remuestreo

Christian Vásquez Velasco

10/7/2021

#data

```
datos <- readxl::read_xlsx("Cosecha arandano.xlsx", sheet = "Hoja1")
str(datos)
```

```
## tibble [6,745 x 10] (S3: tbl_df/tbl/data.frame)
##  $ Variedad             : chr [1:6745] "Biloxi" "Biloxi" "Biloxi" "Biloxi" ...
##  $ SemCal               : num [1:6745] 45 46 47 48 49 25 25 25 26 25 ...
##  $ KgCosechados         : num [1:6745] 71.8 114.8 145 66.9 96.2 ...
##  $ BayasiniciandoaCremas: num [1:6745] 0 0 0 0 29.6 ...
##  $ BayasCremas          : num [1:6745] 0 0 0 0 0 ...
##  $ BayasMaduras         : num [1:6745] 0 0 0 0 0 ...
##  $ BayasCosechables     : num [1:6745] 0 0 0 0 29.6 ...
##  $ Hectareas            : num [1:6745] 2.72 2.72 2.72 2.72 2.72 2 3.2 3.2 2.46 3.2 ...
##  $ Plantasprod          : num [1:6745] 8086 8086 8086 8086 8086 ...
##  $ Hcalor               : num [1:6745] 38 39 27 45 54 20 20 20 44 20 ...
```

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
datos <- mutate(datos, Kg.ha = KgCosechados/Hectareas)

datos <- datos %>%
  filter(!Kg.ha %in% "0")

datos <- datos[,c(7,10,11)]

datos <- na.omit(datos)
RNGkind(sample.kind = "Rounding")
```

```
## Warning in RNGkind(sample.kind = "Rounding"): non-uniform 'Rounding' sampler
## used
```

```
set.seed(100)
datos <- datos[sample(nrow(datos),100,replace = F),]
datos
```

```
## # A tibble: 100 x 3
##    BayasCosechables Hcalor Kg.ha
##               <dbl>  <dbl> <dbl>
##  1             77.9     27 550.
##  2             30.2      2  69.2
##  3             75.8      4 384.
##  4             25.5     20  28.4
##  5             94.5     10 332.
##  6             51.2     58 353.
##  7            118.      24 591.
##  8             91.0      2 444.
##  9            131.       8 568.
## 10             37.6      6  60.6
## # ... with 90 more rows
```

Análisis exploratorio y descriptivo

```
library(summarytools)
```

```
## Registered S3 method overwritten by 'pryr':
##   method      from
##   print.bytes Rcpp
```

```
summarytools::descr(datos)
```

```
## Descriptive Statistics
## datos
## N: 100
##
##                     BayasCosechables   Hcalor     Kg.ha
## ----------------- ------------------ -------- ---------
##              Mean              80.58    21.82    385.63
##           Std.Dev              48.22    21.60    255.97
##               Min               0.00     0.00     28.39
##                Q1              49.81     2.50    151.20
##            Median              77.94    13.50    351.22
##                Q3             106.21    38.00    558.94
##               Max             212.67    75.00   1126.82
##               MAD              41.82    17.05    303.79
##               IQR              56.14    35.25    397.59
##                CV               0.60     0.99      0.66
##          Skewness               0.56     0.78      0.60
```

```
##       SE.Skewness                0.24      0.24      0.24
##          Kurtosis                0.21     -0.53     -0.32
##           N.Valid              100.00    100.00    100.00
##         Pct.Valid              100.00    100.00    100.00
```

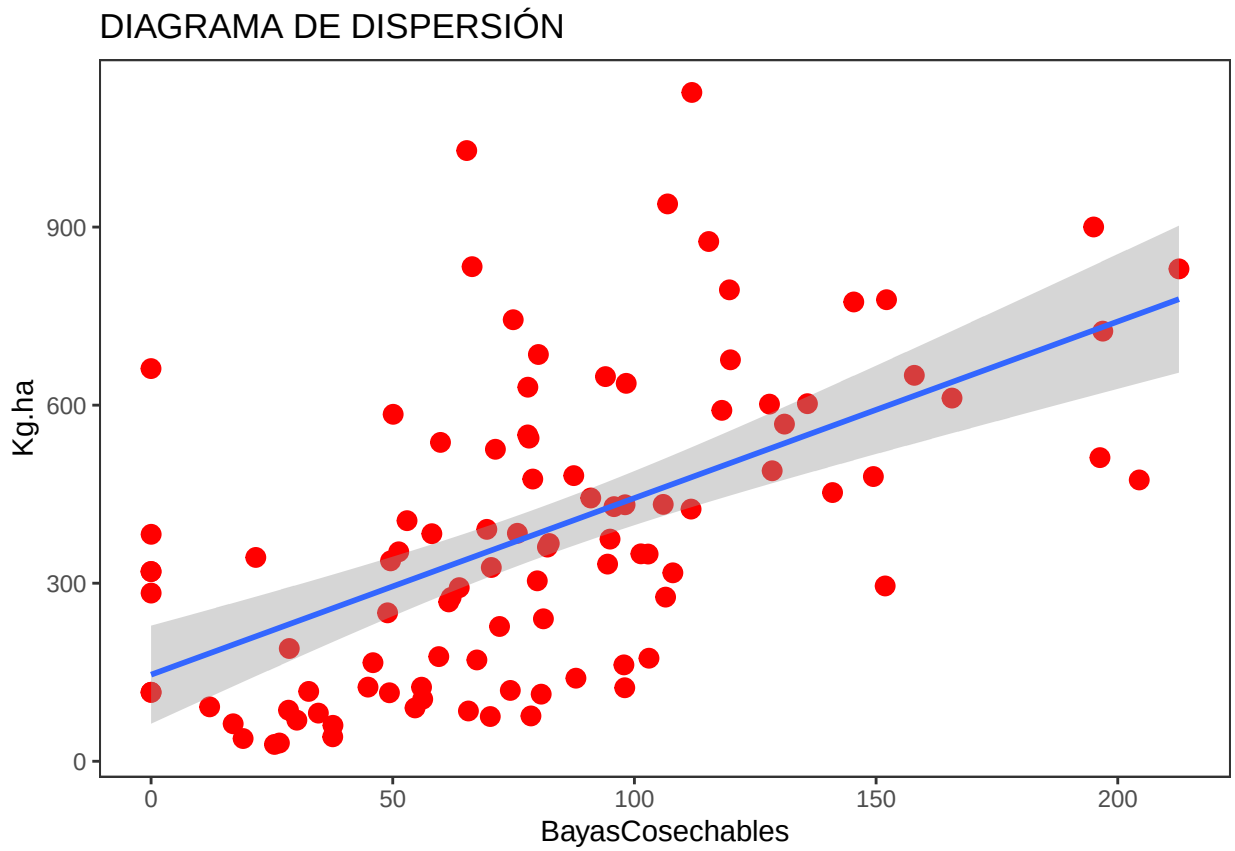
```
#Generando gráfico de dispersión del número de bayas cosechables vs los Kg Cosechados por hectárea
library(ggplot2)
```

```
## Warning in (function (kind = NULL, normal.kind = NULL, sample.kind = NULL) :
## non-uniform 'Rounding' sampler used
```

```
datos %>%
  ggplot(aes(x=BayasCosechables,y=Kg.ha))+
  geom_point(position = "jitter", size=3, colour="red")+
  labs( title = "DIAGRAMA DE DISPERSIÓN")+
  geom_smooth(method = "lm")+
  theme_test()
```

```
## `geom_smooth()` using formula 'y ~ x'

## Warning in (function (kind = NULL, normal.kind = NULL, sample.kind = NULL) :
## non-uniform 'Rounding' sampler used
```



En el siguiente gráfico de diagrama de dispersión, se aprecia que las observaciones del número de bayas cosechables vs los Kg Cosechados por hectárea son muy dispersos asi como también se observa que tenemos

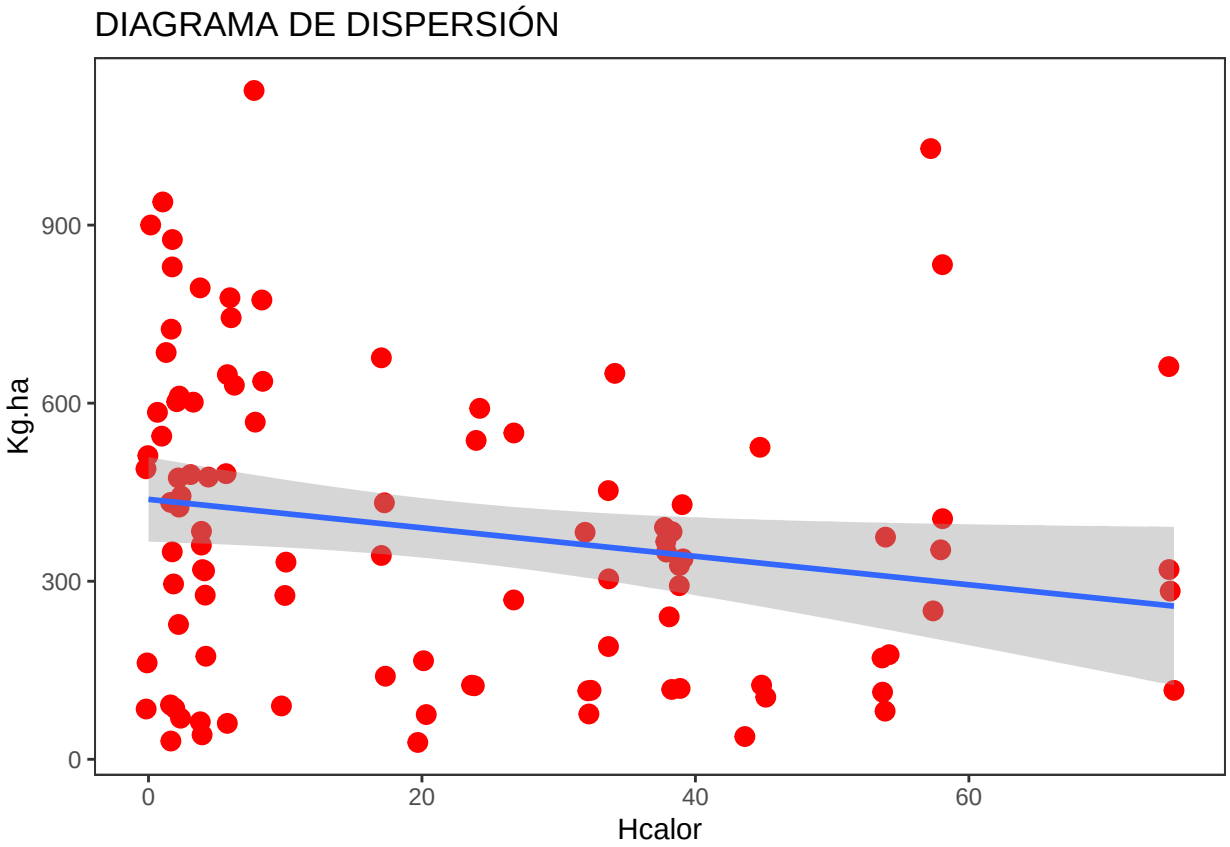
valores atipicos, en el número de BayasCosechales aproximadamente igual a 60 se tienen más de 1000 Kg cocechados por hectárea. Pero estos casos con menor número de bayas y una mayor productividad se pueden deber a que el peso promedio de los frutos en esas válvulas (observaciones) es mayor, por un comportamiento agronómico propio de la variedad. Las plantas de determinadas variedades de arándano y la mayoría de cultivos que presentan menor número de frutos, presentan regularmente un mayor peso promedio de frutos.

Las observaciones mas altas los puntos se alejan más de la recta, pero podriamos suponer que los datos siguen una distribución lineal. Por otro lado, se puede observar que la dispersión de la respuesta "Y" observada, tiene mayor error con respecto a la media condicional de la recta de regresión, a medida que el valor de x aumenta, lo que sería evidencia de un incumplimiento del supuesto de homocedasticidad.

```
#Generando gráfico de dispersión del número de bayas cosechables vs los Kg Cosechados por hectárea
datos %>%
  ggplot(aes(x=Hcalor,y=Kg.ha))+
  geom_point(position = "jitter", size=3, colour="red")+
  labs( title = "DIAGRAMA DE DISPERSIÓN")+
  geom_smooth(method = "lm")+
  theme_test()
```

```
## `geom_smooth()` using formula 'y ~ x'

## Warning in (function (kind = NULL, normal.kind = NULL, sample.kind = NULL) :
## non-uniform 'Rounding' sampler used
```



Se puede observar, que el acumulado de horas de calor (mayores de 24 °C) por semana tienen una relación inversa con los $Kg.ha^{-}1$ de bayas de arándano registrados por válvula, lo que supone que, a medida que el clima se torne más frío (temporada de otoño - invierto) la productividad del arándano es mayor.

Regresión lineal

```
summary(lm(Kg.ha~BayasCosechables+Hcalor,datos))
```

```
##
## Call:
## lm(formula = Kg.ha ~ BayasCosechables + Hcalor, data = datos)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -318.59 -167.46  -39.73  118.38  661.05
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)      112.5071    60.0894   1.872   0.0642 .
## BayasCosechables   3.1548     0.5023   6.281 9.53e-09 ***
## Hcalor             0.8662     1.1212   0.773   0.4416
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 213.5 on 97 degrees of freedom
## Multiple R-squared:  0.3182, Adjusted R-squared:  0.3042
## F-statistic: 22.64 on 2 and 97 DF,  p-value: 8.544e-09
```

Bootstrap

```
boot3<-function(datos,B,estadistico,...){
  n<-nrow(datos)
  p<-ncol(datos)
  estaboot<-matrix(0,B,p)
  for(i in 1:B){
    indices<-sample(1:n,n,T)
    estaboot[i,]<-estadistico(datos[indices,],...)
  }

  esboot<-apply(estaboot,2,mean)
  eeboot<-apply(estaboot,2,sd)
  return(list(esboot=esboot,eeboot=eeboot))
}

coefi<-function(datos,y){
  datos<-as.matrix(datos)
  betas<-lm(datos[,y]~datos[,-y])$coe
  return(betas)
}


boot1<-function(datos,B,estadistico,...){
  n<-nrow(datos)
```

```
  p<-ncol(datos)
  estaboot<-matrix(0,B,p)
  for(i in 1:B){
    indices<-sample(1:n,n,T)
    estaboot[i,]<-estadistico(datos[indices,],...)
  }

  esboot<-mean(estaboot)
  eeboot<-sd(estaboot)
  return(list(esboot=esboot,eeboot=eeboot))
}


r2adj<-function(datos,y){
  datos<-as.matrix(datos)
  r2adj <- summary(lm(datos[,y]~datos[,-y]))$adj.r.squared
  return(r2.adj = r2adj)
}

# coefi(datos,3)

RNGkind(sample.kind="Rounding")
```

```
## Warning in RNGkind(sample.kind = "Rounding"): non-uniform 'Rounding' sampler
## used
```

```
set.seed(99)
boot3(datos,50,coefi,3)
```

```
## $esboot
## [1] 118.9549832   3.1534385   0.4859167
##
## $eeboot
## [1] 48.0906795  0.3733184  1.0592750
```

```
boot1(datos,50,r2adj,3)
```

```
## $esboot
## [1] 0.3033086
##
## $eeboot
## [1] 0.0711541
```

Validación cruzada

```
crossval<-function(datos,K,r,d){
  datos<-as.matrix(datos)
  n<-nrow(datos)
  EVC<-c()
```

```
  resid<-c()
  subm<-floor(n/K)
  resi<-lm(datos[,d]~datos[,-d])$res
  APE<-sum(resi^2)/n
  for(i in 1:r){
    indices<-sample(n,n)
    azar<-datos[indices,]

    for(j in 1:K){
      unid<- ((j-1)*subm+1):(subm*j)
      if (j==K)
      {
        unid<-((j-1)*subm+1):n
      }
      datosp<-azar[unid,]
      datose<-azar[-unid,]
      ye<-datose[,d]
      xe<-datose[,-d]
      betas<-lm(ye~xe)$coef
      r2adj <- summary(lm(ye~xe))$adj.r.squared
      datosp1<-cbind(1,datosp[,-d])
      estim<-datosp1%*%betas
      resid[j]<-sum((datosp[,d]-estim)^2)
    }
    EVC[i]<-sum(resid)/n
  }
  EVCP<-mean(EVC)
  cvEVC<-sd(EVC)*100/EVCP
  sesgo<-EVCP-APE
  return(list(betas = betas, r2.adj = r2adj, APE=APE,EVCP=EVCP,cvEVC=cvEVC,sesgo=sesgo))
}

RNGkind(sample.kind="Rounding")
```

```
## Warning in RNGkind(sample.kind = "Rounding"): non-uniform 'Rounding' sampler
## used
```

```
set.seed(80)
crossval(datos,10,1,3)
```

```
## $betas
##        (Intercept) xeBayasCosechables           xeHcalor
##        101.1063511          3.2515176          0.6110509
##
## $r2.adj
## [1] 0.3495893
##
## $APE
## [1] 44223.92
##
## $EVCP
## [1] 46707.31
```

```
##
## $cvEVC
## [1] NA
##
## $sesgo
## [1] 2483.389
```

Validación cruzada repetida

```
crossval<-function(datos,K,r,d){
  datos<-as.matrix(datos)
  n<-nrow(datos)
  EVC<-c()
  resid<-c()
  subm<-floor(n/K)
  resi<-lm(datos[,d]~datos[,-d])$res
  APE<-sum(resi^2)/n
  for(i in 1:r){
    indices<-sample(n,n)
    azar<-datos[indices,]

    for(j in 1:K){
      unid<- ((j-1)*subm+1):(subm*j)
      if (j==K)
      {
        unid<-((j-1)*subm+1):n
      }
      datosp<-azar[unid,]
      datose<-azar[-unid,]
      ye<-datose[,d]
      xe<-datose[,-d]
      betas<-lm(ye~xe)$coef
      r2adj <- summary(lm(ye~xe))$adj.r.squared
      datosp1<-cbind(1,datosp[,-d])
      estim<-datosp1%*%betas
      resid[j]<-sum((datosp[,d]-estim)^2)
    }
    EVC[i]<-sum(resid)/n
  }
  EVCP<-mean(EVC)
  cvEVC<-sd(EVC)*100/EVCP
  sesgo<-EVCP-APE
  return(list(betas = betas, r2.adj = r2adj, APE=APE,EVCP=EVCP,cvEVC=cvEVC,sesgo=sesgo))
}

RNGkind(sample.kind="Rounding")
```

```
## Warning in RNGkind(sample.kind = "Rounding"): non-uniform 'Rounding' sampler
## used
```

```r
set.seed(80)
crossval(datos,10,5,3)
```

```
## $betas
##        (Intercept) xeBayasCosechables           xeHcalor
##        106.6268053          3.2689411          0.4796198
##
## $r2.adj
## [1] 0.3428871
##
## $APE
## [1] 44223.92
##
## $EVCP
## [1] 46938.86
##
## $cvEVC
## [1] 1.39587
##
## $sesgo
## [1] 2714.946
```

Jacknife

```r
jack2<-function(datos,estadistico,...){
  n<-nrow(datos)
  estjack<-c()
  for(i in 1:n){
    estjack[i]<-estadistico(datos[-i,],...)
  }
  esjack<-mean(estjack)
  eejack<-(n-1)*sd(estjack)/sqrt(n)
  return(list(esjack=esjack,eejack=eejack))
}

coefi1<-function(datos,y){
  datos<-as.matrix(datos)
  betas<-lm(datos[,y]~datos[,-y])$coe
  return(Intercepto = betas[1])
}

coefi2<-function(datos,y){
  datos<-as.matrix(datos)
  betas<-lm(datos[,y]~datos[,-y])$coe
  return(beta1 = betas[2])
}

coefi3<-function(datos,y){
  datos<-as.matrix(datos)
  betas<-lm(datos[,y]~datos[,-y])$coe
```

```
  return(beta2 = betas[3])
}

r2adj<-function(datos,y){
  datos<-as.matrix(datos)
  r2adj <- summary(lm(datos[,y]~datos[,-y]))$adj.r.squared
  return(r2.adj = r2adj)
}

# coefi1(datos,3)

RNGkind(sample.kind="Rounding")
```

```
## Warning in RNGkind(sample.kind = "Rounding"): non-uniform 'Rounding' sampler
## used
```

```
set.seed(99)
jack2(datos,coefi1,3)
```

```
## $esjack
## [1] 112.5046
##
## $eejack
## [1] 53.23454
```

```
jack2(datos,coefi2,3)
```

```
## $esjack
## [1] 3.154972
##
## $eejack
## [1] 0.426692
```

```
jack2(datos,coefi3,3)
```

```
## $esjack
## [1] 0.8656901
##
## $eejack
## [1] 1.264109
```

```
jack2(datos,r2adj,3)
```

```
## $esjack
## [1] 0.3041685
##
## $eejack
## [1] 0.07615932
```

Resumen General

Método	β_0	β_1	β_2	R^2 **ajustado**
Regresion lineal	112.507 ()	3.154 ()	0.866 ()	0.304 ()
Bootstrap	118.954 (48.09)	3.153 (0.37)	0.485 (1.05)	0.303 (0.07)
K - Fold CV	101.106 ()	3.251 ()	0.611 ()	0.349 ()
K - Fold CV Repetido	106.626 ()	3.268 ()	0.479 ()	0.342 ()
Jacknife	112.504 (53.23)	3.154 (0.42)	0.865 (1.26)	0.304 (0.07)

β_1 -> Coeficiente de Bayas cosechables

β_2 -> Coeficiente de Horas de calor