Regresión Lineal versus Kernel

Juan C. Correa

3/10/2021

Para comprender cómo funciona la estrategia típica de la comunidad de R para aprender a usar las herramientas disponibles, vamos a hacer uso de la librería "np".

Comencemos entonces por arrancar un modelo de regresión múltiple paramétrico estándar. La variable dependiente (la que nos interesa analizar) es en este caso "logwage" y la variable independiente es "age".

```
library("np")
```

```
## Nonparametric Kernel Methods for Mixed Datatypes (version 0.60-10)
## [vignette("np_faq",package="np") provides answers to frequently asked questions]
## [vignette("np",package="np") an overview]
## [vignette("entropy_np",package="np") an overview of entropy-based methods]
```

```
data("cps71")
model.par <- lm(logwage ~ age + I(age^2), data = cps71)
summary(model.par)
```

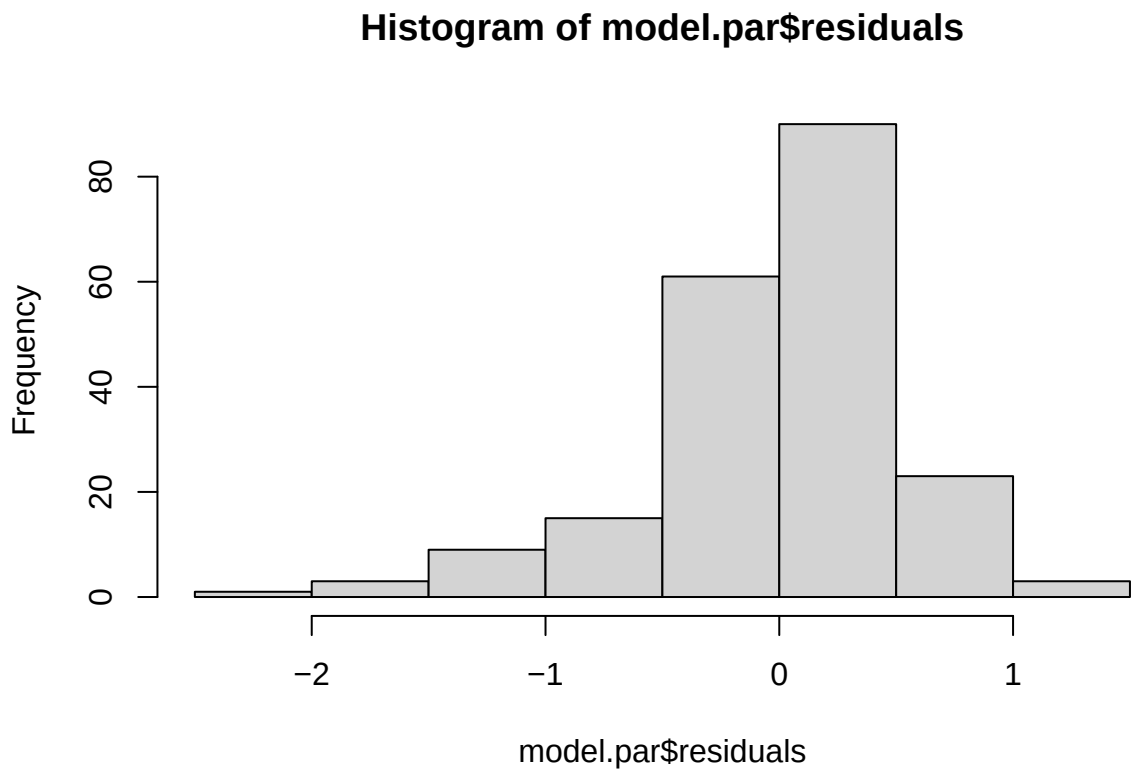
```
##
## Call:
## lm(formula = logwage ~ age + I(age^2), data = cps71)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.4041 -0.1711  0.0884  0.3182  1.3940
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 10.0419773  0.4559986  22.022  < 2e-16 ***
## age          0.1731310  0.0238317   7.265 7.96e-12 ***
## I(age^2)    -0.0019771  0.0002898  -6.822 1.02e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5608 on 202 degrees of freedom
## Multiple R-squared:  0.2308, Adjusted R-squared:  0.2232
## F-statistic:  30.3 on 2 and 202 DF,  p-value: 3.103e-12
```

Del output que se obtiene con la sintaxis previa, obsérvese que el resultado global del modelo anterior arroja un $R^2 = 0.2232$ (F(2, 202) = 30.3, p = 3.103e-12).

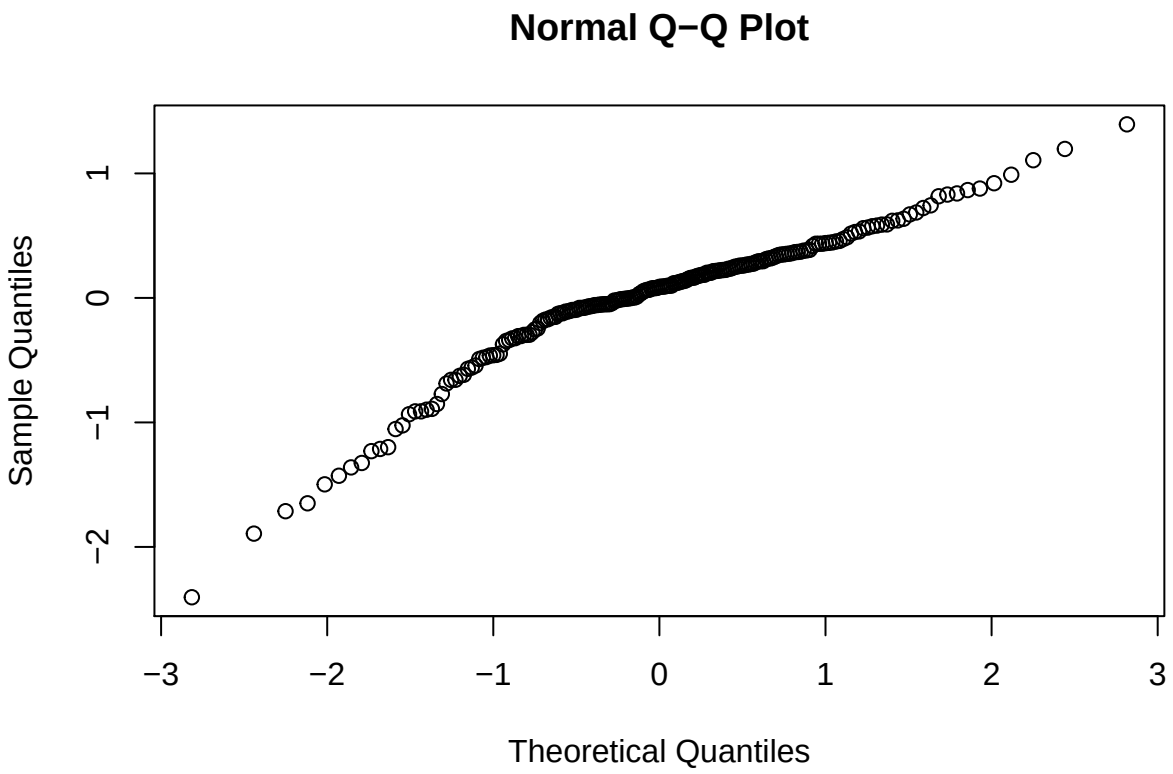
Una manera de chequear que uno de los supuestos de la regresión se cumple es solicitando un histograma a los residuales del ajuste del modelo. Este histograma debería ser semejante al de una distribución normal, o

un qqplot con una nube de puntos que muestre una patrón lineal, o simplemente solicitamos el cálculo de la asimetría y la kurtosis de esos residuales. Veamos.

```
hist(model.par$residuals)
```




```
qqnorm(model.par$residuals)
```




```
library(e1071)
skewness(model.par$residuals)
```

```
## [1] -1.131334
```

```
kurtosis(model.par$residuals)
```

```
## [1] 2.407596
```

Claramente, el modelo anterior está mostrando unos resultados que no cumplen con los principios o supuestos del modelo, y eso nos lleva a buscar otras opciones. Veamos ahora, siguiendo la documentación de la librería np, cómo proceder con una regresión lineal no-paramétrica.

```
model.np <- npreg(logwage ~ age,
regtype = "ll",
bwmethod = "cv.aic",
gradients = TRUE,
data = cps71)
```

```
## Multistart 1 of 1 |Multistart 1 of 1 |Multistart 1 of 1 |Multistart 1 of 1 /Multistart 1 of 1 |Multis
```

```
summary(model.np)
```

```
##
## Regression Data: 205 training points, in 1 variable(s)
##                      age
## Bandwidth(s): 2.805308
##
## Kernel Regression Estimator: Local-Linear
## Bandwidth Type: Fixed
## Residual standard error: 0.5215268
## R-squared: 0.3251639
##
## Continuous Kernel Type: Second-Order Gaussian
## No. Continuous Explanatory Vars.: 1
```

```
plot(cps71$age, cps71$logwage, xlab = "age", ylab = "log(wage)", cex=.1)
lines(cps71$age, fitted(model.np), lty = 1, col = "blue")
lines(cps71$age, fitted(model.par), lty = 1, col = " red")
```