Homework-4

Leo

17/05/2022

First include basic settings and related packages.

```
> library(tidyverse)
> library(skimr)
> library(MASS)
```

Data analysis with dataset ‘bridge.txt’

1st. read data ‘bridge.txt’ in project’s working directory

```
> bridge <- read.table("bridge.txt",header = T)
> head(bridge)
```

```
##   Case  Time DArea CCost Dwgs Length Spans
## 1    1  78.8  3.60  82.4    6     90     1
## 2    2 309.5  5.33 422.3   12    126     2
## 3    3 184.5  6.29 179.8    9     78     1
## 4    4  69.6  2.20 100.0    5     60     1
## 5    5  68.8  1.44 103.0    5     60     1
## 6    6  95.7  5.40 134.4    5     60     1
```

```
> skim(bridge)
```

Table 1: Data summary

Name	bridge
Number of rows	45
Number of columns	7
Column type frequency:	
numeric	7

Table 1: Data summary

Group variables	None

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
Case	0	1	23.00	13.13	1.00	12.00	23.00	34.00	45.0	
Time	0	1	153.31	96.71	46.60	70.30	124.90	199.80	418.1	
DArea	0	1	9.89	10.82	0.85	3.43	5.48	10.36	45.0	
CCost	0	1	303.47	305.32	30.00	99.30	187.30	421.40	1264.1	
Dwgs	0	1	7.38	2.95	3.00	5.00	6.00	9.00	15.0	
Length	0	1	206.49	200.75	25.00	70.00	126.00	285.00	902.0	
Spans	0	1	2.33	1.68	1.00	1.00	2.00	3.00	7.0	

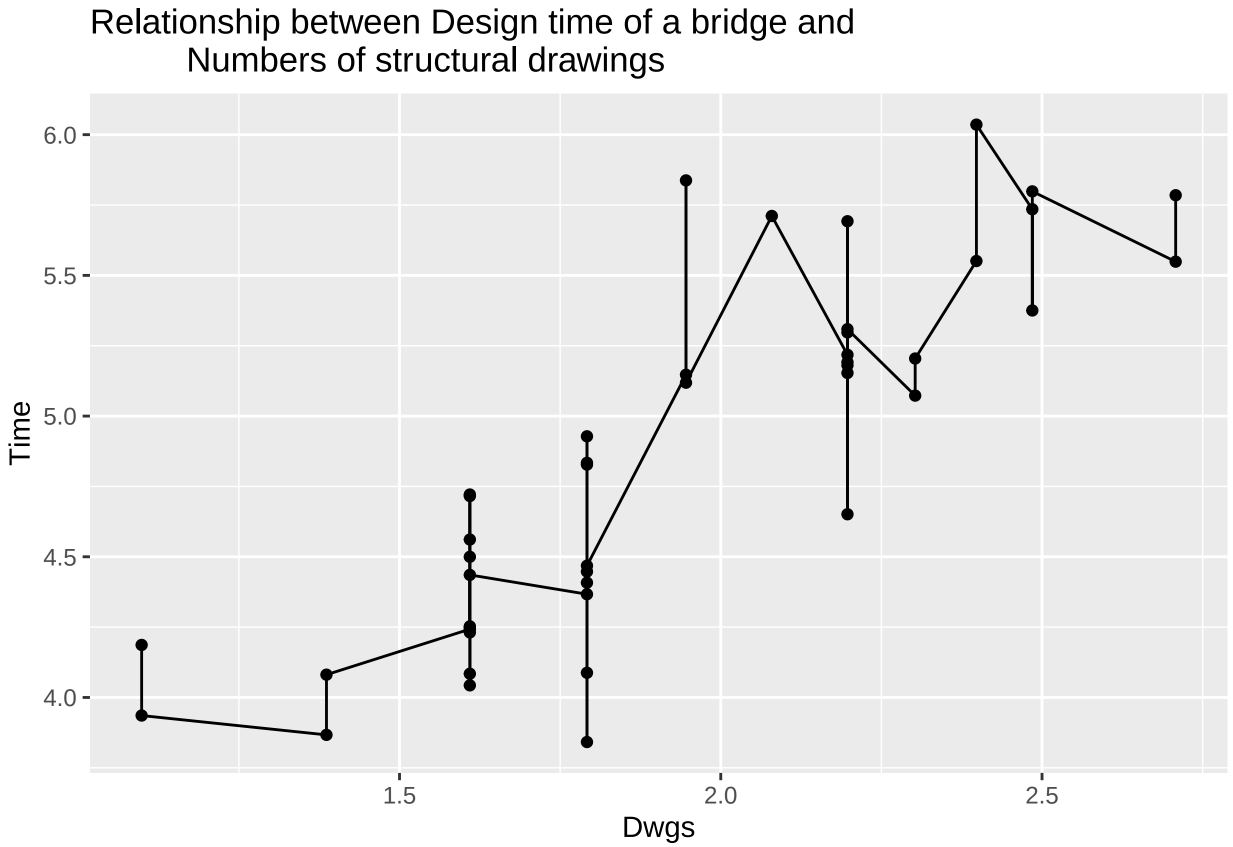
2nd. Delete the variable 'Case' and transform all the variables to the log form

```
> bridge <- bridge[,-1]
> bridge <- log(bridge)
> head(bridge)
```

```
##       Time     DArea    CCost     Dwgs   Length     Spans
## 1 4.366913 1.2809338 4.411585 1.791759 4.499810 0.0000000
## 2 5.734958 1.6733512 6.045716 2.484907 4.836282 0.6931472
## 3 5.217649 1.8389611 5.191845 2.197225 4.356709 0.0000000
## 4 4.242765 0.7884574 4.605170 1.609438 4.094345 0.0000000
## 5 4.231204 0.3646431 4.634729 1.609438 4.094345 0.0000000
## 6 4.561218 1.6863990 4.900820 1.609438 4.094345 0.0000000
```

3rd. EDA examples

```
> #Concerning design time of a bridge, I prefer Numbers of structural drawings to be more relevant. Fir
> bridge%>%
+   ggplot(mapping = aes(x = Dwgs,y = Time)) +
+   geom_point() +
+   geom_line() +
+   ggtitle("Relationship between Design time of a bridge and
+           Numbers of structural drawings")
```



4th. Fit a linear regression model to explain the Design time of a bridge

```
> #Construct the full linear regression model using Time as the response variable
> full.model <- lm(Time ~ .,data = bridge)
> summary(full.model)
```

```
##
## Call:
## lm(formula = Time ~ ., data = 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 ***
## DArea       -0.04564    0.12675  -0.360 0.720705
## CCost        0.19609    0.14445   1.358 0.182426
## Dwgs         0.85879    0.22362   3.840 0.000440 ***
## Length      -0.03844    0.15487  -0.248 0.805296
## 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
```

5th. Variable Selection(Backward Selection)

```
> #Using stepwise selection with BIC
> stepwiseSelection <- stepAIC(full.model,direction = "both",
+                               trace = FALSE,k = log(NROW(bridge)))
> summary(stepwiseSelection)
```

```
##
## Call:
## lm(formula = Time ~ Dwgs + Spans, data = bridge)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.68649 -0.24728 -0.05988  0.26050  0.63759
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.66173    0.26871   9.905 1.49e-12 ***
## Dwgs         1.04163    0.15420   6.755 3.26e-08 ***
## Spans        0.28530    0.09095   3.137  0.00312 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3105 on 42 degrees of freedom
## Multiple R-squared:  0.7642, Adjusted R-squared:  0.753
## F-statistic: 68.08 on 2 and 42 DF,  p-value: 6.632e-14
```

6th. Interpretation

$$log(\hat{Time}) = 2.66 + 1.04 * log(Dwgs) + 0.29 * log(Spans)$$

1.Keep other covariates unchanged, the log(Time) is expected to increase by 1.04% with every unit increase of log(Dwgs).

2.Keep other covariates unchanged, the log(Time) is expected to increase by 0.29% with every unit increase of log(Spans).