Universidad Central de Venezuela
Facultad de Ciencias
Escuela de Computación
Probabilidad y Estadística

Taller Nº 1
Estadística Descriptiva y Regresión Lineal con R

Resumen
Este taller esta dirigido a estudiantes de la materia de Introducción a la Probabilidad y Estadística de la Escuela de Computación de la UCV. Se resolverán problemas de descripción de una muestra y la inferencia estadística a través de modelos lineales. El software estadístico utilizado es el R, y trabajaremos sobre una muestra estadística específicamente creada para este taller.

Objetivos del Taller:

(a) Conocer y manejar los elementos fundamentales del entorno del software estadístico
(b) Experimentar los conceptos de estadística descriptiva sobre una muestra grande.
(c) Aplicar el modelo de Regresión Lineal para hacer inferencias estadísticas sobre mediciones reales.

Objetivos Específicos:

(a) Analizar una muestra de niños y niñas entre 0 y 3 años de edad con medidas de Talla. Peso y Circunferencia Craneal a través de Estadística descriptiva
(b) Generar modelos para contruir los percentiles de cada una de las medidas a partir de modelos lineales
(c) Construir un algoritmo que a partir de unas medidas de Peso, talla o Circunferencia Craneal de una determinada edad de una niña o niño, el mismo retorne entre que percentiles se encuentra.

1. Introducción.

Uno de los instrumentos más utilizados por los pediatras para el monitoreo del crecimiento de los niños en sus primeros años de vida son los gráficos de percentiles de peso talla y diámetro craneal emitidos por la Organización Mundial de la Salud (OMS) como el de ejemplo anexado a este taller. Estos gráficos permiten a los doctores identificar entre que percentiles se encuentra un niño en particular dadas su medidas de talla, peso y circunferencia craneal, así como también, pueden saber si su peso está acorde con su talla.
Estas gráficas se generan a partir del análisis de datos estadísticos muestrales recolectados en varios países del mundo.

2. Paquete estadístico R.

R es un paquete estadístico de software libre, distribuido bajo licencia GPL. Más específicamente se trata de un lenguaje y un entorno de programación para el análisis estadístico. Este programa esta instalado en el Live-CD que hemos preparado para la realización de este taller en los laboratorio docentes da la Facultad. Las versiones de otros sistemas operativos pueden descargase desde la pagina oficial www.r-project.org

3. Importar los datos muestrales.

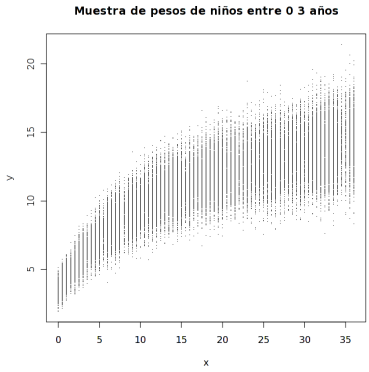
Es recomendable definir el directorio de trabajo, que es donde se deben encontrar los archivos de entrada y donde se guardaran los archivos de salida. Seguidamente cargaremos en un marco de datos la muestra llamada "Muestra.txt". Esta muestra esta compuesta por un listado de mediciones del peso, largo y circunferencia craneal de niños de ambos sexos entre 0 y tres años de edad (medidos en meses)

- *setwd("/root/Desktop/Taller_R")*
- *muestra <- read.csv("Muestra.txt")*
- *attach(muestra)*

4. Representación gráfica de los datos muestrales.

Graficar la nube de datos que representa la relación del peso niños varones con respecto a su edad. Para esto crearemos sol listados, uno para contener las edades y el otro para los pesos.

- *y <- muestra$Pesos[Sexo==1]*
- *x <- muestra$Edades[Sexo==1]*
- *plot(x,y, pch=".", main="Muestra de pesos de niños entre 0 3 años")*

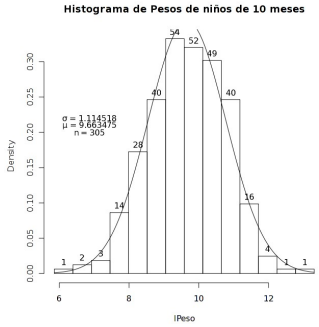


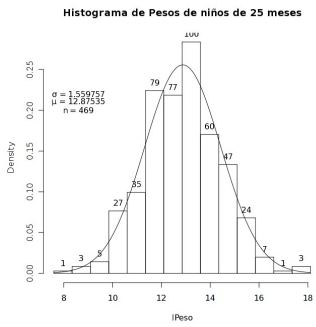
5. Análisis descriptivo de la muestra.

Grafíque el histograma de los pesos de los niños varones de 10 meses de edad. Diga el tamaño de la muestra para esta población y determine su media, desviación estándar y rango muestrales.

- *lPeso <- muestra$Pesos[Sexo==1 & Edades==10]*
- *l <- length(lPeso)*
- *m <- mean(lPeso)*
- *s <- sqrt(var(lPeso))*
- *clases <- seq(min(lPeso), max(lPeso), length.out=15)*
- *h <- hist(lPeso, prob=T,labels=F, breaks=clases, main="Histograma de Pesos de niños de 10 meses")*
- *normal <- function(x) dnorm(x, m, s)*
- *curve(normal, add=T)*
- *text(h$mids,h$density+0.01,h$counts)*
- *text(min(lPeso)+1,0.22,as.expression(substitute(sigma==des,list(des=s))))*
- *text(min(lPeso)+1,0.21,as.expression(substitute(mu==prom,list(prom=m))))*
- *text(min(lPeso)+1,0.20,as.expression(substitute(n==tam,list(tam=l))))*

Repita luego esta operación para los niños varones de 25 meses, compare y comente.





6. Cálculo de los percentiles muestrales.

Generé un archivo de salida con el listado de medias, desviaciones estándar y percentiles de 5%, 10%, 25%, 50%, 75%, 90% y 95% para cada una de las edades y sexo.

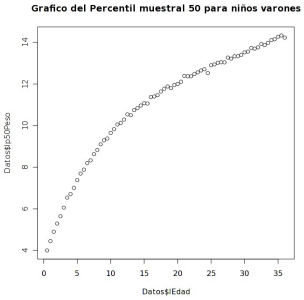
- *Nombres <- c("Sexo","Edad","mPeso","sPeso","p05Peso","p10Peso","p25Peso","p50Peso","p75Peso","p90Peso","p95Peso")*
- *write.table(t(Nombres), file = "Salida.txt", append = TRUE, row.names = FALSE, col.names = FALSE, sep = ",")*
- *for (sexo in 1:2) {*
- *lEdades <- unique(muestra$Edades[Sexo==sexo])*
- *for (edad in lEdades) {*
- *lPesos <- muestra$Pesos[Sexo==sexo & Edades==edad]*
- *mPeso <- mean(lPesos)*
- *sPeso <- sqrt(var(lPesos))*
- *p05Peso <- quantile(lPesos, 0.05)*

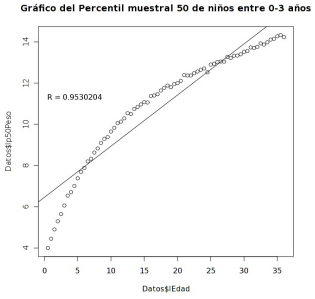
- *p10Peso <- quantile(lPesos, 0.1)*
- *p25Peso <- quantile(lPesos, 0.25)*
- *p50Peso <- quantile(lPesos, 0.50)*
- *p75Peso <- quantile(lPesos, 0.75)*
- *p90Peso <- quantile(lPesos, 0.90)*
- *p95Peso <- quantile(lPesos, 0.95)*
- *Salidas <- c(sexo, edad, mPeso, sPeso,p05Peso,p10Peso, p25Peso, p50Peso, p75Peso, p90Peso, p95Peso)*
- *write.table(t(Salidas), file = "Salida.txt", append = TRUE, row.names = FALSE, col.names = FALSE, sep = ",")*
- *}*
- *}*
- *detach(muestra)*

7. Modelo de Regresión Lineal.

Cargue el archivo generado para crear modelo lineal que relacione en peso de los niños varones en función de las edades. Haremos esto con el percentil muestral 50.

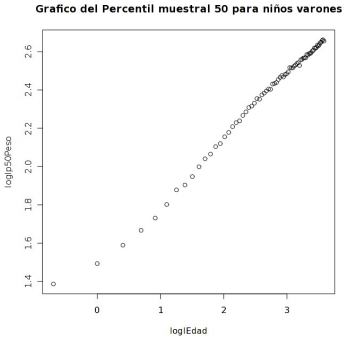
- *muestra <- read.csv("Salida.txt")*
- *attach(muestra)*
- *lp50Peso <- muestra$p50Peso[Sexo==1]*
- *lEdad <- muestra$Edad[Sexo==1]*
- *Datos <- data.frame(lEdad, lp50Peso)*
- *plot(Datos$lEdad, Datos$lp50Peso, main="Gráfico del Percentil muestral 50 de niños entre 0-3 años")*
- *mlAjuste <- lm(lp50Peso~lEdad, data=Datos)*
- *abline(coef(mlAjuste))*
- *r <- cor(Datos$lEdad, Datos$lp50Peso)*
- *text(min(Datos$lEdad)+4,max(Datos$lp50Peso)-3,as.expression(substitute(R==correl,list(correl=r))))*

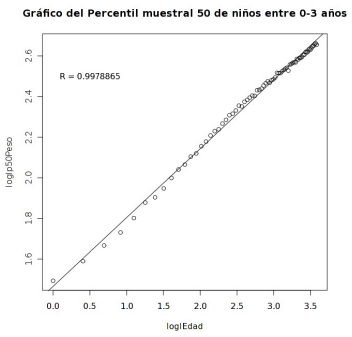




Comente sus observaciones. Repita la operación sobre una espacio transformado mediante el cálculo de los logaritmos, es decir manejando las coordenadas (Log(x),Log(y))

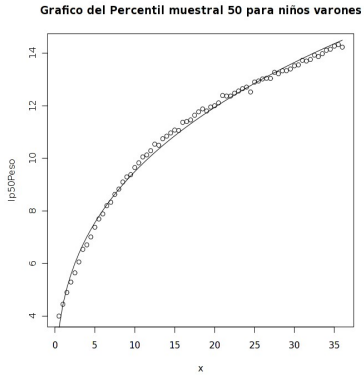
- *e <- exp(1)*
- *loglEdad <- log(lEdad[-1],e) #Suprimiendo el primer elemento para evitar el log(0)*
- *loglp50Peso <- log(lp50Peso[-1],e) #Suprimiendo el primer elemento para evitar el log(0)*
- *logDatos <- data.frame(loglEdad, loglp50Peso)*
- *mlAjuste <- lm(loglp50Peso~loglEdad, data=logDatos)*
- *plot(loglEdad, loglp50Peso, main="Gráfico del Percentil muestral 50 de niños entre 0-3 años")*
- *abline(coef(mlAjuste))*
- *r <- cor(loglEdad, loglp50Peso)*
- *text(0.5,2.5,as.expression(substitute(R==correl,list(correl=r))))*





Comente sus observaciones. Ahora grafiquemos estos resultados en el espacio no transformado (x,y) y calculemos el error promedio obtenido con este modelo.

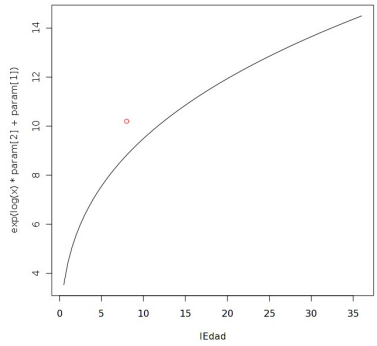
- *param <- coef(mlAjuste)*
- *x <- lEdad*
- *y <- exp(log(x)*param[2]+param[1])*
- *plot(lEdad, lp50Peso, main="Grafico del Percentil muestral 50 para niños varones")*
- *lines(lEdad,y)*



8. Aplicando el modelo.

Consideremos un caso particular de un niño con peso y Edad dados (8 meses, 10.2 Kg). Escriba un función que reciba estos parámetros y determine si dicho niño se encuentra por encima del percentil 50 o no. Grafíque el resultado.

- *SobrePercentil <- function(edad, peso) {*
- *peso >= exp(log(edad)*param[2]+param[1])*
- *}*
-
- *Medidas <- c(8, 10.2)*
- *plot(lEdad,exp(log(x)*param[2]+param[1]),type="l")*
- *points(Medidas[1], Medidas[2], col="red")*
- *SobrePercentil(8, 10.2)*



9. Tarea para la casa.

a) En función del digito con el que termina su número de cédula realize uno de los siguientes ejercicios para la casa. Este será evaluado en un breve interrogatorio posteriormente:

0. Genere un archivo de salida con los coeficientes asociados a los modelos lineales de los percentiles 10%,25%,50%,75% y 90% de las mediciones de Talla en función de la edad para niños varones entre 0 y 3 años de edad. Luego codifique un algoritmo (función) que reciba como parámetros un valor de Sexo, Edad y Talla de un niño determinado y retorne como resultado entre que percentiles se encuentra, graficando el resultado (el punto y los dos percentiles entre los que se encuentra).
1. Genere un archivo de salida con los coeficientes asociados a los modelos lineales de los percentiles 5%,25%,50%,75% y 95% de las mediciones de Talla en función de la edad para niñas hembras entre 0 y 3 años de edad. Luego codifique un algoritmo (función) que reciba como parámetros un valor de Sexo, Edad y Talla de una niña determinada y retorne como resultado entre que percentiles se encuentra, graficando el resultado (el punto y los dos percentiles entre los que se encuentra).
2. Genere un archivo de salida con los coeficientes asociados a los modelos lineales de los percentiles

entre 5% y 95% a intervalos de 5%, de las mediciones de Peso en función de la edad para niños varones entre 0 y 3 años de edad. Luego codifique un algoritmo (función) que reciba como parámetros un valor de Sexo, Edad y Peso de un niño determinado y retorne como resultado entre que percentiles se encuentra, graficando el resultado (el punto y los dos percentiles entre los que se encuentra).

3. Genere un archivo de salida con los coeficientes asociados a los modelos lineales de los percentiles entre 5% y 95% a intervalos de 5%, de las mediciones de Peso en función de la edad para niñas hembras entre 0 y 3 años de edad. Luego codifique un algoritmo (función) que reciba como parámetros un valor de Sexo, Edad y Peso de una niña determinada y retorne como resultado entre que percentiles se encuentra, graficando el resultado (el punto y los dos percentiles entre los que se encuentra).
4. Genere un archivo de salida con los coeficientes asociados a los modelos lineales de los percentiles 10%,25%,50%,75% y 90% de las mediciones de Circunferencia Craneal en función de la edad para niños varones entre 0 y 3 años de edad. Luego codifique un algoritmo (función) que reciba como parámetros un valor de Sexo, Edad y Circunferencia Craneal de un niño determinado y retorne como resultado entre que percentiles se encuentra, graficando el resultado (el punto y los dos percentiles entre los que se encuentra).
5. Genere un archivo de salida con los coeficientes asociados a los modelos lineales de los percentiles 5%,25%,50%,75% y 95% de las mediciones de Circunferencia Craneal en función de la edad para niñas hembras entre 0 y 3 años de edad. Luego codifique un algoritmo (función) que reciba como parámetros un valor de Sexo, Edad y Circunferencia Craneal de una niña determinada y retorne como resultado entre que percentiles se encuentra, graficando el resultado (el punto y los dos percentiles entre los que se encuentra).
6. Genere un archivo de salida con los coeficientes asociados a los modelos lineales de los percentiles entre 10% y 90% a intervalos de 10%, de las mediciones de Talla en función de la edad para niños varones y hembras entre 0 y 3 años de edad. Luego genere las gráficas de percentiles para niños y niñas de la Talla en función de la edad, como las generadas por las organizaciones de salud.
7. Genere un archivo de salida con los coeficientes asociados a los modelos lineales de los percentiles entre 10% y 90% a intervalos de 10%, de las mediciones de Talla en función de la edad para niños varones y hembras entre 0 y 3 años de edad. Luego genere las gráficas de percentiles para niños y niñas de la Talla en función de la edad, como las generadas por las organizaciones de salud.
8. Genere un archivo de salida con los coeficientes asociados a los modelos lineales de los percentiles entre 10% y 90% a intervalos de 10%, de las mediciones de Circunferencia Craneal en función de la edad para niños varones y hembras entre 0 y 3 años de edad. Luego genere las gráficas de percentiles para niños y niñas de la Circunferencia Craneal en función de la edad, como las generadas por las organizaciones de salud.
9. Genere un archivo de salida con los coeficientes asociados a los modelos lineales de los percentiles entre 5% y 95% a intervalos de 10%, de las mediciones de Peso y Talla en función de la edad para niñas hembras entre 0 y 3 años de edad. Luego genere las gráficas de percentiles para niñas del Peso y Talla en función de la edad, como las generadas por las organizaciones de salud.

b) Comente ¿cual sería su propuesta sí se le solicitará opinión sobre la realización de un servicio informático que haga este trabajo sin la necesidad de contar con una muestra en línea como la utilizada en este taller?. ¿Cual es la importancia del uso de modelos matemáticos en este tipo de problemas?