Biometría II - 2024

TP Nº 1: Modelos lineales

Objetivos

1. Repasar conceptos de biometría 1: estadística descriptiva, modelo lineal simple con predictora cuantitativa o categórica.
2. Utilizar R para análisis descriptivos y modelos con una sola variable predictora.
3. Poder construir gráficos informativos e interpretarlos.
4. Interpretar resultados en contexto.

Problema 1. Modelo de regresión lineal: una aplicación.

El daño al material genético por los pesticidas es un tema de preocupación a escala global. El aumento de la frontera agropecuaria y el uso de plaguicidas en la Argentina es un proceso del que aún se desconocen sus consecuencias sobre las comunidades de flora y fauna nativas. El glifosato es un herbicida utilizado para el control de malezas, inhibiendo el crecimiento de las plantas. Las formulaciones comerciales (ej Roundup®) incluyen mezclas para mejorar la eficacia de la acción del herbicida. Poletta y colaboradores (2009) realizaron un trabajo cuyo objetivo fue evaluar la potencial genotoxicidad del herbicida Roundup sobre eritrocitos del yacaré overo *Caiman latirostris*, luego de haber sido expuestos in ovo. Se plantea un experimento con el principal objetivo de relacionar el daño en el material genético de embriones de yacaré con la dosis de Roundup (RU). El experimento (adaptado para este problema) consistió en exponer a 11 huevos de yacaré a distintas concentraciones de RU entre 0 y 1750 ug/huevo. Los huevos fueron asignados al azar a las distintas concentraciones y la dosis de RU fue fijada por lxs investigadores. Al momento de la eclosión se tomaron muestras de sangre y se calculó el daño en el ADN mediante un índice de daño (“DI”, por sus siglas en inglés) y puede considerarse una variable cuantitativa continua. Los datos se encuentran en el archivo yacarés.csv (data.frame de 11 observaciones y 2 variables).

Explorar el data.frame

“Diccionario de variables”:
“RU”: dosis de Roundup (ug/huevo), *integer*;
“DI”: índice de daño (unidades específicas), *numeric*.

Utilizando los datos proporcionados, responder:

1. ¿Es lógico suponer una relación lineal entre ambas variables?
2. Escribir el modelo lineal e interprete en el contexto del problema sus parámetros
3. Ajustar el modelo en R e interprete el resultado de la prueba de hipótesis sobre la pendiente con su $IC_{95\%}$. Informe la ecuación de la recta estimada.
4. Informar el R^2 e interprete su valor en el contexto del problema.

5. Si se repite el experimento, ¿espera obtener los mismos valores para los estimadores de los parámetros (mismos valores de β_0^ y β_1^)?
6. Presentar un gráfico con los valores predichos por el modelo que resuma los principales hallazgos.

-> Puede utilizar como guía el script yacares.R

Problema 2. Importancia del análisis gráfico: el cuarteto de Anscombe.

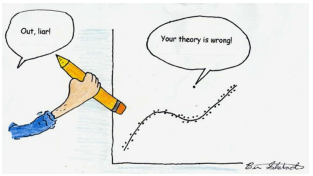
El dataset de Anscombe (1973, Francis Anscombe) comprende cuatro conjuntos de datos, de once pares de puntos cada uno (x, y). Muestra la importancia de graficar para analizar relaciones entre variables (lineal, cuadrática, otra), datos atípicos y observaciones influyentes. En este ejercicio vamos a trabajar con este dataset para relacionar como distintos conjuntos de pares (X, Y) pueden generar la misma (o muy similar) estadística descriptiva pero la manera en que se vinculan las variables pueden ser muy diferentes.

Cuarteto de Anscombe

I		II		III		IV	
x	y	x	y	x	y	x	y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

El archivo Anscombe_script.R , contiene una guía con código para hacer estadística descriptiva, gráficos y tendencias en estos cuatro conjuntos de datos. Visualizar y extraer conclusiones.

Problema 3. Atípicos e influyentes en RLS: ¿cómo se modifican las estimaciones? Algunos ejemplos.



Un outlier (o dato atípico) es una observación (dato) que no sigue el mismo patrón que el resto de los datos. En particular el método de cuadrados mínimos es muy sensible a estas observaciones, pudiendo afectar el ajuste del modelo y/o modificar sustancialmente la estimación de los parámetros del modelo. Un outlier en Y es una observación con un residual muy grande (en valor absoluto). Una observación es influyente si al excluirla del conjunto de datos la estimación cambia notablemente. Es decir, las observaciones influyentes tienen mucho peso en las estimaciones de los parámetros y en general son outliers en X.

El dataset que vamos a utilizar (ejemplo2b.csv) complementa la idea del cuarteto de anscombe, y fue extraído de teóricas de la Prof. Adriana Pérez (FCEyN-UBA), pensado para cursos iniciales de estadística. Muestra el efecto de un sólo dato atípico o influyente sobre la estimación de los parámetros y el ajuste en un modelo de RLS.

Problema 4. Rendimiento de Girasol en el Oeste de la Provincia de Buenos Aires.

Se desea ajustar un modelo para describir la relación entre el rendimiento del girasol (kg/ha) en el oeste de la Provincia de Buenos Aires y la aplicación de un nuevo fertilizante, compuesto por el tradicional más una mezcla novedosa de nitrógeno, fósforo y azufre ("Fert"). Se sospecha que a mayor contenido del nuevo fertilizante mayor será el rendimiento del girasol, siendo esta relación lineal. Para ello, se propone un experimento en el cual se fijan 9 valores de la nueva composición, entre 0 (composición tradicional pura) y un máximo de 80 gr / 100g de mezcla del nuevo fertilizante, asignados al azar 72 lotes de tamaño fijo y de manera balanceada. Al cabo de una temporada se registra el rendimiento de girasol. Supongamos que cada uno/a de ustedes es un investigador/a que realiza el ensayo y obtiene un conjunto de datos (x,y).

Parte A:

1. Obtener sus propios datos
2. Cargar el .txt en RStudio (llamar "datos" al data.frame). El data.frame contiene 72 observaciones (filas) y 2 variables (columnas).

Diccionario de variables:
"Fert": Gramos de fertilizante nuevo / 100 g de mezcla, integer,;
"Rend": Rendimiento de girasol (kg/ha), numeric.
Explorar y corroborar.

3. Realizar un gráfico de dispersión de rendimiento en función de los gramos de nuevo
4. fertilizante / 100 g de mezcla.
5. Ajustar un modelo de RLS a partir del conjunto de datos obtenido.
6. Escribir la ecuación estimada de la recta.
7. Calcular el $IC_{95\%}$ para la pendiente, primero "a mano" y luego con la función confit del R. Interpretar el $IC_{95\%}$ en el contexto del ensayo. ¿El $IC_{95\%}$, construido con sus datos, contiene a β1? ¿Puede saberlo?
8. En base al modelo ajustado, completar la planilla con los datos del curso. ¿Cómo son los valores de los estimadores obtenidos por cada alumnx?
9. Construir un histograma con los valores de β1 estimados (β1^).

-> Puede utilizar como guía el script rendimiento2024.R

Parte B: "la verdad"

1. Sabiendo que los valores de los parámetros poblacionales son: β0 =615 kg/ha y β1 =24 kg * ha^{-1} / gr fert * 100 g $mezcla^{-1}$. ¿Cuántos IC del curso contienen a β1? ¿Resultó - aproximadamente - dentro de lo esperado?

2. Ubicar el valor de β1 en el histograma construido en la parte anterior

Parte C: Simular datos bajo un modelo conocido

1. Siguiendo el script proporcionado, simulamos la obtención de tantas muestras como queramos, para:

Estimar muchas pendientes y calcular los $IC_{95\%}$ para cada valor de la pendiente estimado y ver si se cumple empìricamente que de ____ intervalos ____ van a contener al parámetro y ___ no. Compruebe mediante una tabla y gráfico si se cumple aproximadamente lo esperado

2. Analizar qué ocurre con el error estándar del estimador si aumentamos o disminuimos la varianza del modelo (“el ruido aleatorio”)

3. Modificar el n, generar un modelo con heterocedasticidad, con errores no normales, etc. Analizar qué ocurre con los parámetros estimados, con los supuestos, etc. en casa caso.

Problema 5. Fitorremediación en plantas

La fitorremediación es el proceso a través del cual algunas especies de plantas que son tolerantes a los metales pesados, los acumulan en sus órganos disminuyendo así su concentración en el ambiente que habitan. Para evaluar la capacidad fitorremediadora de la especie herbácea *Thlaspi caerulescens* se realizó un experimento en el cual se pusieron a crecer plántulas de esta especie en 32 frascos con medio de cultivo Hoagland a los que se le asignaron al azar y de manera balanceada una de las 8 concentraciones de cadmio suministradas (5, 10, 25, 50, 100, 200, 300 y 500 µM). Luego de 14 días se cosecharon las plantas y se midió la concentración de Cd acumulada en tallos [mg/kg], obteniéndose los resultados que se encuentran en el archivo Cadmio.txt (data frame de 32 obs y 2 variables)
Diccionario de variables:
"Cd": Concentración de cadmio suministrada (µM), integer;
"Cdenplanta": Concentración de Cd acumulada en planta luego de 15 días (mg/kg), integer.
Explorar y corroborar.

Para este experimento:

1. Indique la cantidad de réplicas y el tipo de variables involucradas.
2. Describa gráfica y estadísticamente los datos.
3. Analice cómo se modifica la concentración de cadmio absorbida por las plantas en relación a la concentración de cadmio ambiental. Plantee el modelo, evalúe los supuestos. ¿Considera que se puede proponer a *Thlaspi caerulescens* como un agente fitorremediador? Informe la magnitud del efecto.
4. Analice cómo se modifica la concentración de cadmio absorbida por las plantas en relación a la concentración de cadmio ambiental, pero ahora considerando a la predictora como categórica. Evalúe los supuestos. Interprete los resultados obtenidos.
5. Discuta ventajas, desventajas y alcances de cada aproximación (predictora cuantitativa o cualitativa). Complete el cuadro.
6. Pronostique para ambos modelos la concentración de cadmio absorbida por las plantas de *Thlaspica erulescens* sometidas a 500 µM de cadmio. ¿Podría predecir la

respuesta esperada a 450 µM? ¿Y si la concentración de cadmio ambiental supera los 600 µM?

→ Puede utilizar como guía el script cadmio 2024.R

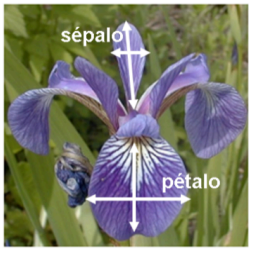
Cuadro comparativos predictora cuantitativa vs predictora cuantitativa:

	cuantitativa	cualitativa
Pregunta de investigación		
Tipo de relación asumida entre X e Y		
Potencia de la prueba en relación al otro enfoque		
Predicciones posibles (SI/NO):		
-> Dosis de 500 µM		
-> Dosis de 450 µM		
->Dosis de 600 µM		

7. ¿Podría haber utilizado el enfoque de variable predictora como cualitativa para resolver los problemas 1 (yacarés) y 4 (rendimiento) de esta guía? ¿por qué?
8. Para complementar este problema recomendamos la lectura del siguiente artículo: "Knowing when to draw the line: designing more informative ecological experiments"

Problema 6. Características morfológicas de plantas del género Iris

La base de datos denominada "iris" (Fisher, 1936) fue recolectada por Edgar Anderson para evaluar la utilización de diferentes características morfológicas de plantas del género Iris para diferenciar entre distintas especies de dicho género. La base consta de 150 datos de cuatro rasgos de flores (longitud y ancho del sépalo, longitud y ancho del pétalo; todas en centímetros), correspondientes a tres especies: Iris setosa, Iris versicolor e Iris virginica.

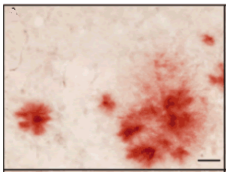



1. Abrir y explorar el data.frame (data.frame iris, disponible en R base)

2. Explorar gráfica y analíticamente los datos (tamaño de muestras, tendencia central, dispersión, datos faltantes, etc.).
3. Explorar gráfica y analíticamente la asociación (correlaciones) entre variables morfométricas.
4. ¿Hay evidencia para decir que las especies difieren en relación a la longitud del sépalo?
5. En caso que difieran, realizar comparaciones a posteriori y analizar la magnitud del efecto.
6. Presentar un gráfico con los resultados.

→ Para la exploración gráfica y correlaciones, puede utilizar como ayuda el script iris Gráficos.R
→ Para el modelo lineal y comparación de medias puede utilizar como guía el script irisCompMedias.R

Problema 7. Terapia con anticuerpos para enfermedad de Alzheimer



La enfermedad de Alzheimer (EA) es una enfermedad neurodegenerativa y representa la forma más común de demencia senil. Aunque las causas de la EA no han sido elucidadas, se sabe que está asociada con el depósito de proteína amiloide-β (Aβ) en la corteza cerebral e hipocampo y con la neuroinflamación. Existe mucho interés en la identificación de terapias que puedan ser eficaces en el tratamiento de esta enfermedad. Entre las nuevas líneas de investigación que persiguen alcanzar un tratamiento para la EA, una de las más prometedoras es la denominada inmunoterapia, que consiste en el tratamiento con anticuerpos anti-Aβ de manera de reducir los depósitos de amiloide beta presentes en corteza frontal e hipocampo de individuos enfermos. Se efectuó un ensayo para evaluar la eficacia del tratamiento con anticuerpos anti-Aβ. Para ello, ratones transgénicos PDAPPAJ20 con EA fueron sometidos a tratamiento con anticuerpos anti-Aβ. Al inicio, 1 mes, 2 meses y 3 meses de tratamiento los ratones fueron sacrificados y por ELISA se determinó la masa total de amiloide beta en cerebro (en ng), obteniéndose los resultados que se encuentran en el archivo terapiaEA.xls

Se aplicó un análisis de regresión simple.

Para este experimento:
1. Indique la cantidad de réplicas y el tipo de variables involucradas.
2. Describa gráfica y estadísticamente los datos.
3. Sabiendo que se aplicó un análisis de regresión simple, analice cómo se modifica la masa total de amiloide beta en el cerebro con el tiempo luego de aplicado el tratamiento con anticuerpos. Informe la magnitud del efecto en el contexto del ensayo.