Curso de Doctorado: "Introducción al lenguaje R. Modelos lineales y fundamentos de programación"

Unidad Académica organizadora:

- Doctorado en Ciencias Biológicas

Responsable Académico:

- Dr. Andrea Arístides Cocucci

Temario a Desarrollar

Unidad 1. Introducción al lenguaje R.

1.1 Introducción e instalación del software R. Instalación de paquetes específicos. Presentación del lenguaje R: lenguaje orientado a objetos. Diferencias con otros programas estadísticos. Tablas de funciones y ayudas de R. R como *lingua franca* en el ámbito científico. RStudio como Interfaz Gráfica de Usuario

1.2 Ingreso y manipulación de bases de datos. Creación de vectores, matrices y marcos de datos. Indexación.

1.3 Trabajo con rutinas. Concatenacion de operaciones con operadores pipa. Comparación de rutinas en R-base y tidyverse (dplyr).

Unidad 2. Modelos lineales.

2.1 Exploración gráfica de datos. Definición de modelos lineales. Relaciones lineales y no lineales, gráficos de dispersión y líneas de mejor ajuste. Regresión lineal simple y Análisis de la Varianza. Método de mínimos cuadrados. Bondad de ajuste. Análisis gráfico. Interpretación de parámetros. Gráficas básicas de diagnóstico.

2.2 Regresión lineal múltiple. Colinealidad y factores de inflación de la varianza. Modelos con interacción: Modelos de ANCOVA. Regresiones con variables dummy.

2.3 Selección de modelos. Compromiso entre sesgo y varianza de un modelo. Datos de entrenamiento y datos de prueba. Validación cruzada. Selección basada en criterios y promediado de modelos. *Ridge regression* y *Lasso*.

Unidad 3. Gráficos y reporte de resultados.

3.1 Manejo de gráficos en R-base: Comandos gráficos de alto y bajo nivel, parámetros gráficos. La función *plot* y sus comandos asociados de bajo nivel. División de la ventana de gráficos. Otras funciones gráficas comunes. Introducción a *lattice*.

3.2 La gramática de los gráficos en ggplot2. Componentes estéticos y uso de diferetentes geoms. Paletas gráficas amigables con el usuario.

3.3 Exportación de gráficos, ventajas y desventajas de los distintos formatos de imagen.

3.4 Introducción a R Markdown: reportes y análisis totalmente reproducibles en html y pdf.

Unidad 4. Introducción a la programación en R y aplicaciones prácticas.

4.1 Construcción de funciones. Almacenamiento de rutinas. Bucles. Funciones condicionales. Las funciones *function*, *for* e *if*. La serie de funciones *apply*. La función *sample*.

4.2 Métodos de remuestreo y randomizaciónes. Aplicaciones en biología: modelos nulos y modelos lineales randomizados.

4.3 Técnicas de bootstrap. Elección de intervalos de confianza.

Objetivos del curso

1. Adquirir destreza en el uso del lenguaje R para su uso en estadística básica y gráficos, así como nociones de programación.

2. Desarrollar una visión global de una amplia clase de modelos, poniendo énfasis en aquellos comúnmente utilizados en ciencias biológicas.

3. Adquirir destreza en la interpretación de análisis estadísticos y gráficos presentados en trabajos científicos.

Justificación

R es un sistema que posee una doble naturaleza de programa de análisis y de lenguaje de programación. Fue creado por Ross Ihaka y Robert Gentelman, como un derivado del lenguaje estadístico S. Sin embargo, a diferencia de su antecesor y de S-Plus, R es un software gratuito, de libre circulación y que puede ser copiado, distribuido y

modificado a voluntad. Adicionalmente posee una serie de ventajas que justifican su enseñanza en el ámbito universitario.

1. Es actualmente el software más comprensivo ya que ofrece más funciones que cualquier otro disponible. Esta versatilidad se debe a que colaboradores de todo el mundo producen paquetes con fines tan distintos como análisis de sonidos, estadística multivariada, cálculo de índices ecológicos y análisis filogenético, entre otros. Actualmente existen casi 20000 paquetes y un gran número es creado anualmente.
2. Además de las funciones construidas por los colaboradores de R, cada usuario puede crear sus propias funciones y rutinas ad hoc combinando las funciones existentes. Esta ventaja surge de la naturaleza abierta del lenguaje R.
3. Existe una amplia comunidad de usuarios, por lo que existe una gran oferta de ayudas para aquellos que se inician en el uso del programa, incluyendo foros de discusión en internet y una creciente bibliografía.
4. Ofrece múltiples aplicaciones para la creación de gráficos de alta calidad.
5. Actualmente R es una *lingua franca* para el intercambio de ideas en el ámbito científico, como puede advertirse en los trabajos que citan este software y en la presencia de rutinas en este lenguaje publicadas como material suplementario en revistas internacionalmente prestigiosas.

Debido a las múltiples aplicaciones que puede tener el programa, en este curso se ha elegido introducir a los alumnos en su uso mediante la enseñanza de funciones básicas destinadas al uso de modelos lineales en las ciencias biológicas. Este tipo de modelos se encuentran entre los de uso más común en varias áreas de investigación en biología, además de que constituyen un ejemplo del funcionamiento de la mayoría de las funciones disponibles en R. Se propone además brindar a los alumnos con herramientas para la adecuada representación gráfica de estos modelos y para comprender los fundamentos de la programación en este lenguaje. Estos fundamentos se aplicarán para la construcción de funciones, rutinas y en métodos de remuestreo y randomización. Finalmente, se pretende que el alumno desarrolle una visión crítica de la aplicación de los modelos estadísticos en el trabajo científico.

Contenidos mínimos

Unidad 1. R como lenguaje orientado a objetos y *lingua franca* en el ámbito científico. Ingreso y manipulación de bases de datos. Trabajo con rutinas en R-base y tidyverse.

Unidad 2. Exploración gráfica de datos. Modelos lineales, interpretación de parámetros, gráficas básicas de diagnóstico. Selección de modelos. Compromiso entre sesgo y varianza, datos de entrenamiento y de prueba.

Unidad 3. Manejo de gráficos en R en loas paquetes *graphics*, *lattice* y *ggplot2*. Exportación de gráficos. Reporte de resultados y de análisis reproducibles a partir de R Markdown.

Unidad 4. Construcción de funciones, bucles y funciones condicionales. Métodos basados en remuestreo y randomizaciónes. Bootstrap.

Nombre de el/los disertante/s (se adjuntan CVs reducidos)

- Dr. Santiago Miguel Benitez-Vieyra
- Dr. Andrea Arístides Cocucci

Destinatarios de la actividad

El curso está destinado a alumnos de doctorado o investigadores en biología y ciencias afines. Se requiere de los alumnos conocimiento de inglés y tendrán preferencia aquellos que posean conocimientos de estadística, obtenidos en cursos de posgrado anteriores, así como también aquellos alumnos del Doctorado en Cs. Biológicas de la esta casa de estudios. Se establece un cupo mínimo de 15 y máximo de 30 alumnos.

Fecha de realización

- 29 de mayo al 16 de junio de 2023 (periodicidad anual).

Duración y programa de actividad diaria

- Duración: 40 Hs. Se establecen dos modalidades alternativas de dictado, a definir según las necesidades del Doctorado en Cs. Biológicas de la UNC y la disponibilidad del disertante.

Modalidad presencial.

Día 1.

9:00 a 11:00. Teórico: Contenido de la materia. Presentación del lenguaje R.

11:00-13:00. Práctico: Instalación y actualización del programa. Instalación de paquetes. Creación de objetos. Indexación.

14:00-16:00. Teórico: Regresión lineal simple y ANOVA. Métodos de mínimos cuadrados y pruebas de bondad de ajuste. Datos de entrenamiento y datos de prueba.

16:00-18:00. Práctico. Ingreso de datos y modelos lineales sencillos.

Día 2.

9:00 a 11:00. Teórico: Regresión lineal múltiple. Métodos para detectar colinealidad.

11:00-13:00. Teórico: Modelos con interacciones. Selección de modelos

14:00-16:00. Práctico: selección de variables y selección de modelos.

16:00-18:00. Práctico: Comparación entre los métodos de selección de modelos.

Día 3.

9:00 a 11:00. Práctico: Manejo de gráficos en R. Exportación de gráficos.

11:00-13:00. Práctico: Representación gráfica de modelos lineales.

14:00-16:00. Práctico: Gráficos: Paquetes *lattice* y *ggplot2*.

16:00-18:00. Práctico: Construcción de rutinas y utilidades de Rstudio para elaborar reportes.

Día 4.

9:00 a 11:00. Práctico: Presentación de resultados, introducción a R Markdown.

11:00-13:00. Práctico: Presentación de resultados, introducción a R Markdown.

14:00-16:00. Práctico: Introducción a la programación en R.

16:00-18:00. Práctico: construcción de una función.

Día 5.

9:00 a 11:00. Teórico: Introducción a técnicas de remuestreo. Aplicaciones biológicas: modelos nulos en ecología y modelos lineales randomizados.

11:00-13:00. Práctico. Manejo de la función sample y contrucción de funciones para modelos nulos y randomizados.

14:00-16:00. Introducción a la técnica de bootstrap.

16:00-18:00. Práctico. Manejo del paquete *boot*.

Modalidad a distancia.

Día 1.

Teórico: ¿Por qué R? Contenido del curso.

Práctico guiado: Introducción al lenguaje. Ingreso y manipulación de datos.

Trabajo práctico: preparación de datos, ingreso y manipulación.

Día 2.

Encuentro sincrónico, resolución de dudas y ejercicios.

Día 3.

Teórico: Introducción a modelos lineales. Predicción: datos de entrenamiento y de prueba.

Práctico guiado: Modelos lineales simples.

Trabajo práctico: Ejercicios sobre modelos lineales.

Día 4.

Encuentro sincrónico, resolución de dudas y ejercicios.

Día 5.

Teórico: Regresión lineal múltiple e interacciones.

Práctico guiado: Modelos lineales múltiples I.

Trabajo práctico: Ejercicios sobre modelos lineales múltiples.

Día 6.

Encuentro sincrónico, resolución de dudas y ejercicios.

Día 7.

Teórico: Colinealidad, selección de modelos y regularización.

Práctico guiado: Modelos lineales múltiples II.

Trabajo práctico: Ejercicios

Día 8.

Encuentro sincrónico, resolución de dudas y ejercicios.

Día 9.

Práctico guiado: Manejo de gráficos en R.

Trabajo práctico: Ejercicios.

Día 10.

Encuentro sincrónico, resolución de dudas y ejercicios.

Día 11.

Práctico guiado: manejo de *lattice* y *ggplot2*.

Trabajo práctico: Ejercicios

Día 12.

Encuentro sincrónico, resolución de dudas y ejercicios.

Día 13.

Práctico guiado: Introducción a *R Markdown*.

Trabajo práctico: Ejercicios.

Día 14.

Encuentro sincrónico, resolución de dudas y ejercicios.

Día 15.

Práctico guiado: Introducción a la programación en R.

Trabajo práctico: Construcción de funciones.

Día 16.

Encuentro sincrónico, resolución de dudas y ejercicios.

Día 17.

Teórico: Introducción a técnicas de remuestreo. Modelos nulos y randomizaciones.

Práctico guiado: Función sample y construcción de funciones para modelos nulos y randomizados.

Trabajo práctico: Ejercicios.

Día 18.

Encuentro sincrónico, resolución de dudas y ejercicios.

Día 19.

Teórico: Bootstrap.

Práctico guiado: Construcción de bootstraps.

Trabajo práctico: Ejercicios.

Día 20.

Encuentro sincrónico, resolución de dudas y ejercicios.

Metodología a utilizar en el dictado

El curso comprende 40 horas de duración, en dos modalidades posibles.

En la ***modalidad presencial*** comprende un dictado intensivo a lo largo de cinco días (aproximadamente ocho horas diarias), repartidas en doce horas de clases teóricas y veintiocho horas de práctica en computadora y de discusión de resultados. En esta modalidad se requiere que cada alumno asista con su computadora personal.

La ***modalidad virtual*** involucra tres semanas con encuentros sincrónicos tres veces por semana durante dos horas (totalizando 18 horas sincrónicas). Se provee al alumno de material grabado con clases teóricas y ejercicios guiados y se espera que el alumno resuelva ejercicios antes de cada encuentro.

Se proveerá al alumno de rutinas en lenguaje R para los trabajos prácticos, así como también de material grabado para las clases en la modalidad a distancia. Para la

evaluación final se desarrollará un trabajo que utilice las metodologías aprendidas, aplicadas a problemas biológicos. Se espera que al final del curso el alumno sea capaz de escribir una rutina de análisis completa en R, aplicada a la solución de estos problemas.

Bibliografía y material didáctico que se proveerá a los asistentes

Boelker, B.M. (2008). *Ecological Models and Data in R.* Princeton University Press, USA.

Burnham, K. & Anderson, D.R. (2002) *Model Selection and Multimodel Inference : A Practical Information-Theoretic Approach.* Springer, New York.

Chambers, J.M. (2008). *Software for Data Analysis. Programing with R.* Springer, USA.

Dalgaard, P. (2008). *Introductory Statistics with R.* Springer, USA.

Gandrud, C. (2013). *Reproducible Research with R and Rstudio*. CRC Press, USA.

Hector, A. (2015). *New statistics with R: An introduction for biologists*. Oxford University Press, UK.

James, G., Witten, D., Hastie, T. & Tibshirani, R. (2013) *An Introduction to Statistical Learning.* Springer, USA.

Kabacoff, R.L. (2022) *R in Action*. 3rd Edition. Dreamtech Press, USA.

Kutner, M., Nachtsheim, C., Neter, J., & Li, W. (2004). *Applied Linear Statistical Models.* McGraw-Hill, UK.

Logan, M. (2010) *Biostatistical Design and Analysis Using R : A Practical Guide.* Wiley-Blackwell, Chichester UK.

Manly, B.J.F. (2007) *Randomization, Bootstrap and Monte Carlo Methods in Biology.* Chapman & Hall, UK.

Paradis, E. (2003). *R for Beginners.* Institut des Sciences de l'Evolution. Université Montpellier II, France.

Quinn, G.P. & Keough, M.J. (2002). *Experimenal Design and Data Analysis for Biologists.* Cambridge University Press, UK.

R Development Core Team (2017) *An Introduction to R.* R Foundation for Statistical Computing, Austria.

Sarkar, D. (2008). *Lattice. Multivariate Data Visualization with R.* Springer, USA.

Sheather, S.J. (2009). *A modern Approach to Regression with R.* Springer, USA.

Spector, P. (2008) *Data Manipulation with R.* Springer, USA.

Xie, Y. (2015). *Dynamic Documents with R and knitr.* CRC Press, USA.

Wickham, H. 2009. ggplot2. Elegant Graphics for Data Analysis. Springer New York, USA.

Wickham, H. & Grolemund, G. 2017. R for Data Science: Import, Tidy, Transform, Visualize, and Model Data. O'Reilly Media, USA.

Zuur, A.F., Ieno, E.N., & Meesters, E.H.W.G. (2009) A Beginners' Guide to R. Springer, USA.

Zuur, A., Ieno, E.N., Walker, N.J., Saveliev, A.A. & Smith, G.M. (2009) Mixed Effects Models and Extensions in Ecology with R. Springer, USA.

- **Material didáctico**

Se proveerá al alumno del material completo de las clases (teóricas, prácticas y repositorio de las rutinas y datos utilizados) a través del sitio de internet http://santiagombv.github.io/CursoR/

Evaluación final, metodología y profesores propuestos para realizarla

- **Evaluación**: SI
- **Tribunal**:

  Dr. Andrea Arístides Cocucci,

  Dr. Arnaldo Pedro Mangeaud,

  Dr. Mariano Pablo Grilli.

- **Aranceles**: $ 7000 para público en general, $5600 para estudiantes del Doctorado en Ciencias Biológicas de la FCEFyN, UNC.
- **Cupo**: 15 alumnos mínimo; 30 máximo.

Presupuesto estimativo y prioridades para la asignación de recursos

- **Honorarios**: El porcentaje a pagar se decidirá entre el Doctorado y Disertante.

Entidad que operará como unidad ejecutora de recursos

- Doctorado en Ciencias Biológicas