Simple Linear Regression:Diagnostics and Assumptions Test

EXST 7014 - Lab 2

January 11, 2019

Table of Contents

Objectives

Simple Linear Regression (SLR) is a common analysis procedure, used to describe the significant relationship between two variables: the dependent (or response) variable, and the independent (or explanatory) variable. In lab 1, SLR was performed to fit a straight line model relating two variables. We learned how to interpret parameter estimates and R^2, and understood the hypothesis test of SLR.

You might notice that a single observation that is substantially different from all other observations can make a large difference in the results of your regression analysis. If a single observation (or small group of observations) substantially changes your results, you would want to know about this and investigate further. In this lab exercise, we will conduct appropriate regression diagnostics to detect outliers (or unusual observations) as well as evaluate some model assumptions.

In this lab exercise, you will get familiar with and understand as listed:

1. Conduct appropriate regression diagnostics to detect outliers (or unusual observations)

2. Evaluate the assumptions of SLR using Residual Plots and the Normality Test.

Part I

Lab Setup

Run the following code to both install and load the required packages.

```
install.packages('olsrr')          # install the package that runs residual
plots and check assumptions
library(olsrr)                     # Load the package

##
## Attaching package: 'olsrr'

## The following object is masked from 'package:datasets':
##
##     rivers
```

The Dataset

The data is from the textbook, Chapter 7, problem 6 and can be obtained from the url:
http://stat.lsu.edu/exstweb/statlab/datasets/fwdata97/FW07P06.txt

The latitude (LAT) and the mean monthly range (RANGE), which is the difference between mean monthly maximum and minimum temperatures, are given for a selected set of US cities. The following program performs a SLR using RANGE as the dependent variable and LAT as the independent variable.

```
# Create an object called 'theData' to store the data

theData <- read.table(header=T, stringsAsFactors = TRUE,text='
       CITY STATE LAT RANGE
       Montgome     AL      32.3     18.6
       Tuscon       AZ      32.1     19.7
       Bishop       CA      37.4     21.9
       Eureka       CA      40.8      5.4
       San_Dieg     CA      32.7      9.0
       San_Fran     CA      37.6      8.7
       Denver       CO      39.8     24.0
       Washingt     DC      39.0     24.0
       Miami        FL      25.8      8.7
       Talahass     FL      30.4     15.9
       Tampa        FL      28.0     12.1
       Atlanta      GA      33.6     19.8
       Boise        ID      43.6     25.3
       Moline       IL      41.4     29.4
       Ft_wayne     IN      41.0     26.5
       Topeka       KS      39.1     27.9
       Louisv       KY      38.2     24.2
       New_Orl      LA      30.0     16.1
       Caribou      ME      46.9     30.1
```

```
        Portland      ME       43.6      25.8
        Alpena        MI       45.1      26.5
        St_cloud      MN       45.6      34.0
        Jackson       MS       32.3      19.2
        St_Louis      MO       38.8      26.3
        Billings      MT       45.8      27.7
        N_PLatte      NB       41.1      28.3
        L_Vegas       NV       36.1      25.2
        Albuquer      NM       35.0      24.1
        Buffalo       NY       42.9      25.8
        NYC           NY       40.6      24.2
        C_Hatter      NC       35.3      18.2
        Bismark       ND       46.8      34.8
        Eugene        OR       44.1      15.3
        Charestn      SC       32.9      17.6
        Huron         SD       44.4      34.0
        Knoxvlle      TN       35.8      22.9
        Memphis       TN       35.0      22.9
        Amarillo      TX       35.2      23.7
        Brownsvl      TX       25.9      13.4
        Dallas        TX       32.8      22.3
        SLCity        UT       40.8      27.0
        Roanoke       VA       37.3      21.6
        Seattle       WA       47.4      14.7
        Grn_bay       WI       44.5      29.9
        Casper        WY       42.9      26.6
        ')


# Scatterplot of Temperature versus Latitude
with(theData, plot(RANGE, LAT, main = 'Scatterplot of Temperature versus 
Latitude'))
```

Part II

Fitting the SLR model

Based on the scatterplot produced above, we assume that an appriopriate regression model relating RANGE and LAT is the linear model given by

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

where Y is the RANGE, X is the LAT, and ϵ is a random error term that is normally distributed with the mean 0 and the unknown variance σ^2.

β_0 is the estimate of the Y-intercept. and β_1 is the estimate of the slope coefficient.

```
# Fit the model

SLR_model <- lm(RANGE ~ LAT, data = theData)
```

```r
summary(SLR_model)



## R Student
rStudent <- rstudent(SLR_model)    # get r-student values  N/B call rStudent
to print the Rstudent scores

# Install.packages('car')
library(car)
outlierTest(SLR_model)  # run test to get possible outliers

## Hat diagonal values
HatDiag <- lm.influence(SLR_model)$hat        # get Hat Diag values
cutoff <- 2*(length(coef(SLR_model))/length(HatDiag))     # cu-off at 2*p/n
a <- theData[which(lm.influence(SLR_model)$hat > cutoff),]   # Print obs with
HAT > 2*p/n
cbind(a, HatDiag[HatDiag > cutoff])
plot(HatDiag, ylab="HatDiag")


# Get ALL INFLUENCE MEASURES DISCUSSED
influence.measures(SLR_model)


## To get individual outliers run the codes below.

## DFFITS
DFFITS_model <- abs(dffits(SLR_model))          # get absolute values of
DFFITS
b <- theData[which(DFFITS_model > 1),]    # get DFFITS > 1
cbind(b, DFFITS = DFFITS_model[DFFITS_model > 1])    # Print obs with DFFITS >
1
plot(dffits(SLR_model), ylab="DFFITS")      # Plot all DFFITS


## DFBETAS
DFBETAS_model <- abs(dfbetas(SLR_model)[, 'LAT'])    # get absolute values of
DFBETAS for LAT
c <- theData[which(DFBETAS_model > 1),]          # get LAT DFBETAS > 1
cbind(c, DFBETA_LAT = DFBETAS_model[DFBETAS_model > 1])    # Print obs with
LAT DFBETAS > 1
plot(DFBETAS_model, ylab="DFBETA (LAT)")    # Plot LAT DFBETAS


## COOK's Distance
COOKS_mod <- cooks.distance(SLR_model)  # get cook's D for all observations
cutoff <- 4/length(COOKS_mod)    # cut off at 4/n
```

```r
d <- theData[which(COOKS_mod > cutoff),]
cbind(d, COOKsD = COOKS_mod[COOKS_mod > cutoff])
```

Part III

Evaluate Assumptions - Residual Analysis

Residual Plot can be used to detect various problems such as non-linear pattern, non-homogeneous variances and outliers.

- If the data is of homogeneity, most of residual points of data scatter around zero.

- If problesm such as courvature or non-homogenous variance are detected in residual plot, we may need to consider fitting a more complicated model.

Shapiro-Wilk Test is conducted on the **RESIDUALS of the fitted model** to check for normiality. If the p-value of this test is less than the signifcant level of 0.05, the null hypothesis is rejected and we conclude that the data was not sampled from a normally distributed population. Otherwise, we fail to reject to rejct the null and conclude that the data is normally distributed.

```r
par(mfrow = c(1,2))
plot(SLR_model, which = 1:2)
```

Alternative plotting using the olsrr package

```r
ols_plot_resid_fit(SLR_model)
ols_plot_resid_qq(SLR_model)
ols_test_normality(SLR_model)  ## Test normality
```

Lab Assignment

Your assignment is to perform necessary analysis using R to answer the following questions.

1. Use **Residual Plot** to check the assumption of homogeneity of variance. Does the data set appear to be homogenous?

2. Use the **olsrr package** or any function you deem appropriate. Does RANGE appear to be normally distributed? Why? Is this relevant to the normality assumption? Why?

3. Using the **lm** function fit the regression model. Write down the regression equation and answer: Does the model fit the data well? Why? Is this relevant to the normality assumption? Why?

4. What is the predicted value of RANGE at LAT=42 ? (Hint use the **predict** function. See example below)

```r
predict(SLR_model, newdata=data.frame(LAT=29))  # remember to change to the required valude of LAT for this question
```

5. Does there appear to be any possible outlier(s)? State the name and value of the statistics that you use to reach your conclusion.