Taller 2 - Modelo Regresión Cuantílica

AUTHORS
Jhon Tascon
Lino Sinisterra
Juan Chacon

PUBLISHED
September 8, 2025

0. Información general

Este trabajo consiste identificar un caso real donde el interés esté en colas de la distribución (poblaciones muy vulnerables o muy favorecidas) y mostrar cómo la regresión cuantil (QR) revela patrones que el promedio (OLS) oculta.

En este caso, haremos el ejercicio con el dataset “Evaluación de la polución y calidad del aire” de kaggle: https://www.kaggle.com/datasets/mujtabamatin/air-quality-and-pollution-assessment. Este dataset se enfoca en envaluaciones de la calidad de aire en varias regiones del mundo. El dataset contiene 5000 muestras y caputra facotres ambientales y demográficos que influyen en los niveles de polución.

La pregunta de investigación es: ¿Cómo influyen los facotres meteorológicos, de densidad de población y cercanía a zonas industriales los diferentes niveles de CO?

1. Análisis exploratorio

Las variables a analizar son las siguientes:

- Temperatura (°C): Temperatura media de la región
- Humedad (%): Humedad relativa registrada en la región
- Concentración PM2.5 (µg/m³): Niveles de partículas finas
- Concentración de PM10 (µg/m³): Niveles de partículas gruesas -> no se trabajará en el estudio
- Concentración de NO2 (ppb): Niveles de dióxido de nitrógeno -> no se trabajará en el estudio
- Concentración de SO2 (ppb): Niveles de dióxido de azufre -> no se trabajará en el estudio
- Cercanía a Zonas Industriales (km): Distancia a la zona industrial más cercana
- Densidad de población (personas/km²): Número de personas por kilómetro cuadrado en la región
- Calidad del aire: niveles de calidad del aire (Bueno, Moderado, Pobre, Peligroso)
- **Concentración de CO (ppm): Niveles de monóxido de carbono -> Variable objetivo**

Los percentiles a trabajar serán: 0.05, 0.25, 0.50, 0.75 y 0.95

A continuación, se hace un análisis exploratorio de los datos:

```
Visualización de los 10 primeros registros del conjunto de datos:
```

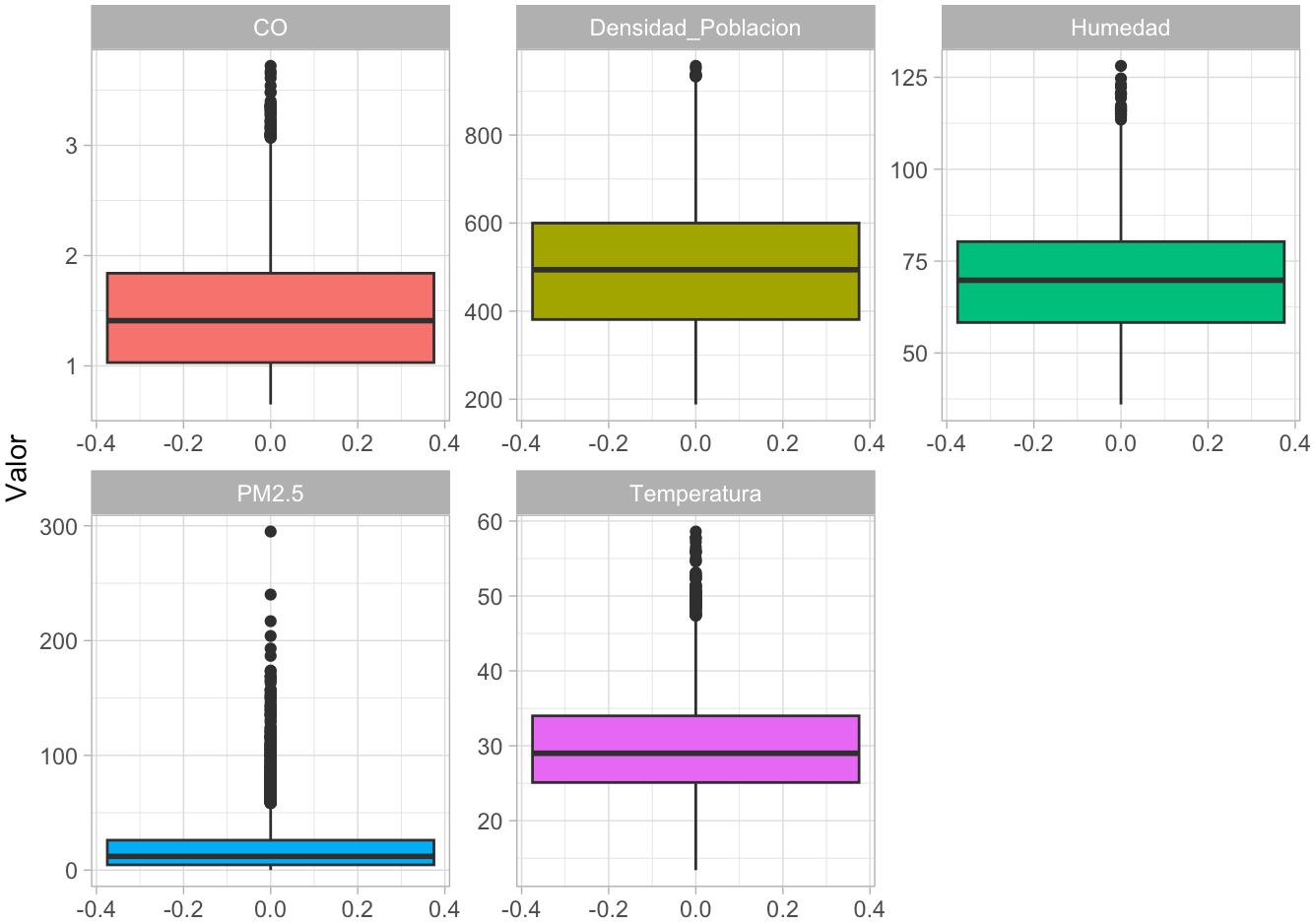
Temperatura	Humedad	PM2.5	CO	Cercania_Areas_Industriales	Densidad_Poblacion	Calidad_Aire
29.8	59.1	5.2	1.72	6.3	319	Moderate

Temperatura	Humedad	PM2.5	CO	Cercania_Areas_Industriales	Densidad_Poblacion	Calidad_Aire
28.3	75.6	2.3	1.64	6.0	611	Moderate
23.1	74.7	26.7	1.63	5.2	619	Moderate
27.1	39.1	6.1	1.15	11.1	551	Good
26.5	70.7	6.9	1.01	12.7	303	Good
39.4	96.6	14.6	1.82	3.1	674	Hazardous
41.7	82.5	1.7	1.80	4.6	735	Poor
31.0	59.6	5.0	1.38	6.3	443	Moderate
29.4	93.8	10.3	2.03	5.4	486	Poor
33.2	80.5	11.1	1.69	4.9	535	Poor

Análisis univariado

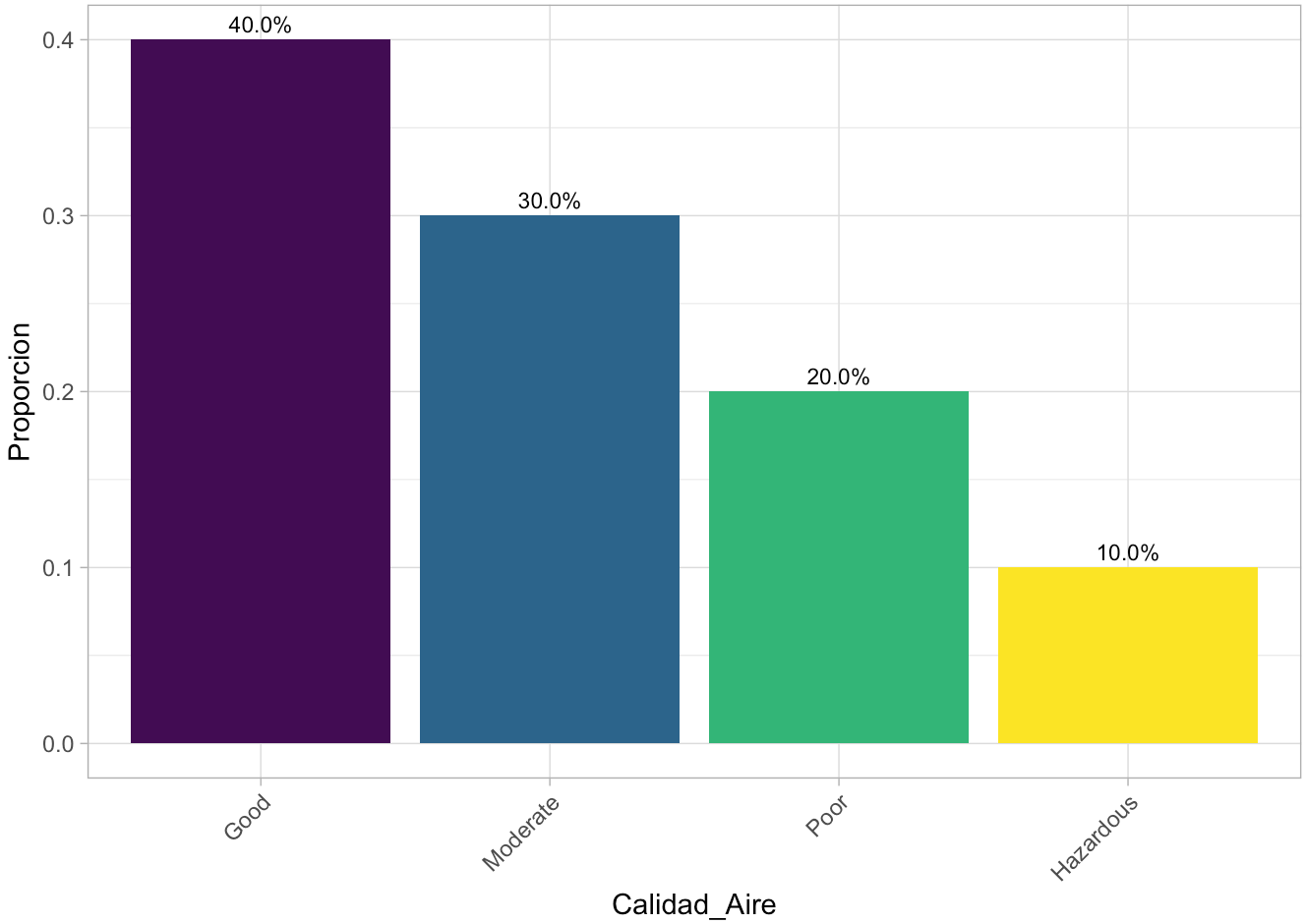
Temperatura	Humedad	PM2.5	CO	Cercania_Areas_Industriales	Densidad_Poblacion	Calidad_Aire
Min. :13.40	Min. : 36.00	Min. : 0.00	Min. :0.65	Min. : 2.500	Min. :188.0	Good :2000
1st Qu.:25.10	1st Qu.: 58.30	1st Qu.: 4.60	1st Qu.:1.03	1st Qu.: 5.400	1st Qu.:381.0	Moderate :1500
Median :29.00	Median : 69.80	Median : 12.00	Median :1.41	Median : 7.900	Median :494.0	Poor :1000
Mean :30.03	Mean : 70.06	Mean : 20.14	Mean :1.50	Mean : 8.425	Mean :497.4	Hazardous: 500
3rd Qu.:34.00	3rd Qu.: 80.30	3rd Qu.: 26.10	3rd Qu.:1.84	3rd Qu.:11.100	3rd Qu.:600.0	NA
Max. :58.60	Max. :128.10	Max. :295.00	Max. :3.72	Max. :25.800	Max. :957.0	NA

Distribución de las variables numéricas



Puntos Clave:

Distribución de las variables categóricas



Puntos Clave:

2. Modelo de regresión cuantílica

2.1. Modelo OLS

```
Call:
lm(formula = CO ~ ., data = df)

Residuals:
     Min       1Q   Median       3Q      Max
-1.40766 -0.11574 -0.00227  0.11109  1.21272

Coefficients:
                            Estimate Std. Error t value Pr(>|t|)
(Intercept)                1.755e+00  3.586e-02  48.958  < 2e-16 ***
Temperatura               -4.776e-04  7.154e-04  -0.668  0.50442
Humedad                   -2.066e-04  2.553e-04  -0.809  0.41846
PM2.5                      3.685e-04  1.417e-04   2.601  0.00933 **
Cercania_Areas_Industriales  1.435e-04  1.573e-03   0.091  0.92728
Densidad_Poblacion         2.566e-05  2.708e-05   0.948  0.34340
Calidad_Aire.L             1.107e+00  1.655e-02  66.934  < 2e-16 ***
```

```
Calidad_Aire.Q              -7.498e-03  7.932e-03  -0.945  0.34456
Calidad_Aire.C               2.035e-03  6.681e-03   0.305  0.76065
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2233 on 4991 degrees of freedom
Multiple R-squared:  0.8331,    Adjusted R-squared:  0.8328
F-statistic:  3113 on 8 and 4991 DF,  p-value: < 2.2e-16
```

Interpretar

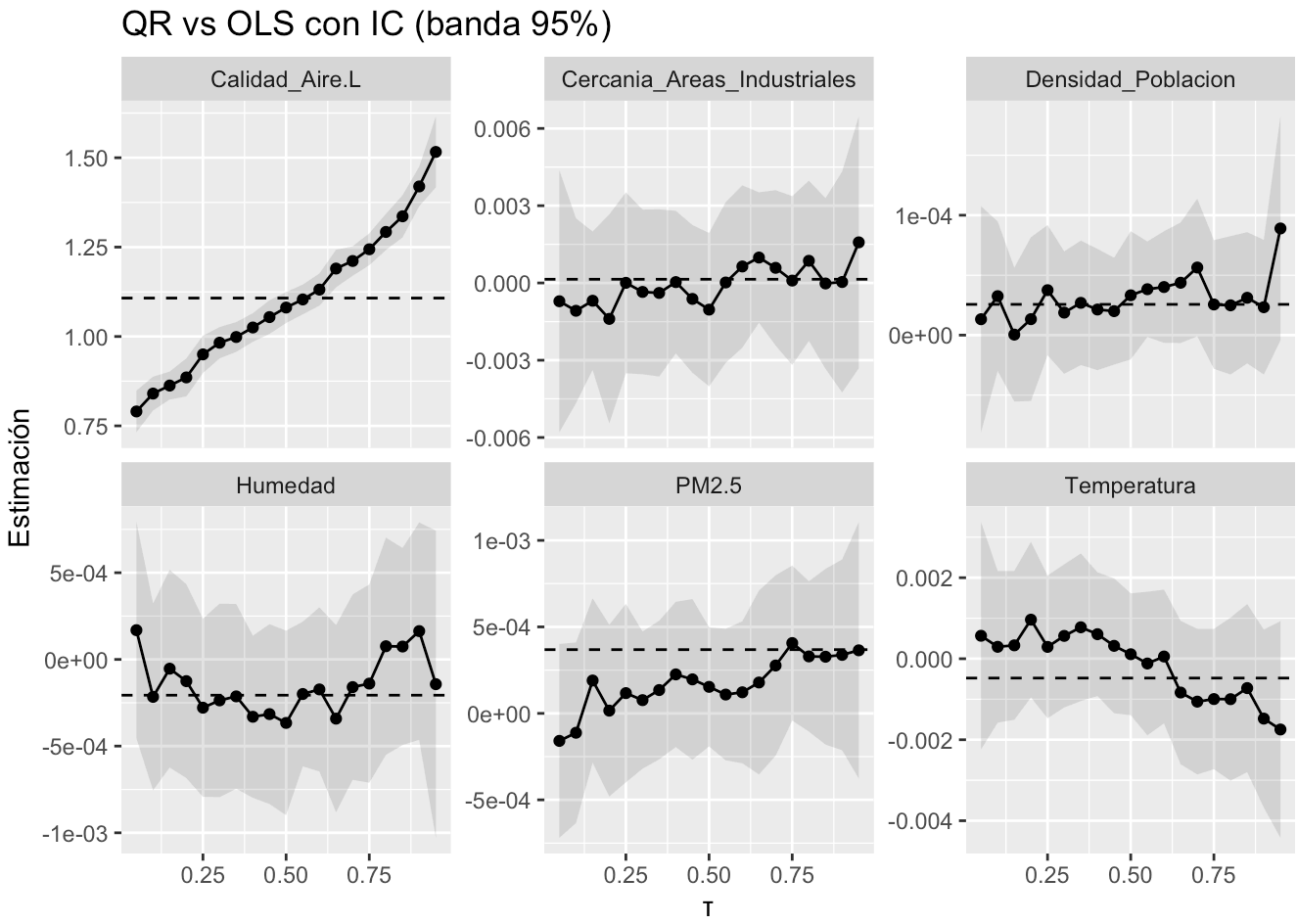
2.2. Modelo QR

```
Call: rq(formula = CO ~ ., tau = taus, data = df)

tau: [1] 0.05

Coefficients:
                            Value    Std. Error t value  Pr(>|t|)
(Intercept)                  1.33734  0.05592   23.91678  0.00000
Temperatura                  0.00057  0.00145    0.38945  0.69696
Humedad                      0.00017  0.00037    0.45331  0.65034
PM2.5                       -0.00016  0.00029   -0.55688  0.57763
Cercania_Areas_Industriales -0.00071  0.00265   -0.26791  0.78878
Densidad_Poblacion           0.00001  0.00005    0.29263  0.76982
Calidad_Aire.L               0.79030  0.03076   25.69342  0.00000
Calidad_Aire.Q               0.02076  0.02125    0.97711  0.32856
Calidad_Aire.C               0.01195  0.02088    0.57260  0.56694

Call: rq(formula = CO ~ ., tau = taus, data = df)

tau: [1] 0.25

Coefficients:
                            Value    Std. Error t value  Pr(>|t|)
(Intercept)                  1.56810  0.04846   32.35881  0.00000
Temperatura                  0.00029  0.00086    0.33827  0.73517
Humedad                     -0.00028  0.00026   -1.08266  0.27901
PM2.5                        0.00012  0.00023    0.51003  0.61006
Cercania_Areas_Industriales  0.00001  0.00189    0.00281  0.99776
Densidad_Poblacion           0.00004  0.00003    1.33149  0.18309
Calidad_Aire.L               0.94994  0.02710   35.04766  0.00000
Calidad_Aire.Q              -0.01108  0.01848   -0.59982  0.54865
Calidad_Aire.C               0.00454  0.01278    0.35509  0.72254

Call: rq(formula = CO ~ ., tau = taus, data = df)

tau: [1] 0.5

Coefficients:
```

```
                            Value    Std. Error t value  Pr(>|t|)
(Intercept)                  1.74914  0.03233   54.11050  0.00000
Temperatura                  0.00011  0.00066    0.16770  0.86683
Humedad                     -0.00037  0.00025   -1.44310  0.14906
PM2.5                        0.00015  0.00020    0.76901  0.44192
Cercania_Areas_Industriales -0.00104  0.00148   -0.69794  0.48525
Densidad_Poblacion           0.00003  0.00002    1.41324  0.15765
Calidad_Aire.L               1.08095  0.02272   47.58143  0.00000
Calidad_Aire.Q              -0.01268  0.01147   -1.10600  0.26878
Calidad_Aire.C              -0.00713  0.00926   -0.76978  0.44147

Call: rq(formula = CO ~ ., tau = taus, data = df)

tau: [1] 0.75

Coefficients:
                            Value    Std. Error t value  Pr(>|t|)
(Intercept)                  1.92954  0.04285   45.02540  0.00000
Temperatura                 -0.00100  0.00090   -1.10803  0.26790
Humedad                     -0.00014  0.00029   -0.47295  0.63627
PM2.5                        0.00041  0.00023    1.75747  0.07890
Cercania_Areas_Industriales  0.00009  0.00140    0.06635  0.94710
Densidad_Poblacion           0.00003  0.00003    0.89686  0.36984
Calidad_Aire.L               1.24396  0.02209   56.32140  0.00000
Calidad_Aire.Q              -0.01820  0.01561   -1.16590  0.24371
Calidad_Aire.C              -0.00574  0.01247   -0.46044  0.64522

Call: rq(formula = CO ~ ., tau = taus, data = df)

tau: [1] 0.95

Coefficients:
                            Value    Std. Error t value  Pr(>|t|)
(Intercept)                  2.16689  0.06207   34.91161  0.00000
Temperatura                 -0.00175  0.00137   -1.27253  0.20324
Humedad                     -0.00014  0.00047   -0.30250  0.76228
PM2.5                        0.00036  0.00038    0.95632  0.33896
Cercania_Areas_Industriales  0.00158  0.00263    0.59883  0.54931
Densidad_Poblacion           0.00009  0.00005    1.92242  0.05461
Calidad_Aire.L               1.51622  0.04713   32.17250  0.00000
Calidad_Aire.Q               0.02164  0.03564    0.60727  0.54370
Calidad_Aire.C               0.01132  0.02388    0.47409  0.63546
```

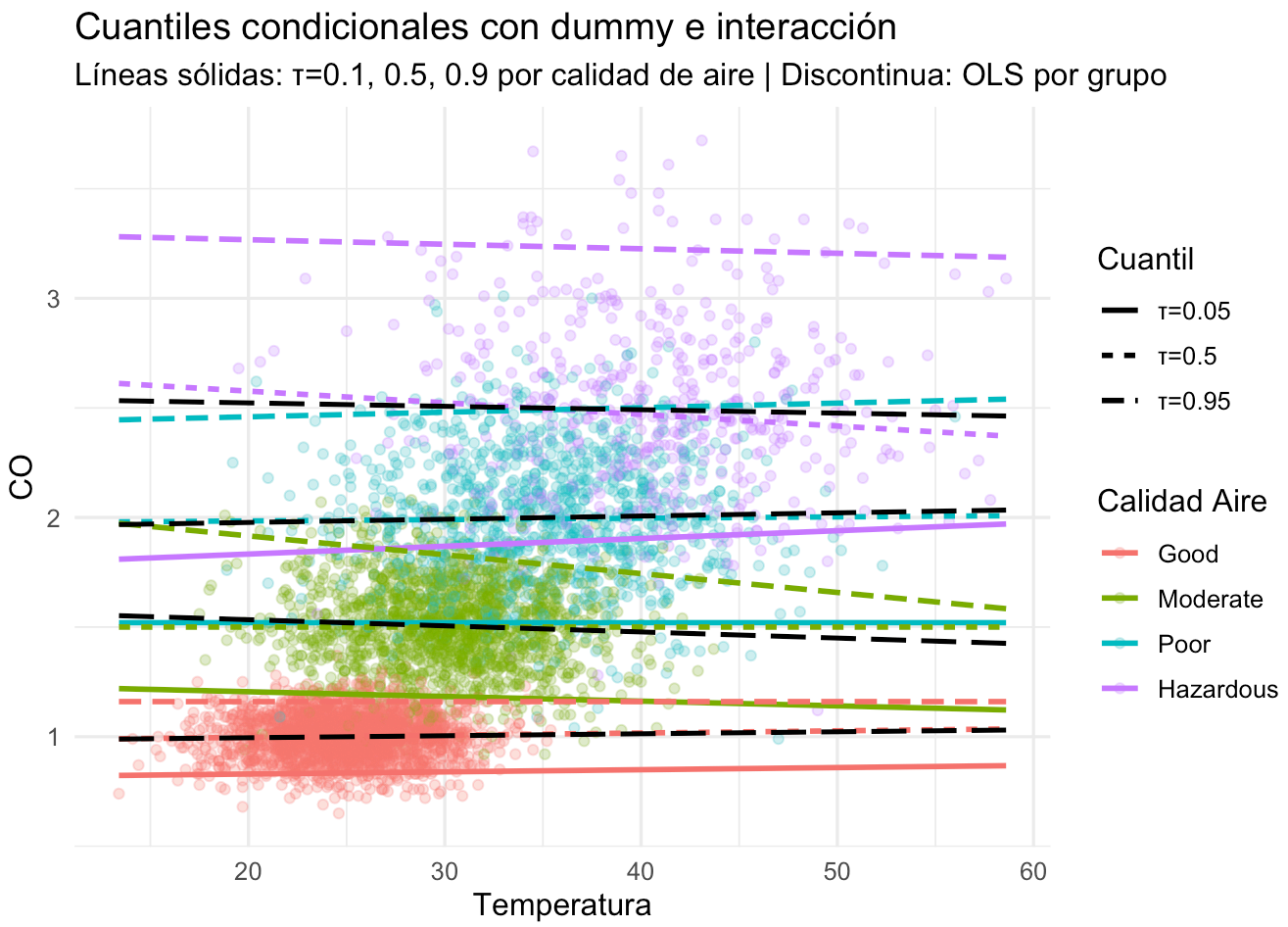
Interpretar

2.3. Resultados gráficos



Interpretar

2.4. Modelo QR con interacción



Interpretar