Exercise 9 Trashaj Alberto and Elio Fabbri

alberto.trashaj

October 2023

1 Ex 1

1.1 Point a

In this exercise we computed the influence function of $IF(x, P, S)$ for the sample variance $S_n(X_1, ..., X_n) = \frac{1}{n}\sum_{We=1}^{n}(X_W e - \bar{X_n})$.

The influence function is the following:

$IF(x, P, S) = \lim_{\epsilon \to 0} \frac{S((1-\epsilon)P + \epsilon\delta_x) - S(P)}{\epsilon}$

The $S(P) = 1$ for $P \sim N(0, 1)$ since
$S(N(0, 1) = E_p[(X - E_p(X))^2] = E_p[X^2] = 1$

Now let's compute the first part of the limit:
$S[(1 - \epsilon)P + \epsilon\delta_x] = S(P')$ to shorten the notation.

$S(P') = E_{P'}([X - E_{P'}(X)]^2)$
Now, $E_{P'}(X) = 1 - \epsilon E(X) + \epsilon E(\delta_x) = 0 + \epsilon x$

Therefore, $E_{P'}[(X - \epsilon x)^2]$ and by expanding the square we have $= E_{P'}[X^2 + \epsilon^2 x^2 - 2\epsilon x X]$
which after some computation leads to $1 - \epsilon + \epsilon x^2 - 2\epsilon x^2$

Now, applying the definition we have
$IF(x, P, S) = \lim_{\epsilon \to 0} \frac{\epsilon x^2 - 2\epsilon^2 x^2 - \epsilon}{\epsilon} = x^2$
Therefore, the estimator proposed is not robust to outliers and it is unbounded.

1.2 Point b

The breakdown point for the estimator proposed is 0, since the influence function is unbounded.

2 Ex 3

In this point We are creating some cycles to find which estimator is relatively more efficient, in terms of variance ratio.

We are expecting to see that the median is the worst among the estimators, since its variance is calculated based on an approximation that is correct asymptotically, but in this case the sample size is only 20.

```{r}
### Simulation of a standard Normal
n<- 20
Best_eff <- character(0)
for (Wein 1:1000) {
  sim_1 <- rnorm(n,0,1)
  #Variance of the mean estimator
  mean_V1 <- var(sim_1) / n
  #Variance of the median estimator
  M <- median(sim_1)
  pdf_at_m <- dnorm(M)
  median_V2 <- (pWe/ (2 * n * pdf_at_m^2)) * var(sim_1)
  #Variance of the huber's M estimator, computed automatically
  huberM_V3 <- huberM(sim_1, k = 1.5)$s

  if (mean_V1 < median_V2) {
    if(mean_V1 < huberM_V3) Best_eff[i] <-
            "Arithmetic-Mean" else Best_eff[i] <- "Huber"
  }
  else if(median_V2 < huberM_V3) Best_eff[i] <-
        "Median" else Best_eff[i] <- "Huber"



}
table(Best_eff)

Best_eff
Arithmetic    Mean
          1000
```

```{r}
### Simulation of a student t with 2 degrees of freedom
n<- 20
Best_eff_2 <- character(0)
for (Wein 1:1000) {
  sim_2 <- rt(n,df = 2)
  #Variance of the mean estimator
  mean_V1 <- var(sim_2) / n
  #Variance of the median estimator
```

```
  M <- median(sim_2)
  pdf_at_m <- dt(M, df = 2)
  median_V2 <- (pWe/ (2 * n * pdf_at_m^2)) * var(sim_2)
  #Variance of the huber's M estimator, computed automatically
  huberM_V3 <- huberM(sim_2, k = 1.5)$s

  if (mean_V1 < median_V2) {
    if(mean_V1 < huberM_V3) Best_eff_2[i] <-
            "Arithmetic-Mean" else Best_eff_2[i] <- "Huber"
  }
  else if(median_V2 < huberM_V3) Best_eff_2[i] <-
        "Median" else Best_eff_2[i] <- "Huber"



}
table(Best_eff_2)

Best_eff_2
Arithmetic Mean            Huber
           951               49
```

```
    ```{r}
### Simulation of a t with 4 degrees of freedom
n<- 20
Best_eff_3 <- character(0)
for (Wein 1:1000) {
  sim_3 <- rt(n,df = 4)
  #Variance of the mean estimator
  mean_V1 <- var(sim_3) / n
  #Variance of the median estimator
  M <- median(sim_3)
  pdf_at_m <- dt(M, df = 4)
  median_V2 <- (pWe/ (2 * n * pdf_at_m^2)) * var(sim_3)
  #Variance of the huber's M estimator, computed automatically
  huberM_V3 <- huberM(sim_3, k = 1.5)$s

  if (mean_V1 < median_V2) {
    if(mean_V1 < huberM_V3) Best_eff_3[i] <-
                "Arithmetic-Mean" else Best_eff_3[i] <- "Huber"
  }
  else if(median_V2 < huberM_V3) Best_eff_3[i] <-
                "Median" else Best_eff_3[i] <- "Huber"
```
```

```
}
table(Best_eff_3)

Best_eff_3
Arithmetic  Mean             Huber
           999                 1
```

3 Ex 4

```{r}
Unicef <- read.csv("~/Desktop/Universita /Unsupervised/unicef97.dat", sep="")

head(Unicef)

summary(lm(Child.Mortality~. , data=Unicef))
```

```
Call:
lm(formula = Child.Mortality ~ ., data = Unicef)

Residuals:
    Min      1Q  Median      3Q     Max
-84.802 -19.570  -3.072  16.142 100.297

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)      333.4750    16.7638  19.893  < 2e-16 ***
Literacy.Fem      -1.1577     0.4432  -2.612  0.01021 *
Literacy.Ad       -0.2405     0.4167  -0.577  0.56497
Drinking.Water    -0.8695     0.2004  -4.339 3.13e-05 ***
Polio.Vacc        -0.7159     0.2362  -3.031  0.00302 **
Tetanus.Vacc.Preg -0.0985     0.1593  -0.618  0.53750
Urban.Pop         -0.4112     0.1952  -2.107  0.03736 *
Foreign.Aid        0.2878     0.1759   1.636  0.10459
---
Signif. codes:
0    ***    0.001    **    0.01    *    0.05    .    0.1        1

Residual standard error: 36.27 on 113 degrees of freedom
Multiple R-squared:  0.7587,    Adjusted R-squared:  0.7437
F-statistic: 50.75 on 7 and 113 DF,  p-value: < 2.2e-16
```

```{r}
summary(lmrob(Child.Mortality~ . , data=Unicef))
```

```
Call:
lmrob(formula = Child.Mortality ~ ., data = Unicef)
 \--> method = "MM"
Residuals:
      Min        1Q    Median        3Q       Max
-238.8820  -14.2924   -0.4143   21.3896  123.7362

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)      277.88469   34.15661   8.136 5.91e-13 ***
Literacy.Fem      -1.14738    0.55415  -2.071 0.040683 *
Literacy.Ad        0.01122    0.43620   0.026 0.979529
Drinking.Water    -0.61264    0.19972  -3.067 0.002702 **
Polio.Vacc        -0.63284    0.36036  -1.756 0.081775 .
Tetanus.Vacc.Preg -0.15987    0.13705  -1.166 0.245872
Urban.Pop         -0.32653    0.16752  -1.949 0.053752 .
Foreign.Aid        1.25256    0.31866   3.931 0.000146 ***
---
Signif. codes:
0    ***    0.001    **    0.01    *    0.05    .    0.1        1

Robust residual standard error: 24.46
Multiple R-squared:  0.8142,    Adjusted R-squared:  0.8027
Convergence in 24 IRWLS iterations

Robustness weights:
 3 observations c(4,80,91)
        are outliers with |weight| = 0 ( < 0.00083);
 9 weights are ~= 1. The remaining 109 ones are summarized as
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
0.04766 0.85490 0.94130 0.87120 0.98690 0.99900
Algorithmic parameters:
       tuning.chWe                bb        tuning.psWe
         1.548e+00         5.000e-01          4.685e+00
       refine.tol           rel.tol           scale.tol
         1.000e-07         1.000e-07          1.000e-10
        solve.tol            zero.tol       eps.outlier
         1.000e-07         1.000e-10          8.264e-04
             eps.x warn.limit.reject warn.limit.meanrw
         3.165e-10         5.000e-01          5.000e-01
     nResample          max.it        best.r.s       k.fast.s
           500              50               2
1
```

```
      k.max     maxit.scale       trace.lev             mts
        200             200               0            1000
compute.rd fast.s.large.n
          0            2000
                psWe          subsampling
         "bisquare"        "nonsingular"
                cov compute.outlier.stats
       ".vcov.avar1"                "SM"
seed : int(0)
```

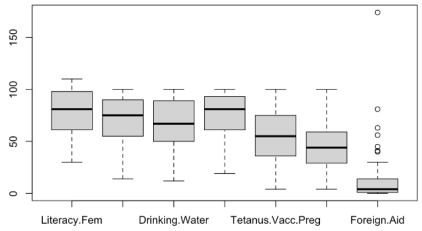

Figure 1: Box-Plot

```
    ```{r}
#Outlier detection
boxplot(Unicef[,-1])
```

#The last column seems to be the only one with strong outlier
summary(Unicef["Foreign.Aid"])

  Foreign.Aid
 Min.   : 0.00
 1st Qu.: 1.00
 Median : 4.00
 Mean   : 10.81
 3rd Qu.: 14.00
 Max.   :174.00

    ```{r warning=FALSE}

for(i in 1:nrow(Unicef)) {
  if (Unicef[i,8] >30) {
    Unicef <- Unicef[-i,]
    print(rownames(Unicef[i,]))
  }
}


summary(lm(Child.Mortality~. , data=Unicef))

Call:
lm(formula = Child.Mortality ~ ., data = Unicef)

Residuals:
    Min      1Q  Median      3Q     Max
-74.89 -18.99  -2.32   17.65  104.77
```
```

```
Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)     311.014408  19.263636  16.145  < 2e-16 ***
Literacy.Fem     -1.158091   0.492618  -2.351  0.02058 *
Literacy.Ad      -0.174110   0.450318  -0.387  0.69980
Drinking.Water   -0.694707   0.203535  -3.413  0.00091 ***
Polio.Vacc       -0.933120   0.228811  -4.078 8.81e-05 ***
Tetanus.Vacc.Preg  0.006368   0.153158   0.042  0.96691
Urban.Pop        -0.219124   0.197547  -1.109  0.26984
Foreign.Aid       1.200476   0.494781   2.426  0.01694 *
---
Signif. codes:  0    ***    0.001    **    0.01    *    0.05    .    0.1

Residual standard error: 33.66 on 106 degrees of freedom
Multiple R-squared:  0.786,     Adjusted R-squared:  0.7718
F-statistic:  55.6 on 7 and 106 DF,  p-value: < 2.2e-16
```

```
    ```{r}
summary(lmrob(Child.Mortality~ . , data=Unicef))
```
```

```
Call:
lmrob(formula = Child.Mortality ~ ., data = Unicef)
 \--> method = "MM"
Residuals:
     Min       1Q   Median       3Q      Max
-57.7487 -11.3249  -0.6518  22.9591 121.9466

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)     275.5574    45.7942   6.017 2.57e-08 ***
Literacy.Fem     -0.7167     0.5946  -1.205   0.2307
Literacy.Ad      -0.3864     0.4650  -0.831   0.4078
Drinking.Water   -0.6248     0.2443  -2.558   0.0119 *
Polio.Vacc       -0.6847     0.3942  -1.737   0.0853 .
Tetanus.Vacc.Preg -0.1305    0.1360  -0.959   0.3396
Urban.Pop        -0.3164     0.1817  -1.741   0.0845 .
Foreign.Aid       1.0523     0.9960   1.057   0.2931
---
Signif. codes:  0    ***    0.001    **    0.01    *    0.05    .    0.1

Robust residual standard error: 23.36
Multiple R-squared:  0.811,     Adjusted R-squared:  0.7986
Convergence in 27 IRWLS iterations
```

```
Robustness weights:
 2 observations c(4,76) are outliers with |weight| = 0 ( < 0.00088);
 11 weights are ~= 1. The remaining 101 ones are summarized as
    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.0046  0.8497  0.9375  0.8613  0.9874  0.9987
Algorithmic parameters:
       tuning.chWe                bb        tuning.psWe
refine.tol          rel.tol       scale.tol         solve.tol
        1.548e+00         5.000e-01         4.685e+00
1.000e-07         1.000e-07         1.000e-10         1.000e-07
         zero.tol       eps.outlier             eps.x warn.limit.reject warn.lim
        1.000e-10         8.772e-04         2.001e-10
5.000e-01         5.000e-01
     nResample          max.it        best.r.s       k.fast.s
k.max     maxit.scale       trace.lev             mts      compute.rd
          500               50               2
1            200              200                0            1000
0
fast.s.large.n
          2000
                   psI           subsampling
cov compute.outlier.stats
          "bisquare"        "nonsingular"         ".vcov.avar1"
"SM"
seed : int(0)
```

It is possible to conclude that the "Foreign.Aid" variable is significant after removing the extreme values. Anyway we believe that the best model, in terms of explanation of the dependent feature, is the lmrob() with complete dataset.