Modelos de Regresión

Ana X. Ezquerro

ana.ezquerro@udc.es, GitHub

Grado en Ciencia e Ingeniería de Datos
Universidad de A Coruña (UDC)

Curso 2020-2021

Tabla de Contenidos

Tema 0: Introducción a la Regresión

0.1. Defínción de modelo de regresión

Un **modelo de regresión** explica la relación funcional entre la media de una variable de interés Y y un conjunto de variables explicativas $(X_1, ..., X_p)$.

$$\mathbb{E}(Y|X_1, ..., X_p) = m(X_1, ..., X_p)$$

Dada una muestra de realizaciones del vector $(p+1)$-dimensional $(X_1, ..., X_p, Y)$, el primer objetivo es determinar la función m que mejor se ajusta a la muestra de acuerdo a algún criterio de optimalidad.

0.2. Estimación paramétrica

En la estimación paramétrica se asume que m pertenece a una familia indexada por un conjunto finito de parámetros y el objetivo es estimar estos parámetros. Por ejemplo, es frecuente asumir que un modelo de regresión sea lineal y estimar sus coeficientes:

$$m(X) = \beta_0 + \beta_1 X$$

Ventajas: Si el modelo supuesto es correcto, es sencillo de interpreta y se estima con alta precisión y bajo coste computacional.

Desventajas: Si el modelo supuesto no es correcto: es inconsistente y proporciona una foto engañosa de la relación funcional buscada.

0.3. Estimación no paramétrica

En la estimación paramétrica, confinar el modelo subyacente a una familia paramétrica puede ser excesivamente rígido (por ejemplo, para explicar desviaciones locales o cambios de tendencia).

Las técnicas no paramétricas cancelan esta restricción, permitiendo que *sean los datos muestrales* quienes guien el proceso de inferencia (se suele decir, "dejan hablar a los datos por sí mismos). Solamente se requiere que m satisfaga la hipótesis de regularidad (que sea continua y diferenciable).

Ventajas: Versatilidad para ajustar relaciones funcionales complejas, robustez frente outliers, gran utilidad para análisis exploratorio.

Inconvenientes: Mayor coste computacional y mal comportamiento en dimensiones altas.

Tema 1: Regresión Lineal Simple (RLS)

1.1. Definición del modelo de regresión lineal simple

El modelo de regresión simple estudia la relación lineal entre **una variable respuesta** Y y **una variable regresora** X. Sea $\{(X_1, Y_1), ..., (X_n, Y_n)\}$ una muestra de realizaciones independientes de (X, Y), asumamos la relación estocástica:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i, \quad \text{con } \varepsilon_i \text{ i.i.d. } \sim N(0, \sigma), \text{ para } i = 1, ..., n$$

- β_0 (intercept) es el valor esperado de Y cuando $X = 0$.
- β_1 (pendiente) es la tasa de cambio del valor esperado de Y por unidad de incremento en X.
- σ es la desviación estándar de las respuestas para un valor arbitrario de X.

1.1.1. Hipótesis estructurales sobre el modelo RLS

Para $i = 1, ..., n$, definimos las siguientes hipótesis estructurales sobre las que se construye el modelo de regresión lineal simple:

- **Linealidad**: $\mathbb{E}(Y|X) = m(X) = \beta_0 + \beta_1 X \iff \mathbb{E}(\varepsilon_i) = 0$
- **Homoscedasticidad**: $\text{Var}(Y|X) = \sigma^2 \iff \text{Var}(\varepsilon_i) = \sigma^2$
- **Normalidad**: $Y|X \sim N(\beta_0 + \beta_1 X, \sigma) \iff \varepsilon_i \sim N(0, \sigma)$
- **Independencia**: $\text{Cov}(Y_i, Y_j) = 0, \ i \neq j \iff \text{Cov}(\varepsilon_i, \varepsilon_j) = 0, \ i \neq j$

1.1.2. Estimación por mínimos cuadrados ordinarios (OLS)

Sea $\{(X_1, Y_1), ..., (X_n, Y_n)\}$ una muestra de n observaciones. Para determinar la recta de la forma $Y = \beta_0 + \beta_1 X$ que minimiza la suma de cuadrados de las diferencias entre las observaciones (Y_i) y los valores pronosticados por dicha recta $(\widehat{Y}_i)$ se deben estimar los parámetros que la definen $(\widehat{\beta}_0, \widehat{\beta}_1)$.

- Muestra: $\{(X_1, Y_1), ..., (X_n, Y_n)\}$
- Valores ajustados por la recta: $\widehat{Y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 X_i$
- Residuos: $e_i = Y_i - \widehat{Y}_i = Y_i - (\widehat{\beta}_0 + \widehat{\beta}_1 X_i)$.
- Suma de los residuos al cuadrado:

$$\text{SS(R)} = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (Y_i - \widehat{Y}_i)^2 = \sum_{i=1}^{n} (Y_i - \widehat{\beta}_0 - \widehat{\beta}_1 X_i)^2$$

A través de un proceso de optimización y búsqueda de mínimos en esa función (véase en el apéndice la estimación por OLS) llegamos a las siguientes conclusiones.

Resultados de la estimación OLS

- Ecuaciones normales de la regresión: $\sum_{i=1}^{n} e_i = 0, \qquad \sum_{i=1}^{n} e_i X_i = 0$
- Estimación de los parámetros de la recta: $\widehat{\beta}_0 = \overline{Y} - \widehat{\beta}_1 \overline{X}, \qquad \widehat{\beta}_1 = \dfrac{S_{XY}}{S_X^2}$
- Recta de regresión por OLS: $\widehat{Y} = \overline{Y} + \dfrac{S_{XY}}{S_X^2}(X - \overline{X})$

1.1.3. Varianza y residuos del modelo de regresión

Los **residuos** del modelo de regresión se definen como la diferencia entre un valor observado de la variable respuesta (Y) y el correspondiente valor ajustado ($\widehat{Y}_i$) por la recta de regresión estimada. Según las hipótesis de normalidad y homoscedasticidad, los residuos tienen media cero y varianza constante.

$$e_i = Y_i - \widehat{Y}_i = Y_i - \widehat{m}(X_i) \sim N(0, \sigma)$$

La **varianza** del modelo de regresión se define como:

$$\text{Var}(Y_i|X_i) = \text{Var}\big(m(X_i) + e_i|X_i\big) = \text{Var}(e_i|X_i) = \sigma^2_{X_i}$$

- Como se asume que la varianza de los errores es **constante** e **independiente** de X, entonces:

$$\text{Var}(e_i|X_i) = \sigma^2_{X_i} = \sigma^2$$

- El parámetro σ^2 mide la dispersión de las respuesta Y_i en torno a la recta que subyace. Cuanto menor es el valor de σ^2 mayor es la calidad del ajuste y la representatividad de la recta.

1.1.4. Estimación de la varianza del modelo

El **estimador máximo-verosímil** (pero sesgado) de σ^2 es:

$$\widehat{\sigma}^2_{\text{MV}} = \frac{1}{n}\text{SS(R)} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \widehat{Y}_i)^2 = \frac{1}{n}\sum_{i=1}^{n} e_i^2$$

Los n residuos del modelo están sujetos a dos restricciones lineales (las ecuaciones normales del ajuste del modelo RLS), por tanto, **tienen $n-2$ grados de libertad**.

El estimado insesgado de la varianza del modelo σ^2 es:

$$\widehat{\sigma}^2 = \text{MSS(R)} = \frac{1}{n-2}\text{SS(R)} = \frac{1}{n-2}\sum_{i=1}^{n} e_i^2$$

1.2. Inferencia sobre el modelo

1.2.1. Distribución de los estimadores de los parámetros

Mediante una serie de demostraciones que parten de los resultados obtenidos en la estimación por mínimos cuadrados, se demuestra que los estimadores $\widehat{\beta}_0$ y $\widehat{\beta}_1$ se pueden expresar como **combinación lineal de las observaciones** Y_i (las cuales son variables normales independientes) y, por tanto, siguen una distribución normal.

$$\widehat{\beta}_1 \sim N\left(\beta_1, \frac{\sigma}{S_X\sqrt{n}}\right) \iff \frac{\widehat{\beta}_1 - \beta_1}{\frac{\sigma}{S_X\sqrt{n}}} \sim N(0,1)$$

$$\widehat{\beta}_0 \sim N\left(\beta_0, \sigma\sqrt{\frac{1}{n} + \frac{\overline{X}^2}{nS_X^2}}\right) \iff \frac{\widehat{\beta}_0 - \beta_0}{\sigma\sqrt{\frac{1}{n} + \frac{\overline{X}^2}{nS_X^2}}} \sim N(0,1)$$

Podemos obtener la distribución de $\widehat{\sigma}^2$ siguiendo un razonamiento similar. Como $e_i \sim N(0,\sigma)$, si tipificamos los residuos y realizamos la suma de estos al cuadrado, deducimos que:

$$(n-2)\frac{\widehat{\sigma}^2}{\sigma} = (n-2)\frac{\text{MSS(R)}}{\sigma^2} = \frac{\text{SS(R)}}{\sigma^2} \sim \chi^2_{n-2}$$

Comentarios

- $\widehat{\beta}_0$, $\widehat{\beta}_1$ y σ^2 son insesgados.
- $\widehat{\beta}_0$, y $\widehat{\beta}_1$ son de varianza mínima entre los estimadores lineales insesgados (Teorema de Gauss-Markov).
- $\widehat{\beta}_0$ y $\widehat{\beta}_1$ no son independientes. $\implies \text{Cov}(\widehat{\beta}_0, \widehat{\beta}_1) = -\frac{\overline{X}\sigma^2}{nS_X^2}$
- $\text{Var}(\widehat{\beta}_1) = \frac{\sigma^2}{nS_X^2}$, por tanto $\text{Var}(\widehat{\beta}_1)$ aumenta con la varianza del modelo (σ^2) y disminuye con n y la varianza de X.

En la práctica, como σ^2 es desconocido, se reemplaza por su estimador $\widehat{\sigma}^2 = \text{MSS(R)}$. Se reesriben las distribuciones de $\widehat{\beta}_0$ y $\widehat{\beta}_1$:

$$\frac{\widehat{\beta}_1 - \beta_1}{\sqrt{\dfrac{\text{MSS(R)}}{nS_X^2}}} \sim t_{n-2}, \qquad \frac{\widehat{\beta}_0 - \beta_0}{\sqrt{\text{MSS(R)}}\sqrt{\dfrac{1}{n} + \dfrac{\overline{X}^2}{nS_X^2}}} \sim t_{n-2}$$

1.2.2. Intervalos de confianza para β_0, β_1 y σ^2

Para $i = 0, 1$ se tiene que:

$$\frac{\widehat{\beta}_i - \beta_i}{\widehat{\sigma}(\widehat{\beta}_i)} \sim t_{n-2}, \quad \text{siendo} \quad \begin{cases} \widehat{\sigma}(\widehat{\beta}_0) = \sqrt{\text{MSS(R)}}\sqrt{\dfrac{1}{n} + \dfrac{\overline{X}^2}{nS_X^2}} \\ \widehat{\sigma}(\widehat{\beta}_1) = \sqrt{\dfrac{\text{MSS(R)}}{nS_X^2}} \end{cases}$$

Sea $t_{g,\alpha}$ el valor tal que $\mathbb{P}(t_g > t_{g,\alpha}) = \alpha$:

$$\text{IC}_{(1-\alpha)}(\beta_i) = \left(\widehat{\beta}_i - t_{n-2,\alpha/2}\ \widehat{\sigma}(\widehat{\beta}_i), \quad \widehat{\beta}_i + t_{n-2,\alpha/2}\ \widehat{\sigma}(\widehat{\beta}_i)\right)$$

Sea $\chi^2_{g,a}$ el valor tal que $\mathbb{P}(\chi^2_g > \chi^2_{g,\alpha}) = \alpha$, entonces:

$$\text{IC}_{(1-\alpha)}(\sigma^2) = \left(0, \ \frac{(n-2)\text{MSS(R)}}{\chi^2_{n-2,1-\alpha}}\right) = \left(0, \ \frac{\text{SS(R)}}{\chi^2_{n-2,1-\alpha}}\right)$$

1.2.3. Contrastes de hipótesis sobre β_0 y β_1

Una vez establecidas las distribuciones de los parámetros, para $i = 0, 1$, podemos resolver el contrastes de hipótesis bilaterales del tipo:

$$\begin{cases} H_0 : \beta_i = \beta^* \\ H_1 : \beta_i \neq \beta^* \end{cases}$$

Utilizando como estadístico de contraste $\dfrac{\widehat{\beta}_i - \beta^*}{\widehat{\sigma}(\widehat{\beta}_i)}$ que, bajo H_0, se distribuye según una t_{n-2}.

Estadísticos de contrastes para β_0 y β_1

$$\widehat{T}_0 = \frac{\widehat{\beta}_0}{\sqrt{\text{MSS(R)}}\sqrt{\dfrac{1}{n} + \dfrac{\overline{X}^2}{nS_X^2}}}, \qquad \widehat{T}_1 = \frac{\widehat{\beta}_1}{\sqrt{\dfrac{\text{MSS(R)}}{nS_X^2}}}$$

Contraste	Región de rechazo	p-valor
$H_0 : \beta_i = 0$ $H_1 : \beta_i \neq 0$	$\widehat{T}_i \notin (-t_{n-2,\alpha/2}, t_{n-2,\alpha/2})$	$p = 2\mathbb{P}(t_{n-2} > \lvert\widehat{T}_i\rvert)$
$H_0 : \beta_i = 0$ $H_1 : \beta_i > 0$	$\widehat{T}_i > t_{n-2,\alpha}$	$p = \mathbb{P}(t_{n-2} > \widehat{T}_i)$
$H_0 : \beta_i = 0$ $H_1 : \beta_i < 0$	$\widehat{T}_i < -t_{n-2,\alpha}$	$p = \mathbb{P}(t_{n-2} < \widehat{T}_i)$

1.2.4. El contraste de regresión

El contraste de regresión chequea si el mejor ajuste linela tiene pendiente no nula.

$$\begin{cases} H_0 : \beta_1 = 0 \\ H_1 : \beta_1 \neq 0 \end{cases} \iff \begin{cases} H_0 : \mathbb{E}(Y|X) = \beta_0 \\ H_1 : \mathbb{E}(Y|X) = \beta_0 + \beta_1 X \end{cases}$$

- Si la prueba no resulta significativa (se acepta H_0), un modelo lineal es inapropiado (X no ayuda a predecir linealmente el valor esperado de Y).
- Si la prueba es significativa (se rechaza H_0), X ayuda a predecir linealmente el valor esperado de Y, pero no significa necesariamente que el ajuste lineal sea el correcto.

1.2.5. Contrastes de hipótesis para σ^2

$$\begin{cases} H_0 : \sigma^2 = \sigma_0^2 \\ H_1 : \sigma^2 \neq \sigma_0^2 \end{cases} \iff \text{Bajo } H_0 : \ (n-2)\frac{\text{MSS(R)}}{\sigma_0^2} \sim \chi^2_{n-2}$$

- Rechazar H_0 a nivel de significación α si: $(n-2)\frac{\text{MSS(R)}}{\sigma_0^2} \notin \left(\chi^2_{n-2,1-\alpha/2},\ \chi^2_{n-2,\alpha/2}\right)$

$$\begin{cases} H_0 : \sigma^2 = \sigma_0^2 \\ H_1 : \sigma^2 > \sigma_0^2 \end{cases} \iff \text{Rechazar } H_0 \text{ si } : \ (n-2)\frac{\text{MSS(R)}}{\sigma_0^2} > \chi^2_{n-2,\alpha}$$

1.3. Análisis de la varianza

1.3.1. Descomposición de la variabilidad

En un modelo de regresión podemos definir los siguientes conceptos relacionados con la varianza de la variable respuesta Y. Para $i = 1, ..., n$.

- **SS(G)**: Suma de diferencias al cuadrado entre $\overline{Y}$ y Y_i.

$$\text{SS(G)} = \sum_{i=1}^{n} (Y_i - \overline{Y})^2$$

- **SS(R)**: Suma de diferencias al cuadrado entre Y_i y $\widehat{Y}_i$.

$$\text{SS(R)} = \sum_{i=1}^{n}(Y_i - \widehat{Y}_i)^2$$

- **SS(E)**: Suma de diferencias al cuadrado entre $\widehat{Y}_i$ y $\overline{Y}$.

$$\text{SS(E)} = \sum_{i=1}^{n}(\widehat{Y}_i - \overline{Y})^2$$

Podemos demostrar con facilidad que SS(G) = SS(R) + SS(E), esto es:

$$\sum_{i=1}^{n}(Y_i - \overline{Y})^2 = \sum_{i=1}^{n} e_i^2 + \sum_{i=1}^{n}(\widehat{Y}_i - \overline{Y})^2$$

1.3.2. Contraste de regresión con el cuadro ANOVA

$$\begin{cases} H_0 : \beta_1 = 0 & \text{No existe relación entre las variables} \\ H_1 : \beta_1 \neq 0 & \text{Existe una relación entre las variables} \end{cases}$$

Sabemos que bajo H_0 la mejor recta que describe el modelo de regresión lineal es horizontal.

$$\widehat{Y} \approx \overline{Y} \implies \text{SS(E)} \approx 0 \implies \text{SS(G)} = \text{SS(R)}$$

Una forma de contrastar esto es empleando el cociente entre SS(E)/SS(R). Cuanto mayor sea este cociente, mayor es la evidencia a favor de rechazar H_0. Como el comportamiento de las sumas de cuadrados (SS) depende de sus grados de libertad, se emplean las SS promediadas por sus grados de libertad.

Fuente de variación	**SS**	**Grados libertad**	**MSS**
Explicada por el ajuste	SS(E)	1	$\text{MSS(E)} = \text{SS(E)}$
No explicada por el ajuste	SS(R)	$n-2$	$\text{MSS(R)} = \dfrac{\text{SS(R)}}{n-2}$
Global	SS(G)	$n-1$	$\text{MSS(G)} = \widehat{S}_Y^2 = \dfrac{\text{SS(G)}}{n-1}$

$$\text{Bajo } H_0,\ F = \frac{\text{MSS(E)}}{\text{MSS(R)}} \sim F_{1,n-2} \implies \begin{cases} \text{Rechazar } H_0 \text{ a nivel } \alpha \text{ si } : \ \widehat{F} > F_{1,n-2,\alpha} \\ p\text{-valor} : \ p = \mathbb{P}(F_{1,n-2} > \widehat{F}) \end{cases}$$

1.4. Bondad de ajuste

1.4.1. Covarianza de dos variables X e Y

La **covarianza poblacional** se define como:

$$\sigma_{XY} = \mathbb{E}\Big((X - \mathbb{E}(X))(Y - \mathbb{E}(Y))\Big) = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y)$$

Y dada una muestra $\{(X_i, Y_i), i = 1, ...n\}$, su estimación es:

$$S_{XY} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y}) = \frac{1}{n}\sum_{i=1}^{n} X_i Y_i - \overline{XY}$$

- $S_{XY}0 \iff \beta_1 > 0$ (dependencia lineal positiva)
- $S_{XY}0 \iff \beta_1 < 0$ (dependencia lineal negativa)
- $S_{XY} \approx 0 \iff \beta_1 \approx 0$ (ausencia de dependencia lineal)

1.4.2. Coeficiente de correlación lineal de Pearson

La correlación poblacional se define como $\rho_{XY} = \dfrac{\sigma_{XY}}{\sigma_X \sigma_Y}$ y se estima a través de:

$$r_{XY} = \frac{S_{XY}}{S_X S_Y}$$

- $\rho_{XY} \in [-1, 1]$.
- Cuanto más cercano esté $|\rho_{XY}|$ de 1, mayor evidencia de dependencia lineal (positiva o negativa).
- Cuanto más cercano esté ρ_{XY} de 0, mayor ausencia de dependencia lineal.

Podemos plantearnos **contrastes de hipótesis** del tipo:

$$\begin{cases} H_0 : \rho_{XY} = 0 & \text{Ausencia de dependencia lineal} \\ H_1 : \rho_{XY} \neq 0 & \text{Existe dependencia lineal} \end{cases}$$

Bajo H_0 y la normalidad de los datos: $\widehat{T}_r = r_{XY}\sqrt{\dfrac{n-2}{1-r_{XY}^2}} \sim t_{n-2}$

- Rechazar H_0 a nivel α si $\widehat{T}_r \notin (-t_{n-2,\alpha/2}, t_{n-2,\alpha/2})$.
- Bajo H_0 si $n > 30 \Longrightarrow \sigma(r_{XY}) \approx \dfrac{1}{\sqrt{n}}$. Por tanto, rechazar H_0 si: $|r_{XY}| > \dfrac{2}{\sqrt{n}}$.
- $R^2 = 1 - \dfrac{(n-2)\text{MSS(R)}}{(n-1)\text{MSS(G)}} = 1 - \dfrac{(n-2)\widehat{\sigma}^2}{(n-1)\widehat{S}_Y^2}$

1.4.3. Coeficiente de determinación

El **coeficiente de determinación** es la proporción de variación de la respuesta explicada por la recta de regresión ajustada.

$$R^2 = \frac{\text{SS(E)}}{\text{SS(G)}} = 1 - \frac{\text{SS(R)}}{\text{SS(G)}}$$

- $0 \leq R^2 \leq 1$
- R^2 evalúa la bondad del ajuste. A mayor R^2, mayor capacidad predictiva del modelo ajustado.
- $R^2 = r_{XY}^2$ (sólo en el caso de RLS).
- R^2 puede ser alto con modelos no lineales, incluso con ajustes lineales pobres.

1.5. Predicción sobre un modelo de RLS

Tras realizar el ajuste de un modelo de RLS de la forma $\widehat{m}(X) = \widehat{\beta}_0 + \widehat{\beta}_1 X$, existen dos objetivos:

1. **Estimar el valor esperado de la respuesta** para un valor específico X_0 de la explicativa.
$$m_0 = m(X_0) = \mathbb{E}(Y|X = X_0) = \beta_0 + \beta_1 X_0$$
2. **Predecir el valor de la respuesta** para un valor específico X_0 de la explicativa.
$$Y_0 = Y|(X = X_0)$$

En ambos casos la solución se obtiene sustituyendo X por X_0 en la recta de regresión estimada.

$$\widehat{m}_0 = \widehat{Y}_0 = \widehat{\beta}_0 + \widehat{\beta}_1 X_0$$

Sin embargo, la resolución de ambos problemas se realiza con distinta precisión.

1.5.1. Estimación de la respuesta media condicionada

En un modelo de RLS, el estimador de m_0 es:

$$\widehat{m}_0 = \widehat{\beta}_0 + \widehat{\beta}_1 X_0 = \overline{Y} + \widehat{\beta}_1 (X - X_0$$

- $\widehat{m}_0$ es insesgado: $\mathbb{E}(\widehat{m}_0) = m_0 = \beta_0 + \beta_1 X_o$
- La varianza de $\widehat{m}_0$ toma la forma estándar de la varianza de una media muestral (σ^2/n), reemplazando n por n_0, que depende de la distancia estandarizada de X_0 a $\overline{X}$.

$$\text{Var}(\widehat{m}_0) = \frac{\sigma^2}{n_0}, \quad \text{donde} \quad n_0 = \frac{n}{1 + \left(\frac{X_0 - \overline{X}}{S_X}\right)^2}$$

- Si X_0 está en el rango de valores observados (interpolación) se tiene que $n_0 \in [1, n]$.
- El valor máximo $n_0 = n$ se alcanza cuando $X_0 = \overline{X}$.
- Si se extrapola, entonces $n_0 \to 0$ y $\text{Var}(\widehat{m}_0) \to \infty$, ya que no hay información muestral sobre la respuesta.

1.5.2. Inferencia sobre la respuesta media condicionada

Se puede demostrar que la distribución de $\widehat{m}_0$ es:

$$\widehat{m}_0 \sim N\left(m_0, \frac{\sigma}{\sqrt{n_0}}\right) \quad \Longleftrightarrow \quad \left(\frac{\widehat{m}_0 - m_0}{\sigma/\sqrt{n_0}}\right) \sim N(0,1)$$

Por tanto un estadístico de contraste toma la forma: $\widehat{T}_{m_0} = \dfrac{\widehat{m}_0 - m_0^*}{\sqrt{\dfrac{\text{MSS(R)}}{n_0}}} \sim t_{n-2}$

- Intervalo de confianza al nivel $100(1-\alpha)\,\%$ para m_0:

$$\left(\widehat{m}_0 - \sqrt{\frac{\text{MSS(R)}}{n_0}} t_{n-2,\alpha/2} \ , \ \widehat{m}_0 + \sqrt{\frac{\text{MSS(R)}}{n_0}} t_{n-2,\alpha/2}\right)$$

- Contraste de hipótesis: $H_0 : m_0 = m_0^*$ vs $H_1 : m_0 \neq m_0^*$.
  - Rechazar H_0 al nivel de significación α si $\widehat{T}_{m_0} \notin (-t_{n-2,\alpha/2}, t_{n-2,\alpha/2})$.
  - p-valor $= 2\mathbb{P}\Big(t_{n-2} > |\widehat{T}_{m_0}|\Big)$.

1.5.3. Predicción para un valor de la explicativa

En un modelo RLS, la predicción Y_0 de la respuesta para $X = X_0$ se estima mediante:

$$\widehat{Y}_0 = \widehat{\beta}_0 + \widehat{\beta}_1 X_0 = \overline{Y} + \widehat{\beta}_1 (X - X_0)$$

- $\widehat{Y}_0$ es insesgado: $\mathbb{E}(\widehat{Y}_0) = Y_0 = \mathbb{E}(Y|X = X_0) = \beta_0 + \beta_1 X_0$.
- La varianza de la predicción (el error cuadrático medio de predicción) es:

$$\text{Var}(\widehat{Y}_0) = \mathbb{E}(\widehat{Y}_0 - Y_0)^2 = \mathbb{E}(\widehat{m}_0 - (m_0 + \varepsilon_0))^2 = \text{Var}(\widehat{m}_0) + \text{Var}(\varepsilon_0) = \frac{\sigma^2}{n_0} + \sigma^2$$

- Nótese que $\text{Cov}(\widehat{m}_0, e_0) = 0$ porque (X_0, Y_0) no es un punto muestral.

1.5.4. Inferencia sobre la predicción para un valor de la explicativa

Se puede demostar que la distribución de $\widehat{Y}_0$ es:

$$\widehat{Y}_0 \sim N\left(Y_0, \sigma\sqrt{1+\frac{1}{n_0}}\right) \iff \frac{\widehat{Y}_0 - Y_0}{\sigma\sqrt{1+\frac{1}{n_0}}} \sim N(0,1)$$

Por tanto el estadístico de contraste toma la forma: $\widehat{T}_{Y_0} = \dfrac{\widehat{Y}_0 - Y_0^*}{\sqrt{\mathsf{MSS(R)}}\sqrt{1+\dfrac{1}{n_0}}} \sim t_{n-2}$

- Intervalo de confianza al nivel $100(1-\alpha)\,\%$ para Y_0:

$$\left(\widehat{Y}_0 - \sqrt{\mathsf{MSS(R)}}\sqrt{1+\frac{1}{n_0}}\cdot t_{n-2,\alpha/2},\ \ \widehat{Y}_0 + \sqrt{\mathsf{MSS(R)}}\sqrt{1+\frac{1}{n_0}}\cdot t_{n-2,\alpha/2}\right)$$

- Contraste de hipótesis:
  - Rechazar H_0 si $\widehat{T}_{Y_0} \notin (-t_{n-2,\alpha/2}, t_{n-2,\alpha/2})$.
  - p-valor: $p = 2\mathbb{P}\left(t_{n-2} > |\widehat{T}_{Y_0}|\right)$

1.6. Diagnosis del modelo

Validar un modelo de RLS ajustado requiere:

1. Chequear gráfica y analíticamente que se satisfacen las hipótesis estructurales (linealidad, homoscedasticidad, normalidad e independencia).
2. Analizar la influencia de registros atípicos que pueden tener un efecto importante en el ajuste y que a menudo se traduce en la violación de alguna de las hipótesis estructurales.
3. Analizar los residuos y sus propiedades, pues las hipótesis estructurales se asientan principalmente en sus características.

1.6.1. Propiedades de los residuos. Leverages

Los residuos del ajuste se definen como aproximaciones a las perturbaciones aleatorias no observables.

$$e_i = Y_i - \widehat{Y}_i = Y_i - \mathbb{E}(Y|X_i) = Y_i - \widehat{m}(X_i), \quad \text{para } i = 1, ..., n$$

1. Por construcción, los e_i suman cero, por tanto, tienen media cero.
2. Los residuos e_i no son independientes, ya que tienen $n-2$ grados de libertad (para $n > 30$ el efecto es irrelevante).

3. Los residuos e_i no son homoscedásticos. La varianza de e_i depende de X_i.

$$\text{Var}(e_i) = \sigma^2 - \frac{\sigma^2}{n_{X_i}} = \sigma^2(1 - h_{ii})$$

donde $h_{ii} = 1/n_{X_i}$ se denomina **leverage** y n_{X_i} depende de la idstancia estandarizada de X_i a $\overline{X}$ (cuanto mayor sea, menos distancia hay entre X_i y $\overline{X}$).

Definición del leverage

El **leverage** del i-ésimo residuo h_{ii} depende de X_i y es tanto más elevado cuanto mayor sea su distancia de $\overline{X}$.

- Se prueba que $1/n \leq h_{ii} \leq 1$.
- Si X_i dista mucho de $\overline{X}$, entonces:

$$h_{ii} \approx 1 \iff \text{Var}(e_i) = 0 \iff \mathbb{E}(e_i) = 0 \iff \text{La recta ajustada pasa por } Y_i$$

1.6.2. Estandarización de los residuos

- **Residuos estandarizados (studentizados internamente)**

$$r_i = \frac{e_i}{\sqrt{\text{MSS(R)}(1 - h_{ii})}} \sim t_{n-2}$$

- **Residuos studentizados (studentizados externamente)**

$$\widehat{t}_i = \frac{e_i}{\sqrt{\text{MSS(R)}_{(i)}(1 - h_{ii})}} \sim t_{n-3}$$

donde $\text{MSS(R)}_{(i)}$ denota la varianza residual de un ajuste sin (X_i, Y_i).

1.6.3. Chequeando linealidad

Hipótesis de linealidad: $\mathbb{E}(Y|X) = m(X) = \beta_0 + \beta_1 X \iff \mathbb{E}(\varepsilon_i) = 0$ para $i = 1, ..., n$.

- Gráficamente: Gráfico de dispersión $(e_i \sim \widehat{Y}_i)$.
- Analíticamente: Si para algunos X_i se dispone de varias respuestas $Y_i = Y_{ij}$.

$$\begin{cases} H_0 : \mathbb{E}(Y|X) = m(X) = \beta_0 + \beta_1 X \\ H_1 : \mathbb{E}(Y|X) = m(X) = \beta_0 + \beta_i \mathbb{I}_i \end{cases} \qquad \text{siendo } \mathbb{I}_i = \begin{cases} 1 & \text{si } X = X_i \\ 0 & \text{si } X \neq X_i \end{cases}$$

1.6.4. Chequeando homoscedasticidad

Hipótesis de homoscedasticidad: $\text{Var}(Y|X) = \sigma^2 \iff \text{Var}(\varepsilon_i) = \sigma^2$, para $i = 1, ..., n$.

- Gráficamente: Gráfico de dispersión $(e_i \sim \widehat{Y}_i)$, observar el efecto de espiral.

- Test de Breausch-Pagan: Contrasta una regresión lineal entre los residuos al cuadrado y la variable explicativa X.

$$\begin{cases} H_0 : \mathbb{E}(\varepsilon^2|X) = \gamma_0 \\ H_1 : \mathbb{E}(\varepsilon^2|X) = \gamma_0 + \gamma_1 X \end{cases} \iff \begin{cases} H_0 : \text{Var}(\varepsilon|X) = \gamma_0 \\ H_1 : \text{Var}(\varepsilon|X) = \gamma_0 + \gamma_1 X \end{cases}$$

  Si la prueba es significativa, se rechaza la hipótesis estructural de homoscedasticidad.

1.6.5. Chequeando normalidad

Hipótesis de normalidad: $Y|X \sim N(\beta_0 + \beta_1 X, \sigma) \iff \varepsilon_i \sim N(0, \sigma)$ para $i = 1, ..., n$.

Pruebas analíticas:

- Kolmogorov-Smirnov-Lilliefors
- Cramér-von Mises
- Anderson-Darling
- Shapiro-Wilk
- Jarque-Bera
- D'Agostino-Pearson

Pruebas gráficas:

- Histograma
- Estimador kernel de la densidad
- Diagrama de cajas
- Gráfica de tallo y hojas
- Gráfico P-PA
- Gráfico Q-Q

1.6.6. Chequeando aleatoriedad

Hipótesis de aleatoriedad: $\text{Cov}(X_i, X_j) = 0 \iff \text{Cov}(\varepsilon_i, \varepsilon_j) = 0$, para $i = 1, ..., n$, $i \neq j$.

- **Coeficiente de autocorrelación**. Chequea la existencia de autocorrelación en residuos ordenados temporalmente.

$$r_k = \Big(\sum_{i=1}^{n-k} e_i e_{i+k}\Big) / \Big(\sum_{i=1}^{n} e_i^2\Big)$$

  - Autocorrelación positiva. Valores por encima o debajo de la media tienden a ser seguidos por valores también ascendentes o descendentes, respectivamente.
  - Autocorrelación negativa. Valores por encima de la media tienden a ir seguidos de valores inferiores a la media, y viceversa.
  - Autocorrelación nula. Ausencia de tendencia temporal.

- **Prueba de Lyung-Box** (donde m es el número de retardos):

$$\begin{cases} H_0 : \text{Cov}(\varepsilon_i, \varepsilon_j) = 0, & i \neq j \\ H_1 : \text{Cov}(\varepsilon_i, \varepsilon_j) \neq 0, & i \neq j \end{cases} \qquad \text{Bajo } H_0 : \; Q = n(n+2) \sum_{i=1}^{m} \frac{r_k^2}{n-2} \sim \chi^2_{m-2}$$

  Rechazar H_0 a un nivel de significación α si $\widehat{Q} > \chi^2_{m-2,\alpha}$

- **Prueba de rachas**. Sea R el número de rachas en los residuos, p el número de residuos de signo positivo y q el número de residuos de signo negativo, bajo H_0:

$$R|(p,q) \sim N(\mu_R, \sigma_R) \quad \text{donde} \quad \mu_R = 1 + \frac{2pq}{p+q} \text{ y } \sigma_R^2 = \frac{2pq(2pq-p-q)}{(p+q)^2(p+q-1)}$$

Rechazar H_0 a un nivel α si: $\left|\frac{\widehat{R} - 0{,}5 - \mu_R}{\sigma_R}\right| > z_{\alpha/2}$.

1.7. Análisis de influencia

1.7.1. Tipos de observaciones atípicas

- Atípico (outlier): Dato muestral extremo en la respuesta. Incrementa el MSS(R), dificulta la interpretación del modelo y posibilita un ajuste erróneo.
- Influyente: Dato con mucho peso en el ajuste. Los ajustes con y sin el dato difieren sustancialmente.
- Con elevado leverage: Dato muestral extremo en la explicativa.

Atípicos y con alto leverage: potencialmente influyentes.

1.7.2. Mediciones de las influencias

- **Leverage de los residuos**. Los datos con elevado leverage son candidatos a ser influyentes.
- **Distancia D de Cook**. Sea $\widehat{Y}_i$ el i-ésimo valor ajustado con todos los datos de la muestra y $\widehat{Y}_{(i)}$ el i-ésimo valor ajustado en un modelo sin el i-ésimo dato:

  $$D_i = \frac{\sum_{i=1}^{n}(\widehat{Y}_i - \widehat{Y}_{(i)})^2}{\widehat{\sigma}^2}$$

  Un elevado valor de D_i indica que el i-ésimo dato es influyente.
- **Estimación de Jacknife**. Evalúa el cambio en los coeficientes estimados cuando se realiza un ajuste omitiendo un dato cada vez. Si un dato al ser omitido genera cambios sustanciales en los coeficientes estimados, entonces son influyentes.

1.8. Transformaciones para obtener linealidad

En general, probaremos a transformar la variable explicativa X cuando los residuosno muestren desviación de las hipótesis de **normalidad** y **homoscedasticidad**. Si, por el contrario, la distribución de los residuos es asimétrica y evidencia heteroscedasticidad, probaremos a transformar la Y o ambas variables.

La familia de transformaciones Box-Cox se usa a menudo usada para corregir **heteroscedasticidad** y **desviaciones de la normalidad** pero es también útil para conseguir linealidad cuando la relación es monótona.

$$Y^* = \begin{cases} \dfrac{Y^\lambda - 1}{\lambda} & \lambda \neq 0 \\ \ln Y & \lambda = 0 \end{cases}$$

El valor óptimo de λ puede aproximarse por máxima verosimilitud.

Algunas relaciones funcionales fácilmente transformables en una relación lineal son las que siguen:

Nombre	Relación	Linealizada
Lineal	$Y = \beta_0 + \beta_1 X$	
Logarítmica	$Y = \beta_0 + \beta_1 \ln X$	
Potencial	$Y = \beta_0 X^{\beta_1}$	$\ln Y = \ln \beta_0 + \beta_1 \ln X$
Exponencial	$Y = \beta_0 \exp(\beta_1 X)$	$\ln Y = \ln \beta_0 + \beta_1 X$
Compuesta	$Y = \beta_0 \beta_1^X$	$\ln Y = \ln \beta_0 + X \ln \beta_1$
Curva-S	$Y = \exp(\beta_0 + \frac{\beta_1}{X})$	$\ln Y = \beta_0 + \frac{\beta_1}{X}$
Crecimiento	$Y = \exp(\beta_0 + \beta_1 X)$	$\ln Y = \beta_0 + \beta_1 X$
Inversa	$Y = \beta_0 + \frac{\beta_1}{X}$	
Cuadrática	$Y = \beta_0 + \beta_1 X + \beta_2 X^2$	

Tema 2: Regresión Lineal Múltiple

2.1. Formulación del modelo de RLM

Sea $\{(X_{11},...,X_{1k},Y_1),...,(X_{n1},...,X_{nk},Y_n)\}$ una muestra de n realizaciones independientes de $(X_1,...,X_k,Y)$, asumimos la relación estocástica:

$$Y_i=\beta_0+\beta_1X_{i1}+\cdots+\beta_kX_{ik}+\varepsilon_i,\quad \text{con } \varepsilon_i \text{ i.i.d } \sim N(0,\sigma),\ \text{ para } i=1,...,n$$

Y definimos el modelo de RLS como:

$$\mathbb{E}(Y|(X_1,...,X_k))=m(X_1,...,X_k)=\beta_0+\sum_{j=1}^{k}\beta_jX_j$$

- β_0: El intercept o valor esperado de Y cuando $X_j=0$ para todo $j=1,...,k$.
- β_j $(j=1,...,k)$: Tasa de cambio del valor esperado de Y por unidad de incremento en X_j cuando X_r permanece constante, para todo $r\neq j$.
- σ: Desviación estándar de las respuestas para n valor arbitrario de $\vec{X}$.

En formato matricial, esto se traduce a:

$$\vec{Y}=\mathbf{X}\vec{\beta}+\vec{\varepsilon}\iff\begin{pmatrix}Y_1\\ \vdots\\ Y_n\end{pmatrix}=\begin{pmatrix}1 & X_{11} & \cdots & X_{1k}\\ \vdots & \vdots & \ddots & \vdots\\ 1 & X_{n1} & \cdots & X_{nk}\end{pmatrix}\begin{pmatrix}\beta_0\\ \vdots\\ \beta_k\end{pmatrix}+\begin{pmatrix}\varepsilon_1\\ \vdots\\ \varepsilon_n\end{pmatrix}$$

2.1.1. Hipótesis estructurales sobre el modelo RLM

- **Linealidad**: $\mathbb{E}(\vec{Y}|\mathbf{X}) = m(\mathbf{X}) = \mathbf{X}\vec{\beta} \iff \mathbb{E}(\vec{\varepsilon}) = \vec{0}$
- **Homoscedasticidad e independencia**: $\text{Var}(\vec{Y}|\mathbf{X}) = \sigma^2\mathbf{I}_n \iff \text{Var}(\vec{\varepsilon}) = \sigma^2\mathbf{I}_n$
- **Normalidad**: $\vec{Y}|\mathbf{X} \sim N_n(\mathbf{X}\vec{\beta}, \sigma^2\mathbf{I}_n) \iff \vec{\varepsilon} \sim N_n(\vec{0}, \sigma^2\mathbf{I}_n)$

Se asume también que $n > k+1$ y rango$(\mathbf{X}) = k$

2.1.2. Criterio de estimación por mínimos cuadrados

Objetivo: Determinar el hiperplano generado por los vectores $\vec{1}, \vec{X}_1, ..., \vec{X}_k$ que minimiza la suma de cuadrados de las diferencias entre las observaciones Y_i y los valores pronosticados por el hiperplano $\widehat{Y}_i$. Esto es, encontrar los $\widehat{\beta}_j$ que minimicen:

$$\text{SS(R)} = \sum_{i=1}^{n}(Y_i - \widehat{Y}_i)^2 = \sum_{i=1}^{n}\left(Y_i - \beta_0 - \sum_{j=1}^{k}\beta_k X_{ij}\right)^2 = (\vec{Y} - \mathbf{X}\vec{\beta})^t(\vec{Y} - \mathbf{X}\vec{\beta})$$

Resultados obtenidos del criterio OLS

Para $i = 1, ..., n$ y $j = 1, ..., k$:

- Ecuaciones canónicas de la regresión: $\mathbf{X}^t\mathbf{X}\widehat{\vec{\beta}} = \mathbf{X}^t\vec{Y}$
- Ecuaciones normales: $\sum_{i=1}^{n} e_i = 0, \quad \text{y} \quad \sum_{i=1}^{n} X_{ij}e_i = 0$
- Estimador de $\vec{\beta}$: $\widehat{\vec{\beta}} = (\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\vec{Y}.$
- Valores ajustados: $\widehat{\vec{Y}} = \mathbf{X}\widehat{\vec{\beta}} = \mathbf{X}(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\vec{Y} = \mathbf{H}\vec{Y}$
- Matriz sobrero: $\mathbf{H} = \mathbf{X}(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t$
- Residuos del modelo ajustado: $e_i = Y_i - \widehat{Y}_i = Y_i - \widehat{\beta}_0 - \sum_{j=1}^{k}\widehat{\beta}_j X_{ij}.$

2.1.3. El modelo de RLM en desviaciones

Premisa: $\sum_{i=1}^{n} e_i = 0 \iff \sum_{i=1}^{n} Y_i = n\widehat{\beta}_0 + \sum_{j=1}^{k}\widehat{\beta}_j \sum_{i=1}^{n} X_{ij} \implies \widehat{\beta}_0 = \overline{Y} - \sum_{j=1}^{k}\widehat{\beta}_j\overline{X}_j.$

Modelo en desviaciones: $\widehat{Y} - \overline{Y} = \sum_{j=1}^{k}\widehat{\beta}_j(X_j - \overline{X}_j)$

Denotando:

- $\tilde{\vec{Y}} = Y - \overline{Y}$ al vector de respuestas centradas.
- $\tilde{\mathbf{X}}$ a la matriz de explicativas centradas sin la columna de unos.
- $\vec{b}$ al vector de coeficientes sin β_0.

Las ecuaciones normales se definen de nuevo de la siguiente manera:

$$\tilde{\mathbf{X}}^t\tilde{\mathbf{X}}\widehat{\vec{\beta}} = \tilde{\mathbf{X}}\tilde{\vec{Y}} \quad \Longleftrightarrow \quad \vec{b} = \mathbf{S}_{\mathbf{XX}}^{-1}\mathbf{S}_{\mathbf{X}\vec{Y}}$$

- Si X_j son incorreladas, $\mathbf{S_{XX}}$ es diagonal. Por tanto, $\widehat{\beta}_j = \dfrac{\mathsf{Cov}(X_j, Y)}{\mathsf{Var}(X_j)}$.
- Si X_j son correladas, sus coeficientes estimados por RLM pueden variar mucho de los obtenidos por RLS.

2.2. Interpretación del modelo

2.2.1. Interpretación de los estimadores $\widehat{\beta}_j$

Lema: $\widehat{\beta}_j$ recoge el efecto diferencial de X_j tras haber eliminado los efectos de $\vec{X}_l$, donde $l \neq j$.

1. Realizar el ajuste RLM que explica la variable Y a partir de $\vec{X}_l$.
$$Y_i = \sum_{l\neq j} \widehat{b}_l X_{il} + e_i^{y\sim l(j)}$$
donde $e_i^{y\sim l(j)}$ recoge la parte de Y no explicada por $\vec{X}_l$.

2. Realizar el ajuste RLM que explica la variable X_j a partir del conjunto de variables $\vec{X}_l$.
$$X_{ij} = \sum_{l\neq j} \widehat{c}_l X_{il} + e_i^{j\sim l(j)}$$
donde $e_i^{j\sim l(j)}$ recoge la parte de X_j no explicada por $\vec{X}_l$.

3. Realizar el ajuste de RLS: $\quad e_i^{y\sim l(j)} = \widehat{d}_j e_i^{j\sim l(j)} + e_i^{y\sim j}$

4. Y se prueba que $\widehat{\beta}_j = \widehat{d}_j$ y que $e_i = e_i^{y\sim j}$.

2.2.2. Enfoque geométrico

Sea $\mathsf{col}(\mathbf{X})$ el hiperplano generado por $\left(\vec{1}, \vec{X}_1, ..., \vec{X}_k\right)$. El objetivo de OLS es encontrar el vector:

$$\widehat{\vec{Y}} = \widehat{\beta}_0\vec{1} + \widehat{\beta}_1\vec{X}_1 + \cdots + \widehat{\beta}_k\vec{X}_k = \mathbf{X}\widehat{\vec{\beta}}$$

minimizando el módulo del vector de residuos:

$$\|\vec{e}\|^2 = \|\vec{Y} - \widehat{\vec{Y}}\|^2 = \|\vec{Y} - \mathbf{X}\widehat{\vec{\beta}}\|^2 = \mathsf{SS(R)}$$

- Entonces, $\vec{e} = \vec{Y} - \mathbf{X}\widehat{\vec{\beta}}$ es ortogonal a col$(\mathbf{X})$ y, por tanto, a todos los vectores que lo generan:

$$\vec{1}^t\vec{e} = \vec{X}_1\vec{e} = ... = \vec{X}_k\vec{e} = 0 \iff \mathbf{X}^t\vec{e} = \vec{0}$$

$$\mathbf{X}^t\vec{e} = \mathbf{X}^t(\vec{Y} - \mathbf{X}\widehat{\vec{\beta}}) = \vec{0} \iff \mathbf{X}^t\vec{Y} = \mathbf{X}^t\mathbf{X}\widehat{\vec{\beta}}$$

- Y el vector de valores ajustados que resulta, $\widehat{\vec{Y}}$, es la proyección ortogonal de $\vec{Y}$ sobre col$(\mathbf{X})$.

$$\widehat{\vec{Y}} = \mathbf{X}\widehat{\vec{\beta}} = \mathbf{X}(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\vec{Y} = \mathbf{H}\vec{Y}$$

- Luego, la matriz sombrero $\mathbf{H} = \mathbf{X}(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t$ es una matriz de **proyección ortogonal** (idempotente y simétrica) sobre col$(\mathbf{X})$.

0

Residuos obtenidos matricialmente

Los residuos son proyección de $\vec{Y}$ sobre el espacio ortogonal al definido por $\mathbf{H}$.

$$\vec{e} = \vec{Y} - \widehat{\vec{Y}} = \vec{Y} - \mathbf{X}\widehat{\vec{\beta}} = \vec{Y} - \mathbf{H}\vec{Y} = (\mathbf{I}_n - \mathbf{H})\vec{Y}$$

Por la ortogonalidad de $\vec{e}$ y $\widehat{\vec{Y}}$ se tiene que: $\|\vec{Y}\|^2 = \|\widehat{\vec{Y}}\|^2 + \|\vec{e}\|^2$.

2.2.3. Grados de libertad del modelo

Teorema: Los grados de libertad de un modelo de RLM son $k+1$, el número de parámetros que lo definen, que coincide con el número de ecuaciones normales.

Como $\widehat{\vec{Y}} = \mathbf{H}\vec{Y}$, se tiene que para cada $i = 1, ..., n$:

$$\widehat{Y}_i = h_{i1}Y_1 + ... + h_{ii}Y_i + ... + h_{in}Y_n$$

donde el leverage de la i-ésima observación (h_{ii}) evalúa la importancia de Y_i en la estimación de $\widehat{Y}_i$.

Teorema 2: El rango y la traza de una matriz idempotente coinciden:

$$\text{tr}(\mathbf{H}) = \sum_{i=1}^{n} h_{ii} = \text{tr}(\mathbf{I}_{k+1}) = k+1$$

La influencia global de las observaciones para estimar sus valores ajustados está restringida al número de parámetros del modelo: $k+1$.

Teorema 3: Los grados de libertad de un ajuste del tipo $\widehat{\vec{Y}} = \mathbf{M}\vec{Y}$, paramétrico o no, se puede entender como la influencia que el total de los Y_i tienen para obtener los $\vec{Y}_i$, que vene dada:

$$\text{tr}(\mathbf{M}) = \sum_{i=1}^{n} m_{ii} = \text{grados de libertad del modelo}$$

2.2.4. Estimación de la varianza del modelo

Premisa: $\vec{e} = (\mathbf{I}_n - \mathbf{H})\vec{Y}$, donde $\mathbf{H}$ es una matriz idempotente y simétrica.

La suma de residuos al cuadrado en un modelo de RLM se define como:

$$\begin{aligned} \text{SS(R)} &= \sum_{i=1}^{n} e_i^2 = \vec{e}^t\vec{e} = \vec{Y}^t(\mathbf{I}_n - \mathbf{H})^t(\mathbf{I}_n - \mathbf{H})\vec{Y} \\ &= \vec{Y}^t(\mathbf{I}_n - \mathbf{H})\vec{Y} = \vec{Y}^t\vec{Y} - \widehat{\vec{\beta}}^t\mathbf{X}^t\vec{Y} \end{aligned}$$

El estimador por máxima verosimilitud (pero sesgado) de $\sigma^2 = \text{Var}(\varepsilon)$ es:

$$\widehat{\sigma}^2_{\text{MV}} = \frac{\text{SS(R)}}{n} = \frac{1}{n}\vec{e}^t\vec{e} = \frac{1}{n}\vec{Y}^t(\mathbf{I}_n - \mathbf{H})\vec{Y}$$

El estimador insesgado de la varianza del modelo de RLM se define como:

$$\widehat{\sigma}^2 = \text{MSS(R)} = \frac{\text{SS(R)}}{n-k-1} = \frac{1}{n-k-1}\sum_{i=1}^{n} e_i^2 = \frac{1}{n-k-1}\vec{e}^t\vec{e}$$

Y bajo las hipótesis estructurales del modelo se prueba que:

$$(n-k-1)\frac{\text{MSS(R)}}{\sigma^2} = \sum_{i=1}^{n}\left(\frac{e_i}{\sigma}\right)^2 \sim \chi^2_{n-k-1}$$

2.3. Propiedades de los estimadores $\widehat{\vec{\beta}}$

2.3.1. Insesgadez

Prmisa: $\widehat{\vec{\beta}} = (\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\vec{Y}$.

Si denotamos $\mathbf{C} = (\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t$, podemos reescribir:

$$\begin{aligned} \widehat{\vec{\beta}} &= \mathbf{C}\vec{Y} = \mathbf{C}(\mathbf{X}\vec{\beta} + \vec{\varepsilon}) = \mathbf{C}\mathbf{X}\vec{\beta} + \mathbf{C}\vec{\varepsilon} \\ &= (\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\vec{\beta} + \mathbf{C}\vec{\varepsilon} = \mathbf{I}_{k+1}\vec{\beta} + \mathbf{C}\vec{\varepsilon} = \vec{\beta} + \mathbf{C}\vec{\varepsilon} \end{aligned}$$

Por tanto, $\widehat{\vec{\beta}}$ s un estimador insesgado de β, pues:

$$\mathbb{E}(\widehat{\vec{\beta}}) = \mathbb{E}(\vec{\beta} + \mathbf{C}\vec{\varepsilon}) = \vec{\beta} + \mathbf{C}\mathbb{E}(\vec{\varepsilon}) = \vec{\beta}$$

2.3.2. Varianza de los estimadores $\widehat{\vec{\beta}}$

Premisa: $\mathbb{E}(\vec{\varepsilon}\vec{\varepsilon}^t) = \sigma^2\mathbf{I}_n$.

La matriz de varianzas-covarianzas de $\widehat{\vec{\beta}}$ toma la forma:

$$\mathbf{Var}(\widehat{\vec{\beta}}) = \begin{pmatrix} \text{Var}(\widehat{\beta}_0) & \text{Cov}(\widehat{\beta}_0,\widehat{\beta}_1) & \cdots & \text{Cov}(\widehat{\beta}_0,\widehat{\beta}_k) \\ \text{Cov}(\widehat{\beta}_0,\widehat{\beta}_1) & \text{Var}(\widehat{\beta}_1) & \cdots & \text{Cov}(\widehat{\beta}_1,\widehat{\beta}_k) \\ \vdots & \vdots & \ddots & \vdots \\ \text{Cov}(\widehat{\beta}_0,\widehat{\beta}_k) & \text{Cov}(\widehat{\beta}_1,\widehat{\beta}_k) & \cdots & \text{Var}(\widehat{\beta}_k) \end{pmatrix}$$

$$= \mathbb{E}\Big[(\widehat{\vec{\beta}}-\vec{\beta})(\widehat{\vec{\beta}}-\vec{\beta})^t\Big] = \mathbb{E}\Big[(\mathbf{C}\vec{\varepsilon})(\mathbf{C}\vec{\varepsilon})^t\Big] = \mathbb{E}[\mathbf{C}\vec{\varepsilon}\vec{\varepsilon}^t\mathbf{C}^t] = \mathbf{C}\mathbb{E}(\vec{\varepsilon}\vec{\varepsilon}^t)\mathbf{C}^t$$

$$= \sigma^2\mathbf{C}\mathbf{C}^t = \sigma^2(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\mathbf{X}(\mathbf{X}^t\mathbf{X})^{-1} = \sigma^2(\mathbf{X}^t\mathbf{X})^{-1}$$

- Siendo $(\mathbf{X}^t\mathbf{X})^{-1} = \big(q_{ij}\big)$, denotamos $\text{Var}(\widehat{\beta}_j) = \sigma^2 q_{jj}$, para $0 \leq j \leq k$.
- En general, $(\mathbf{X}^t\mathbf{X})^{-1}$ no es diagonal, de ahí que los $\widehat{\beta}_j$ no sean independientes.

2.3.3. Distribución de los parámetros $\widehat{\vec{\beta}}$, $\widehat{\beta}_j$ y σ^2

Para $j = 0, 1, ...k$ y $(\mathbf{X}^t\mathbf{X})^{-1} = (q_{ij})$:

Parámetro	Distribución	Estandarización
$\widehat{\vec{\beta}}$	$N_{k+1}\Big(\vec{\beta}, \sigma^2(\mathbf{X}^t\mathbf{X})^{-1}\Big)$	$\dfrac{(\widehat{\vec{\beta}}-\vec{\beta})\mathbf{X}^t\mathbf{X}(\widehat{\vec{\beta}}-\vec{\beta})}{\sigma^2} \sim \chi^2_{k+1}$
$\widehat{\beta}_j$	$N(\beta_j, \sigma\sqrt{q_{jj}})$	$\dfrac{\widehat{\beta}_j-\beta_j}{\sigma\sqrt{q_{jj}}} \sim N(0,1)$
σ^2	$\dfrac{\text{SS(R)}}{\sigma^2} \sim \chi^2_{k+1}$	$(n-k-1)\dfrac{\text{MSS(R)}}{\sigma^2} \sim \chi^2_{k+1}$

2.3.4. Independencia entre $\widehat{\vec{\beta}}$ y MSS(R)

Teorema: Sean $\vec{V}_1 = A_1\vec{Y}$ y $\vec{V}_2 = A_2\vec{Y}$, con $\vec{Y} \sim N_p(\vec{\mu}, \sigma^2\mathbf{I}_p)$. Entonces $\vec{V}_1$ y $\vec{V}_2$ son independientes si, y sólo si, $A_1^tA_2 = 0$.

Como aplicación directa se tiene la independencia de $\widehat{\vec{\beta}} = \mathbf{C}\vec{Y}$ y $\vec{e} = (\mathbf{I}_n - \mathbf{H})\vec{Y}$:

$$\mathbf{C}(\mathbf{I}_n - \mathbf{H}) = \mathbf{C} - \mathbf{C}\mathbf{H} = (\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t - (\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\mathbf{X}(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t = \mathbf{0}$$

2.4. Inferencia sobre el modelo

2.4.1. Bandas de confianza para un conjunto de parámetros $\vec{\beta}_{(r)}$

Sea $\widehat{\vec{\beta}}_{(r)}$ un subconjunto de r coeficientes de $\widehat{\vec{\beta}}$ y $(\mathbf{X}^t\mathbf{X})^{-1}_{(r)}$ la parte de la matriz $(\mathbf{X}^t\mathbf{X})^{-1}$ asociada a estos coeficientes.

- $\widehat{\vec{\beta}}_{(r)} \sim N_r\left(\vec{\beta}_{(r)}, \sigma^2(\mathbf{X}^t\mathbf{X})^{-1}_{(r)}\right)$
- $(n-k-1)\dfrac{\text{MSS(R)}}{\sigma^2} \sim \chi^2_{k+1}$
- $\widehat{\vec{\beta}}$ y $\vec{e}$ independientes $\Longrightarrow \widehat{\vec{\beta}}$ y MSS(R) independientes.

El estadístico de contraste de $\widehat{\vec{\beta}}_{(r)}$ toma la forma:

$$\widehat{F} = \frac{(\widehat{\vec{\beta}}_{(r)} - \vec{\beta}^t_{(r)})\left((\mathbf{X}^t\mathbf{X})^{-1}_{(r)}\right)^{-1}(\widehat{\vec{\beta}}_{(r)} - \vec{\beta}_{(r)})}{r\text{MSS(R)}} \sim F_{r,n-k-1}$$

Y por tanto una región de confianza al nivel $100(1-\alpha)\,\%$ para $\vec{\beta}_{(r)}$ es:

$$(\widehat{\vec{\beta}}_{(r)} - \vec{\beta}^t_{(r)})\left((\mathbf{X}^t\mathbf{X})^{-1}_{(r)}\right)^{-1}(\widehat{\vec{\beta}}_{(r)} - \vec{\beta}_{(r)}) \;\leq\; r\text{MSS(R)}F_{r,n-k-1,\alpha}$$

2.4.2. Bandas de confianza para el parámetro β_j

En este caso, $(\mathbf{X}^t\mathbf{X})^{-1}_{(r=1)}$ coincide con el elemento diagonal q_{jj}, donde $j = 0, ..., k$. El estadístico de contraste se simplifica a:

$$\frac{(\widehat{\beta}_j - \beta_j)^2}{q_{jj}\text{MSS(R)}} \sim F_{1,n-k-1} \iff \frac{\widehat{\beta}_j - \beta_j}{\sqrt{q_{jj}\text{MSS(R)}}} \sim t_{n-k-1}$$

- Región de confianza al nivel $100(1-\alpha)\,\%$ para β_j:

$$\frac{(\widehat{\beta}_j - \beta_j)^2}{q_{jj}\text{MSS(R)}} \leq F_{1,n-k-1,\alpha} \iff -t_{n-k-1,\alpha/2} \leq \frac{\widehat{\beta}_j - \beta_j}{\sqrt{q_{jj}\text{MSS(R)}}} \leq t_{n-k-1,\alpha/2}$$

- Intervalo de confianza al nivel $100(1-\alpha)\,\%$:

$$\text{IC}_{1-\alpha}(\beta_j) = \left(\widehat{\beta}_j - t_{n-k-1,\alpha/2}\sqrt{q_{jj}\text{MSS(R)}},\;\; \widehat{\beta}_j + t_{n-k-1,\alpha/2}\sqrt{q_{jj}\text{MSS(R)}}\right)$$

2.4.3. Intervalo de confianza para la varianza del modelo σ^2

Premisa: $(n-k-1)\dfrac{\text{MSS(R)}}{\sigma^2} \sim \chi^2_{n-k-1}$.

Entonces, si $\chi^2_{gl,\alpha}$ denota el valor tal que $\mathbb{P}(\chi^2_{gl} > \chi^2_{gl,\alpha}) = \alpha$, se tiene:

$$\mathbb{P}\left((n-k-1)\frac{\mathsf{MSS(R)}}{\sigma^2} \geq \chi^2_{n-k-1,1-\alpha}\right) = \mathbb{P}\left(0 \leq \frac{\sigma^2}{(n-k-1)\mathsf{MSS(R)}} \leq \frac{1}{\chi^2_{n-k-1,1-\alpha}}\right) = 1-\alpha$$

El intervalo de confianza para la varianza σ^2 al nivel $100(1-\alpha)\,\%$ es:

$$\mathsf{IC}_{(1-\alpha)}(\sigma^2) = \left(0, \frac{(n-k-1)\mathsf{MSS(R)}}{\chi^2_{n-k-1,1-\alpha}}\right) = \left(0, \frac{\mathsf{SS(R)}}{\chi^2_{n-k-1,1-\alpha}}\right)$$

2.4.4. Contrastes de hipótesis sobre un subconjunto de coeficientes $\vec{\beta}_{(r)}$

$$\begin{cases} H_0 : \vec{\beta}_{(r)} = \vec{\beta}^*_{(r)} \\ H_1 : \vec{\beta}_{(r)} \neq \vec{\beta}^*_{(r)} \end{cases} \qquad \text{Bajo } H_0: \ \widehat{F} = \frac{(\widehat{\vec{\beta}}_{(r)} - \vec{\beta}^t_{(r)})\left((\mathbf{X}^t\mathbf{X})^{-1}_{(r)}\right)^{-1}(\widehat{\vec{\beta}}_{(r)} - \vec{\beta}_{(r)})}{r\mathsf{MSS(R)}} \sim F_{r,n-k-1}$$

- Rechazar H_0 al nivel de significación α si: $\widehat{F} > F_{r,n-k-1,\alpha}$.
- El p-valor viene dado por: $p = \mathbb{P}(F_{r,n-k-1} > \widehat{F})$

2.4.5. Contrastes de hipótesis sobre un β_j

$$\begin{cases} H_0 : \beta_j = \beta^*_j \\ H_1 : \beta_j \neq \beta^*_j \end{cases} \qquad \text{para algún } 0 \leq j \leq k \text{ arbitrario}$$

Bajo H_0: $\widehat{F}_j = \dfrac{(\widehat{\beta}_j - \beta_j)^2}{q_{jj}\mathsf{MSS(R)}} \sim F_{1,n-k-1} \iff \widehat{T}_j = \dfrac{\widehat{\beta}_j - \beta^*_j}{\sqrt{q_{jj}\mathsf{MSS(R)}}} \sim t_{n-k-1}$

- Rechazar H_0 al nivel de significación α si: $\widehat{F}_j > F_{1,n-k-1,\alpha} \iff \widehat{T}_j \notin (-t_{n-k-1,\alpha/2}, t_{n-k-1,\alpha/2})$.
- El p-valor de la prueba es: $p = \mathbb{P}\big(F_{1,n-k-1} > \widehat{F}_j\big) = 2\mathbb{P}\big(t_{n-k-1} > |\widehat{T}_j|\big)$

2.4.6. Contraste de regresión en RLM ($r = k$)

El contraste de regresión chequea si ninguna de las variables explicativas ayuda a explicar linealmente la respuesta media:

$$\text{Contraste de regresión en RLM}: \quad \begin{cases} H_0 : \beta_1 = \beta_2 = ... = \beta_k = 0 \\ H_1 : \exists j \in \{1, ..., k\} / \beta_j \neq 0 \end{cases}$$

Denótese:

- $\vec{b}$ es el vector que resulta de eliminar β_0 de $\vec{\beta}$.
- $(\mathbf{X}^t\mathbf{X})^{-1}_{(11)}$ la matriz $(\mathbf{X}^t\mathbf{X})^{-1}$ sin la fila 1 ni la columna 1.

El estadístico de contraste es:

$$widehatF = \frac{\widehat{\vec{b}}^t \left((\mathbf{X}^t\mathbf{X})^{-1}_{(11)} \right)^{-1} \widehat{\vec{b}}}{k\text{MSS(R)}} \sim F_{k,n-k-1}$$

- Rechazar H_0 al nivel de significación α si: $\widehat{F} > F_{k,n-k-1,\alpha}$.
- El p-valor de la prueba viene dado por: $p = \mathbb{P}\big(F_{k,n-k-1} > \widehat{F}\big)$

2.5. Descomposición de la variabilidad

Igual que en el modelo de RLS, la variabilidad de los valores Y_i se descomponen de la forma:

$$\underbrace{\sum_{i=1}^{n}(Y_i - \overline{Y})^2}_{\text{SS(G)}} = \underbrace{\sum_{i=1}^{n} e_i^2}_{\text{SS(R)}} + \underbrace{\sum_{i=1}^{n}(\widehat{Y}_i - \overline{Y})^2}_{\text{SS(E)}}$$

$$\text{SS(E)} = \sum_{i=1}^{n}(\widehat{Y}_i - \overline{Y})^2 = (\widehat{\vec{Y}} - \overline{Y}\vec{1})^t(\widehat{\vec{Y}} - \overline{Y}\vec{1}) = \widehat{\vec{b}}\tilde{\mathbf{X}}^t\tilde{\mathbf{X}}\widehat{\vec{b}}$$

Como $\widehat{\vec{Y}}$ es la proyección sobre un espacio $k+1$ dimensional que contiene al vector $\vec{1}$, entonces $(\widehat{\vec{Y}} - \overline{Y}\vec{1})$ tiene dimensión k. Es decir, **SS(E) tiene k grados de libertad**.

2.5.1. Contraste de regresión en RLM con ANOVA

Cuanto mayor es el cociente SS(E)/SS(R), mayor es la evidencia en favor de rechazar H_0.

$$\begin{cases} H_0 : \vec{b} = \vec{0} \\ H_1 : \vec{b} \neq \vec{0} \end{cases} \quad \Longrightarrow \quad \widehat{F} = \frac{\text{MSS(E)}}{\text{MSS(R)}} \sim F_{k,n-k-1} \quad \text{y} \quad p = \mathbb{P}(\widehat{F} > F_{k,n-k-1})$$

Fuente de variación	**SS**	**Grados libertad**	**MSS**
Regresión	SS(E)	k	$\text{MSS(E)} = \frac{\text{SS(E)}}{k}$
Residual	SS(R)	$n-k-1$	$\text{MSS(R)} = \frac{\text{SS(R)}}{n-k-1}$
Global	SS(G)	$n-1$	

2.6. Comparación de dos modelos anidados

Un modelo de regresión M1 se dice *anidado* en un modelo M2 si las variables explicativas en M1 son un subconjunto de variables explicativas de M2.

$$\begin{cases} H_0 : Y_i = \beta_{0,(1)} + \sum_{j=1}^{k_1} \beta_{j,(1)} X_{ij} + \varepsilon_{i,(1)} & \text{M1} \\ H_1 : Y_i = \beta_{0,(2)} + \sum_{j=1}^{k_1} \beta_{j,(2)} X_{ij} + \sum_{j=k_1+1}^{k_2} \beta_{j,(2)} X_{ij} + \varepsilon_{i,(2)} & \text{M2} \end{cases}$$

Entonces, podemos decir que los valores de Y pueden ser explicados por el M1 o el M2:

$$\vec{Y} = \widehat{\vec{Y}}_{(1)} + \vec{e}_{(1)} = \widehat{\vec{Y}}_{(2)} + \vec{e}_{(2)}$$

Y siguiendo las premisas dadas en la interpretación geométrica del modelo (véase en el apéndice la descomposición ortogonal de $\vec{Y}$) llegamos a la conclusión de que podemos descomponer el vector $\vec{Y}$ en una suma de vectores ortogonales entre ellos:

$$\vec{Y} = \widehat{\vec{Y}}_{(1)} + (\vec{e}_{(1)} - \vec{e}_{(2)}) + \vec{e}_{(2)}$$

Lo que se puede traducir en:

$$\begin{array}{ccccccc} \text{SS(G)} & = & \text{SS(E)}_{\text{M1}} & + & \text{SS(R)}_{\text{M1}} - \text{SS(R)}_{\text{M2}} & + & \text{SS(R)}_{\text{M2}} \\ \underbrace{\frac{1}{\sigma^2}\left\|\vec{Y} - \overline{Y}\right\|^2}_{n-1} & = & \underbrace{\frac{1}{\sigma^2}\left\|\widehat{\vec{Y}}_{(1)} - \overline{Y}\right\|^2}_{k_1} & + & \underbrace{\frac{1}{\sigma^2}\left\|\vec{e}_{(1)} - \vec{e}_{(2)}\right\|^2}_{k_2-k_1} & + & \underbrace{\frac{1}{\sigma^2}\left\|\vec{e}_{(2)}\right\|^2}_{n-k_2} \end{array}$$

Siendo $\Delta\text{SS(R)} = \vec{e}^{\,t}_{(1)}\vec{e}_{(1)} - \vec{e}^{\,t}_{(2)}\vec{e}_{(2)} = \|\vec{e}_{(1)} - \vec{e}_{(2)}\|^2$.

Por tanto, se deberá rechazar H_0 (el modelo M1) a nivel α si:

$$\widehat{F} = \frac{(\text{SS(R)}_{\text{M1}} - \text{SS(R)}_{\text{M2}})/(k_2 - k_1)}{\text{SS(R)}_{\text{M2}}/(n - k_2)} > F_{k_2-k_1, n-k_2, \alpha}$$

2.7. Medidas de bondad de ajuste

2.7.1. Coeficiente de determinación

El **coeficiente de determinación** evalúa el cociente entre la variabilidad explicada por la RLM ajustada y la variabilidad global.

$$R^2 = \frac{\text{SS(E)}}{\text{SS(G)}} = 1 - \frac{\text{SS(R)}}{\text{SS(G)}}$$

- $0 \leq R^2 \leq 1$.
- A mayor valor de R^2, mayor capacidad predictiva del modelo ajustado.

- A R se le denomina **coeficiente de correlación múltiple** y proporciona el coeficiente de correlación lineal simple entre $\vec{Y}$ e $\widehat{\vec{Y}}$.

R^2 se relaciona con el cociente de varianzas estimadas de la forma:

$$1 - R^2 = \frac{\text{SS(R)}}{\text{SS(G)}} = \frac{(n-k-1)\text{MSS(R)}}{(n-1)\text{MSS(G)}} \Longrightarrow \frac{\text{MSS(R)}}{\text{MSS(G)}} = \frac{n-1}{n-k-1}(1-R^2)$$

Problema: R^2 aumenta al introducir nuevas variables explicativas en el modelo independientemente de si su efecto es o no significativo.

2.7.2. Coeficiente de determinación ajustado

El coeficiente de determinación ajustado resuelve el problema del coeficiente de determinación R^2:

$$R^2_{\text{adj}} = 1 - \frac{\text{MSS(R)}}{\text{SS(G)}}$$

- A mayor valor de R^2_{adj}, mayor capacidad predictiva del modelo.
- Relación entre R^2 y R^2_{adj}:

$$\frac{\text{MSS(R)}}{\text{MSS(G)}} = \frac{n-1}{n-k-1}(1-R^2) \iff R^2_{\text{adj}} = 1 - \frac{n-1}{n-k-1}(1-R^2)$$

- $R^2_{\text{adj}} < R^2$, para $k \geq 1$.
- R^2_{adj} puede ser negativo.
- Obsérvese que $\text{MSS(R)} = \text{MSS(G)}(1-R^2_{\text{adj}})$. Por tanto, si añadimos una nueva variable al modelo, MSS(R) disminuye sólo si R^2_{adj} aumenta.

2.7.3. Coeficiente de correlación parcial

Sea un conjunto $\mathcal{X}$ de k variables: $\mathcal{X} = \{X_1, ..., X_k\}$.

El **coeficiente de correlación parcial** entre X_i y X_j, donde $1 \leq i, j \leq k$, con repecto a $\mathcal{X}$ es el **coeficiente de correlación lineal entre** X_i **y** X_j una vez que se ha cancelado de ambas el efecto lineal de todas las $X_r \in \mathcal{X}$, con $r \neq i, j$, que denotaremos $\mathcal{X}_{ij}$.

Se puede calcular obteniendo el coeficiente de correlación lineal asociado a la regresión lineal simple:

$$e_{i.\mathcal{X}_{ij}} = \beta e_{j.\mathcal{X}_{ij}} + \varepsilon$$

donde $e_{i.\mathcal{X}_{ij}}$ y $e_{j.\mathcal{X}_{ij}}$ son los residuos de las regresiones lineales múltiples de X_i y X_j respecto al resto de las variables en $\mathcal{X}_{ij}$.

En RLM es útil examinar los coeficientes de correlación parcial entre la respuesta Y y cada explicativa X_j con respecto al resto de explicativas $\mathcal{X}_j = \mathcal{X} - \{X_j\}$, que denotaremos por $r_{YX_j,\mathcal{X}_j}$, $k = 1, .., k$.

- El coeficiente de correlación parcial representa el **nivel de relación lineal neto** entre Y y X_j.
- Representa el grado de relación lineal entre ambas variables que no es achacable al resto de variables.
- Si entre Y y X_j se tiene un valor elevado de correlación lineal ero sensiblemente más bajo de correlación parcial, cabe interpretar una **relación espuria** entre ambas o achacable a otras variables.

Los valores de $r_{YX_j,\mathcal{X}_j}$ pueden calcualrse a partir de los estadísticos de Wald.

$$\text{Sea} \quad \widehat{T}_j = \frac{\widehat{\beta}_j}{\sqrt{q_{jj}\text{MSS(R)}}} \implies r_{YX_j,\mathcal{X}_j} = \frac{\widehat{T}_j^2}{\widehat{T}_j^2 + n - k - 1}$$

2.8. Predicción del modelo

2.8.1. Estimación de la respuesta media condicionada

Dado un modelo RLM ajustado, sea $\vec{X}_0 = (1, X_{01}, ..., X_{0K})^t$ un valor específico arbitrario de las variables explicativas del modelo. El valor estimado de la respuesta media de Y cuando $\vec{X} = \vec{X}_0$ es:

$$m_0 = m(\vec{X}_0) = \mathbb{E}(Y|\vec{X}_0) = \vec{X}_0^t\vec{\beta} = \beta_0 + \sum_{j=1}^{k}\beta_j X_{0j}$$

y viene dado por:

$$\widehat{m}_0 = \widehat{m}(\vec{X}_0) = \vec{X}_0^t\widehat{\vec{\beta}} = \widehat{\beta}_0 + \sum_{j=1}^{k}\widehat{\beta}_j X_{0j}$$

- $\widehat{m}_0$ es insesgado por serlo $\widehat{\vec{\beta}}$. En efecto, $\mathbb{E}(\widehat{m}_0) = \vec{X}_0^t\mathbb{E}(\widehat{\vec{\beta}}) = \vec{X}_0^t\vec{\beta} = m_0$.
- La varianza de $\widehat{m}_0$ viene dada por:

$$\begin{aligned}\text{Var}(\widehat{m}_0) &= \mathbb{E}\Big[(\widehat{m}_0 - m_0)^2\Big] = \mathbb{E}\Big[\vec{X}_0^t(\widehat{\vec{\beta}} - \vec{\beta})(\widehat{\vec{\beta}} - \vec{\beta})^t\vec{X}_0\Big] \\ &= \vec{X}_0^t\mathbf{Var}(\widehat{\vec{\beta}})\vec{X}_0 = \sigma^2\underbrace{\vec{X}_0^t(\mathbf{X}^t\mathbf{X})^{-1}\vec{X}_0}_{\nu_{00}} = \sigma^2\nu_{00}\end{aligned}$$

2.8.2. Leverage en RLM

Premisa: $\text{Var}(\widehat{m}_0) = \text{Var}(\widehat{m}(\vec{X}_0)) = \sigma^2\nu_{00}$, con $\nu_{00} = \vec{X}_0^t(\mathbf{X}^t\mathbf{X})^{-1}\vec{X}_0$.

Si $\vec{X}_0$ coincide con el i-ésimo dato muestral: $\vec{X}_0 = \vec{X}_i = (1, X_{i1}, ..., X_{ik})^t$, para $1 \leq i \leq n$, entonces:

$$\nu_{ii} = \vec{X}_i^t(\mathbf{X}^t\mathbf{X})^{-1}\vec{X}_i = h_{ii}$$

Es decir, ν_{ii} es el i-esimo elemento diagonal de la matriz sombrero, el i-ésimo leverage.

Denotando por $\tilde{\vec{X}}_i = (X_{i1}, ..., X_{ik})^t$, por $\overline{\vec{X}}$ al vector de medias y $\mathbf{S}_{XX}$ a la matiz de varinazas de las k variables explicativas respectivamente, se prueba que:

$$h_{ii} = \vec{X}_i^t(\mathbf{X}^t\mathbf{X})^{-1}\vec{X}_i = \frac{1}{n}\left[1 + (\tilde{\vec{X}}_i - \overline{\vec{X}})^t\mathbf{S}_{XX}^{-1}(\tilde{\vec{X}}_i - \overline{\vec{X}})\right]$$

donde $\text{DM}_i = (\tilde{\vec{X}}_i - \overline{\vec{X}})^t\mathbf{S}_{XX}^{-1}(\tilde{\vec{X}}_i - \overline{\vec{X}})$ es el cuadrado de la distancia de Mahalanobis entre $\tilde{\vec{X}}_i$ y $\overline{\vec{X}}$.

Además, si $\vec{X}_0 = \vec{X}_i$ entonces $\nu_{00} = h_{ii}$:

- $\dfrac{1}{n} \leq h_{ii} \leq 1 \implies \dfrac{\sigma^2}{n} \leq \text{Var}(\widehat{Y}_i) \leq \sigma^2$.
- Varianza mínima = precisión máxima para estimar la respuesta media en $\tilde{\vec{X}}_i = \overline{\vec{X}}$.
- Mayor varianza = menor precisión = $\tilde{\vec{X}}_i$ se aleja de $\overline{\vec{X}}$.

Denotando $n_0 = \dfrac{1}{\nu_{00}}$ se puede reescribir:

$$\text{Var}(\widehat{m}(\vec{X}_0)) = \sigma^2\nu_{00} = \frac{\sigma^2}{n_0}$$

Como en RLS, n_0 se interpreta como el **número efectivo o equivalente** de datos muestrales para estimar $\widehat{m}(\vec{X}_0)$. Las cotas de los leverage no se aplican a ν_{00}, pues dependen de un X_0 arbitrario, de modo que ν_{00} será arbitrariamente grande en puntos muy alejados de las variables explicativas (extrapolación).

2.8.3. Inferencia sobre la media condicionada en RLM

Bajo las hipótesis estructurales del modelo de RLM, $\widehat{m}_0$ es una combinación lineal de normales:

$$\frac{\widehat{m}_0 - m_0}{\sigma\sqrt{\nu_{00}}} \sim N(0,1) \iff \frac{\widehat{m}_0 - m_0}{\sqrt{\nu_{00}\text{MSS(R)}}} \sim t_{n-k-1}$$

- Intervalo de confianza al $100(1-\alpha)\,\%$ para $m_0 = \mathbb{E}(Y|\vec{X} = \vec{X}_0)$:

$$\left(\widehat{m}_0 - \sqrt{\frac{\text{MSS(R)}}{n_0}}t_{n-k-1,\alpha/2},\ \widehat{m}_0 + \sqrt{\frac{\text{MSS(R)}}{n_0}}t_{n-k-1,\alpha/2}\right)$$

- Contrastes de hipótesis:

$$\begin{cases} H_0 : m_0 = m_0^* \\ H_1 = m_0 \neq m_0^* \end{cases}, \quad \text{donde } \widehat{T}_{m_0} = \frac{\widehat{m}_0 - m_0^*}{\sqrt{\nu_{00}\text{MSS(R)}}}$$

  - Rechazar H_0 al nivel α si: $\widehat{T}_{m_0} \notin (-t_{n-k-1,\alpha/2}, t_{n-k-1,\alpha/2})$.
  - p-valor: $p = 2\mathbb{P}(t_{n-k-1} > |\widehat{T}_{m_0}|)$

2.8.4. Predicción de una nueva observación

En RLM, la predicción Y_0 de la respuesta apra un nuevo valor $\vec{X} = \vec{X}_0$ coincide con la estimación de la respuesta media en $\vec{X} = \vec{X}_0$.

$$\widehat{Y}_0 = \vec{X}_0\widehat{\vec{\beta}} = \widehat{\beta}_0 + \sum_{j=1}^{k} \widehat{\beta}_j X_{0j}$$

- $\widehat{Y}_0$ es insesgado por serlo $\widehat{\vec{\beta}}$: $\mathbb{E}(\widehat{Y}_0) = \vec{X}_0^t\mathbb{E}(\widehat{\vec{\beta}}) = \vec{X}_0^t\vec{\beta} = \mathbb{E}(Y_0)$
- El error cuadrático medio de predicción es:

$$\begin{aligned}\mathbb{E}(\widehat{Y}_0 - Y_0)^2 &= \mathbb{E}(\widehat{Y} - m_0 + m_0 - Y_0)^2 = \mathbb{E}(\widehat{Y}_0 - m_0)^2 + \mathbb{E}(Y_0 - m_0)^2 \\ &= \mathbb{E}(\widehat{m}_0 - m_0)^2 + \mathbb{E}(\varepsilon_0)^2 = \mathsf{Var}(\widehat{m}_0) + \mathsf{Var}(\varepsilon_0) \\ &= \sigma^2\nu_{00} + \sigma^2 = \sigma^2(\nu_{00} + 1) = \sigma^2\Big(1/n_0 + 1\Big)\end{aligned}$$

- Nótese que $\mathsf{Cov}(\widehat{m}_0, \varepsilon_0) = 0$ porque (X_0, Y_0) no es un punto muestral.

2.8.5. Inferencia sobre una predicción en RLM

Bajo las hipótesis estructurales del modelo de RLM, el error de predicción:

$$e_0 = \widehat{Y}_0 - Y_0 \sim N(0, \sigma\sqrt{1+\nu_{00}}) \iff \frac{\widehat{Y}_0 - Y_0}{\sqrt{(1+\nu_{00})\mathsf{MSS(R)}}} \sim t_{n-k-1}$$

- Intervalo de predicción al nivel $100(1-\alpha)\,\%$ para Y_0:

$$\left(\widehat{Y} - \sqrt{(1+\nu_{00})\mathsf{MSS(R)}}t_{n-k-1,\alpha/2},\ \ \widehat{Y} + \sqrt{(1+\nu_{00})\mathsf{MSS(R)}}t_{n-k-1,\alpha/2}\right)$$

- Contrastes de hipótesis:

$$\begin{cases} H_0 : Y_0 = m_0^* \\ H_1 = Y_0 \neq m_0^* \end{cases}, \quad \text{donde } \widehat{T}_{Y_0} = \frac{\widehat{Y}_0 - Y_0^*}{\sqrt{(1+\nu_{00})\mathsf{MSS(R)}}}$$

  - Rechazar H_0 al nivel α si: $\widehat{T}_{Y_0} \notin (-t_{n-k-1,\alpha/2}, t_{n-k-1,\alpha/2})$.
  - p-valor: $p = 2\mathbb{P}(t_{n-k-1} > \widehat{T}_{Y_0})$.

2.9. El problema de la multicolinealidad

El problema de la multicolinealidad surge cuando algunas o todas las variables explicativas X_j están altamente correlacionadas entre sí. En tal caso, los efectos sobre la respuesta de las variables explicativas están confundidos y es muy complejos separarlos, lo que conduce a estimaciones de los coeficientes β_j inestables y con varianza elevada.

- Los estimadores $\widehat{\beta}_j$ presentan varianzas muy elevadas.
- Los estimadores $\widehat{\beta}_j$ son muy dependientes entre sí.

2.9.1. Ilustración del efecto de la multicolinealidad con $k = 2$

$$\text{Modelo de RLM:} \quad Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \varepsilon_i.$$

Modelo en desviaciones:

$$Y_i - \overline{Y} = \beta_1(X_{i1} - \overline{X}_1) + \beta_2(X_{i2} - \overline{X}_2) + \varepsilon_i \iff \tilde{Y}_i = \beta_1 \tilde{X}_{i1} + \beta_2 \tilde{X}_{i2} + \varepsilon_i$$

Por tanto,

$$\tilde{\mathbf{X}}^t\tilde{\mathbf{X}} = \begin{pmatrix} \sum_{i=1}^n \tilde{X}_{i1}^2 & \sum_{i=1}^n \sum_{l=1}^n \tilde{X}_{i1}\tilde{X}_{l2} \\ \sum_{i=1}^n \sum_{l=1}^n \tilde{X}_{i1}\tilde{X}_{l2} & \sum_{i=1}^n \tilde{X}_{i2} \end{pmatrix} = n \begin{pmatrix} S_1^2 & S_{12} \\ S_{21} & S_2^2 \end{pmatrix}$$

Como $S_{12} = rS_1S_2$ (por definción) y $\det(\tilde{\mathbf{X}}^t\tilde{\mathbf{X}}) = nS_1^2S_2^2(1-r^2)$:

$$\text{Var}(\widehat{\vec{b}}) = \sigma^2(\tilde{\mathbf{X}}^t\tilde{\mathbf{X}})^{-1} = \frac{\sigma^2}{n} \begin{pmatrix} \frac{1}{S_1^2(1-r^2)} & \frac{-r}{S_1S_2(1-r^2)} \\ \frac{-r}{S_1S_2(1-r^2)} & \frac{1}{S_2^2(1-r^2)} \end{pmatrix}$$

Concluyendo, si el cuadrado del coeficiente de correlación lineal entre X_1 y X_2 es próximo a uno, la varianza de la estimación del efecto de esta variable también es muy grande y las estimaciones muy dependientes.

$$\text{Var}(\widehat{\beta}_j) = \frac{\sigma^2}{nS_j^2(1-r^2)}, \quad \text{Cov}(\widehat{\beta}_1, \widehat{\beta}_2) = \frac{-r\sigma^2}{nS_1S_2(1-r^2)}$$

Se prueba que para el caso general k:

$$\text{Var}(\widehat{\beta}_j^{\text{RLM}}) = \text{Var}(\widehat{\beta}_j^{\text{RLS}})(1 - R_{(j)}^2)^{-1}, \quad \text{para } j = 1, ..., k$$

donde $R_{(j)}$ es el coeficiente de correlación múltiple entre X_j y el resto de variables explicativas.

2.9.2. Efectos de una elevada multicolinealidad

- Estimaciones poco precisas de los coeficientes β_j asociados a las explicativas altamente correlacionadas.
- Pequeños cambios en el vector de respuestas o añadir/quitar una explicativa puede suponer variaciones sustanciales en los valores estimados $\widehat{\beta}_j$.
- Resultados aparentemente contradictorios. El test de regresión conjunto es significativo pero todas las pruebas individuales sobre los coeficientes individuales resultan no significativas.
- La predicción y el ajuste del modelo no se ve afectado. El ajuste es útil para predecir, pero no es válido para un análisis estructural del modelo.
- Incorrecta especificación del modelo, de modo que no se puede extraer más información de la muestra de la que contiene.

2.10. Detección de la multicolinealidad

2.10.1. Matriz de correlaciones $\mathbb{R}$

$$\mathbb{R} = \begin{pmatrix} 1 & r_{12} & \cdots & r_{1(k-1)} & r_{1k} \\ & 1 & \cdots & r_{2(k-1)} & r_{2k} \\ & & \ddots & \vdots & \vdots \\ & & & 1 & r_{(k-1)k} \\ & & & & 1 \end{pmatrix}, \quad \text{con} \quad r_{ij} = \text{Corr}(X_i, X_j)$$

- Problema: Puede existir una relación lineal muy alta con más de dos variables explicativas y la matriz de correlaciones no detecta esa correlación significativa.
- Solución: Acudir a la **matriz de correlaciones parciales** o a la matriz $\mathbb{R}^{-1}$, que tiene en cuenta la dependencia conjunta.

2.10.2. Factores de inflación de la varianza (FIV)

El FIV_j mide el **factor de incremento en la varianza** de $\widehat{\beta}_j$ en RLM con respecto a su varianza en RLS y se define para $j = 1, ..., k$ como:

$$\text{FIV}_j = \frac{1}{1 - R^2_{(j)}}$$

Y la tolerancia, como el inverso del FIV:

$$\text{Tol}_j = \text{FIV}^{-1} = 1 - R^2_{(j)}$$

- $\text{FIV}_j > 1, \forall j$.
- Si la correlación entre X_j y el resto de variables explicativas es muy alta, $R^2_{(j)} \approx 1$. Por tanto, FIV_j es muy elevado.
- $\text{FIV}_j = \text{diag}_j(\mathbb{R}^{-1})$
- Umbral para asumir alta multicolinealidad: $\text{FIV}_j > 10$ o $\text{FIV}_j > (1 - R^2)^{-1}$.

2.10.3. Índice de condicionamiento (IC)

La multicolinealidad se traduce en proximidad a la singularidad (rango menor que k) de la matriz $\mathbf{X}^t\mathbf{X}$ o $\mathbb{R}$. El **índice de singularidad** de una matriz es llamado índice de condicionamiento:

$$\text{IC} = \sqrt{\frac{\text{máx}_{1 \leq j \leq k} \lambda_j}{\text{mín}_{1 \leq j \leq k} \lambda_j}}$$

- $\text{IC} > 30$: Multicolinealidad elevada.

- $10 \leq \text{IC} < 30$: Multicolinealidad moderada.
- $\text{IC} < 10$: Matriz bien condicionada.

2.10.4. PCA de variables explicativas

El anáisis de componentes principales (PCA) de las variables explicativas conduce a combinaciones lineal incorreladas de las variables explicativas que explican la varianza total de las variables. Si un número pequeño de componente principales explica un porcentaje elevado de varianza global, entonces es indicativo de que hay variables explicativas correlacionadas y por tanto alta multicolinealidad.

2.10.5. Tratamiento. Elminación de variables y ECM

El **error cuadrático medio** del vector de estimadores $\widehat{\vec{\beta}}$ es:

$$\begin{aligned}\text{ECM}(\widehat{\vec{\beta}}) &= \mathbb{E}\Big((\widehat{\vec{\beta}} - \vec{\beta})^t(\widehat{\vec{\beta}} - \vec{\beta})\Big) = \sum_{j=0}^{k} \mathbb{E}(\widehat{\beta}_j - \beta_j)^2 \\ &= \sigma^2 \text{tr}(\mathbf{X}^t\mathbf{X})^{-1} = \sigma^2 \sum_{j=0}^{k} \frac{1}{\lambda_j}\end{aligned}$$

donde λ_j son los autovalores de la matriz $\mathbf{X}^t\mathbf{X}$. Si $\mathbf{X}^t\mathbf{X}$ es casi singular, algún λ_j es próximo a cero y el $\text{ECM}(\widehat{\vec{\beta}})$ puede ser grande.

- Eliminar una variable X_i correlada con X_j minora la varianza estimada del efecto de X_j, pero en contrapartida la estimación de su efecto es sesgada.
- Será interesante si la reducción de la varianza conduce a una reducción del ECM.

Ejemplo con $k = 2$:

Relación correcta	Se elimina X_2
$Y = \beta_1 X_1 + \beta_2 X_2 + \varepsilon_{12}$	$Y = \beta_1^* X_1 + \varepsilon_1$
$\mathbb{E}(\widehat{\beta}_1) = \beta_1$	$\mathbb{E}(\widehat{\beta}_1^*) = \beta_1 + \beta_2 r_{12}\frac{S_2}{S_1}$
$\text{Var}(\widehat{\beta}_1) = \frac{\sigma^2}{nS_1^2(1-r_{12}^2)}$	$\text{Var}(\widehat{\beta}_1^*) = \frac{\sigma^2}{nS_1^2}$
$\text{ECM}(\widehat{\beta}_1) = \frac{\sigma^2}{nS_1^2(1-r_{12}^2)}$	$\text{ECM}(\widehat{\beta}_1^*) = \left(\beta_2 r_{12}\frac{S_2}{S_1}\right)^2 + \frac{\sigma^2}{nS_1^2}$

Por tanto:

$$\text{ECM}(\widehat{\beta}_1^*) < \text{ECM}(\widehat{\beta}_1) \iff \frac{1}{1-r_{12}^2} > \left(\beta_2 r_{12}\frac{S_2}{S_1}\right)^2 \iff 1 > \left(\frac{\beta_2}{\text{sd}(\widehat{\beta}_2)}\right)^2$$

- Si $r_{12}^2 \approx 1$, $\widehat{\beta}_1^*$ es sesgado, pero con menor ECM que $\widehat{\beta}_1$.
- Eliminar regresoras X_i con estadístico t menor que uno podría mejorar en promedio el ECM de los estimadores de los efectos del resto de explicativas.

2.11. Error de especificación

2.11.1. Omisión de variables relevantes

- Estimaciones sesgadas de los coeficientes β_j.
- El sesgo en la estimación β_j es tanto mayor como su correlación con las variables emitidas. Si es incorrelada con ellas, entonces no existe tal sesgo.
- La varianza de los estimadores β_j queda afectada:
  - Si X_j es incorrelada con las omitidas, los residuos contienen el efecto de las omita. Se incrementa el MSS(R) y Var$(\widehat{\beta}_j)$.
  - Si X_j es correlada con las omitidas, Var$(\widehat{\beta}_j)$ disminuye al cancelar el FIV$_j$.
- **Consecuencias**:
  - MSS(R) sesgada por exceso.
  - Contrastes sesgados hacia la falta de relación.
  - Residuos afectados en su estructura.

2.11.2. Inclusión de variables irrelevantes

- Los estimadores de los efectos β_j son insesgados.
- Var$(\widehat{\beta}_j)$ puede aumentar si las variables incluidas son altamente correladas con X_j, pues incrementan el FIV$_j$.
- Si las variables incluidas son incorreladas con X_j, no incrementan la varianza de $\widehat{\beta}_j$, pero el ajuste no es eficiente al invertir grados de libertad en parámetros no relevantes.
- **Consecuencias**
  - MSS(R) es insesgado.
  - Contrastes sesgados hacia la falta de relación.
  - Los residuos no se ven afectados en su estructura.

2.12. Análisis de influencia

2.12.1. Residuos parciales

Los **residuos parciales** se emplean para examinar la relación entre Y y cada regresora X_j una vez se han eliminado los efectos del resto de regresoras. Su construcción sigue los siguientes pasos:

- Ajuste RLM: $\vec{Y}^{(j)} = \widehat{\beta}_0^{(j)} + \sum_{i \neq j} \widehat{\beta}_i^{(j)} X_i$.
- Residuos parciales respcto a X_j: $e^{(j)} = Y - \widehat{Y}^{(j)}$

Cuando X_j está incorrelada con el resto de regresoras, los coeficientes $\widehat{\beta}_i^{(j)} \approx \widehat{\beta}_i$, siendo $\widehat{\beta}_i$ el i-ésimo coeficiente en el ajuste completo, y el cómputo de los residuos parciales puede aproximarse por:

$$e^{(j)} \approx Y - \widehat{\beta}_0 + \sum_{i \neq j} \widehat{\beta}_i X_i = e + \widehat{\beta}_j X_j$$

Además de los gráficos considerados en RLS, es oportuno visualizar e interpretar:

- Gráfico de residuos frente a cada una de las variables explicativas. Pueden ayudar a identificar si la falta de linealidad o la presencia de heteroscedasticidad son achacables a una variable regresora concreta.
- Gráficos de residuos parciales $e^{(j)}$ frente a la explicativa correspondiente X_j al objeto de intuir como X_j incide en la respuesta. Son útiles para comprobar si X_j contribuye linealmente a explicar la respuesta Y (deberíamos apreciar una tendencia lineal tras la nube).

2.12.2. Robustez a priori del modelo

Los puntos heterogéneos respecto a las X_j presentan elevado **leverage**.

$$h_{ii} = \text{diag}_i(\mathbf{H}), \quad \forall i = 1, ..., n$$

- $\frac{1}{n} \leq h_{ii} \leq 1$
- $\sum_{i=1}^{n} h_{ii} = \text{tr}(\mathbf{H}) = k + 1$
- Puede asumirse que un dato es potencial outlier si: $h_{ii} > \overline{h} + 3S_h$.
- Bajo normalidad de las X_j, considerar el umbral: $h_{ii} > 2\overline{h} = 2(k+1)/n$.

2.12.3. Robustez a posteriori. Distancia de Cook

Para $i = 1, ..., n$, la distancia de Cook en el dato muestral i-ésimo $(\vec{X}_i, Y_i)$ se define como:

$$D_i = \frac{(\widehat{\vec{\beta}} - \widehat{\vec{\beta}}_{(i)})^t \mathbf{X}^t \mathbf{X} (\widehat{\vec{\beta}} - \widehat{\vec{\beta}}_{(i)})}{(k+1)\text{MSS(R)}} = \frac{(\widehat{\vec{Y}} - \widehat{\vec{Y}}_{(i)})^t (\widehat{\vec{Y}} - \widehat{\vec{Y}}_{(i)})}{(k+1)\text{MSS(R)}}$$

donde el subíndice (i) en el vector denota que se obtuvo sin el i-ésimo dato muestral.

- D_i mide el **cambio estandarizado** en el vector de coeficientes estimados y en el de predicciones.
- $\widehat{\vec{\beta}}_{(i)}$ está en la región de confianza al nivel $100(1-\alpha)\,\%$ si:

  $$D_i = \frac{(\widehat{\vec{\beta}} - \widehat{\vec{\beta}}_{(i)})^t \mathbf{X}^t \mathbf{X} (\widehat{\vec{\beta}} - \widehat{\vec{\beta}}_{(i)})}{(k+1)\text{MSS(R)}} \leq F_{k+1, n-k-1, 1-\alpha}$$

  En otro caso, la observación i es influyente.
- D_i se relaciona con los residuos estandarizados de la siguiente forma:

  $$D_i = \frac{r_i^2}{k+1} \left(\frac{h_{ii}}{1 - h_{ii}} \right)$$

2.12.4. Robustez a posteriori. DFBETAS y DFFITS

Los **estadísticos DFBETAS** evalúan el impacto de eliminar el i-esimo dato muestral sobre la estimación individual de cada coeficiente β_j, para $i = 1, ...n, j = 1, ..., k$.

$$\text{DFBETAS}_{j,(i)} = \frac{\widehat{\beta}_j - \widehat{\beta}_{j,(i)}}{\sqrt{\text{MSS(R)}_{(i)} q_{jj}}}, \quad \text{donde } q_{jj} = \text{diag}_j(\mathbf{X}^t\mathbf{X})$$

Un umbral recomendado para considerar al i-ésimo dato muestral influyente en la estimación de β_j es: $|\text{DFBETAS}_{j,(i)}| > 2/\sqrt{n}$.

Los **estadísticos DFFITS** evalúan el impacto de eliminar el i-ésimo dato muestral en la predicción individual de ese dato:

$$\text{DFFITS}_i = \frac{\widehat{Y}_i - \widehat{Y}_{i,(i)}}{\sqrt{\text{MSS(R)}_{(i)} h_{ii}}}$$

Umbral recomendado: $|\text{DFFITS}| > \frac{2}{\sqrt{(k+1)/n}}$.

2.13. Criterios de selección de variables

Asumida una relación lineal y dado un conjunto amplio de explicativas, un primer objetivo razonable es seleccionar el **mejor subconjunto de variables explicativas** en orden a producir un ajuste correcto.

Obtendremos mejores predicciones con más variables explicativas (minoramos el sesgo), pero en contrapartida el modelo ajustado tendrá menor precisión (porque la varianza crece). El objetivo es seleccionar el **subconjunto de variables regresoras que alcance un buen compromiso entre sesgo y varianza**, entre capacidad predictiva y precisión en el proceso de inferencia.

2.13.1. Estadístico C_p de Mallows

El estadístico C_p de Mallows compara el ECM de predicción del modelo completo (con k regresoras) con modelos que incluyen un subconjunto $(p-1)$ de regresoras:

$$C_p = \frac{\text{SS(R)}(p)}{\text{MSS(R)}} - (n - 2p)$$

donde $\text{SS(R)}(p)$ es la suma de cuadrados residual de un modelo con $p-1$ regresoras.

Si el modelo con p parámetros es adecuado, $\text{MSS(R)} \approx \sigma^2$, y por tanto:

$$\mathbb{E}(C_p) = \mathbb{E}\left(\frac{\text{SS(R)}(p)}{\text{MSS(R)}} - (n-2p)\right) \approx \frac{(n-p)\sigma^2}{\sigma^2} - (n-2p) = p$$

Luego si el modelo adecuado tiene p parémetros, entonces $C_p \approx p$.

2.13.2. Los estadísticos AIC y BIC

El criterio de información de Akaike (AIC) está basado en la verosimilitud del modelo ajustado incluyendo una penalización por el número de parámetros que lo definen.

$$\text{AIC} = 2p - 2\ln L(\widehat{\vec{\beta}})$$

donde p es el número de parámetrs del modelo y $L(\widehat{\vec{\beta}})$ es el máximo de la función de verosimilitud.

El BIC es una versión de AIC que tiene en cuenta el número de datos:

$$\text{BIC} = p\log n - 2\ln L(\widehat{\vec{\beta}})$$

Cuanto menor valor de AIC/BIC, mejor modelo.

2.13.3. El error de predicción PRESS

El estadístico PRESS (*Prediction Residual Error Sum of Squares*) proporciona una **medida de error de predicción global** del modelo calculada por LOOCV.

$$\text{PRESS} = \sum_{i=1}^{n}\left(Y_i - \widehat{Y}_{(i)}\right)^2 = \sum_{i=1}^{n}\left(\frac{e_i}{1-h_{ii}}\right)^2$$

donde $\widehat{Y}_{(i)}$ la predicción del i-ésimo dato con el modelo ajustado sin el i-ésimo dato.

- Modelos con un valor pequeño de PRESS generan un menor error de predicción.
- Un modelo sobre-parametrizado generará:
  - Errores de predicción pequeños sobre datos muestrales (residuos pequeños).
  - Errores de predicción grandes sobre nuevos datos (elevado PRESS).

2.14. Estrategias de selección de variables

2.14.1. Selección del mejor subconjunto de variables (*Best Subset Selection*)

1. Partir de un modelo $\mathcal{M}_0$ sin regresoras, basado en el intercept.
2. Para cada $j = 1, ..., k$, elegir el mejor modelo $\mathcal{M}_j$ con j variable regresoras entre los $\binom{k}{j}$ posibles (en base al R^2).
3. Seleccionar el mejor entre $\mathcal{M}_0, \mathcal{M}_1, ..., \mathcal{M}_k$ (en base a alguno de los criterios mencionados).

Este método conlleva un alto coste computacional. Además, si k es grande:

- Aumenta la posibilidad de hallar modelos que entrenen bien pero predigan mal sobre otros datos.
- Puede conducir a sobreajuste y alta varianza de los coeficientes estimados.

2.14.2. Selección por eliminación progresiva de variables (*Backward Stepwise*)

En cada paso se elimina la variabl que **menos mejora** aporta al ajuste.

1. Partir de un modelo completo $\mathcal{M}_k$ con todas las regresoras.
2. Para $j = k, ..., 1$:
   a) Considerar todos los j modelos que eliminan un regresora de $\mathcal{M}_j$.
   b) Elegir el mejor de los nuevos j modelos de acuerdo a algún criterio y denotarlo por $\mathcal{M}_{j-1}$.
3. Seleccionar el mejor entre $\mathcal{M}_0, \mathcal{M}_1, ..., \mathcal{M}_k$ en base a alguno de los criterios mencionados.

Desventajas:

- Poco eficiente (tal vez sólo algunas pocas variables sean significativas).
- Si k es grande o hay regresoras correlacionadas puede conducir al problema de multicolinealidad.
- Excelente para evitar excluir variables relevantes.

2.14.3. Selección por introducción progresiva de variables (*Forward Stepwise*)

En cada paso se añade la variable que **mayor mejora** aporta en el ajuste.

1. Partir de un modelo $\mathcal{M}_0$ sin regresoras, basado en el intercept.
2. Para $j = 0, ..., k-1$:
   a) Considerar todos los $k - j$ modelos que incrementan una regresora en $\mathcal{M}_j$.
   b) Elegir el mejor de los nuevos $k - j$ modelos de acuerdo a un criterio y denotarlo $\mathcal{M}_{j+1}$.
3. Seleccionar el mejor entre $\mathcal{M}_0, \mathcal{M}_1, ..., \mathcal{M}_k$ en base a alguno de los criterios mencionados.

Ventajas y desventajas:

- Computacionalmente eficiente.
- No es capaz de elimianr variables en etapas posteriores cuando la introducción de otras nuevas hacen innecesaria su presencia.

2.14.4. Regresión paso a paso (*Stepwise Regression*)

Trata de evitar los inconvenientes de la introducción progresiva de variables manteniendo su eficiencia computacional.

Para ello, en cada paso, tras la entrada de una nueva variable establece una revisión del conjunto de variables seleccionando y evalúa si alguna debe salir.

2.15. Validación del modelo

2.15.1. Contraste de validación

Supóngase que se ha realizado un ajuste: $\vec{Y}_1 = \mathbf{X}_1\vec{\beta}_1 + \vec{\varepsilon}_1$ con n_1 datos (Modelo M1) y se dispone de $n_2 > k+1$ datos nuevos. Se pueden plantear las siguientes hipótesis:

$$\text{Modelo simple M1 : } \begin{cases} \vec{Y}_1 = \mathbf{X}_1\vec{\beta}_1 + \vec{\varepsilon}_1 \\ \vec{Y}_2 = \mathbf{X}_2\vec{\beta}_2 + \vec{\varepsilon}_2 \end{cases} \qquad \text{por tanto } \vec{\beta}_1 = \vec{\beta}_2$$

$$\text{Modelo complejo M2 : } \begin{cases} \vec{Y}_1 = \mathbf{X}_1\vec{\beta}_1 + \vec{\varepsilon}_1 \\ \vec{Y}_2 = \mathbf{X}_2\vec{\beta}_2 + \vec{\varepsilon}_2 \end{cases} \qquad \text{por tanto } \vec{\beta}_1 \neq \vec{\beta}_2$$

- En M1 se unen los $n_1 + n_2$ datos y se ajusta una RLM a todos ellos ($\vec{\beta}_1 = \vec{\beta}_2$). Se obtiene una SS(R)_{M1} con $n_1 + n_2 - k - 1$ grados de libertad.
- En M2 se ajusta por separado cada conjunto ($\vec{\beta}_1 \neq \vec{\beta}_2$), resultando $\text{SS(R)}_{\text{M2(1)}}$ con $n_1 - k - 1$ grados de libertad y $\text{SS(R)}_{\text{M2(2)}}$ con $n_2 - k - 1$ grados de libertad.

Se rechaza M1 en favor de M2 con un nivel de significación α si:

$$\widehat{F} = \frac{\Big(\text{SS(R)}_{\text{M1}} - \text{SS(R)}_{\text{M2(1)}} - \text{SS(R)}_{\text{M2(2)}}\Big)/(k+1)}{\Big(\text{SS(R)}_{\text{M2(1)}} + \text{SS(R)}_{\text{M2(2)}}\Big)/(n_1 + n_2 - 2(k+1))} > F_{k+1,n_1+n_2-2(k+1),\alpha}$$

Este contrastes es más potente para **detectar errores de especificación asociados a sesgos** en los parámetros por omisión de variables relevantes o errores en los datos. En caso de que $n_2 \leq k+1$, se puede considerar:

$$\text{Modelo simple M1 : } \begin{cases} \vec{Y}_1 = \mathbf{X}_1\vec{\beta}_1 + \vec{\varepsilon}_1 \\ \vec{Y}_2 = \mathbf{X}_2\vec{\beta}_2 + \vec{\varepsilon}_2 \end{cases} \qquad \text{por tanto } \vec{\beta}_1 = \vec{\beta}_2$$

$$\text{Modelo complejo M2 : } \begin{cases} \vec{Y}_1 = \mathbf{X}_1\vec{\beta}_1 + \vec{\varepsilon}_1 \\ \vec{Y}_2 = \vec{\theta} + \vec{\varepsilon}_2 \end{cases} \qquad \text{donde } \vec{\theta} \text{ es un vector de constantes}$$

Se rechaza M1 en favor de M2 con un nivel de significación α si:

$$\widehat{F} = \frac{\Big(\text{SS(R)}_{\text{M1}} - \text{SS(R)}_{\text{M2(1)}} - \text{SS(R)}_{\text{M2(2)}}\Big)/n_2}{\text{SS(R)}_{\text{M2(1)}}/(n_1 - k - 1)} > F_{n_2,n_1-k-1,\alpha}$$

Este contraste es más potente para **detectar cambios fundamentales en la especificación del modelo** (la muestra de n_2 datos requiere un modelo más complejo que M1).

2.15.2. Validación cruzada

El proceso de validación cruzada consiste en dividir una muestra en dos (de forma aleatoria) y usando una de las submuestras para ajustar el modelo (**muestra de entrenamiento**) y la otra para evaluar su comportamiento (**muestra test o de validación**).

Una desventaja de este procedimiento es que el ajuste se realiza con solo una porción de los datos, incrementando así el **sesgo** de la estimación. En consecuencia, la **tasa de error** en la muestra de validación puede ser muy variable.

Las vías más habituales para relizar validación cruzada son las que se exponen a continuación.

2.15.3. K-fold cross-validation

Requiere prefijar un parámetro k, que indica el número de grupos aproximadamente de igual tamaño en los que se fragmentará el conjunto de datos. Una vez fijado k, el algoritmo procede de la siguiente manera:

1. Distribuir al azar los datos en k grupos.
2. Para cada uno de los k grupos.
   a) Ajustar una RLM con los datos de los $k-1$ grupos restantes.
   b) Predecir la respuesta para los datos del grupo seleccionado con el modelo ajustado.
   c) Evaluar el error comparando valores reales y predicciones.
3. Promediar los errores obtenidos con los k grupos (ajustes).

El resultado es una predicción por validación cruzada para la totalidad de observaciones. Cuanto menor es k, mayor tamaño de las submuestras de entrenamiento, menor sesgo y mayor varianza de los resultados. En la práctica es muy común elegir $= 3, 5, 10$ 0 20.

2.15.4. Leave-One-Out Cross-Validation (LOOCV)

Tomar el caso extremo de k-fold Cross-Validation: $k = n$.

1. Se separa un dato i y se ajusta el modelo con los $n-1$ datos restantes.
2. Se usa el ajuste para predecir el dato separado $\widehat{Y}_{(i)}$.
3. Se evalúa el error de predicción:
$$e_{(i)} = Y_i - \widehat{Y}_{(i)}$$

Si el **error global** que se evalúa al terminar el proceso LOOCV es el error cuadrático, entonces se obtiene el estadístico PRESS:

$$\sum_{i=1}^{n} e_{(i)}^2 = \sum_{i=1}^{n} (Y_i - \widehat{Y}_{(i)})^2$$

2.15.5. Coeficiente de robustez de un ajuste

$$B^2 = \frac{\text{SS(R)}}{\text{PRESS}} = \frac{\sum_{i=1}^{n} (Y_i - \widehat{Y}_i)^2}{\sum_{i=1}^{n} (Y_i - \widehat{Y}_{(i)})^2}$$

- $B^2 \in (0, 1)$ y representa una medida de la robustez del ajuste.
- B^2 se acercará a 1 en la medida en que $\widehat{Y}_{(i)} \approx \widehat{Y}_i$ y a 0 si estas predicciones son muy diferentes.

2.16. Regresoras cualitativas

2.16.1. Modelo lineal con una regresora cualitativa y una cuantitativa

Sea $\{(X_1, Z_1, Y_1), ..., (X_n, Z_n, Y_n)\}$ una muestra aleatoria simple de (X, Z, Y). Definimos $d-1$ variables ficticias (dummy): $\{I_2, ..., I_d\}$, indicadoras del atributo de una observación.

$$I_j(Z) = \mathbb{I}(Z = \mathcal{Z}_j) = \begin{cases} 1 & \text{si } Z = \mathcal{Z} \\ 0 & \text{si } Z \neq \mathcal{Z} \end{cases} \quad \text{para } j = 2, ..., d$$

- Si $Z_i = \mathcal{Z}_1$, entonces $I_j(Z_i) = 0, \forall 2 \leq j \leq j$.
- Si $Z_i = \mathcal{Z}_j$, entonces $I_j(Z_i) = 1$ y $I_k(Z_i)$, $\forall 1 < j, k \leq d, k \neq j$.

Asumiendo que el efecto de X sobre Y no depende del atributo de Z_1 se tiene:

Modelo matemático

Para $i = 1, ..., n$.

$$Y_i = \beta_0 + \beta_1 X_i + \sum_{j=2}^{d} \tau_j I_j(Z_i) + \varepsilon_i, \quad \text{donde } \varepsilon_i \text{ i.i.d. } \sim N(0, \sigma)$$

$$\mathbb{E}(Y|X = X_i, Z = \mathcal{Z}_j) = \beta_0 + \beta_1 X_i + \tau_j$$

- β_0: El intercept o valor esperado de Y cuando $X = 0$ y $Z = \mathcal{Z}_1$.
- β_1: Tasa de cambio del valor esperado de Y por unidad de incremento en X con independencia del valor del atributo Z_i.
- $\tau_j, j = 2, ..., d$: Efeto medio incremental sobre el valor esperado de Y cuando $X = 0$ y Z pasa de ser $\mathcal{Z}_1$ a $\mathcal{Z}_j$.
- σ: Desviación estándar de las respuestas para valores arbitrario de X y Z.

2.16.2. Estimador OLS de los efectos de los atributos

Los estimadores OLS de la diferencia de dos efectos de la variable cualitativa Z vienen dados por:

$$\widehat{\tau}_j - \widehat{\tau}_k = \overline{Y}_{Z=\mathcal{Z}_j} - \overline{Y}_{Z=\mathcal{Z}_k} - \widehat{\beta}_1(\overline{X}_{Z=\mathcal{Z}_j} - \overline{X}_{Z=\mathcal{Z}_k})$$

En particular se prueba que:

$$\widehat{\beta}_0 = \overline{Y}_{Z=\mathcal{Z}_1} - \widehat{\beta}_1(\overline{X}_{Z=\mathcal{Z}_1} - \overline{X})$$

de donde se sigue que, para $2 \leq j \leq d$:

$$\widehat{\tau}_j = \overline{Y}_{Z=\mathcal{Z}_j} - \overline{Y}_{Z=\mathcal{Z}_1} - \widehat{\beta}_1(\overline{X}_{Z=\mathcal{Z}_k} - \overline{X}_{Z=\mathcal{Z}_1})$$

2.17. Regresión polinómica con una variable

Modelo de regresión polinómica

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + ... + \beta_k X^k + \varepsilon_i, \quad \text{con } \varepsilon_i \text{ i.i.d. } \sim N(0, \sigma)$$

Haciendo $Z_j = X^j, j = 1, ..., k$:

$$Y = \beta_0 + \beta_1 Z_1 + \beta_2 Z_2 + ... + \beta_k Z_k + \varepsilon_i, \quad \text{con } \varepsilon_i \text{ i.i.d. } \sim N(0, \sigma)$$

- **Problema**: Encontrar k.
- **Solución**: Aumentar progresivamente k hasta que β_k no difiera significativamente de 0.
- **Peligros**: Caer en un sobreajuste (modelar la perturbación aleatorio) y provocar multicolinealidad.

Para evitar sobreajuste, examinar la linealidad de los residuos tras cada ajuste. Si se ajusta un polinomio de orden r cuando el orden correcto es $k > r$, los residuos toman la forma:

$$e_i = Y_i - \widehat{Y}_i = \varepsilon_i + \sum_{j=0}^{r} (\beta_j - \widehat{\beta}_j) X^j + \sum_{j=r+1}^{k} \beta_j X^j$$

Y las diferencias $(\beta_k - \widehat{\beta}_j)$ serán pequeñas, de modo que los residuos estarán dominados por los términos no incluidos en el ajuste y mostrarán no linealidad.

2.17.1. Elección del grado del polinomio

1. Estimar el modelo con las X en desviaciones (se atenúa la dependencia).

$$\widehat{Y} = \widehat{\beta}_0 + \widehat{\beta}_1 (X - \overline{X}) + \widehat{\beta}_2 (X - \overline{X})^2 + \cdots + \widehat{\beta}_k (X - \overline{X})^k$$

2. Utilizar polinomios ortogonales:

$$Y = \gamma_0 + \gamma_1 P_1(X) + \gamma_2 P_2(X) + \cdots + \gamma_k P_k(X) + \varepsilon$$

donde los términos $P_j(X)$ son polinomios de X de orden j cumpliendo la condición de orgonalidad:

$$\sum_{i=1}^{n} P_j(X_i) P_h(X_i) = 0, \forall j \neq h$$

Por la ortogonalidad se prueba que los coeficientes estimados toman la forma:

$$\widehat{\gamma}_j = \frac{\sum_{i=1}^{n} Y_i P_i(X_i)}{\sum_{i=1}^{n} P_j^2(X_i)}$$

Tema 3: Regresión Logística

3.1. Planteamiento del problema

Supongamos un problema de regresión donde se dispone de k variables regresoras pero la variable respuesta Y es binaria (sólo puede tomar de valores 0 o 1). Dada una observación $\vec{X}_i = (1, X_{i1}, ..., X_{ik})^t$ se tiene un nuevo enfoque del modelo de regresión:

$$\mathbb{E}(Y|\vec{X}_i) = \vec{\beta}^t\vec{X}_i = 1 \cdot \mathbb{P}(Y = 1|\vec{X}_i) + 0 \cdot \mathbb{P}(Y = 0|\vec{X}_i) = \mathbb{P}(Y = 1|\vec{X}_i) = p_i$$

La predicción $\widehat{\vec{Y}}_i$ estima la probabilidad de $Y = 1$ cuando las regresoras toman los valores $\vec{X}_i$.

Para conseguir que $\vec{\beta}^t\vec{X}_i = \beta_0 + \sum_{j=1}^{k} \beta_j X_{ij} \in [0, 1]$ se aplica una **transformación** $\chi : \mathbb{R} \to [0, 1]$ para que:

$$p_i = \chi(\vec{\beta}^t\vec{X}_i) \in [0, 1]$$

Si se elige como función χ la distribución normal, entonces el modelo resultante se denomina **modelo probit**:

Modelo probit: $\mathbb{E}(Y|\vec{X}_i) = p_i = \mathbb{P}(Y = 1|\vec{X}_i) = \Phi(\vec{\beta}^t\vec{X}_i)$

Si se elige como función χ la distribución logística: $F(z) = \dfrac{e^z}{1+e^z} = \dfrac{1}{1+e^{-z}}$. Entonces el modelo resultantes se denomina **modelo logístico**.

Modelo logístico: $\mathbb{E}(Y|\vec{X}_i) = p_i = \mathbb{P}(Y = 1|\vec{X}_i) = \dfrac{1}{1+e^{-\vec{\beta}^t\vec{X}_i}} = \dfrac{e^{\vec{\beta}^t\vec{X}_i}}{1+e^{\vec{\beta}^t\vec{X}_i}}$

3.2. El modelo logístico

El modelo logístico se define como:

$$\mathbb{E}(Y|\vec{X}_i) = p_i = \mathbb{P}(Y = 1|\vec{X}_i) = \frac{1}{1+e^{-\vec{\beta}^t\vec{X}_i}}$$

Y modeliza la esperanza de una variable de Bernoulli (Y) condicionada a los valores que toman un conjunto de variables explicativas $(\vec{X}_i)$. En este modelo la esperanza condicionada de Y es una **función no lineal** de las variables regresoras.

Para facilitar la interpretación del modelo podemos dar un nuevo enfoque:

$$p_i = \mathbb{P}(Y = 1|\vec{X}_i) = \frac{1}{1+e^{-\vec{\beta}^t\vec{X}_i}}$$

$$1 - p_i = \mathbb{P}(Y = 0|\vec{X}_i) = 1 - \frac{1}{1 + e^{-\vec{\beta}^t \vec{X}_i}} = \frac{e^{-\vec{\beta}^t \vec{X}_i}}{1 + e^{-\vec{\beta}^t \vec{X}_i}} = \frac{1}{1 + e^{\vec{\beta}^t \vec{X}_i}}$$

Y por tanto:

$$\ln \frac{p_i}{1 - p_i} = \vec{\beta}^t \vec{X}_i$$

Odds

Los odds caracterizan la distribución de Y en términos del cociente entre la probabilidad de éxito y la probabilidad de fracaso. Se definen como:

$$\text{odds}(Y) = \frac{\mathbb{P}(Y = 1)}{\mathbb{P}(Y = 0)} = \frac{p}{1 - p}$$

Conocidos los odds, se obtienen las probabilidades de éxito y de fracaso:

$$p = \mathbb{P}(Y = 1) = \frac{\text{odds}(Y)}{1 + \text{odds}(Y)}$$

- $\text{odds}(Y) \in [0, \infty)$.
- $\text{odds}(Y) = 0 \iff \mathbb{P}(Y = 1) = 0$.
- $\text{odds}(Y) \to \infty \iff p \to 1$.

Si aplicamos el concepto de odds al modelo logístico, podemos dar un nuevo enfoque a la interpretación de los efectos de los coeficientes. Para $i = 1, ..., n$, $j = 1, ..., k$:

$$\text{odds}_i = \text{odds}(Y|\vec{X}_i) = \frac{p_i}{1 - p_i} = \exp(\vec{\beta}^t \vec{X}_i) = e^{\beta_0} e^{\beta_1 X_{i1}} \cdots e^{\beta_k X_{ij}}$$

- e^{β_0} es el odds de la respuesta cuando $X_{ij} = 0$, $\forall 1 \leq j \leq k$.
- $e^{\beta_j}, 1 \leq j \leq k$, es el **incremento multiplicativo** en el odds para un incremento unitario en X_j, manteniendo iguales los valores del resto de regresoras.

Odds-Ratio

Sea X una variable regresora binaria (que toma valores 0 o 1) y Y la variable respuesta también binaria. Se introduce el concepto de la **razón de odds** (odds-ratio), que indica en **qué medida** la ocurrencia de $Y = 1$ es más probable si $X = 1$ que si $X = 0$.

$$\text{OR} = \frac{\mathbb{P}(Y = 1|X = 1)/\mathbb{P}(Y = 0|X = 1)}{\mathbb{P}(Y = 1|X = 0)/\mathbb{P}(Y = 0|X = 0)}$$

$$= \frac{\left(\frac{e^{\beta_0+\beta_1}}{1 + e^{\beta_0+\beta_1}}\right) / \left(\frac{1}{1 + e^{\beta_0+\beta_1}}\right)}{\left(\frac{e^{\beta_0}}{1 + e^{\beta_0}}\right) / \left(\frac{1}{1 + e^{\beta_0}}\right)} = \frac{e^{\beta_0+\beta_1}}{e^{\beta_0}} = e^{\beta_1}$$

- Si OR $= 1$, X no incide en las odds.
- Si OR > 1, la probabilidad de $Y = 1$ frente a $Y = 0$ es OR-veces mayor cuando $X = 1$ que cuando $X = 0$.
- Si OR < 1, la probabilidad de $Y = 1$ frente a $Y = 0$ es OR-veces menor cuando $X = 1$ que cuando $X = 0$.

Si X es cualitativa con d atributos se crean $d - 1$ variables ficticias y $\text{OR}_j = e^{\beta_j}, j = 2, ..., d$ y compara el odd para el atributo j con el odd del atributo 1.

Transformación logit (log-odds)

Otra forma de reescribir el modelo logístico es empleando la **transformación logit** o **log-odds**, que conduce a expresar el **logaritmo de los odds** como función lineal de las variables regresoras, facilitando aún más su interpretación.

$$\eta_i = \text{logit}_i = \ln \frac{p_i}{1 - p_i} = \vec{\beta}^t \vec{X}_i = \beta_0 + \sum_{j=1}^{k} \beta_j X_{ij}$$

- $\eta_i \to -\infty \iff p_i \to 0$.
- $\eta_i \to \infty \iff p_i \to 1$.
- Si $p_i = 0{,}5 \implies \text{odds}_i = 1 \implies \eta_i = 0$.
- Si $p_i < 0{,}5 \implies \text{odds}_i < 1 \implies \eta_i < 0$.
- Si $p_i > 0{,}5 \implies \text{odds}_i > 1 \implies \eta_i > 0$.

Formulación del modelo logístico

Sea $\{(X_{11}, ..., X_{1k}, Y_1), ..., (X_{n1}, ..., X_{nk}, Y_n)\}$ una m.a.s. de n realizaciones independientes de $(X_1, ..., X_k, Y) = (\vec{X}^t, Y)$, siendo Y una variable binaria que toma valores 0 y 1.

Modelo de regresión logística

$$Y_i | (\vec{X} = \vec{X}_i) \sim \text{Be}(p_i), \text{ con } p_i = \frac{1}{1 + e^{-\vec{\beta}^t \vec{X}_i}}, \ i = 1, ..., n$$

De este modelo se derivan las siguientes hipótesis estructurales:

- **Linealidad** de los logits.

$$\eta_i = \ln \frac{p_i}{1 - p_i} = \vec{\beta}^t \vec{X} = \beta_0 + \sum_{j=1}^{k} \beta_j X_{ij} \iff \mathbb{E}(Y | \vec{X} = \vec{X}_i) = \frac{1}{1 + e^{-\vec{\beta}^t \vec{X}_i}}$$

- **Respuesta binaria** para todo $1 \neq i \neq n$.
- **Independencia** de las n observaciones muestrales.

Estimación de los parámetros del modelo logístic

Como $Y_i|(\vec{X} = \vec{X}_i) \sim \text{Be}(p_i)$ con $p_i = p(\vec{X}_i)$, para $i = 1, ..., n$, la función de log-verosimilitud del modelo toma la forma:

$$\begin{aligned} l(\beta) &= \sum_{i=1}^{n} \ln \left(p(\vec{X}_i)^{Y_i} (1 - p(\vec{X}_i))^{1-Y_i} \right) \\ &= \sum_{i=1}^{n} \left[Y_i \ln p(\vec{X}_i) + (1 - Y_i) \ln(1 - p(\vec{X}_i)) \right] \end{aligned}$$

3.3. Inferencia y bondad del ajuste

Inferencia con el estadístico de Wald

Con muestras grandes los estadísticos máximo-verosímiles $\widehat{\vec{\beta}}$ son asintóticamente normales, por tanto es viable emplear el estadístico de Wald.

Estadístico de contraste de Wald

$$\begin{aligned} W &= (\widehat{\vec{\beta}} - \vec{\beta})^t \text{Var}(\vec{\beta})^{-1} (\widehat{\vec{\beta}} - \vec{\beta}) \\ &= (\widehat{\vec{\beta}} - \vec{\beta})^t (\mathbf{X}\widehat{\mathbf{V}}\mathbf{X})^{-1} (\widehat{\vec{\beta}} - \vec{\beta}) \approx \chi^2_{k+1} \end{aligned}$$

donde $\widehat{\mathbf{V}}$ es una matriz diagonal $n \times n$ cuyo i-ésimo elemento es $\widehat{p}_i(1 - \widehat{p}_i)$.

Para chequear el contraste de hipótesis: $\begin{cases} H_0 : \beta_j = 0 \\ H_1 : \beta_j \neq 0 \end{cases}$, $j = 0, ..., k$. Se puede emplear:

$$W = \frac{\widehat{\beta}_j}{\widehat{\sigma}(\widehat{\beta}_j)} \approx N(0, 1) \Longrightarrow W_j^2 \approx \chi^2_1$$

- A menudo se sobreestima el error estándar de $\widehat{\beta}_j$ y las pruebas de Walk son conservadoras y los correspondientes intervalos de confianza erróneamente amplios.
- Con tamaños muestrales pequeños, la prueba basada en la **razón de verosimilitudes** es más faible que el test de Wald.
- Los intervalos de confianza basados en una variante de máxima-verosimilitud (profile-likelihood) son en general más precisos.

Razón de verosimilitudes (Deviance)

Estadístico del test de razón de verosimilitudes

$$D = -2\ln\left(\frac{\text{verosimilitud del modelo ajustado}}{\text{verosimilitud del modelo saturado}}\right) = 2\Big(\ln L(\tilde{\vec{\beta}}_S) - \ln L(\widehat{\vec{\beta}}_S)\Big)$$

donde el **modelo saturado** es un modelo con tantos parámetros como datos muestrales y su estimador se denota $\tilde{\vec{\beta}}_S$.

- La deviance indica en qué medida la verosimilitud del modelo saturado excede a la verosimilitud del modelo planteado.
- La deviance será baja para un modelo apropiado pero elevada en otro caso.
- Bajo H_0 (el modelo propuesto es el adecuado), asintóticamente $D \sim \chi^2_{n-k-1}$.

En regresión logística, como Y es binaria y $L(\tilde{\vec{\beta}}_S) = 1$:

$$-2\ln L(\widehat{\vec{\beta}}) = -2\sum_{i=1}^{n}\left[Y_i \ln \widehat{p}(\vec{X}_i) + (1-Y_i)\ln(1-\widehat{p}(\vec{X}_i))\right]$$

Comparación de dos modelos anidados

El cambio en la deviance se emplea en regresión logística para chequear si un modelo M2 es preferido a un modelo M1 más sencillo anidado en M2.

$$\begin{cases} H_0 : \eta_i = \text{logit}_i = \beta_{0,(1)} + \sum_{j=1}^{k_1}\beta_{j,(1)}X_{ij} & \text{M1} \\ H_1 : \eta_i = \text{logit}_i = \beta_{0.(2)} + \sum_{j=1}^{k_1}\beta_{j,(2)}X_{ij} + \sum_{j=k_1+1}^{k_2}\beta_{j,(2)}X_{ij} & \text{M2} \end{cases}$$

Equivalentemente;

$$\begin{cases} H_0 : \beta_j = 0, \ \ \forall k_1 < j \leq k_2 & \text{M1} \\ H_1 : \beta_j \neq 0, \ \ \forall k_1 < j \leq k_2 & \text{M2} \end{cases}$$

Si D_1 y D_2 denotan las deviance de los modelos ajustados para M1 y M2 respectivamente, se tiene que, bajo H_0:

$$D_1 - D_2 \sim \chi^2_{k_2-k_1}$$

Bajo H_0, esta diferencia de deviances debería ser pequeña. Se rechaza M1 en favor de M2 con un nivel α si $D_1 - D_2 > \chi^2_{k_2-k_1,1-\alpha}$.

En el caso particular:

$$\begin{cases} H_0 : \beta_j = 0 \\ H_1 : \beta_j \neq 0 \end{cases}$$

Rechazar H_0 a un nivel α si $D_{(-j)} - D > \chi^2 1, 1-\alpha$, siendo $D_{(-j)}$ la deviance del modelo sin la j-ésima variable y D la deviance del modelo completo.

3.4. Predicción del modelo logístico

Estimación de la media y la predicción condicionales

La estimación de la media condicional para un valor observado $\vec{X}_o$ se denota como $\widehat{p}_0$ y puede calcularse a partir del logit estimado:

$$\widehat{\eta} = \widehat{\vec{\beta}}^t \vec{X}_0$$

Estimación de la media condicionada: $\widehat{\mathbb{E}(Y|\vec{X}_0)} = \widehat{p}_0 = \dfrac{e^{\widehat{\eta}_0}}{1+e^{\widehat{\eta}_0}}$

La predicción individual de Y para un valor observador $\vec{X}_0$ se denota como $\widehat{Y}_0$ y puede calcularse como sigue:

$$\text{Sabiendo que: } \widehat{Y}|\vec{X}_0 = \begin{cases} 1 & \text{con probabilidad } \widehat{p}_0 = \dfrac{e^{\widehat{\eta}_0}}{1+e^{\widehat{\eta}_0}} \\ 0 & \text{con probabilidad } 1-\widehat{p}_0 \end{cases}$$

Predicción condicionada: $\widehat{Y}|\vec{X}_0 = 1$ si $\widehat{p}_0 > 1/2$ y 0 en otro caso.

Intervalo de confianza para la media condicional

Para n grande:

$$\frac{\widehat{\eta}_0 - \eta_0}{\widehat{\sigma}(\widehat{\eta}_0)} \approx N(0,1)$$

Por tanto, un intervalo de confianza para los logit estimados es:

$$\text{IC}_{1-\alpha}(\eta_0) = \left(\widehat{\eta}_0 - z_{1-\alpha/2}\widehat{\sigma}(\widehat{\eta}_0),\ \ \widehat{\eta}_0 + z_{1-\alpha/2}\widehat{\sigma}(\widehat{\eta}_0)\right)$$

Aplicando la transformada inversa de los logits obtenemos un intervalo de confianza para la media condicional.

$$\text{IC}_{100(1-\alpha)\,\%}(\widehat{p}_0) = \left(\frac{e^{\widehat{\eta}_0 - z_{1-\alpha/2}\widehat{\sigma}(\widehat{\eta}_0)}}{1+e^{\widehat{\eta}_0 - z_{1-\alpha/2}\widehat{\sigma}(\widehat{\eta}_0)}}\right) = \left(\frac{e^{\widehat{\eta}_0 + z_{1-\alpha/2}\widehat{\sigma}(\widehat{\eta}_0)}}{1+e^{\widehat{\eta}_0 + z_{1-\alpha/2}\widehat{\sigma}(\widehat{\eta}_0)}}\right)$$

3.5. Bondad del ajuste

χ^2 de Pearson y Deviance

- **Deviance nula**, D_0: Deviance de un modelo con sólo el intercept. Toma el mayor valor pues es el modelo más lejano al saturado. Generaliza SS(G) en RL.
- **Deviance residual**, D_R: Deviance del modelo ajustado. Generaliza la SS(R) en RL.

Supongamos el siguiente contraste de hipótesis:

$$\begin{cases} H_0 : \eta_i = \beta_0^* \\ H_1 : \eta_i = \beta_0 + \sum_{j=1}^{k} \beta_j X_{ij} \end{cases} \qquad \text{Bajo } H_0: \quad D_0 - D_R \approx \chi^2_{n-k}$$

Pseudo-R^2 o R^2 de McFadden

$$R^2 = 1 - \frac{D}{D_0}$$

- D_0 es la verosimilitud del modelo nulo.
- $D_0 - D$ puede verse como la verosimilitud explicada por el modelo respecto a la verosimilitud explicada por el modelo nulo.
- Por tanto $100R^2\,\%$ es el **porcentaje de verosimilitud** explicada con respecto a la del modelo nulo.
- Cuanto más grande (más próximo a 1), mejor ajuste.

Matriz de confusión

		Valor real	
		$Y_i = 0$	$Y_i = 1$
Predicción	$\widehat{Y}_i = 0$	Verdadero Negativo (VN)	Falso Negativo (FN)
	$\widehat{Y}_i = 1$	Falso Positivo (FP)	Verdadero Positivo (VP)

$$\text{Tasa de Clasificación Correcta (TCC)} = \frac{\text{VN} + \text{VP}}{n}$$

AIC, BIC y Prueba de Hosmer-Lemeshown

El test de Hosmer-Lemeshown realiza una partición de los datos en G grupos y examina mediante una prueba de bondad de ajuste χ^2 de Pearson si en esos grupos las proporciones observadas de valores $Y = 1$ e $Y = 0$ son semejantes a las proporciones predichas $\widehat{Y} = 1$ y $\widehat{Y} = 0$ respectivamente.

Si $\mathbb{P}(Y = 1)$ es alta, en un grupo es de esperar que la proporción de valores predichos iguales a 1 sea alta. Si el valor de la prueba supera el cuantil de la $\chi^2_{G-1,1-\alpha}$, entonces se concluye que el modelo no se ajusta a la realidad con una significación α.

3.6. Análisis de los residuos

- **Residuos convencionales** (`type='response'`)

$$e_i = Y_i - \widehat{Y}_i = Y_i - \widehat{p}_i$$

- **Residuos estandarizados o Pearson** (`type='pearson'`)

$$r_i = \frac{Y_i - \widehat{Y}_i}{\widehat{\sigma}(\widehat{Y}_i)} = \frac{Y_i \widehat{p}_i}{\sqrt{\widehat{p}_i(1-\widehat{p}_i)}}, \quad \text{asintóticamente } r_i \approx N(0,1)$$

- **Residuos deviante** (`type='deviance'`)

$$d_i = \begin{cases} \sqrt{-2\ln \widehat{p}_i} & \text{si } Y_i = 1 \\ \sqrt{-2\ln(1-\widehat{p}_i)} & \text{si } Y_i = 0 \end{cases}$$

Nótese que:

$$D = -2\ln L(\widehat{\vec{\beta}}) = -2\sum_{i=1}^{n}\Big[Y_i \ln \widehat{p}_i + (1-Y_i)\ln(1-\widehat{p}_i)\Big] = \sum_{i=1}^{n} d_i^2$$

Luego d_i^2 es la contribución de Y_i a la deviance del modelo ajustado. En este sentido son análogos a los residuos convencionales en regresión lineal.

3.7. Diagnosis del modelo

Chequeando linealidad

Hipótesis de linealidad de los logit: $\eta_i = \text{logit}_i = \ln \dfrac{p_i}{1-p_i} = \vec{\beta}^t \vec{X}_i = \beta_0 + \sum_{j=1}^{k} \beta_j X_{ij}$

Si el modelo es correcto, $D = \sum_{i=1}^{n} d_i^2 \chi^2_{n-k-1}$, de donde se sigue que los residuos deviance d_i deben ser aproximadamente $N(0,1)$ en muestras grandes.

Recuérdese que en regresión logística la hipótesis de linealidad se establece sobre los logit, y no sobre las respuestas Y_i. Gráficamente se debe examinar la nube de punto $\{(\widehat{\eta}_i, d_i), i = 1, ..., n\}$ y corroborar que se trata de una muestra aleatoria sin tendencia.

En otro caso, igual que en RLM, examinar la conducta de los residuos frente a cada regresora $\{(X_i, d_i)\}$ para determinar si la falta de linealidad es achacable a alguna de ellas.

Si falla la hipótesis de linealidad podría ser útil considerar alguna transformación de las regresoras problemáticas, añadir iteracciones...

A: Apéndice

A.1. Estimación por mínimos cuadrados ordinarios (OLS)

Tenemos por tanto, para $i = 1, ..., n$:

- Muestra: $\{(X_1, Y_1), ..., (X_n, Y_n)\}$
- Valores ajustados por la recta: $\widehat{Y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 X_i$
- Residuos: $e_i = Y_i - \widehat{Y}_i = Y_i - (\widehat{\beta}_0 + \widehat{\beta}_1 X_i)$.
- Suma de los residuos al cuadrado:

$$\text{SS(R)} = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (Y_i - \widehat{Y}_i)^2 = \sum_{i=1}^{n} (Y_i - \widehat{\beta}_0 - \widehat{\beta}_1 X_i)^2$$

Para minimizar la suma de los cuadrados de los residuos, aplicamos el método de las derivadas parciales para hallar el mínimo en la función SS(R) que depende de los parámetros $\widehat{\beta}_0$ y $\widehat{\beta}_1$.

1. Calculamos las derivadas parciales de SS(R) ($\text{SS(R)}_{\widehat{\beta}_0}$, $\text{SS(R)}_{\widehat{\beta}_1}$) y las igualamos a 0.

$$\begin{aligned} \text{SS(R)}_{\widehat{\beta}_0} &= \sum_{i=1}^{n} 2(Y_i - \widehat{\beta}_0 - \widehat{\beta}_1 X_i)(-1) = 0 \\ \text{SS(R)}_{\widehat{\beta}_1} &= \sum_{i=1}^{n} 2(Y_i - \widehat{\beta}_0 - \widehat{\beta}_1 X_i)(-X_i) = 0 \end{aligned}$$

   Hallando además las **ecuaciones normales de la regresión**:

$$\sum_{i=1}^{n} e_i = 0, \qquad \sum_{i=1}^{n} e_i X_i = 0$$

2. Desarrollamos para hallar el mínimo.

$$\begin{aligned} \text{SS(R)}_{\widehat{\beta}_0} &= \sum_{i=1}^{n} Y_i - n\widehat{\beta}_0 - \widehat{\beta}_1 X_i = 0 \\ \text{SS(R)}_{\widehat{\beta}_1} &= \sum_{i=1}^{n} X_i Y_i - \widehat{\beta}_0 \sum_{i=1}^{n} X_i - \widehat{\beta}_1 \sum_{i=1}^{n} X_i^2 = 0 \end{aligned}$$

3. Dividimos entre n, y, conociendo las siguientes definiciones:

$$\text{Medias muestrales : } \quad \overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i, \qquad \overline{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$$

$$\text{Varianza muestral : } \quad S_X^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \overline{X}^2$$

$$\text{Covarianza muestral : } \quad S_{XY} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y}) = \frac{1}{n}\sum_{i=1}^{n} X_i Y_i - \overline{XY}$$

Podemos completar la estimación deduciendo que:

$$\widehat{\beta}_0 = \overline{Y} - \widehat{\beta}_1 \overline{X}, \qquad \widehat{\beta}_1 = \frac{S_{XY}}{S_X^2}$$

A.2. Descomposición ortogonal del vector $\vec{Y}$

Dadas las premisas de la interpretación geométrical del modelo de RLM:

$$\begin{cases} \widehat{\vec{Y}}_{(1)} \in \text{col}(\mathbf{X}_{(1)}) \ \wedge \ \widehat{\vec{Y}}_{(2)} \in \text{col}(\mathbf{X}_{(2)}) \\ \vec{e}_{(1)} \perp \text{col}(\mathbf{X}_{(1)}) \ \wedge \ \vec{e}_{(2)} \perp \text{col}(\mathbf{X}_{(2)}) \\ \text{col}(\mathbf{X}_{(1)}) \subset \text{col}(\mathbf{X}_{(2)}) \Longrightarrow \vec{e}_{(2)} \perp \text{col}(\mathbf{X}_{(1)}) \end{cases}$$

$$\vec{Y} = \widehat{\vec{Y}}_{(1)} + \vec{e}_{(1)} = \widehat{\vec{Y}}_{(2)} + \vec{e}_{(2)}$$

Podemos deducir que:

1. Como $\vec{e}_{(2)} \perp \text{col}(\mathbf{X}_{(1)}) \wedge \widehat{\vec{Y}}_{(1)} \in \text{col}(\mathbf{X}_{(1)}) \implies \vec{e}_{(2)} \perp \widehat{\vec{Y}}_{(1)}$
2. Como $\vec{e}_{(1)} - \vec{e}_{(2)} = \widehat{\vec{Y}}_{(2)} - \widehat{\vec{Y}}_{(1)} \in \text{col}(\mathbf{X}_{(2)}) \implies \vec{e}_{(2)} \perp \vec{e}_{(1)} - \vec{e}_{(2)}$
3. Como $(\vec{e}_{(1)} - \vec{e}_{(2)})\widehat{\vec{Y}}_{(1)} = \vec{e}_{(1)}\widehat{\vec{Y}}_{(1)} - \vec{e}_{(2)}\widehat{\vec{Y}}_{(1)} = 0 \iff \widehat{\vec{Y}}_{(1)} \perp (\vec{e}_{(1)} - \vec{e}_{(2)})$