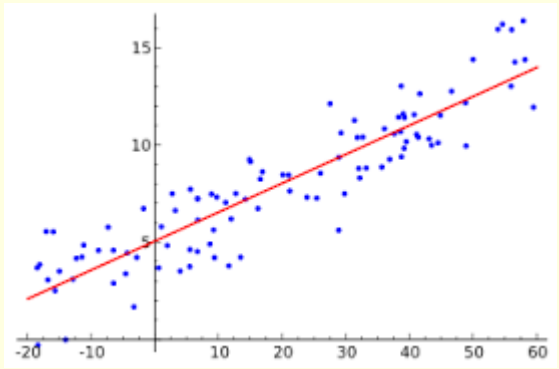
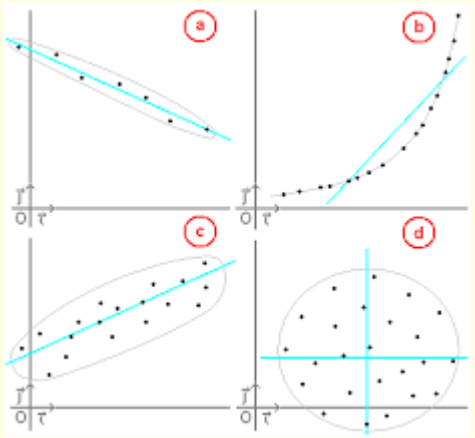
Universidad Autónoma de Estado de México

Facultad de Economía

Unidad de Aprendizaje: Estadística inferencial

Regresión lineal





Material didáctico: visual elaborado por Fidelmar Sandoval Durán

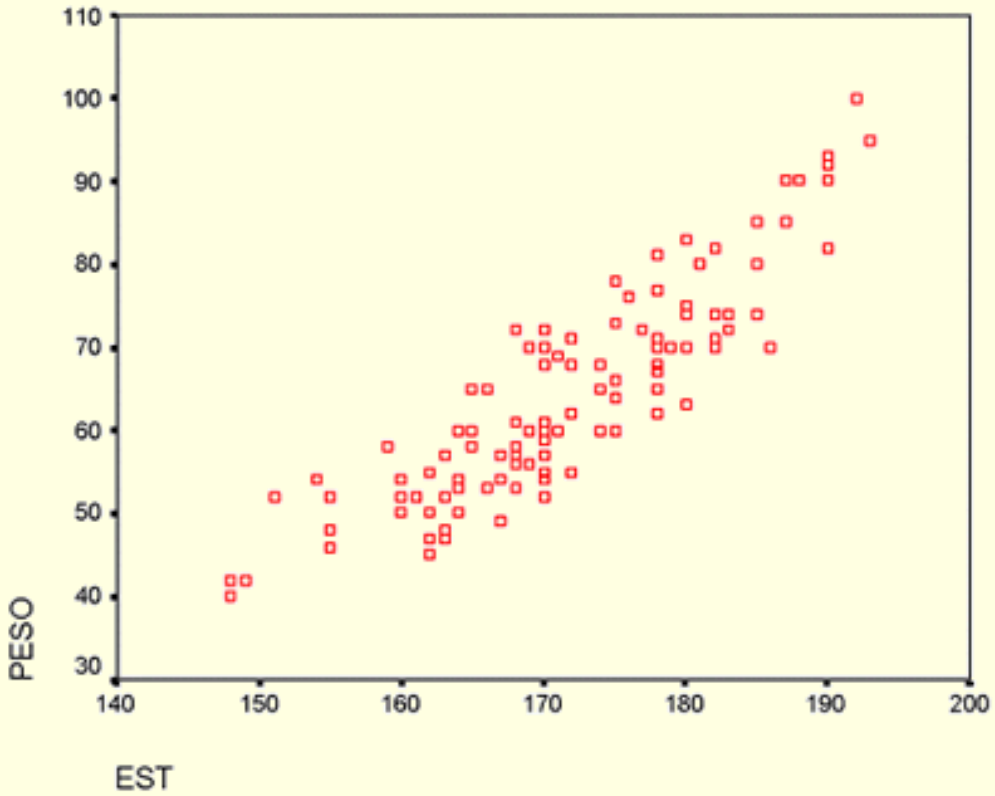
Objetivos

- Conocerá como trazar una línea de regresión
- Sabrá como probar hipótesis acerca de la línea de regresión
- Sabrá como realizar un análisis ANOVA

- Cuando dos variables se relacionan, se debe cuantificar la relación entre ellas.
- Al hacer esto, podemos estimar el valor de una variable, si conocemos el valor de la otra.
- Este método se llama regresión.

Diapositivas	SUGERENCIAS METODOLÓGICAS Y TECNICAS
1-4	Es recomendable, antes de comenzar la presentación, que comente con el grupo las expectativas frente al tema y para que creen que les servirá. • Aclare que esta es una actividad con un enfoque marcadamente práctico. El interés es que Ustedes terminen conociendo y manejando una herramienta que les será útil para su práctica profesional.
5-17	Comente con el grupo los objetivos y aclare cada duda que surja. Comente con precisión y recoja la principales ideas de los alumnos. en algún lado para que estén presentes durante el trabajo.
18-20	Aquí, además de que tracen diagramas pida ejemplos y realice la actividad de equipos colaborativos.
21-36	Recuerden y afiance el metodo de pruebas de hipotesis.

Grafica de puntos dispersos



Material didáctico: visual elaborado por Fidelmar Sandoval Durán

- La gráfica de puntos dispersos muestra la relación entre las variables.
- Nuestro objetivo es trazar una línea, que mejor describa la relación entre X y Y.
- Se puede trazar una línea con una regla, que una los puntos, pero es improbable que obtengamos una misma línea y cada una de ellas, da diferente descripción de la relación entre X y Y.

- Cada distancia vertical es la diferencia entre el valor observado para la variable dependiente (en el eje y) y el valor de la línea trazada para el correspondiente valor del eje x.
- La distancia vertical entre los valores observados y los trazados es conocida como residual.

- La línea que mejor traza los datos se le conoce como línea de regresión.
- Da una estimación del valor promedio de y por algún valor de x. En general decimos que es una regresión de y sobre x.
- Se puede pensar en la línea de regresión como una línea que une los valores medios de y por cada valor de x.

- La expresión matemática para la línea de regresión es la ecuación:

$$y = \alpha + \beta x$$

donde α es la intersección de la línea con el eje y,

β es la pendiente de la línea.

- Regresión de los cuadrados mínimos da una línea de mejor trazo con una intersección y una pendiente determinada.

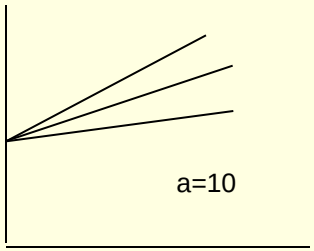
- Podemos trabajar sobre la pendiente de la línea tomando dos puntos a lo largo de la línea.

Por ejemplo, tomamos los puntos 1 y 2 de la gráfica de abajo.

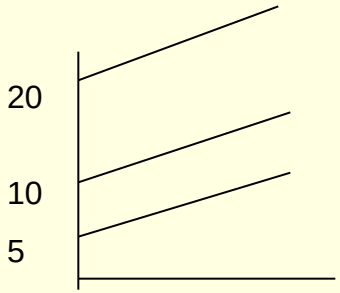
Punto 1 tiene los valores x=4, y= 16

Punto 2 tiene los valores x=8, y=22

- Esta gráfica corresponde a un valor fijo de a= 10 y un valor de b diferente.
- Muestra tres líneas que corresponden a un valor fijo de a y un valor diferente de y.

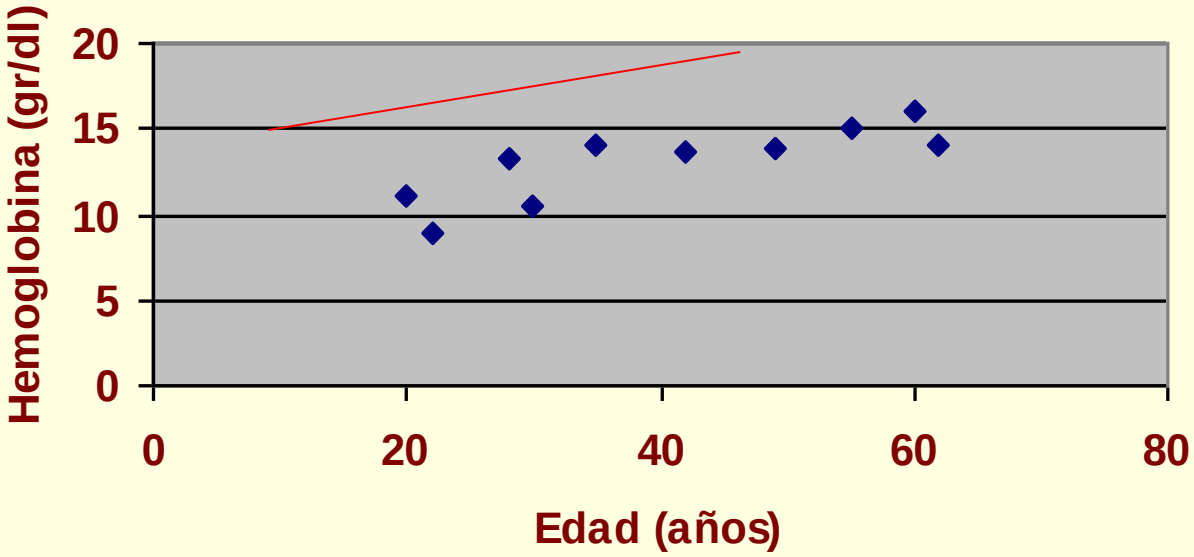


Esta gráfica corresponde a un valor fijo de b y un valor diferente de a.



- Una vez que se obtiene la línea de regresión, podemos usarla para dar un resumen de la relación entre la variable explicativa y respuesta (independiente, dependiente).
- Podemos decir:
  Por una unidad de incremento en x, y se incrementa por un cierto valor (el valor de b).
  $y = a + bx$

Relación entre edad y hemoglobina



$y = 7.9 + 0.136x$

Material didáctico: visual elaborado por Fidelmar Sandoval Durán

- Hasta ahora hemos visto sólo la descripción de la relación entre dos variables con una línea de regresión, donde *a* (la intersección) y *b* (la pendiente) son estimadas de los puntos de los datos de la muestra.
- La ecuación de regresión describiendo la relación entre dos variables en la población se escribe:
  $y = a + bx$
  Así, *a* es una estimación de α y *b* es una estimación de β.

	Población	Muestra
Intercepción	α	a
Pendiente	β	b

- La línea de regresión da una estimación de la relación entre las dos variables x y, y en la población.

- De la misma forma que hemos usado la inferencia para hacer conclusiones acerca de medias y proporciones, usaremos la línea de regresión para llegar a conclusiones acerca de la relación entre dos variables cuantitativas en la población.

- Si tomamos diferentes muestras de la población, con cada muestra podemos obtener una línea de regresión trazada por el método de los cuadrados mínimos.

- En la población hay una relación lineal entre dos variables y cada muestra puede ser ligeramente diferente.

- En la muestra $y = a + bx$.
- En la población $y = \alpha + \beta x$.
- Hay tres suposiciones subyacentes en el método de regresión lineal:

1. La variable respuesta, y, tiene una distribución Normal en cada x
2. La variabilidad de y deberá ser la misma a través de x
3. La relación entre x y deberá ser lineal.

- La pendiente b es de fundamental interés en el análisis de regresión.
- Nos da la más importante información acerca de la relación entre x y, esto es, el cambio promedio en y por una unidad de cambio en x.
- Obteniendo el error estándar de b, podemos calcular el intervalo de confianza y realizar una prueba de hipótesis sobre b.

- Podemos calcular la prueba de hipótesis acerca de la verdadera pendiente β, la pendiente de la relación lineal entre dos variables en la población.
  - Hipótesis nula
    - La hipótesis nula es que la pendiente en la población es cero.
    - Esto está implícito cuando decimos que no hay relación lineal entre altura y madurez ósea.
    - Ho: b = 0

- Hipótesis alternativa
  - La hipótesis alternativa es que la pendiente en la población no es cero. Si esto es verdad, podemos decir que hay una relación lineal entre estatura y madurez ósea.
  - H1: $b \neq 0$

- Para probar la hipótesis nula dividimos la estimación de *b* entre su error estándar y comparamos el resultado en la distribución t con n - 2 grados de libertad.

- Evaluación de un análisis de regresión involucra la comparación de la varianza de los residuales y la variación en los datos explicada por la línea de regresión.
- Esto se puede mostrar en una tabla de análisis de varianza.
- Este análisis se le llama ANOVA.

- Regresión
  - La gráfica muestra la relación entre x y, con cuatro puntos.
  - Se traza la línea de regresión y se analiza las diferentes partes de la variación en la relación entre x y, para evaluar la regresión

- La diferencia entre la suma total de cuadrados y la suma de los cuadrados de los residuales (la variación que permanece después de que es trazada una línea a través de los puntos) es la variación que es explicada por la regresión de y sobre x.

- ¿Qué es la suma de cuadrados de regresión?
  - La línea de regresión trazada explica la proporción de la variabilidad en la variable respuesta mientras que los residuales indican la cantidad de variabilidad sin explicación.
  - Una línea de regresión que describe bien los datos y explica la mayoría de la variación es preferible.

- La suma de cuadrados muestran cuanto de la variación es explicada por la línea de regresión y cuánto es explicada por los residuales.
- Esto se muestra en un análisis de varianza a través de la tabla ANOVA.

Tabla ANOVA

Fuente	Suma de cuadrados	Grados de libertad	Media de suma de cuadrados	F	Valor de p
Regresión	45	1	45	22.5	0.042
Residual	4	2	2		
Total	49	3			

El enfoque del análisis de varianza es comparar las dos fuentes de variación (regresión y residual) para saber cuál explica mejor la variación en la variable respuesta.

Para hacer esto, usamos una prueba que compara la variación en regresión y la variación residual, conocida como la prueba F.

- La razón de usar una prueba F es que la razón de dos varianzas tiene una distribución de muestreo conocida como distribución F.
  - La suma de cuadrados debido a la línea de regresión tiene un grado de libertad.
  - La suma de cuadrados debido a la variación residual (inexplicable) tiene n-2 grados de libertad.

- Para tomar en cuenta los grados de libertad, calculamos la media de la suma de cuadrados, dividiendo la suma de cuadrados entre los grados de libertad.

- Media de la suma de cuadrados = Suma de cuadrados/grados de libertad

- Podemos calcular el valor de F como la razón de la media suma de cuadrados.
- La prueba F, basada en ANOVA, es una forma alternativa de probar la hipótesis nula, $\beta = 0$.
- Es equivalente al cuadrado de la prueba de t sobre la pendiente b.

- La prueba F y la prueba t son para probar la hipótesis nula de que x no tiene relación con y.

- El valor de F es referido a las tablas de la distribución F con 1 y n-2 grados de libertad, para obtener el valor correspondiente de p.

Análisis de varianza (ANOVA)

- *¿Qué concluimos del valor de p?*
  - El valor de p nos dice la probabilidad de observar una relación lineal en la muestra si la hipótesis nula fuera verdad y no hubiera relación lineal en la población.
  - Así, para un valor de p bajo podemos rechazar la hipótesis nula y decir que hay una relación lineal en la población y la línea de regresión traza bien los datos.

- R^2
  - Hemos trabajado en casi todos los términos de una tabla ANOVA.
  - Sólo falta calcular el porcentaje de la variación total explicada por la línea de regresión.
  - Es una forma general de evaluar qué bien la línea de regresión traza los datos.

- ¿Cuánto de la variación total de la variable respuesta puede ser explicada por la línea de regresión?
  - Llamamos a este valor R^2 y lo calculamos como la razón de la suma de cuadrados de la regresión dividida entre la total suma de cuadrados.
- R^2 = Suma de cuadrados de regresión/Total suma de cuadrados x100

- Suposiciones para la regresión
  - Recuerde las suposiciones que están subyacentes al método de regresión lineal:
    - La variable respuesta deberá estar normalmente distribuida
    - La variabilidad de y deberá ser la misma a través de todos los valores de x
    - Deberá haber una relación lineal entre x y.

- Precauciones
  - Es posible obtener una línea de regresión de cualquier gráfica de puntos dispersos pero una regresión lineal deberá sólo ser aplicada donde existe una relación lineal.
  - Una asociación lineal entre dos variables no significa que una causa a la otra.
  - Puede ser necesario ajustar para confundidores potenciales.

Bibliografía

LIND, Douglas y MARCHAL, William y MASON, Robert. Estadística para administración y economía. Alfaomega. Colombia 11ava edición. 2004 Cap.13 y 14

CORDOVA, Jorge Herramientas Estadísticas para la Gestión en Salud. JC ediciones. Versión electrónica (formato CD) Mayo 2003.

HILDEBRAND, David y OTT, Lyman. Estadística Aplicada a la administración y a la economía. Adidison Wesley Iberoamericana sa. 1997. Cap. 13,14 y 15.