CargarDatos

Ignacio Acosta, Mauro Loprete y Sofía Itté

04 de June

Carga de datos

```
Datos <- read.table(
    "UScrime.txt",
    header = TRUE,
    dec = ","
)
```

Primera idea: Incluir todas las variables :

```
lm(
    y ~ . ,
    data = Datos
) %>% summary()
```

```
##
## Call:
## lm(formula = y ~ ., data = Datos)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -395.74  -98.09   -6.69  112.99  512.67
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -5984.2876  1628.3184  -3.675 0.000893 ***
## M               8.7830     4.1714   2.106 0.043443 *
## So             -3.8035   148.7551  -0.026 0.979765
## Ed             18.8324     6.2088   3.033 0.004861 **
## Po1            19.2804    10.6110   1.817 0.078892 .
## Po2           -10.9422    11.7478  -0.931 0.358830
## LF             -0.6638     1.4697  -0.452 0.654654
## M.F             1.7407     2.0354   0.855 0.398995
## Pop            -0.7330     1.2896  -0.568 0.573845
## NW              0.4204     0.6481   0.649 0.521279
## U1             -5.8271     4.2103  -1.384 0.176238
```

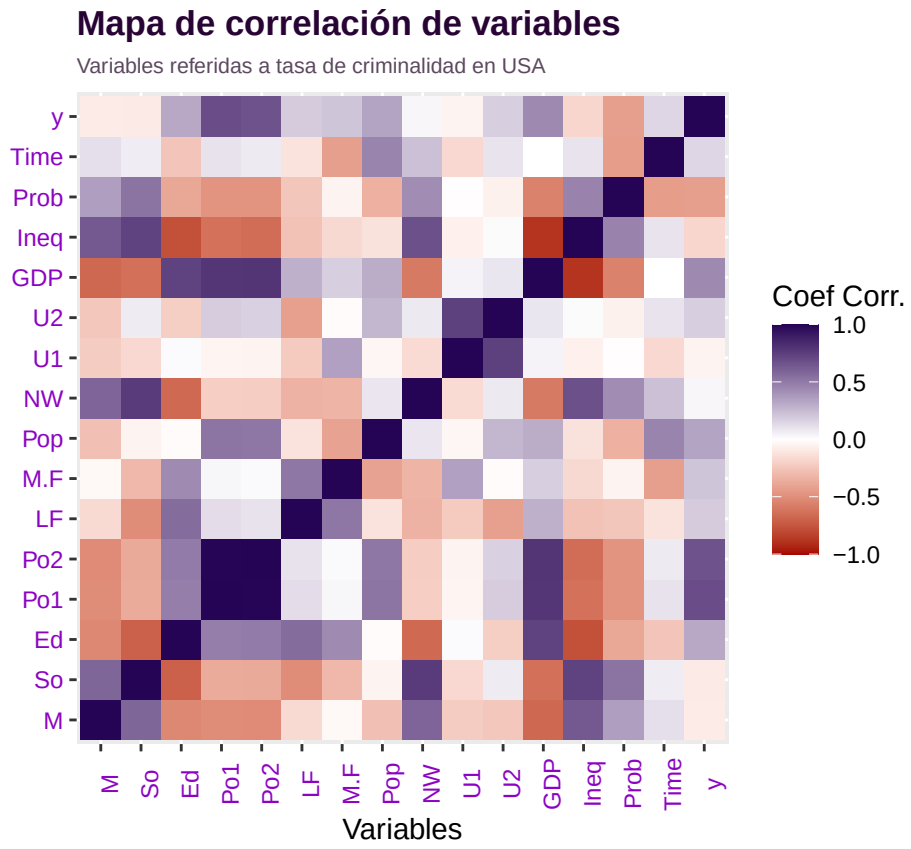
```
## U2             16.7800     8.2336   2.038 0.050161 .
## GDP             0.9617     1.0367   0.928 0.360754
## Ineq            7.0672     2.2717   3.111 0.003983 **
## Prob        -4855.2658  2272.3746  -2.137 0.040627 *
## Time           -3.4790     7.1653  -0.486 0.630708
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 209.1 on 31 degrees of freedom
## Multiple R-squared:  0.8031, Adjusted R-squared:  0.7078
## F-statistic: 8.429 on 15 and 31 DF,  p-value: 3.539e-07
```

En este caso se puede ver que pocas variables son significativas, pero el modelo en general si lo es ...

Pasos siguientes :

- Hay que empezar el proceso de ... (Backward/Forward/Stepwise)
- Diagnostico ...

```
qplot(x=Var1,
      y=Var2,
      data = melt(cor(Datos, use = "p")),
      fill = value,
      geom="tile"
)+scale_fill_gradient2(limits = c(-1, 1),
                       low = "#A50303", high = "#250455")+
  theme(aspect.ratio = 1,
        plot.title = element_text(
          face = "bold",
          color = "#280434FF"),
        plot.subtitle = element_text(
          size=8,
          color="#5C485F"
        ),
        axis.text.x = element_text(
          color = "#8C04C2",
          angle=90
        ),
        axis.text.y = element_text(
          color="#8C04C2"
        )
        )+
  labs(title="Mapa de correlación de variables",
       subtitle = "Variables referidas a tasa de criminalidad en USA",
       x="Variables",
       y="",
       fill="Coef Corr.")
```



Histogramas

```
# Multiple plot function, la hizo Dios.

multiplot <- function(..., plotlist=NULL, file, cols=1, layout=NULL) {
  library(grid)

  # Make a list from the ... arguments and plotlist
  plots <- c(list(...), plotlist)

  numPlots = length(plots)

  # If layout is NULL, then use 'cols' to determine layout
  if (is.null(layout)) {
    # Make the panel
    # ncol: Number of columns of plots
    # nrow: Number of rows needed, calculated from # of cols
    layout <- matrix(seq(1, cols * ceiling(numPlots/cols)),
                     ncol = cols, nrow = ceiling(numPlots/cols))
  }

 if (numPlots==1) {
    print(plots[[1]])

 } else {
    # Set up the page
    grid.newpage()
```

```
    pushViewport(viewport(layout = grid.layout(nrow(layout), ncol(layout))))

    # Make each plot, in the correct location
    for (i in 1:numPlots) {
      # Get the i,j matrix positions of the regions that contain this subplot
      matchidx <- as.data.frame(which(layout == i, arr.ind = TRUE))

      print(plots[[i]], vp = viewport(layout.pos.row = matchidx$row,
                                      layout.pos.col = matchidx$col))
    }
  }
}

HistM<-ggplot(Datos,aes(x=M))+
  geom_histogram(binwidth = 5,
                 fill="#8F3A84FF",
                 colour="black")+
  theme(aspect.ratio = 1,
        plot.subtitle = element_text(
          size=6.5
        ),
        plot.title = element_text(
          face="bold",
          size=10,
          color = "#280434FF"
        ),
        axis.text.x = element_text(
          color = "#8C04C2"
        ),
        axis.text.y = element_text(
          color="#8C04C2"
        ))+
  labs(title="Histograma variable M",
       subtitle="Número de hombres entre 14 y 24
       años por 1000 habitantes",
       x="Cantidad de Hombres",
       y="Cantidad de Estados")


BarpSo<-ggplot(data=Datos, aes(x=factor(So))) +
  geom_bar(stat="count",
           fill="#8F3A84FF",
           colour="black")+
  theme(plot.subtitle = element_text(
          size=6.5
        ),
        plot.title = element_text(
          face = "bold",
          size=10,
          color = "#280434FF"
        ),
        axis.text.x = element_text(
          color = "#8C04C2"
```

```
        ),
        axis.text.y = element_text(
          color="#8C04C2"
        ),
        aspect.ratio=1)+
  labs(title = "Barplot variable So",
       subtitle = "¿Es este un estado sureño?",
       x="Respuesta",
       y="Cantidad de respuestas")

HistEd<-ggplot(Datos,aes(x=Ed))+
  geom_histogram(binwidth = 5,
                 fill="#8F3A84FF",
                 colour="black")+
  theme(aspect.ratio = 1,
        plot.subtitle = element_text(
          size=6.5
        ),
        plot.title = element_text(
          size=10,
          face="bold",
          color = "#280434FF"
        ),
        axis.text.x = element_text(
          color = "#8C04C2"
        ),
        axis.text.y = element_text(
          color="#8C04C2"
        ))+
  labs(title="Histograma variable Ed",
       subtitle="Escolaridad del Estado",
       x="Puntuación",
       y="Cantidad de Estados")

HistPo1<-ggplot(Datos,aes(x=Po1))+
  geom_histogram(binwidth = 5,
                 fill="#8F3A84FF",
                 colour="black")+
  theme(aspect.ratio = 1,
        plot.subtitle = element_text(
          size=6.5
        ),
        plot.title = element_text(
          face="bold",
          size=10,
          color = "#280434FF"
        ),
        axis.text.x = element_text(
          color = "#8C04C2"
        ),
        axis.text.y = element_text(
          color="#8C04C2"
        ))+
```

```
  labs(title="Histograma variable Po1",
       subtitle = "1960",
       x="Gasto",
       y="Cantidad de Estados")

HistPo2<-ggplot(Datos,aes(x=Po2))+
  geom_histogram(binwidth = 5,
                 fill="#8F3A84FF",
                 colour="black")+
  theme(aspect.ratio = 1,
        plot.subtitle = element_text(
          size=6.5
        ),
        plot.title = element_text(
          face="bold",
          size=10,
          color = "#280434FF"
        ),
        axis.text.x = element_text(
          color = "#8C04C2"
        ),
        axis.text.y = element_text(
          color="#8C04C2"
        ))+
  labs(title="Histograma variable Po2",
       subtitle = "1959",
       x="Gasto",
       y="Cantidad de Estados")

HistLF<-ggplot(Datos,aes(x=LF))+
  geom_histogram(binwidth = 5,
                 fill="#8F3A84FF",
                 colour="black")+
  theme(aspect.ratio = 1,
        plot.subtitle = element_text(
          size=6.5
        ),
        plot.title = element_text(
          face="bold",
          size=10,
          color = "#280434FF"
        ),
        axis.text.x = element_text(
          color = "#8C04C2"
        ),
        axis.text.y = element_text(
          color="#8C04C2"
        ))+
  labs(title="Histograma variable LF",
       subtitle = "TPFLM de 14 a 24 años",
       x="Gasto",
       y="Cantidad de Estados")
```

```r
HistMF<-ggplot(Datos,aes(x=M.F))+
  geom_histogram(binwidth = 15,
                 fill="#8F3A84FF",
                 colour="black")+
  theme(aspect.ratio = 1,
        plot.subtitle = element_text(
          size=6.5
        ),
        plot.title = element_text(
          face="bold",
          size=10,
          color = "#280434FF"
        ),
        axis.text.x = element_text(
          color = "#8C04C2"
        ),
        axis.text.y = element_text(
          color="#8C04C2"
        ))+
  labs(title="Histograma variable M.F",
       subtitle = "Cantidad de hombres cada 1000 mujeres",
       x="Cantidad de Hombres",
       y="Cantidad de Estados")

HistPop<-ggplot(Datos,aes(x=Pop))+
  geom_histogram(binwidth = 15,
                 fill="#8F3A84FF",
                 colour="black")+
  theme(aspect.ratio = 1,
        plot.subtitle = element_text(
          size=6.5
        ),
        plot.title = element_text(
          face="bold",
          size=10,
          color = "#280434FF"
        ),
        axis.text.x = element_text(
          color = "#8C04C2"
        ),
        axis.text.y = element_text(
          color="#8C04C2"
        ))+
  labs(title="Histograma variable Pop",
       subtitle = "Tamaño de la población en cienmiles",
       x="Pop",
       y="Cantidad de Estados")

HistNW<-ggplot(Datos,aes(x=NW))+
  geom_histogram(binwidth = 15,
                 fill="#8F3A84FF",
                 colour="black")+
  theme(aspect.ratio = 1,
```

```
        plot.subtitle = element_text(
          size=6.5
        ),
        plot.title = element_text(
          face="bold",
          size=10,
          color = "#280434FF"
        ),
        axis.text.x = element_text(
          color = "#8C04C2"
        ),
        axis.text.y = element_text(
          color="#8C04C2"
        ))+
  labs(title="Histograma variable NW",
       subtitle = "Cantidad de no caucásicos cada 1000 habitantes",
       x="Cantidad de no caucásicos",
       y="Cantidad de Estados")

HistU1<-ggplot(Datos,aes(x=U1))+
  geom_histogram(binwidth = 10,
                 fill="#8F3A84FF",
                 colour="black")+
  theme(aspect.ratio = 1,
        plot.subtitle = element_text(
          size=6.5
        ),
        plot.title = element_text(
          face="bold",
          size=10,
          color = "#280434FF"
        ),
        axis.text.x = element_text(
          color = "#8C04C2"
        ),
        axis.text.y = element_text(
          color="#8C04C2"
        ))+
  labs(title="Histograma variable U1",
       subtitle = "TD cada 1000 habitantes,
hombres de 14 a 24 ",
       x="U1",
       y="Cantidad de Estados")

HistU2<-ggplot(Datos,aes(x=U2))+
  geom_histogram(binwidth = 10,
                 fill="#8F3A84FF",
                 colour="black")+
  theme(aspect.ratio = 1,
        plot.subtitle = element_text(
          size=6.5
        ),
        plot.title = element_text(
```

```
          face="bold",
          size=10,
          color = "#280434FF"
        ),
        axis.text.x = element_text(
          color = "#8C04C2"
        ),
        axis.text.y = element_text(
          color="#8C04C2"
        ))+
  labs(title="Histograma variable U1",
       subtitle = "TD cada 1000 habitantes,
hombres de 35 a 39 ",
       x="U2",
       y="Cantidad de Estados")

HistGDP<-ggplot(Datos,aes(x=GDP))+
  geom_histogram(binwidth = 50,
                 fill="#8F3A84FF",
                 colour="black")+
  theme(aspect.ratio = 1,
        plot.subtitle = element_text(
          size=6.5
        ),
        plot.title = element_text(
          face="bold",
          size=10,
          color = "#280434FF"
        ),
        axis.text.x = element_text(
          color = "#8C04C2"
        ),
        axis.text.y = element_text(
          color="#8C04C2"
        ))+
  labs(title="Histograma variable GDP",
       subtitle = "PIB per cápita",
       x="GDP",
       y="Cantidad de Estados")+
  scale_x_continuous(breaks = c(200,300,400,500,600,700))+
  scale_y_continuous(breaks = c(2,4,6,8,10,12,14,16))

HistIneq<-ggplot(Datos,aes(x=Ineq))+
  geom_histogram(binwidth = 10,
                 fill="#8F3A84FF",
                 colour="black")+
  theme(aspect.ratio = 1,
        plot.subtitle = element_text(
          size=6.5
        ),
        plot.title = element_text(
          face="bold",
          size=10,
```

```
          color = "#280434FF"
        ),
        axis.text.x = element_text(
          color = "#8C04C2"
        ),
        axis.text.y = element_text(
          color="#8C04C2"
        ))+
  labs(title="Histograma variable Ineq",
       subtitle = "Desigualdad de Ingresos",
       x="Ineq",
       y="Cantidad de Estados")+
   scale_x_continuous(breaks = c(100,150,200,250,300))+
   scale_y_continuous(breaks = c(2,4,6,8,10,12))

HistProb<-ggplot(Datos,aes(x=Prob))+
  geom_histogram(binwidth = 0.02,
                 fill="#8F3A84FF",
                 colour="black")+
  theme(aspect.ratio = 1,
        plot.subtitle = element_text(
          size=6.5
        ),
        plot.title = element_text(
          face="bold",
          size=10,
          color = "#280434FF"
        ),
        axis.text.x = element_text(
          color = "#8C04C2",
          angle = 90
        ),
        axis.text.y = element_text(
          color="#8C04C2"
        ))+
  labs(title="Histograma variable Prob",
       subtitle = "Probabilidad de encarcelamiento",
       x="P",
       y="Cantidad de Estados")+
  scale_x_continuous(breaks = c(0,0.02,0.04,0.06,0.08,0.10,0.12))

HistTime<-ggplot(Datos,aes(x=Time))+
  geom_histogram(binwidth = 8,
                 fill="#8F3A84FF",
                 colour="black")+
  theme(aspect.ratio = 1,
        plot.subtitle = element_text(
          size=6.5
        ),
        plot.title = element_text(
          face="bold",
          size=10,
          color = "#280434FF"
```

```
      ),
      axis.text.x = element_text(
        color = "#8C04C2"
      ),
      axis.text.y = element_text(
        color="#8C04C2"
      ))+
labs(title="Histograma variable Time",
     subtitle = "Tiempo promedio de estadía en cárceles estatales",
     x="Tiempo",
     y="Cantidad de Estados")
```

##Analisis de Histogramas

Los histogramas de las variables M, Po1, Nw, Po2, M.F, Pop, U1, U2, Prob y Time cuentan con una distribución asimétrica. Cuenta con menor variabilidad entre el valor mínimo y la mediana, dejado asi mayor variabilidad de observaciones entre la mediana y el valor máximo generando una cola hacia la derecha.

Por otro lado, los histogramas de las variables Ed y GDP también cuentan con una distribución asimétrica, pero en este caso cuenta con menor variabilidad entre la mediana y el valor máximo, dejando mayor variabilidad entre el valor mínimo y la mediana generando asi una cola hacia la izquierda.

La variable LF cuenta con una distribución casi simétrica ya que su media y su mediana solo difieren en 1. Tiene un intervalo modal que va desde el valor 530 a 590.

La variable ineq cuenta con una distribución asimétrica, tiene un intervalo modal entre los valores de . . . a. . . . Cuenta con menor variabilidad entre el primer y el tercer cuartil.

Por último, el histograma de la variable Y. Cuenta con una distribución asimétrica, tiene un intervalo modal entre los valores de 500 y 800 ofensas reportadas en casi 13 Estados. Podemos afirmar que la mayoría de las observaciones están concentradas entre 342 que es su valor mínimo y 831 que es la mediana. Dado que su valor máximo es el 1993, el histograma cuenta con una cola hacia la derecha.