Cómputo Estadístico (Tarea 2)

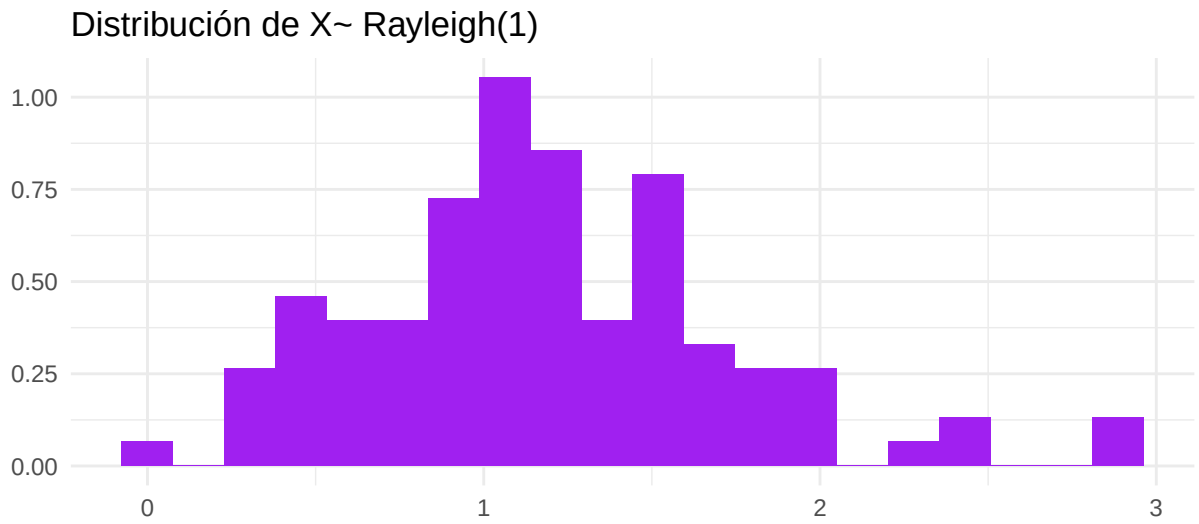
J. Antonio García Ramirez

10 de septiembre de 2018

1. *Generación de datos simulados y aplicación de los métodos de selección de subconjuntos (selección de los mejores subconjuntos y selección paso a paso).*

a) *Usa una función en* ***R*** *para generar una variable predictora* X *de longitud* $n = 100$*, así como un vector de ruido* ϵ *de tamaño* $n = 100$*.*

Retomando la tarea anterior, correspondiente a GLM, usare la distribución de $Rayleigh(\sigma = 1)$ para generar la v.a. X, en las siguientes gráficas podemos apreciar la distribución de la muestra generada.

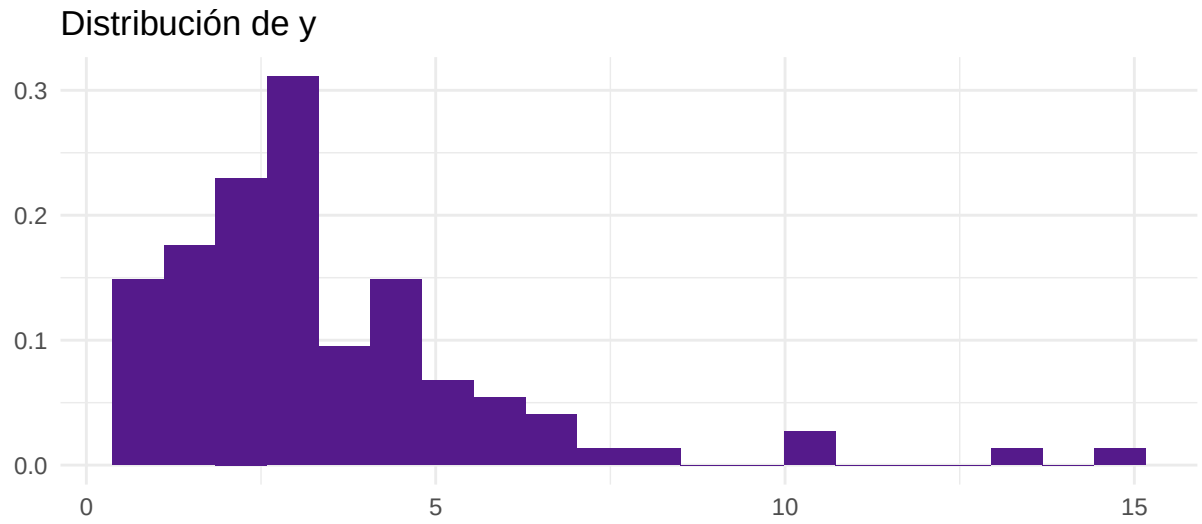


b) *Genera un vector de respuestas* Y *de longitud* $n = 100$ *de acuerdo al modelo*

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \epsilon$$

Donde $\beta_0, \beta_1, \beta_2$ *y* b_3 *son constantes de tu elección.*

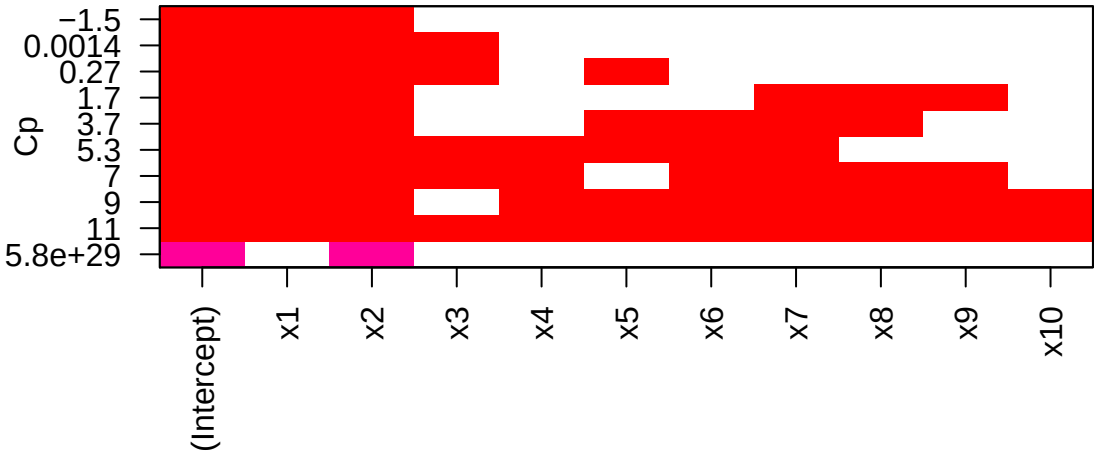
Y en la siguiente gráfica podemos apreciar la distribución de la variable de respuesta Y



c) *Utiliza la función regsubsets() para realizar la selección de los mejores subconjuntos con el fin de elegir el mejor modelo que contenga los predictores* $X, X^2, ..., X^{10}$. *¿Cuál es el mejor modelo obtenido según el* C_p , BIC *y el* R_2 *ajustado?, Muestra algunas gráficas que proporcionen evidencia de tu respuesta y reporta los coeficientes del mejor modelo obtenido.*

Considerando el criterio C_p tenemos la siguiente gráfica que muestra el desempeño utilizando este estadístico:

```
library(leaps)
z <- data.frame(x=x)
for ( i in 2:10) z[, as.character(i)] <- z$x**i
names(z) <- paste0('x',1:10)
modelos <- regsubsets(y~., data =  z, method = 'exhaustive', nvmax = 10)
#summary(modelos)
plot(modelos, scale="Cp", col=rainbow(10))
```



Donde podemos ver que el estadístico C_p se incrementa de -1.5 a $.0014$ cuando pasamos del conjunto de predictores $\{intercepto, X, X^2\}$ al conjunto $\{intercepto, X, X^2, X^3\}$ solo por confirmar observemos el RSS asociado a los dos conjuntos anteriores.

```
summary(lm(y ~ x+x2, data=z))$sigma
```

```
## [1] 8.124959e-16
```

```
#summary(lm(y ~x3 + x4+x5+x6+x7+x8+x9+x10-1, data=z))
  #salida para diferenciar de colores
summary(lm(y ~ x + x2 +x3, data=z))$sigma
```

```
## [1] 8.071218e-16
```

Donde se cumple que al incrementar el número de predictores el RSS disminuye, concluimos que según el criterio del estadístico C_p y la búsqueda exhaustiva de subconjuntos el mejor modelo es el que incluye a los dos primeros predictores y el intercepto. En la siguiente sección mostramos los coeficientes con los que se generó el modelo Y y las estimaciones $\hat{\beta}$, que son iguales a presición de millonésimas.

```
betas #coeficientes con los que se genero Y
```

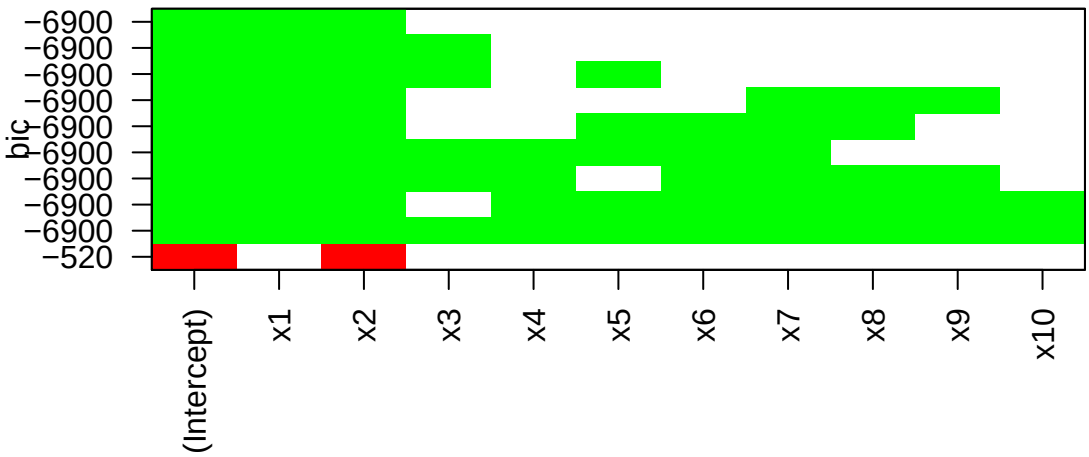
```
## (Intercept)           x          x2          x3
##   0.4504854   0.8142518   0.9287772   0.1474810
```

```
coef(lm(y ~ x+x2, data=z))#coeficientes estimados
```

```
## (Intercept)           x          x2
##   0.4504854   0.8142518   0.9287772
```

Continuando con el enfoque exhaustivo, pero considerando el criterio bayesiano, BIC tenemos el siguiente gráfico informativo:

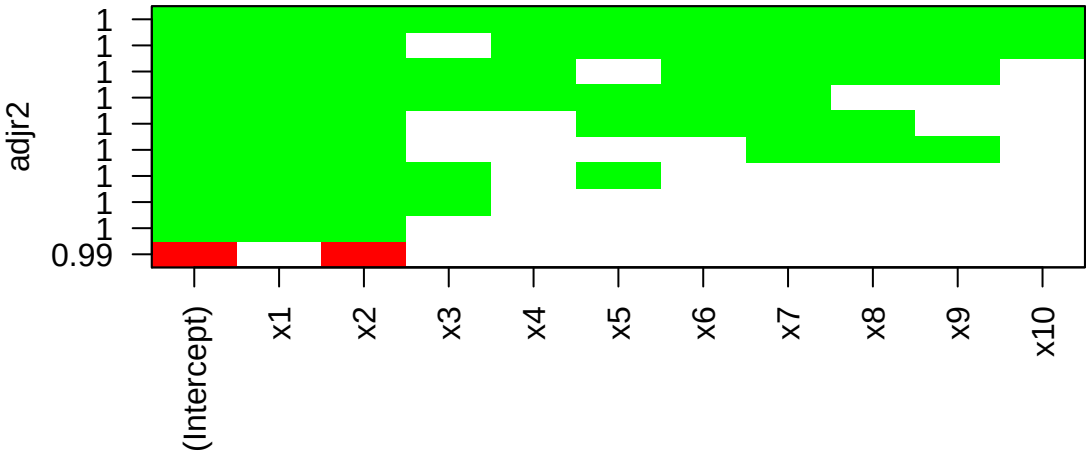
```r
plot(modelos, scale="bic", col=c('green', 'red'))
```



Donde nuevamente el modelo con los dos primeros predictores es el recomendado por el estadístico BIC, por lo que los coeficientes estimados son los mismos que en la sección anterior.

Finalmente repitiendo la metodología para el estadístico R^2 ajustado tenemos el desempeño en la siguiente imagen, donde podemos apreciar que este estadístico es el menos específico ya que vale la unidad en la mayoría de los casos; en esta circunstancia y como conocemos el modelo generador para Y evauaremos los modelo con el tres primeros predictores y por parsimonia el modelo con los dos primeros predictores:

```r
plot(modelos, scale="adjr2", col=c('green', 'red'))
```




```
summary(lm(y~ x1+x2+x3, data=z))
```

```
## Warning in summary.lm(lm(y ~ x1 + x2 + x3, data = z)): essentially perfect
## fit: summary may be unreliable
```

```
##
## Call:
## lm(formula = y ~ x1 + x2 + x3, data = z)
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -7.265e-15 -4.630e-17  8.950e-17  1.967e-16  1.038e-15
##
## Coefficients:
##              Estimate Std. Error   t value Pr(>|t|)
## (Intercept) 4.505e-01  5.121e-16 8.797e+14   <2e-16 ***
## x1          8.143e-01  1.331e-15 6.118e+14   <2e-16 ***
## x2          9.288e-01  1.028e-15 9.035e+14   <2e-16 ***
## x3          3.500e-16  2.310e-16 1.515e+00    0.133
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.071e-16 on 96 degrees of freedom
## Multiple R-squared:      1,  Adjusted R-squared:      1
## F-statistic: 1.759e+32 on 3 and 96 DF,  p-value: < 2.2e-16
```

```
summary(lm(y~ x1+x2, data=z))
```

```
## Warning in summary.lm(lm(y ~ x1 + x2, data = z)): essentially perfect fit:
## summary may be unreliable
```

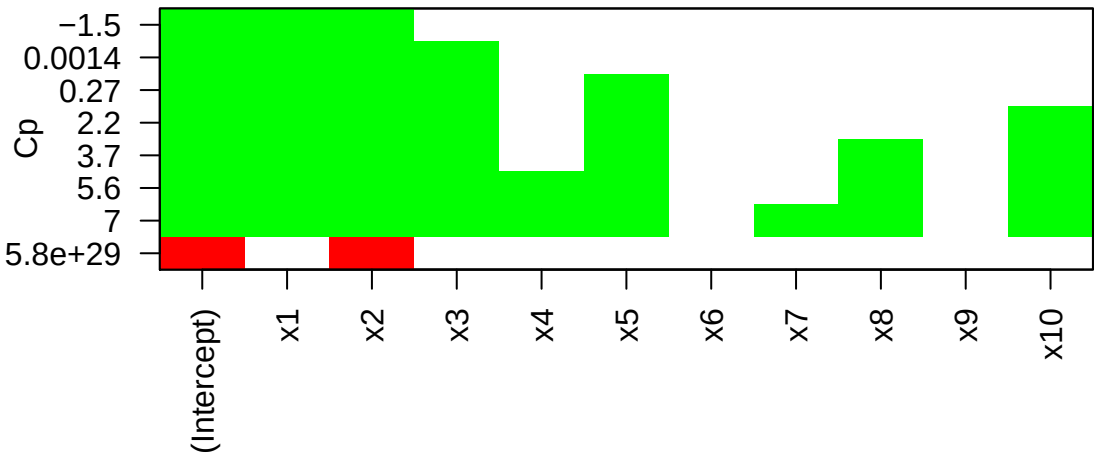
```
##
## Call:
## lm(formula = y ~ x1 + x2, data = z)
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -7.321e-15 -2.220e-17  1.044e-16  2.208e-16  9.900e-16
##
## Coefficients:
##              Estimate Std. Error   t value Pr(>|t|)
## (Intercept) 4.505e-01  3.336e-16 1.350e+15   <2e-16 ***
## x1          8.143e-01  5.049e-16 1.613e+15   <2e-16 ***
## x2          9.288e-01  1.774e-16 5.235e+15   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.125e-16 on 97 degrees of freedom
## Multiple R-squared:      1,  Adjusted R-squared:      1
## F-statistic: 2.604e+32 on 2 and 97 DF,  p-value: < 2.2e-16
```

De donde observamos que el coeficiente del tercer regresor no es significativo y teniendo el mismo R^2 ajustado, optamos por el modelo con solo los primeros dos regresores nuevamente.

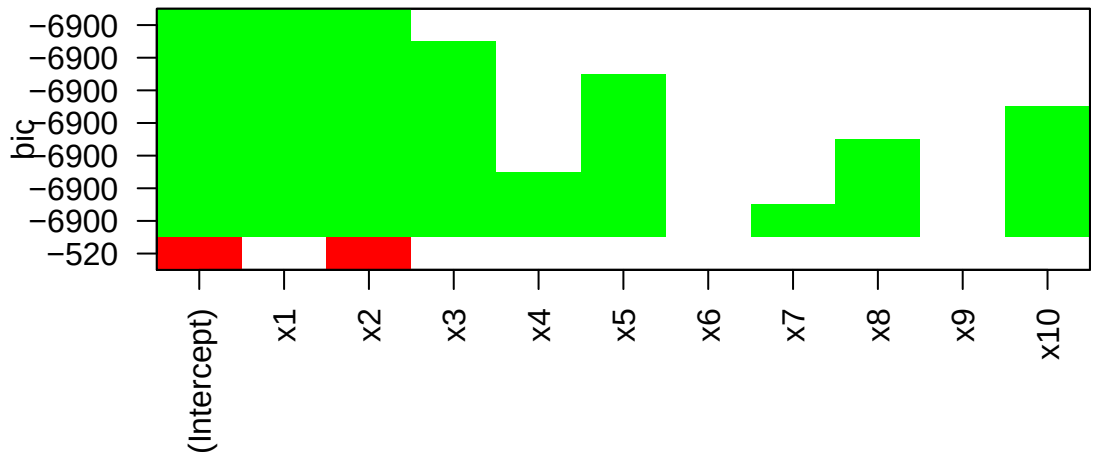
d) *Repite c), usando la selección paso a paso adelante y la selección paso paso atrás. ¿Cómo se compara tu respuesta con los resultados obtenidos en c)?*

Usando la selección hacia adelante tenemos nuevamente, que el mejor modelo es el que contienen a los dos primeros regresores y el intercepto (con los tres estadísticos).

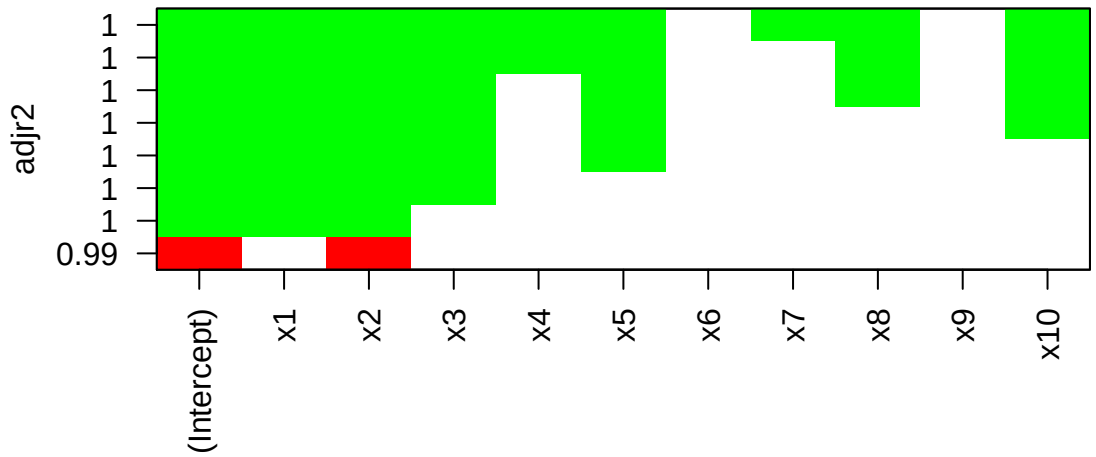
```
modelos <- regsubsets(y~ ., data =z, method = 'forward')
plot(modelos, scale="Cp", col=c('green', 'red'))
```



```
plot(modelos, scale="bic", col=c('green', 'red'))
```



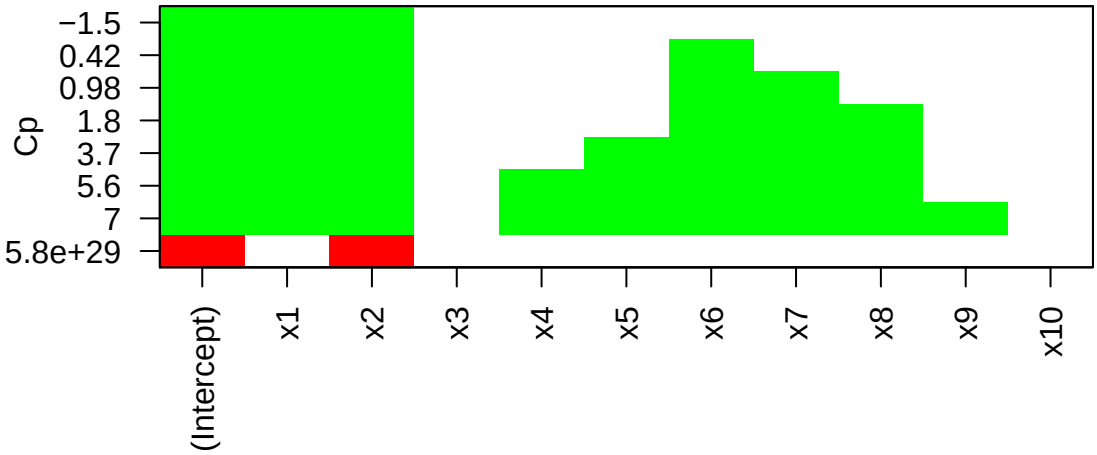

```
plot(modelos, scale="adjr2", col=c('green', 'red'))
```



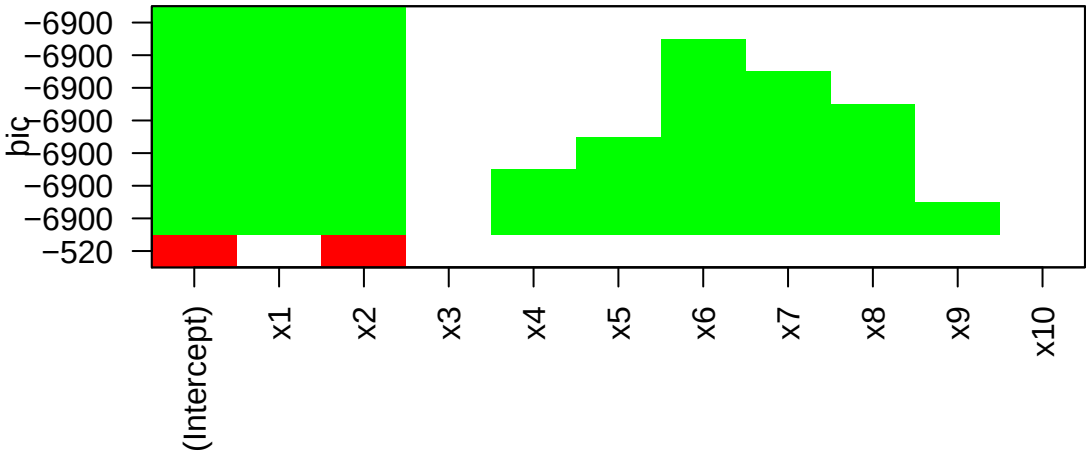
En vista de que son los mismos modelos que en la sección anterior los parámetros estimados son los mismos y por brevedad no los incluyo en el reporte.

Usando la selección hacia atrás tenemos resultados identicos para los tres estadísticos.

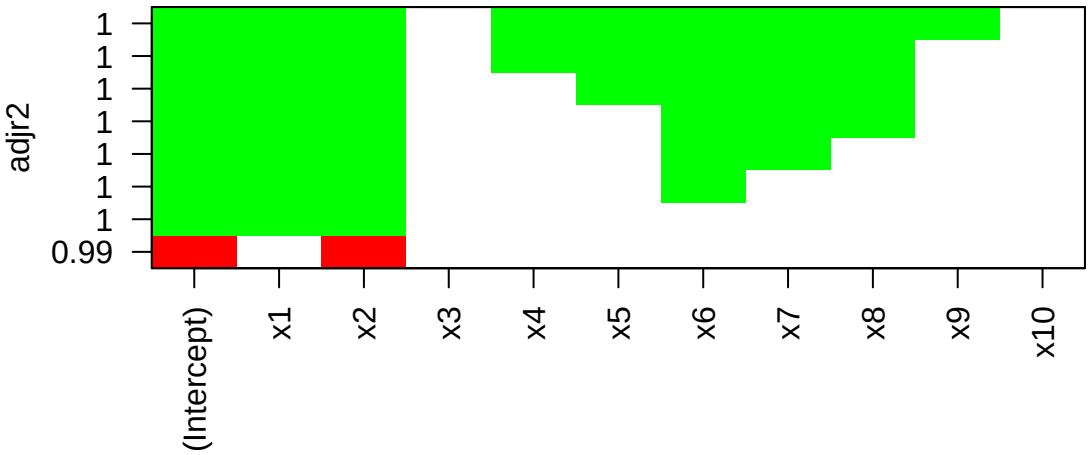
```
modelos <- regsubsets(y~ ., data =z, method = 'backward')
plot(modelos, scale="Cp", col=c('green', 'red'))
```




```
plot(modelos, scale="bic", col=c('green', 'red'))
```




```
plot(modelos, scale="adjr2", col=c('green', 'red'))
```



En vista de que son los mismos modelos que en las secciones anteriores los parámetros estimados son los mismos y por brevedad no los incluyo en el reporte.

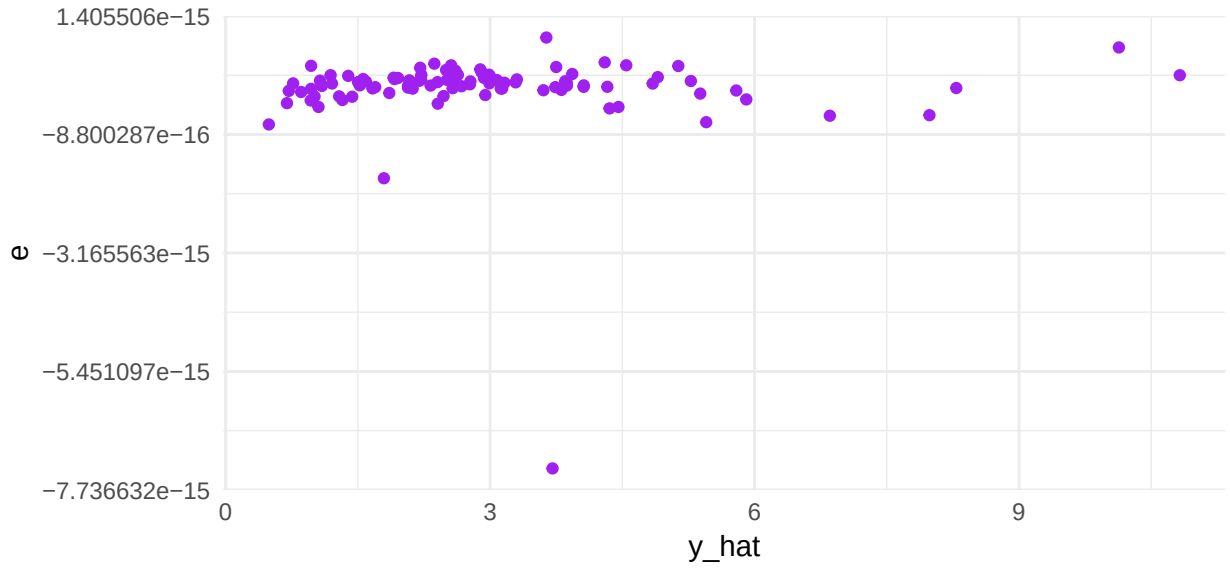
Como conclusión tenemos que los tres criterios son consistentes. Si bien el criterio R^2 no es tan automatizable, pues en nuestro ejemplo este criterio sugiere a primera vista un modelo con tres variables (y parece ser menos estable) no es gran ventaja frente al sencillo con solo los dos primeros predictores y el intercepto, de manera general sugiere los mismos conjuntos de variables que los otros dos criterios.

Parte importante de los resultados fue la generación del modelo Y pues elegimos una distribución con colas pesadas, si bien en la práctica se suelen transformar las variables que se identifican como tales, estas se presentan con mucha frecuencia (como los ingresos declarados ante Hacienda, los ingresos personales, etc) y en contra de nuestra intuición (que sugeriría un modelo con la variable X^3) el predictor X ajusta bien pero no perfectamente.

En un trabajo futuro preferiría al criteria bayesiano pues presenta mayor sensibilidad o penalización y considera en la penalización de manera más activa el tamaño de muestra y el número de regresores.

Terminamos mostrando los errores del modelo seleccionado, que si bien no son normales y muestran correlación y no independencia... recordemos que son los residuos de un modelo que busca reducir error de predicción (pronóstico) no inferencia (conocer acerca del modelo generador de las observaciones).

```
modelo <- lm(y~x1+x2, data=z)
e <- modelo$residuals
a <- data.frame(y_hat=fitted(modelo), res=e)
a$index <- 1:dim(a)[1]
ggplot(a, aes(y_hat, e)) + geom_point(aes(colour=I('purple'))) +
  theme_minimal()
```



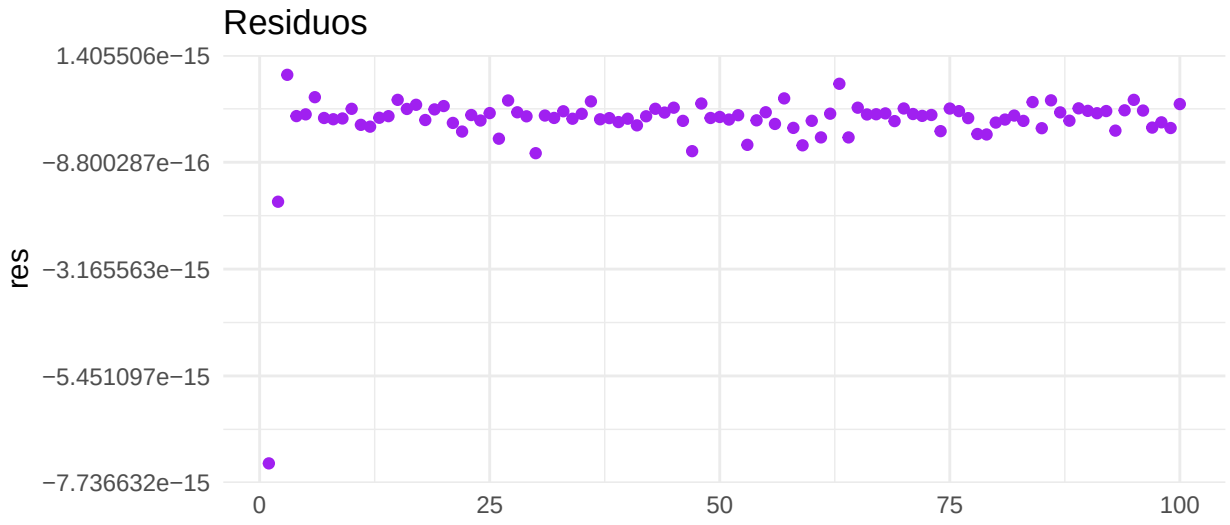

```
cor(a$res, a$y_hat)
```

```
## [1] -4.71545e-16
```

```
library(nortest)
ad.test(e)
```

```
##
##  Anderson-Darling normality test
##
## data:  e
## A = 16.602, p-value < 2.2e-16
```

```
ggplot(a, aes(index, res) )+  geom_point(aes(colour=I('purple'))) +
  theme_minimal() + ggtitle('Residuos ') +xlab('')
```




```
library(qqplotr)
set.seed(0)
ggplot(data = a, mapping = aes(sample = res , color = I('#619CFF')) ) +
  stat_qq_line() +  stat_qq_point() +
  geom_qq_band(bandType = "ts", mapping = aes(fill = "TS"), alpha = 0.1) +
  labs(x = "Theoretical Quantiles", y = "Sample Quantiles")+ theme_minimal() + ggtitle('QQ-normal, resi
```