Regresión Lineal Múltiple

Introducción

- Cualitativo o Cuantitativo

  - El método cualitativo lo usaremos cuando se desea profundizar en un fenómeno, interpretarlo, conocer su significado en cuanto a su entorno, respecto a lo que dicen sus participantes.

  - En este método es interesante tener una visión más constructivista del fenómeno lograr interpretarlo, escuchando y utilizando información útil.

  - La información se desprende de las personas y de su interacción con el entorno y la experiencia que ellos tengan sobre el fenómeno determinado

Introducción

- Cualitativo o Cuantitativo
  - Además los métodos cualitativos son importantes si quiero profundizar conclusiones que se han visto en el campo cuantitativo
  - Un método mixto es un procedimiento donde podemos recolectar, analizar, mezclar información cualitativa en cuanto al método de estudio que nos ayudará en la pregunta de investigación

Introducción

- Cualitativo o Cuantitativo
  - Lo que queremos de alguna manera tener tanto información cualitativa como cuantitativa para un mejor entendimiento del problema
  - Queremos una imagen más clara que la da la literatura pues nos permiten dos elementos importantes, lo mejor de ambos métodos.

Introducción

- Para qué nos sirven los datos cuantitativos

?

Introducción

- Entre otras cosas, sirve para:
  - Resumir información de números “grandes" de personas.
  - Identificar patrones, tendencias en las poblaciones de interés
  - Establecer similitudes y diferencias entre distintos grupos
  - Identificar cuando dos variables co-varían

Introducción

- Entre otras cosas, sirve para:
  - Sólo en determinadas situaciones (casos específicos) identificar las causas de el o los fenómenos de análisis
  - Responder preguntas del tipo:
    - ¿Cuándo ocurre Y?
    - ¿Hay una asociación entre X e Y?
    - ¿Si pasa X, que se espera que pase en Y?

Introducción

- Es muy difícil entender el por qué de las cosas
  - El por qué está dado por la teoría
  - Los datos dan evidencia a favor o en contra de las asociaciones que nuestra teoría nos dice que probemos, pero son solo tan buenas como nuestra teoría
  - Los resultados estadísticos son sobre el promedio, siempre el promedio
  - Los datos NO nos dicen que ocurrió en cada caso específico

Introducción

- Un breve repaso
  - X = Variable independiente
  - Y = Variable dependiente
  - X -> Y Una hipótesis
  - $Y = \alpha + \beta_1 X +$ Una regresión lineal

Introducción

- Tipos de variables
  - Continuas (ej. edad, tiempo, calor, PIB)
  - Dicotómicas (SI-NO, ABIERTO-CERRADO, DIA-NOCHE)
  - Categóricas (ej. género, partidos políticos, compañías, drogas, tipos de estafas)

Introducción

- Por qué importa determinar el tipo de variables
  - Porque la selección de tipos de estadísticas que debemos utilizar depende directamente del tipo de variable que tenemos
    - Como variables dependientes: el tipo de variable define el tipo de método de estimación debemos usar (OLS, Logit, Poisson, mlogit, ologit, etc)
    - Como variables independientes: el tipo de variable define como debemos interpretar los coeficientes

Introducción

- Estadísticas descriptivas

Secuencia	Nombre	Edad	Horas Trabajo semanal
1	Hugo	35	51
2	Paco	33	44
3	Luis	30	44
4	Pepe	30	43
5	Ramón	38	46
6	Andrea	38	48
7	Loreto	29	43
8	Juan	31	38
9	Alina	34	48

Introducción

- Estadísticas descriptivas

Media $= \frac{35+33+30+30+38+38+29+31+34}{9} = 33,11111$ $\quad \overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{x_1 + x_2 + x_3 \ldots x_n}{n}$

Mediana $= 29, 30, 30, 31, 33, 34, 35, 38, 38 = 33$ $\quad (\frac{n+1}{2})^{ava}$ $\quad o$ $\quad \overline{x}[(\frac{n}{2})^{ava}, (\frac{n}{2}+1)^{ava}]$

Moda $= 30$ y $38 =$ es bi-modal $\quad$ Valor más común en un conjunto de datos

Varianza $= \frac{(35-33,1)^2+(33-33,1)^2+(30-33,1)^2+\ldots+(31-33,1)^2+(34-33,1)^2}{9} = 11,61111$ $\quad \sigma^2 = \frac{\sum(x-\overline{x})^2}{n}$ $\quad o$ $\quad \frac{\sum(x-\overline{x})^2}{(n-1)}$

Desviación estándar $= \sqrt{var} = 3,407508$ $\quad \sigma = \sqrt{\frac{\sum(x-\overline{x})^2}{n}}$ $\quad o$ $\quad \sqrt{\frac{\sum(x-\overline{x})^2}{(n-1)}}$

Introducción

- Cómo podemos saber si dos poblaciones son distintas
  - Covarianza
  - Correlaciones

Introducción

- La covarianza es una medida de la variabilidad conjunta de dos variables aleatorias.
  - Si los valores altos de una variable coinciden con los valores altos de una segunda variable, y lo mismo ocurre con los valores bajos (en otras palabras las variables presentan un comportamiento similar) la covarianza es positiva.
  - En el caso contrario, que los valores altos de una variable coincidan con los valores bajos de otra, la covarianza es negativa.

Introducción

$$\text{cov}(X, Y) = \frac{1}{n} \sum_{i=1}^{n} (x_i - E(X))(y_i - E(Y))$$

donde

$$E(X) = \overline{X}$$

Introducción

- Una correlación es una asociación estadística entre dos variables.
  - Sin embargo, el uso habitual en estadística hace referencia a que tan cerca están dos variables de tener una relación lineal.
  - Por las propiedades de su ecuación, la correlación produce valores entre -1 y 1:
    - 1 implica una correlación perfecta y positiva;
    - -1 perfecta y negativa;
    - 0 no existe correlación.

Introducción

$$\rho_{X,Y} = \operatorname{corr}(X, Y) = \frac{\operatorname{cov}(X, Y)}{\sigma_X \sigma_Y} = \frac{E[(X - \mu_X)(Y - \mu_Y)]}{\sigma_X \sigma_Y},$$

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{(n-1)s_x s_y} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

Introducción

- Varianza
  - Es una medida de dispersión que representa la variabilidad de una serie de datos respecto a su media
- Desviación Estándar
  - Medida de dispersión, que indica qué tan dispersos están los datos con respecto a la media.
  - El símbolo σ (sigma) se utiliza frecuentemente para representar la desviación estándar de una población, mientras que s se utiliza para representar la desviación estándar de una muestra.

Regresión Lineal Simple

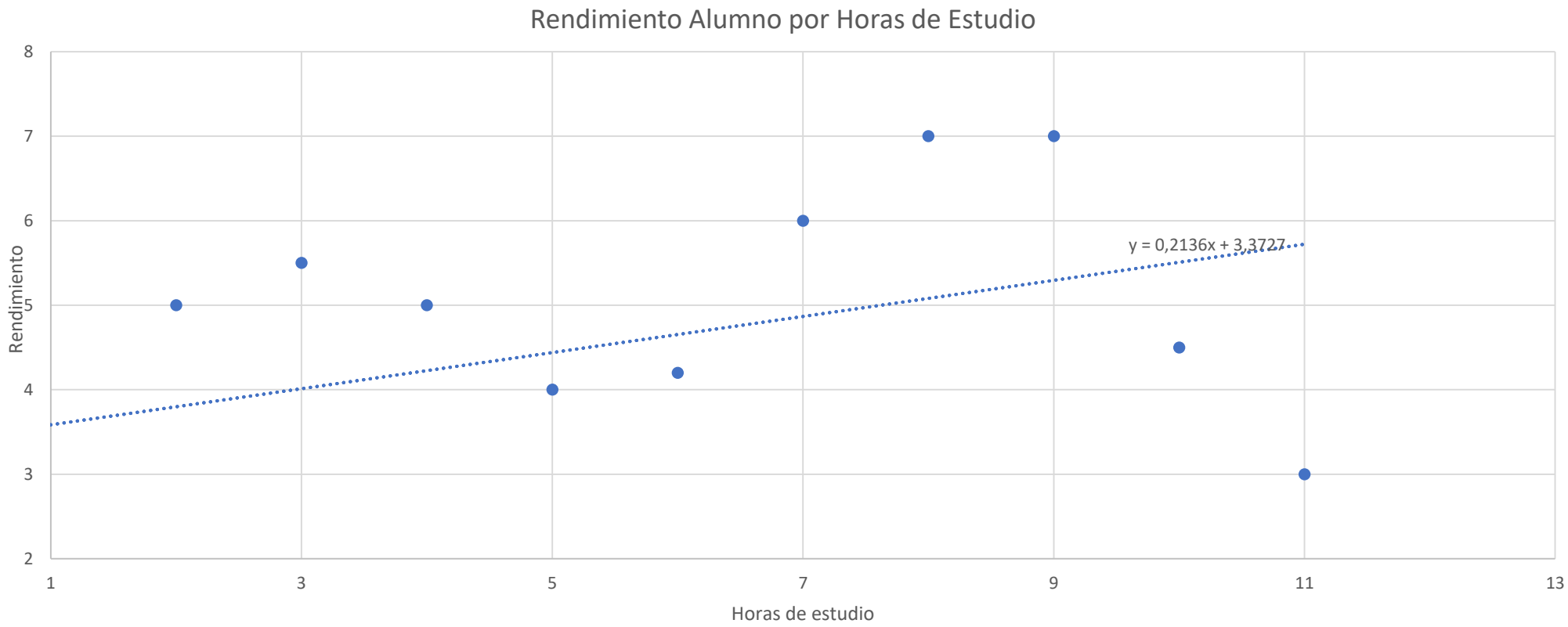
- La regresiones lineales simples del tipo OLS (not so - Ordinary Least Squares) estiman una ecuación del tipo:
  - $Y = \alpha + \beta X + \varepsilon$
  - Variable Dependiente = [parte sistemática] + [parte aleatoria]
- Con el objetivo de establecer un modelo estadístico que nos permita interpretar lo que ocurre en (y predecir) un fenómeno de interés.

Regresión Lineal Simple

.

Horas de Paro en la semana	Horas de Estudio	Nota
8	12	5
5	13	5,5
4	11	5
12	9	4
15	10	4,2
7	14	6
6	14	7
8	15	7
10	9	4,5
11	7	3

Regresión Lineal Simple



Regresión Lineal Simple

- Consideraciones en un Regresión Lineal Simple
  - Las observaciones Yi son estadísticamente independientes unas de otras.
  - Las observaciones Yi son una muestra aleatoria de una población donde Yi tiene una distribución normal con media μ_i y varianza σ^2.
    - Tomar en consideración que la varianza σ^2 se asume igual para todas las unidades i . En otras palabras que no depende de Xi , presunción conocida como homoskedasticidad.
    - La presunción de que la distribución de la población es normal, aunque regularmente incluida, no es estrictamente necesaria para algunos casos.

Regresión Lineal Simple

- Consideraciones en un Regresión Lineal Simple
  - La media de μ_i de Yi para cada unidad de i depende del valor de la variable independiente Xi , a través de una función lineal del tipo:

$$\mu_i = \alpha + \beta Xi, \text{ donde los parámetros } \alpha \text{ y } \beta \text{ son desconocidos}$$

Regresión Lineal Simple

- Consideraciones en un Regresión Lineal Simple
- En la estimación hay un elemento aleatorio no observable ε_i que se asume tiene las siguientes propiedades:
    - Todos los ε_i son estadísticamente independientes unos de otros.
    - La media (valor esperado) de ε_i es 0 para todos los i , sin depender de Xi . $E(\varepsilon_i) = 0$
- La varianza de ε_i es σ^2 para todos los i , independiente de Xi. $Var(\varepsilon_{i)} = \sigma^2$
- Los ε_i están distribuidos normalmente, aunque esta presunción no es estrictamente necesaria

Regresión Lineal Simple

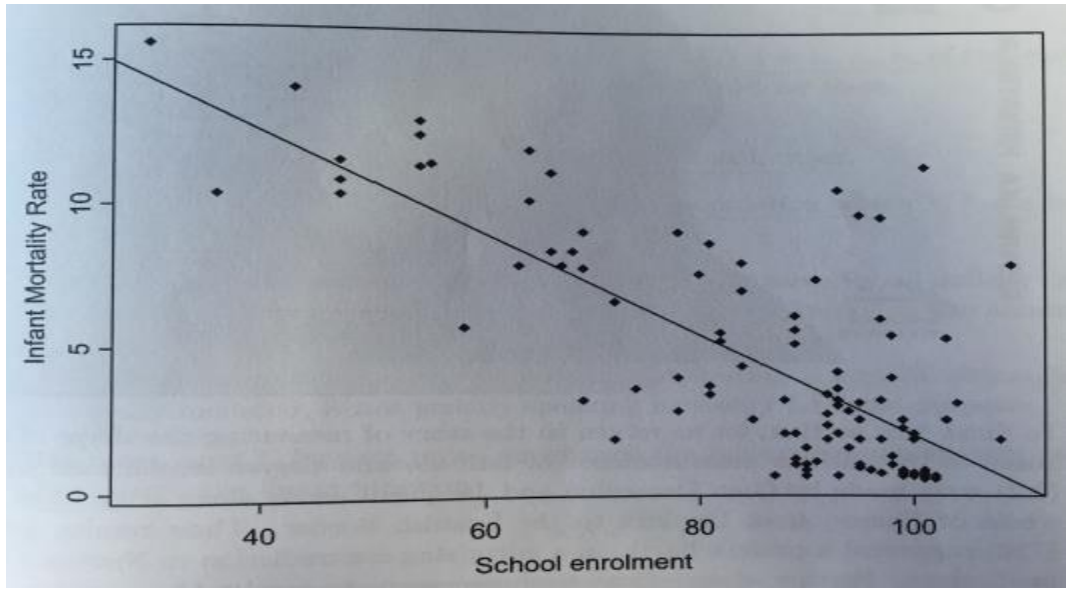
- Interpretación
- Una regresión lineal tiene tres parámetros α, β y σ^2 (este último denominado ε) regresión lineal. Los parámetros α y β son conocidos como coeficientes de la regresión y se interpretan de la siguiente forma:

- α, conocido como intercepto o constante de una regresión. Como el valor de α se da cuando X es igual a 0, y este valor, regularmente no es un valor interesante, α no se suele interpretar.

- β, en cambio, es interesante y corresponde al cambio en el valor esperado de Y cuando X aumenta o disminuye en una unidad. El signo refleja el valor de asociación

Regresión Lineal Simple

- El tercer parámetro en una regresión es σ^2 que corresponde a la varianza de la distribución de Y dado X. También se le conoce como la varianza de error o la varianza residual y su raíz cuadrada se le conoce como el error condicional o desviación estándar residual.
- De esta forma σ^2 y su correspondiente σ indican que tan concentrados están los valores de Y alrededor de esa medida (Y| x= 85)



Regresión Lineal Simple

- Una vez realizada la regresión, y estimados los parámetros de α y β se pueden calcular los valores estimados.

$$\hat{Y} = \hat{\alpha} + \hat{\beta}X$$

- Estos valores se comparan con los valores reales y Yi.

- La diferencia es conocida como los residuos de la muestra

- OLS minimiza esos residuos entregando la línea que mejor se aproxima a los datos

Es bueno recordar …

- "All models are wrong because the world, especially the social world, is an exceedingly complext place, full of local detail. As social scientists we do not want to reproduce that local detail in our models. What we are trying to do is capture the essentials and leave out the inessentials. A model that was one hundred percept correct would be of no value because it would be as complex as reality itself and if we could understand reality in all its complexity we would have no need for models!" (Kuha, 2011:26)

"Essentially, all models are wrong, but some are useful" (Box, 1987: 424)

Es bueno recordar …

Todos los modelos son incorrectos porque el mundo, especialmente el mundo social, es un lugar extremadamente complejo, lleno de detalles.

Como científicos sociales no queremos reproducir ese detalle local en nuestros modelos.

Lo que estamos intentando hacer es capturar lo esencial y dejar de lado lo no esencial.

Un modelo 100% la realidad no tendría ningún valor porque sería tan complejo como la realidad misma y si pudiéramos entender la realidad en toda su complejidad no tendríamos necesidad de modelos

Regresión Lineal Múltiple

$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \ldots \beta_n X_n + \epsilon$$

Regresión Lineal Múltiple

- Una extensión natural del modelo de regresión lineal simple consiste en considerar más de una variable explicativa.
- Los modelos de regresión múltiple estudian la relación entre
  - Una variable de interés Y, y
  - Un conjunto de variables explicativas o regresoras $X_1, X_2, \ldots, X_p$
- En el modelo de regresión lineal múltiple se supone que la función de regresión que relaciona la variable dependiente con las variables independientes es lineal, es decir:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \varepsilon$$

Regresión Lineal Múltiple

- Ejemplo:

  Uno de los problemas que tratan disciplinas como la Ecología o la Biología de la Conservación es identificar factores que influyen en variables como la riqueza de una especie (medida como el N° de individuos de la especie en un superficie dada)

Regresión Lineal Múltiple

- Ejemplo:

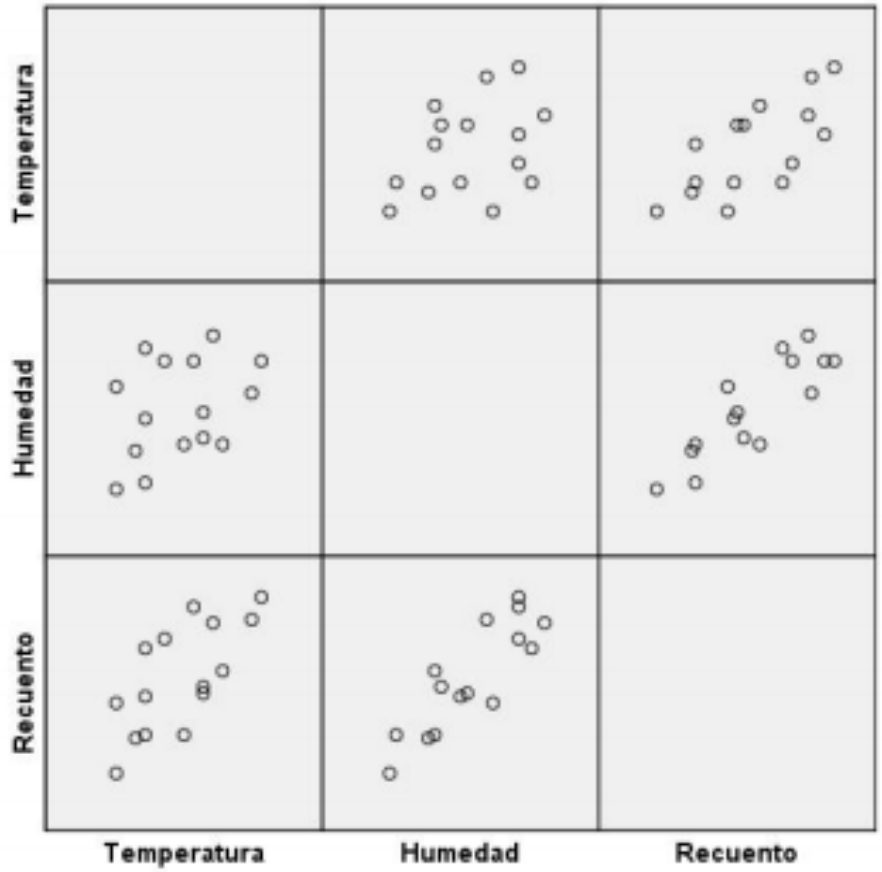
  En un estudio sobre la población de un parásito se hizo un recuento de parásitos en 15 localizaciones con diversas condiciones ambientales. Los datos obtenidos son los siguientes:

Temperatura	15	16	24	13	21	16	22	18	20	16	28	27	13	22	23
Humedad	70	65	71	64	84	86	72	84	71	75	84	79	80	76	88
Recuento	156	157	177	145	197	184	172	187	157	169	200	193	167	170	192

Regresión Lineal Múltiple

- Ejemplo:

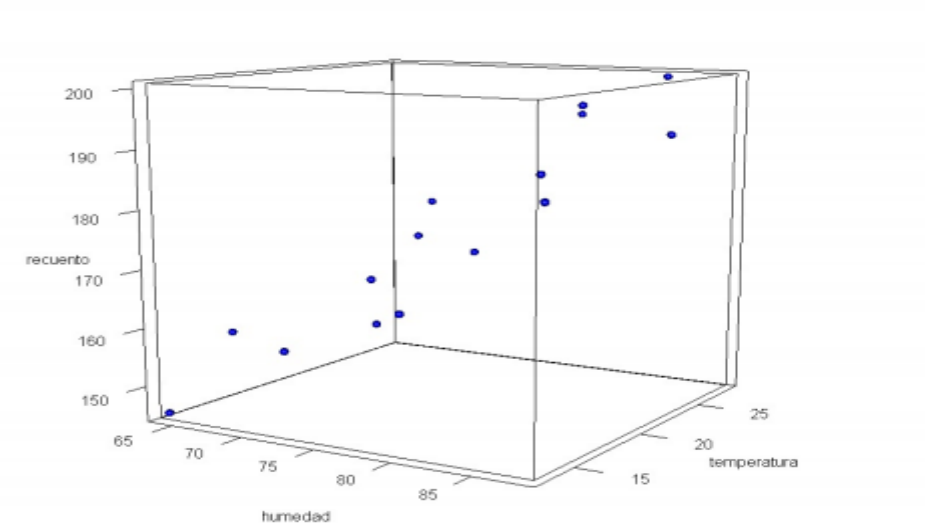
	Temperatura	Humedad	Recuento	var
1	15,00	70,00	156,00	
2	16,00	65,00	157,00	
3	24,00	71,00	177,00	
4	13,00	64,00	145,00	
5	21,00	84,00	197,00	
6	16,00	86,00	184,00	
7	22,00	72,00	172,00	
8	18,00	84,00	187,00	
9	20,00	71,00	157,00	
10	16,00	75,00	169,00	
11	28,00	84,00	200,00	
12	27,00	79,00	193,00	
13	13,00	80,00	167,00	
14	22,00	76,00	170,00	
15	23,00	88,00	192,00	
16				
17				



Regresión Lineal Múltiple

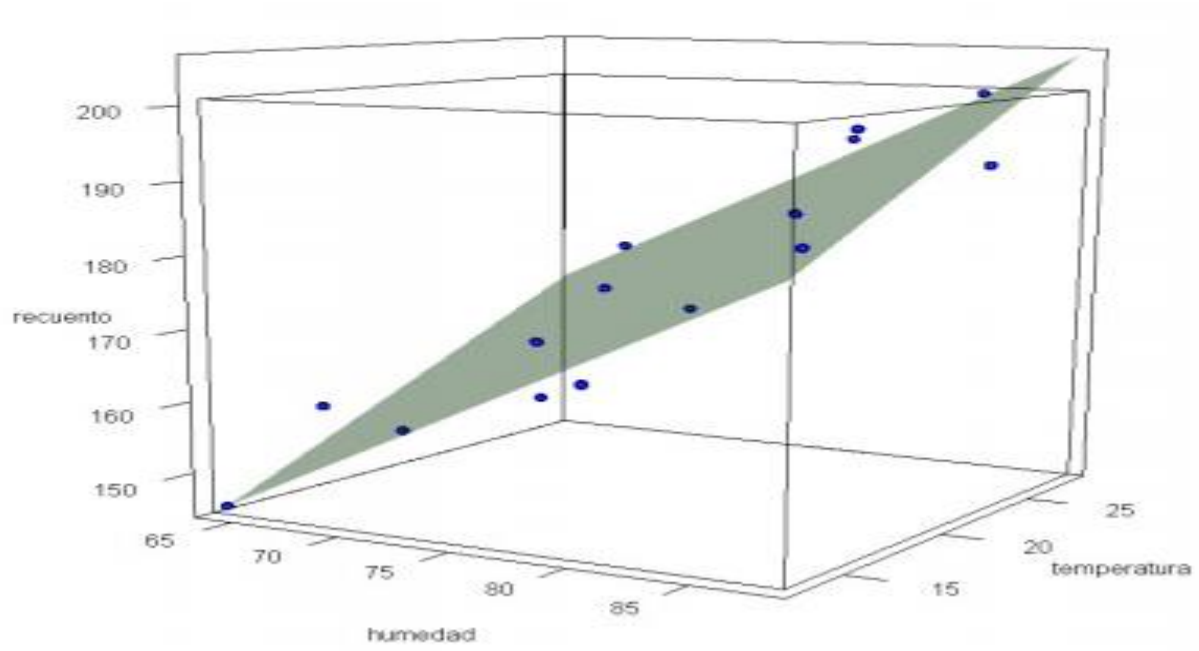
- Ejemplo: Parece que la humedad y la temperatura son dos factores que afectan a la riqueza de la especie. ¿por qué no usamos los datos disponibles e intentamos explicar el comportamiento de la riqueza de parásitos a partir de ambas variables?

$$\text{Recuento} = \beta_0 + \beta_1 \text{Temperatura} + \beta_2 \text{Humedad} + \epsilon$$



Regresión Lineal Múltiple

$$\text{Recuento} = 25.7115 + 1.5818\text{Temperatura} + 1.5424\text{Humedad}$$



Regresión Lineal Múltiple

- Supuestos para el modelo

1. Los coeficientes parciales de la regresión lineal múltiple son lineales.
2. Los residuos se distribuyen normalmente con media cero y varianza constante.
3. No debe haber multicolinealidad entre la variables explicativas. Las variables explicativas no deben estar correlacionadas.
4. No debe haber heteroscedasticidad.
5. Cuando se trabaja con datos temporales, no debe haber autocorrelación entre los residuos.

Regresión Lineal Múltiple

- Supuestos para el modelo

1. El primer supuesto se puede probar con el análisis de Varianza o con las pruebas para cada coeficiente.
2. El segundo supuesto se puede probar con el test de Kolmogorov-Smirnov-Lillifort si n >30, en caso contrario se usa la prueba de Shapiro-Wilk.
3. Para probar la multicolinealidad se usa los coeficientes de correlación de Pearson (significancia).
4. Para la heteroscedasticidad se usa la Prueba de White.
5. Para la autocorrelación se usa la prueba de Durbin-Watson.

Regresión Lineal Múltiple

- Supuestos para el modelo

- Con el objeto de averiguar si hay correlación entre las variables explicativas, se puede usar el coeficiente de correlación de Pearson:

$$r = \frac{Cov(X_i, X_j)}{\sigma_{x_i} \sigma_{x_j}} \qquad -1 \leq r \leq 1$$

Regresión Lineal Múltiple

- Supuestos para el modelo

- Para averiguar el grado de explicación (asociación) que tiene cada variable sobre la explicativa en términos neto, se puede utilizar el coeficiente de correlación parcial.

$$r_{12.3} = \frac{r_{12} - r_{13} r_{23}}{\sqrt{(1 - r_{13}^2)(1 - r_{23}^2)}} \qquad r_{13.2} = \frac{r_{13} - r_{12} r_{23}}{\sqrt{(1 - r_{12}^2)(1 - r_{23}^2)}}$$

Regresión Lineal Múltiple

- Supuestos para el modelo

- Para averiguar el grado de explicación (asociación) que tiene cada variable sobre la explicativa en términos neto, se puede utilizar el coeficiente de correlación parcial.

$$r_{12.3} = \frac{r_{12} - r_{13} r_{23}}{\sqrt{(1 - r_{13}^2)(1 - r_{23}^2)}} \qquad r_{13.2} = \frac{r_{13} - r_{12} r_{23}}{\sqrt{(1 - r_{12}^2)(1 - r_{23}^2)}}$$

Regresión Lineal Múltiple

- El coeficiente de determinación R^2 nos indica el grado de ajuste de la recta de regresión a los valores de la muestra.

- R^2 toma valores entre 0 y 1

- Si el ajuste es bueno R^2 será cercano a 1

- Si el ajuste es malo será R^2 cercano a 0