Homework 4

Lydia Strebe

July 20, 2019

Problem 1

```
Mdat<-read.csv("http://users.stat.umn.edu/~sandy/alr4ed/data/MinnWater.csv")
perCapitalUse=(Mdat$muniUse/Mdat$muniPop)*10^6
Mdat2=cbind(Mdat,perCapitalUse)
Mdat_mod=glm(log(Mdat2$perCapitalUse)~year+muniPrecip,data=Mdat2)
summary(Mdat_mod)
```

```
##
## Call:
## glm(formula = log(Mdat2$perCapitalUse) ~ year + muniPrecip, data = Mdat2)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -0.09766  -0.03057   0.01086   0.02871   0.07577
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.5040361  2.5925575   1.352    0.191
## year         0.0002155  0.0012969   0.166    0.870
## muniPrecip  -0.0102590  0.0020845  -4.922 7.21e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 0.001933179)
##
##     Null deviance: 0.087427  on 23  degrees of freedom
## Residual deviance: 0.040597  on 21  degrees of freedom
## AIC: -77.062
##
## Number of Fisher Scoring iterations: 2
```

a) The units for the intercept are natural log of thousands of gallons. The units for $\hat{\beta}_1$ are also natural log of thousands of gallons. The units for $\hat{\beta}_2$ are natural log of thousands of gallons per inch.

b) The intercept: Assuming the year is 0 and the average precipitation is 0 inches, the average water usage per person in thousands of gallons would be

```
exp(3.5040361)
```

```
## [1] 33.24938
```

$\hat{\beta}_1$: For every additional year, the average water usage per person in thousands of gallons (10^3) increases by

```
(exp(0.0002155)-1)*100

## [1] 0.02155232
```

percent.

$\hat{\beta}_2$: For every additional inch of precipitation in the metropolitan area, the average water usage per person in thousands of gallons (10^3) decreases by

```
 (1-exp(-0.0102590))*100

## [1] 1.020656
```

percent.

c) muniPrecip is more practically signficant than year. An additional year seems to change the average amount of water per person by a few gallons, not very practically significant. An additional inch of precipitation seems to change the average amount of water per person by a few hundred gallons of water, somewhat practically significant.

d) The only predictor that is statistically significant is muniPrecip since its p-value is much lower than 0.05. The other predictors have p-values greater than 0.10.

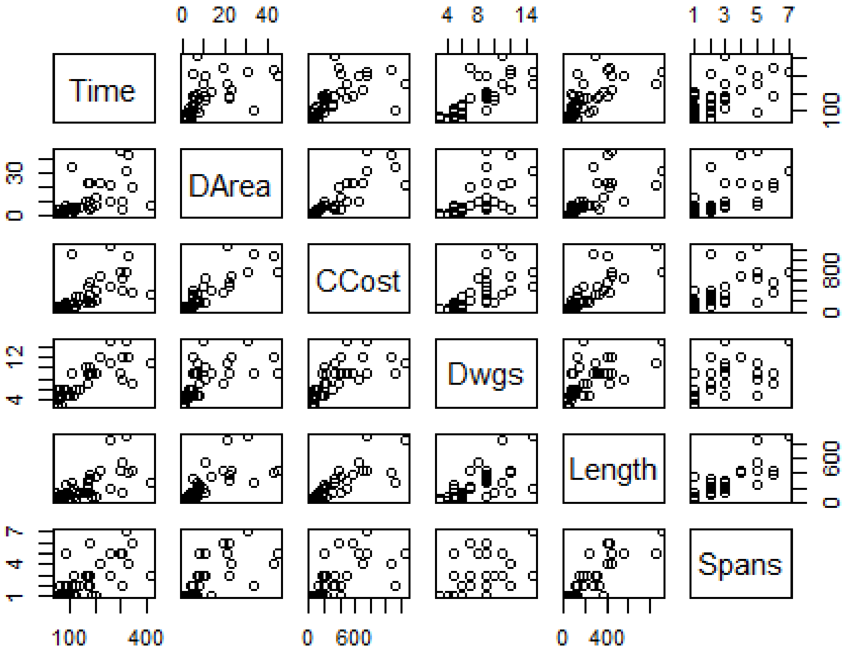
Problem 2

When there is a categorical variable with g groups, if we create g dummy variables, the columns will be linearly dependent. We would be able to express one of the columns as a linear combination of the other columns. That is why we always have g-1 dummy variables.

Problem 3

a)

```
Bridge=read.table("http://gattonweb.uky.edu/sheather/book/docs/datasets/bridge.
txt", header=TRUE)
new_Bridge=Bridge[,-1]
plot(new_Bridge)
```



The predictors look somewhat linearly related, but not all of the assumptions appear to hold. For example, equal variance does not seem to hold. The variance appears to spread as the values get further from the origin.

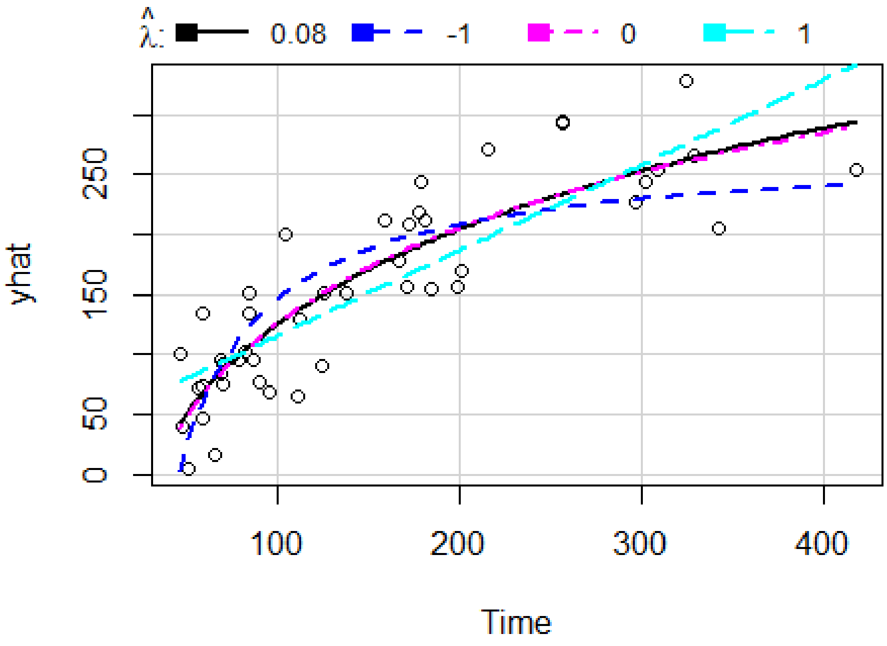
b)

```
library(car)

summary(powerTransform(new_Bridge[,2:6]))

## bcPower Transformations to Multinormality
##        Est Power Rounded Pwr Wald Lwr Bnd Wald Upr Bnd
## DArea    -0.1330           0      -0.3080       0.0420
## CCost    -0.1623           0      -0.3545       0.0299
## Dwgs     -0.2994           0      -0.8153       0.2165
## Length   -0.1893           0      -0.3966       0.0180
## Spans    -0.4225           0      -0.9395       0.0944
##
## Likelihood ratio test that transformation parameters are equal to 0
##  (all log transformations)
##                                 LRT df    pval
## LR test, lambda = (0 0 0 0 0) 7.625856  5 0.1781
##
## Likelihood ratio test that no transformations are needed
##                                 LRT df        pval
## LR test, lambda = (1 1 1 1 1) 253.1015  5 < 2.22e-16
```

```
bridge_mod=lm(Time~log(DArea)+log(CCost)+log(Dwgs)+log(Length)+log(Spans),data=
new_Bridge)
invResPlot(bridge_mod)
```

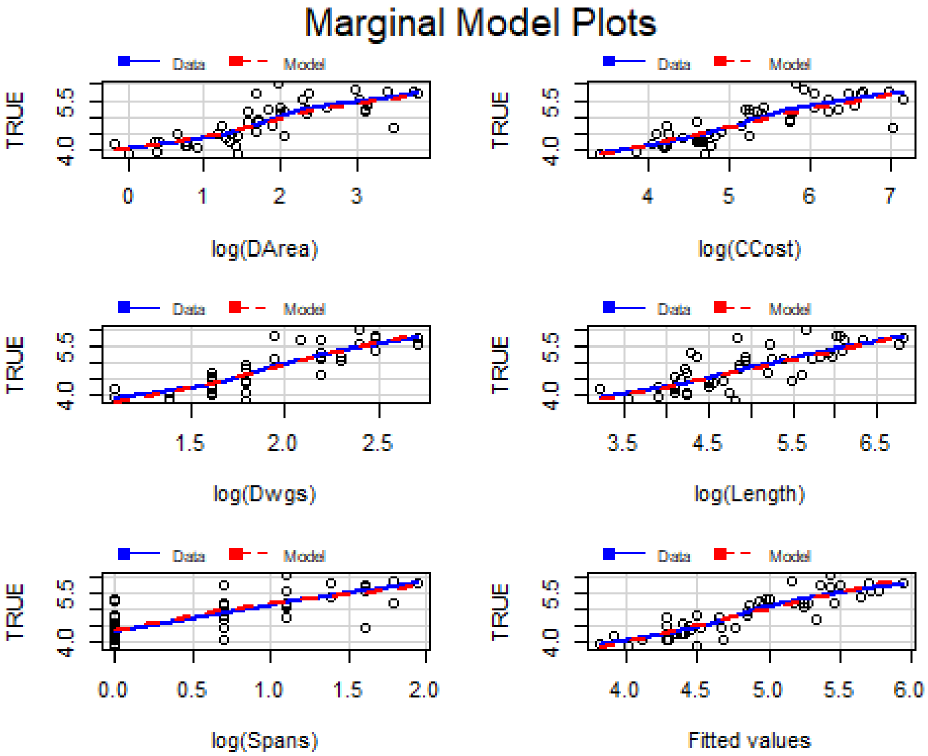


```
##        lambda      RSS
## 1  0.08115424 68107.53
## 2 -1.00000000 88988.87
## 3  0.00000000 68243.23
## 4  1.00000000 84705.50
```

According to our analysis, the most appropriate transformation for the predictor and response variables is the natural log function.

c)

```
bridge_mod2=lm(log(Time)~log(DArea)+log(CCost)+log(Dwgs)+log(Length)+log(Spans)
,data=new_Bridge)
mmps(bridge_mod2)
```

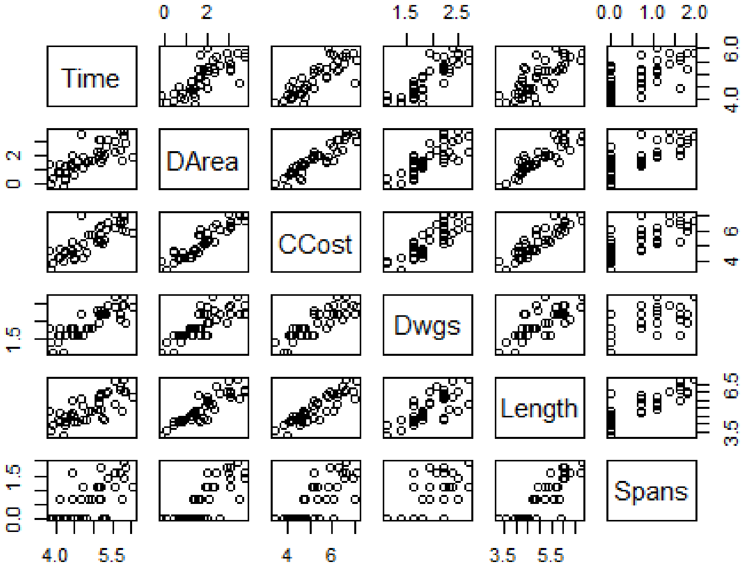


The model and the loess fit are very similar for all regressors, suggesting that the model is an adequate fit for the data.

d)

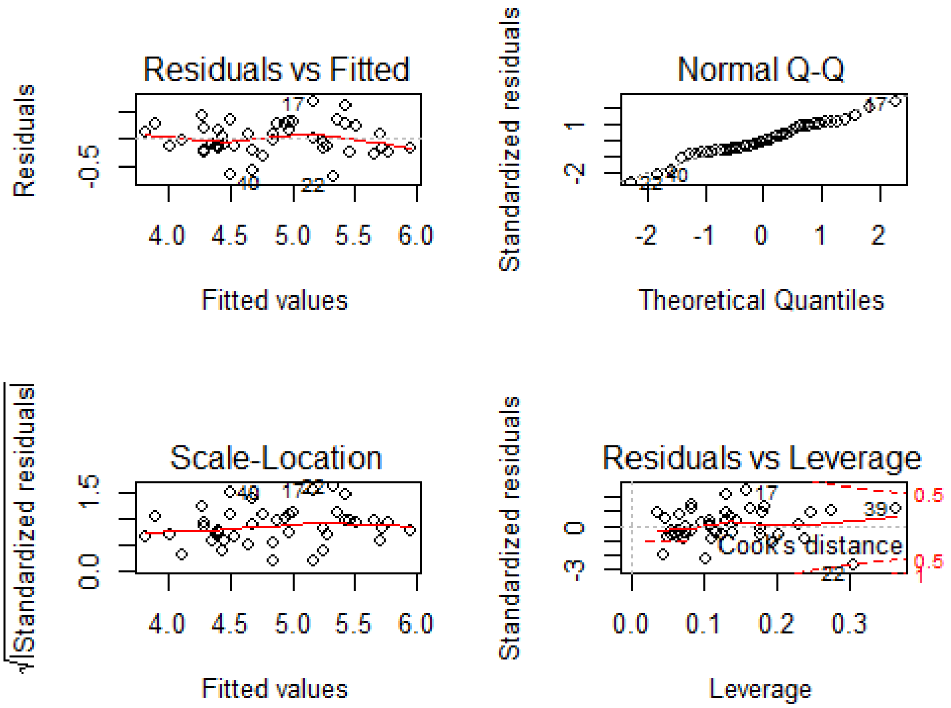
(1)

```
plot(log(new_Bridge))
```



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

(2)
plot(bridge_mod2)
```



```
par(mfrow=c(1,1))
```

From these plots, it looks like our model meets all the assumptions: Linearity, normality, and equal variance (we cannot assess independence from the plots). Additionally, all the points seem to be within Cook's distance.

(3)

```
summary(bridge_mod2)

## 
## Call:
## lm(formula = log(Time) ~ log(DArea) + log(CCost) + log(Dwgs) + 
##     log(Length) + log(Spans), data = new_Bridge)
## 
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -0.68394 -0.17167 -0.02604  0.23157  0.67307 
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept)  2.28590    0.61926   3.691 0.000681 ***
## log(DArea)  -0.04564    0.12675  -0.360 0.720705    
## log(CCost)   0.19609    0.14445   1.358 0.182426    
## log(Dwgs)    0.85879    0.22362   3.840 0.000440 ***
## log(Length) -0.03844    0.15487  -0.248 0.805296    
## log(Spans)   0.23119    0.14068   1.643 0.108349    
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 0.3139 on 39 degrees of freedom
## Multiple R-squared:  0.7762, Adjusted R-squared:  0.7475 
## F-statistic: 27.05 on 5 and 39 DF,  p-value: 1.043e-11
```

Based on the summary output of the model, plot 1 is log(Length) and plot 2 is log(Dwgs). We can see that the estimated coefficient of log(Dwgs) is larger, positive, and more statistically significant than the estimated coefficient of log(Length) (holding all other variables constant).