**Ejercicio Propuesto 6 (Resuelto)**

**En 5 zonas de la provincia de Granada (Ladihonda y Fazares, zonas muy secas y Cortijuela, Molinillo y Fardes, zonas húmedas) se hacen una serie de mediciones sobre las hojas de las encinas a lo largo de 3 años consecutivos: 1995, muy seco y 1996 y 1997, muy lluviosos.**

**El objetivo es medir la simetría fluctuante en dichas hojas como indicador de stress en la planta. Bajo condiciones de stress (sequía, herbivoría, limitación por nutrientes...), la hipótesis es que la asimetría aumente. Contamos con la siguiente información:**

* **Localización árboles: 5 zonas, dos en zonas muy secas (Hoya Guadix-Baza, Ladihonda y Fazares) y tres en zonas con mayor precipitación (Cortijuela, Molinillo, Fardes). En esta última, Fardes, son árboles situados en la ladera de un río (presumiblemente poco afectados por años más o menos secos).**
* **Años de climatología diferente: 1995 año muy seco y años 1996 y 1997, años muy lluviosos.**
* **Situación de la hoja: Canopy (copa de los árboles) y Sprouts (rebrotes, hojas nuevas que salen desde la parte inferior del tronco).**

**Disponemos de un total de 2101 casos, cedidos por el Departamento de Ecología de la Universidad de Granada (España), de los que hemos seleccionado aleatoriamente una muestra de tamaño 15 que se presenta en la siguiente tabla:**



**Se pide:**

**a) ¿Se puede admitir que la longitud de las hojas de encina se distribuye normalmente?**

**b) ¿Se puede admitir que la longitud media de las hojas es igual a 30 cm a un nivel de significación del 5%?  (Suponiendo que la varianza es conocida)**

**c) Suponiendo que las asimetrías de las hojas sigan una distribución Normal; comprobar mediante un contraste de hipótesis si existen diferencias significativas en la asimetría de las hojas teniendo en cuenta la situación de la hoja en el árbol.**

**d) A un nivel de significación del 5%, ¿es representativo el ajuste lineal entre la longitud y la asimetría? ¿Cuál sería la expresión del modelo? ¿Cuánto explica el modelo?**

**Solución**

**a) ¿Se puede admitir que la longitud de las hojas de encina se distribuye normalmente?**

Introducimos los datos en RStudio

longitud <- c(26.51,30.17, 34.24, 31.04, 34.99, 30.48, 25.07, 25.04, 29.16,35.12, 25.41, 27.02, 23.04, 27.69, 34.71)



Calculamos la media y la desviación típicamean(longitud)

sd(longitud)



Se contrasta la normalidad mediante el contraste de Kolmogorov-Smirnov

ks.test(longitud, y = pnorm, 29.31267, 4.062451, alternative = "two.sided")



Mediante la prueba de *Kolmogorov-Smirnov* obtenemos que el p-valor es 0.8173, mayor que el nivel de significación 0.05, por lo tanto, no se puede rechazar la hipótesis nula y admitimos que la longitud de las hojas sigue una distribución Normal.

**b) ¿Se puede admitir que la longitud media de las hojas es igual a 30 cm**

Nos piden el siguiente contraste de hipótesis



t.test(longitud, alternative = "two.sided", mu = 30, conf.level = 0.95)



El valor del nivel crítico o p-valor (Sig. (bilateral)) es 0.5229, mayor que el nivel de significación 0.05, por lo que no se rechaza la hipótesis nula y admitimos que la longitud media de las hojas de encina es igual a 30 cm.

**c) Suponiendo que la asimetría de las hojas sigan una distribución Normal; comprobar mediante un contraste de hipótesis si existen diferencias significativas en la asimetría de las hojas teniendo en cuenta la situación de la hoja en el árbol.**

asimetriaCanopy<- c(0.028, 0.010, 0.080, 0.340, 0.087, 0.040, 0.010, 0.021, 0.135, 0.010, 0.094, 0.153)

asimetriaSprouts<- c(0.156, 0.172, 0.077)



Para poder realizar el contraste de igualdad de medias de dos poblaciones independientes es necesario contrastar previamente la igualdad de varianzas.



donde  representa la varianza de la asimetría en el Canopy y  representa la varianza de la asimetría en el Sprouts.

var.test(asimetriaCanopy, asimetriaSprouts, alternative = "two.sided", conf.level = 0.95)



Analizando la información relativa al contraste de hipótesis que se incluye en la salida de var.test, vemos que el valor del estadístico de contraste es 3.4611. La distribución F de Snedecor que sigue el estadístico de contraste tiene 11 grados de libertad en el numerador y 2 en el denominador. El p-valor asociado al contraste es 0.4908. Como este valor es superior al nivel de significación (que para este ejemplo es 0.05), no podemos rechazar la hipótesis nula que hemos planteado. Es decir, se puede considerar que las varianzas de ambas poblaciones son iguales.

Una vez aceptada la igualdad de varianzas, realizamos en contraste de diferencia de medias de dos poblaciones normales independientes.

t.test(asimetriaCanopy, asimetriaSprouts, alternative = "two.sided", mu = 0, var.equal = TRUE)



En la salida se incluye el valor del estadístico de contraste (-0.88477), los grados de libertad de la distribución t de Student que sigue el estadístico de contraste (13) y el p-valor (0.3924). Como el p-valor es mayor que el nivel de significación fijado (0.05), no rechazamos la hipótesis nula y se deduce que las partes de la planta (Canopy y Sprouts) no influyen en la asimetría de las hojas.

**d) A un nivel de significación del 5%, ¿es representativo el ajuste lineal entre la longitud y la asimetría? ¿Cuál sería la expresión del modelo? ¿Cuánto explica el modelo?**

Introducimos los datos de la asimetría en R Studio

longitud <- c(26.51,30.17, 34.24, 31.04, 34.99, 30.48, 25.07, 25.04, 29.16, 35.12, 25.41, 27.02, 23.04, 27.69, 34.71)

asimetría<- c(0.028, 0.010, 0.080, 0.340, 0.087, 0.040, 0.010, 0.021, 0.135, 0.010, 0.094, 0.153, 0.156, 0.172, 0.077)



reg\_lin <- lm(asimetría ~ longitud)

reg\_lin



Por defecto, la salida que muestra la función lm incluye únicamente las estimaciones para los parámetros, en nuestro caso y . Por tanto, el modelo lineal puede escribirse del siguiente modo:

asimetría = 0.119847 - 0.000875 \* longitud

Podemos obtener más información sobre el modelo de regresión que hemos calculado aplicando la función summary al objeto que contiene los datos de la regresión, al cual hemos llamado reg\_lin en este ejemplo.

summary(reg\_lin)



Esta salida contiene una información más completa sobre el análisis. Así, por ejemplo, encontramos información sobre los residuos (en el apartado Residuals), que se definen como la diferencia entre el verdadero valor de la variable dependiente y el valor que pronostica el modelo de regresión. Cuanto más pequeños sean estos residuos mejor será el ajuste del modelo a los datos y más acertadas serán las predicciones que se realicen a partir de dicho modelo.

En la tabla Coefficients encontramos los valores de los parámetros que aparecían en la salida por defecto junto a su error estándar. Cada parámetro aparece acompañado del valor de un estadístico t de Student y un p-valor que sirven para contrastar la significación del parámetro en cuestión, es decir, para resolver los siguientes contrastes de hipótesis:

 

Lo que se pretende mediante estos contrastes es determinar si los efectos de la constante y de la variable independiente son realmente importantes para de explicar la variable dependiente o si, por el contario, pueden considerarse nulos.

En nuestro ejemplo, los p-valores que nos ayudan a resolver estos contrastes son 0.514 y 0.887, ambos mayores que 0.05. Así, considerando un nivel del significación del 5%, no rechazamos la hipótesis nula en ambos contrastes, de manera que podemos suponer ambos parámetros no son significativamente distintos de 0. Por lo tanto que concluimos que longitud no es válida para predecir la asimetría según un modelo lineal.

Por último, en la parte final de la salida, encontramos el valor de R² (Multiple R-squared) y de R² ajustado (Adjusted R-squared), que son indicadores de la bondad del ajuste de nuestro modelo a los datos. R² oscila entre 0 y 1, de manera que, valores de R² próximos a 1 indican un buen ajuste del modelo lineal a los datos. Por otro lado, R² ajustado es similar a R², pero penaliza la introducción en el modelo de variables independientes poco relevantes a la hora de explicar la variable dependiente. Por tanto, R² ajustado <= R². En nuestro ejemplo, R² = 0.001609 y R² ajustado = -0.07519, por lo que podemos concluir que el modelo lineal no se ajusta a nuestros datos. El porcentaje de variación de la asimetría de las hojas de la encina explicado por el modelo de regresión lineal es igual al 0.1 %, siendo ésta una cantidad claramente insatisfactoria.

La última línea de la salida incluye un estadístico F de Snedecor y el p-valor correspondiente que se utilizan para resolver el siguiente contraste:



que se conoce habitualmente como contraste ómnibus. Mediante este contraste se comprueba si, de forma global, el modelo lineal es apropiado para modelizar los datos. En nuestro ejemplo, el p-valor asociado a este contraste 0.8871 es mayor que 0.05 por lo que, al 5% de significación no podemos rechazar la hipótesis nula y afirmar que, efectivamente, el modelo lineal no es adecuado para nuestro conjunto de datos