Logic and Organization

The purpose and objective of the analysis is made clear, and the analysis addresses the objective(s) in a focused and logical manner. Use informative headings.

This criterion is linked to a Learning OutcomeTechnical Depth and/or Breadth

Depth: Besides the normal approaches that an ordinary analyst can perform, critical thinking and creativity was reflected in the report, both in problem formulation and solution.

Breadth: To solve a given problem, an overview of commonly used tools is provided. Tools are properly chosen, and the pros and cons of choices are well discussed.

Professional-looking, Appropriate use of figures/tables. All axes and legend of figures properly labelled.

Title

1. Purpose

The purpose of this post is to show 1) how to examine factors associated with egg weight of a bird species, the Vinous-throated parrotbill. 2) how to compare two methods in prediction accuracy. There are several factors that might contribute to the egg size of the bird shown below with the name of the variables in the dataset.

- Parent\_age\_class: Age of parents that lay eggs. Measuring age in wild bird precisely is not easy, so I measured in categorical way. They are either young, which means it is their 1st year of laying eggs, or old, which means when they are older than 1year when they lay eggs.

- Tarsus: length of legs in (mm), continuous variable

- Bill: length of bill (mm), continuous variable

- Wing: length of wing span (mm), continuous variable

- Tail: Length of tail feather (mm), continuous variable

- Temperature: averaged temperature of 4 days before the date of laying eggs

- ld: Date of egg laying. 0 means the first date of egg laying and the other number means the number of days passed from the first date of egg laying (0). Continuous variable.

- cs: Clutch size, which means the number of eggs in a nest. On average, the birds lay 5 eggs but sometimes lay 4 or 6 eggs. There are only three values in this measurement, 4,5 and 6, so it is close to categorical variable.

- weight: Averaged weight of the eggs in a nest. (g). continuous variable

Among the variables, weight is the dependent variables and all other variables are independent variables.

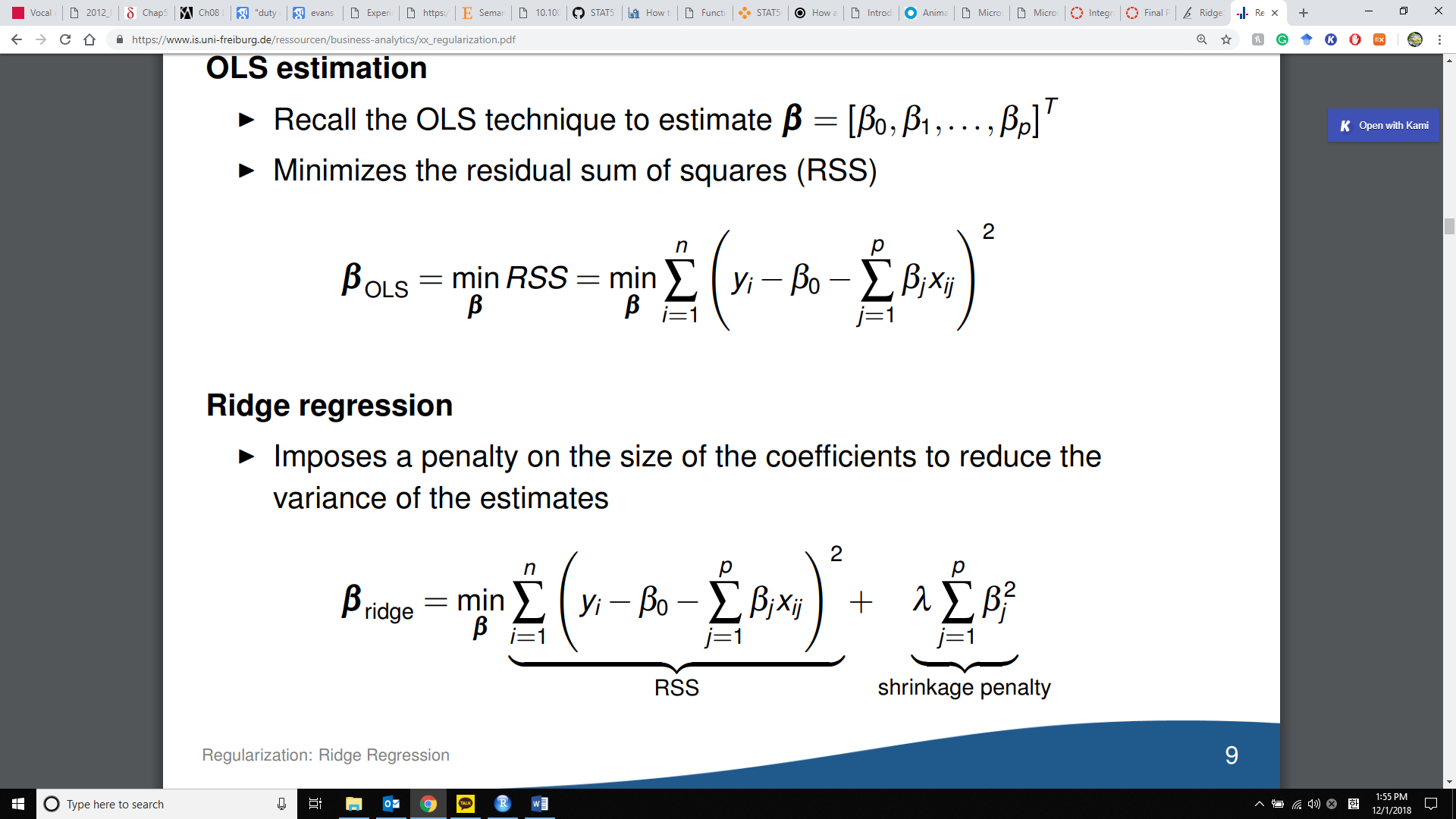
2. Overview of analysis techniques

The typical analysis that most first-year graduate students in Business analytics department or graduate students in other fields would be multiple regression. However, there are problems of using multiple regression in this data, because of multicollinearity. For example, 4 measurements in body size (Tarsus, Bill, wing and tail) will be all correlated. Also, laying date will be correlated with temperature because the data has been collected through early March to August. As the laying date increases, temperatures will also be increased as moving from spring to summer. To deal with this collinearity, I used shrinkage methods that effectively reduce coefficient estimates towards zero, the ridge regression and the lasso regression.

Both ridge and lasso regression have a tuning parameter called lambda, which is a critical penalty term to reduce coefficients. The role of lambda is to adjust value of coefficient estimates and results of both regression methods subject to change according to the lambda. Hence it is crucial to set an appropriate lambda to get the best results.

Ridge regression is an improved version of ordinary least square analysis. When lambda equals to 0, it is exactly the same analysis with ordinary least square. As the lambda increases, the shrinkage penalty increases, and coefficient estimates get closer to zero. One thing to note is that, even though the coefficients reduces toward zero, these do not reach to actual zero. That means all independent variables will be included no matter how high the lambda is.

Lasso regression works also similarly with ordinary least square analysis. The main difference with ridge regression is that it use l2 norm and lasso regression is based on l1 norm. l2 norm means the squared value and l1 norm means the absolute value that without being squared.



From https://www.is.uni-freiburg.de/ressourcen/business-analytics/xx\_regularization.pdf

3. Pros and cons of choice

If all variables contribute to the prediction of dependent variable, then ridge regression works better, whereas if only some variables contribute, then lasso regression is appropriate.

Benefit of ridge regression:

* When the variance is high, it can reduce it effectively.

Disadvantage of ridge regression:

* Ridge regression does not shrink coefficient to zero, so it does not have variable selection property that lasso regression has.

Benefit of lasso:

* Since the number of independent variables is decreased, we can get more parsimonious results and easy to interpret.

Limitation of lasso:

* Subject to produce biased results
* It lacks consistency to produce same results
* If variables are highly correlated, it will make one of the coefficient of variable 0, which in turn can result in losing information.

Post2

1. Purpose

The purpose of this post is to show how to examine presence of hidden subgroup and the subgroups match with a specific group category that is already known.

I am studying the vocal behavior of Tufted titmice, commonly found species in Tennessee.

The titmice produce various types of calls that convey threat information to other birds to alert them.

There are 10 acoustic measurements of the titmice calls and 1 threat context.

The variables in the dataset are:

* Threat: Ordinal level of predation threat with three levels: low, medium and high. This will be regarded as a categorical variable.
* No\_notes: The total number of calls produced
* Duration: The temporal duration of calls measured in millisecond
* Distomax: is temporal duration to measure the time taken to reach the maximum amplitude from begging of the call.
* RMS: loudness of the call with 3dB deducted from the peak amplitude
* Peakfreq: Peak frequency (high or low tone) of a call in Khz
* Peakamp: Peak amplitude (loudness) of a call
* Fundfreq: fundamental frequency in Khz
* Minfreq: minimum frequency in Khz
* Maxfreq: Maximum frequency in Khz
* Entropy: The level of tonality (clear and pure tone like flute sound, or harsh sound such as cat’s hissing call)

Among 12 variables, threat will be used as dependent variables and all other variables will be dependent variables.

2. Overview of analysis techniques

2.1 Hierarchical clustering

Hierarchical clustering is a technique by which group similar data points together by using hierarchical tree model, a dendrogram.

(Dendrogram plot here)

Dendrogram is a figure that shows how each data point is grouped. Bottom of the plot shows all data points called leaves. They are grouped by clades (branch), and the leaves that are combined in the same height of branch are in the same cluster. Height of the branch indicates more dissimilarity between clusters.

There are largely 2 ways of merging groups: agglomerative and divisive.

Agglomerative is basically a bottom-up approach. It starts merging leaves into a bigger cluster based on similarity, until all nodes and leaves are merged into one root.

Divisive clustering is a top-down approach. It starts dividing one root into nodes and leaves based on dissimilarity until all nodes get separated into leaves.

There are 3 ways of linkage

* single linkage:

Calculate all pairwise distance and identify the closest distance between 2 clusters. Fast in calculation and works well for data of which order is important. Also good for separating few isolated groups.

* Complete linkage

Calculate all pairwise distance and identify the farthest distance between groups to maximize dissimilarity between clusters.

Distance is usually calculated with Euclidean distance, but there are many other options such as manhattan, bray-curtis.. etc.

* Average linkage calculate all pairwise distance with average distance between 2 clusters

Advantage of using hierarchical clustering: It does not require to fix the number of the cluster from the beginning, unlike K-means clustering.

Cons:

2.2 PCA (principal component analysis)

PCA is a technique to reduce dimension by calculating principal components that effectively represent variability with smaller numbers of variables. Normal regression analyses just use the variables themselves, but PCA creates new variables (PCs), each of which consists of original variables by differing intensity and direction.

Suppose there are P variables X1, X2,.., Xp and you want to make scatter plots for every pair of variables. The number of cases of choosing 2 variables out of P is P(p-1)/2, which seems to be too many.

PCA is based on an idea that not all variables contribute for explaining variation of dependent variable. It aims to find out the most important variables. What does the PC look like?

Let’s take an example of PC1, the PC that explains the largest variance.

PC1= (equation here)

----- are “loadings” SS=1 \_x\_=0

Before running PCA, all variables should be centered or scaled because PCA focuses on variance.

Since I am not going to use PCA as a separate method in this posting, and just calculate PCs that are going to be used in K-means clustering, I am not going to show how to interpret results.

2.3 K-means clustering

K-means clustering divide observations into K clusters that do not overlap to find out data points that have similar traits.

Minimize variance of distance among clusters

Partitioning in the way that increase inner similarity of observations in the same cluster

It requires data to take Euclidean distance

How K-means cluster works

Decide the number of K. we can choose any natural number from 1.

K Centroids (the mean center of cluster) will be randomly assigned first, so all centroids will almost overlap.

a. In the next step, nearest data points from each centroid will be assigned (labeled) as a cluster that belongs to a cluster.

b. Then, centroid will be calculated again and moved to the point, where is the mean of the cluster.

As procedure a and b are iterated repeatedly, the centroids will become more stable and eventually stay in the same place.

3. Pros and cons

4. Analysis

4.1 hierarchical model

Post3

1. Purpose

In this post, I will show how to build a model with variable selection methods using stepwise selection. I will generate models to predict important factors associated with house price.

Variables are

I first excluded unnecessary, irrelevant variables with house price.