Problem Set 7

cdurso

March 07, 2018

These questions were rendered in R markdown through RStudio (https://www.rstudio.com/wp-content/uploads/2015/02/rmarkdown-cheatsheet.pdf, http://rmarkdown.rstudio.com).

Please complete the following tasks, in R where applicable. Please generate a solution document in R markdown and upload the rendered .doc, .docx, or .pdf document. You may add hand computations to a .doc or .docx if you prefer. In the rendered document, please show your code. That is, don't use "echo=FALSE".

In either case, your work should be based on the data's being in the same folder as the R files. Please turn in your work on Canvas. Your solution document should have your answers to the questions and should display the requested plots.

1. Suppose the $Y_i = mX_i + b + \varepsilon_i$ where the ε_i's are independent, identically distributed $Normal\left(0, \sigma^2\right)$ random variables. Define M and B as in the notes. What is the expected value of $Y_i - (X_i M + B)$?(10 points)

Response

From the definitions, $Y_i - (X_i M + B) = mX_i + b + \varepsilon_i - (X_i M + B) = X_i(m - M) + (b - B) + \varepsilon_i$. For the expected value, we can think if it as $E[X_i(m - M) + (b - B)] + E[\varepsilon_i] = E[X_i(m - M)] + E[(b - B)] + 0$ since this is the mean of the identical distribution.Now, since each X_i is constant, and $E[M] = m$ and $E[B] = b$, we have $X_i(E[M] - E[m]) + E[b] - E[B] = X_i(m - m) + (b - b) = 0$.

2. (1 point each) In the following, please use the"ads.txt"data, obtained from http://lib.stat.cmu.edu/DASL/Datafiles/tvadsdat.html then reformatted to load correctly using "read.table" with the option "header=TRUE". The data description follows: "This data appeared in the Wall Street Journal. The advertisement were selected by an annual survey conducted by Video Board Tests, Inc., a New York ad-testing company, based on interviews with 20,000 adults who were asked to name the most outstanding TV commercial they had seen, noticed, and liked. The retained impressions were based on a survey of 4,000 adults, in which regular product users were asked to cite a commercial they had seen for that product category in the past week.

FIRM: Firm name

SPEND: TV advertising budget, 1983 ($ millions)

MILIMP: Millions of retained impressions per week"

- 2.a. Use maximum likelihood regression to model MILIMP as a linear function of SPEND. Provide any diagnostic plots you feel are relevant. Is the relationship linear? Is the assumption of independent, identically distributed Normal errors justifiable? (15 points)

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.2.1 --
```

```
## v ggplot2 2.2.1     v purrr   0.2.4
## v tibble  1.4.1     v dplyr   0.7.4
## v tidyr   0.7.2     v stringr 1.2.0
## v readr   1.1.1     v forcats 0.2.0
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
```

```
## x dplyr::lag()    masks stats::lag()
```

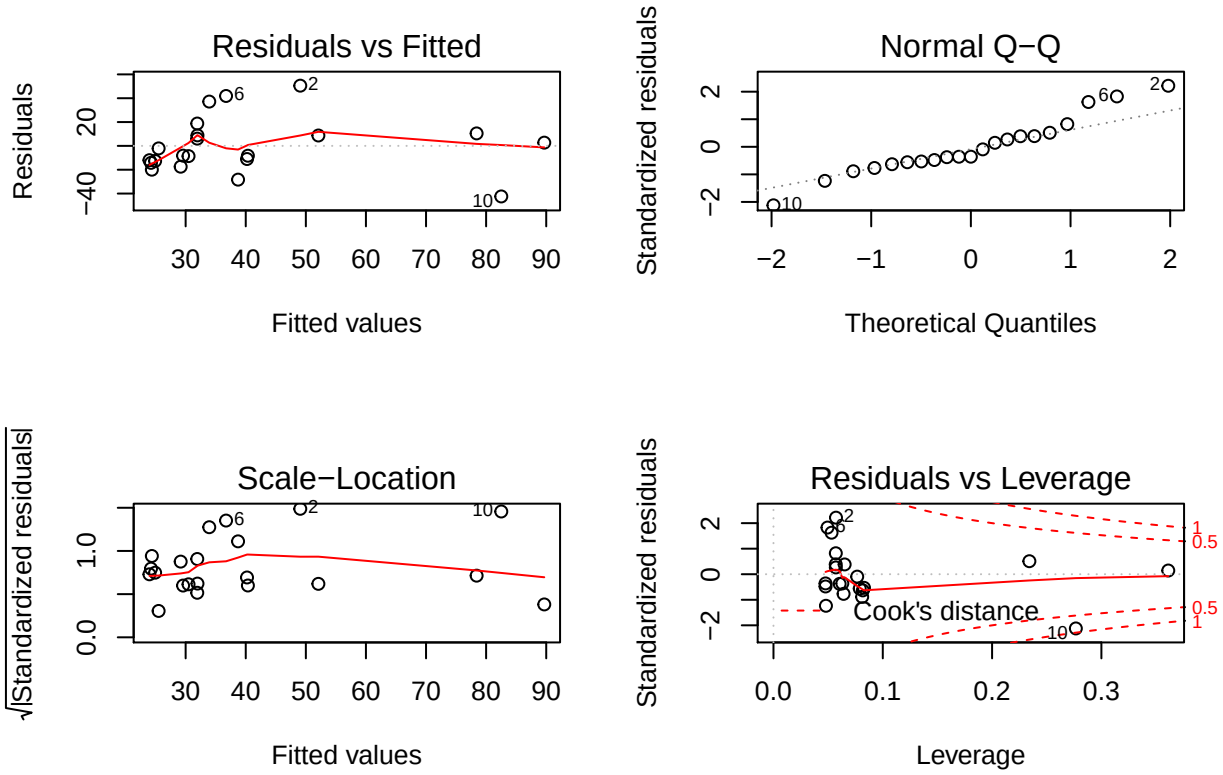
```
dataAds <- read.table("ads.txt", header = TRUE)

milimp <- dataAds$MILIMP
spend <- dataAds$SPEND

#milimp "~ = is modeled as" linear function of spend

lmfit <- lm(milimp~spend)
lmfit
```

```
##
## Call:
## lm(formula = milimp ~ spend)
##
## Coefficients:
## (Intercept)        spend
##     22.1627       0.3632
```

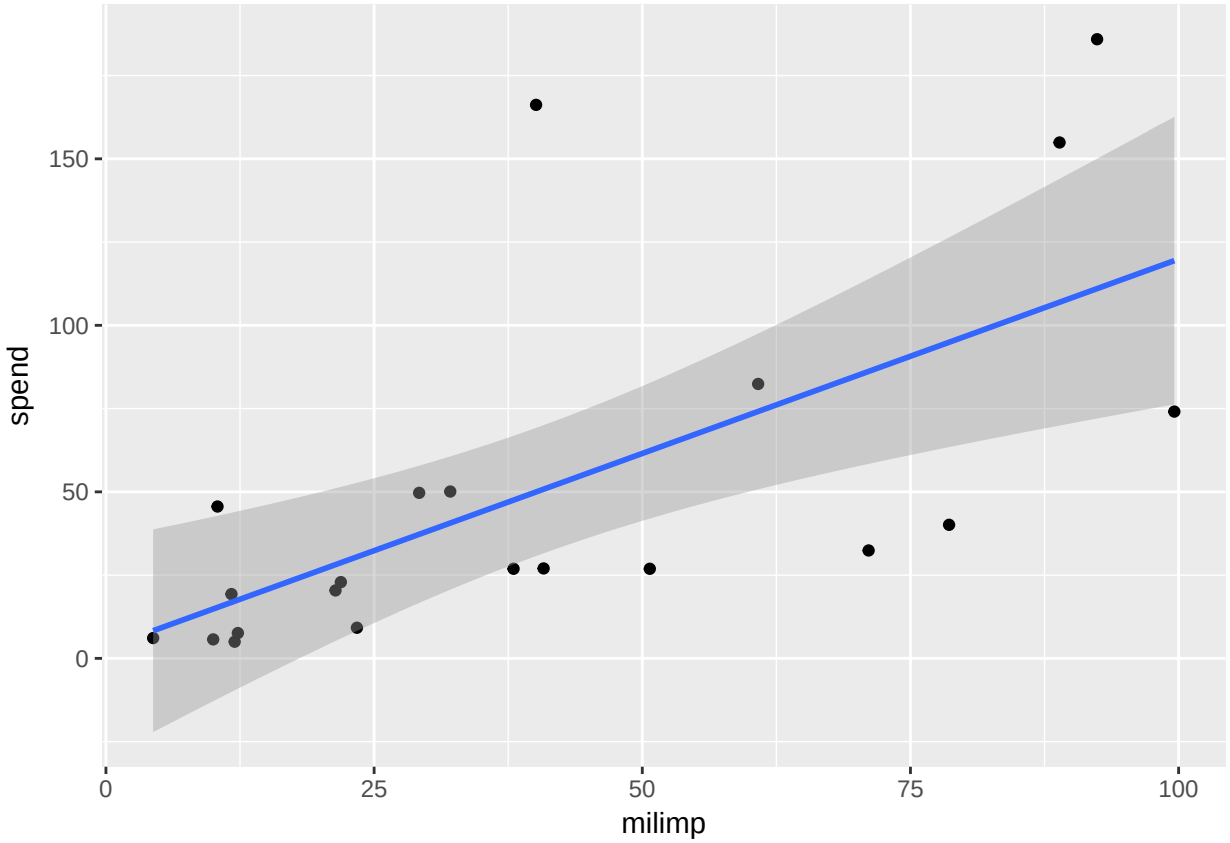
```
summary(lmfit)
```

```
##
## Call:
## lm(formula = milimp ~ spend)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -42.422 -12.623  -8.171   8.832  50.526
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 22.16269    7.08948   3.126  0.00556 **
## spend        0.36317    0.09712   3.739  0.00139 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 23.5 on 19 degrees of freedom
## Multiple R-squared:  0.424,  Adjusted R-squared:  0.3936
## F-statistic: 13.98 on 1 and 19 DF,  p-value: 0.001389
```

```
par(mfrow=c(2,2))
plot(lmfit)
```




```
g <- ggplot(lmfit, aes(x = milimp, y=spend))+geom_point()+stat_smooth(method = "lm")
g
```



Response

The data independent, identically distributed Normal errors is justifiable; the data is not a fan nor a funnel nor a curve. It is The relationship is linear due to the lack of obvious shape.

- 2.b. Use maximum likelihood regression to model log(MILIMP) as a linear function of log(SPEND). Provide any diagnostic plots you feel are relevant. Is the relationship linear? Is the assumption of independent, identically distributed Normal errors justifiable? (15 points)

```
logM <- log(milimp)
logS <- log(spend)

LogLmFit <- lm(logM~logS)
LogLmFit
```

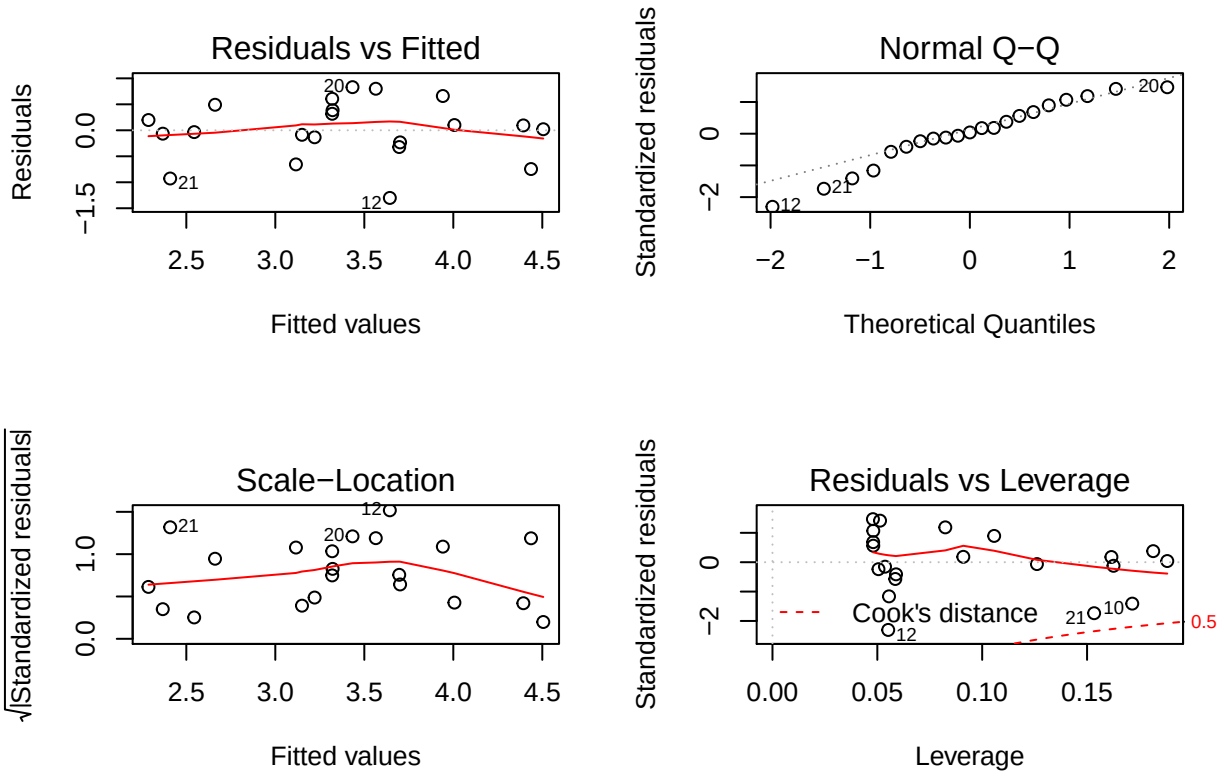
```
##
## Call:
## lm(formula = logM ~ logS)
##
## Coefficients:
## (Intercept)         logS
##      1.2999       0.6135
```

```
summary(LogLmFit)
```

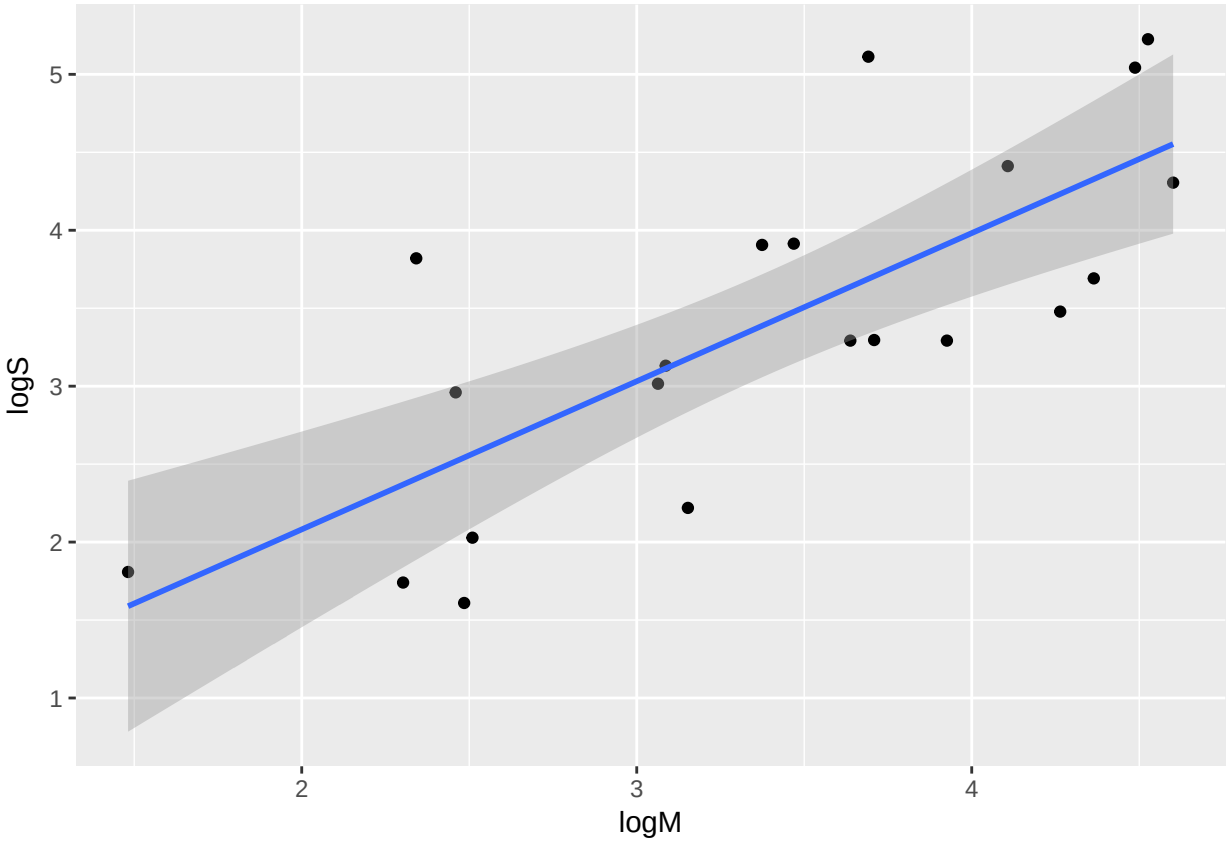
```
##
## Call:
```

```
## lm(formula = logM ~ logS)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.30158 -0.23227  0.02061  0.38680  0.83035
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   1.2999     0.4236   3.069  0.00632 **
## logS          0.6135     0.1191   5.153 5.66e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.581 on 19 degrees of freedom
## Multiple R-squared:  0.5829, Adjusted R-squared:  0.561
## F-statistic: 26.55 on 1 and 19 DF,  p-value: 5.655e-05
```

```
par(mfrow=c(2,2))
plot(LogLmFit)
```




```
g <- ggplot(LogLmFit, aes(x = logM, y = logS))+geom_point()+stat_smooth(method = "lm")
g
```



Response

The data independent, identically distributed Normal errors is probably justifiable; the data is not a fan nor a funnel nor a curve. The relationship is linear, as evidenced by the Residuals vs. Fitted graph not having an obvious shape.

- 2.c Which of the two regressions above looks like a better fit to the data? (5 points)

Response

The model in 2B appears to be a better fit, as the Multiple R-Squared value for the model in 2B is much higher than the Multiple R-Squared value for the model in 2A.

- 2.d For the model you chose in 2.c, what is the predicted MILIMP for SPEND=75? (5 points)

```
pred<-0.6135*log(75)+1.299
pred<-exp(pred)
```

Response

Using the slope of 0.6135 and intercept of 1.299, we would predict a value of 51.8201442 for SPEND=75.