Assignment #1

Gustavo Esparza, Lindsay Brock, and Brian Schetzsle

September 5, 2019

Chapter 1: Algorithms and Inference

This chapter begins by focusing on the Algorithmic and Inferential division of Statistical Analysis methods. In order to make a clear distinction between the two, we are presented with a quintessential statistical method: Averaging.

First, we will look at the basic formula for a single mean value of n observations denoted by $x_1, \dots, x_n$:

$$\bar{x} = \sum_{i=1}^{n} \frac{x_i}{n}$$

Thus, we are provided with a succinct result that allows us to measure our data in a quantitative manner. However, we are still left with the question of how accurate our measure truly is. This traditional average is accompanied by the traditional Standard Error:

$$\hat{se} = \left[\sum_{i=1}^{n} (x_i - \bar{x})^2 / (n(n-1)) \right]^{1/2}$$

In this elementary example we have seen the Statistical Algorithm of Averaging accompanied by the Statistical Inference of the Standard Error.

A broader point is established, using this basic example of averages and standard errors. Here, the standard error gives us inference of the accuracy of the averaging algorithm. However, we could also be interested in the accuracy of the standard error calculation. Thus, in this line of thinking, we have recontextualized the standard error as an algorithm that requires its own inferential analysis. The main takeaway from this is the Algorithm always comes first and the Inference functions as a follow-up attempt to assess accuracy. Given this nature, we can observe modern technological advancements allowing for Algorithms and Methodology to evolve and improve at a rapid pace. Inference, on the other hand, is slowly but surely attempting to assess the validity of these new techniques.

To further demonstrate the interplay between our two aspects, we will focus on a **Regression** example that attempts to study Kidney function.

We can model the given data via a least squares linear regression line defined by

$$y = \hat{\beta}_0 + \hat{\beta}_1 x$$

where least squares refers to the minimization of the sum of squared deviations given by

$$\sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$

Using this algorithm, we can assess accuracy via the inferential standard error that is an extended version of the standard error encountered in our basic average example. Based on the standard error inferential analysis, we may opt to choose another regression model that could better fit our provided data set such as a quadratic model:

$$y = \hat{\beta}_0 + \hat{\beta}_1 x + \hat{\beta}_2 x^2$$

These attempts could continue on and on in the pre-computer age, but modern innovation such as the **Lowess curve** can quickly determine a more precise curve for the given data. Again, this modern algorithm for regression calls for a modern inferential analysis to access the underlying accuracy. As the classical standard error computations are no longer valid for the lowess curve, we instead use a computational inferential engine known as the **bootstrap**. Bootstrap resampled from our data set with replacement allows for certain data points to be repeated or omitted. Applying our lowess curve to the new bootstrapped dataset allows us to check the validity of our curve with computer technology as opposed to classical methods that are no longer valid.

This regression exercise, along with another example involving Hypothesis Testing, provide a clear illustration of how modern advancements have affected Statistical Algorithms as well as the Inferential methods that attempt to assess their validity and accuracy. A final observation that is quite interesting regarding this division of Algorithms and Inference relates back to how the modern age is utilizing these tools. Simply put, algorithms can be found in just about any discipline but Inference can be viewed as a statistics-focused attempt to assess these algorithms.

Chapter 2: Frequentist Inference

Now moving away from explaining the counterparts of statistical methods, this chapter focuses in on a specific form of inference known as **Frequentist Inference**.

Once again, we will use the basic principle of an Average and the corresponding standard error to illustrate our point. Given a true population parameter θ (in our case we will let $\theta = \mu$), we will also have an observed estimate $\hat{\theta} = t(x)$ ($t(x) = \bar{x}$). Provided these definitions,the accuracy of the observed estimate $\hat{\theta}$ is the probabilistic accuracy of $\hat{\Theta} = t(X)$ as an estimator of θ. Put into basic concepts, $\hat{\theta}$ is a single number but $\hat{\Theta}$ represents a range of values that defines measures of accuracy based on its' spread. This difference implies that $\hat{\Theta}$ is largely concerned with future trials that have not been observed and will impact the estimate being measured. The previous example and definition was quite general, but we will now focus further into frequentism in practice.

The textbook working definition of frequentism is that the probabilistic properties of a procedure of interest are derived and then applied verbatim to the procedure's output for the observed data. This definition is explicitly stating that we attempt to make inference about the true distribution F, when F is unknown. There are excellent examples describing how this issue is circumvented such as Maximum Likelihood and bootstrap. Here, we will use the **plug in principle** with our reliable average algorithm to show how frequentism can be used in practice.

From our distribution F, we have the standard error of our estimate, $\bar{X}$ defined as:

$$se(\bar{X}) = [var_F(X)/n]^{1/2}$$

Since we are only able to use the observed x_i values, we can estimate the population variance with the following formula:

$$\widehat{var_F} = \sum (x_i - \bar{x})^2/(n-1)$$

We then **plug in** this substitution to obtain our standard error estimate. Thus, we have made inference about the unknown via our specific observed data.

Moving along from frequentism in classical practice, we will cover two concepts related to **frequentist Optimality**. In the prior example, we have essentially been given the mean as the parameter to make frequentist inference from. Now, we have the dilemma of choosing what source to make inference from. This is the essence of frequentist optimality: finding the best choice of $t(x)$ given a model F.

The first example is the Maximum Likelihood Estimation, which is briefly discussed as the estimate that minimizes asymptotic standard error. This section appears to be primarily focused on the Neyman Pearson Lemma. This lemma begins by assuming we have two distribution options to choose from and we assign a value of 0 or 1 depending on whether we chose the correct distribution for a given value. This is summarized via the following formula:

$$\alpha = Pr_{f_0}[t(x) = 1], \beta = Pr_{f_1}[t(x) = 0]$$

We then let $L(x)$ be the likelihood ratio

$$L(x) = f_1(x)/f_0(x)$$

and the testing rule $t_c(x)$ be

$$t_c(x) = \begin{cases} 1 & \text{if} \quad \log L(x) \geq c \\ 0 & \text{if} \quad \log L(x) < c. \end{cases}$$

Thus, we are now provided with a test for which minimized α and β values provide the best option for which distribution to choose for our observed data in order to make inference. This chapter ends by discussing the current limitations that Frequentist Optimality Theory is facing due to the larger data sets and more complicated inferential questions being encountered. Although there is no modern alternative, the author closes the discussion on a hopeful note for the future.

Chapter 3: Linear Regression

This next chapter will focus on review of linear regression. It is important to gain a clear understanding of this topic as linear regression is an approach for supervised learning that will be covered in future chapters. Newer approaches to supervised learning stem from this type of modeling.

Section 1: Simple Linear Regression

$$Y \approx \beta_0 + \beta_1 X$$

Above is the basic equation, also called a model, for *simple linear regression.* This type of regression predicts a quantitative response Y using a single predictor X, assuming an approximately linear relationship between the two. The constants β_0, intercept, and β_1, slope, are the unknown model coefficients or parameters that will need to be estimated using $\hat{\beta_0}$ and $\hat{\beta_1}$, respectively. These estimations will be used in the equation below to compute $\hat{y}$ which is a prediction of Y based on $X = x$.

$$\hat{y} = \hat{\beta_0} + \hat{\beta_1} x$$

As β_0 and β_1 are the unknown population coefficients, we must use a sample of data with n paired observations, $(x_1, y_1), ..., (x_n, y_n)$, to find their best estimations. We can define these best coefficient estimations as those that produce $\hat{y}s$ that are as close to the original data points as possible. This leads us to the term *residual*, or $e_i = y_i - \hat{y_i}$, which will be used to find the *residual sum of squares* (RSS) described in the equation below. As our goal is to find the coefficient estimates that provide the closest $\hat{y}s$, the appropriate estimates will be those that minimize the RSS. Minimizing *least squares* is not the only approach to consider but for the purposes of this chapter, it will be the method used.

$$RSS = e_1^2 + e_2^2 + \cdots + e_n^2$$

The equations to find the least squares coefficient estimates, or the minimized $\hat{\beta}_0$ and $\hat{\beta}_1$ are as follows:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} \qquad and \qquad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \qquad where \qquad \bar{y} = \frac{1}{n}\sum_{i=1}^n y_i \qquad and \qquad \bar{x} = \frac{1}{n}\sum_{i=1}^n x_i$$

Now that a set of coefficient estimates have been calculated their accuracy must also be assessed, and to understand what accuracy means in the context of these estimates the difference between the *population regression line* and the *least squares line* must be recognized. Adding an error term to the original simple linear regression formula gives us the population regression line model below. It is the equation that represents the true relationship between the population X and Y. The error term is used to account for measurement error, other variables that might cause a variation in Y, and the true relationship between X and Y not actually being linear.

$$Y = \beta_0 + \beta_1 X + \epsilon$$

As the population regression line is unobserved because the true relationship can't be known, the least squares line can be computed using the estimated coefficients based on the observed data of the sample. In other words, it is an estimation of the population regression line. Since this equation is only an estimation, whether or not is it *biased* must be considered. Bias is the over or under estimation of the coefficients based on the sampled data and with each new set of observations $\hat{\beta}_0$ and $\hat{\beta}_1$ will differ. After taking multiple samples and averaging the $\hat{\beta}_0 s$ and $\hat{\beta}_1 s$, these averages become close to the population coefficients resulting in unbiased estimators. The difference between the single estimates of $\hat{\beta}_0$ and $\hat{\beta}_1$ and the actual β_0 and β_1 is called the *standard error* which becomes smaller as the sample size gets larger. The equations for the standard error squared, or the variance, are below. Since the actual standard error of the population is not known the *residual standard error* must be calculated using $RSE = \sqrt{RSS/(n-2)}$. This RSE is an estimate of the standard deviation of the error term discussed in the previous paragraph.

$$SE(\hat{\beta}_0)^2 = \sigma^2 \left[\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^n (x_i - \bar{x})^2}\right] \qquad and \qquad SE(\hat{\beta}_1)^2 = \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

Note: $\sigma^2 = Var(\epsilon)$; To use equations, we need to assume ϵ_i for each observation are uncorrelated with σ^2.

Knowing these RSEs can be useful for many reasons, one being their use in computing *confidence intervals*, or a range of values that will contain the true value of the unknown parameter a certain percentage of the time, usually that percentage being 95%. The interval will be calculated as $[\hat{\beta} - 2 * SE(\hat{\beta}), \quad \hat{\beta} + 2 * SE(\hat{\beta})]$ for both $\hat{\beta}_0$ and $\hat{\beta}_1$.

Another way in which these residual standard errors can be useful is in performing *hypothesis tests* of the coefficients, using null and alternative hypotheses. Although many different tests can be performed, the most frequent tests a null hypothesis that there is no relationship between the X and Y variables against the alternative hypothesis that there is a relationship.

$$H_0 : \beta_1 = 0 \qquad against \qquad H_a : \beta_1 \neq 0$$

To test the null hypothesis above, it must be determined how far away $\hat{\beta}_1$ is from β_1 using the residual standard error. This computes a $t - statistic$ which measures how many standard deviations $\hat{\beta}_1$ is away from 0. This t-statistic (below) will then be used to calculate the *p-value*, which indicates just how unlikely we are to observe a relationship between the predictor and response due to chance. When the value is small, usually below a cut-off value of 0.01 or 0.05, the null hypothesis is rejected indicating there is a relationship.

$$t = \frac{\hat{\beta}_1 - 0}{SE(\hat{\beta}_1)}$$

Once it has been determined that there is a relationship between the predictor and response variables the fit of the model (proposed relationship) must be evaluated and is usually done using the RSE and the R^2 statistic. An alternative form of the RSE better suited for evaluation of fit is given below. When the value for RSE is very small it is an indication of a well fit model, but it can be difficult to determine how small that value should be. This is where the R^2 statistic comes into play as an additional measure of fit. It is a proportion of the variance of the response explained by the predictor and its value will always be between 0 and 1. When the value is close to 1 most of the variance is accounted for by the predictor and the model is determined to be a good fit. An appropriate value can also be difficult to determine for R^2 but it provides more information than only the RSS. Squaring the correlation between X and Y can also be used in place of R^2, but this can only be used when there is a single predictor.

$$RSE = \sqrt{\frac{1}{n-2}RSS} \qquad where \qquad RSS = \sum_{i=1}^{n}(y_i - \hat{y_i})^2$$

$$R^2 = 1 - \frac{RSS}{\sum_{i=1}^{n}(y_i - \bar{y_i})^2}$$

Section 2: Multiple Linear Regression

As you may well know, in most regression cases there will be more than one predictor. This section reviews the *multiple linear regression model.*

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \epsilon$$

An individual coefficient in the model above can be described as the average effect a one unit increase in the predictor has on the response, holding all other predictors constant. Just as in the previous sections, the population coefficients are not known but must be estimated to make predictions using the formula below, and the same approach of minimizing the least squares is used to find their estimates. It is also important to mention that in the case of multiple predictors, the $\hat{\beta}s$ do not have such simple equations as before but rather require matrix algebra which can be done using statistical software.

$$\hat{y} = \hat{\beta_0} + \hat{\beta_1}x_1 + \hat{\beta_2}x_2 + \cdots + \hat{\beta_p}x_p$$

Just as in simple linear regression, a hypothesis test must be performed to determine if there is a relationship between the response and predictors, the null being that all coefficients equal to zero (no relationship) and the alternative that at least one is not (there is at least one relationship). The null hypothesis can also be written to include only one or a subset of the predictor variables. Instead of the t-statistic, the *F-statistic* is used (below). When the value is close to one, it is determined that there is no relationship between the response and predictions, and if the value is far greater than one, there is some relationship. Statistical software can also be used to calculate the F-statistic and in turn, the p-value for each of the predictor variables but the overall F-statistic must also be assessed as it adjusts for a large number of predictors.

$$F = \frac{([\sum_{i=1}^{n}(y_i - \bar{y_i})^2] - RSS)/p}{RSS/(n-p-1)}$$

Once the relationship has been established, it is important to decide which variables are important (those actually associated with the response) also referred to as *variable selection.* There are multiple statistics that may be used to judge the quality of a model and the included variables, such as Mallow's C_p, Akaike information criterion (AIC), Bayesian information criterion (BIC), and adjusted R^2, but to decide upon these models variable by variable can become too tedious. There are 2^p models (p = number of predictor variables) to be considered and as you can see this number can become large very quickly. Luckily there are other methods such as Forward selection, Backward selection, and Mixed selection, discussed in this section, that can help to determine a smaller set of models to evaluate.

Next, after the model has been selected, the fit must be assessed. This is accomplished much in the same way as simple linear regression with only a few minor differences. R^2 is calculated using the correlation between the response and fitted linear model, with the same interpretation as previously discussed, and RSE is now equal to $\sqrt{\frac{1}{n-p-1}RSS}$. Additionally, it may be helpful to plot the data when assessing fit as it provides a picture rather than simply numbers.

Finally, the model has come together and is ready for predictions but with predictions comes uncertainty as the model was built using estimations of the population. Three types of uncertainty were discussed in this chapter, uncertainty of the estimated coefficients, which can be visualized using the confidence interval, *model bias*, and the error within the model, which can be visualized using a prediction interval.

Section 3: Other Considerations in the Regression Model

Section 3.3 deals with other considerations in a regression model setting. While sections 3.1 and 3.2 assumed all the predictors in a model are quantitative, section 3.3 starts by addressing what to do with predictors that are qualitative. The book doesn't delve too deeply into what qualitative data is, focusing on nominal examples like gender and ethnicity, but we know qualitative data can also be ordinal, which is numeric but cannot be manipulated with mathematical operations. If a qualitative predictor has only two possible values, or levels, it is easy to incorporate into a linear regression by setting up a *dummy variable* that has two corresponding numeric values, like (0,1) or (-1,1). The numeric values are arbitrary but do impact the interpretation of the coefficients in the model. Qualitative predictors with more than two levels require an extension of the dummy variable encoding; each level gets its own dummy variable (one level, chosen arbitrarily, is the baseline and its dummy variable is excluded from the model). The coefficients are then interpreted as the difference in each level from the baseline, which is the intercept of the model, β_0.

Linear regression models also assume that the relationship between the predictors and response are *additive* and *linear*. The additive assumption means that changes in one predictor are independent of the other predictors.The linear assumption means that changes in a predictor have a constant effect on the response, regardless of the value of the predictor. Often these restrictive assumptions are violated. A way to deal with the additive assumption is to include interaction terms in the model. An interaction term is essentially a new predictor that is the product of two observed predictors; its inclusion allows changes in one predictor to affect the response both directly and through an interaction with another predictor. To deal with violations of the linear assumption, this section suggests including powers of the predictors as well, resulting in a polynomial regression.

This section concludes with a brief summary of potential problems from fitting a linear regression. **Non-linearity of the response-predictor relationships** is a problem that occurs when the response is not truly a linear function of the predictors. This problem was addressed in the previous sub-section and the book suggests again including transformed predictors, such as X^2. If **error terms are correlated**, then variance of the predictors is underestimated and significance of the coefficients can be erroneously concluded. **Non-constant variance** of the error term in the data can also lead to incorrect estimation of significance of coefficients; a possible remedy is to apply a concave transformation of the response, like the log function, or to fit the model using weighted least squares. **Outliers** also impact the significance of coefficients and measurements of fit of the model. Outliers can be removed if they were caused by a problem in data collection, but they may also indicate a deficiency in the model. **High-leverage points** are unusual values of the predictors and have an outsize influence in the model. Problems with these data points can invalidate the fit of the entire model. Finally, collinearity is when two predictors are closely related to each other and can cause problems in a model because their individual effects on the response are difficult to separate. This can lead to some predictors appearing not statistically significant. Sometimes sets of predictors can be highly correlated even though no pair has high correlations; this is called multicollinearity and causes the same problem in the model.

Section 4: The Marketing Plan

Section 3.4 provides an example of how a methodological approach to data analysis can answer questions about the relationship in an advertising dataset.

Section 5: Comparison - Linear Regression, K-Nearest Neighbors

Section 3.5 introduces non-parametric models, an alternative to linear regression and specifically focuses on the K-Nearest Neighbors (KNN) Regression. Here, the response for a given set of predictors is the average of the k points closest to those predictors. The KNN Regression is desirable because it makes far fewer assumptions about the functional form of $f(X)$ but performs worse than a linear regression, such as when K is too small or when the number of predictors is large. This latter case is called the *curse of dimensionality* and leads to poor performance because data tends to spread out in higher dimensions. KNN regression also suffers from lack of interpretability; it's difficult to say how changes in the predictors impact the response generally.

$$\hat{f}(x_0) = \frac{1}{K} \sum_{x_i \in N_0} y_i$$

Section 6: Lab - Linear Regression

This final section provides a lab using R for the reader to practice the topics described in Chapter 3. It begins with a discussion on the libraries needed for datasets and functions to be used, as well as simple and multiple linear regression functions (specifically lm). Delving deeper into these functions, it provides practice with interaction terms, non-linear transformations of the predictors, and qualitative predictors. Finally, it provides the user with an introduction to writing their own functions.

References

- *Computer Age Statistical Inference*, 1st Edition, by Efron and Hastie,Cambridge;

  Summary of Chapters 1 & 2
- *An Introduction to Statistical Learning*, 1st Edition, by James et al., Springer;

  Summary of Chapter 3