Tarea 4

Rodrigo Carvajal Pizarro (rcarvaja@astro.puc.cl)
https://github.com/racarvajal/Tarea4_Astro_Stats.git

Resolución

Problema 1

a) El método de *Maximum Likelihood Estimation* (MLE) para extraer los mejores valores de los parámetros en un modelo involucra optimizar la Función de Verosimilitud en estos parámetros.

Como los coeficientes a determinar son bastantes, se trabajará de forma matricial. De este modo, el modelo tiene la siguiente forma:

$$\vec{Y} = \mathbb{M}\vec{\theta} + \vec{L} + \vec{\epsilon}$$

En que $\vec{Y}$ es el vector de los datos entregados (la variable dependiente, el flujo). $\mathbb{M}$ corresponde a la Matriz de Diseño, que contiene a la variable independiente del sistema (En este caso, incluye variaciones de los tiempos medidos, para dar cuenta de que el modelo está configurado por un polinomio y un tránsito). Esta matriz será:

$$\mathbb{M} = \begin{pmatrix} 0 & 1 & t_1 & t_1^2 & \cdots & t_1^p \\ 0 & 1 & t_2 & t_2^2 & \cdots & t_2^p \\ \vdots & \vdots & \vdots & \vdots & & \vdots \\ -1 & 1 & t_l & t_l^2 & \cdots & t_l^p \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 1 & t_k & t_k^2 & \cdots & f \end{pmatrix}$$

Por otro lado, el vector $\vec{\theta}$ es el que contiene los valores (a determinar) de los parámetros del modelo. Incluye el parámetro δ y los coeficientes del polinomio.

$$\vec{\theta} = \begin{pmatrix} \delta \\ c_0 \\ c_1 \\ \vdots \\ c_p \end{pmatrix}$$

$\vec{L}$ corresponde a un vector de k números 1. Este vector es el término constante en el modelo de tránsito. Por último, $\vec{\epsilon}$es el vector que induce la aleatoriedad al sistema. Es un vector de k números 1 multiplicados por el valor ϵ que, siendo constante, se extrae de una distribución Normal con media 0 y varianza $\sigma^2 = (30ppm)^2$.

Para poder utilizar el método MLE, debe calcularse la verosimilitud del modelo. Como el error es gaussiano, es más sencillo trabajar con la verosimilitud. Esto se expresa en que la diferencia entre los datos obtenidos y los valores expresados por el modelo corresponden al error del modelo (ϵ).

$$\vec{Y_k} - \vec{Y}_{model} = \vec{Y_k} - \mathbb{M}\vec{\theta} - \vec{L} = \vec{\epsilon} \sim \mathcal{N}(0, \sigma^2)$$

Y esta diferencia puede llevarse a una verosimilitud (matricial) con la siguiente forma:

$$\mathcal{L} = (\vec{Y_k} - \mathbb{M}\vec{\theta} - \vec{L})^T \mathbb{V}^{-1} (\vec{Y_k} - \mathbb{M}\vec{\theta} - \vec{L})$$

En que $\mathbb{V}$ es la matriz que tiene el valor de la varianza en su diagonal. Esta verosimilitud debe ser maximizada para los valores de $\vec{\theta}$

$$\frac{\partial \mathcal{L}}{\partial \theta} = 0$$

$$\frac{\partial \mathcal{L}}{\partial \theta} = 2\mathbb{M}^T\mathbb{V}^{-1}(\vec{Y}_k - \mathbb{M}\vec{\theta} - \vec{L}) = 0$$

Lo que lleva a

$$\vec{\theta} = (\mathbb{M}^T\mathbb{V}^{-1}\mathbb{M})^{-1}\mathbb{M}^T\mathbb{V}^{-1}(\vec{Y} - \vec{L})$$

Estos parámetros deben usarse en el modelo y se tendrá una estimación de su comportamiento.

Si se deja el análisis matricial y se toma de un punto de vista escalar, el modelo no es lineal en los parámetros. Esto ocurre porque el modelo de tránsito es válido en un intervalo del tiempo solamente. Con esta consideración, habrá dos discontinuidades en los tiempos $t = 0{,}4$ y $t = 0{,}7$. Esta dificultad puede ser superada si se considera que cada intervalo como una entidad independiente. Dentro de cada uno de ellos, el modelo será completamente lineal.

b) El código escrito trabaja de forma matricial con el modelo propuesto en la sección anterior. Además, la matriz de diseño es llenada de forma separada para el modelo de tránsito y los coeficientes del polinomio (de grado $p = 5$). Los coeficientes (y el modelo) que entrega el código pueden verse representados en la Figura 1.

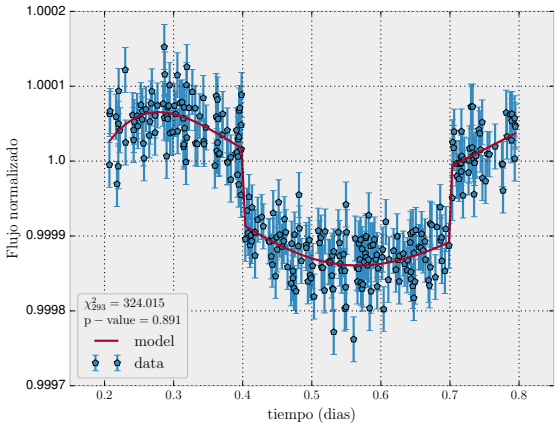

Figura 1: Datos entregados y modelo ajustado para $p = 5$

Puede verse que el modelo muestra una evolución polinomial en el tiempo y, además, un valle (no continuo en sus bordes) que representa el tránsito planetario.

c) Para realizar un test de chi-cuadrado, se necesita modificar (transformar) la variable que se ha estado utilizando, $\vec{\theta}$, para llevarla a una que distribuya como chi-cuadrado.

Una transformación posible es

$$X^2 = (\vec{Y}_k - \mathbb{M}\vec{\theta} - \vec{L})^T\mathbb{V}^{-1}(\vec{Y}_k - \mathbb{M}\vec{\theta} - \vec{L}) \sim \chi^2_{N-\text{par}} = \chi^2_{300-7}$$

La distribución será como una chi-cuadrado con 293 grados de libertad, que corresponden a los 300 datos con los que se cuenta menos los siete parámetros que se busca determinar (uno para el tránsito planetario y seis coeficientes para el polinomio de grado 5).

Una vez obtenidos los parámetros, calcular este valor es sencillo y, en este caso es:

$$X^2 = 324{,}015$$

Corresponde, ahora, obtener el p-valor para este resultado. Debe recordarse que el p-valor es el probabilidad de que, bajo la hipótesis nula, se obtengan valores de X^2 como el calculado anteriormente o más extremos. Escrito en una fórmula, es

$$\text{p} - \text{value} = \mathbb{P}(\chi^2_{300-7} \geq X^2)$$

$$\text{p} - \text{value} = 1 - \mathbb{P}(\chi^2_{300-7} < X^2)$$

Utilizando los valores de la Distribución Cumulativa de Probabilidad (CDF) para una distribución chi-cuadrado de 293 grados de libertad, se llega a:

$$\text{p} - \text{value} = 1 - 0{,}109$$

$$\text{p} - \text{value} = 0{,}891$$

Este valor (que, también está presente en la Figura 1) se interpreta, por lo tanto, como que, bajo la hipótesis nula, se rechaza el modelo en un 11 %. Llevando este resultado a una interpretación muy ligera, se puede decir que el modelo obtenido (con los parámetros mencionados anteriormente) es bueno en general.

Problema 2

a) Ahora deben utilizarse los parámetros encontrados en la pregunta anterior pero considerando un modelo polinomial solamente. De este nuevo modelo, se simulan 10^3 conjuntos de datos (con un error aleatorio y gaussiano). En cada una de las realizaciones de este experimento, se debe calcular el p-valor entre los datos simulados y el modelo polinomial. El resultado de estas simulaciones puede verse en la Figura 2

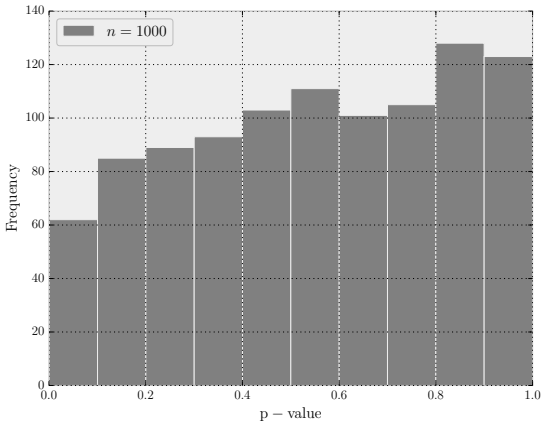

Figura 2: Distribución de p-valores para modelo polinomial

A la luz de la figura, se ve que los p-values tienen una distribución que crece con el p-valor. En este caso, significa que, existe una mayor probabilidad de encontrar p-valores altos (de no rechazar la hipótesis nula) que valores bajos (rechazar hipótesis nula).

Según se verá en las secciones siguientes de esta pregunta, la distribución aquí presentada no es la esperada. Se espera una distribución uniforme para los p-valores.

b) En la Tarea 2, se demostró que si se tiene una variable aleatoria que es función de cualquier otra variable aleatoria que tenga una función cumulativa de distribución monótonamente creciente, esa función tendrá una distribución uniforme. Esto es lo que se conoce como *Probability Integral Transform*.

Utilizando la demostración de la Tarea ya mencionada, se puede ver que, si Y corresponde a una función de la variable aleatoria X:

$$F_Y(y) = \mathbb{P}(Y \leq y) = \mathbb{P}(F(X) \leq y)$$
$$F_Y(y) = \mathbb{P}(X \leq F^{-1}(y)) = F_X(F^{-1}(y)) = y$$

Por lo tanto, este es el resultado esperado en la parte anterior de esta pregunta; la distribución de p-valores deberá ser uniforme.

c) Ahora, lo que se quiere comprobar es cómo cambia la distribución de p-valores para diferentes grados del polinomio del modelo de los datos. Si se repite el experimento de la parte (a) para tres valores de p (2, 5, 15), se obtiene lo representado en la Figura 3

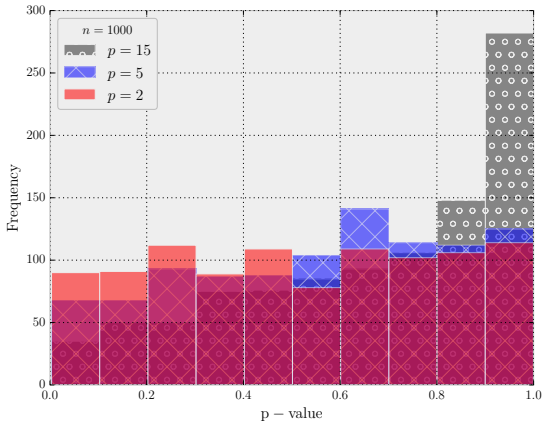

Figura 3: Distribución de p-valores para modelo polinomial con diferentes grados

En esta sección, sí se obtienen resultados consistentes, en mayor medida, con los discutidos en la sección (c).

Para un grado muy alto del polinomio, la distribución de p-valores deja de ser uniforme. Esto se explica porque hay menos probabilidades que la hipótesis nula ("el modelo con el grado indicado es perfecto") sea verdadera. Y ese es uno de los requisitos para poder aplicar la *Probability Integral Transform*.

Para grados más bajos, la distribución tiende hacia la Uniforme. Esto ocurre pues las condiciones para aplicar la *Probability Integral Transform* empiezan a cumplirse.

d) Ahora, se realiza la misma experiencia que en la sección (a) pero incorporando el modelo de tránsito. El resultado de esta experiencia puede observarse en la Figura 4.

En la Figura, puede observarse la misma distribución que en la sección (a). Ésta no es la distribición esperada según se hace notar en la sección (b); una distribución uniforme.

Ante este resultado (similitud de distribución para ambos modelos), puede decirse que, a través del estudio de los p-valores, no se puede establecer una comparación entre una u otra manera de modelar los datos. El uso de los p-valores solo es útil para descartar (o no) la validez de un modelo por sí solo.

e) (1) El p-valor no corresponde a la probabilidad de que la hipótesis nula sea cierta. Se refiere solo a la probabilidad de ocurrencia de los resultados obtenidos bajo la hipótesis nula. Con los p-valores, se puede rechazar una hipótesis, pero no se puede confirmar su validez

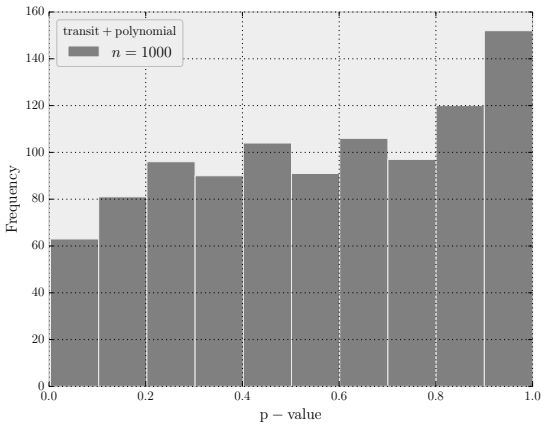

Figura 4: Distribución de p-valores para modelo completo: tránsito + polinomio

(2) No es correcto comparar p-values para seleccionar modelos. Como se indica en las secciones anteriores, para modelos diferentes, los p-valores se distribuirán de manera similar. Cualquier resultado de un test chi-cuadrado solo permite establecer la validez de un modelo frente a los datos, pero no frente a otros modelos.

Problema 3

a) Para calcular el Akaike Information Criterion (AIC), se utiliza la fórmula:

$$\text{AIC} = -2\log\left[\mathcal{L}^o\right] + 2k + \frac{2k(k+1)}{N-k+1}$$

En que $\mathcal{L}^o$ es el valor de la verosimilitud evaluada en los parámetros que la maximizan, k es el número de parámetros y N es el número de datos.

Por otro lado, el Bayesian Information Criterion (BIC), se extrae de:

$$-2\log\left[\mathcal{L}^o\right] + k\log\left(N\right)$$

En que $\log\left[\mathcal{L}^o\right]$ corresponde, ahora, a la probabilidad posterior del modelo evaluada en los parámetros que la maximizan.

Cabe destacar que la probabilidad posterior requiere conocer la probabilidad a priori y su factor de normalización. Estos factores pueden ser difíciles de obtener. En este caso, se asumirá que $\log\left[\mathcal{L}^o\right]$, en el BIC, será la verosimilitud y los resultados que se obtengan de ella solo se podrán analizar cualitativamente (a través de su tendencia) pero no cuantitativamente.

Para diferentes grados del polinomio (número de parámetros), los valores de AIC y BIC pueden verse en la Figura 5.

Ya que los valores de AIC y BIC son diferentes en magnitud y como el BIC no está calculado de forma completa, se muestran ambos indicadores luego de ser normalizados. Así, ambos criterios variarán entre 0 y 1 y será menos complejo compararlos.

Puede verse en los gráficos que ambos criterios no se comportan como se espera: no muestran un valor alto para pequeños p y para altos p y un valle en cantidades intermedias de parámetros. Se intentaron diferentes estrategias para obtener estos valores, pero todas dieron la misma tendencia.

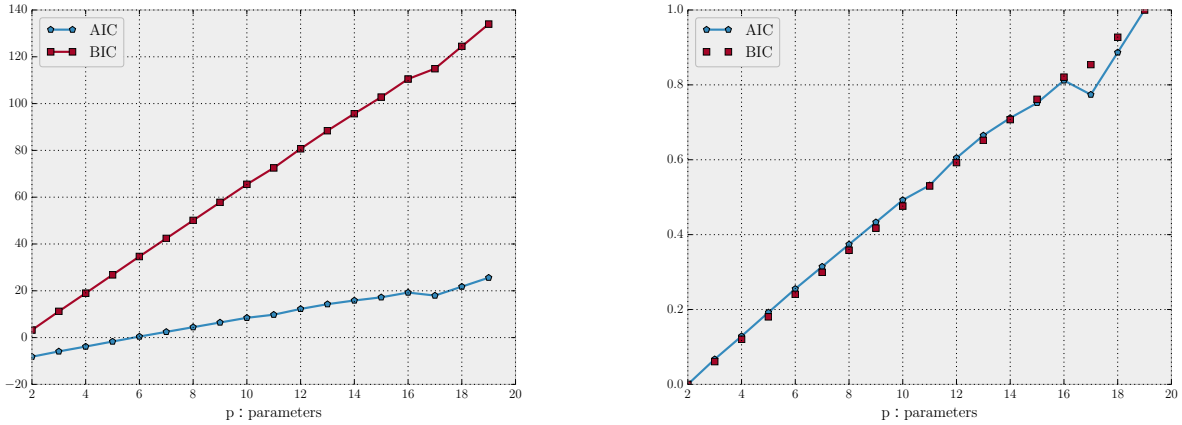

Figura 5: Distribución de p-valores para modelo completo: tránsito + polinomio. Izquierda: Valores Originales. Derecha: Valores Normalizados

b) El método de Validación Cruzada (*Cross Validation*) también es útil para escoger un número óptimo de parámetros a utilizar en el modelo estudiado. Se basa en el entrenamiento del modelo utilizando solo una fracción de los datos. Luego, este modelo se valida con los datos faltantes.

En particular, el método *K-fold Cross Validation* divide la muestra en K subconjuntos con igual cantidad de puntos adyacentes (en la variable independiente). Se toma cada uno de estos subconjuntos y se utilizan para validar el entrenamiento alcanzado al usar los otros $K-1$ subconjuntos.

Este proceso se realiza K veces y se promedian los resultados para cada variación del modelo (en número de parámetros).

La medida de qué tan útil es un modelo en el proceso de validación se hace a través de el Error Cuadrático Medio (MSE, por sus siglas en inglés). Cuando se conoce la distribución subyacente, corresponde a la esperanza de las diferencias cuadradas de los parámetros y sus estimadores. Pero como la distribución original de los datos no es conocida, se utiliza la siguiente expresión:

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(\hat{Y} - Y)^2$$

En que n es el número de puntos a utilizar en la validación (el tamaño de cada uno de los K subconjuntos), $\hat{Y}$ son los datos que se extraen del modelo de entrenamiento que serán comparados con Y, los datos originales. El resultado de este procedimiento se puede ver en la Figura 6.

En el gráfico, puede verse que no hay un valor mínimo que se ajuste con lo esperado (ocurre lo mismo anteriormente). Tal como se hizo en las secciones anteriores, varias estrategias se intentaron para obtener un resultado consistente con la teoría, pero no hubo éxito en ello.

De todos modos, se puede mencionar que se espera que el mejor número de parámetros que entregue el uso del método *K-fold Cross Validation* se acerque a aquel entregado por el cálculo del AIC y BIC en las mismas circunstancias.

c) Es importante mencionar las diferencias existentes entre el uso del AIC, el BIC y *K-fold Cross Validation* para seleccionar modelos. A pesar de entregar resultados similares, sus diferencias radican en qué pregunta buscan responder (a través de la que se llega a un valor óptimo para los parámetros).

El uso de AIC busca seleccionar el modelo que, más adecuadamente, describe los datos. En cambio, BIC intenta encontrar el modelo real para los datos estudiados. Y *K-fold Cross Validation* busca el modelo que menos diferencia genera entre éste y los puntos a considerar, aunque también puede predecir el comportamiento de datos extra que se agreguen al estudio.

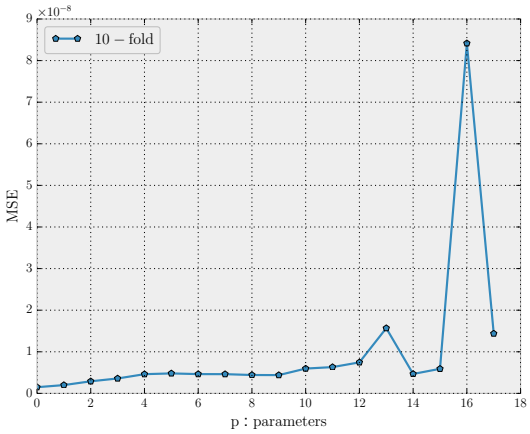

Figura 6: MSE en función del número de parámetros para *K-fold Cross Validation*

También, puede mencionarse que AIC considera de forma muy certera la relación existente entre sesgo y varianza de un modelo lineal. *K-fold Cross Validation* compara los mismos efectos pero en modelos generales (no, necesariamente, lineales).