P. Joser Atauchi ∗ Joel Olarte

Análisis Bioestadístico con R

Primera Edición

P. Joser Atauchi

Investigador Asociado
Museo de Historia Natural de Cusco (MHNC)
Cusco, Perú

Joel Olarte

Profesor
Universidad Nacional de San Antonio Abad del Cusco
Cusco, Perú

Capítulo 1

Análisis de Regresión Lineal

En este capítulo revisaremos conceptos básicos de regresión lineal y mostraremos como aplicar código R usando datos ecológicos, además discutiremos acerca de la validación y las limitaciones que tienen las regresiones lineales en el campo de las ciencias biológicas. A menudo los científicos obtienen datos de dos o más variables en cada muestreo o experimento. Por ejemplo, los ecólogos pueden registrar datos de abundancias de una especie en particular, riqueza de especies en bosques distintos, características del suelo como pH, tipo de suelo. Estos datos son llamados **bivariados** cuando se tienen dos variables aleatorias y **multivariados** cuando se tiene más de dos variables aleatorias registradas.

Despierta una gama de preguntas relevantes que requiera la obtención de dichos datos basados en su naturaleza y las relaciones biológicas y estadísticas entre las variables.

Los próximos tres capítulos consideran distintos procedimientos estadísticos para describir la relación que existe entre las variables en estudio. Distintas técnicas y procedimientos para analizar datos complejos y multivariados también serán tratados.

1.1. Modelo de Regresión Lineal

Una simple relación entre una variable y otra en una población es una regresión lineal simple. Constantemente hablaremos de ajustar y validar modelos. Estos términos se emplea en situaciones donde se puede definir una variable de respuesta (variable dependiente: Yi) y dos o más variables predictivas (variables independientes: $X_1, X_2, X_3, ...X_n$). Un valor para cada variable de respuesta y predictiva es obtenido a partir del muestreo o experimentos unitarios de las poblaciones.

Un modelo de regresión simple bivariante está definido por:

$$y = \alpha + \beta X_i + \varepsilon_i \tag{1.1}$$

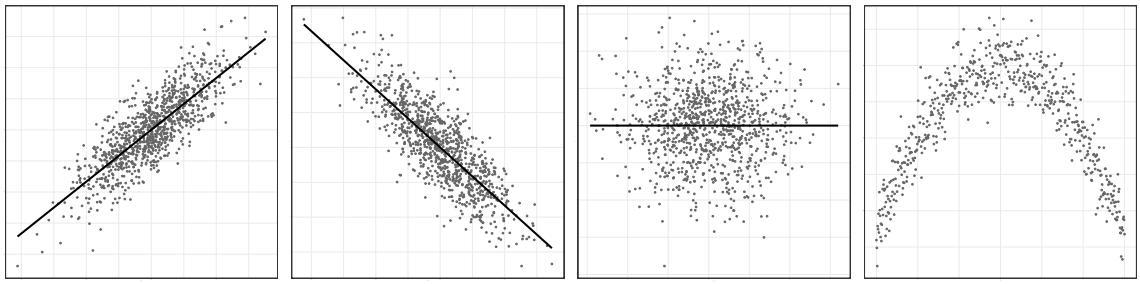
En la ec. 1.1, Y_i representa la variable de respuesta (o dependiente) y X_i la variable predictiva (o independiente). La información que no se puede explicar con el modelo se almacena en los residuales o error (ε_i), estos valores asumen una distribución normal y la suma de sus términos es 0 ($\sum_{i=1}^{n} \varepsilon_i = 0$). Los parámetros de la población α y β son el intercepto y la pendiente, respectivamente [1] . En la práctica, en muchos casos, el principal interés es estimar β (la pendiente) para responder si hay una relación entre X y Y.

1.1.1. Estimación de Parámetros

Pendiente de Regresión

El parámetro β es conocido como *Coeficiente de regresión*. Científicos a menudo concuerdan que es el parámetro más importante de la recta de regresión por qué esta medida define la fuerza en que la variable predictiva se relaciona con la variable de repuesta.

La pendiente de una recta de regresión puede ser positiva, negativa o cero (Fig. 1.1). Si $\beta > 0$, la tendencia creciente, es decir a valores mayores de la variable explicativa le corresponde valores mayores de la variable de respuesta; Si $\beta < 0$, la tendencia es decreciente, valores mayores de la variable explicativa le corresponde valores menores de la variable de respuesta, o viceversa; Si $\beta = 0$, entonces la variable de respuesta permanecería constante (estacionario) cuando la variable explicativa es dinámica. En este caso se dice que no hay regresión.



(a) **(b)** **(c)** **(d)**

Figura. 1.1: Las pendientes de una recta de regresión puede ser: **(a)** Positivo, **(b)** Negativo, **(c)**Ningúno, **(d)** No lineal.

[1] α y β son comúnmente usados en los textos de estadística, sin embargo el uso de β_0 y β_1, también son asignados para el intercepto y la pendiente, respectivamente. No debería provocar equivocaciones la nomenclatura de Greek que asigna α y β al tipo de error I y II.

Intercepto

Un número infinito de lineas puede tener el mismo valor de pendiente, todo ellos paralelos (Fig. 1.2a). El intercepto puede no ser nuestro mayor inveteres en un análisis de regresión por que el rango de nuestras observaciones raramente incluye valores de X (variable productiva) igual a 0; y no se sabe con certeza si nuestra muestra contiene los valores extremos de la población en estudio. Por lo tanto, no debe pensarse en realizar extrapolaciones más allá de los datos obtenidos.

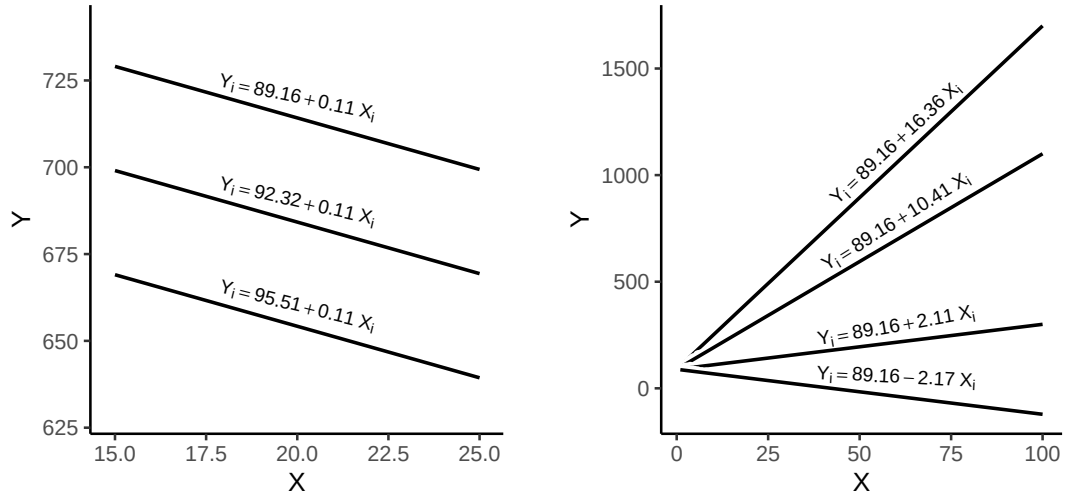


Figura. 1.2: Simulación de los parámetros fijos de una regresión lineal: **(a)** Pendiente fijo. **(b)** Intercepto fijo.

En la figura 1.2 se puede demostrar que la información única de α o β puede llevarnos a describir errónea e incompleta la función de regresión. Por cada pendiente dada fijo ($\beta = 0{,}11$), existe infinitas rectas paralelas (Fig. 1.2a); y para cada intercepto dado ($\alpha = 89{,}16$) existe finitas rectas no paralelas que determina la fuerza de relación entre las variables (Fig. 1.2b). Pero, si especificamos ambos parámetros α y β obtenemos una única recta de regresión.

1.1.2. Intervalo de Confianza

Una vez encontrado el estimador puntual del modelos de regresión ($\alpha,\ \beta$), podemos prestar atención a la distribución de muestreo, al error estándar y al intervalo de confianza para la pendiente. El IC para una regresión lineal está definida por:

$$\hat{\beta} \pm t_{0{,}05,n-2,\alpha/2} SE(\hat{\beta}_1) \tag{1.2}$$

Donde, $t_{0{,}05,n-2,\alpha/2}$ es el percentil de distribución t de Student con $n-2$ grados de libertad que deja a su derecha un área $\alpha/2$. Un intervalo de confianza no tiene sentido alguno si la pendiente es 1.

1.2. Supuestos y Validación

En el campo de las ciencias biológicas, los datos rara vez se modelan o ajustan adecuadamente usando modelos lineales simples. Al aplicar un modelo de regresión lineal a los datos, implícitamente se está asumiendo una serie de supuestos. La validación de modelos de regresión es la verificación de los supuestos. En esta sección revisaremos aspectos de la (a) Normalidad, (b) Homogeneidad, e (c) Independencia.

En la practica, muchas veces se olvidan de verificar que supuestos están asumiendo en sus modelos. Pero, ¿Qué tan malo es un modelo de regresión lineal que viola todos los supuestos?. La respuesta es simple, debe de rechazar el modelo. A partir de este punto puede realizar transformaciones y re-analizar para validad los supuestos u optar por otros métodos analíticos.

1.2.1. Normalidad

Este supuesto trata que, la población de los valores de Y_i y el error (ε_i) son normalmente distribuidos para cada nivel de la variable predictora (X_i). La forma correcta de verificar este supuesto sería realizar un histograma de todas las observaciones para un valor particular de X, pero no muchas veces se logra realizar sub muestreos para realizar esta tarea. Entonces, ¿Cómo puedo validar el supuesto de normalidad si solo cuento con un muestreo?. La respuesta es analizar los residuales del modelo de regresión. Los residuales almacenan la información que no pudo ser explicada con el modelo. Una mala practica es aceptar o rechazar la normalidad usando unicamente un histograma de los datos.

Algunas investigaciones justifican que la violación al supuesto de normalidad es un problema serio (REFERENCIAS), debido a que es una consecuencia de la teoría de limite central. También, a menudo se encuentra en la literatura que el supuesto de normalidad no es necesario verificar o se asume normalidad en largos tamaños de muestra ($n > 30$) (Whitehead et al. [2016])

1.2.2. Homogeneidad de Varianza

El supuesto de homogeneidad de varianza (u Heterocedasticidad) es muy importante, el efecto de su violación puede traer problemas y grandes efectos sobre la estimación de parámetros, intervalos de confianza, prueba de hipótesis sobre la pendiente. A menudo, las varianzas heterogéneas provienen de poblaciones con distribución sesgada o también puede deberse al pequeño número de observaciones extremas, máximos y mínimos de la distribución.

En la práctica, se usan dos maneras de probar la homogeneidad: (i) Inspecciones gráficas y (ii) pruebas estadísticas. Se muestra ambas técnicas en las próximas secciones, sin embargo cabe hacer algunas aclaraciones. Si bien los métodos de inspección gráfico es muy criticado para probar la homogeneidad debido a que asume aspecto de subjetividad que usar una prueba estadística. Por lo contrario, las pruebas más usadas y conocidas como la prueba de Barlet para probar la homogeneidad es muy sensible a la no-normalidad.

1.2.3. Independencia

Es increíble que en la literatura uno encuentre este gran problema de independencia de datos. En ecología y en el campo de las humanidades, las pruebas más utilizadas son la prueba t y la prueba F, estas pueden ser invalidadas si se viola este supuesto. En enfoques como los modelos de nicho ecológico aplicado a la transmisión de enfermedades esto podría es un problema muy grave, ya que nuestros resultados podrían resultar en salvar vidas o no. En esta sección no trataremos de procedimientos para obtener datos con independiente [2], nos enfocaremos a identificar si nuestras datos son independiente o no.

Entonces, ¿Qué que debemos hacer para identificar si los datos son independientes?. La respuesta es, revisar los residuales de un modelo incorrecto. Suponga que ajusta una recta lineal en un conjunto de datos que muestra una tendencia no lineal. Consiguientemente, observa el patrón de residuos de la variable X. Lo que observa es un patrón de puntos positivos y negativos, por lo tanto se puede decir que se trata de un modelo inadecuado que provoca la violación de este supuesto.

Mis datos no son independientes ¿Que puedo hacer?. Linearizar la relación, un procedimiento que podría ser utilizado es transformar la variable X, sin embargo esto es muy discutido debido a que, al linearizar la relación se pierde la naturaleza de los datos.

1.3. Datos artificiales

Por ejemplo, consideremos las mediciones de la longitud del ala del Gallito-Hormigero de Pecho Rufo *Formicarius rufipectus* (Formicariidae) durante 24 días, medidas desde el tercer día de eclosión del huevo. Las edades (medidas en [*Días*]) son: 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25,y 26; y la longitud del ala (medidas en [*Cm*]) son: 1.4, 1.4, 1.5, 1.8, 2.2, 2.4, 2.8, 3.1, 3.2, 3.2, 3.9, 4.1, 4.7, 4.7, 4.7, 4.8, 5.0, 5.1, 5.7, 6.3, 6.7, 7.2, 7.8, 8.7.

La figura (1.3) mostró la relación entre la Longitud del ala y la edad del ave para una especie en particular. A partir de la figura se puede decir que la longitud del ala aumenta con la edad.

Usaremos un ejemplo hipotético con datos artificiales para mostrar los fundamentos y el código de R para verificar los supuestos y ajustar un modelo adecuado.

```
> Edad <- c(3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16,17,18,19,
            20, 21, 22, 23, 24, 25, 26)
> Ala <- c(1.4, 1.4, 1.5, 1.8, 2.2, 2.4, 2.8, 3.1, 3.2, 3.2, 3.9, 4.1,
           4.7, 4.7, 4.7, 4.8,5.0,5.1,5.7,6.3,6.7,7.2,7.8,8.7)
> Test1 <- lm(Age~Wing)
> op <- par(mfrow = c(2, 2))
> plot(Test1, add.smooth = FALSE)
> par(op)
```

[2] Textos que analizan diseño de experimentos deberían ser revisasdos para tener una idea

La primera y segunda linea de código construyen el objeto `Edad` y `Ala` con el número de días después de la eclosión del huevo y la medición de la longitud del ala. La función `lm` ajusta un modelo lineal y la `par(op)` configura la venta del resultado a su valores por defecto.

El gráfico de validación de modelos son (i) residuals versus fitted para verificar la homogeneidad, (ii) a QQ-plot o histograma de residuos para validad la normalidad, y (iii) residuals versus Leverage para validad la independencia (Fig. 1.3).

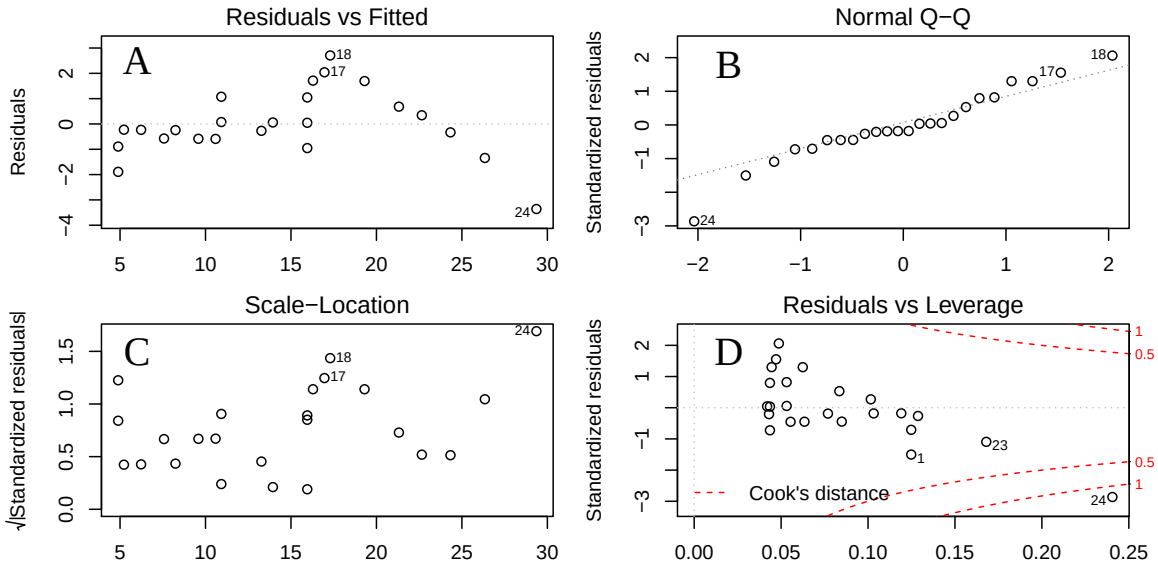


Figura. 1.3: Las pendien

La figura 1.3 es una salida típica al ejecutar la función `plot` al objeto `"lm"` en R. La violación de la homogeneidad de varianza puede ser detectada en la Fig. 1.3A y Fig. 1.3C si se muestra cualquier patrón de dispersión sobre los residuos. La violación de la normalidad puede ser detectada en la Fig. 1.3B, si la distribución de los puntos cae sobre la linea se puede asumir que los datos son normales. La violación sobre la independencia puede ser verificada en la Fig. 1.3D, se observa si el conjunto de datos tiene valores extremos de la variable predictiva.

La Fig. 1.3A muestra una clara violación de la homogeneidad y la Fig. 1.3D viola el supuesto de independencia. Para este conjunto de datos podemos decir que existe dos supuestos que han sido violados. En la Fig. 1.3 se observa los puntos de dispersión de la edad vs longitud del ala, y la recta de regresión lineal.

El gráfico puede ser obtenido mediante el siguiente código:

```
> plot(x=Edad, y=Ala, type='p',  xlab='Edad [Dias]',
       ylab = 'Long. del ala [Cm]')
> abline(Test1)
```

Ya se ha visto que la función `plot` utiliza los objetos `Edad` y `Ala` para construir el gráfico `xy`, y la función `abline` dibuja la recta de regresión del objeto `Test1`. El objeto `Test1` tiene más información de lo que parece, exploraremos los datos más relevantes para este modelo.

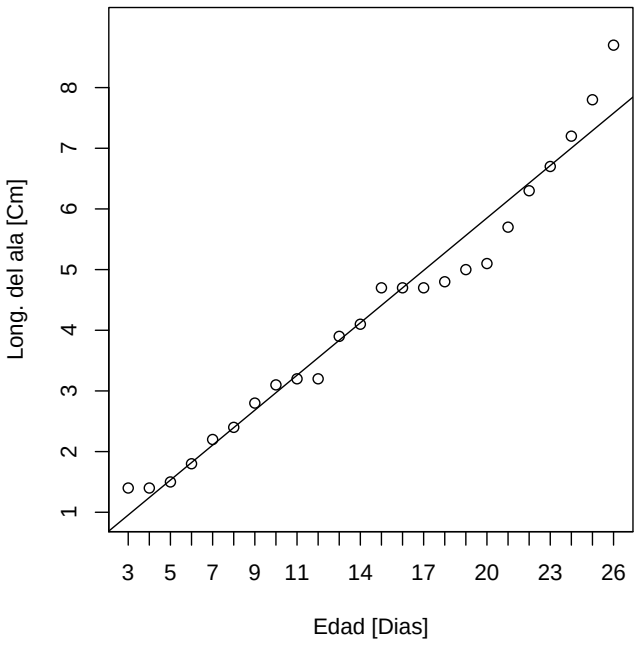


Figura. 1.4: Diagrama de dispersión: Longitud del Ala vs Edad de los pichones para *Formicarius rufipectus*

```
> summary(Test1)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.09067    0.18702   0.485    0.633
Age          0.28800    0.01164  24.743   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
Residual standard error: 0.3947 on 22 degrees of freedom
Multiple R-squared:  0.9653,Adjusted R-squared:  0.9637
F-statistic: 612.2 on 1 and 22 DF,  p-value: < 2.2e-16
```

Al construir un modelo de regresión lineal tenemos en mente encontrar los parámetros de la población (pendiente y el intercepto). En el resumen del modelo (`Test1`) podemos observar parámetros como el intercepto, la pendiente, el error estándar y el valor de p. Además, resultados como R^2 y R^2 ajustado son útiles al momento de redactar algún reporte. el modelo para la longitud del ala está dada por:

$$\hat{Y}_{Long.\ del\ ala} = 0{,}09 + 0{,}288 \times Edad_i$$

Por último, la pendiente son significativos ($p < 0{,}001$) con un nivel de significancia de 5 % y el modelo explica el 96.3 % la variación de los datos. Todo sería perfecto si no si hubieran violado ningún supuesto. Los dos supuestos violados (independencia y homogeneidad) hacen que debamos invalidad el modelo encontrado.

1.4. Datos de Defaunación

A lo largo del tiempo se ha usado el recurso natural como fuente de proteínas y energía por parte de las comunidades locales. Sin embargo, el crecimiento acelerado de las poblaciones ha traído un mejoramiento de las armas para caza y mayor acceso a lugar más alejados. La caza de especies (principalmente mamíferos y aves) tiene efectos directos e indirecto como la disminución de la abundancia de especies y por otro lado, el aumento en abundancia de especies poco aprovechadas por las comunidades locales. Koerner et al. [2017], examina el efecto que tiene la presión por la caza sobre la composición y estructura de aves y mamíferos tropicales. Usan la distancia a las comunidad locales como indicador de presión por la caza. Todos lo detalles del estudio pueden encontrase en su artículo. Para responder esta pregunta, usaron una regresión lineal para examinar la relación entre la variable de respuesta (Riqueza de especies) y la variable predictiva (Presión por la caza). Aquí usaremos una parte de sus datos (presión por la caza y riqueza) para repetir sus análisis.

Para un análisis de relaciones necesitamos definir dos variables (respuesta y predictiva). La variable de repuesta es la riqueza, con número de especies como unidades. La variable predictiva es la presión por la caza que tiene kilómetros(km) como unidades. Claramente podemos definir nuestro modelo lineal simple que esta dado por:

$$\hat{Y}_S \sim N(0, \sigma^2)$$
$$\hat{Y}_S = \alpha + \beta \times X_{Dist} + \varepsilon_i$$

La primera linea define que la riqueza (S) tienen una distribución normal y la segunda, el modelo de regresión lineal.

Aves

Primero, usaremos el modelo definido para el caso de la riqueza de Aves. El código de R para construir este modelo para la riqueza de aves, es:

```
> library(dataECO); data(Caza)
> AveCaza0 <- lm(Distance ~ Rich_BirdSpecies, data=Caza)
> op <- par(mfrow=c(2,2),mar=c(4,4,1.5,1.5))
> plot(AveCaza0, add.smooth = FALSE)
> par(op)
```

El modelo ajustado para la presión por la caza y la riqueza de aves se construye usando la función `lm`, con los objetos `Distance~Rich_BirdSpecies`.

La Fig. 1.5, se usa para verificar los supuestos. Recordemos que, los paneles **A** y **C** son usados para verificar la homogeneidad de varianza, el panel **B** usado para verificar la normalidad, y el panel **D** para la independencia. Entonces, en la Fig. 1.5A y C, vemos un patrón de forma que ciertas observaciones se agrupan en distancia categóricas violando el supuesto de homogeneidad de varianza. En la Fig. 1.5B, vemos que los puntos forman deformaciones en relación a la recta, violando el supuesto de normalidad. Por último, 1.5D no muestra valore extremos de la distribución por consiguiente se dice que los datos son independiente.

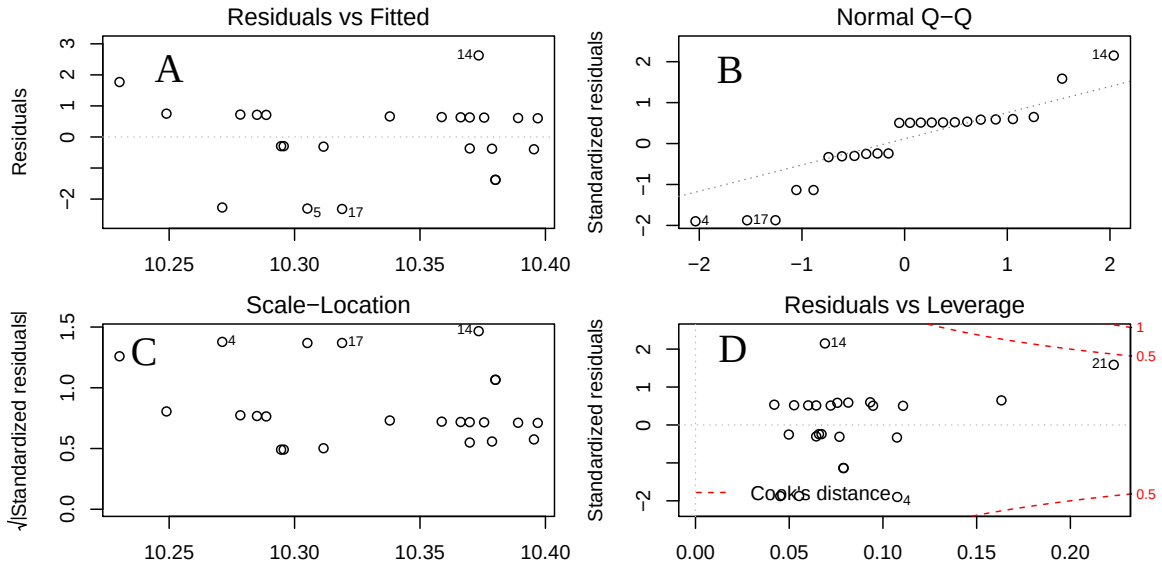


Figura. 1.5: Las pendien

Entonces, posterior a la validación del modelo podemos decir que la regresión lineal para la presión por la caza y la riqueza de aves viola dos supuestos, la homogeneidad de varianza y la normalidad.

La función `summary` del objeto `AveCaza0` muestra información relevante para el estudio.

```
> summary(AveCaza0)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 10.415654   0.502250  20.738 6.24e-16 ***
Distance    -0.006929   0.036249  -0.191     0.85
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.266 on 22 degrees of freedom
Multiple R-squared:  0.001658,Adjusted R-squared:  -0.04372
F-statistic: 0.03654 on 1 and 22 DF,  p-value: 0.8502
```

Tenemos el intercepto ($\alpha = 10{,}42$), la pendiente ($\beta = -0{,}006$), probabilidad ($p = 0{,}85$). Ahora, concentrémonos en el $R^2 = -0{,}044$, este valor pequeño para este parámetro nos dice que solo el 4 % de los datos es explicado por esta relación. Valores por debajo de 75 % se debería concluir que la relación no es consistente.

$$\hat{Y}_{Riqueza\ de\ Aves} = 14{,}42 - 0{,}006 \times X_{Distancia}$$

Este sería nuestro modelo que responde a la riqueza de aves y la presión por caza. Sin embargo, recordemos que este modelo ha violado dos supuesto —Homogeneidad y Normalidad—. Por lo que, este modelo es inválido.

Mamíferos

El modelo usaremos como variable de respuesta a la riqueza de Mamíferos. El código de R para construir este modelo para la riqueza de Mamíferos, es:

```
> library(dataECO); data(Caza)
> MamiferosCaza0 <- lm(Distance ~ Rich_MammalSpecies, data=Caza)
> op <- par(mfrow=c(2,2),mar=c(4,4,1.5,1.5))
> plot(MamiferosCaza0, add.smooth = FALSE)
> par(op)
```

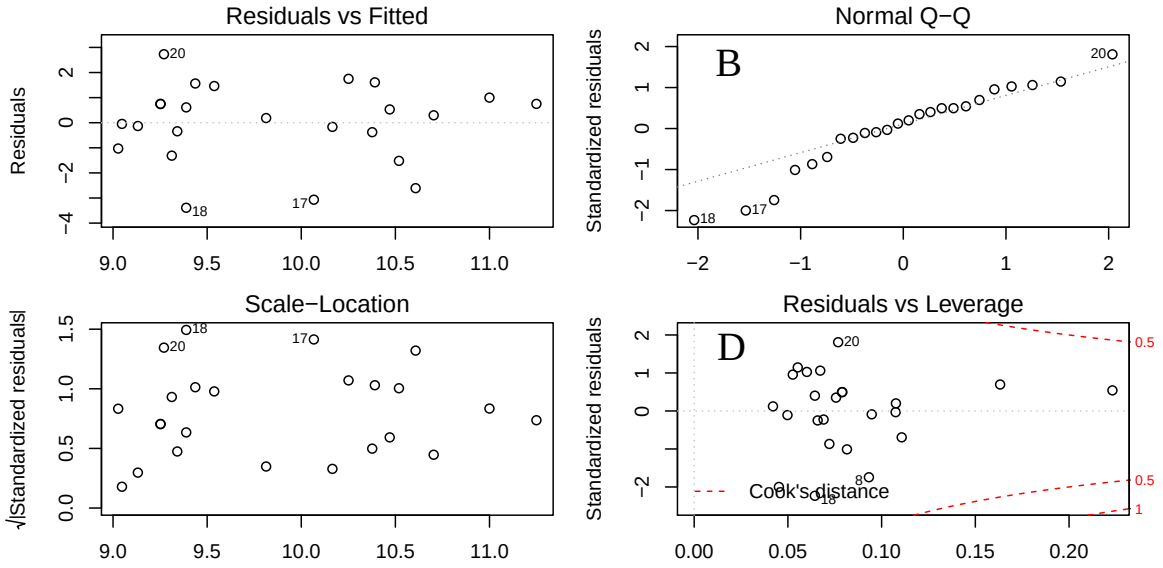


Figura. 1.6: Las pendien

El modelo ajustado para la presión por la caza y la riqueza de mamíferos se construye usando la función `lm`, con los objetos `Distance~Rich_MammalSpecies`.

Analicemos el modelo, La Fi.g 1.6A y C, vemos que no existe un patrón apreciable, entonces no viola el supuesto de homogeneidad de varianza. En la Fig. 1.6B, vemos que los puntos forman deformaciones en la curva que es dibujada por los puntos, violando el supuesto de normalidad. Por último, 1.6D no muestra valores extremos ni cercanos a 0.5, podemos decir que son datos independientes. Entonces, posterior a la validación vemos que sola viola un supuesto el de normalidad.

En el ejemplo anterior de la presión por la caza y las aves, vimos que violaba dos supuestos el de normalidad y la homogeneidad, y decidimos invalidar el modelo. Pero, que

pasa en este caso particular de la presión por la caza y las aves. Realizaremos dos procesos (1) definir el modelo asumiendo la no normalidad de los datos y (2) realizaremos una transformación a la variable explicativa.

```
> summary(MamiferosCaza0)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  8.77788    0.62267  14.097  1.7e-12 ***
Distance     0.09235    0.04494   2.055   0.0519 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.57 on 22 degrees of freedom
Multiple R-squared:  0.161,Adjusted R-squared:  0.1229
F-statistic: 4.223 on 1 and 22 DF,  p-value: 0.05195
```

El resumen del modelo `MamiferosCaza0` vemos que explica el 12 % de los datos $R2 = 0{,}123$, con una pendiente $beta = 0{,}092$ y un intercepto de $\alpha = 8{,}777$. El modelo de regresión esta dado, por:

$$\hat{Y}_{Riqueza\ de\ Mamiferos} = 8{,}777 + 0{,}092 \times X_{Distancia}$$

Transformación de datos: Mamíferos

Posterior al modelo de validación identificamos que los datos no son normales. Realizaremos una transformación de la variable explicativa (Distancia a la comunidad local). En el capítulos XX se muestra varios enfoques que se puede utilizar al momento transformar datos. Para la variable explicativa (Distancias [km]), usaremos una transformación logarítmica en base 10.

Los siguientes códigos aplica modelos de las ecuaciones XX-XXX en R y se compara usando la métrica de AIC.

```
> library(quantreg)
> Caza$Distance_trans <- log10(Caza$Distance)
> MamiferosCaza1 <- lm(Rich_MammalSpecies ~ Distance_trans, data=Caza)
> MamiferosCaza2 <- rlm(Rich_MammalSpecies ~ Distance_trans, data=Caza)
> MamiferosCaza2 <- rlm(Rich_MammalSpecies ~ Distance_trans, data=Caza,
                        psi = psi.bisquare)
> MamiferosCaza3 <- rq(Rich_MammalSpecies ~ Distance_trans, data=Caza)
```

La primera línea crea una columna con nombre `Distance_trans` a la cuál se asigno los valores logarítmicos en base 10 de la variable `Distance`. Las transformaciones logarítmicas son usados frecuentemente para mediciones como tamaño, altura, distancias, etc.

El primer modelo MamiferosCaza0 solamente usamos una regresión lineal entre la riqueza de mamíferos y la transformación de la variable Distance. El primero modelo MamiferosCaza1 y el MamiferosCaza2 utilizan modelos lineales robustos y se diferencian por el uso del parámetro psi=psi.bisquare que remueve los valores extremos u outliers. El modelo MamiferosCaza3 utiliza regresiones por cuantiles. Se puede utilizar el mismo código de arriba para ajustar otros modelos con especificaciones que puede elegir el usuario.

La Fig. 1.7 es construido el siguiente código, a cada gráfico se le agrega la recta de regresión estimada usando la función abline sobre cada modelo construido.

```
> op <- par(mfrow = c(2, 2), mar=c(2,4,2,1.2))
> plot(Caza$Rich_BirdSpecies~Caza$Distance_trans,
       main='MamiferosCaza0',cex.main=0.8)
> abline(MamiferosCaza0)
> plot(Caza$Rich_BirdSpecies~Caza$Distance_trans,
       main='MamiferosCaza1',cex.main=0.8)
> abline(MamiferosCaza1)
> plot(Caza$Rich_BirdSpecies ~ Caza$Distance_trans,
       main='MamiferosCaza2',cex.main=0.8)
> abline(MamiferosCaza2)
> plot(Caza$Rich_BirdSpecies~Caza$Distance_trans,
       main='MamiferosCaza3', cex.main=0.8)
> abline(MamiferosCaza3)
> par(op)
```

Selección de Modelo

Se tiene cuatro modelos MamiferosCaza0-3, se observa que los modelos 1, 2 y 3 construyen rectas muy similares. Entonces, ¿Cómo sabemos que modelo elegir?. Se puede usar distintos enfoques para seleccionar un modelo de regresión (1) se puede comparar usando pendientes, (2) criterios de información, (3) pruebas estadísticas para regresiones. Aquí, usaremos criterios información de Akaike (AIC). Cuando se comparan modelos usando máxima verosimilitud a un conjunto de datos, el menor valor de AIC se le otorga al modelo con mejor ajuste.

La siguiente linea de código utiliza la función AIC para calcular el valor de Akaike para los cuatro modelos construidos lineas arriba.

```
> AIC(MamiferosCaza0,MamiferosCaza1,MamiferosCaza2,MamiferosCaza3)
                df    AIC
MamiferosCaza0  3   93.66805
MamiferosCaza1  3   94.06370
MamiferosCaza2  3   94.54131
MamiferosCaza3  2   91.60761
```

Concentrémonos en los resultados de la tabla de AIC. A partir de estos valores, se debe seleccionar el modelo con menor valor de AIC que sería MamiferosCaza3 para este con-

junto de datos. Por último construiremos su definición extrayendo los parámetros usando la función `summary` al modelo seleccionado.

```
> summary(MamiferosCaza3)
Coefficients:
               coefficients lower bd upper bd
(Intercept)     7.91039      6.51049 10.37390
Distance_trans  2.34215      0.19164  3.77671
```

Nuestro modelo de regresión que responda la variación de la riqueza de mamíferos esta dado con `tau=0.5`, por:

$$\hat{Y}_{Riqueza\ de\ Mamiferos} = 7{,}91 + 12{,}342 \times X_{Distancia}$$

Además es posible construir su intervalo de confianza usando los parámetros `lower bd` y `upper bd` para mostrar como la recta se podría comportar a los largo de su intervalo de confianza.

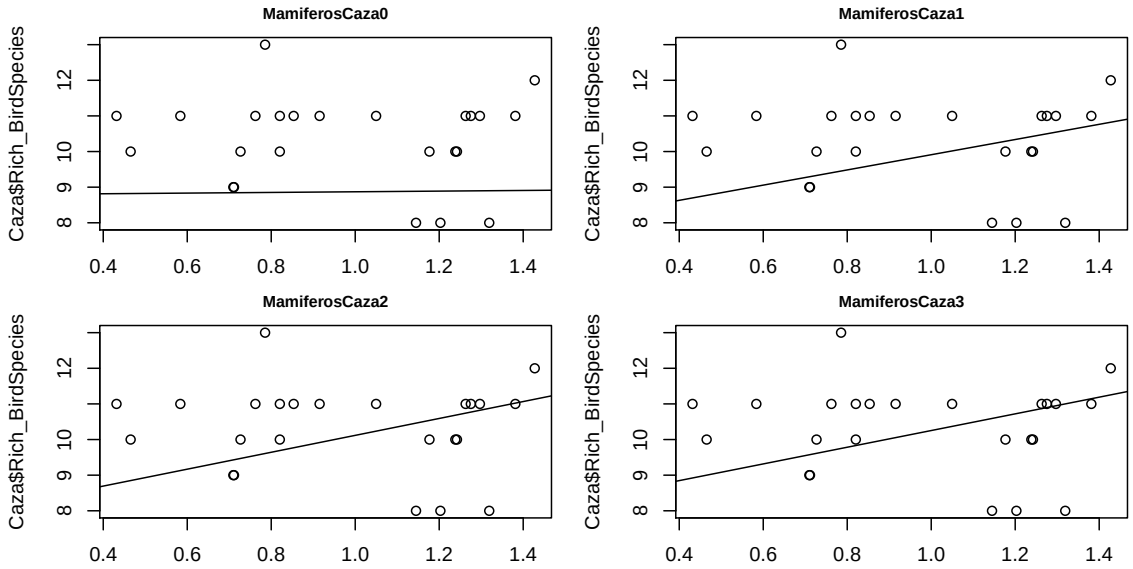


Figura. 1.7: Las pendien

1.5. Datos Libélulas

a

1.6. Que debo reportar

Bibliografía

S. E. Koerner, J. R. Poulsen, E. J. Blanchard, J. Okouyi, and C. J. Clark. Vertebrate community composition and diversity declines along a defaunation gradient radiating from rural villages in gabon. *Journal of Applied Ecology*, 54(3):805–814, 2017.

G. D. Ruxton and M. Neuhäuser. When should we use one-tailed hypothesis testing? *Methods in Ecology and Evolution*, 1(2):114–117, 2010.

S. R. Whitehead, M. F. O. Quesada, and M. D. Bowers. Chemical tradeoffs in seed dispersal: defensive metabolites in fruits deter consumption by mutualist bats. *Oikos*, 125(7):927–937, 2016.