Assignment 1

The overall objective of this assignment is to develop the core techniques that can be used to examine a basic data set, including scatterplots, and basic linear and logistic regression.

Part 1. The objective in this part is to perform a series of linear regressions using a simulated data set named “Carseats”, contained in the “Carseats\_org.csv file.

1. Read in Carseats\_org.csv. You may, or may not need to remove the first column ( if it is only an ID indicator that simply counts each observation, remove it. It is common to receive datasets with an ID column, and also without. It is always a good idea to view the dataset in a text editor before loading it into R or Python or MATLAB or whatever software being used.

2. What is the dimension of the dataset? What variables are numeric, or integers ? Which ones are factors, or categorical variables?

3. Produce a scatterplot of sales vs all other variables, then plot the sales of carseats using only 3 or 4 variables only on one plot, so that the individual relationships are easier to see

4. Plot the sales by each categorical variable

5. Perform a linear regression, using the lm command, on the Carseats data, with Sales as the dependent variable, using all variables plus the following “interaction terms”: ( use the formula interface and use syntax such as: +var1:var2 in the formula for each pair of interaction terms.

a. Income and Advertising

b. Income and CompPrice

c. Price and Age

6. Which variables ( and interaction pairs ) are significant ? Do you find anything surprising or counterintuitive in the results? Describe any response variable or lack of response that seems surprising.

7. Repeat the above but using the “leaps” package to perform subset selection, using the full model using ONLY the following variables:

Price

Population

CompPrice

Income

Education

Age

Advertising

ShelveLoc

Plus the following interaction term:

Income and Age

Income and Advertising

Income and CompPrice

Use the regsubsets command in “leaps” to do the full model ( do NOT specify really.big=T as it will require a long time to complete, along with forward and backward subset selection. Use the R file, Subset\_select.R as a guide for syntax.

8. Use the following model selection criteria to determine which of the resulting models is “best” according to RSS, adjusted R squared, Mallow’s Cp, and BIC.

The result of the regressions is a model object, so you can query the proper coefficient of the model to determine which model, of the multiple models fit, produces the best metric.

Create a table, or simply list, for the full model, and forward and backward selection, which model # is the best, then write out the model, that is put the model in the form y=intercept + coefficient\*variable, for the best model using the BIC measure. NOTE: these may be the same model, or may differ slightly.

Info on the Carseats data follows:

Carseats data info:

|  |  |
| --- | --- |
|  |  |

**Sales of Child Car Seats**

**Description**

A simulated data set containing sales of child car seats at 400 different stores.

**Usage**

Carseats

**Format**

Sales

Unit sales (in thousands) at each location

CompPrice

Price charged by competitor at each location

Income

Community income level (in thousands of dollars)

Advertising

Local advertising budget for company at each location (in thousands of dollars)

Population

Population size in region (in thousands)

Price

Price company charges for car seats at each site

ShelveLoc

A factor with levels Bad, Good and Medium indicating the quality of the shelving location for the car seats at each site

Age

Average age of the local population

Education

Education level at each location

Urban

A factor with levels No and Yes to indicate whether the store is in an urban or rural location

US

A factor with levels No and Yes to indicate whether the store is in the US or not

**Source**

Simulated data

**\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\***

Part 2.

In this part you are to also use scatterplots to examine the variables in the WCGS data set, which records the onset of heart disease along with variables that “may” be important. Use a logistic regression model here. Use the glm command with family=binomial. While any results are not conclusive proof of causes of heart disease they do provide indications of where to focus future research and exploration using more thorough hypothesis testing. Here, we’ll only use basic plots and logistic regression ( because we wish to determine a categorical result, whether the onset of heart disease is likely, or not, given certain indicators ).

9. Use logistic regression to build a model, with the presence of heart disease ( found in the chd variable) based on height (height) and number of cigarettes consumed each day (cigs). That is, perform logistic regression with chd as the response and height and cigs as predictors.

10. Use logistic regression to repeat the above but this time use the number of cigarettes consumed each day, and add in both blood pressure variables ( sbp, dbp )

11. Use logistic regression to determine the probability of having heart disease (chd) based only on sdp, dbp, and chol

12. Print out the summary statistics, using the “summary” command, for each of the models created in 9, 10, 11 above.

13. Use the ilogit function to plot these predictions (for each of 9, 10, and 11 above). This function is contained in the “faraway” package and also, if you do not wish to load another package for one function, will be placed on Canvas. Have a look at ilogit.R to see how to define a function of your own in R and the added caution of removing data points with an NA ( or not available ) value in order to prevent plotting errors. Note: you can either create a vector of points ( holding the other variables fixed, and plot, or you can use the “curve” function in R to simplify the process.

14. Can you draw any conclusions from the curves that result ? Note, this is NOT conclusive, but it does indicate LIKELY causes.

15. Deliverables: include you code and print-out of results. This can be a pdf file or simply a Jupyter Notebook that contains your code and results., either is acceptable. Or, if so inclined, copy and paste your plots and code into a Word document, though this method will grow tiresome over time and using a Jupyter Notebook or creating a pdf file is a skill you should develop.

Wcgs data info:

|  |  |
| --- | --- |
|  |  |

**Western Collaborative Group Study**

**Description**

3154 healthy young men aged 39-59 from the San Francisco area were assessed for their personality type. All were free from coronary heart disease at the start of the research. Eight and a half years later change in this situation was recorded.

**Usage**

data(wcgs)

**Format**

age

age in years

height

height in inches

weight

weight in pounds

sdp

systolic blood pressure in mm Hg

dbp

diastolic blood pressure in mm Hg

chol

Fasting serum cholesterol in mm %

behave

behavior type which is a factor with levels A1 A2 B3 B4

cigs

number of cigarettes smoked per day

dibep

behavior type a factor with levels A (Agressive) B (Passive)

chd

coronary heat disease developed is a factor with levels no yes

typechd

type of coronary heart disease is a factor with levels angina infdeath none silent

timechd

Time of CHD event or end of follow-up

arcus

arcus senilis is a factor with levels absent present

**Details**

The WCGS began in 1960 with 3,524 male volunteers who were employed by 11 California companies. Subjects were 39 to 59 years old and free of heart disease as determined by electrocardiogram. After the initial screening, the study population dropped to 3,154 and the number of companies to 10 because of various exclusions. The cohort comprised both blue- and white-collar employees. At baseline the following information was collected: socio-demographic including age, education, marital status, income, occupation; physical and physiological including height, weight, blood pressure, electrocardiogram, and corneal arcus; biochemical including cholesterol and lipoprotein fractions; medical and family history and use of medications; behavioral data including Type A interview, smoking, exercise, and alcohol use. Later surveys added data on anthropometry, triglycerides, Jenkins Activity Survey, and caffeine use. Average follow-up continued for 8.5 years with repeat examinations

**Source**

Statistics for Epidemiology by N. Jewell (2004)