Pontificia Universidad Católica de Chile
Facultad de Matemáticas
Departamento de Estadística

Segundo Semestre 2025

EYP2307 / EYP230i - ANÁLISIS DE REGRESIÓN
TAREA 2

Introducción

La realización de pronósticos precisos de la demanda de gas a corto y mediano plazo (1 y 4 semanas) es fundamental para garantizar un suministro eficiente y sostenible en Chile. Dado que el país presenta características geográficas y climáticas diversas, con distintas zonas que abarcan el norte, centro y sur, es esencial anticipar las variaciones en el consumo de gas para responder a las necesidades específicas de cada región.

En el norte, las temperaturas suelen ser más cálidas, mientras que en el centro y sur del país, especialmente en invierno, se experimentan bajas temperaturas que aumentan significativamente la demanda de gas para calefacción. El pronóstico a 1 semana permite a las empresas distribuidoras ajustar sus operaciones diarias y responder a cambios abruptos en las condiciones climáticas, mientras que el pronóstico a 4 semanas es clave para la planificación estratégica, logística y el aseguramiento de reservas de gas.

La capacidad de predecir la demanda de gas con precisión no solo garantiza un servicio eficiente para los clientes residenciales e industriales, sino que también permite optimizar los costos de almacenamiento, transporte y distribución. En un país con una geografía tan extensa y diversa como Chile, esta planificación es crucial para asegurar un abastecimiento confiable y minimizar los riesgos de escasez o sobreabastecimiento.

Objetivo

El objetivo de esta tarea es que cada grupo, de a lo más tres integrantes, construya un modelo de regresión múltiple para pronosticar la demanda de gas diaria de una zona particulares de chile, cuya información se encuentra en el archivo `Tarea_1_2025_02.xlsx`. Se espera que los grupos seleccionen cuidadosamente las variables que mejor expliquen la variabilidad en la demanda de gas de una de las siguientes series: comercial, industrial o granel. Se recomienda incorporar al data frame variables macroeconómicas como inflación, crecimiento, desempleo entre otras.

Descripción de las Variables

La base de datos proporcionada contiene las siguientes variables:

- **date**: Fecha de la semana correspondiente a la observación, en formato "YYYY-MM-DD".
- **day**: Año de la observación.
- **month**: Número de mes del año (1 a 12).
- **year**: Año de la observación.
- **comercial**: Demanda de gas comercial diaria, medida en kilogramos (kg).
- **granel**: Demanda de gas a granel diaria, medida en kilogramos (kg).
- **industrial**: Demanda de gas industrial diaria, medida en kilogramos (kg).
- **cloudcov_min**: Cobertura mínima de nubes observada durante el día (proporción entre 0 y 1).
- **cloudcov_max**: Cobertura máxima de nubes observada durante el día (proporción entre 0 y 1).
- **cloudcov_avg**: Cobertura promedio de nubes durante la semana (proporción entre 0 y 1).
- **temp_min**: Temperatura mínima registrada durante el día (en grados Celsius).
- **temp_max**: Temperatura máxima registrada durante el día (en grados Celsius).

- **temp_avg**: Temperatura promedio registrada durante el día (en grados Celsius).
- **rain**: Total de precipitación acumulada durante el día (en mm).
- **dayhumidity_min**: Humedad relativa mínima observada durante el día (proporción entre 0 y 1).
- **dayhumidity_avg**: Humedad relativa promedio durante el día (proporción entre 0 y 1).
- **dayhumidity_max**: Humedad relativa máxima observada durante el día (proporción entre 0 y 1).
- **visibility_min**: Visibilidad mínima registrada durante el día (en kilómetros).
- **visibility_avg**: Visibilidad promedio durante el día (en kilómetros).
- **visibility_max**: Visibilidad máxima registrada durante el día (en kilómetros).

Procedimiento general para la construcción del modelo

Cada grupo deberá seguir, al menos, las siguientes etapas en la construcción de su modelo de regresión múltiple:

1. **Selección serie de demanda.** Escoger una de las series de interés (`comercial`, `industrial` o `granel`) como variable respuesta. Separar el data frame en un 80 % de registros escogidos al azar para entrenar y el 20 % para test.

2. **Análisis exploratorio.** Realizar un análisis descriptivo de la serie de demanda seleccionada y de los posibles regresores (climáticos y macroeconómicos), incluyendo gráficos, diagramas de dispersión y medidas de resumen básicas.

3. **Evaluación de transformaciones (Box–Cox).** Evaluar, para la variable respuesta y para los regresores cuantitativos relevantes, la conveniencia de aplicar transformaciones de tipo Box–Cox con el fin de:

   - Mejorar la aproximación a la normalidad de los errores.
   - Estabilizar la varianza.
   - Mejorar la linealidad en la relación entre la respuesta y los regresores.

   Justificar brevemente las transformaciones finalmente seleccionadas (o la decisión de no transformar).

4. **Construcción y comparación de modelos.** Proponer un conjunto de modelos candidatos (por ejemplo, mediante esquemas *forward* o *backward*) y comparar su desempeño utilizando, al menos, los siguientes criterios:

   - Criterio de Información de Akaike (AIC).
   - Criterio de Información Bayesiano (BIC).
   - Medidas de ajuste basadas en proporción de variabilidad explicada (por ejemplo, R_1 y R^2 ajustado).
   - Comparación de modelos anidados mediante test F, según lo visto en clases.

   De manera opcional, los grupos podrán aplicar **Análisis de Componentes Principales (PCA)** sobre familias de variables altamente correlacionadas (por ejemplo, variables climáticas) para construir regresores derivados que resuman la información principal.

5. **Evaluación de supuestos e heterocedasticidad.** Analizar los residuos del modelo final para evaluar los supuestos clásicos de regresión lineal como la normalidad y presencia de heterocedasticidad.

6. **Corrección de heterocedasticidad (HC3).** En caso de evidenciar heterocedasticidad en los errores, recalcular los errores estándar de los coeficientes utilizando la matriz de varianzas-covarianzas robusta tipo HC3. Comparar las conclusiones inferenciales (significancia de los coeficientes, intervalos de confianza) bajo Mínimos Cuadrados Ordinarios y bajo la corrección HC3, comentando brevemente las diferencias.

7. **Validación con datos fuera de muestra (año 2019).** Además de la partición utilizada para la construcción y selección del modelo, cada grupo deberá evaluar el desempeño del modelo final utilizando la información del año **2019**, que no formó parte del proceso de ajuste (ver 2da hoja archivo excel).

- Generar predicciones para el año 2019 utilizando el modelo seleccionado.
- Comparar los valores observados y predichos mediante gráficos y métricas de error relevantes (por ejemplo, MAPE, RMSE o MAE).
- Comentar si la capacidad predictiva fuera de muestra es coherente con la obtenida durante el proceso de selección y validación interna.

8. **Extensión del informe.** El informe final deberá tener una extensión máxima de **4 páginas**, incluyendo portada y anexos. Se evaluará tanto la claridad en la presentación de resultados como la correcta justificación de las decisiones metodológicas.

Rúbrica de Evaluación (100 puntos)

La corrección del informe considerará los siguientes criterios:

- **1. Análisis exploratorio (15 pts).** Claridad en la descripción inicial de la serie de demanda seleccionada y de los regresores climáticos y macroeconómicos. Uso apropiado de gráficos, tablas y estadísticas descriptivas. Interpretación coherente y alineada con el fenómeno observado.
- **2. Transformaciones y diagnóstico preliminar (10 pts).** Evaluación correcta del procedimiento Box–Cox. Justificación clara de las transformaciones aplicadas o de la decisión de no transformar. Evidencia de haber analizado linealidad y varianza.
- **3. Construcción y comparación de modelos (25 pts).** Presentación de diferentes modelos candidatos ajustados sobre la partición de entrenamiento (80 %). Uso correcto de AIC, BIC, R_1, R^2 ajustado y test F para comparar modelos. Uso explícito y adecuado de la partición de validación (20 %) para evaluar y comparar el desempeño de los modelos. Selección bien argumentada del modelo final. Uso opcional de PCA: *se evaluará positivamente si se usa correctamente*, pero no penaliza no incluirlo.
- **4. Evaluación de supuestos y heterocedasticidad (10 pts).** Revisión clara y correcta de los supuestos del modelo final mediante análisis residual. Discusión sobre normalidad y heterocedasticidad. Conclusiones bien fundamentadas.
- **5. Corrección robusta HC3 (10 pts).** Aplicación correcta de errores estándar robustos HC3. Comparación entre inferencia clásica y robusta. Discusión pertinente sobre eventuales cambios en significancia.
- **6. Predicción fuera de muestra: año 2019 (15 pts).** Generación de predicciones para 2019 usando el modelo final. Comparación de observados vs. predichos mediante gráficos y métricas (MAPE, RMSE o MAE). Comentario razonado sobre la capacidad predictiva fuera de muestra y su coherencia con la validación interna.
- **7. Presentación, claridad y extensión (15 pts).** Informe ordenado, bien articulado y dentro del máximo de **4 páginas** (incluyendo portada y anexos). Uso adecuado de figuras y tablas integradas en el texto. Redacción clara con menos de 5 faltas ortográficas. Argumentación consistente y profesional.