BSTAT 3322 - H07

by Craig W. Slinkman

November 23, 2015

In this homework there are 10 questions. In each question you are asked to

* Identify the response and predictor variables.
* Use ggplot2 to draw a suitable diagram.
* Identify the nature of the relationship between the response and the predictor variables.
* Form the scatterplots only discuss how strong and how useful it will be to use the predictor variable to forecast the response variable.

## 1. Fire and Theft in Chicago

The following data concerns the relationship between the number of fires in a zip code and the number of thefts in the same zip code. Does it appear that number of thefts be 1000 units is linearly related to the number of fires? The data is stored in the file .

## 2. Radiation exposure and cancer

To measure the health consequences of this contamination, an index of exposure was calculated for each of the nine Oregon counties having frontage on either the Columbia River or the Pacific Ocean. This particular index was based on several factors, including the county's stream distance from Hanford and the average distance of its population from any water frontage. As a covariate, the cancer mortality rate was determined for each of these same counties.

The data give the index of exposure and the cancer mortality rate during 1959-1964 for the nine Oregon counties affected. Higher index values represent higher levels of contamination.

|  |  |
| --- | --- |
| Variable | Description |
| County | Name of county |
| Exposure | Index of exposure |
| Mortality | Cancer mortality per 100,000 man-years |

The data file is a text file and you will need to read it with function. I show you how do this below.

Hanford <- read.table( "Hanford.txt" )  
colnames( Hanford ) <- c( "County", "Exposure", "Mortality" )  
Hanford

## County Exposure Mortality  
## 1 County Exposure Mortality  
## 2 Umatilla 2.49 147.1  
## 3 Morrow 2.57 130.1  
## 4 Gilliam 3.41 129.9  
## 5 Sherman 1.25 113.5  
## 6 Wasco 1.62 137.5  
## 7 HoodRiver 3.83 162.3  
## 8 Portland 11.64 207.5  
## 9 Columbia 6.41 177.9  
## 10 Clatsop 8.34 210.3

## 3. Charting and stockprice changes

An investor believes that the signs of the zodiac in relation to his birth date control the returns of the stocks he has in his porfirio. To obtain the data for this problem load and run the scripts . This will create a csv file that you can read with your R-markdown chunk.

## 4. Price and ages of Mazda cars in Melpborne Austrailia

The two columns of the data are the prices and year purchased for 124 Mazda cars, as taken from the classified section of the Melbourne Age during the course of 1991. Hence the age of the car at the time can be calculated and used to model car price.

### Comments

* The data file will need to be read with a read table statement. The read,table statement assumes the data is character so we need to transform it to numeric before computing the age.

Mazadas <- read.table( "Mazadas.txt ")  
colnames( Mazadas ) <- c( "Year", "Price" )  
Mazadas$Year <- as.numeric( Mazadas$Year )  
Mazadas$Price <- as.numeric( Mazadas$Price )  
Mazadas$age <- 1991 - Mazadas$Year

Source

Melbourne Age, various issues during 1991.

Lee, A. (1994) Data Analysis: An introduction based on R. Auckland: Department of Statistics, University of Auckland.

# 5. Capital and Rental Values of Auckland Properties

The data show the capital value and annual rental value of 96 domestic properties in Auckland in 1991. The aim was to explore their relationship in the hope of being able to predict capital value from rental value, thus the latter is the explanatory variable in this case. The data file is .

You will need to read this data with the function. You should also divide both the response and the predictor variable by 1,000 to make your plot look better. The data file is .

### Description

The data show the capital value and annual rental value of 96 domestic properties in Auckland in 1991. The aim was to explore their relationship in the hope of being able to predict capital value from rental value, thus the latter is the explanatory variable in this case.

rm( list=ls() )  
Units <- read.csv( "Rents.csv" )  
colnames( Units ) <- c( "Capital", "Rents")  
Units$Capital <- as.numeric( Units$Capital )  
Units$Rents <- as.numeric( Units$Rents )  
Units <- Units[,-2]  
colnames( Units ) <- c( "Capital", "Rents")  
head( Units )

## Capital Rents  
## 1 61500 6656  
## 2 67500 6864  
## 3 75000 4992  
## 4 75000 7280  
## 5 76000 6656  
## 6 77000 4576

### Source

Lee, A. (1994) Data Analysis: An introduction based on R. Auckland: Department of Statistics, University of Auckland. Data courtesy of Sage Consultants Ltd.

# 6. Atlanta Commute Times

Install the Lock5Data package if you have not already installed it. Load the package with the OR functions. Then load the with the data function. Analyze the relationship between commute distance and compute time. The variable the government is interested is commute time.

# 7. Homes for Sale in California

The dataset in the package contains a random sample of 30 houses for sale in California. We are interested in whether we can use number of bathrooms Baths to predict number of bedrooms Beds in houses in California.

Lock, Robin H. (2012-10-10). Statistics: Unlocking the Power of Data, First Edition (Page 550). Wiley Higher Ed. Kindle Edition.

# 8. 1.4 Old Faithful

(Data file: in package ) The data file gives information about eruptions of Old Faithful Geyser during October 1980. Variables are the Duration in seconds of the current eruption, and the Interval, the time in minutes to the next eruption. The data were collected by volunteers and were provided by the late Roderick Hutchinson. Apart from missing data for the period from midnight to 6 a.m., this is a complete record of eruptions for that month.

Old Faithful Geyser is an important tourist attraction, with up to several thousand people watching it erupt on pleasant summer days. The park service uses data like these to obtain a prediction equation for the time to the next eruption. Draw the relevant summary graph for predicting interval from duration and summarize your results.

# 9. Predicting Windfarm Wind Speeds

(Data file: in R-package ) Energy can be produced from wind using windmills. Choosing a site for a wind farm, the location of the windmills, can be a multi-millionaire dollar gamble. If wind is inadequate at the site, then the energy produced over the lifetime of the wind farm can be much less than the cost of building and operation. Prediction of long-term wind speed at a candidate site can be an important component in the decision to build or not to build. Since energy produced varies as the square of the wind speed, even small errors can have serious consequences. The data in the data frame provides measurements that can be used to help in the prediction process. Data were collected every 6 hours for the year 2002, except that the month of May 2002 is missing. The values are the calculated wind speeds in meters per second at a candidate site for building a wind farm. These values were collected at a tower erected on the site. The values are wind speeds at a reference site, which is a nearby location for which wind speeds have been recorded over a very long time period. Airports sometimes serve as reference sites, but in this case, the reference data comes from the National Center for Environmental Modeling (NCAR, 2013). The reference is about 50 km southwest of the candidate site. Both sites are in the northern part of South Dakota. The data were provided by Mark Ahlstrom and Rolf Miller of WindLogics.

Weisberg, Sanford (2013-11-25). Applied Linear Regression (Wiley Series in Probability and Statistics) (Kindle Locations 1484-1493). Wiley. Kindle Edition.

# MLB Sabermetrics

Suppose we want determine the average number of runs that can be purchased for a million dollars. The code below finds all players who have more than 100 at bats in the 2014 MLB season and the players salary for the 2014 season. The R-script below uses the Lahman MLB baseball database package stored as data frames. The data frame is stored as file named and will be found in this project folder.

###########################################################  
# Clear the environment. #  
###########################################################  
#  
rm( list=ls())  
#  
#  
###########################################################  
# We attempt to require the package. If the require #  
# fails we install the package and require it again #  
###########################################################  
#  
success <- require( Lahman )

## Loading required package: Lahman

if( !success){  
 install.packages( "Lahman" )  
 require( Lahman )  
}  
  
success <- require( sqldf )

## Loading required package: sqldf  
## Loading required package: gsubfn  
## Loading required package: proto  
## Loading required package: RSQLite  
## Loading required package: DBI

if( !success){  
 install.packages( 'sqldf' )  
 require( sqldf )  
}  
#  
#  
###########################################################  
# We obtain the team tables for the same time period. #  
###########################################################  
#  
data( Teams )  
TeamWins <-   
 Teams[Teams$yearID > 1999, c( 1,3, 9) ]  
head( TeamWins )

## yearID teamID W  
## 2326 2000 ANA 82  
## 2327 2000 ARI 85  
## 2328 2000 ATL 95  
## 2329 2000 BAL 74  
## 2330 2000 BOS 85  
## 2331 2000 CHA 95

tail( TeamWins )

## yearID teamID W  
## 2770 2014 SFN 88  
## 2771 2014 SLN 90  
## 2772 2014 TBA 77  
## 2773 2014 TEX 67  
## 2774 2014 TOR 83  
## 2775 2014 WAS 96

#  
#  
##########################################################  
# Now we get the player salaries for the same time. We #  
# Group the salaries by year and team for hoining to #  
# the Teams table. #  
##########################################################  
#  
data( Salaries )  
Salaries <- Salaries[ Salaries$yearID > 1999, ]  
head( Salaries )

## yearID teamID lgID playerID salary  
## 12264 2000 ANA AL anderga01 3250000  
## 12265 2000 ANA AL belchti01 4600000  
## 12266 2000 ANA AL botteke01 4000000  
## 12267 2000 ANA AL clemeed02 215000  
## 12268 2000 ANA AL colanmi01 200000  
## 12269 2000 ANA AL davanje02 200000

tail( Salaries )

## yearID teamID lgID playerID salary  
## 24753 2014 WAS NL stammcr01 1375000  
## 24754 2014 WAS NL storedr01 3450000  
## 24755 2014 WAS NL strasst01 3975000  
## 24756 2014 WAS NL werthja01 20000000  
## 24757 2014 WAS NL zimmejo02 7500000  
## 24758 2014 WAS NL zimmery01 14000000

TeamSalaries <-  
 sqldf( "Select yearID,  
 teamID,  
 sum( salary ) / 1000000. AS teamSalary  
 FROM Salaries  
 GROUP BY yearID,  
 teamID" )

## Loading required package: tcltk

head( TeamSalaries )

## yearID teamID teamSalary  
## 1 2000 ANA 51.46417  
## 2 2000 ARI 81.02783  
## 3 2000 ATL 84.53784  
## 4 2000 BAL 81.44743  
## 5 2000 BOS 77.94033  
## 6 2000 CHA 31.13350

tail( TeamSalaries )

## yearID teamID teamSalary  
## 447 2014 SFN 20.0000  
## 448 2014 SLN 120.6930  
## 449 2014 TBA 72.6891  
## 450 2014 TEX 112.2551  
## 451 2014 TOR 109.9201  
## 452 2014 WAS 131.9837

###########################################################  
# Now we join the TeamSalaries with the TeamWins tables #  
# using a natrual join. #  
###########################################################  
#  
SalaryWins <-  
 sqldf( "  
 Select \*   
 FROM TeamWins W,  
 TeamSalaries S  
 WHERE W.yearID = S.yearID  
 AND W.teamID = S.teamID" )  
  
head( SalaryWins )

## yearID teamID W yearID teamID teamSalary  
## 1 2000 ANA 82 2000 ANA 51.46417  
## 2 2000 ARI 85 2000 ARI 81.02783  
## 3 2000 ATL 95 2000 ATL 84.53784  
## 4 2000 BAL 74 2000 BAL 81.44743  
## 5 2000 BOS 85 2000 BOS 77.94033  
## 6 2000 CHA 95 2000 CHA 31.13350

tail( SalaryWins )

## yearID teamID W yearID teamID teamSalary  
## 445 2014 SFN 88 2014 SFN 20.0000  
## 446 2014 SLN 90 2014 SLN 120.6930  
## 447 2014 TBA 77 2014 TBA 72.6891  
## 448 2014 TEX 67 2014 TEX 112.2551  
## 449 2014 TOR 83 2014 TOR 109.9201  
## 450 2014 WAS 96 2014 WAS 131.9837

write.csv( SalaryWins, file="SalaryWin.csv" )  
rm( list=ls() )

Weisberg, Sanford (2013-11-25). Applied Linear Regression (Wiley Series in Probability and Statistics) (Kindle Locations 671-674). Wiley. Kindle Edition.