**Chapter 1: Concepts and Examples of Research**

Preview Questions

1. What is regression analysis?
2. What are the key concepts of applied regression analysis?
3. What are multivariable techniques?
4. What’s the difference between experimental, quasi-experimental, and observational research?
5. What is a response variable?
6. What is a predictor variable?
7. What are examples of the type of research that can be done using regression analysis?

Reading Summary

Key concepts in empirical research

* Empirical research examines the relationship among a number of variables.
* Variables are factors that are measured for observational units or subjects.
* Multivariable methods use statistical techniques to examine the relationship among at least three variables.
* Regression analysis is a type of multivariable method.

Types of research

* Experimental, in which observational units or subjects are randomly assigned levels of predictor variables.
* Quasi-experimental, in which observational units or subjects are assigned levels of predictor variables but not in a random manner.
* Observational, in which observational units are not assigned levels of predictor variables.

Characteristics of research

* Multivariable methods are applicable to all three types of research.
* Each type of research provides a different level of confidence one can have in the results.
* The response variable is the dependent variable, which is influenced by predictor variables.
* Predictor variables are independent variables, which influence the response variable.
* Typically one (1) response variable and one (1) or more predictor variables.
* Potential for drawing definitive conclusions: observational < quasi-experimental < experimental
* Difficulty of implementation: observational < quasi-experimental < experimental

Measuring variables

* Error is unavoidable in measuring variables.
* Error in measuring variables gives rise to the need for statistical design and analysis.

Types of statistical inference

* Estimation comprises describing the characteristics and strength of the relationship among variables by quantifying them.
* Hypothesis testing comprises prosing explanations about the relationship among variables, stating probabilities about the reasonableness of such explanations, and drawing conclusions based on the stated probabilities.

**Chapter 2: Classification of Variables and the Choice of Analysis**

Preview Questions

1. What is gappiness?
2. What is level of measurement precision?
3. What is meant by descriptive orientation?
4. How do you overlap variable classification systems?
5. How do you choose a method of analysis?

Reading Summary

Approaches to classifying variables

* Classifications for variables help in deciding which methods to use for an analysis.
* Three methods for classifying variables are:
  + By the gappiness
  + By the level of measurement precision
  + By the descriptive orientation

Gappiness

* Gappiness refers to whether or not gaps exist between successive observations of the values of a variable.
* Discrete variables have gaps.
* Non-numeric data may be numerically coded as discrete variables.
* Continuous variables DO NOT have gaps (i.e., between any two values another value can potentially exist).
* Data on discrete variables are represented by a line chart to display sampling frequency.
* Data for continuous variables are grouped into intervals (e.g., histogram) to display sampling frequency.
* Discrete variables can be treated like continuous variables for analysis purposes when the values of a variable are not far apart and cover a wide range of numbers.
* Continuous variables are sometimes treated like discrete variables for analysis purposes.
* Considerations when deciding whether to categorize continuous variables:
  + Makes data collection easier
  + Simplifies the presentation of results
  + Information is lost
* Considerations for deciding when to categorize continuous variables:
  + At the time of collection
    - Less expensive
    - Less time consuming
    - Less precise
    - More likely to introduce human error (i.e., classification error)
  + At the time of analysis
    - Less error prone
    - Enables consideration of various classification schemes
* Errors
  + Classification error is a factor with discrete variables.
  + Measurement error is a factor with continuous variables.

Level of measurement precision

* Three (3) levels of measurement precision
  + Nominal (i.e., categorical) indicates different categories for the variable.
  + Ordinal indicates different categories for the variable and the order of the categories matters.
  + Interval indicates different categories for the variable, the order of the categories matters, and the distance between categories has meaning.
* Ratio variables or ratio-scale variables are interval variables in which the scale has a true zero.
* Measurement error for ratio-scale variables often have a non-normal distribution and are proportional to the size of the measurement.
* An interval scale is also ordinal and nominal.
* An ordinal scale is also nominal.

Descriptive orientation

* Descriptive orientation indicates whether a variable is meant to describe or be described by other variables.
  + Response or dependent variables are typically denoted by letter Y.
  + Predictor, regressor, or independent variables are typically denoted by letter X.
* Control variables are independent variables that affect relationships among other independent variables in a study but are of no interest.
* Control variables are sometimes referred to as nuisance variables, covariates, or confounders.

Overlap of Classification Schemes

* Any variable can be labeled according to each scheme.
* See Figure 2.5 (p. 12) for diagram of classification scheme overlap.
* All nominal variables are discrete but NOT all discrete variables are nominal.
* All continuous variables are ordinal and interval but NOT all ordinal and interval variables are continuous.

Choosing a method of analysis

* There are four considerations:
  + Purpose of the research
  + Mathematical characteristics of the variables
  + Statistical assumptions about the variables
  + Data collection method (i.e., sampling procedure)
* See Table 2.1 (p. 13) regarding guidance for choosing analysis methods
* Methods not covered
  + Nonparametric methods, which don’t require the data to fit a normal distribution
  + Cluster analysis

**Chapter 3: Basic Statistics Review**

Preview Questions

1. What are descriptive statistics?
2. What are random variables?
3. How are random variables distributed?
4. What are sampling distributions?
5. What is statistical inference?
6. How is statistical inference done?
7. What are error rates?
8. How is the power of an analysis determined?
9. What is the impact of sample size on an analysis?

Reading Summary

Basic overview

* Statistical analysis includes collecting, classifying, summarizing, and analyzing data.
* The text focuses on summarizing and analyzing data.
* Statistical inference is drawing valid conclusions about a population based on information about a sample from that population.
* A population is any set of items or measurements.
* A sample is any subset taken from a population.
* Descriptive statistics should be reviewed before making statistical inferences.
* Statistical inference
  + Two types are estimation and hypothesis testing
  + Based on certain assumptions about the distribution of random variables

Descriptive Statistics

* Descriptive statistics are measures computed from a set of data designed to describe aspects of the data.
* Most common types of descriptive statistics:
  + Central tendency (mean, median, and mode), which indicates average value of a variable.
  + Variability (dispersion), which indicates the extent to which the values of a variable differ from one another.
* Sample mean uses all observations in its calculation, but median and mode do not.
* Most common measures of variability:
  + Sample variance is the averaged squared deviation about the mean (Eq. on p. 17)
  + Sample variance (S2), which is expressed in squared units of the variable of interest
  + Sample standard deviation (S) = square root of sample variance, which is expressed in the same units of the variable of interest
* Plots of the data are a convenient way to examine data and are often revealing
  + Histogram (see Figure 3.1 on p. 18), which can be converted to a stem-and-leaf diagram
  + Stem-and-leaf diagram(see Figure 3.1 on p. 18)
  + Schematic plot, which is also called a box-and-whiskers plot
* Box-and-whiskers plot
  + Bottom line of box is the 25th percentile of data
  + Middle line of box is the 50th percentile of data
  + Top line of box is the 75th percentile of data
  + Top line - Bottom line = interquartile range (ICR)
  + + in box indicates the mean
  + Vertical lines extend from box as far as the data extend up to 1.5 ICRs
  + O beyond the vertical lines indicate moderate outliers

Random variables and distributions

* A random variable is a variable in which the observed values can be considered to result of a random experiment and cannot be anticipated with any certainty.
* Stochastic means randomly determined sequence of observations.
* Random variables denoted by capital letters.
* The probability distribution of a random variable is the pattern of the relative frequencies of all possible values in a population.
* The probability distribution is represented by a table, graph, or equation.
* For discrete random variables, the height of the lines of the line chart represents the probabilities of each possible outcome.
  + The sum of all probabilities for a random variable equal 1.
* For continuous random variables, the area under the curve between two points represents the probability associated with that range of values.
  + The total area under the curve equals 1.
  + The probability of any one particular value is 0.

Binomial distribution, B(n,π)

* The number of occurrences of a particular event in a series of n trials, where:
  + The trials are conducted in the same way.
  + There are only two possible outcomes for each trial, which is represented as π and 1-π
  + The outcome of any one trial is NOT affected by the outcome of any other trial.
  + The probability of success (π) remains the same for each trial.
* X ᴖ B(n,π) 🡪 random variable X as distributed by the binomial distribution with a probability of success of π in n trials.

Normal distribution, N(μ,σ)

* Symmetrical bell curve
* μ is the mean of the distribution
* σ is the standard deviation of the distribution

Standard normal distribution

* μ = 0, σ = 1
* To standardize X ᴖ N(μX,σX)
  + Z = (X - μX) / σX
  + X = μX + σX Z
* P(a X b) = P[{(a - μX) / σX  } Z {(b - μX) / σX }]
  + Look up equivalent probability statement about Z in the N(0,1) table
* When X is normally distributed and the sample size is moderately large, the same rule applies to the mean
  + Z = [( - μX) / ] = ( - μX) / ( σX / )
  + = [μX + ( σX / ) Z] = μX + ( σX Z / )
* Z function can be used to approximated binomial distribution B(n,π) when n > 20
  + μ = nπ
  + σ =

Normal probability plot

* Assesses how well sample data adheres to a normal distribution
* Plot ordered data values against corresponding percentiles from an estimated normal distribution.
* Cumulative relative frequencies of a normal distribution plot as a straight line.

Assessing normality

* Skewness indicates the amount of asymmetry of a distribution
  + Represented by sk(X)
  + Skewness is the average of cubed deviations about the mean (Eq. on p. 23)
* Kurtosis indicates the heaviness of the tail relative to the middle of a distribution
  + Represented by Kur(X)
  + The average of the fourth power of the deviations about the mean (Eq. on p. 23).
  + Is always non-negative.
  + Kurtosis for standardized normal distribution is 3.
  + -3 <= Kur(Z) < 0 for flat distributions with short tails.
  + Kur(Z) ≈ 0 for moderate to large random samples from a normal distribution.
  + Kur(Z) > 0 for heavy-tailed distributions.

Sampling distributions

* Student’s t distribution
  + Symmetric about 0 like the standard normal distribution
  + T = ( - μX) / [SX/]
  + Alternative to Z = (Ẋ - μX) / σẊ = (Ẋ - μX) / ( σX / )
  + When population is unknown and is estimated by
  + SX/ is the estimated standard error of
  + T has the t distribution with n-1 degrees of freedom
    - X is normally distributed
    - and are calculated using a random sample from the normal distribution
  + is the pooled sample variance to two normally distributed populations with the same standard deviation {i.e., N(μ1,σ) and N(μ2,σ)} (Eq. on p. 26)
* Chi-Square distribution,
  + Non-symmetric distribution skewed to the right
  + Describes the behavior of the non-negative random variable (n-1)S2 / σ2 
    - Has chi-square distribution with n-1 degrees of freedom , n-1
* F distribution
  + Non-symmetric distribution skewed to the right
  + Modeling probability distribution of the ratio of independent estimators of two population variances
  + Has distribution with n1-1 numerator degrees of freedom and n2-1 denominator degrees of freedom

Statistical Inference: Estimation

* Quantifying the specific value of an unknown population parameter (i.e., likely set of values for the parameter)
* Estimate using random variable
* is a point estimator of
* Point estimator takes the form of a formula or rule
* Procedure to estimate parameter
  1. Select a random sample from the population
  2. Calculate the point estimator of the parameter
  3. Associate a measure of its variability (e.g., confidence interval)
* is a fixed unknown constant and the lower and upper limits of the confidence interval are random quantities that vary from sample to sample.
* Confidence interval indicates the percent of all intervals from many repeated sets of samples of the same size that are expected to contain the parameter value
* Confidence interval = Point estimate [(Percentile of the t distr)(Est std error of the estimate)]

Statistical Inference: Hypothesis Testing

* Making a decision about a hypothesized value of an unknown population parameter (i.e., is the estimated value for the parameter different enough from the null hypothesis to conclude that the null hypothesis is unlikely to be true).
* Procedure for testing a statistical null hypothesis
  1. Check assumptions to justify selected testing procedure
  2. State the null hypothesis and alternative hypothesis
  3. Specify the significance level
  4. Specify the null hypothesis test statistic and distribution
  5. Form the decision rule
  6. Compute the value of the test statistic from the observed data
  7. Decide whether to reject or maintain the null hypothesis

Errors Rates, Power, and Sample Size

* Type 1 error is a false positive error
  + Reject the null hypothesis when it is in fact true
  + Accept the alternative hypothesis when it is in fact false
* Type 2 error is a false negative error
  + Accept the null hypothesis when it is in fact false
  + Reject the alternative hypothesis when it is in fact true
* See tables on p. 35
* The power of the test is the probability of selecting the alternative hypothesis when it is true, which is given by 1- where is the probability of a Type 2 error.
* We want the probability of Type 1 and Type 2 errors to be small.
* Sometimes possible to determine the sample size necessary to ensure that the probability of a Type 2 error is no larger than some desired value.

**Chapter 4 introduction to Regression Analysis**

Preview Questions

1. What is the difference between association versus causality?
2. What is a statistical model?
3. What is a deterministic model?

Reading Summary

Overview of regression analysis

* Regression analysis is used to evaluate the relationship of one or more independent variables with a single continuous dependent variable.
* Determine extent, direction, and strength of associations.
* Used when you can’t control independent variables.
* Also applicable to controlled experiments.

Applications of regression analysis

* Characterize the relationship
* Develop a quantitative formula
* Control for the effects of other variables that might also have a relationship with the dependent variable
* Determine which of several independent variables are important for predicting a dependent variable
* Determine the best mathematical model for describing an association
* Compare several derived regression relationships
* Evaluate the interactive effects of the two or more independent variables on a dependent variable
* Calculate a valid and precise estimate of one or more regression coefficients

Association versus causality

* An estimate may be biased
  + Method used to select subjects
  + Errors in the information or data
  + Other variables that can account for the observed association
* Statistically significant associations DO NOT establish causality
* For causality to exist
  + Change in one variable produces a change in another variable
  + Experimental proof is required
  + Obtaining experimental proof might be impractical, infeasible, or unethical

Methods for establishing causality

* Sufficient component cause model (SCC)
* Directed acyclic graph (DAG)
* Path analysis
* Structured equation modeling
* Qualitative evaluation of combined results from several studies
  + Strength of association
  + Consistency of findings
  + Specificity of the association of the suspected cause
  + The cause precedes the effect
  + Meaningful changes in the suspected cause are associated with changes in the dependent variable
  + Hypothesized causal relationship is consistent with current theoretical knowledge
  + Findings are coherent with known facts about the outcome variable
  + Findings are supported by experimental or quasi-experimental evidence
  + Similar to another situation where a causal link has been established

Statistical versus deterministic models

* Deterministic models assume an ideal setting
* Statistical models allow for the possibility of error

**Chapter 5: Straight-line Regression Analysis**

Preview Questions

1. How do you create a regression with a single-independent variable?
2. What are the mathematical properties of a straight line?
3. What are the statistical assumptions for a straight line model?
4. How do you determine the best fitting straight line?
5. How do you measure the quality of the straight line fit and the estimate of the variance?
6. How do you make inferences about the slope and intercept?
7. What are interpretations of tests for slope and intercept?
8. How do you determine the mean value of Y at a specified value of X?
9. How do you predict a new value of Y at a given value of X?
10. How do you determine if a straight line model is appropriate?
11. What is BRFSS Analysis?

Reading Summary

Preview

* Special case of k=1 variables.
* Researchers often evaluate one variable at a time even when several independent variables are eventually considered all together.

Regression with a single independent variable

* n pairs of observations X1,Y1… Xn,Yn
* Plot on a scatter diagram
* Basic questions addressed in regression analysis
  + What is the most appropriate mathematical model for the data?
  + How do we determine the best-fitting model for the data?

General strategies for regression analysis

* Forward method
  + Start with a simply structured model and ad more complexity as necessary.
* Backward method
  + Start with a complex model and successively simplify.
* Theory-driven method
  + Start with a model suggest by experience or theory and revise towards or away from complexity as required
* Exploratory approach
  + Try multiple strategies and select the most reasonable one.
  + Requires additional care to ensure reliable conclusions.
* Forward and backward methods don’t necessarily lead to the same final model.

Forward method

1. Assume a straight line model.
2. Find the best fitting line.
3. Determine whether the resulting line helps to describe the dependent variable.
4. Determine whether the assumption of the line is appropriate.
5. If the model is not appropriate, select a slightly more complex model and repeat steps 2-4.
6. Continue until an appropriate model is found.

Mathematical properties of a straight line

* Mathematical model does NOT consider Y as a random variable.
* is the y intercept
* is the slope
* Plot any two points then connect with a line

Statistical assumptions for a straight-line model

* Straight line only approximates the true state of the phenomenon under consideration.
* Since the line is determined from sample data and not the population were required to consider how to estimate unknown population parameters.
* For any value of X, Y is a random variable with a certain probability distribution of finite mean and variance.
* The Y values are statistically independent of each other.
  + This is often violated when different observations are made on the same individual at different times.
  + Can result in invalid statistical conclusions.
  + Requires special methods.
* The mean value of Y is a straight line function of X.
  + , where E denotes a random variable with mean = 0 for any X
  + Since E has a mean of 0, y must have the mean
  + Statistical model
* E is how distant an individual observational unit’s response can be from the population regression line (i.e., error from the expected average).
  + Best fitting line will have small errors
* The variance for Y is the same for any fixed X (i.e., homoscedasticity).
* For any fixed value of X, Y has a normal distribution.
  + Regression analysis conclusions will generally be reliable and accurate even if the normality assumption is slightly violated.
  + Give considerable leeway before deciding that normality assumption is violated.
  + Can use transformations to obtain usable data if raw data violates normality assumption but must check to make sure that other assumptions are not violated

Important concepts

* E has a normal distribution with mean = 0 and variance σ2 for all observations
* Y is a random variable
* X is assumed to be measured without error
* are parameters of unknown specific values for a population
* E is a random unobservable variable
* Point estimate () of E at the value X is calculated as = Y -
* is called a residual.
* If there are n(X,Y) pairs then there are n residuals
* Normally distributed random variable is referred to as a Gaussian distribution

Analytical approaches for determining the best-fitting straight line

* Least-squares method
  + Mathematical algorithm
  + Line that minimizes the sum of the squares of the errors.
  + 0 and 1 are least squares estimate of and
  + Referred to as:
    - Sum of squares about the regression line
    - Residual sum of squares
    - Sum of squares due to error (SSE)
  + Outliers affect the least-squares estimates.
* Minimum variance method
  + 0 and 1 that are unbiased for their unknown population counterparts and have minimum variance among all unbiased estimators of and
  + Yields same solution as least-squares method.
  + See equation 5.7 on p. 57

Measure of the quality of the straight-line fit and estimate of variance

* SSE is a measure of the quality of the straight-line fit and estimate of variance.
* SSE increases as the fit gets worse.
* A lot of random variation in Y (i.e., large variance) increases SSE.
* If the assumption of a straight line model is not appropriate then SSE may be large.
* SSE can be used to estimate variance S2 = SSE

Inferences about the slope and intercept

* Test statistic, T, is used to test the null hypothesis for the slope of the straight line.
* See Eq. on p. 61.
* See rules for rejecting null hypothesis on bottom of p. 62
* See Table 5.2 on p. 63 and Table 5.3 on p. 64.

Interpretations of tests for slope and intercept

* Test for zero slope, H0: =0
  + If null hypothesis is NOT rejected, one of the following is true:
    - X provides little, if any, help in predicting Y
    - The underlying relationship between X and Y is NOT linear
  + If null hypothesis is rejected, one of the following is true:
    - X provides significant information about predicting Y
    - There is statistical evidence of a linear component even thought a better model might have a nonlinear term
* Test for zero intercept, H0: =0
  + If the null hypothesis is NOT rejected, may be able to remove from the model
  + Rarely of interest in most studies

Mean value of Y at a specified value of X

* See Eq. 5.11 on p. 67 for test statistic, T, for this hypothesis.
* See Eq. 5.13 on p. 67 for confidence interval of population mean at given X=X0
* A confidence bands is the confidence interval for the regression line over the entire range of X values
  + Plot upper and lower confidence limits for several values of X
  + Sketch the curves that connect the points

Prediction of a new value of Y at X0

* Random variable Y has a prediction interval (PI) instead of a confidence interval.
* Describe individual predictions over the entire range of X values
* See Eq. 5.15 on p. 70
* An estimator of an individual response should have more variability than an estimator of a population mean response.
  + Two sources of error
  + Individual error measured by variance σ2
  + Error in estimating the population mean using the point estimator Y at a given X0
* Prediction bands are calculated in a manner analogous to confidence bands.

Assessing the appropriateness of the straight-line model

* Tests for lack of fit
* Best understood in terms of polynomial regression models

**Chapter 6: The Correlation Coefficient and Straight-Line Regression Analysis**

Preview Questions

1. What is the definition of the correlation coefficient?
2. How is r used as a measure of association?
3. What is the bivariate normal distribution?
4. When is the bivariate normal distribution applied?
5. How do you measure the strength of the straight-line relationship?
6. What does r2 not measure?
7. How do you test hypotheses and confidence intervals for the correlation coefficient?
8. How do you test for the equality of two correlations?
9. What is the minimum value of r2 for a correlation to be considered significant?

Reading Summary

Definition of correlation coefficient

* Denoted by r
* Measures how two random variables are linearly associated
* Can be positive, negative, or zero
* Correlation coefficient has no dimensions
* Sample correlation coefficient, r = = 1
* Eq. for SSXY, SSX, SSY located on p. 108
* S = sample standard deviation

Correlation coefficient as a measure of association

* The larger the absolute value of r, the stronger the association
* A positive r indicates a positive association
* A negative r indicate a negative association
* A value of r close to zero indicates little if any linear association
* Association means there is a lack of statistical independence between the variables
* Sample correlation coefficient, r is an estimate of the population correlation coefficient, ρXY , which is an unknown population parameter.
* ρXY = which implies that σXY = (ρXY )(σX )(σY)
* Reference lines at the X mean and Y mean on a scatter plot help visual analysis of the data
* More data points in the positive quadrants than the negative quadrants is a good indicator of a positive association.
* More date points in the negative quadrants than the positive quadrants is a good indicator of a negative association.

The Bivariate Normal Distribution

* Joint density function
* Bell shaped density function in three dimensions
  + Univariate normal distribution looks like a cross section of a bell
  + Bivariate normal distribution looks like the bell itself
* The distribution of Y for a fixed X is the univariate normal distribution
  + Conditional distribution of Y at X
  + Random variable denoted as YX
* Mean of the YX = μ Y|X = μY + (ρXY )( ) (X - μX)
* Variance of YX = = ( )(1 - )
* = = 1 -

r2 and the Strength of the Straight-Line Relationship

* r2 = = 1 -
* Measures improvement in prediction of the random variable versus simply using the mean of the random variable.
* Measures the percent of a value of one variable that can be explained by another variable.
* Use because r can be negative whereas r2 is between 0 and 1.

What r2 Does NOT Measure

* It is NOT a measure of the magnitude of the slope of the regression line.
* It is NOT a measure of the appropriateness of the straight-line model.

Test of null hypothesis that the population correlation coefficient of two variables is zero

* H0: ρXY = 0
* Mathematically equivalent to H0: β1 = 0
* *T* = , which has t-distribution of n-2 degrees of freedom when H0 is true.
* Test statistic *T* increases in value as the sample size increases

Test of null hypothesis that the population correlation coefficient of two random variables is equal to a non-zero number

* H0: ρXY = ρ0, where ρ0 ≠ 0
* HA: ρ0 ≠ 0 is a two-tailed alternative ( Z1-α/2)
* The distribution or r is skewed when the population correlation coefficient is non-zero.
* Transform r into a statistic that often is approximately normal using the Fisher’s Z transformation
* Z = ln
* Mean of the Fisher’s Z transformation is ln
* Variance of the Fisher’s Z transformation is , if n≥20

Calculate the confidence interval for the population correlation coefficient

* ln ±

Test the null hypothesis that the population correlation coefficients are of two variables are equal

* H0: ρ1 = ρ2 using independent random samples
* HA: ρ1 ≠ ρ2 is a two-tailed alternative ( Z1-α/2)
* Random samples are both large
* Approximate using the Fisher’s Z transformation
* Calculate the Z-values for each variable
* Calculate the test statistic Z =

Test the null hypothesis that the population correlation coefficient of variable 1 and variable 2 is the same as that of variable 1 and variable 3

* H0: ρ1,2 = ρ1,3 using a single random sample
* Sample correlation coefficients calculated using the same data set
* Eq. for the test statistic Z located on p. 121 (eq. 6.14)

How Large Should r2 Be For A Correlation To Be Considered Significant

* No strict statistical guideline
* Low r2 does not necessarily invalidate certain regression analysis conclusions
* An association may exist but it may not be causal (i.e., etiologically relevant).
* Corresponding estimated effect size can help determine etiological relevance.
* Variables may be meaningful (i.e., etiologically relevant) and statistically significant but have low predictive value
  + Study sample is heterogeneous (high diversity)
  + Outcome is complex (multiple, interacting causes)
  + Some independent variables are uncommon (variable is rarely present)

**Chapter 7: The Analysis-of-Variance (ANOVA) Table**

Preview Questions

1. What is the ANOVA table?
2. For what is the ANOVA table used?

Reading Summary

Preview

* Applies to all forms of regression analysis.
* Consists of several estimates of variance.
* Estimates used to answer the inferential questions
* Inferential questions for straight-line regression analysis
  + Is the straight-line model appropriate?
  + Is the true slope of the best-fit line zero?
  + What is the strength of the straight-line relationship?
* Regression analysis and analysis of variance are closely related.

The ANOVA Table for Straight-Line Regression

* Mean-square term is the sum of squares divided by the degrees of freedom.
* F-statistic is the regression mean square by the residual mean square.
* SSY called the total unexplained variation or total sum of squares about the mean.
* SSE measures the amount of variation in the observed Y remaining after accounting for the linear effect of X.
* SSY-SSE is the variation due to regression.
* SSY-SSE also called the sum of squares due to regression.
* Total unexplained variation = variation due to regression + residual variation after regression (see Eq. 7.1 on p. 130), which holds for any regression.
* Mean-square error (residual) value is estimate of sample standard deviation, S2Y|X
* S2Y|X is an estimate of the population standard deviation, σ, if straight-line regression model is appropriate
* Mean-square regression value is estimate of population standard deviation ONLY IF variable X DOES NOT help predict variable Y (i.e., the null hypothesis slope = 0 is true).
* Mean-square residual and mean-square regression are statistically independent of each other.

**Chapter 8: Multiple Regression Analysis: General Considerations**

Preview Questions

1. What is multiple regression analysis?
2. How does multiple regression analysis differ from straight-line regression analysis?
3. How do you construct multiple regression models?
4. What are the assumptions of multiple regression analysis?
5. How do you determine the best estimate of the multiple regression equation?
6. How do you use the ANOVA table for multiple regression analysis?

Reading Summary

Preview

* Multiple regression is an extension of straight-line regression.
* Straight line regression uses only one independent variable.
* Multiple regression uses more than one independent variable.
* Multiple regression is more difficult than straight-line regression for the following reasons:
  + More difficult to choose the best model.
  + More difficult to visualize how the fitted model looks.
  + More difficult to translate the model results to real-life.
  + Manual computations are virtually impossible.

Multiple Regression Models

* Graphical representation of a multiple regression equation is a hypersurface in (k+1) dimensional space.
* For special case of k=2 (i.e., two independent variables) the regression equation is the surface (i.e., plane) described by the mean value of the dependent variable for various combinations of the independent variables.

Assumptions of Multiple Regression Analysis

* For each combination of independent variables, the dependent variable Y is random with a certain probability distribution having a finite mean and variance.
* The Y observations are statistically independent of one another.
* The mean value of Y for each combination of independent variables is a linear function.
* The regression equation is also called the response surface or regression surface.
* The variance of Y is the same for any fixed combination of independent variables.
* The variable Y is normally distributed for any fixed combination of independent variables.
  + Required for general inference making but not the fitting of the regression model.
  + If normality does not hold, transform the variables so that Y is approximately normally distributed.
  + Other regression methods are required if the Y variable is nominal or ordinal.
* The independent variables are assumed to be measured without error.

Determining the Best Estimate of the Multiple Regression Equation

* Basic approaches to estimate a multiple regression equation:
  + Least-squares approach
  + Minimum-variance approach
  + Maximum likelihood approach
* Both approaches produce the same result in straight-line regression.
* is the multiple correlation coefficient.
* Multiple regression is related to the multivariate normal distribution.

The ANOVA Table for Multiple Regression

* SSE is the sum of the squares due to error (i.e., residual sum of squares)
* SSE represents the amount of Y variation left unexplained after the independent variables have been used to predict Y.
* SSY-SSE is the regression sum of squares
* SSY-SSE measures the reduction in variation due to the independent variables in the regression equation.
* SSY is the total sum of squares
* Total sum of squares = regression sum of squares + residual sum of squares  
  SSY = (SSY-SSE)+SSE
* MS ≡ mean-square terms
* F =
* R2 =
* 0 ≤ R2 ≤ 1 and always increases as more variables are added to the model.

BRFSS Analysis Example

* Regression degrees of freedom is the number of predictor variables.
* Residual degrees of freedom = n-k-1 = (1,049 observations) – (3 predictor variables) – 1 = 1,045

**Chapter 14: Regression Diagnostics**

Preview Questions

1. How do you diagnose problems in data?
2. How do you detect outliers and violations of model assumptions?
3. How do you address violations of regression assumptions?
4. What is collinearity?

Reading Summary

Preview

* Regression diagnostics are statistical techniques for detecting conditions that can lead to invalid regression results.
  + Outliers in the data
  + Violations of regression assumptions
  + The presence of collinearity

Approaches to Diagnosing Problems in Data

* Become familiar with the characteristics of the data
  + Type of subject or experimental unit
  + Data collection procedure
  + Unit of measurement for each variable
  + Plausible range of values for each variable
  + Typical value for each variable
* Perform simple descriptive analysis
  + Numeric variables
    - Examine five largest and smallest values
      * Recording errors
      * Outliers in the data
    - Review outliers
      * Recover correct data
      * Set value to missing in analysis (impossible and highly implausible values)
      * Judge plausibility of value
  + Categorical variables
    - frequency tables
  + Continuous variables
    - Calculate measures of central tendency and range of data
    - Review scatter plots
    - Calculate Pearson correlation between pairs of variables
    - Prepare partial regression plots
      * Y for k-1 variables (see Eq. on p. 342)
      * Xk as a function of all other variables (see Eq. on p. 342)
      * Plot (-) versus (k-k)
      * Plot represents the original Y and Xk variables adjusted for the remaining (k-1) predictors
    - Calculate pairwise correlations between independent variables

Residual Analysis: Detecting Outliers and Violations of Model Assumptions

* Residuals should:
  + Be independent of each other
  + Have a mean of zero
  + Have a common variance
  + Standardized residuals, zi = , where S is the square root of the mean squared error
  + Studentized residuals, ri (Eq. on p. 348)
    - Each should follow a Student’s t-distribution with (n-k-1) degrees of freedom if regression assumptions hold
  + Jackknife residuals
* Evaluating outliers
  + Criteria used
    - Reasonableness
    - Response extremeness
    - Predictor extremeness
  + Leverages
    - Denoted by hi
    - Measure of extremeness of an observation with respect to the independent variables.
    - hi = ; 0 ≤ hi ≤ 1
    - Scrutinize al observations where hi >
  + Jackknife Residuals
    - Denoted by r(-i)
    - Outliers can mask their effects by pulling the fitted regression surface away from the main body of the data
    - See Eq. on p. 349
    - Ordinary mean square error will be larger than the mean square error computed with the *i*th observation deleted.
    - Jackknife residual will be large if an observation is an outlier in the response variable Y or the predictor space X1 … Xk
    - Has a t-distribution with n-k-2 degrees of freedom if regression assumptions hold
    - Scrutinize observations where r(-i) > 95th percentile of the relevant t-distribution
  + Cook’s Distance
    - Measures the extent to which the estimates of the regression coefficients change when an observation is deleted from the analysis.
    - See Eq. on p. 350
    - Scrutinize observations where di > 1
  + Assessing the Linearity, Homoscedasticity, and Independence Assumptions
    - Plot residuals against predicted values (i.e., residual plots)
    - Variance of residuals not constant suggest violation of homoscedasticity assumption
    - Time-related effects may suggest
      * Time is a covariate
      * Violation of the independence assumption
    - Create residual plots versus each predictor to identify the specific predictors that are of issue.
    - Patterns may be due to small sample size or sparse data for certain predictor value combinations
    - Focus on gross assumption violations
  + Assessing the Normality Assumption
    - Skewness
    - Kurtosis
    - Goodness-of-fit tests

Strategies for Addressing Violations of Regression Assumptions

* Transformations
  + Reasons
    - Stabilize the variance of the dependent variable (homoscedasticity assumption)
    - Normalize the dependent variable (normal distribution assumption)
    - Linearize the regression model (nonlinear regression coefficients)
  + Process
    - Fit linear regression equation to the transformed version of the dependent variable
    - Conduct all procedures again
    - Possible to violate formerly upheld assumptions when trying to correct other violated assumptions
  + Log transformation (Y’ = log Y, where Y>0)
    - stabilize variance when variance increases with increasing Y
    - normalize dependent variable for positively skewed Y residuals
    - linearize the regression model if relationship of Y to some independent variable suggests increasing slope
  + Square root transformation (Y’ = , where Y≥0)
    - stabilize variance when variance is proportional to the mean of Y
    - Dependent variable has Poisson distribution
  + Square transformation (Y’ = Y2)
    - Stabilize variance when variance decreases with the mean of Y
    - Normalize dependent variable for negatively skewed Y residuals
    - Linearize model if relationship of Y to some independent variable suggests decreasing slope
  + Arcsin transformation (Y’ = arcsin = sin-1 , where Y≥0)
    - Stabilize variance if Y is proportion or rate
* Weighted least-squares method
  + Assumptions of variance homogeneity or independence do not hold

Collinearity

* Strong relationship among independent variables
  + produces unreliable parameter estimates and standard errors
* Two-variable model
  + If r2X1,X2 is nearly 1.0 then collinearity is likely present.
* Multiple regression analysis
  + Variance inflation factor (VIF) used to measure collinearity
  + VIFj = , where j = 1, 2,… ,k
  + Tolerancej =
  + Scrutinize when VIFj > 10
* Resolving collinearity
  + Use computational algorithms
  + Choice of measurement and choice of origin (i.e., scaling)
  + Use alternate computational methods and exact collinearities among the predictors
    - Eigenvalues
    - Condition index, condition number, variance proportions (Eq. on p. 367)
    - CN≥30 indicates moderate to severe collinearity
    - Variance proportion of concern when two or more loadings greater than 0.9 on component with a large condition index
* Collinearity diagnostics
  + Perform simple descriptive analysis
  + Examine VIF values for each predictor
  + Examine condition index and variance proportion statistics
* Treating collinearity problems
  + Eliminate one or more of the predictors in the collinear set
    - DO NOT make decision on basis of P-values
    - May consider cost and data collection difficulty
    - Least scientifically interesting predictors
  + Use orthogonal polynomials
  + Collecting data that break the pattern of collinearity
  + Using centered data (transformed predictor variables by subtracting the mean value from each value of the variable in question)
  + Regression on principal components (original predictors replaced by mutually uncorrelated variables)
  + Ridge regression (perturbing the eigenvalues of the original predictor values cross-product matrix)

**Chapter 9: Statistical Inference in Multiple Regression**

Preview Questions

1. How do you test for the overall significance of the regression?
2. What is the partial F-test?
3. What is the multiple partial F-test?
4. How do you use partial F-tests?
5. What other methods are used for multiple regression?

Reading Summary

Preview

* Three basic types of tests
  + Overall test
    - Evaluates whether the entire set of independent variables contribute significantly to predicting Y.
  + Test for addition of a single variable
    - Evaluates whether the addition of a single independent variable add significantly to the prediction of Y over what is predicted by a model with other independent variables included.
  + Test for addition of a group of variables
    - Evaluates whether the addition of a group of independent variables add significantly to the prediction of Y over what is predicted by a model with other independent variables included.
* Tests are statistical tests of hypotheses stated in terms of the unknown regression coefficients of the model.
* Tests are F-tests but can be expressed as t-tests in some cases.
* F-tests have an F-distribution when the null hypothesis is true.
* F-tests are ratios of two independent estimates of variance, F = /
* estimates if H0 is true.
* estimates whether or not H0 is true.
* F = 1 if H0 is true; F > 1 if H0 is not true.
* Each test is a comparison of two models
  + Full model: Y = β0 + β1X1 + β2X2 + E
  + Reduced model: Y = β0 + β1X1 + E
  + H0: β2=0

Test for Significant Overall Regression

* F =
* Mean Square Regression = (SSY – SSE)/k  
   = (sum of the squares of Y – sum of residual squares) / (number of independent variables)  
   and SSY-SSE = R2
* Mean Square Residuals = SSE/(n-k-1)  
   = sum of the residual squares / (sample size – number of independent variables – 1)  
   and SSE = 1 - (SSY-SSE)
* Compare F with the critical point Fk, n-k-1, 1-α , where α is the significance level.

Partial F-Test

* Used to answer three questions
  + Does X1 alone significantly predict Y ?
  + Does the addition of X2 add significantly to predicting Y after we control (i.e., account) for the contribution of X1 ?
  + Does the addition of X3 add significantly to predicting Y after we control (i.e., account) for the contribution of X1 and X2 ?
* SS(X\*|X1 … Xp) = Regression SS(X1 … Xp, X\*) – Regression SS(X1 … Xp)  
   = Residual SS(X1 … Xp) – Residual SS(X1 … Xp, X\*)
* F = SS(X\*|X1 … Xp) / Mean Square Residual for the model with all variables
* H0 has n-p-2 degrees of freedom
* Compare F with the critical point F1, n-p-2, 1-α , where α is the significance level.
* T-test can be used for a variable added last.
  + F1,w = (Tw)2
* T = \*/
* Reject H0: β\*=0 if
  + |T| > tn-p-2, 1-α/2 🡪 two-sided test, HA: β\*≠0
  + T > tn-p-2, 1-α 🡪 upper one-sided test (right tail), HA: β\*>0
  + T < -tn-p-2, 1-α 🡪 lower one-sided test (left tail), HA: β\*<0
* Strategies for finding the best model
  + Deleting study variables one at a time (backwards strategy), usually beginning with the variable that has the smallest partial F value that is not significant.

Multiple Partial F-Test

* Used to assess the contribution of adding two or more independent variables after we control for the contribution made by other variables already in the model.
* Test procedure is straightforward extension of the partial F-test.
* H0: = = 0
* Useful for
  + Assessing groups of variables with some common trait
  + Assessing two-way product terms (also called interaction variables) such as X1X2
* Testing groups of variables can reduce the total number of tests to be performed.

Strategies for Using Partial F-Test

* Variables-added-in-order tests
  + Useful when order in which predictors enter model is important
* Variables-added-last tests
  + Useful when all variables considered important

Additional Inference Methods for Multiple Regression

* Inferences about the Y-intercept are sometimes important
  + Usually done with intercept-added-last test
* Recommended to construct confidence interval for regression coefficients.
  + \* ± tn-k-1, 1-
* Comparing the estimated value of Y for some specified set of independent variable values
  + Hypothesis tests T-statistic (see Eq. 9.8 on p. 183)
  + Confidence intervals (see Eq. 9.9 on p. 183)
* Compare a prediction interval for a new, unobserved value of Y for some specified set of independent variable values.
* Statistical inferences about a linear sum of regression coefficients.
  + e.g., varying dosages of two drugs

**Chapter 10: Correlations: Multiple, Partial, and Multiple Partial**

Preview Questions

1. What’s the difference between a multiple correlation, partial correlation, and a multiple partial correlation?
2. How do you interpret a correlation matrix?
3. What is the multiple correlation coefficient?
4. What is the relationship of Y given variables X1 through Xk in a regression model to the multivariate normal distribution?
5. What is an alternate way to represent the regression model?

Reading Summary

Preview

* Sample correlation r is an estimate of the population correlation ρ.
* The more highly positive r, the more positive the linear correlation.
* The more highly negative r, the more negative the linear correlation.
* If r is close to zero, there is little if any correlation.
* r = ()
* r2 measures the strength of the linear relation and takes on values between 0 and 1.
* r2 =
* Conditional distribution of Y given X is N(μY|X , σ2Y|X)
* μ Y|X = μY + (ρXY )( ) (X - μX)
* = = 1 -
* Concept can be extended to the multiple regression case.

Correlation Matrix

* Describes all zero-order correlation coefficients between all possible pairs of variables.
* Each correlation describes the strength of the linear relationship between the two variables.
* They DO NOT describe:
  + The overall relationship of the dependent variable to the independent variables.
  + The relationship between variables after controlling for all other variables.
  + The relationship between the dependent variable and the combined effects one group of variables after controlling for another group of variables.

Multiple Correlation Coefficient

* Measure of the overall linear relationship between one dependent variable and several independent variables.
* Is always a positive value.

Relationship of RY|X1,X2,…Xk to the Multivariate Normal Distribution

* Sample multiple correlation coefficient is an estimator of a population correlation coefficient.
* Analogous result to two variables case assuming that joint distribution is multivariate normal.
* The conditional distribution of Y given X1 … Xk is a univariate normal distribution.
* RY|X1,X2,…Xk is the correlation of which is always a positive value.
* RY|X1,X2,…Xk is an estimate of ρY|X1,X2,…Xk

Partial Correlation Coefficient

* Measure of the strength of the linear relationship between two variables after we control for the effects of all other variables.
* Denoted by rYX|C1…Cq and sometimes referred to as the full partial correlation coefficient
* Use partial F-test to test whether adding a variable to the regression model significantly improves the model given that certain other variables are already in the model.
  + Calculate F statistic = F(X|C1…Cq)
  + Reject null hypotheses is F statistic is greater than the appropriate critical value.
  + Critical value determined based on numerator DOF, denominator DOF, and selected confidence level.
  + H0: ρYX|C1,C2,…Cq = 0
* For r2YX|C in context of residual sum of squares, see Eq. 10.2 on p. 208
* For rYX|C in context of correlation coefficients, see Eq. 10.3 on p. 208
* The partial correlation between Y and X, controlling C is the correlation of the residuals of the straight-line regressions of Y on C and of X on C, rYX|C = rY-Y^, X-X^
* Semipartial correlation coefficient is the correlation between two variables when only one of them has been adjust for a third variable, denoted as rY(X|C)
* For semipartial correlation coefficient equations, see Eq. 10.5 and 10.6 on p. 210.

Alternative Representation of the Regression Model

* Regression model can be expressed in terms of partial correlation coefficients and conditional variances (see Eq. 10.7 on p. 212).

Multiple Partial Correlation

* Describes the overall relationship between the dependent variable and two or more independent variables while controlling for other variables (see Eq. 10.8 on p. 213).
* Rarely of interest in most cases.
* Used mostly for testing hypotheses about a group of higher-order terms.
* For F statistic calculation, see Eq. 10.9 on p. 214.

The following null hypotheses all state the same thing in different ways:

* Prediction hypotheses
  + H0: Adding variables to the smaller model to form the larger model does NOT significantly improve the prediction of Y.
  + H0: The population regression coefficients for the variables in the larger model are all equal to zero.
* Association hypotheses
  + H0: The population multiple partial correlation between Y and variables added to produce the larger model is zero once we control for the variable in the smaller model.
  + H0: The value of the population squared multiple correlation coefficient for the larger model is NOT greater than the value of that parameter for the smaller model.

**Chapter 11: Confounding and Interaction in Regression**

Preview Questions

1. What is interaction in regression?
2. What is confounding in regression?
3. How do you deal with interaction and confounding in regression?

Reading Summary

Preview

* Goals of regression
  1. Predict the dependent variable using a set of independent variable.
  2. Quantify the relationship between the dependent variable and one or more independent variable.
* Confounding and interaction are most relevant to the second goal.

Overview

* For continuous variables, the measure of association is usually a regression coefficient.
  + Additional variables not of particular interest but are accounted for in the model are called extraneous variables, control variables, or covariates.
  + Is the estimate of association between a dependent variable and a set of independent variables meaningfully different depending on whether or not we ignore extraneous variables? 🡪 Confounding
  + Is the estimate of the association between a dependent variable and a set of independent variables meaningfully different for different values of extraneous variables? 🡪 Interaction
* Assessing confounding in regression generally involves comparing a crude estimate of association excluding the extraneous variables with an adjusted estimate of association including the extraneous variables.
* Assessing interaction in regression generally involves describing the relationship between the dependent variable and the set of independent variables at different levels of the extraneous variables.
* Assess interaction before assessing confounding.
* Use a summary (adjusted) estimate that controls for confounding only when there is no meaningful interaction.

Interaction in Regression

* Complete factorial experiment is when you collect observations for all combinations of setting for the independent variables (i.e., factors).
* When there is no interaction you can evaluate the relationship between the dependent variable and the independent variables independently of one another.
  + The effect of changing the level of one of the independent variables only shifts the straight line describing the relationship between the dependent variable and the other independent variable up or down.
  + Response curves of the dependent variable versus each independent variable are parallel (i.e., parallelism).
  + See Eq. 11.1 on p. 229.
* When there is interaction, the independent variables do not operate independently of one another.
  + See Eq. 11.2 on p. 230.
* Interaction in regression model described by product terms.
* Approaches for including product terms in regression model:
  + Include only interactions that are reasonable a priori (i.e., based on theory prior to analysis)
  + Include a full set of interactions
    - The higher the order of interactions the more difficult it is to interpret the meaning of the model.
    - If a higher order interaction is specified in a model, then all lower order interactions must be specified (i.e., hierarchically well-formulated [HWF])
    - A model with an intercept term CANNOT contain more than n-1 independent variables, where n is the total number of observations in the data.
  + Include only interactions with the primary factors
    - Approaches
      * Test globally for any interaction then identify particular interaction terms of importance.
      * Begin with higher-order terms then proceed to lower-order terms if higher-order terms are not significant.
  + Interaction versus effect modification
    - Interaction is a statistical property of a mathematical model.
    - Effect modifiers are control variables that modify an outcome depending on their values.
    - Effect modification (also called effect measure modification) is the meaning we assign to the presence of interaction among variables under consideration.
    - The terms effect modification and interaction used interchangeably.

Confounding in Regression

* Assumes no interaction is present.
* Controlling for one extraneous variable
  + Crude estimate is Y = β0 + T + E (Eq. 11.6)
  + Adjusted estimate is Y = β0 + T + β2C + E (Eq. 11.5)
  + Confounding present if ≠ 🡪 subjective determination of meaningful difference
  + and estimate β1
* Recommendations for determining which variables for which to control.
  + One approach is to control for any variable as a confounder that changes the crude estimate by some pre-specified amount determined by clinical judgement.
  + Create list of eligible variables (risk factors) based on prior knowledge and theoretical research.
  + Only consider variables known to be reasonably associated with the dependent variable.
  + Determine confidence interval for β1 with and without control variables (i.e., precision)
    - Confounding takes precedence over precision.
* Controlling for several extraneous variables
  + ≠
  + Only a subset of C1,C2,…,Cp may be required for adequate control.
  + Consider possible gains in precision.
  + No subset should be considered unless if gives nearly the same adjusted-effect estimate as obtained when controlling for all Cp’s or produces a large increase in precision.

**Chapter 12: Dummy Variables in Regression**

Preview Questions

1. What is a dummy variable?
2. How are dummy variables different from control variables?
3. What are the rules for defining dummy variables?
4. How do you compare two straight-line regression equations?
5. What are the most important questions to answer when comparing two straight lines?
6. What methods are used to compare two straight lines?
7. How do you interpret tests comparing two straight lines?
8. How do you compare four regression equations?
9. How do you compare several regression equations involving two nominal variables?

Reading Summary

Preview

* Methods of regression analysis can be generalized to include categorical predictors (nominal and ordinal variables).

Definitions

* A dummy variable is also called an indicator variable.
* A dummy variable takes on a finite number of values.
* Values taken on by dummy variables only indicate categories of interest, meaningful measurements.

Rule for Defining Dummy Variables

* Avoids collinearity
* If the nominal independent variable of interest has k categories and the regression model contains an intercept (i.e., constant term) then:
  + Exactly k-1 dummy variables must be defined
* If the nominal independent variable of interest has k categories and the regression model DOES NOT contain an intercept (i.e., constant term) then:
  + Exactly k dummy variables must be defined
* In reference cell coding each dummy variable takes on only values of 1 and 0.

Questions for Comparing Two Straight Lines

* +Three basic questions:
  + Are the two slopes the same regardless of the values of the intercepts?
  + Are the two intercepts the same regardless of the slopes?
  + Are the two lines the same (i.e., coincident)?

Methods for Comparing Two Straight Lines

* Two basic methods
  + Treat the data for each category separately and fit two separate regression equations.
    - Conduct appropriate two-sample t-tests.
  + Define a dummy variable to distinguish between the categories.
    - Single regression equation Y = β0 + β1X + β2Z + β3XZ + E
    - For Z=0: YA = β0 + β1X + E
    - For Z=1: YB = β0 + β1X + β2 + β3X + E = (β0 + β2) + (β1 + β3)X + E
    - β0A = β0 and β0B = (β0 + β2)
    - β1A = β1 and β1B = (β1 + β3)

Using Separate Regression Fits to Compare Two Straight Lines

* Testing for parallelism
  + H0: β1A= β1B
  + T = (1A- 1B) /
  + To interpret T statistic, see equations at bottom of p. 264
* Testing for equal intercepts
  + H0: β0A= β0B
  + Use the same T-statistic formula for parallelism substitution the appropriate variables.
  + To interpret T statistic, see equations in middle of p. 266
* Testing for coincidence
  + Slope and intercepts are equal
  + One method is to pool all observations of the categories for regression
  + Alternatively, evaluate null hypothesis for equal intercepts and slopes; if either is rejected then conclude there is no coincidence.
    - Does not precisely test for coincidence
    - There is more chance for a type 1 error

Using a Single Regression Equation to Compare Two Straight Lines

* Testing for parallelism
  + β1A = β1 and β1B = (β1 + β3) 🡪 H0: β3 = 0
  + Use partial F-test with mean square residual for model containing only X and Z
  + F-statistic is the square of the T-statistic
* Testing for equal intercepts
  + β0A = β0 and β0B = (β0 + β2) 🡪 H0: β2 = 0
  + Use partial F-test with mean square residual for the full model
* Testing for coincidence
  + H0: β2 = β3 = 0
  + Use simple F-test comparing single regression equation and the reduced equation.

Comparison of the Two Methods

* The tests for parallel lines are exactly equivalent.
* The tests for coincident lines differ; method using dummy variables is generally preferred.
* The method using dummy variables is usually easier to implement in statistical software packages.

Testing Strategies and Interpretation When Comparing Two Straight Lines

* Prefer backwards strategy
  + Start with largest model of interest then reduce through a sequence of hypothesis tests.
  + See flow diagram in Fig. 12.3 on p. 273
* Process
  + Test for coincidence and stop if not significant.
  + Test for parallelism.
  + Test for same intercept using variables-added-last F-statistic.
* Other Dummy Variable Models
  + Effect coding
  + See Equations 12.9 on p. 273 and 12.10 on p. 274
  + Test of null hypothesis differs from the above methods because coefficients of null hypothesis are not equal to 0.

Comparing Four Regression Equations

* Define the requisite number of dummy variables.
* See Eq. 12.11 on p. 275
* Hypotheses are of interest:
  + All four regression equations are coincident.
    - Test statistic is the multiple partial F-statistic shown on p. 276.
  + All four regression equations are parallel.
    - Test statistic is the multiple partial F-statistic shown on p. 276.
* Can use a model with effect coding dummy variables.
  + See Eq. 12.12 on p. 277.

Comparing Several Regression Equations Involving Two Nominal Variables

* X1 and X2 are continuous variable; Q and SC which uses a dummy variable Z for effect coding.
* See Eq. 12.13 on p. 278.
* Hypotheses of interest:
  + All eight regression equations are coincident.
  + All eight regression equations are parallel.
  + YX1 and YX2 regression equations are coincident.
  + YX1 and YX2 regression equations are parallel.
  + YX1 and YX2 regression equations are parallel controlling for SC.
  + All four SC equations are parallel controlling for X1.

**Chapter 13: Analysis of Covariance and Other Methods for Adjusting Continuous Data**

Preview Questions

1. What is the adjustment problem?
2. What is analysis of covariance?
3. Why is the assumption of parallelism a potential drawback?
4. How do you conduct analysis of covariance with several groups and several covariates?
5. How do you conduct analysis of covariance with several nominal independent variables?
6. What do we need to be cautious about when conducting analysis of covariance?

Reading Summary

Preview

* Reasons for controlling for certain variables:
  + Assess interaction
  + Correct for confounding
  + Increase the precision of the estimate
* Basic approach
  + Fit regression model with study factors of interest, important control variables, and product terms of these variables (**e.g., control variables?**) if necessary
  + Determine the effects of the study factors adjusted for the control variables
* Process is done using analysis of covariance (ANOCOVA) technique
  + Study factors are treated as nominal variables through dummy variables
  + Variables being controlled (covariates) are continuous interval variables
  + Dependent variable is a continuous interval variable
  + Assumes there is no interaction of covariates with study variables (should be assessed)

Adjustment Problem

* Questions
  + Is the true-straight line relationship Y and X the same for populations NA and NB?
  + Do the mean Y’s for NA and NB differ after taking into account the possible confounding effects of different distributions between the two populations?
* First question considered in Chapter 12 (Y = β0 + β1X + β2Z + β3XZ + E)
  + The lines are coincident (H0: β2 = β3 = 0)
  + The lines are parallel but not coincident (H0: β3 = 0 and H0: β2 ≠ 0)
  + The lines are not parallel (H0: β3 ≠ 0)
* Conclusions from answering the first question enable us to make inferences about the answer to the second question.
* Question arises regarding the most appropriate method to adjust for different for NA and NB

Analysis of Covariance

* Y = β0 + β1X + β2Z + β3XZ + E 🡪 β0 + β1X + β2Z + E because of parallelism
* X is the covariate
* Z is a dummy variable that indexes to two groups to be compared
* Adjusted means for the two groups is the predicted Y values when Z = 0 and Z = 1 when X is set equal to the overall mean X for the two groups.
* Partial F-test of the H0: β2 = 0 is used to determine if the adjusted means are significantly different.
* Two alternative formulas for computing the adjusted means (see Eq. 13.5 on p. 311)

Assumption of Parallelism

* The regression lines in the ANACOVA method may have different slopes
* Conduct a test for parallelism before performing ANACOVA
  + Test H0: β3 = 0 for the complete model
* If not parallel, then make no adjustment between the means for the two groups
  + The two regression lines describe very different relationships.

Analysis of Covariance for Several Groups and Several Covariates

* ANACOVA can be used to provide adjusted means when
  + There are several groups (s) and 🡪 s-1 dummy variables
  + It’s necessary to adjust simultaneously for several covariates (q) 🡪 q covariates
  + Regression model is shown in Eq. 13.6 on p. 314
* To test for different adjusted means
  + H0: βq+1 = βq+2 = … = βq+(s+1) = 0
  + Use multiple partial F-test

Analysis of Covariance for Several Nominal Independent Variables

* When interested in the effects on a dependent variable of nominal predictor variables with each having two levels
* Want to include other variables known to influence the dependent variable
* Consider the four combinations for the two predictor variables and compute the corresponding adjusted means

Comments and Cautions

* ANACOVA artificially assumes that all groups have the same set of mean covariate values
  + Equivalent to assuming a common covariate distribution based on the combined sample over all groups
  + Means and entire distribution of covariates in combined sample assumed to be the same as in each group
* ANACOVA is inappropriate when there is non-parallelism
  + No adjustment of means when there is interaction 🡪 quantify the nature of the interaction
* Validity and Precision
  + Validity is achieved by adjusting for confounding; should be considered first.
  + If no confounding in data, can still control for covariates to gain precision.
    - The smaller the variances of the estimators of association (i.e., confidence intervals), the greater the precision
* Alternatives to ANACOVA
  + We can fit a model containing the covariates and the study variables without making the study variable categorical (i.e., nominal)
  + However, it’s impossible to obtain adjusted means for groups unless the study variables are defined categorical even if there is no interaction
  + Predicted values based on the best regression model can be treated as adjusted values because covariates are being taken into account
  + Adjusted means for distinct values of continuous study variables can be obtained by computing predicted values using the overall mean covariate values of the best model.
  + If all study variables and covariates are treated as categorical we can treat the regression model as a two-way analysis-of-variance (ANOVA) model with unequal cell numbers
    - Might be inappropriate if the underlying measurement scales of some of the covariates are continuous but arbitrarily categorized.

**Chapter 15: Polynomial Regression**

Preview Questions

1. What are polynomial models?
2. How do you perform the least-squares procedure for fitting a parabola?
3. How do you interpret the ANOVA table for second-order polynomial regression?
4. What inferences are associated with second-order polynomial regression?
5. How do you fit and test higher-order models?
6. What is a lack-of-fit test?
7. What are orthogonal polynomials?
8. What strategies are used to choose a polynomial model?

Reading Summary

Chapter Preview

* The polynomial model is a special case of the multiple regression model.
  + Often used when only one basic independent variable is to be considered.
  + Any polynomial model can be represented on two-dimensional graph rather than as a surface in higher-dimensional space because only one basic independent variable is being used.
* Determine whether prediction can be improved significantly by increasing the complexity of the fitted straight-line model.
* Second-order polynomial is the simplest extension of the straight-line model.
* Rename X as X1 and X2 as X2
* Methods for fitting models and inference are essentially the same as those for multiple regression.

Polynomial Models

* y = c0 + c1X + c2X2 + … + ckXk
* Parabolic model
  + k = 2
  + μY|X = β0 + β1X + β2X2
  + Y = β0 + β1X + β2X2 + E

Least-squares Procedure for Fitting a Parabola

* Minimize the sum of the squares of deviations (errors)

ANOVA Table for Second-Order Polynomial Regression

* Variables-added-last test for each term should be avoided
* Use variables-added-in-order tests

Inferences Associated with Second-Order Polynomial Regression

* Three basic inferential questions
  + Is the overall regression significant (versus using as the estimate)?
  + Does the second-order model provide significantly more predictive power than the straight-line model?
  + If the second-order model is more appropriate than a straight-line model, should we add higher-order terms to the second-order model?
* Test for overall regression and strength of the overall parabolic relationship.
  + H0: β1 = β2 = 0
  + Use F-statistic
  + Use R2 as quantitative measure of how well the model predicts the dependent variable.
* Test for the addition of the X2 term to the model
  + H0: β2 = 0
  + Use F-statistic
  + Use R2 as quantitative measure of how well the model predicts the dependent variable.
* Testing for adequacy of the second-order model
  + Use lack-of-fit test
  + Partial or multiple partial F-test for additional terms
* Fitting and testing higher-order models
  + Same methods apply to all higher-order polynomial models
  + Consider the number of relative extrema (bends of the curve)
    - Number of bends = k-1
    - Substantial theoretical evidence should exist to support using more than 3 terms.
  + Quantity of data limits the maximum order of a polynomial that may be fitted.
    - One less than the number of distinct X-values.
    - Replicate observations count only once.
* Lack-of-fit tests
  + LOF test evaluates a model more complex than the one under primary consideration
  + Classical LOF test can only be applied if there are replicate observations
  + H0: regression coefficients for the highest-order model all equal zero
  + Use a multiple partial F-test of the null hypothesis
  + Use only orthogonal polynomials for LOF test
* Orthogonal polynomials
  + Natural polynomials are simple polynomials by themselves.
  + Orthogonal polynomials are defined in terms of the simple polynomials
  + Use orthogonal polynomials to avoid collinearity inherent in using natural polynomials
    - e.g., X1\* = a01 + a11X ; X2\* = a02 + a12X + a22X2
  + Each simple polynomial can be rewritten as a linear combination of the orthogonal
  + Simple polynomials are highly correlated with one another but the orthogonal polynomials are pairwise uncorrelated.
  + Transforming from natural to orthogonal polynomial
    - Use Table A.7 only if
      * the predictor values are equally spaced
      * the same number of replicates occurs at each value
    - Otherwise, use a computer program
  + **What does it mean to center X-variable to reduce collinearity?**
    - Centering does NOT solve collinearity problem for higher-order models.

Strategies for Choosing a Polynomial Model

* Forward-selection strategy
  + Can produce misleading results
  + Can lead to under-fitting data (final model of an order less than required)
* Backward-elimination strategy
  + Avoids under-fitting bias
  + May lead to over-fitting data (final model of an order higher than required)
  + Lose some statistical power, but loss is usually negligible.
  + **What does it mean to lose statistical power?**
* Conduct residual analysis iteratively throughout the model selection process.
  + Need for higher-order model often appears as a nonlinear trend in residuals.
  + Use plot of jackknife residuals against X.
  + **What are jackknife residuals?**

**Chapter 16: Selecting the Best Regression Equation**

Preview Questions

1. What are the steps in selecting the best regression equation when the goal is prediction?
2. How do you select the most valid model?

Reading Summary

Chapter Preview

* The meaning of “best-fitting” line depends partly on the overall goal for modeling.
* Two basic goals:
  + Find a model that provides the best prediction of the dependent variable given a set of independent variables (i.e., prediction).
  + Obtain accurate estimates for one or more regression coefficient parameters in a model and then make inferences about the parameters of interest (i.e., validity).

Steps in Selecting the Best Regression Equation: Prediction Goal

* Specify the maximum model.
* Specify a criterion for selecting a model.
* Specify a strategy for selecting variables.
* Conduct the specified analysis.
* Evaluate the reliability of the model chosen.

Step 1: Specifying the Maximum Model

* The model having the most predictor variables being considered at any point in the process of model selection.
* All other possible models can be created by deleting predictor variables from the maximum model.
  + Restriction of the maximum model.
* Minimizes the chance for a Type II error (i.e., false negative which is accepting the null hypothesis that the regression coefficient equals zero).
* Over-fitting may introduce harmful collinearity.
* Reliability pushes towards a small maximum model.
  + Avoid Type I error (i.e., false positive which is including a predictor that has a population regression coefficient equal to zero.)
  + Don’t want to include practically unimportant but statistically significant predictors.
* d.f. error = n – k – 1 > 0 OR n > k + 1
  + n is the number of observations
  + k is the number of predictors
* There is at least one perfect collinearity when there are negative error degrees of freedom
* Rules of thumb
  + n –k – 1 ≥ 10 OR n ≥ 10 + k + 1
  + n ≥ 5k OR n ≥ 10k

Step 2: Specify a Criterion for Selecting a Model

* A selection criterion is an index that can be computed for each candidate model and used for comparison.
* No single criterion is always best; consider more than one.
* Types of differences
  + Numerical differences
  + Statistically significant differences
  + Scientifically important differences
* Don’t just rely on the sample squared multiple correlation R2
  + Always increase with the addition of predictors
  + Always largest for the maximum model
* Use F-test statistic for comparing the full and restricted models
* Mallow’s Cp is another alternative criterion.
  + Helps decide how many variables to include in the best model
  + Cp = (k-p)Fp + (2p – k + 1)

Step 3: Specify a Strategy for Selecting Variables

* Basic methods
  + Forward-selection method
  + Backward-elimination method
* All possible regressions procedure
  + Preferred over any other selection strategy.
  + Only method guaranteed to find the most predictive model with ideal number of variables.
  + Can become impractical when there is a large number of variables.
  + Process
    - Fit each possible regression equation associated with all possible sets of the independent variables.
    - Assemble the fitted models into sets.
    - Order the models within each set according to some criterion.
    - Use various criteria to select the best model.
* Backward-elimination procedure
  + Determine the fitted regression equation containing all independent variables.
  + Determine the partial F-statistic for every variable in the model as though it were the last variable to enter the model.
  + Determine the P-values associated with the test statistics.
  + Evaluate the variable with the lowest observed partial F-statistic.
  + Compare the P-value with the selected significance level
    - Remove variable P-value is greater than selected significance level, recalculate the regression for the remaining variables, repeat backward-elimination
    - Stop if the variable P-value is less than the selected significance level
* Forward-selection procedure
  + Select the first variable to enter the model
    - Variable most highly correlated with the dependent variable.
    - Largest F-statistic and smallest P-value of all possible single-variable models.
  + Fit the associated straight-line regression equation.
  + Perform F-test.
    - Stop if F-statistic is not significant.
    - Include variable in model if F-statistic is significant.
  + Determine the partial F-statistic and P-value associated with each remaining variable for the regression equation containing the variables previously included in the model and each remaining variable separately.
  + Evaluate the variable with the smallest partial F-statistic.
  + Perform partial F-test
    - Stop if partial F-statistic is not significant.
    - Include variable in model if partial F-statistic is significant.
  + Repeat steps 4-6 above.
* Stepwise regression procedure
  + Modified version of forward-selection procedure with re-examination of the variables added into the model in previous steps at every step along the way.
  + Variable that was included early in the process may become unimportant later in the process because of its relationship with other variables added to the model.
  + Conduct partial F-test as variable-added-last for each variable in the model at each step.
* Chunkwise methods
  + Chunks are sets of predictor variables that are logically related and equally important relative to the other variable in the chunk.
  + Impose an order on chunk selection simplifies the analysis.
  + Incorporates prior scientific knowledge into the analysis.
  + Reduces the number of possible models to be evaluated.

Step 4: Conducting the Analysis

* Check the goodness of fit of the model chosen.
* Demonstrate that the model chosen is reasonable for the data.

Step 5: Evaluating Reliability with Split Samples

* Use split sample to ensure that the model can be reliably applied to other samples.
* Most rigorous approach is to use a new data set randomly selected from the population under study.
  + Expensive
  + Not always feasible
* Randomly assign each observation to one of two groups.
  + Analysis group
  + Holdout group
* Stratified random splitting
  + Find pairs of subjects that are as similar as possible
  + Randomly assign pair members to each of the two groups
  + Tends to produce unrealistically optimistic evaluation of model reliability.
* Calculate cross-validation correlation (2)
* Calculate shrinkage = R2(1) - (2)
  + Shrinkage ≥ 0.90 indicates the model is unreliable
  + Shrinkage < 0.10 indicates a very reliable model

Selecting the Most Valid Model

* Use variant of the backward-elimination procedure.
  + Select variables to include in the maximum model.
  + Fit the maximum model.
  + Determine interaction effects in the maximum model.
  + Evaluate confounding in the maximum model.
  + Determine which subsets of control variables produce the most precise estimated regression coefficients for the main variables of interest.
* Alternative notation
  + EVW model
    - E variables are the main independent variables.
    - V variables are potential confounder variables.
    - W variables are the potential effect-modifiers of the associations between the E variables and the dependent variable.
      * Only variables involved in product terms with the main E variable are considered to be W variables
  + Model must be hierarchically well-formulated (HWF)
    - All W variables in EW product terms must also be included as lower-order V variables.
  + May need to force certain control variables into the model for political reasons.
  + Consider precision
    - Width of confidence interval for regression coefficients when there are no EW terms.
    - Certain V variables may negligibly increase precision and thus be deemed unnecessary in the final model.