Proyecto final de Extremo

Guevara Díaz Karlo Jair (Licenciatura en Matemáticas).

Lunes 29 de Noviembre del 2021.

1. Introducción

Al querer analisar datos es muy comun encontrarse con series de tiempo, en particular es muy comun querer analizar series de tiempo no independientes. Por lo que es bueno contar con herramientas para tratar con estos datos. Por lo que en este trabajo presentaremos algunas para tratar con series de tiempo no independientes. Estas herramientas son los modelos AR(p), AM(q), ARMA(p,q) y teria de extremos aplicada a seres de tiempo. Pero antes de ver estos metodos, explicaremos lo que son las series de tiempo, una pequeña explicacion de lo que es el ruido blanco. Tambien veremos la funcion de autocorelacion que nos ayudar para determinar cuando es viable aplicar estos metodos y presentaremos algunos ejemplo de como se ajustan estos metodos a datos reales y resultados obtenidos usando estos metodos.

2. Desarrollo

Series de tiempo

Iniciaremos dando una breve explicación de lo que son las series de tiempo, dando una definición de estas, presentando algunos tipos de seres de tiempo, destacando algunas características de algunas series de tiempo y por ultimo dando un par de ejemplos de series de tiempo.Una serie temporal es un proceso estocástico uninvariante que consiste en un conjunto de variables aleatoreas indexadas por el tiempo. Si hay $T \subset [0, \infty)$ variables, se denota por $\{X_t | t \in T\}$ o $\{X_t\}_{t \in T}$.
También existen las series de tiempo multivariantes (en este caso T podría ser $T = \{1, ..., n\} \times \{1, 2, 3\}$ para algun $n \in \mathbb{N}$), donde se analizan varias series temporales a la vez. Por ejemplo la venta de gasolina, la venta de carros y la venta de vuelos ya que se sabe que hay una estrecha relación entre estas 3 series de tiempo por lo que el comportamiento de una afecta al comportamiento de las otras.
Nos enfocaremos en el analisis de las series de tiempo univariadas. Tomaremos series de tiempo indexadas en un subconjunto de los naturales i.e $\{X_t \ | t \in [n]\}$ con $[n] = \{1, ..., n\}$ para algún $n \in \mathbb{N}$ o $\{X_t\}_{t \in \mathbb{N}}$.
Un ejemplo de una serie de tiempo univariadas seria, el nivel medio global del mar al tiempo t. Por lo que X_t seria el nivel medio global del mar al tiempo t y $\{X_t(w)\}_{t \in [n]}$ (con w fijo) seria una muestra de la serie de tiempo, donde su gráfica esta dada por

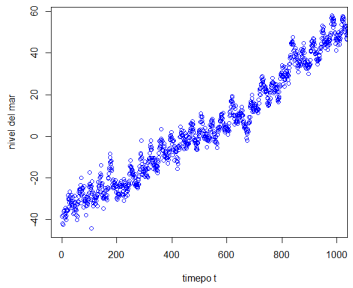

Figura 1

Definición **Definición** Otra serie de tiempo que podemos obtener de los mismos muestra o datos seria la de las diferencias, con esto nos referimos al aumento del nivel medio global del mar. Lo que se obtiene de la siguiente manera $X_t(w) = X_{t+1}(w) - X_t(w)$, donde su gráfica esta dada por

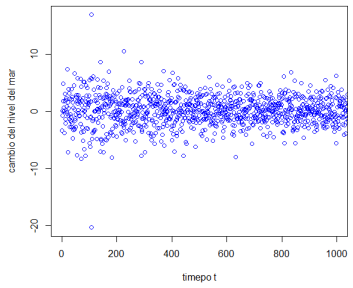

Figura 2

Lo usual en este tipo de gráficas es que se presenten de la siguiente forma,

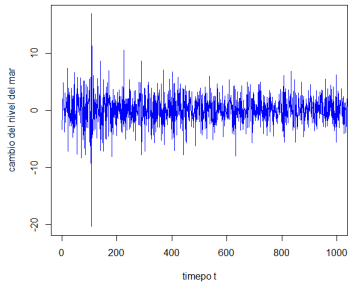

Figura 3

Notemos que, aun que esta ultima serie de tiempo proviene de los mismos datos que la primera que presentamos, veremos que tiene características muy diferente entre si.Veremos contención que la segunda serie de tiempo tiene propiedad que nos ayudaran a estudiarla con métodos como AR(p), MA(q) o ARMA(p,q).

Estacionalidad

Presentaremos algunas propiedades que nos ayudaran a manipular seres de tiempo no independientes. Estas propiedades nos ayudan a realizar los ajustes de algunos modelos que veremos, por ejemplo una propiedad es la estacionalidad. Usaremos dos tipos de estacionalidad muy importante. La estacionariedad estricta y la estacionariedad en covarianza. La estacionalidad estricta se define como

Definición Una serie de tiempo $\{X_t\}_{t\in T}$, se dice que es estacionaria en el sentido estricto si $(X_{t_1}, ..., X_{t_n})$ y $(X_{t_1+k}, .., X_{t_n+k})$ son iguales en distribución para cualquier $n, t_1, ..., t_n \in \mathbb{N}$.

Mientras la estacionalidad en covarianza se define como

Definición Una serie de tiempo se dice, $\{X_t\}_{t\in T}$ se dice estacionaria en covarianza si $E(X_t) = \mu$ para algún $\mu \in \mathbb{R}$ y para cualquier $t \in T$, $V(X_t) = \sigma < \infty$ y $cov(X_t, X_s) = E(X_t - \mu)E(X_s + \mu) = \gamma_k < \infty$ para cualquier $k = |t - s|$.

Donde la ultima definición se refiere a que la varianza solo depende del numero de periodos de separación.

Función de autocorrelación

La función de autocorrelacion nos da información sobre el grado de asociación lineal existente entre dos variables aleatorias del proceso separadas k periodos. Dada una observacion de una serie de tiempo $\{X_t\}_{t\in[N]}$(una base de datos) se define define la función de autocorrelación por $p : [n] \cup \{0\} \to \mathbb{R}$, con $n < N$ donde N es el numero de observaciones tal que $p(k) = \frac{cov(X_t, X_{t+k})}{\sqrt{V(X_t)V(X_{t+k})}}$.

La gráfica de F, se le conoce como correlogramas.

Ruido blanco

El ruido blanco es una serie de tiempo que puede ser algo simple pero sera muy útil he importante para el análisis del series de tiempo ya que es la base para la construcción de los modelos $AR(p)$, $AM(p)$, $ARMA(p, q)$ y $ARIMA(p, q)$, los cuales definiremos y analizaremos mas adelante. La definición de ruido blanco es la siguiente.

Definición Una serie de tiempo $\{a_t\}_{t\in T}$ se dice ruido blanco si cada a_t tiene media 0 y covarianza constante para toda $t \in T$.

En particular si la covarianza es finita, entonces el proceso de ruido blanco es estacionario. Un ejemplo de ruido blanco es $\{a_t\}$ donde cada a_t tiene distribución normal $(0, 1)$. La siguiente es una imagen de 100 simulaciones de este ruido blanco,

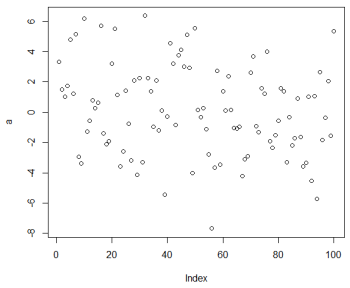

Figura 4

Proceso AR(p)

El proceso autorregresivo de orden p, expresa X_t en función lineal de su pasado hasta el retardo tp mas un ruido blanco. La definición de $AR(p)$ es la siguiente

Definición Una serie de tiempo $\{X_t\}_{t\in T}$ es un proceso $AR(p)$ si $X_t = c+\sum_{j=1}^{p}\theta_j X_{t-j}+a_t$, para constantes $c, \theta_1, .., \theta_j \in \mathbb{R}$ y a_t es ruido blanco para toda t.

Tomemos un proceso auto regresivo de orden 1, por lo que $X_t = \theta X_{t-1} + a_t$. Desarrollando la exprecion obtenemos lo siguiente,

$$\begin{aligned} X_t &= \theta X_{t-1} + a_t \\ &= \theta(\theta X_{t-2} + a_t) + a_t \end{aligned}$$

repitiendo llegamos a que $X_t = \theta^k X_{t-k} + \sum_{j=0}^{k-1}\theta^j a_{t-j}$

Procederemos a probar algunas propiedades del proceso $AR(p)$.

Teorema 1 Un proceso autorregresivo de orden p es estacionario si, y solo si, el modulo de las raíz del polinomio $1-\theta_1 L-\theta_2 L^2-...-\theta_p L^p$ esta fuera del circulo unitario (dicho polinomio se conoce como el polinomio auto regresivo).

Calculemos la esperanza de un proceso AR(p), donde $\theta_1, ..., \theta_p$ cumplen que el modulo de las raíz del polinomio auto regresivo esta fuera del circulo unitario.

Sea $\{X_t\}_{t\in T}$ un proceso AR(p) tal que $X_t = c+\sum_{j=0}^{n}\theta_j X_{t-j}+a_t$ con $c, \theta_1, .., \theta_j \in \mathbb{R}$ donde $\theta_1, ..., \theta_p$ son tales que el modulo de las raíz del polinomio auto regresivo esta fuera del circulo unitario y $\{a_t\}$ es ruido blanco.

Por el teorema 1 tenemos que la serie de tiempo es estacionaria, por lo que $E(X_t) = E(X_{t-j})$ para cualquier $t \in T$ y $j \leq t$. Usando lo anterior y recordando que $E(a_t) = 0$, tenemos que

$$\begin{aligned} E(X_t) &= E(c+\sum_{j=0}^{p}\theta_j X_{t-j}+a_t) \\ &= E(c)+\sum_{j=0}^{p}\theta_j E(X_{t-j})+E(a_t) \\ &= c+\sum_{j=0}^{p}\theta_j E(X_t)+0, \end{aligned}$$

despejando $E(X_t)$ de lo anterior, obtenemos que

$$(1-\sum_{j=0}^{p}\theta_j)E(X_t) = c$$

como el modelo es estacionario, tenemos que $1-\sum_{j=0}^{n}\theta_j \neq 0$ y esto implica que $E(X_t) = \frac{c}{1-\sum_{j=0}^{n}\theta_j}$.

Ejemplos de modelo AR(1)

Notemos que en la practica si queremos simular un modelo $AR(p)$, necesitamos dar los valores $X_0, ..., X_{p-1}$. Vemos los siguientes dos ejemplo de un modelo $AR(1)$. Tomemos

$$X_t = 0 + a_t - (0.9)X_{t-1}$$
$$Y_t = 0 + a_t + (0.9)X_{t-1}$$

donde $X_0 = 0Y_0$ y a_t tiene distribuccion normal con media 0 y varianza $\sigma^2 = 1$. Una simulación de tamaño 100 de estos procesos nos genera la siguientes imágenes.

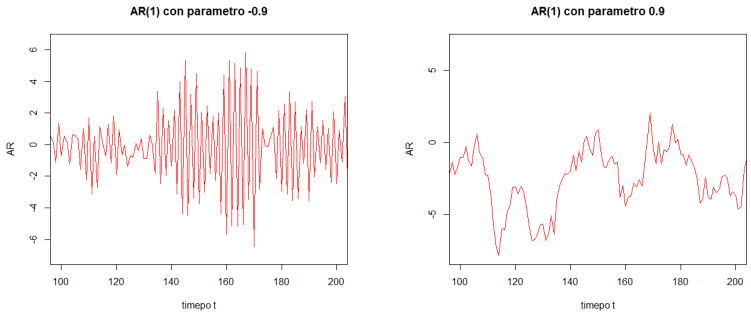

Figura 5

Sus respectivos correlogramas, están dados por

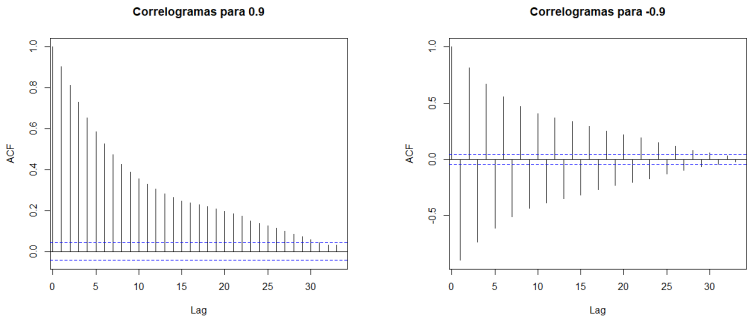

Figura 6

Procesos de Medias Móviles de orden q

El modelo medias móviles finito de orden q, es la aproximación natural al modelo lineal general.

Definición Un proceso $\{X_t\}_{t\in[n]}$ es un proceso de medias móviles de orden n si es de la forma $X_t = c + \sum_{j=1}^{n} \theta_j a_{t-j}$ donde a_t es ruido blanco y c es una constante.

Notemos que este proceso se tiene que X_t es suma de $n+1$ proceso estacionarios, por lo que el proceso es estacionario. También podemos notar que si $\{X_t\}_{t\in[n]}$ es un proceso $AR(1)$ i.e $X_t = \theta X_{t-1} + c + a_t$ con $X_0 = x_0$, entonces

$$\begin{aligned} X_t &= c + \theta X_{t-1} + a_t \\ &= c + \theta(\theta X_{t-2} + a_t + c) + a_t \\ &. \\ &. \\ &. \\ &= \theta^t X_0 + \sum_{j=0}^{t-1}(a_t + c)\theta^j \\ &= \theta^t X_0 + \sum_{j=0}^{t-1} a_t\theta^j + \sum_{j=0}^{t-1} c\theta^j \end{aligned}$$

Por lo que tomando $C = \theta^t X_0 + \sum_{j=0}^{t-1} c\theta^j$, podemos expresar el proceso $AR(1)$ como un proceso de medias móviles.

Ejemplos de procesos MA(q)

Proceso autorregresivo de media móvil de orden(p,q)

Lo antes explicado nos lleva de forma natural al proceso autorregresivo de media móvil de orden (p,q), el cual se define de la siguiente forma.

Definición Una serie de tiempo $\{X_t\}_{t\in T}$ es un proceso autorregresivo de medias móviles de orden (p,q) si es de la forma $X_t = \sum_{j=1}^{p} \theta_j X_{t-j} + a_t + \sum_{j=1}^{q} a_{t-j}\gamma_j + c$ donde $\theta_1, ..., \theta_p, \gamma_1, ..., \gamma_q, c \in \mathbb{R}$ y a_t es ruido balnco para todo t.

Una forma intuitiva de interpretar este proceso es $ARMA(p, q) = AR(p) + MA(q)$. Usanto el teorema , tenemos el siguiente teorema

Teorema 1 Un proceso autorregresivo de medias móviles ARMA(p, q) es estacionario si, y solo si, la norma de las raíces del polinomio autorregresivo esta fuera del circulo unidad.

Esto se puede ver usando el Teorema 1 y notando que estacionalidad del proceso ARMA(p,q) esta determinada por el componente autorregresivo (AR(p)) ya que el componente de medias móviles es estacionario.

Calculemos la media de un proceso ARMA(p,q).

Sea $\{X_t\}_{t\in T}$ un proceso ARMA(p,q) tal que $X_t = \sum_{j=1}^{p} \theta_j X_{t-j} + a_t + \sum_{j=1}^{q} a_{t-j}\gamma_j + c$.

Observemos que

$$\begin{aligned} E(X_t) &= E(\sum_{j=1}^{p} \theta_j X_{t-j} + a_t + \sum_{j=1}^{q} a_{t-j}\gamma_j + c) \\ &= E(\sum_{j=1}^{p} \theta_j X_{t-j}) + E(a_t) + E(\sum_{j=1}^{q} a_{t-j}\gamma_j) + E(c) \\ &= E(\sum_{j=1}^{p} \theta_j X_{t-j}) + c \end{aligned}$$

Lo que nos lleva al resultado obtenido en el proceso $AR(p)$. Por ende la esperanza del proceso es $\frac{c}{1-\sum_{j=0}^{n}\theta_j}$.

Ejemplos de procesos ARMA(p,q)

Ajuste del proceso AR(p), MA(q) y ARMA(p,q) en R

Dada una serie de tiempo, las características que debe cumplir para poder ajustar uno de estos procesos son, estacionalidad (estricta o en covarianza), media cero, que sea no anticipable i.e que el presente no venga determinado por el futuro y invertible lo que significa que el presente venga determinado por el pasado de forma convergente lo que se puede representar como $\sum_{j=1}^{infty} \theta_j < \infty$ donde los θ_j son tales que $Y_t = \sum_{j=1}^{\infty} \theta_j Y_{t-j}$. Esto ultimo es equivalente a que el correlogramas decrezca rápidamente a 0.

Considera la base de datos del aumento nivel medio global del mar,

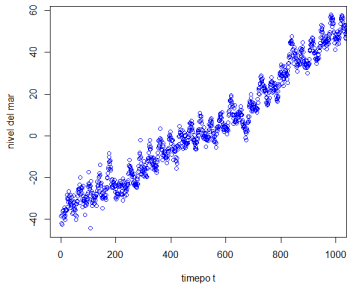

Figura 7

la cual claramente no cumplirá las condiciones de estacionalidad o de media 0. Pero al ver la serie de tiempo de sus diferencias,

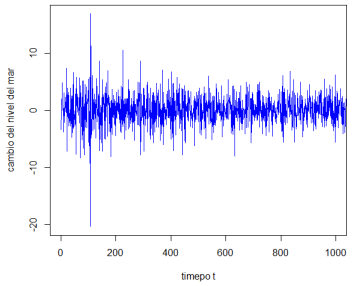

Figura 8

podemos notar que esta serie de tiempo muy posiblemente si cumpla las condiciones necesarias para ajustar los procesos $AR(p),\ MA(q)$ y $ARMA(p,q)$. Por lo que procederemos a ver que esta serie de tiempo cumple con las condiciones necesarias para ajustar los procesos $AR(p),\ MA(q)$ y $ARMA(p,q)$ y posteriormente los ajustaremos.

Guardando nuestros datos en "dif" y usando la función de R "k = acf(dif)", esta hace una prueba de hipótesis de si hay o no correlación. Al usar el comando "k$type" obtenemos "correlation", por lo que hay evidencia a favor de que hay correlación. Esta misma función nos brinda el siguiente correlograma

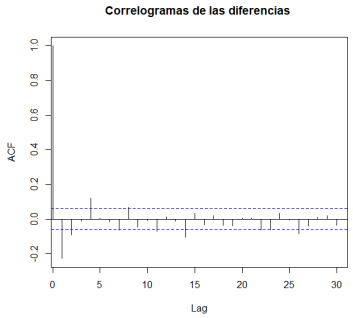

Figura 9

esto nos indica que los datos se comportan como un proceso ARMA(p,q). Por lo que ahora buscaremos ajustar los parámetros $\theta_1, ..., \theta_p$ y $\psi_1, ..., \psi_q$ del proceso. Usando la función de R "c = auto.arima(dif)" y usando el comando "c$coef" obtenemos

Figura 10

Lo que nos indica que el el proceso ARMA(p,q) que mejor ajusta a los datos es ARMA(4,1) con parámetros $\theta_1 = 0.11952744, \theta_2 = -0.04775842, \theta_3 = 0.01596729, \theta_4 0.12763155$ y $\psi_1 = -0.39039621$. Comparando el ajuste que hicimos con los datos originales obtenemos la siguiente imagen

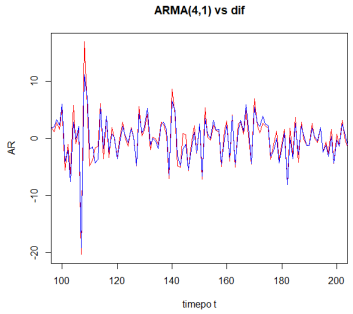

Figura 11

Usando la función "tsdiag(c)" obtenemos que,

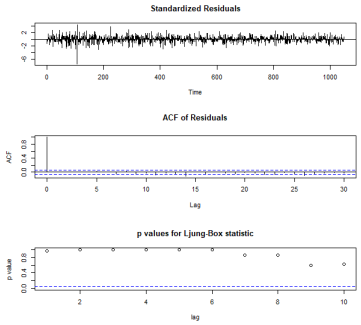

Figura 12

lo que nos indica que el proceso se ajusta bien a la serie de tiempo dada.

Algo importante que podemos notar de este ajuste, es que no ajusta bien los valores máximos de los datos véase la siguiente figura. Lo que nos lleva a pensar en otro métodos para ajustar la distribución de los máximos de los datos.

Extremos en series de tiempo no independientes

Como ya hicimos notar anteriormente, los ajustes de procesos AR(p). AM(q) y ARMA(p,q) no suelen modelar o aproximas de una buena manera a los maximos de los datos de las series de tiempo. Por lo que optamos por otras herramientas para analizar el comportamiento de dichos máximos.

Dicha herramienta sera la teoría de extremos aplicada a series de tiempo estacionarias con dependencia (como hemos estado haciendo). Aun que ciertamente podríamos ajustar una distribución de máximo a los datos y argumentar que se ajusta bien. Tendríamos el problema de que las series de tiempo estacionarias es dependiente, por lo que no podríamos aplicar la teoría que conocemos y el ajuste a un que pueda parecer bueno no tendría teoría que lo respalde y que somos hombres o payasos ?.

Por lo que en esta sección, presentaremos la teoría que respalda dicho ajuste. Comenzaremos con definir las propiedades D, $D(u_n)$ y $D'(u_n)$.

Definición Sea $\{X_t\}_{t\in T}$ una succión de variables aleatorias. Se dice que la condición D se cumple si para cualquier enteros positivos $i_1, ..., i_p$ y $j_1, ..., j_q$ tales que $|j_1 - i_p| \geq k$ y cualquier $u \in \mathbb{R}$, se cumple que

$$|F_{i_1,...,i_p,j_1,...,j_q}(u) - F_{i_1,...,i_p}(u)F_{j_1,...,j_q}(u)| \leq g(k)$$

donde g es tal que $g(k) \to 0$ si $k \to \infty$ y $F_{s_1,...,s_p}(u) = P(X_{s_1} \leq u, ..., X_{s_r} \leq u)$ para cualquier $s_1, ..., s_r \in T$.

Definición Sea $\{X_t\}_{t\in T}$ una succión de variables aleatorias. Se dice que la condición $D(u_n)$ se cumple si para cualquier enteros positivos $i_1 < ... < i_p < j_1 < ... < j_q \leq n$ tales que $|j_1 - i_p| \geq k$ y cualquier $u \in \mathbb{R}$, se cumple que

$$|F_{i_1,...,i_p,j_1,...,j_q}(u_n) - F_{i_1,...,i_p}(u_n)F_{j_1,...,j_q}(u_n)| \leq \alpha_{n,k_n}$$

donde $\alpha_{n,k_n} \to 0$ si $n \to \infty$ y $\{k_n\}$ es tal que $k_n = O(n)$.

Definición Sea $\{X_t\}_{t\in T}$ una succión de variables aleatorias y $\{u_n\}_{n\geq 0}$. Se dice que $\{X_t\}_{t\in T}$ cumple la condición $D'(u_n)$ si

$$\lim_{k\to\infty}(\limsup_{n\to\infty} n\sum_{j=2}^{[\frac{n}{k}]} P(X_1 > u_n, X_j > u_n)) = 0$$

Teorema : Sea $\{X_t\}_{t\geq 0}$ una succión de variables aleatorias. Si $\{u_n\}_{n\geq 0}$ es una succión tal que $\{X_t\}_{t\geq 0}$ cumple las condiciones $D(u_n)$ y $D'(u_n)$, entonces

$$\lim_{n\to\infty} P(N_n \leq u_n) = e^{-\tau}$$

para $\tau < \infty$.

En general esto nos dice que las condiciones D y D', son condiciones suficientes para poder aplicar los resultados de teoría de extremos, como dominios de atracción.

Aplicación de extremos en series de tiempo

Tomemos los datos $\{X_t\}_{t\in T}$ tal que X_t como el nivel medio global del mar a tiempo t y dif como antes. Buscaremos calcular el tiempo esperado de retorno de N en $\{X_t\}_{t\ inT}$ y N' en dif. Esto usando lo probado anteriormente.

Algunas evidencia a favor de que se cumple D y D' son los correlogramas de los datos que nos indica que la dependencia de los datos decrece rápidamente a 0.

Por lo antes argumentado por demos aplicar la teoría de extremos a series de tiempo.

Calculemos el tiempo esperado de retorno del evento un dato pase el umbral 50.

Para esto buscamos calcular $\frac{1}{\overline{F}(50)}$. Buscaremos primero aproximar $\overline{F}(50)$ usando la aproximación de la generalizada de pareto i.e $\overline{F}(50) \approx \overline{F}(u)\overline{P}_{a(u),\xi}$ (esto es posible por que sabemos que los datos toman una distribución de extremos) para algún u.

Por lo que primero buscaremos un u tal que la cola de los datos se aproxime a la pareto.
Tomando la región '(-7.568,-7.566)' en la funcion 'gpd.fitrange' obtenemos

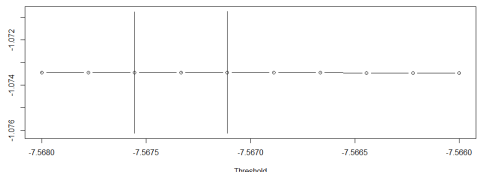

Figura 13

Lo que nos indica que '-7.568' es una buena u para ajustar a una DGP. Usando 'gpd.fitgpd.fit(max,-7.568)' obtenemos los valores del ajuste son $(a(u), gi) = (70.297140, -1.073435)$. Evaluando obtenemos que $\overline{P}_{70.297140,-1.073435}(50 + 7.568) = 0.1397422$ y calculando el valor $\overline{F}(-7.568)$ con la cola empírica obtenemos que $\hat{F}(-7.568) = 0.75$. Por lo tanto $\overline{F}(50)$ se aproxima por 0.1060113. Y por lo tanto el tiempo medio de retorno de dicho evento es $\frac{1}{0.1060113} = 9.432957$.

Esto es un ejemplo, pero en realidad podemos aplicar toda la teoría de extremos que hemos visto en el curso.