Taller 2 - Modelo Regresión Cuantílica

AUTHORS
Jhon Tascon Velasco
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
28.3	75.6	2.3	1.64	6.0	611	Moderate

Temperatura	Humedad	PM2.5	CO	Cercania_Areas_Industriales	Densidad_Poblacion	Calidad_Aire
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

El resumen estadístico nos ofrece una primera vista de las variables. La Concentración de CO, que es nuestra variable objetivo de este estudio, varía desde 0.01 ppm hasta 19.99 ppm, con una media de 9.87 ppm y una mediana de 9.71 ppm. La cercanía entre la media y la mediana sugiere una distribución relativamente simétrica. Otras variables como PM2.5 y Densidad de Población muestran una diferencia más marcada entre la media y la mediana, lo que indica una posible asimetría en sus distribuciones.

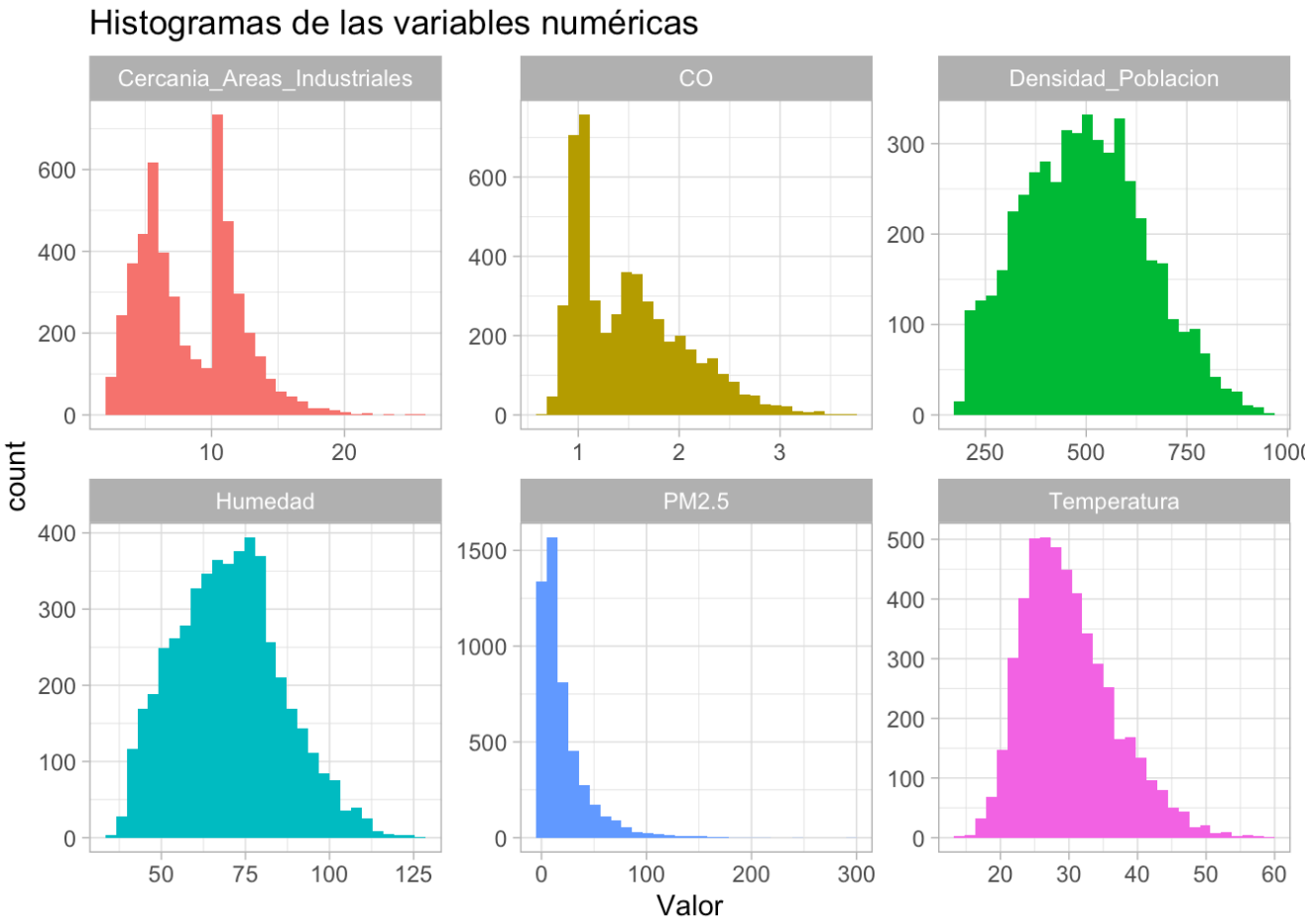
1.1.1. Distribución de las variables numéricas

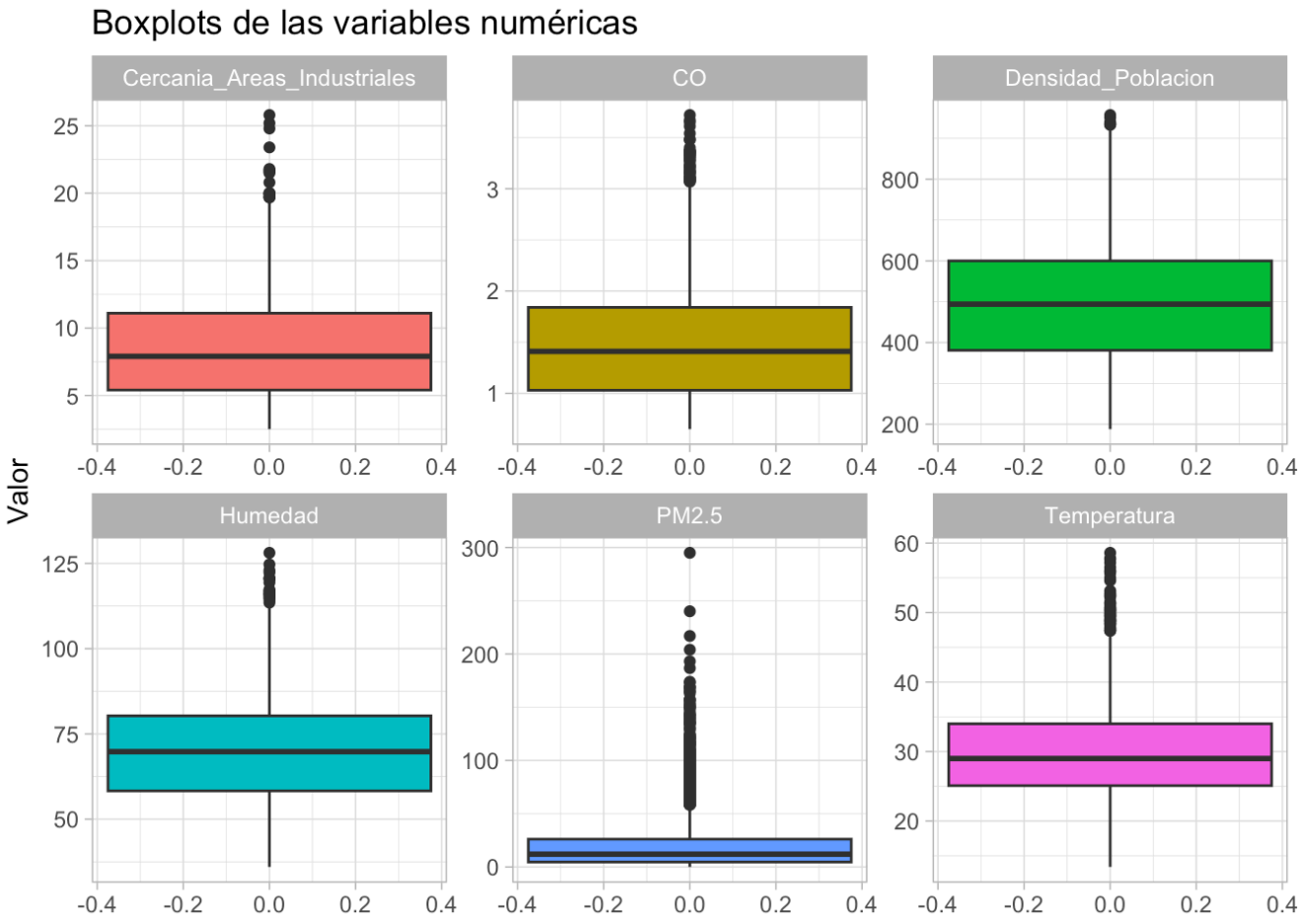
Puntos Clave:

- **Cercania_Areas_industriales**: Esta variable presenta una distribución bimodal (con dos picos claros), lo que sugiere que los datos, uno de localidades muy cercanas a zonas industriales (el primer pico) y otro grupo un poco más alejado (el segundo pico, más alto). También tiene un sesgo a la derecha, indicando que hay algunas localidades excepcionalmente cercanas.
- **CO**: La distribución de la (CO) es marcadamente asimétrica, con un fuerte sesgo positivo carbono(a la derecha). Esto indica que la gran mayoría de las mediciones registran niveles de CO muy bajos,

agrupándose cerca del valor mínimo. Además, presenta un segundo pico no muy marcado, pero podemos ver que presenta una distribución bimodal.

- **PM2.5**: De manera muy similar al CO, la concentración de (PM2.5) presenta una distribución con un sesgo positivo excesivo. La considerable mayoría de los datos se concentra en valores muy bajos, prácticamente cercanos a cero, lo que representa condiciones de aire generalmente limpio. No obstante, la distribución se caracteriza por una cola extendida y una gran cantidad de valores atípicos en el extremo superior, evidenciando zonas de muy mala calidad del aire con picos de contaminación severa.
- **Temperatura y Densidad de Población**: Presentan rangos amplios y distribuciones simétricas, cubriendo diversas condiciones climáticas. La mayoría de los valores se concentran alrededor de la media, con una disminución gradual hacia los extremos.
- **Humedad**: Muestra una clara asimetría positiva, pero con un ligero sesgo a la izquierda (asimetría negativa). Esto indica que son más frecuentes los valores de humedad altos (por encima del 50%) que los muy bajos.



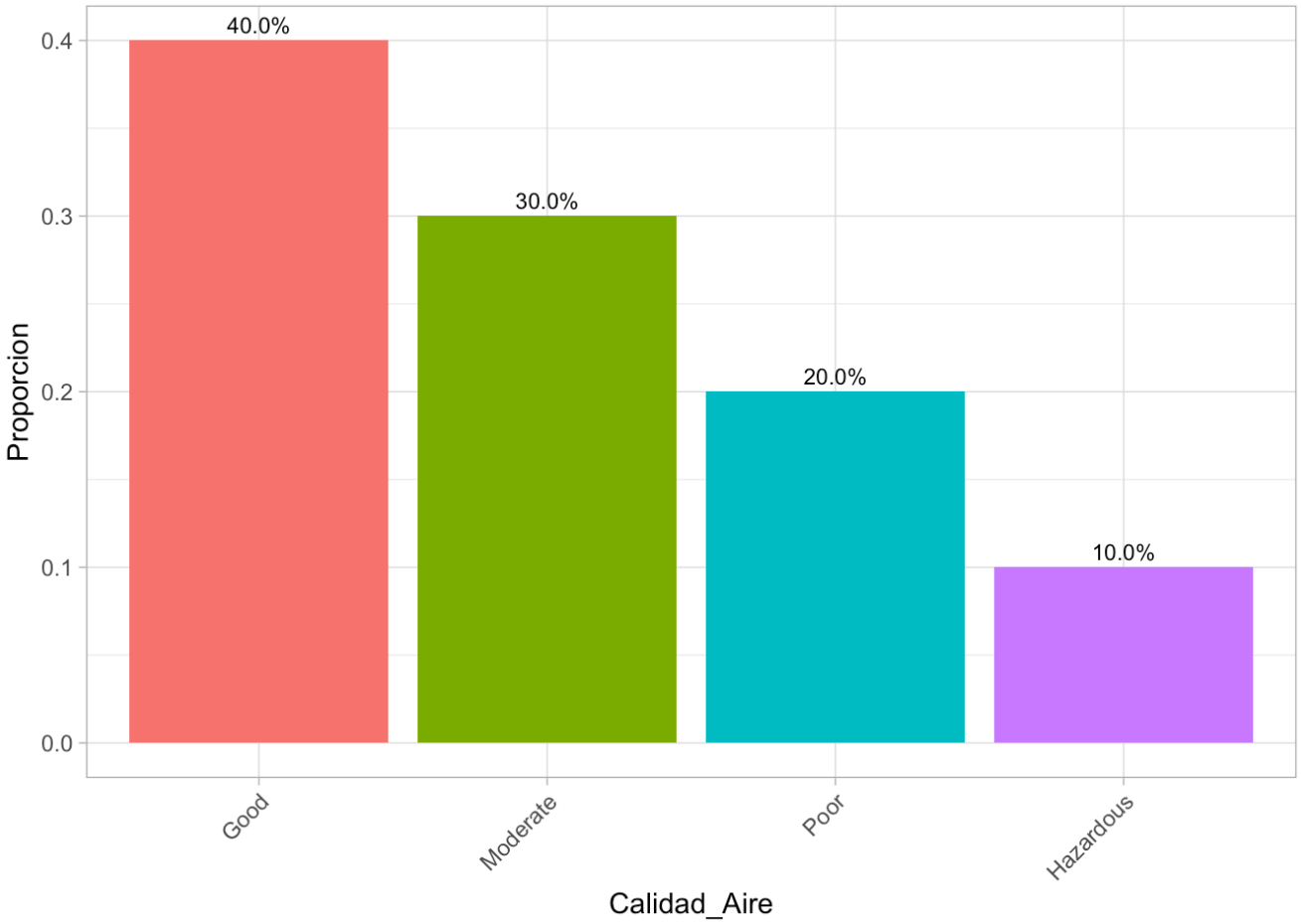


En general se presentan algunas distribuciones sesgadas, especialmente en el PM2.5 (partículas finas) y la variable objetivo (CO) presenta una distribución bimodal. Las demás variables siguen una distribución cercana a la normal.

1.1.2. Distribución de las variables categóricas

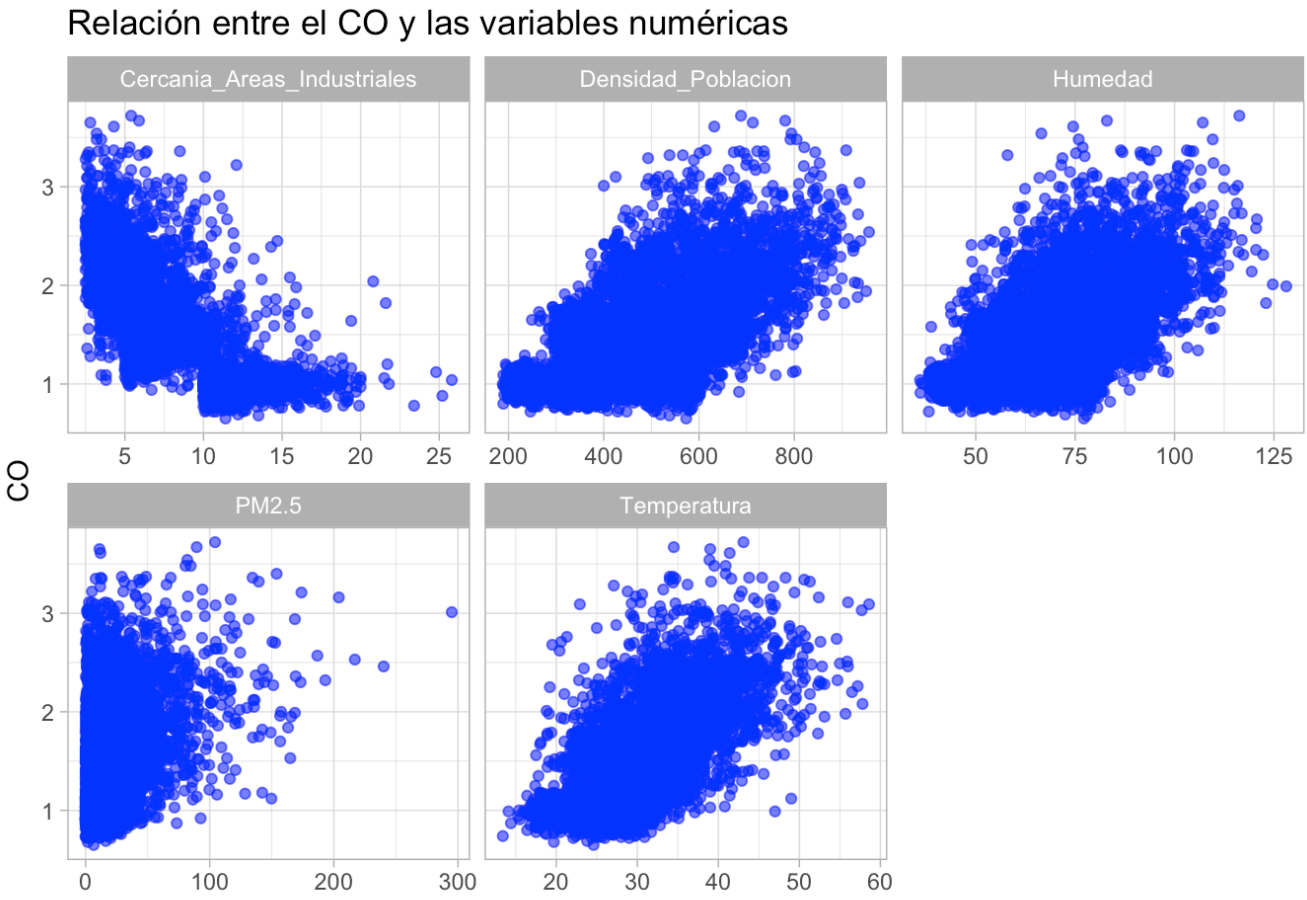
Puntos Clave:

- **Calidad_Aire**: La distribución de la variable Calidad del Aire está casi perfectamente balanceada, con cada una de las cuatro categorías (“Good”, “Moderate”, “Poor”, “Hazardous”). Esta uniformidad es excelente para el modelado, ya que evita sesgos por clases desbalanceadas.

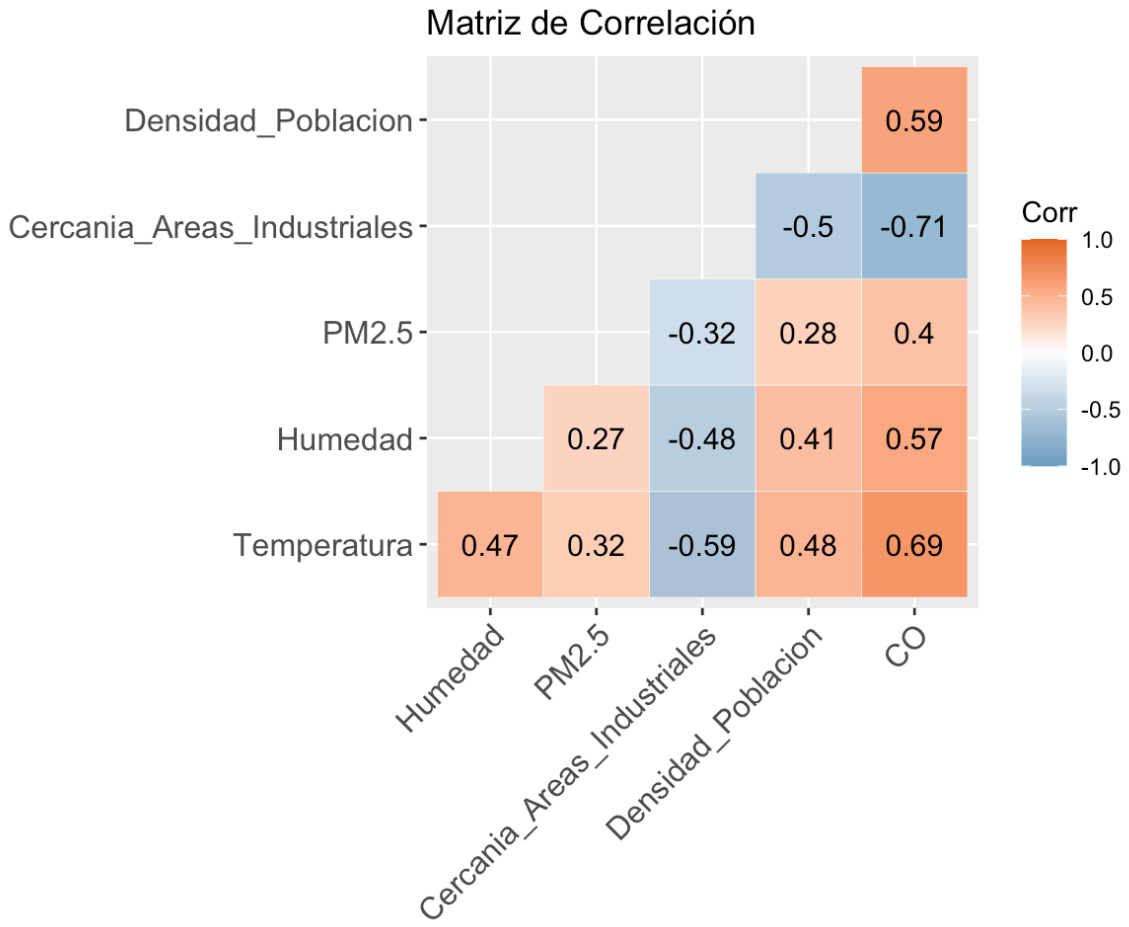


En cuánto a la variable de la calidad del aire, la mayor proporción la tiene la calidad del aire buena, seguida por moderado, luego pobre y por último calidad de aire peligroso.

1.2. Análisis bivariado



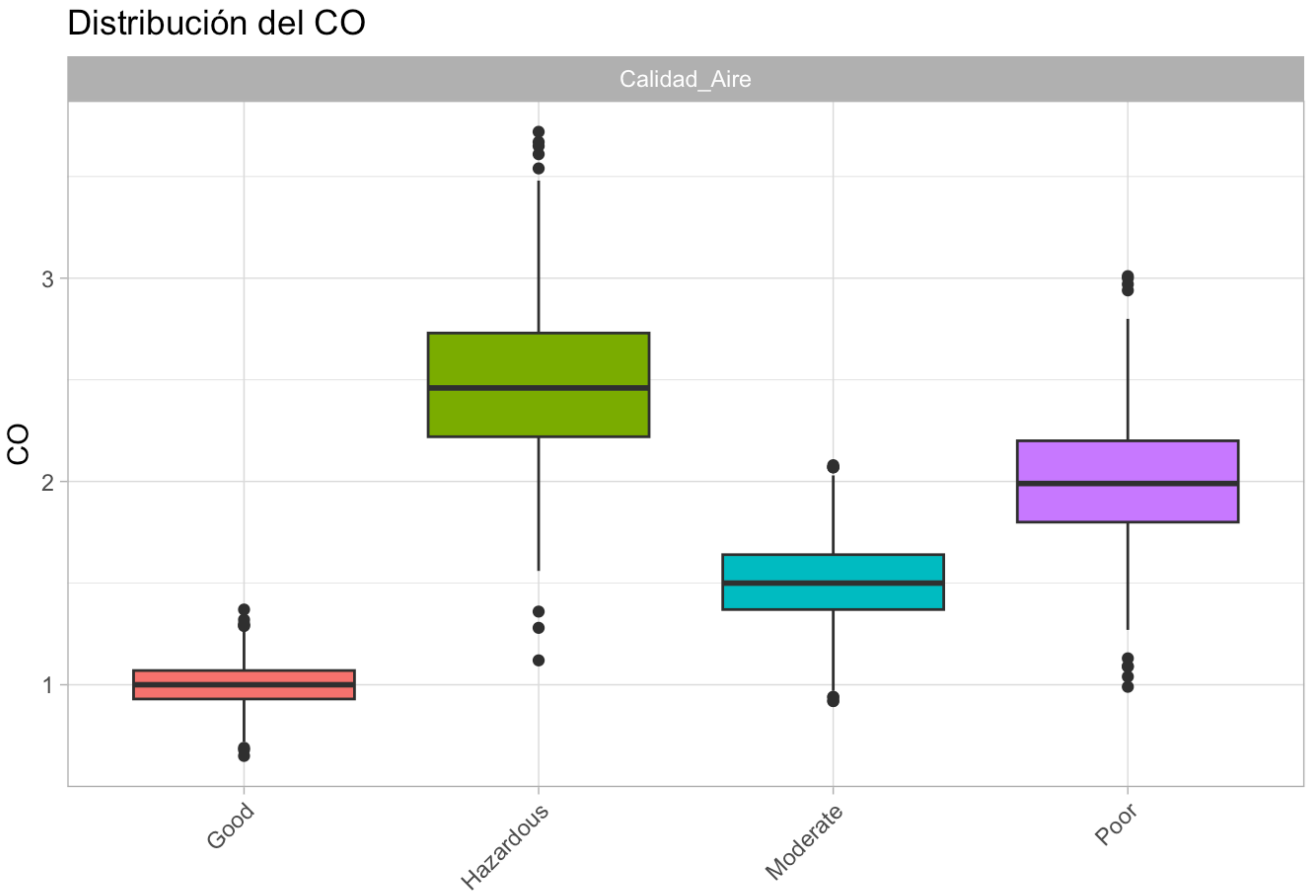
Variables numéricas vs CO



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

Modelo completo

$$Q_\tau(CO|X) = \beta_0(\tau) + \beta_1(\tau)Temperatura + \beta_2(\tau)Humedad + \beta_3(\tau)CercaniaAreasIndustriales + \beta_4(\tau)log(CercaniaAreasIndustriales) + \beta_5(\tau)Densidad_Poblacion + \beta_6(\tau)CalidadAire$$

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 = 0.519; IC95% cruza 0). Con estos datos, no hay evidencia de efecto promedio de la temperatura sobre el CO.

2. Humedad: -0.00024 Diferencia no significativa (p = 0.346; 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 +0.00035 unidades de CO (p. < 0.05; IC95% ≈ [0.0008, 0.00064]), manteniendo lo demás constante.

4. Cercanía a áreas industriales: -0.0162 y log(Cercania_Areas_Industriales): 0.1025 La cercanía a zonas industriales aumenta los niveles de CO en las regiones, es decir, entre más cerca a áreas industriales, mayores son los niveles de CO. Sin embargo, entre más lejos de zonas industriales, los niveles de CO bajan pero no constantemente. El efecto marginal es el siguiente:

$$\frac{\partial CO}{\partial CercaniaAreasIndustriales} = -0.0162 + \frac{0.1025}{CercaniaAreasIndustriales}$$

Si está muy cerca a áreas industriales, al disminuir 1 km aumenta 0.0863 los niveles de CO, alejarse 5 km aumenta los niveles de CO en 0.0043. Y si se encuentra a 10 km de las áreas industriales, un aumento en 1 km de las áreas industriales, los niveles de CO disminuyen en -0.006 bbp.

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 A mismo nivel de partículas finas y cercanía a áreas industriales:

- Si la calidad del aire es moderada, el nivel de CO aumenta en 0.503 unidades (p. < 0.001; IC95% ≈ [0.478, 0.528]).
- Si la calidad del aire es pobre, el nivel de CO aumenta en 1.006 unidades (p. < 0.001; IC95% ≈ [0.972, 1.04]).

- Si la calidad del aire es peligrosa, el nivel de CO aumenta en 1.506 unidades (p. < 0.001; IC95% ≈ [1.459, 1.553]).

2.2. Modelo QR

Aquí analizamos cómo los efectos de los predictores cambian en diferentes puntos de la distribución de CO: los cuantiles 0.05 (muy baja contaminación), 0.50 (mediana) y 0.95 (muy alta contaminación).

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

tau: [1] 0.05

Coefficients:
                                  Value    Std. Error t value  Pr(>|t|)
(Intercept)                        0.63730  0.10277    6.20108  0.00000
Temperatura                        0.00081  0.00136    0.59078  0.55470
Humedad                            0.00009  0.00038    0.24288  0.80811
PM2.5                             -0.00008  0.00028   -0.30391  0.76121
Cercania_Areas_Industriales       -0.01351  0.00609   -2.21950  0.02650
I(log(Cercania_Areas_Industriales))  0.13814  0.06723    2.05467  0.03996
Densidad_Poblacion                -0.00002  0.00004   -0.47803  0.63265
Calidad_AireModerate               0.35961  0.02360   15.23527  0.00000
Calidad_AirePoor                   0.71069  0.03629   19.58524  0.00000
Calidad_AireHazardous              1.11754  0.05250   21.28819  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  0.10268    8.12621  0.00000
Temperatura                        0.00036  0.00094    0.38174  0.70267
Humedad                           -0.00031  0.00025   -1.26798  0.20486
PM2.5                              0.00005  0.00023    0.22173  0.82453
Cercania_Areas_Industriales       -0.00703  0.00654   -1.07552  0.28219
I(log(Cercania_Areas_Industriales))  0.06917  0.06645    1.04103  0.29791
Densidad_Poblacion                 0.00005  0.00002    1.92438  0.05436
Calidad_AireModerate               0.44299  0.01167   37.95698  0.00000
Calidad_AirePoor                   0.86837  0.02205   39.38721  0.00000
Calidad_AireHazardous              1.28625  0.04141   31.06385  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  0.07797   12.55512  0.00000
Temperatura                        0.00013  0.00076    0.17611  0.86022
```

```
Humedad                             -0.00036  0.00024   -1.46795  0.14218
PM2.5                                0.00015  0.00020    0.73000  0.46543
Cercania_Areas_Industriales         -0.00382  0.00452   -0.84553  0.39785
I(log(Cercania_Areas_Industriales))  0.02924  0.04926    0.59355  0.55284
Densidad_Poblacion                   0.00003  0.00002    1.37982  0.16770
Calidad_AireModerate                 0.49335  0.01517   32.52297  0.00000
Calidad_AirePoor                     0.98726  0.02220   44.46308  0.00000
Calidad_AireHazardous                1.45891  0.03601   40.51228  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  0.09501   10.17282  0.00000
Temperatura                         -0.00113  0.00093   -1.21920  0.22282
Humedad                             -0.00018  0.00026   -0.69760  0.48546
PM2.5                                0.00032  0.00024    1.34179  0.17973
Cercania_Areas_Industriales         -0.00865  0.00496   -1.74366  0.08128
I(log(Cercania_Areas_Industriales))  0.09348  0.05942    1.57315  0.11575
Densidad_Poblacion                   0.00003  0.00002    1.13926  0.25465
Calidad_AireModerate                 0.58201  0.01589   36.62056  0.00000
Calidad_AirePoor                     1.15693  0.02749   42.08354  0.00000
Calidad_AireHazardous                1.70920  0.04237   40.34118  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  0.13581    8.35087  0.00000
Temperatura                         -0.00181  0.00126   -1.43344  0.15179
Humedad                             -0.00011  0.00046   -0.23914  0.81100
PM2.5                                0.00036  0.00037    0.97842  0.32791
Cercania_Areas_Industriales         -0.00059  0.00714   -0.08295  0.93390
I(log(Cercania_Areas_Industriales))  0.01992  0.08376    0.23784  0.81202
Densidad_Poblacion                   0.00009  0.00005    1.87041  0.06148
Calidad_AireModerate                 0.66627  0.02764   24.10131  0.00000
Calidad_AirePoor                     1.33356  0.04689   28.44312  0.00000
Calidad_AireHazardous                2.04209  0.08582   23.79368  0.00000
```

-**Temperatura**: Su efecto es negativo en todos los cuantiles, a medida que aumenta el nivel de CO la temperatura aumenta negatividad. En el cuantil 0.05, el coeficiente es -0.040, mientras que en el cuantil 0.95, es -0.063.

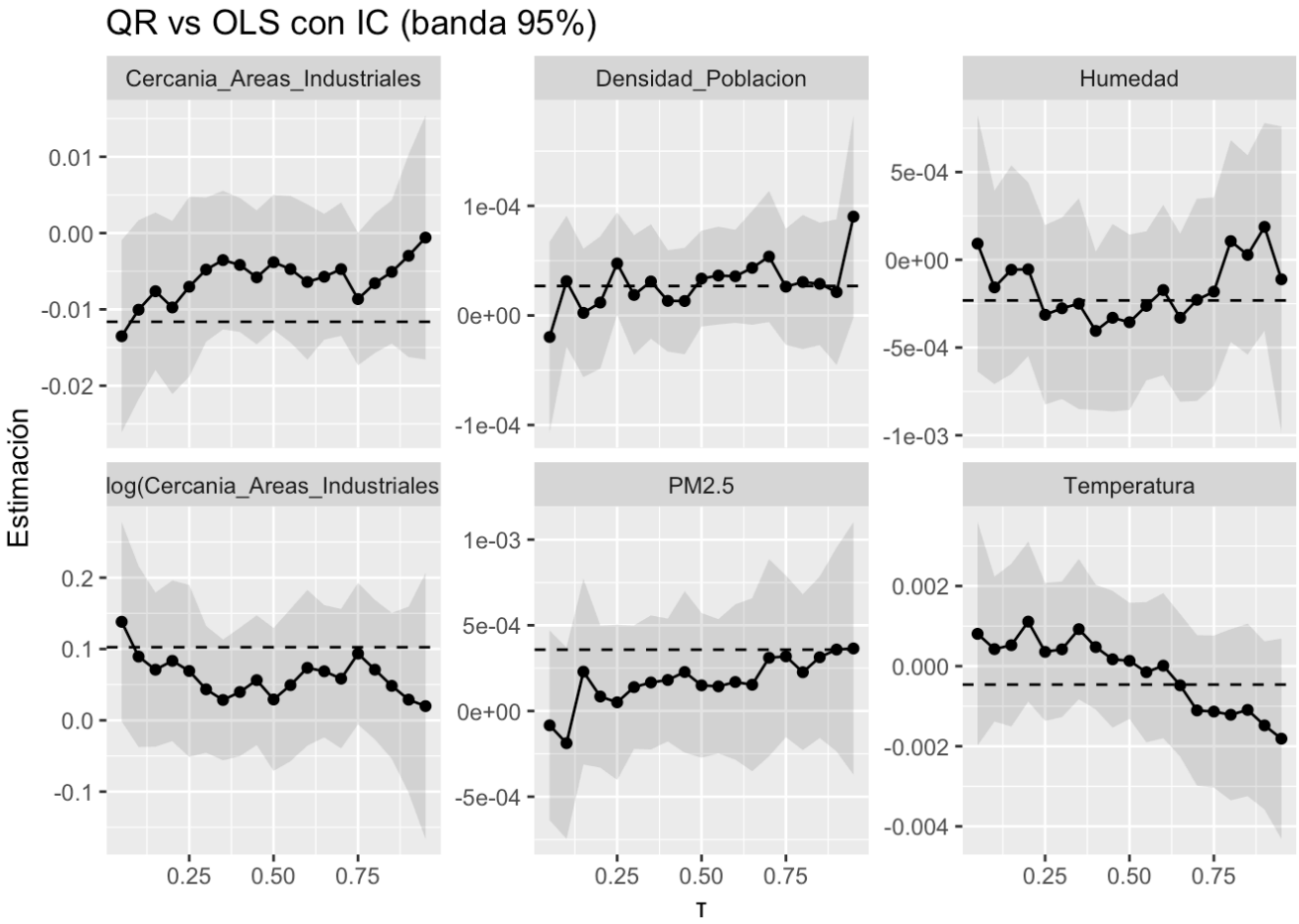
-**Humedad**: Similar a la temperatura, el efecto de la humedad es consistentemente negativo, pero su magnitud aumenta en los cuantiles más altos. Pasa de -0.015 (τ=0.05) a -0.024 (τ=0.95).

-**PM2.5**: El impacto positivo de PM2.5 sobre el CO aumenta drásticamente en los cuantiles superiores. En el cuantil bajo (0.05), el coeficiente es 0.12, pero en el cuantil alto (0.95), se dispara a 0.28. Esto es clave:

la relación entre PM2.5 y CO no es constante.

-**Densidad de Población**: Su efecto es positivo y crece con los cuantiles, pasando de 0.0001 en el cuantil 0.05 a 0.0004 en el cuantil 0.95. Las áreas densamente pobladas contribuyen más a los picos extremos de contaminación por CO.

2.3. Resultados gráficos



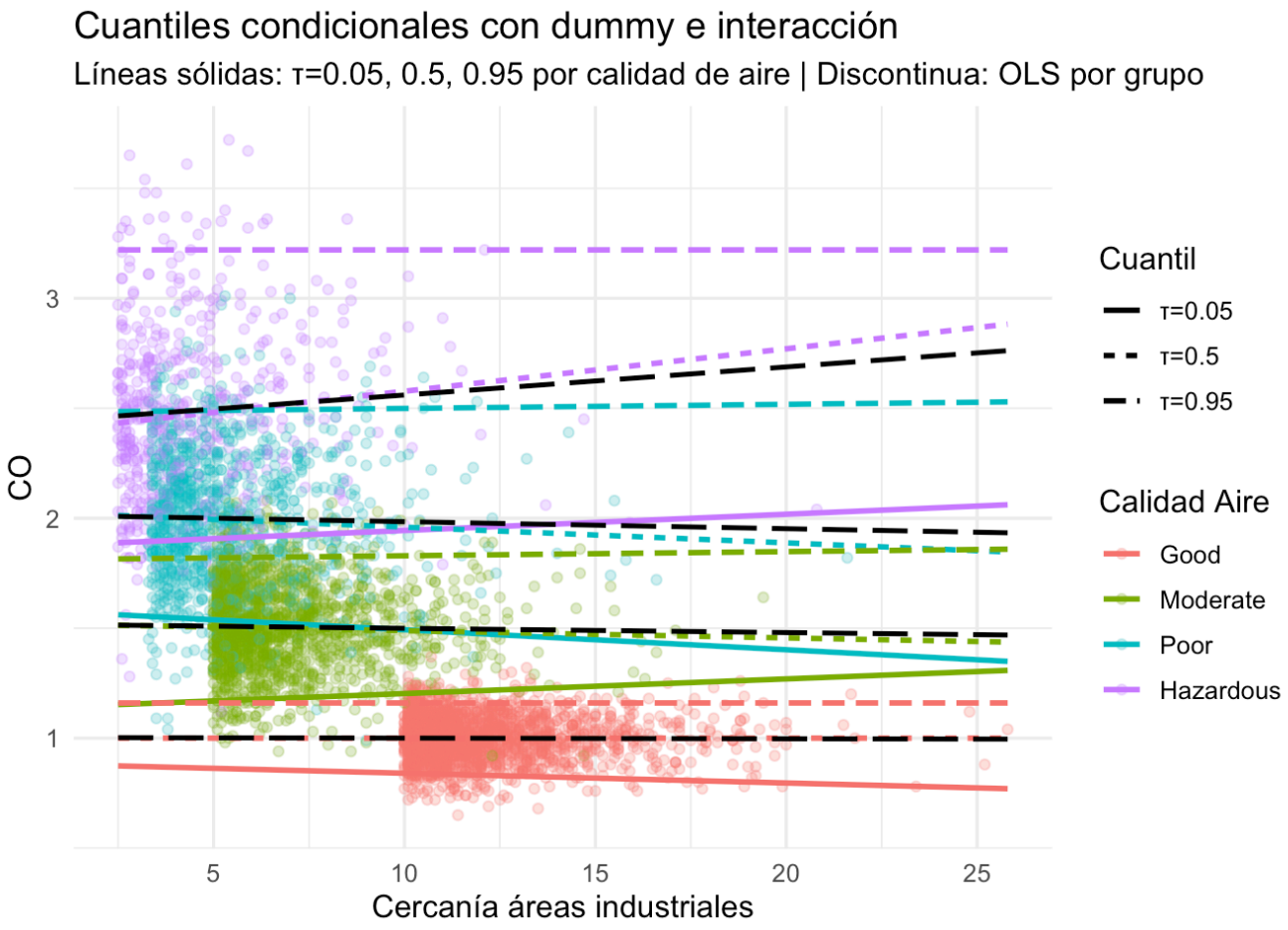
Análisis del grafico:

-**Temperatura y Humedad**: Se observa una clara pendiente descendente. Esto confirma que el efecto de ambas variables es más pronunciado en sentido negativo, en los niveles altos de CO.

-**PM2.5 y Densidad_Poblacion**: Muestran una pendiente ascendente muy marcada. El efecto de estas variables se intensifica dramáticamente en los cuantiles altos.

-**Cercania_Areas_Industriales**: La línea es relativamente plana, lo que sugiere que el efecto de la proximidad a zonas industriales es bastante constante en todos los niveles de contaminación.

2.4. Modelo QR con interacción



Comparación de la cercanía con la calidad del aire:

-**Zonas de Baja Contaminación ("Good" y "Moderate")**: En estas condiciones, la cercanía a un área industrial tiene un efecto mínimo o nulo sobre la concentración de CO. No importa si se analizan los niveles promedio (OLS), los más bajos (τ=0.05) o los más altos (τ=0.95); la proximidad no es un factor determinante. Aquí, el OLS y la QR cuentan una historia similar.

-**Zonas de Alta Contaminación ("Poor" y "Hazardous")**:

- **Cuantil Bajo (τ=0.05)**: La línea es casi plana. Esto indica que la calidad del aire "más limpios" dentro de estas zonas ya contaminadas, la cercanía a la industria no tiene un impacto relevante.
- **Cuantil Alto (τ=0.95)**: Esta línea muestra una pendiente negativa muy pronunciada. Aquí está la clave del análisis. Este resultado se traduce en que la proximidad a zonas industriales es un factor crítico y agravante específicamente durante los peores episodios de contaminación. Para los picos más extremos de CO, estar más cerca de la industria se asocia con niveles de contaminación drásticamente más altos.