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 (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:
```

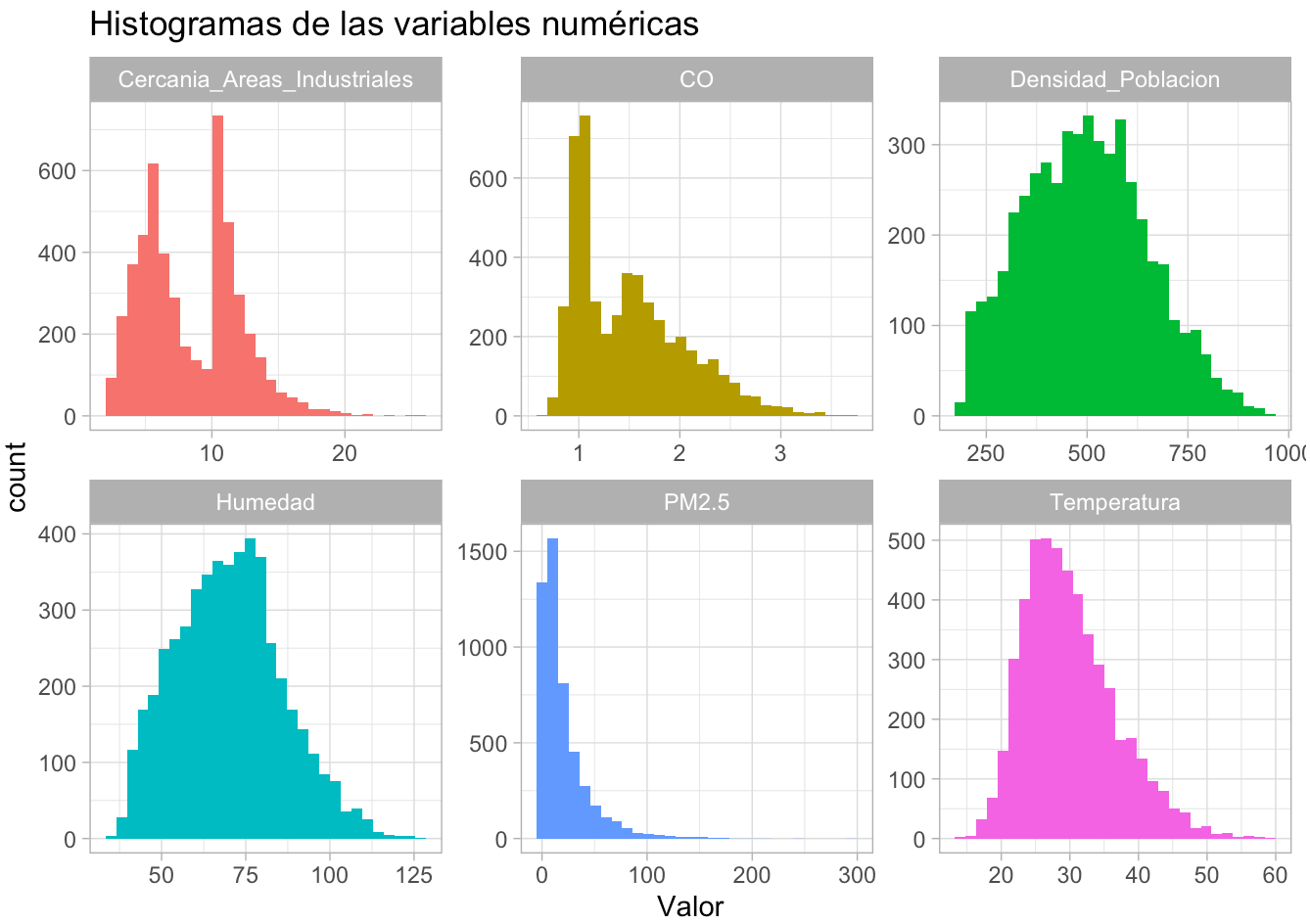
Temperatura	Humedad	PM2.5	CO	Cercania_Areas_Industriales	Densidad_Poblacion	Calidad_Aire
29.8	59.1	5.2	1.72	6.3	319	Moderate

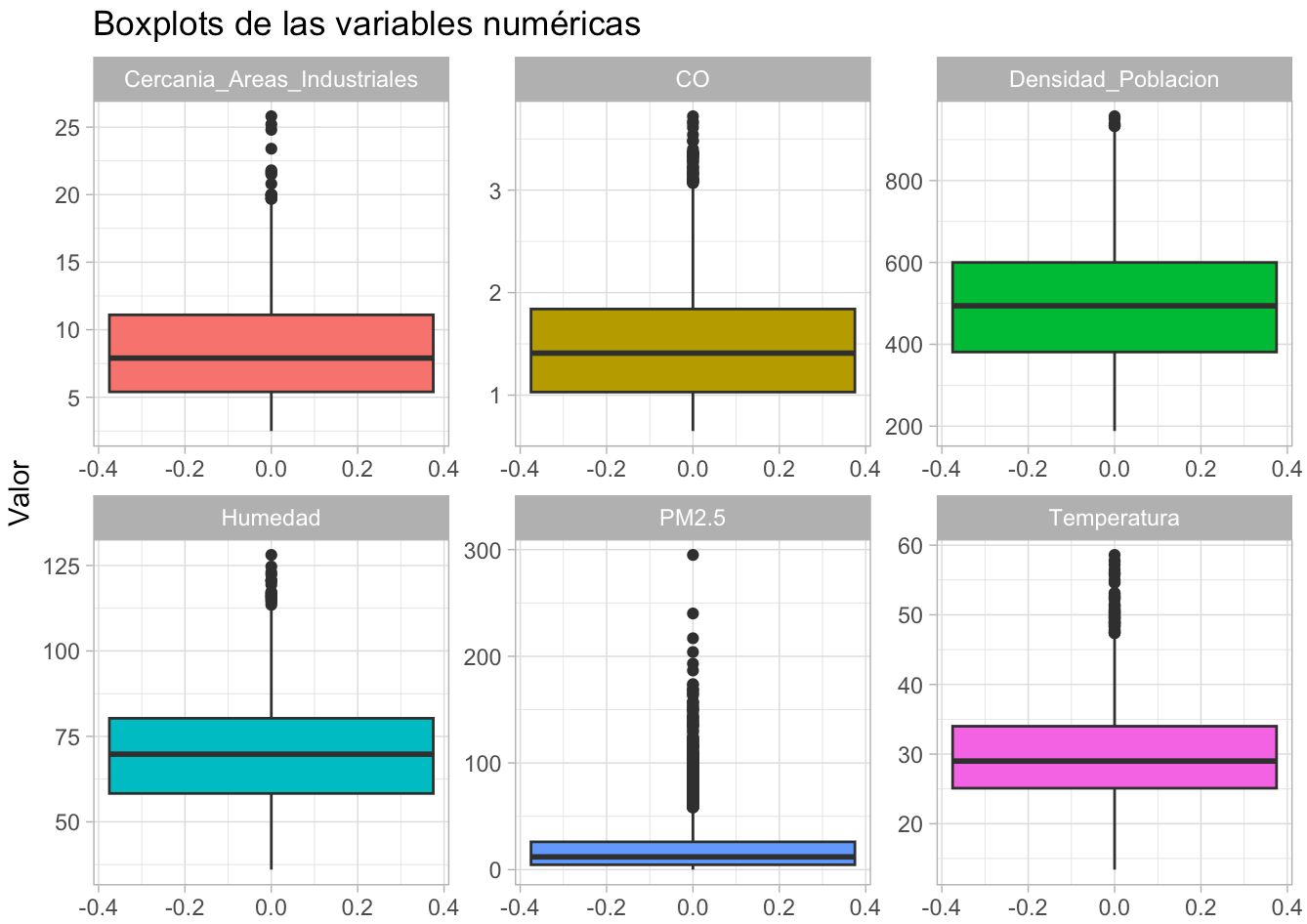
Temperatura	Humedad	PM2.5	CO	Cercania_Areas_Industriales	Densidad_Poblacion	Calidad_Aire
28.3	75.6	2.3	1.64	6.0	611	Moderate
23.1	74.7	26.7	1.63	5.2	619	Moderate
27.1	39.1	6.1	1.15	11.1	551	Good
26.5	70.7	6.9	1.01	12.7	303	Good
39.4	96.6	14.6	1.82	3.1	674	Hazardous
41.7	82.5	1.7	1.80	4.6	735	Poor
31.0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

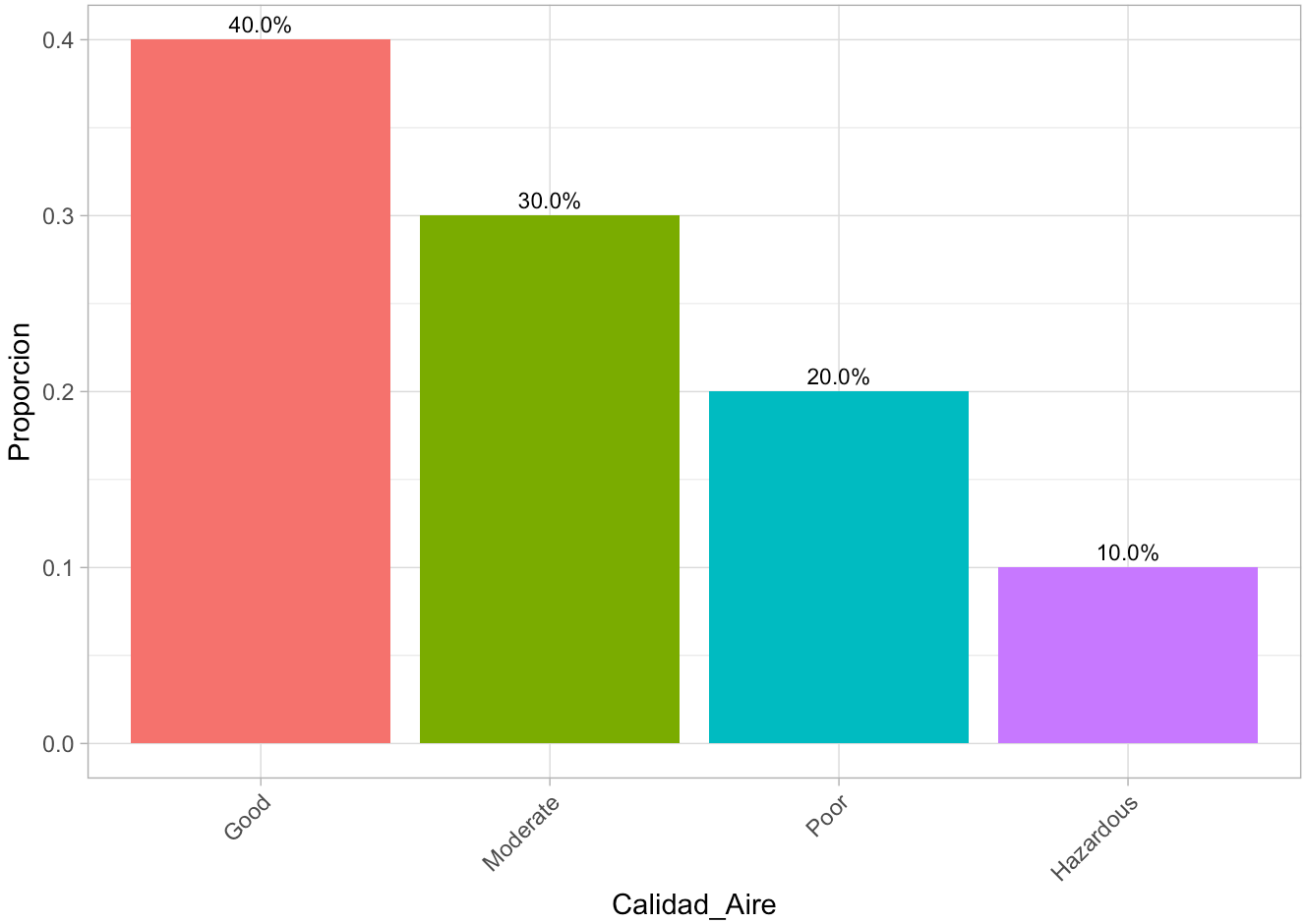
1.1.1. Distribución de las variables numéricas


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



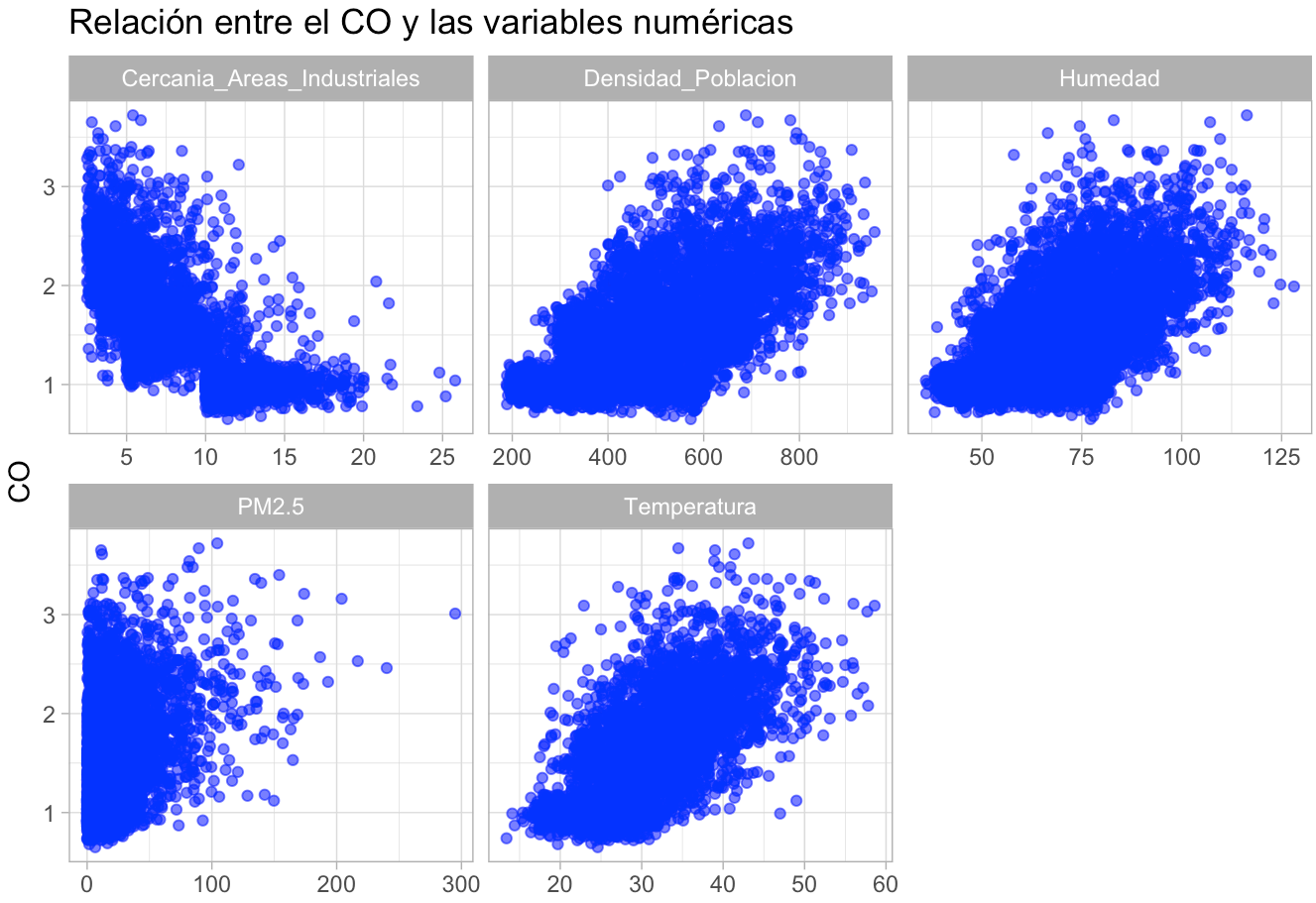
En general se presentan algunas distribuciones sesgadas, especialmente en el PM2.5 (partículas finas) y la variable objetivo (CO) presenta una distribución bimodal. Las demás variables siguen una distribución cercana a la normal.

1.1.2. Distribución de las variables categóricas

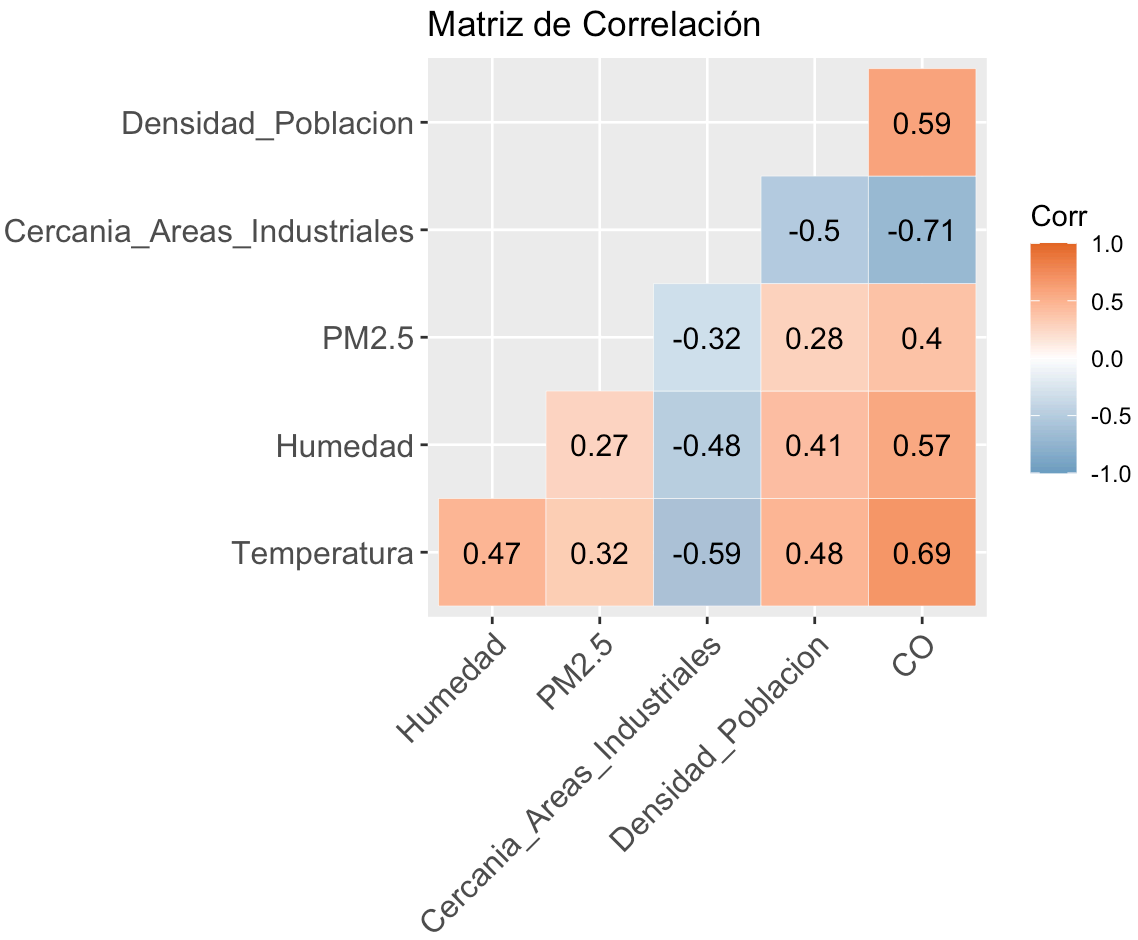


En cuánto a la variable de la calidad del aire, la mayor proporción la tiene la calidad del aire buena, seguida por moderado, luego pobre y por último calidad de aire peligroso.

1.2. Análisis bivariado



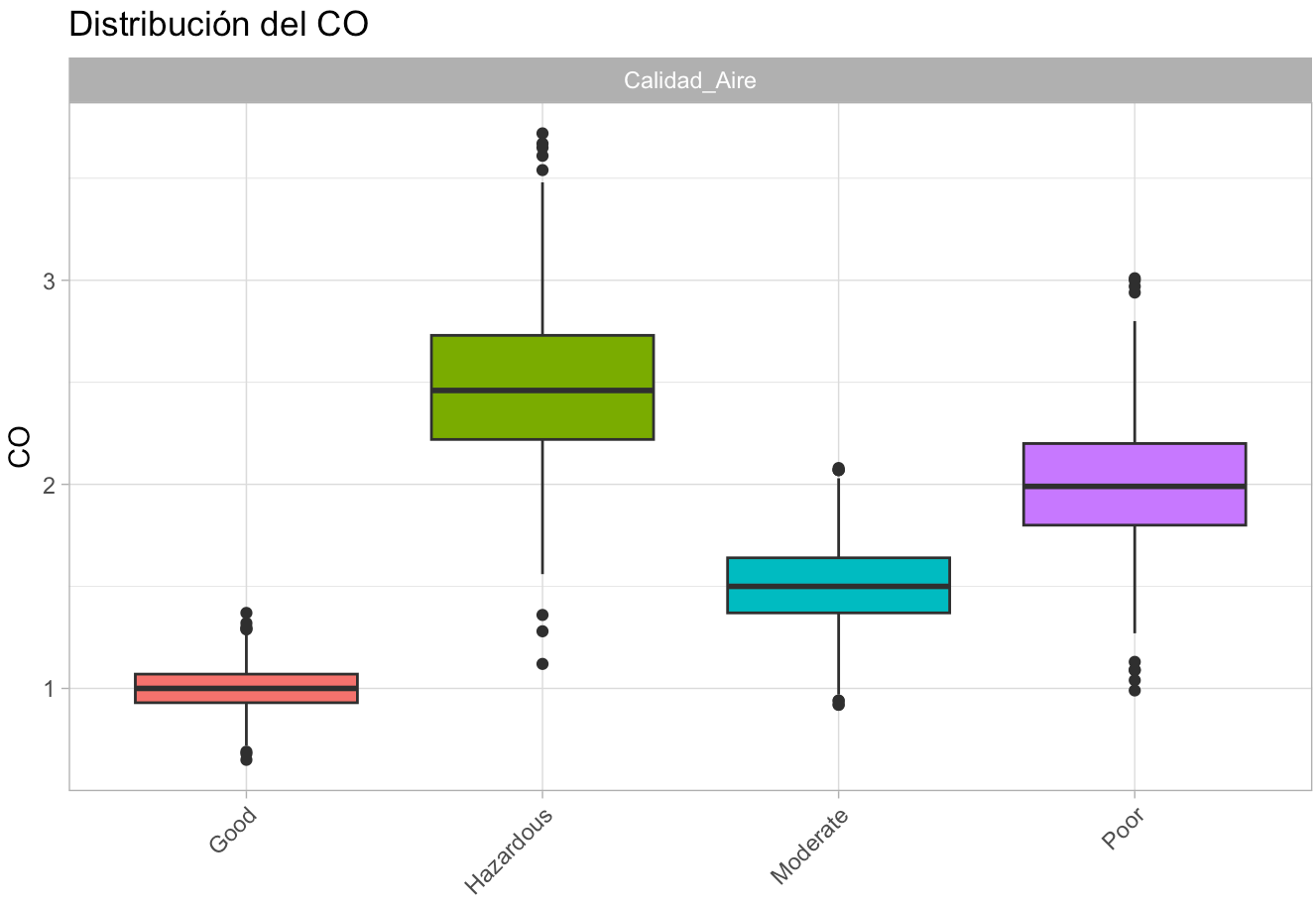
Variables numéricas vs CO



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

	Estimate	Std. Error	t value
(Intercept)	0.8967186387	4.953989e-02	18.1009422
Temperatura	-0.0004610962	7.148229e-04	-0.6450495
Humedad	-0.0002316166	2.552339e-04	-0.9074678
PM2.5	0.0003579116	1.416390e-04	2.5269286
Cercania_Areas_Industriales	-0.0116213933	4.247786e-03	-2.7358708
I(log(Cercania_Areas_Industriales))	0.1024828117	3.437738e-02	2.9811114
Densidad_Poblacion	0.0000269053	2.705775e-05	0.9943659
Calidad_AireModerate	0.5033227767	1.226695e-02	41.0307886
Calidad_AirePoor	1.0058697323	1.736921e-02	57.9110651
Calidad_AireHazardous	1.5059459244	2.371248e-02	63.5085880

	Pr(>|t|)
(Intercept)	5.597556e-71
Temperatura	5.189247e-01
Humedad	3.642033e-01

```
PM2.5                                1.153714e-02
Cercania_Areas_Industriales          6.243551e-03
I(log(Cercania_Areas_Industriales))  2.885910e-03
Densidad_Poblacion                   3.200930e-01
Calidad_AireModerate                2.218134e-317
Calidad_AirePoor                     0.000000e+00
Calidad_AireHazardous                0.000000e+00
```

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% ≈ [0.74, 1.05]), manteniendo lo demás constante.

2.2. Modelo QR

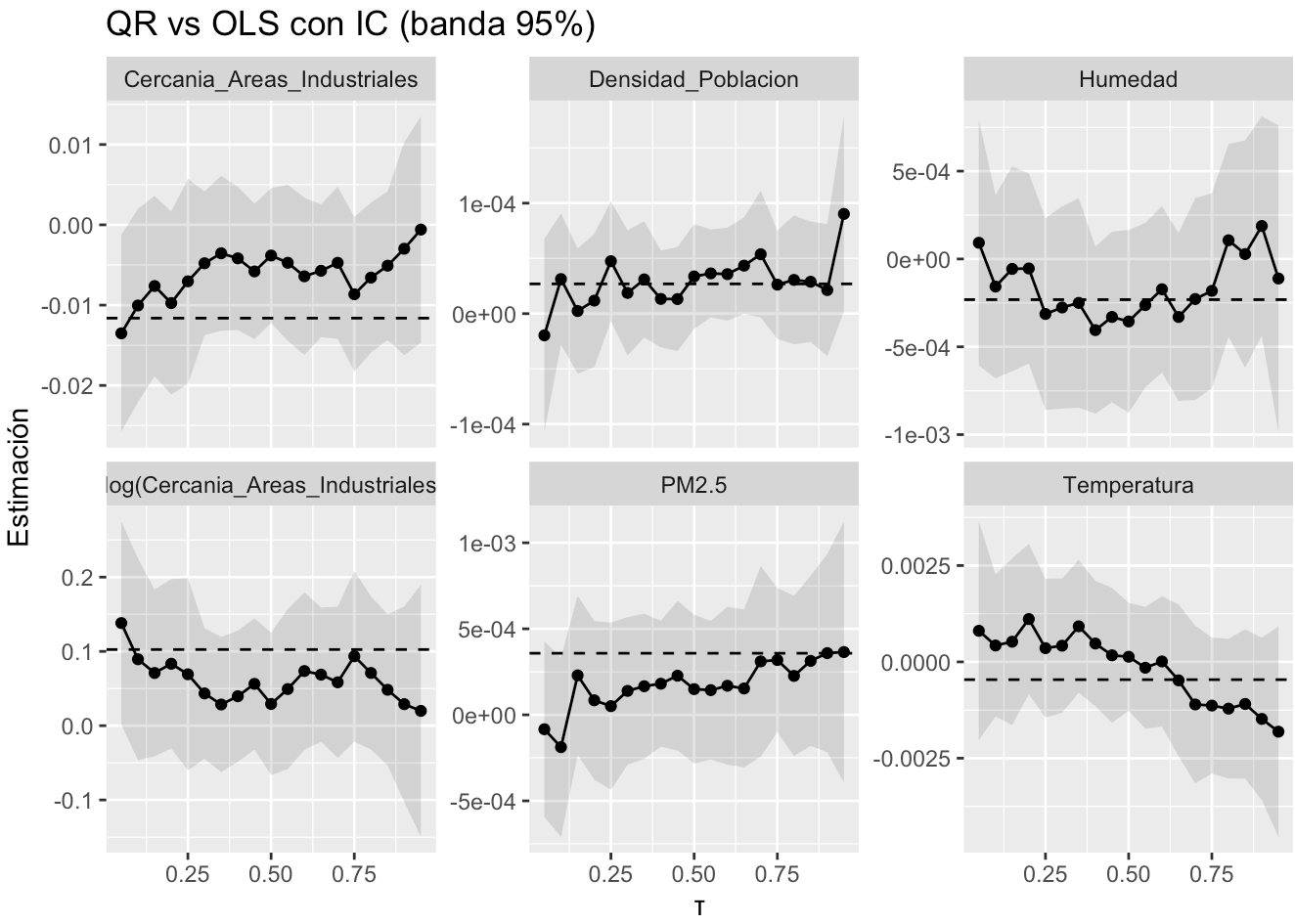
```
Call: rq(formula = form, tau = taus, data = df)

tau: [1] 0.05

Coefficients:
                                   Value    Std. Error t value  Pr(>|t|)
(Intercept)                         0.63730  0.10180    6.26056  0.00000
Temperatura                         0.00081  0.00143    0.56348  0.57314
Humedad                             0.00009  0.00039    0.23737  0.81238
PM2.5                              -0.00008  0.00027   -0.30626  0.75942
Cercania_Areas_Industriales        -0.01351  0.00574   -2.35463  0.01858
I(log(Cercania_Areas_Industriales))  0.13814  0.06441    2.14469  0.03203
Densidad_Poblacion                 -0.00002  0.00004   -0.45170  0.65150
Calidad_AireModerate                0.35961  0.02499   14.38834  0.00000
Calidad_AirePoor                    0.71069  0.03278   21.68378  0.00000
Calidad_AireHazardous               1.11754  0.05181   21.56815  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  0.10459    7.97835  0.00000
Temperatura                         0.00036  0.00092    0.38811  0.69795
Humedad                            -0.00031  0.00029   -1.09322  0.27435
PM2.5                               0.00005  0.00022    0.23181  0.81670
Cercania_Areas_Industriales        -0.00703  0.00687   -1.02404  0.30587
I(log(Cercania_Areas_Industriales))  0.06917  0.06888    1.00423  0.31532
Densidad_Poblacion                  0.00005  0.00003    1.83891  0.06599
Calidad_AireModerate                0.44299  0.01367   32.41399  0.00000
```

```
Calidad_AirePoor                     0.86837  0.02410   36.02698  0.00000
Calidad_AireHazardous                1.28625  0.04490   28.64658  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  0.08522   11.48702  0.00000
Temperatura                         0.00013  0.00082    0.16245  0.87096
Humedad                            -0.00036  0.00027   -1.29533  0.19527
PM2.5                               0.00015  0.00021    0.71089  0.47719
Cercania_Areas_Industriales        -0.00382  0.00485   -0.78802  0.43072
I(log(Cercania_Areas_Industriales))  0.02924  0.05336    0.54795  0.58375
Densidad_Poblacion                  0.00003  0.00003    1.33868  0.18074
Calidad_AireModerate                0.49335  0.01488   33.16141  0.00000
Calidad_AirePoor                    0.98726  0.02327   42.43184  0.00000
Calidad_AireHazardous               1.45891  0.03557   41.01376  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  0.09243   10.45599  0.00000
Temperatura                        -0.00113  0.00094   -1.20411  0.22860
Humedad                            -0.00018  0.00026   -0.70584  0.48032
PM2.5                               0.00032  0.00022    1.45878  0.14469
Cercania_Areas_Industriales        -0.00865  0.00476   -1.81746  0.06921
I(log(Cercania_Areas_Industriales))  0.09348  0.05632    1.65994  0.09699
Densidad_Poblacion                  0.00003  0.00003    0.90597  0.36500
Calidad_AireModerate                0.58201  0.01553   37.46857  0.00000
Calidad_AirePoor                    1.15693  0.02766   41.82530  0.00000
Calidad_AireHazardous               1.70920  0.04383   38.99397  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  0.13876    8.17321  0.00000
Temperatura                        -0.00181  0.00125   -1.44818  0.14763
Humedad                            -0.00011  0.00047   -0.23433  0.81474
PM2.5                               0.00036  0.00041    0.88757  0.37482
Cercania_Areas_Industriales        -0.00059  0.00764   -0.07754  0.93820
I(log(Cercania_Areas_Industriales))  0.01992  0.08988    0.22163  0.82461
Densidad_Poblacion                  0.00009  0.00005    1.94685  0.05161
Calidad_AireModerate                0.66627  0.02536   26.26938  0.00000
```

Calidad_AirePoor	1.33356	0.04206	31.70248	0.00000
Calidad_AireHazardous	2.04209	0.07836	26.06131	0.00000

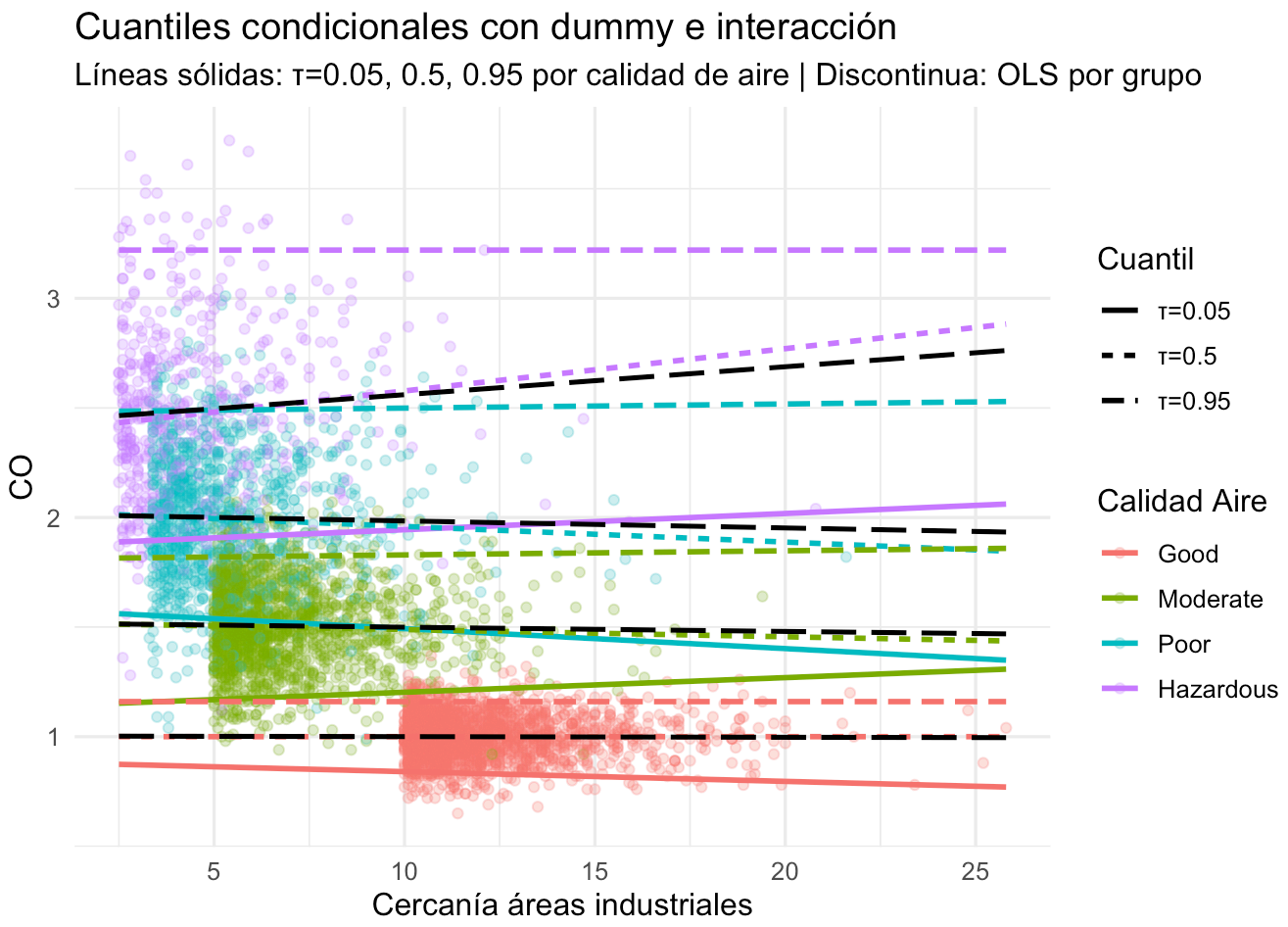
Interpretar

2.3. Resultados gráficos



Interpretar

2.4. Modelo QR con interacción



Interpretar