logit.
* Probability less than 0.5 corresponds to a negative logit.
* Coefficient of zero means probability does NOT change as independent variable changes.
  + Not simple slope
* Significance of independent variable is determined by Wald scale.
* Exp(B) shown in SPSS output is used for odds ratio.
* If variable is significant, the confidence interval for Exp(B) will NOT contain 1.0
* Model fit determined by Omnibus test and Hosmer-Lemeshow test.
  + Omnibus test: Chi-square = (-2LLInitial) – (-2LLFinal)
    - Looking for large chi-square and small p-value
  + Hosmer-Lemshow test
    - Looking for small chi-square and large p-value
* H0: Model = Data 🡪 Goodness-of-fit of the model
* In logistic regression, likelihood ratio (LR) chi-square is multiplied by -2log   
  (i.e., -2 log likelihood[-2LL])
  + Looking for small -2LL
* For continuous independent variable, use the beta coefficient (β) to interpret the relationship rather than the odds ratio.
* SPSS cutoff point based on frequencies.

Assignment 04

* Due Thursday, November 14, 2018 at 12:00 pm noon but no later than Friday November 15, 2018 at 5:00 PM.
* Does not have to be very long.

Presentation 02 Order

* First group on November 28, 2018
  + Yit Mui Khoo
  + Destiny Brooks
  + Renee Jonas
  + Darren O’Brien
  + Nathan House
  + Saeed Asiri
  + Malcolm Townes
  + Laura Gardner
* Second group on December 05, 2018
  + Tahani Al-Rejeh
  + Diana Ramirez
  + James Hillis
  + Joe Fetter
  + Courtney
  + Ashley

Binary Logistic Regression

Classification cutoff (in SPSS)

* Affects the value of the coefficients.
* It should correspond to the percentage of the category of interest.

Hosemer-Lemeshow Test

* H0 is the model is a good fit for the data.
* Small chi-square with large p-value is desired to not reject H0 (i.e., reject HA).
* If that’s the case, the model is a good fit.

Omnibus Test of Model Coefficients

* Comparing the difference between the -2 Log Likelihood without the additional variables included from the last iteration and the -2 Log Likelihood with the variables included.
* H0 is the variables DO NOT improve the model fit.
* A small (or reduced) -2 Log Likelihood and a small p-value is desired to reject H0.
* If that’s the case, the addition of the variables improved the model fit.