PSY2014

Henriette Johansen

UiO

PSY2014

Kvantitativ metode



# 1 Introduction

|  |  |  |
| --- | --- | --- |
| Page | Term | Description |
| 14 | Data |  |
| 15 | Databases |  |
| 16 | The general social survey (GSS) |  |
| 15 | Statistics | Consists of a body of methods for obtaining and analysing data.  Provides methods for:  1. design: planning how to gather data from a research study to investigate questions of interest to us.  2. description: summarising the data obtained in the study.  3. inference: making predictions based on the data, to help us deal with uncertainty in an objective manner. |
| 16 | Descriptive statistics | Use it to reduce the data to a simpler and more understandable form without distorting or losing much information.  Summarises the information in the collection of data. |
| 16 | Inference | Using the data to make predictions. |
| 16 | Statistical inferences | Predictions made using the data. |
| 16 | Subjects |  |
| 17 | Population | Total set of subjects of interest in a study. |
| 17 | Sample | The subset of the population on which the study collects data. |
| 17 | Inferential statistics | Provide predictions about a population, based on the data from a sample of that population. |
| 17 | Parameter | The corresponding numerical summary for the population. Numerical summary of the population. |
| 19 | Data file |  |

# 7 Comparison of two groups

|  |  |  |
| --- | --- | --- |
| Page | Term | Description |
| 191 | Binary variable |  |
| 191 | Bivariate statistical methods |  |
| 191 | Respons variable | An outcome variable about which comparisons are made |
| 191 | Explanatory variable | Variable defining the group |
| 192 | Longitudinal study |  |
| 192 | Dependent samples |  |
| 192 | Independent samples |  |
| 192 | Cross-sectional study |  |
| 193 |  |  |
|  |  |  |

# 9 Linear regression and correlation

|  |  |  |
| --- | --- | --- |
| Page | Term | Description |
| 259 | Regression analysis |  |
| 259 | Linear relationships |  |
| 259 | Response variable |  |
| 259 | Explanatory variable |  |
| 260 | Linear function |  |
| 260 | y-intercept | alpha |
| 260 | Slope | beta |
| 260 | Regression coefficients |  |
| 262 | Positive relationship |  |
| 262 | Negative relationship |  |
| 262 | Model | Simple approcimation for the relationship between variables in the population. |
| 262 | Scatterplot |  |
| 264 | Prediction equation |  |
| 265 | Regression outlier |  |
| 265 | Influential observation | If removing it results in a large change in the prediction equation. |
| 266 | Residuals | Predicting errors  For an observation, the difference between an observed value and the predicted value of the response variable, |
| 267 | Sum of squares errors (SSE) |  |
| 267 | Least square estimates |  |
| 267 | Least squares line |  |
| 267 | Residual square of sums |  |
| 268 | Deterministic |  |
| 268 | Conditional distribution |  |
| 268 | Probabilistic model |  |
| 268 | Expected value of y |  |
| 268 | Regression function |  |
| 268 | Regression coefficients |  |
| 269 | Conditional standard deviation |  |
| 270 | Residual mean square | Its square root is the estimate of the conditional standard deviation of y. |
| 270 | ANOVA table |  |
| 270 | Degrees of freedom (df) |  |
| 270 | Residual standard error | For the square root of the mean square error |
| 271 | Standard deviation | . To reflect its conditional form, the conditional distribution, that SD is sometimes denoted as |
| 271 | Total sum of squares (TSS) | The sum of squares in the numerator of |
| 271 | Marginal distribution | Shows the overall variability in y-values |
| 271 | Conditional distribution | Shows how y varies at a fixed value of x |
| 272 | The slope *b* of the prediction equation | Tells us the direction of the association.  The slope *b* doesn’t directly indicate whether the association is the same in each case, because we can make *b* as large as we like by an appropriate choice of units.  The slope is useful for comparing effects of two predictors having the same units. |
| 272 | The strength of the association | Is the same in each case if the variables and data are the same |
| 272 | Correlation | The correlation is a standardised version of the slope.  The correlation is useful for comparing associations for variables having different units.  Correlation describes linear relationships.  Correlation implies regression toward the mean. |
| 273 | Standardised regression coefficient |  |
| 273 | Pearson correlation |  |
| 273 | Properties of the correlation | * Valid only when a straight-line model is sensible for the relationship between x and y. since r is the proportional to the slope of a linear prediction equation, it measures the strength *of the linear association*. * . The correlation, unlike the slope *b*, must fall between -1 and +1. * r has the same sign as the slope *b*. * r=0 for those lines having b=0. * when all the sample points fall exactly on the prediction line. * The larger the absolute value of r, the stronger the linear association. * The correlation, unlike the slope *b*, treats x and y symmetrically. * The value of r does not depend on the variables’ units. |
| 275 | Regression toward the mean |  |
| 276 | The mean of association | Has four elements:   * Rule 1: A rule for predicting without using x. * Rule 2: A rule for predicting y using information about x. * Prediction errors: A summary measure of prediction error for each rule. * The difference in the amount of error with the two rules is . Converting this reduction in error to a proportion provides the definition below. |
| 276 | Proportional reduction in error |  |
| 276 | Sum of squared errors/residual sum of squares |  |
| 277 | Definition of measure |  |
| 277 | r-squared/coefficient of determination |  |
| 277 | Properties of r-squared | * Since , r^2 falls between 0 and 1. * The minimum possible value for SSE is 0, in which case r^2 = TSS/TSS = 1. * When the least squares slope *b*=0, the y-intercept *a* equals * Like the correlation, r^2 measures the strength of linear association. * r^2 does not depend on the units of measurement, and it takes the same value when x predicts y as when y predicts x. |
| 278 | Assumptions for statistical inference | * Randomisation, such as a simple random sample in a survey. * The mean of y is related to x by the linear equation . * The conditional standard deviation is identical at each x-value. * The conditional distribution of y at each value of x is normal. |
| 280 | Correlation matrix |  |
| 283 | P-value | A small P-value for suggests that the regression line has a nonzero slope. |
| 283 | Confidence interval for |  |
| 283 | t-score | The value, with df=n-2, for the desired confidence level. |
| 284 | Sums of squares (SS) |  |
| 284 | Regression sum of squares/model sum of squares | TSS – SSE. This difference is the numerator of the measure. It represents the amount of the total variation TSS in y that is explained by x in using the least squares line. The ratio of this sum of squares to TSS equals . |
| 285 |  | Need not to correspond to independence if the assumption of a linear regression model is violated. For this reason, you should always construct a scatterplot to check this fundamental assumption. |
| 285 | Normality assumption | Is relatively unimportant |
| 285 | Central limit theorem | Implies that sample slopes and correlations have approximately normal sampling distributions. |
| 286 | Factors influencing the correlation | * Outliers * The range of x-values sampled (when a sample has a much narrower range of variation in x than the population, the sample correlation tends to underestimate drastically (in absolute value) the population correlation) |
| 288 | Error term () | Represents the error that results from using the mean value of y at a certain value to x to predict the individual observation. |
| 288 | Population residual () | The difference between the observation y and the mean of all possible observations on y at that value of x. |
| 289 | The linear regression equation | describes the form of the relationship. |
| 289 | Scatterplot |  |
| 289 | The least squares | Estimates if the y-intercept and the slope provide the prediction equation closest to the data, minimising the sum of squared residuals. |
| 289 | Correlation r | And its square describes the strength of the linear association. |
| 289 | Null hypothesis of independence |  |
| 289 | Multiple regression model | A generalisation that permits several explanatory variables in the model. |
| 289 | Categorical predictors |  |
| 289 | Quantitative predictors |  |
| 290 | Categorical response variables |  |

# 10 Introduction to multivariate relationships

|  |  |  |
| --- | --- | --- |
| Page | Term | Description |
| 299 | Statistical control |  |
| 299 | Causal relationships | x 🡪 y |
| 300 | Causal relati0nship criteria | * Association between variables * An appropriate time order * The elimination of alternative explanations |
| 300 | Association | Association does not imply causation. |
| 300 | Time order |  |
| 301 | Alternative explanation |  |
| 301 | Observational studies | Can never prove that one variable is a cause of another. |
| 301 | Anecdotal evidence |  |
| 302 | Randomised experiment | Inherently control other variables in a probabilistic sense. |
| 302 | Controlled variable | When its influence is removed.  We control a variable by holding its value constant. |
| 303 | Statistical control |  |
| 303 | Experimental control |  |
| 304 | Control variable |  |
| 304 | Partial tables | Separate tables that display the relationships within the fixed levels of the control variable. |
| 304 | Bivariate table | Contains data only on two variables. All other data are ignored. None is controlled. |
| 305 | Sample variation |  |
| 305 | Lurking variable |  |
| 306 | Spurious associations |  |
| 307 | Socioeconomic status (SES) |  |
| 307 | Chain relationships |  |
| 307 | Intervening (mediator) variables |  |
| 308 | Multiple causes |  |
| 308 | Independent causes |  |
| 309 | Suppressor variable | Occasionally two variables show no association until a third variable is controlled. That controlled variable is the suppressor variable. |
| 310 | Statistical interaction |  |
| 311 | Simpson’s paradox | After controlling for a variable, each association in a partial table has the opposite direction as the bivariate association. |
| 311 | Confounding variable |  |
| 311 | Omitted variable bias |  |
| 312 | Categorical control variables |  |
| 313 | Partial association |  |

11 Multiple regression and correlation

|  |  |  |
| --- | --- | --- |
| Page | Term | Description |
| 319 | Multiple regression model | Can have multiple explanatory variables. |
| 320 | Statistical interaction |  |
| 320 | Bivariate model | Model with a single predictor |
| 320 | Multiple regression function | More difficult to portray graphically than the bivariate regression function. |
| 321 | Simpson’s paradox | The partial association has the opposite direction from the bivariate association. |
| 322 | Multiple regression | A slope describes the effect on an explanatory variable while *controlling* effects of the explanatory variables in the model. |
| 322 | Partial regression coefficients |  |
| 323 | Statistical interaction |  |
| 324 | Prediction equation |  |
| 324 | Residuals | Predicton errors |
| 324 | Sum of squared errors (SSE) | Also called residual sum of squares (RSS) |
| 324 | The least squares criterion | The prediction equation has the *smallest* SSE value of all possible equations of the form |
| 325 | Scatterplot matrix |  |
| 326 | Partial regression plot (added-variable plot) | Displays the relationship between the response variable and an explanatory variable after removing the effects of the other predictors in the multiple regression model. |
| 329 | Multiple correlation | The sample multiple correlation for a regression model, denoted by R, is the correlation between the observed y-values and the predicted -values.  The larger value of R the better the predictions of y by the set of explanatory variables.  The larger value of R the better  The predicted values cannot correlate with negatively with the observed values. |
| 329 | R-squared: The coefficient of multiple determination | The square of the multiple correlation *R*.  Uses the proportional reduction in error concept, generalising for bivariate models.   * Rule 1 * Rule 2 * Prediction Errors * Definition of Measure: The proportional reduction in error from using the prediction equation instead of to predict y is R-squared.   Properties of are similar to . |
| 331 | Adjusted | Less biased estimate |
| 331 | Regression sum of squares |  |
| 331 | Multicollinearity | When does not increase much, this does not mean that the additional variables are uncorrelated with y. It means merely that they don’t add much new power for predicting y, given the values of the explanatory variables already in the model.  The explanatory variables are highly correlated, no one having much unique explanatory power. |
| 333 | F-distribution | For hypotheses about the p predictors, the test statistic is:  Like the chi-squared distribution, the *F* distribution can take only nonnegative values and it is somewhat skewed to the right.  The mean of the F distribution is approximately equal to 1. The larger the -value, the larger the ratio |
|  |  |  |
| 336 | Estimation of |  |
| 337 | F test statistic | Alternative formula for the F test statistic:  This gives us the same value as the F test statistic formula based on . |
| 337 | Interaction | For quantitative variables, interaction exists between two explanatory variables in their effects on y when the effect of one variable changes as the level of the other variable changes. |
| 338 | Cross-product terms |  |
| 340 | Centring | Centring the explanatory variables before using them in a model allowing interaction has two benefits.   * The estimates of the effects of and are more meaningful, being effects at the mean rather than at 0. * The estimates and their standard errors are similar as in the no-interaction model. |
| 341 | Complete model | Full model with all the explanatory variables. |
| 341 | Reduced model | Model containing only some of these variables. The reduced model is said to be nested within the complete model, being a special case of it.  They are identical if the partial regression coefficients for the extra variables in the complete model all equal 0. |
| 342 | Test statistic for comparing two regression models |  |
| 343 | T test method | Limited to testing one parameter at a time. The F test can test several regression parameters together to analyse whether at least one of them is nonzero. F tests are equivalent to t tests only when contains a single parameter. |
| 343 | Partial correlation | Based on the ordinary correlations between each pair of variables.  Has properties similar to those of the ordinary correlation between two variables. |
| 343 | First order partial correlations |  |
| 344 | Proportional reduction in error (PRE) |  |
| 344 | Squared partial correlation |  |
| 346 | Second-order partial correlation | Controls two variables |
| 346 | Standardised regression coefficient | Represents the values the regression coefficient can tak when the units are such that y and the explanatory variables all have equal standard deviations, such as when we use standardised variables.  We can obtain the standard regression coefficients from the unstandardised coefficients. |
| 347 | Beta weights |  |

5 Statistical Inference: Estimation

|  |  |  |
| --- | --- | --- |
| Page | Term | Description |
| 115 | Confidence interval |  |
| 115 | Maximum likelihood |  |
| 115 | Bootstrap |  |
| 115 | Point estimate (estimate) | A single number that is the best guess for the parameter value. |
| 115 | Interval estimate (confidence interval) | An interval for numbers around the point estimate that we believe contains the parameter value.  Helps us gauge the precision of a point estimate. |
| 115 | Margin of error |  |
| 115 | Estimator | Refers to a particular type of statistic for estimating a parameter and estimate refers to its value for a particular sample. |
| 116 | Unbiased estimator | If its sampling distribution centres around the parameter. |
| 116 | Biased estimator | Tends to underestimate the parameter, on the average, or it tends to overestimate the parameter. |
| 116 | Efficient estimator | Estimator having standard error that is smaller than those of the other estimators. |
| 117 | Maximum likelihood |  |
| 117 | Confidence interval | An interval of numbers within which the parameter is believed to fall. |
| 117 | Confidence level | The probability that this method produces an interval that contains the parameter. |
| 117 | Margin of error | Form of confidence interval: Point estimate margin of error |
| 118 | Proportions of observations in the categories |  |
| 118 | Central limit theorem |  |
| 119 | 95% confidence interval |  |
| 123 | The width of a confidence interval | * Increases as the confidence level increases * Decreases as the sample size increases |
| 123 | Error probability | Error probability = 1 – confidence level |
| 126 | t-score | When n is small, the error can be sizeable and to account for this increased error, we must replace the z-score by a slightly larger score with a wider confidence interval. |
| 126 | t-distribution | The t-score comes from a bell-shaped distribution that is slightly more spread out than the standard normal distribution. |
| 126 | Degrees of freedom |  |
| 129 | Confidence interval for population mean |  |
| 135 | Sample size for estimating a population proportion | You need to guess or take the safe approach of setting . |
| 135 | Sample size for estimating a population mean | You need to guess the population standard deviation . |
| 136 | Complexity of analysis |  |
| 138 | Artificial observations |  |
| 138 | Estimation methods | * Maximum likelihood * Bootstrap |
| 138 | Maximum likelihood methods of estimation |  |
| 139 | Maximum likelihood estimate | The value of the parameter that is most consistent with the observed data, in the following sense: if the parameter equalled that number, the observed data would have greater chance of occurring than if the parameter equalled any other number. |
| 139 | Likelihood function |  |
| 139 | Maximum likelihood estimators | Three desirable properties:   * Efficient * Consistent * Approximately normal sample distributions |
| 140 | Bootstrap | This method treats the sample distribution as if it were the true population distribution and approximates by simulation the unknown sampling distribution. |
| 143 | True standard error |  |

6 Statistical inference: Significance Tests

|  |  |  |
| --- | --- | --- |
| Page | Term | Description |

13 Statistical Inference: Significance Tests

|  |  |  |
| --- | --- | --- |
| Page | Term | Description |
| 399 | Covariate |  |
| 399 | Analysis of covariance |  |
| 399 | Linear mixed model |  |
| 399 | Random effects |  |
| 400 | No interaction |  |
| 400 | Equality of slopes | The regression lines are parallel. |
| 406 | Inference |  |
| 407 | Test for individual partial effects |  |
| 413 | Adjusted means (least squares means) | Represents the expected value for y at the means of the explanatory variables for the combined population. |
| 414 | Sample adjusted mean of y | …for a group is the prediction equation for that group evaluated at |
| 415 | Unadjusted mean |  |
| 415 | Overall mean |  |
| 416 | Bonferroni method/Bonferroni multiple comparison approach | Compares all pairs of means simultaneously with a fixed confidence level. This method extends directly to multiple comparison if adjusted means. |
| 416 | Covariance matrix |  |
| 418 | Linear mixed model |  |
| 418 | Mixed effects models |  |
| 418 | Random effect | The coefficient of a dummy variable represents a random effect for a particular subject.  We treat the random effects as unobserved random variables rather than as parameters. |
| 418 | Fixed effects | Ordinary parameters. |
| 418 | Linear mixed models |  |
| 418 | Linear adjective | Refers to the effects in the right-hand side of the regression model equations having an additive rather than multiplicative structure. |
| 418 | Dummy variables | When some explanatory variables are categorical, we use dummy variables for them. |
| 418 | Error term |  |
| 419 | Linear mixed model | The random effects are not restricted to individual subjects. they can represent clusters of subjects that are similar in some way.  Unlike ordinary repeated-measures ANOVA, it permits some observations to be missing. For any particular subject, it uses the available data. Depending on what causes the observations to be missing, resulting estimates of fixed effects may or may not be unbiased. |
| 419 | Random intercept |  |
| 419 | Compound symmetry |  |
| 419 | Autoregressive structure |  |
| 419 | Unstructured structure | Makes no assumptions about the correlation pattern. |
| 419 | Model-fitting process |  |
| 419 | Assumed correlation structure | Implies correlations among the repeated responses. |
| 419 | Intraclass correlation |  |
| 421 | ANOVA for repeated measurement of subject |  |
| 421 | Sphericity structure |  |
| 421 | Random slope |  |
| 422 | Assumption of “missing at random” |  |
| 422 | Missing at random | The probability an observation is missing does not depend on the value of the unobserved response. |

14 Model building with multiple regression

|  |  |  |
| --- | --- | --- |
| Page | Term | Description |
| 431 | Multicollinearity |  |
| 431 | Generalised linear model |  |
| 431 | Gamma distribution |  |
| 431 | Exponential increase/decrease |  |
| 432 | General guidelines for selecting explanatory variables | * Include the relevant variables to make the model useful for theoretical purposes, so you can address hypotheses posed by the study, with sensible control and mediating variables * Include enough variables to obtain a good predictive power * Keep the model simple |
| 432 | Most popular automated variable selection methods | * Backward elimination * Forward selection * Stepwise regression |
| 432 | Backward elimination | Begins by placing all of the explanatory variables under consideration in the model. It deletes one at a time until reaching a point where the remaining variables all make significant partial contributions to predicting y. tha variable deleted at each stage is the least significant, having the largest P-value in the significance test for its effect.  Explanatory variable, controlled for other variables, with the biggest P-value is removed. |
| 434 | Forward selection | Begins with none of the potential explanatory variables in the model. Adds one variable at a time to the model until no remaining variable not yet in the model makes a significant partial contribution to predicting y.  Explanatory variable with the smallest P-value is added. |
| 434 | Stepwise regression | A modification of forward selection that drops variables from the model if they lose their significance as other variables are added. |
| 436 | Explanatory research | Has a theoretical model to test using multiple regression. Automated selection procedures are usually not appropriate, because theory dictates which variables are in the model. |
| 436 | Exploratory research | Has the goal of not examining theoretically specified relationships but merely finding a good set of explanatory variables. This approach searches for explanatory variables that give a large R-squared. |
| 437 | Cross-validation |  |
| 437 | Predicted residual sum of squares (PRESS) | The smaller the value of PRESS, the better the predictions tend to be, in a summary sense. According to this criterion, the best-fitting model is the one with the smallest value of PRESS. |
| 437 | Akaike information criterion (AIC) | Attempts to find a model for which the tend to be the closest in an average sense. The best model is the one with the smallest AIC. |
| 438 | Regression diagnostics |  |
| 438 | Homoscedasticity | The conditional distribution of y has constant standard deviation throughout the range of values of the explanatory variables. |
| 438 | Studentised residual |  |
| 442 | Time series |  |
| 442 | Longitudinal data |  |
| 442 | Observation’s influence | Depends on two factors:   * How far its y-value falls from the overall trend in the data * How far the values of the explanatory variables fall from their means. |
| 442 | Leverage of the observation | Formula for the leverage in bivariate model, the leverage for observation I simplifies to: |
| 442 | DFFIT | Summarises the effect on the fit of deleting the observation. |
| 442 | DFBETA | Summarises the effect on the model parameter estimates of removing the observation from the data set. |
| 443 | Cook’s distance |  |
| 445 | Multicollinearity/collinearity | Causes inflated standard errors for estimates of regression parameters. |
| 445 | Variance inflation |  |
| 445 | Variance inflator factor (VIF) | Represents the multiplicative increase in the variance (squared standard error) of the estimator due to being correlated with the other explanatory variables. |
| 446 | Remedial measures | Can help reduce the effects of multicollinearity. |
| 447 | Factor analysis | Method for creating artificial variables from the original ones in such a way that the new variables can be uncorrelated. |
| 447 | Generalised linear models (GLM) | Generalises ordinary regression in two ways:   * Y can have a distribution other than the normal. * It can model a function of the mean. |
| 448 | Distributions for y | * Poisson * Negative binominal |
| 448 | Gamma distribution |  |
| 448 | Link function | Links the mean of the response variable to positive numbers. |
| 448 | Log link |  |
| 448 | Loglinear models |  |
| 448 | Logit link |  |
| 448 | Logistic regression model |  |
| 448 | Identity link |  |
| 448 | “nonnormal” data |  |
| 449 | Maximum likelihood estimation method |  |
| 449 | Weighted least squares | Gives more weight to observations over regions that show less variability. |
| 449 | Heteroscedasticity |  |
| 449 | Gamma distributions |  |
| 449 | Scale parameter |  |
| 450 | Shape parameter |  |
| 451 | Polynominal regression | Includes a diverse set of functional patterns, including straight lines. |
| 451 | Polynominal regression function |  |
| 451 | Degree of the polynominal function |  |
| 452 | Quadratic regression model |  |
| 452 | Cubic function |  |
| 452 | Convex functions |  |
| 452 | Concave functions |  |
| 455 | Exponential regression |  |
| 455 | Parsimony |  |
| 456 | Exponential regression function | Has the form |
| 458 | Antilog |  |
| 459 | Multiplicative change | The parameter in the mean of y for a one-unit increase in x. |
| 459 | Additive change | The parameter in the linear model |
| 460 | Robust variances estimates |  |
| 461 | Sandwich estimate |  |
| 461 | Robust standard error estimate |  |
| 461 | Generalised estimating equations (GEEs) |  |
| 461 | Random effects |  |
| 461 | Generalised linear mixed model |  |
| 461 | Nonparametric regression |  |
| 461 | Generalised additive modelling |  |
| 462 | Nonparametric smoothing methods | * LOESS * Kernel methods |

8 Analysing association between categorical variables

|  |  |  |
| --- | --- | --- |
| Page | Term | Description |
| 227 | Association |  |
| 227 | Statistical independence | A type of lack of independence in a population. |
| 227 | Chi-squared test | Determining whether two categorical variables are statistically independent or associated. |
| 227 | Residual analysis |  |
| 227 | Contingency tables | Dispays the number of subjects observed at all combinations of possible outcomes for the two variables. |
| 228 | Marginal distributions |  |
| 228 | Conditional distributions |  |
| 228 | Joint distribution |  |
| 229 | Statistical independence | Two categorical variables are statistically independent if the population conditional distributions on one of them are identical at each category of the other. |
| 229 | Statistical dependence | Id the conditional distributions are not identical. |
| 230 | Chi-squared test of independence |  |
| 230 | Expected frequencies |  |
| 230 | Observed frequencies |  |
| 231 | Multinominal distribution | Generalises the binominal distribution from two categories to several categories. |
| 232 | Chi-squared probability distribution (Pearson chi-squared statistic) | Relates to z statistic by  Main properties:   * It is concentrated on the positive part of the real line. * It is skewed to the right * The precise shape of the distribution depends on the degrees of freedom (). |
| 236 | Fisher’s exact test |  |
| 236 | Homogenous response variable |  |
| 236 | Test of homogeneity | The chi-squares test of independence is often referred to as this. |
| 236 | The five parts of the chi-squared test of independence | 1. Assumptions 2. Hypotheses 3. Test statistic 4. P-value 5. Conclusion   The chi-squared test of independence tells us nothing about the nature or strength of the association. |
| 237 | Residual analysis |  |
| 237 | Residual |  |
| 237 | Standardised residual |  |

12 Regression with categorical predictors: analysis of variance methods

|  |  |  |
| --- | --- | --- |
| Page | Term | Description |
| 363 | Analysis of variance (ANOVA) | Inferential method for testing equality of several means. |
| 363 | Repeated-measures ANOVA |  |
| 363 | Dummy variables |  |
| 368 | Confidence intervals for pairwise comparisons of means | A confidence interval for is  In this formula, is the residual mean square in the regression model for *g* groups. The t-value for the chosen confidence level has df=N-g. |
| 369 | Multiple comparison methods | They fix the probability that *all* intervals contain the true differences of population means *simultaneously*, rather than individually. |
| 369 | Multiple comparison error rate |  |
| 369 | Bonferroni multiple comparison method |  |
| 370 | Tukey method |  |
| 370 | Studentized range |  |
| 370 | Analysis of variance (ANOVA) | A test of independence between the quantitative response variable and the categorical explanatory variable that defines the groups. |
| 371 | Between-groups estimate |  |
| 371 | Within-groups estimate |  |
| 371 | F-test statistics |  |
| 372 | Between-group sum of squares (SS) |  |
| 372 | Within-groups SS | Residual SS (denoted SSE) |
| 373 | Within-group sum of squares |  |
| 373 | Between-group sum of squares |  |
| 374 | Total sum of squares |  |
| 374 | Group sum of squares |  |
| 374 | Kruskal-Wallis test | Uses only the ordinal information in the data. |
| 375 | Two-way ANOVA |  |
| 375 | One-way ANOVA |  |
| 375 | Main effects |  |
| 375 | Mean square (MS) | The test statistic is the ratio of the mean square: |
| 376 | Residual MS |  |
| 380 | Partial sum of squares / type III sums of squares |  |
| 381 | Factorial ANOVA |  |
| 381 | Repeated-measures of variance |  |
| 383 | Sphericity |  |
| 384 | Compound symmetry | Special case of sphericity |
| 384 | Mauchly’s test |  |
| 384 | Greenhouse-Geisser adjustment |  |
| 384 | Blocking |  |
| 384 | Random effect |  |
| 384 | Fixed effect |  |
| 386 | Within-subjects factor |  |
| 386 | Between-subjects factor |  |
| 386 | Crossed subjects |  |
| 386 | Nested subjects |  |
| 390 | MANOVA | Multivariate analysis of variance |
| 390 | Wilks’ lambda |  |
| 390 | Likelihood-ratio test |  |
| 390 | Mixed model |  |