

Notas de Clase Estadística III 3009137

Capítulo 2: Repaso regresión lineal múltiple (MRLM)

Nelfi González Alvarez
Profesora Asociada

Departamento de Estadística
Universidad Nacional de Colombia, Sede Medellín



UNIVERSIDAD NACIONAL DE COLOMBIA

2025

Índice general

2. Repaso regresión lineal múltiple (MRLM)	30
2.1. Definición modelo de regresión lineal	30
2.2. Consideraciones	31
2.3. Pasos en el análisis de regresión	32
2.4. Regresión lineal en R	33
2.5. Representación matricial del MRLM y estimación por mínimos cuadrados ordinarios	34
2.6. Sumas de cuadrados, estimador de varianza y ANOVA del MRLM	35
2.7. Inferencias sobre los coeficientes de regresión	36
2.8. Respuesta media y valores futuros de la respuesta	37
2.9. El R^2 , el coeficiente de correlación múltiple y el R^2_{adj}	39
2.10. Evaluación de los supuestos en un MRLM	41
2.10.1. Supuesto de media cero	41
2.10.2. Supuesto de varianza constante	41
2.10.3. Supuesto de independencia	43
2.10.4. Supuesto de normalidad	44
2.10.5. Observaciones atípicas u outliers en la variable respuesta	44
2.11. Regresión lineal con variables indicadoras	45
2.11.1. Regresión lineal con un predictor cuantitativo y otro cualitativo	45
2.11.2. Ejemplo	48
2.11.3. Código R usado en el ejemplo de la Sección 2.11.2	53
Bibliografía	57

Índice de figuras

2.1.	Recta en el modelo de regresión lineal simple como la función de la media condicional de $Y x$. En este ejemplo se ha asumido que $Y x \sim N(62 + 3.5x, \sigma^2)$, luego, la recta de regresión corresponde a $E[Y x] = 62 + 3.5x$ y en cada nivel de x se tiene la misma varianza para Y alrededor de la respectiva media condicional.	32
2.2.	Diagrama de flujo proceso de modelación.	33
2.3.	Interpretación del R^2 : En los cuatro casos el modelo ajustado es $Y = \beta_0 + \beta_1 X + E$, con $E_i \stackrel{iid}{\sim} N(0, \sigma^2)$. (a) la verdadera relación estadística es lineal: $\mu_{Y x} = 650 + 2x$ y el ajuste arroja un R^2 cercano a 1; (b) La verdadera relación no es lineal: $\mu_{Y x} = 30000 + 2x + 0.2x^2$ aunque el ajuste arroja R^2 cercano a 1; (c) La verdadera relación no es lineal: $\mu_{Y x} = 92500 - 100x + 0.2x^2$ y el ajuste da un R^2 de casi cero, sin embargo, en este caso, no se puede decir que no existe relación estadística entre X y Y sino que la relación es no lineal; (d) El verdadero modelo es $Y_i = 5000 + E_i$, $E_i \stackrel{iid}{\sim} N(0, \sigma^2)$, es decir, no hay asociación estadística de Y con X , sin embargo, se ajustó el MRL asumiendo que $\mu_{Y x} = \beta_0 + \beta_1 x$, y su ajuste da un R^2 pequeño, como era de esperarse y la estimación del modelo $Y_i = \beta_0 + E_i$, $E_i \stackrel{iid}{\sim} N(0, \sigma^2)$ da $\hat{\beta}_0 = 4995.508$ muy próximo a la media verdadera de Y	40
2.4.	(a) Gráfico esperado de residuos vs. un predictor X_j cuando en el modelo no hay anomalías; (b) Gráfico esperado de residuos vs. \hat{y} cuando en el modelo no hay anomalías.	42
2.5.	Panel superior: Ejemplo del caso donde el modelo lineal entre Y y X_j no es adecuado, pero la varianza es constante, donde (a) residuos vs. X_j y (b) residuos vs. \hat{y} . Panel inferior: Ejemplo del caso donde el modelo lineal entre Y y X_j no es adecuado, ni la varianza es constante, donde (c) residuos vs. X_j y (d) residuos vs. \hat{y}	42
2.6.	Patrones donde el modelo de regresión no es carente de ajuste pero la varianza no es constante. Patrón de embudo: (a) residuos vs. X_j ; (b) residuos vs. \hat{y} . Patrón de balón de fútbol americano: (c) residuos vs. X_j ; (d) residuos vs. \hat{y} . Patrón de embudo no lineal: (e) residuos vs. X_j ; (f) residuos vs. \hat{y}	43
2.7.	Densidades de la distribución poblacional y patrones en gráficos de probabilidad normal sobre una muestra proveniente de tales distribuciones. (a) y (e) con una distribución normal de media cero; (b) y (f) con una distribución no normal y asimétrica a derecha; (c) y (g) con una distribución no normal asimétrica a izquierda; (d) y (h) con una distribución no normal, simétrica pero de colas pesadas.	45
2.8.	(a) Ilustración caso 1, con $c = 3$ y nivel de referencia el 3ro: La relación lineal de Y vs. X cambia con niveles de la variable cualitativa. (b) Ilustración caso 2, con $c = 3$ y nivel de referencia el 3ro: El efecto medio de X sobre Y no cambia con niveles de la variable cualitativa, pero la media de Y no es igual para todos los niveles de la variable cualitativa	47
2.9.	Gráfico de dispersión de las ventas Y vs. los gastos en publicidad X_1 , identificando la Sección.	49
2.10.	Captura consola R: Ajuste y tabla de parámetros estimados en el modelo 1.	49
2.11.	Captura consola R: Ajuste y tabla de parámetros estimados en el modelo 2.	50
2.12.	(a) Gráfico de dispersión de Y vs. X_1 con rectas ajustadas por Sección según modelo 1. (b) Gráfico de dispersión de Y vs. X_1 con rectas ajustadas por Sección según modelo 2	51
2.13.	Gráficos para análisis de supuestos en el modelo 2: (a) \hat{E} vs. \hat{Y} ; (b) \hat{E} vs. X_1 ; (c) \hat{E} vs. X_2 (Sección); (d) Gráfico de probabilidad normal con los residuos.	52
2.14.	Captura consola R: Test Shapiro Wilk y predicciones con el modelo 2	53
2.15.	Visualización del archivo DATOSPROBLEMAGASTOSPUBLICIDAD.csv	54

Índice de tablas

2.1. Fórmulas R en modelos de regresión lineal	34
2.2. Sumas de cuadrados en modelos de regresión lineal	35
2.3. Tabla ANOVA del modelo de regresión lineal múltiple	36
2.4. Prueba de hipótesis e intervalo de confianza (I.C) sobre los parámetros β_j	37
2.5. Inferencias sobre $\mu_{Y \mathbf{x}_0}$	38
2.6. Inferencias sobre la respuesta futura Y_0 en $\mathbf{x}_0 = (1, x_{01}, x_{0,2}, \dots, x_{0k})^T$	39
2.7. Modelo de Y vs X_1 , en cada nivel de X_2 , caso 1	46
2.8. Valor esperado de Y vs X_1 , en cada nivel de X_2 , caso 1, y diferencia de medias con respecto al nivel de referencia	46
2.9. Modelo de Y vs X_1 , en cada nivel de X_2 , caso 2.	47
2.10. Valor esperado de Y vs X_1 , en cada nivel de X_2 , caso 2, y diferencia de medias con respecto al nivel de referencia	47
2.11. Datos observados, ejemplo RLM con variables indicadoras	48
2.12. Tabla de parámetros estimados y ec. ajustada según modelo caso 1 tomando como referencia la Sección C	50
2.13. Modelos y ecuaciones ajustadas por Sección según modelo caso 1 tomando como referencia la Sección C	50
2.14. Tabla de parámetros estimados y ec. ajustada para modelo según caso 2, tomando como referencia la Sección C	50
2.15. Modelos y ecuaciones ajustadas por Sección según modelo caso 2 tomando como referencia la Sección C	50
2.16. Ecuación de pronósticos puntuales y resultados de predicción con el modelo 2 en los puntos $X_1 = 6$, Sección A y $X_1 = 11$, Sección C	53

Capítulo 2

Repaso regresión lineal múltiple (MRLM)

2.1. Definición modelo de regresión lineal

Un modelo de regresión es un medio formal para expresar dos aspectos importantes de una relación estadística (Kutner et. al., 2005):

1. Una tendencia de una variable dependiente Y que cambia cuando una o más variables independientes cambian en una forma sistemática.
2. Una dispersión de los puntos alrededor de una relación estadística.

Así, el análisis de regresión es un proceso a través del cual es derivada una relación estadística o predictiva entre una variable respuesta y un conjunto de variables predictoras o explicatorias, usando datos referentes a estas variables.

Considere el caso en el cual se desea modelar la variabilidad total de una variable respuesta de interés, en función de sus relaciones lineales con dos o más variables predictoras, formuladas simultáneamente en un único modelo. Suponemos en principio que las variables predictoras guardan poca asociación lineal entre sí, es decir, cada variable predictora aporta información independiente de las demás predictoras presentes en el modelo, lo que implica que hasta cierto grado, la información aportada por cada una no es redundante.

Para la especificación del modelo y de su ajuste, considere la siguiente nomenclatura. Sean,

- Y : Variable aleatoria denotando la respuesta o variable dependiente. y corresponde a una realización de esta variable.
- X_1, \dots, X_k : Llamadas variables predictoras, independientes, regresoras o covariables.
- E : Error aleatorio
- $\beta_j, j = 0, 1, \dots, k$: Parámetros de la regresión. β_0 es el intercepto de la función o superficie de regresión y β_j es el coeficiente de regresión asociado a la variable predictora X_j ; también es llamado coeficiente de regresión parcial.
- $\hat{\beta}_j$: Estimador del parámetro β_j
- \hat{E} : Variable aleatoria correspondiendo al residual del modelo ajustado. Su valor observado es denotado por \hat{e} .
- \hat{Y} : Respuesta estimada. Variable aleatoria que representa la estimación de $E[Y|\mathbf{x}]$ con $\mathbf{x} = [X_1, X_2, \dots, X_k]^T$. Su valor observado es denotado por \hat{y} .

La forma clásica del modelo de regresión lineal con k variables predictoras, sobre n observaciones, es como sigue

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \dots + \beta_k X_{ik} + E_i, \quad E_i \stackrel{iid}{\sim} N(0, \sigma^2), \quad (2.1)$$

con $i = 1, \dots, n$, denotando la i -ésima unidad de observación en una muestra de tamaño n , de modo que X_{ij} corresponde al valor de la variable X_j en la i -ésima observación y Y_i el valor de la correspondiente respuesta. Además, $\stackrel{iid}{\sim}$ denota que los errores E_1, E_2, \dots, E_n son independientes y se distribuyen con idéntica distribución.

2.2. Consideraciones

Estadísticamente (2.1) establece que, bajo los supuestos de normalidad, independencia y varianza constante de los errores (Kutner et. al., 2005),

1. La variable respuesta es una variable aleatoria cuyos valores se observan mediante la selección de los valores de las variables predictoras en un espacio de valores de interés. Se supone que cada valor de Y está constituido por un valor real y una componente aleatoria
2. Por lo anterior, las variables predictoras no son consideradas como variables aleatorias, sino como un conjunto de valores fijos que representan los puntos de observación, que se seleccionan con anticipación y se miden sin error. Sin embargo si esto último no se cumple, el método de mínimos cuadrados ordinarios puede seguir siendo válido si los errores en los valores de las variables X_j son pequeños en comparación con los errores aleatorios.
3. Los datos que se observan constituyen una muestra representativa de un medio acerca del cual se desea generalizar. Si no es así, no es apropiado realizar inferencias (extrapolaciones) en un rango de los datos por fuera del considerado.
4. El modelo de regresión es lineal en los parámetros. Es decir, ningún parámetro de la regresión aparece como el exponente o es dividido o multiplicado por otro parámetro, o transformado por alguna función matemática. Sin embargo, la superficie de ajuste puede tener una curvatura (no ser lineal en alguna X_j y/o en Y), caso en el cual mediante una transformación conveniente de las variables (X_j y/o Y), es posible aplicar las técnicas de regresión lineal sobre estas nuevas variables.
5. Si la ecuación de regresión seleccionada es correcta, cualquier variabilidad en la variable respuesta que no puede ser explicada exactamente por dicha ecuación, es debida a un error aleatorio.
6. Los errores aleatorios $E_i \sim N(0, \sigma^2), i = 1, 2, \dots, n$ y además, son estadísticamente independientes, por tanto:

$$\text{Cov}(E_i, E_j) = 0, \forall i, j, i \neq j \quad (2.2)$$

$$\text{Cov}(Y_i, Y_j) = 0, \forall i, j, i \neq j \quad (2.3)$$

En consecuencia, los valores de la variable respuesta también son estadísticamente independientes y por tanto no correlacionados.

7. La varianza de los errores aleatorios es $\sigma^2, \forall i, i = 1, 2, \dots, n$ pero desconocida. Dado que se asume que las variables predictoras no son variables aleatorias, la varianza de los Y_i también es $\sigma^2, \forall i$ y por tanto no depende del punto de observación, es decir, del valor de las X_j . Pero en el caso que esta última suposición no pueda aplicarse, entonces el método de regresión empleado será el de mínimos cuadrados ponderados.

Con estas consideraciones, podemos afirmar que la distribución de la respuesta dadas los predictoras es:

$$Y|X_{i1}, X_{i2}, \dots, X_{ik} \stackrel{\text{ind}}{\sim} N(\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \dots + \beta_k X_{ik}, \sigma^2), \quad (2.4)$$

donde la respuesta media está dada por

$$E[Y|X_1, X_2, \dots, X_k] = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_k X_k, \quad (2.5)$$

la cual describe a un hiperplano en el espacio de $k + 1$ dimensiones llamado *superficie de regresión o superficie de respuesta*.

Nota 2.1. Si el número de predictoras es solo $k = 1$, el modelo en (2.1) es conocido como el modelo de regresión lineal simple (MRLS),

$$Y_i = \beta_0 + \beta_1 X_{i1} + E_i, \quad E_i \stackrel{\text{iid}}{\sim} N(0, \sigma^2), \quad (2.6)$$

este modelo supone una relación de tendencia lineal entre la variable respuesta y el predictor lineal, en donde las medias de la distribución de $Y|X_1$ caen sobre la recta de regresión, como ilustra la Figura 2.1.

Nota 2.2. Recuerde que el término *modelo lineal* significa que el modelo es una función lineal en los parámetros $\beta_j, j = 0, 1, \dots, k$, lo cual no hace referencia a la forma de la superficie de respuesta.

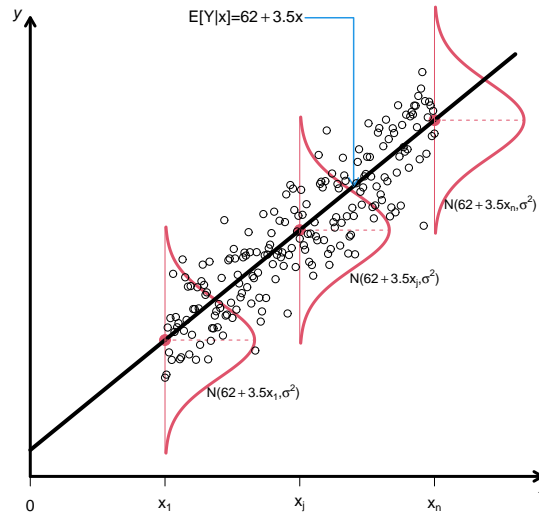


Figura 2.1: Recta en el modelo de regresión lineal simple como la función de la media condicional de $Y|x$. En este ejemplo se ha asumido que $Y|x \sim N(62 + 3.5x, \sigma^2)$, luego, la recta de regresión corresponde a $E[Y|x] = 62 + 3.5x$ y en cada nivel de x se tiene la misma varianza para Y alrededor de la respectiva media condicional.

2.3. Pasos en el análisis de regresión

De acuerdo a Kutner et. al., 2005, el trabajo de análisis de regresión requiere que el analista sea hábil en el uso de las herramientas computacionales para estimar modelos, desarrollar interpretaciones de los resultados, analizar residuales, y permitir la mejora del modelo a través de experimentaciones que incluyan observaciones futuras y traducir sus resultados en términos comprensibles para el usuario final del modelo. Así, el análisis de regresión contempla una serie de tareas que pueden resumirse en las siguientes, las cuales son esquematizadas en la Figura 2.2:

1. Comprensión del problema. Es el paso más importante del análisis de regresión. Consiste en la formulación de las preguntas que se desean resolver a través de la modelación e implica la identificación de las variables relevantes y la justificación del método de análisis estadístico.
2. Con las variables potenciales identificadas en la formulación del problema, realizar análisis exploratorio de los datos mediante diagramas de dispersión para establecer el tipo de función de regresión apropiada.
3. Aplicar transformaciones para estabilizar varianza o para simplificar la modelación.
4. Desarrollar uno o más modelos de regresión tentativos.
5. Ajustar los modelos tentativos.
6. Evaluar los modelos ajustados
 - a) Analizar de residuales para:
 - Verificar si el modelo es adecuado: Gráfico de residuos vs. x para chequear ausencia de patrones sistemáticos, test de carencia de ajuste.
 - Verificar si los supuestos sobre el término de error se cumplen: Gráficos y test de probabilidad normal; gráficos de residuos vs. valores ajustados de la respuesta para chequear varianza constante y ausencia de patrones sistemáticos.
 - b) Diagnóstico de observaciones atípicas e influyentes.
7. Para los modelos que pasen las pruebas en 6:
 - Evaluar calidad del ajuste.

- Hacer predicciones: Solo dentro del espacio de valores considerados para las variable predictoras o valores cercanos a éste y evaluar la calidad de pronósticos.
8. Escoger el mejor modelo: Entre modelos estadísticamente válidos con un buen ajuste y y útiles para la predicción o para los propósitos para los cuales se formuló la modelación.
 9. Interpretar los parámetros del modelo seleccionado a la luz de los datos.
 10. Hacer las inferencias de interés.
 11. Reportar resultados.

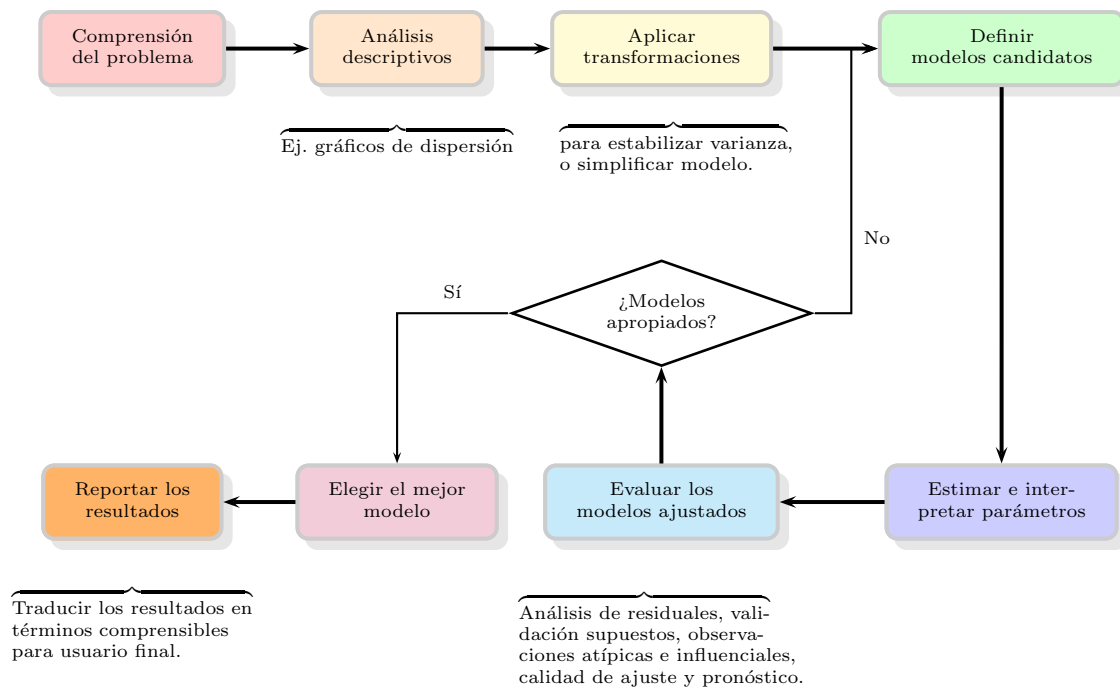


Figura 2.2: Diagrama de flujo proceso de modelación.

2.4. Regresión lineal en R

En R la función disponible para regresión lineal múltiple clásica es la función `lm(...)` en la cual se formula el modelo usando la sintaxis de fórmulas admisibles. La Tabla 2.1 ejemplifica la sintaxis básica para la formulación y ajuste de algunos modelos de regresión lineal mediante la función R `lm()`. Otras funciones importantes en este curso son las siguientes (para detalles sobre cada una de estas funciones consulte la ayuda R correspondiente):

- `summary()`: Sobre un objeto clase `lm` obtiene la tabla de parámetros estimados.
- `coef()`: Sobre un objeto clase `lm` obtiene el vector con los parámetros estimados.
- `confint()`: Sobre un objeto clase `lm` obtiene una matriz en la cual en cada fila se dan los límites de los intervalos de confianza para los parámetros del modelo, por defecto, del 95 % de confianza.
- `residuals()`: Sobre objeto clase `lm` obtiene los valores de los residuos de ajuste.
- `fitted()`: Sobre un objeto clase `lm` obtiene los valores ajustados en la variable respuesta.
- `predict()`: Sobre un objeto clase `lm` obtiene las predicciones puntuales y por intervalos de predicción. También permite calcular valores ajustados y los intervalos de confianza para $\mu_{Y|x}$.
- `qqnorm()`, `qqline()`: Para gráfico de probabilidad normal.
- `shapiro.test()`: Para test de normalidad Shapiro-Wilk.

Tabla 2.1: Fórmulas R en modelos de regresión lineal

Fórmula R	Modelo a ajustar	Corrida con <code>lm()</code>
$Y \sim X$, o bien, $Y \sim 1 + X$	$Y_i = \beta_0 + \beta_1 X_i + E_i$ $E_i \stackrel{\text{iid}}{\sim} N(0, \sigma^2)$	<code>lm(Y ~ X)</code> , o bien, <code>lm(Y ~ 1 + X)</code>
$Y \sim -1 + X$, o bien, $Y \sim 0 + X$	$Y_i = \beta_1 X_i + E_i$ $E_i \stackrel{\text{iid}}{\sim} N(0, \sigma^2)$	<code>lm(Y ~ -1 + X)</code> , o bien, <code>lm(Y ~ 0 + X)</code>
$\log(Y) \sim X$	$\log(Y_i) = \beta_0 + \beta_1 X_i + E_i$ $E_i \stackrel{\text{iid}}{\sim} N(0, \sigma^2)$	<code>lm(log(Y) ~ X)</code>
$Y \sim X1 + X2 + X3 + X4$	$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \beta_4 X_{i4} + E_i$ $E_i \stackrel{\text{iid}}{\sim} N(0, \sigma^2)$	<code>lm(Y ~ X1 + X2 + X3 + X4)</code>
$Y \sim X + I(X^2)$	$Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + E_i$ $E_i \stackrel{\text{iid}}{\sim} N(0, \sigma^2)$	<code>lm(Y ~ X + I(X^2))</code>
$Y \sim X1 * X2$	$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i1} * X_{i2} + E_i$ $E_i \stackrel{\text{iid}}{\sim} N(0, \sigma^2)$	<code>lm(Y ~ X1 * X2)</code>

2.5. Representación matricial del MRLM y estimación por mínimos cuadrados ordinarios (MCO)

El uso del álgebra matricial es la clave para el procedimiento de estimación por mínimos cuadrados. Para ello, note que cuando se tienen observaciones $(x_{i1}, x_{i2}, \dots, x_{ik}, y_i)$, $i = 1, 2, \dots, n$, para el modelo lineal dado en (2.1), en realidad se obtiene un sistema de n ecuaciones con $k+1$ incógnitas correspondiendo al intercepto β_0 y los coeficientes de regresión parcial β_j , $j = 1, \dots, k$, donde las variables Y y las X_j toman los valores observados y_i y x_{ij} , respectivamente:

$$\begin{aligned}
 y_1 &= \beta_0 + \beta_1 \cdot x_{11} + \beta_2 \cdot x_{12} + \dots + \beta_k \cdot x_{1k} + E_1 \\
 y_2 &= \beta_0 + \beta_1 \cdot x_{21} + \beta_2 \cdot x_{22} + \dots + \beta_k \cdot x_{2k} + E_2 \\
 &\vdots \\
 y_n &= \beta_0 + \beta_1 \cdot x_{n1} + \beta_2 \cdot x_{n2} + \dots + \beta_k \cdot x_{nk} + E_n
 \end{aligned} \tag{2.7}$$

Tal sistema expresado en forma matricial corresponde a: $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{E}$, donde: el vector de parámetros, de respuestas y de error son, respectivamente,

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{bmatrix}, \quad \mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \quad \mathbf{E} = \begin{bmatrix} E_1 \\ E_2 \\ \vdots \\ E_n \end{bmatrix}. \tag{2.8}$$

La matriz \mathbf{X} llamada matriz de diseño, con los valores de las variables predictoras en cada observación corresponde a:

$$\mathbf{X} = \begin{bmatrix} 1 & x_{11} & x_{12} & \dots & x_{1k} \\ 1 & x_{21} & x_{22} & \dots & x_{2k} \\ \vdots & \vdots & \vdots & \dots & \vdots \\ 1 & x_{n1} & x_{n2} & \dots & x_{nk} \end{bmatrix} \tag{2.9}$$

Observe que en \mathbf{X} la primera columna es un vector de 1's desde que en cada ecuación el intercepto siempre tiene como coeficiente a 1, en tanto que las demás columnas corresponden a los vectores de valores observados de las variables predictoras X_j , $j = 1, 2, \dots, k$, vectores colocados en el orden en que las variables entran en la ecuación del modelo.

En la estimación del modelo por el método de mínimos cuadrados ordinarios (MCO), el objetivo es obtener estimaciones de los parámetros de regresión hallando el vector $\boldsymbol{\beta} \in \mathbb{R}^{k+1}$, que minimice la suma de los cuadrados de los errores,

$$S(\boldsymbol{\beta}) = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) = \sum_{i=1}^n [y_i - (\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_k x_{ik})]^2 = \sum_{i=1}^n E_i^2. \tag{2.10}$$

Denotaremos por $\hat{\beta}$ al vector de parámetros estimados, es decir,

$$\hat{\beta} = \arg \min_{\beta} S(\beta). \quad (2.11)$$

Para la solución de mínimos cuadrados calculamos el gradiente $\partial S(\beta) / \partial \beta$, y lo igualamos al vector de ceros $\mathbf{0}_{k+1}$, obteniendo el sistema de **ecuaciones normales de mínimos cuadrados** para el modelo lineal general, que se expresa en forma matricial como,

$$(\mathbf{X}^T \mathbf{X}) \beta = \mathbf{X}^T \mathbf{y} \quad (2.12)$$

El vector de parámetros estimados es la solución del sistema de ecuaciones normales y es igual a,

$$\hat{\beta} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y} = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \hat{\beta}_2 \\ \vdots \\ \hat{\beta}_k \end{bmatrix} \quad (2.13)$$

desde que $\mathbf{X}^T \mathbf{X}$ sea invertible. El valor observado de la respuesta ajustada en la i -ésima observación es igual a

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + \cdots + \hat{\beta}_k x_{ik}, \quad (2.14)$$

corresponde a una realización de la variable aleatoria \hat{Y}_i , en tanto que la diferencia $\hat{\varepsilon}_i = y_i - \hat{y}_i$, es el **residuo del ajuste observado** en la i -ésima observación, el cual es una realización de la variable aleatoria

$$\hat{E}_i = Y_i - \hat{Y}_i, \quad (2.15)$$

de modo que en una muestra aleatoria de tamaño n , tenemos el vector aleatorio de residuos, $\hat{\mathbf{E}} = (\hat{E}_1, \dots, \hat{E}_n)^T$.

2.6. Sumas de cuadrados, estimador de varianza y ANOVA del MRLM

En la regresión lineal son definidas varios tipos de sumas, entre ellas las denominadas sumas de cuadrados, las cuales permiten construir estadísticos de interés en varias de las inferencias que son relevantes en el modelo estimado. Estas sumas son presentadas en la Tabla 2.2.

Tabla 2.2: Sumas de cuadrados en modelos de regresión lineal

Suma	Ecuación	Grados de libertad	Interpretación
SST	$\sum_{i=1}^n (y_i - \bar{y})^2$	$n - 1$	Suma de cuadrados totales: La variabilidad total observada en la variable respuesta
SSE	$\sum_{i=1}^n (y_i - \hat{y}_i)^2$	$n - k - 1$	Suma de cuadrados de residuos: la variabilidad de la respuesta que no es explicada por el modelo de regresión
SSR	$\sum_{i=1}^n (\hat{y}_i - \bar{y})^2$	k	Suma de cuadrados de regresión: la variabilidad de la respuesta que es explicada por el modelo de regresión
Con $\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i$ (promedio muestral de la respuesta) y \hat{y}_i la respuesta estimada según (2.14).			

Bajo los supuestos del modelo de regresión, puede demostrarse que

$$E[SSE] = (n - k - 1) \sigma^2, \quad (2.16)$$

de modo que un estimador insesgado para σ^2 corresponde a

$$MSE = \frac{SSE}{n - k - 1}. \quad (2.17)$$

El MSE es conocido como la suma de cuadrados medios de residuos y como vemos corresponde a la suma de cuadrados de residuos dividida por sus grados de libertad.

El análisis de varianza o ANOVA, consiste en la descomposición de la variabilidad total observada en la variable respuesta, es decir de la SST, en la suma de componentes o fuentes de variabilidad, de acuerdo al modelo propuesto.

Recuerde que el modelo de regresión lineal plantea que la respuesta es igual a la suma de una componente real no aleatoria (la función de regresión) y un error aleatorio E . Se espera que la superficie ajustada explique en forma significativa la variabilidad observada en Y . Bajo los supuestos del modelo, la variabilidad total muestral de la respuesta satisface la siguiente descomposición,

$$\underbrace{\sum_{i=1}^n (Y_i - \bar{Y})^2}_{\text{Variabilidad total (SST)}} = \underbrace{\sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2}_{\text{Variabilidad explicada (SSR)}} + \underbrace{\sum_{i=1}^n (Y_i - \hat{Y}_i)^2}_{\text{Variabilidad no explicada (SSE)}} \quad (2.18)$$

En virtud de la anterior igualdad, podemos también establecer la siguiente identidad para los grados de libertad (g.l) de las sumas de cuadrados:

$$\underbrace{\text{g.l (SST)}}_{n-1} = \underbrace{\text{g.l (SSR)}}_k + \underbrace{\text{g.l (SSE)}}_{n-k-1}. \quad (2.19)$$

Tabla 2.3: Tabla ANOVA del modelo de regresión lineal múltiple

Fuente	Suma de cuadrados	g.l	Cuadrados medios	Cuadrados medios esperados	Estadístico F_0	Valor P
Regresión	SSR	k	$\text{MSR} = \text{SSR}/k$	$E[\text{MSR}] = \sigma^2 + \frac{Q_R^2}{k}$	MSR/MSE	$P(f_{k,n-k-1} > F_0)$
Error	SSE	$n - k - 1$	$\text{MSE} = \text{SSE}/(n - k - 1)$	$E[\text{MSE}] = \sigma^2$		
Total	SST	$n - 1$	$\text{MST} = \text{SST}/(n - 1)$			
Donde $Q_R^2 = \beta_R^T (\mathbf{X}_c^T \mathbf{X}_c) \beta_R$, con $\beta_R = (\beta_1, \beta_2, \dots, \beta_k)^T$, es decir, el vector de parámetros sin intercepto, \mathbf{X}_c es la matriz de dimensión $n \times k$, de valores centrados de las X_j , es decir, su componente i, j corresponde a $x_{ij} - \bar{x}_j$, siendo \bar{x}_j el promedio muestral de la variable X_j . Recuerde que: k es el número de predictores en el modelo.						

Antes de proceder con las inferencias en cada parámetro de la regresión o incluso sobre la respuesta media, evaluamos la significancia del MRLM: *si la porción de la variabilidad observada en la respuesta que es explicada por el MRLM es significativa*. Esto es realizado mediante el test ANOVA, el cual propone el siguiente juego de hipótesis:

$$\begin{aligned} H_0 : \beta_1 = \beta_2 = \dots = \beta_k = 0; \\ H_1 : \beta_j \neq 0, \text{ para al menos un } j, j = 1, 2, \dots, k, \end{aligned} \quad (2.20)$$

de modo que bajo H_0 el MRLM se reduce a $Y_i = \beta_0 + E_i$, con $E_i \stackrel{\text{iid}}{\sim} N(0, \sigma^2)$, este es el modelo nulo o reducido, indicando que ninguno de los k predictores contribuye a explicar la variabilidad en la respuesta, en tanto que bajo H_1 el modelo es el modelo completo con los k predictores y bajo esta hipótesis, el modelo es estadísticamente significativo porque por lo menos uno de los k predictores contribuye a explicar significativamente la variabilidad en la respuesta (pero tenga cuidado, esto no implica que el modelo sea correcto en su especificación ni que con él se logren buenas predicciones). El test ANOVA suele presentarse a través de la tabla denominada tabla ANOVA, la cual se muestra en la Tabla 2.3. Bajo H_0 en (2.20) y los supuestos del modelo de regresión, el estadístico de la prueba es

$$F_0 = \frac{\text{MSR}}{\text{MSE}} \sim f_{k,n-k-1}. \quad (2.21)$$

con $\text{MSR} = \text{SSR}/k$, el *cuadrado medio de regresión*. F_0 es una razón que compara los estimadores de los valores esperados $E[\text{MSR}]$, el cual es estimado por el MSR, y $E[\text{MSE}]$, el cual es estimado por el MSE. Estos valores esperados valen σ^2 bajo la no significancia del MRLM, pero bajo H_1 , $E[\text{MSR}]$ es como muestra la Tabla 2.3 y resulta mayor que σ^2 , en tanto que F_0 tomaría valores estadísticamente más grandes de los esperados bajo H_0 . Por tanto, H_0 en el test ANOVA es rechazado para valores de F_0 estadísticamente grandes bajo la distribución $f_{k,n-k-1}$, lo cual, en términos probabilísticos es equivalente a que $P(f_{k,n-k-1} > F_0)$ sea pequeño.

2.7. Inferencias sobre los coeficientes de regresión

Cuando el modelo es estadísticamente significativo, es de interés determinar la significancia individual de los predictores incluidos, que en el modelo lineal dado en (2.1), corresponde a probar la significancia del parámetro β_j asociado al predictor X_j . Esto es realizado bajo la premisa de independencia entre los predictores y que es posible variar marginalmente el valor de cada uno, en tanto que las demás variables pueden mantenerse fijas. De otro lado, un test de significancia sobre el intercepto β_0 solo se recomienda cuando este parámetro sea interpretable, esto

es, cuando la coordenada $\mathbf{x} = (0, 0, \dots, 0)^T$ haya sido observada en el conjunto de datos usados en el ajuste del modelo.

El test de significancia individual del parámetro β_j consiste en probar

$$H_0 : \beta_j = 0 \text{ vs. } H_1 : \beta_j \neq 0. \quad (2.22)$$

Puede mostrarse que bajo los supuestos del MRLM, las distribuciones marginales para los parámetros estimados corresponden a,

$$\hat{\beta}_j \sim N(\beta_j, \sigma^2 C_{jj}), \quad j = 0, 1, \dots, k. \quad (2.23)$$

donde las cantidades C_{jj} son positivas y corresponden a los elementos en la diagonal principal de la matriz $(\mathbf{X}^T \mathbf{X})^{-1}$. Conociendo estas distribuciones ahora podemos realizar para cada β_j su respectivo test de significancia enunciado en (2.22). Teniendo en cuenta además que σ^2 es desconocido, el estadístico de prueba con su distribución bajo H_0 y el criterio de rechazo del test son como sigue:

$$\text{Estadístico de prueba: } T_0 = \frac{\hat{\beta}_j}{\sqrt{\text{MSE } C_{jj}}} \stackrel{H_0}{\sim} t_{n-k-1}; \quad (2.24)$$

$$\text{Rechazo de } H_0 \text{ con valor P: si } P(|t_{n-k-1}| > |T_0|) \text{ es pequeño;} \quad (2.25)$$

$$\text{Rechazo de } H_0 \text{ con región crítica a un nivel de significancia } \alpha: \text{ si } |T_0| > t_{\alpha/2, n-k-1}. \quad (2.26)$$

Nota 2.3. A la raíz cuadrada de la varianza estimada del estimador $\hat{\beta}_j$ se le denomina el error estándar de ese estimador denotado por $\text{s.e}(\hat{\beta}_j)$, de modo que

$$\text{s.e}(\hat{\beta}_j) = \sqrt{\text{MSE } C_{jj}} \quad (2.27)$$

La Tabla 2.4 resume otras pruebas de hipótesis que pueden realizarse sobre cada uno de los parámetros e indica además, cómo se construyen intervalos de nivel de confianza $(1 - \alpha)\%100$ para cada uno de los parámetros.

Tabla 2.4: Prueba de hipótesis e intervalo de confianza (I.C) sobre los parámetros β_j .

Test bilateral	Estadístico de prueba	Criterio de rechazo	I.C del $(1 - \alpha)\%100$
$H_0 : \beta_j = \beta_{j0}$ $H_1 : \beta_j \neq \beta_{j0}$	$T_0 = \frac{\hat{\beta}_j - \beta_{j0}}{\sqrt{\text{MSE } C_{jj}}} \sim t_{n-k-1}$	con nivel α : si $ T_0 > t_{\alpha/2, n-k-1}$ con valor P, si: $P(t_{n-k-1} > T_0)$ es pequeño.	$\hat{\beta}_j \pm t_{\alpha/2, n-k-1} \times \sqrt{\text{MSE } C_{jj}}$
con $\beta_{j0} = 0$ en el test de significancia.			
Tests unilaterales	Estadístico de prueba	Criterio de rechazo	
$H_0 : \beta_j = \beta_{j0}$ $H_1 : \beta_j > \beta_{j0}$	$T_0 = \frac{\hat{\beta}_j - \beta_{j0}}{\sqrt{\text{MSE } C_{jj}}} \sim t_{n-k-1}$	con nivel α : si $T_0 > t_{\alpha, n-k-1}$ con valor P, si: $P(t_{n-k-1} > T_0)$ es pequeño.	
$H_0 : \beta_j = \beta_{j0}$ $H_1 : \beta_j < \beta_{j0}$	$T_0 = \frac{\hat{\beta}_j - \beta_{j0}}{\sqrt{\text{MSE } C_{jj}}} \sim t_{n-k-1}$	con nivel α : si $T_0 < -t_{\alpha, n-k-1}$ con valor P, si: $P(t_{n-k-1} < T_0)$ es pequeño.	

2.8. Respuesta media y valores futuros de la respuesta

Para el MRLM en (2.1), considere la respuesta media en el punto $\mathbf{x}_0 = (1, x_{01}, x_{02}, \dots, x_{0k})^T$ ¹, es decir

$$\mu_{Y|\mathbf{x}_0} = \mathbf{x}_0^T \boldsymbol{\beta} = \beta_0 + \beta_1 x_{01} + \beta_2 x_{02} + \dots + \beta_k x_{0k}. \quad (2.28)$$

¹El número 1 es incluido en el vector de predictores para poder escribir la respuesta media mediante el producto punto entre el vector de valores de los predictores y el vector de parámetros. También puede pensarse como el valor del predictor asociado al intercepto en la matriz de diseño.

La respuesta futura en el mismo punto, bajo el MRLM, es la variable aleatoria Y_0 tal que

$$Y_0 = \mathbf{x}_0^T \boldsymbol{\beta} + E_0 = \beta_0 + \beta_1 x_{01} + \beta_2 x_{02} + \cdots + \beta_k x_{0k} + E_0, \text{ con } E_0 \sim N(0, \sigma^2), \quad (2.29)$$

de donde

$$Y_0 \sim N(\mathbf{x}_0^T \boldsymbol{\beta}, \sigma^2). \quad (2.30)$$

Tanto la estimación de la respuesta media como la predicción de la respuesta en el punto \mathbf{x}_0 , son obtenidas mediante el MRLM ajustado con una muestra de tamaño n que no incluye el valor de la respuesta futura,

$$\text{Respuesta media estimada en } \mathbf{x}_0: \hat{\mu}_{Y|\mathbf{x}_0} = \hat{\beta}_0 + \hat{\beta}_1 x_{01} + \hat{\beta}_2 x_{02} + \cdots + \hat{\beta}_k x_{0k} \quad (2.31)$$

$$\text{Predicción respuesta futura en } \mathbf{x}_0: \hat{Y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_{01} + \hat{\beta}_2 x_{02} + \cdots + \hat{\beta}_k x_{0k}, \quad (2.32)$$

es decir, $\hat{\mu}_{Y|\mathbf{x}_0} = \hat{Y}_0$, sin embargo, cuando consideramos a (2.31) como la estimación de $\mu_{Y|\mathbf{x}_0}$, estamos apuntando a la estimación del valor en el cual se espera que la distribución de la respuesta tienda a centrarse en repetidos ensayos aleatorios en los que $\mathbf{x} = (1, X_1, X_2, \dots, X_k)^T$ se ha fijado en \mathbf{x}_0 . Por su parte, cuando consideramos a (2.32) como la predicción de una nueva observación de la respuesta cuando $\mathbf{x} = \mathbf{x}_0$, apuntamos al valor que pudiera ocurrir para $Y|\mathbf{x}_0$ en un nuevo ensayo, independiente de aquellos ensayos con los cuales se realizó el análisis de regresión. Por supuesto se asume que el modelo de regresión subyacente aplicable para los datos muestrales con los cuales se estimaron los parámetros continúa siendo apropiado para la nueva observación (Kutner et. al., 2005). Para las inferencias se parte del siguiente resultado, el cual es obtenido bajo los supuestos del MRLM:

$$\hat{\mu}_{Y|\mathbf{x}_0} \sim N\left(\mathbf{x}_0^T \boldsymbol{\beta}, \sigma^2 \mathbf{x}_0^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}_0\right). \quad (2.33)$$

Sobre la respuesta media podemos hacer inferencias bien sea mediante intervalos de confianza o bien mediante pruebas de hipótesis. La Tabla 2.5 resume estas inferencias.

Tabla 2.5: Inferencias sobre $\mu_{Y|\mathbf{x}_0}$.

Pruebas de hipótesis	
Hipótesis H_0	Estadístico de prueba
$H_0 : \mu_{Y \mathbf{x}_0} = c$	$T_0 = \frac{\mathbf{x}_0^T \boldsymbol{\beta} - c}{\sqrt{\text{MSE } \mathbf{x}_0^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}_0}} \sim t_{n-k-1}$
Hipótesis H_1	Criterio de rechazo de H_0
$H_1 : \mu_{Y \mathbf{x}_0} \neq c$	con nivel α : si $ T_0 > t_{\alpha/2, n-k-1}$ con valor P: si $P(t_{n-k-1} > T_0)$ es pequeño.
$H_1 : \mu_{Y \mathbf{x}_0} > c$	con nivel α : si $T_0 > t_{\alpha, n-k-1}$ con valor P: si $P(t_{n-k-1} > T_0)$ es pequeño.
$H_1 : \mu_{Y \mathbf{x}_0} < c$	con nivel α : si $T_0 < -t_{\alpha, n-k-1}$ con valor P: si $P(t_{n-k-1} < T_0)$ es pequeño.
Intervalo de confianza de nivel $(1 - \alpha)100\%$ en el punto \mathbf{x}_0 :	
$\mathbf{x}_0^T \hat{\boldsymbol{\beta}} \pm t_{\alpha/2, n-k-1} \sqrt{\text{MSE } \mathbf{x}_0^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}_0}$	

Por su parte, sobre la respuesta futura, la inferencia de interés en este curso es la predicción puntual y por intervalo. Recuerde que la respuesta futura y_0 en el punto \mathbf{x}_0 es el valor puntual que pudiera tomar la variable aleatoria $Y_0 = Y|(\mathbf{x} = \mathbf{x}_0)$, en un nuevo ensayo, valor resultante bajo la distribución dada en (2.30). Como previamente se indicó, la predicción puntual \hat{Y}_0 de tal valor futuro, se obtiene con el modelo de regresión ajustado en la muestra de ajuste, de acuerdo a la ecuación (2.32). Esta ecuación define una variable aleatoria que es independiente de Y_0 puesto que el valor futuro de la respuesta no hace parte de la muestra con la cual se obtienen las estimaciones de los parámetros. Específicamente bajo los supuestos del MRLM, se tiene que:

- La predicción \hat{Y}_0 es una variable aleatoria cuya distribución es la misma que la de $\hat{\mu}_{Y|\mathbf{x}_0}$, dada en (2.33).
- La predicción \hat{Y}_0 es estadísticamente independiente de Y_0 , por tanto $\text{Cov}(Y_0, \hat{Y}_0) = 0$.
- Como consecuencia de las dos propiedades anteriores, el error de predicción en \mathbf{x}_0 , dado por $e_0 = Y_0 - \hat{Y}_0$, es una variable aleatoria que se distribuye también normal,

$$e_0 \sim N\left(0, \sigma^2 \left[1 + \mathbf{x}_0^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}_0\right]\right). \quad (2.34)$$

- Bajo los supuestos del MRLM, teniendo en cuenta además que σ^2 es desconocido y que su estimador insesgado es el MSE, puede demostrarse que

$$T_0 = \frac{e_0}{\sqrt{\text{MSE} [1 + \mathbf{x}_0^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}_0]}} = \frac{Y_0 - \hat{Y}_0}{\sqrt{\text{MSE}(1 + h_{00})}} \sim t_{n-k-1}, \quad (2.35)$$

$$\text{con } h_{00} = \mathbf{x}_0^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}_0.$$

Es con base en el estadístico T_0 dado en (2.35), que se deduce el intervalo de predicción (I.P) de nivel $(1 - \alpha)100\%$ para la respuesta futura Y_0 , a partir de la probabilidad

$$P(-t_{\alpha/2, n-k-1} \leq T_0 \leq t_{\alpha/2, n-k-1}) = P\left(-t_{\alpha/2, n-k-1} \leq \frac{Y_0 - \hat{Y}_0}{\sqrt{\text{MSE}(1 + h_{00})}} \leq t_{\alpha/2, n-k-1}\right) = 1 - \alpha.$$

La Tabla 2.6 resume las inferencias sobre la respuesta futura.

Tabla 2.6: Inferencias sobre la respuesta futura Y_0 en $\mathbf{x}_0 = (1, x_{01}, x_{02}, \dots, x_{0k})^T$

Pronóstico puntual	Estadístico	Intervalos de predicción del $(1 - \alpha) 100\%$
$\hat{Y}_0 = \hat{\beta}_0 + \sum_{j=1}^k \hat{\beta}_j x_{0j}$	$T_0 = \frac{\hat{Y}_0 - Y_0}{\sqrt{\text{MSE}[1 + h_{00}]}} \sim t_{n-k-1}$	$\hat{Y}_0 \pm t_{\alpha/2, n-k-1} \sqrt{\text{MSE}[1 + h_{00}]}$
Con $h_{00} = \mathbf{x}_0^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}_0$		

Nota 2.4. Observe que

- Un intervalo de confianza (I.C) para la respuesta media en $\mathbf{x} = \mathbf{x}_0$, proporciona un rango de valores en el cual se estima, con una confianza de $(1 - \alpha) 100\%$, puede encontrarse el valor esperado o media de la distribución de la respuesta cuando \mathbf{x} es fijado en \mathbf{x}_0 .
- Un intervalo de predicción (I.P) para un valor futuro de la respuesta en $\mathbf{x} = \mathbf{x}_0$, proporciona una predicción del rango de valores en cual, con una confianza de $(1 - \alpha) 100\%$, podría ocurrir el valor de la respuesta en un nuevo ensayo con \mathbf{x} fijado en \mathbf{x}_0 .
- Desde que para todo \mathbf{x}_0 la varianza del error de predicción es mayor que la varianza del estimador de la respuesta media, los I.P para la respuesta futura en $\mathbf{x} = \mathbf{x}_0$, son más amplios que los I.C para la respuesta media en $\mathbf{x} = \mathbf{x}_0$, lo que indica que es menos precisa y por tanto más incierta la predicción puntual que la estimación de la media de la respuesta en cada valor particular en que se fije a \mathbf{x} . Además, la incertidumbre es mucho mayor a mayor distancia del punto \mathbf{x}_0 al centro de los datos usados en la estimación del modelo.

Nota 2.5. Es muy importante valorar la precisión de los pronósticos a través de los I.P, esto se hace evaluando tanto la amplitud del intervalo (diferencia entre los límites superior e inferior) como la cobertura (proporción de los I.P calculados que contienen el verdadero valor que se observe para la variable respuesta), de forma que a menor longitud y mayor cobertura promedios, mejor es la calidad de los pronósticos por I.Ps.

2.9. Coeficiente de determinación Múltiple ó R^2 , el coeficiente de correlación múltiple y el R^2_{adj}

Retomemos la descomposición ANOVA de la variabilidad total observada en la respuesta, dada en la ecuación (2.18). El coeficiente de determinación muestral, denotado por R^2 , es un estadístico que aparece en los resultados de la regresión lineal y proviene de la razón

$$R^2 = \frac{\text{SSR}}{\text{SST}} = 1 - \frac{\text{SSE}}{\text{SST}} = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2}, \quad (2.36)$$

por tanto, podemos interpretar el R^2 como la *proporción de la variabilidad total observada en la variable respuesta, que es explicada por la relación lineal con las variables predictoras consideradas*. Esta cantidad ha sido utilizada erróneamente como medida para evaluar la bondad del ajuste lineal, pues si bien valores cercanos a 1 indican una mayor asociación lineal, no necesariamente garantiza que los supuestos básicos del modelo lineal se estén cumpliendo y menos que no haya carencia de ajuste lineal (Montgomery et. al., 2021; Kutner et. al., 2005). Kutner et. al., 2005, lista como los tres errores de interpretación más comunes del R^2 , los siguientes:

1. *Creer que un R^2 alto indica que el modelo puede hacer predicciones útiles.* Hay casos donde se tiene un R^2 alto y sin embargo, los intervalos de predicción para la variable respuesta son muy amplios indicando poca precisión del pronóstico. O bien, el modelo solo es apropiado dentro del espacio de valores muestrales usados en el ajuste pero no por fuera de éste.
2. *Creer que un R^2 alto indica que la superficie de regresión ajustada tiene buen ajuste.* Por ejemplo, en la regresión lineal simple (RLS), hay casos en los cuales se ajusta una recta obteniendo un R^2 alto cuando la verdadera relación no es lineal. Ver Figura 2.3(b).
3. *Creer que un R^2 cercano a cero indica que Y no está relacionada con las variables X_j .* Por ejemplo, en la RLS, cuando existe una relación no lineal entre X y Y , puede ocurrir que al ajustar considerando linealidad, el R^2 dé cercano a cero. Ver Figura 2.3(c).

En el modelo (2.1), $R = \sqrt{R^2}$ es llamado el coeficiente de correlación múltiple, puesto que mide la relación entre la variable respuesta Y y un conjunto de predictores X_1, X_2, \dots, X_k .

Por otra parte, también suele darse como resultado de la calidad del ajuste el R^2 ajustado, denotado por R^2_{adj} , el cual es una medida que tiene en cuenta la magnitud del modelo, es decir el número de predictores presentes, y que penaliza al modelo en virtud de los grados de libertad en la estimación de la varianza. Su ecuación es como sigue,

$$R^2_{adj} = 1 - \frac{\text{MSE}}{\text{MST}} = 1 - \frac{n-1}{n-k-1} (1 - R^2). \quad (2.37)$$

A diferencia del R^2 , el R^2_{adj} puede decrecer con el ingreso de nuevas variables predictoras al modelo de regresión, incluso, $R^2_{adj} \in (-\infty, 1]$, mientras que $R^2 \in [0, 1]$, por ello el R^2_{adj} no puede interpretarse como la proporción de la variabilidad observada en la respuesta que es explicada por el modelo. Como medida de ajuste, entre modelos que ajusten a la respuesta en la misma escala, es mejor aquél cuyo R^2_{adj} sea el mayor.

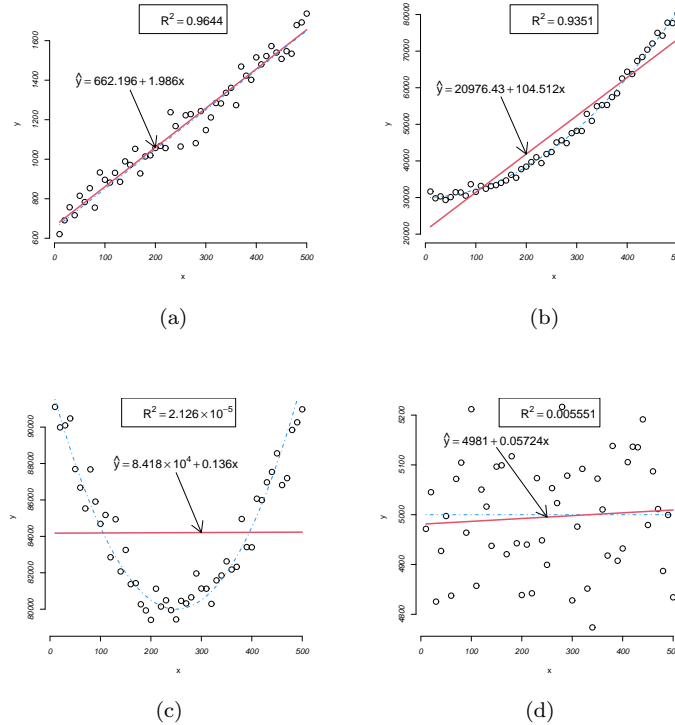


Figura 2.3: Interpretación del R^2 : En los cuatro casos el modelo ajustado es $Y = \beta_0 + \beta_1 X + E$, con $E_i \stackrel{iid}{\sim} N(0, \sigma^2)$. (a) la verdadera relación estadística es lineal: $\mu_{Y|x} = 650 + 2x$ y el ajuste arroja un R^2 cercano a 1; (b) La verdadera relación no es lineal: $\mu_{Y|x} = 30000 + 2x + 0.2x^2$ aunque el ajuste arroja R^2 cercano a 1; (c) La verdadera relación no es lineal: $\mu_{Y|x} = 92500 - 100x + 0.2x^2$ y el ajuste da un R^2 de casi cero, sin embargo, en este caso, no se puede decir que no existe relación estadística entre X y Y sino que la relación es no lineal; (d) El verdadero modelo es $Y_i = 5000 + E_i$, $E_i \stackrel{iid}{\sim} N(0, \sigma^2)$, es decir, no hay asociación estadística de Y con X , sin embargo, se ajustó el MRL asumiendo que $\mu_{Y|x} = \beta_0 + \beta_1 x$, y su ajuste da un R^2 pequeño, como era de esperarse y la estimación del modelo $Y_i = \beta_0 + E_i$, $E_i \stackrel{iid}{\sim} N(0, \sigma^2)$ da $\hat{\beta}_0 = 4995.508$ muy próximo a la media verdadera de Y .

2.10. Evaluación de los supuestos en un MRLM

Recuerde que en el modelo de regresión se han impuesto las siguientes condiciones sobre el término de error:

- Los errores son variables aleatorias normales de media cero.
- Los errores se distribuyen con igual varianza.
- Los errores son mutuamente independientes.

Las desviaciones del modelo pueden ser estudiadas a través de los residuales (Kutner et. al., 2005). Algunos de los tipos de desviaciones que pueden presentarse son:

- La función de regresión presenta carencia de ajuste en una o más de sus variables explicatorias.
- Los errores no tienen varianza constante.
- Los errores no son independientes.
- Los errores no son normales.
- El modelo ajusta bien pero unas pocas observaciones son outliers o atípicas.

Ahora bien, puede demostrarse que bajo la validez de los supuestos sobre los errores de ajuste, el vector de residuos del ajuste es normal n -variado de la siguiente manera,

$$\hat{\mathbf{E}} \sim N_n(\mathbf{0}_{n \times 1}, \sigma^2 [\mathbf{I}_n - \mathbf{H}]), \quad (2.38)$$

donde \mathbf{I}_n es la matriz identidad de orden n y $\mathbf{H} = \mathbf{X} (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T$ es conocida como **la matriz hat o matriz sombrero**, la cual es una matriz $n \times n$ de proyección ortogonal, tal que transforma el vector de respuesta observado en el vector de respuesta estimado, $\hat{\mathbf{Y}} = \mathbf{H}\mathbf{Y}$. De (2.38), es claro que los residuos de ajuste no son ni independientes ni de varianza constante, ya que su matriz de varianzas-covarianzas es $\sigma^2(\mathbf{I}_n - \mathbf{H})$, sin embargo para n grande, esta no homogeneidad en varianza y la no independencia se reducen, de modo que podemos proceder en los diagnósticos y chequeo de supuestos del modelo de la siguiente manera:

2.10.1. Supuesto de media cero

El supuesto de media cero no se rechaza desde que el modelo de regresión no sea carente de ajuste, es decir, si la función de regresión propuesta representa adecuadamente la media de la respuesta en función de las variables predictoras, y en ese caso la dispersión de los residuos debe ocurrir alrededor de cero sin importar el nivel de valores de la respuesta estimada o de las variables predictoras:

- El gráfico de \hat{E}_i vs. \hat{y}_i , el cual debe mostrar una nube de puntos centrada verticalmente en cero y libre de patrones de tendencia (ver Figura 2.4(b)), puesto que bajo los supuestos sobre los errores del MRLM, los residuos y la respuesta estimada son incorrelacionadas, esto es,

$$\text{Corr}(\hat{E}_i, \hat{Y}_j) = 0, \quad \forall i, j, \quad (2.39)$$

de hecho, el vector de residuos y de la respuesta estimada son independientes bajo los supuestos del MRLM.

- Ordinariamente también se construyen los gráficos de residuos vs. cada X_j , en los que si el modelo introduce de forma apropiada a estas variables predictoras en la ecuación, no se espera algún patrón en estos gráficos, por el contrario, los residuos deben mostrar una dispersión sin ninguna tendencia con relación a la línea horizontal en cero (ver Figura 2.4(a)), ya que bajo los supuestos del modelo y su correcta especificación, los residuos deben tener media en cero sin importar el valor de los predictores. Tendencias en estos gráficos son indicativos de desviaciones con respecto a la relación asumida entre Y y la X_j asociada al gráfico de residuos donde es observada esta situación.

2.10.2. Supuesto de varianza constante

Para chequear el supuesto de varianza constante también resultan útiles los gráficos de residuales versus valores ajustados de la respuesta y versus cada variable X_j , de modo que con n grande, sin carencia de ajuste del modelo y validez del supuesto de varianza constante en los errores de ajuste, los gráficos de \hat{E}_i vs. \hat{y}_i y de \hat{E}_i vs. X_j deben mostrar una dispersión homogénea de los residuos alrededor de cero, Ver gráficas en la Figura 2.4. También existen pruebas de homogeneidad de varianza para modelos de regresión lineal como el test Brown-Forsythe y el test de Breusch - Pagan (ver Kutner et. al., 2005), pero estas pruebas requieren la validez de los supuestos de independencia y de normalidad de los errores del modelo.

El patrón deseable en los gráficos de residuales es ilustrado en la Figura 2.4. Las gráficas en la Figura 2.6 ilustran el caso en el cual el modelo es adecuado en cuanto a la forma de la función de regresión, sin embargo, la varianza no es constante. En la Figura 2.5 se ilustra el caso cuando el modelo no es adecuado (hay carencia de ajuste), aunque la varianza es constante (gráficas del panel superior) y el caso en el cual el modelo no es adecuado y tampoco se cumple que la varianza sea constante (gráficas del panel inferior).

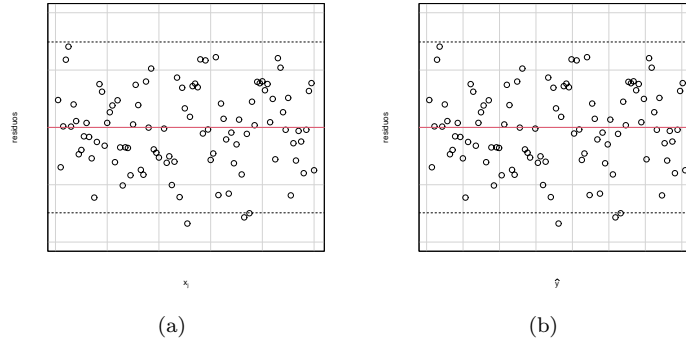


Figura 2.4: (a) Gráfico esperado de residuos vs. un predictor X_j cuando en el modelo no hay anomalías; (b) Gráfico esperado de residuos vs. \hat{y} cuando en el modelo no hay anomalías.

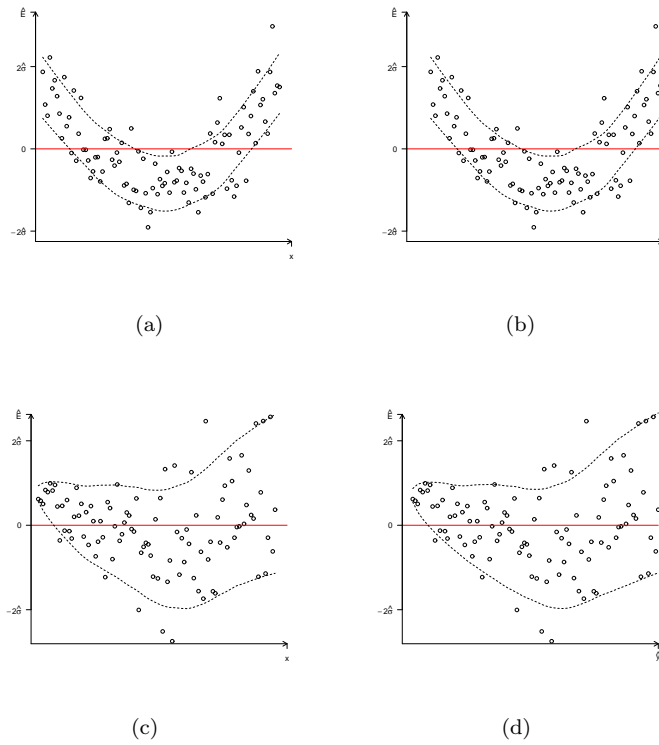


Figura 2.5: Panel superior: Ejemplo del caso donde el modelo lineal entre Y y X_j no es adecuado, pero la varianza es constante, donde (a) residuos vs. X_j y (b) residuos vs. \hat{y} . Panel inferior: Ejemplo del caso donde el modelo lineal entre Y y X_j no es adecuado, ni la varianza es constante, donde (c) residuos vs. X_j y (d) residuos vs. \hat{y} .

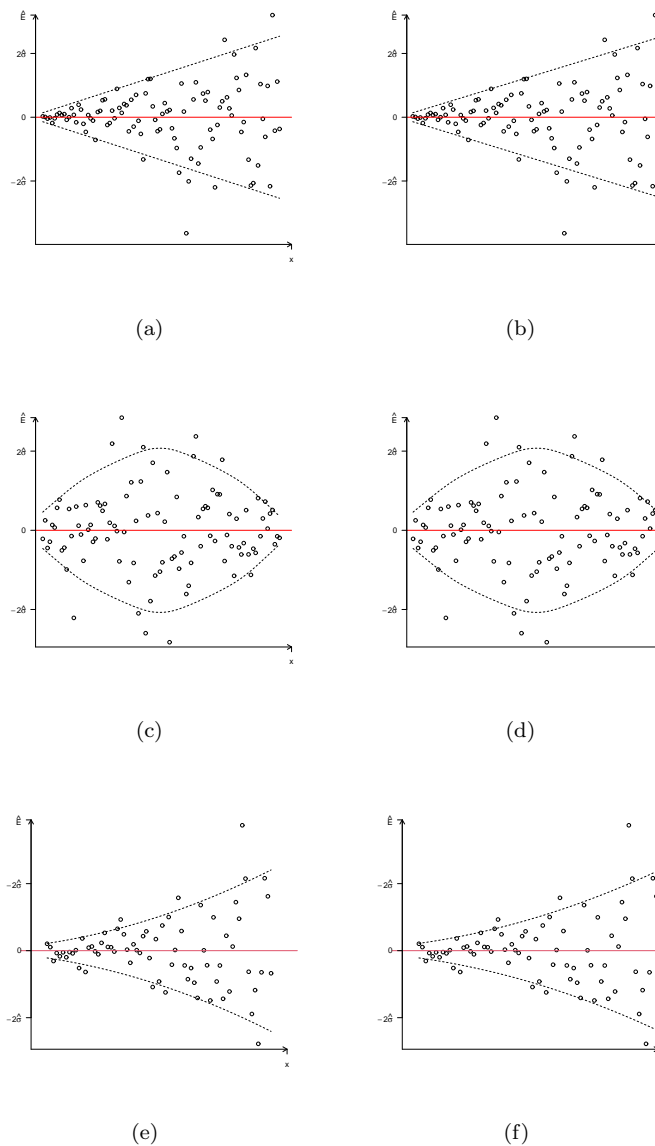


Figura 2.6: Patrones donde el modelo de regresión no es carente de ajuste pero la varianza no es constante. Patrón de embudo: (a) residuos vs. X_j ; (b) residuos vs. \hat{y} . Patrón de balón de fútbol americano: (c) residuos vs. X_j ; (d) residuos vs. \hat{y} . Patrón de embudo no lineal: (e) residuos vs. X_j ; (f) residuos vs. \hat{y} .

2.10.3. Supuesto de independencia

Si los errores de ajuste son independientes, entonces son incorrelacionados. Ahora bien, aunque los residuos de ajuste, que se suponen representan en la muestra a los verdaderos errores del modelo, no son variables aleatorias incorrelacionadas, pues involucran a los \hat{y}_i obtenidos con la misma función de regresión ajustada, la correlación entre estos es pequeña cuando el tamaño de muestra es grande comparado con el número de parámetros de regresión, y por tanto el efecto de la dependencia entre los residuos es despreciable, y en consecuencia ningún patrón particular debe observarse en la gráfica de residuos vs. orden de observación (cuando se tiene conocimiento del orden en que han sido tomadas las observaciones) aunque no siempre en esta gráfica se logran identificar patrones aún cuando existe correlación. También, conocido el orden en que fueron recolectadas las observaciones, es posible analizar la serie de tiempo de los residuales y aplicar por ejemplo el test de Durbin Watson, el cual será estudiado más adelante. En el ámbito de las series de tiempo, existen métodos más efectivos para evaluar si la serie de tiempo de los errores de un modelo de regresión es un ruido blanco², como el test de Lung-Box y las funciones de autocorrelación (ACF por su sigla en inglés) y de autocorrelación parcial (PACF por su sigla en inglés), las cuales también serán presentadas más

²En series de tiempo, un ruido blanco es un proceso estocástico que se caracteriza por tener media constante e igual a cero, varianza constante y los variables del proceso son independientes y por tanto incorrelacionadas.

adelante en esta asignatura.

Nota 2.6. Recuerde que

- La correlación es una medida de *dependencia lineal* entre dos variables aleatorias.
- Correlación nula (o incorrelación) entre dos variables aleatorias no implica independencia, pues pudiera ser que las variables presentan asociación no lineal.
- Bajo independencia estadística entre dos variables aleatorias, no existe asociaciones ni de tipo lineal o no lineal, de modo que la correlación es cero.
- En los modelos de regresión se aplican pruebas de incorrelación bajo el supuesto de independencia, de modo que este supuesto se mantendrá a menos que se encuentre evidencia de correlación no nula, es decir, si hay correlación entre dos variables se rechaza la independencia, pero si no se encuentra evidencia muestral de correlación, se mantiene el supuesto de independencia (aunque no se ha probado con certeza que es correcto tal supuesto).
- El supuesto de independencia entre los errores de ajuste del MRLM se asumió válido en el curso de Estadística II, considerando que el diseño muestral que fue seguido y la forma en que se obtuvieron los datos garantizan la independencia entre las observaciones.

2.10.4. Supuesto de normalidad

Si los errores son normales, se espera que en la gráfica de probabilidad normal obtenida con los residuos de ajuste, no se encuentren desviaciones significativas con respecto a la recta de probabilidad normal (en este gráfico se evalúa si la nube de los puntos en la escala normal se puede ajustar por la línea recta del modelo de los cuantiles muestrales versus los cuantiles normales), y además que cualquier test de bondad de ajuste distribucional no rechace el supuesto de normalidad. Entre las pruebas de normalidad, la más popular es el test de Shapiro-Wilk que para el caso de la regresión es aplicado sobre los residuos de ajuste, sin embargo, ya sea con el gráfico de probabilidad normal o con un test de normalidad, no debe perder de vista que la prueba se formula para los errores del modelo de regresión y las conclusiones también deben formularse para los errores del modelo, es decir, concluir si se ha encontrado o no evidencia muestral en contra del supuesto de normalidad de los errores E_i del modelo de regresión:

$$H_0 : E_i \sim N(0, \sigma^2), \quad (2.40)$$

$$H_1 : E_i \not\sim N(0, \sigma^2). \quad (2.41)$$

Nota 2.7. Tenga en cuenta que:

- Como cualquier test de bondad de ajuste, los tests de normalidad exigen que el conjunto de valores sobre los que se aplican provengan de una muestra aleatoria³.
- La correlación entre variables en una muestra puede afectar significativamente el desempeño de los tests de bondad de ajuste distribucional.
- El supuesto de independencia debería verificarse antes de la evaluación de normalidad.

Ver gráficas en la Figura 2.7 sobre patrón correcto y patrones con desvío de la normalidad, en el gráfico de probabilidad normal.

Nota 2.8. Puede haber desacuerdo entre el test y el gráfico de normalidad. En tal caso tener en cuenta que el gráfico de probabilidad debe ser la herramienta principal para la decisión final.

2.10.5. Observaciones atípicas u outliers en la variable respuesta

Observaciones atípicas en la respuesta Y son aquella que en algún aspecto están separadas del resto de los datos y por tanto pueden afectar los resultados del ajuste del modelo de regresión. Nos interesa identificarlas para luego determinar si se tratan de observaciones malas (por errores de registro o medición) que pueden ser descartadas, o si realmente son datos correctos pero extraños que no deben ser eliminados del conjunto de datos. Para detectar observaciones atípicas construimos *residuales escalados* (los residuales divididos por una estimación de su error estándar, Ver Kutner et. al., 2005, o bien Montgomery et. al., 2021). Sin embargo, en este curso no trabajaremos con los residuales escalados y simplemente utilizaremos en los gráficos de residuos líneas horizontales trazadas al nivel de $\pm 2\hat{\sigma}$ (ver Figura 2.4), de modo que aquellos residuos por fuera de tales límites serán considerados como provenientes de observaciones sospechosas de ser atípicas. Recuerde que bajo una distribución $N(0, \sigma^2)$, aproximadamente el 95 % de las observaciones se encuentran entre -2σ y 2σ .

³Recuerde que un conjunto de variables aleatorias, W_1, W_2, \dots, W_n , constituyen una muestra aleatoria de tamaño n si y solo si son mutuamente independientes e idénticamente distribuidas.

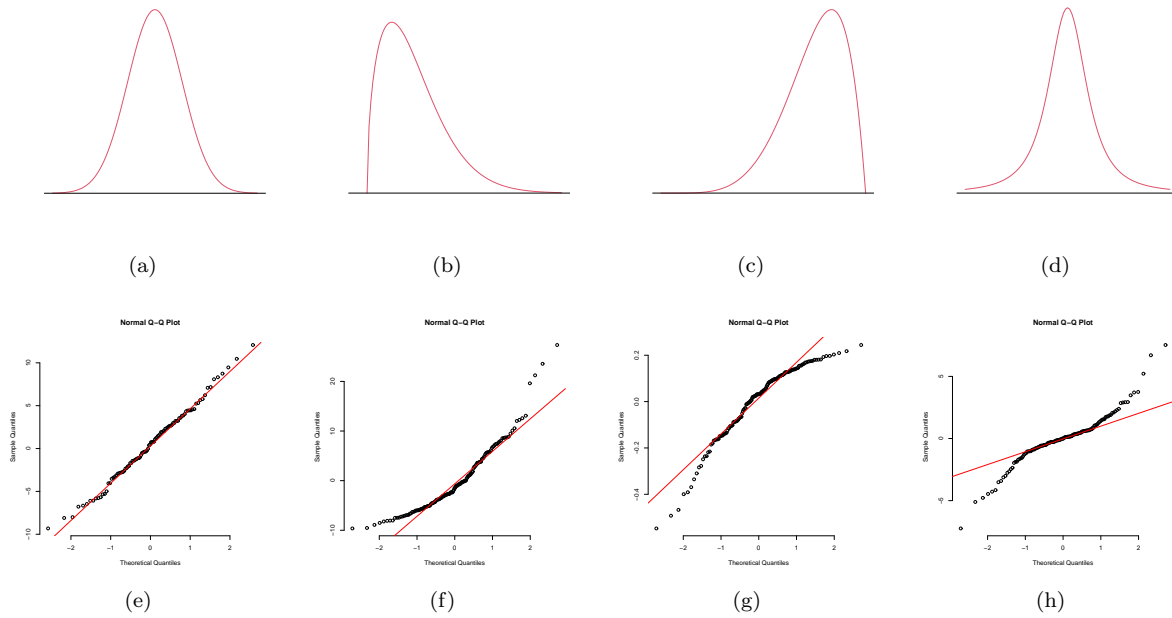


Figura 2.7: Densidades de la distribución poblacional y patrones en gráficos de probabilidad normal sobre una muestra proveniente de tales distribuciones. (a) y (e) con una distribución normal de media cero; (b) y (f) con una distribución no normal y asimétrica a derecha; (c) y (g) con una distribución no normal asimétrica a izquierda; (d) y (h) con una distribución no normal, simétrica pero de colas pesadas.

2.11. Regresión lineal con variables indicadoras

Cuando se presenta la ecuación general del MRLM, las variables predictoras X_j son asumidas de tipo cuantitativo, pero existen casos donde una o más de las variables predictoras pudieran ser categóricas o cualitativas, es decir, definidas en una escala nominal u ordinal, por ejemplo, género, nivel de escolaridad, nivel de calidad de un producto o de satisfacción con un servicio, etc., y por tanto, sus valores no contienen información numérica por lo que no tiene sentido realizar operaciones aritméticas con los valores de tales variables, aun cuando estos sean definidos con números. En estos casos, se debe definir una variable indicadora para cada nivel o categoría de la variable cualitativa a ser considerada en el modelo. Una variable indicadora sólo toma dos posibles valores: 1 Si la característica o nivel de interés es observado y 0 en caso contrario. Por tanto, si el predictor cualitativo posee c niveles o categorías, se podrían definir igual número de variables indicadoras, digamos I_1, I_2, \dots, I_c ,

$$I_j = \begin{cases} 1 & \text{si en la unidad experimental es observada la categoría } j \\ 0 & \text{si en la unidad experimental no es observada la categoría } j \end{cases}, \quad j = 1, 2, \dots, c. \quad (2.42)$$

Sin embargo, desde que sobre el i -ésimo individuo o unidad de observación en la muestra de ajuste sólo es posible observar una de las c categorías (es decir, sólo una de las variables indicadoras puede tomar el valor 1 en el individuo u observación i), entonces $\sum_{j=1}^c I_{ij} = 1$, con I_{ij} el valor de la j -ésima variable indicadora en el i -ésimo individuo u observación, lo que implica que en principio sólo $c - 1$ de estas indicadoras serían necesarias en el modelo de regresión, además, para evitar que en la matriz de diseño \mathbf{X} asociada al modelo de regresión con intercepto β_0 se presente dependencia lineal entre las columnas de las c variables indicadoras y la primera columna de 1's que contiene \mathbf{X} , se omitirá del modelo una de las variables I_j .

Nota 2.9. Tenga en cuenta que:

- El nivel cuya variable indicadora es excluida del MRLM es llamado el *nivel de referencia*.
- Variar el nivel de referencia mientras se mantenga la misma estructura de regresión no altera las estimaciones de la respuesta, ni las predicciones del modelo, y por tanto, tampoco cambia la calidad del ajuste y del pronóstico, solo cambia la interpretación de los parámetros del modelo de regresión.

2.11.1. Regresión lineal con un predictor cuantitativo y otro cualitativo

A continuación, se considera el caso cuando en el modelo de regresión existe una variable predictora cuantitativa X_1 y otro predictor de tipo cualitativo X_2 con c niveles y tomaremos como nivel de referencia al último nivel. Definimos

las variables indicadoras para los primeros $c - 1$ niveles, como sigue,

$$I_j = \begin{cases} 1 & \text{si el nivel observado de } X_2 \text{ es el nivel } j \\ 0 & \text{si el nivel observado de } X_2 \text{ es diferente del nivel } j \end{cases}, \quad j = 1, 2, \dots, c-1.$$

En este problema es posible plantear dos tipos de modelos:

Caso 1: La relación lineal de Y vs. X_1 difiere según el nivel observado en X_2 .

Caso 2: El efecto promedio de X_1 sobre Y no depende de X_2 pero la media de Y es diferente según el nivel observado en X_2 .

En el primer caso el modelo estadístico implica plantear una ecuación donde la recta que describe la relación lineal de Y vs. X_1 cambia en su intercepto y pendiente según el nivel observado en X_2 . La ecuación general de este modelo, considerando las primeras $c - 1$ indicadoras, corresponde a

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 I_{i1} + \beta_3 I_{i2} + \dots + \beta_c I_{i,c-1} + \beta_{1,1} X_{i1} * I_{i1} + \beta_{1,2} X_{i1} * I_{i2} + \dots + \beta_{1,c-1} X_{i1} * I_{i,c-1} + E_i, \\ E_i \stackrel{\text{iid}}{\sim} N(0, \sigma^2) \quad (2.43)$$

Este modelo define c rectas de regresión simple no paralelas, de Y vs. X_1 , una en cada categoría de la variable cualitativa X_2 , (ver Figura 2.8(a)), como se establece en la siguiente tabla:

Tabla 2.7: Modelo de Y vs X_1 , en cada nivel de X_2 , caso 1

Nivel j de X_2	Valor indicadoras	Modelo en el nivel j de X_2	Intercepto	Pendiente
1	$I_1 = 1, I_j = 0, j \neq 1$	$Y_i = (\beta_0 + \beta_2) + (\beta_1 + \beta_{1,1})X_{i1} + E_i, E_i \stackrel{\text{iid}}{\sim} N(0, \sigma^2)$	$\beta_0 + \beta_2$	$\beta_1 + \beta_{1,1}$
2	$I_2 = 1, I_j = 0, j \neq 2$	$Y_i = (\beta_0 + \beta_3) + (\beta_1 + \beta_{1,2})X_{i1} + E_i, E_i \stackrel{\text{iid}}{\sim} N(0, \sigma^2)$	$\beta_0 + \beta_3$	$\beta_1 + \beta_{1,2}$
\vdots	\vdots	\vdots	\vdots	\vdots
$c - 1$	$I_{c-1} = 1, I_j = 0, j \neq (c - 1)$	$Y_i = (\beta_0 + \beta_c) + (\beta_1 + \beta_{1,c-1})X_{i1} + E_i, E_i \stackrel{\text{iid}}{\sim} N(0, \sigma^2)$	$\beta_0 + \beta_c$	$\beta_1 + \beta_{1,c-1}$
c	$I_j = 0, j = 1, 2, \dots, c - 1$	$Y_i = \beta_0 + \beta_1 X_{i1} + E_i, E_i \stackrel{\text{iid}}{\sim} N(0, \sigma^2)$	β_0	β_1

La función de regresión que corresponde al valor esperado de Y dado X_1 y X_2 es,

$$E[Y_i|X_1, X_2] = \beta_0 + \beta_1 X_{i1} + \beta_2 I_{i1} + \beta_3 I_{i2} + \dots + \beta_c I_{i,c-1} + \beta_{1,1} X_{i1} * I_{i1} + \beta_{1,2} X_{i1} * I_{i2} + \dots + \beta_{1,c-1} X_{i1} * I_{i,c-1}, \quad (2.44)$$

luego, también se puede discriminar según el nivel de X_2 como muestra la siguiente tabla.

Tabla 2.8: Valor esperado de Y vs X_1 , en cada nivel de X_2 , caso 1, y diferencia de medias con respecto al nivel de referencia

Nivel j de X_2	$E[Y_i X_1, X_2 = j]$	$E[Y_i X_1, X_2 = j] - E[Y_i X_1, X_2 = c]$
1	$(\beta_0 + \beta_2) + (\beta_1 + \beta_{1,1})X_{i1}$	$\beta_2 + \beta_{1,1}X_1$
2	$(\beta_0 + \beta_3) + (\beta_1 + \beta_{1,2})X_{i1}$	$\beta_3 + \beta_{1,2}X_1$
\vdots	\vdots	\vdots
$c - 1$	$(\beta_0 + \beta_c) + (\beta_1 + \beta_{1,c-1})X_{i1}$	$\beta_c + \beta_{1,c-1}X_1$
c	$\beta_0 + \beta_1 X_{i1}$	0

En el caso 2 el modelo estadístico implica una ecuación donde la recta que describe la relación lineal de Y vs. X_1 cambia solo en su intercepto según el nivel observado en X_2 . La ecuación general de este modelo, considerando las primeras $c - 1$ indicadoras, corresponde a

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 I_{i1} + \beta_3 I_{i2} + \dots + \beta_c I_{i,c-1} + E_i, \quad E_i \stackrel{\text{iid}}{\sim} N(0, \sigma^2) \quad (2.45)$$

Este modelo define c rectas de regresión simple paralelas y no necesariamente coincidentes, de Y vs. X_1 (ver Figura 2.8(b)), como se explica en la Tabla 2.9. Por su parte, la función de regresión que corresponde al valor esperado de Y dado X_1 y X_2 es,

$$E[Y_i|X_1, X_2] = \beta_0 + \beta_1 X_{i1} + \beta_2 I_{i1} + \beta_3 I_{i2} + \dots + \beta_c I_{i,c-1}, \quad (2.46)$$

la cual se reduce en cada nivel de X_2 como muestra la Tabla 2.10.

Tabla 2.9: Modelo de Y vs X_1 , en cada nivel de X_2 , caso 2.

Nivel j de X_2	Valor indicadores	Modelo en el nivel j de X_2	Intercepto	Pendiente
1	$I_1 = 1, I_j = 0, j \neq 1$	$Y_i = (\beta_0 + \beta_2) + \beta_1 X_{i1} + E_i, E_i \stackrel{\text{iid}}{\sim} N(0, \sigma^2)$	$\beta_0 + \beta_2$	β_1
2	$I_2 = 1, I_j = 0, j \neq 2$	$Y_i = (\beta_0 + \beta_3) + \beta_1 X_{i1} + E_i, E_i \stackrel{\text{iid}}{\sim} N(0, \sigma^2)$	$\beta_0 + \beta_3$	β_1
\vdots	\vdots	\vdots	\vdots	\vdots
$c-1$	$I_{c-1} = 1, I_j = 0, j \neq (c-1)$	$Y_i = (\beta_0 + \beta_c) + \beta_1 X_{i1} + E_i, E_i \stackrel{\text{iid}}{\sim} N(0, \sigma^2)$	$\beta_0 + \beta_c$	β_1
c	$I_j = 0, j = 1, 2, \dots, c-1$	$Y_i = \beta_0 + \beta_1 X_{i1} + E_i, E_i \stackrel{\text{iid}}{\sim} N(0, \sigma^2)$	β_0	β_1

Tabla 2.10: Valor esperado de Y vs X_1 , en cada nivel de X_2 , caso 2, y diferencia de medias con respecto al nivel de referencia

Nivel j de X_2	$E[Y_i X_1, X_2 = j]$	Diferencia con nivel de referencia
1	$(\beta_0 + \beta_2) + \beta_1 X_{i1}$	β_2
2	$(\beta_0 + \beta_3) + \beta_1 X_{i1}$	β_3
\vdots	\vdots	\vdots
$c-1$	$(\beta_0 + \beta_c) + \beta_1 X_{i1}$	β_c
c	$\beta_0 + \beta_1 X_{i1}$	0

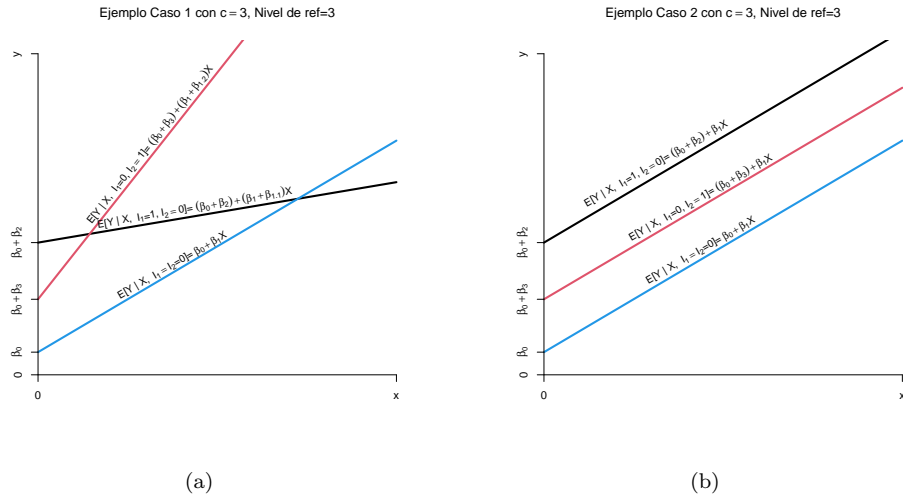


Figura 2.8: (a) Ilustración caso 1, con $c = 3$ y nivel de referencia el 3ro: La relación lineal de Y vs. X cambia con niveles de la variable cualitativa. (b) Ilustración caso 2, con $c = 3$ y nivel de referencia el 3ro: El efecto medio de X sobre Y no cambia con niveles de la variable cualitativa, pero la media de Y no es igual para todos los niveles de la variable cualitativa

¿Cuál de los dos tipos de modelos plantear en un caso dado? Un análisis del gráfico de dispersión de Y vs. X_1 donde cada observación sea identificada según el nivel de la variable categórica X_2 , resulta útil para determinar cuál de los dos tipos de modelos considerar, evaluando si por nivel de X_2 las posibles rectas pueden ser o no paralelas. Otra posibilidad sería plantear el modelo bajo el caso 1 y probar si no existe efecto significativo de X_2 sobre el efecto promedio que X_1 tiene sobre Y , lo cual es lo mismo que probar que las c rectas tienen la misma pendiente, lo cual implica probar que

$$\begin{aligned}
 H_0 : \beta_{1,1} &= \beta_{1,2} = \dots = \beta_{1,c-1} = 0, \\
 H_1 : \beta_{1,j} &\neq 0 \text{ para al menos un } j = 1, 2, \dots, c-1.
 \end{aligned} \tag{2.47}$$

Nota 2.10. El anterior test de hipótesis es probado bajo la construcción de un estadístico de prueba tipo F, así,

$$F_0 = \frac{[\text{SSE}(\text{MR}) - \text{SSE}(\text{MF})] \div [\text{dfe}(\text{MR}) - \text{dfe}(\text{MF})]}{\text{MSE}(\text{MF})} \sim f_{\text{dfe}(\text{MR}) - \text{dfe}(\text{MF}), \text{dfe}(\text{MF})} \tag{2.48}$$

donde,

- SSE(MR) es la suma de cuadrados de residuos del modelo reducido (MR): el modelo al que se reduce la ecuación de regresión del caso 1 al eliminar los términos $\beta_{1,j}X_{i1}I_{ij}$, por tanto el MR es el modelo bajo el caso 2.
- SSE(MF) es la suma de cuadrados de residuos del modelo completo (MF): el modelo bajo el caso 1 y MSE(MF) es su MSE.
- dfe(MR) son los grados de libertad del SSE(MR) y dfe(MF) son los grados de libertad del SSE(MF).
- $f_{\text{dfe(MR)}-\text{dfe(MF)}, \text{dfe(MF)}}$ denota la distribución f_{ν_1, ν_2} con $\nu_1 = \text{dfe(MR)} - \text{dfe(MF)}$ grados de libertad en el numerador y $\nu_2 = \text{dfe(MF)}$ grados de libertad en el denominador.

2.11.2. Ejemplo

Un gran almacén realizó un experimento para investigar los efectos de los gastos por publicidad sobre las ventas semanales de sus secciones de ropa (variable X_2) para caballeros (A), para niños (B) y para damas (C). Se seleccionaron al azar 5 semanas para observación en cada sección, y un presupuesto para publicidad (X_1 , en cientos de dólares) se asignó a cada una de las secciones. Las ventas semanales (Y , en miles de dólares), los gastos de publicidad en cada uno de las tres secciones en cada una de las cinco semanas del estudio se listan en la Tabla 2.11.

Tabla 2.11: Datos observados, ejemplo RLM con variables indicadoras

Sección	Publicidad	Ventas semanales
A	5.2	9
A	5.9	10
A	7.7	12
A	7.9	12
A	9.4	14
B	8.2	13
B	9.0	13
B	9.1	12
B	10.5	13
B	10.5	14
C	10.0	18
C	10.3	19
C	12.1	20
C	12.7	21
C	13.6	22

1. Realice un gráfico de dispersión para analizar la relación entre las ventas y los gastos de publicidad según las secciones y globalmente.
2. Modelo 1: Tomando como nivel de referencia la Sección C, postule y ajuste un modelo de regresión para estudiar los efectos que las secciones del almacén puedan tener sobre la relación de las ventas versus los gastos de publicidad. Halle las ecuaciones de las rectas ajustadas que relacionan las ventas con la publicidad en cada sección. Tomando como indicadoras de las secciones A y B, a I_1 e I_2 , respectivamente:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 I_{i1} + \beta_3 I_{i2} + \beta_{1,1} X_i I_{i1} + \beta_{1,2} X_i I_{i2} + E_i, \quad E_i \stackrel{\text{iid}}{\sim} N(0, \sigma^2)$$

3. Modelo 2: Con nivel de referencia la Sección C, postule y ajuste un MRLM en donde se considere que en promedio el efecto de los gastos en publicidad sobre las ventas es el mismo para las tres secciones, pero la media de las ventas es diferente. Interprete parámetros estimados y analice residuos.

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 I_{i1} + \beta_3 I_{i2} + E_i, \quad E_i \stackrel{\text{iid}}{\sim} N(0, \sigma^2)$$

4. Con base en el modelo anterior, prediga los gastos de publicidad en los puntos de predicción indicados en la tabla siguiente.

Obs.	X_1	SEC
1	6	A
2	11	C

El código R usado para la obtención de los resultados es presentado al final del capítulo.

La Figura 2.9 muestra el gráfico de dispersión pedido, en el cual podemos analizar si la relación entre los ingresos por ventas y los gastos de publicidad depende de la Sección. Efectivamente, esta relación no puede representarse con una sola recta ignorando las secciones, pues de acuerdo a los niveles de esta variable, pueden cambiar la pendiente y/o el intercepto. En principio pudiera ser por tanto más apropiado el modelo 1 que el modelo 2.

Nota 2.11. Para el ajuste de estos modelos en R no es necesario crear las variables indicadoras, pues en la función `lm()` puede hacerse uso de variables de clase **factor**, e internamente la función `lm()` crea e incluye en el modelo las indicadoras para los niveles diferentes al nivel de referencia. Por defecto, el nivel de referencia es el primero de los niveles en orden alfanumérico, que R encuentre en la variable tipo **factor**. Si ese nivel no corresponde al que se desea como nivel de referencia, lo ajustamos previamente mediante la función R `relevel()`, como se muestra en el código R para este ejemplo.

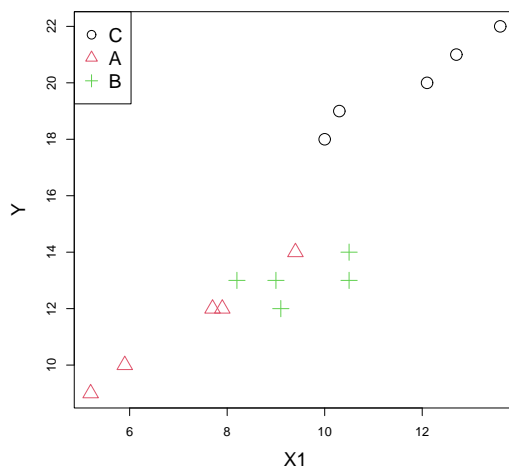


Figura 2.9: Gráfico de dispersión de las ventas Y vs. los gastos en publicidad X_1 , identificando la Sección.

Los siguientes son los resultados del ajuste del modelo 1. Note en la Figura 2.10 cómo es especificado en la función `lm()` la fórmula del modelo 1: $Y \sim X_1 * SEC$, donde Y , X_1 , SEC son los objetos R creados previamente con los valores de las variables Y , X_1 y X_2 definida esta última como un objeto clase **factor** de niveles A, B, y C, siendo C el nivel de referencia. En la salida R de la tabla de parámetros estimados observamos que internamente la función `lm()` creó las indicadoras I_1 e I_2 bajo los nombres `SECA` y `SECB` respectivamente, en tanto que los términos de interacción $X_1 * I_1$ y $X_1 * I_2$ son creados con los nombres `X1:SECA` y `X1:SECB`, respectivamente. La interpretación de los resultados R del ajuste de este modelo son presentados en las Tablas 2.12 y 2.13.

```
R Console
Archivo  Editar  Misceláneo  Paquetes  Ventanas  Ayuda

> modelo1=lm(Y~X1*SEC)
> summary(modelo1)

Call:
lm(formula = Y ~ X1 * SEC)

Residuals:
    Min       1Q   Median       3Q      Max
-0.87683 -0.22516  0.04366  0.14985  0.64418

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   8.2747     1.7957   4.608 0.001276 **
X1             0.9988     0.1519   6.575 0.000102 ***
SECA          -5.2429     2.0724  -2.530 0.032243 *
SECB           1.4888     2.8494   0.522 0.613946
X1:SECA        0.1603     0.2068   0.775 0.458127
X1:SECB       -0.6566     0.2780  -2.362 0.042452 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4709 on 9 degrees of freedom
Multiple R-squared:  0.9916,    Adjusted R-squared:  0.9869
F-statistic: 211.4 on 5 and 9 DF,  p-value: 4.782e-09
```

Figura 2.10: Captura consola R: Ajuste y tabla de parámetros estimados en el modelo 1.

Tabla 2.12: Tabla de parámetros estimados y ec. ajustada según modelo caso 1 tomando como referencia la Sección C

Parámetro	Estimación	Error Estándar	T_0	$P(t_9 > T_0)$
β_0	8.2747	1.7957	4.608	0.0013
β_1	0.9988	0.1519	6.575	0.0001
β_2	-5.2429	2.0724	-2.530	0.0322
β_3	1.4888	2.8494	0.522	0.6139
$\beta_{1,1}$	0.1603	0.2068	0.775	0.4581
$\beta_{1,2}$	-0.6566	0.2780	-2.362	0.0425
$\hat{Y}_i = 8.2747 + 0.9988X_{i1} - 5.2429I_{i1} + 1.4888I_{i2} + 0.1603X_{i1}I_{i1} - 0.6566X_{i1}I_{i2}$ $\sqrt{\text{MSE}} = 0.4709$, $R^2 = 0.9916$, $R^2_{adj} = 0.9869$ Resultados para test Anova del modelo: $F_0 = 211.4$, $P(f_{5,9} > F_0) = 4.782 \times 10^{-9}$				

Tabla 2.13: Modelos y ecuaciones ajustadas por Sección según modelo caso 1 tomando como referencia la Sección C

Sección	Modelo	Ecuación ajustada
A ($I_1 = 1, I_2 = 0$)	$Y_i = (\beta_0 + \beta_2) + (\beta_1 + \beta_{1,1})X_{i1} + E_i$, $E_i \stackrel{\text{iid}}{\sim} N(0, \sigma^2)$	$\hat{Y}_i = 3.0318 + 1.1591X_{i1}$
B ($I_1 = 0, I_2 = 1$)	$Y_i = (\beta_0 + \beta_3) + (\beta_1 + \beta_{1,2})X_{i1} + E_i$, $E_i \stackrel{\text{iid}}{\sim} N(0, \sigma^2)$	$\hat{Y}_i = 9.7635 + 0.3422X_{i1}$
C ($I_1 = 0, I_2 = 0$)	$Y_i = \beta_0 + \beta_1X_{i1} + E_i$, $E_i \stackrel{\text{iid}}{\sim} N(0, \sigma^2)$	$\hat{Y}_i = 8.2747 + 0.9988X_{i1}$

Los siguientes son los resultados del ajuste del modelo 2. Note en la Figura 2.11 cómo es especificado en la función `lm()` la fórmula del modelo 2: $Y \sim X_1 + \text{SEC}$, donde Y , X_1 , SEC son de nuevo, los objetos R creados previamente con los valores de las variables Y , X_1 y X_2 definida esta última como un objeto clase **factor** de niveles A, B, y C, siendo C el nivel de referencia. En la salida R de la tabla de parámetros estimados observamos que internamente la función `lm()` creó las indicadoras I_1 e I_2 bajo los nombres **SECA** y **SECB** respectivamente. La interpretación de los resultados R del ajuste de este modelo son presentados en las Tablas 2.14 y 2.15.

```

R Console
Archivo  Editar  Misceláneo  Paquetes  Ventanas  Ayuda

> modelo2=lm(Y~X1+SEC)
> summary(modelo2)

Call:
lm(formula = Y ~ X1 + SEC)

Residuals:
    Min       1Q   Median       3Q      Max
-1.00202 -0.33520 -0.00202  0.29767  1.21398

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    8.6888      1.4455   6.011 8.79e-05 ***
X1              0.9635      0.1210   7.966 6.80e-06 ***
SECA          -4.2451      0.6671  -6.363 5.34e-05 ***
SECB          -4.8033      0.4714 -10.190 6.12e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.6044 on 11 degrees of freedom
Multiple R-squared:  0.983,    Adjusted R-squared:  0.9784
F-statistic: 212 on 3 and 11 DF,  p-value: 5.186e-10

> |

```

Figura 2.11: Captura consola R: Ajuste y tabla de parámetros estimados en el modelo 2.

Tabla 2.14: Tabla de parámetros estimados y ec. ajustada para modelo según caso 2, tomando como referencia la Sección C

Parámetro	Estimación	Error Estándar	T_0	$P(t_{11} > T_0)$
β_0	8.6888	1.4455	6.011	8.79×10^{-5}
β_1	0.9635	0.1210	7.966	6.80×10^{-6}
β_2	-4.2451	0.6671	-6.363	5.34×10^{-5}
β_3	-4.8033	0.4714	-10.190	6.12×10^{-7}
$\hat{Y}_i = 8.6888 + 0.9635X_{i1} - 4.2451I_{i1} - 4.8033I_{i2}$ $\sqrt{\text{MSE}} = 0.6044$, $R^2 = 0.983$, $R^2_{adj} = 0.9784$ Resultados para test Anova del modelo: $F_0 = 212$, $P(f_{3,11} > F_0) = 5.186 \times 10^{-10}$				

Tabla 2.15: Modelos y ecuaciones ajustadas por Sección según modelo caso 2 tomando como referencia la Sección C

Sección	Modelo	Ecuación ajustada
A ($I_1 = 1, I_2 = 0$)	$Y_i = (\beta_0 + \beta_2) + \beta_1X_{i1} + E_i$, $E_i \stackrel{\text{iid}}{\sim} N(0, \sigma^2)$	$\hat{Y}_i = 4.4437 + 0.9635X_{i1}$
B ($I_1 = 0, I_2 = 1$)	$Y_i = (\beta_0 + \beta_3) + \beta_1X_{i1} + E_i$, $E_i \stackrel{\text{iid}}{\sim} N(0, \sigma^2)$	$\hat{Y}_i = 3.8855 + 0.9635X_{i1}$
C ($I_1 = 0, I_2 = 0$)	$Y_i = \beta_0 + \beta_1X_{i1} + E_i$, $E_i \stackrel{\text{iid}}{\sim} N(0, \sigma^2)$	$\hat{Y}_i = 8.6888 + 0.9635X_{i1}$

Observe de las Tablas 2.13 y 2.15 que mientras en el modelo 1 hay tres rectas ajustadas por Sección que difieren tanto en intercepto como en pendiente, en el modelo 2 las tres rectas ajustadas son paralelas difiriendo solo en el intercepto. Ver Figuras 2.12(a) y 2.12(b).

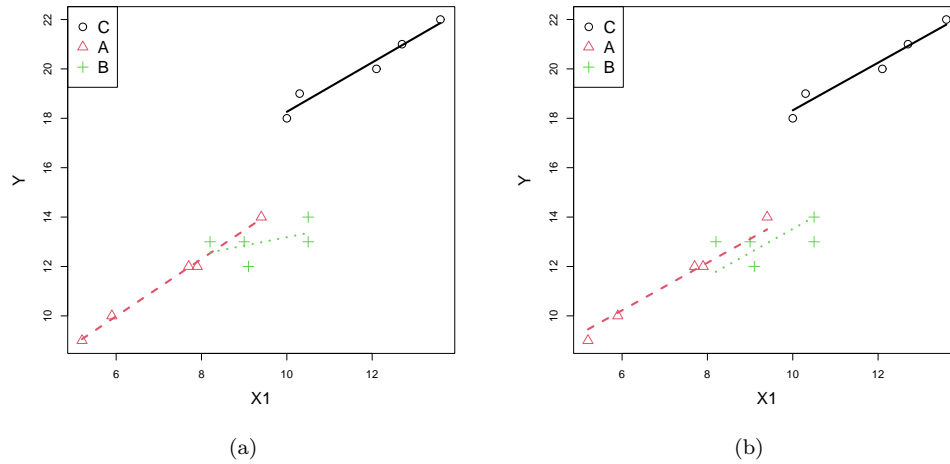


Figura 2.12: (a) Gráfico de dispersión de Y vs. X_1 con rectas ajustadas por Sección según modelo 1. (b) Gráfico de dispersión de Y vs. X_1 con rectas ajustadas por Sección según modelo 2

Dado que los modelos 1 y 2 difieren en el número de parámetros, para comparar la calidad de ajuste usamos bien sea el $\sqrt{\text{MSE}}$ o bien el R_{adj}^2 . Estas medidas son comparables desde que ambos modelos ajustan en la misma escala a la variable respuesta. El ajuste resulta un poco mejor con el modelo 1 (menor $\sqrt{\text{MSE}}$ y mayor R_{adj}^2). Cualitativamente (Ver gráficas en Figura 2.12), con el modelo 1 parece haber mejor ajuste de las observaciones en la Sección A que con el modelo 2; por otra parte, no hay mucha diferencia entre las rectas ajustadas para las observaciones en la Sección C, mientras que hay gran diferencia entre las rectas ajustadas para la Sección B, en este caso es difícil juzgar si es más adecuado el ajuste de la recta bajo el modelo 2 que bajo el modelo 1.

A continuación, vamos a interpretar las estimaciones de los parámetros en el modelo 2. Tenga en cuenta que las interpretaciones que se dan en el modelo 2 no son aplicables en el caso del modelo 1, donde se asume que la relación lineal entre Y y X_1 cambia según la Sección (o sea, en el modelo donde las rectas son diferentes tanto en pendiente como en intercepto). Según la tabla de parámetros ajustados del modelo 2, se tiene que (recuerde que cada unidad en Y representa mil dólares y cada unidad en X_1 representa cien dólares):

- $\beta_1 \neq 0$, pues su valor P , $P(|t_{11}| > |T_0|) = 6.80 \times 10^{-6}$, indica la significancia de este parámetro y por tanto podemos hacer la siguiente conclusión con $\hat{\beta}_1 = 0.9635$: Se estima que en promedio por cada cien dólares que se incremente el gasto en publicidad (X_1) en una semana (en cualquiera de las tres secciones) habrá un incremento en 963.5 dólares en las ventas.
- $\beta_2 \neq 0$, pues su valor P , $P(|t_{11}| > |T_0|) = 5.34 \times 10^{-5}$, indica la significancia de este parámetro y por tanto podemos hacer la siguiente conclusión con $\hat{\beta}_2 = -4.2451$: Se estima que para un mismo nivel de gasto en publicidad semanal (X_1), el promedio de ventas en la Sección A es menor (debido al signo menos) al promedio de ventas en la Sección de referencia (Sección C) en 4245.1 dólares.
- $\beta_3 \neq 0$, pues su valor P , $P(|t_{11}| > |T_0|) = 6.12 \times 10^{-7}$, indica la significancia de este parámetro y podemos hacer la siguiente conclusión con $\hat{\beta}_3 = -4.8033$: Se estima que para un mismo nivel de gasto en publicidad semanal (X_1), el promedio de ventas en la Sección B es menor (debido al signo menos) al promedio de ventas en la Sección de referencia (Sección C) en 4803.3 dólares.

Nota 2.12. En el modelo 2, el valor esperado de Y según Sección y las diferencias de estos valores esperados con respecto al del nivel de referencia, es como se indicó en la Tabla 2.10 tomando $c = 3$, de donde,

Nivel j de X_2	$E[Y_i X_1, X_2 = j]$	$E[Y_i X_1, X_2 = j] - E[Y_i X_1, X_2 = 3]$
1 (Sección A)	$(\beta_0 + \beta_2) + \beta_1 X_{i1}$	β_2
2 (Sección B)	$(\beta_0 + \beta_3) + \beta_1 X_{i1}$	β_3
3 (Sección C)	$\beta_0 + \beta_1 X_{i1}$	0

De aquí que las estimaciones para β_2 y β_3 sean interpretadas en términos de la estimación de la diferencia entre el promedio de las ventas en la sección correspondiente vs. el promedio en la sección de referencia, que para el ejemplo es la C, dado un mismo valor del gasto semanal en publicidad (X_1).

Para el análisis de residuos con fines de evaluar supuestos de que los errores del modelo 2 son normales, de media cero y de varianza constante, considere los gráficos de residuos y de probabilidad normal que se muestran en la Figura 2.13, además de los resultados del Test Shapiro-Wilk. Las líneas horizontales en los gráficos de residuales pasan por $-2\sqrt{\text{MSE}}$, 0 y $2\sqrt{\text{MSE}}$. Recuerde que en el modelo de regresión el MSE estima a σ^2 , de modo que $\pm 2\sqrt{\text{MSE}} = \pm 2\hat{\sigma}$ y que bajo una distribución $N(0, \sigma^2)$ aproximadamente el 95 % de sus valores se encuentra entre $-2\sigma^2$ y $2\sigma^2$. Tenga en cuenta también que por el diseño muestral que se adoptó para la recolección de los datos, es posible asumir la independencia entre las observaciones y por tanto entre los errores del ajuste del modelo 2.

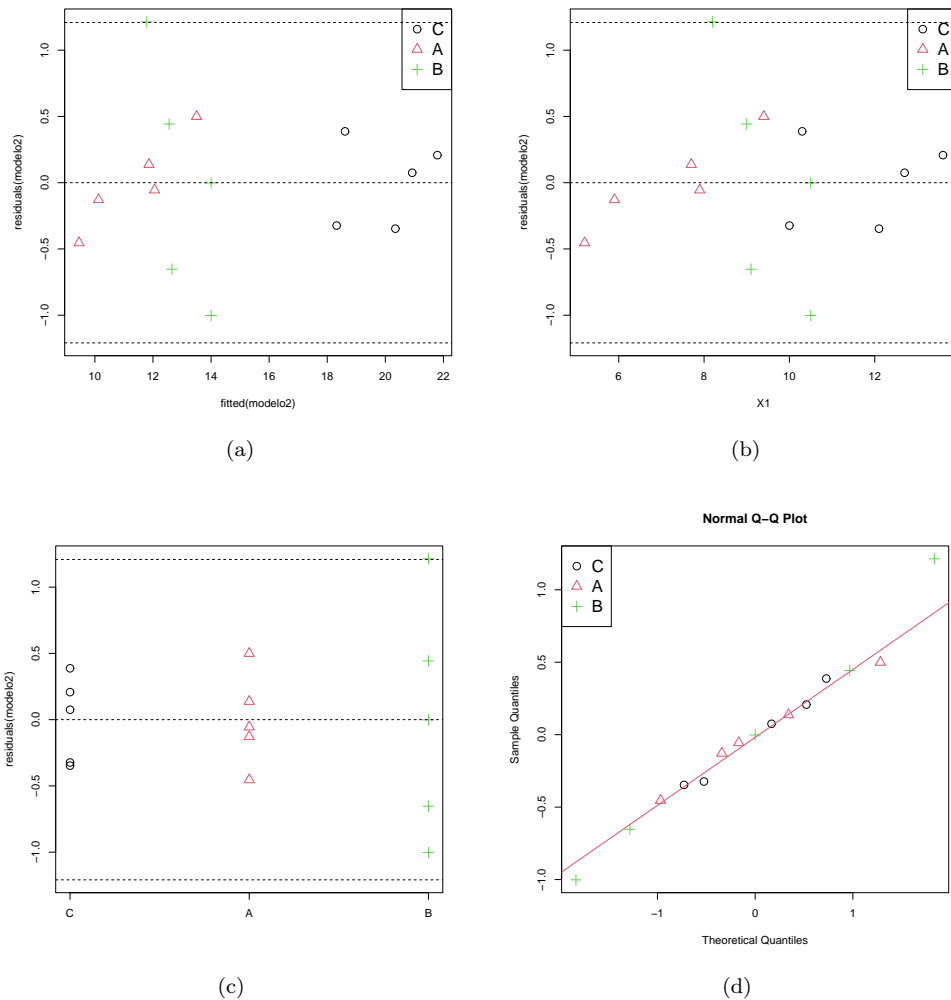


Figura 2.13: Gráficos para análisis de supuestos en el modelo 2: (a) \hat{E} vs. \hat{Y} ; (b) \hat{E} vs. X_1 ; (c) \hat{E} vs. X_2 (Sección); (d) Gráfico de probabilidad normal con los residuos.

De las Figuras 2.13(a), 2.13(b) y 2.13(c), se concluye que

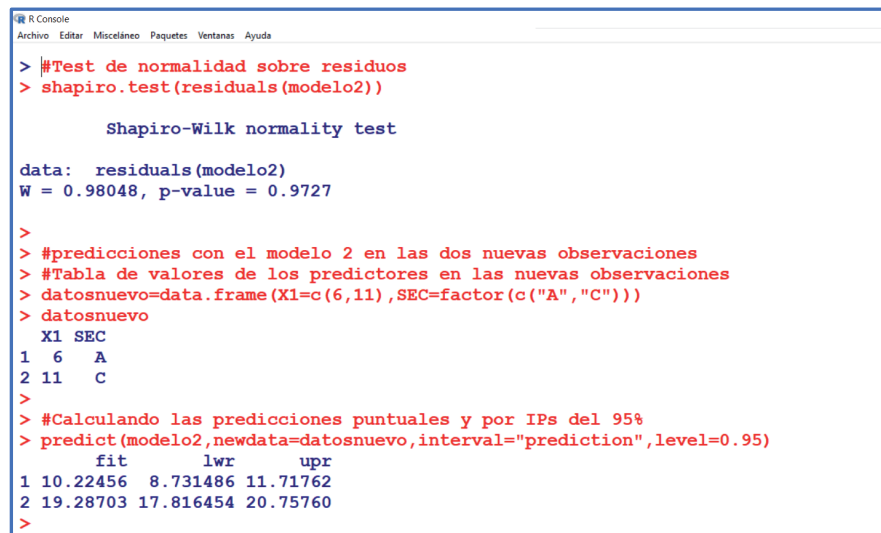
- Dado que es mayor la dispersión respecto a la línea horizontal que pasa por cero para los residuos correspondientes a la sección B (símbolo $+$) que la que muestran los residuos para las otras dos secciones, se concluye que no hay validez del supuesto de varianza constante para los E_i según la sección.
- Observando los residuales para la sección A (símbolo Δ), parece que existe un patrón de tendencia lineal en estos puntos con pendiente positiva, en lugar de una dispersión aleatoria con respecto a cero, entonces hay indicios de que posiblemente el modelo 2 está mal especificado (el modelo es carente de ajuste), y por tanto, no cumple que para los errores de ajuste, dado que la Sección es la A, la media es cero. Recuerde que el modelo 2 asume que las

pendientes de las rectas de Y vs. X_1 en cada sección son iguales, pero éste puede ser un supuesto erróneo. Mire de nuevo el gráfico de dispersión de los datos presentados en la Figura 2.9 (aunque también con pocos datos, la presencia de observaciones atípicas puede sesgar la modelación).

- Con respecto al supuesto de normalidad para los errores E_i , el gráfico de probabilidad en la Figura 2.13(d), no muestra una evidencia fuerte en contra de $H_0 : E_i \sim N(0, \sigma^2)$. Por su parte, el test Shapiro Wilk, cuya ejecución en R y resultado es visible en la Figura 2.14, arroja un estadístico $W_0 = 0.9805$ con valor P: $P(W \leq 0.9805) = 0.9727$, con lo cual $H_0 : E_i \sim N(0, \sigma^2)$ no es rechazada.

Nota 2.13. Desde que el diseño muestral y la forma en que se obtuvieron los datos en este problema permite asumir la independencia y por tanto la incorrelación entre los errores, es decir, $\text{corr}(E_i, E_j) = 0, \forall i \neq j, i, j = 1, 2, \dots, n$, entonces podemos proceder con la evaluación del supuesto de normalidad con la gráfica y el test Shapiro-Wilk.

Finalmente, veamos los resultados de las dos predicciones pedidas en el numeral 4. En la Figura 2.14 vemos además de la ejecución del test Shapiro Wilk, el procedimiento R para el cálculo de las predicciones para los puntos $X_1 = 6$, Sección A (nivel 1 de X_2), y $X_1 = 11$, Sección C (nivel 3 de X_2). En la Tabla 2.16 se muestran los resultados editados de las predicciones puntuales y por intervalos del 95 % (las columnas LIP, LSP) en estos dos puntos y las ecuaciones con las cuales pueden hallarse los pronósticos puntuales del modelo 2 en las Secciones A y C en cualquier valor x_{01} de la variable X_1 . Recuerde que estas ecuaciones son obtenidas reemplazando en la ecuación ajustada del modelo 2 los valores (0 ó 1) de las variables indicadoras I_1 e I_2 , según la Sección a considerar.



```

R Console
Archivo  Editar  Misceláneo  Paquetes  Ventanas  Ayuda

> ##Test de normalidad sobre residuos
> shapiro.test(residuals(modelo2))

      Shapiro-Wilk normality test

data:  residuals(modelo2)
W = 0.98048, p-value = 0.9727

>
> #predicciones con el modelo 2 en las dos nuevas observaciones
> #Tabla de valores de los predictores en las nuevas observaciones
> datosnuevo=data.frame(X1=c(6,11), SEC=factor(c("A", "C")))
> datosnuevo
  X1 SEC
1  6  A
2 11  C
>
> #Calculando las predicciones puntuales y por IPs del 95%
> predict(modelo2, newdata=datosnuevo, interval="prediction", level=0.95)
      fit      lwr      upr
1 10.22456  8.731486 11.71762
2 19.28703 17.816454 20.75760
>

```

Figura 2.14: Captura consola R: Test Shapiro Wilk y predicciones con el modelo 2

Tabla 2.16: Ecuación de pronósticos puntuales y resultados de predicción con el modelo 2 en los puntos $X_1 = 6$, Sección A y $X_1 = 11$, Sección C

Sección	Ec. pronóstico puntual en $X_1 = x_{01}$	Valor de x_{01} (*)	Pronóstico (\hat{Y}_0) (**)	LIP (**)	LSP (**)
A ($I_1 = 1, I_2 = 0$)	$\hat{Y}_0 = 4.4437 + 0.9635x_{01}$	6	10.22456	8.731486	11.71762
C ($I_1 = I_2 = 0$)	$\hat{Y}_0 = 8.6888 + 0.9635x_{01}$	11	19.28703	17.816454	20.75760
(*) Cientos de dólares					
(**) Miles de dólares					

De acuerdo con estos resultados, en la sección de caballeros (A) el modelo 2 pronostica que si se gastan \$600 dólares en publicidad en una semana, las ventas en tal semana serán de \$10224.56 dólares y que el verdadero valor de las ventas podrá estar entre \$8731.49 y \$11717.62 dólares con un 95 % de confianza, mientras que en la sección de damas (C) el modelo 2 pronostica que si se gastan \$1100 dólares en publicidad en una semana, las ventas correspondientes serán de \$19287.03 dólares y que con un 95 % de confianza, el verdadero valor de las ventas podrá estar entre \$17816.45 y \$20757.60 dólares.

2.11.3. Código R usado en el ejemplo de la Sección 2.11.2

A continuación se describe paso a paso el código R usado en la Sección anterior. Este código también está disponible en el Script R de nombre PROGRAMARREPASORLMCONINDICADORAS.R.

Código R 2.1. *Lectura de los datos por teclado: La función `scan()` usada dentro de la función `data.frame()`, intermedia el ingreso de los datos, indicando que tres variables son leídas: `SEC` de tipo alfanumerica y `X1`, `Y` de tipo numerica.*

```
datos=data.frame(scan(what=list(SEC="",X1=0,Y=0)))
A 5.2 9
A 5.9 10
A 7.7 12
A 7.9 12
A 9.4 14
B 8.2 13
B 9.0 13
B 9.1 12
B 10.5 13
B 10.5 14
C 10.0 18
C 10.3 19
C 12.1 20
C 12.7 21
C 13.6 22

#Transformando en un variable tipo factor a la variable SEC guardada en el data.frame datos
datos$SEC=as.factor(datos$SEC)

datos
```

Código R 2.2. *Lectura de los datos desde archivo tipo .csv: Otra forma de leer las observaciones es mediante la función `read.table()`. Los datos guardados en el archivo de nombre `DATOSPROBLEMAGASTOSPUBLICIDAD.csv`, tienen la estructura que muestra la Figura 2.15. Consulte en el Capítulo 1 cómo se determinan los valores de los argumentos de la función `read.table()` según las características del archivo a leer. Al ejecutar este código, se abre una ventana para buscar y seleccionar el archivo.*

	A	B	C	D	E	F	G	H	I
1	Un gran almacén realizó un experimento para investigar los efectos de los gastos por publicidad								
2	sobre las ventas semanales de sus secciones de ropa para caballeros (A), para niños								
3	(B) y para damas (C). Se seleccionaron al azar 5 semanas para observación en cada sección,								
4	y un presupuesto para publicidad (X1, en cientos de dólares) se asignó a cada una de las								
5	secciones. Las ventas semanales (en miles de dólares), los gastos de publicidad en cada uno								
6	de las tres secciones en cada una de las cinco semanas del estudio se listan en la siguiente tabla.								
7	SEC	X1	Y						
8	A	5.2	9						
9	A	5.9	10						
10	A	7.7	12						
11	A	7.9	12						
12	A	9.4	14						
13	B	8.2	13						
14	B	9.0	13						
15	B	9.1	12						
16	B	10.5	13						
17	B	10.5	14						
18	C	10.0	18						
19	C	10.3	19						
20	C	12.1	20						
21	C	12.7	21						
22	C	13.6	22						

Figura 2.15: Visualización del archivo `DATOSPROBLEMAGASTOSPUBLICIDAD.csv`

```
datos=read.table(file.choose(),header=T,skip=6,sep=";",dec=".",colClasses=c("factor","numeric","numeric"))
datos
```

Código R 2.3. *Disponibilizando las variables guardadas en el objeto `datos` y ajuste del nivel de referencia de la variable sección (`SEC` es su nombre `R`) en el nivel `C`: El objeto `datos` dentro del cual existen las variables leídas, es de la clase `data.frame`, y para acceder a sus variables se hace uso de la función `attach()`. la variable de nombre `SEC`, es de tipo `factor` con niveles, `A`, `B`, `C` y por defecto `R` toma como nivel de referencia, el primero de ellos según orden alfanumérico, es decir el nivel `A`, pero queremos cambiarlo por el nivel `C`, esto es hecho con la función `relevel()`.*

```
#Haciendo disponibles las variables del data.frame "datos"
attach(datos)

#verifique valores y orden de los niveles que R adjudico por defecto a la Secciones (nivel 1: A, nivel 2: B, nivel 3: C)
SEC

#Definiendo como nivel de referencia a la seccion C (el orden de los niveles queda asi: nivel 1: C, nivel 2: A, nivel 3: B)
SEC=relevel(SEC,ref="C")
SEC
```

Código R 2.4. Representación gráfica de los datos por gráfico de dispersión. Observe el uso de las funciones `plot()`, `as.numeric()`, `legend()`, `levels()` y sus respectivos argumentos. Note cómo con los argumentos `pch=as.numeric(SEC)`, `col=as.numeric(SEC)` se identifican los puntos de acuerdo a los niveles de la variable `SEC` (la sección).

```
plot(X1,Y,pch=as.numeric(SEC),col=as.numeric(SEC),xlab="X1",ylab="Y",cex=2,cex.lab=1.5)
legend("topleft",legend=levels(SEC),pch=c(1:3),col=c(1:3),cex=1.5)
```

Código R 2.5. Ajuste del modelo 1. Internamente la función `lm()` excluye el nivel C (nivel de referencia) y crea las indicadoras para las Secciones A y B, así como los productos de estas variables con el predictor cuantitativo `X1`. Note que en la fórmula R los nombres de las variables `X1` y `SEC` se separan con el símbolo `*`.

```
modelo1=lm(Y~X1*SEC)
summary(modelo1)
```

Código R 2.6. Gráfico de dispersión con las rectas ajustadas en el modelo 1, según cada sección A, B y C. Observe el uso de las funciones `plot()`, `lines()`, `legend()`, note cómo con los argumentos `pch=as.numeric(SEC)`, `col=as.numeric(SEC)` se identifican los puntos de acuerdo a los niveles de la variable `SEC` (la sección), también, cómo se separan de los valores ajustados (obtenidos con la función `fitted()`) y de la variable `X1` aquellos que corresponden a cada sección para trazar las respectivas rectas ajustadas.

```
win.graph()
plot(X1,Y,pch=as.numeric(SEC),col=as.numeric(SEC),xlab="X1",ylab="Y",cex=1.5,cex.lab=1.5)
legend("topleft",legend=levels(SEC),pch=1:3,col=1:3,cex=1.5) #Leyenda colocada es esquina superior izquierda
lines(X1[SEC=="C"],fitted(modelo1)[SEC=="C"],col=1,lty=1,lwd=3) #Recta ajustada en la Seccion C
lines(X1[SEC=="A"],fitted(modelo1)[SEC=="A"],col=2,lty=2,lwd=3) #Recta ajustada en la Seccion A
lines(X1[SEC=="B"],fitted(modelo1)[SEC=="B"],col=3,lty=3,lwd=3) #Recta ajustada en la Seccion B
```

Código R 2.7. Ajuste del modelo 2. De nuevo, con la fórmula programada dentro de la función `lm()`, ésta excluye el nivel C (nivel de referencia) y crea las indicadoras para las Secciones A y B. Note que en la fórmula R los nombres de las variables `X1` y `SEC` se separan con el símbolo `+`.

```
modelo2=lm(Y~X1+SEC)
summary(modelo2)
```

Código R 2.8. Gráfico de dispersión con las rectas ajustadas en el modelo 2, según cada sección A, B y C. De nuevo, observe el uso de las funciones `plot()`, `lines()`, `legend()`, cómo con los argumentos `pch=as.numeric(SEC)`, `col=as.numeric(SEC)` se identifican los puntos de acuerdo a los niveles de la variable `SEC` (la sección) y cómo se separan de los valores ajustados (obtenidos con la función `fitted()`) y de la variable `X1` aquellos que corresponden a cada sección para trazar las respectivas rectas ajustadas.

```
win.graph()
plot(X1,Y,pch=as.numeric(SEC),col=as.numeric(SEC),xlab="X1",ylab="Y",cex=1.5,cex.lab=1.5)
legend("topleft",legend=levels(SEC),pch=1:3,col=1:3,cex=1.5) #Leyenda colocada es esquina superior izquierda
lines(X1[SEC=="C"],fitted(modelo2)[SEC=="C"],col=1,lty=1,lwd=3) #Recta ajustada en la Seccion C
lines(X1[SEC=="A"],fitted(modelo2)[SEC=="A"],col=2,lty=2,lwd=3) #Recta ajustada en la Seccion A
lines(X1[SEC=="B"],fitted(modelo2)[SEC=="B"],col=3,lty=3,lwd=3) #Recta ajustada en la Seccion B
```

Código R 2.9. Gráficos de residuos, de probabilidad normal y test Shapiro Wilk para evaluar supuestos en el modelo 2. Observe el uso de la función `plot()` en los gráficos de residuos vs. valores ajustados de la respuesta y de residuos versus valores del predictor X_1 , y cómo con los argumentos `pch=as.numeric(SEC)`, `col=as.numeric(SEC)` se identifican los puntos de acuerdo a los niveles de la variable SEC (la sección). También observe el uso de la función `abline()`, para trazar sobre estos gráficos las líneas horizontales identificando los límites $\pm 2\sqrt{MSE}$ para identificar residuos de posibles observaciones atípicas. Tenga en cuenta que debido a que el predictor SEC es un factor, el gráfico de residuos vs. los niveles de esta variable es realizado mediante la función `stripchart()` en lugar de la función `plot()`.

Nota 2.14. Los límites $\pm 2\sqrt{MSE}$ son obtenidos mediante `-2*summary(modelo2)$sigma` y `2*summary(modelo2)$sigma`

```
#Límites para identificar obs atípicas en los graficos de residuos
lim1=-2*summary(modelo2)$sigma
lim2=2*summary(modelo2)$sigma

#Rango para eje vertical de graficos de residuos
minres=min(lim1,residuals(modelo2),lim2)
maxres=max(lim1,residuals(modelo2),lim2)

#Residuales vs. valores ajustados, con representacion de las secciones
win.graph()
plot(fitted(modelo2),residuals(modelo2),ylim=c(minres,maxres),pch=as.numeric(SEC),col=as.numeric(SEC),cex=1.5)
abline(h=c(lim1,0,lim2),lty=2)
legend("topright",legend=levels(SEC),pch=c(1:3),col=c(1:3),cex=1.5)

#Residuales vs. predictor cuantitativo X1, con representacion de las secciones
win.graph()
plot(X1,residuals(modelo2),ylim=c(minres,maxres),pch=as.numeric(SEC),col=as.numeric(SEC),cex=1.5)
abline(h=c(lim1,0,lim2),lty=2)
legend("topright",legend=levels(SEC),pch=c(1:3),col=c(1:3),cex=1.5)

#Residuales vs. niveles predictor cualitativo SEC
win.graph()
stripchart(residuals(modelo2)~SEC,vertical=TRUE,ylim=c(minres,maxres),pch=c(1,2,3),col=c(1,2,3),cex=1.5)
abline(h=c(lim1,0,lim2),lty=2)

#Grafico de probabilidad normal sobre residuos modelo 2, con representacion de las secciones
win.graph()
qqnorm(residuals(modelo2),pch=as.numeric(SEC),col=as.numeric(SEC),cex=1.5)
qqline(residuals(modelo2),col=2)
legend("topleft",legend=levels(SEC),pch=c(1:3),col=c(1:3),cex=1.5)

#Test de normalidad sobre residuos del modelo 2
shapiro.test(residuals(modelo2))
```

Código R 2.10. Calculo de las predicciones puntuales y por intervalos del 95 % en los puntos ($X_1 = 6$, Sección A), y ($X_1 = 11$, Sección C). Observe que se define un objeto `data.frame` de nombre `datosnuevo` con los valores para los predictores X_1 y SEC en los puntos a predecir, manteniendo el mismo nombre R y tipo de las variables predictoras guardadas en el objeto inicial `datos`. Las predicciones son obtenidas mediante la función `predict()`; observe sus argumentos.

```
#Tabla de valores de los predictores en las nuevas observaciones
datosnuevo=data.frame(X1=c(6,11),SEC=factor(c("A","C")))
datosnuevo

#Calculando las predicciones puntuales y por IPs del 95%
predict(modelo2,newdata=datosnuevo,interval="prediction",level=0.95)
```

Código R 2.11. Desanclando (ocultando) con la función `detach()` las variables guardadas en el `data.frame` de nombre `datos`, que fueron inicialmente disponibilizadas con `attach()`.

```
detach(datos)
```

Bibliografía

- Kutner, M. H., Nachtsheim, C. J., Neter, J. and Li, W. (2005). *Applied Linear Statistical Models*, 5th ed. McGraw-Hill Irwing, New York.
- Montgomery, D. C., Peck, E. A. and Vining, G. G. (2021). *Introduction to Linear Regression Analysis*, 6th ed. Wiley, New Jersey.