Pontificia Universidad Catolica de Chile

EYP230I — Analisis de Regresion

Tarea 1

Modelos de Regresion Aplicados a la Demanda de Arriendos de Bicicletas

Luna Garces, Tomas Perez, Benjamin Thareau

13 de octubre de 2025

Introduccion y contexto del problema

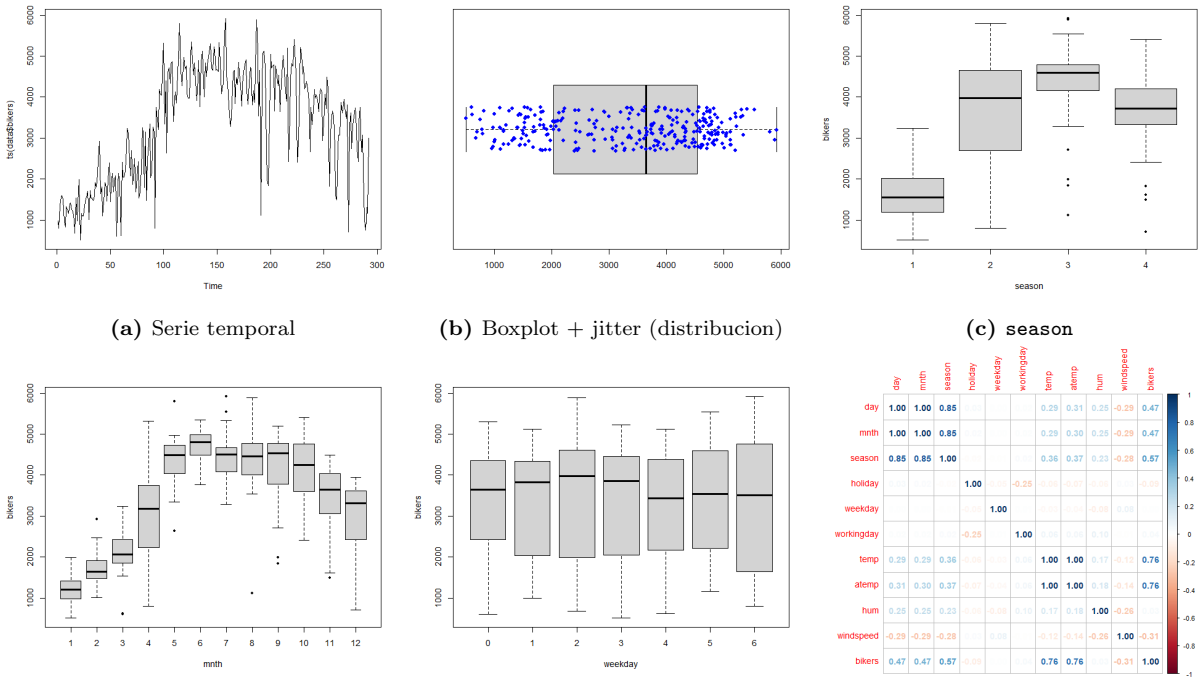
El objetivo de este trabajo es modelar y explicar la variación diaria en la cantidad de ciclistas (`bikers`) registrada por un sistema de bicicletas compartidas. Para ello se construiremos un modelo de regresión lineal con variables meteorológicas y calendarizadas, buscando dos cosas complementarias:

- comprender los factores que mas influyen en la demanda diaria
- evaluar la capacidad predictiva del modelo sobre datos no vistos.

El dataset original esta en resolución horaria. Nosotros modificamos el dataset **agregando la columna a nivel diario**: sumamos los conteos de usuarios (`bikers`) en el día y promediamos las variables meteorológicas (`temp`, `atemp`, `hum`, `windspeed`); la situación del tiempo se resume con la **moda** (`weathersit_mode`). Finalmente, dividimos aleatoriamente la muestra en 80 % entrenamiento y 20 % prueba para evaluar el desempeño fuera de muestra. En esta tarea veremos el éxito del análisis con MAPE en el conjunto de prueba.

Resultados del análisis exploratorio

La unidad de análisis es **diaria**. Variables de calendario: `season`, `mnth`, `weekday`, `holiday` y `workingday`. Condiciones meteorológicas: `temp`, `atemp`, `hum`, `windspeed` (promedios diarios) y situación del tiempo: `weathersit_mode`. A continuación se muestran gráficos resumen:



(a) Serie temporal **(b)** Boxplot + jitter (distribucion) **(c)** `season`

(d) `mnth` **(e)** `weekday` **(f)** Matriz de correlaciones

Figura 1: Resumen de gráficos del análisis exploratorio

En conjunto, las seis visualizaciones confirman patrones claros: **(a)** la serie temporal muestra **estacionalidad** marcada, con alzas en meses cálidos y caídas en meses frios; **(b)** la distribución marginal (boxplot + jitter) muestra alta dispersión pero sin outliers extremos tras la agregación diaria; **(c)** por `season` se observan niveles maximos en primavera/verano y mínimos en invierno; **(d)** por `mnth` aparece un patrón con pico aprox. en 6–8; **(e)** por `weekday` las diferencias son acotadas; **(f)** la matriz de correlaciones destaca la fuerte relación positiva de `temp`/`atemp` con `bikers` (y entre si), mientras `hum` y `windspeed` se relacionan negativamente.

Otro punto importante en destacar es que quitamos la columna bikers, que es la que se busca predecir.

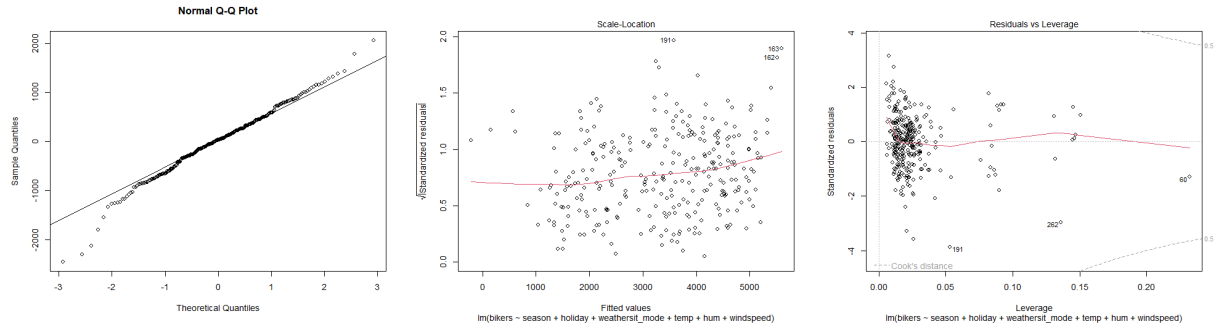
Selección de variables y modelo final

Se partió de un modelo *completo* (todas las variables diarias) y se aplicaron procedimientos de selección por AIC en direcciones *forward* y *backward*. Ambos llegaron a la misma especificacion, eliminando `mnth`, `weekday`, `workingday` y `atemp`. Esta ultima se descarto por alta colinealidad con `temp` (VIF $\approx$ 12–16 en el completo); en el modelo final los VIF quedan proximos a 1–1.3. Escogimos por *backward* porque parte del modelo saturado y permite evaluar con mayor claridad el aporte marginal de cada predictor, también porque es mas estable que *forward* cuando hay predictores correlacionados, y alcanzo el mismo AIC mínimo que *forward* la **Formula del modelo final (*backward*):**

$$\texttt{bikers}_t = \beta_0 + \beta_1\,\texttt{season}_t + \beta_2\,\texttt{holiday}_t + \beta_3\,\{\texttt{weathersit\_mode} = \text{cloudy/misty}\}_t + \beta_4\,\{\texttt{weathersit\_mode} = \text{light rain/snow}\}$$

Tenemos que

- **Estacionalidad** (`season`, $\hat{\beta}_1 > 0$): mas ciclistas en estaciones cálidas.
- **Feriados** (`holiday`, $\hat{\beta}_2 < 0$, margen de significancia): leve reducción de la demanda.
- **Situación del tiempo**: respecto de *clear*, `cloudy/misty` ($\hat{\beta}_3 < 0$) y especialmente `light rain/snow` ($\hat{\beta}_4 < 0$) disminuyen fuertemente `bikers`.
- **Meteorología continua**: `temp` ($\hat{\beta}_5 > 0$) aumenta la demanda; `hum` ($\hat{\beta}_6 < 0$) y `windspeed` ($\hat{\beta}_7 < 0$) la reducen.



(a) Residuals vs. Fitted (Grafico 1) **(b)** Normal Q–Q (Grafico 2) **(c)** Residuals vs. Leverage (Grafico 3)

Figura 2: Diagnostico del modelo final (Gráficos 1–3).

Diagnostico de supuestos. Los residuales exhiben leve curvatura y heterocedasticidad creciente a valores altos ajustados (Graficó 1), también colas algo mas pesadas que la normal pero buen alineamiento central (Graficó 2), y pocos puntos con alta influencia sin sobrepasar el umbral de Cook 0.5 (Graficó 3). Estos indicios no invalidan el modelo, pero sugieren que transformaciones leves o varianzas robustas podrían mejorar la inferencia.

Métricas Comparativas (Entrenamiento vs. Validación)

Evaluamos el desempeño del modelo final (*backward*) tanto en datos de entrenamiento como de validación. Los resultados indican un ajuste global satisfactorio, con un $R^2 = 0{,}776$ en entrenamiento que se mantiene en $R^2 = 0{,}747$ en validación, explicando aproximadamente el 75 % de la variabilidad en la demanda de bicicletas. Podemos observar ambos resultados en el siguiente cuadro.

Conjunto	R^2	R^2_{aj}	MAPE	RMSE
Entrenamiento (80 %)	0.7767	0.7712	20.39 %	654.96
Validación (20 %)	0.7478	0.7206	19.73 %	669.13

Cuadro 1: Métricas de desempeño del modelo final, (donde se puede confirmar la robustez del modelo.

El modelo demuestra una notable capacidad de generalización, con un error relativo promedio del 20 % y mínima variación en MAPE entre conjuntos, lo que confirma que aprende patrones genuinos en lugar de memorizar ruido. Si bien estos resultados son satisfactorios, sugieren oportunidades de mejora mediante la incorporación de no linealidades, interacciones clima-estación o modelos que capturen mejor la dependencia temporal residual.

Conclusiones

Hallazgos principales. El modelo lineal explico adecuadamente los factores que determinan la demanda diaria de bicicletas, mostrando un claro patrón estacional con mayor uso en las estaciones calidas. Las variables climáticas fueron las mas influyentes: la temperatura incrementa la demanda, mientras que las condiciones *cloudy/misty* y especialmente *light rain/snow* la reducen de forma significativa. También se observaron efectos negativos de la humedad, el viento y los feriados. El ajuste del modelo fue solido ($R^2 = 0{,}776$; $R^2_{\text{aj}} = 0{,}771$) y el MAPE fue similar entre entrenamiento (20.4 %) y prueba (19.7 %), por lo que es una buena generalización sin sobre-ajuste.

Limitaciones. Se detectaron incumplimientos de los supuestos de regresión, como heterocedasticidad, colas mss pesadas que la normal y cierta autocorrelación temporal. También podrían existir relaciones no lineales o interacciones no consideradas (por ejemplo, entre temperatura y estación). La agregación diaria reduce la variabilidad omite posibles eventos específicos que podrían mejorar el modelo.

Mejoras futuras. Se recomienda utilizar errores robustos o WLS para corregir la heterocedasticidad, e incorporar términos autorregresivos de la demanda para capturar la dinámica temporal. Ademas, introducir *splines* o interacciones (`season`×`temp`, `weathersit`) permitiría modelar no linealidades. Seria útil comparar con modelos alternativos y aplicar validación temporal.