Estadística [continuación]

Bondad de un ajuste (I)

- Volvamos una vez más al ejemplo de la regresión lineal con likelihood en donde tenemos:

Modelo lineal con M parámetros:

$$y = x^T \theta$$

Probabilidad gaussiana:

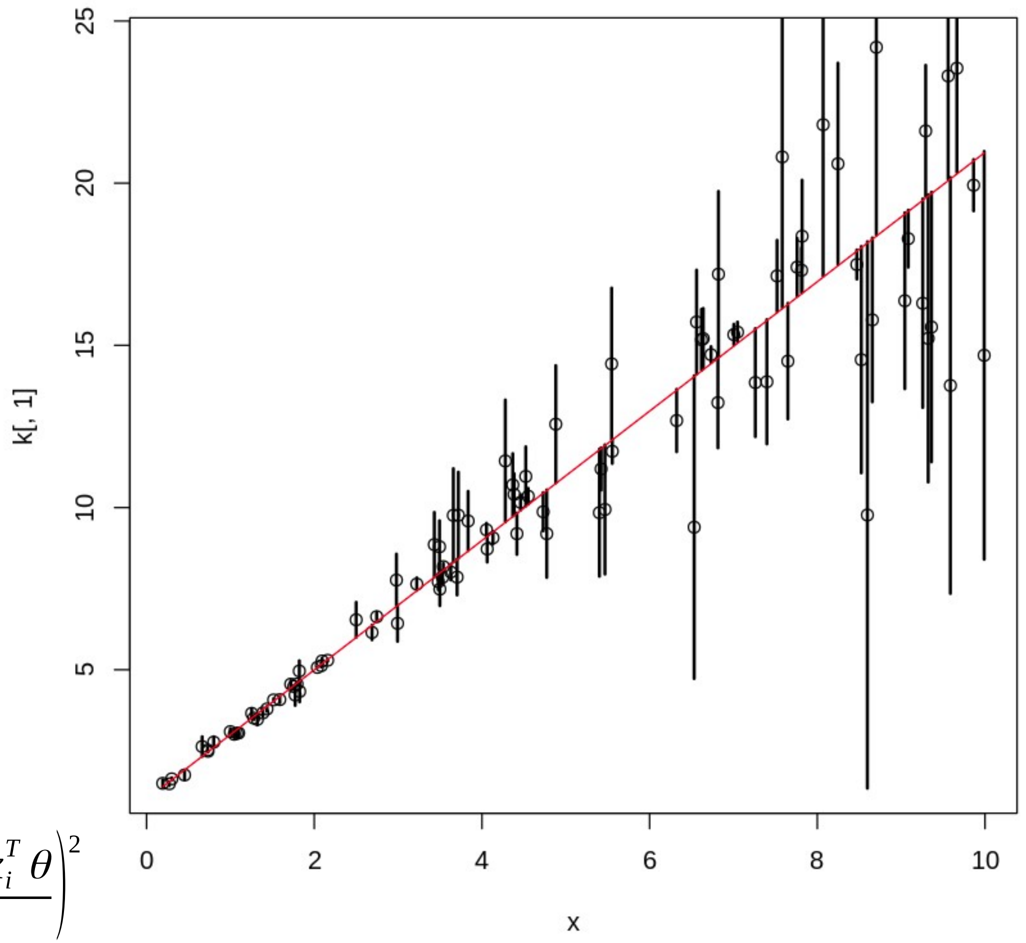
$$p(y_i|x_i) = \frac{1}{\sqrt{2\pi}\sigma_i} e^{\frac{-1}{2}\frac{(y_i - x_i^T\theta)^2}{\sigma_i^2}}$$

Likelihood:

$$L(\theta;(x_i, y_i)) = \prod_i \frac{1}{\sqrt{2\pi}\sigma_i} e^{\frac{-1}{2}\frac{(y_i - x_i^T\theta)^2}{\sigma_i^2}}$$

q = -2 log(L):

$$q(\theta;(x_i, y_i)) = 2\log(\sqrt{2\pi})\sum_i \log(\sigma_i) + \sum_i \left(\frac{y_i - x_i^T\theta}{\sigma_i}\right)^2$$



Bondad de un ajuste (II)

- Supongamos ahora que minimizamos la función q y encontramos el valor de θ_{min}
- Si nuestro modelo lineal es correcto entonces esta variable debería distribuirse como gauss(0, 1)

$$z_i = \frac{y_i - x_i^T \theta_{min}}{\sigma_i}$$

- En efecto, si hemos asumido que la probabilidad de cada y_i era:

$$p(y_i|x_i) = \frac{1}{\sqrt{2\pi}\sigma_i} e^{\frac{-1}{2}\frac{(y_i - x_i^T \theta_{min})^2}{\sigma_i^2}} \Rightarrow p(z_i|x_i) = \frac{1}{\sqrt{2\pi}} e^{\frac{-z_i^2}{2}}$$

- Y de esta forma tenemos que la función que aparece en la expresión de q es:

$$\sum_i \left(\frac{y_i - x_i^T \theta}{\sigma_i} \right)^2 = \sum_i z_i^2$$

- Es decir, esa cantidad es la suma al cuadrado de un conjunto de variables distribuidas normalmente.

Bondad de un ajuste (III)

- Vamos a llamar a esa cantidad de la siguiente forma "chi cuadrado":

$$\chi^2 = \sum_i^N \left(\frac{y_i - x_i^T \theta_{min}}{\sigma_i} \right)^2 = \sum_i^N z_i^2$$

- La distribución estadística de esta cantidad (su pdf) es una función conocida.

$$pdf(\chi^2) = \frac{1}{2^{\nu/2} \Gamma(\nu/2)} e^{-\chi^2/2} (\chi^2)^{(\nu/2-1)}$$

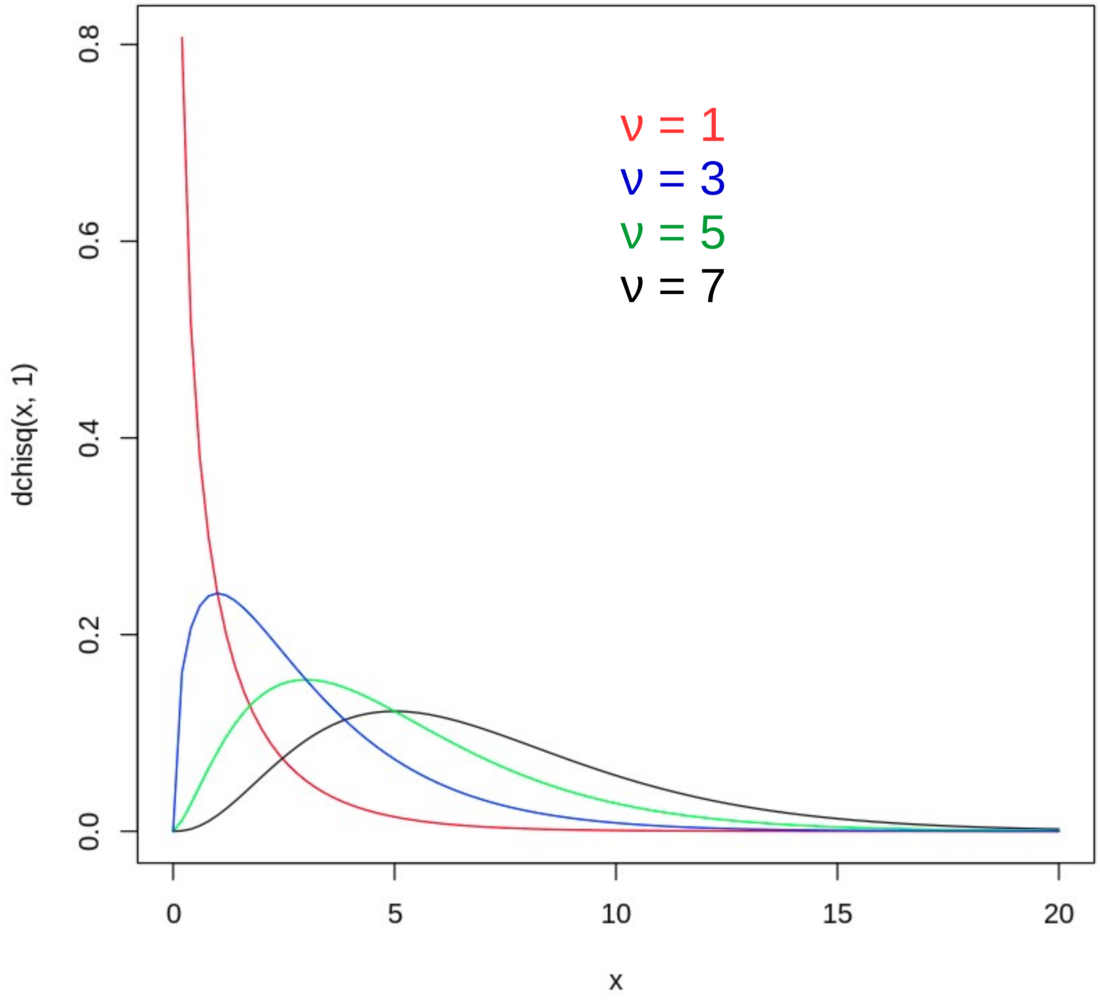
- En donde ν es el "número de grados de libertad" → número de puntos menos número de parámetros.

$$\nu = N - M$$

- Es decir que si el modelo es correcto y nuestras pdf son gaussianas conocemos la distribución de χ^2

Distribución chi²

- La función χ^2 tiene como valor esperado (media) ν, y su varianza es 2ν.

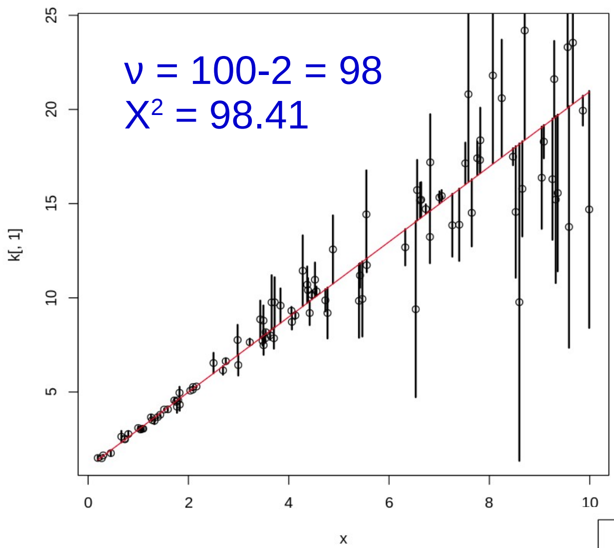


Test de hipótesis

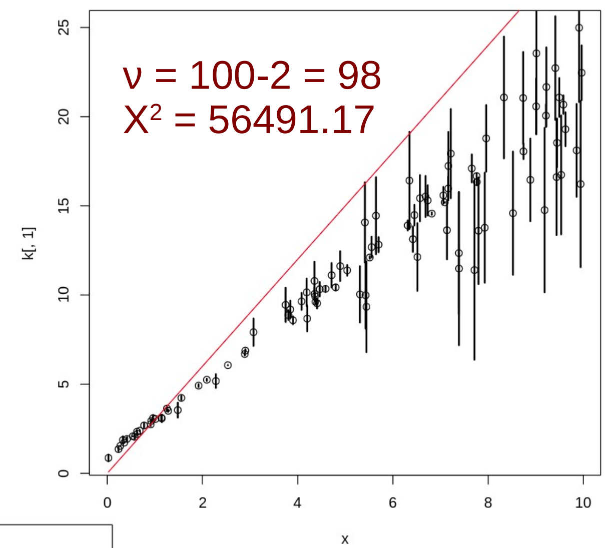
- ¿Cómo podemos utilizar esta distribución para asignar la bondad de un ajuste?
- Podemos ejecutar lo que se conoce como un Test de Hipótesis en el que tenemos dos hipótesis.
  - H0 → Hipótesis Cero → Nuestro modelo lineal verdaderamente describe los datos.
  - H1 → Hipótesis Uno → Nuestro modelo lineal NO describe los datos (H0 es falsa)
- ¿Qué consecuencias tiene esto?
  - Si H0 es cierta sabemos que nuestro chi2 se distribuye con la función pdf(chi2) que hemos visto.
  - Si H1 es cierta no sabemos cómo se distribuye chi2. En general será diferente a pdf(chi2).
- ¿Y cómo podemos usar esto para determinar la bondad de un ajuste?
  - Si el valor que obtenemos para chi2 es un valor "normal" dentro de pdf(chi2) podríamos estar dispuestos a creernos que el ajuste es bueno, y que efectivamente H0 es cierto.
  - Si por el contrario obtenemos un valor chi2 totalmente "anormal", tenderemos a rechazar H0 y a pensar que seguramente chi2 viene dado por una distribución diferente.

Test de hipótesis. Ejemplo

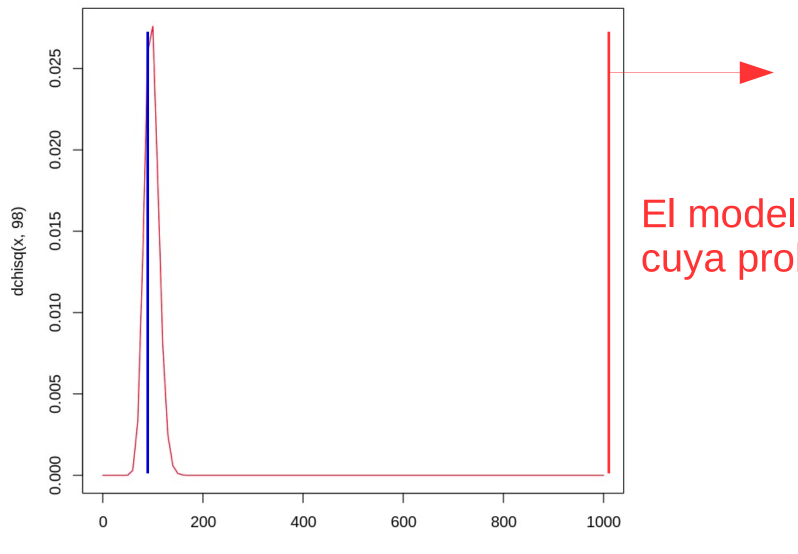
Modelo correcto



Modelo incorrecto



El modelo correcto nos da un chi2 cuya probabilidad es razonablemente alta.

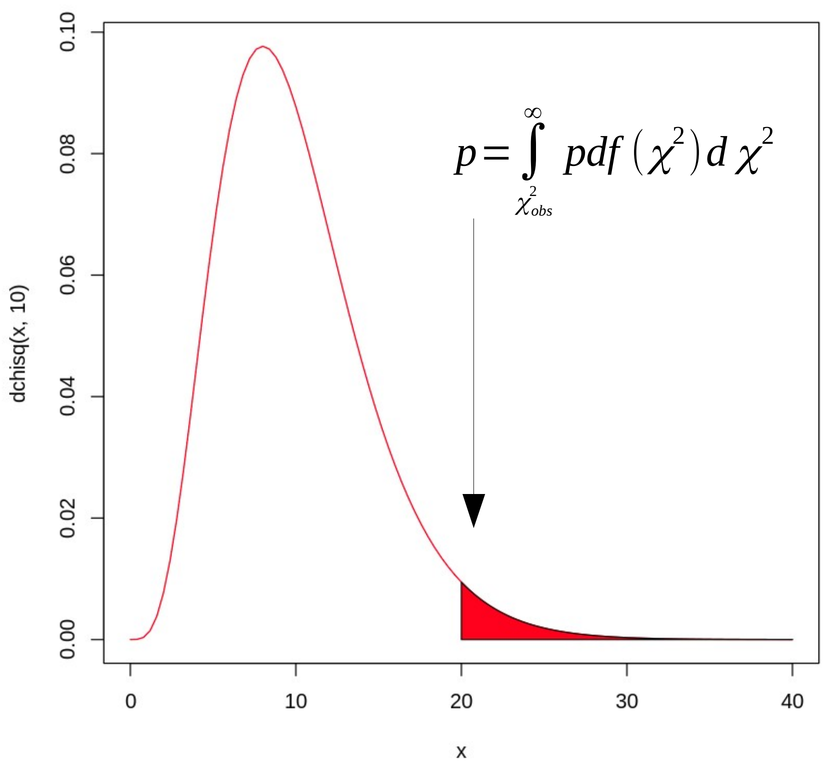


El modelo incorrecto nos da un chi2 cuya probabilidad es cero.

Conclusión: tendemos a pensar que el primer ajuste es bueno y el segundo es malo

Test de hipótesis. Cuantificación

- Este sistema resulta útil pero todavía deja una duda: ¿a partir de qué “chi2” decimos que es bueno?
- Esta pregunta no tiene respuesta: somos nosotros los que establecemos ese criterio.
- Para ello nos servimos del concepto de “p-value”:
- ¿Cuál es la probabilidad de que yo obtenga un valor de chi2 como el observado o mayor?



Llamemos al p-value observado “alpha”

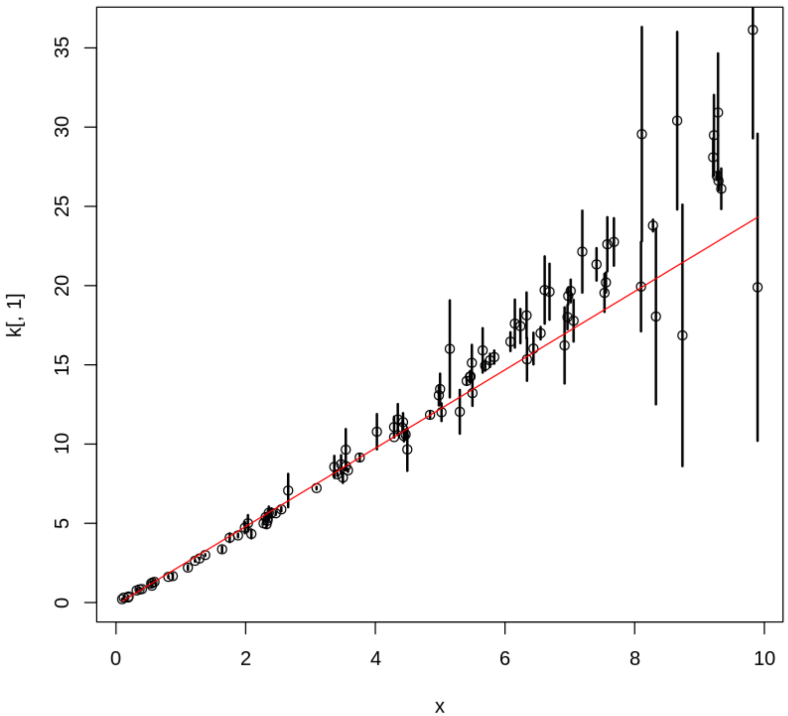
Si alpha es pequeño tendemos a rechazar el modelo, porque siendo cierta H0 sólo en un 100 * α% de las veces hubiésemos obtenido un valor así o mayor.

Solemos decir que rechazamos el modelo con una confianza del 100 * (1-α)%.

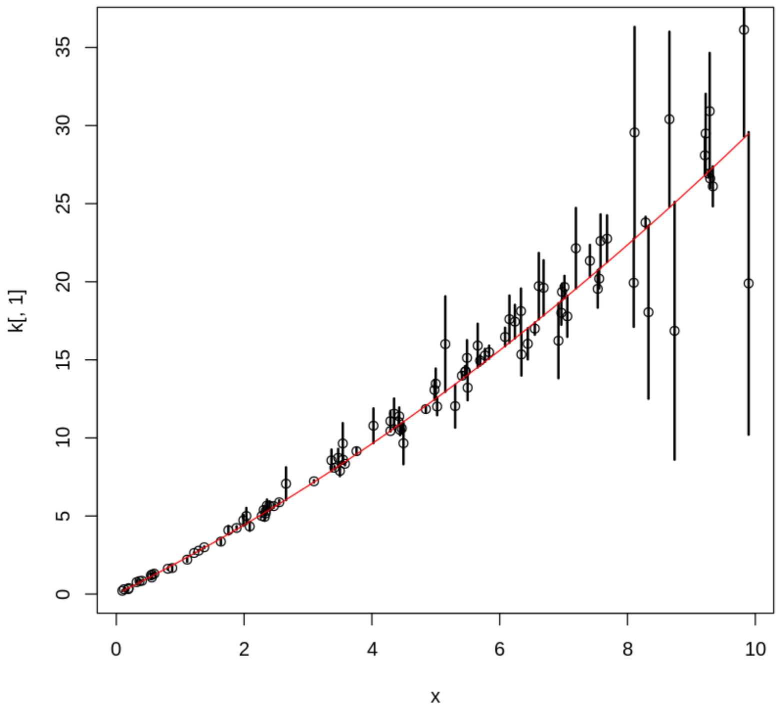
Test de hipótesis. Comparación de modelos.

- Supongamos ahora que disponemos de 2 modelos diferentes y queremos saber cuál ajusta mejor.
- Ejemplo: hemos generado datos que siguen una ley cuadrática y hemos ajustado con 2 modelos:

$Y = a + bx$



$Y = a + bx + cx^2$



Test de hipótesis. Comparación de modelos. (II)

- Por lo tanto vamos a trabajar con dos hipótesis diferentes:
  - H0 → Asume que el modelo más sencillo y = a + b x es el correcto.
  - H1 → Asume que el modelo más complejo y = a + b x + c x^2 es el correcto.
- Consideremos ahora la siguiente cantidad conocida como el likelihood ratio:

$$\frac{L(H_{0,}(x_i, y_i))}{L(H_{1,}(x_i, y_i))} = \frac{L((a,b)_{max},(x_i, y_i))}{L((a,b,c)_{max},(x_i, y_i))}$$

- El símbolo “max” implica que son los valores de los parámetros que maximizan cada likelihood.
- Podemos considerar también la cantidad siguiente:

$$q = -2\log\left(\frac{L(H_{0,}(x_i, y_i))}{L(H_{1,}(x_i, y_i))}\right) = -2\log\left(\frac{L((a,b)_{max},(x_i, y_i))}{L((a,b,c)_{max},(x_i, y_i))}\right)$$

Comparación de modelos. Teorema de Wilk.

- Supongamos que para unos mismos datos tenemos dos hipótesis y modelos:
  - H0 → Asume que el modelo más sencillo con θ_i con i=1,2,..., N parámetros es el correcto.
  - H1 → Asume que el modelo más complejo con θ_i con i=1,2,..., N, N+1,...,M parámetros es el correcto.
- Si H0 es cierta y el número de datos es grande entonces esta cantidad:

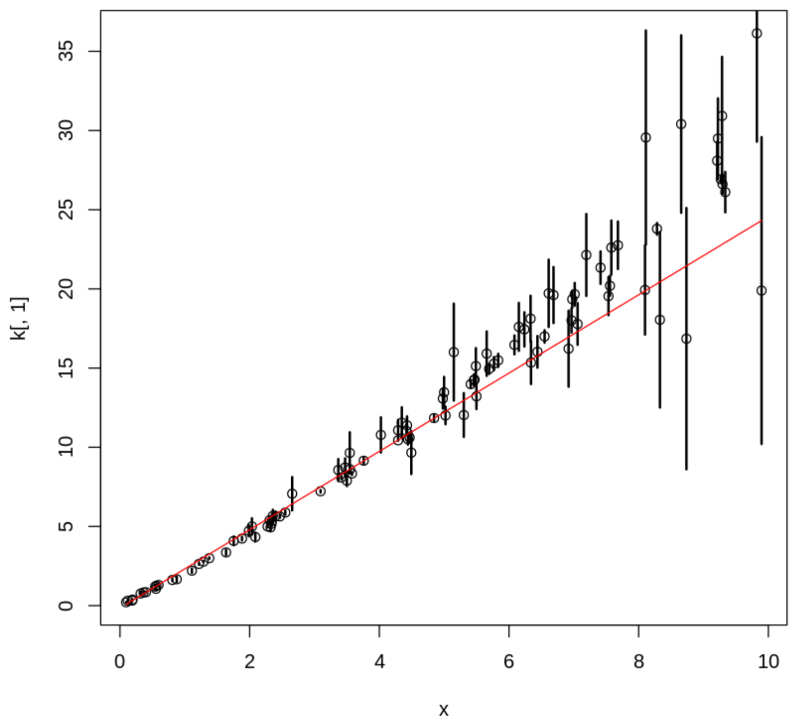
$$q=-2\log\left(\frac{L(H_0,(x_i,y_i))}{L(H_1,(x_i,y_i))}\right)=-2\log\left(\frac{max\, L(\theta_{i=1,\ldots,N},(x_i,y_i))}{max\, L(\theta_{i=1,\ldots,N,N+1,\ldots,M},(x_i,y_i))}\right)$$

- Se distribuye como una función chi2 con M – N grados de libertad.
- Si el valor que obtenemos para q es muy atípico en esta distribución podemos descartar H0.
- Podemos asignar un p-value y un nivel de confianza de la misma forma que se hizo previamente.

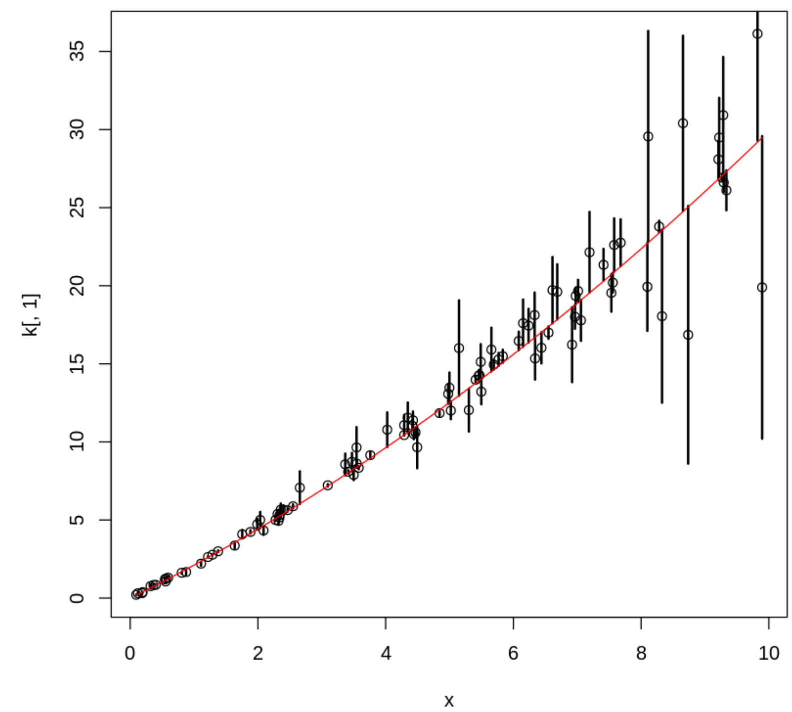
Comparación de modelos. Teorema de Wilk. Ejemplo.

- Volvamos al ejemplo anterior, en donde la diferencia en el número de grados de libertad es 1:

$Y = a + bx$



$Y = a + bx + cx^2$

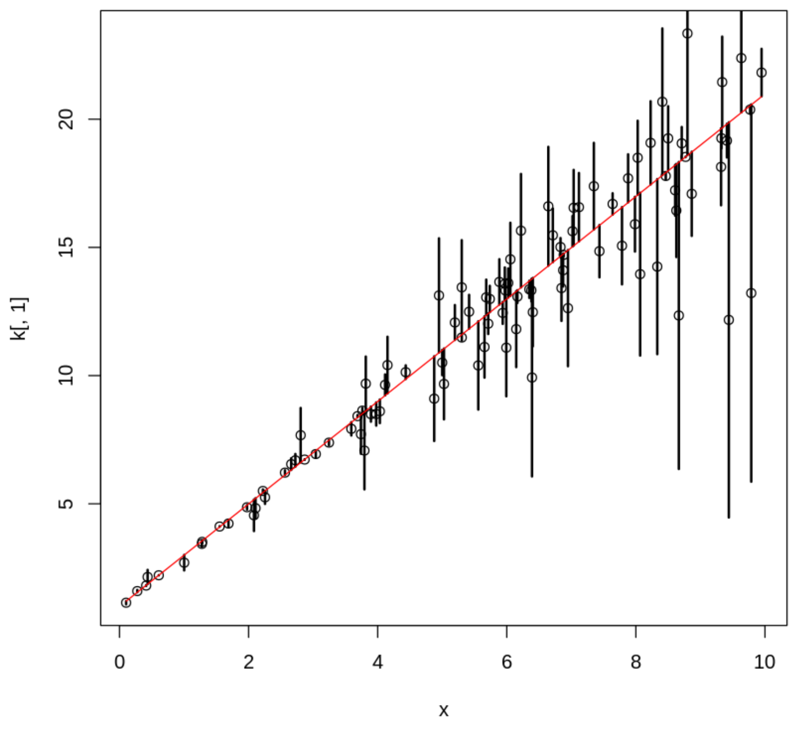


- El valor de q es 8262.254 → la probabilidad de sacar ese valor o mayor es ~ 0 → descartamos H0.

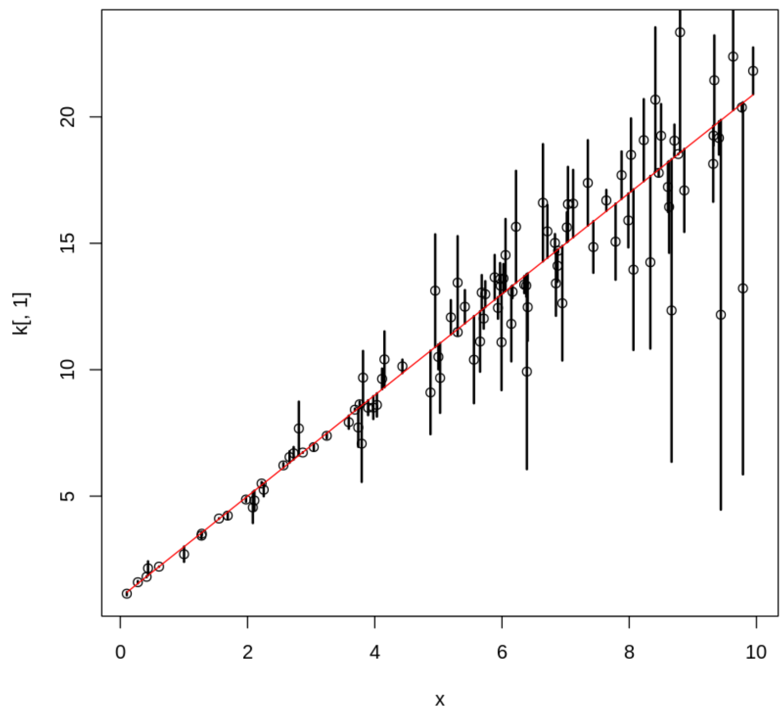
Comparación de modelos. Teorema de Wilk. Ejemplo.

- Ahora pongamos un modelo verdaderamente lineal (H0 es cierta):

$Y = a + bx$



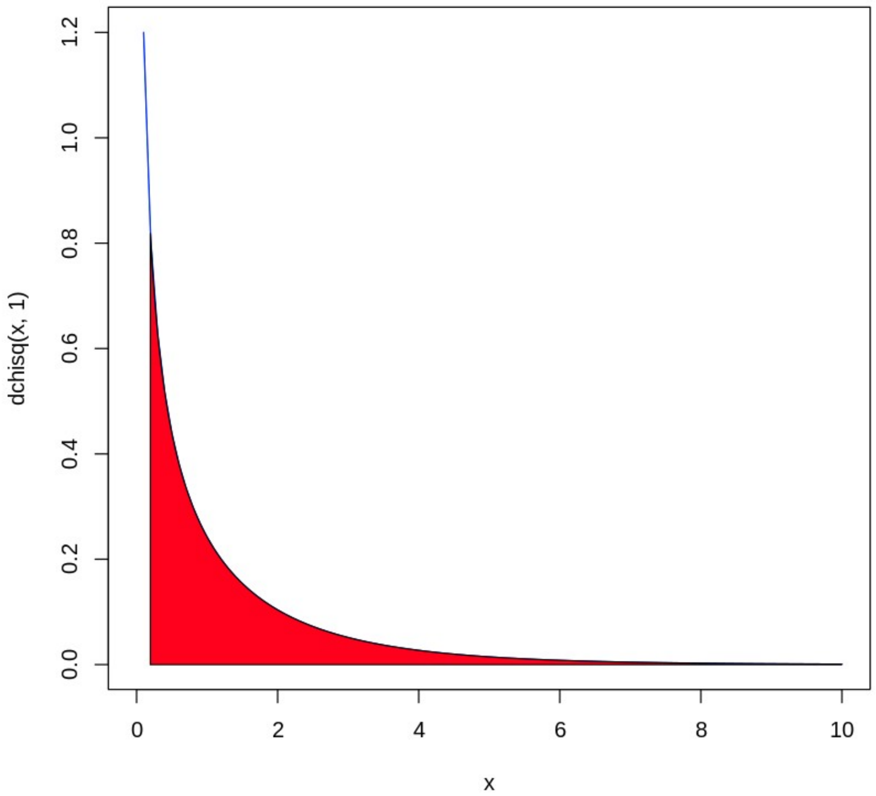
$Y = a + bx + cx^2$



- El valor de q es 0.1955284 → la probabilidad de sacar ese valor o mayor es ~ 0.93.

Comparación de modelos. Teorema de Wilk. Ejemplo.

- Veamos qué posición ocupa este valor en la distribución de chi2 para 1 grado de libertad.



- Se trata de un valor muy probable y por lo tanto aceptamos H0 con un 93% confidence level.

Ejercicio 7

1) Escribe una función que tenga como input un vector x con valores distribuidos uniformemente, y unos parámetros a, b, m y n; y que devuelva como output una matriz cuya primera columna sea $y = a + b\,x$ más un término estocástico sacado de una gaussiana con sigma = $c + d*x^2$; y la segunda columna el error sigma = $m + n*x^2$.
2) Escribe otra función que haga lo mismo que la anterior pero con un parámetro más "c" de tal forma que haga lo mismo pero con un modelo $y = a + b\,x + c\,x^2$.
3) Usando como valores $a = 1$, $b = 2$, $m = 0.1$ y $n = 0.04$ para el modelo de la primera función: calcula los parámetros para las que el likelihood es máximo asumiendo un modelo con dos parámetros $y = a + b\,x$. Calcula el valor del chi2 y calcula el nivel de confianza con el que rechazaríamos este ajuste.
4) Repite el paso anterior con el mismo modelo pero usando la segunda función.
5) Usando la primera de las funciones anteriores con los valores dados en 3) calcula el valor del likelihood ratio "q" para un modelo con 2 y 3 parámetros. ¿Con qué confidence level podemos aceptar H0? Comenta el resultado: ¿qué modelo es mejor el primero o el segundo?.