

Informe de
**ASESORAMIENTO
y TRANSFERENCIA**

087-23

NO-2023-140893326-APN-DNI#INIDEP
24/11/2023

Manual de uso del script lpr.R v1.0 para el estudio de relaciones Longitud-Peso en RStudio

Adrián Jové y Aníbal Aubone

Jové, A., Aubone, A.. 2023. Manual de uso del script lpr.R v1.0 para el estudio de relaciones Longitud-Peso en RStudio. Inf ASES INIDEP N° XXX/XX, xx pp.





Manual de uso del script lpr.R v1.0 para el estudio de relaciones Longitud-Peso en RStudio

Adrián Jové, Aníbal Aubone

Instituto Nacional de Investigación de Desarrollo Pesquero

Resumen

En este trabajo se presenta un Script en R para estimar los coeficientes de la función que relaciona el peso medio y la longitud, función que se asume potencial. La variable peso se asume con distribución de probabilidades log-normal para cada longitud. El modelo se linealiza y se estiman los coeficientes por el Método de Máxima Verosimilitud (Regresión Lineal Simple). Se valida estadísticamente la hipótesis de distribución de probabilidades log-normal de los errores. Se evalúa la isometría mediante una prueba estadística. Se prueba estadísticamente la igualdad de coeficientes entre dos regresiones.

Palabras Clave

Regresión lineal simple, regresión potencial, máxima verosimilitud, comparación estadística de coeficientes.

Introducción

La longitud y el peso de organismos marinos son datos esenciales para muchas investigaciones. La importancia de contar con una estimación del peso medio por clase de longitud o por edad es crucial en la evaluación de recursos pesqueros. La biomasa poblacional se estima a partir de pesos medios de los individuos por edad o por clase de longitud. Sin embargo, el peso individual es muchas veces difícil de ser registrado, por carecer de instrumental adecuado, o situaciones difíciles como ocurren en una campaña de investigación a bordo de un buque (se requiere de una balanza compensada). Es por ello, que poder estimar una función que describa la relación del peso medio para cada longitud, facilita la estimación del peso para diferentes individuos (longitudes). La función que relaciona el peso medio para una longitud se estima a nivel poblacional y no individualmente.

La relación entre la longitud y el peso medio es también utilizada para definir un índice de condición corporal, que permite evaluar un individuo y su estado de condición respecto de una condición media poblacional.

Los coeficientes de la relación Longitud-Peso pueden variar entre poblaciones diferentes, o por sexos o por diferentes periodos de tiempo. Es por ello que también resulta importante poder contar con herramientas estadísticas para realizar una comparación.

El objetivo de este trabajo es presentar un Script realizado en R para realizar un análisis pormenorizado de la relación Longitud-Peso. A su vez se presenta una síntesis metodológica del análisis.

Materiales y métodos

El modelo que relaciona la longitud y el peso, suele ser potencial en términos medios.

$$E(P/L) = aL^b; \text{ con } a > 0, b > 0.$$



donde P es el peso y L la longitud estándar. $E(P/L)$ es la esperanza matemática de la variable aleatoria peso, condicionada a la longitud L .

Al considerar errores aleatorios se obtiene:

$$P = aL^b e_L, \text{ con } \ln(e_L) \sim N(0; \sigma_L^2)$$

Teniendo en cuenta que el índice de masa corporal se define: $IMC = \frac{P}{L^3} = a e_L$ y $E(IMC) = a$, la consideración de un intervalo de confianza para a , sirve para evaluar el estado de condición de un individuo.

De acuerdo a Huxley & Teisser (1936), si el exponente b es igual a uno el *crecimiento es isométrico*, si es mayor que uno se denomina *alometría positiva* y si es menor que uno, *alometría negativa*. La alometría es una característica específica de la especie y las variaciones intraespecíficas se pueden relacionar con efectos determinados genéticamente, así como por el sexo, el estadio de madurez, o el período de desove.

Los coeficientes de la relación Longitud-Peso pueden variar entre sexos. En el caso de peces, el exponente se suele aproximar al valor de 3. Es por ello que resulta importante realizar un test de hipótesis para validar si se puede considerar el exponente $b = 3$.

El tamaño de la muestra debe ser adecuado por lo menos para definir adecuadamente el valor medio del peso y también el de la longitud. Para ello puede utilizarse el programa Nmin1 v111023 (Aubone, 2023) para determinar un tamaño mínimo de muestreo común para la longitud y el peso.

Para trabajar con el modelo potencial $P = aL^b e_L$ y realizar la estimación de los coeficientes a y b , se linealiza el mismo aplicando el logaritmo natural, quedando un modelo de Regresión Lineal Simple (Draper y Smith, 1980), con posibles errores heterocedásticos (diferencias en las varianzas de los errores). Entonces se obtiene:

$$P = aL^b e_L \\ \ln(P) = \ln(a) + b \ln(L) + \ln(e_L)$$

Donde $a > 0, b > 0, \ln(e_L) \sim N(0, \sigma_L^2)$.

La tarea ahora es estimar el valor de $\ln(a)$ y b . La estimación de los coeficientes se realiza por el Método de Máxima Verosimilitud.

Se debe previamente, verificar los supuestos del problema: homocedasticidad (igualdad de las varianzas de los errores) y la independencia entre datos. La normalidad de los errores (una vez estimados los coeficientes) se puede validar con la prueba estadística de Kolmogorov-Smirnov (Daniel, 1990).

Homocedasticidad

La homocedasticidad de los errores, puede validarse mediante la prueba de comparación de varianzas de Bartlett (Goicoechea, 2018). Para ello tener en cuenta que la varianza de error para cada longitud, es la varianza del $\ln(P_L)$. Para ello, si no hay repeticiones de datos para la misma longitud (y suficientes), había que agrupar longitudes similares para obtener las réplicas necesarias para la estimación de una varianza de $\ln(P_L)$. Si las varianzas de los errores no fueran estadísticamente iguales, se recomienda ponderar por las dispersiones: $\frac{\ln(P)}{\sigma_L} = \ln(a) \frac{1}{\sigma_L} + b \frac{\ln(L)}{\sigma_L} + \varepsilon_L$, donde ahora para toda longitud será $var(\varepsilon_L) = 1$. Y $\varepsilon_L \sim N(0; 1)$, un problema de Regresión Lineal Múltiple homocedástico. El Script realiza esto automáticamente.



Prueba estadística de distribución normal de los errores (Kolmogorov-Smirnov)

Una de las posibles pruebas de normalidad es la prueba de Kolmogorov-Smirnov (K-S) (Daniel, 1990). Esto se puede hacer mediante una rutina en Rstudio (lpr.R) junto con los datos obtenidos y plasmados en una planilla de cálculo MS EXCEL. Inicialmente, se ingresan los datos originales de L y P en dos columnas, junto con la columna de varianzas (si el problema es homocedástico, se ingresa una columna de varianzas igual a 1). La rutina lpr.R posee en la línea 96 una función llamada “ks.test” que se encargará de realizar la prueba de normalidad arrojando el p -valor. Un criterio para aceptar la normalidad es considerar que el p -valor obtenido es mayor a 0,05 (nivel de significación elegido).

Resumen de la regresión

Los resultados de la regresión se presentan en una tabla. También se presenta un gráfico de dispersión de los puntos y la recta de regresión estimada.

Para obtener un intervalo de confianza del 95% para $\widehat{\ln(a)}$:

$$[\widehat{\ln(a)} - t(0,05; n - 2)s_{\widehat{\ln(a)}}; \widehat{\ln(a)} + t(0,05; n - 2)s_{\widehat{\ln(a)}}]$$

Y para obtener un intervalo de confianza del 95% para a :

Ahora, se desea buscar un estimador de a y su varianza junto con su coeficiente de variación, dada la estimación de a .

Para estimar a se procede calcular: $\hat{a} = e^{\widehat{\ln(a)} + \frac{s_{\widehat{\ln(a)}}^2}{2}}$, donde $s_{\widehat{\ln(a)}}^2$ es la varianza de $\widehat{\ln(a)}$ obtenida de la matriz de covarianza.

Para estimar la varianza del estimador de a :

$$\text{var}(\hat{a}) = e^{s_{\widehat{\ln(a)}}^2 + 2\widehat{\ln(a)}} \left(e^{s_{\widehat{\ln(a)}}^2} - 1 \right) \text{ y } s_{\hat{a}} = \sqrt{\text{var}(\hat{a})}$$

Para obtener un intervalo de confianza aproximado para a se calculan los antilogaritmos de los límites del intervalo de confianza para $\ln(a)$

Y para un intervalo del 95% para \hat{b} , tenemos

$$[\hat{b} - t(0,05; n - 2)s_{\hat{b}}; \hat{b} + t(0,05; n - 2)s_{\hat{b}}]$$

donde $\widehat{\ln(a)}$ es la estimación de $\ln(a)$, \hat{b} es la estimación de b ,

$s_{\widehat{\ln(a)}}$, $s_{\hat{b}}$ son las dispersiones o desvíos estándar de $\widehat{\ln(a)}$ y \hat{b} y $t(0,05; n - 2)$ es el valor observado de la distribución t - student con probabilidad 0,05 y $n - 2$ grados de libertad.

Para obtener las dispersiones $s_{\widehat{\ln(a)}}$ y $s_{\hat{b}}$, se utilizan los resultados directamente del script de Rstudio. También se calcula

$$s^2 = \frac{1}{n - 2} \sum_{t=1}^n \hat{e}_t^2$$

donde s^2 es el estimador insesgado de la varianza de error σ^2 . \hat{e}_t son los errores estimados.

Para estimar los coeficientes de variación:



$$cv \hat{b} = \frac{\sqrt{\text{var}(\hat{b})}}{\hat{b}}; cv \hat{a} = \frac{\sqrt{\text{var}(\hat{a})}}{\hat{a}}$$

Así, ya tienen estimados los coeficientes de la función potencial buscada junto con los coeficientes de variación.

Prueba estadística de $b=3$

Para verificar la hipótesis nula $H0: b = 3$, vs $H1: b \neq 3$:

$$t_{obs} = \frac{|\hat{b} - 3|}{s_{\hat{b}}}$$

A partir de este valor t_{obs} , se obtiene el valor $p = Prob(t > t_{obs})$, para una prueba t-Student para dos colas. Si $p < 0,025$, entonces se rechaza la hipótesis nula (con un nivel de significación de 0,05), sino se acepta la misma. Si se acepta la hipótesis nula, se debe volver a hacer la regresión sobre $\ln(a)$ y fijando $b = 3$.

En este último caso, se plantea la regresión:

$$\ln(P) = \ln(a) + 3 \ln(L) + \ln(e_L).$$

$$\text{Se obtiene: } \text{var}(\widehat{\ln(a)}) = \frac{s^2}{n} = \frac{\frac{1}{n-1} \sum_{t=1}^n \hat{e}_t^2}{n} = \frac{\sum_{t=1}^n \hat{e}_t^2}{n(n-1)}.$$

Para obtener un intervalo de confianza del 95% para $\widehat{\ln(a)}$:

$$[\widehat{\ln(a)} - t(0,05; n - 1)s_{\widehat{\ln(a)}}; \widehat{\ln(a)} + t(0,05; n - 1)s_{\widehat{\ln(a)}}].$$

Prueba de comparación de coeficientes

Dos ejemplos de pruebas estadísticas para verificar la igualdad de los coeficientes entre dos regresiones, pueden ser la prueba Fisher (tradicional prueba de comparación de coeficientes) (Draper y Smith, 1980, pág. 105) o Chi-cuadrado (cociente de verosimilitudes) (Moreno y Miravalls Sierra, 2016). La prueba estadística se basa en hipótesis nula es $H0: \ln(a_1) = \ln(a_2)$ y $b_1 = b_2$ y la hipótesis alternativa $H1: \ln(a_1) \neq \ln(a_2)$ o $b_1 \neq b_2$, para los coeficientes de dos regresiones lineales simples. En caso de que el valor de la probabilidad p obtenido sea mayor a 0,05 se acepta la hipótesis de que los coeficientes son similares (con el nivel de significación de 0,05) y así se considera que las muestras provienen de la misma población. En caso contrario, se rechaza a este nivel de significación. Todo este procedimiento está realizado en el script lpr.R. Puede cambiarse el nivel de significación.

Procedimiento para el usuario del Script lpr.R para R Studio

1. Instalar R desde la página <https://iimyc.gob.ar/cursor/> y también RStudio. Este trabajo se realizó en la versión 4.3.1 de R.
2. Una vez copiado el script lpr.R (es decir la rutina del programa), se debe importar primero la primer planilla de Excel llamada "planilla1" con las columnas llamadas "L", "P" y "var" correspondiente a los datos de las longitudes, pesos y varianzas de la primera muestra y otra



- planilla de la segunda muestra con las columnas llamadas “L”, “P” y “var” si la hubiese. Estos archivos Excel, se deben llamar “planilla1” y “planilla2”, pues el script se realizó de manera tal que en la rutina se llama a la planilla y se realizan los pasos bajo este nombre.
- Una forma de importar las planillas, es haciendo clic en la pestaña “Import Dataset”, luego hacer clic en “From Excel”, seguido de “Browse” y allí localizar el archivo. Estas ubicaciones colocarlas en las líneas 29 y 30 del script. Así, el programa podrá trabajar con las planillas.
 - Para ejecutar la rutina, en cada renglón se puede presionar “Control+Enter” o sino hacer clic en “Run”. El programa corre en este caso renglón por renglón. Otra forma de llevar a cabo el script es arrastrar parte de la rutina y presionar “Control+Enter”. Se recomienda correr la rutina por renglón para verificar los pasos que realiza el programa.
 - Si el usuario quiere hacer el estudio de la relación con una sola muestra, deberá ejecutar la rutina hasta la línea 187.
 - En el caso de tener la segunda muestra, se deberá ejecutar hasta la línea 323, donde se obtendrá el detalle de la segunda regresión.
 - Para realizar la regresión total junto con la prueba de comparación de coeficientes, ejecutando el resto del script lpr.R se unirá la “planilla1” junto con la “planilla2” y se obtendrán los detalles de la regresión total y una tabla descriptiva que verifica la comparación de coeficientes.
 - En el caso de aceptar la hipótesis nula $b = 3$, se debe considerar el script lp3r.R. En la línea 29 del script se colocará la ubicación de la planilla en la que se aceptó el supuesto que se puede considerar $b = 3$. Ejecutando la rutina, se podrá obtener el ajuste y tabla de resumen para $\widehat{\ln(a)}$ y \hat{a} .
 - Los gráficos que se generan que son los gráficos de dispersión con recta ajustada, histogramas de frecuencias absolutas de residuos, tablas de estadística descriptiva de L y P, tablas de resumen de la regresión y de comparación de coeficientes, aparecen en la pestaña “Plots”. Una forma de poder guardarlos en la PC es hacer clic en la pestaña “Export” y tenerlo en el formato que uno desee. En el caso que el archivo tenga una extensión grande y no se pueda guardar como imagen, se puede solucionar haciendo clic en “Save as Web Page”, luego se aparecerá este archivo en el navegador y posteriormente haciendo clic en “imprimir”, se generará un archivo PDF.

Bibliografía

- Aubone, A. 2023. Programa Nmin1 v:111023. Estimación del tamaño mínimo de muestreo con múltiples objetivos. BIOMAT-INIDEP/CEMIM-UNMdP.
- Jové, A. 2023. Programa lpr.R y lp3r.R v220823. Programas para R para estimar los coeficientes de la función que relaciona el peso medio y la longitud. BIOMAT/INIDEP.
- Huxley, J. and Teissier, G. (1936) Terminology of Relative Growth. Nature, 137, 780-781. <http://dx.doi.org/10.1038/137780b0>.
- Draper, M.R. and Smith, H. 1980. Applied Regression Analysis, Second Edition. Wiley Series in Probability and Mathematical Statistics. 709 pp.
- Daniel, W. 1990. Applied Nonparametric Statistics. Second Edition. PWS-KENT Publishing Company. 635 pp.
- Goicoechea, H. 2018. <https://www.fbc.unl.edu.ar/laboratorios/ladaq/wp-content/uploads/2016/06/2-Estadistica-basica-2018.pdf>, UNL.
- Moreno, G.J. y Miravalls Sierra, E. 2016, Estadística I, Apuntes UAM. 160pp