# Mini-project R script

Ruiming Nie

2023-02-18

## Mini-project

### Loading and check data

```
#clean the workspace and set working directory
rm(list=ls())
dev.off()
setwd("/Users/ruimingnie/Desktop/R/data")
require(readxl)
```

```
## Loading required package: readxl
```

```
dog<-read_excel("Data.xlsx")
#check the data and find the NA value
sum(is.na(dog))
str(dog)
names(dog)
head(dog)
mean(dog$`Total (Million)`)
sd(dog$`Total (Million)`)
var(dog$`Total (Million)`)
```

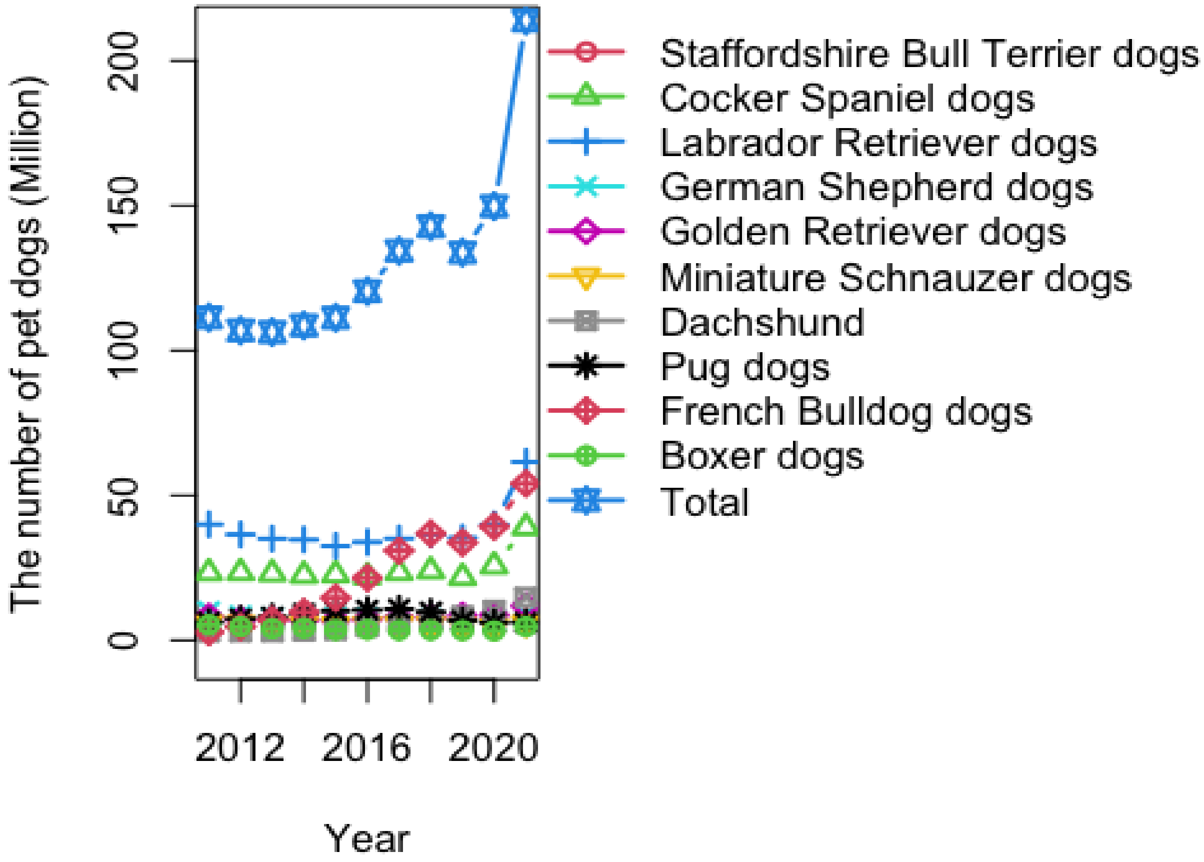
```
## null device
##           1
## [1] 0
## tibble [11 × 27] (S3: tbl_df/tbl/data.frame)
##  $ Year                              : num [1:11] 2011 2012 2013 2014 2015
...
##  $ Staffordshire Bull Terrier dogs: num [1:11] 7.11 6.24 5.77 4.94 4.56
...
##  $ Cocker Spaniel dogs             : num [1:11] 23.3 23.3 22.9 22.4 22.6
...
##  $ Labrador Retriever dogs         : num [1:11] 40 36.5 35 34.7 32.5 ...
##  $ German Shepherd dogs            : num [1:11] 9.89 8.5 7.95 7.93 7.78 ...
##  $ Golden Retriever dogs           : num [1:11] 8.08 7.08 7.12 6.98 6.93
...
##  $ Miniature Schnauzer dogs        : num [1:11] 5.92 5.8 5.58 5.48 5.3 ...
##  $ Dachshund                       : num [1:11] 2.86 2.85 2.87 3.13 3.45
...
##  $ Pug dogs                        : num [1:11] 6.22 7.36 8.07 9.24 10.09
...
##  $ French Bulldog dogs             : num [1:11] 2.77 4.65 6.99 9.67 14.61
```

```
...
##  $ Boxer dogs                      : num [1:11] 5.28 4.62 4 4.15 3.48 ...
##  $ Total                           : num [1:11] 111 107 106 109 111 ...
##  $ Total (Million)                 : num [1:11] 111 107 106 109 111 120 134
143 134 150 ...
##  $ GDP                             : num [1:11] 29961 30195 30552 31290
31786 ...
##  $ Annual earnings in 1000         : num [1:11] 26.1 26.5 27 27.2 27.6 ...
##  $ Annual earnings                 : num [1:11] 26095 26472 27011 27215
27615 ...
##  $ Annual expenditure on pets      : num [1:11] 4686000 4583000 4924000
5696000 6195000 ...
##  $ Cost                            : num [1:11] 42.2 42.8 46.5 52.3 55.8
...
##  $ Without                         : num [1:11] 2909 3039 3000 3007 3031
...
##  $ 65+                             : num [1:11] 16.6 16.9 17.3 17.6 17.8
...
##  $ Depression                      : num [1:11] 15 19 31 31 32 32 34 38 38
39 ...
##  $ Single child                    : num [1:11] 3549 3717 3676 3631 3598
...
##  $ Two children                    : num [1:11] 3042 3045 3105 3150 3186
...
##  $ Three or more children          : num [1:11] 1156 1134 1134 1155 1177
...
##  $ Family with child               : num [1:11] 4198 4179 4239 4305 4363
...
##  $ Child                           : num [1:11] 4.2 4.18 4.24 4.3 4.36 ...
##  $ Education Index                 : num [1:11] 0.872 0.866 0.912 0.92
0.911 0.911 0.913 0.918 0.928 0.924 ...
##  [1] "Year"                             "Staffordshire Bull Terrier dogs"
##  [3] "Cocker Spaniel dogs "             "Labrador Retriever dogs"
##  [5] "German Shepherd dogs "            "Golden Retriever dogs "
##  [7] "Miniature Schnauzer dogs"         "Dachshund"
##  [9] "Pug dogs "                        "French Bulldog dogs"
## [11] "Boxer dogs"                       "Total"
## [13] "Total (Million)"                  "GDP"
## [15] "Annual earnings in 1000"          "Annual earnings"
## [17] "Annual expenditure on pets "      "Cost"
## [19] "Without"                          "65+"
## [21] "Depression"                       "Single child"
## [23] "Two children"                     "Three or more children"
## [25] "Family with child"                "Child"
## [27] "Education Index"
## # A tibble: 6 × 27
##    Year Staffo…¹ Cocke…² Labra…³ Germa…⁴ Golde…⁵ Minia…⁶ Dachs…⁷ Pug d…⁸
Frenc…⁹
##   <dbl>    <dbl>   <dbl>   <dbl>   <dbl>   <dbl>   <dbl>   <dbl>   <dbl>
<dbl>
```

```
## 1  2011     7.11    23.3    40.0    9.89    8.08    5.92    2.86    6.22
2.77
## 2  2012     6.24    23.3    36.5    8.50    7.08    5.80    2.85    7.36
4.65
## 3  2013     5.77    22.9    35.0    7.95    7.12    5.58    2.87    8.07
6.99
## 4  2014     4.94    22.4    34.7    7.93    6.98    5.48    3.13    9.24
9.67
## 5  2015     4.56    22.6    32.5    7.78    6.93    5.30    3.45   10.1
14.6
## 6  2016     4.21    21.9    33.9    7.75    7.23    5.44    4.58   10.4
21.5
## # … with 17 more variables: `Boxer dogs` <dbl>, Total <dbl>,
## #   `Total (Million)` <dbl>, GDP <dbl>, `Annual earnings in 1000` <dbl>,
## #   `Annual earnings` <dbl>, `Annual expenditure on pets ` <dbl>, Cost
<dbl>,
## #   Without <dbl>, `65+` <dbl>, Depression <dbl>, `Single child` <dbl>,
## #   `Two children` <dbl>, `Three or more children` <dbl>,
## #   `Family with child` <dbl>, Child <dbl>, `Education Index` <dbl>, and
## #   abbreviated variable names ¹`Staffordshire Bull Terrier dogs`, …
## [1] 130.8182
## [1] 31.65065
## [1] 1001.764
```

### The time series graph about each pet dog and total

```
##the time series graph
#plot the trends for each dog type and total pet dog number
par(mar=c(4,4,1,14),mfrow=c(1,1))
matplot(x=dog[1], y=dog[2:12],  type = "b",pch=1:11,lwd = 2, lty =1,col
=2:12, ylim = c(-5,210),
        xlab = "Year", ylab = "The number of pet dogs (Million)")
#Add legend
legend(par('usr')[2], par('usr')[4], xpd=NA,
       legend = colnames(dog)[2:12], pch = 1:11, col=2:12,
       lty =1,lwd = 2, bty = "n")
```



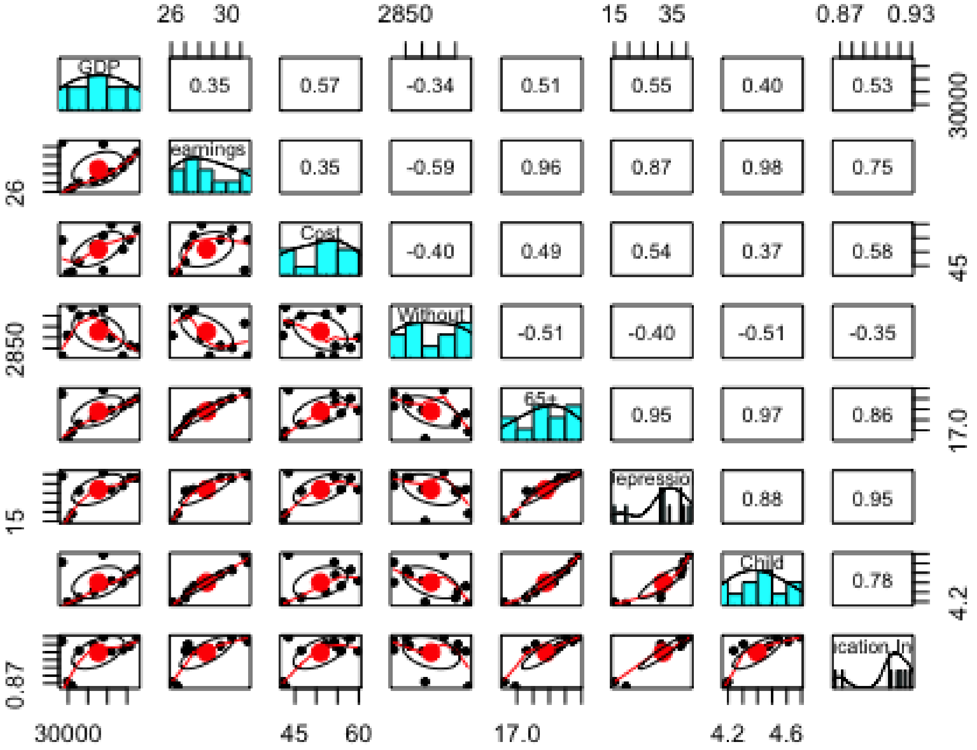
## Factors affect the pet dogs number

### 1. Colinearity

```
#The factors that affect the pet dogs number
require(usdm)

## Loading required package: usdm

## Loading required package: sp

## Loading required package: raster

require(psych)

## Loading required package: psych

##
## Attaching package: 'psych'

## The following object is masked from 'package:raster':
##
##     distance

require(lmerTest)
```

```
## Loading required package: lmerTest

## Loading required package: lme4

## Loading required package: Matrix

##
## Attaching package: 'lme4'

## The following object is masked from 'package:raster':
##
##     getData

##
## Attaching package: 'lmerTest'

## The following object is masked from 'package:lme4':
##
##     lmer

## The following object is masked from 'package:stats':
##
##     step

require(sjPlot)

## Loading required package: sjPlot

require(factoextra)

## Loading required package: factoextra

## Loading required package: ggplot2

##
## Attaching package: 'ggplot2'

## The following objects are masked from 'package:psych':
##
##     %+%, alpha

## Welcome! Want to learn more? See two factoextra-related books at
https://goo.gl/ve3WBa

require(ggpubr)

## Loading required package: ggpubr

##
## Attaching package: 'ggpubr'

## The following object is masked from 'package:raster':
##
##     rotate
```

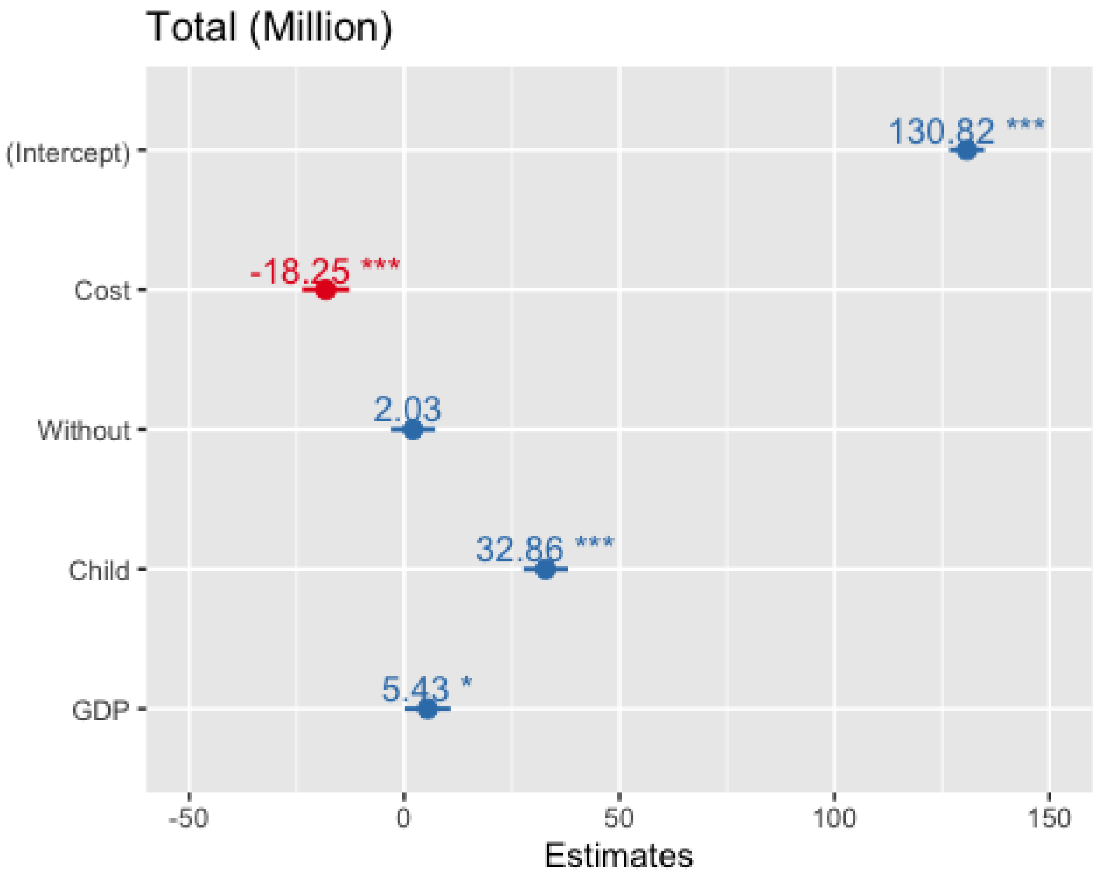
```r
#check the relations between factors
pairs.panels(dog[,c(14,15,18,19,20,21,26,27)])
```



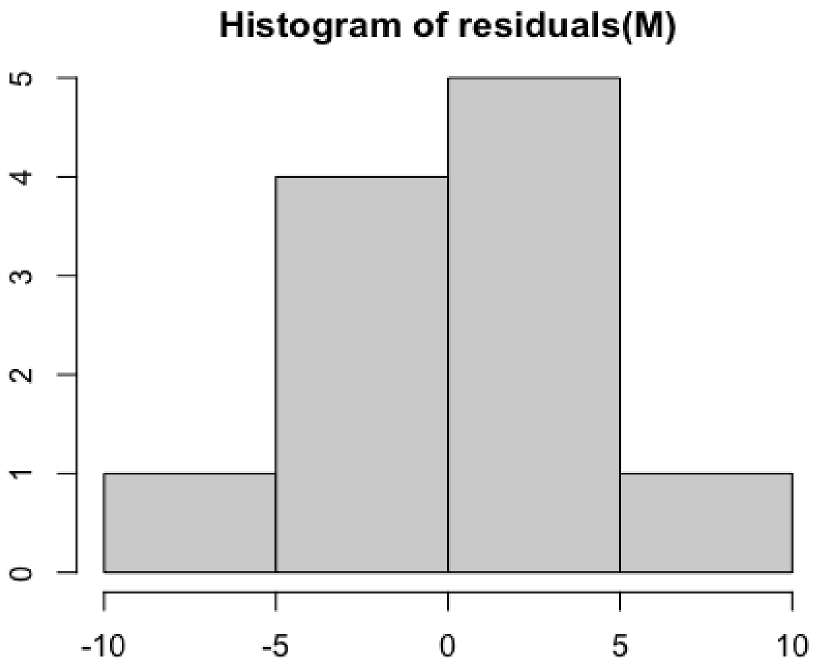
```r
#convert the dataset type to data.frame before using VIF function
dog1<- as.data.frame(dog)
##remove the collinearity by using VIF (threshold =3)
vif(dog1[,c(14,15,18,19,20,21,26,27)])

##                 Variables        VIF
## 1                     GDP   2.739377
## 2 Annual earnings in 1000  57.024364
## 3                    Cost   2.476824
## 4                 Without   2.905004
## 5                     65+ 135.228951
## 6              Depression  71.494703
## 7                   Child  54.172501
## 8         Education Index  20.668093

#remove 65+
vif(dog1[,c(14,15,18,19,21,26,27)])

##                 Variables       VIF
## 1                     GDP  2.276350
## 2 Annual earnings in 1000 50.754779
## 3                    Cost  2.087685
## 4                 Without  2.876099
## 5              Depression 35.793356
## 6                   Child 31.873437
## 7         Education Index 16.810946

#remove earning
vif(dog1[,c(14,18,19,21,26,27)])

##         Variables       VIF
## 1             GDP  1.770335
## 2            Cost  2.003664
## 3         Without  1.610329
## 4      Depression 21.559528
## 5           Child  6.245731
## 6 Education Index 12.655907

#remove  depression
vif(dog1[,c(14,18,19,26,27)])

##         Variables      VIF
## 1             GDP 1.650136
## 2            Cost 2.002719
## 3         Without 1.590022
## 4           Child 3.459040
## 5 Education Index 3.900411

#remove education index
vif(dog1[,c(14,18,19,26)]) #all values are smaller than 3

##annual cost per dog, number of family without children, number of family
with children, GDP per capita are final variables

##   Variables      VIF
## 1       GDP 1.588898
## 2      Cost 1.614216
## 3   Without 1.459108
## 4     Child 1.480460
```

### 2. Multiple regression model

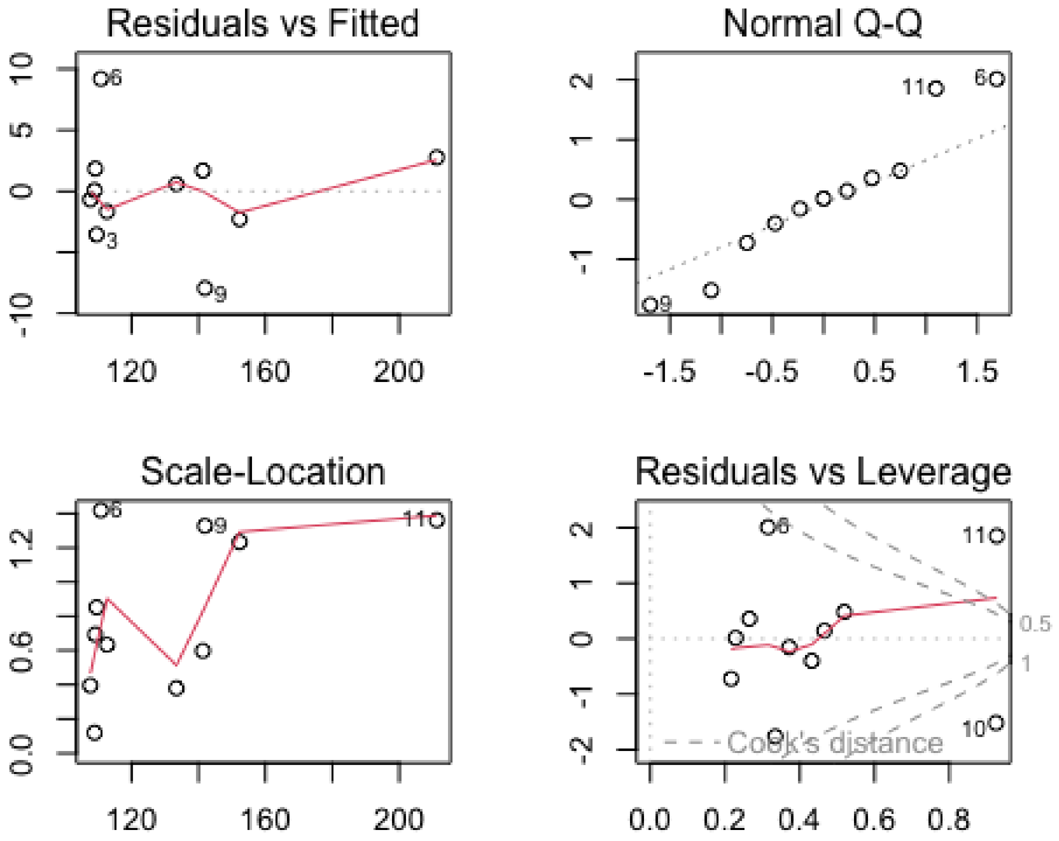
```
#Linear model-- scale() make the units simple
##Multiple continuous explanatory variables on different scales, scale()
function to z-standardize them
M<- lm(`Total (Million)`~scale(Cost)+scale(Without)+
     scale(Child)+ scale(GDP), data = dog1)

#Model interpretion
plot_model(M, show.values = TRUE, show.intercept = TRUE)
```



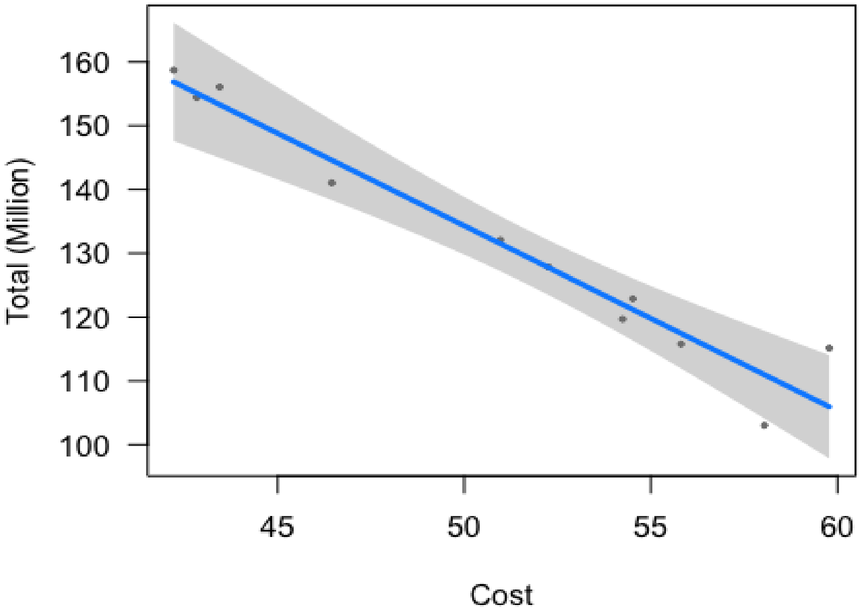
```
summary(M)
#Summary table
library(parameters)
model_parameters(M, summary = TRUE)

##
## Uncertainty intervals (equal-tailed) and p-values (two-tailed) computed
##   using a Wald t-distribution approximation.

##
## Call:
## lm(formula = `Total (Million)` ~ scale(Cost) + scale(Without) +
##     scale(Child) + scale(GDP), data = dog1)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -7.942 -1.982  0.069  1.767  9.193
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)     130.818      1.665  78.578 2.86e-10 ***
## scale(Cost)     -18.248      2.218  -8.225 0.000174 ***
## scale(Without)    2.030      2.109   0.962 0.373050
## scale(Child)     32.855      2.125  15.465 4.62e-06 ***
## scale(GDP)        5.431      2.201   2.468 0.048617 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.522 on 6 degrees of freedom
## Multiple R-squared:  0.9817, Adjusted R-squared:  0.9696
## F-statistic: 80.64 on 4 and 6 DF,  p-value: 2.402e-05
##
## Parameter   | Coefficient |   SE |           95% CI |  t(6) |      p
## ----------------------------------------------------------------------
## (Intercept) |      130.82 | 1.66 | [126.74, 134.89] | 78.58 | < .001
## Cost        |      -18.25 | 2.22 | [-23.68, -12.82] | -8.23 | < .001
## Without     |        2.03 | 2.11 | [ -3.13,   7.19] |  0.96 | 0.373
## Child       |       32.86 | 2.12 | [ 27.66,  38.05] | 15.46 | < .001
## GDP         |        5.43 | 2.20 | [  0.05,  10.82] |  2.47 | 0.049
##
## Model: `Total (Million)` ~ scale(Cost) + scale(Without) + scale(Child) +
scale(GDP) (11 Observations)
## Residual standard deviation: 5.522 (df = 6)
## R2: 0.982; adjusted R2: 0.970

#Model validation
#plot  residuals distribution
par(mfrow=c(1,1), mar=c(3,3,2,2))
hist(residuals(M))
```

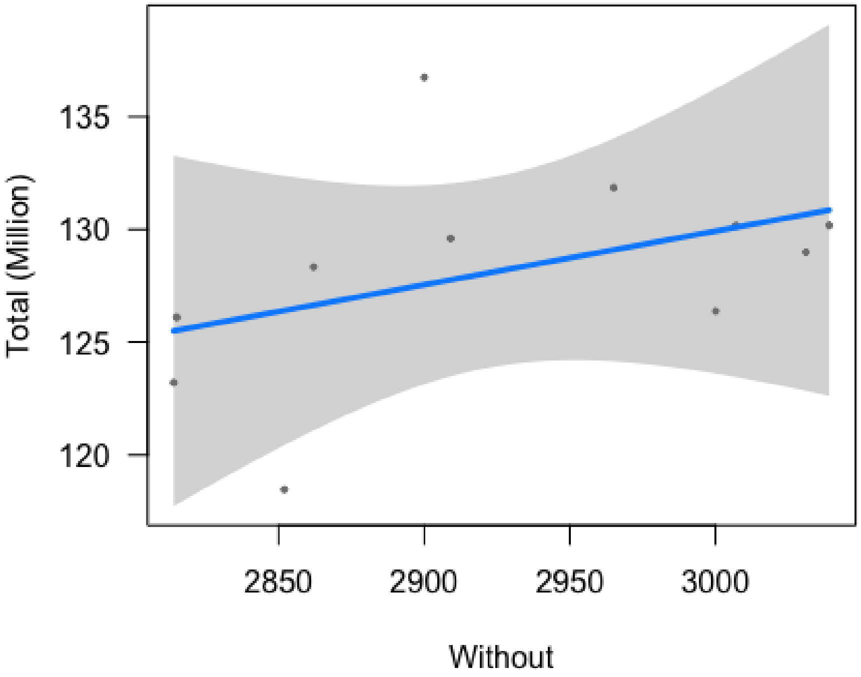


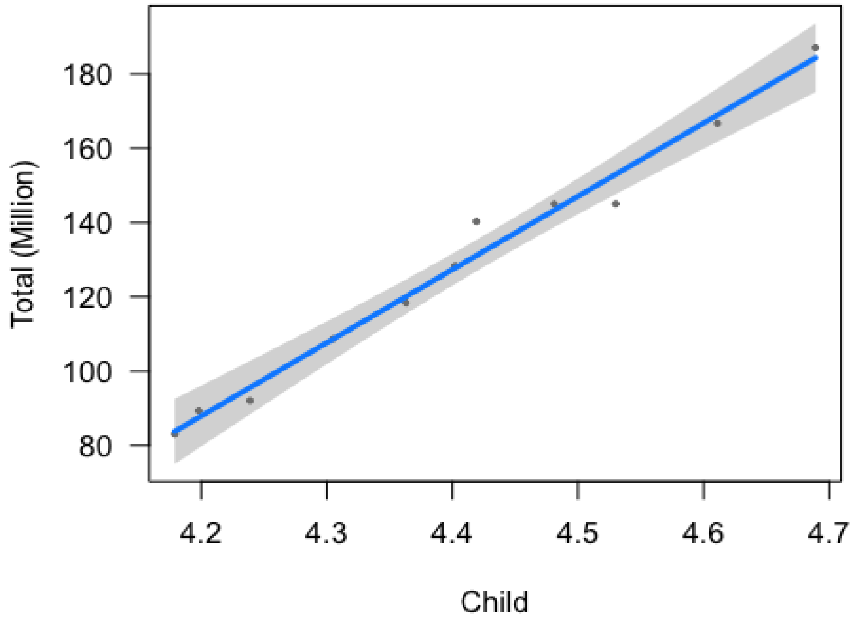
```r
#Model diagnostics
par(mfrow=c(2,2), mar=c(3,3,2,2))
plot(M) #no assumptions were violated
```

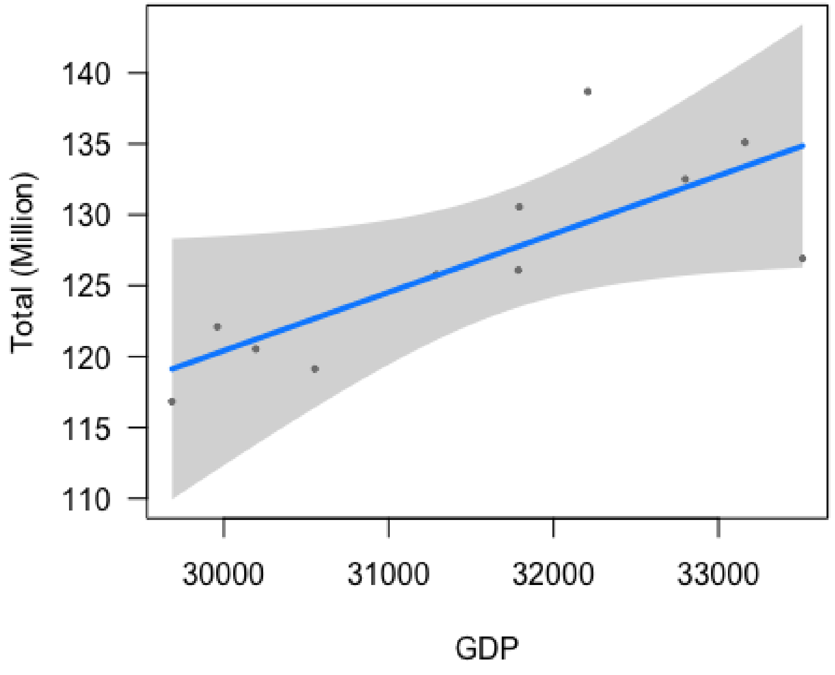


```r
#multiple regression model visulisation by using visreg
library(visreg)
par(mfrow=c(1,1), mar=c(4,4,2,2))
visreg(M)
```


Total (Million)
160
150
140
130
120
110
100
45
50
55
60
Cost


Total (Million)
135
130
125
120
2850
2900
2950
3000
Without

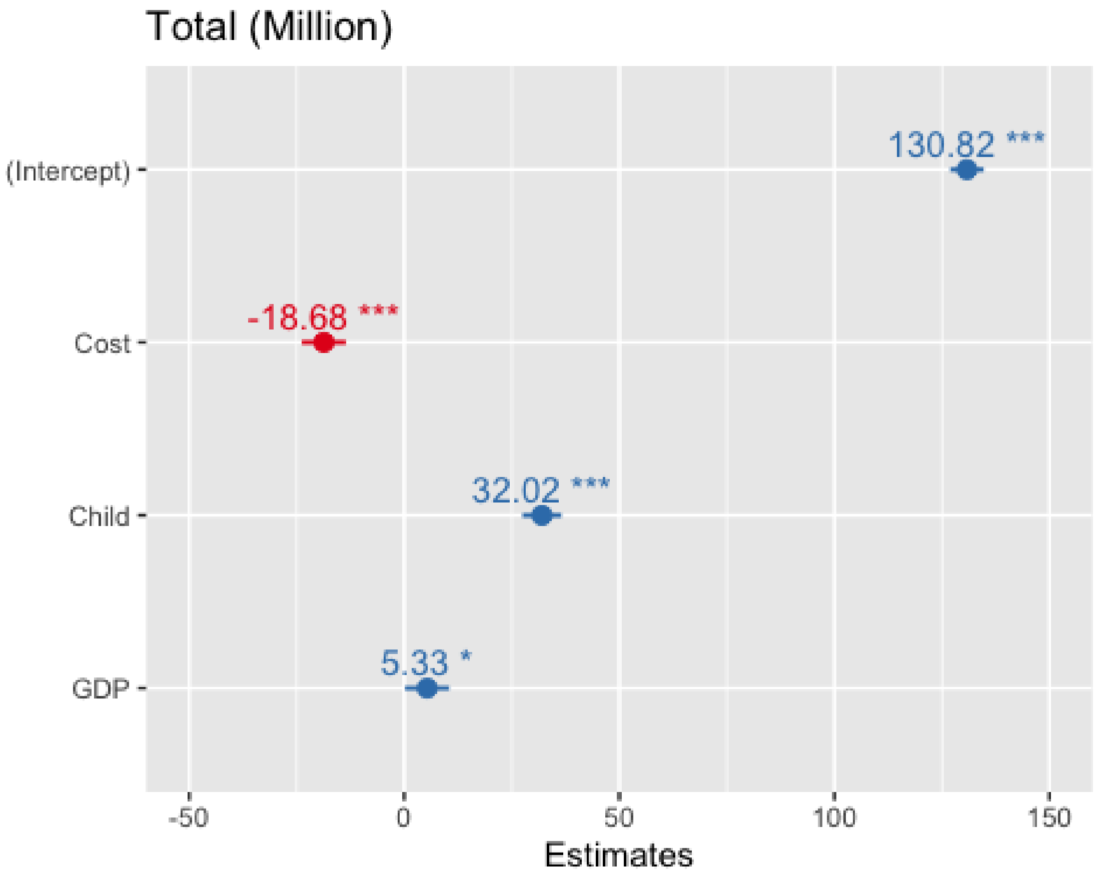




### 3. Model selection: Information Criteria (AIC)

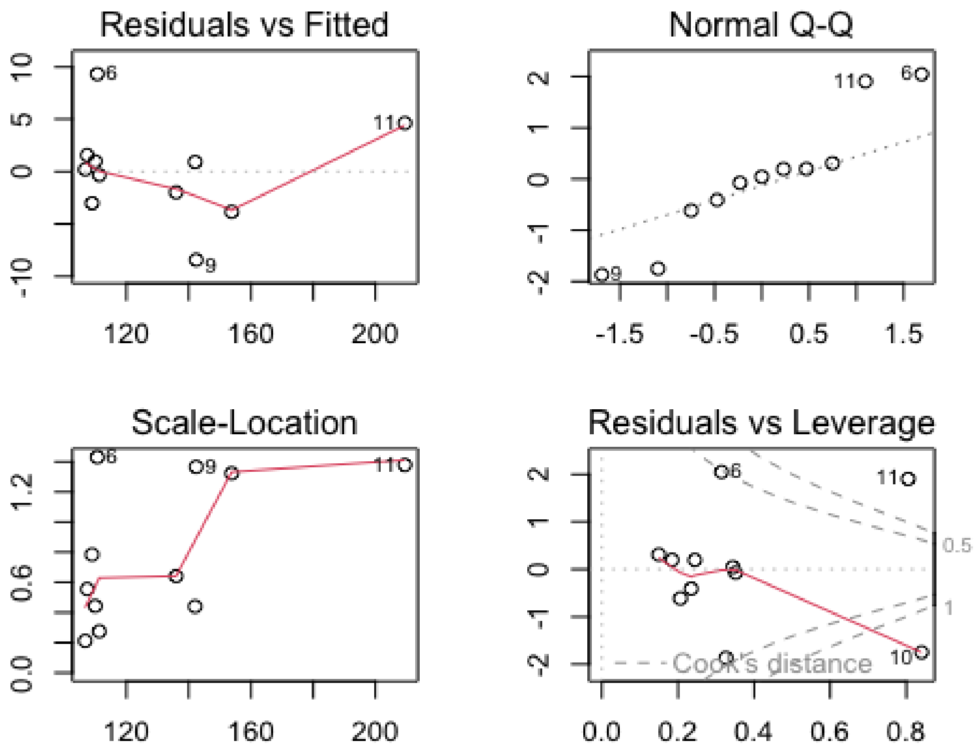
```
# Model selection--- AIC
## scope here is indicating the lower (null model) and upper (maximal model)
M1<-step(M, direction = "backward", scope = list(lower=~1,
upper=~scale(Cost)+scale(Without)+scale(Child)+ scale(GDP)))

## Start:  AIC=40.92
## `Total (Million)` ~ scale(Cost) + scale(Without) + scale(Child) +
##     scale(GDP)
## 
##                   Df Sum of Sq    RSS    AIC
## - scale(Without)  1      28.2  211.2 40.502
## <none>                         182.9 40.923
## - scale(GDP)      1     185.6  368.6 46.629
## - scale(Cost)     1    2062.8 2245.7 66.508
## - scale(Child)    1    7291.3 7474.3 79.735
## 
## Step:  AIC=40.5
## `Total (Million)` ~ scale(Cost) + scale(Child) + scale(GDP)
## 
##                 Df Sum of Sq    RSS    AIC
## <none>                       211.2 40.502
## - scale(GDP)    1     179.0  390.1 45.255
## - scale(Cost)   1    2256.5 2467.6 65.544
## - scale(Child)  1    8322.5 8533.7 79.193
```

```
M2<-lm(`Total (Million)`~scale(Cost)+scale(Child)+ scale(GDP), data = dog1)

summary(M2)
#plot the model
plot_model(M2, show.values = TRUE, show.intercept = TRUE)
```



```
##
## Call:
## lm(formula = `Total (Million)` ~ scale(Cost) + scale(Child) +
##     scale(GDP), data = dog1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.4266 -2.4949  0.2006  1.2690  9.3093
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   130.818      1.656  78.996 1.37e-11 ***
## scale(Cost)   -18.683      2.160  -8.649 5.52e-05 ***
## scale(Child)   32.017      1.928  16.610 7.00e-07 ***
## scale(GDP)      5.326      2.187   2.436    0.045 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.492 on 7 degrees of freedom
## Multiple R-squared:  0.9789, Adjusted R-squared:  0.9699
## F-statistic: 108.4 on 3 and 7 DF,  p-value: 3.14e-06
```

```
#model validation
par(mfrow=c(2,2), mar=c(3,3,2,2))
plot(M2)
```



```
#compare two models
anova(M2,M)

## Analysis of Variance Table
##
## Model 1: `Total (Million)` ~ scale(Cost) + scale(Child) + scale(GDP)
## Model 2: `Total (Million)` ~ scale(Cost) + scale(Without) + scale(Child) +
##     scale(GDP)
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1      7 211.16
## 2      6 182.93  1    28.234 0.9261  0.373

## Why do not choose the selected model -- (1)lost a vital coefficient
(household without children), which will affect my final discussion.
(2)although it doesn't make the model worse, it doesn't make it much better
either (AIC difference is quite small)

#check AIC
AIC(M)-AIC(M2)

## [1] 0.421142
```