Estadística multivariada: Regresión y ANOVA

Versión PDF

II-1123 Estadística para Ingeniería Industrial II

Steven García Goñi

steven.garciagoni@ucr.ac.cr

14 de marzo de 2026

Agenda

- Preguntas generadoras
- Introducción
- Fuentes de datos
- Pasos en una regresión
- ANOVA
- Multicolinealidad
- Interpretación

Preguntas generadoras

- ¿Qué es una regresión?
- ¿Por qué son importantes los supuestos de aplicación de regresión?
- ¿Qué relación guardan los estadísticos de regresión con lo aprendido en intervalos de confianza?
- ¿Cómo se relacionan ANOVA y regresión?
- ¿Qué es multicolinealidad?

Regresión Lineal Simple (RLS)

Introducción

- > **All models are wrong, but some are useful - *George E. P. Box***

- El análisis de regresión es una técnica estadística para investigar y modelar la relación entre variables.

- Las aplicaciones de las regresiones son numerosas y se presentan prácticamente en todos los campos.

Regresión

- En su forma más simple se representa por:
- $$y = \beta_0 + \beta_1 x + \varepsilon$$

  — lo cual corresponde a un modelo de regresión lineal simple.
- Por convencionalismo, a x_i se le conoce como variable **independiente**, predictor o regresor.
- Por otro lado, y es conocida como variable **dependiente** o respuesta.

Regresión

- Otra forma de escribir el modelo anterior es en función de la esperanza matemática.
- $$E(y|x) = E(\beta_0 + \beta_1 x + \varepsilon)$$
$$\rightarrow \hat{y} = \beta_0 + \beta_1 x$$

  — Note que desaparece el término aleatorio ε, la razón se explica más adelante.

- Donde β_0 es el intercepto y β_1 es la pendiente.

Regresión

- La diferencia entre los valores observados de y y el resultado esperado de la regresión $\hat{y}$ se llama error o residuo ($e = y - \hat{y}$), el cual estima al error aleatorio ε.
- El error ε se supone que debe seguir una distribución normal con media cero ($\mu = 0$) y desviación estándar constante σ.
- $$\varepsilon \sim N(0, \sigma^2)$$

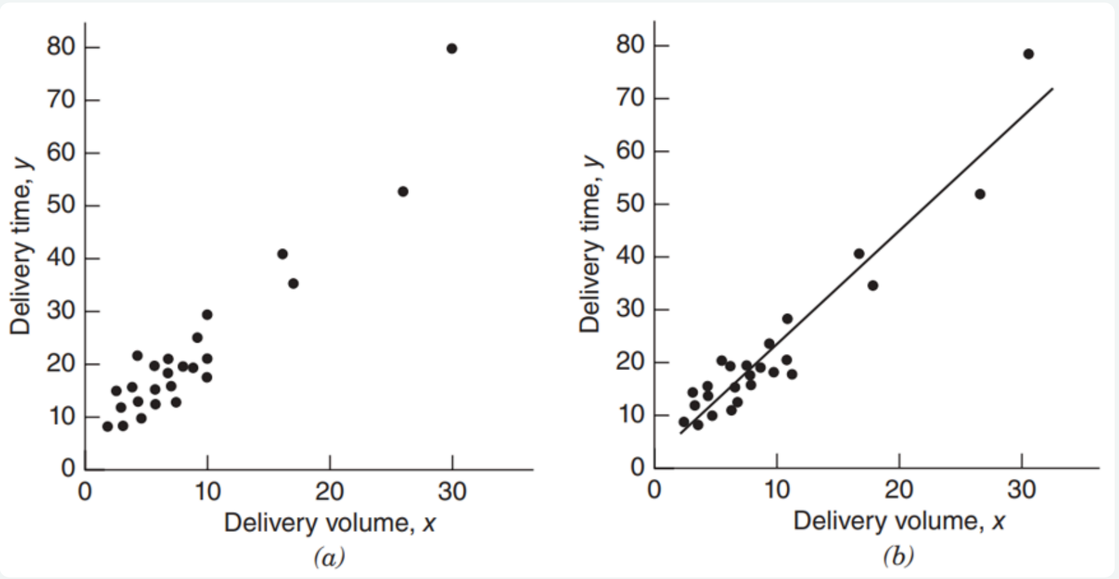
  — Note que la esperanza de una distribución normal es igual a μ y que en este caso $\mu = 0$, por eso desaparece el término ε en el slide anterior.

- El **no cumplir con este requisito tiene sus consecuencias**. Más adelante lo estudiaremos.

Regresión

- El error (e) explica por qué **no todos los puntos** caen sobre la línea recta en la imagen de la derecha. Lo cual corresponde a la dispersión observable.

Imagen tomada de Montgomery et al.(2012)



Regresión

- Cuando hay más de un regresor, el modelo se conoce como **regresión lineal múltiple**.
- $$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \varepsilon$$
- El objetivo de un análisis de regresión es determinar los parámetros desconocidos en el modelo de regresión.

  — ¿Cuáles son esos parámetros? Pues β_i.

Regresión lineal simple (RLS)

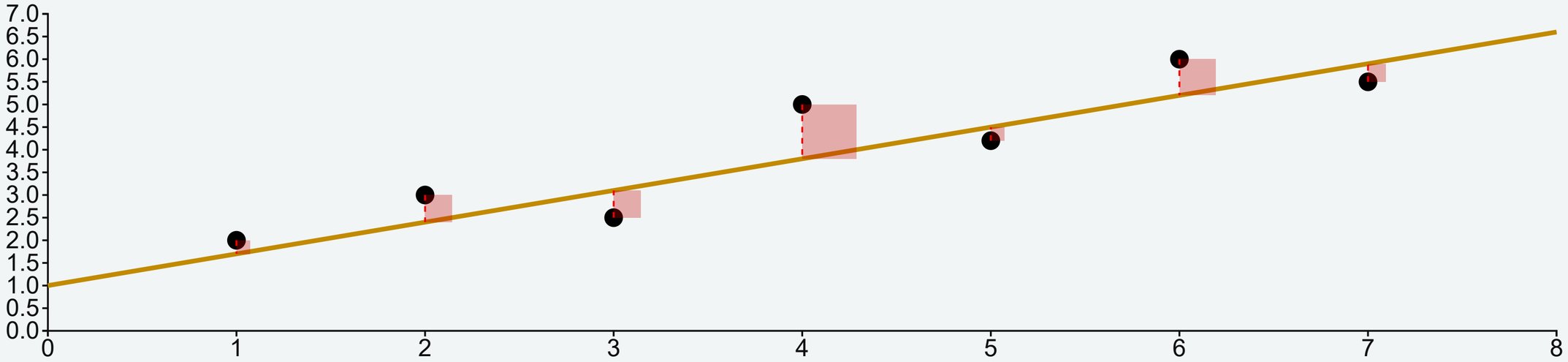
Estimación de β_i

- El método que se emplea para estimar una regresión lineal (simple o múltiple) es el de mínimos cuadrados ordinarios (**MCO**). Se parte del hecho de que los β_i son desconocidos y se estiman a partir de una muestra.

  — MCO busca minimizar $\sum e_i^2$, en el siguiente slide hay una representación visual de esta ecuación.

- > La calidad de los modelos depende de la calidad de los datos, ya sean de fuentes secundarias o primarias. Pueden existir sesgos de medición, omisión de variables o correlaciones espurias.

Estimadores MCO - Explicación visual



Intercepto β_0 Pendiente β_1

SSE = 3.14

Inspirado por: **setosa.io — Ordinary Least Squares Regression**

Estimadores MCO

- Para una regresión lineal simple de la forma $\hat{y} = \beta_0 + \beta_1 x$

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}}$$
$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

- Donde:

$$S_{xx} = \sum_{i=1}^{n} (x_i - \bar{x})^2$$
$$S_{xy} = \sum_{i=1}^{n} (x_i - \bar{x}) \cdot (y_i - \bar{y})$$

Estimadores MCO - Forma matricial

- El modelo de regresión lineal simple se puede escribir matricialmente de la siguiente forma

$$y = X\beta + \varepsilon$$

- Y los estimadores MCO se pueden obtener de forma matricial empleando

$$\hat{\beta} = (X^T X)^{-1} X^T y$$

Ejemplo 01

- La resistencia al cizallamiento (*deformación mecánica producida por fuerzas paralelas, opuestas y de igual magnitud que actúan sobre un material, haciendo que sus capas internas se deslicen entre sí*) de la unión entre los dos tipos de propelentes es una característica importante de calidad en el ensamble de un motor. Se sospecha que la fuerza de cizallamiento está relacionada con la edad, en semanas, del lote de materia prima.
- Para ello, se han recogido 20 observaciones sobre la resistencia al cizallamiento y la edad de la materia prima.

  — Puede acceder a los datos y la solución de este y otros ejemplos en este **Excel**.

Ejemplo 01

Observación	Resistencia(y)	Edad(x)	Observación	Resistencia(y)	Edad(x)
1	2158.70	15.50	11	2165.20	13.00
2	1678.15	23.75	12	2399.55	3.75
3	2316.00	8.00	13	1779.80	25.00
4	2061.30	17.00	14	2336.75	9.75
5	2207.50	5.50	15	1765.30	22.00
6	1708.30	19.00	16	2053.50	18.00
7	1784.70	24.00	17	2414.40	6.00
8	2575.00	2.50	18	2200.50	12.50
9	2357.90	7.50	19	2654.20	2.00
10	2256.70	11.00	20	1753.70	21.50

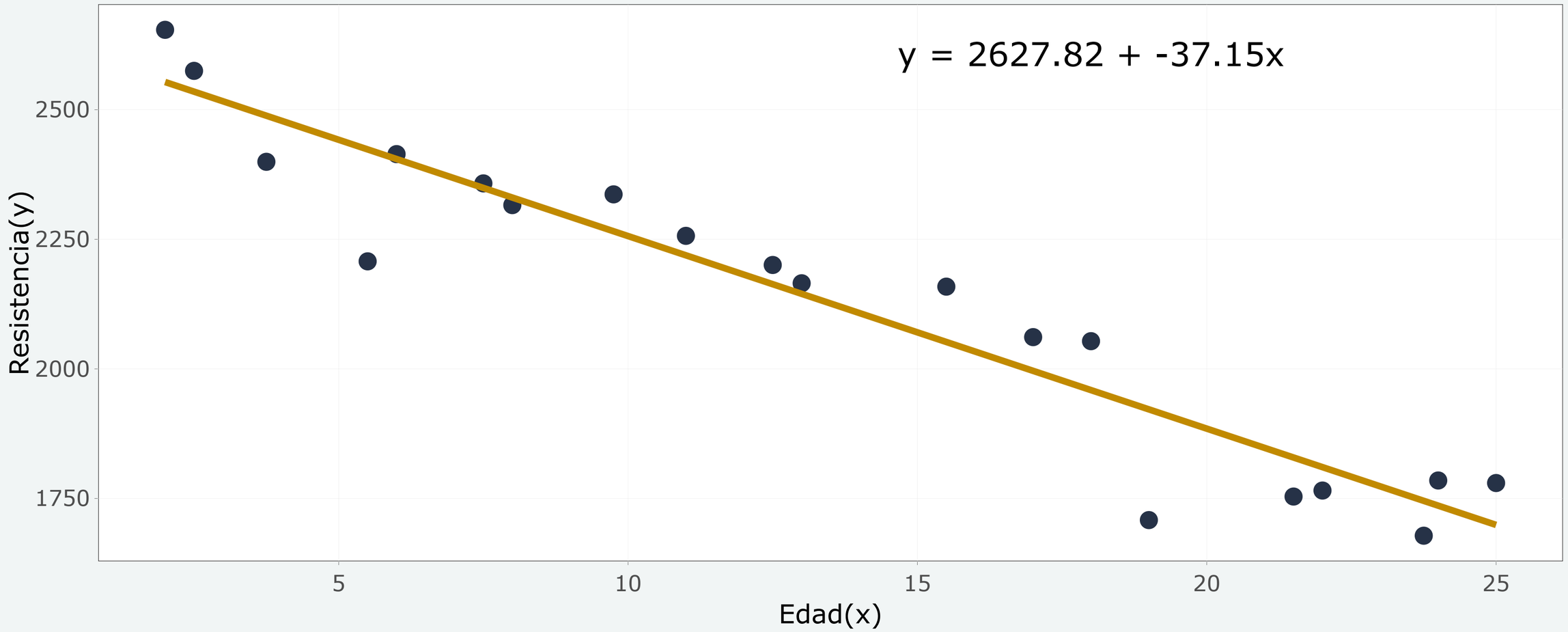
Actividad en grupos

- Estime la recta de regresión con MCO, si conoce de álgebra de matrices, resuélvalo también de esa forma.
- Respuesta:

$$\hat{y} = 267.82 - 37.15x$$

- Obtenga, además, un gráfico de dispersión.
  - En Excel, el gráfico de dispersión permite obtener la ecuación de regresión

Ejemplo 01



Ejemplo 01

- Ahora, para el mismo conjunto de datos, obtenga una estimación de la respuesta, es decir $\hat{y}$.
- Con base en esta, estime el error (también conocido como **residuos**) y obtenga la suma de todos los residuos.

—

$$e = y - \hat{y}$$

- En teoría este valor debe ser cero o aproximadamente cero.De forma matricial, lo puede obtener de esta manera:

—

$$\hat{\varepsilon} = y - X\hat{\beta}$$

Regresión

- Así como se han obtenido los valores de la esperanza $E(y|x)$, también se pueden obtener los valores de la varianza $V(y|x)$.
- En adición a la estimación de β_i, la estimación de σ^2 es necesaria para probar un contraste de hipótesis.

Regresión

- Se introduce el concepto de suma de cuadrados, por ejemplo, la suma de cuadrados del error (SS_e) es:
- $$SS_e = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} e_i^2$$
- La varianza es, entonces, $s^2 = \frac{SS_e}{df_e} = MS_e$, donde df son grados de libertad y MS_e es el cuadrático medio del error. Tome en cuenta que los términos SS, MS y df son acrónimos en inglés. Sus equivalentes al español son SC, CM y gl.

Ejemplo 01

- Obtenga, para los datos del ejemplo 01, la SS_e.

  — $166\,254.86$

- Obtenga además el valor de s^2

  — $9\,236.38$

- Y el valor de s

  — 96.11

- > Resuelva también con álgebra matricial.

Otras sumas de cuadrados

- La suma de cuadrados totales
- $$SS_T = \sum_{i=1}^{n} y_i^2 - n\bar{y}^2$$

  — Esto es matemáticamente equivalente a $SS_T = \sum_{i=1}^{n} (y_i - \bar{y})^2$

- **Ejemplo 01**
- Obtenga la suma de cuadrados totales SS_T para el ejemplo desarrollado.

  — $SS_T = 1\,693\,737.60$

Pruebas de hipótesis e IC

- A menudo estamos interesados en probar hipótesis y construir intervalos de confianza acerca de los parámetros del modelo construido. Para ello vamos a recurrir al ya conocido `t-test` (Prueba t) y se aplica al intercepto y la pendiente. La hipótesis de prueba es

$$H_o : \beta_i = 0$$
$$H_i : \beta_i \neq 0$$

- Siempre o casi siempre se hace de esta manera pues se busca determinar si el coeficiente β_i tiene o no un efecto "verdadero" sobre la variable de respuesta y.

> Nota: tome en cuenta que pueden existir implicaciones prácticas que no estén siendo detectadas por pruebas estadísticas.

Pruebas de hipótesis e IC

- Cuando se rechaza la hipótesis nula se dice que el **coeficiente es significativo**. Recuerde además, que, para la media:
- Intervalo de confianza:

  — $$\theta \pm EE \cdot t_{\frac{\alpha}{2}}$$

- Prueba de hipótesis:

  — $$t_0 = \frac{\hat{\beta}_i - \mu}{EE}$$

  — donde μ es típicamente igual a cero ($\mu = 0$).

Error estándar (EE)

- ¿Cómo obtenemos el error estándar (EE)?

  — En inglés SE.

- El error estándar se obtiene como:

  — $$EE_{\beta_1} = \sqrt{\frac{MS_e}{S_{xx}}}$$

  — $$EE_{\beta_0} = \sqrt{MS_e \cdot \left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)}$$

Error estándar (EE)

- De forma **matricial**
- $$EE = \sqrt{[s^2(X^TX)^{-1}]_{jj}}$$
- Para facilitar la interpretación, se aclara que el jj se refiere a la diagonal de la matriz resultante. El tamaño de la diagonal debe ser igual a la cantidad de coeficientes β_i.

Ejemplo 01

- Continue con el ejemplo 01, usando tanto las fórmulas regulares como las matriciales.

$$EE_{\beta_1} = 2.88$$

$$EE_{\beta_0} = 44.18$$

Hasta el momento hemos construido gran parte de la tabla de la derecha. Obtenga el valor t (t_0) y el Valor P.

Término	gl	Coeficiente	EE	t_0	Valor P
β_0	1	2627.82	44.18	59.47	0
β_1	1	-37.15	2.89	-12.86	0
Error	18	NA	NA	NA	NA

- Con esta información, construya ahora el intervalo de confianza y concluya sobre la prueba y el estudio. Para el ejemplo utilice un 95 % de confianza.

IC	Inferior	Superior
β_0	2535.00	2720.65
β_1	-43.22	-31.08

ANOVA

- Lo que acabamos de llevar a cabo es una prueba de hipótesis **sobre la esperanza matemática** de los coeficientes de regresión, pero también se puede hacer sobre las **estimaciones de la varianza**.
- Esto se conoce como análisis de varianza (ANOVA) y para ello, se retoman los conceptos de suma de cuadrados.

ANOVA

- $$SS_T = SS_R + SS_e$$
- Donde:
- $$SS_R = \hat{\beta}_1 \cdot S_{xy}$$

$$SS_T = \sum_{i=1}^{n} y_i^2 - n\bar{y}^2$$

$$SS_e = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} e_i^2$$

ANOVA

- Al dividir cada una de las sumas de cuadrados (**SS**) por sus grados de libertad se obtiene el cuadrático medio (**MS**), el cual es una **estimación de la varianza**.
- La prueba de hipótesis que se sigue para dos varianzas es la prueba F
- $$F_i = \frac{CM_i}{CM_e}$$

ANOVA

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F_0
Regression	$SS_{\mathrm{R}} = \hat{\beta}_1 S_{xy}$	1	MS_{R}	$MS_{\mathrm{R}}/MS_{\mathrm{Res}}$
Residual	$SS_{\mathrm{Res}} = SS_{\mathrm{T}} - \hat{\beta}_1 S_{xy}$	$n-2$	MS_{Res}	
Total	SS_{T}	$n-1$		

Ahora, construya la tabla ANOVA para el ejemplo 01.

Término	GL	SS	MS	F	Valor P
Regresión	1	1527482.7	1527482.740	165.38	0
Residuos	18	166254.9	9236.381	NA	NA
Total	19	1693737.6	89144.084	NA	NA

¿Qué tan bueno es el modelo?

- En regresión lineal múltiple se avanza más en este concepto: bondad de ajuste de una regresión.

  — Como se interpreta y los cuidados que hay que tener.

- No obstante, aquí tenemos un indicador de bondad de ajuste, el R^2, el cual expresa la proporción de variabilidad que está siendo explicada por el modelo.

- $$R^2 = \frac{SS_R}{SS_T} = \frac{1\,527\,482.74}{1\,693\,737.60} = 0.9018 = 90.18\%$$

Entonces, ¿Qué es ANOVA?

- ANOVA es un **caso especial de la regresión**, son *básicamente* lo mismo.

  — ANOVA puede expresarse como un modelo de regresión con variables categóricas.

- En ANOVA los predictores suelen llamarse factores.
- En lugar de aplicarse una prueba t, se aplica una prueba F.

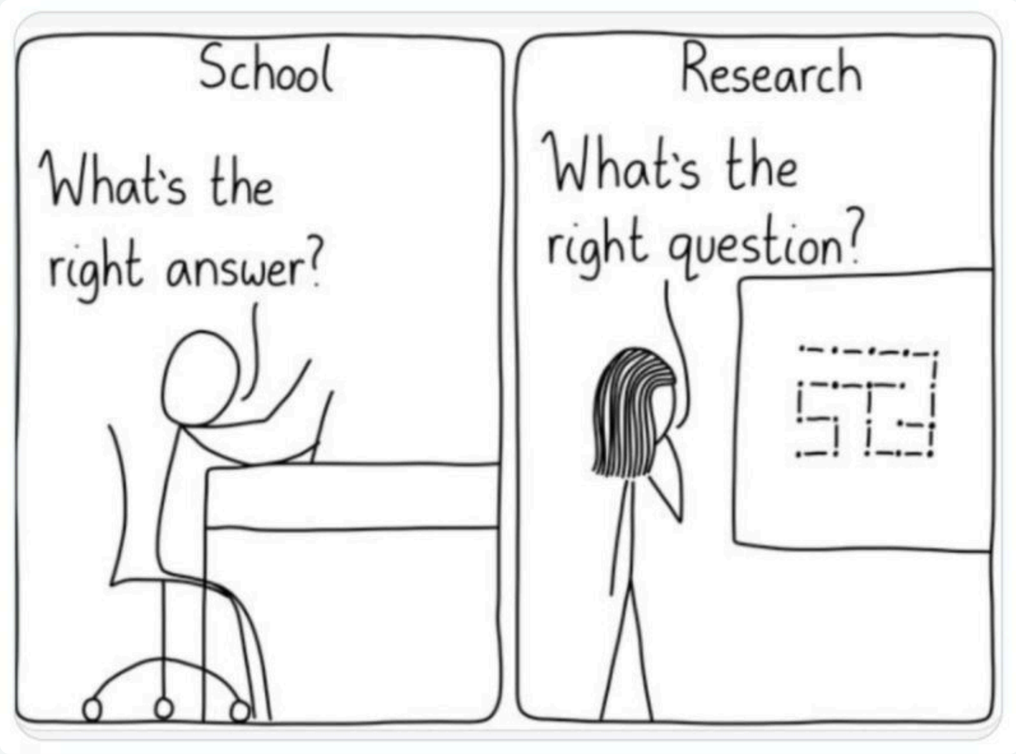
  — Ambas, t y F están íntimamente relacionadas con la distribución normal.

- ANOVA típicamente tiene más utilidad cuando el/los predictores son categorías.

¿Cuándo usar regresión o ANOVA?

- > Basado en el material de **PhD. Rosana Ferrero**

- Cada vez son más importantes las preguntas que se hacen que las respuestas que se obtienen, en este sentido, ¿qué está evaluando?.
- Si el interés reside en **comparar grupos**, es más apropiado un ANOVA. Si por el contrario se desea estudiar la **asociación entre variables** es mejor utilizar una regresión.



¿Cómo se relacionan?

- Cuando los predictores son categóricos, ANOVA y regresión son “matemáticamente idénticos” (en términos de las inferencias que se pueden obtener del estadístico de prueba).
- Cuando son continuos, ANOVA es una medida de la **significancia de la regresión**.
- Prácticamente cualquier software va a presentar regresión y ANOVA en un mismo reporte.
  — Por ejemplo, Minitab presenta primero la regresión y por debajo un ANOVA.
- En el caso del Ejemplo 01, como el predictor es continuo podemos contrastar la hipótesis de que la regresión es significativa
  — Como el valor P es ~ 0, se puede decir que la regresión es significativa a prácticamente cualquier nivel de confianza.

Utilidad

- Para sacar el máximo provecho de ANOVA como caso especial de regresión debemos avanzar hacia la regresión lineal múltiple.
- Se necesita entender, además, el concepto de variable *dummy*.

Variables categóricas (*dummy*)

- Es importante comprender que regresión y ANOVA solo "trabajan" con variables numéricas y **no con categorías**.
- Las variables *dummy* solucionan este problema, convirtiendo las n categorías en $n - 1$ variables cuantitativas binarias.
- Son $n - 1$ porque de incluirse el regresor x_n se generaría una combinación lineal, como las estudiadas en álgebra. Y eso haría a la regresión colineal.

  — El concepto de colinealidad se abordará luego.

Ejemplo 02

- Suponga una situación en la que se requiere evaluar la resistencia de dos tipos distintos de telas.
- Las telas son A y B, para A los resultados son:
  - A: 8, 9, 7, 10, 6, 8, 9
  - B: 5, 6, 5, 4, 6, 5, 7
- Realice una regresión y un ANOVA para estas variables
- Los datos y la solución se encuentran en el mismo **Excel** anterior.

Ejemplo 02 - Codificación

- El primer paso es convertir las variables categóricas a “continuas”. Para ello hay que escoger una categoría de referencia que NO va a formar parte de la regresión. Por conveniencia muchos software seleccionan **la primera en orden alfabético**.
- Una vez hecho esto, resuelva siguiendo la misma secuencia de pasos del ejemplo 01.

Ejemplo 02

Datos

Tela(x)	Resistencia(y)
A	8
A	9
A	7
A	10
A	6
A	8
A	9
B	5
B	6
B	5
B	4
B	6
B	5
B	7

Variables codificadas

Tela(x)_A	Tela(x)_B	Resistencia(y)
1	0	8
1	0	9
1	0	7
1	0	10
1	0	6
1	0	8
1	0	9
0	1	5
0	1	6
0	1	5
0	1	4
0	1	6
0	1	5
0	1	7

Eliminación de la colinealidad

Tela(x)_B	Resistencia(y)
0	8
0	9
0	7
0	10
0	6
0	8
0	9
1	5
1	6
1	5
1	4
1	6
1	5
1	7

Ejemplo 02

Regresión

Término	Estimador	EE	Valor T	Valor P
(Intercept)	8.142857	0.4441609	18.333124	0.0000000
`Tela(x)`B	-2.714286	0.6281384	-4.321159	0.0009941

ANOVA

Término	GL	SS	MS	Valor F	Valor P
`Tela(x)`	1	25.78571	25.785714	18.67241	0.0009941
Residuals	12	16.57143	1.380952	NA	NA

Interpretación

- En este caso β_1 representa a la tela B. Entonces se puede decir que cuando se use la tela B se espera una reducción en el **promedio** de la resistencia de 2.71, pues el estimador es negativo.
- En este sentido A, es una **categoría de referencia** que toma el valor de cero (0).
  - Promedio de $A = 8.1429 - 2.7143 \cdot 0 = 8.1429$
  - Promedio de $B = 8.1429 - 2.7143 \cdot 1 = 5.4286$
- Observe el rombo en el siguiente gráfico, este corresponde a la media por cada tela, si está viendo esta presentación en su versión web, pose el cursor sobre el mismo y contraste los valores mostrados contra los obtenidos.

Interpretación

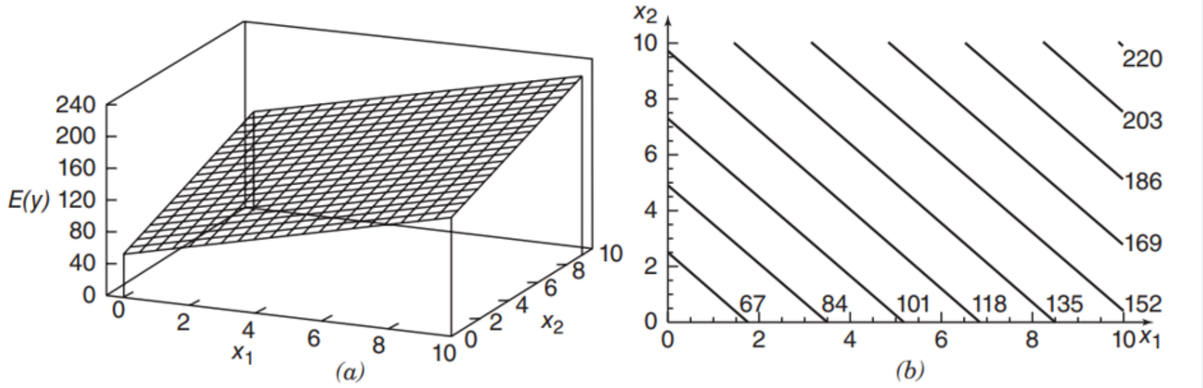
Regresión lineal múltiple (RLM)

Regresión lineal múltiple

Cuando un modelo de regresión involucra más de una variable predictora. Por ejemplo de la forma:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \varepsilon$$

Tomado de Montgomery et al.(2012)



Interacciones

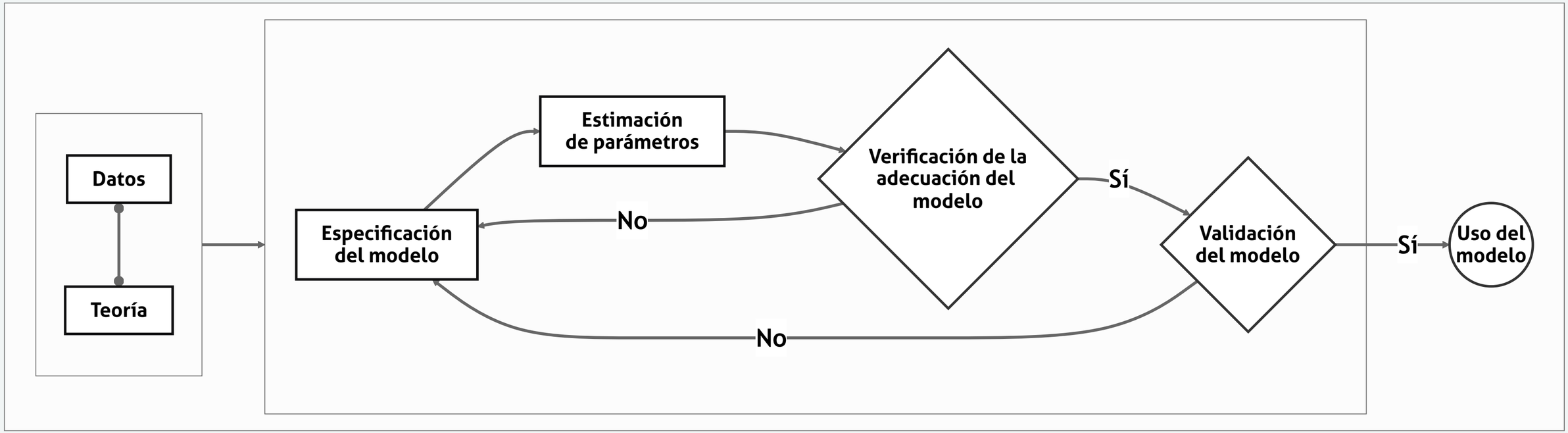
- Es importante que tome en cuenta que los modelos de regresión lineal múltiple también pueden incluir interacciones, de la forma:

- $$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \varepsilon$$

- Incluir interacciones puede ayudar a mejorar la precisión de los modelos, pero éstas deben de tener un sentido, enmarcado en un contexto.

Pasos de una RLM

Pasos en la construcción de una RLM



Para ver el diagrama en pantalla completa, diríjase a este **enlace**.

Tomado y adaptado de Montgomery et al. (2012).

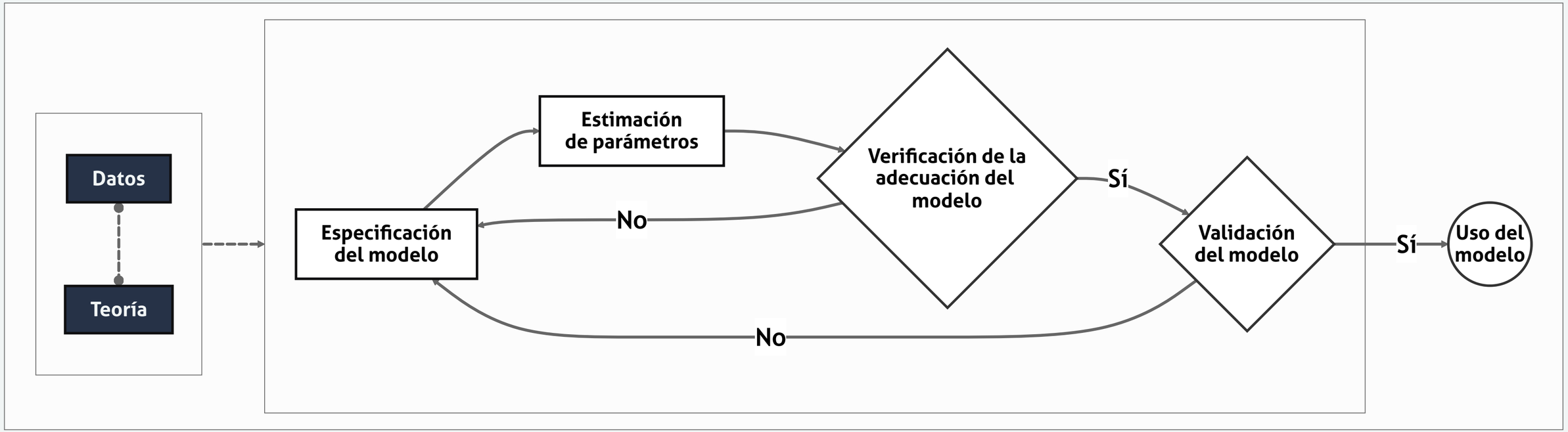
Importante

- A partir de este punto el abordaje de las técnicas se realiza con apoyo de software estadístico.
- La lógica de las fórmulas es la misma, pero la complejidad de cálculo aumenta.
- Cada paso de la secuencia anterior será abordado mediante el Ejemplo 03, que se muestra en el siguiente slide.

Ejemplo 03

- Un ingeniero industrial está modelando el tiempo de entrega (y) en minutos de una bebida en diferentes locales de un centro comercial según el número de cajas a entregar (x_1) y la distancia en ft (x_2) que tiene que recorrer la persona.
- La intención con este ejemplo, es crear una regresión lineal múltiple desde el inicio.
- La base de datos se encuentra en este **Excel**.

Datos + Teoría



Tomado y adaptado de Montgomery et al. (2012).

Para ver el diagrama en pantalla completa, diríjase a este **enlace**.

Datos + Teoría

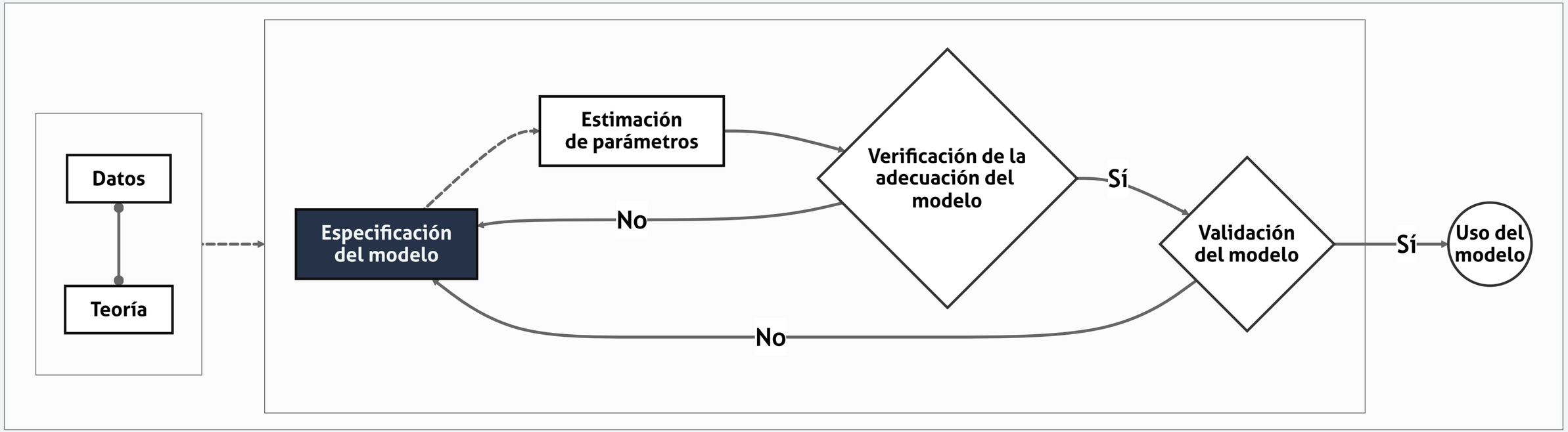
- Resaltan dos aspectos importantes, la entrada del proceso es datos + teoría. Y esto es un detalle VITAL en el proceso estadístico.
- No en todas las ocasiones se le va a proveer un conjunto de datos para que haga la regresión, creer que siempre será así es muy inocente. En muchas ocasiones es usted en su rol como persona ingeniera quien tiene que decidir cuántos y cuáles datos recoger.

    — Por ello la relevancia del tema de muestreo.

¿Por qué es importante?

- La estadística es una **herramienta**, NO un objetivo
- El **contexto** define el propósito de la estadística. Cada uno tiene sus propias necesidades, preguntas y limitaciones y esto influye en cómo se aplican y analizan las técnicas estadísticas.
- Los resultados estadísticos deben interpretarse en función del contexto. Un mismo dato puede tener implicaciones **muy diferentes** dependiendo de la situación.
- Sin contexto, la estadística puede ser malinterpretada o incluso **manipulada** para respaldar conclusiones erróneas.

Especificación del modelo



Para ver el diagrama en pantalla completa, diríjase a este **enlace**.

Tomado y adaptado de Montgomery et al. (2012).

Especificación del modelo

- Con base en los datos y **la teoría**, se especifica el modelo de regresión que se desea estimar.
- Algunas herramientas computacionales le permiten "probar" múltiples modelos, esto, desde luego, es una forma válida de hacer las cosas, pero no se debe obviar que la teoría es la que dicta si un resultado tiene sentido o no.

  — Además, para probar esto, debe haber recolectado los datos. ¿Recolectaría datos sin estar seguro? ¿Invertiría dinero en medir variables que no va a usar?

- Por ejemplo, nos podríamos preguntar si es necesario (y lógico según la teoría) que dos variables interactúen.En este caso, podríamos querer un modelo de este tipo:

- $$y_{ij} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + e_{ij}$$

Especificación del modelo

- Nótese que el diagrama que se le provee es **iterativo** o cíclico, es decir, no se tiene que *"casar"* con el primer modelo que especifique.
- Dos o tres pasos hacia delante en el diagrama tiene la oportunidad de regresar a la especificación del modelo para mejorar su ajuste.

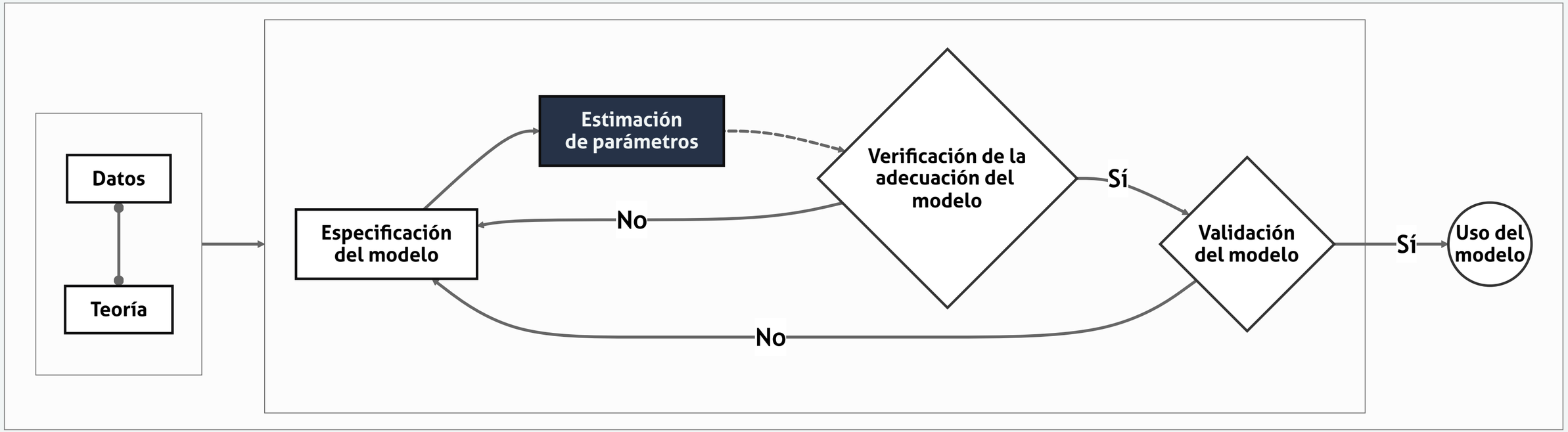
Ejemplo 03 - Especificación

- Para el caso de tiempo de entrega, se especifica un modelo, con base en la **teoría aplicable**, con esta forma:

- $$y[min] = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + e$$

- Donde el objetivo es minimizar la respuesta y x_1 y x_2 son el número de cajas y la distancia recorrida, respectivamente. Se aclara que el min en la ecuación es de minutos, no de minimizar.

Estimación de parámetros



Para ver el diagrama en pantalla completa, diríjase a este **enlace**.

Tomado y adaptado de Montgomery et al. (2012).

Estimación de parámetros

- Los parámetros (β_i) especificados en el modelo **son desconocidos** y deben ser estimados con base en una **muestra**.

  — De nuevo, recuerde las bases de muestreo.

- Ésta estimación se realiza mediante MCO (Mínimos Cuadrados Ordinarios). El detalle de las fórmulas es trivial, ya que su estimación por software está más que extendida.

  — Además, ya fue abordado el caso en RLS.

Una anotación relevante

- Hoy por hoy el método predominante o más utilizado (desde los 70's, formalizados por John Nelder y Robert Wedderburn) para la estimación de parámetros es el de máxima verosimilitud (ML – Maximum Likelihood).
- Los estimadores **ML y MCO son iguales siempre que se cumplan los supuestos** de regresión (que más adelante serán estudiados), principalmente el de normalidad.

  — Maximizar la verosimilitud produce exactamente el mismo estimador que minimizar $\sum e_i^2$. Por ello, MCO sigue válido y ampliamente utilizado.

- No obstante, en Modelos Lineales Generales (GLM), donde el error no es normal (binomial, Poisson, Exponencial, etc.) se utiliza ML para la estimación de los coeficientes de regresión.

Ejemplo 03 - Estimación

- Los coeficientes estimados son:

Término	Estimador	EE	Valor T	Valor P
(Intercept)	2.341	1.097	2.135	0.044
x1	1.616	0.171	9.464	0.000
x2	0.014	0.004	3.981	0.001

- Cuya ecuación de regresión es:

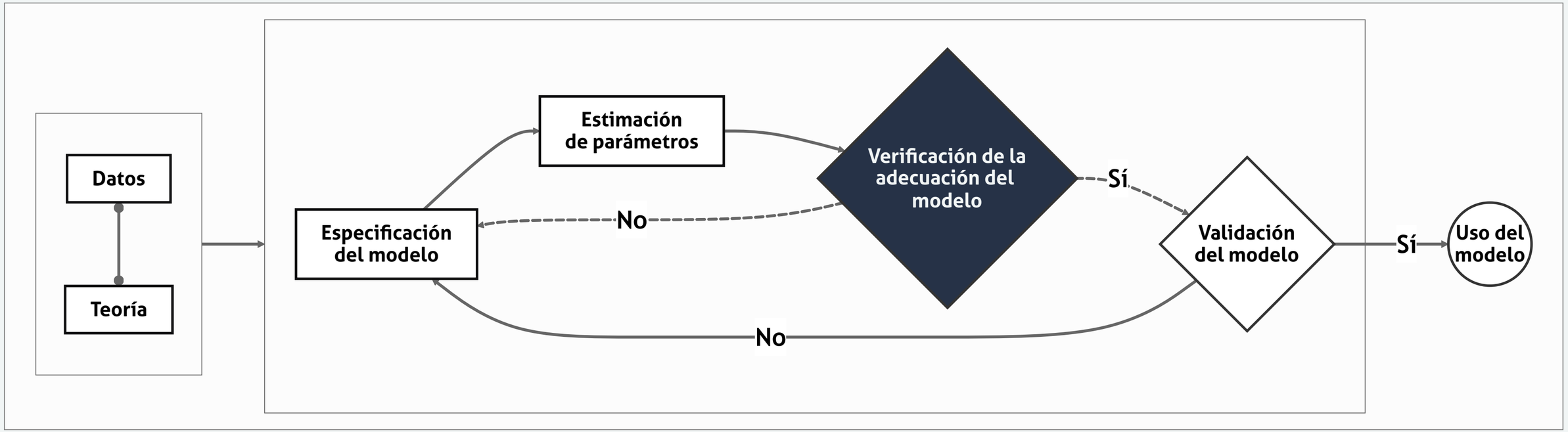
$$\widehat{\mathrm{y}} = 2.341 + 1.616(\mathrm{x1}) + 0.014(\mathrm{x2})$$

Ejemplo 03 - Estimación

- El resultado de ANOVA es:

Término	Estimador	EE	Valor T	Valor P	NA
x1	1	5382.409	5382.409	506.619	0.000
x2	1	168.402	168.402	15.851	0.001
Residuals	22	233.732	10.624	NA	NA

Verificación de la adecuación del modelo



Para ver el diagrama en pantalla completa, diríjase a este **enlace**.

Tomado y adaptado de Montgomery et al. (2012).

Verificación de la adecuación del modelo

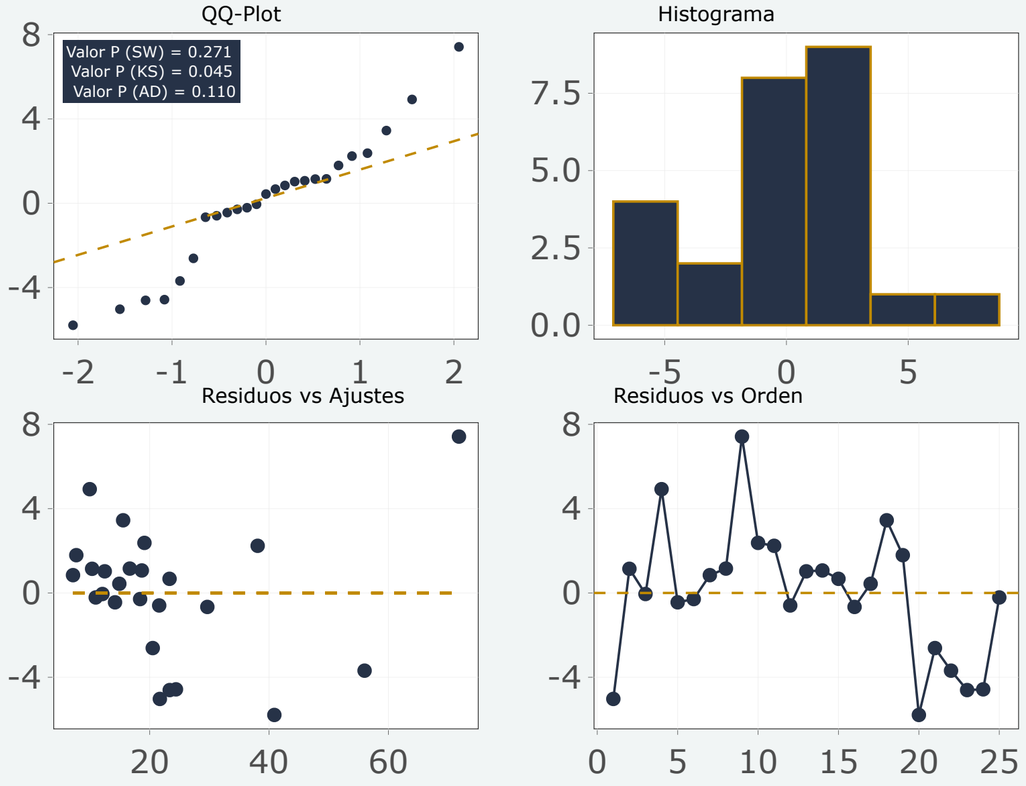
- Este apartado se refiere, entre otras cosas, a varios **supuestos** que se deberían cumplir:
  - La relación entre x y y debería ser **lineal**, al menos aproximadamente.
  - El error debe tener media 0 ($\mu = 0$).
  - El error debe tener varianza constante σ^2 (**Homocedasticidad**)
  - Los errores deben ser **independientes** (no correlacionados)
  - Los errores deben estar **normalmente** distribuidos.

Supuestos

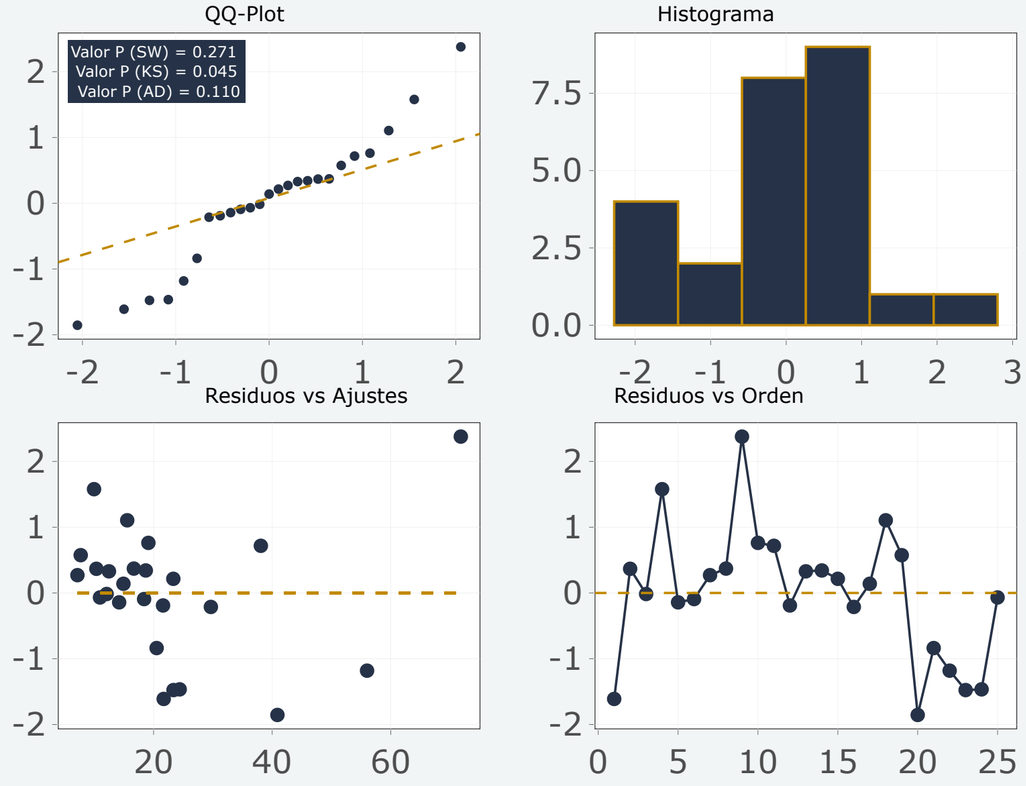
- ¿Qué es un error o un residuo?
  - En simple, es la diferencia entre el valor observado (y) y el valor estimado ($\hat{y}$).
  - En ocasiones puede ser conveniente presentar los **residuales estandarizados**.
    - Estandarizado se refiere a que se le aplica la transformación z.
    - Como es una transformación lineal, los resultado que va a obtener son los mismos, salvo que posiblemente algunos softwares o algoritmos grafiquen histogramas ligeramente diferentes.
  - Esto ayuda a detectar datos extraños (fuera de 3 desviaciones), basándose en la distribución normal.
- Una forma de verificar los supuestos es mediante pruebas gráficas.
  - Y se recomienda empezar por esto.

Ejemplo 03 - Supuestos

Residuos regulares



Residuos estandarizados



De forma gráfica

Normalidad

- Se puede observar en el QQ-plot y en el histograma

Los residuales deben seguir la recta del gráfico de probabilidad (o QQ-plot). El histograma debe tener la forma de una distribución normal.

Homocedasticidad

- Se puede observar en el gráfico de residuos vs ajustes.

Los residuales graficados contra el valor ajustado no deben presentar patrones, como formas de trompeta, en cualquier dirección o ambas de forma simultánea. Sino que debe observarse un comportamiento aleatorio.

Independencia

- Se puede observar en el gráfico de residuos vs orden.

Los residuales graficados contra el orden en que se tomaron no deben presentar patrones con forma de tendencia. Sino que debe observarse un comportamiento aleatorio.

De forma analítica

Normalidad

También existen las pruebas analíticas. Algunas de ellas son:

- Anderson-Darling
- Ryan-Joiner
- Shapiro-Wilk
- Cramer-Von Mises
- Kolmogorov-Smirnov

Recuerde la clase de bondad de ajuste para distribuciones.

Homocedasticidad

- Bartlett
- Levene
- Comparaciones múltiples
- Breush-Pagan

Independencia

- Breusch-Godfrey
- Durbin-Watson

Importante

- Las pruebas analíticas **no son de aplicación indiscriminada**. Existen condiciones bajo las cuales interesaría aplicar una u otra. Cuando se vaya a aplicar una de estas **se debe entender** los motivos por los cuales fue seleccionada.
- En ocasiones es más que suficiente con la forma gráfica.

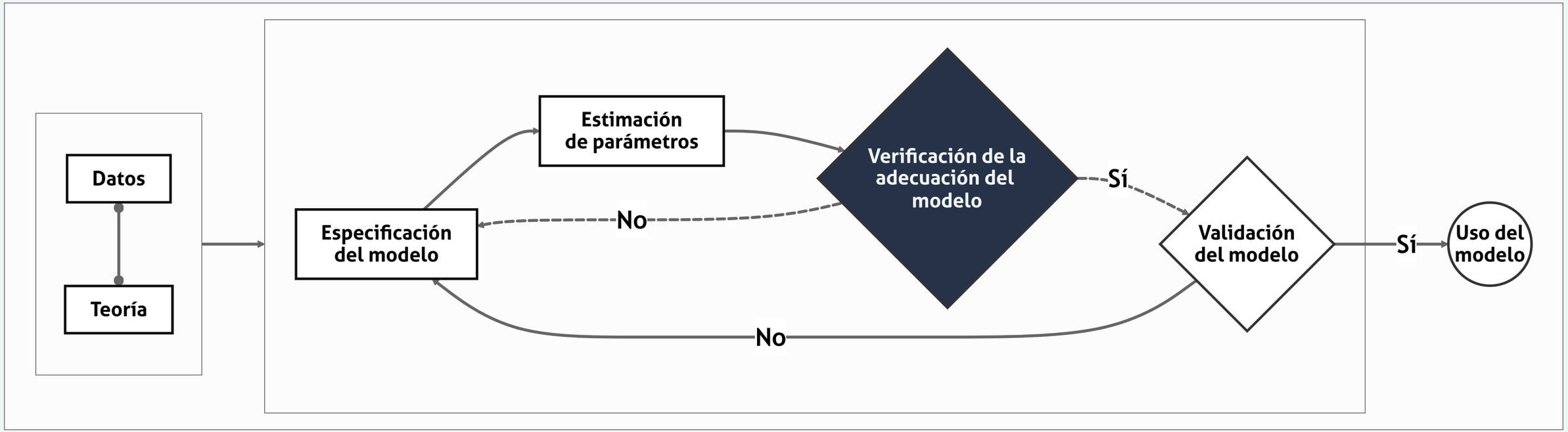
¿Por qué son importantes los supuestos?

Primero ...

- Hay que comprender que es relativamente común no cumplir con los supuestos de aplicación de regresión y ANOVA.
- La forma más sencillo de corregirlo, es entender lo que provoca el incumplimiento.
- Antes de hacer algún cambio drástico, **puede** ser suficiente con volver a especificar el modelo (agregar o quitar términos en la ecuación).

  — Recuerde el diagrama de pasos de regresión. Si el modelo no es adecuado, se busca especificar otro modelo. No debería ser su primera opción tratar de cambiar los datos.

Verificación de la adecuación del modelo



Para ver el diagrama en pantalla completa, diríjase a este **enlace**.

Tomado y adaptado de Montgomery et al. (2012).

¿Por qué se deben cumplir los supuestos?

- En el caso de la **normalidad**, no cumplir con este supuesto es un indicativo de que la esperanza matemática de los residuales no es ya, necesariamente cero ni con varianza σ^2.
- Cuando este supuesto no se cumple es una señal de que el modelo ajustado "no es bueno" en ciertos intervalos, por ejemplo en las colas, donde suelen desviarse los valores de la línea teórica central.
- A la postre, **el incumplimiento de la normalidad suele afectar la estimación del error estándar (SE)**.

  — Recuerde que el SE o EE es vital en el cálculo del Valor P, si el EE se calcula mal, entonces el valor P también.

¿Por qué se deben cumplir los supuestos?

- Como medida remedial, primero se debe investigar la causa del incumplimiento y corregirlo de ser posible (desde la forma en que se recolectaron los datos, hasta la especificación del modelo). Es importante saber determinar si de los datos y la teoría se espera que NO HAYA un comportamiento normal, lo cual está bien que así sea.
- Puede recurrir (*con mucha cuidado y prácticamente como última opción*) a transformaciones no lineales como Box-Cox (tema de otro curso o de estudio individual si se requiriese).
- También existen los modelos lineales generales (la regresión que estudiamos es uno de estos modelos), aunque con mayor dificultad de cálculo, se prefieren sobre las transformaciones.

¿Por qué se deben cumplir los supuestos?

- Interpretar con cautela:

- > *Por lo general, si los demás supuestos se cumplen, el tamaño de muestra es adecuado y la recolección de datos es adecuada, los resultados no se ven afectados sustancialmente por desviaciones* **ligeras** *a la normalidad.*

- Por eso no vale la pena, en regresión, penalizar en exceso el incumplimiento de la normalidad.

¿Por qué se deben cumplir los supuestos?

- En el caso de la **homocedasticidad**, al incumplirla habría consecuencias en el método de estimación de MCO. Los estimadores de los coeficientes seguirían siendo insesgados, pero **la estimación de los errores estándar de estos parámetros no sería válida**.

  — Puede notarse al ver la fórmula matricial estudiada en RLS.

- Por esta razón los intervalos de confianza construidos y por tanto el valor P no se pueden interpretar normalmente.
- Como medida remedial se pueden seguir los mismos consejos que con normalidad.

¿Por qué se deben cumplir los supuestos?

- Si los residuales están relacionados entre si (autocorrelación), no se cumpliría con el supuesto de **independencia**. En este caso los estimadores MCO siguen siendo insesgados pero los errores estándar son inconsistentes.
- Incumplir con este supuesto es el "peor" de los escenarios, ya que es el más difícil de solucionar y en ocasiones es imposible.
- Una medida remedial es recurrir a modelos de auto-regresión y análisis de series de tiempo.

  — Pero solo es válida si el estudio es intrinsecamente dependiente, no subsana errores de diseño que provoquen este incumplimiento.

Recordemos

- Ahora bien, tratemos de entender a que se refiere con que no se puede interpretar bien al valor P.
- En simple, cada coeficiente de la regresión está siendo sometido a una prueba t.
- $$\theta \pm EE \cdot t_{\frac{\alpha}{2}}$$
- El valor P, depende del EE.
- Por eso, tanto en regresión como en ANOVA no se debe leer solo el valor P al final de la tabla. Sino que se analiza cada parte como un elemento que conforma un todo.

Ejemplo comparativo

- Sin entrar en detalles de que es un modelo Poisson, se aclara que este es el correcto.

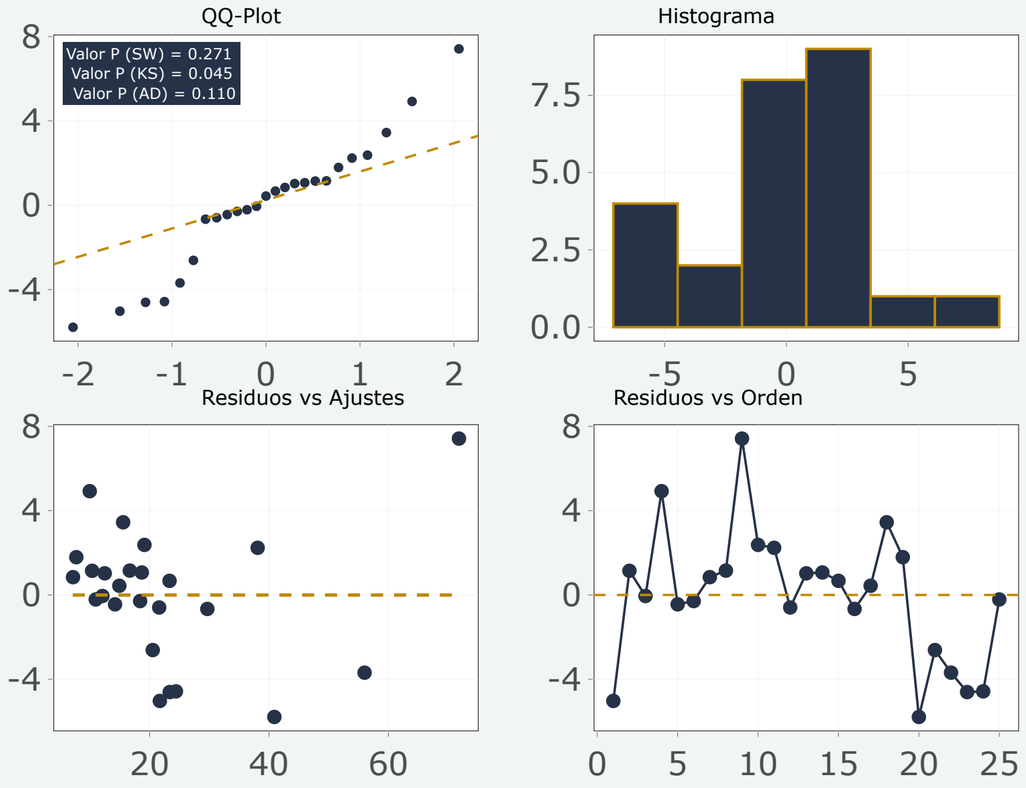
  — Este un perfecto ejemplo de que es un modelo lineal general.

- En el primer modelo, la regresión que estamos estudiando, no está cumpliendo los supuestos.
- Note la diferencia entre los EE y en consecuencia en los valores P. **Los coeficientes se estiman razonablemente bien**, pero los **errores estándar (EE) están mal estimados** y por tanto, todo de ahí en adelante se estima mal también. Lo que lo puede llevar a conclusiones erróneas.

Regresión normal				
Término	**Coeficiente**	**EE**	**Valor T**	**Valor P**
β_0	7.455	0.543	13.721	0.0000026
β_1	4.909	0.729	6.735	0.0002687
Regresión Poisson				
Término	**Coeficiente**	**EE**	**Valor T**	**Valor P**
β_0	7.452	0.884	8.428	0.0000000
β_1	4.935	1.089	4.531	0.0000059
Diferencia porcentual				
β_0	0.039%	38.548%	62.801%	
β_1	0.531%	33.079%	48.634%	

Ejemplo 03 - Adecuación del modelo

Gráficos



- Dado que hay algunas dudas sobre la normalidad, se procede a realizar pruebas analíticas. En este caso, en pos de la didáctica, se realizan 3 de ellas. ¿Cuál deberíamos analizar?
- Con un 90 % de confianza, se puede rechazar la H_o de que los residuos siguen la distribución normal. Tanto por la forma de los gráficos, como por la prueba SW.

Ejemplo 03 - Re-especificación del modelo

- ¿Cómo lo soluciono? La forma más recomendada es la de re-especificar el modelo y re-estimar los parámetros.
- Esto involucra incluir otras variables o agregar interacciones. En este caso es razonable pensar que puede haber una interacción entre x_1: cantidad de cajas y x_2: distancia recorrida.
- $$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \varepsilon$$

Ejemplo 03 - Re-estimación del modelo

- Los coeficientes estimados son:

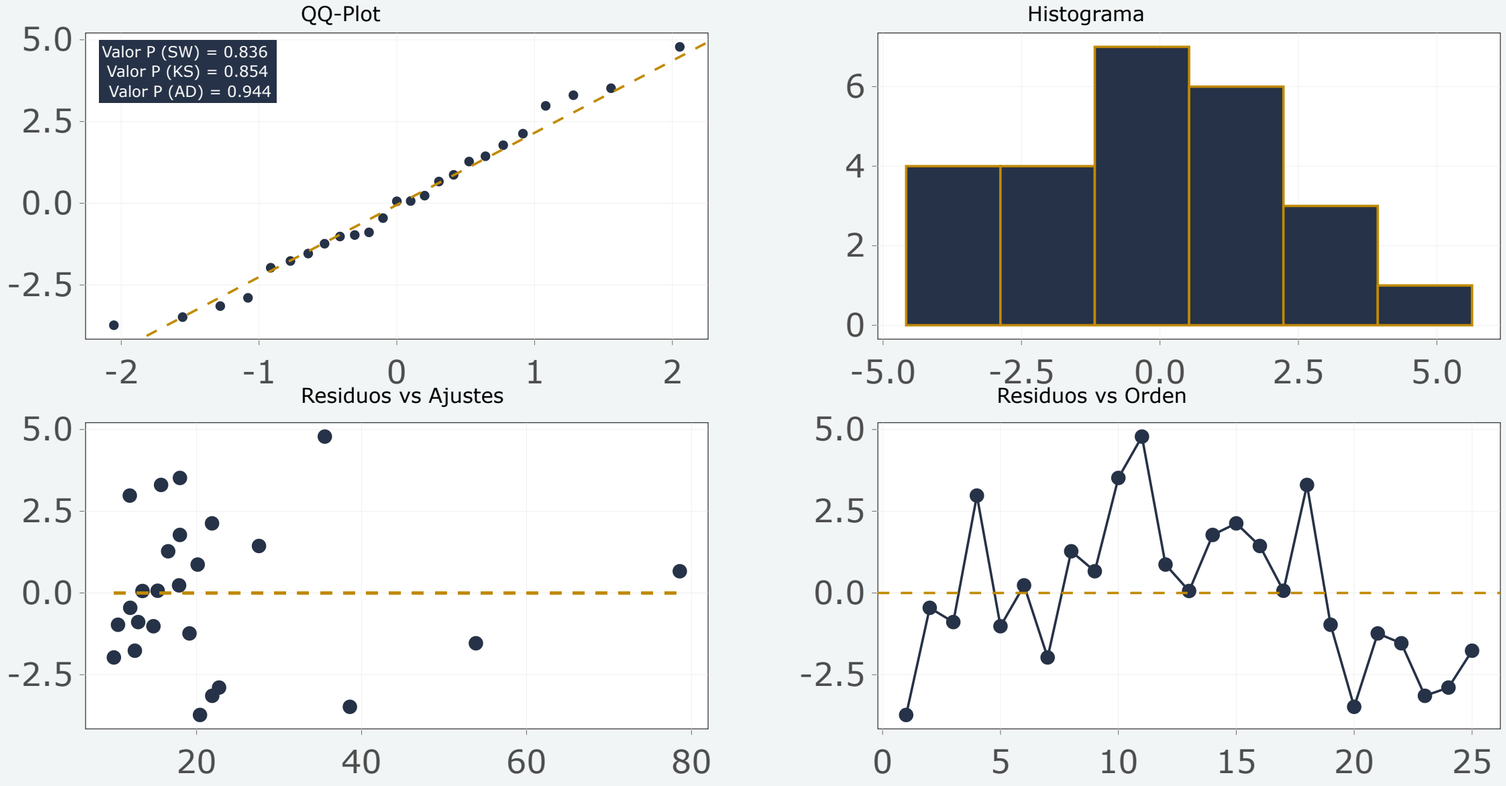
Término	Estimador	EE	Valor T	Valor P
(Intercept)	7.139	1.400	5.100	0.0
x1	1.014	0.191	5.304	0.0
x2	0.006	0.003	1.723	0.1
x1:x2	0.001	0.000	4.240	0.0

- Cuya ecuación de regresión es:

$$\widehat{\mathrm{y}} = 7.139 + 1.014(\mathrm{x1}) + 0.006(\mathrm{x2}) + 0.001(\mathrm{x1} \times \mathrm{x2})$$

Ejemplo 03 - Adecuación del modelo

Gráficos



- Ahora observe el comportamiento de los residuos en este nuevo modelo.
- Discuta con la persona docente y estudiantes.

¿Qué más puedo hacer sino se cumplen los supuestos?

Alternativas ante el incumplimiento de supuestos

- Lo que se aborda en esta sección tiene como fin el evitar que la persona estudiante piense que "se acaba el mundo" cuando esto sucede.
- Las alternativas aquí mencionadas no serán todas, necesariamente, abordadas en el curso o incluso en la carrera.
    — No obstante, es importante que sepa que puede hacer cuando esto le ocurra, para que pueda proceder con el estudio individual.

Alternativas

- **Transformaciones**
  - Útil para problemas de normalidad y homocedasticidad
  - Tiene sus amplias desventajas, sobre todo de interpretación.
  - Aunque existen transformaciones que combaten esas desventajas.
- Las más comunes son Box - Cox y Johnson. Con base en criterios de complejidad, se prefiere a la primera sobre la segunda.
- Aunque son son la alternativa más popular, pero no se deben aplicar a la ligera.
  - Antes de aplicar una transformación que cambie la interpretación de los datos, es mejor plantearse usar el modelo correcto.
  - Antes de aplicar una transformación, debe someterse a juicio si los datos son intrínsecamente no normales o si devienen de un problema de muestreo, precisión de la medición, entre otros.

Alternativas

- **ANOVA no paramétrico**
  - — Se conoce como Kruskall-Wallis.
  - — Útil para problemas de normalidad y homocedasticidad
  - — Muchas desventajas:
    - — Cantidad de factores (no existen ANOVA no paramétricos para más de dos factores).
    - — El tamaño de los intervalos generados (muy anchos).
  - — Este tópico será abordado con detalle en este curso.

Alternativas

- **Regresiones robustas**
  - — Útil para problemas de normalidad y homocedasticidad
  - — Desventaja:
    - — Complejos de comprender y estimar
  - — En este curso solo se mencionan como una alternativa:
    - — Huber
    - — Regresión con kernels
    - — Loess
  - — Recuerde que en este curso abordamos la prueba de Levene, la cual es “resistente” al incumplimiento de la normalidad. Estas siguen la misma idea.

Alternativas

- **Modelos autorregresivos**
  - — Útil para problemas con la independencia
  - — Tienen mucha complejidad, una forma de solucionarlo es con modelos de series de tiempo con intervención.
    - — No se aborda en este curso
  - — Generalmente solo se recurre a ellos cuando los datos son intrínsecamente dependientes en el tiempo, y NUNCA por errores en la recolección de datos.
    - — Si hubiese errores de este tipo, se invalidan los resultados.

Alternativas

- **Modelos lineales generalizados**
  - Útil para problemas de normalidad y homocedasticidad.
  - Aunque un poco más complejo que regresión y ANOVA clásico, es la opción más viable y la que se prefiere sobre las transformaciones u otros de naturaleza similar.
  - No se aborda en este curso, pero su mención es importante, incluye modelos binarios, de Poisson, exponenciales, entre otros. Por prácticamente cada distribución de probabilidad estudiada en estos cursos, se puede hacer un modelo lineal general.
  - En otros cursos estudiará los modelos binarios como técnicas de clasificación de datos (Machine Learning).

Alternativas

- **Estadística Bayesiana**
  - — Útil para problemas de normalidad y homocedasticidad y convergencia en modelos muy complejos
  - — Es un enfoque bastante más complejo, pero cada vez más popular.
  - — No se aborda en este curso y salvo algunas aplicaciones, no se aborda en la carrera.

Continuamos con la adecuación del modelo

Indicadores de bondad de ajuste

- Como parte de la verificación de la adecuación del modelo se tienen índices de bondad de ajuste que se pueden clasificar en absolutos, relativos y comparativos.
- En ese curso se van a estudiar algunos de ellos.

Coeficiente determinación R^2

- Es una medida con la que se puede saber que tanto está acertando o fallando el modelo. Este mide que tan grandes son los residuos.
- Este es un estadístico que debe analizarse con cuidado, pues **siempre crece a medida que se añaden términos en el modelo**, aun cuando estos no tengan un aporte real o significativo sobre la variable.
- $$R^2 = 1 - \frac{SS_e}{SS_T}$$

Coeficiente determinación R^2

- El valor de R^2 es engañoso, pues como se dijo anteriormente, siempre crece, sin importar si el predictor que se incluye es valio o no.
- Para eso casos existe una versión ajustada
- $$R^2_{adj} = 1 - \frac{MS_e}{MS_T}$$
- El R^2_{adj} penaliza la inclusión de predictores o variables que no aportan al modelo. Este valor es siempre menor que R^2.

Coeficiente determinación R^2

- Por ello se recomienda que su interpretación se realice de forma conjunta, tanto R^2 como R^2_{adj}. Una diferencia grande ($\sim > 3\%$) entre estos dos valores implicaría que en el modelo hay (o mejor dicho puede haber) términos no significativos y que por ende puede ser prudente eliminarlos para obtener un modelo limpio.
- Básicamente, se busca aplicar el principio de parsimonia. Por tanto se aconseja siempre presentar modelos "limpios".

¿Cómo se interpreta?

- Clásicamente, se indica que los valores de R^2 deben ser típicamente altos (>80 %). Si nos falta contexto, esto no es ninguna mentira, pero ahí está la cuestión...
- Los datos son de "carne y hueso" y sufren de las vicisitudes del mundo real.

¿Cómo se interpreta?

- **Caso No.1**
- Hacemos un experimento físico, que debería responder a una cierta fórmula ($PV = nRT$), en condiciones de aislamiento total y absoluto. En los experimentos obtenemos un dato real (experimental) y lo comparamos el que da la fórmula.

  — ¿Qué R^2 debería esperar?

- Hemos controlado que no le afecte nada, así que tendría que ser un ajuste casi perfecto. **Un valor <0.9 sería casi catastrófico**. Pues estaríamos siendo incapaces de medir un fenómeno físico.

¿Cómo se interpreta?

- **Caso No.2**
- Usamos los datos del censo, queremos analizar en qué medida los años de estudio pueden predecir un salario. Obviamente no hemos podido aislar el resto de variables que influyen sobre el sueldo (género, edad, carreras, etc).

  — ¿Qué R^2 debería esperar?

- No se puede explicar toda la renta de una persona con solo los estudios, así que un $R^2 = 0.3$ ya sería impactante.

  — Con una sola variable (entre miles) se consigue explicar un 30 % de variabilidad.

Como conclusión

- El valor de R^2 no es algo que haya que interpretar siempre igual para todos los conjuntos de datos. No podemos aislarlo del contexto ni de las expectativas.
- Podríamos hablar de más cosas:
  - Si $R^2 > 0$, ya existe una asociación, que podrá ser débil, pero existente.
  - El R^2 se ve influido por el tipo de relación asumida: lineal, cuadrática, exponencial, etc... En este curso las relaciones son lineales.
- Pero el contexto es lo más importante. Si los datos pueden estar influidos por muchos factores, un R^2 **no tan alto puede decirnos mucho**.

Coeficiente determinación R^2

- Es un indicador que puede fungir de forma relativa y comparativa. Por ejemplo:
- **Relativo**
  — Por ejemplo, un estudio dice que en fenómenos físicos se espera un valor mayor a 98 %. Por lo que mi modelo debe obtener un valor de R^2 similar para que sea válido.
- **Comparativo**
  — El modelo con 3 variables explica un 85 % de la variabilidad, pero al agregar una variable 4, aumenta a 90 %.

Índices de información

- Son índices **comparativos** que buscan que el modelo sea lo más sencillo posible, por ello, tienen incorporado una penalización por complejidad.
- Existen varios, pero todos se interpretan igual

  — **Entre más bajo sean en comparación con otro modelo, mejor**.

Índices de información

- Es importante anotar que la comparación solo puede hacerse en modelos anidados. Por ejemplo, en el **Ejemplo 03** tenemos dos especificaciones de modelo.

  — El que tiene interacción y el que no.

- El segundo está anidado en el primero. Por tanto, se pueden comparar con estos índices para decidir cuál de los dos es mejor.

  — Son útiles en el sentido iterativo con el que se está abordando esta sesión.

- La idea en simple: qué tan bien el modelo explica los datos, comparado con los otros, pero castigando los modelos demasiado complejos.

Índices de información

- Los índices de información más populares son:
  - — **AIC**: Criterio de información de Akaike
  - — **AICc**: Criterio de información de Akaike corregido
  - — **BIC**: Criterio de información Bayesiano
- Estos están presentes en varios softwares estadísticos. Ambos son similares, pero BIC tiene un factor de penalización más fuerte cuando se incluyen muchas variables. Y el AICc es especialmente útil con muestras pequeñas.
- El detalle de las fórmulas se obvia voluntariamente, pues necesitan una comprensión detallada de la logverosimilitud. Por lo que se enfatiza en su interpretación.

Globales del modelo

- El estadístico F que se obtiene de ANOVA para toda la regresión, contrasta la hipótesis de si el modelo complejo explica significativamente más que un modelo sin predictores.
- En simple, si hay ganancia en hacer la regresión.

Errores de predicción

- Estos se enfocan en cuánto se equivocan las predicciones del modelo.
- En algunos software como Minitab existe el PRESS (*Prediction Sum of Squares*). Un valor de PRESS menor indica un mejor modelo predictivo.
- También se puede encontrar el valor de S, que es $S = \sqrt{CM_e}$, es una medida de la desviación estándar de la regresión.

Resumen de indicadores de bondad de ajuste

- En la práctica se suelen combinar varios de estos criterios. Por ejemplo:
  - Analizar el R^2 para explicar la varianza, el *PRESS* para evaluar la capacidad de predicción y el AIC/BIC para comparar y decidir entre dos modelos competentes y rivales.

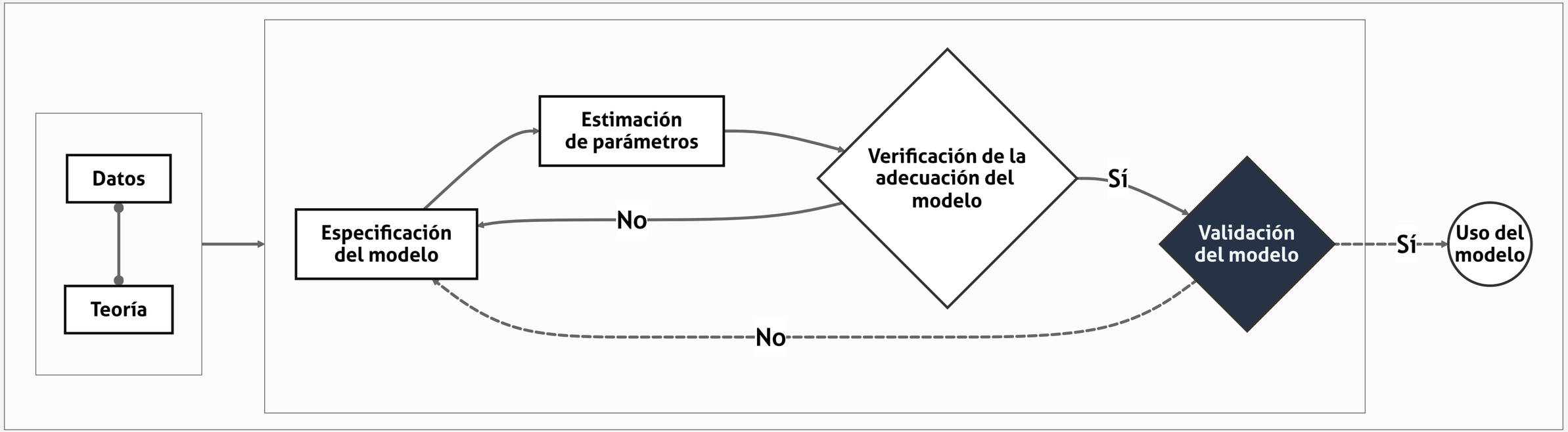
Ejemplo 03 - Adecuación del modelo

- Obtenga los indicadores de bondad de ajuste de los dos modelos estimados (con interacción y sin interacción). Compárelos y decida sobre uno de ellos.
- En general, se prefiere el modelo con interacción, ¿verdad?

  — Explique y detalle el por qué

Índice	Con interacción	Sin interacción
R^2	0.978	0.960
R^2_{adj}	0.975	0.956
s	2.449	3.259
Valor F (Regresión)	314.555	261.235
Valor P (Regresión)	0.000	0.000
AIC	121.367	134.829
BIC	127.462	139.705

Validación del modelo



Tomado y adaptado de Montgomery et al. (2012).

Para ver el diagrama en pantalla completa, diríjase a este **enlace**.

Validación del modelo

- Se hace una distinción entre la adecuación del modelo y la validez del modelo.
- Aunque este último paso con frecuencia es extenso e implica algunas técnicas que superan el alcance del curso.
- Está dirigida a determinar **si el modelo va a funcionar en la forma en la que fue concebido** en las condiciones o ambiente de operación.

Validación del modelo

- La validez del modelo se puede determinar de varias formas:
  - **Contrastando los resultados obtenidos contra la experiencia previa, la teoría aplicable, entre otros**. Por ejemplo, que la regresión indique que la gravedad es negativa no tendría lógica.
  - **Recolectando nuevos datos** y contrastando estos valores contra los valores predichos por la ecuación del modelo.
  - **Partición de datos**, es deci rpartir los datos recolectados en dos partes. Se emplea una para hacer la estimación y la otra parte para validar que el modelo funcione.

Ejemplo 03 - Validación del modelo

- Con la información disponible, valide el modelo seleccionado del Ejemplo 03.
- Discuta en grupos. ¿Considera que se debe reespecificar el modelo?
- Si no, el modelo está listo para ser usado.

Multicolinealidad

Introducción

- Vale la pena aclarar que el análisis de multicolinealidad se puede realizar como parte de la revisión de la **adecuación** del modelo.
- El uso e interpretación de los modelos de regresión múltiple dependen a menudo (explícita o implícitamente) en la estimación de los coeficientes individuales.
- Se supone que los regresores (x) deben ser independientes entre si y dependientes con la variable de respuesta (y). Es por eso que los regresores deben escogerse cuidadosamente.
- De cumplirse la primera premisa, estos serían ortogonales.
- Cuando los regresores son ortogonales, los análisis e inferencias pueden realizarse fácilmente.

Multicolinealidad

- No obstante, en ocasiones, los regresores pueden estar cerca de ser linealmente dependientes y esto provocaría que las conclusiones a las que se arriban sean erróneas.
- Esto es lo que se conoce como multicolinealidad. Esta provoca que las estimaciones sean inestables.
- Cuando dos predictores están casi alineados, el modelo tiene problemas para decidir cuál de ellos merece el crédito, y la incertidumbre del coeficiente explota (se infla).
- A menudo la multicolinealidad se analiza con el FIV: Factor de Inflación de la Varianza.

FIV

- Es un diagnóstico de la multicolinealidad. No obstante el primer diagnóstico debe ser la teoría aplicable al construir el modelo. La fórmula para una variable x_j es:
- $$FIV_j = \frac{1}{1 - R_j^2}$$
- Donde R_j^2 es el coeficiente de determinación obtenido al hacer una regresión donde x_j se explica usando todas las demás variables explicativas. En simple, x_j pasa de ser un regresor a la variable de respuesta.
- Valores de $FIV > 10$ implica problemas serios con la multicolinealidad. Valores de $FIV > 5$ indican un posible problema y debe analizarse con cautela.

FIV

- Tome en cuenta que con frecuencia, al incluir interacciones los FIV explotan, es una consecuencia natural y estructural del modelo. Es importante entender que la interacción no es una dirección o variable nueva, sino una especie de “curvatura”.

  — En resumen, en presencia de interacciones el FIV se vuelve menos informativo.

- También tome en cuenta que un modelo con interacción y sin interacción se interpretan de formas distintas.

What if...?

- **¿Qué pasa si la colinealidad está presente?**
- El paso natural es eliminar la variable problemática (re-especificar el modelo), pero a veces no es posible o deseable, en ese caso se puede recurrir a al menos dos opciones:
    - — Regresión Ridge
    - — Análisis de componentes principales

Ejemplo 03 - Multicolinealidad

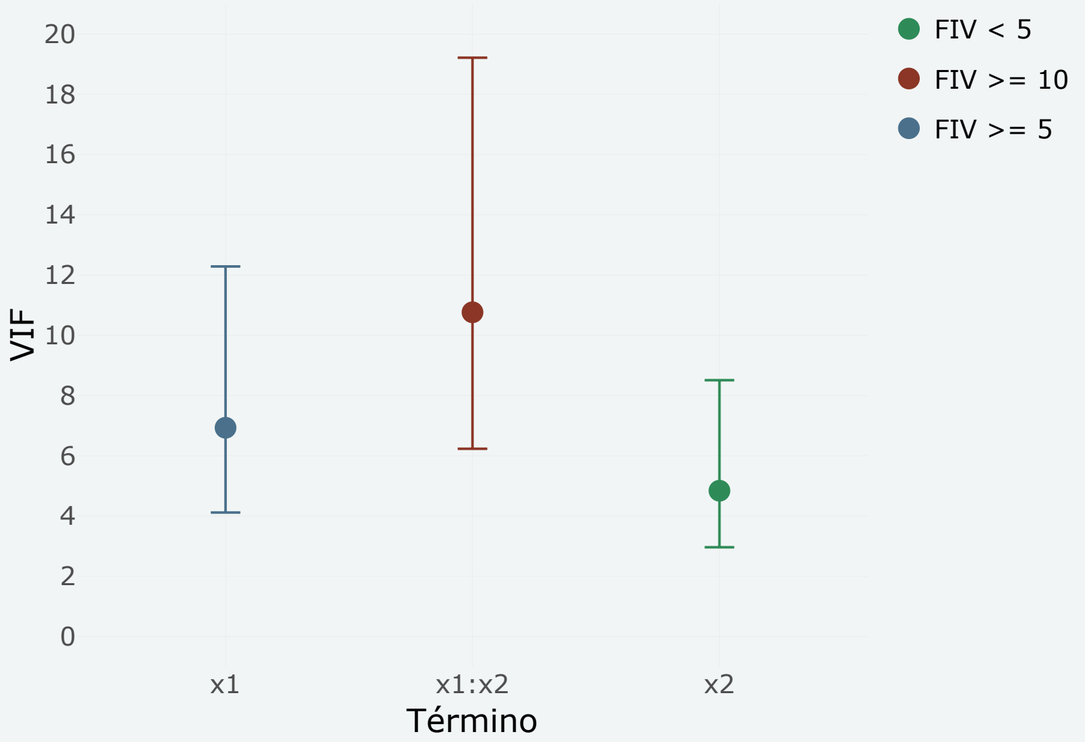
- Realice el análisisde multicolinealidad para los dos modelos planteados en el Ejemplo 03. Recuerde que el FIV es más revelante para el modelo sin interacción que en el con interacción.
- Interprete los resultados.

Variable	Sin	Con
x_1	3.118	6.933
x_2	3.118	4.842
x_1x_2	NA	10.765

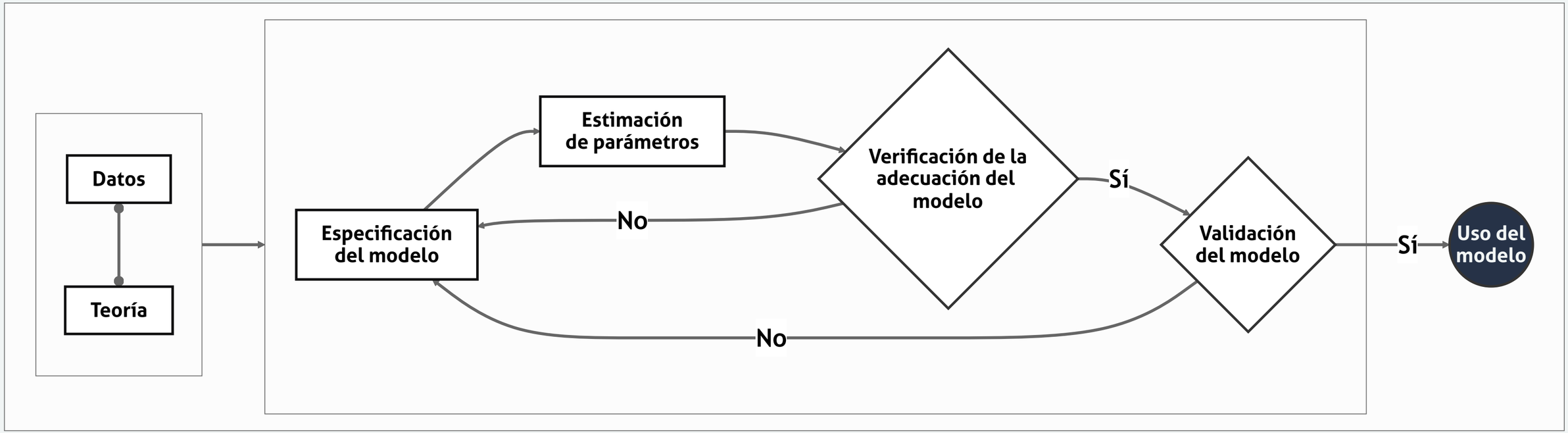
Ejemplo 03 - Multicolinealidad

Sin interacción

Con interacción



Uso del modelo - Interpretación



Tomado y adaptado de Montgomery et al. (2012).

Para ver el diagrama en pantalla completa, diríjase a este **enlace**.

Interpretación de resultados

- Esta sección se refiere a como se interpretan correctamente los resultados obtenidos una vez **todo el proceso para obtener una regresión ha sido realizado**.
- Solo se realiza cuando el modelo está limpio, todos los supuestos se cumplen, el modelo se ha validado, etc.
- Por tanto, esta sección NO se refiere a que solo se interpretan estos resultados

Interpretación de resultados

Sin interacción

$$y = 2.34 + 1.62 \cdot x_1 + 0.0144 \cdot x_2 + e$$

- La interpretación que se da a cada coeficiente de correlación parcial (β_i) a excepción del intercepto (β_0) es la siguiente:
  - El tiempo de entrega se aumenta en 1.62 minutos por cada caja (x_1) adicional, manteniendo los demás regresores constantes.
  - El tiempo de entrega se aumenta en 0.0144 minutos por cada ft (x_2) adicional que se deba recorrer, manteniendo los demás regresores constantes.

Con interacción

$$y = 7.14 + 1.014 \cdot x_1 + 0.0058 \cdot x_2 + 0.000742 \cdot x_1 x_2 + e$$

- El tiempo de entrega aumenta en $1.014 + 0.000742x_2$ minutos por cada caja adicional (x_1), manteniendo constante la distancia recorrida.
  - Es decir, a mayor distancia, mayor es el tiempo adicional de entregar una caja extra.
- El tiempo de entrega aumenta en $0.0058 + 0.000742x_1$ minutos por cada ft adicional, manteniendo constante el número de cajas.
  - Cuanto más cajas se lleven, más costoso en tiempo es recorrer distancia adicional.

Interpretación de resultados

Sin interacción

$$y = 2.34 + 1.62 \cdot x_1 + 0.0144 \cdot x_2 + e$$

- La interpretación de β_0 se debe realizar con cuidado (y en ocasiones ni siquiera es necesario realizarla), pues sus resultados pueden no ser coherentes.
- Si fuese estrictamente necesario obtener un β_0 interpretable, habría que hacer una operación particular de centramiento.
- La interpretación es la siguiente:

  — Cuando se transporten cero cajas y se recorra una distancia de cero metros, el tiempo de entrega estimado es de 2.34 minutos.

Con interacción

$$y = 7.14 + 1.014 \cdot x_1 + 0.0058 \cdot x_2 + 0.000742 \cdot x_1 x_2 + e$$

- La interpretación es la siguiente:

  — Cuando se transporten cero cajas y se recorra una distancia de cero metros, el tiempo de entrega estimado es de 7.14 minutos.

ANOVA

ANOVA de dos factores

- Como se ha detallado previamente, ANOVA y regresión son básicamente lo mismo, ANOVA es un caso especial de regresión que tiene mayor utilidad cuando se trabaja con variables categóricas.
- Sigue los mismos supuestos que regresión, tiene los mismos indicadores de bondad de ajuste.

ANOVA de dos factores

Tiene la siguiente forma funcional:

$$y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \varepsilon_{ijk}$$

donde:

- y_{ijk}: cada una de las observaciones.
- μ: la media conjunta.
- α_i: efecto del i-ésimo nivel del factor A.
- β_j: efecto del j-ésimo nivel del factor B.
- $(\alpha\beta)_{ij}$: efecto de la interacción entre el i-ésimo nivel del factor A y el j-ésimo nivel del factor B.
- ε_{ijk}: desviaciones, con respecto a la media conjunta, de los valores y_{ijk}.

Grados de libertad

- Los grados de libertad en ANOVA, para los predictores, se asignan de la siguiente forma: $k - 1$.
- Siendo k la cantidad de niveles en la categoría. Por ejemplo, si pruebo tres grosores de lámina, los grados de libertad serán 2.
- Los grados de libertad totales son siempre $n-1$, siendo n la cantidad de observaciones. Los grados de libertad del error son lo que sobra de la asignación para los predictores.

Prueba de hipótesis

- En ANOVA se realiza el siguiente contraste de hipótesis:

$$H_o : \text{Todas las medias por grupo son iguales}$$
$$H_i : \text{Al menos una de las medias por grupo es diferente a las demás}$$

- La respuesta a la pregunta: ¿Cuál de las medias es diferente? se responde más adelante. Claramente esta pregunta solo tiene sentido si la variable tiene más de dos categorías, por ejemplo: Color (rojo, azul y amarillo).

Ejemplo 04 - ANOVA

- Una empresa de materiales de construcción quiere estudiar la influencia que tienen el grosor y el tipo de templado sobre la resistencia máxima de unas láminas de acero.
- Para ello miden el estrés hasta la rotura (variable cuantitativa dependiente) para dos tipos de templado (lento y rápido) y tres grosores de lámina (8 mm, 16 mm y 24 mm).
- Use un 90 % de confianza estadística.
- Como en los otros ejemplos, los datos se encuentran en Excel.

Ejemplo 04 - ANOVA

- Obtenga la tabla ANOVA para la situación descrita. Incluya desde un inicio la interacción. Analice los resultados obtenidos.

  — ¿Es el efecto de interacción significativo?

- Obtenga también indicadores de bondad de ajuste.

- > Como parte de su estudio individual, verifique los supuestos del modelo y determine las acciones a seguir.

ANOVA

Término	GL	SS	MS	Valor F	Valor P
Templado	1	112.68	112.68	380.08	0.00
Grosor	2	10.41	5.21	17.56	0.00
Templado:Grosor	2	1.60	0.80	2.70	0.09
Residuals	24	7.11	0.30	NA	NA

Indicadores de bondad de ajuste

R^2	R^2_{adj}	s	Valor F (Regresión)	Valor P (Regresión)
0.946	0.935	0.544	84.123	0

Ejemplo 04 - ANOVA

- Con base en los resultados obtenidos, hay evidencia para decir que la resistencia es distinta en función de los niveles de grosor y de templado.
- Ahora, el interés de la organización es seleccionar la configuración que de la mayor resistencia. ¿Cómo obtengo esos valores? ¿Cómo doy una respuesta?
- Para ello vamos a usar una técnica llamada **comparaciones múltiples**. Muchos softwares estadísticos ofrecen una batería de herramientas para tomar esta decisión, entre ellas:
    - Tukey
    - Fisher
    - Bonferroni
    - Sidak

Ejemplo 04 - Tukey

- Es una prueba bastante conservadora que calcula el intervalo de confianza **simultáneo** al nivel de confianza especificado. Controla bien el error tipo I.
- Es muy útil cuando el tamaño de los grupos está balanceado.
- **En el ejemplo**
  - Esto quiere decir que todos los grosores son diferentes. Siendo todos diferentes, se escoje el de menor media, en este caso el grosor 8 tiene una media de 14.151.
  - Interprete para el Templado.

Grosor

Contraste	Estimado	EE	GL	t	P
Grosor8 - Grosor16	-0.85	0.24	24	-3.49	0.01
Grosor8 - Grosor24	-1.44	0.24	24	-5.89	0.00
Grosor16 - Grosor24	-0.58	0.24	24	-2.40	0.06

Templado

Contraste	Estimado	EE	GL	t	P
Lento - Rápido	-3.88	0.2	24	-19.5	0

Ejemplo 04 - Fisher

- Aumenta el riesgo de error tipo I, pues calcula todos los intervalos de confianza de forma individual, manteniendo el α para cada intervalo
- Básicamente aplica varios t-test entre pares, pero no corrige por las comparaciones múltiples (Como si lo hace Tukey)
- **En el ejemplo**
  - Se llegan a las mismas conclusiones que con Tukey, pero observe como cambian las estimaciones de los valores P.

Grosor

Contraste	Estimado	EE	GL	t	P
Grosor8 - Grosor16	-0.85	0.24	24	-3.49	0.00
Grosor8 - Grosor24	-1.44	0.24	24	-5.89	0.00
Grosor16 - Grosor24	-0.58	0.24	24	-2.40	0.02

Templado

Contraste	Estimado	EE	GL	t	P
Lento - Rápido	-3.88	0.2	24	-19.5	0

Ejemplo 04 - Bonferroni

- Calcula, como Tukey, intervalos de confianza simultáneos.
- Es simple de aplicar, pero produce resultados más conservadoras.
- Reduce la probabilidad de falsos positivos, a costa de perder potencia.
- **En el ejemplo**
  - Se llegan a las mismas conclusiones que con Tukey y Fisher, pero observe como cambian las estimaciones de los valores P.
  - Analice el comportamiento que se aprecia al aplicar las pruebas.

Grosor

Contraste	Estimado	EE	GL	t	P
Grosor8 - Grosor16	-0.85	0.24	24	-3.49	0.01
Grosor8 - Grosor24	-1.44	0.24	24	-5.89	0.00
Grosor16 - Grosor24	-0.58	0.24	24	-2.40	0.07

Templado

Contraste	Estimado	EE	GL	t	P
Lento - Rápido	-3.88	0.2	24	-19.5	0

Ejemplo 04 - Sidak

- Es, grosso modo, una variante más refinada de Bonferroni, lo que la hace ligeramente menos conservadora.
- **En el ejemplo**
  - Puede observar como los resultados son sumamente similares a los de Bonferroni.

Grosor

Contraste	Estimado	EE	GL	t	P
Grosor8 - Grosor16	-0.85	0.24	24	-3.49	0.01
Grosor8 - Grosor24	-1.44	0.24	24	-5.89	0.00
Grosor16 - Grosor24	-0.58	0.24	24	-2.40	0.07

Templado

Contraste	Estimado	EE	GL	t	P
Lento - Rápido	-3.88	0.2	24	-19.5	0

Ejemplo 04 - Regresión

- Resuelva este ejercicio como una regresión lineal múltiple, no como un ANOVA y discuta con compañeros y docente sobre si la respuesta obtenida es equiparable a los resultados que se obtienen de ANOVA y comparaciones múltiples.

ANOVA no paramétrico

Kruskall - Wallis

- Tiene la siguiente prueba de hipótesis sobre la mediana de más de dos grupos independientes que se aplica cuando no se cumplen los supuestos de normalidad o de varianza homogénea.
- La idea básica que se sigue es que evalúa si las medianas de los grupos son iguales.
- $$H_o : \widetilde{X}_A = \widetilde{X}_B = \widetilde{X}_C$$
$$H_i : \text{Al menos una es diferente}$$

Ejemplo 05 - Kruskall - Wallis

- Se quiere evaluar el resultado de tres dietas en el peso de pollos en un granja.
- Los datos se entregan en el Excel respectivo.
- Determine si alguno de los pesos es diferente.
- **Solución**
  - Con los resultados obtenidos, a un 95 % de confianza, tengo evidencia suficiente para rechazar la hipótesis nula de que las dietas son iguales.

```
    Kruskal-Wallis rank sum test

data:  Peso by Dieta
Kruskal-Wallis chi-squared = 7.1141, df = 2, p-value = 0.02852
```

Ejercicio integrador

- Se realizó un estudio por parte de personas estudiantes de la carrera de Bach. en Ciencias del Movimiento Humano (CIMOHU), cuyo objetivo era examinar el efecto que tiene la posición de salida (+ y -) y el tipo de calentamiento (A, B y C) en el recorrido de 100 metros planos. Se trabajó con una población de 17 a 24 años, todos estudiantes masculinos de la sede Rodrigo Facio de la Universidad de Costa Rica.
- Como variable de respuesta se mide el tiempo en segundos (s) que cada persona tarda en recorrer los 100 metros. Se asume que todas las mediciones son correctas y que los instrumentos de medición utilizados son adecuados.

Ejercicio integrador

- En resumen, se puede decir que la fuente de datos es correcta y que los datos se recolectaron de forma adecuada. La hipótesis que se sigue, basada en la teoría de las CIMOHU indica que si debiera existir diferencia entre el tiempo de recorrido de los 100 m planos dado los regresores estudiados.
- Se le hace entrega de los datos, así como del ejercicio parcialmente resuelto (**enlace**).
- La idea es que la persona estudiante lo realice sin ver la solución y que de ser necesario se guíe con la resolución aportada, pero que a su vez complete los elementos faltantes.

Bibliografía

- Montgomery, D.; Peck, E.; Vining, G. (2012). Introduction to Linear Regression Analysis (5th Ed). Wiley.

  — *Capítulo 1, 2, 3, 4, 9 y 11*

Estadística multivariada: Regresión y ANOVA
II-1123 Estadística para Ingeniería Industrial II

Gracias por su atención

Steven García Goñi

steven.garciagoni@ucr.ac.cr

Dudas o correcciones requeridas pueden solicitarse al correo