Modelos de regresión lineal

1. Introducción

La regresión lineal es una medida de correlación que, en términos muy generales, consiste en encontrar una recta entre los puntos de nuestros datos que sea la más representativa de ellos. Esta recta va a ser aquella que se encuentre a la menor distancia posible de cada uno de los puntos. Si tenemos un gráfico como el siguiente, lo que buscamos es una recta que pase lo más cerca posible de cada punto:

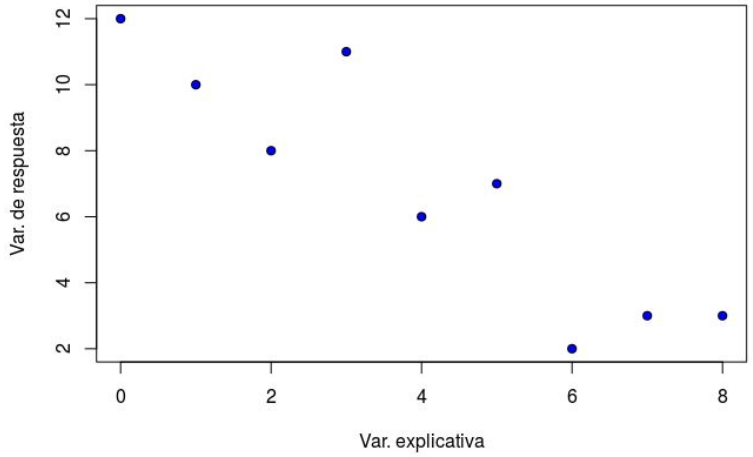

Fig. 1: Gráfico de un set de datos.

Esta recta se encuentra con el método de los mínimos cuadrados. La ecuación que se usa para encontrar la función de esta recta es la siguiente:

$$y = a + bx$$

Donde y es la variable dependiente, a es el intercepto al origen, b es la pendiente de la recta y x es la variable independiente. Ahora bien, para calcular la pendiente necesitamos primero conocer el coeficiente de correlación de Pearson, o sea, *r*. Recuerden que las correlaciones son una forma de medir la dependencia entre variables y *que haya una correlación no implica que haya una relación causal*. Otro dato a tener en cuenta es que para que Pearson nos sea útil, necesitamos que nuestros datos tengan una distribución normal, lo que podemos chequear con un test de Shapiro (para más información sobre la correlación de Pearson, ver el *handout* de estadística descriptiva). La forma de calcular a y b es:

$$b = r \frac{SDy}{SDx}$$

$$a = \overline{y} - b\overline{x}$$

Lo que se logra con el método de los mínimos cuadrados es minimizar la magnitud de la diferencia entre los puntos y la recta, es decir, reducir la distancia entre cada punto de nuestros datos y la recta, de modo que obtengamos la recta más representativa. Esta distancia entre cada punto y la recta se denomina residuo.

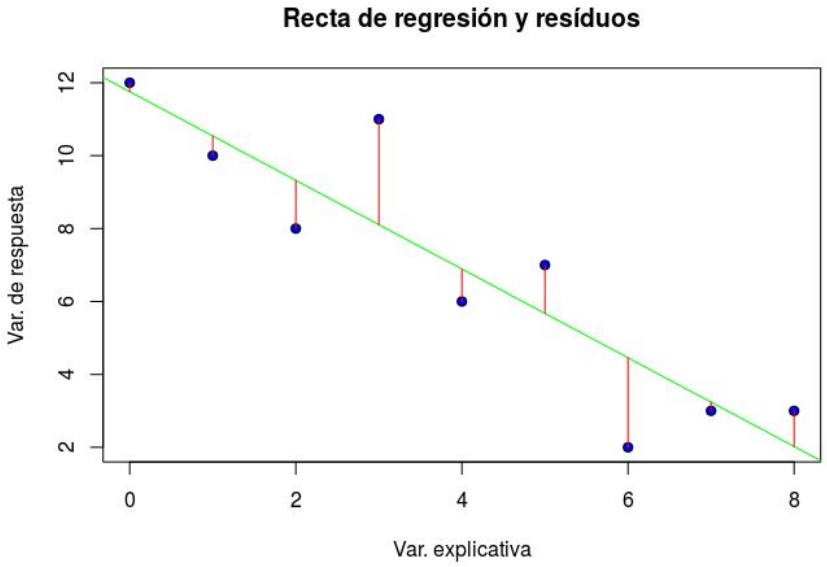

Fig. 2: recta de regresión y residuos de datos de la Fig. 1

Por lo que vimos hasta ahora, podemos concluir que un modelo de regresión nos sirve para predecir el valor de una variable a partir del valor de otra. Se utiliza con variables continuas y ordinales. Hay que tener en cuenta sin embargo que este modelo puede hacer predicciones sin sentido, como dar números negativos cuando no es posible (por ejemplo, en tiempos de respuesta). Es por ello que siempre el lingüista tiene que saber interpretar los datos que tiene y los resultados de los test estadísticos que aplica.

2. Regresión lineal simple

Ahora que ya tenemos una imagen de lo que es una regresión lineal, vamos a entrar en detalle al modelo de regresión lineal simple. Es similar a la función de la recta que ya vimos:

$$y_i = (\beta_0 + \beta_1 x_i) + \varepsilon_i$$

Donde y_i representa el valor de la variable dependiente o respuesta, β_0 es el intercepto al origen (a), β_1 es la pendiente de la recta (b), x_i representa el valor de la variable independiente o explicativa y finalmente ε_i representa el error, que se asume que es 0. En otras palabras, representa la influencia de otra variable que no conocemos en nuestra variable dependiente pero que, en nuestro experimento, nos resulta incontrolable o aleatoria. También se le llama perturbación aleatoria y es una variable ruido blanco, esto implica que se espera que sea nula, que su varianza va a ser constante y su covarianza va a ser nula. En el próximo apartado lo vamos a ver con más detalle.

2.1 Supuestos

Antes de pasar a la implementación en R, vamos a enumerar los supuestos. Como vimos con los tests estadísticos anteriores, siempre que aplicamos un test hay una serie de supuestos que se tienen que cumplir para que nuestros resultados sean confiables.

- Linealidad
  La fórmula de nuestro modelo debe ser lineal. Dicho de otro modo, la variable respuesta o dependiente se relaciona de manera lineal con la(s) variable(s) predictora(s) o independiente(s).
  En un gráfico, esto lo vemos como una distribución de los puntos de manera lineal alrededor de la recta:

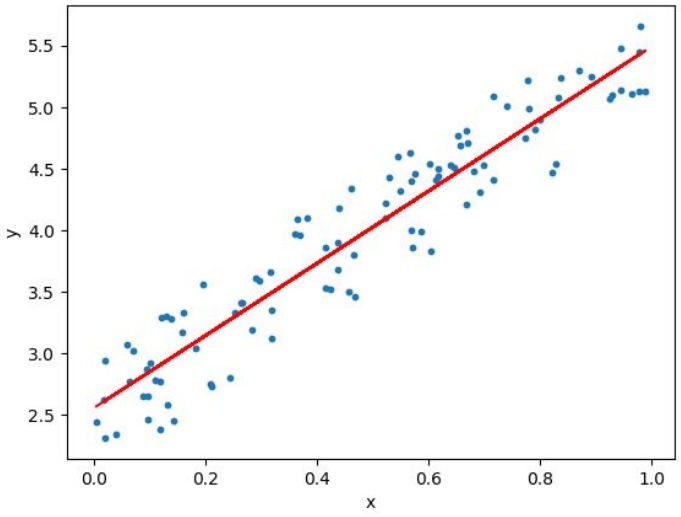

Fig. 3: Datos que cumplen el supuesto de linealidad.

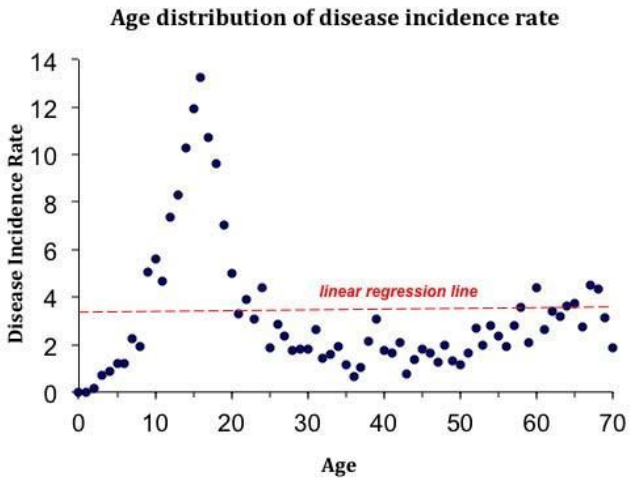

Fig. 4: Datos que no cumplen el supuesto de linealidad.

- Homocedasticidad
  Este supuesto estipula que las varianzas deben ser constantes. Es decir, la distribución los puntos debe ser homogénea alrededor de la recta de regresión. Si giramos un gráfico de manera que la recta de regresión quede horizontal lo vemos más fácilmente:

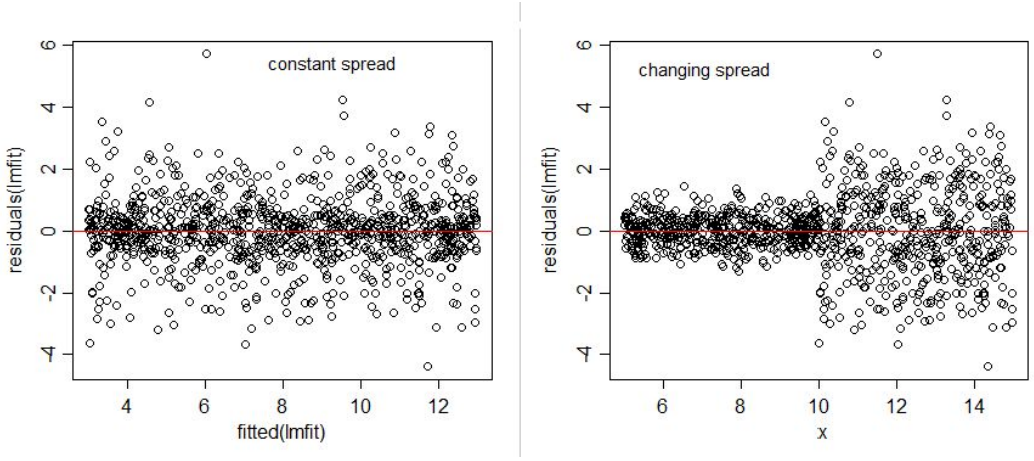

Fig. 5: izq: datos homocedásticos (varianzas constantes), der: datos heterocedásticos (varianzas no constantes).

- Independencia o no colinealidad
  Cuando existe más de un predictor es muy importante que sean independientes entre sí, es decir, que uno no prediga al otro. Por ahora sólo vimos un modelo con un predictor, pero es un supuesto importante para seguir avanzando.
  Por ejemplo, si queremos predecir la inteligencia de una persona en función de la fluencia verbal y para medir la fluencia utilizamos medidas como sílabas por segundo y palabras por segundo, tenemos dos variables predictoras colineales porque están relacionadas entre sí. Otro caso en el que no se cumple este supuesto sería tener varias respuestas por sujeto en nuestro set de datos.
  La independencia de los predictores es algo que debe chequearse desde el diseño experimental. Si no se cumple este supuesto nuestros resultados no van a ser confiables.

- Normalidad de los residuos
  El último supuesto es que la distribución de los residuos es normal. Sin embargo hay quienes sostienen que la regresión lineal es bastante robusta cuando no se cumple este supuesto. Para chequear la normalidad de los residuos podemos hacer un histograma con los residuos o un Q-Q plot:

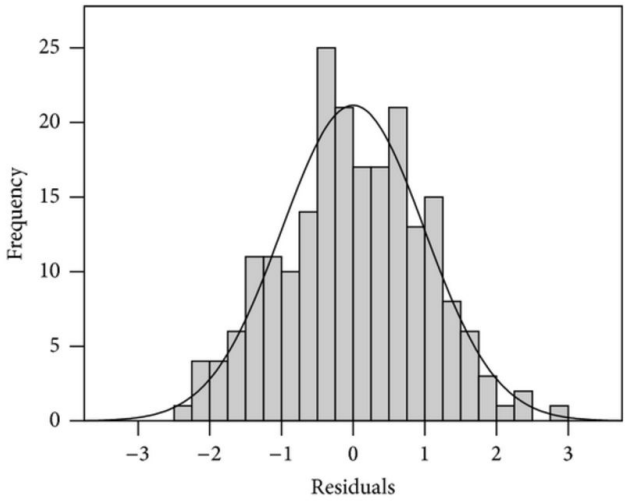

Fig. 6: Histograma de residuos con distribución normal.

Este último punto implica que chequeamos los supuestos después de tener un modelo lineal, pues los residuos sólo pueden ser verificados cuando tenemos un modelo a partir del cual computarlos.

2.2 Regresión lineal simple en R

Vamos a ver cómo implementar esto en R (script adjunto). En nuestro ejemplo vamos a ver si en un dataset hay una correlación en la frecuencia escrita de una palabra en función del promedio de la frecuencia de bigramas en esa palabra. Primero chequeamos normalidad con Shapiro y linealidad con Pearson. Luego, graficamos la recta de regresión y obtenemos el siguiente gráfico:

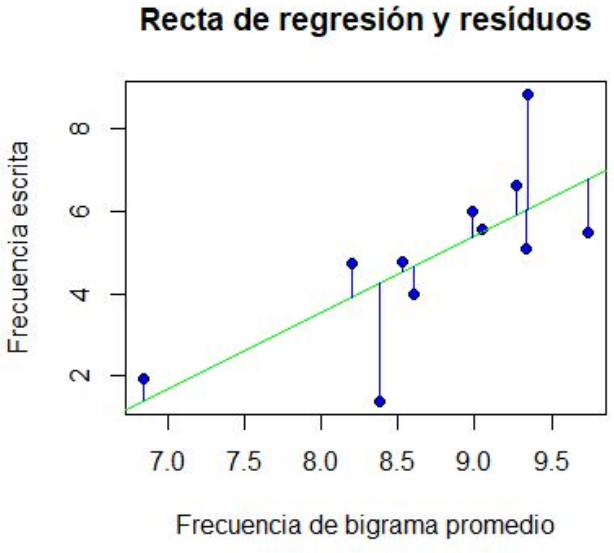

Fig. 7: Recta de regresión y residuos.

Usamos la función lm(dataset$VD ~ dataset$VI) para analizar los coeficientes. Otra forma de escribirlo es lm(VD~VI, data=dataset). Si usamos la función summary(), obtenemos el siguiente output (recordemos que siempre tenemos que cumplir los supuestos), abajo se indica cómo leerlo:

```
Call:
lm(formula = EnglishLR$WrittenFrequency ~
EnglishLR$MeanBigramFrequency)

Residuals:
    Min      1Q  Median      3Q     Max
-2.8837 -0.8033  0.2575  0.6668  2.7951

Coefficients:
                            Estimate Std. Error t value Pr(>|t|)
(Intercept)                 -11.2472     5.3718  -2.094   0.0658 .
EnglishLR$MeanBigramFrequency   1.8513     0.6115   3.028   0.0143 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.53 on 9 degrees of freedom
Multiple R-squared:  0.5046,   Adjusted R-squared:  0.4495
F-statistic: 9.166 on 1 and 9 DF,  p-value: 0.0143
```

Fig. 8: Output de lm()

Call: nos muestra el modelo que nosotros pusimos.

Residuals: es una especie summary de los residuos. Recordemos que los residuos son la diferencia entre cada punto y su lugar estimado por la recta de regresión. Esperamos que los números sean pequeños.

Coefficients: es una de las partes más importantes del output. Vamos a tener un Intercept y luego una fila por cada variable independiente. En este caso, como es una regresión lineal simple, tenemos sólo uno. Intercept puede definirse como lo que queda de lado cuando se promedia la variable dependiente y la independiente. Para cada coeficiente tenemos varios tipos de información:

- Estimate: es el peso dado a la variable. Dicho mal y pronto, nos muestra cuánto va a variar nuestra variable dependiente cuando la independiente varíe en 1.
- Std. Error: nos dice qué tan preciso es nuestro Estimate.
- t value: se usa para ver si el coeficiente es significativamente distinto a 0.
- Pr(>|t |): nos muestra el nivel de significancia.

Residual standard error: es el SD de los residuos, buscamos que el número sea pequeño.

Multiple/ Adjusted R-squared: como tenemos solo una variable independiente la diferencia no nos importa por ahora. Lo que nos indica es cuánto de la varianza es explicado por el modelo. En el Adjusted lo que hace es tomar en cuenta el número de variables independientes, y nos va a servir para la regresión múltiple.

F-statistic: nos dice si el peso de al menos una de las variables (en este caso sólo tenemos una) es significativamente distinto a 0. Si el p-value nos da mayor a 0.05 significa que nuestro modelo no está haciendo nada realmente.

Como ya pudieron ver, nuestro output dio bastante bien, porque está armado para eso. Sin embargo es muy importante que sepamos que con la cantidad de datos que tenemos en este ejemplo los resultados no son muy confiables.

Nota: para ver cómo interpretar el output en una regresión lineal donde la variable independiente no es numérica, ver Gries (2013) páginas 264 a 276.

3. Regresión polinomial

La regresión polinomial es un caso especial de regresión lineal en el que se reemplaza el modelo lineal por una función polinomial. Esto nos permite dar un ajuste no lineal a los datos. Esta fórmula se obtiene al elevar cada uno de los predictores originales a una potencia, es decir, agregando un parámetro:

$$Lineal: \ y = a + bx$$

$$Polinomial: y = a + bx \ + cx^2 + dx^3 ...$$

Este modelo sigue siendo lineal porque a pesar de que los datos de x estén elevados al cuadrado o al cubo, el coeficiente a no lo está. Las funciones polinomiales pueden presentar una gran variedad de formas:

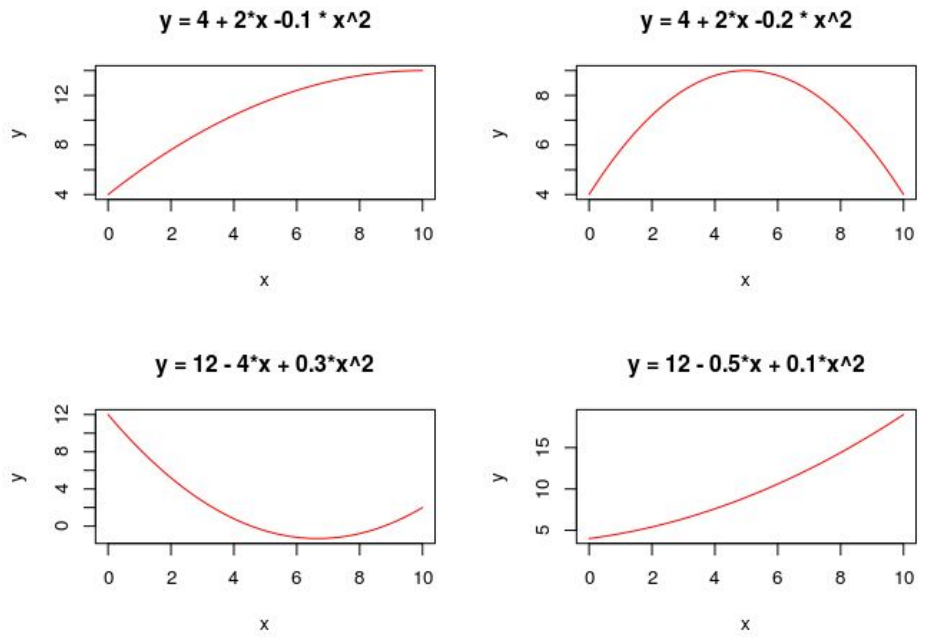

Fig. 9: ejemplos de regresión polinomial con una variable.

En la figura 9 vemos modelos cuadráticos con 3 parámetros (el intercepto, pendiente de x y pendiente de x^2), pero como se ve, incluso al añadir una sola variable elevada al cuadrado, podemos obtener distintos tipos de curvas, que pueden ajustarse a diversos tipos de datos. Si

quisiéramos pasar el modelo del ejemplo de la frecuencia (`EnglishLRCo= lm(EnglishLR$WrittenFrequency ~ EnglishLR$MeanBigramFrequency`) a uno polinomial, la forma de implementarlo sería: `EnglishLRCo= lm(EnglishLR$WrittenFrequency ~ EnglishLR$MeanBigramFrequency+ I(EnglishLR$MeanBigramFrequency^2))`. De todos modos, en este caso no es necesario ya que nuestros datos son claramente lineales.

Si bien la regresión polinomial puede ser de mucha ayuda para casos en los que los datos no son estrictamente lineales, tiene algunos problemas que es importante tener en cuenta:

- Si se agregan variables, es decir, cuanto más aumenta la complejidad de la fórmula, se hace más difícil de manejar.
- La regresión polinomial tiende a ajustarse drásticamente (overfitting), por lo que hay que tener cuidado al usar polinomios de mayor grado (lo que reduce el error) porque puede ser excesivo, lo que conlleva a una pérdida del poder explicativo del modelo. Hay que asegurarse de que la curva se ajuste bien a los datos del problema que estamos tratando.

Para saber si las predicciones que hacen dos modelos distintos tienen una diferencia significativa, podemos compararlos mediante el uso de una ANOVA. También podemos hacer uso del criterio de información de Akaike (AIC), que se utiliza como medida de la información perdida al usar un modelo para aproximar la realidad. Por ende, un menor valor de AIC significa un mejor ajuste del modelo a los datos. Sin embargo, hay que ser cuidadosos porque, como ya lo dijimos anteriormente, un modelo demasiado ajustado no es deseable.

4. Regresión múltiple

Un modelo de regresión lineal múltiple es un modelo de regresión lineal en el que en lugar de tener una sola variable independiente o predictora tenemos dos o más. Es por ello que es un método multifactorial. El modelo de regresión múltiple tiene la siguiente forma:

$$y_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_n x_n + \varepsilon_i$$

Como ven, ya no hay una sola x .. β_0 es el valor de y cuando todos los coeficientes son iguales a cero. Cada uno de los otros coeficientes es lo que se le agrega al intercepto cuando ese predictor aumenta en una unidad, y todos los otros coeficientes se mantienen al nivel del intercepto.

Graficar una regresión múltiple se hace más difícil ya que tenemos que agregar una dimensión por cada predictor.

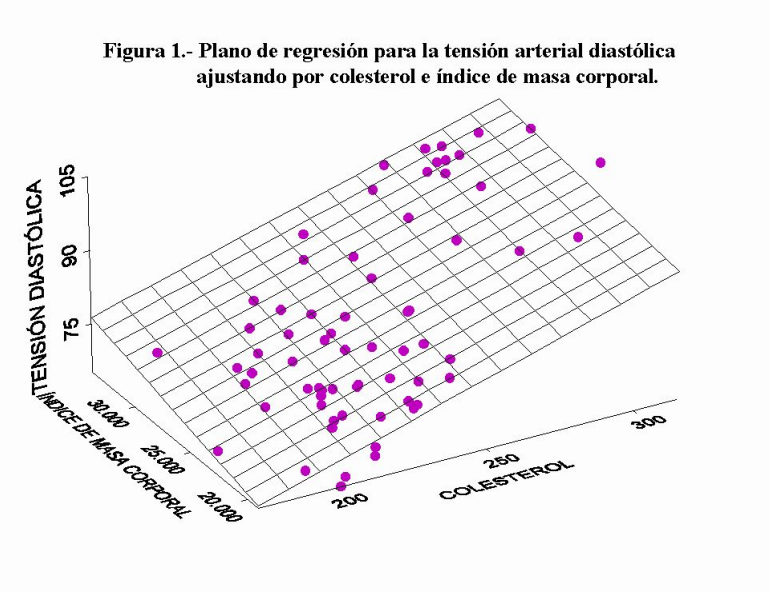

Fig. 10: ejemplo de visualización en tres dimensiones

Si tenemos más de dos predictores, ya no es tan fácil. Lo que se hace generalmente es presentar un histograma para cada variable:

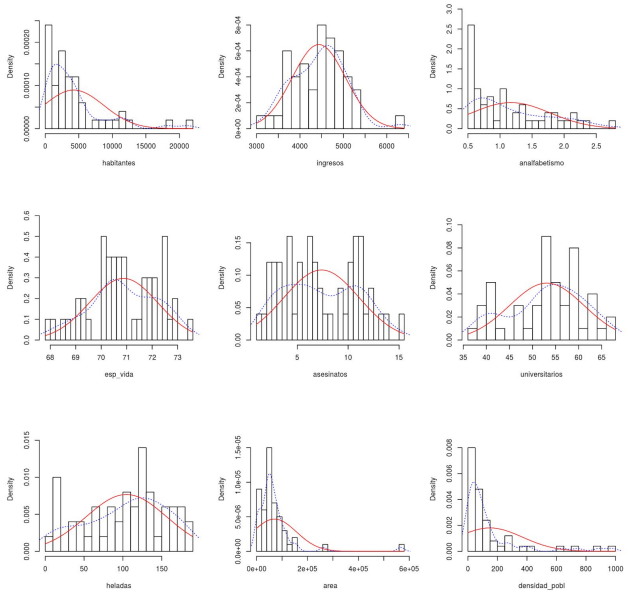

Fig. 11: ejemplo de visualización en múltiples histogramas

La regresión lineal múltiple nos sirve tanto para predecir los valores de la variable dependiente como para evaluar la influencia de los predictores en ella.

4.1 Supuestos

En la regresión múltiple, además de los supuestos que ya vimos, se agregan otros:

4.1.1.Colinealidad entre predictores

No debe haber colinealidad entre las variables independientes. Una forma de evaluar si existe colinealidad entre dos variables es con el factor de inflación de la varianza (VIF):

$$VIF\beta_i = \frac{1}{1 - R^2}$$

Donde R^2 se obtiene de la regresión de cada predictor sobre los otros. Los valores de referencia son:

- VIF=1: ausencia de colinealidad.
- 1<VIF<5:colinealidad puede estar afectando.
- 5<VIF<10: muy posiblemente haya problemas con la colinealidad.

Hay que tener en cuenta que hay otros modos de calcular esto, y los valores de referencia son los que se usan generalmente.

4.1. 2. *Parsimonia*

No califica exactamente como supuesto pero es muy importante tenerlo en cuenta a la hora de elegir un modelo. Lo que postula la parsimonia es que el mejor modelo será aquel que pueda explicar con mayor precisión la variabilidad de la variable independiente con menos predictores.

4.2 Interacciones

Cuando tenemos más de una variable independiente, se pueden dar varios escenarios:

- Las variables independientes pueden ser aditivas o independientes. En este caso, el efecto que tienen ambas es igual a la suma que tiene cada una por separado. Este sería el caso que venimos viendo hasta ahora.
- Las variables independientes también pueden interactuar. Dos variables interactúan si el efecto de ambas sobre la variable dependiente no es predecible a partir de los efectos individuales de cada una sobre la variable dependiente. En estos casos se necesita un término de interacción que "corrija" la predicción. En otras palabras, hay que agregar un nuevo predictor a nuestro modelo que sea la interacción de las dos variables. Importante: no confundir interacción con colinealidad.

En un modelo con dos predictores si queremos incluir la interacción lo hacemos de la siguiente manera:

$$y_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \varepsilon_i$$

También se puede hacer cuando tenemos más predictores, el método es agregar un nuevo coeficiente por cada interacción (las dos x que le siguen son las variables que interactúan). Una forma de saber si dos variables interactúan es graficando con la función

`interaction.plot(VI1, VI2, VD)`.Si tenemos líneas paralelas no hay interacción, si no son paralelas es porque interactúan:

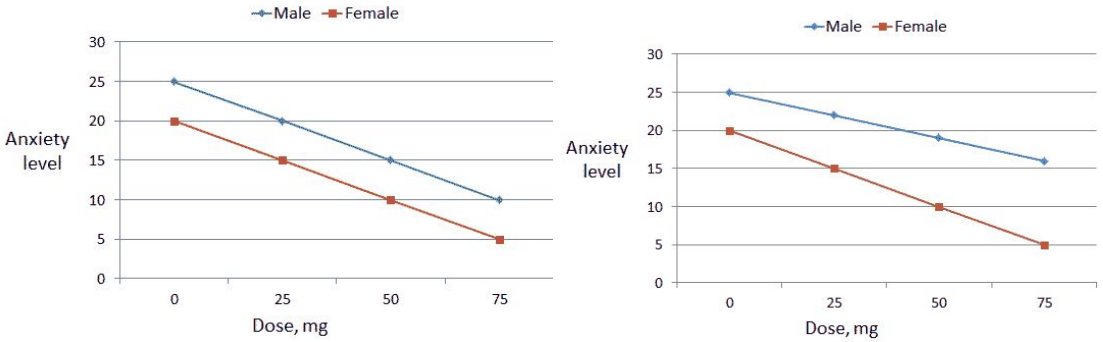

Fig. 11: izq sin interacción, der con interacción.

5. Referencias

Gries, S. (2013) *Statistics for Linguistics with R, a practical introduction*. Walter de Gruyter GmbH, Berlin/Boston. Capítulo 5.

Winter, B. (2016) L*inear models and linear mixed effects models in R: Tutorial 1 and 2*. University of California, Merced, Cognitive and Information Sciences. California, Estados Unidos.

Recursos online:

http://www.ccg.unam.mx/~vinuesa/R4biosciences/docs/Tema9_regresion.html#diagnostico-del-modelo-de-regresion-y-validacion-de-supuestos

http://www.learnbymarketing.com/tutorials/linear-regression-in-r/

https://rpubs.com/Joaquin_AR/226291

https://rpubs.com/Joaquin_AR/254575

https://www.econometrics-with-r.org/8-3-interactions-between-independent-variables.html