MÉTODOS ESTADÍSTICOS. ANÁLISIS DE REGRESIÓN

Curso de nivelación
Maestría en Estadística Aplicada

Facultad de Ciencias Económicas y Estadística

Lic. Noelia Castellana
Lic. Mara Catalano

Unidad 1 y 2

Presentación

Presentación

- **Docentes**
  - ✓ Lic. Mara Catalano
  - ✓ Lic. Noelia Castellana

- **Días y horarios**
  - ✓ Viernes 07/06: **14 a 18**
  - ✓ Sábado 08/06: 9 a 13
  - ✓ Viernes 21/06: **14 a 18**
  - ✓ Sábado 22/06: 9 a 13
  - ✓ Viernes 28/06: **14 a 18**
  - ✓ Sábado 29/06: 9 a 13
  - ✓ Viernes 05/07: **14 a 18**
  - ✓ Sábado 06/07: 9 a 13

- **Evaluación**: Trabajo práctico grupal

  Fecha de entrega 1° parte: 21/06/2019

  Fecha de entrega 2° parte y exposición: 06/07/2019

Programa

- **Unidad 1**

Estudios observacionales y experimentales. Tipo de variables y roles. Relaciones entre variables. Modelos estadísticos: lineales y no lineales. Introducción al análisis de regresión. Objetivos y usos.

- **Unidad 2**

Regresión lineal simple. Análisis descriptivo preliminar. Estimación e inferencia. Partición de la suma de cuadrados total (ANOVA). Medidas descriptivas de la relación entre las variables en un modelo de regresión.

- **Unidad 3**

Regresión lineal múltiple. Análisis descriptivo preliminar. Enfoque matricial. Estimación e inferencia. Partición de la suma de cuadrados total (ANOVA). Principio de la suma de cuadrados extra y su uso en pruebas de hipótesis. Multicolinealidad y sus efectos. Causas. Diagnósticos. Soluciones a la multicolinealidad.

- **Unidad 4**

Comprobación de la adecuación del modelo. Definición de residuos y métodos gráficos correspondientes. Gráficos de regresión parcial y residuos parciales. Pruebas de hipótesis formales. Soluciones al incumplimiento de los supuestos.

- **Unidad 5**

Modelos con regresores cuantitativos y cualitativos. Concepto de variables indicadoras. Modelos de regresión con una o más variables indicadoras. Usos de las variables indicadoras. Métodos de regresión por segmentos.

- **Unidad 6**

Construcción de modelos de regresión: efectos de una especificación incorrecta del modelo. Criterios para evaluar submodelos. Técnicas para seleccionar las variables explicativas: todas las regresiones posibles y métodos de selección automáticos.

- **Unidad 7**

Detección de valores atípicos y estudio de su influencia. Diagnósticos para detectar los valores atípicos. Matriz H y residuos estudentizados. Influencia sobre la ecuación de regresión estimada. Medidas de influencia.

Bibliografía

- NETER, J., KUTNER, M., NACHTSHEIM, C. y WASSERMAN, W. (2005) "*Applied linear statistical models*". Irwin

- MONTGOMERY, D. Y PECK, E. (2012) "*Introduction to linear regression analysis*". Wiley, NY.

- MONTGOMERY, D., PECK, E. Y VINING, G. (2004) "*Introducción al análisis de regresión lineal*". CECSA, México.

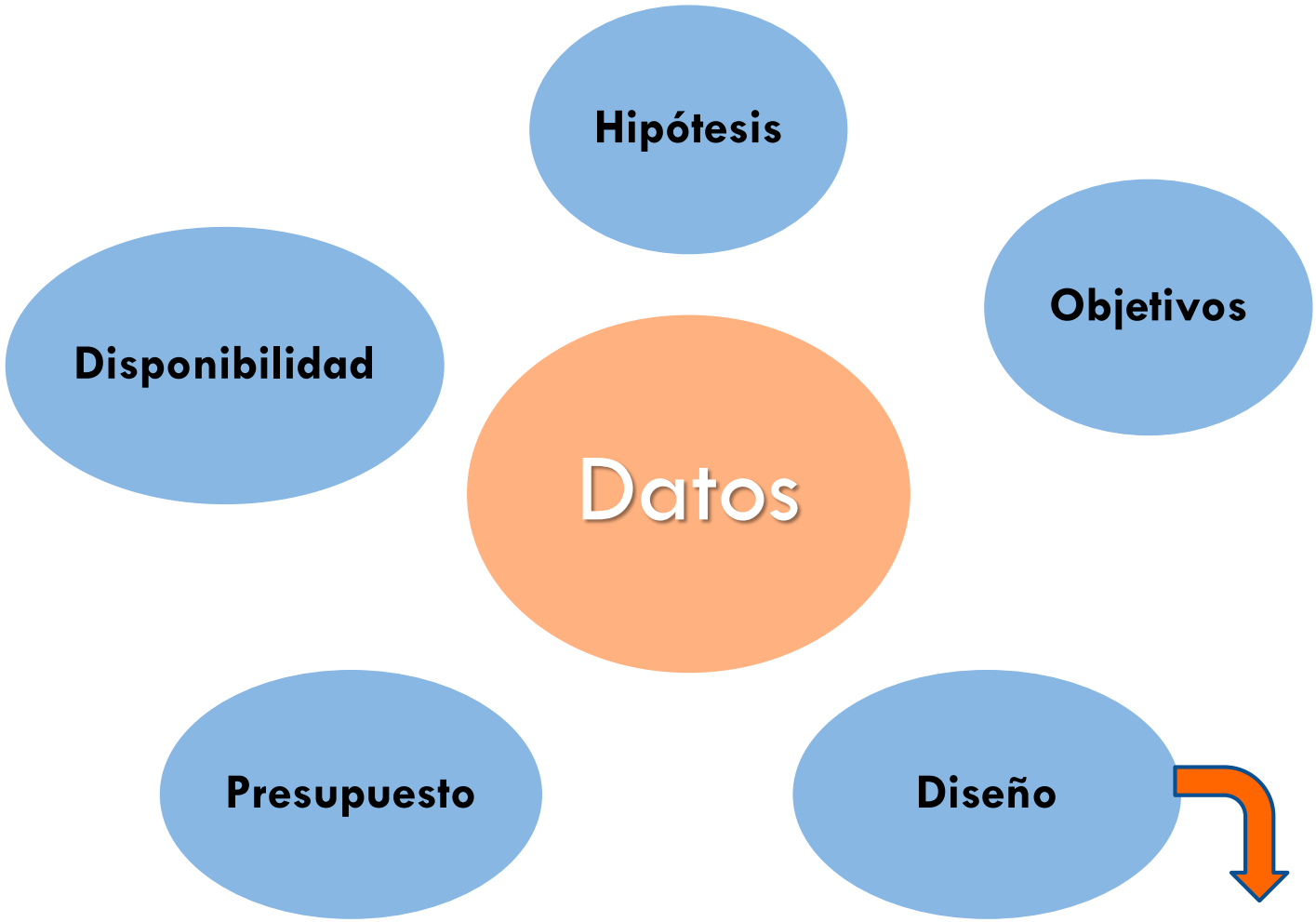
Unidad 1

Introducción

- Recolección de datos
- Variables
- Modelos estadísticos
- Introducción al Análisis de Regresión

Recolección de datos

- **¿Cómo se obtienen los datos**



Recolección de datos

- **¿Cómo se obtienen los datos**

Estudios observacionales	Estudios experimentales
• El investigador sólo observa • Se **observan** las variables sin intervenir en el proceso que las genera • **Prospectivos** • **Restrospectivos**	• El investigador interviene • Se **controla** el proceso de generación de los datos

Recolección de datos

- **Ejemplo-Botellas de vidrio**

En una planta industrial donde se fabrican botellas de vidrio se sospecha que la **cantidad de botellas defectuosas** por lote podría estar relacionada, entre otras causas, con:

- ✓ la **temperatura** del horno de fundición
- ✓ la **proporción de arena** que contiene la mezcla

Recolección de datos

- **¿Cómo es el proceso de fabricación del vidrio?**

1) A partir de las **materias primas** (arena, piedra caliza, vidrio reciclado y carbonato de sodio) se genera una "**mezcla**" que es **fundida** en un horno a altas temperaturas.

2) El **vidrio fundido** se enfría y se corta en "**gotas**".
3) Las **gotas** se distribuyen en las **máquinas** que le dan la **forma de envase de vidrio** y se genera el lote.

4) Finalmente **el lote** pasa a través del **templador,** para fortalecer el vidrio.
5) Se realizan los controles de calidad.

Recolección de datos

- **Ejemplo-Botellas de vidrio**

Para recolectar datos de las tres variables de interés se puede llevar a cabo tres tipos de estudios:

- ✓ **Observacional retrospectivo**
- ✓ **Observacional prospectivo**
- ✓ **Experimental**

Recolección de datos

- **Observacional retrospectivo:**

Se **revisan los registros** de cada proceso del **último año** considerando las siguientes variables **para cada lote** de botellas :

- ✓ Cantidad de botellas defectuosas
- ✓ Temperatura del horno al que fue fundida la mezcla
- ✓ % de arena que contenía la mezcla

Ventajas?

Desventajas?

Recolección de datos

- **Observacional prospectivo:**

Se decide **empezar a registrar** en detalle las características más relevantes del proceso durante los **próximos meses,** teniendo en cuenta para **cada lote** de botellas :

- ✓ Cantidad de botellas defectuosas
- ✓ Temperatura del horno al que fue fundida la mezcla
- ✓ % de arena que contenía la mezcla

Ventajas?

Desventajas?

Recolección de datos

- **Experimental:**

Se diseña un **experimento** en donde **en cada proceso de fundición** se tendrán en cuenta las **diferentes temperaturas** que el horno puede alcanzar (1475ºC, 1500ºC) y los **diferentes % de arena** que puede contener la mezcla (70%, 73%, 75%) dejando fija, en forma proporcional, los demás componentes.

Se tendrán entonces **6 combinaciones** posibles.

Recolección de datos

- **Experimental:**

Arena 70%- Temp 1475ºC	Arena 70%- Temp 1500ºC
Arena 73%- Temp 1475ºC	Arena 73%- Temp 1500ºC
Arena 75%- Temp 1475ºC	Arena 75%- Temp 1500ºC

Durante 36 días a cada proceso diario se le asignará, en forma aleatoria, una combinación. Por lo tanto, cada combinación estará presente 6 veces en cada lote.

Luego para cada proceso, se registrará la cantidad de botellas defectuosas en el lote.

Ventajas?

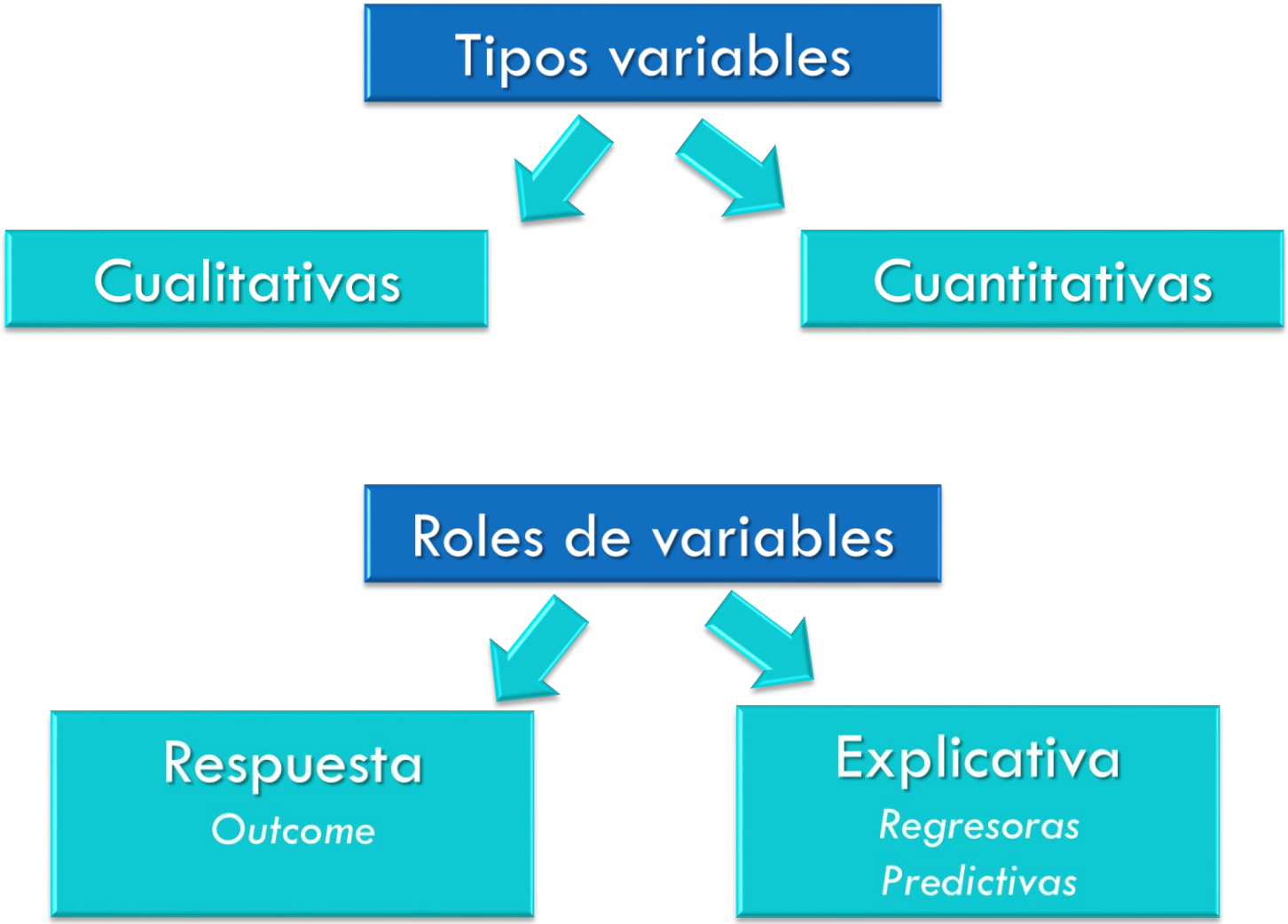
Desventajas?

Unidad 1

Introducción

- Recolección de datos
- Variables
- Modelos estadísticos
- Introducción al Análisis de Regresión

Variables


Tipos variables
Cualitativas
Cuantitativas
Roles de variables
Respuesta
Outcome
Explicativa
Regresoras
Predictivas

Variables

Análisis de las variables en estudio

Univariado

- ✓ Medidas descriptivas
- ✓ Gráficos
- ✓ Intervalo de confianza
- ✓ Test de hipótesis

Multivariado

En este curso:
1 respuesta y p explicativas

Variables

- Se pueden distinguir **dos tipos** de **relaciones**

Funcional	Estadística
• La relación es exacta • Modelo matemático	• La relación no es perfecta • Todos los puntos no caen sobre la curva que describe la relación • Modelo estadístico

Variables - Ejemplo

Funcional

Circunferencia = π × diámetro

Y= $50 × X
Y= total vendido
X= cantidad unidades vendidas
$50= precio unitario

Fahrenheit = 9/5*Celsius + 32

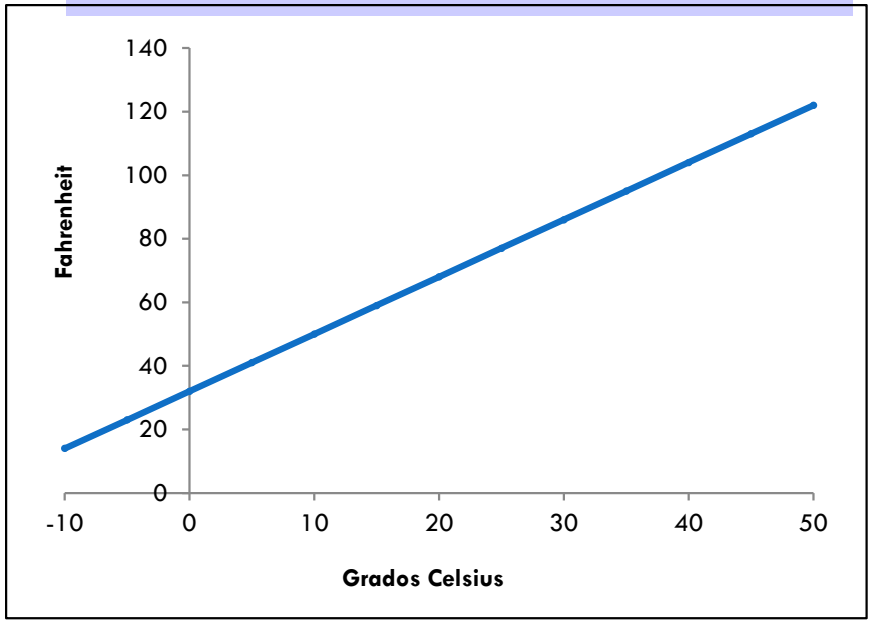
Estadística

En un estudio realizado en la Facultad de Cs Económicas y Estadística se desea evaluar la relación entre **la nota del parcial** obtenida por un alumno y su **nota final.**
Para 11 alumnos se registra:
-la nota del parcial
-la nota del final

Variables - Ejemplo

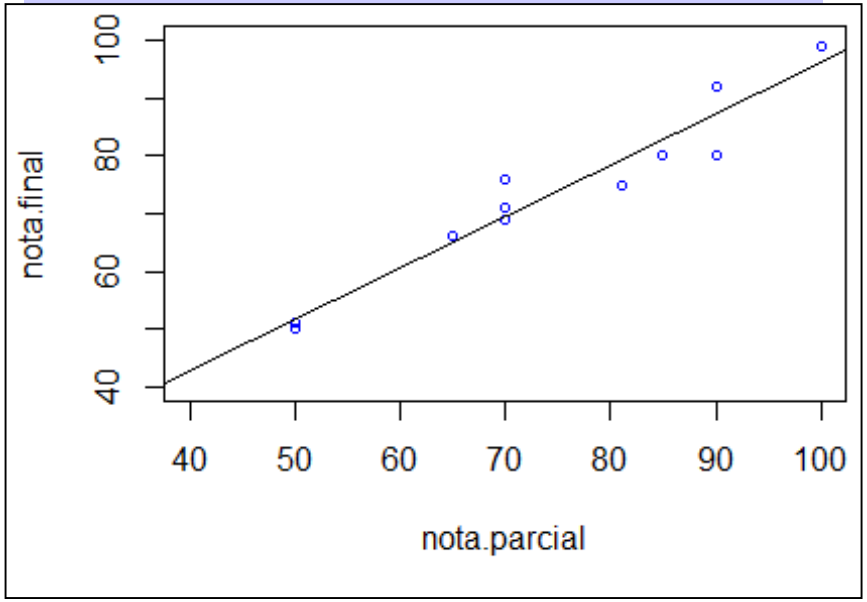
Funcional

Fahrenheit = 9/5*Celsius + 32



Estadística

Alumnos según nota parcial y nota final



Gráficos de dispersión

Unidad 1

Introducción

- Recolección de datos
- Variables
- Modelos estadísticos
- Introducción al Análisis de Regresión

Modelos estadísticos

¿Qué es un modelo?

- **Representación** de lo que se percibe como el **mecanismo** que **generan** los **datos.**

$$y_i = f(\mathbf{x}_i) + \varepsilon_i$$

Modelos estadísticos

Un modelo está constituido por:

- **Ecuación matemática**: que idealiza la relación entre las variables. Esta ecuación está compuesta por:
  - ✓ Variables
  - ✓ La forma en que las variables están relacionadas

- **Especificaciones** realizadas sobre algunas de las variables que intervienen en la ecuación.

Modelos estadísticos - Tipos

Lineales

Los **parámetros del modelo** están involucrados de manera **lineal**:

- ✓ Aparecen elevados a potencia uno
- ✓ No están divididos ni multiplicados por otros parámetros.

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

Los **variables del modelo** pueden estar involucradas de manera lineal y no lineal:

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \varepsilon_i$$

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 z + \beta_3 xz + \varepsilon_i$$

No lineales

Los **parámetros del modelo** están involucrados de manera **no lineal**:

- ✓ Pueden aparecen elevados a potencia diferente de uno
- ✓ Pueden estar divididos o multiplicados por otros parámetros.

$$y_i = \frac{1}{\beta_0 + \beta_1 x_i + \varepsilon_i}$$

Los **variables del modelo** pueden estar involucradas de manera lineal y no lineal:

$$y_i = \frac{1}{\beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \varepsilon_i}$$

Modelos estadísticos- Tipos

Modelos linealizable

Modelos no lineales que con una transformación se convierten en modelos lineales

$$y_i = \beta_0 e^{\beta_1 x_i} + \varepsilon_i$$

$$\ln(y_i) = \ln\left(\beta_0 e^{\beta_1 x_i} \varepsilon_i\right) =$$

$$= \ln(\beta_0) + \beta_1 x_i + \ln(\varepsilon_i) =$$

$$= \beta_0^* + \beta_1 x_i + \varepsilon_i^*$$

Unidad 1

Introducción

- Recolección de datos
- Variables
- Modelos estadísticos
- Introducción al Análisis de Regresión

¿Qué es el análisis de regresión?

¿Qué es el análisis de regresión?

- Metodología estadística
- Estudia y modela la relación entre dos o más variables **cuantitativas**
- Diferencia entre **roles** de variables: respuesta y explicativa
- Utiliza la **relación** entre las variables cuantitativas de tal manera que **una de las variables** (respuesta) puede **predecirse** a través de los valores observados de la o las otras (explicativas).

Análisis de regresión

- **Ejemplos**

✓ ¿El **consumo de grasas** está relacionado con la **cantidad de colesterol** ?

✓ ¿Cómo influyen en la **nota final** de una materia las **notas parciales,** la **cantidad de trabajos prácticos entregados,** el **número de clases a las que el alumno no asiste** y el **sexo**?

✓¿Las **ventas** de un producto dependen de la **inversión en publicidad realizada**?

Análisis de regresión

- **Relación no implica causalidad**

Que la relación entre las variables sea muy fuerte no significa que una de ellas sea la causa de la otra.

Ejemplo: peso – altura

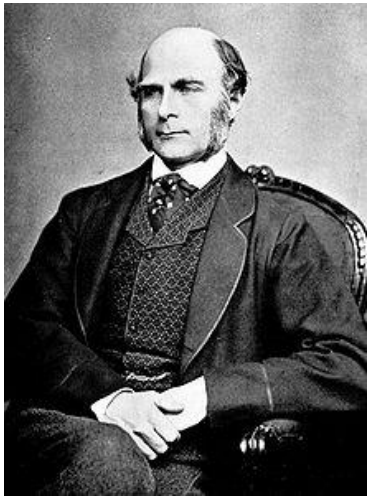
n° de palabras de un libro – n° de páginas de un libro

Análisis de regresión

- **Un poco de historia….**

✓**A principios del siglo XIX:** Legendre y Gauss presentaron publicaciones del método de mínimos cuadrados en aplicaciones en astronomía.

✓**A finales del siglo XIX,** Francis Galton estudió la relación entre la altura de padres e hijos. Observó que las alturas de los hijos de padres altos tienden a ser altas pero no tan altas como la de sus padres, "regresaban" hacia la media. A este fenómeno lo llamó "***regresión a la mediocridad***" y es de aquí surgió el término **regresión.**



Análisis de regresión

- **Objetivos y usos**

- ✓ **Describir** el fenómeno que se está estudiando
- ✓ **Estimar** los parámetros de un modelo
- ✓ **Predecir** los valores de una variable en función de otras variables
- ✓ **Controlar** si la relación entre las variables sigue siendo la misma

Análisis de regresión

- **SIMPLE:** sólo considera una sola variable explicativa

- **MÚLTILPLE:** considera más de una variable explicativa

Unidad 2

Regresión lineal simple.

- Análisis exploratorio.
- Planteo formal del modelo.
- Estimación y test de hipótesis.
- Partición de la suma de cuadrados total (ANOVA).
- Medidas descriptivas de la relación entre las variables en un modelo de regresión.
- Evaluación de los supuestos.

Análisis exploratorio

- **Univarido:**

Tabla de distribución de frecuencias

Gráficos

Medidas

- **Bivariado:**

Gráfico de dispersión

Coeficiente de correlación de Pearson

Ejemplo

Se desea evaluar la relación entre la **nota obtenida en el examen parcial** y la **nota obtenida en el examen final**. Los datos de 11 alumnos de un curso se presentan en la siguiente tabla:

Alumno	Nota final	Nota parcial
1	76	70
2	99	100
3	66	65
4	92	90
5	69	70
6	80	85
7	71	70
8	51	50
9	50	50
10	80	90
11	75	81

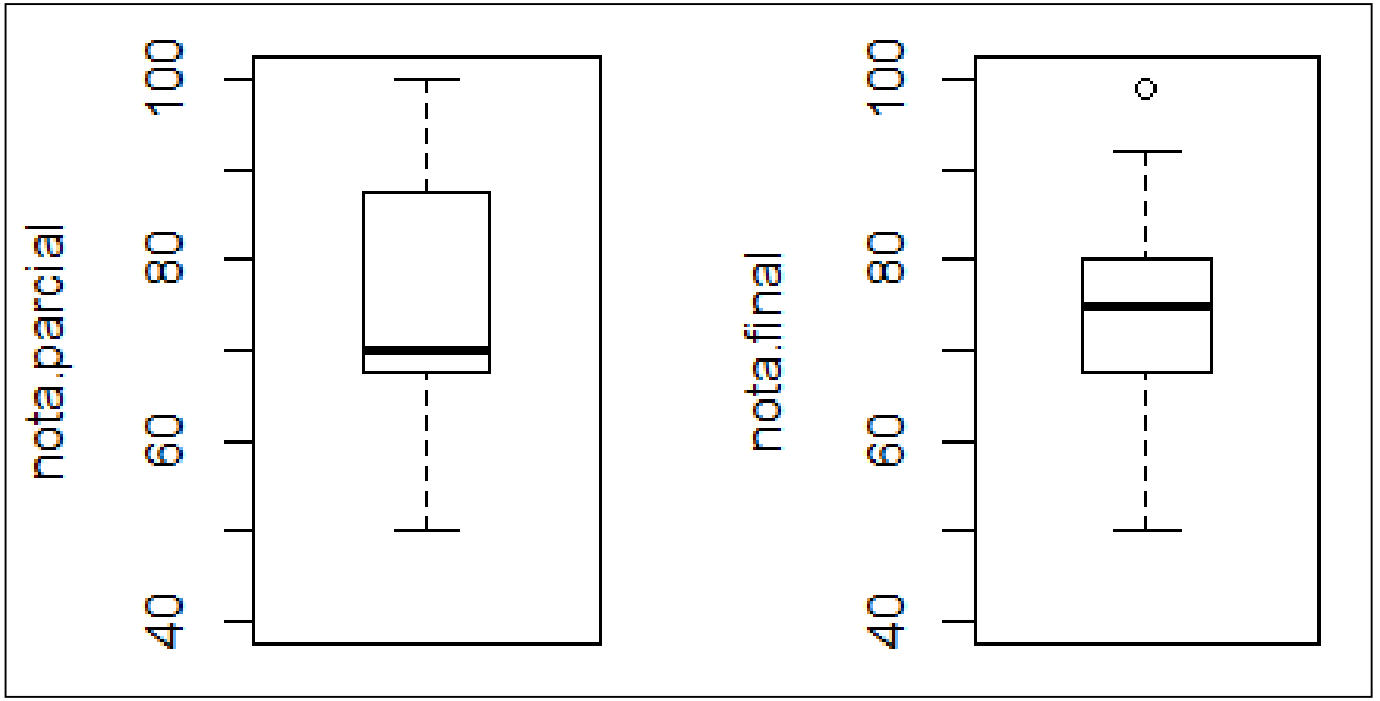
Ejemplo

- ¿Cuál es la unidad análisis?
- ¿Cuántas unidades se consideran?
- ¿Cuál es la variable respuesta “Y”?
- ¿Cuál es la variable explicativa “X”?

Análisis exploratorio: Ejemplo

- **Univariado:**

	Mínimo	Máximo	Media	Desvío
Nota parcial	50	100	74,64	16,23
Nota final	50	99	73,55	14,92

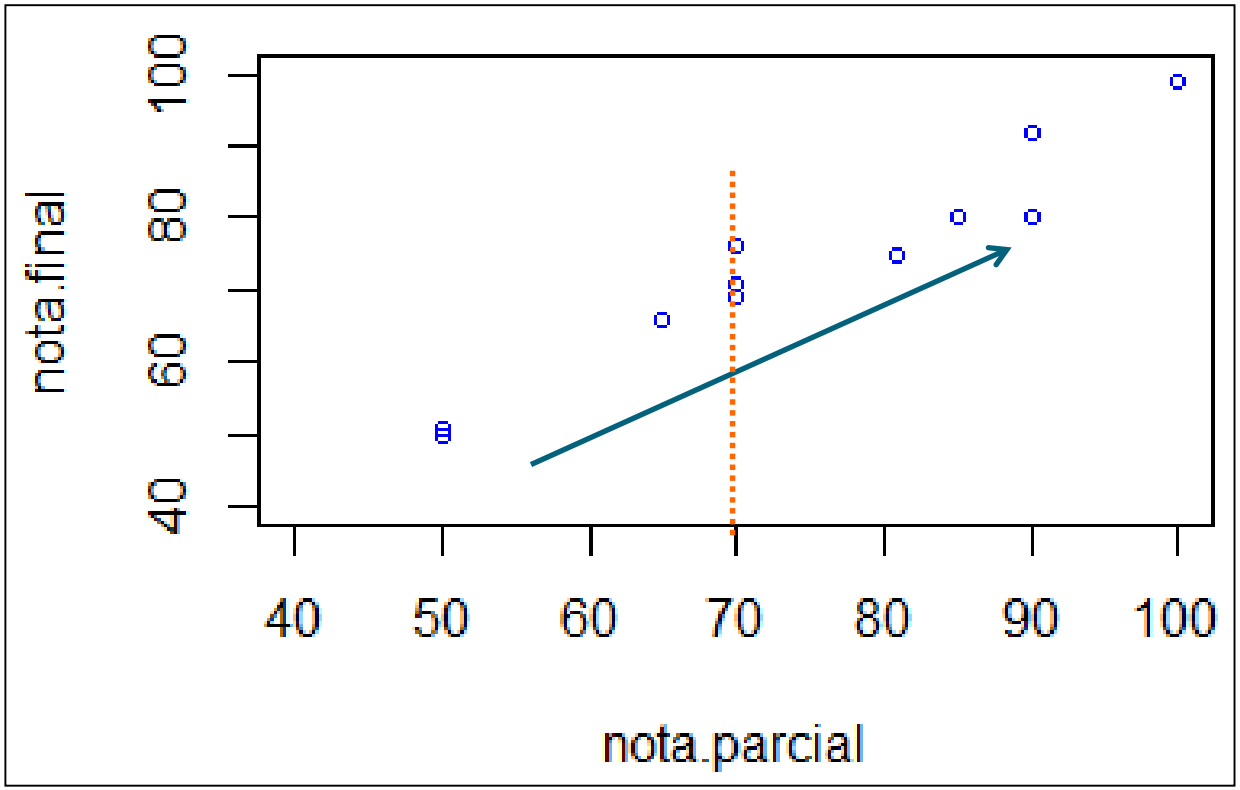


Análisis exploratorio: Ejemplo

- **Bivariado:**

Coef de regresión lineal
0.96

Gráfico de dispersión



Tendencia

A medida que la nota en el examen parcial aumenta, la nota en el examen final también .

Para una misma nota obtenida en el examen parcial, puede haber varios resultados de notas obtenidas en el examen final

Unidad 2

Regresión lineal simple.

- Análisis exploratorio.
- Planteo formal del modelo.
- Estimación y tests de hipótesis.
- Partición de la suma de cuadrados total (ANOVA).
- Medidas descriptivas de la relación entre las variables en un modelo de regresión.
- Evaluación de los supuestos.

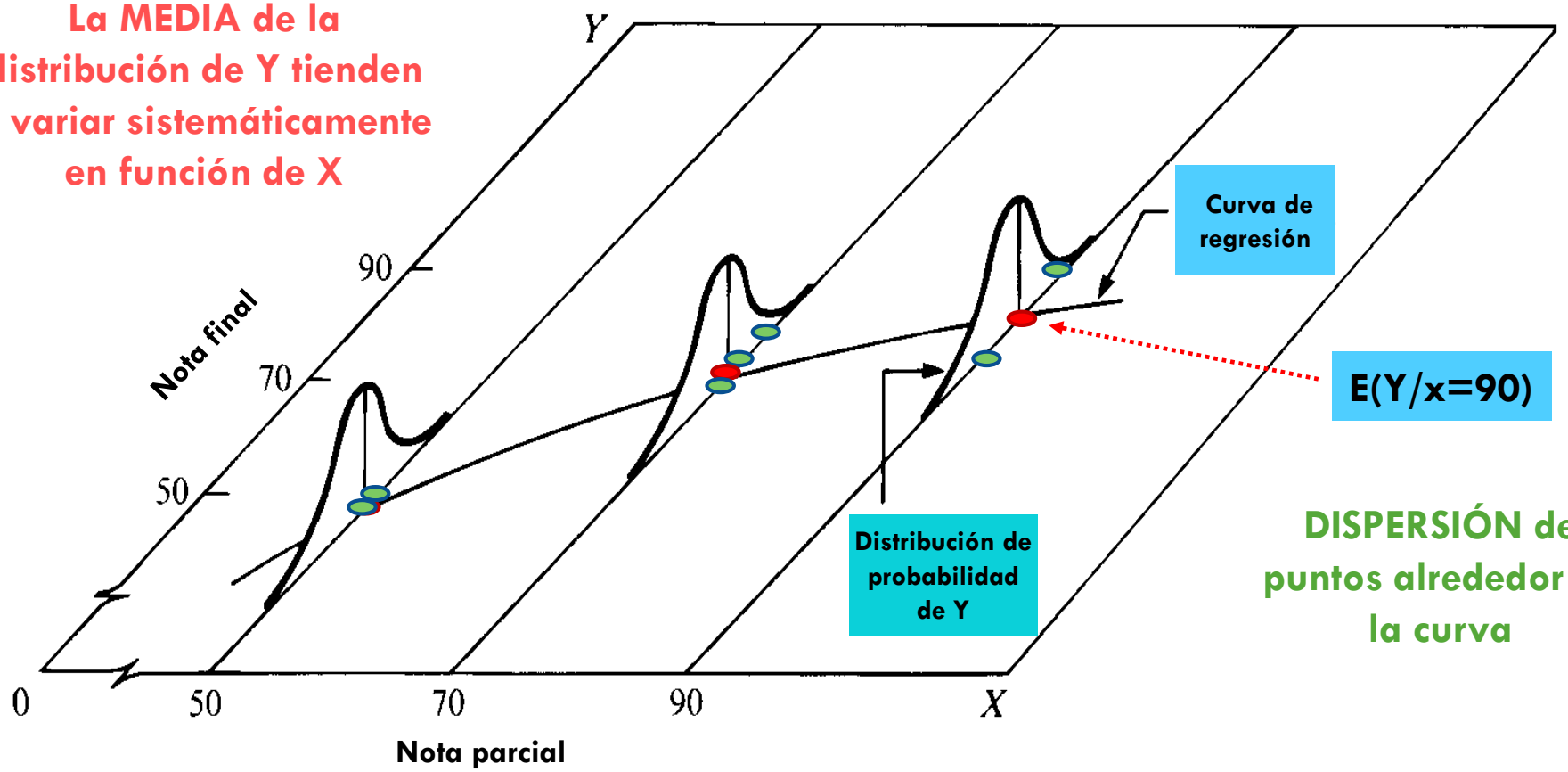
Planteo formal del modelo

- Sean x_i (i=1,2,..,n) los distintos valores de la variable X.
- **Estos valores son seleccionados por el investigador.**
- **Para cada uno** de ellos existe una **distribución** de **probabilidad** de valores de la respuesta **Y** que está caracterizada por su Esperanza y Variancia.

Planteo formal del modelo

Representación pictórica de un modelo de regresión

La MEDIA de la distribución de Y tienden a variar sistemáticamente en función de X



DISPERSIÓN de puntos alrededor de la curva

Planteo formal delo modelo

- **Modelo de regresión**

Es un medio formal para expresar los dos elementos esenciales de una relación estadística:

- ✓ La **tendencia** de la respuesta a variar con las variables explicativas de una manera sistemática
- ✓ Una **dispersión** de puntos alrededor de una curva de la relación estadística

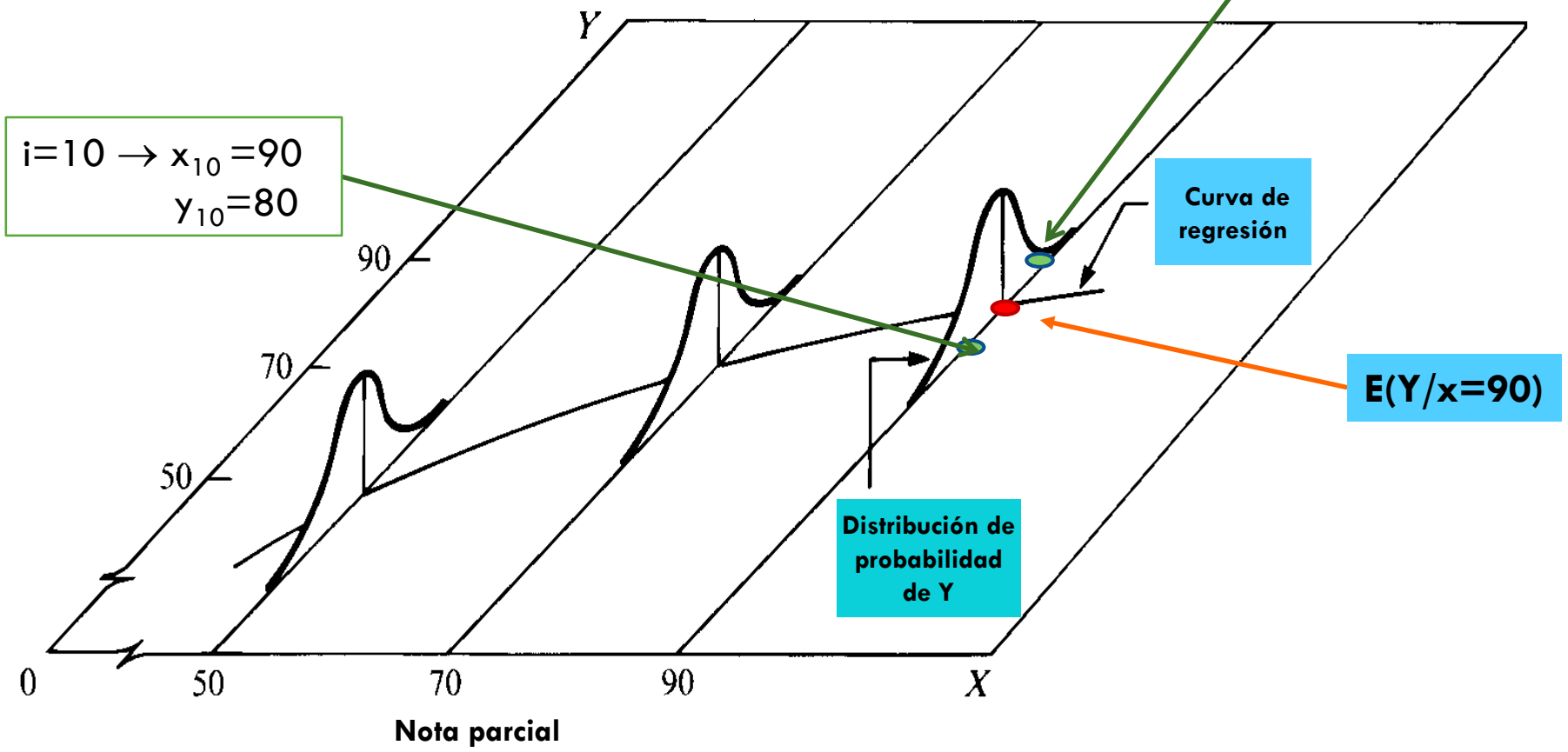
Estas dos características se manifiestan en el modelo de regresión postulando que:

- ✓ Hay una distribución de probabilidad de Y para cada X=x
- ✓ La media de esa distribución varía de alguna manera sistemática con x.

Planteo formal del modelo - Ejemplo

Caso X=90

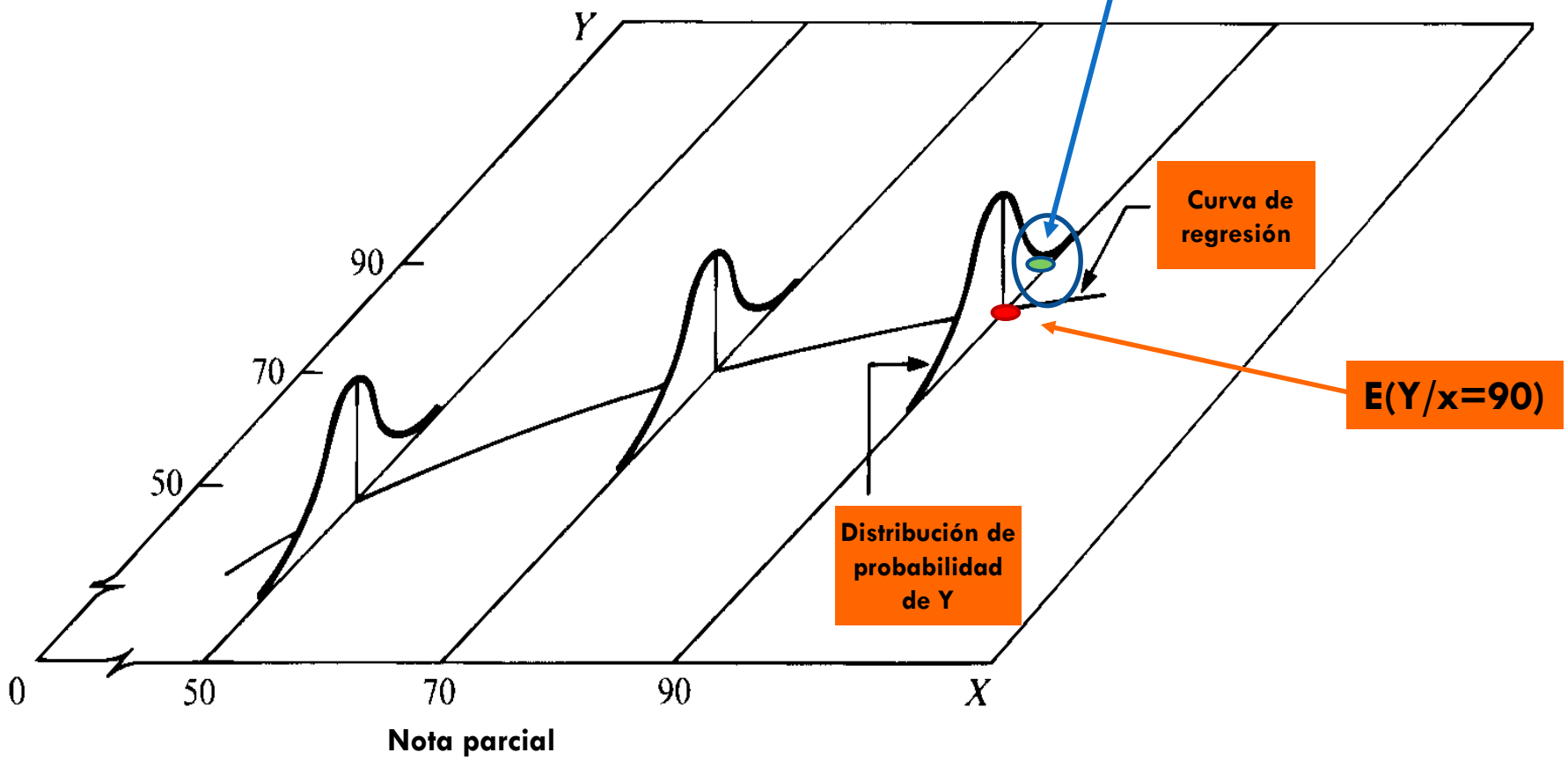
Hubo **dos alumnos** que obtuvieron 90 puntos en la nota parcial pero diferentes notas en el examen final:



Planteo formal del modelo - Ejemplo

¿Cómo se puede expresar el valor y_4=92?

$y_4 = E(y/x=90) + \varepsilon_4$



Planteo formal del modelo - Ejemplo

La **nota final de un alumno,** dada su nota parcial, se puede escribir como:

$$y_i = E(y_i/x_i) + \varepsilon_i$$

la **nota final promedio** de todos los alumnos que comparten esa nota parcial

cantidad aleatoria, que indica esa discrepancia.

Planteo formal del modelo

Si la **relación** entre E(yi/xi) y xi es aproximadamente **lineal,** entonces se puede elegir la **ecuación de una recta** para modelar la relación.

$$y_i = E(y_i/x_i) + \varepsilon_i \qquad i=1,2,\ldots n$$

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

y_i: valor de la variable respuesta para la unidad i

x_i: valor de la variable respuesta para la unidad i

β_0 y β_1: coeficientes de regresión (parámetros)

ε_i: error aleatorio

Se supone que: $E(\varepsilon_i)=0$, $V(\varepsilon_i)=\sigma^2$ $\forall$ i, y ε_i y $\varepsilon_{i'}$ no están correlacionados

Planteo formal del modelo

$$y_i = E(y_i/x_i) + \varepsilon_i$$

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

Este modelo es:

- ✓ **Simple:** sólo hay una variable predictora
- ✓ **Lineal en los parámetros:** ningún parámetro aparece elevado a ningún exponente ni está multiplicado o dividido por otro parámetro
- ✓ **Lineal en la variable predictora:** la variable predictora aparece sólo a la primera potencia.

↓

Modelo de primer orden

Características del modelo

- **y_i es una variable aleatoria**

La y_i es la suma de dos componentes:

✓ El término constante $\beta_0+\beta_1 x_i$ SISTEMATICA

✓ El término aleatorio ε_i ALEATORIA

$\Rightarrow$ y_i es una variable aleatoria.

- **La respuesta y_i se aproxima al valor de la función de regresión por la cantidad dada por el error ε_i**

Características del modelo

Media de y_i/x_i: $E(y_i/x_i)=\beta_0+\beta_1 x_i$

$E(y_i/x_i)=E(\beta_0+\beta_1 x_i+\varepsilon_i)=\beta_0+\beta_1 x_i+E(\varepsilon_i)=\beta_0+\beta_1 x_i$

La media de y depende del valor de x.

Si $\beta_1>0 \rightarrow E(y_i/x_i)$ crece cuando x crece

Si $\beta_1<0 \rightarrow E(y_i/x_i)$ decrece cuando x decrece

Variancia de y_i/x_i: $V(y_i)=\sigma^2 \ \forall i$

$V(y_i/x_i)=V(\beta_0+\beta_1 x_i+\varepsilon_i)=V(\varepsilon_i)=\sigma^2$

El modelo de regresión asume que las distribuciones de Y tienen la misma variancia independientemente del valor de X.

y_i e $y_{i'}$ no están correlacionados $i \neq i'$

Como ε_i y $\varepsilon_{i'}$ están no correlacionados $\Rightarrow$ y_i e $y_{i'}$ tampoco.

Características del modelo

- **Interpretación de los parámetros**

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

✓ β_1 : **pendiente** de la línea de regresión.

Indica el **cambio** en la **media** de la distribución de probabilidad de Y por **incrementos unitarios** en X.

✓ β_0 : **ordenada al origen** de la línea de regresión.

Cuando el alcance del modelo incluye a x=0, β_0 da la media de la distribución de probabilidad de Y en x=0.

En caso contrario β_0 no tiene un significado particular como un término separado del modelo.

Unidad 2

Regresión lineal simple.

- Análisis exploratorio.
- Planteo formal del modelo.
- Estimación y tests de hipótesis.
- Partición de la suma de cuadrados total (ANOVA).
- Medidas descriptivas de la relación entre las variables en un modelo de regresión.
- Evaluación de los supuestos.

Unidad 2

Regresión lineal simple.

- Estimación y tests de hipótesis.
  - Estimación puntual de los parámetros del modelo
  - Estimación por IC y test de hipótesis de β_1 y β_0
  - Estimación puntual y por IC de la respuesta media y predicción de nueva observaciones

¿Cómo estimamos el modelo?

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \qquad i=1,2,\ldots n$$

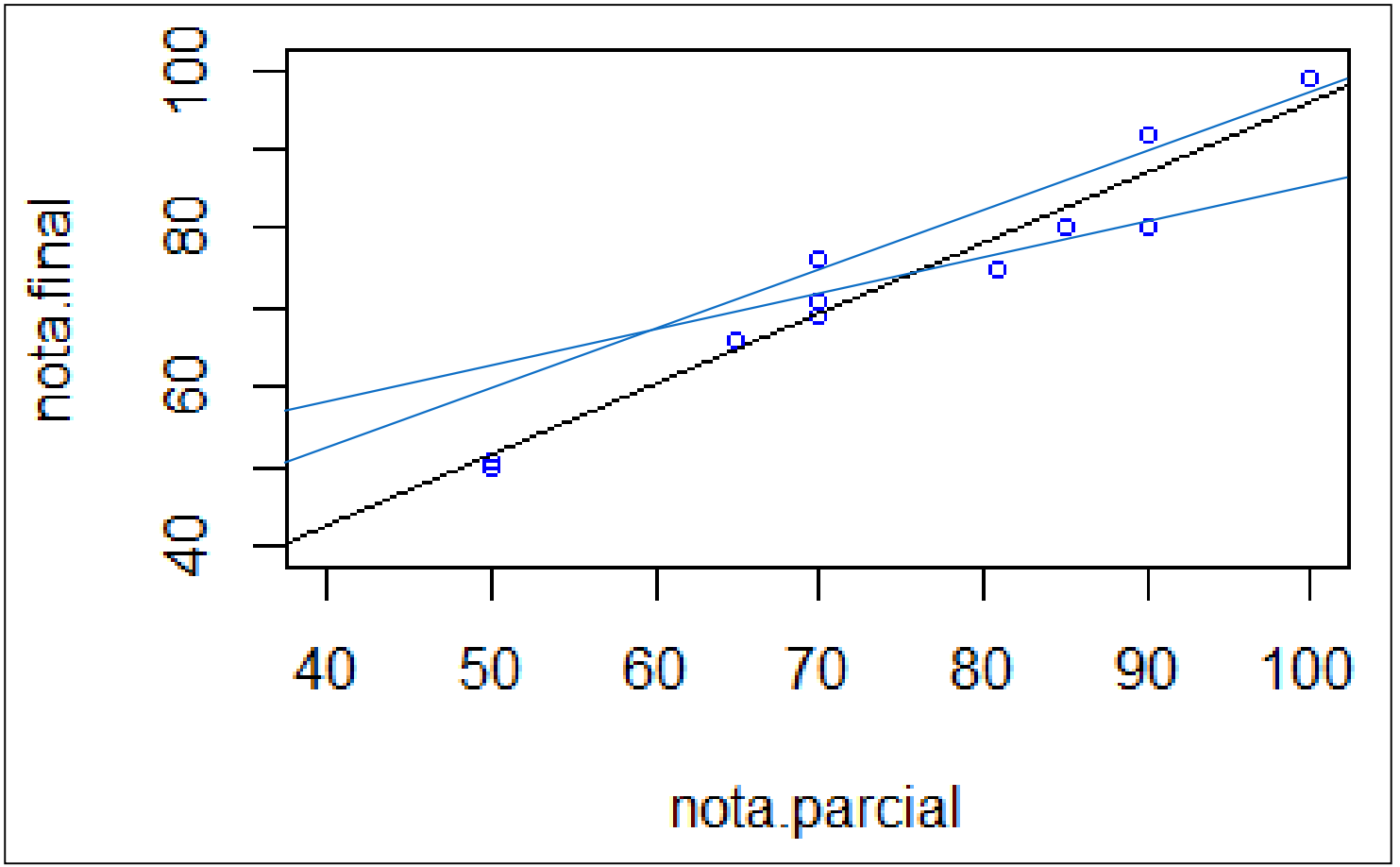
Parámetros del modelo (desconocidos)

β_0 y β_1: coeficientes de regresión

σ^2:variancia del error aleatorio

Para estimar el modelo primero necesitamos estimar sus parámetros

¿Cuál será la mejor recta de regresión?



¿Cuál será la mejor recta de regresión?

- **Método de mínimos cuadrados**

Mejor recta de regresión

Aquella que diste lo menos posible de todos los puntos. Es decir, que la diferencia entre los valores observados y la recta sea mínima.

$$y_i - E(y_i/x_i) = y_i - (\beta_0 + \beta_1 x_i) \quad \text{MINIMA}$$

Son desconocidos

Estimación puntual de los coeficientes de regresión

- **Método de mínimos cuadrados**

Función a **minimizar:** $Q = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$

Se hallan aquellos de valores β_0 y β_1 que minimizan Q: b_0 y b_1.

b_0 y b_1 son los estimadores de β_0 y β_1 $\left(\hat{\beta}_0 = b_0, \hat{\beta}_1 = b_1\right)$.

¿Cómo se obtienen?

- ✓ Se **deriva** Q con respecto a β_0 **y** β_1
- ✓ Cada derivada se la **iguala a cero**
- ✓ Se **despeja** b_0 y b_1

Estimación puntual de los coeficientes de regresión

- **Método de mínimos cuadrados**

Estimador de β_0

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

$$-2\sum_{i=1}^{n} (y_i - b_0 - b_1 x_i) = 0$$

$$\sum_{i=1}^{n} y_i = n b_0 + b_1 \sum_{i=1}^{n} x_i$$

Estimador de β_1

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

$$-2\sum_{i=1}^{n} x_i (y_i - b_0 - b_1 x_i) = 0$$

$$\sum_{i=1}^{n} x_i y_i = b_0 \sum_{i=1}^{n} x_i + b_1 \sum_{i=1}^{n} x_i^2$$

Ecuaciones normales

Estimación puntual de los coeficientes de regresión

- **Método de mínimos cuadrados**

Estimador de β_0

$$b_0 = \bar{y} - b_1 \bar{x}$$

Estimador de β_1

$$b_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

$$b_1 = \frac{\sum_{i=1}^{n} x_i y_i - n\,\bar{x}\,\bar{y}}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}$$

Estimación puntual de los coeficientes de regresión

- **Propiedades de los estimadores mínimos cuadrados**

✓ **Insesgados:** $E(b_0)=\beta_0$ y $E(b_1)=\beta_1$

✓ De **variancia mínima** entre todos los estimadores lineales insesgados.

Estimación puntual de los coeficientes de regresión - Ejemplo

- **Estimación puntual de los coeficientes de regresión**

✓ **$b_1 = 0{,}886$**

A medida que la nota obtenida en el parcial aumenta en 1 punto, la nota final **promedio** aumenta en 0,886 puntos.

O bien..

A medida que la nota obtenida en el parcial aumenta en 10 puntos, la nota final **promedio** aumenta en 8,86 puntos.

✓ **$b_0 = 7{,}418$**

La nota final promedio para un alumno que obtuvo 0 puntos en la nota parcial es 7,418 puntos. **¿Tiene sentido interpretarlo?**

Estimación puntual de la variancia de y

- **Estimador de** $\sigma^2 = Var(\varepsilon_i) = Var(y_i)$

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-2} = \frac{\sum_{i=1}^{n} e_i^2}{n-2} = \frac{SCE}{n-2} = CME$$

Se pierden 2 grados de libertad por las ***estimaciones*** *de* β_0 *y* β_1

Cada tiene su propia distribución de probabilidad y por lo tanto, su propio estimador de la media

Recordemos...

$$s^2 = \frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-1}$$

- **Propiedades de el estimador de la variancia**

✓ **Insesgado:** $E\left(\hat{\sigma}^2\right) = \sigma^2$

Estimación puntual de la variancia de y - Ejemplo

Estimación de la variancia de y

$$\hat{\sigma}^2 = CME = 17,63$$

Unidad 2

Regresión lineal simple.

- Estimación y tests de hipótesis.
  - Estimación puntual de los parámetros del modelo
  - Estimación por IC y test de hipótesis de β_1 y β_0
  - Estimación puntual y por IC de la respuesta media y predicción de nueva observaciones

Intervalo de confianza – Test de hipótesis

- Para realizar **intervalos de confianza** (IC) y **pruebas de hipótesis** se necesita un **supuesto** sobre la forma de la distribución de los ε_i.

$$\varepsilon_i \sim N\left(0, \sigma^2\right)$$

$$y_i \sim N\left(\beta_0 + \beta_1 x_i, \sigma^2\right)$$

Intervalo de confianza – Test de hipótesis

Respecto a β_0

- **Distribución muestral de b_0**

  - ✓ Se refiere a los **diferentes valores de b_0** obtenidos extrayendo muestras de tamaño n, manteniendo constantes los valores de X de muestra a muestra.

  - ✓ $b_0 \sim N(\beta_0, V(b_0))$

$$V(b_0) = \sigma^2 \left[\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \right]$$

Intervalo de confianza – Test de hipótesis

Respecto a β_0

- **Intervalo de confianza**

$$IC_{\beta_0,(1-\alpha)100\%} = \left(b_0 - t_{n-2,1-\frac{\alpha}{2}} s(b_0), b_0 + t_{n-2,1-\frac{\alpha}{2}} s(b_0) \right)$$

donde $$s(b_0) = \sqrt{\hat{V}(b_0)} = \sqrt{CME\left[\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right]}$$

Intervalo de confianza – Test de hipótesis

Respecto a β_0

- **Prueba de hipótesis**

Hipótesis: $H_0) \ \beta_0 = 0 \qquad H_A) \ \beta_0 \neq 0$

Estadística de prueba: $$t = \frac{b_0}{s(b_0)} \overset{H_0}{\sim} t_{n-2}$$

Regla de decisión: rechazo H_0 si $|t_{obs}| > t_{n-2;1-\frac{\alpha}{2}}$ o valor $p \leq 0,05$

Intervalo de confianza – Test de hipótesis

Respecto a β_1

- **Distribución muestral de b_1**

  - ✓ Se refiere a los **diferentes valores de b_1** obtenidos extrayendo muestras de tamaño n, manteniendo constantes los valores de X de muestra a muestra.

  - ✓ $b_1 \sim N(\beta_1, V(b_1))$

$$V(b_1) = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

Intervalo de confianza – Test de hipótesis

Respecto a β_1

- **Intervalo de confianza**

$$IC_{\beta_1,(1-\alpha)100\%} = \left(b_1 - t_{n-2,1-\frac{\alpha}{2}} s(b_1), b_1 + t_{n-2,1-\frac{\alpha}{2}} s(b_1) \right)$$

donde $$s(b_1) = \sqrt{V(b_1)} = \sqrt{\frac{CME}{\sum_{i=1}^{n} (x_i - \bar{x})^2}}$$

Intervalo de confianza – Test de hipótesis

Respecto a β_1

- **Prueba de hipótesis**

Hipótesis: $H_0) \ \beta_1 = 0 \qquad H_A) \ \beta_1 \neq 0$

Test de Regresión

Estadística de prueba: $t = \dfrac{b_1}{s(b_1)} \overset{H_0}{\sim} t_{n-2}$

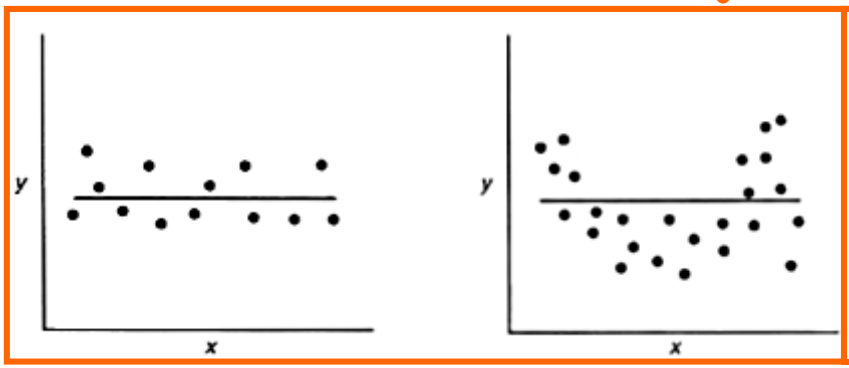
Regla de decisión: rechazo H_0 si $|t_{obs}| > t_{n-2;1-\frac{\alpha}{2}}$ o valor $p \leq 0,05$

Intervalo de confianza – Test de hipótesis

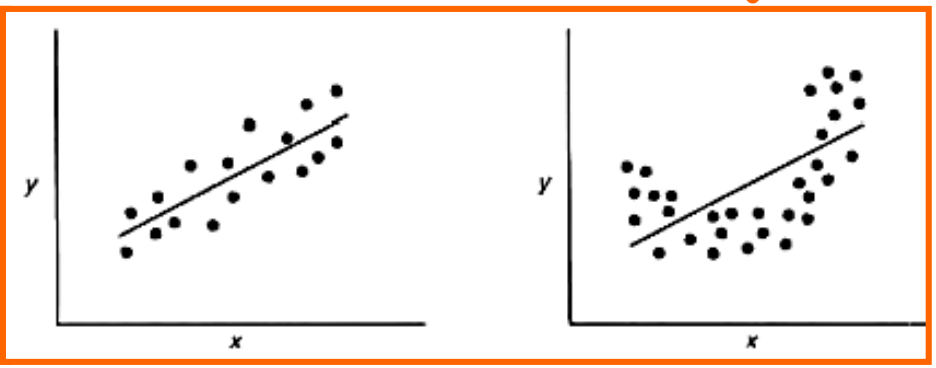
Respecto a β_1

- **Prueba de hipótesis** *Hipótesis*: $H_0) \beta_1 = 0$ $H_A) \beta_1 \neq 0$

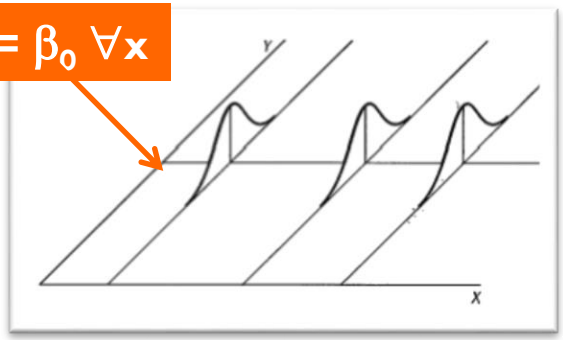
Casos en los que no se rechaza H_0



Casos en los que se rechaza H_0



$E(y_i) = \beta_0 \; \forall x$



Intervalo de confianza - Ejemplo

- **Intervalos de confianza**

$$IC_{\beta_0,95\%} = (7,418 - 2,262 \,.\, 6,235 \;\; ; \;\; 7,418 + 2,262 \,.\, 6,235)$$
$$= (-0,686 \;\; ; \;\; 21,521)$$

Con una confianza del 95% es de esperar que cuando x=0 la media de y se encuentre entre -0,686 y 21,521.

$$IC_{\beta_1,95\%} = (0,886 - 2,262 \,.\, 0,081 \;\; ; \;\; 0,886 + 2,262 \,.\, 0,081)$$
$$= (0,703 \;\; ; \;\; 1,069)$$

Con una confianza del 95% es de esperar que la pendiente de la recta de regresión se encuentre entre 0,703 y 1,069.

Test de hipótesis - Ejemplo

- **Prueba de hipótesis**

Hipótesis: $H_0) \beta_0 = 0 \quad H_A) \beta_0 \neq 0$

$$t_{obs} = 1{,}19 < t_{9;0,975} = 2,262 \Rightarrow \text{no rechazo } H_0$$

Hipótesis: $H_0) \beta_1 = 0 \quad H_A) \beta_1 \neq 0$

Test de Regresión

$$t_{obs} = 10,87 > t_{9;0,975} = 2,262 \Rightarrow \text{rechazo } H_0$$

Unidad 2

Regresión lineal simple.

- Estimación y tests de hipótesis.
  - Estimación puntual de los parámetros del modelo
  - Estimación por IC y test de hipótesis de β_1 y β_0
  - Estimación puntual y por IC de la respuesta media y predicción de nueva observaciones

¿Cómo estimamos la recta de regresión?

Estimador puntual de la media de y

$$\hat{E}(y_i/x_i) = \hat{\beta}_0 + \hat{\beta}_1 x_i = b_0 + b_1 x_i$$

Predictor puntual de y

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i = b_0 + b_1 x_i$$

¿Cómo estimamos la recta de regresión? - Ejemplo

Estimación de la media de y

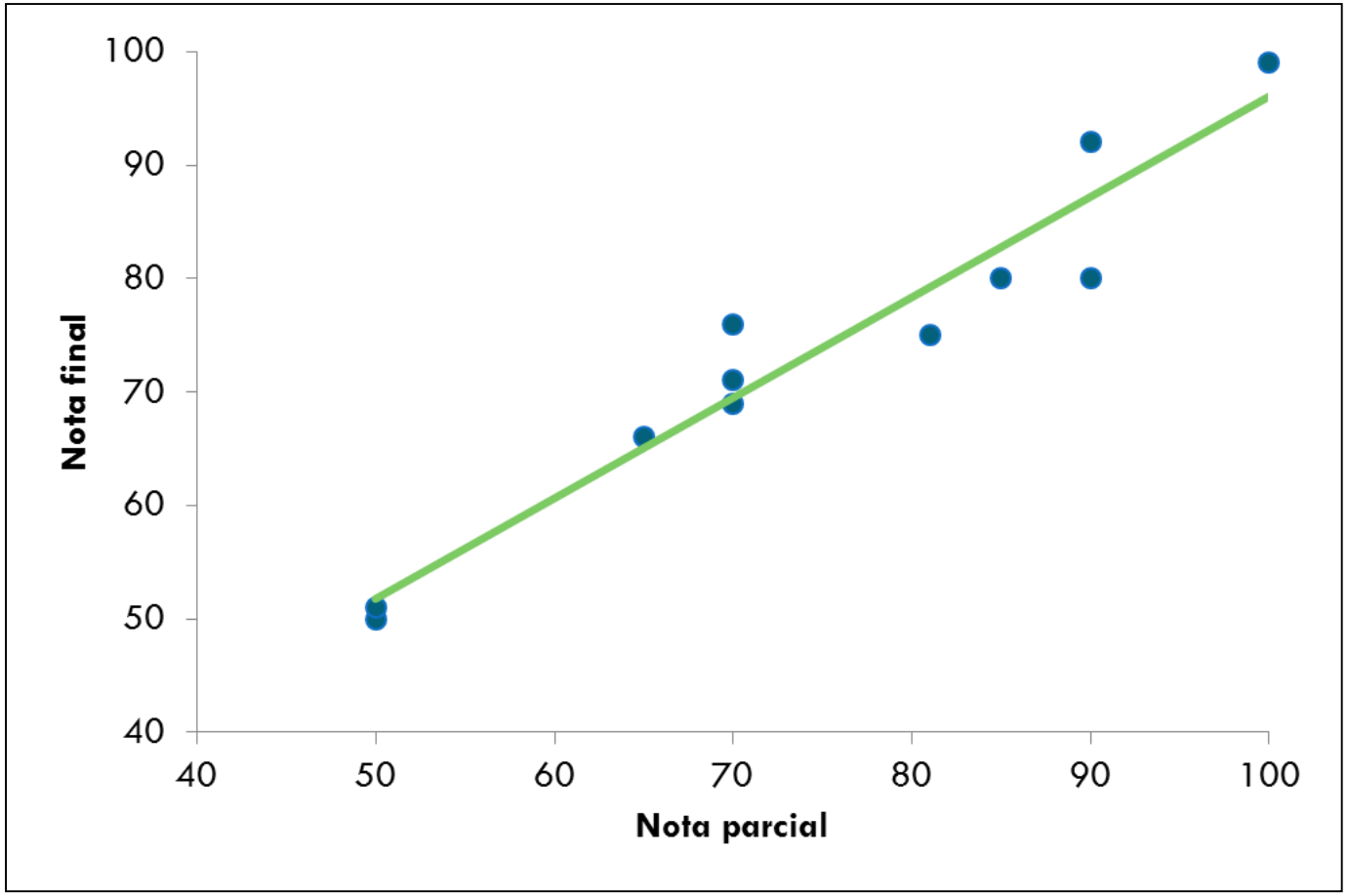
$$\hat{E}(y_i/x_i) = 7,418 + 0,886x_i$$

Predicción de y

$$\hat{y}_i = 7,418 + 0,886x_i$$

¿Cómo estimamos la recta de regresión?: Ejemplo

¿Cuál es la recta de regresión estimada para este ejemplo?



Intervalo de confianza

- **Distribución muestral de** $\hat{E}(y_i/x)$

$$\hat{E}(y_i/x) \sim N\left(E\left[\hat{E}(y_i/x)\right], V\left[\hat{E}(y_i/x)\right]\right)$$

$$E\left[\hat{E}(y_i/x)\right] = E\left[b_0 + b_1 x_i\right] = \beta_0 + \beta_1 x$$

$$V\left[\hat{E}(y_i/x)\right] = V\left[b_0 + b_1 x_i\right] = \sigma^2 \left[\frac{1}{n} + \frac{(x - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right]$$

Intervalo de confianza

- **Intervalo de confianza de** $E(y/x)$

$$IC_{E(y/x),(1-\alpha)100\%} = \left(\hat{E}(y/x) \pm t_{n-2,1-\frac{\alpha}{2}} s\left(\hat{E}(y/x)\right) \right)$$

donde $$s\left(\hat{E}(y/x)\right) = \sqrt{V\left(\hat{E}(y/x)\right)} = \sqrt{CME\left[\frac{1}{n} + \frac{(x-\bar{x})^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}\right]}$$

Estimación puntual y por intervalo de confianza – Ejemplo

¿Qué **nota final** se espera obtener **en promedio** cuando se obtuvo en la prueba parcial 70 puntos?

$$\hat{E}(y/70) = 7,418 + 0,886 . 70 = 69,438$$

$$IC_{E(y/70),95\%} = (66,448 \; ; \; 72,427)$$

Intervalo de predicción

- **Intervalo de predicción de $\hat{y}_i/x$**

$$IP_{\hat{y}/x,(1-\alpha)100\%} = \left((\hat{y}/x) \pm t_{n-2,1-\frac{\alpha}{2}} s\left((\hat{y}/x)\right) \right)$$

donde $$s\left((\hat{y}/x)\right) = \sqrt{V\left((\hat{y}/x)\right)} = \sqrt{CME\left[1 + \frac{1}{n} + \frac{(x-\bar{x})^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}\right]}$$

Predicción puntual y por intervalo – Ejemplo

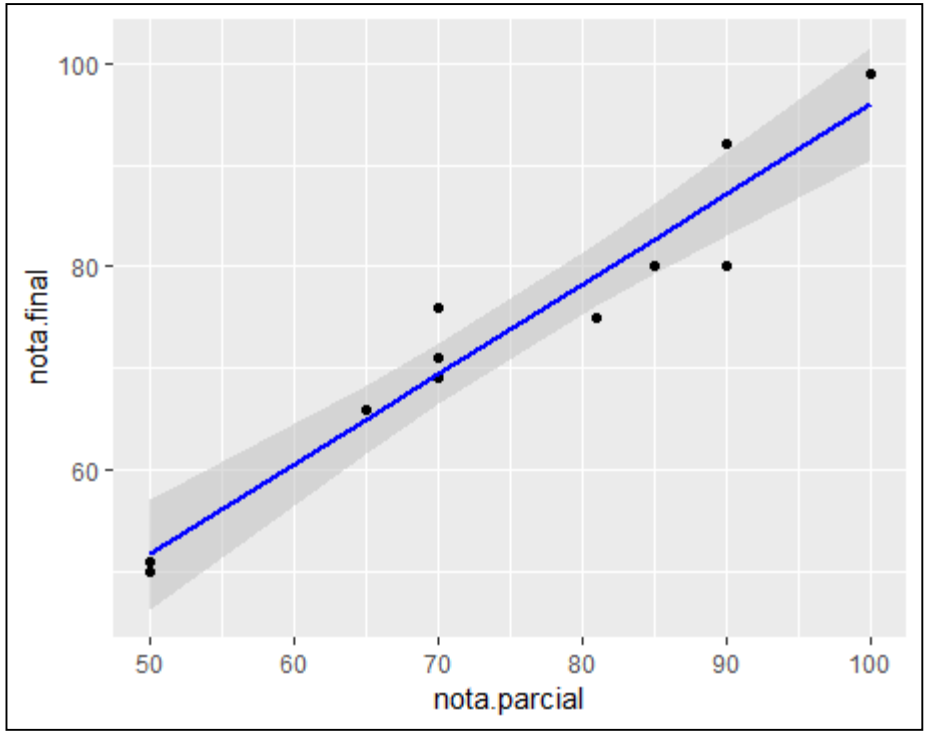
¿Qué **nota final** se espera obtener cuando se obtuvo en la prueba parcial 70 puntos?

$$y/70 = 7,418 + 0,886.70 = 69,438$$

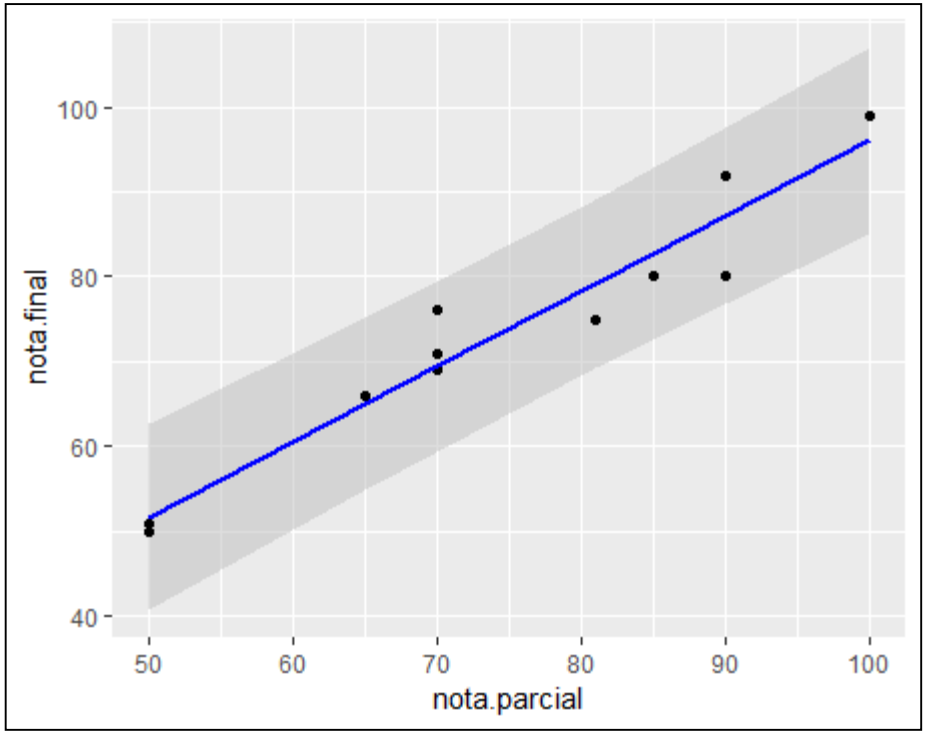
$$IP_{y/70,95\%} = (59,480 \; ; \; 79,395)$$

Intervalos - Ejemplo

Bandas de confianza del 95%



Bandas de predicción del 95%



Unidad 2

Regresión lineal simple.

- Análisis exploratorio.
- Planteo formal del modelo.
- Estimación y tests de hipótesis.
- Partición de la suma de cuadrados total (ANOVA).
- Medidas descriptivas de la relación entre las variables en un modelo de regresión.
- Evaluación de los supuestos.

ANOVA

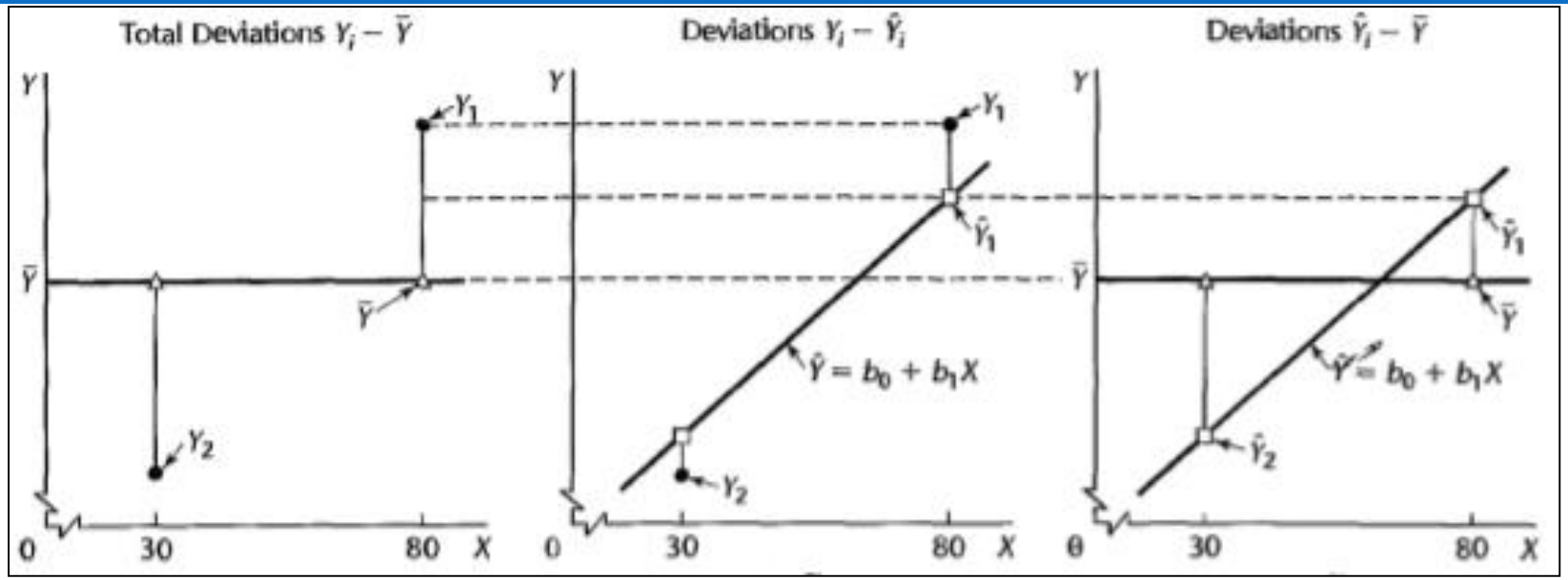
- **El enfoque del Análisis de la Variancia (ANOVA) para el Análisis de Regresión**

  - ✓ Consiste en **particionar** la **variabilidad** total de la **respuesta** en distintas **componentes.**

  - ✓ La **variación total** de la respuesta Y se piensa como la **desviación** de cada y_i respecto de la media . $\overline{y}$

  - ✓ Se utiliza como una medida de la **variación de las observaciones,** sin tener en cuenta la variable explicativa X.

ANOVA



$$y_i - \bar{y} = (y_i - \hat{y}_i) + (\hat{y}_i - \bar{y})$$

Desviación total	Desviación de cada observación respecto al valor ajustado	Desviación de cada valor ajustado respecto a la media

ANOVA

$$\sum_{i=1}^{n}(y_i-\bar{y})^2=\sum_{i=1}^{n}\left[(y_i-\hat{y}_i)+(\hat{y}_i-\bar{y})\right]^2=$$

$$=\sum_{i=1}^{n}\left[(y_i-\hat{y}_i)^2+(\hat{y}_i-\bar{y})^2+2(y_i-\hat{y}_i)(\hat{y}_i-\bar{y})\right]=$$

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

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

$$=\underbrace{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}_{\text{Variabilidad residual no explicada por la regresión}}+\underbrace{\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2}_{\text{Variabilidad explicada por la regresión}}$$

Variabilidad Total

Cuadro ANOVA

Fuente de variación	Suma de cuadrados	Grados de libertad	Cuadrados medios	E(CM)
Regresión ajustada	$SCR_m = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$	1	$SCR_m = CMR_m$	$\sigma^2 + \beta_1^2 \sum_i (x_i - \bar{x})^2$
Error	$SCE = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$	n-2	$CME = \frac{SCE}{n-2}$	σ^2
Total corregido	$SCT_m = \sum_{i=1}^{n}(y_i - \bar{y})^2$	n-1		

ANOVA

- **Test de regresión a partir del análisis de la variancia**

Hipótesis: $H_0)\ \beta_1 = 0 \qquad H_A)\ \beta_1 \neq 0$

Estadística de prueba: $F = \frac{CMR_m}{CME} \underset{H_0}{\sim} F_{1,n-2}$

Distribución F de Fisher- Snedecor

Regla de decisión: rechazo H_0 si $F_{obs} > F_{n-2;\alpha}$ o valor $p \leq 0,05$

¿Por qué no es $\frac{\alpha}{2}$?

Tests de Regresión

- **Relación entre la estadística t y la estadística F de los dos test de regresión**

$$t^2 = F$$

ANOVA - Ejemplo

Fuente de variación	Suma de Cuadrados	Grados de libertad	Cuadrado medio	F
Regresión ajustada	2068,06	1	2068,06	**117,3**
Error	158,66	9	17,63	
Total ajustado	2226,72	10		

Hipótesis: $H_0) \beta_1 = 0 \quad H_A) \beta_1 \neq 0$

$t^2 = (10,83)^2$

Decisión: $F_{obs} = 117,3 > F_{1,9,0.05} = 5,12$

Conclusión: *En base a la evidencia muestral y con un nivel de significación de 0.05 es de esperar que la nota parcial aporte significativamente a la explicación de la nota final*

Unidad 2

Regresión lineal simple.

- Análisis exploratorio.
- Planteo formal del modelo.
- Estimación y tests de hipótesis.
- Partición de la suma de cuadrados total (ANOVA).
- Medidas descriptivas de la relación entre las variables en un modelo de regresión.
- Evaluación de los supuestos.

Medidas descriptivas de asociación

- **Coeficiente de determinación**

$$R^2 = \frac{SCR_m}{SCT_m} = 1 - \frac{SCE}{SCT_m}$$

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

- ✓ **Proporción** de la variación total de Y que es explicada por la relación lineal entre X e Y o por la regresión.
- ✓ Cuanto más cercano sea a 1 → mayor es la asociación lineal entre X e Y

Medidas descriptivas de asociación

- **Coeficiente de correlación**

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}} = b_1 \frac{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}}{\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

$$-1 \le r \le 1$$

$$r = \pm\sqrt{R^2}$$

- ✓ Medida cuantitativa de la **fuerza** y la **dirección** de la **relación lineal** entre las variables X e Y.

En regresión simple:

- ✓ Lleva el **mismo signo** que la **pendiente** de la recta de regresión ajustada

Medidas descriptivas de asociación – Ejemplo

$$R^2 = \frac{2068,06}{2226,72} = 0,93$$

- El 93% de la variabilidad total de Y que es explicada por la relación lineal entre la nota parcial y la nota final.

$$r = \sqrt{0,93} = 0,96$$

- Existe una relación lineal fuerte y positiva entre la nota parcial y la nota final.

Unidad 2

Regresión lineal simple.

- Análisis exploratorio.
- Planteo formal del modelo.
- Estimación y tests de hipótesis.
- Partición de la suma de cuadrados total (ANOVA).
- Medidas descriptivas de la relación entre las variables en un modelo de regresión.
- Evaluación de los supuestos.

Supuestos

- Distribución normal de errores
- Media de errores igual a cero
- Variancia de errores constante
- Errores no correlacionados
- Linealidad de la variable regresora

Residuos

Residuo ordinario

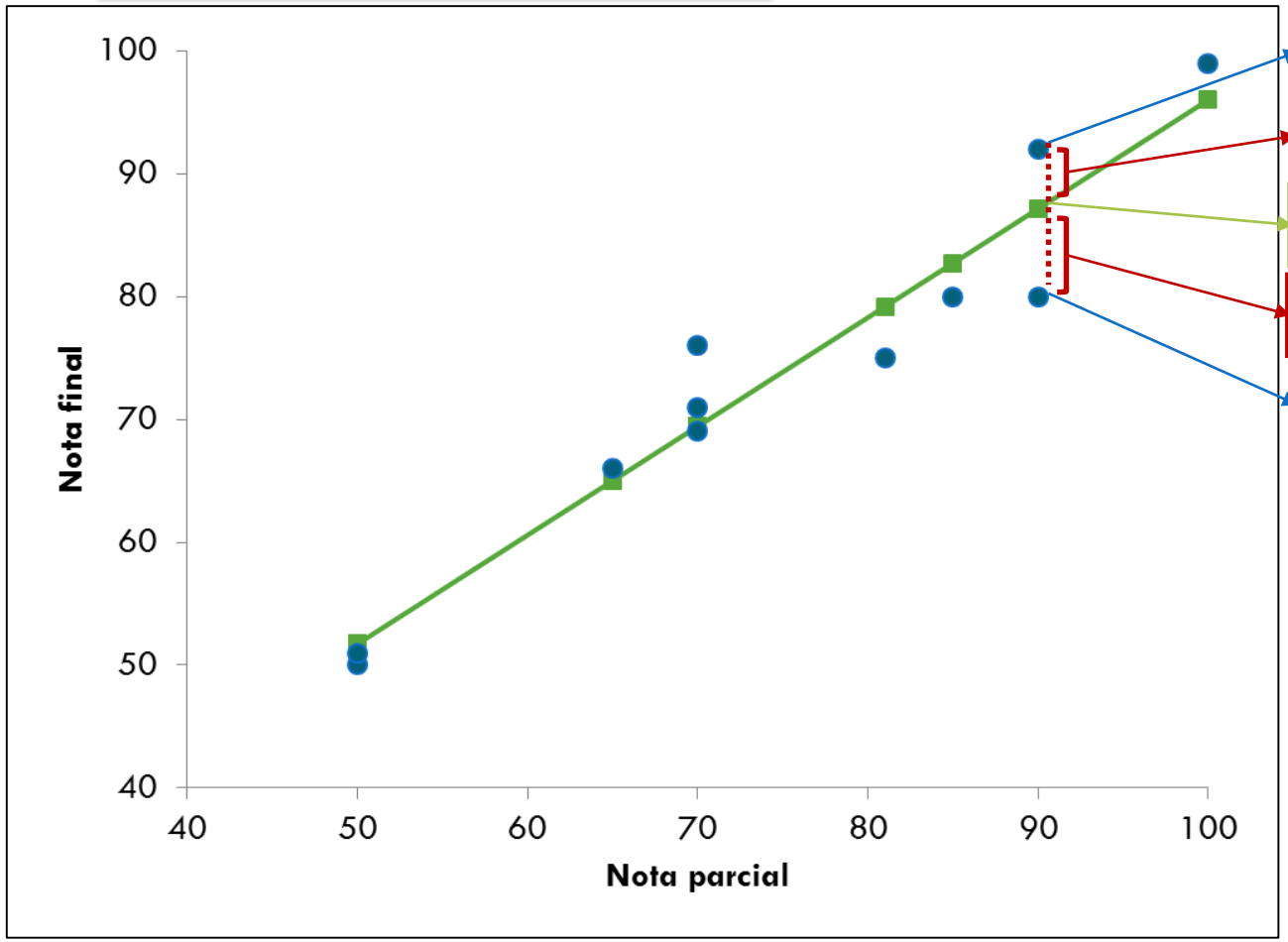
$$e_i = y_i - \hat{y}_i$$

- Diferencia entre el valor observado y el valor estimado
- Estimador del error aleatorio $\varepsilon_i = y_i - E(y_i)$
- Si el modelo es apropiado, los *residuos* van a reflejar las propiedades de los *errores*.

Es por ello que se analizan los residuos para evaluar si se cumplen los supuestos que se hacen sobre los errores

Residuos - Ejemplo

Residuo ordinario



$y_4 = 92$

$e_4 = y_4 - \hat{y}_4 = 4{,}84$

$\hat{y}_{10} = 87{,}16$

$e_{10} = y_{10} - \hat{y}_{10} = -7{,}16$

$y_{10} = 80$

Residuos

Residuo estandarizado o semiestudentizado

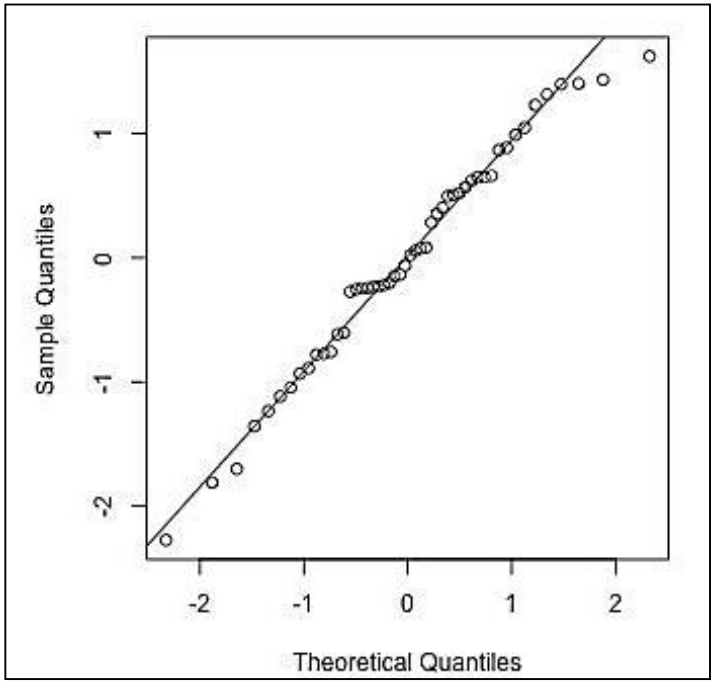
$$e_i^* = \frac{e_i - \bar{e}}{\sqrt{CME}} = \frac{e_i}{\sqrt{CME}}$$

$$CME = \frac{\sum (e_i - \bar{e})^2}{n-2} = \frac{SCE}{n-2}$$

- Estos residuos tienen media cero y la variancia es aprox. 1
- Un residuo estandarizado grande $|e_i| > 3$ muestra un valor atípico

Distribución normal de errores

Gráfico probabilístico normal de e_i^*



Si los puntos se ajusta a la recta

Test de Anderson-Darling (e_i)

H_0) Los errores tienen distribución normal

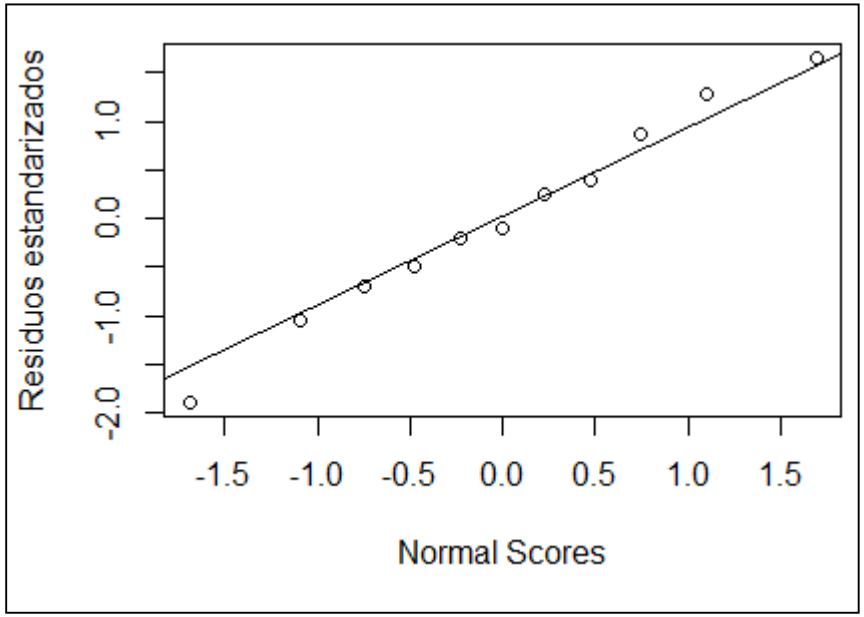
H_1) Los errores no tienen distribución normal

Regla decisión: Si valor p>α

Sugiere que los errores tienen distribución Normal

Distribución normal de errores - Ejemplo

Gráfico probabilístico normal de e_i^*



Test de Anderson-Darling (e_i)

H_0) Los errores tienen distribución normal

H_1) Los errores no tienen distribución normal

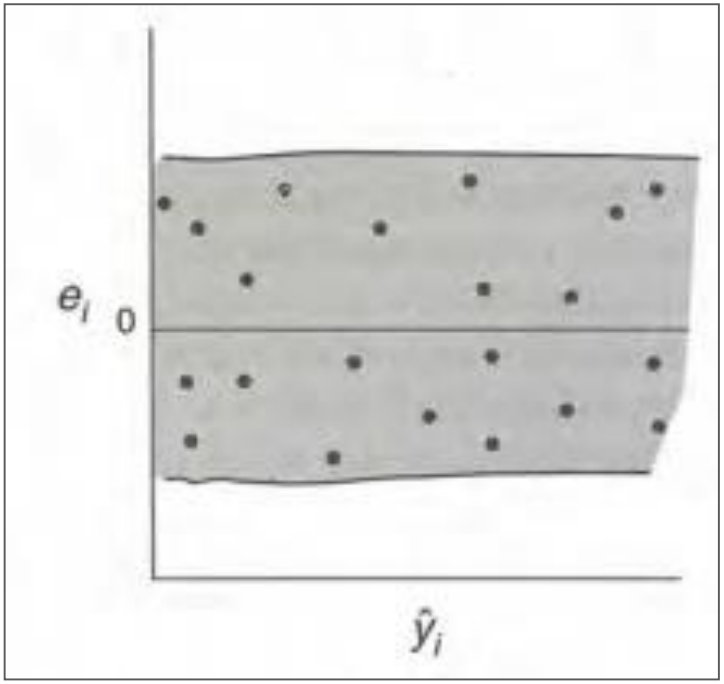
Los puntos se ajusta a la recta

El valor p=0,9949>0,05

Sugiere que los errores tienen distribución Normal

Media de los errores igual a 0 y variancia constante

Gráfico de e_i^* vs valores ajustados

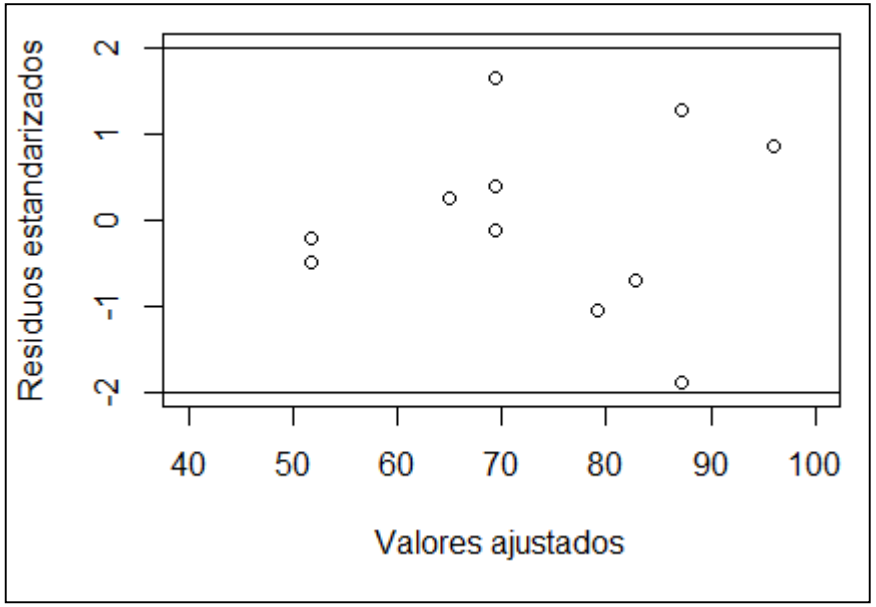


Si los puntos caen dentro de una banda horizontal alrededor del cero

Sugiere que los errores tienen media cero y variancia constante

Media de los errores igual a 0 y variancia constante - Ejemplo

Gráfico de e_i^* vs valores ajustados

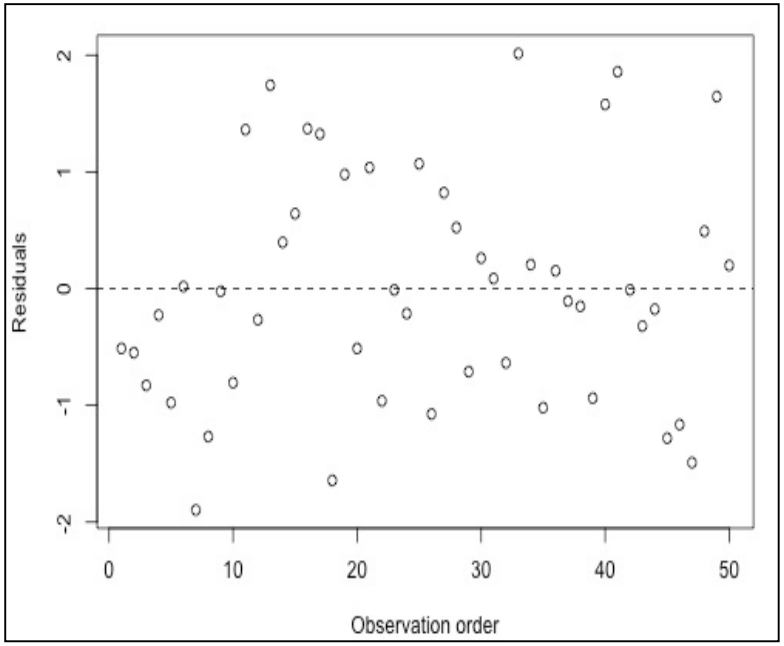


Los puntos caen dentro de una banda horizontal alrededor del cero

Sugiere que los errores tienen media cero y variancia constante

Errores no correlacionados

Gráfico de e_i^* vs secuencia temporal



Si los puntos no muestran un patrón, se presentan aleatoriamente

Test de Durbin- Watson (e_i)

H_0) Los errores no están correlacionados

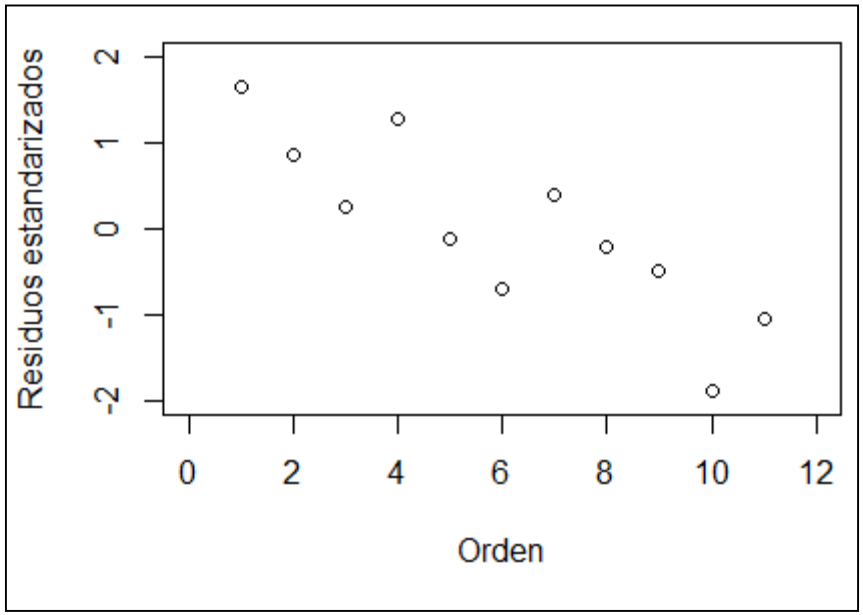
H_1) Los errores están correlacionados

Regla decisión: Si valor $p>\alpha$

Sugiere que los errores no están correlacionados

Errores no correlacionados - Ejemplo

Gráfico de e_i^* vs secuencia temporal



Los puntos muestran un patrón poco claro

Test de Durbin- Watson (e_i)

H_0) Los errores no están correlacionados

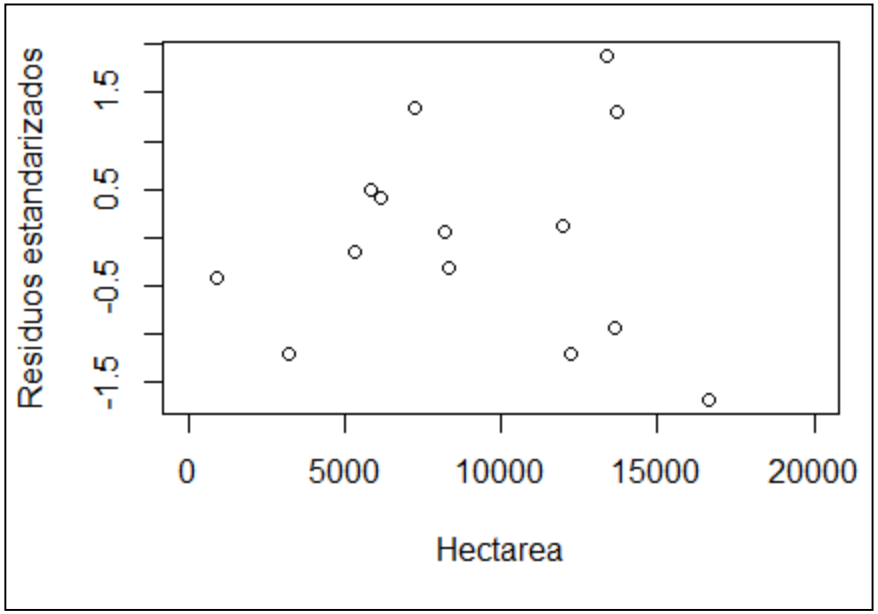
H_1) Los errores están correlacionados

El valor p=0,0115<0,05

Sugiere que los errores están correlacionados

Linealidad de la regresora

Gráfico de e_i^* vs variable explicativa

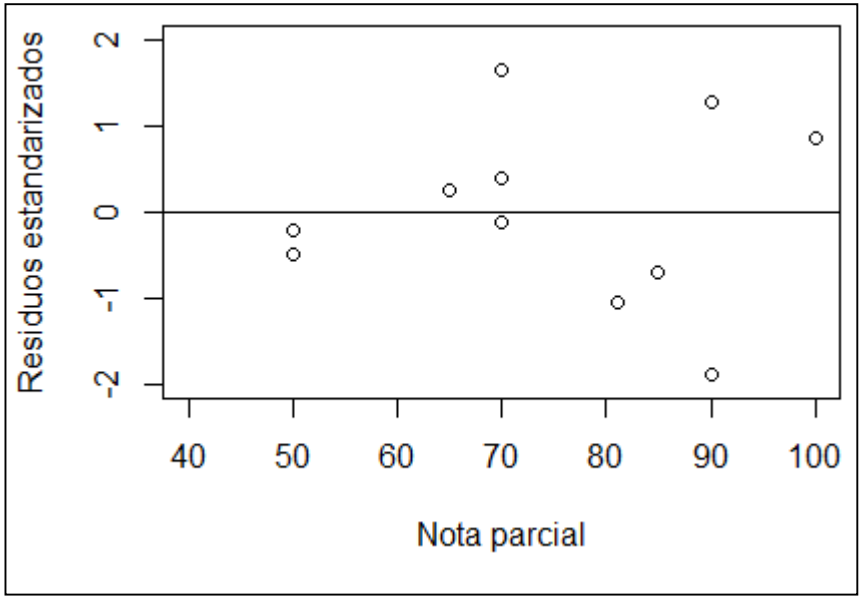


Si los puntos no muestran un patrón, se presentan aleatoriamente

Sugiere que la relación propuesta entre Y y X es correcta

Linealidad de la regresora - Ejemplo

Gráfico de e_i^* vs variable explicativa



Los puntos no muestran un patrón, se presentan aleatoriamente

Sugiere que la relación propuesta entre Y y X es correcta