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 Áreas 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 (ppb): 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:
```

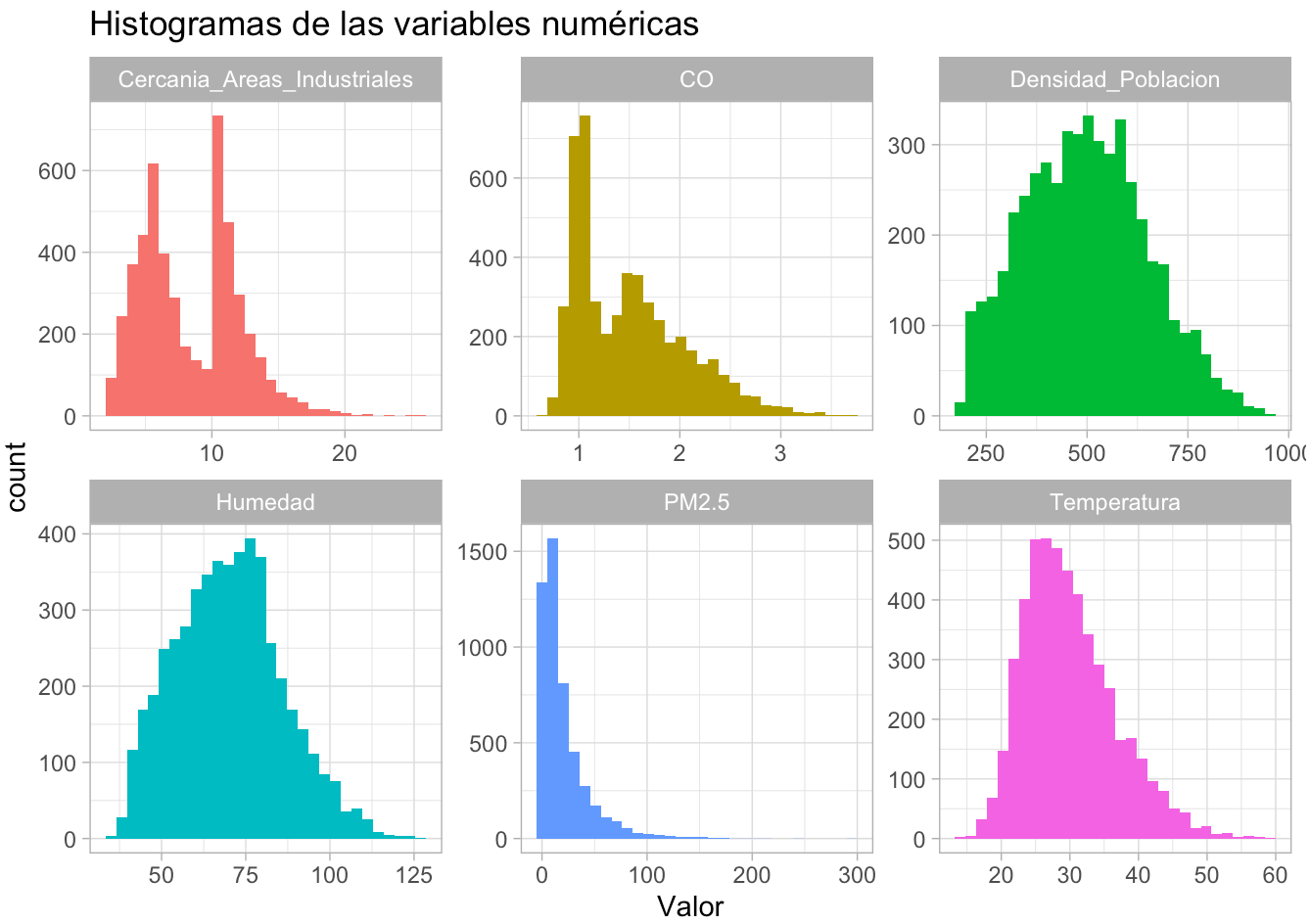
Temperatura	Humedad	PM2.5	CO	Cercania_Areas_Industriales	Densidad_Poblacion	Calidad_Aire
29.8	59.1	5.2	1.72	6.3	319	Moderate

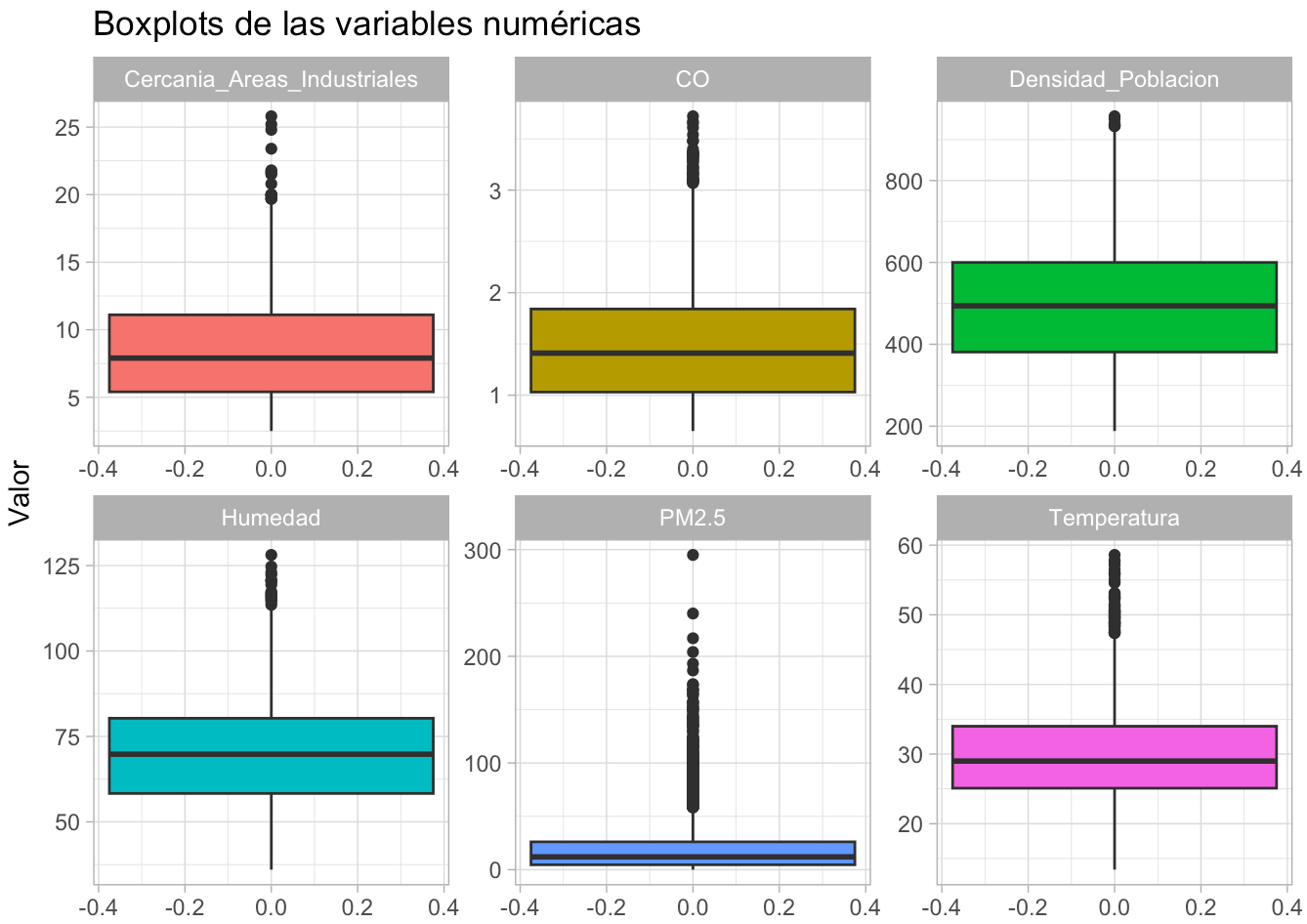
Temperatura	Humedad	PM2.5	CO	Cercania_Areas_Industriales	Densidad_Poblacion	Calidad_Aire
28.3	75.6	2.3	1.64	6.0	611	Moderate
23.1	74.7	26.7	1.63	5.2	619	Moderate
27.1	39.1	6.1	1.15	11.1	551	Good
26.5	70.7	6.9	1.01	12.7	303	Good
39.4	96.6	14.6	1.82	3.1	674	Hazardous
41.7	82.5	1.7	1.80	4.6	735	Poor
31.0	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

1.1. 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

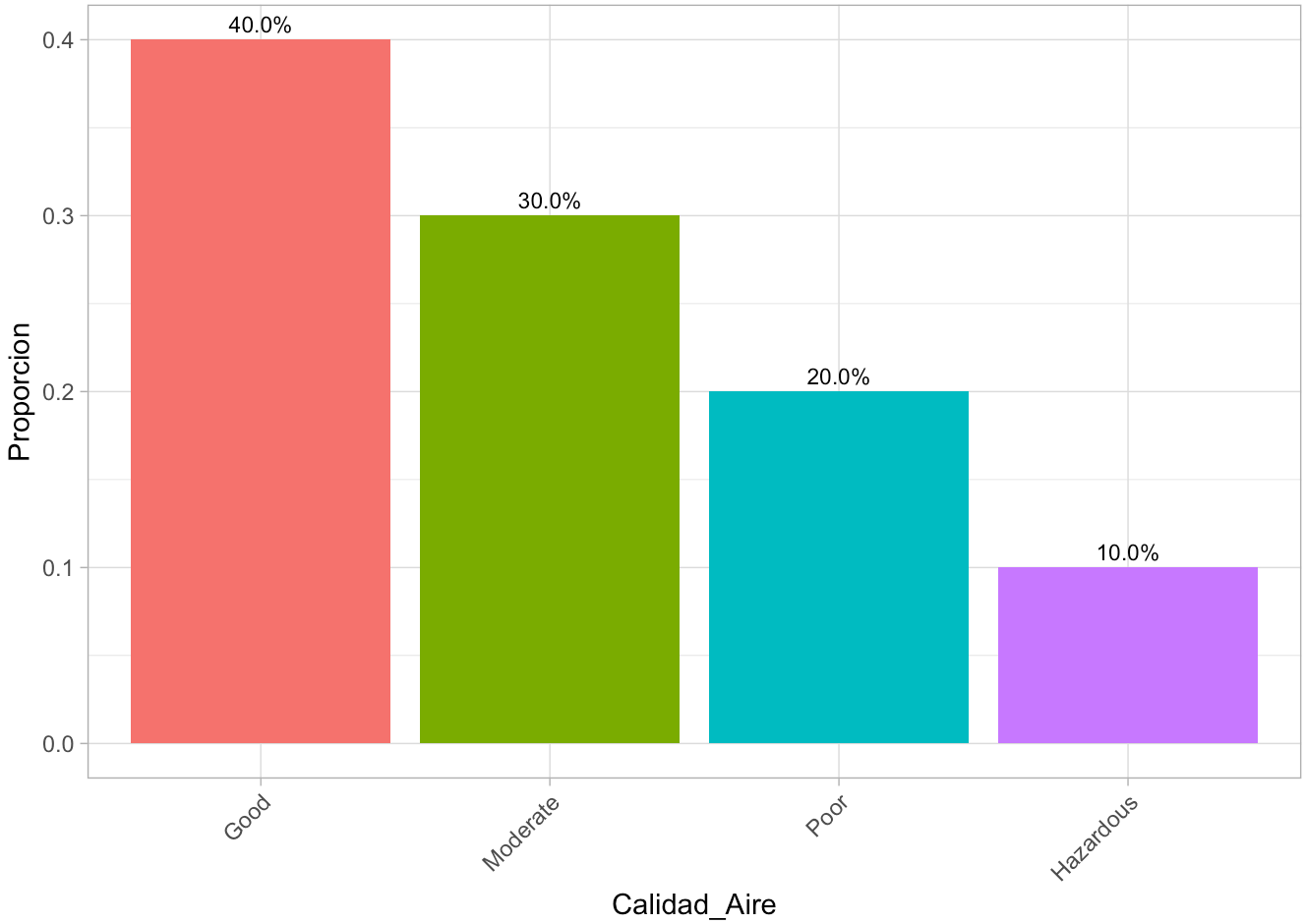
1.1.1. Distribución de las variables numéricas


Histogramas de las variables numéricas
Cercania_Areas_Industriales
CO
Densidad_Poblacion
Humedad
PM2.5
Temperatura
count
Valor



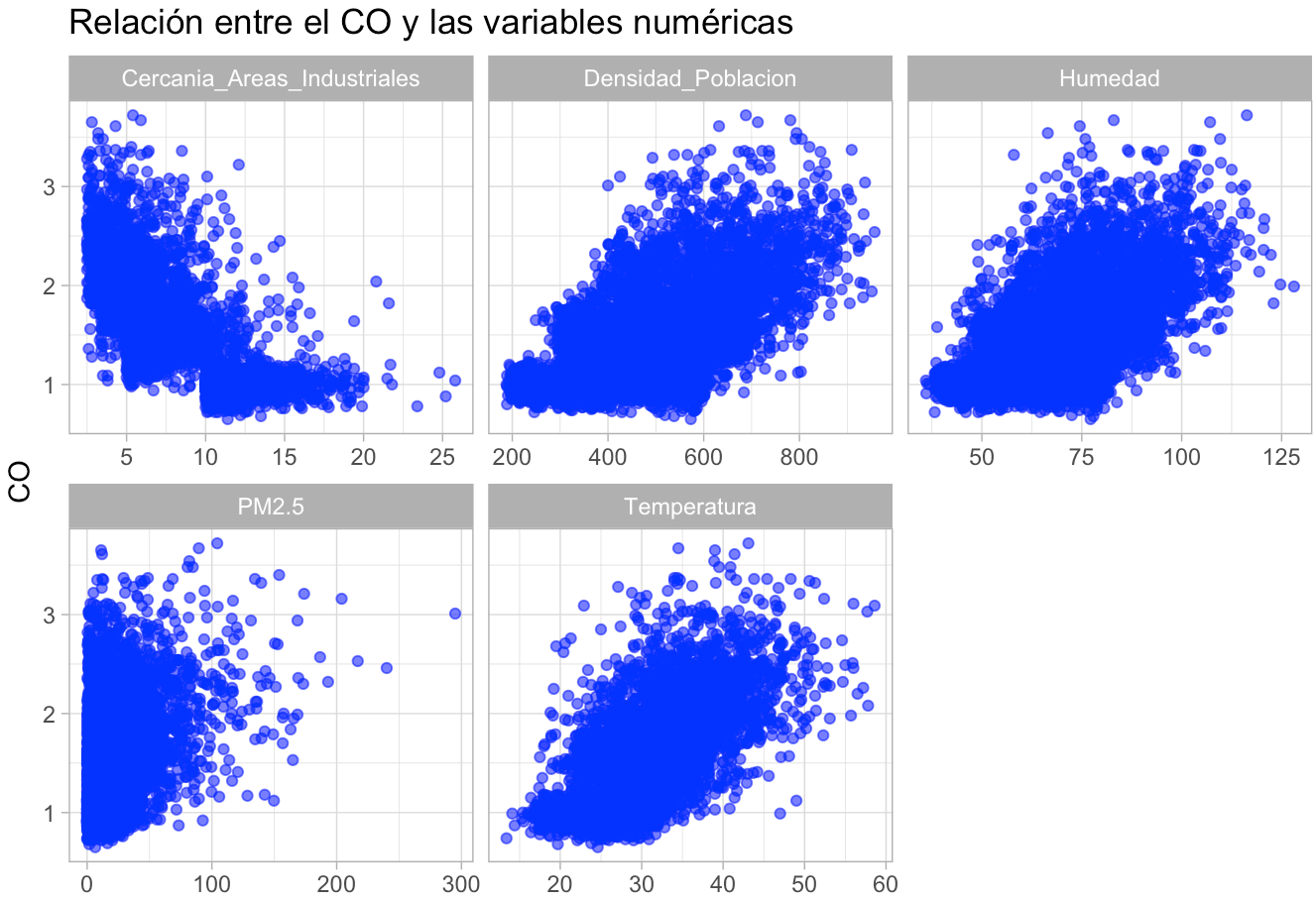
En general se presentan algunas distribuciones sesgadas, especialmente en el PM2.5 (partículas finas) y la variable objetivo (CO) presenta una distribución bimodal. Las demás variables siguen una distribución cercana a la normal.

1.1.2. Distribución de las variables categóricas

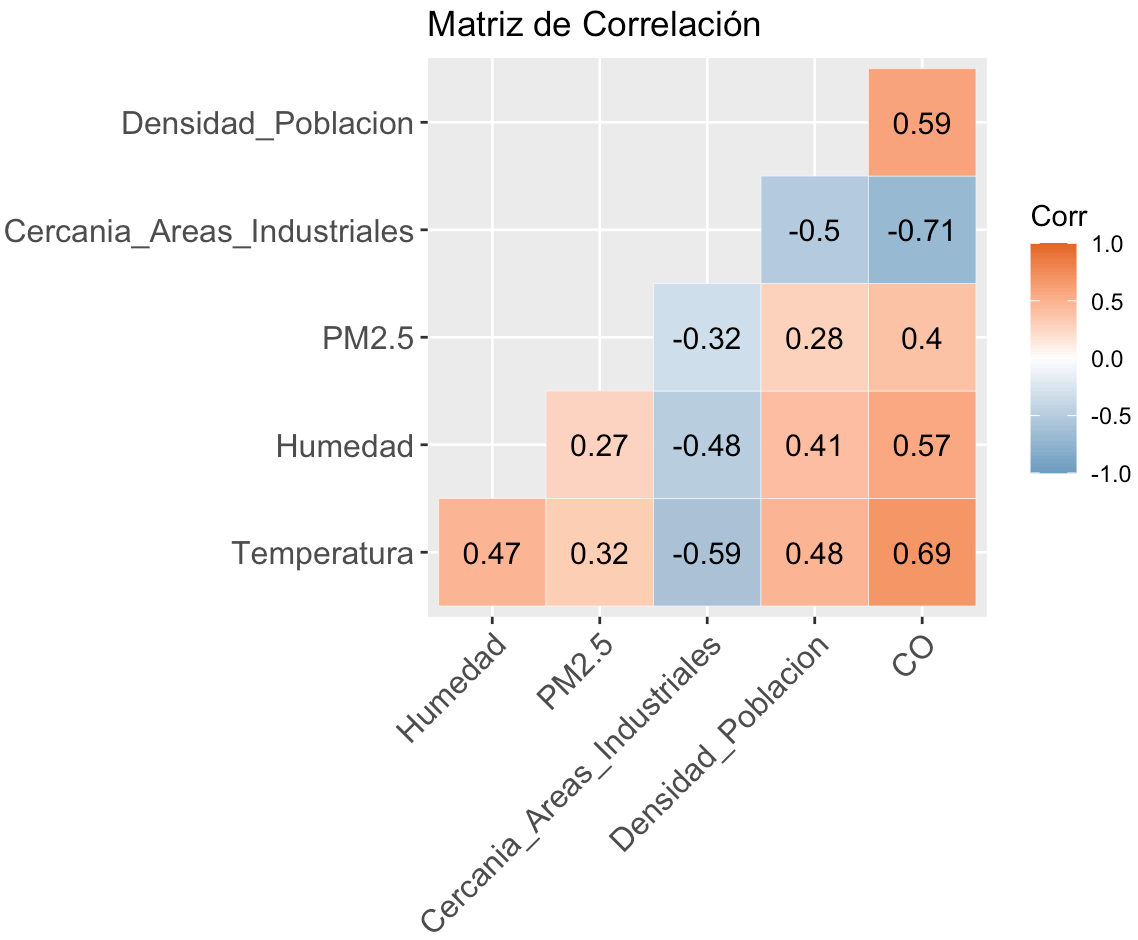


En cuánto a la variable de la calidad del aire, la mayor proporción la tiene la calidad del aire buena, seguida por moderado, luego pobre y por último calidad de aire peligroso.

1.2. Análisis bivariado



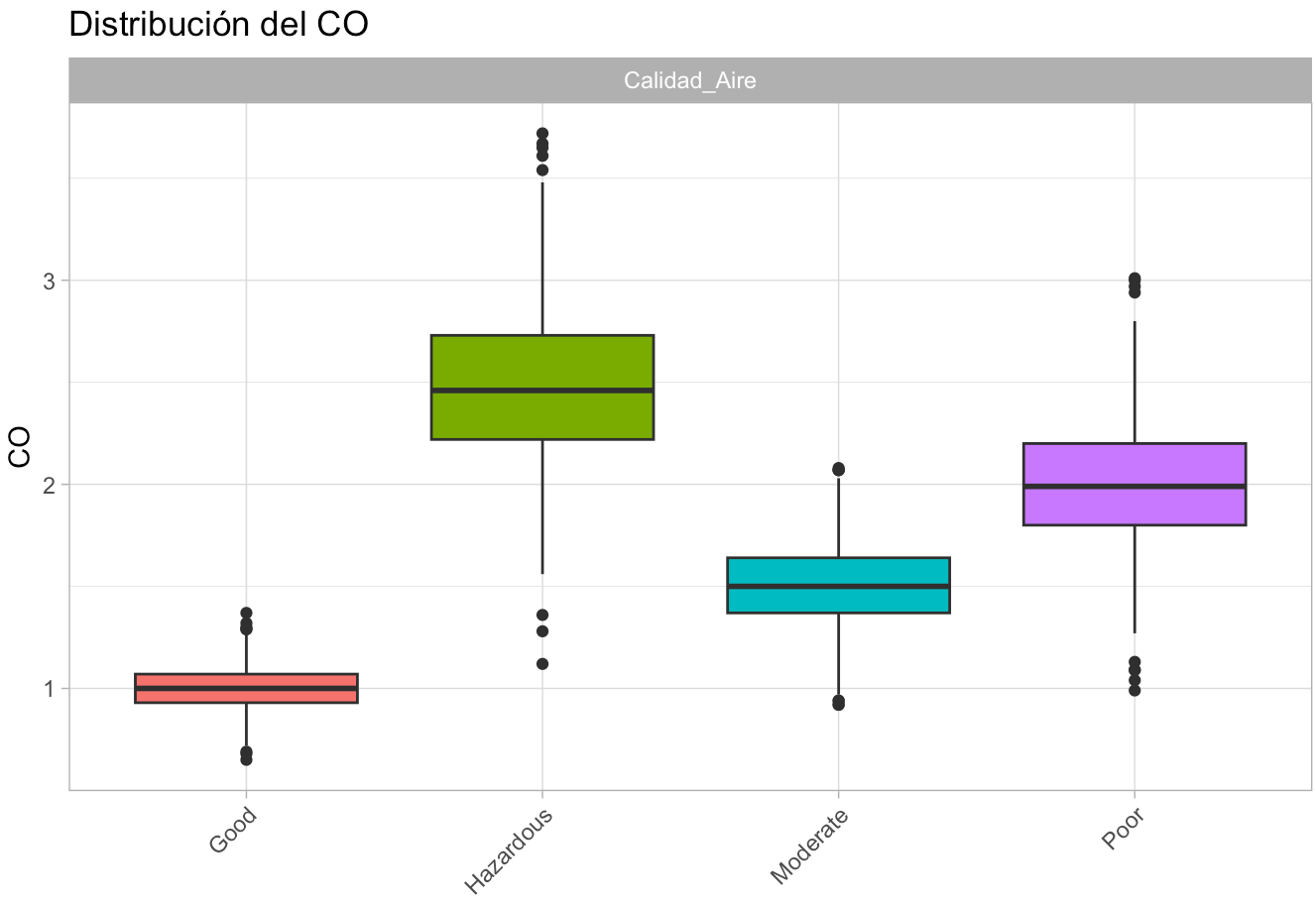
Variables numéricas vs CO



El CO con las demás variables numéricas presentan relaciones medias-altas, especialmente con la temperatura (relación positiva), húmedad (positiva), densidad de población (positiva) y cercanía a áreas industriales (negativa).

- **Temperatura**: En cuanto a la relación con la temperatura, se puede ver que a mayor temperatura, las concentraciones de CO tienden a ser mayores.
- **Húmedad**: muy similar a la temperatura en cuanto a su relación. Sin embgargo, en zonas medias de húmedad (60%-90%) hay una mayor variabilidad de concentración de CO.
- **Densidad de población**: comportamiento similar a los anteriores. Sin embargo, en zonas de baja concentración de población (alrededor de 200 a 300 personas/km), la concentración de CO tiende a ser constante, moviéndose entre 0.5 y 1.5 ppb.
- **Cercanía a áreas industriales**: en este caso, la relación es inversa: entre más cercano esté una región a zonas industriales, la concentración de CO tiende a ser mayor. La relación no es del todo lineal.

Variables categóricas vs CO



En cuanto a la calidad del aire, se ve que el CO influye en gran medida: en zonas donde la calidad del aire es buena o moderada, los niveles de CO son menores, mientras en regiones con calidad de aire pobre o peligroso, el nivel de CO es mayor.

2. Modelo de regresión cuantílica

2.1. Modelo OLS

```
Call:
lm(formula = form, data = df)

Residuals:
     Min       1Q   Median       3Q      Max
-1.42083 -0.11555 -0.00125  0.11103  1.19823

Coefficients:
                                Estimate Std. Error t value Pr(>|t|)
(Intercept)                    8.967e-01  4.954e-02  18.101  < 2e-16 ***
Temperatura                   -4.611e-04  7.148e-04  -0.645  0.51892
Humedad                       -2.316e-04  2.552e-04  -0.907  0.36420
PM2.5                          3.579e-04  1.416e-04   2.527  0.01154 *
Cercania_Areas_Industriales   -1.162e-02  4.248e-03  -2.736  0.00624 **
```

```
I(log(Cercania_Areas_Industriales))  1.025e-01  3.438e-02   2.981  0.00289 **
Densidad_Poblacion                   2.691e-05  2.706e-05   0.994  0.32009
Calidad_AireModerate                 5.033e-01  1.227e-02  41.031  < 2e-16 ***
Calidad_AirePoor                     1.006e+00  1.737e-02  57.911  < 2e-16 ***
Calidad_AireHazardous                1.506e+00  2.371e-02  63.509  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2231 on 4990 degrees of freedom
Multiple R-squared:  0.8333,    Adjusted R-squared:  0.833
F-statistic:  2773 on 9 and 4990 DF,  p-value: < 2.2e-16
```

1. Temperatura: -0.00046 Diferencia no significativa (p = 0.504; IC95% cruza 0). Con estos datos, no hay evidencia de efecto promedio de la temperatura sobre el CO.

2. Humedad: -0.00023 Diferencia no significativa (p = 0.418; IC95% cruza 0). Con estos datos, no hay evidencia de efecto promedio de la humedad sobre el CO.

3. Partículas finas (PM2.5): 0.00036 Cada nivel adicional de partículas finas en el aire, se asocia con +0.00036 unidades de CO (p. < 0.05; IC95% ≈ [0.0008, 0.00064]), manteniendo lo demás constante.

4. Cercanía a áreas industriales: -0.012 y log(Cercania_Areas_Industriales): 0.103 La cercanía a zonas industriales aumenta los niveles de CO en las regiones. Sin embargo, entre más lejos de zonas industriales, los niveles de CO bajan pero no de la misma manera. El efecto marginal es el siguiente:

$$\frac{\partial CO}{\partial CercaniaAreasIndustriales} = -0.012 + \frac{0.103}{CercaniaAreasIndustriales}$$

Si está muy cerca a áreas industriales, si la distancia aumenta 1 km aumenta 0.091 los niveles de CO, alejarse 5 km aumenta los niveles de CO en 0.0086, lejos de areás industriales, los niveles de CO disminuyen en -0.0109 si la distancia aumenta a 100 km.

Ambos son significativos (p. < 0.01)

5. Densidad de población: -0.000027 Diferencia no significativa (p = 0.32; IC95% cruza 0). Con estos datos, no hay evidencia de efecto promedio de la densidad de población sobre el CO.

6. Calidad del aire Interpretar

2.2. Modelo QR

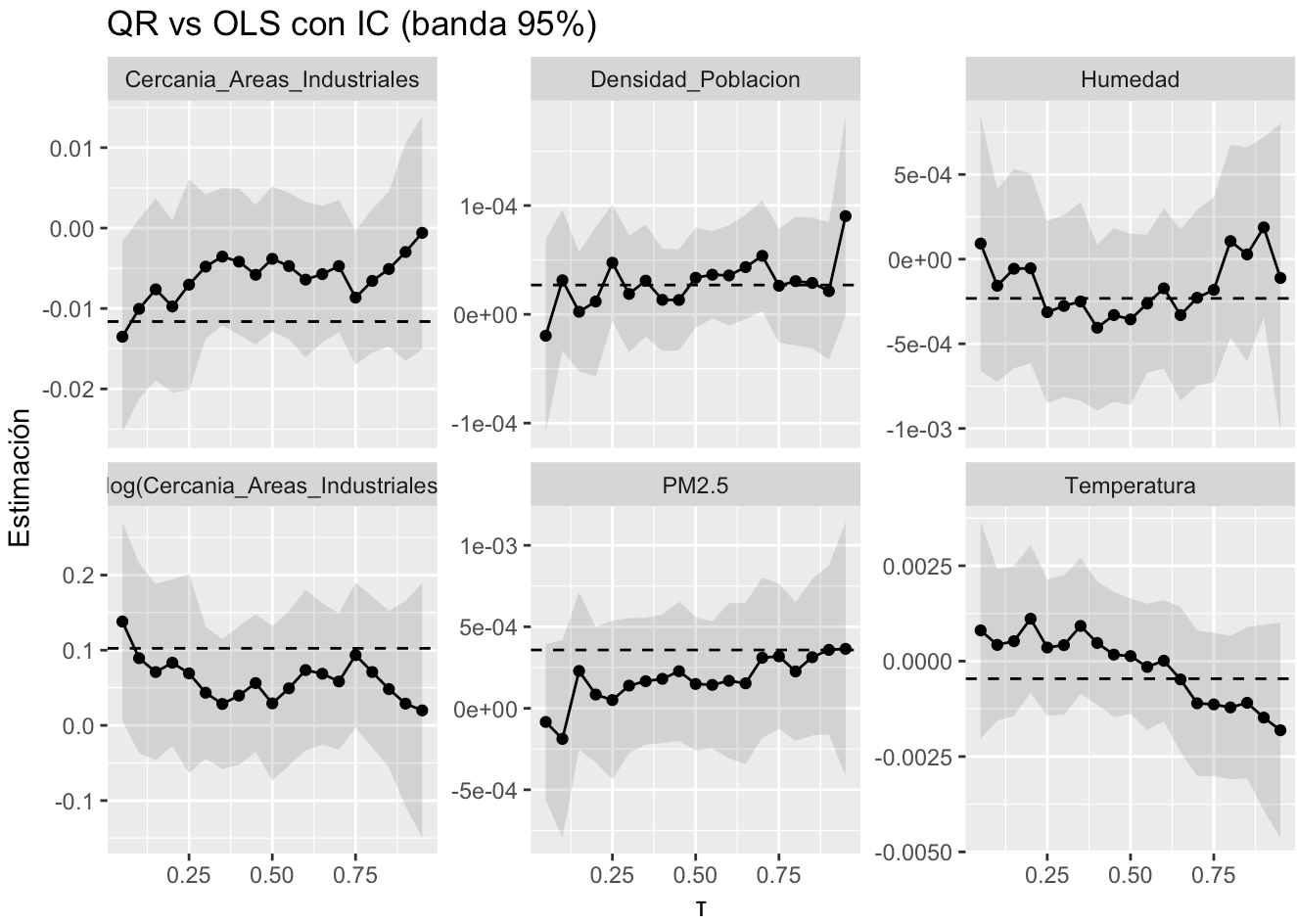
```
Call: rq(formula = form, tau = taus, data = df)

tau: [1] 0.05

Coefficients:
                                Value    Std. Error t value  Pr(>|t|)
(Intercept)                      0.63730  0.11252    5.66383  0.00000
Temperatura                      0.00081  0.00153    0.52557  0.59921
Humedad                          0.00009  0.00036    0.25962  0.79517
```

```
PM2.5                              -0.00008  0.00029   -0.29322  0.76937
Cercania_Areas_Industriales        -0.01351  0.00626   -2.15677  0.03107
I(log(Cercania_Areas_Industriales))  0.13814  0.06903    2.00128  0.04542
Densidad_Poblacion                 -0.00002  0.00004   -0.43972  0.66016
Calidad_AireModerate                0.35961  0.02355   15.26879  0.00000
Calidad_AirePoor                    0.71069  0.03502   20.29475  0.00000
Calidad_AireHazardous               1.11754  0.05507   20.29296  0.00000

Call: rq(formula = form, tau = taus, data = df)

tau: [1] 0.25

Coefficients:
                                   Value    Std. Error t value  Pr(>|t|)
(Intercept)                         0.83442  0.10102    8.26020  0.00000
Temperatura                         0.00036  0.00092    0.38996  0.69658
Humedad                            -0.00031  0.00026   -1.18634  0.23554
PM2.5                               0.00005  0.00024    0.21248  0.83174
Cercania_Areas_Industriales        -0.00703  0.00600   -1.17258  0.24102
I(log(Cercania_Areas_Industriales))  0.06917  0.06375    1.08514  0.27791
Densidad_Poblacion                  0.00005  0.00003    1.75112  0.07999
Calidad_AireModerate                0.44299  0.01236   35.83794  0.00000
Calidad_AirePoor                    0.86837  0.02544   34.13777  0.00000
Calidad_AireHazardous               1.28625  0.04232   30.38979  0.00000

Call: rq(formula = form, tau = taus, data = df)

tau: [1] 0.5

Coefficients:
                                   Value    Std. Error t value  Pr(>|t|)
(Intercept)                         0.97887  0.07652   12.79230  0.00000
Temperatura                         0.00013  0.00079    0.16777  0.86677
Humedad                            -0.00036  0.00026   -1.35952  0.17404
PM2.5                               0.00015  0.00019    0.78159  0.43449
Cercania_Areas_Industriales        -0.00382  0.00451   -0.84656  0.39728
I(log(Cercania_Areas_Industriales))  0.02924  0.04879    0.59930  0.54900
Densidad_Poblacion                  0.00003  0.00002    1.40398  0.16039
Calidad_AireModerate                0.49335  0.01418   34.80046  0.00000
Calidad_AirePoor                    0.98726  0.02071   47.66609  0.00000
Calidad_AireHazardous               1.45891  0.03395   42.96958  0.00000

Call: rq(formula = form, tau = taus, data = df)

tau: [1] 0.75

Coefficients:
                                   Value    Std. Error t value  Pr(>|t|)
(Intercept)                         0.96649  0.08796   10.98829  0.00000
Temperatura                        -0.00113  0.00089   -1.27518  0.20230
Humedad                            -0.00018  0.00027   -0.68309  0.49458
```

```
PM2.5                                0.00032  0.00024    1.32385  0.18561
Cercania_Areas_Industriales         -0.00865  0.00474   -1.82363  0.06827
I(log(Cercania_Areas_Industriales))  0.09348  0.05534    1.68928  0.09123
Densidad_Poblacion                   0.00003  0.00003    0.90730  0.36429
Calidad_AireModerate                 0.58201  0.01392   41.80527  0.00000
Calidad_AirePoor                     1.15693  0.02652   43.63298  0.00000
Calidad_AireHazardous                1.70920  0.03807   44.89509  0.00000

Call: rq(formula = form, tau = taus, data = df)

tau: [1] 0.95

Coefficients:
                                    Value    Std. Error t value  Pr(>|t|)
(Intercept)                          1.13414  0.14272    7.94673  0.00000
Temperatura                         -0.00181  0.00135   -1.34661  0.17817
Humedad                             -0.00011  0.00046   -0.24178  0.80896
PM2.5                                0.00036  0.00039    0.94250  0.34598
Cercania_Areas_Industriales         -0.00059  0.00794   -0.07458  0.94055
I(log(Cercania_Areas_Industriales))  0.01992  0.09141    0.21793  0.82749
Densidad_Poblacion                   0.00009  0.00004    2.10899  0.03499
Calidad_AireModerate                 0.66627  0.02786   23.91254  0.00000
Calidad_AirePoor                     1.33356  0.04372   30.50395  0.00000
Calidad_AireHazardous                2.04209  0.08468   24.11616  0.00000
```

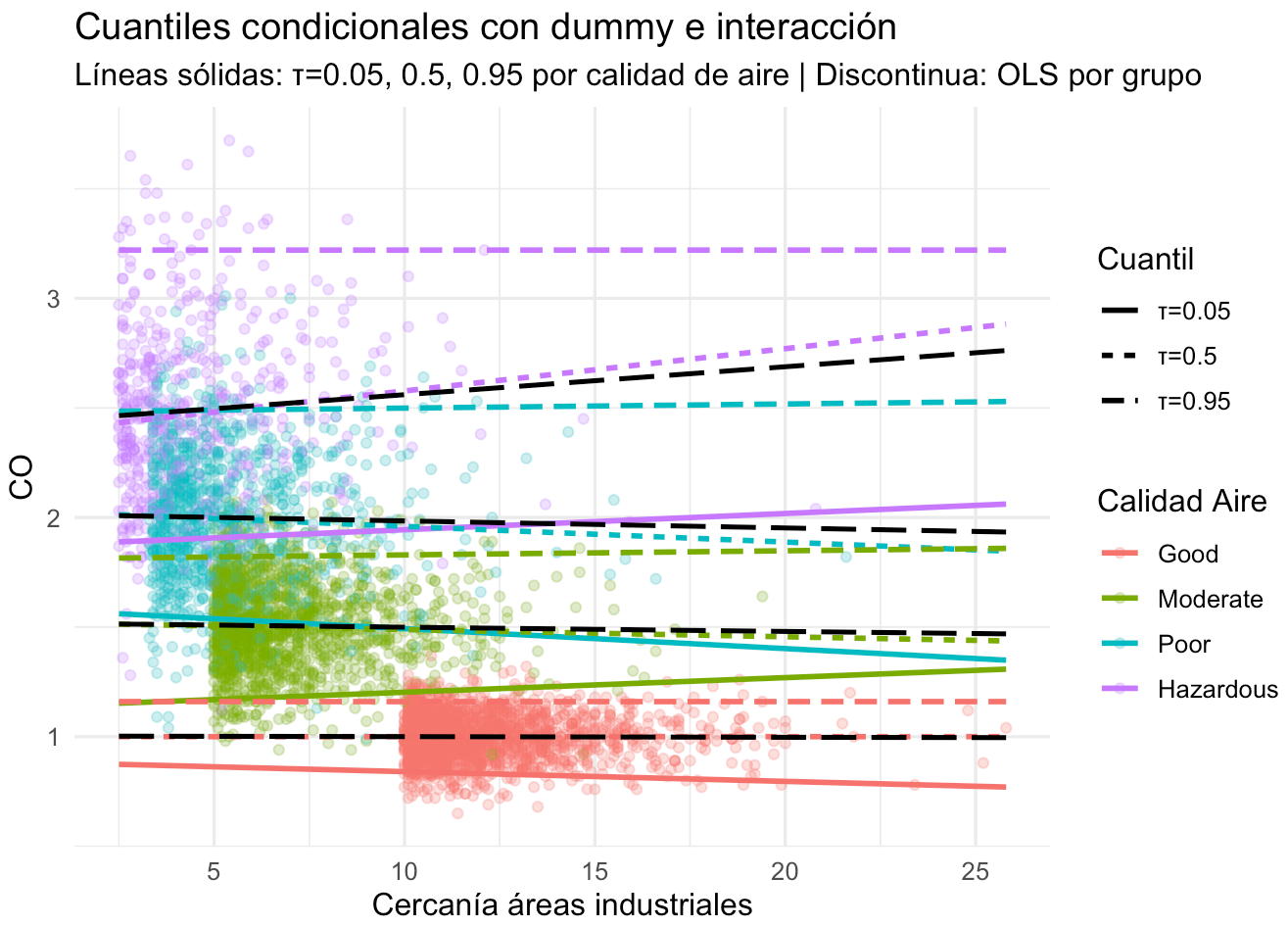
Interpretar

2.3. Resultados gráficos



Interpretar

2.4. Modelo QR con interacción



Interpretar