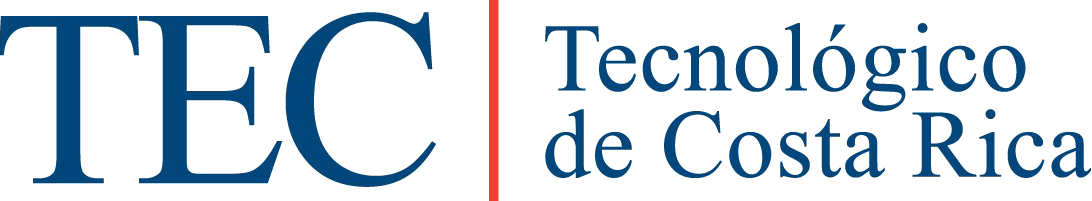
Especificaciones del proyecto Fase III

Esteban Ballestero Alfaro

1/6/2021



# I PARTE: OTRAS PRUEBAS DE HIPÓTESIS EN R

En esta sección se presentarán algunos casos relacionados con pruebas de hipótesis de bondad de ajuste, ANOVA´s e independencia. Deben posterior a la lectura del caso, identificar el tipo de prueba que mejor se adapte a la situación.

## CASO 1:

Utilice la base de datos KidsFeet paquete mosaicData como evidencia para para realizar un estudio de normalidad de la longitud del pie (length) de los niños (**B:** boys) es igual que el de las niñas (**G:** girls). Para emitir su conclusión, siga el siguiente proceso:

* Para este ejercicio, explique como la siguiente instrucción permite agrupar los datos para niños y niñas:

feetsplit <- split(feet$length,feet$sex)

1. Para cada datset (niños y niñas), haga un estudio gráfico de la normalidad. Para esto elabore el gráfico que contenga tanto la función de densidad para cada caso como su respectiva curva normal teórica a partir de su media y desviación estándar. ¿Se puede intuir una posible normalidad para los datos?. explique brevemente.
2. Para cada dataset (niños y niñas) elabore un gráfico QQ-plot y haga la interpretación en cada caso, desde el punto de vista si es posible asumir normalidad.
3. Realice las pruebas formales de normalidad S-W test, A-D test,K-S-L test.
4. ¿Es posible aplicar D’Agostino-Pearson?. En caso de ser posible, realice la prueba.
5. Haga un análisis de la custoris y simetría de los datos usando las funciones respectivas del paquete moments. No es necesario que realice el gráfico tal como se mostró en clases, solo observe los resultados para una posterior interpretación

## CASO 2:

Realice todo lo solicitado el **Caso 1**, pero utilizando los datos de biceps de la tabla medidas\_cuerpo. Para obtener la base de datos, realice los siguiente:

1. Descargue la base de datos desde la url como sigue:

url <- 'https://raw.githubusercontent.com/fhernanb/datos/master/medidas\_cuerpo'  
variable <- read.table(file=url, header=T)

1. Recuerde que puede darle el nombre que desee a la variable.
2. Analice cada prueba de normalidad como se solicitó en el **Caso 1**, utilizando todos los datos de la varible biceps, **NO** haga separación alguna por sexo u otra categoría en este análisis.

Una vez realizado todos los cálculos de las pruebas de normalidad para los **Casos 1 y 2**, complete la siguiente tabla con la información que corresponda en cada una de las celdas:

|  |  |  |  |
| --- | --- | --- | --- |
| **Tipo de prueba** | **Longitud del pie** | | **Medidas de biceps** |
| **Niños** | **Niñas** |
| **Función de densidad versus curva normal** | **[Anote lo observado]** | **[Anote lo observado]** | **[Anote lo observado]** |
| **Conclusión** |  |  |  |
| **QQ-Plot** | **Se observa que los datos se alinean a la recta, aunque hay ciertos datos que no la hacen, pero concluimos que sí se justifica la normalidad de los datos.** | **Se observa que los datos están muy inconsistentes, debido a que hay algunos alineados a la recta y otros que están muy alejados. Concluimos que no hay normalidad con este gráfico.** | **Se observa que los datos están muy inconsistentes, debido a que hay algunos alineados a la recta y otros que están muy alejados. Concluimos que no hay normalidad con este gráfico.** |
| **Conclusión** |  |  |  |
| **S-W test** | **[Valor P]** | **[Valor P]** | **[Valor P]** |
| **Conclusión** |  |  |  |
| **A-D test** | **[Valor P]** | **[Valor P]** | **[Valor P]** |
| **Conclusión** |  |  |  |
| **K-S-L test** | **[Valor P]** | **[Valor P]** | **[Valor P]** |
| **Conclusión** |  |  |  |
| **D’ Agostino - Pearson** | **[Valor P]** | **[Valor P]** | **[Valor P]** |
| **Conclusión** |  |  |  |
| **Curtosis y simetría** | **[Valores]** | **[Valores]** | **[Valores]** |
| **Conclusión** |  |  |  |
| **Conclusión general sobre normalidad de los datos** |  |  |  |

## 

## CASO 3:

La siguiente tabla resume los datos de obtenidos de víctimas de crímenes elegidas al azar (según datos del Departamento de Justicia de USA):

|  |  |  |  |
| --- | --- | --- | --- |
|  | HOMICIDIO | ROBO | ASALTO |
| El criminal era un extraño | 12 | 379 | 727 |
| El criminal era un conocido o pariente | 39 | 106 | 642 |

Con los datos anteriores, ¿sería posible considerarque el tipo de delito es independiente de la condición del delincuente?, o por el contrario, ¿existe alguna relación entre el tipo de delito con respecto al quien comete el acto? Use un nivel de significancia de 5%.

Además, complete a partir de los resultados de la prueba, lo que se solicita en la siguiente tabla:

|  |  |
| --- | --- |
| **Resumen de la prueba** | |
| Tipo de prueba a utilizar (independencia, bondad de ajuste, ANOVA) | Independencia |
| Valor observado | X-squared = 1339.3 |
| Grados de libertad y qué representan | 2 |
| Valor P | 2.2e-16 |
|  | El delito es independiente de la condición del delincuente |
|  | El delito no es independiente de la condición del delincuente |
| **Conclusión:**  Para este caso, tomando en cuenta todos los resultados obtenidos, se concluye que se rechaza H0 , debido a que no se encuentra suficiente evidencia para aceptarla. Para fundamentar la respuesta se destaca que el valor de P es menor que el valor de nuestro “Alpha” que es 0.05, por ende, se rechaza H0. Lo que supone la conclusión de esta prueba es que existe alguna relación entre el tipo de delito con respecto al quien comete el acto. | |

## CASO 4:

Analice con detenimiento la siguiente situación:

La seguridad de los automóviles se determina mediante diversas pruebas. Una de ellas consiste en hacer chocar un automóvil contra una barrera fija a 35 con un maniquí colocado en el asiento del conductor.

A una de las medidas utilizadas para cuantificar el impacto del choque sobre el conductor se le conoce como **Desaceleración de pecho** y se mide en unidades de fuerza de gravedad (). Los valores más grandes indican mayores cantidades de desaceleración, las cuáles pueden provocar lesiones graves en los conductores. La siguiente tabla muestra mediciones de desaceleraciones de pecho obtenidas a partir de pruebas de choques de diferentes tipos de vehículos:

|  |  |  |
| --- | --- | --- |
| Autos compactos | Autos medianos | Autos grandes |
| 44 | 41 | 32 |
| 43 | 49 | 37 |
| 44 | 43 | 38 |
| 54 | 41 | 45 |
| 38 | 47 | 37 |
| 43 | 44 | 33 |
| 42 | 37 | 38 |
| 45 | 34 | 45 |
| 50 |  | 43 |
|  |  | 42 |

Con los datos anteriores, ¿es posible considerar que el tamaño del automóvil puede variar en cuanto a la seguridad de sus pasajeros o por el contrario, es igualmente riesgoso? Use un nivel de significancia del 5%.

1. Con los datos que imprime la prueba, complete la información requerida en la siguiente tabla:

|  |  |
| --- | --- |
| **Resumen de la prueba** | |
| Tipo de prueba a utilizar (independencia, bondad de ajuste, ANOVA) | ANOVA |
| Valor observado | 3.552 |
| Grados de libertad y qué representan | 2 , que representan el auto.tipo y 24 que presentan los datos o los residuales |
| SSE | 535.6 |
| SSA | 158.5 |
| SST | 535.6+158.5= 694.1 |
| y (use la columna de Mean Sq) | = 79.26 y = 22.31 |
| Valor P | 0.0445 |
|  | La seguridad no varía considerando el tamaño del automóvil : u1 = u2 = u3 |
|  | La seguridad varía considerando el tamaño del automóvil : al menos dos de las medias no son iguales |
| **Conclusión:**  Para este caso, tomando en cuenta todos los resultados obtenidos, se concluye que se rechaza H0 , debido a que no se encuentra suficiente evidencia para aceptarla. Para fundamentar la respuesta se destaca que el valor de P es menor que el valor de nuestro “Alpha” que es 0.05, por ende, se rechaza H0. Aunque cabe destacar que la significancia del rechazo no es muy grande, ya que el valor de P obtenido se acerca mucho al valor establecido para el “Alpha”. Lo que supone la conclusión de esta prueba es que es igualmente riesgoso y que el tamaño del automóvil no varía en cuanto a la seguridad de sus pasajeros y también, al menos dos de las medias no son iguales. | |

1. En caso de detectar alguna diferencia en cuanto a la seguridad que puede brindar al tamaño del vehículo, ¿es posible en cuáles existe una verdadera diferencia significativa? De ser así, realice la prueba que corresponda e interprete el resultado.

R/

# II PARTE: MODELOS DE REGRESIÓN LINEAL

En esta sección se analizarán casos relacionados con los modelos de Regresión Lineal Simple (RLS) y múltiple (RLM), así como los modelos de Regresión no Lineal Simple (RNLS).

## CASO 5:

Cargue los datos **EdadPesoGrasas.txt** como se muestra a continuación:

grasas <- read.table('http://verso.mat.uam.es/~joser.berrendero/datos/EdadPesoGrasas.txt', header = TRUE)

Para dichos datos, realice lo siguiente:

1. Elabore un análisis de correlación completo entre todas las variables y haga una interpretación de lo que observa. Debe utilizar al menos un recurso gráfico y uno numérico.
2. Seleccione las dos variables con mayor coeficiente de correlación y genere un modelo RLS de mejor ajuste para ambas variables. Defina claramente quién es la variable respuesta y quién la variable explicativa.
3. Analice la calidad del modelo generado a partir de su coeficiente de correlación y de determinación.
4. Genere un gráfico que muestre tanto los datos de dispersión como la recta de mejor ajuste. Interprete lo que observa en el gráfico.
5. Realice la prueba de normalidad para los residuos e indique si se cumple la condición.
6. Calcule el intervalo de confianza de 95% para los coeficientes del modelo.
7. Para una edad de 27 años, determine un IC de 95% para y un intervalo de predicción para los valores de Y asociados a dicha edad. Interprete el resultado para ambos casos.
8. ¿Es significativa la linealidad para estas variables?, ¿es posible considerar dependencia lineal entre estas? Haga la prueba de hipótesis correspondiente para responder a estas preguntas.

## CASO 6:

Supóngase que el departamento de ventas de una empresa quiere estudiar la influencia que tiene la publicidad a través de distintos medios de comunicación, sobre el número de ventas de un producto. Se dispone de un conjunto de datos que contiene los ingresos (en millones) conseguido por ventas en 200 regiones, así como la cantidad de presupuesto, también en millones, destinado a anuncios por radio, TV y periódicos en cada una de ellas.

Los datos se encuentran en el archivo DatosVentas.csv.

Para estos datos, realice los siguiente:

* Instale y cargue las librerías bbmle y DALEX

### GENERACIÓN DE LOS MODELOS

1. Genere un primer modelo denominado mod0 el cuál será un modelo RLS, usando como variables respuesta ventas y como variables exploratorias a tv. Una vez calculados los coeficientes en R para las variables y el intercepto, escriba la ecuación del modelo estimado.
2. Genere, usando la función lm de R, un primer Modelo RLM donde las variables predictoras o exploratorias sean tv, radio y periodico, la variable respuesta ventas. Escriba la ecuación provisional del modelo que se desea determinar. Guarde el modelo en una variable que debe llamar mod1

Una vez calculados los coeficientes en R para las variables y el intercepto, escriba la ecuación del modelo estimado.

De los 4 coeficientes calculados en R, ¿cuáles de ellos son estadísticamente significativos (aportan al modelo) y cuáles no lo son? Explique su respuesta, apoyándose en los datos del modelo obtenidos en R.

1. Seguidamente, genere un segundo Modelo RLM usando nuevamente la función lm, pero en este caso contemple **SOLO** las variables cuyos coeficientes resultaron ser estadísticamente significativos en el análisis anterior. Guarde este modelo en una variable que debe llamar mod2. Según los datos, ¿son todos los coeficientes significativos? Justifique su respuesta.
2. Usando los mismos datos, genere un Modelo de RNLM de la forma . Se recomienda usar nuevamente la función lm y guarde el modelo en una variable que deberá llamar mod3. Una vez que se tengan los datos del modelo, indique si los coeficientes calculados son estadísticamente significativos o no, justificando adecuadamente su respuesta.

### SELECCIONANDO EL MEJOR MODELO

Para esta sección, deberá seleccionar el mejor modelo entre los 4 ya generados en la I Parte de esta prueba, es decir: **mod0, mod1, mod2 y mod3**.

Para hacer una adecuada escogencia, debe utilizar 4 criterios a saber:

1. Coeficiente de correlación y coeficiente de determinación
2. Criterio de Información de Akaike (**AIC**), el cual, representa una medida de la calidad relativa de un modelo estadístico, para un conjunto dado de datos. Como tal, el **AIC** proporciona un medio para la selección del modelo. AIC maneja un trade-off entre la bondad de ajuste del modelo y la complejidad del modelo, basado en el principio de que un mejor modelo debe además de la precisión, ser simple en cuanto a que debe utilizar la menor cantidad de variables predictoras (Principio de parsimonia).
3. Los residuales, que están conformados por todas las diferencias entre el valor obtenido en la muestra con respecto al estimado por el modelo. Recuerde que mientras más grande sean los residuales, más se alejan los puntos del modelo y este se vuelve menos confiable.
4. Comparación de modelos mediante Análisis de Varianzas, es decir, un ANOVA para los modelos. Para este estudio deberá usar la función anova() del paquete stats.

Una vez definida la ruta, realice lo siguiente:

1. Calcule los coeficientes de correlación y de determinación de los modelos **mod0, mod1, mod2 y mod3** y haga una comparación entre estos modelos a partir de este criterio, ¿Cuál de los 3 modelos es mejor? Explique su escogencia.

Esta información deberá ser presentada en la siguiente tabla:

|  |  |  |
| --- | --- | --- |
| **Ecuación del modelo** | **Coeficiente de correlación**  **()** | **Coeficiente de determinación**  **()** |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |

1. Tome el modelo 2 (mod2) y explique el significado de los coeficientes de correlación y de determinación para ese caso.
2. Usando la función AICctab del paquete bbmle, calcule los valores del indicador AIC para cada uno de los modelos y con base en ellos, defina el orden de estos de mejor a peor. El criterio que debe usar es **menor AIC, mejor el modelo**. Instrucción para R:

AICctab(mod1, mod2, mod3, base = T, delta = T, sort = T, weights = T, nobs = #)

*Nota:* debe completar el parámetro nobs (número de observaciones), según la corresponda a la base de datos.

1. El análisis de los residuales se trabajará **solo** con los modelos mod0, mod2 y mod3 debido a que el modelo 1 y 2 resultaron ser similares. Para este trabajo debe usar algunas funciones del paquete DALEX. Seguidamente se detallan los pasos a seguir:

**Cree los Explain Models**

exp\_lm0 <- explain(mod0, data = base de datos, label = "lm", y = variable respuesta)  
  
exp\_lm2 <- explain(mod2, data = base de datos, label = "lm", y = variable respuesta)  
  
exp\_lm3 <- explain(mod3, data = base de datos, label = "lm", y = variable respuesta)

**Cree los Performance Models**  a partir de los Explain Models

lm0 <- model\_performance(exp\_lm0)  
lm2 <- model\_performance(exp\_lm2)  
lm3 <- model\_performance(exp\_lm3)

**Grafique los modelos**

plot(lm0,lm2, lm3)

Interprete el gráfico y escoja el mejor modelo, usando el criterio: **curva menor o más baja, es mejor**. Explique a partir del gráfico, por qué este criterio tiene sentido para justificar que un modelo es mejor que otro.

1. Finalmente, y como último criterio, realice el análisis de varianza. Para escoger el mejor modelo solamente debe tomar en cuenta el valor de Residual Sum of Squares(**RSS**), donde a menor valor, mejor es el modelo, es decir, los puntos están más ajustados.

Considere la siguiente instrucción para realizar este análisis:

anova(mod0,mod1, mod2, mod3)

# III PARTE: MODELOS DE REGRESIÓN NO LINEAL

## CASO 7:

Para esta sección deberá generar dos modelos de RNLS usando los datos de bones de la base de datos jaws, la cual contiene información sobre la longitud de la mandíbula de los venados, según la edad.

1. Para iniciar cargue el archivo jaws.txt. Asegúrese de que este archivo esté en la misma carpeta donde están los archivos creados en RStudio para el proyecto.

var <- read.table("nombre.txt",header=T)

1. Luego, construya una gráfico de dispersión para los datos:

ggplot(datos,aes(x = ..., y = ...)) +  
 geom\_point()  
# en lugar de geom\_point(), también puede usar geom\_jitter()

1. Usando la función nlm() del paquete stats, van a generar un modelo de RNLS de la forma $ y = a(1- e^{-c x })$. A este modelo asígnele el nombre de model1. Use como valores iniciales **a = 120** y **c = 0.064**. Escriba la ecuación del modelo usando los valores para los parámetros obetinos en R.
2. Realice el gráfico del modelo 1 (model1) junto con su gráfico de dispersión.
3. Deberán crear un segundo modelo a partir de los modelos estudiados en clase. Analice con detenimiento la forma de los datos y haga una selección apropiada.
4. Una vez escogido el modelo, haga un proceso de linealización y estime los parámetros a partir de un modelo de RLS.
5. Use los valores de las estimaciones de los coeficientes del modelo de RLS obtenidos en el punto anterior, como los valores iniciales para generar el modelo de RNLS a partir de la función nlm().
6. Genere el Modelo de RNLS y escriba la ecuación resultante. Además, realice el gráfico respectivo de este modelo junto con el gráfico de dispersión de los datos.
7. Utilice la función *anova* para comparar ambos modelos. ¿cuál modelos es el mejor?, gráficamente, ¿se puede llegar a esa misma conclusión?

# Fases del proyecto y detalles de entrega

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Fase 1** | **Fase 2** | **Fase 3** |
| Valor total | 15% | 15% | 20% |
| Entrega de informe detallado del análisis realizado que involucra las partes: | Partes I y II  Valor 12% | Parte III  Valor 12% | Parte IV y V  Valor 16% |
| Fecha de entrega del informe: | A más tardar:  **17 de abril** | A más tardar:  **14 de mayo** | A más tardar:  **2 de junio** |
| Exposición del trabajo realizado | Valor 3% | Valor 3% | Valor 4% |

## Observaciones generales:

* Cada grupo deberá crear una Bitácora de minutas de cada una de las reuniones, donde en cada una de ellas aparezca el número de minuta, fecha, presentes, hora de inicio, agenda, acuerdos y hora de cierre. La primera minuta deberá contener una distribución de tareas por persona y un plan de trabajo que visibilice el alcance de la meta en el plazo. Se adjuntará una plantilla de minuta, por si desean considerarla como base:
* El informe debe crearse en **RStudio** y exportarse como documento de **Word**. Este informe debe subirse en la sección de tareas que se creará en el grupo general, pero sólo se le dará acceso a las personas que realizaron las entregas anteriores. No es necesario que se adjunten los documentos *rmd*, solo en caso de que estos sean solicitados.
* Todos los cálculos numéricos o estadísticos, estimaciones, gráficos de análisis, trabajo con la base de datos, deberá realizarse en *RStudio*.
* La autoevaluación, coevaluación y exposición, se rigen bajo las misas reglas explicadas para las entregas anteriores.
* La entrega de esta fase III será para el **lunes 14 de junio a las 8:00am** en la plataforma de Teams.