Taller de Consultoria - TP1

Gonzalo Barrera Borla

8/25/2019

Setup

```
library(fitdistrplus) # ajuste exploratorio de distribuciones
library(tidyverse) # manipulación de datos en general, graficos
library(broom) # limpieza y estructuracion de resultados de regresiones
```

Problema 1

> Se dan las duraciones (medidas en ciclos hasta la ruptura) de una muestra de rodamientos ("rulemanes"). Describir las características principales de la muestra (posición, dispersión, asimetría), y buscar una distribución adecuada.

Los funcionales de locación más habituales son la media μ y la mediana η, digamos. Reportamos sus estimadores puntuales muestrales, $\hat{\mu} = \bar{x} = n^{-1} \sum_i x_i$ y $\hat{\eta} = x^{(\frac{n}{2})}$ (donde $x^{(i)}$ denota el i-ésimo elemento de la muestra ordenada). Para la dispersión, es razonable usar el la raíz cuadrada del estimador puntual insesgado de la varianza, $s^2 = (n-1)^{-1} \sum_i (x_i - \bar{x})^2$. Para la asimetría γ, construimos el estimador b reemplazando en la definición de γ a cada momento por su estimador insesgado:

$$\gamma = E\left[\left(\frac{X-\mu}{\sigma}\right)^3\right] \quad \Rightarrow \quad \hat{\gamma} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{s}\right)^3$$

media	mediana	dispersión	asimetría
72.387	61.68	38.363	0.847

Comparamos nuestra funcion de asimetría con la implementación de un paquete bien conocido de R para comprobar su coherencia:

```
casi_iguales <- function(x, y, tol=1e-6) { abs(x-y) <= tol }
stopifnot(casi_iguales(
  e1071::skewness(df1$duracion),
  asimetria(df1$duracion)))
```

La distribución más habitual para modelar la vida media de individuos/componentes es la Weibull, que generaliza la bien conocida distribución exponencial para considerar tasas de fallo no necesariamente constantes en el tiempo. Se dice que $X \sim \text{Weibull}(k, \lambda)$ con parámetros de *forma* $k > 0$ y *escala lambda* > 0 si la densidad de X está dada por:

$$f(x; \lambda, k) = \begin{cases} \frac{k}{\lambda}\left(\frac{x}{\lambda}\right)^{k-1} e^{-(x/\lambda)^k} & x \geq 0, \\ 0 & x < 0, \end{cases}$$

Nótese que cuando $k = 1$, la tasa de fallos es constante, y $X \sim \text{Exp}(\lambda^{-1})$.

El estimador de máxima verosimilitud de λ dado k (referencia) es

$$\widehat{\lambda^k} = \frac{1}{n}\sum_{i=1}^{n} x_i^k$$

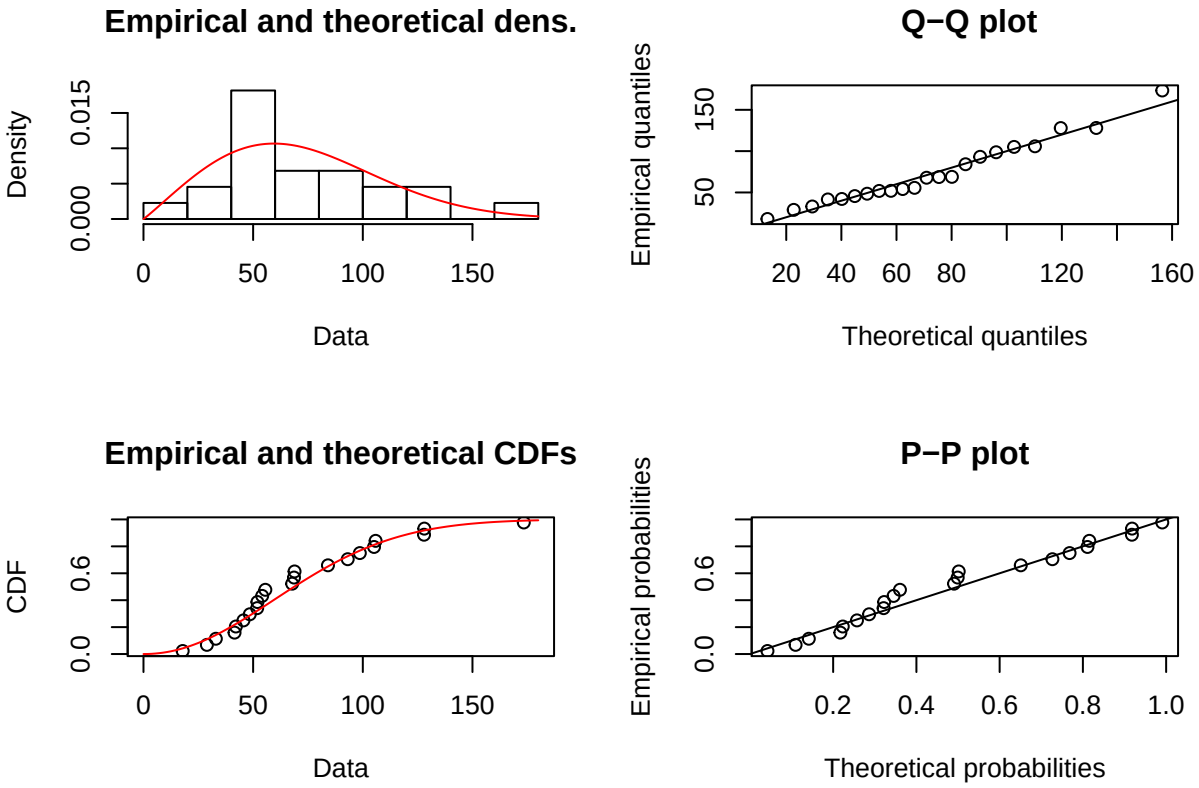
Mientras el el EMV de k es la solución para k de la siguiente ecuación, que ha de ser encontrado numéricamente:

$$0 = \frac{\sum_{i=1}^{n} x_i^k \ln x_i}{\sum_{i=1}^{n} x_i^k} - \frac{1}{k} - \frac{1}{n}\sum_{i=1}^{n} \ln x_i$$

A continuación, implementamos la búsqueda antes descrita y la comparamos con el resultado de una implementación estándar, `fitdistrplus::fitdist`.

```
ajustar_weibull <- function(x, rango_forma = c(0, 10)) {
  eq_k <- function(x, k){ sum((x^k) * log(x)) / sum(x^k) -1/k -mean(log(x)) }
  # Busco numéricamente el k que minimiza la distancia a 0 de eq_k
  k <- optimise(function(k){abs(eq_k(x, k))}, interval = rango_forma)$minimum
  lambda <- mean(x^k)^(k^-1)
  return(c(forma = k, escala = lambda))
}
```

La implementación propia es casi idéntica a la estándar con una diferencia del -0.0072% para el parámetro de forma y 0.002 para el de escala. ¡Nada mal! A continuación, aprovechamos los gráficos por defecto del objeto `fitdist`, en particular el plot cuantial-cuantil ("Q-Q") para convencernos de que el ajuste es decente:



En conclusión, postulamos que las duraciones X_i de los $i = 1, \ldots, n$ rodamientos están distribuidas con $X_i \overset{iid}{\sim} \text{Weibull}(k = 2.063, \lambda = 82.079)$.

Problema 2

> Se dan: el punto de ebullición del agua (PE) (en grados Fahrenheit) y la presión atmosférica (PA) (en pulgadas de mercurio), medidos a distintas alturas en los Alpes. Plantear un modelo que describa cómo varía PE en función de PA. ¿Con cuánta precisión se puede estimar PE en función de PA?. Comentar cualquier característica de los datos.

Aparentemente confiables fuentes en la internet apuntan a que un modelo físico del punto de ebullición del agua con la presión atmosférica se puede derivar de la "relación de Clausius-Clapeyron" (referencia). En particular, a bajas temperaturas (id est, a temperaturas por debajo del punto crítico, 647 K para el agua) y asumiendo que la entalpía de vaporización L del agua permanece constante), dados dos puntos (P_1, T_1), (P_2, T_2) en la curva de coexistencia entre agua y vapor, se cumple que:

$$\ln \frac{P_2}{P_1} = -\frac{L}{R}\left(\frac{1}{T_2} - \frac{1}{T_1}\right)$$

donde R es la constante de los gases ideales. Reorganizando un poco los términos obtenemos

$$\begin{aligned} PE^{-1} &= \frac{R \cdot \ln P_0}{L} + \frac{1}{T_0} - \frac{R}{L} \ln PA \\ PE^{-1} &= b + m \ln PA, \text{ digamos} \end{aligned}$$

Es decir que existe una relación lineal entre la inversa del punto de ebullición y el logaritmo natural de la presión atmosférica a la que fue medido, donde

$R = 8,3145 \text{ J K}^{-1} \text{ mol}^{-1}$	es la constante de los gases ideales
$P_0 = 101.325 \text{ Pa}$	es la presión atmosférica estándar
$T_0 = 373,15 \text{ K}$	es el punto de ebullición del agua a P_0
$L = 40.608 \text{ J mol}^{-1}$	es la entalpía de vaporización del agua

Y los coeficientes correspondientes deberían ser

$b = \frac{R \cdot \ln P_0}{L} + \frac{1}{T_0}$	$\approx 0.00504 \text{ K}^{-1}$
$m = \frac{-R}{L}$	$\approx -2.048 \times 10^{-4} \text{ K}^{-1}$

Este modelo provee una forma determinística de estimar PE en función de PA, para lo cual ni siquiera hace falta realizar regresión alguna sobre los datos. Al estimar por regresión lineal los coeficientes b, m, lo que estaremos haciendo no es derivar de cero una relación entre PE, PA, sino: - (a) determinar si los datos recolectados proveen evidencia para el modelo propuesto, o menos arrogantemente, si asumimos correcto el modelo termodinámico, - (b) evaluar si la recolección de los datos fue fidedigna.

Al graficar los datos en forma "cruda", se observa que:

- los datos "crudos" parecen seguir una tendencia lineal, y
- la doceava observación es la única evidentemente por fuera de dicha tendencia.

A continuación entonces, compararemos 4 modelos, surgidos a partir de ajustar (a) un modelo lineal "ingenuo" y (b) el modelo "físico" antedicho, usando (i) todas las observaciones y (ii) todas salvo la #12. Para ello, transformamos las unidadas al estándar internacional (grados Kelvin para temperaturas y Pascales para la presión):

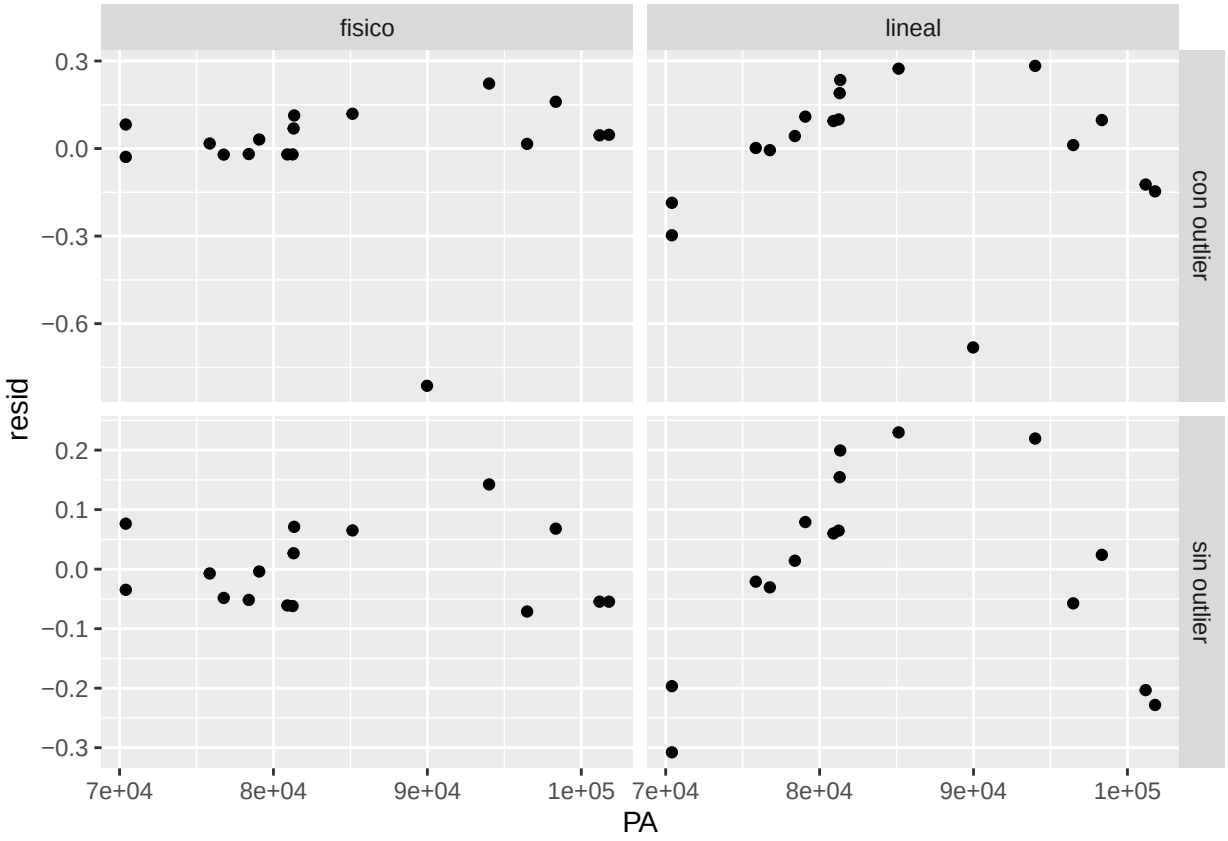
```
inHg_a_Pa <- function(x) {x*3386.39}
fahrenheit_a_kelvin <- function(x) { (x - 32)*5/9 + 273.15 }
```

nombre_modelo	outliers	R2_ajustado	p_valor
lineal	con outlier	0.9941	2.528e-18
lineal	sin outlier	0.9973	1.359e-19
fisico	con outlier	0.9950	7.384e-19
fisico	sin outlier	0.9996	3.441e-25

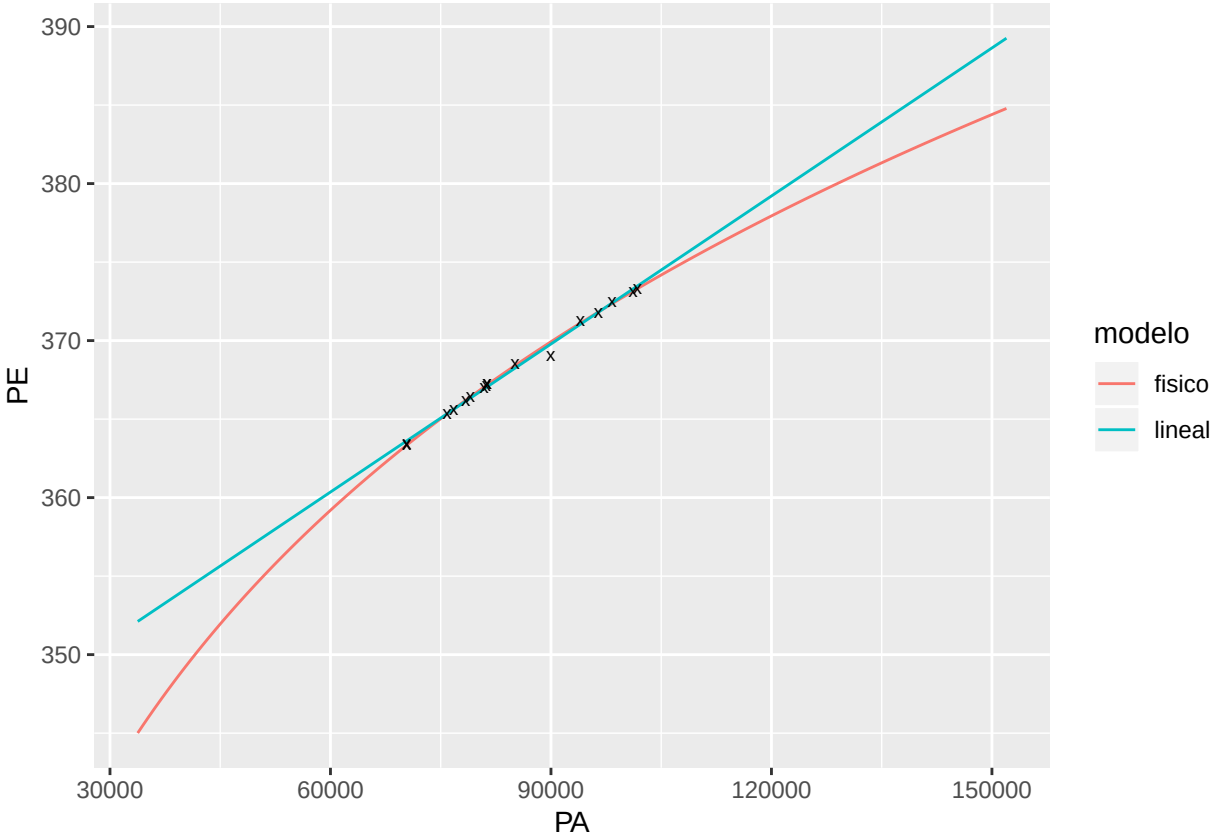
Tanto R^2_{adj} como el p-valor del test de significación global de la regresión son medidas adimensionales de la calidad de ajuste de un modelo, así que es razonable compararlas aún cuando los modelos ajustados tienen distintas unidades. Se observa que

- tanto el modelo lineal como el físico proveen una muy buena estimación de los valores de PE en función de PA,
- el modelo físico ajusta mejor que el lineal, se incluya o no la observación 12, y
- el ajuste del modelo físico mejora espectacularmente al quitarla, mientras que el modelo lineal mejora marginalmente.

Así considerado, podríamos decir que al menos *sobre el soporte de los datos* tanto el modelo lineal como el físico proveen una aproximaci'on decente del fenómeno, y si no se requiere demasiada exactitud o se vive en 1940 y el poder de cómputo escasea, se puede usar el modelo lineal. Graficando los residuos, sin embargo, se observa claramente que los correspondientes al modelo lineal tienen una clara estructura, se incluya la observación 12 o no. Por su parte, en el modelo físico, salvo por el *outlier* #12, los residuos no presentan estructura alguna. En ambos casos, eliminar el *outlier* "centra" mejor los residuos alrededor del 0, sin eliminar la tendencia de los mismos.



Al extrapolar los modelos por fuera del soporte de los datos, es evidente por qué la aproximación lineal fue tan buena: ¡los datos fueron tomados exactamente en el rango en que ambos modelos dan predicciones similares!



Por último, comparemos las constantes calculadas teóricamente con las estimaciones empíricas:

$$b \approx 0.00504 \text{ K}^{-1}, \qquad m \approx -2.048 \times 10^{-4} \text{ K}^{-1}$$
$$\hat{b} \approx 0.004976 \text{ K}^{-1}, \qquad \hat{m} \approx -1.992 \times 10^{-4} \text{ K}^{-1}$$

es decir que el ajuste fue casi perfecto. El modelo final, expresado en grados Kelvin y Pascales, será

$$PE = (0.00504 - 2.048 \times 10^{-4} \ln PA)^{-1}$$

Problema 3

> Se investiga el efecto de la presión aplicada durante la manufactura del papel, en el "factor de ruptura" (la fuerza necesaria para desgarrarlo). Bajo cada valor de la presión P, se manufacturó un lote de papel; de cada lote se eligieron 4 hojas, a cada una de las cuales se midió el factor de ruptura R. Se desea predecir R en función de P.

Aprovechando que tenemos mediciones repetidas para cada uno de los 5 valores de P, comparamos la raíz cuadrada del estimador global de la varianza $s_0 = 7.47$, con la de los estimadores de la varianza para cada valor de P:

P	s
35.0	3.30

P	s
49.5	7.97
70.0	7.72
99.0	4.80
140.0	2.63

Aún con pocos datos, se intuye que la varianza en R no es la misma para todo P. Pareciera haber *heterocedasticidad*, pero la relación entre P y la varianza de R no es lineal: s es máximo para presiones "medias" de fabricación.

Asumiendo que las mediciones de cada par (P, R) son independientes entre sí, la matriz de covarianzas será diagonal, y en vez de utilizar $\Sigma = \sigma^2 \mathrm{I}_n$, podemos considerar una matriz $\Sigma = diag(\sigma_1^2, ..., \sigma_n^2)$, y estimar las varianzas de cada observación, con el estimador insesgado de la varianza para cada nivel de presión antes calculado.

A continuación ajustamos ambos modelos, considerando (a) observaciones iid en general, y matriz de covarianza $\sigma^2 \mathrm{I}_n$, y (b) observaciones iid *en cada nivel de P*, con $\Sigma = diag(\sigma_1^2, ..., \sigma_n^2)$. Este último modelo equivale a realizar un ajuste de mínimos cuadrados pesados, con pesos $w_i = \sigma_i^{-2}$

modelo	adj.r.squared	p.value
ordinario	0.4692775	0.0005159
pesado	0.7563785	0.0000004

Aunque ambos modelos son buenos, el p-valor para la regresión global del segundo modelo es más de 3 órdenes de magnitud más pequeño. La evidencia parece justificar el uso de una regresión pesada. El modelo final quedará

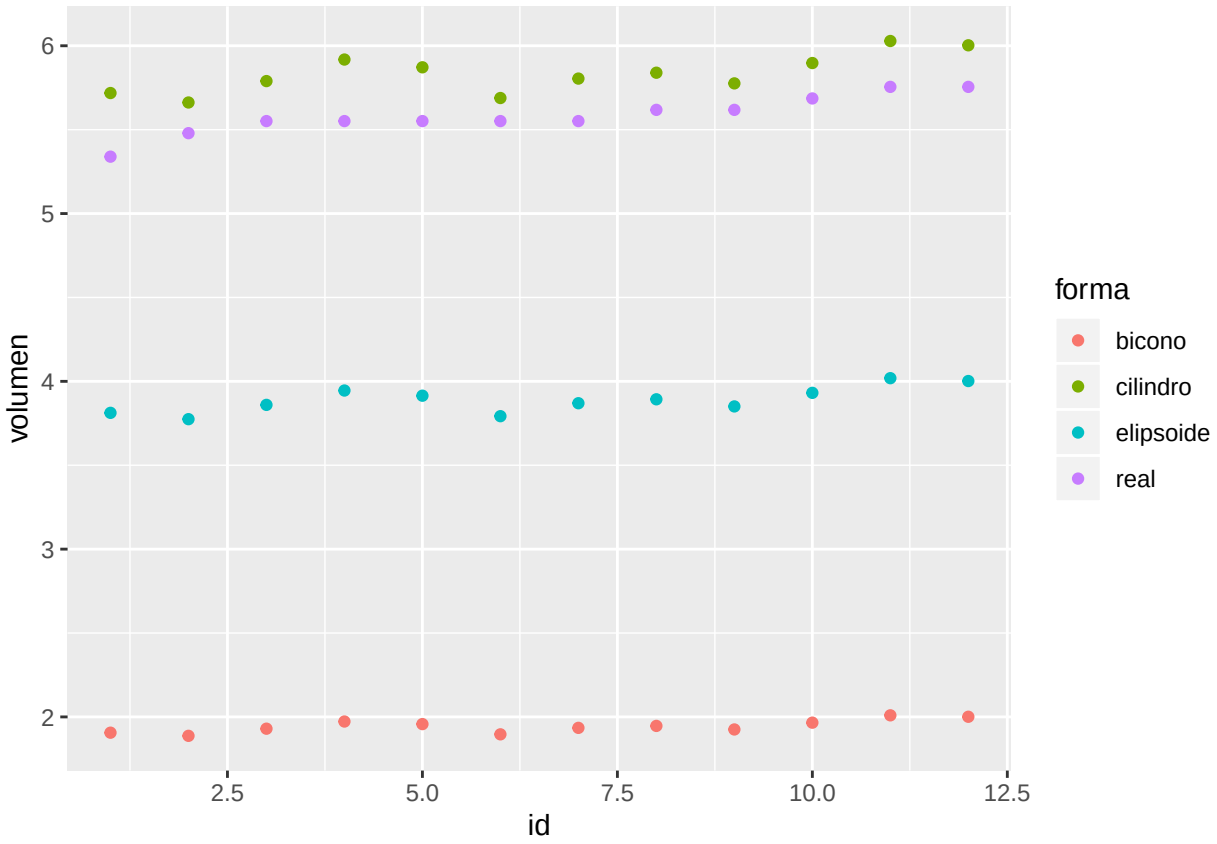
$$R = 119.47 + -0.14 atm^{-1} \times P$$

Problema 4

> La siguiente tabla da, para 12 huevos de gallina, la longitud L (o sea, el mayor diámetro), la mayor sección circular (el mayor diámetro perpendicular a L), ambas en pulgadas; y el volumen V. Interesa predecir V en función de L y M.

Preston (1973) (link) hace un tratamiento bastante exhaustivo de cómo calcular el volumen de un huevo, que a continuación resumimos.

Supongamos que el mayor diámetro de un huevo es M, y su largo es $L = a + b$, donde a, b son las dos partes en que se divide el largo a la altura del máximo diámetro. El volumen de todo huevo está acotado superiormente por un ciindro perfecto de diámetro M, y por debajo por un bicono de máximo diámetro M y alturas a, b. Si el huevo fuese cilíndrico, su volumen será $\frac{\pi}{4} M^2 L$, y si fuese bicónico, $\frac{\pi}{12} M^2 L$. En un escenario más realista e intermedio, en que el huevo está formado por dos medios elipsoides, su volumen es $\frac{\pi}{6} M^2 L$. Nótese que en ningún caso la asimetría (*id est*, cuán lejos de $L/2$ están a y b) hace diferencia alguna, pero sí es clave saber la forma dominante (bicono, elipsoide, cilindro). Los huevos de colibrí son más bien romos, casi cilíndricos, mientras que los de zampullín son casi bicónicos. En el siguiente gráfico, exhibimos los posibles volúmenes de cada huevo según la suposición de forma:



El formato más razonable parece ser un cilindro, así que si el volumen del huevo de gallina está dado por Íla fórmula $V = kM^2L$, $k \approx \pi/4$. Sin embargo, una simple regresión lineal sobre M y L que incluya un término cuadrático sobre M ya tiene un error cuadrático medio mucho menor que nuestro modelo de huevo cilíndrico:

ecm_lm	ecm_cil
0.0043	0.0674

Una forma directa de mejorar el modelo, es usar una regresión lineal *sin ordenada*, sobre una covariable "sintética", $V = k \times (M^2 \cdot L)$ y estimar empíricamente k.

modelo	adj.r.squared	p.value
polinomico	0.5164	0.0319
fisico	0.9998	0.0000

¡Y cómo mejora! Evidentemente, con un R^2_{adj} tan cercano a 1, algo debemos haber hecho bien. Es interesante notar que si comparásemos los dos modelos según su error cuadrático medio, ,el "polinómico" ingenuo da 0.0043 y el "físico" basado en una teoría real sobre la forma de los huevos de aves, da 0.0051. Las predicciones del modelo polinómico tienen menor error cuadrático medio, pero el modelo físico es tanto más parsimonioso (ajusta 1 sólo parámetro en lugar de 4), que termina siendo ampliamente preferible. Así, concluimos que el mejor modelo para predecir V es $V = 0.752 \times M^2 \times L$.