Quantile Regression

Joaquín Amat Rodrigo j.amatrodrigo@gmail.com

Julio 2017

Índice

Versión PDF: https://github.com/JoaquinAmatRodrigo/Estadistica-con-R

Introducción

La regresión lineal por mínimos cuadrados, así como muchas de sus adaptaciones, se emplean como método para estimar la media de una variable respuesta condicionada a uno o varios predictores. Es decir, parte de la premisa de que la media de la variable respuesta depende del valor que tomen otras variables. En determinadas situaciones, puede ocurrir que la media no sea el parámetro más informativo y que en su lugar se desee conocer un determinado cuantil, por ejemplo el cuantil 0.5 (la mediana). Es aquí donde la regresión de cuantiles (*quantile regression*) interviene. Si bien la matemática es diferente a la empleada en regresión por mínimos cuadrados ordinarios, la interpretación final es muy similar. Se obtienen coeficientes de regresión que estiman el efecto que tiene cada predictor sobre un cuantil específico de la variable respuesta.

El poder realizar regresión sobre cualquier parte de la distribución permite conocer la influencia de los predictores desde el mínimo al máximo rango de la variable respuesta. Esto es especialmente útil en modelos de regresión en los que no se cumple la condición de varianza constante, ya que esto significa que no hay un único ratio de cambio (pendiente) que represente bien a toda la variable respuesta a la vez.

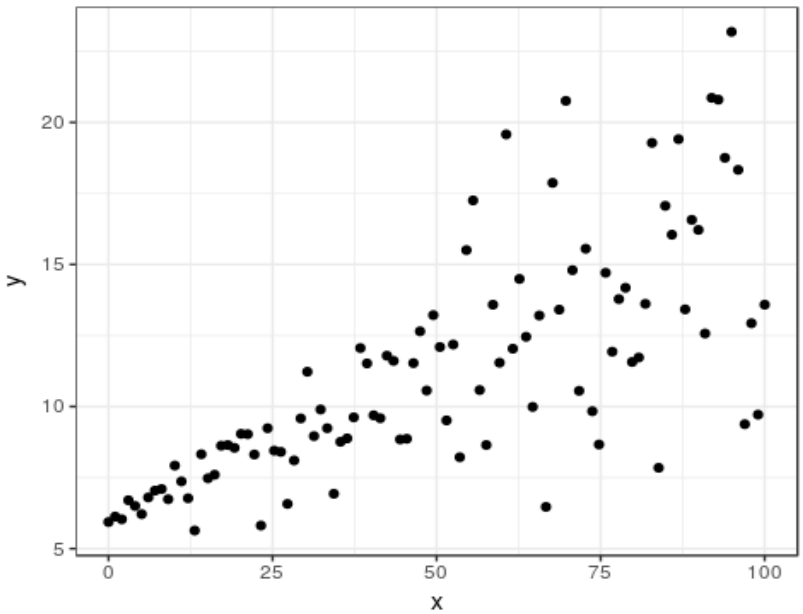
Cuantil: El cuantil de orden τ de una distribución ($0 < \tau < 1$) es el valor de la variable X que marca un corte de modo que una proporción τ de valores de la población es menor o igual que dicho valor. Por ejemplo, el cuantil de orden 0,36 deja un 36% de valores por debajo y el cuantil de orden 0,50 se corresponde con la mediana de la distribución.

En la práctica, la proporción de observaciones menor o igual a un determinado plano de regresión del cuantil no es exactamente igual a τ, debido a la varianza muestral.

Ejemplo

Supóngase las siguientes dos variables simuladas de tal forma que su varianza no es constante.

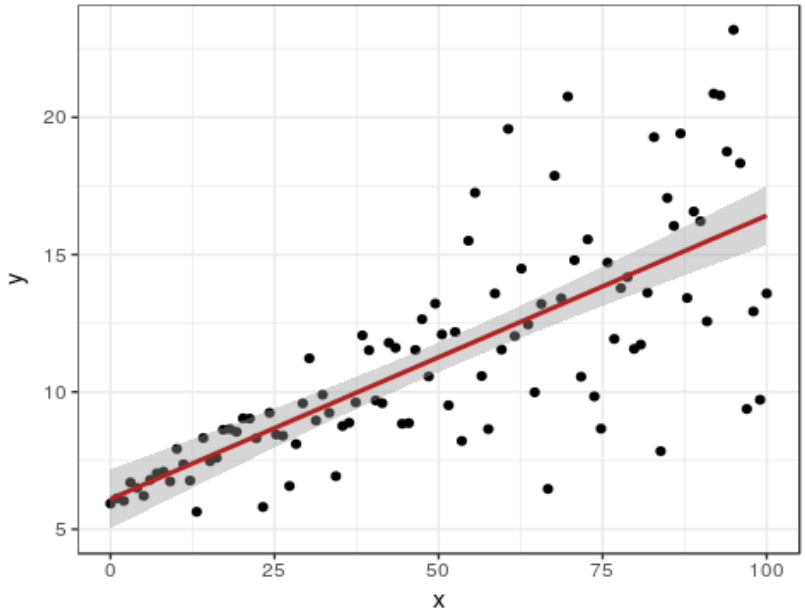
```
library(ggplot2)
# Predictor
x <- seq(0, 100, length.out = 100)
# Varianza no constante
varianza <- 0.1 + 0.05*x
# Intercept y pendiente
b_0 <- 6
b_1 <- 0.1
# Error aleatorio normal con varianza no constante
set.seed(1)
error <- rnorm(100, mean = 0, sd = varianza)
# Variable respuesta
y <- b_0 + b_1*x + error
datos <- data.frame(x, y)
ggplot(data = datos, aes(x = x, y = y)) +
  geom_point() +
  theme_bw()
```



La representación gráfica muestra una clara falta de homocedasticidad, la varianza de la variable respuesta aumenta con forme aumenta el valor del predictor. Esto supone una

violación de las condiciones necesarias para la regresión lineal por mínimos cuadrados ordinales, haciendo que sus resultados tengan una utilidad limitada.

```
ggplot(data = datos, aes(x = x, y = y)) +
  geom_point() +
  geom_smooth(method = "lm", colour = "firebrick", se = TRUE) +
  theme_bw()
```



A pesar de que la regresión lineal por mínimos cuadrados ordinarios va a calcular la estimación de la media de forma no sesgada, generando una estimación tan buena como es posible, para valores altos de X (por ejemplo X = 75) los límites de confianza van a ser mucho más optimistas de lo que realmente muestran los datos. Existen diferentes formas de intentar minimizar este problema, como por ejemplo la regresión por Weighted Least Squares. Esta sección se centra en la regresión por cuantiles.

La función `rq()` del paquete `quantreg` permite ajustar modelos de regresión para cualquier cuantil.

```
library(quantreg)
# regresión del quantil 0.9
modelo_q09 <- rq(formula = y ~ x, tau = 0.9,  data = datos)
summary(modelo_q09)
```

```
## Warning in rq.fit.br(x, y, tau = tau, ci = TRUE, ...): Solution may be
## nonunique
```

```
##
## Call: rq(formula = y ~ x, tau = 0.9, data = datos)
##
## tau: [1] 0.9
##
## Coefficients:
##             coefficients lower bd upper bd
## (Intercept) 6.04168      5.93265  6.39328
## x           0.16125      0.15591  0.17826
```

El `summary()` de un modelo `rq` devuelve los coeficientes de regresión de cada predictor junto con los correspondientes intervalos de confianza calculados por el método *rank* (existen 4 métodos adicionales). Acorde al modelo, por cada unidad que se incrementa el predictor *X*, el cuantil 0.9 de la variable respuesta aumenta en promedio 0.16 unidades. Además, como el intervalo de confianza no contiene el *0*, hay evidencias de que el predictor influye de forma significativa.

Para obtener el *p-value* asociado a cada predictor, se tiene que recurrir a los otros métodos disponibles:

```
summary(object = modelo_q09, se = "boot")
```

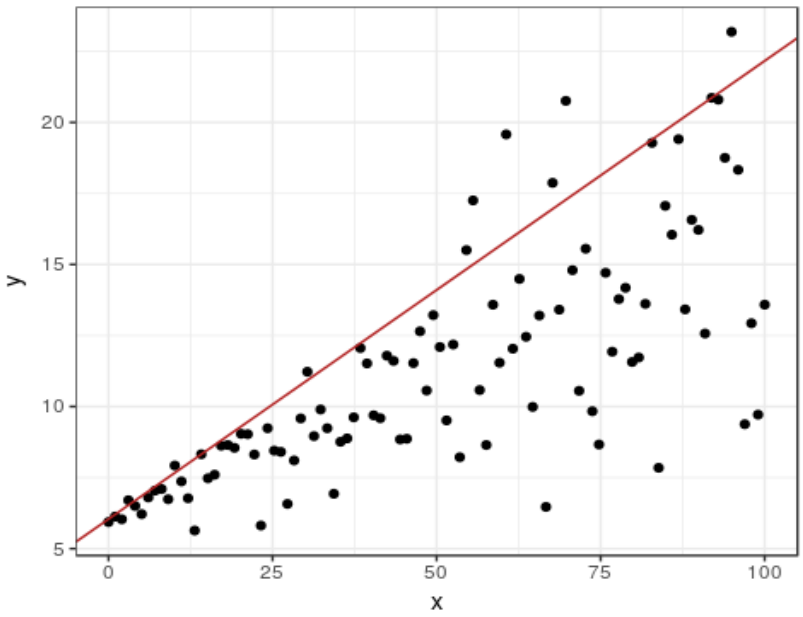
```
##
## Call: rq(formula = y ~ x, tau = 0.9, data = datos)
##
## tau: [1] 0.9
##
## Coefficients:
##             Value    Std. Error t value  Pr(>|t|)
## (Intercept)  6.04168  0.16870   35.81280  0.00000
## x            0.16125  0.01083   14.89372  0.00000
```

```
summary(object = modelo_q09, se = "nid")
```

```
##
## Call: rq(formula = y ~ x, tau = 0.9, data = datos)
##
## tau: [1] 0.9
##
## Coefficients:
##             Value    Std. Error t value  Pr(>|t|)
## (Intercept)  6.04168  0.26581   22.72912  0.00000
## x            0.16125  0.01632    9.88007  0.00000
```

Empleando las estimaciones de *intercept* y del predictor se puede representar la recta de regresión del cuantil.

```
ggplot(data = datos, aes(x = x, y = y)) +
  geom_point() +
  geom_abline(intercept = coef(modelo_q09)[1], slope = coef(modelo_q09)[2],
              colour = "firebrick") +
  theme_bw()
```



El paquete gráfico `ggplot2` contiene la función `geom_quantile()` que ajusta y representa directamente la recta de regresión.

```
ggplot(data = datos, aes(x = x, y = y)) +
  geom_point() +
  geom_quantile(quantiles = 0.9) +
  theme_bw()
```

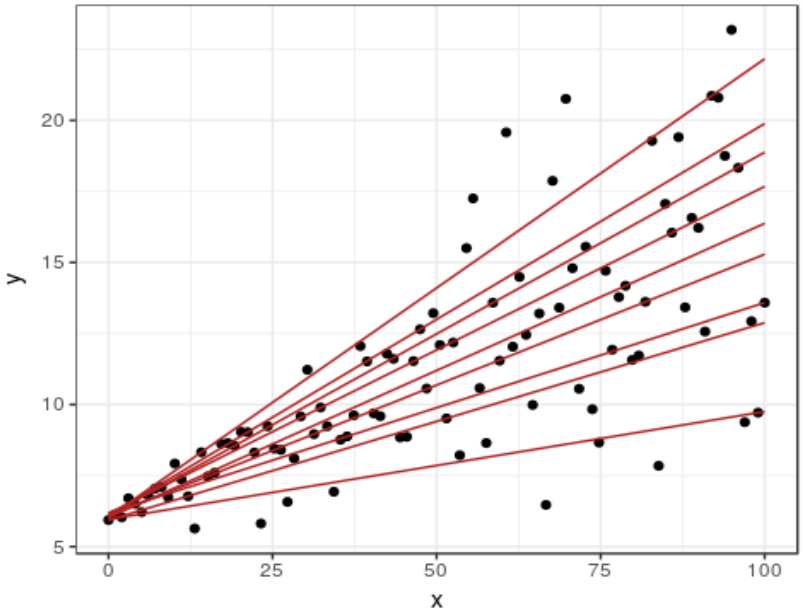
`rq()` permite ajustar varios cuantiles a la vez indicando la secuencia en el argumento `tau`.

```
modelo_q <- rq(formula = y ~ x, tau = c(0.1,0.2,0.3,0.4,0.5,0.6,0.7,0.8,0.9),
               data = datos)
summary(object = modelo_q, se = "boot")
```

```
##
## Call: rq(formula = y ~ x, tau = c(0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7,
##     0.8, 0.9), data = datos)
##
## tau: [1] 0.1
##
## Coefficients:
##             Value    Std. Error t value  Pr(>|t|)
## (Intercept)  5.95736  0.52419   11.36485  0.00000
## x            0.03796  0.01459    2.60118  0.01073
##
## Call: rq(formula = y ~ x, tau = c(0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7,
##     0.8, 0.9), data = datos)
##
## tau: [1] 0.2
##
## Coefficients:
##             Value    Std. Error t value  Pr(>|t|)
## (Intercept)  5.93735  0.18568   31.97671  0.00000
## x            0.06943  0.00825    8.41136  0.00000
##
## Call: rq(formula = y ~ x, tau = c(0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7,
##     0.8, 0.9), data = datos)
##
## tau: [1] 0.3
##
## Coefficients:
##             Value    Std. Error t value  Pr(>|t|)
## (Intercept)  6.19112  0.18881   32.79103  0.00000
## x            0.07395  0.00804    9.20083  0.00000
##
## Call: rq(formula = y ~ x, tau = c(0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7,
##     0.8, 0.9), data = datos)
##
## tau: [1] 0.4
##
## Coefficients:
##             Value    Std. Error t value  Pr(>|t|)
## (Intercept)  6.08146  0.22313   27.25464  0.00000
## x            0.09208  0.01102    8.35354  0.00000
##
## Call: rq(formula = y ~ x, tau = c(0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7,
##     0.8, 0.9), data = datos)
##
```

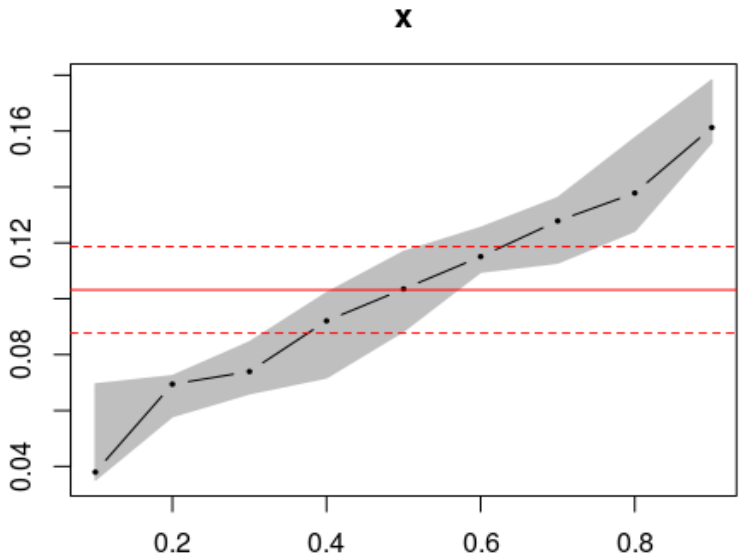
```
## tau: [1] 0.5
##
## Coefficients:
##             Value    Std. Error t value  Pr(>|t|)
## (Intercept)  6.02410  0.18650   32.30089  0.00000
## x            0.10351  0.01064    9.72588  0.00000
##
## Call: rq(formula = y ~ x, tau = c(0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7,
##     0.8, 0.9), data = datos)
##
## tau: [1] 0.6
##
## Coefficients:
##             Value    Std. Error t value  Pr(>|t|)
## (Intercept)  6.16841  0.17689   34.87151  0.00000
## x            0.11507  0.00796   14.45453  0.00000
##
## Call: rq(formula = y ~ x, tau = c(0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7,
##     0.8, 0.9), data = datos)
##
## tau: [1] 0.7
##
## Coefficients:
##             Value    Std. Error t value  Pr(>|t|)
## (Intercept)  6.09508  0.14699   41.46629  0.00000
## x            0.12784  0.00764   16.73064  0.00000
##
## Call: rq(formula = y ~ x, tau = c(0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7,
##     0.8, 0.9), data = datos)
##
## tau: [1] 0.8
##
## Coefficients:
##             Value    Std. Error t value  Pr(>|t|)
## (Intercept)  6.10539  0.20845   29.28953  0.00000
## x            0.13783  0.01246   11.06294  0.00000
##
## Call: rq(formula = y ~ x, tau = c(0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7,
##     0.8, 0.9), data = datos)
##
## tau: [1] 0.9
##
## Coefficients:
##             Value    Std. Error t value  Pr(>|t|)
## (Intercept)  6.04168  0.18763   32.19953  0.00000
## x            0.16125  0.01171   13.77635  0.00000
```

```
ggplot(data = datos, aes(x = x, y = y)) +
geom_point() +
geom_quantile(quantiles=c(0.1,0.2,0.3,0.4,0.5,0.6,0.7,0.8,0.9),color="firebrick") +
theme_bw()
```



Se puede observar que la pendiente de la recta de cada cuantil es distinta, lo que significa que el predictor *X* influye de forma distinta a cada cuantil de la variable respuesta. Para estudiar como varia esta influencia y su significancia, se pueden representar la pendiente para cada cuantil.

```
plot(summary(modelo_q), parm = "x")
```



Cada punto representa el coeficiente de regresión estimado de un cuantil. La línea continua roja es el coeficiente de regresión estimado para el predictor utilizando mínimos cuadrados ordinarios (`lm(y ~ x, data = datos)`) y las líneas discontinuas sus límites de confianza del 95%. Para este ejemplo, los cuantiles 0.1, 0.2, 0.3, 0.7, 0.8 y 0.9 se diferencian significativamente del valor obtenido por mínimos cuadrados ordinarios.

Además de conocer si un cuantil específico se ve significativamente influenciado por un predictor (*p-value* calculado en el summary), puede ser de interés determinar si los coeficientes de regresión de dos cuantiles son distintos. Para ello se ajustan modelos individuales con cada cuantil y se comparan mediante la función `anova()`.

```
modelo_q05 <- rq(formula = y ~ x, tau = 0.5, data = datos)
modelo_q045 <- rq(formula = y ~ x, tau = 0.45, data = datos)
modelo_q07 <- rq(formula = y ~ x, tau = 0.7, data = datos)

anova(modelo_q05, modelo_q045)
```

```
## Quantile Regression Analysis of Deviance Table
##
## Model: y ~ x
## Joint Test of Equality of Slopes: tau in {  0.5 0.45  }
##
##   Df Resid Df F value Pr(>F)
## 1  1      199  2.3477 0.1271
```

```
anova(modelo_q05, modelo_q07)
```

```
## Quantile Regression Analysis of Deviance Table
##
## Model: y ~ x
## Joint Test of Equality of Slopes: tau in {  0.5 0.7  }
##
##   Df Resid Df F value   Pr(>F)
## 1  1      199   10.03 0.001783 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

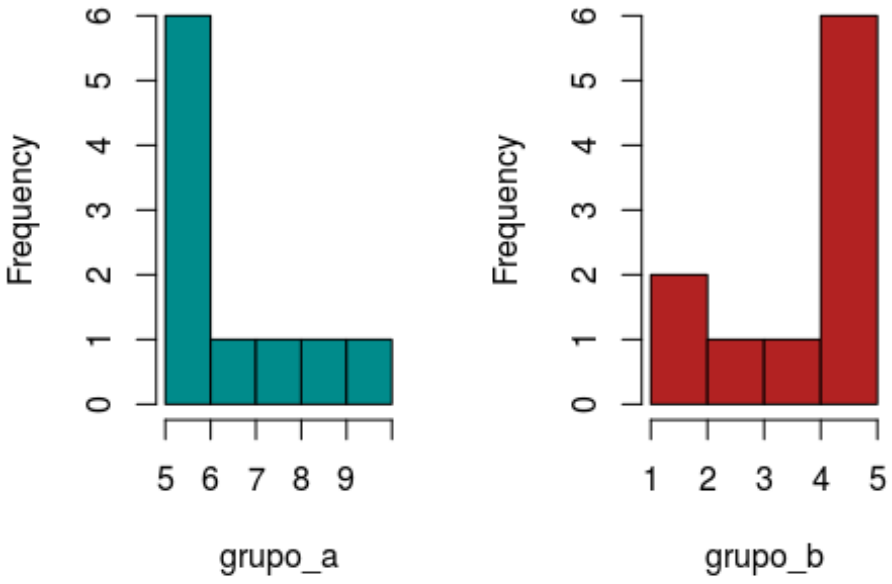
Los test *F* confirman que no hay diferencia significativa entre los coeficientes de regresión de los cuantiles 0.45 y 0.5 pero sí entre 0.5 y 0.7.

Regresión de cuantiles para comparar medianas

Cuando se quiere comparar dos poblaciones pero los datos no siguen una distribución normal, por ejemplo por que presentan colas asimétricas, y además el tamaño muestral es pequeño, suele ser adecuado comparar medianas en lugar de medias. Con frecuencia se recurre al test de Mann–Whitney–Wilcoxon para contrastar la hipótesis nula de dos medianas son iguales. Este test solo puede utilizarse con este fin si se cumple la exigente condición de que la única diferencia entre las poblaciones es su localización, el resto de características (asimetría, dispersión...) tienen que ser idénticas. La regresión de cuantiles permite comparar medianas (cuantil 0.5) sin necesidad de que se cumpla esta condición.

Supóngase que se dispone de las siguientes muestras y de que se desea conocer si existe diferencia significativa entre las poblaciones de origen.

```
grupo_a <- c(5, 5, 5, 5, 5, 5, 7, 8, 9, 10)
grupo_b <- c(1, 2, 3 ,4, 5, 5, 5, 5, 5, 5)
datos <- data.frame(grupo = rep(c("a", "b"), each = 10),
                    valor = c(grupo_a, grupo_b))
par(mfrow = c(1,2))
hist(grupo_a, col = "cyan4", main = "")
hist(grupo_b, col = "firebrick", main = "")
```



En vista de que el tamaño muestral es pequeño y de que ambos grupos muestran una clara asimetría, el *t-test* queda descartado. Una posible alternativa es comparar las medianas. Como la asimetría de los grupos tiene dirección opuesta, el test de *Mann–Whitney–Wilcoxon* no puede emplearse con esta finalidad. En su lugar se recurre a la regresión de cuantiles.

```
modelo_q50 <- rq(valor ~ grupo, tau = 0.5, data = datos)
```

```
## Warning in rq.fit.br(x, y, tau = tau, ...): Solution may be nonunique
```

```
summary(modelo_q50, se = "boot")
```

```
##
## Call: rq(formula = valor ~ grupo, tau = 0.5, data = datos)
##
## tau: [1] 0.5
##
## Coefficients:
##             Value   Std. Error t value Pr(>|t|)
## (Intercept) 5.00000 1.05125    4.75624 0.00016
## grupob      0.00000 1.19201    0.00000 1.00000
```

El resultado no muestra evidencias en contra de la hipótesis nula de que las medianas de ambos grupos son iguales. De hecho, la media de ambos grupos es la misma.

```
median(grupo_a)
```

```
## [1] 5
```

```
median(grupo_b)
```

```
## [1] 5
```

Si se emplea de *Mann–Whitney–Wilcoxon*, debido a la diferencia en forma de las distribuciones, el resultado es significativo, lo que llevaría a conclusiones erróneas.

```
wilcox.test(grupo_a, grupo_b, paired = FALSE)
```

```
## Warning in wilcox.test.default(grupo_a, grupo_b, paired = FALSE): cannot
## compute exact p-value with ties

##
##  Wilcoxon rank sum test with continuity correction
##
## data:  grupo_a and grupo_b
## W = 82, p-value = 0.007196
## alternative hypothesis: true location shift is not equal to 0
```

Regresión de cuantiles y regresión OLS en distribuciones normales

Una de las propiedades de la distribución normal es que el valor de la media y mediana son iguales. Esto significa que el resultado de comparar las medias (regresión por mínimos cuadrados ordinarios) y medianas (regresión del cuantil 0.5) tiende a ser el mismo a medida que aumenta el tamaño muestral. Las pequeñas diferencias se deben a variaciones del muestreo.

```
# Predictor
x <- seq(0, 100, length.out = 100)
# Intercept y pendiente
b_0 <- 6
b_1 <- 0.1
# Error aleatorio normal
set.seed(1)
error <- rnorm(100, mean = 0, sd = 4)
# Variable respuesta
y <- b_0 + b_1*x + error
datos <- data.frame(x, y)

modelo_ols <- lm(y ~ x, data = datos)
summary(modelo_ols)
```

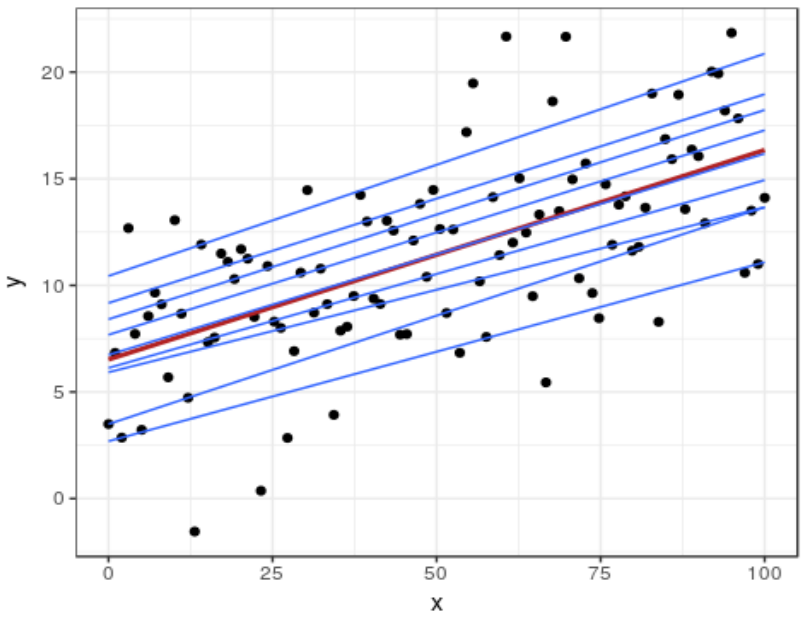
```
##
## Call:
## lm(formula = y ~ x, data = datos)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -9.360 -2.423  0.062  2.341  9.190
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  6.52486    0.71676   9.103 1.07e-14 ***
## x            0.09821    0.01238   7.931 3.56e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.611 on 98 degrees of freedom
## Multiple R-squared:  0.3909, Adjusted R-squared:  0.3847
## F-statistic:  62.9 on 1 and 98 DF,  p-value: 3.558e-12
```

```
modelo_q50 <- rq(y ~ x, tau = 0.5, data = datos)
summary(modelo_q50)
```

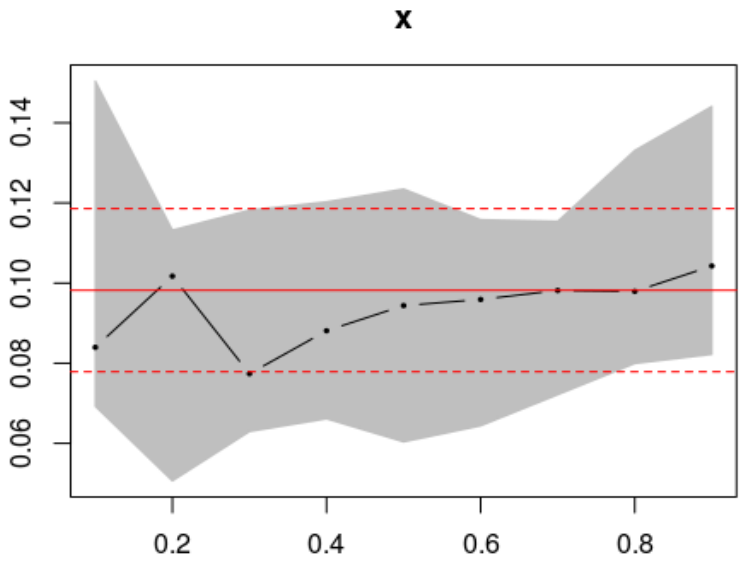
```
## Warning in rq.fit.br(x, y, tau = tau, ci = TRUE, ...): Solution may be
## nonunique

##
## Call: rq(formula = y ~ x, tau = 0.5, data = datos)
##
## tau: [1] 0.5
##
## Coefficients:
##             coefficients lower bd upper bd
## (Intercept) 6.74025      5.03377  8.54760
## x           0.09438      0.06034  0.12350
```

De hecho, si la distribución es normal, la pendiente de la regresión de todos los cuantiles es la misma e igual a la obtenida mediante regresión por mínimos cuadrados ordinarios. Cualquier desviación se debe a variaciones de muestreo.

```
ggplot(data = datos, aes(x = x, y = y)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE, color = "firebrick") +
  geom_quantile(quantiles = c(1:9)/10) +
  theme_bw()
```



```
plot(summary(rq(y ~ x, tau = c(1:9)/10, data = datos)), parm = "x")
```



La superposición de los coeficientes de regresión de cada cuantil con el coeficiente de regresión de la media por OLS y su intervalo del 95% muestra que no hay diferencias significativas entre ellos. De hecho, la comparación entre cualquier par de cuantiles no debería ser significativa si la población es de tipo normal.

```
modelo_q01 <- rq(formula = y ~ x, tau = 0.1, data = datos)
modelo_q05 <- rq(formula = y ~ x, tau = 0.5, data = datos)
modelo_q09 <- rq(formula = y ~ x, tau = 0.9, data = datos)

anova(modelo_q01, modelo_q05)
```

```
## Quantile Regression Analysis of Deviance Table
##
## Model: y ~ x
## Joint Test of Equality of Slopes: tau in {  0.1 0.5  }
##
##   Df Resid Df F value Pr(>F)
## 1  1      199  0.1164 0.7333
```

```
anova(modelo_q01, modelo_q09)
```

```
## Quantile Regression Analysis of Deviance Table
##
## Model: y ~ x
## Joint Test of Equality of Slopes: tau in {  0.1 0.9  }
##
##   Df Resid Df F value Pr(>F)
## 1  1      199  0.2477 0.6192
```

Bibliografía

A gentle introduction to quantile regression for ecologists, Brian S. Cade and Barry R. Noon

http://data.library.virginia.edu/getting-started-with-quantile-regression/

Quantile regression in r: a vignette, Roger Koenker