Project 1

Comp 4442

Guidelines

1. For this project, you may use any reasonable resources as long as you give a citation. You know what honest work is - let this be your guide.
2. You may work with 1 other person if you like, but you must turn your work separately.
3. Your answers should be as complete as you are able to make them. Try to frame your work in a way that you could give a short presentation of your efforts.
4. You may always ask me for assistance.
5. This is due 05/??/2019

Overview of Topics

1. ANOVA - Used when we want to compare the means of some property of different groups.
2. Linear regression - continuous
3. Logistic regression - Trying to predict a binary outcome with continuous data.
4. Lasso vs Ridge - two methods that let us choose coefficients in a more dynamic way
5. Cross-validation - A method for breaking up data into pieces, and using this to evaluate a particular model choice.

An exercise in ANOVA

The simulated data in the file "dat1.RData" represent the bids for copies of the same product auctioned on five different platforms. The dollar value of the bids are in the variable "bid". An identifier for the platform is in the variable "platform". You may assume that the bids in the data are independent samples from the population of bids for the product on the given platform.

Q1, part 1

Please perform visual diagnostics appropriate to preparation for an ANOVA to investigate whether the bids are consistent with the null hypothesis that the bid amounts on each platform are drawn from populations with equal means. You may optionally include relevant statistical tests.

```
load(file="dat1.RData")

#head(dat1,20)
```

Q1, part 2

Please perform an ANOVA to investigate whether the bids are consistent with the null hypothesis that the bid amounts on each platform are drawn from populations with equal means. Output the summary of the ANOVA model. Please give your interpretation of the results in the context of your response to part 1.

Q1, part 3

Please construct a column plot of the means for each platform, including error bars with the property that non-overlapping error bars indicate a difference in mean significant at the $p \leq 0.05$ level.

Question 2

We will look at this one together in class. The simulated data in the file "dat2.RData" represent the results of a drug trial. Subjects with one of four comorbidities, "comorb", were recruited, then randomly assigned to one of three treatments, "treatment", and the amount of a biomarker assessed, "amt".

Q2, part 1

Both the amount of the biomarker and the log of the amount are of medical interest. Which variable (if any) from "amt" and "log(amt)" is best suited to an ANOVA to investigate whether the data are consistent with the null hypothesis that the mean level of the biomarker is equal for each treatment population, each comorbidity population, and each interaction? Please explain your conclusion.

```
load("dat2.RData")

#head(dat2,20)
```

Q2, part 2

Please perform a 2-factor ANOVA with interaction using "amt" as the response variable and "treatment" and "comorb" as the grouping variables. Please provide an interpretation of the results, taking into account your response to part 1. (10 points)

Multiple linear regression

We have seen multiple linear regression now where we form a model of the form:

$$y = \alpha + \sum_{k=1}^{n} \beta_k x_j$$

We did this by minimizing the residual sum or errors, either using calculus or linear algebra. Given a data set that we can use MLR on, we shouldn't assume that we should use all of the information involved. The question of what information can be thrown away is a question of great importance.

Exercise 3 Will rescaling a variable by a constant change only the coefficient of the rescaled variable or possibly others as well? Demonstrate with an example

Best-subset regression

Suppose we want to model some quantity y and we have a choice of n independent variables. Then there is an algorithm known as best-subset regression that returns, for each $k \in \{2, 3, \ldots, n\}$ the subset of k independent that has the smallest residual sum of squares. While this algorithm gives the 'best' choice of descriptors for a given k, it does not tell us which value of k to use. That is, it doesn't tell us how many variables to ignore, it just tells us which ones to use for whatever choice of k we pick.

Ridge and Lasso

The idea behind subset selection is throwing away entire classes of information. This is sometimes very useful, but it is a very choppy process. Ridge and Lasso regression try to reduce the effect of 'lesser' variables rather than totally discard them (well, lasso removes some variables)

The coefficients, for the case of multiple linear regression, using the ridge method are found by choosing the vector $\vec{b}$ that minimizes the following:

$$\sum_{j=1}^{m}(y_j - \alpha - \sum_{k=1}^{n} x_{jk}\beta_j)^2 + \lambda \sum_{k=1}^{n} \beta_k^2$$

Here, λ is a parameter of our choosing.

For the lasso method, the coefficients are found by minimizing:

$$\sum_{j=1}^{m}(y_j - \alpha - \sum_{k=1}^{n} x_{jk}\beta_j)^2 + \lambda \sum_{k=1}^{n} |\beta_k|$$

Exercise 4 What is significant about the case $\lambda = 0$ in the methods described above?

Recall that the linear algebra expression for the coefficients in a linear model are given by

$$\vec{\beta} = (A^T A)^{-1} A^T \vec{y}$$

Exercise 5 Show that the coefficients in ridge are given by

$$\vec{\beta} = (A^T A - \lambda I)^{-1} A^T \vec{y}$$

where λ is the parameter in the ridge regression, A is the matrix of observations (of independent variables), and $\vec{y}$ is the vector of observed dependent variable.

Exercise 6

Super Conductors

In Canvas is file called super_train.csv. This file contains information about the different properties of various super conductors.

1. Make a multiple linear regression of this information and give a quick rundown. Remove any variables you think unnecassary and use ANOVA to compare some of your choices.
2. Attempt running the best-subset algorithm and report your results.
3. Use Lasso and Ridge on this data set, using cross-validation to pick a value for λ. How does this model compare

Exercise 7

Parkinson

In Canvas is file called parkinsons_updrs.csv. This file contains information about the different properties of various super conductors.

1. Make a logistic regression using this information and give a quick rundown.
2. Use Lasso and Ridge on this data set, using cross-validation (or any method I suppose) to pick a 'good' value for λ.

```r
library(leaps)
library(tidyverse)
```

```
## -- Attaching packages --------------------------- tidyverse 1.2.1 --
## v ggplot2 3.1.0       v purrr   0.3.2
## v tibble  2.1.1       v dplyr   0.8.0.1
## v tidyr   0.8.3       v stringr 1.4.0
## v readr   1.3.1       v forcats 0.4.0
## -- Conflicts ------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```r
library(caret)
```

```
## Loading required package: lattice
##
## Attaching package: 'caret'
## The following object is masked from 'package:purrr':
##
##     lift
```

```r
library(glmnet)
```

```
## Loading required package: Matrix
##
## Attaching package: 'Matrix'
## The following object is masked from 'package:tidyr':
##
##     expand
## Loading required package: foreach
##
## Attaching package: 'foreach'
## The following objects are masked from 'package:purrr':
##
##     accumulate, when
## Loaded glmnet 2.0-16
```

```r
park_data = read.csv("parkinsons_updrs.csv")
#head(park_data,4)
```