**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