

# 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 capturando factores 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:

Descripción de variables del estudio

Variables	Unidades	Descripción
Temperatura	°C	Temperatura media de la región
Humedad	%	Humedad relativa registrada en la región
Concentración PM2.5	µg/m³	Niveles de partículas finas
Concentración PM10	µg/m³	Niveles de partículas gruesas (no se trabajará en el estudio)
Concentración NO2	ppb	Niveles de dióxido de nitrógeno (no se trabajará en el estudio)
Concentración 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	Categorías	Clasificación: Bueno, Moderado, Pobre, Peligroso
Concentración de CO	ppb	Variable objetivo: Niveles de monóxido de carbono

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
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 ppb hasta 19.99 ppb, con una media de 9.87 ppb y una mediana de 9.71 ppb. 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.

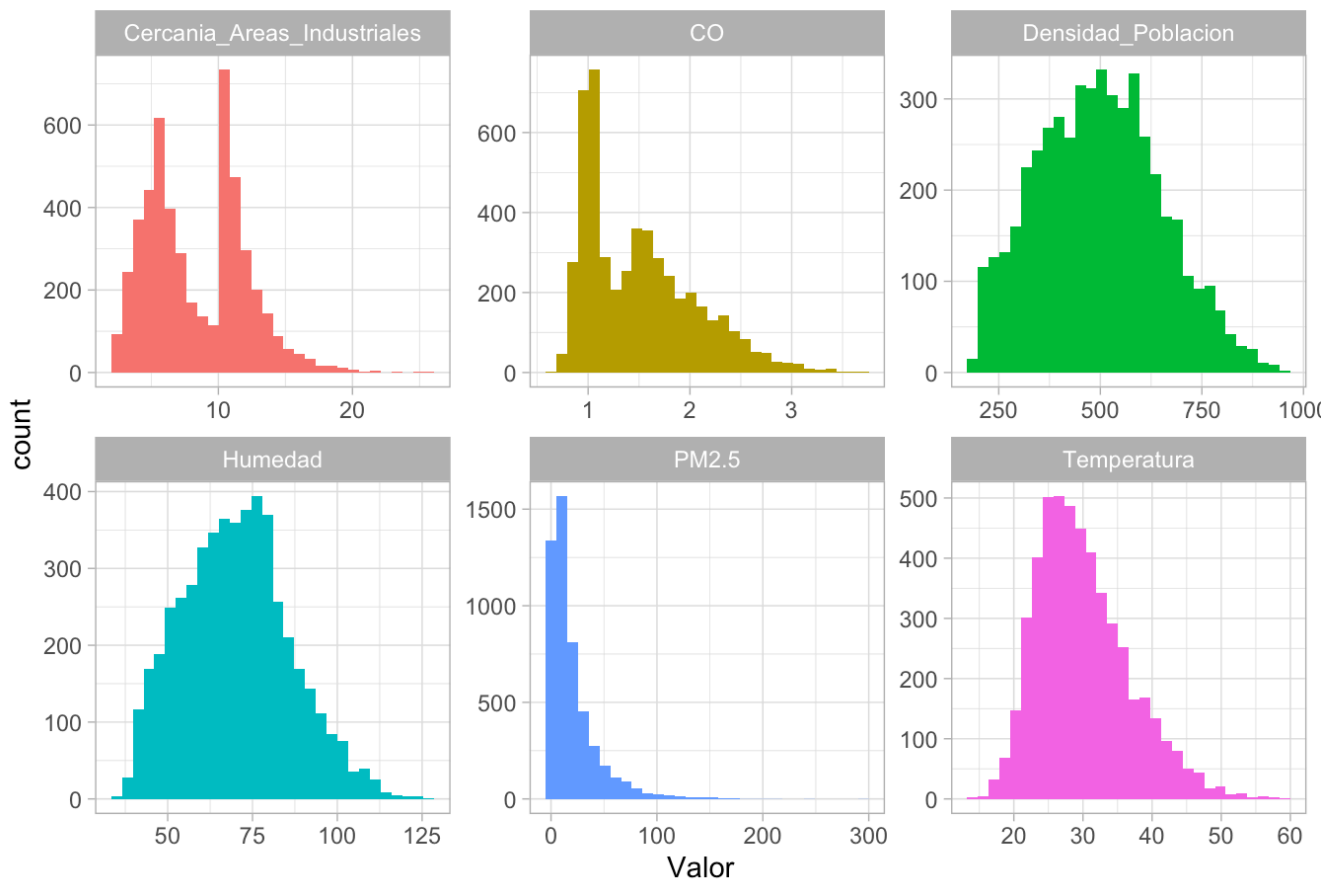
#### 1.1.1. Distribución de las variables numéricas

Puntos Clave:

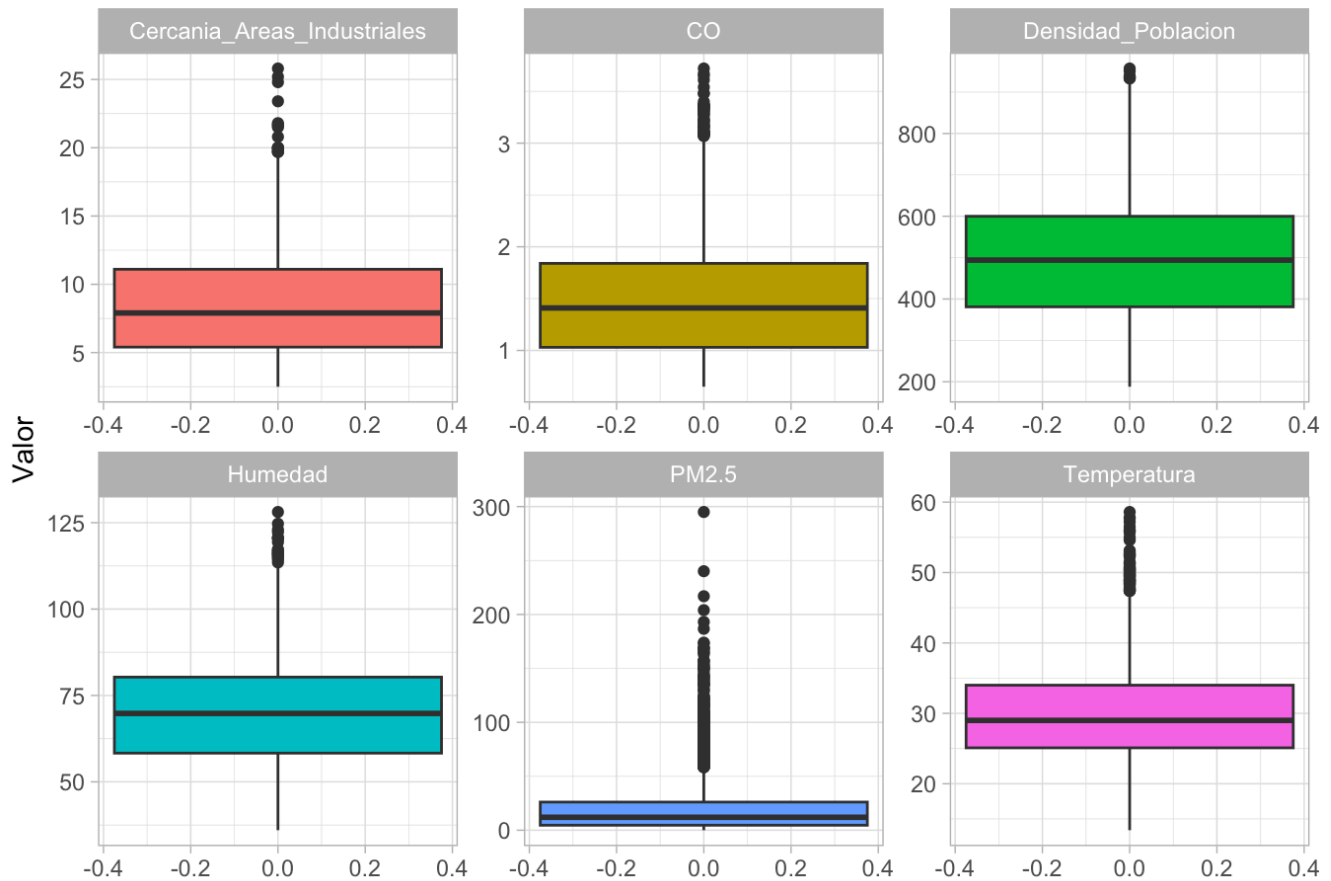
- **Cercania\_Areas\_industriales:** Esta variable presenta una distribución bimodal (con dos picos claros), lo que sugiere que los datos, uno de localidades muy cercanas a zonas industriales (el primer pico) y otro grupo un poco más alejado (el segundo pico, más alto). También tiene un sesgo a la derecha, 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.

Histogramas de las variables numéricas



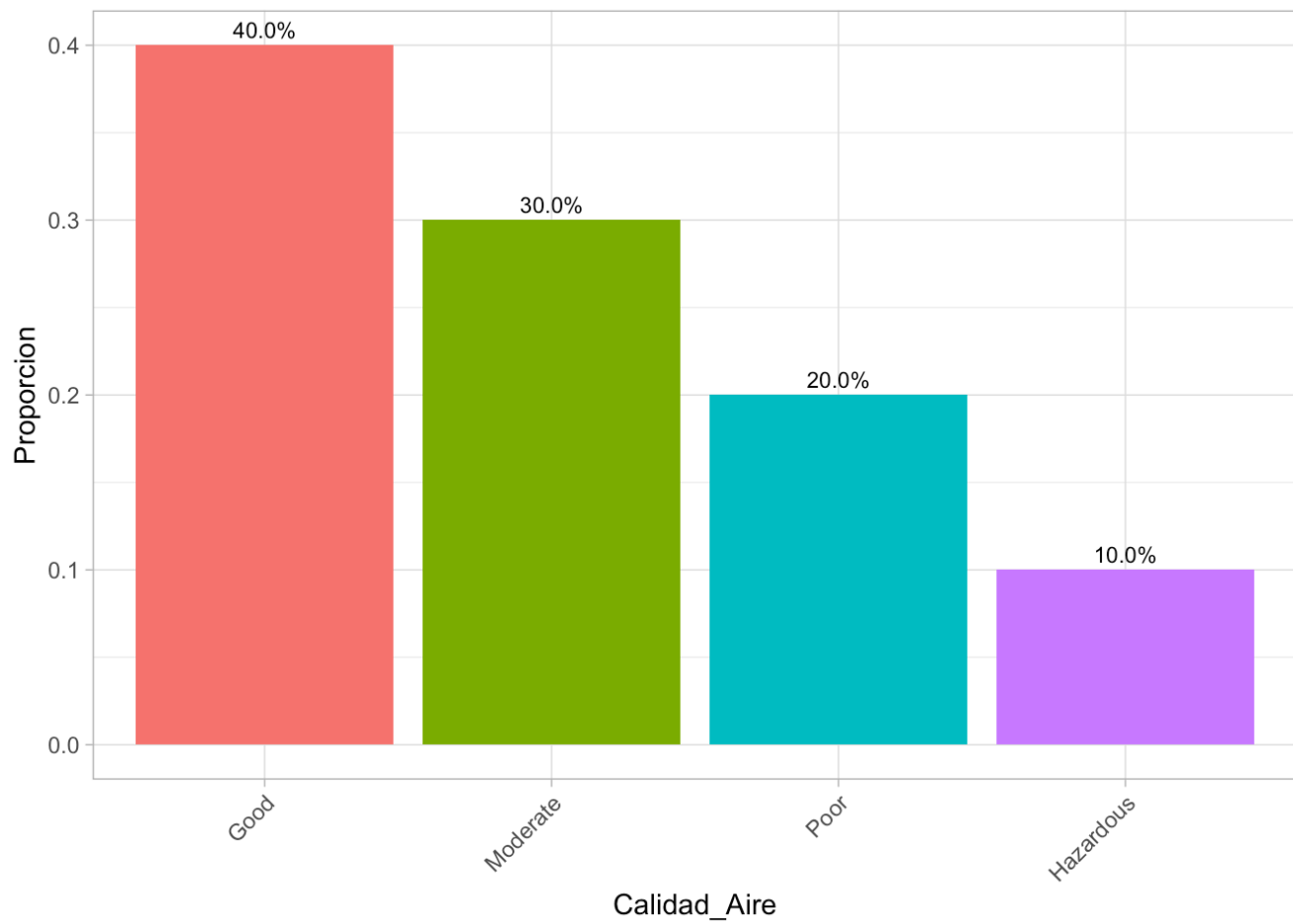
## Boxplots de las variables numéricas



### 1.1.2. Distribución de las variables categóricas

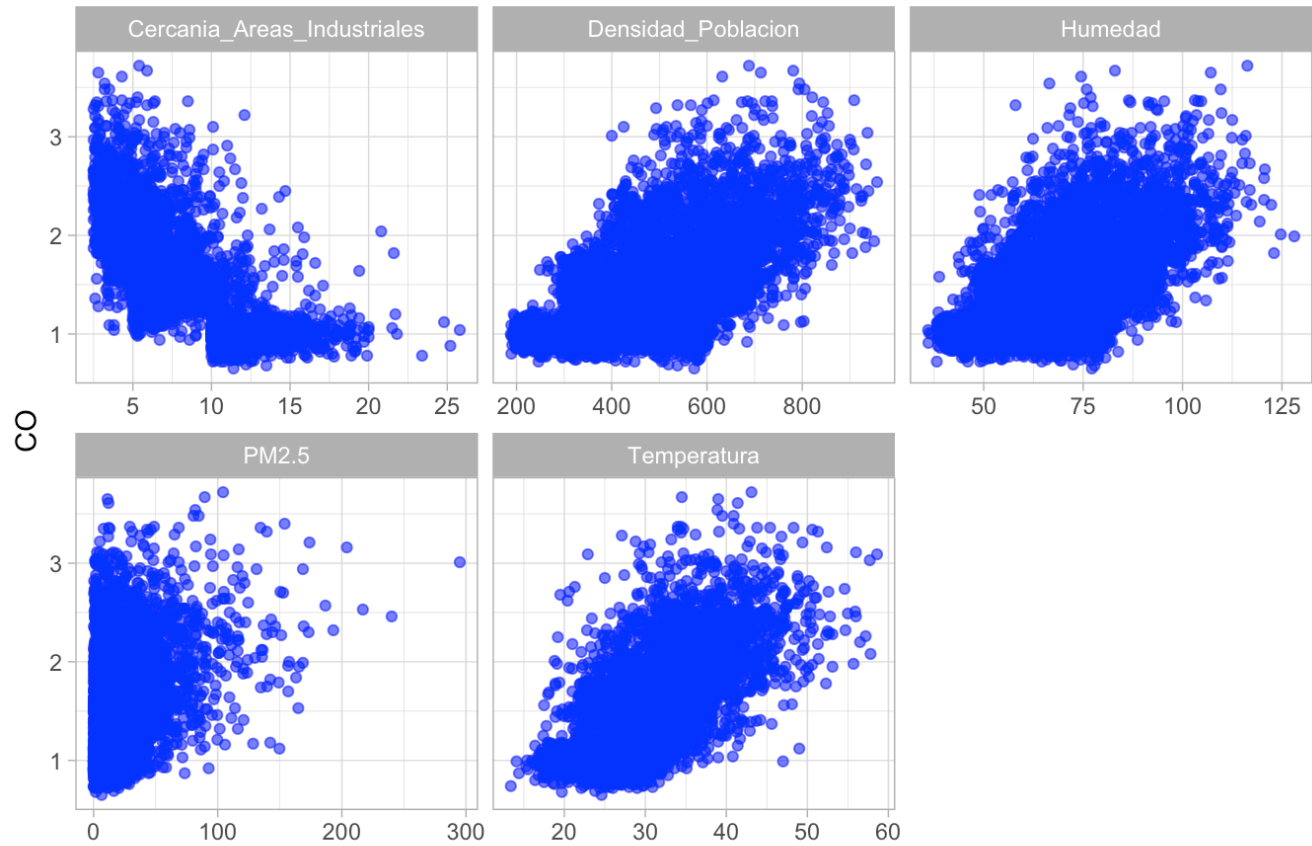
Puntos Clave:

- Calidad\_Aire:** La variable Calidad del Aire presenta una distribución heterogénea pero relativamente balanceada entre sus cuatro categorías. La mayor proporción corresponde a la categoría Good (40%), seguida por Moderate (30%), lo que indica que en la mayoría de las observaciones la calidad del aire se encuentra en niveles aceptables o moderados. Sin embargo, también se evidencia que un 20% de los registros corresponden a condiciones Poor y un 10% a situaciones Hazardous, reflejando episodios de contaminación con riesgos significativos para la salud. Esta distribución es adecuada para el análisis, ya que garantiza variabilidad suficiente en los datos sin presentar un desbalance extremo entre clases, aunque la menor representación de la categoría Hazardous requiere especial atención para evitar que sea subestimada en los modelos predictivos.

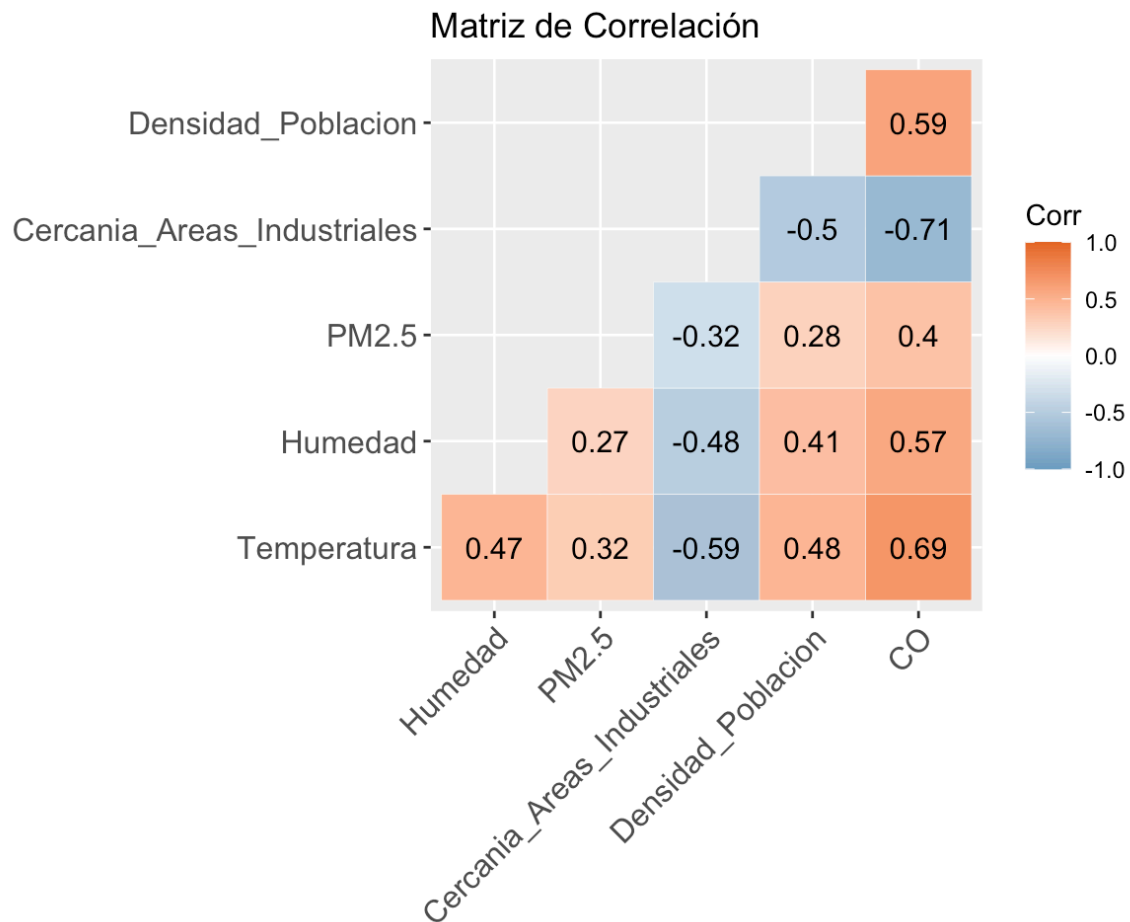


## 1.2. Análisis bivariado

Relación entre el CO y las variables numéricas



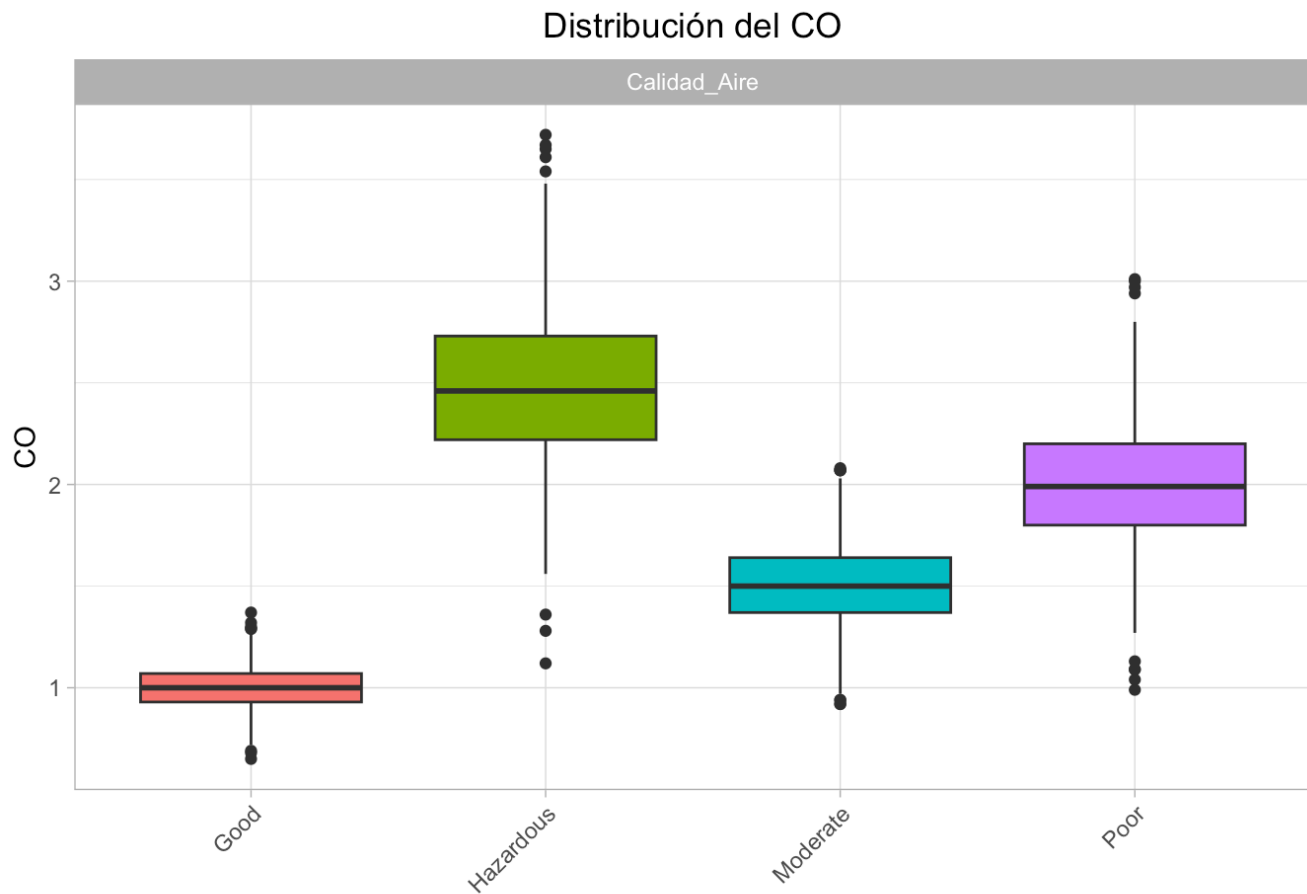
Variables numéricas vs CO



El CO con las demás variables numéricas presentan relaciones medias-altas, especialmente con la temperatura (relación positiva), humedad (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 embargo, en zonas medias de humedad (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

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

### 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: -0.00023** 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%  $\approx [1.459, 1.553]$ ).

## 7. Medidas globales del Modelo

A nivel global, el modelo tiene los siguientes valores:

- Residual standard error (0.2231): en promedio, el modelo se equivoca  $\pm 0.22$  unidades al predecir.
- $R^2 = 0.8333$ : el modelo explica aproximadamente el 83% de la variabilidad de la variable dependiente  
→ muy buen ajuste.
- Adjusted  $R^2$  (0.833): prácticamente igual, lo que indica que no hay sobreajuste.
- F-statistic (2773,  $p < 2.2e-16$ ): el modelo en su conjunto es altamente significativo.

## 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	0.10744	5.93147	0.00000
Temperatura	0.00081	0.00138	0.58233	0.56037
Humedad	0.00009	0.00038	0.24524	0.80628
PM2.5	-0.00008	0.00028	-0.30120	0.76327
Cercania_Areas_Industriales	-0.01351	0.00606	-2.23129	0.02571
I(log(Cercania_Areas_Industriales))	0.13814	0.06806	2.02966	0.04244
Densidad_Poblacion	-0.00002	0.00005	-0.43306	0.66499
Calidad_AireModerate	0.35961	0.02222	16.18435	0.00000
Calidad_AirePoor	0.71069	0.03520	20.18792	0.00000
Calidad_AireHazardous	1.11754	0.05242	21.31754	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	0.08953	9.32026	0.00000
Temperatura	0.00036	0.00087	0.41163	0.68063
Humedad	-0.00031	0.00029	-1.08446	0.27821
PM2.5	0.00005	0.00024	0.21137	0.83260
Cercania_Areas_Industriales	-0.00703	0.00592	-1.18867	0.23463
I(log(Cercania_Areas_Industriales))	0.06917	0.05926	1.16722	0.24318
Densidad_Poblacion	0.00005	0.00003	1.56704	0.11717
Calidad_AireModerate	0.44299	0.01368	32.38812	0.00000
Calidad_AirePoor	0.86837	0.02399	36.19619	0.00000
Calidad_AireHazardous	1.28625	0.04172	30.83197	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	0.07802	12.54694	0.00000
Temperatura	0.00013	0.00079	0.16813	0.86649
Humedad	-0.00036	0.00025	-1.41183	0.15806
PM2.5	0.00015	0.00019	0.76949	0.44164
Cercania_Areas_Industriales	-0.00382	0.00442	-0.86475	0.38722
I(log(Cercania_Areas_Industriales))	0.02924	0.04890	0.59795	0.54990
Densidad_Poblacion	0.00003	0.00002	1.48083	0.13872
Calidad_AireModerate	0.49335	0.01562	31.58664	0.00000
Calidad_AirePoor	0.98726	0.02328	42.40412	0.00000
Calidad_AireHazardous	1.45891	0.03609	40.41979	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	0.08702	11.10701	0.00000
Temperatura	-0.00113	0.00086	-1.32524	0.18515
Humedad	-0.00018	0.00030	-0.61356	0.53953
PM2.5	0.00032	0.00023	1.36350	0.17279
Cercania_Areas_Industriales	-0.00865	0.00463	-1.86836	0.06177
I(log(Cercania_Areas_Industriales))	0.09348	0.05487	1.70369	0.08850
Densidad_Poblacion	0.00003	0.00003	0.95747	0.33838
Calidad_AireModerate	0.58201	0.01530	38.04533	0.00000
Calidad_AirePoor	1.15693	0.02779	41.63030	0.00000
Calidad_AireHazardous	1.70920	0.04119	41.49866	0.00000

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

```
tau: [1] 0.95
```

```
Coefficients:
```

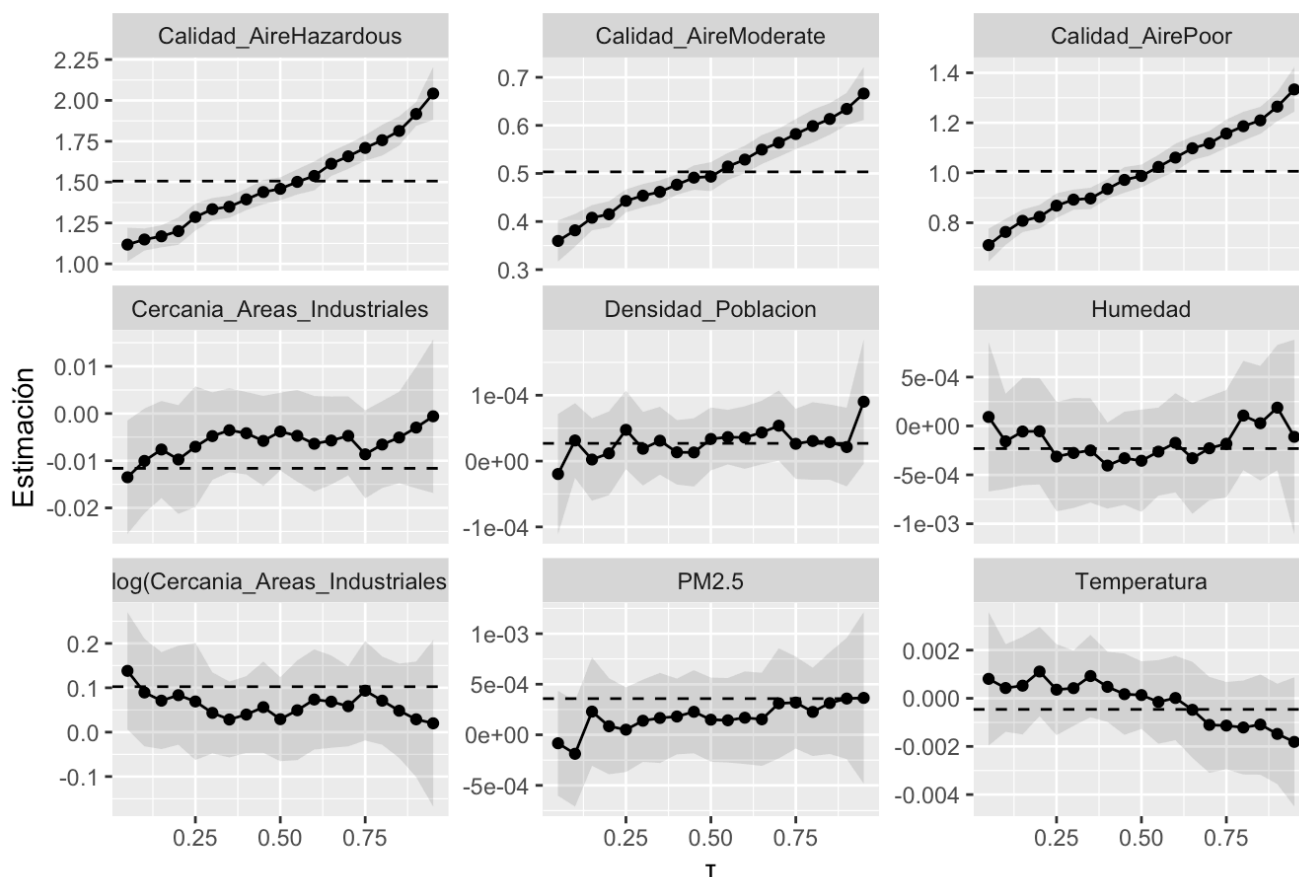
	Value	Std. Error	t value	Pr(> t )
(Intercept)	1.13414	0.13675	8.29371	0.00000
Temperatura	-0.00181	0.00132	-1.37400	0.16950
Humedad	-0.00011	0.00044	-0.25285	0.80040
PM2.5	0.00036	0.00037	0.99497	0.31980
Cercania_Areas_Industriales	-0.00059	0.00755	-0.07848	0.93745
I(log(Cercania_Areas_Industriales))	0.01992	0.08764	0.22729	0.82021
Densidad_Poblacion	0.00009	0.00004	2.16070	0.03077
Calidad_AireModerate	0.66627	0.02830	23.54093	0.00000
Calidad_AirePoor	1.33356	0.04194	31.79591	0.00000
Calidad_AireHazardous	2.04209	0.07750	26.35082	0.00000

Coeficientes de regresión cuantil por variable y cuantiles

Variable	Q0.05	Q0.25	Q0.5	Q0.75	Q0.95	OLS
Temperatura	0.00081	0.00036	0.00013	-0.00113	-0.00181	-0.00046
Humedad	0.00009	-0.00031	-0.00036	-0.00018	-0.00011	-0.00023
PM2.5	-0.00008	0.00005	0.00015	0.00032	0.00036	0.00036
Cercanía_Areas_Industriales	-0.01351	-0.00703	-0.00382	-0.00865	-0.00059	-0.01620
l(log(Cercanía_Areas_Industriales))	0.13814	0.06917	0.02924	0.09348	0.01992	0.10250
Densidad_Población	-0.00002	0.00005	0.00003	0.00003	0.00009	0.00003
Calidad_AireModerate	0.35961	0.44299	0.49335	0.58201	0.66627	0.50300
Calidad_AirePoor	0.71069	0.86837	0.98726	1.15693	1.33356	1.00600
Calidad_AireHazardous	1.11754	1.28625	1.45891	1.70920	2.04209	1.50600

## 2.3. Resultados gráficos

QR vs OLS con IC (banda 95%)



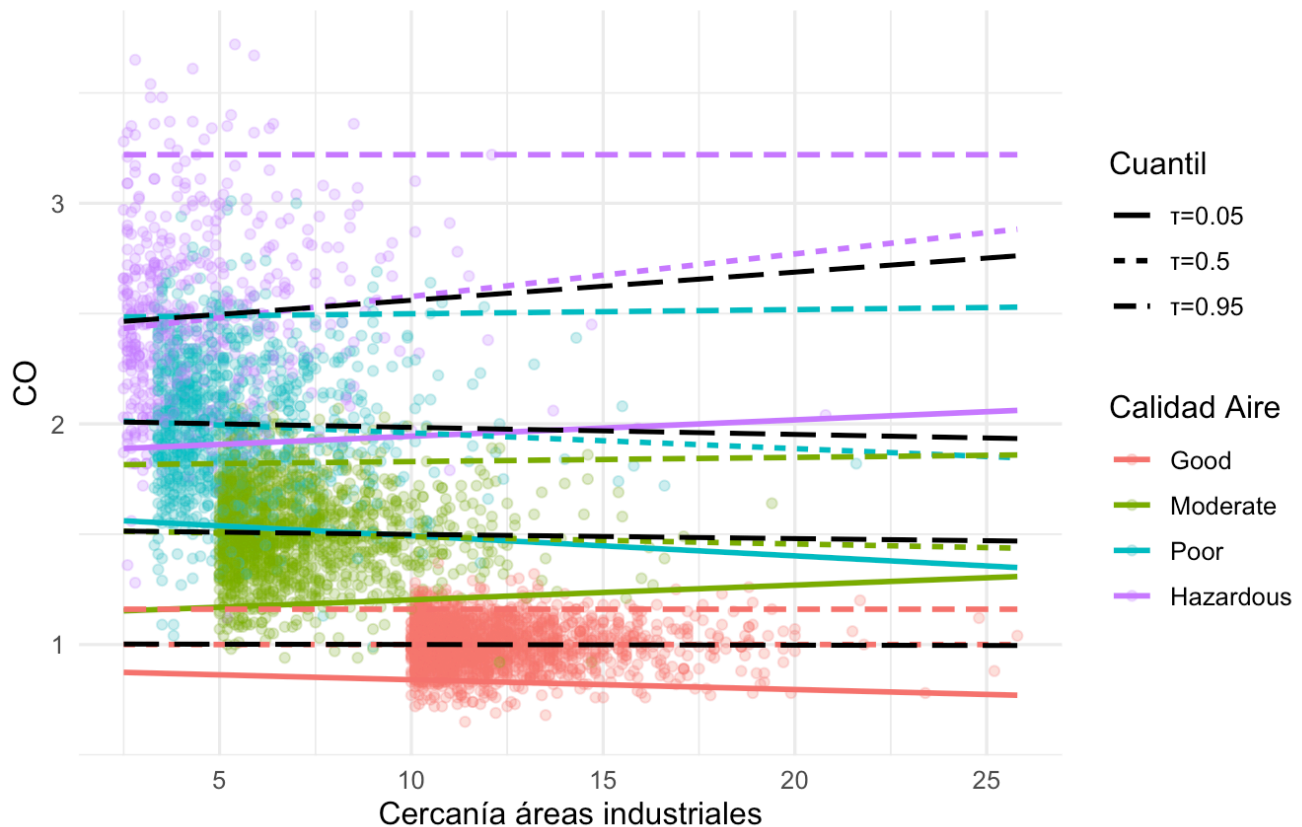
- **Temperatura:** Su efecto va decreciendo a través de los cuantiles. Sin embargo, **no es significativo** en todos los tau. En cuanto al modelo OLS, **sobreestima** en la base y **subestima** en la cúpula.
- **Humedad:** Su impacto es positivo en el cuantil 0.05, pero negativo creciente a partir del cuantil 0.25. Sin embargo, al igual que la temperatura, **no es significativo** en todos los tau.

- **PM2.5:** Tiene un impacto negativo en la base y va creciendo. Sin embargo, **no es significativo** en todos los tau.
- **Cercanía Areas Industriales:** la parte lineal presenta una tendencia creciente en los taus, teniendo un impacto negativo (a mayor distancia, menores los niveles de CO). Sin embargo, esta variable **solo es significativa** en los cuantiles 0.05 y 0.75. La parte logarítmica presenta una tendencia decreciente en los taus, sin embargo **no es significativa** en ningún tau.
- **Densidad de Población:** su efecto es negativo en la base y tiende a crecer muy poco a través de los taus. Sin embargo, **no es significativa**.
- **Calidad del aire:** en todos los casos, la calidad del aire es significativa y es creciente en todos los taus. Además, con base en los gráficos, su crecimiento es constante (lineal), pero hay una brecha un poco mayor en la cúpula, lo que indica que los niveles de CO son cada vez mayores a medida que la calidad del aire va empeorando.

## 2.4. Modelo QR con interacción

Cuantiles condicionales con dummy e interacción

Líneas sólidas:  $\tau=0.05, 0.5, 0.95$  por calidad de aire | Discontinua: OLS por grupo



Comparación de la cercanía con la calidad del aire:

- **Zonas de mejor calidad de aire ("Good" y "Moderate"):** En estas condiciones, la cercanía a un área industrial tiene un efecto mínimo-medio sobre los niveles de CO. No importa si se analizan los niveles promedio (OLS), los más bajos ( $\tau=0.05$ ) o los más altos ( $\tau=0.95$ ); la proximidad no es un factor determinante. Aquí, el OLS y la QR cuentan una historia similar. Sin embargo, las pendientes en calidad de aire moderada muestran una inclinación positiva, aunque no muy pronunciada,

indicando que la cercanía industrial empieza a tener un efecto perceptible, especialmente en los valores más altos de CO.

- **Zonas de baja calidad de aire ("Poor" y "Hazardous"):** en estas zonas, la interacción entre calidad de aire y la cercanía es mayor, lo que indica que la cercanía a zonas industriales se asocia claramente con un aumento en los niveles de CO. En zonas en el que la calidad de aire es pobre (poor), la proximidad a fuentes industriales contribuye significativamente a la contaminación por CO, afectando incluso los valores bajos (cuantil 0.05), mientras que, cuando la calidad del aire ya es peligrosa (Hazardous), la exposición a fuentes industriales es crítica. Incluso a distancias moderadas, los niveles de CO son peligrosamente altos.

Los dos puntos anteriores nos dicen que, la interacción entre la calidad del aire y la cercanía no es lineal ni uniforme, y la influencia de la cercanía es más fuerte en regiones en el que la calidad del aire es pobre o peligrosa.