study

My study notes will draw heavily from the required texts and multimedia. I will also draw from external sources that I find to be adept at explaining a particular topic. If something is referenced here, it is because I found it to be very useful in understanding a topic.

Standard Mathematical and Statistical Notation

Notes below are from the following sources; [Bhatti 2011].

Vector and Matrix Notation

- A *scalar* is a number. *Scalars* are represented by lower case letters from the beginning of the alphabet such as a, b, c etc.
- A *vector* is a $n \times 1$ array defined with the mathematical operations of addition and multiplication. The standard convention is for all vectors to be column vectors, i.e. they are 'long' with n rows and 1 column. Vectors are represented as **bold** faced lower case letters frequently from the end of the alphabet, such as $\mathbf{x}, \mathbf{u}$, and $\mathbf{v}$. the ith entry of a vector $\mathbf{u}$ is denoted by $\mathbf{u}[i] = u_i$.
- A *matrix* is a $n \times m$ array defined with the mathematical operations of addition and multiplication. Matricies are represented by a bold face upper cased letter such as $\mathbf{A}, \mathbf{W}, \mathbf{X}$, etc. The (i, j)th entry of a matrix $\mathbf{A}$ is denoted by $\mathbf{A}[i, j] = a_{ij}$.
- The transpose of a $(n \times 1)$ column vector $\mathbf{a}$ is the $(1 \times n)$ row vector $\mathbf{a}^T = [a_1 \dots a_n]$. Sometimes the transpose $\mathbf{a}^T$ is denoted by $\mathbf{a}'$.
- The transpose of a $(n \times m)$ matrix $\mathbf{A}$ is the $(m \times n)$ matrix $\mathbf{A}^T$ where $\mathbf{A}[i, j] = \mathbf{A}^T[j, i]$. When a matrix is transposed, the rows become the columns and the columns become the rows.
- It is preferred to use the T notation $\mathbf{a}^T$ instead of the "prime notation" $\mathbf{a}'$.

Random Variable Notation

Random variables are how you develop calculus based probability theorey.Random variables are the unkown statistical experiement that generate data. Statistical theory is based upon the concept of random sampling.

- Random variables are denoted by capital letters from the end of the alphabet such as U, V, X, Y, or Z.
- The *observed value* of a random variable is denoted by the lower cased counterpart such as u, v, x, y, or z.
- When we have a *random sample* of independent and identically distributed (iid) random variables, we will index the variables in a set such as $X_1, X_2, \dots, X_n$ for the random variables and $x_1, x_2, \dots, x_n$ for the observed values.
- Random variables are used to devleop statistical estimators. Observes values of random variables are used to compute statistical estimates.
- Random variable notation can become convoluded when we move to multivariate random variables. Pay attention to how an author presents these concepts in text.

'Distribution'

The term *distribution* is used throughout all statistical applications and discussions. Loosly speaking, the term distribution is meant to describe how a group of values are related to either each other or to the range of values on which they are defined (their *support*).

The term *distribution* is used rather sloppily. If you don't understand the context you wont understand the use. The term *distribution* is mapped to many related concepts. In general the term *distribution* is related to the characterization of a random variable, or data generated by a random variable.

There are many mathematical notations for characterizing a statistical distribution. The choice of characterization will depend on the context and the existence of the characterization. A random variable can be characterized by any of the following functions.

- The *cumulative distribution function* (cdf), denoted by $F(x) = Pr(X \leq x)$. the cdf will exist for all random variables, and in general is why we use the term "distribution" so looslely throughout statistics. cdf exists for all random variables. From data you can always estimate a distribution function.
- The *probability density function* (pdf) for continuous random variables, denoted by $f(x)$, or the *probability mass function* (pmf) for discrete random variables, denoted by $p(x)$. Note that neither of these functions are guaranteed to exist. A random variable that can be described with a cdf will not always possess a pdf or pmf.
- Transformation functions such as the moment generating function $m(t) = \mathbb{E}[\exp(tX)]$ and the characteristic function $\phi(t) = \mathbb{E}[\exp(itX)]$. Transformation functions are not used when working with data, they may be used to develop conceptual underpinnings of modeling.
- Specialized representations for particular applications such as the *hazard fucntion* $h(t) = \frac{f(t)}{S(t)}$ and the *survival function* $S(t) = 1 - F(t)$ used in Survival Analysis. Survival function is a simple map of the cdf. The hazard function allows you to get a generic representation of a survival function.
- In data analysis distributions can be analyzed using the empirical cdf, the histogram, the Quantile-Quantile plot, and the Kolmogorov-Smirnov test.
- If you need to assess the distribution of residuals in linear regression and compare that to the assumption that they are normally distributed.

Mathematical Expectation

Mathematical Expertation is the theoretical averaging of a random variable with respect to its distribution function. In this sense the pdf of pmf act as a weight function that allows you to find the "center" of the distribution.

For a continuous random variable X with pdf function $f(x)$, the mathematical expectation of X can be computed by

$$\mathbb{E}[X] = \int x f(x) dx$$

For a discrete random variable X with pmf function $p(x) = Pr(X = x)$, the mathematical expectation of X can be computed by

$$\mathbb{E}[X] = \sum_x x p(x)$$

$\mathbb{E}[X]$ is also referred to as the first moment of X.

Expectation, Variance, and Covariance as Mathematical Operators

Let X denote a random variable. Consider the affine transformation $aX + b$.

- $\mathbb{E}[aX + b] = a\mathbb{E}[X] + b$
- $\text{Var}[aX + b] = a^2\text{Var}[X]$

Let X and Y be random variables with a joint distribution function. (In the continuous case we would denote this joint distribution function by the join density function $f(x, y)$.) Consider the linera transformtions aX and bY.

- $\mathbb{E}[aX + bY] = a\mathbb{E}[X] + b\mathbb{E}[Y]$
- $\text{Var}[aX + bY] = a^2\text{Var}[X] + b^2\text{Var}[Y] + ab\text{Cov}[X, Y]$

Here the reader should note that in general $\text{Cov}[aX + b, cY + d] = ac\text{Cov}[X, Y]$. If X and Y are independent random variables, then $\text{Cov}[X, Y] = 0$. The converse of this statement is not true except when both X and Y are normally distributed. In general $\text{Cov}[X, Y] = 0$ does not imply that X and Y are indepdented random variables.

Study Questions for Ordinary Least Squares Regression

Question: When we refer to a 'simple linear regression', to what type of model are we referring? How does a 'simple linear regression' differ from a 'multiple regression'?

Response from [Montgomery et al. 2012] pages 2 and 4.

$$y = \beta_0 + \beta_1 x + \varepsilon$$

The equation abovce is called a **linear regression model**. Customarily x is called the independent variable and y is called the dependent variable. However, this often causes confustion with the concept of statistical independence, so we refer to x as the **predictor** or **regressor**and y as the **response** variable. Because the equation above involves only one **regressor** variable, it is called a **simple linear regression model**.

In general the response variable y may be related to k regressors, $x_1, x_2, \ldots, x_k$, so that:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_x + \varepsilon$$

This is called a **multiple linear regression model** because more than one regressor is involved. The adjetive linear is employed to indicate that the model is linear in the parameters $\beta_0, \beta_1, \ldots, \beta_k$, not because the y is a linear function of the x's.

Question: In statistics, and in this course, we use the term 'regression' as a general term. What do we mean by the term 'regression'? What is the objective of a 'regression model'?

Response from [Montgomery et al. 2012] pages 1.

Regression analysis is a _statistical technique **for investigating and** modeling the relationship between variables___. The goal of regression analysis is to determine the values of parameters for a function that causes the function to best fit a set of data. Regression analysis helps one understand how the typical value of the dependent variable changes when any one of the independent variables is varied, while the other independent variables are held fixed [Wikipedia 2015].

Question: What do we mean by 'linear regression'? represent a linear regression? Which equations represent a linear regression?

(a) $y = \beta_0 + \beta_1 x_1$
(b) $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2^2$
(c) $y = \beta_0 + \exp(\beta_1) x_1$

Both (a) and (c) are representations of linear regression because they only have a single **regressor**. (b), with the presence of x_2^2 has two **regressors**.

Question: Before building statistical models, it is a common and preferred practice to perform an Exploratory Data Analysis (EDA). What constitutes an EDA for a simple linear regression model? Is this EDA satisfactory for a multiple regression model, or do we need to change or extend the EDA? As we move forward in this course we will also learn about logistic regression models and survival regression models, will these methods need their own EDA or is EDA general to all statistical models?

Question: In the simple linear regression model what is the relationship between R-squared and the correlation coefficient rho?

Question: How do we interpret a regression coefficient in OLS regression?

Question: Frequently, as a form of EDA for OLS regression we make a scatterplot between the response variable Y and a predictor variable X. As an assumption of OLS, the response variable Y must be continuous. However, the predictor variable X could be continuous or discrete. When the predictor variable is discrete, does a scatterplot still make sense? If not, what type of visual EDA does make sense? Does the appropriateness of the scatterplot make sense if the discrete variable takes on many discrete values (such as the set of integers, think of dollar amounts rounded to the nearest dollar) versus only a few discrete values(such as a coded categorical variable which only takes the values 1, 2, or 3)?

Question: The simple linear regression model is a special case of 'Multiple Regression' or 'Ordinary Least Squares'(OLS) regression. (We will typically use the term OLS regression.) What are the assumptions of OLS regression? In the final step of a regression analysis we perform a 'check of model adequacy'. What model diagnostics do we use to validate our fitted model against the model assumptions of OLS regression?

Question: How are the parameters, i.e. the model coefficients, estimated in OLS regression? How does this relate to maximum likelihood estimation? How do you show the relationship between OLS regression and maximum likelihood estimation?

Question: What is the overall F-test? What is the null hypothesis and what is the alternate hypothesis? The overall F-test is also called the 'test for a regression effect'. Why is it called this?

Question: What is the difference between R-squared and adjusted R-squared? How is each measure computed, and which measure should we prefer? How does the interpretation of R-squared change as we move from the simple linear regression model to the multiple regression model?

Question: The simple linear regression model $Y = b_0 + b_1 * X_1$ has three parameters. Two of the parameters are b_0 and b_1. What is the third parameter?

Question: What is a sampling distribution? What theoretical distribution do the parameter estimates have in OLS regression? What distribution do we use in practice? Why do we use a different distribution in practice?

Question: The final step of a regression analysis is a 'check of model adequacy'. This 'check of model adequacy' or 'goodness-of-fit' is a very important step in regression analysis. Why? Which quantities in the regression output are affected when the fitted model deviates from the underlying assumptions of OLS regression?

Question: Nested Models: Given two regression models M1 and M2, what does it mean when we say that 'M2 nests M1'?

Question: What is the Analysis of Variance Table for a regression model? How do we interpret it and what statistical tests and quantities can be computed from it?

Question: When the intercept is excluded in a regression model, how does the computation and the interpretation of R-squared change? Fit a no intercept model in SAS and check the SAS output for any noted differences.

Question: How do we interpret the diagnostic plots output by the PLOTS(ONLY)=(DIAGNOSTICS) option in PROC REG in SAS?

Question: Why do we plot each predictor variable against the residual as a model diagnostic?

Question: Why do we perform transformations in the construction of regression models? Name at least two reasons.

Question: What is multicollinearity and how does it affect the parameter estimates in OLS regression? How do we diagnose multicollinearity?

Question: What is a Variance Inflation Factor (VIF) and how does it relate to multicollinearity?

Question: Given a set of predictor variables $X_1, \ldots, X_n$, which are determined to show a high degree of multicollinearity between some of the variables, how should we choose a subset of these predictor variables to reduce the degree of multicollinearity and improve our OLS regression performance?

Question: Variable Selection: How does forward variable selection work? How does backward variable selection work? How does stepwise variable selection work?

Study Questions for Multivariate Analysis

Principle Components Analysis

Question: Principal Components Analysis (PCA): What is principal components analysis? How does PCA eliminate the problem of multicollinearity? What does it mean for X1 and X2 to be orthogonal? In order to better understand orthogonality, take the building prices data set and perform these steps:

(a) Perform a PROC CORR on X1-X9.
(b) Create nine orthogonal predictor variables using PCA. Call these variables Z1-Z9.
(c) Perform a PROC CORR on Z1-Z9.

Question: Principal Components Analysis is described as a method of ‘dimension reduction’. How does PCA reduce the dimension of a statistical problem? How do you select the reduced dimension for your problem.

Factor Analysis

Question: Are the factor scores always orthogonal? Are they orthogonal after a rotation?

Question: If two analysts perform a factor analysis, are they likely to arrive at the same result? If the same two analysts perform a principal components analysis, are they likely to get the same result?

Question: What is the first step in performing a factor analysis?

Question: In the context of factor analysis, what is the communality of factors?

Cluster Analysis

Question: What is the difference between hierarchical and non-hierarchical clustering?

Question: What is linkage? What types of linkage are there?

Question: How do we examine the goodness-of-fit of a cluster analysis or two comparative cluster analyses?

Question: Do the data need to be treated before we perform a cluster analysis?

References

Bhatti, D.C. 2011. Statistical preliminaries and mathematical notation. http://nwuniversity.adobeconnect.com/p2u4z1zop3a/.

Montgomery, D.C., Peck, E.A., and Vining, G.G. 2012. *Introduction to linear regression analysis.* John Wiley & Sons.

Wikipedia. 2015. Regression analysis — wikipedia, the free encyclopedia. /url\protect\T1\textbracelefthttp://en.wikipedia.org/w/index.php?title=Regression_analysis&oldid=647603059\protect\T1\textbraceright.