HW7

Getong Zhong

2022-12-16

Question 1

(a)

The first concern is there is a weak level of linear relationship between the two variables. The second concern is the correlation between the two variables are really low that might be hard to present a valid linear model.

(b)

From the table we can see that the adj-R square is really low, 0.046, which means only 4.6% of the variation of the data are explained by the model. Besides, the p value of the predictors are greater than 0.05, which means it is not statistically significant. However when we look at the graph, we can observe a log shaped plot between the predictors and dependent variable. Therefore, it indicates that although there are not a linear relationship between these two variable, there is still potential relationship between Y and X.

Question 2

(a)

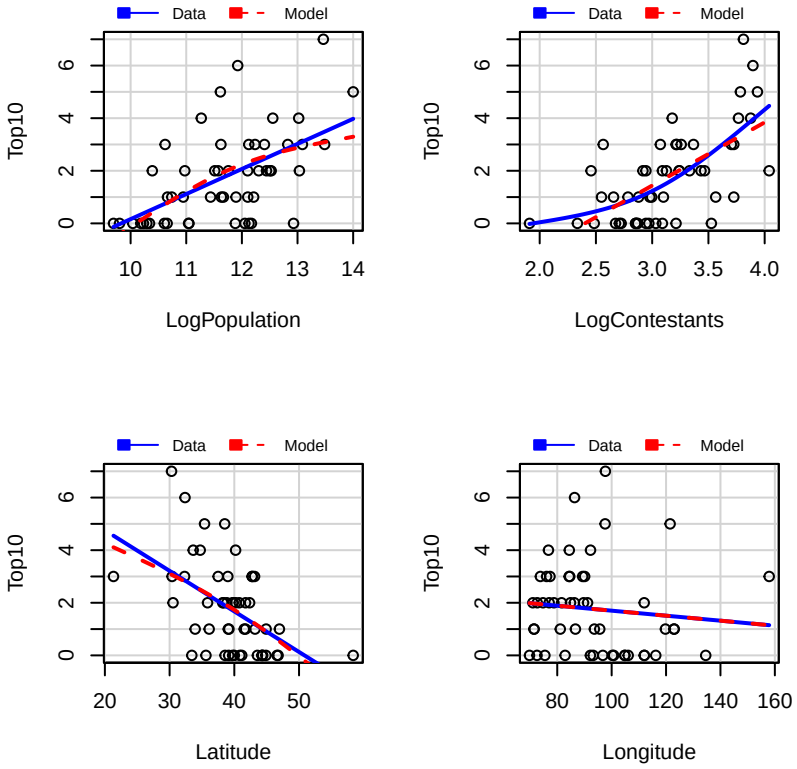
```
library(car)
data1 <- read.table("C:/Users/tonyg/Desktop/Academic/Grad/HUDM 5126/MissAmericato2008.txt"
                    , header = TRUE)

attach(data1)
model1 <- glm(Top10 ~ LogPopulation + LogContestants + LogTotalArea + Latitude + Longitude)
summary(model1)
```

```
##
## Call:
## glm(formula = Top10 ~ LogPopulation + LogContestants + LogTotalArea +
##     Latitude + Longitude)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.5605  -0.7787   0.0023   0.7419   3.0750
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept)    -4.682228   3.294448  -1.421  0.16214
## LogPopulation   0.467997   0.235114   1.991  0.05263 .
## LogContestants  1.596804   0.547194   2.918  0.00548 **
## LogTotalArea   -0.196070   0.170520  -1.150  0.25629
## Latitude       -0.064592   0.037686  -1.714  0.09342 .
## Longitude       0.006435   0.012280   0.524  0.60282
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 1.555903)
##
##     Null deviance: 149.176  on 50  degrees of freedom
## Residual deviance:  70.016  on 45  degrees of freedom
## AIC: 174.89
##
## Number of Fisher Scoring iterations: 2
```

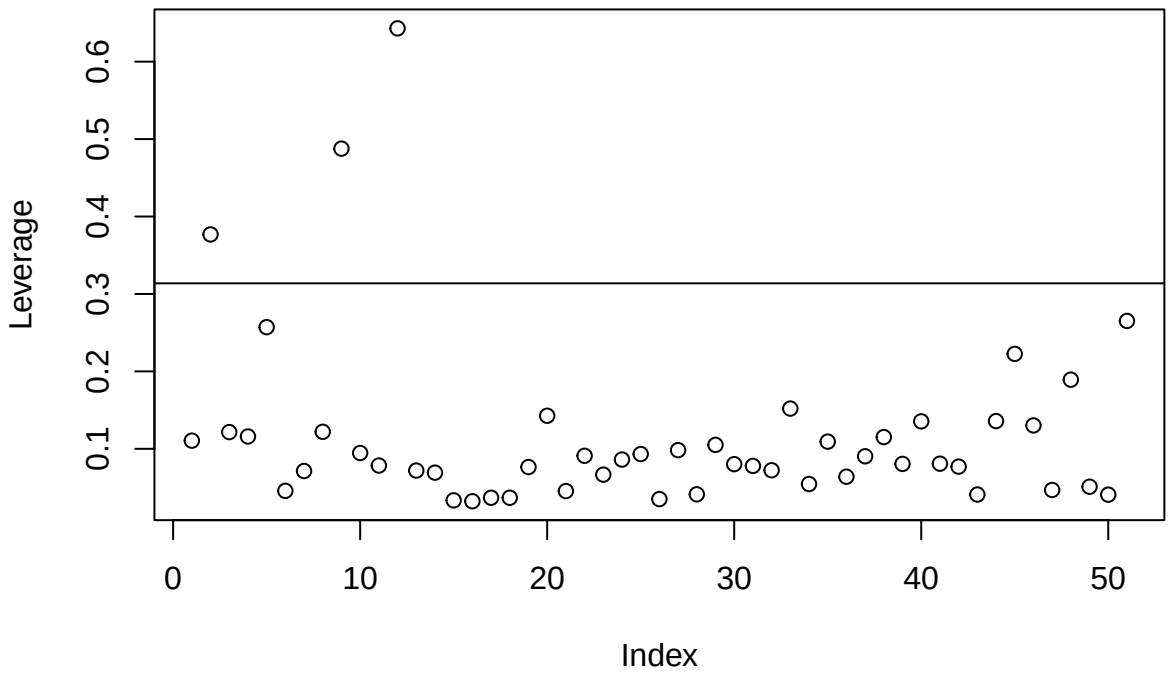
```
detach(data1)
```

```
attach(data1)
par(mfrow = c(2, 3))
mmp(model1, LogPopulation,
    ylab = "Top10", xlab = "LogPopulation")
mmp(model1, LogContestants,
    ylab = "Top10", xlab = "LogContestants")
mmp(model1, LogTotalArea,
    ylab = "Top10", xlab = "LogTotalArea")
mmp(model1, Latitude,
    ylab = "Top10", xlab = "Latitude")
mmp(model1, Longitude,
    ylab = "Top10", xlab = "Longitude")
detach(data1)
```



From the marginal model plots we can see that the non-parametric fits on the original response and the fitted values do not match well, which proves that the model is not valid. ### (b) Check for high leverage points

```
plot(hatvalues(model1), ylab = "Leverage")
abline(h = 16/nrow(data1))
```




```
high_lev_index = which(hatvalues(model1) > 16/nrow(data1))
high_lev_index
```

```
##  2  9 12
##  2  9 12
```

check for bad leverage points

```
outlier_index = which(abs(rstandard(model1)) > 2)
outlier_index
```

```
##  1 26 44
##  1 26 44
```

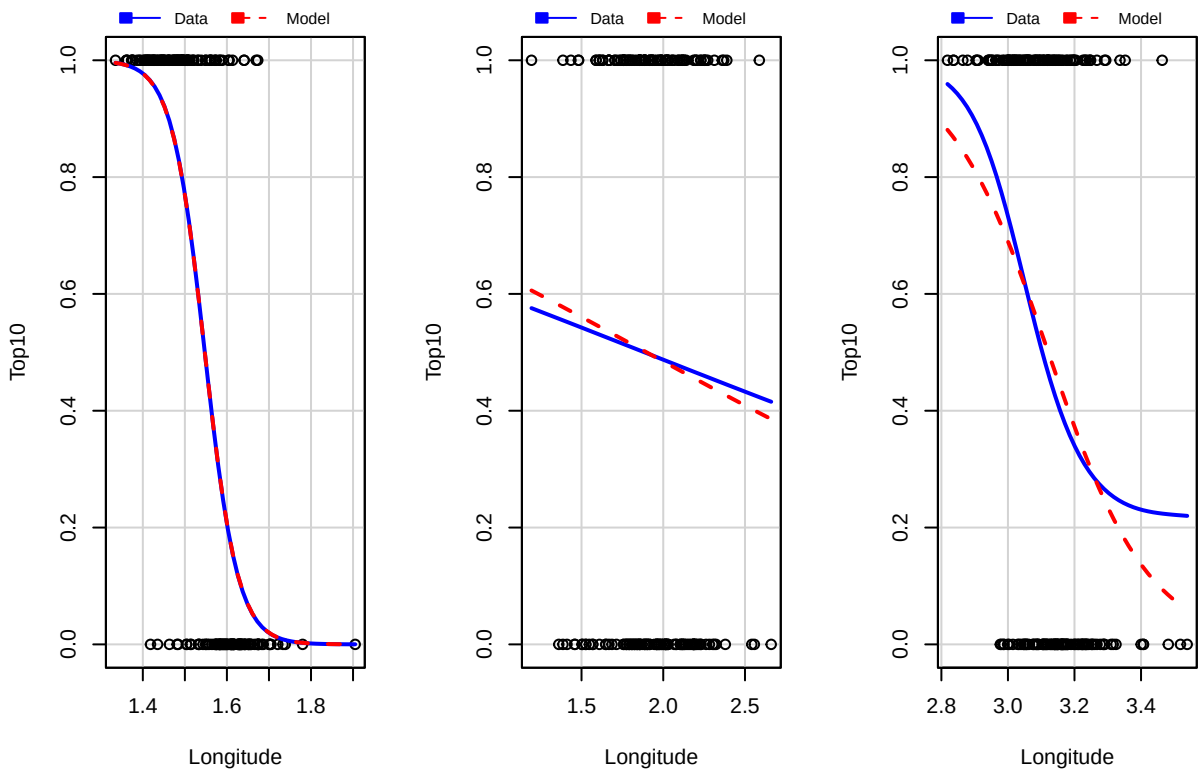
Compare the outcome with the high leverage points, we can conclude that there are no bad leverage point. ### (c) From the F test there only LogPopulation and LogContestants are statistically significant to the model. The coefficient in the model means that with 1 unit of increase in all population, contestants , Total area, Latitude and logitude, there will have a change of 0.467 * log of population plus 1.59 * log of Contestants plus 0.19 * log of Total area minus 0.064 * Latitude plus 0.006 * Longitude in Y. ## Question 3

```
data2 <- read.table("C:/Users/tonyg/Desktop/Academic/Grad/HUDM 5126/ais.txt", header = TRUE)
```

```r
model2 <- glm(Sex ~ RCC + WCC + BMI,family="binomial", data = data2)
summary(model2)
```

```
##
## Call:
## glm(formula = Sex ~ RCC + WCC + BMI, family = "binomial", data = data2)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -2.56012  -0.52459  -0.02467   0.52242   2.67643
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) 29.29399    3.92411   7.465 8.32e-14 ***
## RCC         -5.34706    0.72118  -7.414 1.22e-13 ***
## WCC          0.15505    0.12190   1.272  0.20338
## BMI         -0.22911    0.08723  -2.626  0.00863 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 280.01  on 201  degrees of freedom
## Residual deviance: 147.01  on 198  degrees of freedom
## AIC: 155.01
##
## Number of Fisher Scoring iterations: 5
```

```r
par(mfrow = c(1,3))
mmp(model2, log(data2$RCC), ylab = "Top10", xlab = "Longitude")
mmp(model2, log(data2$WCC), ylab = "Top10", xlab = "Longitude")
mmp(model2, log(data2$BMI), ylab = "Top10", xlab = "Longitude")
```



Question 4

(a)

I think it is not a valid data. From the marginal model plots we can see that the non-parametric fits on the original response and the fitted values do not match well. From the f test of the coefficient we can see that not all the predictors are statistically significant.

(b)

I think the correlation between age and blood pressure could add to the model, because age and blood pressure might have certan relationship if we add such cross predictor to the model can help reduce the Multicollinearity of the model. also certain transformation of insignificant predictors x1 and x4 can also been added to the model.

(c)

From the result of mariginal model plots we can the a major imporvment in model 8.7 than model 8.6. All the predictors in the model are statistically significant predictors, therefore the model is valid.

(d)

coeffcient for x3 means 1 unit increase in x3 will lead to 0.903863 unit of increase in Y.

Question 5

(a)

```
data3<- read.csv("C:/Users/tonyg/Desktop/Academic/Grad/HUDM 5126/Fundraising.csv")
attach(data3)
model3 <- glm (TARGET_B ~ homeowner.dummy + NUMCHLD + INCOME +
                 gender.dummy + WEALTH + HV + Icmed + Icavg + IC15
               + NUMPROM + RAMNTALL + LASTGIFT + totalmonths + TIMELAG
               + AVGGIFT, family="binomial")
detach(data3)
summary(model3)
```

```
##
## Call:
## glm(formula = TARGET_B ~ homeowner.dummy + NUMCHLD + INCOME +
##     gender.dummy + WEALTH + HV + Icmed + Icavg + IC15 + NUMPROM +
##     RAMNTALL + LASTGIFT + totalmonths + TIMELAG + AVGGIFT, family = "binomial")
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -1.82742  -1.15743   0.06352   1.15125   1.94959
##
## Coefficients:
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)      1.816e+00  4.449e-01   4.083 4.45e-05 ***
## homeowner.dummy  7.488e-02  9.199e-02   0.814  0.41567
## NUMCHLD         -2.929e-01  1.116e-01  -2.623  0.00871 **
## INCOME           6.805e-02  2.552e-02   2.667  0.00766 **
## gender.dummy     7.385e-02  7.512e-02   0.983  0.32560
## WEALTH           2.173e-02  1.761e-02   1.234  0.21725
## HV               1.195e-04  6.051e-05   1.975  0.04823 *
## Icmed            1.301e-03  9.157e-04   1.421  0.15539
## Icavg           -1.944e-03  9.931e-04  -1.958  0.05026 .
## IC15             1.440e-04  4.320e-03   0.033  0.97341
## NUMPROM          4.211e-03  2.277e-03   1.849  0.06439 .
## RAMNTALL        -7.960e-05  3.218e-04  -0.247  0.80463
## LASTGIFT        -8.012e-03  7.178e-03  -1.116  0.26437
## totalmonths     -5.939e-02  9.754e-03  -6.088 1.14e-09 ***
## TIMELAG          5.853e-03  6.700e-03   0.874  0.38235
## AVGGIFT         -8.774e-03  1.056e-02  -0.831  0.40615
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 4325.2  on 3119  degrees of freedom
## Residual deviance: 4223.3  on 3104  degrees of freedom
## AIC: 4255.3
##
## Number of Fisher Scoring iterations: 4
```

Question 6

(a)

```
data4 <- read.table("C:/Users/tonyg/Desktop/Academic/Grad/HUDM 5126/banknote.txt",
                    header = TRUE)
model4 <- glm(Y ~ Bottom + Diagonal, family="binomial", data=data4)
```

```
## Warning: glm.fit: algorithm did not converge
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```
summary(model4)
```

```
##
## Call:
## glm(formula = Y ~ Bottom + Diagonal, family = "binomial", data = data4)
##
## Deviance Residuals:
##       Min         1Q     Median         3Q        Max
## -5.51e-04  -2.00e-08   0.00e+00   2.00e-08   6.58e-04
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)   99422.9  5433597.5   0.018    0.985
## Bottom          688.7    37796.3   0.018    0.985
## Diagonal       -751.8    41093.2  -0.018    0.985
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2.7726e+02  on 199  degrees of freedom
## Residual deviance: 1.0424e-06  on 197  degrees of freedom
## AIC: 6
##
## Number of Fisher Scoring iterations: 25
```