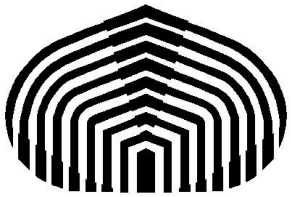


Universidad Simón Bolívar

Departamento de Cómputo Científico y Estadística

CI3321 – Estadística para Ingeniería

Trimestre Enero-Marzo 2012

INFORME II

Integrantes: Fátima Querales 07-41388

Vicente Santacoloma 08-11044

Sartenejas, 26 de marzo de 2012

INTRODUCCIÓN

En la vida cotidiana se presentan hechos que dependen de distintos factores y se necesita explicar o evidenciar este acontecimiento a través de un método matemático llamado regresión lineal.

En el presente proyecto se realizará el análisis de regresión lineal necesario para estudiar la influencia de 5 variables (Agriculture, Examination, Education, Catholic, Infant.Mortality) que evalúan las medidas de fertilidad e indicadores socio-económicos para algunas de las provincias de habla francesa de Suiza.

Para llevar a cabo este evento se elaborará un análisis descriptivo para observar el comportamiento de los datos por medio de diagramas de dispersión, seguidamente, a través de análisis de residuos, se encontrará cual de las variables independientes tiene mayor influencia sobre la variable en estudio. Luego, se presentarán varios modelos que expliquen la variabilidad, y a partir de métodos estadísticos se tendrán las pruebas necesarias para elegir el mejor modelo.

RESUMEN

Se tiene un registro de los datos de 47 provincias de habla francesa de Suiza en el año 1888, referente al estudio de la medida de fertilidad e indicadores socio – económicos de cada una. Las variables que representan a los indicadores socio – económicos están todas en porcentaje y son: Agriculture correspondiente a la población que se basa en la agricultura, Examination que establece los reclutas que obtuvieron el máximo puntaje en los exámenes el ejército, Education indica la población que ha recibido educación más allá de la escuela primaria, Catholic la cantidad de católicos y por último la variable Infant.Mortality referida a la mortalidad infantil.

Se realizó un análisis descriptivo, numérico y gráfico de las mismas consiguiendo algunas más influyentes que otras con respecto a su ajuste lineal. Además, se estudiaron las relaciones halladas entre cada una de las variables, y se hallaron diferentes modelos lineales. Por medio de la realización del análisis de residuo se encontró un modelo ajustado mejorado que representara la función de la mejor manera. Además se realizaran algunas pruebas de hipótesis como la diferencia de medias, así como también se empleará pruebas

de bondad de ajuste y método de ANOVA para hacer algunas inferencias sobre este grupo de datos.

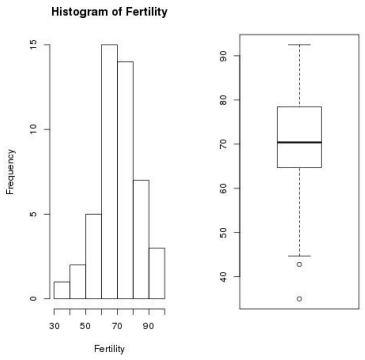
Como herramienta fundamental para la obtención de todos los valores se utilizó el software R, para la estimación y diagnostico de los modelos de regresión lineal hallados.

PREGUNTA 1

	Fertility	Agriculture	Examination	Education	Catholic	Infant.Mortality
Min.	35.00	1.20	3.00	1.00	2.150	10.80
1st Qu.:	64.70	35.90	12.00	6.00	5.195	18.15
Median	70.40	54.10	16.00	8.00	15.140	20.00
Mean	70.14	50.66	16.49	10.98	41.144	19.94
3rd Qu.	78.45	67.65	22.00	12.00	93.125	21.70
Max.	92.50	89.70	37.00	53.00	100.000	26.60

FERTILITY

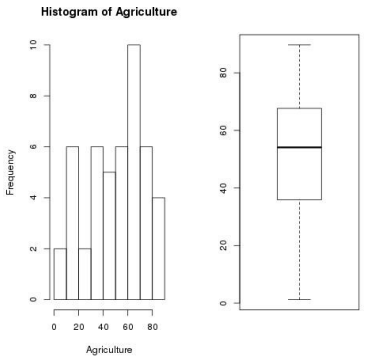
Fertility: Variable cuantitativa que representa una medida común de fertilidad estándar.



El histograma muestra que la mayoría de las provincias tienen un porcentaje de fertlidad entre el 50% y el 80% obtenidos a partir de una muestra la muestra de 47 provincias. Si se compara el boxplot con el histograma se puede apreciar que la distribución establece que la mayoría de las provincias tienen un nivel de fertilidad que esta que está por encima del 50% y por debajo del 80%. Además se observan dos

AGRICULTURE

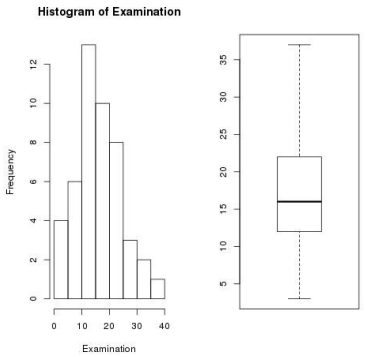
Agriculture: Variable cuantitativa que representa el porcentaje de hombres involucrados en la agricultura como ocupación.



En el gráfico se puede observar que la población dedicada a la agricultura en aproximadamente 10 países se establecen alrededor del 60% y que hay 4 grupos de 6 provincias que se establecen cada 20%. En este boxplot se puede observar que los datos no parecen distribuirse de forma simétrica alrededor de su mediana y no hay datos atípicos

EXAMINATION

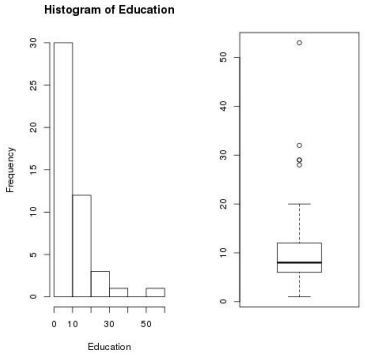
Examination: Variable cuantitativa que representa el porcentaje de reclutas que reciben la marca más alta en el examen del ejército.



El histograma muestra que la mayoría de las provincias tienen entre el 10% y 30% de reclutas que obtuvierón el máximo puntaje en los exámenes del ejército. Al comparar el boxplot con el histograma se puede apreciar que la distribución de las provincias referente a la variable Examination si se encuentra por encima del 10% y por debajo del 30%. Además no hay una distribución simétrica de los datos

EDUCATION

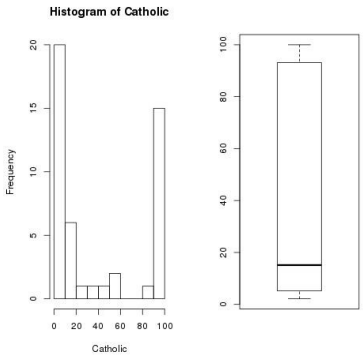
Education: Variable cuantitativa que representa el porcentaje de reclutas con educación más allá de la escuela primaria.



La mayoría de los países tienen una educación por debajo del 20 %. Además se tiene una mínima cantidad de provincias que se ubican por el 50%. Para este boxplot se tiene que los datos no se distribuyen de manera simétrica alrededor de su mediana, pero se presentan dos observaciones mayores que el resto, las cuales se consideran atípicas y representan pocas provincias cuyo porcentaje en educativo esta alrededor del 30 y el 50%. Relacionandolo con el histograma se tiene que efectivamente la mayoría de las provincias

CATHOLIC

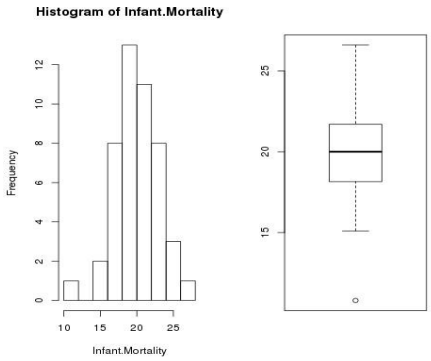
Catholic: Variable cuantitativa que representa el porcentaje de católicos protestantes.



Con respecto a esta variable, se puede observar que poco menos del 50 % de las provincias se ubica en 10 %. Mientras que aproximadamente el 31% de las provincias están alrededor del 90%. Las provincias restantes que son minorías están repartidas entre el 10% y el 90%. No hay provincias entre el rango del 60% al 80%. Los datos típicos están distribuidos en el boxplot de maneta

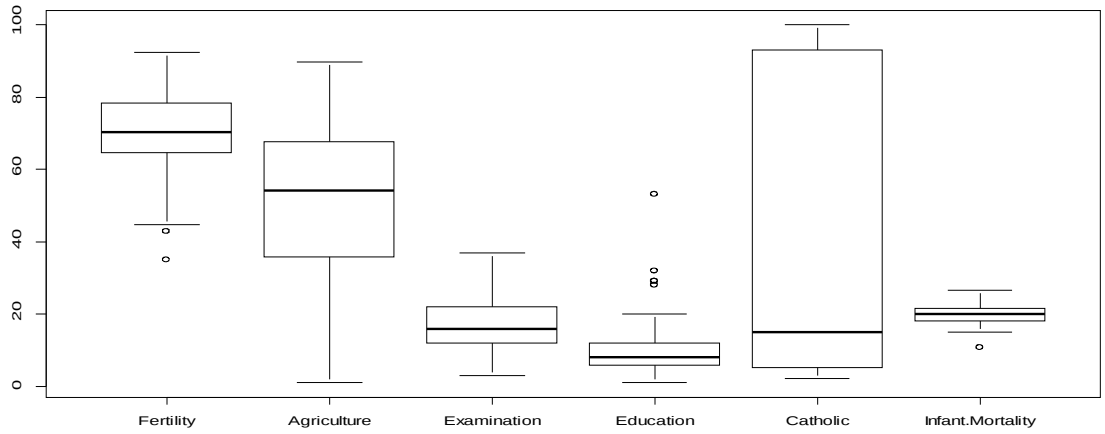
INFANT MORTALITY

Infant.Mortality: Variable cuantitativa que representa el número de nacidos vivos que viven menos de un año.



Más del 50 % de los países tienen una mortalidad infantil por encima del 15%. Además se tiene una mínima cantidad de provincias que se ubican por el 10%. Si se compara el boxplot con el histograma se apoyan las observaciones de que la mayoría de los países tienen más del 15% en mortalidad infantil. Además, se presenta un dato atípico equivalente a una provincia con

DIAGRAMA DE BOXPLOT

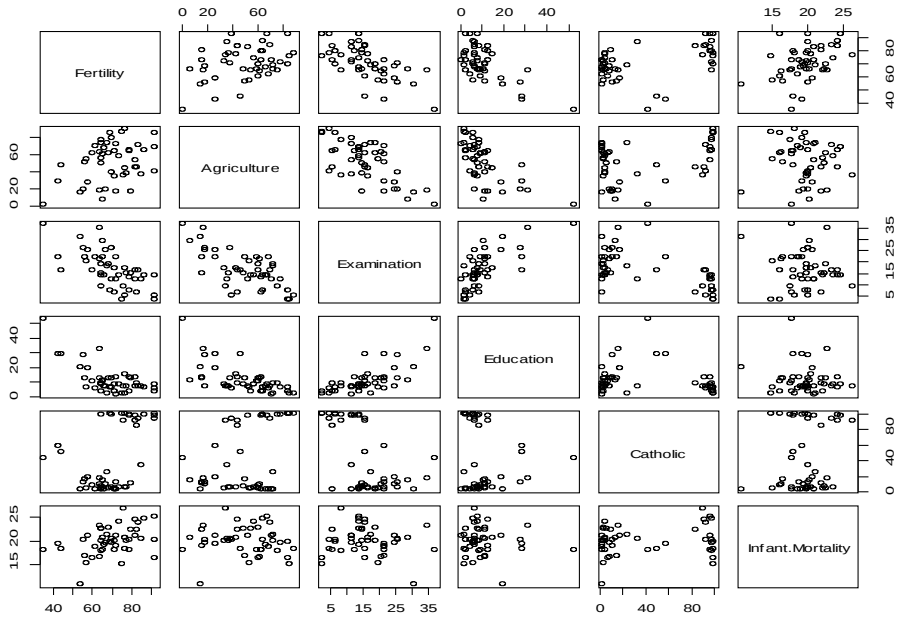


PREGUNTA 2

MATRÍZ DE CORRELACIÓN

	Fertility	Agriculture	Examination	Education	Catholic	Infant.Mortality
Fertility	1.0000000	0.3530791 8	-0.6458827	-0.6637888 6	0.4636847	0.41655603
Agriculture	0.3530792	1.0000000 0	-0.6865422	-0.6395225 2	0.4010951	-0.06085861
Examination	-0.6458827	-0.6865422 1	1.0000000	0.6984153 0	-0.5727418	-0.11402160
Education	-0.6637889	-0.6395225 2	0.6984153	1.0000000 0	-0.1538589	-0.09932185
Catholic	0.4636847	0.4010950 5	-0.5727418	-0.1538589 2	1.0000000	0.17549591
Infant.Mortality	0.4165560	-0.0608586 1	-0.1140216	-0.0993218 5	0.1754959	1.00000000

DIAGRAMA DE DISPERSIÓN



Teniendo como base la matriz de correlación de los datos, así como los diagramas de dispersión de las variables se observa que la variable Fertility está negativamente correlacionada con las variables Education y Examination. Es decir, las poblaciones que poseen un alto índice de educación se relacionan con un índice menor de fertilidad. Además, a medida que aumenta la fertilidad también aumenta la mortalidad infantil. De igual manera, se puede observar que las regiones con mayor porcentaje de personas agrícolas y católicas también presentan un mayor índice de fertilidad.

Además, las regiones con mayor proporción de personas dedicadas a la agricultura tienden a tener menores índices de educación y mayores índices de católicos. Las regiones con mayor educación presentan una menor proporción de católicos y bajo índices de mortalidad infantil.

PREGUNTA 3

MODELO 1:

lm(formula = f ~ a + ex + ed + c + im)
Residuals:

Min	1er Q	Media na	3er Q	Max

-15.2743	-5.2617	0.5032	4.1198	15.3213

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	66.91518	10.70604	6.250	1.91e-07 ***
A	-0.17211	0.07030	-2.448	0.01873 *
Ex	-0.25801	0.25388	-1.016	0.31546
Ed	-0.87094	0.18303	-4.758	2.43e-05 ***
C	0.10412	0.03526	2.953	0.00519 **
Im	1.07705	0.38172	2.822	0.00734 **

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7.165 on 41 degrees of freedom
Multiple R-squared: 0.7067, Adjusted R-squared: 0.671
F-statistic: 19.76 on 5 and 41 DF, p-value: 5.594e-10

En la columna Estimate se puede observar el valor del Bi, en la columna t-value se tienen los valores del estadístico t para cada uno de los valores estimados y en la columna p-value se encuentran los valores correspondientes a los p-valores.

Los p-valores de las variables sirven para rechazar la hipótesis nula, es decir, es el nivel más pequeño de significancia alfa para el cual la hipótesis nula debe ser rechazada. Para nuestro caso, se rechaza dicha hipótesis si el p-valor de la variable es menor que el alfa establecido de 5%. En la última columna se puede ver que la variable ex tiene valor grande de p-valor, por lo que no se rechazaría (se acepta) la hipótesis nula de que los coeficientes son iguales a cero.

Considerando este resultado, la variable ex no debe ser incluida en el modelo de regresión lineal. Por lo tanto eliminamos el modelo 1 y se procede a evaluar el modelo 2, que no posee la variable ex.

MODELO 2:

lm(formula = f ~ a + ed + c + im)

Residuals:

Min	1er Q	Mediana	3er Q	Max
-14.6765	-6.0522	0.7514	3.1664	16.1422

Coefficients:

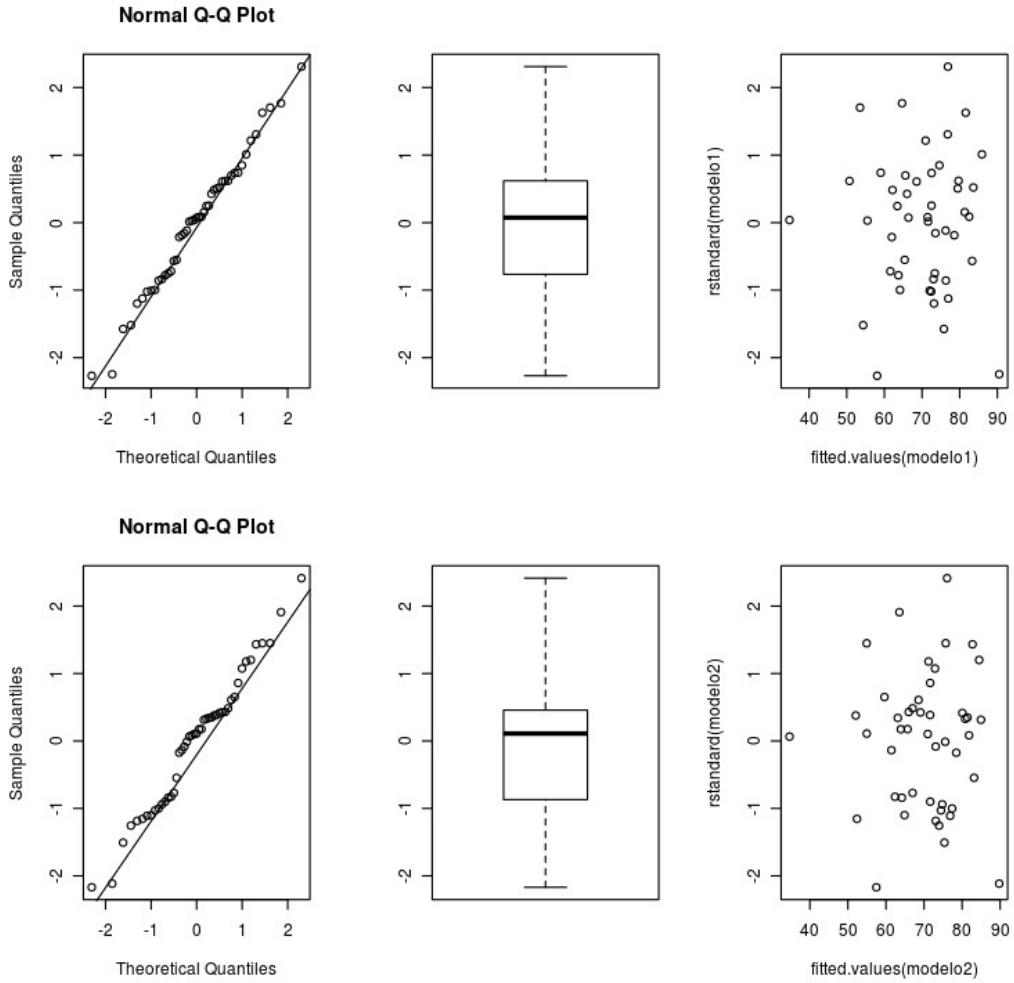
	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	62.10131	9.60489	6.466	8.49e-08 ***
a	-0.15462	0.06819	-2.267	0.02857 *
Ed	-0.98026	0.14814	-6.617	5.14e-08 ***
c	0.12467	0.02889	4.315	9.50e-05 ***
im	1.07844	0.38187	2.824	0.00722 **

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7.168 on 42 degrees of freedom
Multiple R-squared: 0.6993, Adjusted R-squared: 0.6707
F-statistic: 24.42 on 4 and 42 DF, p-value: 1.717e-10

El análisis es similar al modelo 1, por lo tanto para el modelo 2, se concluye que las hipótesis nulas de todas las variables se rechazan, por lo que todas las variables participan en la función de regresión del modelo 2. Se puede suponer que el modelo 2 es el modelo más acertado para la fertilidad. Además, otra manera de observar que tan bueno es el modelo 2 con respecto al modelo 1 es a partir de la comparación de los R^2 y el R^2 ajustado hallados de ambos modelos.

A partir de los valores obtenidos para cada modelo se analizan los errores de las variables para cada uno (análisis de residuos). Se demostrara que los errores de cada modelo siguen una distribución N(0,1) (normalidad) y además se podrá conocer si el modelo cumple homocedasticidad e independencia. A continuación se presentan las gráficas de estas condiciones para cada modelo:



La primera fila del gráfico es referente al modelo 1 y se observa que los residuos siguen una distribución normal. La relación entre las variables es lineal y los datos se ajustan a la recta, por lo tanto la condición de normalidad se cumple al igual que la homocedasticidad, pero el p-valor de una de sus variables es mayor que 0.05, por lo tanto el modelo ya no cumple.

En la segunda fila del gráfico se tiene la representación del modelo 2, el cual a pesar que sus variables rechazan la hipótesis nula, los errores

estandarizados no cumplen con que deben distribuirse equitativamente alrededor de la recta que mejor se adapta a los errores estandarizados, sin embargo vemos que cumple con la homocedasticidad.

Se realizara un modelo 3 que cumpla con que la variable agricultura sea excluida del modelo de regresión lineal. Esto es realizado para mejorar la distribución de la grafica 1 del modelo 2.

MODELO 3:

lm(formula = f ~ ed + c + im)

Residuals:

Min	1er Q	Mediana	3er Q	Max
-14.4781	-5.4403	-0.5143	4.1568	15.1187

Coefficients:

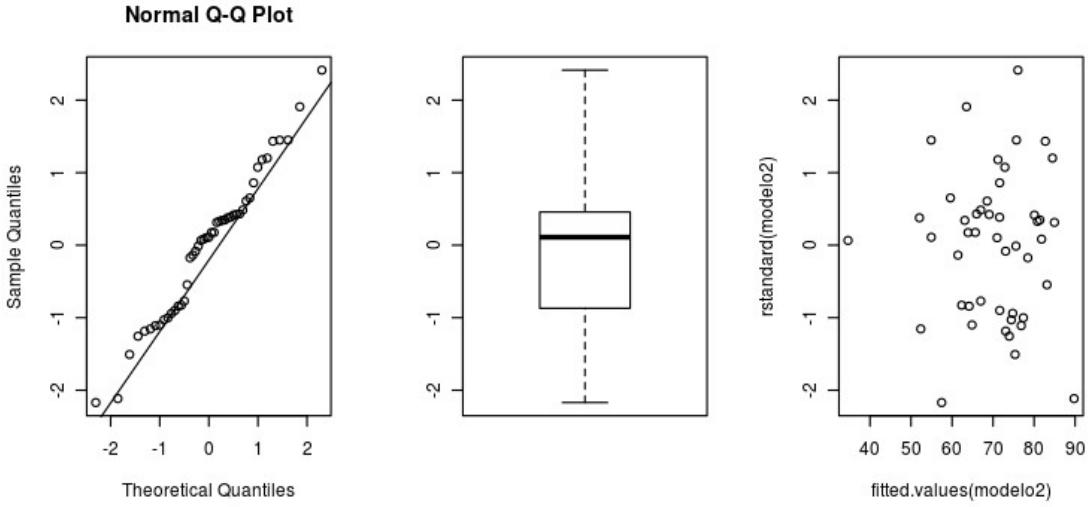
	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	48.67707	7.91908	6.147	2.24e-07 ***
ed	-0.75925	0.11680	-6.501	6.83e-08 ***
c	0.09607	0.02722	3.530	0.00101 **
im	1.29615	0.38699	3.349	0.00169 **

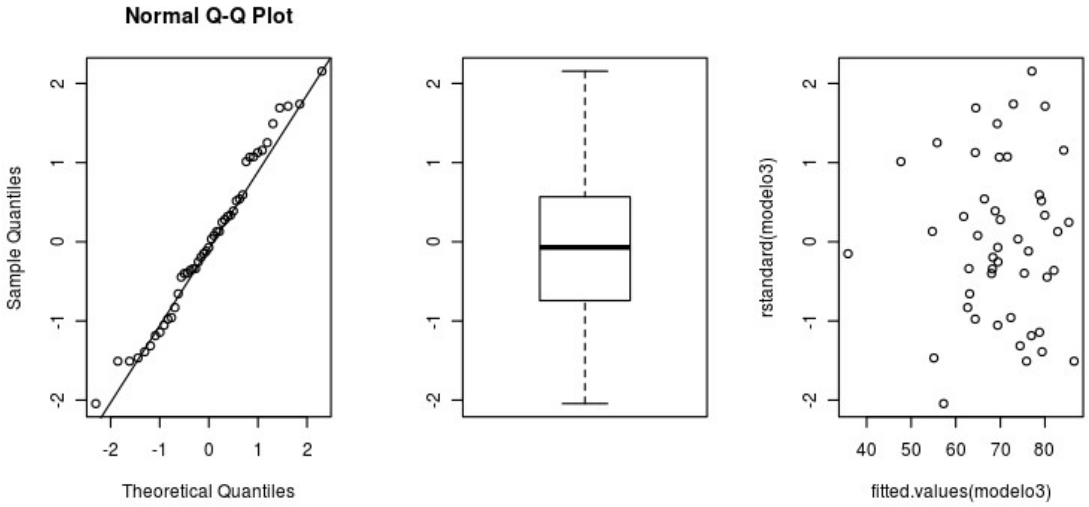
Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7.505 on 43 degrees of freedom
Multiple R-squared: 0.6625, Adjusted R-squared: 0.639
F-statistic: 28.14 on 3 and 43 DF, p-value: 3.15e-10

Por medio de este modelo, se puede apreciar que los p-valores de las variables cumplen con que son menores que 0.05, por lo tanto forman parte de la función de regresión. Además, la diferencia entre Multiple R-squared y Adjusted R-squared es menor que en los 2 modelos anteriores.

A continuación se presentan las gráficas de los errores de estas condiciones para los modelos 2 y 3 respectivamente:





Se puede apreciar que en la gráfica 1 del modelo 3 los puntos se distribuyen mejor alrededor de la recta y además cumple con la homocedasticidad. Por lo tanto, se puede suponer que el mejor modelo de regresión lineal para la fertilidad es el modelo 3.

PREGUNTA 4

La ecuación queda como:

$$FERTILITY_i = 48.67707 - 0.75925 * Ed_i + 0.09607 * C_i + 1.29615 * \mathrm{Im}_i$$

Donde el valor de los coeficientes equivale a la columna Estimate del modelo seleccionado y el 48.67707 corresponde al punto donde la recta corta al eje de FERTILITY y los demás son los coeficientes respectivos de cada variable.

PREGUNTA 5

Modelo 1

Predicción:

	fit	lwr	upr
1	26.24275	-0.2310307	52.71654
2	35.98204	12.0964844	59.86759
3	72.24987	51.4217201	93.07802
4	78.44163	57.6783427	99.20492

Confianza:

	fit	lwr	upr
1	26.24275	8.180451	44.30506
2	35.98204	21.985425	49.97865
3	72.24987	64.555893	79.94384
4	78.44163	70.925001	85.95827

Modelo 2

Predicción:

	fit	lwr	upr
1	24.40058	-1.596605	50.39777
2	35.72415	11.866659	59.58164
3	72.82001	52.062917	93.57710
4	78.39431	57.647244	99.14137

Confianza:

	fit	lwr	upr
1	24.40058	7.027877	41.77329
2	35.72415	21.755003	49.69330
3	72.82001	65.282502	80.35751
4	78.39431	70.884462	85.90415

Modelo 3

Predicción:

	fit	lwr	upr
1	31.69704	6.067747	57.32633
2	43.51343	20.520933	66.50592

3	67.39184	46.759855	88.02383
4	75.20823	53.874268	96.54218

Confianza:

	fit	**lwr**	**upr**
1	31.69704	15.95874	47.43534
2	43.51343	32.58236	54.44450
3	67.39184	63.32830	71.45539
4	75.20823	68.42802	81.98843

PREGUNTA 6

Como el investigador afirma que el índice de fertilidad de las provincias con un porcentaje de educación mayor e igual a 9.6% es menor, en promedio, a los índices de fertilidad de las provincias con un porcentaje de educación menor a 9.6%, se tiene que:

Ho: u1 –u2 = 0

Ha: u1 – u2 < 0

Como los datos son pequeños, tendremos para elegir, tres diferentes contrastes de hipótesis para diferencias de medias. Estos son para varianzas conocidas, varianzas desconocidas e iguales y varianzas desconocidas y desiguales. Para este caso de estudio las varianzas son desconocidas. Quedará verificar si éstas son iguales o desiguales. Para ello se empleará el contraste sobre igualdad de varianzas.

A continuación se hará el análisis en base a la salida obtenida a través de la función programada en R **(VER EN ANEXO PARTE 1)**.

Verificando si las varianzas son iguales o desiguales (esto es, **Ho:** $\sigma_1 = \sigma_2$,

Ha: $\sigma_1 \neq \sigma_2$) mediante var.test se obtuvo:

F = 1.831031

Intervalo de Confianza = (0.7991662, 4.5237714)

p.valor = 0.15086

Como F no cae en la región de rechazo no podemos rechazar la Ho. Haciendo un análisis del p.valor como este no es ni muy pequeño ni muy grande no tenemos fundamento para aceptar o rechazar. Siguiendo un análisis poco robusto se puede suponer en base al α (0.025) de la prueba, que por ser menor al p.valor, las varianzas son iguales (Esta suposición no tiene ningún fundamento claro pues el α depende de lo elegido por el estadístico y no por los datos del experimento).

De la suposición anterior de que las varianzas son iguales, aplicamos el contraste de hipótesis para varianzas desconocidas e iguales, obteniendo los siguientes resultados:

Region de Rechazo= (∞,-2.01410338888085)

Estadistico Z: -3.5664697430162

p.valor = 0.999563901509002

Como el estadístico cae dentro de la región de rechazo, rechazamos la Ho.

PREGUNTA 7

Se quiere probar que:

Ho: Los datos de Infant.Mortality se ajustan a una distribución normal.

Ha: Los datos de Infant.Mortality no se ajustan a una distribución normal.

Ejecutando el código en R que contempla pruebas de bondad de ajuste **(VER EN ANEXO PARTE 2)**, se obtiene los siguientes resultados:

X2.obs = 3.275482

X2.alpha = 9.487729

p.valor = 0.5128263

Como X2.obs no cae dentro de la región de rechazo (9.487729, ∞), se tiene que en base a la evidencia no podemos rechazar la hipótesis nula de que los datos se ajustan a una distribución normal.

PREGUNTA 8

Planteando las hipótesis:

Ho: uG1 = uG2 = uG3 = uG4

Ha: Algún ui es distinto de los otros de los

Mediante el código propuesto en R **(VER ANEXO PARTE 3)**, se obtuvo la siguiente tabla de anova de acuerdo a la agrupación especificada en el enunciado:

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Group	3	23.7	7.886	0.925	0.437
Residuals	43	366.6	8.525		

Donde Group es el Tratamiento.

De acuerdo al p-valor obtenido

Suponiendo un α = 0.05. Calculamos F 0.05; 43; 3 = 2.821628. Luego como F = 0.925 se tiene que no cae en la región de rechazo. Por lo tanto no podemos rechazar la **Ho**. Además viendo el p-valor se puede llegar a la misma conclusión ya que con su valor no es ni muy pequeño ni muy grande, no se puede ni aceptar, ni rechazar, por lo que en base a la evidencia solo podemos no rechazar la **Ho**.

CONCLUSIÓN

Luego de la realización de este proyecto acerca del de pruebas de hipótesis y modelos lineales para un registro de los datos de 47 provincias de habla francesa de Suiza en el año 1888 se logró tener los siguientes resultados:

- Los modelos lineales son fundamentales para predecir el valor que va a tomar una variable.
- El diagrama de correlación permite identificar fácil y claramente la dependencia entre variables.
- Mediante pruebas de bondad de ajuste se puede probar si un conjunto de datos se ajustan a una determinada distribución, bien sea continua o discreta.
- El método de ANOVA permite de una manera sencilla calcular diferencia de en el porcentaje promedia de una determinada variable para un grupo de datos.

BIBLIOGRAFÍA

Dennis D. Wackerly, William Mendenhall, Richard L. Scheaffer. ***"Estadística Matemática con Aplicaciones"***. Editorial: CENGAGE Learning. Edicion: 7ma.

ANEXO

PARTE 1

```
# Funcion para pruebas de hipotesis de diferencias de medias para muestra

# pequenas.

#

# Devuelve:

# true si se rechaza la hipotesis nula.

# false si no se puede rechazar la hipotesis nula.

#

# Parametros:

# x1 = datos del grupo 1

# x2 = datos del grupo 2

# m1 = media del grupo 1

# m2 = media del grupo 2

# s1 = desviacion estandar del grupo 1 (puede ser la real o aproximada segun el caso)

# s2 = desviacion estandar del grupo 2 (puede ser la real o aproximada segun el caso)

# n1 = tamano del grupo 1

# n2 = tamano del grupo 2

# tipo de prueba: 1 => Ha: Z > Zalfa, 2 => Ha: Z < Zalfa, 3 => Ha: |Z| > Zalfa

# varConocida: 1 si las varianzas son conocidas. 0 si no lo son.

# alpha: nivel de significancia (0.05 por defecto)

# u0: valor del lado derecho de la prueba de hipotesis nula. Por defecto se confideran

# iguales las dos medias
```

```
PHDiferenciaMedias <- function(x1,x2,m1,m2,s1,s2,n1,n2,tipo,varConocida,alpha = 0.05,u0=0)
{
  if(n1 <= 0 || n2 <= 0 || !(1 <= tipo & tipo <=3) || !(varConocida != 0 || varConocida != 1)) {
    cat("Datos Invalidos")
    return()
  }
  if(n1 >= 30 | n2 >= 30) {
    cat("No Aplicable a la Hipotesis de Diferencia de Medias para Muestras Pequenas")
    return()
  }
  if(varConocida == 0) {
    varTest = var.test(x1,x2)
    F = unique(unlist(varTest$statistic))
    p.valor = unique(unlist(varTest$p.value))
    intConf = unique(unlist(varTest$conf.int))
    varIguales = 0
    if(F <= intConf[1] || intConf[1] >= F)
      varIguales = 0
    else if(p.valor <= 0.0005 || p.valor >= 0.9 || p.valor >= alpha)
      varIguales = 1
    if(varIguales == 1) {
      Z = m1-m2-u0
      Z = Z/sqrt((((n1-1)*s1^2+(n2-1)*s2^2)/(n1+n2-2))*((1/n1)+(1/n2)))
      p.valor = pt(Z,n1+n2-2,lower.tail=FALSE)
      if(tipo == 3) {
```

```
        alpha = alpha/2
    p.valor = p.valor*2
   }
   RR = qt(alpha,n1+n2-2,lower.tail=FALSE)
  }
  else {
   Z = m1-m2-u0
   Z = Z/sqrt(s1^2/n1+s2^2/n2)
   v = (s1^2/n1+s2^2/n2)^2
   v = v/((s1^2/n1)^2*(1/(n1-1))+(s2^2/n2)^2*(1/(n2-1)))
   v = v-2
   p.valor = pt(Z,v,lower.tail=FALSE)
   if(tipo == 3) {
        alpha = alpha/2
    p.valor = p.valor*2
   }
   RR = qt(alpha,v,lower.tail=FALSE)
  }
}
else {
  Z = m1-m2-u0
  Z = Z/sqrt((s1^2/n1)+(s2^2/n2))
  p.valor = pnorm(Z, mean = 0, sd = 1, lower.tail = FALSE)
  if(tipo == 3) {
   alpha = alpha/2
   p.valor = p.valor*2
  }
  RR = qnorm(alpha, mean = 0, sd = 1, lower.tail = FALSE)
}
if(tipo == 3)
```

```
    print(c('Region de Rechazo: ',-RR,RR))
  else if(tipo == 1)
    print(c('Region de Rechazo: ', -RR))
  else
    print(c('Region de Rechazo: ',  RR))
  print(c('Estadistico Z: ',Z))
  print(c('p-valor: ',p.valor))
  abs(Z) > abs(RR)
}
x1 = Fertility[Education >= 9.6]
x2 = Fertility[Education <  9.6]
m1 = mean(x1)
m2 = mean(x2)
n1 = length(x1)
n2 = length(x2)
s1 = sd(x1)
s2 = sd(x2)
tipo = 1
alpha = 0.025
varConocida = 0
b = PHDiferenciaMedias(x1,x2,m1,m2,s1,s2,n1,n2,tipo,varConocida,alpha)
```

PARTE 2

```
# Construyendo la tabla de distribucion de frecuencias para Infant.Mortality
n = length(Infant.Mortality)
k = sqrt(n)
k = round(k)
l = max(Infant.Mortality) - min(Infant.Mortality)
l = l / k
f = matrix(nrow = k, ncol = 1)
```

```
class = matrix(nrow = k, ncol = 2)
class[1,1] = min(Infant.Mortality)
class[1,2] = class[1,1] + l
for (i in 2:k) {
  class[i,1] = class[i-1,2]
  class[i,2] = class[i,1] + l
}
for (i in 1:k) {
  c = 0
  for(j in 1:n) {
    if (class[i,1] <= Infant.Mortality[j] && Infant.Mortality[j] < class[i,2]) {
      c = c + 1
    }
  }
  f[i] = c
}

# Pruebas de Bondad de Ajuste

# Ho: Los datos de Infant.Mortality se ajustan a una distribucion normal
# Ha: Los datos de Infant.Mortality no se ajustan a una distribucion normal

alpha = 0.05
r = 3
m = matrix(nrow = k, ncol = 1)
for (i in 1:k) {
  m[i] = class[i,1] + (l/2)
}

mean = sum(f * m)/n
```

```
variance = (sum(f * m^2) - ((sum(f * m))^2)/n) / (n-1)

#variance = sum(f * (m-mean)^2)/n

sd = sqrt(variance)


p = pnorm(class[,2],mean,sd) - pnorm(class[,1],mean,sd)


X2.obs = sum((f - n * p)^2 / (n * p))

X2.alpha = qchisq(1 - alpha,k-r)


# Calculando el p-valor:

p.value = 1 - pchisq(X2.obs,k-r)
```

PARTE 3

```
q1 = quantile(Education,.25)

q2 = quantile(Education,.50)

q3 = quantile(Education,.75)


G1 = Infant.Mortality[Education < q1]

G2 = Infant.Mortality[q1 <= Education & Education < q2]

G3 = Infant.Mortality[q2 <= Education & Education < q3]

G4 = Infant.Mortality[Education >= q3]

n1 = length(G1)

n2 = length(G2)

n3 = length(G3)

n4 = length(G4)

Infant.Mortality = c(G1,G2,G3,G4)

Group = factor(rep(LETTERS[1:4],c(n1,n2,n3,n4)))

Infant.Mortality.df = data.frame(Group,Infant.Mortality)

aov.Infant.Mortality = aov(Infant.Mortality ~ Group, Infant.Mortality.df)

summary = summary(aov.Infant.Mortality)
```