DOWNLOAD PDF REGRESSION METHODS

Chapter 1 : Linear regression - Wikipedia

It is the most common method used for fitting a regression line. It calculates the best-fit line for the observed data by minimizing the sum of the squares of the vertical deviations from each data point to the line.

Regression Methods Regression is used for forecasting by establishing a mathematical relationship between two or more variables. We are interested in identifying relationships between variables and demand. If we know that something has caused demand to behave in a certain way in the past, we would like to identify that relationship so if the same thing happens again in the future, we can predict what demand will be. For example, there is a relationship between increased demand in new housing and lower interest rates. Correspondingly, a whole myriad of building products and services display increased demand if new housing starts increase. The rapid increase in sales of VCRs has resulted in an increase in demand for video movies. The simplest form of regression is linear regression, which we used previously to develop a linear trend line for forecasting. Now we will show how to develop a regression model for variables related to demand other than time. Linear Regression Linear regression is a mathematical technique that relates one variable, called an independent variable, to another, the dependent variable, in the form of an equation for a straight line. A linear equation has the following general form: Because we want to use linear regression as a forecasting model for demand, the dependent variable, y, represents demand, and x is an independent variable that causes demand to behave in a linear manner. To develop the linear equation, the slope, b, and the intercept, a, must first be computed using the following least squares formulas: Football attendance accounts for the largest portion of its revenues, and the athletic director believes attendance is directly related to the number of wins by the team. The business manager has accumulated total annual attendance figures for the past eight years: Given the number of returning starters and the strength of the schedule, the athletic director believes the team will win at least seven games next year. Develop a simple regression equation for this data to forecast attendance for this level of success. The computations necessary to compute a and b using the least squares formulas are summarized in the accompanying table. Note that y is given in 1,s to make manual computation easier. Observing the regression line relative to the data points, it would appear that the data follow a distinct upward linear trend, which would indicate that the forecast should be relatively accurate. In fact, the MAD value for this forecasting model is 1. Correlation Correlation in a linear regression equation is a measure of the strength of the relationship between the independent and dependent variables. The formula for the correlation coefficient is The value of r varies between A value of r near zero implies that there is little or no linear relationship between variables. We can determine the correlation coefficient for the linear regression equation determined in Example This value for the correlation coefficient is very close to 1. Another measure of the strength of the relationship between the variables in a linear regression equation is the coefficient of determination. It is computed by squaring the value of r. It indicates the percentage of the variation in the dependent variable that is a result of the behavior of the independent variable. A value of 1. Regression Analysis with Excel The development of the simple linear regression equation and the correlation coefficient for our example was not too difficult because the amount of data was relatively small. However, manual computation of the components of simple linear regression equations can become very time-consuming and cumbersome as the amount of data increases. A12 " entered in cell D7 and shown on the formula bar at the top of the spreadsheet. A linear regression forecast can also be developed directly with Excel using the "Data Analysis" option from the Tools menu we accessed previously to develop an exponentially smoothed forecast. We first enter the cells from Exhibit Next enter the x value cells, A5: The output range is the location on the spreadsheet that you want to put the output results. This range needs to be large 18 cells by 9 cells and not overlap with anything else on the spreadsheet. Clicking on "OK" will result in the spreadsheet shown in Exhibit Note that the "Summary Output" has been slightly moved around so that all the results could be included on the screen in Exhibit The "Summary Output" in Exhibit The essential items that we are interested in are the intercept and slope labeled "X Variable 1" in the "Coefficients" column at the bottom of the spreadsheet, and the "Multiple R" or correlation coefficient value shown under "Regression Statistics.

Multiple Regression Another causal method of forecasting is multiple regression, a more powerful extension of linear regression. Linear regression relates demand to one other independent variable, whereas multiple regression reflects the relationship between a dependent variable and two or more independent variables. A multiple regression model has the following general form: For example, the demand for new housing y in a region might be a function of several independent variables, including interest rates, population, housing prices, and personal income. Development and computation of the multiple regression equation, including the compilation of data, is more complex than linear regression. The only means for forecasting using multiple regression is with a computer. To demonstrate the capability to solve multiple regression problems with Excel spreadsheets we will expand our State University athletic department example for forecasting attendance at football games that we used to demonstrate linear regression. Instead of attempting to predict attendance based on only one variable, wins, we will include a second variable for advertising and promotional expenditures as follows: We will use the "Data Analysis" option add-in from the Tools menu at the top of the spreadsheet that we used in the previous section to develop our linear regression equation, and then the "Regression" option from the "Data Analysis" menu. The resulting spreadsheet with the multiple regression statistics is shown in Exhibit Then we enter the "Input X Range" as A4: B12 as shown in Exhibit The regression coefficients for our x variables, wins and promotion, are shown in cells B27 and B Thus the multiple regression equation is formulated as This equation can now be used to forecast attendance based on both projected football wins and promotional expenditure. This would seem to suggest that the number of wins has a more significant impact on attendance than promotional expenditures. However, as we have already noted, the number of wins would appear to probably account for a larger part of the variation in attendance. Gas orders must be specified to suppliers at least 24 hours in advance. Vermont Gas Systems has storage capacity available for a buffer inventory of only one hour of gas use so an accurate daily forecast of gas demand is essential. Vermont Gas Systems uses regression to forecast daily gas demand. In its forecast models, gas demand is the dependent variable, and factors such as weather information, industrial customer demand, and changing end-use consumer demand are independent variables. During the winter customers use more gas for heat, making an accurate weather forecast a very important factor. Detailed three-day weather forecasts are provided to Vermont Gas Systems five times per day from a weather forecasting service. Individual regression forecasts are developed for 24 large-use industrial and municipal customers such as factories, hospitals, and schools. End-use demand is the total potential capacity of all natural gas appliances in the system. It changes daily as new customers move into a new house, apartment, or business adding new appliances or equipment to the system. Another factor related to end-use demand is water temperature, which will decrease by as much as 25 degrees Fahrenheit within the city water system during the winter. End-use demand and water temperature changes have minimal affect on a daily basis, but their impact is significant over several weeks. To compensate for these factors, the utility uses only the most recent 30 days of demand data in developing its forecast models and updates the models on a weekly basis. The results of the forecast model are interpreted by Vermont Gas Systems and supplemented with its individual knowledge of the supply chain distribution system and customer usage to develop an overall, accurate daily forecast of gas demand. A problem often encountered in multiple regression is multicollinearity, or the amount of "overlapping" information about the dependent variable that is provided by several independent variables. This problem usually occurs when the independent variables are highly correlated, as in this example, in which wins and promotional expenditures are both positively correlated i. Possibly the athletic department increased promotional expenditures when it thought it would have a better team that would achieve more wins. Multicollinearity and how to cope with it is a topic that is beyond the scope of this text and this brief section on multiple regression; however most statistics texts will discuss this topic in detail. What is the difference between linear and multiple regression? Define the different components y, x, a, and b of a linear regression equation. The Chamber of Commerce publishes guidelines for sales forecasting in small businesses. Summarize these guidelines in a one-page report. Which of the recommendations are unique to small businesses? Visit the web site of the Forecasting Business Connection. What services does it offer to assist companies in their forecasting efforts? What types of forecasting techiniques do most of the companies offer? Link to several company sites you have to find the

URLs and print out examples of their forecasting "product". Click Internet Exercises for the list of internet links for these exercises.

Chapter 2 : Introduction to Correlation and Regression Analysis

Familiar methods such as linear regression and ordinary least squares regression are parametric, in that the regression function is defined in terms of a finite number of unknown parameters that are estimated from the data.

Published online Oct Published by Oxford University Press. This article has been cited by other articles in PMC. Penalized regression methods have been adopted widely for high-dimensional feature selection and prediction in many bioinformatic and biostatistical contexts. While their theoretical properties are well-understood, specific methodology for their optimal application to genomic data has not been determined. Through simulation of contrasting scenarios of correlated high-dimensional survival data, we compared the LASSO, Ridge and Elastic Net penalties for prediction and variable selection. Furthermore, we found that in a simulated scenario favoring the LASSO penalty, a univariate pre-filter made the Elastic Net behave more like Ridge regression, which was detrimental to prediction performance. We demonstrate the real-life application of these methods to predicting the survival of cancer patients from microarray data, and to classification of obese and lean individuals from metagenomic data. Based on these results, we provide an optimized set of guidelines for the application of penalized regression for reproducible class comparison and prediction with genomic data. A parallelized implementation of the methods presented for regression and for simulation of synthetic data is provided as the pensim R package, available at http: Supplementary data are available at Bioinformatics online. In this setting, ordinary regression is subject to overfitting and instability of coefficients Harrell et al. Two penalization methods, and a hybrid of these, are most commonly used. Ridge regression Hoerl and Kennard, uses a penalty on the L2 norm of the coefficients, which introduces bias in the prediction error in exchange for reduced variance. However, ridge regression keeps all variables in the model and thus cannot produce a parsimonious model from many variables. LASSO regression Tibshirani, ; penalizes the L1 norm, which tends to reduce many coefficients to exactly zero and thus performs variable selection in addition to prediction. However, the LASSO has been noted to be inferior to Ridge regression for prediction in lower dimensional situations, and tends to select only one of a group of collinear variables, which may not always be desirable Zou and Hastie, These three variants of penalized regressionâ€”LASSO, Ridge and Elastic Netâ€”have since been applied to a variety of phenotype prediction tasks using genomic data for example, Sharma et al. We present a comprehensive assessment and optimization of these methods, using two contrasting configurations of simulated genomic data and two genome-scale experimental datasets. In particular, we show that successive 1D tuning of the Elastic Net restricts the search for tuning parameters sufficiently that it can yield inferior prediction to a single-penalty counterpart. A univariate pre-filter is commonly used to reduce dimensionality and computation time, but we demonstrate a simulated situation in which the pre-filter can reduce predictive performance of the Elastic Net by reducing the importance of the L1 penalty relative to the L2 penalty. We applied these optimized regression procedures in two very differing genomic settings: We use a cross-validation strategy for assessment of model prediction Molinaro et al. Both examples proved favorable to the L2 penalty, and models trained by Ridge regression and Elastic Net showed independent predictive ability, whereas models trained by the LASSO did not. We emphasize evidence of overfitting in both simulated and real experimental data, and summarize methods for realistic assessment of prediction accuracy with limited sample size. Based on results from this study and best practices for high-dimensional model validation, we conclude with an end-to-end methodology for effective application of penalized regression to diverse genomic data for prediction and variable selection. We further apply the guidelines for penalized regression developed from these synthetic data to real expression data and to penalized logistic regression for metagenomic data from the gut microbiomes and obesity status of subjects in the MetaHIT study Qin et al. Simulated variables were standard normal distributed, with covariance matrix specified by the within-group correlations in Table 1 and between-group covariance of zero. Survival times for patients were then sampled from a random exponential distribution with decay rate hj and censored on a uniform U 2,10 distribution as, for example, in Gui and Li Results without censoring for Simulated genomic predictor variables associated with a survival outcome Variable.

Chapter 3 : PROC PLS: Regression Methods :: SAS/STAT(R) User's Guide

Generally, statistical regression is collection of methods for determining and using models that explain how a response variable (dependent variable) relates to one or more explanatory variables (predictor variables).

View Blog Should you use linear or logistic regression? There are hundreds of types of regressions. Here is an overview for data scientists and other analytic practitioners, to help you decide on what regression to use depending on your context. Many of the referenced articles are much better written fully edited in my data science Wiley book. Oldest type of regression, designed years ago; computations on small data could easily be carried out by a human being, by design. Can be used for interpolation, but not suitable for predictive analytics; has many drawbacks when applied to modern data , e. A better solution is piecewise-linear regression, in particular for time series. Used extensively in clinical trials, scoring and fraud detection, when the response is binary chance of succeeding or failing, e. Suffers same drawbacks as linear regression not robust, model-dependent , and computing regression coeffients involves using complex iterative, numerically unstable algorithm. Can be well approximated by linear regression after transforming the response logit transform. Some versions Poisson or Cox regression have been designed for a non-binary response, for categorical data classification , ordered integer response age groups , and even continuous response regression trees. A more robust version of linear regression, putting constrainsts on regression coefficients to make them much more natural, less subject to over-fitting, and easier to interpret. Click here for source code. Similar to ridge regression, but automatically performs variable reduction allowing regression coefficients to be zero. Consists in performing one regression per strata, if your data is segmented into several rather large core strata, groups, or bins. Regression in unusual spaces: Used when all variables are binary, typically in scoring algorithms. It assumes that you have some prior knowledge about the regression coefficients. However, in practice, the prior knowledge is translated into artificial conjugate priors - a weakness of this technique. Similar to linear regression, but using absolute values L1 space rather than squares L2 space. This is the new type of regression, also used as general clustering and data reduction technique. It solves all the drawbacks of traditional regression. Ideal for black-box predictive algorithms. Other Solutions Data reduction can also be performed with our feature selection algorithm. An example of such blending is hidden decision trees. Categorical independent variables such as race, are sometimes coded using multiple binary dummy variables.

Chapter 4 : Regression: Methods

Featured on this site are the online notes on Regression Methods reorganized and supplemented by Dr. Iain Pardoe, based on original notes by Dr. Laura Simon and Dr. Derek Young. In addition, in the Resources section, there are Worked Examples Using Minitab that demonstrate how to perform many of the methods used in regression and Video.

Reviews Summary Handbook of Regression Methods concisely covers numerous traditional, contemporary, and nonstandard regression methods. The handbook provides a broad overview of regression models, diagnostic procedures, and inference procedures, with emphasis on how these methods are applied. The organization of the handbook benefits both practitioners and researchers, who seek either to obtain a quick understanding of regression methods for specialized problems or to expand their own breadth of knowledge of regression topics. This handbook covers classic material about simple linear regression and multiple linear regression, including assumptions, effective visualizations, and inference procedures. It presents an overview of advanced diagnostic tests, remedial strategies, and model selection procedures. Finally, many chapters are devoted to a diverse range of topics, including censored regression, nonlinear regression, generalized linear models, and semiparametric regression. Features Presents a concise overview of a wide range of regression topics not usually covered in a single text Includes over 80 examples using nearly 70 real datasets, with results obtained using R Offers a Shiny app containing all examples, thus allowing access to the source code and the ability to interact with the analyses Table of Contents Introduction. The Basics of Regression Models. Matrix Notation in Regression. Advanced Regression Diagnostic Methods. Measurement Errors and Instrumental Variables Regression. Correlated Errors and Autoregressive Structures. Crossvalidation and Model Selection Methods. Biased Regression Methods and Regression Shrinkage. Piecewise and Nonparametric Methods. Regression Models with Censored Data. Regression Models with Counts as Responses. He has over ten years of experience as a statistician, including positions in industry, government, and academia. During this time, he has also taught online courses in regression methods for Penn State University and the University of Kentucky. His research interests include finite mixture models, tolerance regions, and statistical computing. Reviews "Covering a wide range of regression topics, this clearly written handbook explores not only the essentials of regression methods for practitioners but also a broader spectrum of regression topics for researchers. Complete and detailed, this unique, comprehensive resource provides an extensive breadth of topical coverage, some of which is not typically found in a standard text on this topic. In addition, assumptions and applications of linear models as well as diagnostic tools and remedial strategies to assess them are addressed. Numerous examples using over 75 real data sets are included, and visualizations using R are used extensively. Also included is a useful Shiny app learning tool; based on the R code and developed specifically for this handbook, it is available online. This thoroughly practical guide will be invaluable for graduate collections. Gougeon, Choice Connect "The list of calculated examples contains virtually every possible field of application of statistics, a small subset of them reads as follows:

Chapter 5 : Overview of regression methods | Statistics

Regression analysis is a quantitative research method which is used when the study involves modelling and analysing several variables, where the relationship includes a dependent variable and one or more independent variables. In simple terms, regression analysis is a quantitative method used to.

Correlation and simple regression formulas Linear regression analysis is the most widely used of all statistical techniques: Then the equation for computing the predicted value of Yt is: This formula has the property that the prediction for Y is a straight-line function of each of the X variables, holding the others fixed, and the contributions of different X variables to the predictions are additive. The slopes of their individual straight-line relationships with Y are the constants b1, b2, â€¦, bk, the so-called coefficients of the variables. That is, bi is the change in the predicted value of Y per unit of change in Xi, other things being equal. The coefficients and intercept are estimated by least squares, i. The first thing you ought to know about linear regression is how the strange term regression came to be applied to models like this. They were first studied in depth by a 19th-Century scientist, Sir Francis Galton. Galton was a self-taught naturalist, anthropologist, astronomer, and statistician--and a real-life Indiana Jones character. He was famous for his explorations, and he wrote a best-selling book on how to survive in the wilderness entitled "The Art of Travel: From the Practical to the Peculiar. They provide many handy hints for staying alive--such as how to treat spear wounds or extract your horse from quicksand--and introduced the concept of the sleeping bag to the Western World. Click on these pictures for more details: Galton was a pioneer in the application of statistical methods to measurements in many branches of science, and in studying data on relative sizes of parents and their offspring in various species of plants and animals, he observed the following phenomenon: The same is true of virtually any physical measurement and in the case of humans, most measurements of cognitive and physical ability that can be performed on parents and their offspring. Here is the first published picture of a regression line illustrating this effect, from a lecture presented by Galton in The R symbol on this chart whose value is 0. Galton termed this phenomenon a regression towards mediocrity , which in modern terms is a regression to the mean. It is a purely statistical phenomenon. Unless every child is exactly as the same size as the parent in relative terms i. Return to top of page. Regression to the mean is an inescapable fact of life. Your children can be expected to be less exceptional for better or worse than you are. Your score on a final exam in a course can be expected to be less good or bad than your score on the midterm exam, relative to the rest of the class. The key word here is "expected. We have already seen a suggestion of regression-to-the-mean in some of the time series forecasting models we have studied: This is not true of random walk models, but it is generally true of moving-average models and other models that base their forecasts on more than one past observation. The intuitive explanation for the regression effect is simple: The best we can hope to do is to predict only that part of the variability which is due to the signal. Hence our forecasts will tend to exhibit less variability than the actual values, which implies a regression to the mean. Another way to think of the regression effect is in terms of selection bias. Suppose that we select a sample of professional athletes whose performance was much better than average or students whose grades were much better than average in the first half of the year. The fact that they did so well in the first half of the year makes it probable that both their skill and their luck were better than average during that period. In the second half of the year we may expect them to be equally skillful, but we should not expect them to be equally lucky. So we should predict that in the second half their performance will be closer to the mean. Meanwhile, players whose performance was merely average in the first half probably had skill and luck working in opposite directions for them. We should therefore expect their performance in the second half to move away from the mean in one direction or another, as we get another independent test of their skill. However, the actual performance of the players should be expected to have an equally large variance in the second half of the year as in the first half, because it merely results from a redistribution of independently random luck among players with the same distribution of skill as before. A nice discussion of regression to the mean in the broader context of social science research can be found here. Justification for regression assumptions Why should we assume that relationships between variables are linear? Because linear relationships are the simplest non-trivial relationships that can be imagined hence the easiest to work with , and Because the "true" relationships between our variables are often at least approximately linear over the range of values that are of interest to us, and This is a strong assumption, and the first step in regression modeling should be to look at scatterplots of the variables and in the case of time series data, plots of the variables vs. And after fitting a model, plots of the errors should be studied to see if there are unexplained nonlinear patterns. This is especially important when the goal is to make predictions for scenarios outside the range of the historical data, where departures from perfect linearity are likely to have the biggest effect. If you see evidence of nonlinear relationships, it is possible though not guaranteed that transformations of variables will straighten them out in a way that will yield useful inferences and predictions via linear regression. And why should we assume that the effects of different independent variables on the expected value of the dependent variable are additive? This is a very strong assumption, stronger than most people realize. It implies that the marginal effect of one independent variable i. In a multiple regression model, the estimated coefficient of a given independent variable supposedly measures its effect while "controlling" for the presence of the others. However, the way in which controlling is performed is extremely simplistic: Many users just throw a lot of independent variables into the model without thinking carefully about this issue, as if their software will automatically figure out exactly how they are related. Even automatic model-selection methods e. They work only with the variables they are given, in the form that they are given, and then they look only for linear, additive patterns among them in the context of each other. A common practice is to include independent variables whose predictive effects logically cannot be additive, say, some that are totals and others that are rates or percentages. You need to collect the relevant data, understand what it measures, clean it up if necessary, perform descriptive analysis to look for patterns before fitting any models, and study the diagnostic tests of model assumptions afterward, especially statistics and plots of the errors. You should also try to apply the appropriate economic or physical reasoning to determine whether an additive prediction equation makes sense. Here too, it is possible but not guaranteed that transformations of variables or the inclusion of interaction terms might separate their effects into an additive form, if they do not have such a form to begin with, but this requires some thought and effort on your part. And why should we assume the errors of linear models are independently and identically normally distributed? This assumption is often justified by appeal to the Central Limit Theorem of statistics, which states that the sum or average of a sufficiently large number of independent random variables--whatever their individual distributions--approaches a normal distribution. Much data in business and economics and engineering and the natural sciences is obtained by adding or averaging numerical measurements performed on many different persons or products or locations or time intervals. Insofar as the activities that generate the measurements may occur somewhat randomly and somewhat independently, we might expect the variations in the totals or averages to be somewhat normally distributed. It is again mathematically convenient: This family includes the t distribution, the F distribution, and the Chi-square distribution. But here too caution must be exercised. Even if the unexplained variations in the dependent variable are approximately normally distributed, it is not guaranteed that they will also be identically normally distributed for all values of the independent variables. Perhaps the unexplained variations are larger under some conditions than others, a condition known as "heteroscedasticity". For example, if the dependent variable consists of daily or monthly total sales, there are probably significant day-of-week patterns or seasonal patterns. In such cases the variance of the total will be larger on days or in seasons with greater business activity--another consequence of the central limit theorem. It is also not guaranteed that the random variations will be statistically independent. This is an especially important question when the data consists of time series: A very important special case is that of stock price data, in which percentage changes rather than absolute changes tend to be normally distributed. This implies that over moderate to large time scales, movements in stock prices are lognormally distributed rather than normally distributed. A log transformation is typically applied to historical stock price data when studying growth and volatility. See the geometric random walk page instead. You still might think that variations in the values of portfolios of stocks would tend to be normally distributed, by virtue of the central limit theorem, but the central limit theorem is actually rather slow to bite on the lognormal distribution because it is so asymmetrically long-tailed. A sum of 10 or 20 independently and identically lognormally distributed variables has a distribution that is still quite close to lognormal. Because the assumptions of linear regression linear, additive relationships with i. Its output contains no more information than is provided by its inputs, and its inner mechanism needs to be compared with reality in each situation where it is applied. Correlation and simple regression formulas A variable is, by definition, a quantity that may vary from one measurement to another in situations where different samples are taken from a population or observations are made at different points in time. In fitting statistical models in which some variables are used to predict others, what we hope to find is that the different variables do not vary independently in a statistical sense , but that they tend to vary together. In particular, when fitting linear models, we hope to find that one variable say, Y is varying as a straight-line function of another variable say, X. In other words, if all other possibly-relevant variables could be held fixed, we would hope to find the graph of Y versus X to be a straight line apart from the inevitable random errors or "noise". A measure of the absolute amount of variability in a variable is naturally its variance, which is defined as its average squared deviation from its own mean. Equivalently, we can measure variability in terms of the standard deviation, which is defined as the square root of the variance. The standard deviation has the advantage that it is measured in the same units as the original variable, rather than squared units. Our task in predicting Y might be described as that of explaining some or all of its variance--i. Why is it not constant? That is, we would like to be able to improve on the naive predictive model: More precisely, we hope to find a model whose prediction errors are smaller, in a mean square sense, than the deviations of the original variable from its mean. In using linear models for prediction, it turns out very conveniently that the only statistics of interest at least for purposes of estimating coefficients to minimize squared error are the mean and variance of each variable and the correlation coefficient between each pair of variables. The coefficient of correlation between X and Y is commonly denoted by rXY, and it measures the strength of the linear relationship between them on a relative i. The correlation coefficient is most easily computed if we first standardize the variables, which means to convert them to units of standard-deviations-from-the-mean, using the population standard deviation rather than the sample standard deviation, i. P is the Excel function for the population standard deviation. Here and elsewhere I am going to use Excel functions rather than conventional math symbols in some of the formulas to illustrate how the calculations would be done on a spreadsheet. Now, the correlation coefficient is equal to the average product of the standardized values of the two variables within the given sample of n observations: The average of the values in the last column is the correlation between X and Y. S in Excel, but the population statistic is the correct one to use in the formula above. If the two variables tend to vary on the same sides of their respective means at the same time, then the average product of their deviations and hence the correlation between them will be positive, since the product of two numbers with the same sign is positive.

Chapter 6 : Regression Methods in Statistical Process Control

deal with linear regression and a follow-on note will look at nonlinear regression. Regression analysis is used when you want to predict a continuous dependent variable or response from a number of independent or input variables.

All Modules Introduction to Correlation and Regression Analysis In this section we will first discuss correlation analysis, which is used to quantify the association between two continuous variables e. Regression analysis is a related technique to assess the relationship between an outcome variable and one or more risk factors or confounding variables. The outcome variable is also called the response or dependent variable and the risk factors and confounders are called the predictors, or explanatory or independent variables. In regression analysis, the dependent variable is denoted "y" and the independent variables are denoted by "x". The term "predictor" can be misleading if it is interpreted as the ability to predict even beyond the limits of the data. Also, the term "explanatory variable" might give an impression of a causal effect in a situation in which inferences should be limited to identifying associations. The terms "independent" and "dependent" variable are less subject to these interpretations as they do not strongly imply cause and effect. Correlation Analysis In correlation analysis, we estimate a sample correlation coefficient, more specifically the Pearson Product Moment correlation coefficient. The correlation between two variables can be positive i. The sign of the correlation coefficient indicates the direction of the association. The magnitude of the correlation coefficient indicates the strength of the association. A correlation close to zero suggests no linear association between two continuous variables. We use risk ratios and odds ratios to quantify the strength of association, i. The analogous quantity in correlation is the slope, i. And "r" or perhaps better R-squared is a measure of how much of the variability in the dependent variable can be accounted for by differences in the independent variable. The analogous measure for a dichotomous variable and a dichotomous outcome would be the attributable proportion, i. Therefore, it is always important to evaluate the data carefully before computing a correlation coefficient. Graphical displays are particularly useful to explore associations between variables. The figure below shows four hypothetical scenarios in which one continuous variable is plotted along the X-axis and the other along the Y-axis. Scenario 3 might depict the lack of association r approximately 0 between the extent of media exposure in adolescence and age at which adolescents initiate sexual activity. Example - Correlation of Gestational Age and Birth Weight A small study is conducted involving 17 infants to investigate the association between gestational age at birth, measured in weeks, and birth weight, measured in grams. We wish to estimate the association between gestational age and infant birth weight. In this example, birth weight is the dependent variable and gestational age is the independent variable. The data are displayed in a scatter diagram in the figure below. Each point represents an x,y pair in this case the gestational age, measured in weeks, and the birth weight, measured in grams. Note that the independent variable is on the horizontal axis or X-axis , and the dependent variable is on the vertical axis or Y-axis. The scatter plot shows a positive or direct association between gestational age and birth weight. Infants with shorter gestational ages are more likely to be born with lower weights and infants with longer gestational ages are more likely to be born with higher weights. The formula for the sample correlation coefficient is where Cov x,y is the covariance of x and y defined as are the sample variances of x and y, defined as The variances of x and y measure the variability of the x scores and y scores around their respective sample means , considered separately. The covariance measures the variability of the x,y pairs around the mean of x and mean of y, considered simultaneously. To compute the sample correlation coefficient, we need to compute the variance of gestational age, the variance of birth weight and also the covariance of gestational age and birth weight. We first summarize the gestational age data. The mean gestational age is: To compute the variance of gestational age, we need to sum the squared deviations or differences between each observed gestational age and the mean gestational age. The computations are summarized below. The variance of gestational age is: Next, we summarize the birth weight data. The mean birth weight is: The variance of birth weight is computed just as we did for gestational age as shown in the table below. The variance of birth weight is: Next we compute the covariance, To compute the covariance of gestational age and birth weight, we need to multiply the deviation from the mean gestational age by the deviation from the mean birth weight for each participant i. Notice that we simply copy the deviations from the mean gestational age and birth weight from the two tables above into the table below and multiply. The covariance of gestational age and birth weight is: We now compute the sample correlation coefficient: Not surprisingly, the sample correlation coefficient indicates a strong positive correlation. In practice, meaningful correlations i. There are also statistical tests to determine whether an observed correlation is statistically significant or not i. Procedures to test whether an observed sample correlation is suggestive of a statistically significant correlation are described in detail in Kleinbaum, Kupper and Muller. Boston University School of Public Health.

Chapter 7 : Optimized application of penalized regression methods to diverse genomic data

The two basic types of regression are linear regression and multiple linear regression, although there are non-linear regression methods for more complicated data and analysis.

Numerous extensions have been developed that allow each of these assumptions to be relaxed i. Generally these extensions make the estimation procedure more complex and time-consuming, and may also require more data in order to produce an equally precise model. Example of a cubic polynomial regression, which is a type of linear regression. The following are the major assumptions made by standard linear regression models with standard estimation techniques e. This essentially means that the predictor variables x can be treated as fixed values, rather than random variables. This means, for example, that the predictor variables are assumed to be error-freeâ€"that is, not contaminated with measurement errors. Although this assumption is not realistic in many settings, dropping it leads to significantly more difficult errors-in-variables models. This means that the mean of the response variable is a linear combination of the parameters regression coefficients and the predictor variables. Note that this assumption is much less restrictive than it may at first seem. Because the predictor variables are treated as fixed values see above , linearity is really only a restriction on the parameters. The predictor variables themselves can be arbitrarily transformed, and in fact multiple copies of the same underlying predictor variable can be added, each one transformed differently. This trick is used, for example, in polynomial regression , which uses linear regression to fit the response variable as an arbitrary polynomial function up to a given rank of a predictor variable. This makes linear regression an extremely powerful inference method. In fact, models such as polynomial regression are often "too powerful", in that they tend to overfit the data. As a result, some kind of regularization must typically be used to prevent unreasonable solutions coming out of the estimation process. Common examples are ridge regression and lasso regression. Bayesian linear regression can also be used, which by its nature is more or less immune to the problem of overfitting. In fact, ridge regression and lasso regression can both be viewed as special cases of Bayesian linear regression, with particular types of prior distributions placed on the regression coefficients. This means that different values of the response variable have the same variance in their errors, regardless of the values of the predictor variables. In practice this assumption is invalid i. This is to say there will be a systematic change in the absolute or squared residuals when plotted against the predictive variables. Errors will not be evenly distributed across the regression line. Heteroscedasticity will result in the averaging over of distinguishable variances around the points to get a single variance that is inaccurately representing all the variances of the line. In effect, residuals appear clustered and spread apart on their predicted plots for larger and smaller values for points along the linear regression line, and the mean squared error for the model will be wrong. Typically, for example, a response variable whose mean is large will have a greater variance than one whose mean is small. In fact, as this shows, in many casesâ€"often the same cases where the assumption of normally distributed errors failsâ€"the variance or standard deviation should be predicted to be proportional to the mean, rather than constant. Simple linear regression estimation methods give less precise parameter estimates and misleading inferential quantities such as standard errors when substantial heteroscedasticity is present. However, various estimation techniques e. Bayesian linear regression techniques can also be used when the variance is assumed to be a function of the mean. It is also possible in some cases to fix the problem by applying a transformation to the response variable e. This assumes that the errors of the response variables are uncorrelated with each other. Actual statistical independence is a stronger condition than mere lack of correlation and is often not needed, although it can be exploited if it is known to hold. Bayesian linear regression is a general way of handling this issue. Lack of perfect multicollinearity in the predictors. For standard least squares estimation methods, the design matrix X must have full column rank p; otherwise, we have a condition known as perfect multicollinearity in the predictor variables. This can be triggered by having two or more perfectly correlated predictor variables e. It can also happen if there is too little data available compared to the number of parameters to be estimated e. At most we will be able to identify some of the parameters, i. See partial least squares regression. Methods for fitting linear models with multicollinearity have been developed; [5] [6] [7] [8] some require additional assumptions such as "effect sparsity"â€”that a large fraction of the effects are exactly zero. Note that the more computationally expensive iterated algorithms for parameter estimation, such as those used in generalized linear models , do not suffer from this problem. Beyond these assumptions, several other statistical properties of the data strongly influence the performance of different estimation methods: The statistical relationship between the error terms and the regressors plays an important role in determining whether an estimation procedure has desirable sampling properties such as being unbiased and consistent. This illustrates the pitfalls of relying solely on a fitted model to understand the relationship between variables. A fitted linear regression model can be used to identify the relationship between a single predictor variable xj and the response variable y when all the other predictor variables in the model are "held fixed". This is sometimes called the unique effect of xj on y. In contrast, the marginal effect of xj on y can be assessed using a correlation coefficient or simple linear regression model relating only xj to y; this effect is the total derivative of y with respect to xj. Care must be taken when interpreting regression results, as some of the regressors may not allow for marginal changes such as dummy variables , or the intercept term , while others cannot be held fixed recall the example from the introduction: It is possible that the unique effect can be nearly zero even when the marginal effect is large. This may imply that some other covariate captures all the information in xj, so that once that variable is in the model, there is no contribution of xj to the variation in y. Conversely, the unique effect of xj can be large while its marginal effect is nearly zero. This would happen if the other covariates explained a great deal of the variation of y, but they mainly explain variation in a way that is complementary to what is captured by xj. In this case, including the other variables in the model reduces the part of the variability of y that is unrelated to xj, thereby strengthening the apparent relationship with xj. The meaning of the expression "held fixed" may depend on how the values of the predictor variables arise. If the experimenter directly sets the values of the predictor variables according to a study design, the comparisons of interest may literally correspond to comparisons among units whose predictor variables have been "held fixed" by the experimenter. Alternatively, the expression "held fixed" can refer to a selection that takes place in the context of data analysis. In this case, we "hold a variable fixed" by restricting our attention to the subsets of the data that happen to have a common value for the given predictor variable. This is the only interpretation of "held fixed" that can be used in an observational study. The notion of a "unique effect" is appealing when studying a complex system where multiple interrelated components influence the response variable. In some cases, it can literally be interpreted as the causal effect of an intervention that is linked to the value of a predictor variable. However, it has been argued that in many cases multiple regression analysis fails to clarify the relationships between the predictor variables and the response variable when the predictors are correlated with each other and are not assigned following a study design. Simple and multiple linear regression[edit] Example of simple linear regression , which has one independent variable The very simplest case of a single scalar predictor variable x and a single scalar response variable y is known as simple linear regression. Nearly all real-world regression models involve multiple predictors, and basic descriptions of linear regression are often phrased in terms of the multiple regression model. Note, however, that in these cases the response variable y is still a scalar. Another term, multivariate linear regression, refers to cases where y is a vector, i. General linear models[edit] The general linear model considers the situation when the response variable is not a scalar for each observation but a vector, yi. Conditional linearity of E.

Chapter 8 : Chapter 10, Head 6

Regression Methods All of the predictive methods implemented in PROC PLS work essentially by finding linear combinations of the predictors (factors) to use to predict the responses linearly. The methods differ only in how the factors are derived, as explained in the following sections.

In this course we make use of a variety of methods for regression modeling. Below, we first define some concepts that can be used to understand the major distinctions between various approaches to regression. Then we review some specific regression methods along with their key properties. Before proceeding, note that regression itself is somewhat difficult to define in a way that differentiates it from the rest of statistics. In most cases, regression focuses on a conditional distribution, e. Any analysis focusing on a conditional distribution can be seen as a form of regression analysis. Major concepts Single index models: Depending on the context, this can mean any of the following: Regression for independent observations: Most of the basic regression methods are suitable for samples of independent observations. More advanced regression methods can be used when the observations are known to be dependent. Heteroscedasticity can be accommodated by some regression procedures e. Poisson regression works best if the mean and variance are equal. Some regression procedures, like ordinary least squares, work best when the population is homoscedastic, but can still give meaningful results with some loss of power if the population is heteroscedastic. If the conditional variance of the data is greater than the conditional variance of the population model being fit to the data, there is overdispersion. If the conditional variance of the data is less than the population model, there is underdispersion. This is one reason that data may be non-independent. The marginal regression function remains an object of interest when the data are dependent, even though it does not capture the relationship between the independent and dependent variables in full. It emphasizes the fact that in many data sets, there are complex inter-relationships between the observations that are not explained by the covariates. Multilevel models can also be viewed as a way to model variances and covariances, although these are modeled through random effects, rather than directly. For linear models, conditional and marginal effects are the same. But in nonlinear models the two types of effects differ. Most methods for nonlinear regression target either the marginal effects, or the conditional effects, but not both. In most cases the conditional effect will be numerically larger than the marginal effect. When referring to this type of marginal effect, the marginal and conditional effects differ even in a linear model. This terminology is used inconsistently to refer to different things in different settings. Another important distinction to make is between the various regression model structures, and different ways for fitting a regression model structure to data. There are many ways to fit this model to data, including least squares, and many forms of penalized least squares. However all of these fitting algorithms are fitting the same class of models to the data. For brevity, here we are focusing much more on the model structures, not the various statistical estimation nd fitting procedures used to fit the models to data. Some specific regression analysis methods Least squares: It is optimally used when the conditional mean function is linear in the covariates, and the conditional variance is constant. Both of these restrictions can be worked around, however. The link function allows the expected value of the response variable to be expressed as a known transformation of the linear predictor. GEE is an extension of GLM that allows for certain types of statistical dependencies between the observations. The fitting and inference in a GEE is robust in that the working dependence model can be misspecified, and the estimates and inferences will still be valid this can be stated in more precise terms but we will not do that here. These unobserved random effects can be viewed as missing information that reflects additional structure in the population not captured through the covariates. It is a very rich framework that can be used to account for a variety of structures in the population that are difficult to model in other ways, including clustering, multilevel nested clustering, crossed clustering, and heterogeneous partial associations e. Structurally, they are very similar to linear mixed models, and in practice, can be interpreted in a similar way, except for the important distinction that in a multilevel GLM, the marginal and conditional mean structures differ which is not the case for a multilevel linear model. Other forms of regression: The most familiar forms of this technique are single-level conditional logistic and Poisson regression. In both cases, we can have clustered data which would more often be handled using mixed effects or GEE , but by conditioning on the observed total of the outcome values within each group, the observations become conditionally independent, and can be rigorously fit using a single-level likelihood approach. Variance regression â€” this is a class of approaches that parametrically model the variance along with the mean, e. Additive regression â€" this is a way to restrict the general kernel regression technique to avoid the curse of dimensionality. The model is additive over the covariates, which is a strong restriction, but generalizes classical linear models by allowing each covariate to be transformed in an arbitrary way. Dimension reduction regression â€" this is a very unique and distinct class of regression approaches that posit a multi-index structure and an unknown link function. Generalized method of moments GMM â€" this is a technique for efficiently estimating the parameters of nonlinear models using only the moments. It is mainly used when it is important to estimate regression effects without requiring a full distribution to be specified. Multivariate regression â€" these are techniques for regressing a vector of dependent variables on a vector of independent variables.

Chapter 9 : Handbook of Regression Methods - CRC Press Book

Regression Analysis for Proportions When the response variable is a proportion or a binary value (0 or 1), standard regression techniques must be modified. STATGRAPHICS provides two important procedures for this situation: Logistic Regression and Probit Analysis.

How to select the right Regression Model? What is Regression Analysis? This technique is used for forecasting, time series modelling and finding the causal effect relationship between the variables. Regression analysis is an important tool for modelling and analyzing data. Why do we use Regression Analysis? There are multiple benefits of using regression analysis. They are as follows: Regression analysis also allows us to compare the effects of variables measured on different scales, such as the effect of price changes and the number of promotional activities. There are various kinds of regression techniques available to make predictions. But before you start that, let us understand the most commonly used regressions: Linear Regression It is one of the most widely known modeling technique. Linear regression is usually among the first few topics which people pick while learning predictive modeling. This task can be easily accomplished by Least Square Method. To know more details about these metrics, you can read: Model Performance metrics Part 1 , Part 2. There must be linear relationship between independent and dependent variables Multiple regression suffers from multicollinearity, autocorrelation, heteroskedasticity. Linear Regression is very sensitive to Outliers. Multicollinearity can increase the variance of the coefficient estimates and make the estimates very sensitive to minor changes in the model. The result is that the coefficient estimates are unstable In case of multiple independent variables, we can go with forward selection, backward elimination and step wise approach for selection of most significant independent variables. Here the value of Y ranges from 0 to 1 and it can represented by following equation. And, it is logit function. However, we have the options to include interaction effects of categorical variables in the analysis and in the model. The equation below represents a polynomial equation: While there might be a temptation to fit a higher degree polynomial to get lower error, this can result in over-fitting. Always plot the relationships to see the fit and focus on making sure that the curve fits the nature of the problem. Here is an example of how plotting can help: Especially look out for curve towards the ends and see whether those shapes and trends make sense. Higher polynomials can end up producing wierd results on extrapolation. Some of the most commonly used Stepwise regression methods are listed below: Standard stepwise regression does two things. Forward selection starts with most significant predictor in the model and adds variable for each step. Backward elimination starts with all predictors in the model and removes the least significant variable for each step. The aim of this modeling technique is to maximize the prediction power with minimum number of predictor variables. It is one of the method to handle higher dimensionality of data set. By adding a degree of bias to the regression estimates, ridge regression reduces the standard errors. It can be represented as: The complete equation becomes: In a linear equation, prediction errors can be decomposed into two sub components. Look at the equation below. In this equation, we have two components. This is added to least square term in order to shrink the parameter to have a very low variance. In addition, it is capable of reducing the variability and improving the accuracy of linear regression models. Look at the equation below: Elastic-net is useful when there are multiple features which are correlated. Lasso is likely to pick one of these at random, while elastic-net is likely to pick both. How to select the right regression model? Life is usually simple, when you know only one or two techniques. One of the training institutes I know of tells their students â€” if the outcome is continuous â€” apply linear regression. If it is binary â€” use logistic regression! However, higher the number of options available at our disposal, more difficult it becomes to choose the right one. A similar case happens with regression models. Data exploration is an inevitable part of building predictive model. This essentially checks for possible bias in your model, by comparing the model with all possible submodels or a careful selection of them. Cross-validation is the best way to evaluate models used for prediction. Here you divide your data set into two group train and validate. A simple mean squared difference between the observed and predicted values give you a measure for the prediction accuracy. Regression regularization methods Lasso, Ridge and ElasticNet works well in case of high dimensionality and multicollinearity among the variables in the data set. End Note By now, I hope you would have got an overview of regression. These regression techniques should be applied considering the conditions of data. One of the best trick to find out which technique to use, is by checking the family of variables i. In this article, I discussed about 7 types of regression and some key facts associated with each technique. Did you find this article useful?