FACULTAD DE MATEMÁTICAS
PONTIFICIA UNIVERSIDAD
CATÓLICA DE CHILE

EYP1113 - Probabilidad y Estadística

Capítulo 8: Regresión Lineal

Ricardo Aravena C. - Cristian Capetillo C. - Ingrid Guevara R.
Bladimir Morales T. - Ricardo Olea O. - Daniel Saavedra M.

Facultad de Matemáticas
Departamento de Estadística
Pontificia Universidad Católica de Chile

Segundo Semestre 2024

Contenido I

Introducción

La Inferencia vista anteriormente, puede ser abordada desde el punto de vista de Modelos Estadísticos.

Así por ejemplo, si $Y_1, \ldots, Y_n$ es una muestra aleatoria de una distribución Normal con media μ y varianza σ^2 ambos parámetros desconocidos.

Este experimento se puede escribir en términos del siguiente modelo:

$$Y_i = \mu + \varepsilon_i \quad i = 1, \ldots, n$$

donde ε_i tienen distribución normal con media cero y varianza σ^2.

Introducción

Al permitir que la media de Y varíe de manera funcional con respecto a una covariable X_i de la siguiente manera:

$$Y_i = \mu(X_i) + \varepsilon_i \quad i = 1, \ldots, n$$

Obtenemos el modelo de regresión simple.

Introducción

Se llama a

$$y_i = \mathrm{E}\left(Y_i \mid x_i\right) = \mu\left(x_i\right)$$

a la curva de regresión de Y sobre x.

Si la relación funcional es lineal en los parámetros, es decir,

$$\mu\left(X_i\right) = \beta_0 + \beta_1 X_i,$$

entonces el modelo se llama regresión lineal simple, y la curva de regresión esta dada por $y_i = \beta_0 + \beta_1 x_i$.

En cambio si

$$\mu\left(X_i\right) = \beta_0 X_i^{\beta_1},$$

el modelo sería de regresión No lineal simple, y la curva de regresión esta dada por $y_i = \beta_0 x_i^{\beta_1}$.

Regresión Lineal Simple

Consideremos el modelo de regresión lineal simple,

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \quad i = 1, \ldots, n$$

Supuestos:

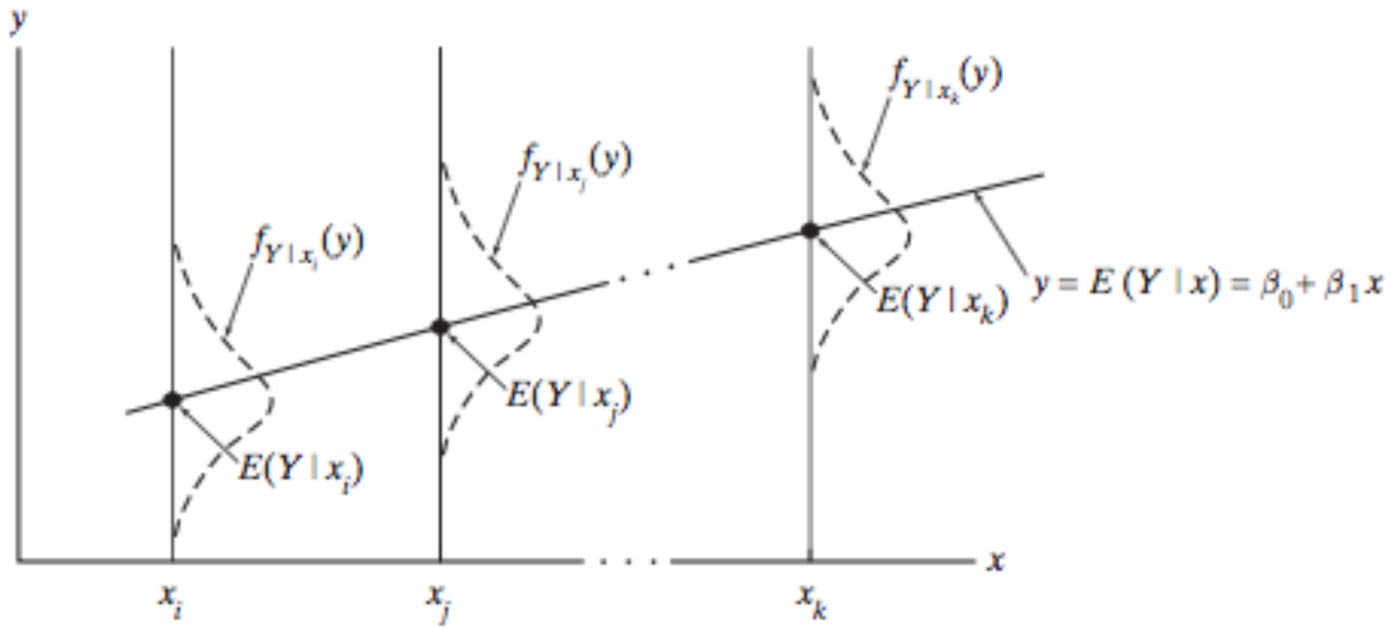
1. Linealidad: La media condicional de Y sobre x es lineal

$$y = \mathrm{E}(Y \mid x) = \beta_0 + \beta_1 x$$

2. Homocedasticidad: La varianza asociada a $f_{Y \mid x}(y)$ es la misma para todo x e iguala σ^2.
3. Independencia: Las distribuciones condicionales son variables aleatorias independientes para todo x.
4. Normalidad: $f_{Y \mid x}(y)$ tiene distribución normal para todo x.

Regresión Lineal Simple

La regresión lineal simple bajo los supuestos anteriores se ilustra en la siguiente figura:



Regresión Lineal Simple

Interpretación de los parámetros del modelo:

- β_0: intercepto, $\beta_0 = \mathrm{E}(Y \mid X = 0)$.
- β_1: pendiente, corresponde a la variación de $\mathrm{E}(Y \mid X = x)$ cuando x aumenta en una unidad.

Máxima Verosimilitud

Bajo los supuestos $(1)-(2)-(3)-(4)$ se tiene que $Y_i \mid x_i$ tiene distribución Normal con media $\mathrm{E}(Y_i \mid x_i) = \beta_0 + \beta_1 x_i$ y varianza σ^2 y además son independientes, entonces la función de verosimilitud de la muestra está dada por

$$L = \prod_{i=1}^{n} f_{Y|x_i}(y_i) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}(y_i-\beta_0-\beta_1 x_i)^2}$$

$$= (2\,\pi\sigma^2)^{-n/2} \exp\left\{-\frac{1}{2\,\sigma^2}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1\,x_i)^2\right\}$$

Regresión Lineal Simple

Estimación del Modelo

Los estimadores máximos verosímiles de los parámetros β_0, β_1 y σ^2 están dados por

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} x_i y_i - n\,\overline{y}\,\overline{x}}{\sum_{i=1}^{n} x_i^2 - n\,(\overline{x})^2}, \qquad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i \right)^2$$

Mínimos Cuadrados

Bajo los supuestos $(1)-(2)-(3)$, El método de mínimos cuadrados estimará los parámetros tales que minimicen la suma la distancia al cuadrado entre los valores observados de y_i y los asumidos por el ajuste de regresión, es decir, minimizar la función Δ^2 dada por

$$\Delta^2 = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$

Luego los EMCO (Estimadores de Mínimos Cuadrados Ordinarios) de β_0 y β_1, coinciden con los EMV.

Regresión Lineal Simple

Estimación del Modelo

Notar que el método de mínimos cuadrados no arroja estimación para σ^2.

Sin embargo, se estima a través de $s^2_{Y\,|\,x}$, que es un estimador insesgado de σ^2 dado por

$$s^2_{Y\,|\,x} = \frac{1}{n-2} \sum_{i=1}^{n} \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i \right)^2$$

Regresión Lineal Simple

Estimación del Modelo

Propiedades

Por supuestos $(1)-(2)-(3)$ los EMV *y* EMCO de β_0 y β_1 se tienen las siguientes propiedades:

- Insesgamiento

$$\mathrm{E}\left(\hat{\beta}_0\right) = \beta_0 \text{ y } \mathrm{E}\left(\hat{\beta}_1\right) = \beta_1$$

- Varianza

$$\mathrm{Var}\left(\hat{\beta}_0\right) = \frac{\sigma^2 \sum_{i=1}^{n} x_i^2}{n \sum_{i=1}^{n} (x_i - \bar{x})^2} \quad \text{y} \quad \mathrm{Var}\left(\hat{\beta}_1\right) = \frac{\sigma^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

Teorema de Gauss Markov

Dado los supuestos $(1)-(2)-(3)$ los EMCO de β_0 y β_1 son los mejores estimadores lineales y con menor varianza entre los estimadores lineales e insesgados.

Regresión Lineal Simple

Estimación del Modelo

Si además agregamos el supuesto (4) se tiene que $\hat{\beta}_0$ y $\hat{\beta}_1$ distribuyen Normal, por lo tanto

$$\hat{\beta}_0 \sim \text{Normal}\left(\beta_0, \frac{\sigma^2 \sum_{i=1}^n x_i^2}{n \sum_{i=1}^n (x_i - \bar{x})^2}\right)$$

$$\hat{\beta}_1 \sim \text{Normal}\left(\beta_1, \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2}\right)$$

Y además,

- $\hat{\beta}_1$, $\bar{Y}$, $\hat{\sigma}^2$ son muatuamente independientes.
- $$\sum_{i=1}^n \left(\frac{y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i}{\sigma}\right)^2 = \frac{n\hat{\sigma}^2}{\sigma^2} = \frac{(n-2)\, s^2_{Y\,|\,x}}{\sigma^2} \sim \chi^2(n-2).$$

Regresión Lineal Simple

Inferencia

A partir de lo anterior se puede hacer inferencia sobre los parámetros del modelo, y poder construir IC o realizar test de hipótesis acerca de ellos.

Sea $\mathrm{E}(Y \mid x) = \beta_0 + \beta_1 x$ el modelo de regresión lineal simple, y $\hat{\beta}_0, \hat{\beta}_1$ los EMV de β_0 y β_1.

Entonces el estadístico,

$$\frac{\hat{\beta}_1 - \beta_1}{s_{Y \mid x} \Big/ \sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2}} \sim \text{t-Student}(n-2)$$

con $s^2_{Y \mid x} = \frac{1}{n-2} \sum_{i=1}^{n} \left(Y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i \right)^2$, estimador insesgado de σ^2.

Regresión Lineal Simple

Inferencia

También se tiene que

$$\frac{\hat{\beta}_0 - \beta_0}{\frac{s_{Y\,|\,x}\sqrt{\sum_{i=1}^{n} x_i^2}}{\sqrt{n\sum_{i=1}^{n}(x_i-\bar{x})^2}}} \sim \text{t-Student}(n-2)$$

y

$$\frac{(n-2)s^2_{Y\,|\,x}}{\sigma^2} \sim \chi^2(n-2)$$

Regresión Lineal Simple

Análisis de la Varianza

En un modelo de regresión existen dos fuentes que explican la variación de los valores observados de Y (variación total)

- Una fuente es debido la regresión, representada por la x,

$$\text{SCR} = \sum_{i=1}^{n} \left(y_i' - \bar{y}\right)^2$$

- Otra fuente es la variación de y_i que no ha sido explicada en el modelo por las x_i,

$$\text{SCE} = \sum_{i=1}^{n} \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right)^2$$

Regresión Lineal Simple

Análisis de la Varianza

De esta manera, la variación total de Y, dada por SCT, puede ser escrita como

$$\text{SCT} = \quad \text{SCR} \quad + \quad \text{SCE}$$

$$\sum_{i=1}^{n} (y_i - \bar{y})^2 = \sum_{i=1}^{n} (y_i' - \bar{y})^2 + \sum_{i=1}^{n} \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i \right)^2$$

Regresión Lineal Simple

Análisis de la Varianza

Tabla ANOVA

Fuente	gl	SC	CM	F
Regresión	1	SCR	$\frac{\text{SCR}}{1}$	$\frac{\text{MCR}}{\text{MCE}}$
Error	$n-2$	SCE	$\frac{\text{SCE}}{n-2}$	
Total	$n-1$	SCT		

Con $F = \frac{\text{MCR}}{\text{MCE}} \sim F(1, n-2)$

Regresión Lineal Simple

Coeficiente de Determinación

Coeficiente de determinación R^2:

$$R^2 = \frac{SCR}{SCT} = \frac{\sum_{i=1}^{n} (y_i' - \bar{y})^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2} = 1 - \frac{SCE}{SCT}$$

Coeficiente de determinación R^2 ajustado:

$$r^2 = 1 - \frac{s^2_{Y|x}}{s^2_Y} = 1 - \frac{(n-1)}{(n-2)} \frac{SCE}{SCT} = \overline{R}^2$$

Ambos se interpretan como la proporción de variación total que es explicada por el modelo de regresión lineal.

Regresión Múltiple

Definición del Modelo

El modelo de regresión múltiple (MRLM) se define de la siguiente forma:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots + \beta_k X_{ki} + \varepsilon_i \qquad i = 1, \ldots, n$$

donde Y es la variable dependiente, X_j, $j = 1, \ldots, k$ son las covariables del modelo, y los β_j son coeficientes constantes del modelo, y las ε_i son variables aleatorias tales que cumplen con:

$$\mathrm{E}(\varepsilon_i) = 0$$
$$\mathrm{Var}(\varepsilon_i) = \sigma^2$$
$$\mathrm{Cov}(\varepsilon_i, \varepsilon_j) = 0 \quad \forall i \neq j$$

El objetivo es poder predecir $\mathrm{E}(Y \mid x_1, \ldots, x_k)$ a partir de k variables inde- pendientes observadas: x_j

Regresión Múltiple

Definición del Modelo

Observaciones:

- El modelo tiene $k+2$ parámetros a estimar: $\beta_0, \beta_1, \ldots, \beta_k, \sigma^2$
- El coeficiente β_j, con $j = 1, \ldots, k$ corresponde a la variación de $E(Y \mid x_1, \ldots, x_k)$, cuando x_j aumenta en una unidad y el resto de las variables no cambian.
- β_0 : corresponde al valor medio $\mathrm{E}(Y \mid x_1, \ldots, x_k)$ cuando todas las covariables x_j son cero.
- Al igual que en el caso MRLS, el MRLM debe ser lineal en los parámetros β_j, y no necesariamente en las variables X_j.

Regresión Múltiple

Estimación del modelo

Dado el modelo de Regresión Lineal Múltiple definido en (1), las estimaciones de mínimos cuadrados de los coeficientes $\beta_0, \beta_2, \ldots, \beta_K$ son los valores $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_K$ para los que la suma de los cuadrados de las desviaciones entre el valor observado y_i y los asumidos por el ajuste de regresión,

$$SCE = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_{i1} - \beta_2 x_{i2} - \cdots - \beta_k x_{ik})^2$$

es la menor posible.

Regresión Múltiple

Estimación del modelo

Para determinar los EMCOS de $\beta_0, \ldots, \beta_k$, se deriva SCE parcialmente respecto a $\beta_0, \beta_1, \ldots, \beta_k$ obtenièndose las siguientes $(k+1)$ ecuaciones normales que se deben resolver:

$$\frac{\partial SCE}{\partial \beta_0} = 0, \quad \frac{\partial SCE}{\partial \beta_1} = 0, \quad \ldots\ldots \frac{\partial SCE}{\partial \beta_k} = 0$$

Regresión Múltiple

Estimación del modelo

La solución $\left(\hat{\beta}_0, \ldots, \hat{\beta}_K\right)$ satisface el sistema lineal de $K+1$ ecuaciones,

$$n\hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^{n} x_{1i} + \ldots + \hat{\beta}_K \sum_{i=1}^{n} x_{Ki} = \sum_{i=1}^{n} y_i$$

$$\hat{\beta}_0 \sum_{i=1}^{n} x_{ji} + \hat{\beta}_1 \sum_{i=1}^{n} x_{1i}x_{ji} + \ldots + \hat{\beta}_k \sum_{i=1}^{n} x_{ki}x_{ji} = \sum_{i=1}^{n} y_i x_{ji}$$

$$j = 1, .., K$$

Si se considera una expresión matricial para el MRLM, entonces se puede obtener una expresión simple para los estimadores MCO.

Regresión Múltiple

Estimación del modelo

Notación Matricial del Modelo

$$
\begin{aligned}
Y_1 &= \beta_0 + \beta_1 X_{11} + \beta_2 X_{21} + \cdots + \beta_K X_{K1} + \varepsilon_1 \\
Y_2 &= \beta_0 + \beta_1 X_{12} + \beta_2 X_{22} + \cdots + \beta_K X_{K2} + \varepsilon_2 \\
&\vdots \\
Y_n &= \beta_0 + \beta_1 X_{1n} + \beta_2 X_{2n} + \cdots + \beta_K X_{K2} + \varepsilon_n
\end{aligned}
$$

Este sistema de ecuaciones puede expresarse matricialmente de la siguiente forma:

Regresión Múltiple

Estimación del modelo

$$\begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix} = \begin{bmatrix} 1 & X_{11} & X_{21} & \cdots & X_{K1} \\ 1 & X_{12} & X_{22} & \cdots & X_{K2} \\ 1 & X_{13} & X_{23} & \cdots & X_{K3} \\ \vdots & & & & \\ 1 & X_{1n} & X_{2n} & \cdots & X_{Kn} \end{bmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

$$\boldsymbol{Y} = \boldsymbol{X}\beta + \varepsilon$$

Donde ε e $\boldsymbol{Y}$ son vectores de $n \times 1$, $\boldsymbol{X}$ es una matriz de $n \times (K+1)$ y el Rango de $\boldsymbol{X}$ debe ser de rango columna completo $(K+1)$

Regresión Múltiple

Estimación del modelo

Luego

$$\begin{aligned} SCE &= \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_{i1} - \beta_2 x_{i2} - \cdots - \beta_k x_i k)^2 \\ &= (\boldsymbol{Y} - \boldsymbol{X}\beta)^t (\boldsymbol{Y} - \boldsymbol{X}\beta) \end{aligned}$$

Debemos derivar SCE parcialmente respecto a $\beta_0, \beta_1, \ldots, \beta_K$ e igualar 0, esto es:

$$X^t X \beta = X^t Y \Rightarrow \hat{\beta} = \left(X^t X\right)^{-1} X^t Y$$

De esta manera, la regresión ajustada de Y sobre $X_1, X_2, \ldots, X_K$ está dada por:

$$y_i' = \mathrm{E}\left(Y \mid \widehat{x_1, \ldots, x_K}\right) = \hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \hat{\beta}_2 x_{2i} + \ldots + \hat{\beta}_K x_{Ki}$$

Regresión Múltiple

Estimación del modelo

Estimación de σ^2

Dado el modelo de regresión poblacional múltiple

$$Y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_K x_{Ki} + \varepsilon_i$$

y los supuestos habituales de la regresión, sea σ^2 la varianza común del término de error, ε_i. Entonces, una estimación insesgada de esta varianza es

$$s^2_{Y|x} = \frac{SCE}{n - k - 1}$$

donde k es el número de variables predictoras.

Regresión Múltiple

Estimación del modelo

Los EMCOS de $\boldsymbol{\beta}$ tienen las siguientes propiedades:

- $\hat{\boldsymbol{\beta}}$ es insesgado, es decir, $\mathrm{E}(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta}$.
- $\mathrm{Var}(\hat{\boldsymbol{\beta}}) = \sigma^2 \left(\boldsymbol{X}^t \boldsymbol{X}\right)^{-1}$.
- Si se asume Normalidad en ε se tiene que
  - $(n-k-1)\dfrac{s^2_{Y\,|\,x}}{\sigma^2} \sim \chi^2_{n-k-1}$
  - $\hat{\boldsymbol{\beta}} \sim \text{Normal}\left(\boldsymbol{\beta}, \sigma^2 \left(\boldsymbol{X}^t \boldsymbol{X}\right)^{-1}\right)$

  Luego para cada i, $\hat{\beta}_i \sim N\left(\beta_i, \sigma^2 c_{ii}\right)$, donde c_{ii} corresponde al elemento ii de la matriz $\left(\boldsymbol{X}^t \boldsymbol{X}\right)^{-1}$

Regresión Múltiple

Inferencia en el modelo

La desviación estándar de $\hat{\beta}_i$: $\sigma\sqrt{c_{ii}}$ puede ser estimada por

$$se_{\hat{\beta}_i} = s_{Y\,|\,x}\sqrt{c_{ii}}$$

A partir de lo anterior, se pueden construir intervalos de confianza y test de hipótesis para β.

Bajo normalidad, se puede demostrar que

$$\frac{\hat{\beta}_i - \beta_i}{se_{\hat{\beta}_i}} \sim t_{n-k-1}$$

Regresión Múltiple

Coeficiente de Determinación y Análisis de la Varianza

Al igual que en MRLS, la variabilidad del modelo puede dividirse en los componentes

$$SCT = SCR + SCE$$

las que se definen de la siguiente manera

$$\begin{aligned} SCT &= \sum_{i=1}^{n} \left(Y_i - \bar{Y}\right)^2 \\ &= \sum_{i=1}^{n} \left(y_i' - \bar{Y}\right)^2 + \sum_{i=1}^{n} \left(Y_i - y'\right)^2 \end{aligned}$$

Regresión Múltiple

Coeficiente de Determinación y Análisis de la Varianza

Coeficiente de Determinación: R^2

Esta descomposición puede interpretarse como

$$\text{Variabilidad Muestral Total} = \text{Variabilidad Explicada} + \text{Variabilidad No Explicada}$$

El coeficiente de determinación, R^2, de la regresión ajustada es la proporción de la variabilidad muestral total explicada por la regresión

$$R^2 = \frac{SCR}{SCT} = 1 - \frac{SCE}{SCT}$$

y se deduce que

$$0 \leq R^2 \leq 1$$

Coeficiente de Determinación Ajustado: $\bar{R}^2$

El coeficiente de determinación ajustado, $\bar{R}^2$, se define de la forma siguiente:

$$r^2 = 1 - \frac{\text{SCE}/(n-K-1)}{\text{SCT}/(n-1)} = \bar{R}^2$$

Utilizamos esta medida para tener en cuenta el hecho de que las variables independientes irrelevantes provocan una pequeña reducción de la suma de los cuadrados de los errores.

Por lo tanto, el $\bar{R}^2$ ajustado permite comparara mejor los modelos de regresión múltiple que tienen diferentes números de variables independientes.

Coeficiente de Correlación Múltiple

El coeficiente de correlación múltiple es la correlación entre el valor predi- cho y el valor observado de la variable dependiente.

$$R = \text{Cor}(y', Y) = \sqrt{R^2}$$

y es igual a la raíz cuadrada del coeficiente múltiple de determinación. Utilizamos R como otra medida de la fuerza de la relación entre variable dependiente y las variables independientes.

Por lo tanto, es comparable a la correlación entre Y y X en la regresión simple.

Regresión Múltiple

Coeficiente de Determinación y Análisis de la Varianza

Al igual que en el MRLS se puede construir la Tabla de Análisis de la Varianza (ANOVA)

Tabla ANOVA

Fuente	gl	SC	CM	F
Regresión	k	SCR	$\frac{\text{SCR}}{k}$	$\frac{\text{MCR}}{\text{MCE}}$
Error	$n-k-1$	SCE	$\frac{\text{SCE}}{n-k-1}$	
Total	$n-1$	SCT		

Con $F = \frac{\text{MCR}}{\text{MCE}} \sim F(k, n-k-1)$

Selección de Modelo

- *Método jerárquico*: Se introducen unos predictores determinados en un orden determinado.
- *Método de entrada forzada*: se introducen todos los predictores simultáneamente.
- *Método paso a paso (stepwise)*: emplea criterios matemáticos para decidir qué predictores contribuyen significativamente al modelo y en qué orden se introducen.

Forward – Backward – mixto

Selección de Modelo

El método paso a paso requiere de algún criterio matemático para determinar si el modelo mejora o empeora con cada incorporación o extracción.

Existen varios parámetros empleados, de entre los que destacan el Cp, AIC, BIC, test F y R^2 ajustado.

Multicolinealidad

Para determinar la existencia de colinealidad o multicolinealidad entre los predictores de un modelo de regresión, se han desarrollado numerosas reglas prácticas que tratan de determinar en qué medida afecta a la estimación y contraste de un modelo:

- Si el coeficiente de determinación R^2 es alto pero ninguno de los predictores resulta significativo, hay indicios de colinialidad.
- Calcular una matriz de correlación en la que se estudia la relación lineal entre cada par de predictores.
- Generar modelos de regresión lineal simple entre cada uno de los predictores frente al resto. Si en alguno de los modelos el coeficiente de determinación R^2 es alto, estaría señalando a una posible colinialidad.

Multicolinealidad

- Tolerancia (TOL) y Factor de Inflación de la Varianza (VIF). Se trata de dos parámetros que vienen a cuantificar lo mismo (uno es el inverso del otro).

  El VIF de cada predictor se calcula según la siguiente fórmula:

$$VIF_{\hat{\beta}_j} = \frac{1}{1 - R^2}$$

$$\text{Tolerancia}_{\hat{\beta}_j} = \frac{1}{VIF_{\hat{\beta}_j}}$$

  Donde R^2 se obtiene de la regresión del predictor X_j sobre los otros predictores.

Multicolinealidad

i. $VIF = 1$ (Ausencia total de colinealidad)
ii. $1 < VIF < 5$ (La regresión puede verse afectada por cierta colinealidad)
iii. $5 < VIF < 10$ (Causa de preocupación)
iv. El termino tolerancia es $1/VIF$ por lo que los límites recomendables están entre 0.1 y 1.

Independencia

Los valores de cada observación son independientes de los otros, esto es especialmente importante de comprobar cuando se trabaja con mediciones temporales.

Se recomienda representar los residuos ordenados acorde al tiempo de registro de las observaciones, si existe un cierto patrón hay indicios de auto-correlación. Función acf() de R.

También se puede emplear el test de hipótesis de Durbin-Watson y Box-Ljung.

Outliers, Leverage e influyentes

Outlier: Observaciones que no se ajustan bien al modelo. Residuo es excesivamente grande. En una representación bidimensional se corresponde con desviaciones en el eje Y.

Observación con alto leverage: Observación con un valor extremo para alguno de los predictores. En una representación bidimensional se corresponde con desviaciones en el eje X. Son potencialmente puntos influyentes.

Observación influyente: Observación que influye sustancialmente en el modelo, su exclusión afecta al ajuste. No todos los outliers tienen por qué ser influyentes.

En R se dispone de la función outlierTest() del paquete car y de las funciones influence.measures(), influencePlot() y hatvalues() para identificar las observaciones más influyentes en el modelo.

Outliers, Leverage e influyentes

Distancia de Cook: Medida muy utilizada que combina, en un único valor, la magnitud del residuo y el grado de leverage. Valores de Cook mayores a 1 suelen considerarse como influyentes.

Cambio en los coeficientes de regresión: Se trata de un proceso iterativo en el que cada vez se excluye una observación distinta y se reajusta el modelo. En cada iteración se registra la diferencia en los coeficientes de regresión con y sin la observación, dividida entre el SE del predictor en el modelo sin la observación.

$$\text{Dfbetas}_i = \frac{\hat{\beta} - \hat{\beta}_i}{SE_{\hat{\beta}_i}}$$

Outliers, Leverage e influyentes

Al tratarse de un valor estandarizado, es sencillo identificar que observaciones influyen más y en que magnitud.

$$|\text{Dfbetas}| > \frac{2}{\sqrt{n}}$$

La función dfbeta() realiza esta comparación.

Aplicación

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.

Aplicación

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.

Aplicación

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 en una zona particular de chile, cuya información se encuentra en el archivo `demanda_gas.xlsx`.

Se espera que se seleccionen cuidadosamente las variables que mejor expliquen la variabilidad en la demanda de gas de manera individual, teniendo en cuenta la importancia de evitar utilizar simultáneamente variables que representan factores similares, como la temperatura promedio, máxima y mínima.

Aplicación

Información

- **date**: Fecha de la semana correspondiente a la observación, en formato "YYYY-MM-DD".
- **year**: Año de la observación.
- **week**: Número de la semana del año (1 a 52/53).
- **residentialgasdemand**: Demanda de gas residencial total durante la semana, medida en kilogramos (kg).
- **cloudcov_min**: Cobertura mínima de nubes observada durante la semana (proporción entre 0 y 1).
- **cloudcov_max**: Cobertura máxima de nubes observada durante la semana (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 la semana (en grados Celsius).
- **temp_max**: Temperatura máxima registrada durante la semana (en grados Celsius).
- **temp_avg**: Temperatura promedio registrada durante la semana (en grados Celsius).
- **rain**: Total de precipitación acumulada durante la semana (en mm).
- **rain_n**: Número de días de la semana en que se registró lluvia.
- **public_holiday_n**: Número de días feriados durante la semana.
- **dayhumidity_min**: Humedad relativa mínima observada durante la semana (proporción entre 0 y 1).
- **dayhumidity_avg**: Humedad relativa promedio durante la semana (proporción entre 0 y 1).
- **dayhumidity_max**: Humedad relativa máxima observada durante la semana (proporción entre 0 y 1).
- **visibility_min**: Visibilidad mínima registrada durante la semana (en kilómetros).
- **visibility_avg**: Visibilidad promedio durante la semana (en kilómetros).
- **visibility_max**: Visibilidad máxima registrada durante la semana (en kilómetros).
- **n_day**: Número total de días registrados en la semana (puede ser útil para casos de datos incompletos).
- **month**: Mes correspondiente a la variable fecha que se encuentra en la variable date.
- **season**: Estación del año a la que corresponde la observación semanal (ej. "summer", "fall").