Homework 4 R Script

Alexander Hernandez

11/17/2022

```r
library(PASWR)
```

```
## Loading required package: lattice
```

```r
library(MASS)
library(ISLR)
```

```
## Warning: package 'ISLR' was built under R version 4.2.2
```

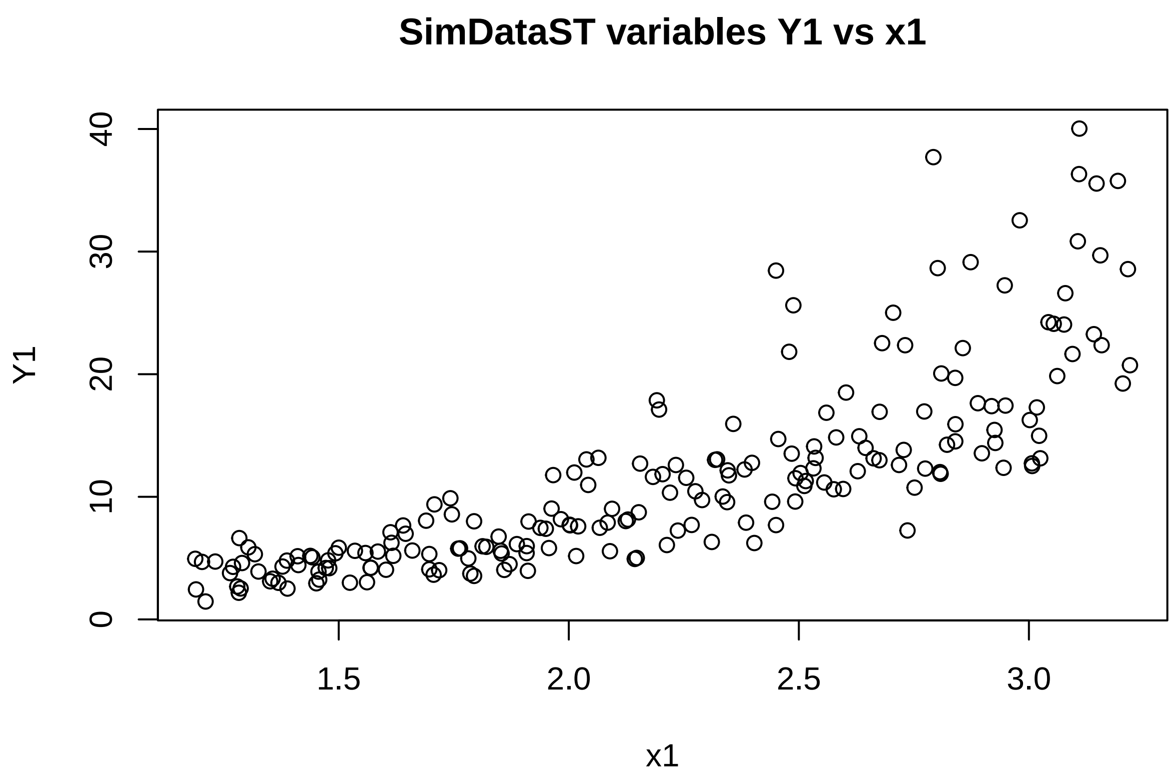
```r
library(UsingR)
```

```
## Loading required package: HistData

## Loading required package: Hmisc

## Loading required package: survival

## Loading required package: Formula

## Loading required package: ggplot2

##
## Attaching package: 'Hmisc'

## The following objects are masked from 'package:base':
##
##     format.pval, units

##
## Attaching package: 'UsingR'

## The following object is masked from 'package:survival':
##
##     cancer
```

1) SimDataST

a) Import the data in R and draw a scatterplot using x1 and Y1

```
attach(SimDataST)
plot(x1, Y1,
     main = "SimDataST variables Y1 vs x1")
```



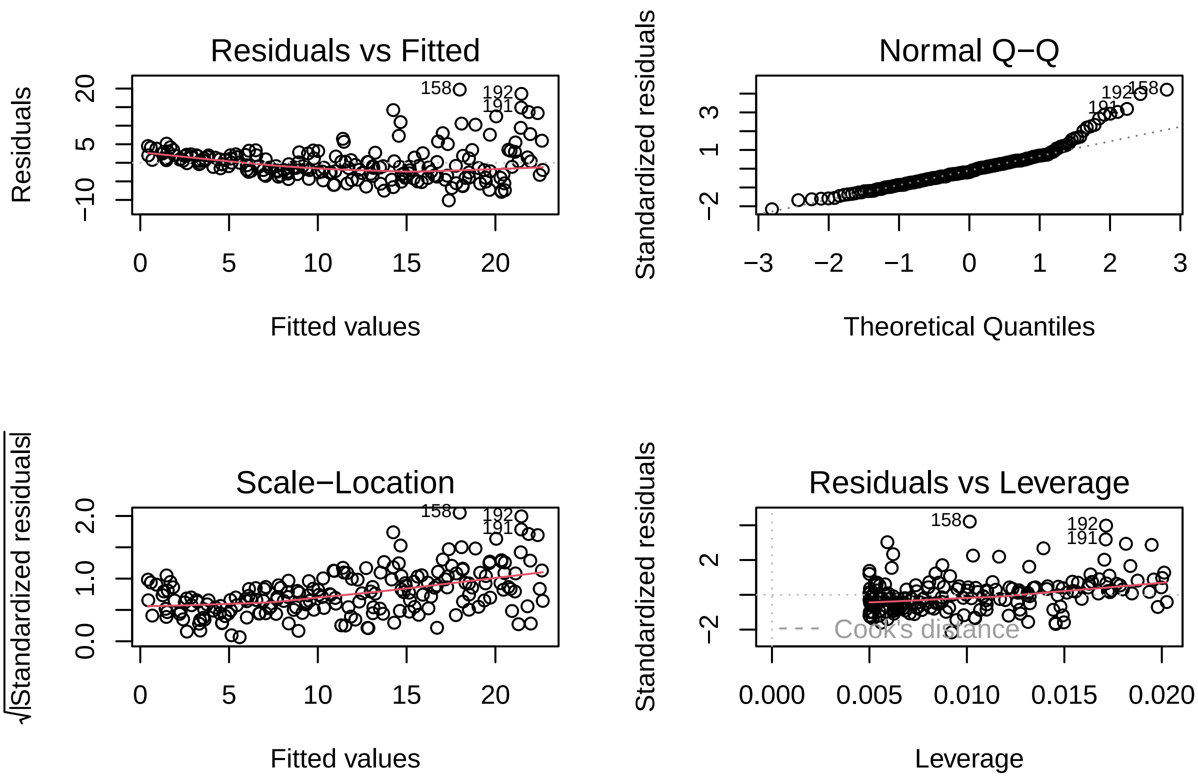
b) Fit a simple linear regression model using Y1 as response and x1 as regressor. Assess the residual plots of the model

```
model1b = lm(Y1 ~ x1)
summary(model1b)
```

```
##
## Call:
## lm(formula = Y1 ~ x1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -10.123  -2.918  -0.826   1.996  19.713
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -12.5720     1.2980  -9.686   <2e-16 ***
## x1           10.9452     0.5702  19.194   <2e-16 ***
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.708 on 198 degrees of freedom
## Multiple R-squared:  0.6504, Adjusted R-squared:  0.6487
## F-statistic: 368.4 on 1 and 198 DF,  p-value: < 2.2e-16
```

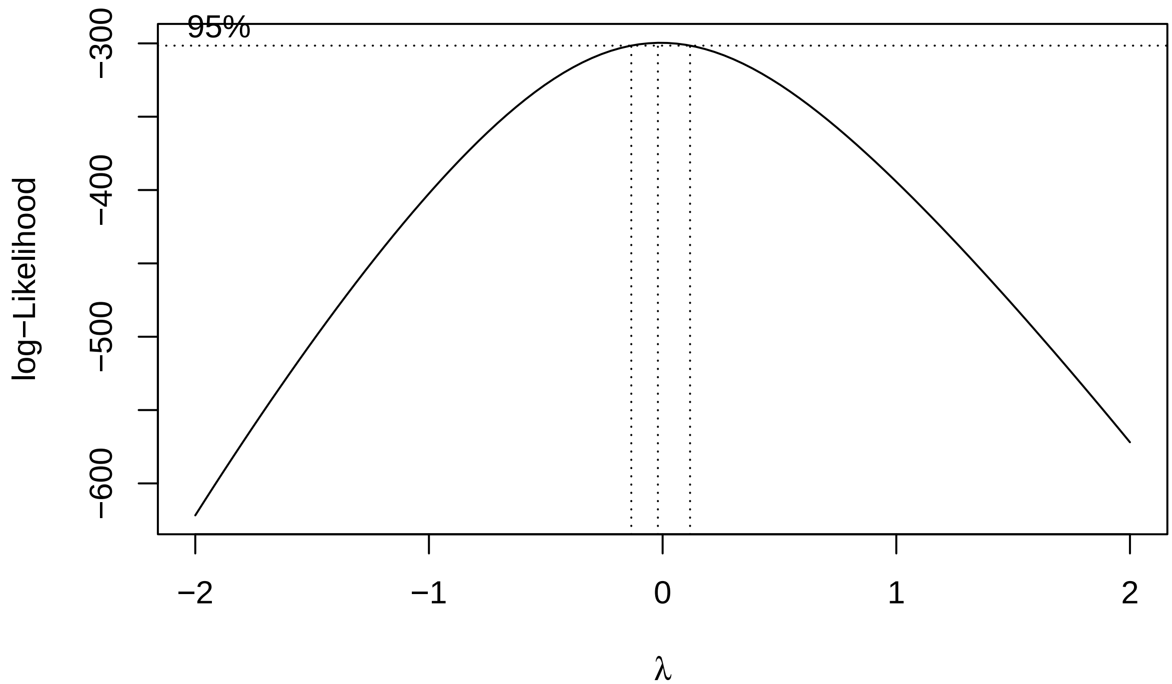
```
# Y1 = -12.5720 + 10.9452(x1)

par(mfrow=c(2,2))
plot(model1b)
```




```
# The residual plots do not look valid.
```

c) Determine a lambda value using Box-Cox transformation to improve the model

```
boxcox(model1b)
```




```
# A lambda value around 0.1 would improve the model
```

d) Fit a simple regression model after transformation

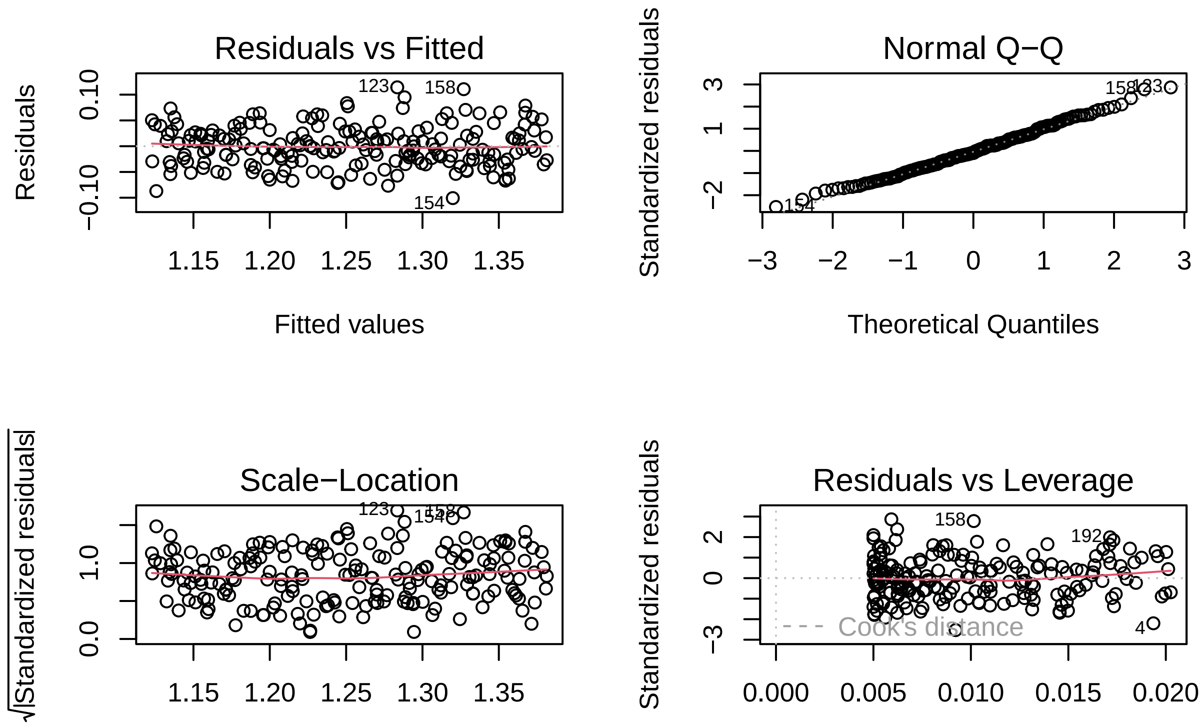
```
model1d = lm(Y1**0.1 ~ x1)
summary(model1d)
```

```
##
## Call:
## lm(formula = Y1^0.1 ~ x1)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.100724 -0.028771 -0.003729  0.025765  0.114189
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 0.971612   0.011021   88.16   <2e-16 ***
## x1          0.127278   0.004842   26.29   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.03998 on 198 degrees of freedom
## Multiple R-squared:  0.7773, Adjusted R-squared:  0.7761
## F-statistic: 690.9 on 1 and 198 DF,  p-value: < 2.2e-16
```

```
# Y1 = -12.5720 + 10.9452(x1)
```

e) Compare the results in b and d. Was the transformation worth it?

```
par(mfrow=c(2,2))
plot(model1d)
```




```
# The plots have improved and thus the transformation is worth it
```

2) Cars93

a) How many variables are included in the dataset

```
names(Cars93)
```

```
##  [1] "Manufacturer"       "Model"              "Type"
##  [4] "Min.Price"          "Price"              "Max.Price"
##  [7] "MPG.city"           "MPG.highway"        "AirBags"
## [10] "DriveTrain"         "Cylinders"          "EngineSize"
## [13] "Horsepower"         "RPM"                "Rev.per.mile"
## [16] "Man.trans.avail"    "Fuel.tank.capacity" "Passengers"
## [19] "Length"             "Wheelbase"          "Width"
## [22] "Turn.circle"        "Rear.seat.room"     "Luggage.room"
## [25] "Weight"             "Origin"             "Make"
```

```
length(names(Cars93))
```

```
## [1] 27
```

b) Fit a regression model for MPG.city using the numerical variables EngineZie, Weight, Passengers, and Price

```
attach(Cars93)
model2b = lm(MPG.city ~ EngineSize +
                        Weight +
                        Passengers +
                        Price)

model2b
```

```
##
## Call:
## lm(formula = MPG.city ~ EngineSize + Weight + Passengers + Price)
##
## Coefficients:
## (Intercept)   EngineSize       Weight   Passengers        Price
##   46.389413     0.196119    -0.008207     0.269622    -0.035804
```

```
# MPG.city = 46.389 + 0.196(EngineSize) - 0.008(Weight) +
#             0.270(Passengers) - 0.036(Price)
```

c) Which variables are marked as statistically significant by the marginal t-test?

```
summary(model2b)
```

```
##
## Call:
## lm(formula = MPG.city ~ EngineSize + Weight + Passengers + Price)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.1207 -1.9098  0.0522  1.1294 13.9580
##
## Coefficients:
```

```
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 46.389413   2.097516  22.116  < 2e-16 ***
## EngineSize   0.196119   0.588880   0.333    0.740
## Weight      -0.008207   0.001343  -6.111 2.63e-08 ***
## Passengers   0.269622   0.424951   0.634    0.527
## Price       -0.035804   0.049179  -0.728    0.469
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.06 on 88 degrees of freedom
## Multiple R-squared:  0.7165, Adjusted R-squared:  0.7036
## F-statistic: 55.59 on 4 and 88 DF,  p-value: < 2.2e-16
```

```r
# Weight is the only variable marked as statistically significant
```

d) Which model is selected by AIC criteria?

```r
model2d = lm(MPG.city ~ Weight)

AIC(model2b, k=5)
```

```
## [1] 496.7923
```

```r
AIC(model2d, k=2)
```

```
## [1] 474.6028
```

```r
# 496.7923 and 474.6028
# Given two models, one using EngineSize, Weight, Passengers, and Price and
# another using only the statistically significant variable Weight,
# the model containing only Weight has a lower AIC and thus is chosen.
```

3) Home Ownership to Family Income

a) Fit a simple logistic regression model for the subject data and display with the scatterplot

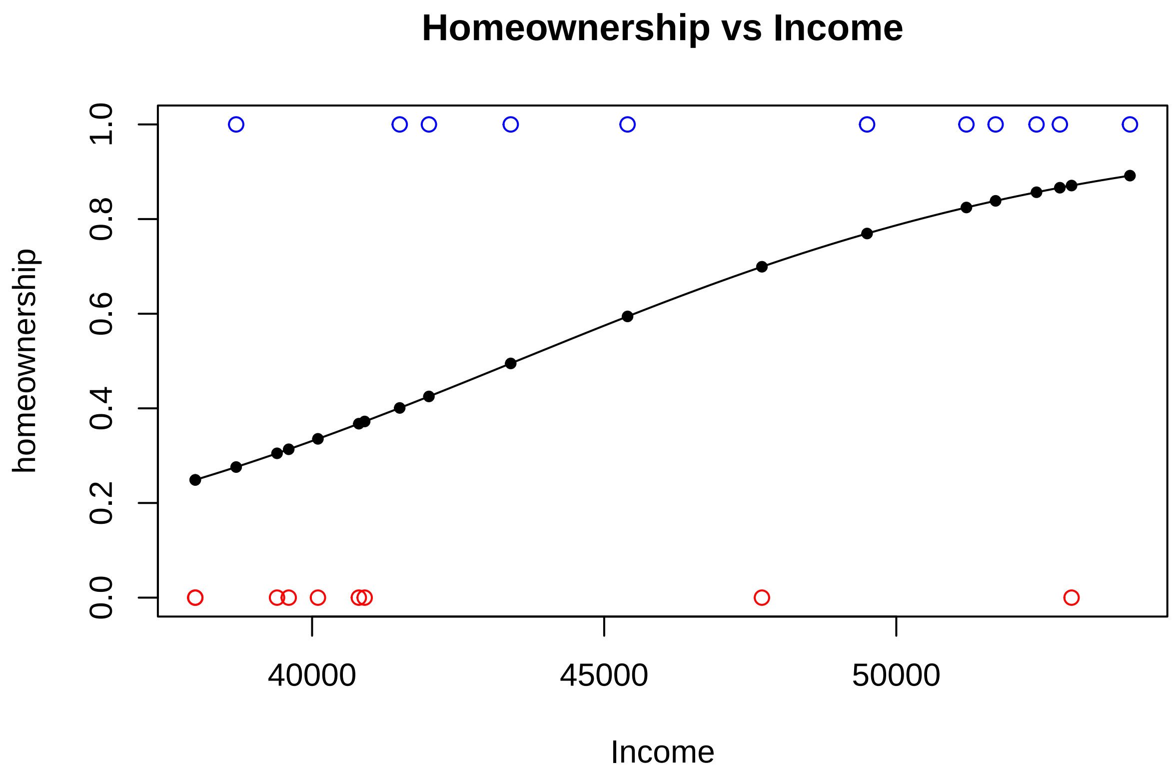
```
homedata = read.csv("C:\\repos\\STAT 50001\\Homework 4\\Homedata.csv")
homedata
```

```
##    Income homeownership
## 1   38000             0
## 2   51200             1
## 3   39600             0
## 4   43400             1
## 5   47700             0
## 6   53000             0
## 7   41500             1
## 8   40800             0
## 9   45400             1
## 10  52400             1
## 11  38700             1
## 12  40100             0
## 13  49500             1
## 14  38000             0
## 15  42000             1
## 16  54000             1
## 17  51700             1
## 18  39400             0
## 19  40900             0
## 20  52800             1
```

```
attach(homedata)
model3 = glm(homeownership ~ Income,
                     family = binomial(logit))
model3
```

```
##
## Call:  glm(formula = homeownership ~ Income, family = binomial(logit))
##
## Coefficients:
## (Intercept)       Income
##  -8.7395139    0.0002009
##
## Degrees of Freedom: 19 Total (i.e. Null);  18 Residual
## Null Deviance:       27.53
## Residual Deviance: 22.43     AIC: 26.43
```

```
# homeownership = [1 + exp( 8.7395139 - 0.0002009(Income) )]^-1

plot(homedata, main = "Homeownership vs Income",
     col=ifelse(homeownership==0, "red", "blue"),
)
curve(predict(model3, data.frame(Income=x), type="resp"), add=TRUE)
points(Income, fitted(model3), pch=20)
```



b) What is the estimated porbability that a family with an income of $45,000 owns a house?

```
predict(model3, data.frame(Income=45000), type="resp")
```

```
##         1
## 0.5747456
```

```
# There is a 0.5747 chance a family of income $45,000 owns a house.
```

4) Defaulting on a Credit Card Versus Annual Income and Balance

```
attach(Default)
model4 = glm(default ~ balance,
             family = binomial(logit))
model4
```

```
##
## Call:  glm(formula = default ~ balance, family = binomial(logit))
##
## Coefficients:
## (Intercept)      balance
##  -10.651331     0.005499
##
## Degrees of Freedom: 9999 Total (i.e. Null);  9998 Residual
## Null Deviance:       2921
## Residual Deviance: 1596  AIC: 1600
```

```
# default = [1 + exp( 10.651331 - 0.005499(balance) )]^-1
```

5) KeepKidsHealthy - fetal smoking and malnutrition on premature births

a) Extract the variables of interest: gestation, smoking status, mother's height and weight, and birth weight of the babies

```
b = babies[ , c("gestation", "smoke", "ht", "wt1", "wt")]
attach(b)
```

b) Clean the data set as there are some missing values coded as 9, 99, or 999

```
bedit = b
is.na(bedit) = bedit == 9 | bedit == 99 | bedit == 999
bedit = na.omit(bedit)
```

c) Calculate the BMI of mothers

```
bedit["BMI"] = (bedit["wt1"] * 0.453592) / (bedit["ht"] * 0.0254)
head(bedit["BMI"], 5)
```

```
##        BMI
## 1 28.80315
## 2 37.66912
## 3 32.08851
## 5 33.31708
## 6 26.78693
```

d) Create indicator variable (1 for premature and 0 for not premature) babies

```
bedit["premature"] = with(bedit, ifelse(bedit["gestation"] < 259, 1, 0))
```

e) Fit a logistic regression model with smoke and BMI as a predictor variable and premature as a response variable

```
attach(bedit)
```

```
## The following objects are masked from b:
##
##     gestation, ht, smoke, wt, wt1
```

```
model5 = glm(premature ~ smoke + BMI,
             family = binomial(logit))
model5
```

```
##
## Call:  glm(formula = premature ~ smoke + BMI, family = binomial(logit))
##
## Coefficients:
## (Intercept)        smoke          BMI
##    -3.43570      0.09376      0.02491
##
## Degrees of Freedom: 1152 Total (i.e. Null);  1150 Residual
## Null Deviance:       636.8
```

```
## Residual Deviance: 634.6     AIC: 640.6
```

```
# homeownership = [1 + exp( 3.53570 - 0.09376(smoke) - 0.02491(BMI) )]^-1
```