Laboratorio 1
Turno C

Clase 4

Mediciones indirectas II

Regresiones – Ajuste por cuadrados mínimo

(24/04/2021)

A veces necesitamos determinar una magnitud y a partir de un experimento como función de otra magnitud x

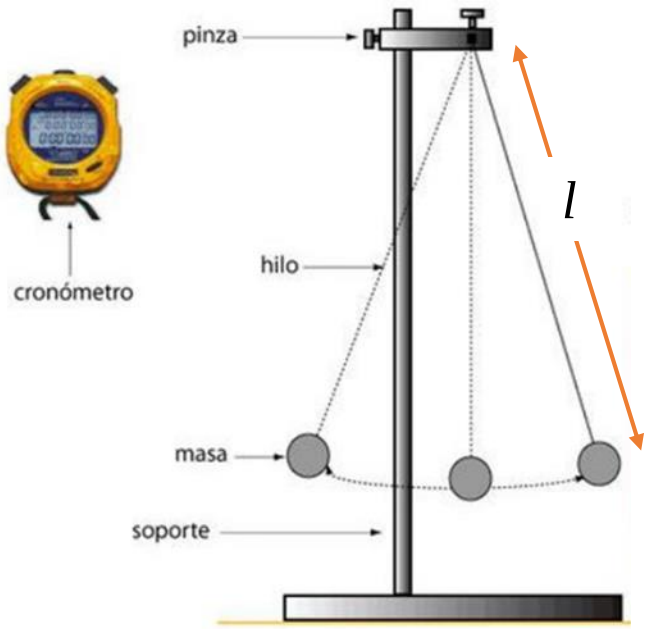
En vez de hacer la medición varias veces de la magnitud x → Se hacen N mediciones del par (x_i , y_i) i = 1,2,…,N

Objetivo : Encontrar la función (modelo) $y = f(x)$

Supongamos que la función $y = f(x)$ que describe los pares (x_i , y_i) es lineal

$$y(x) = a + bx$$

Se quieren encontrar los valores más probables de los coeficientes a y b



Experiencia de péndulo de longitud variable

modelo

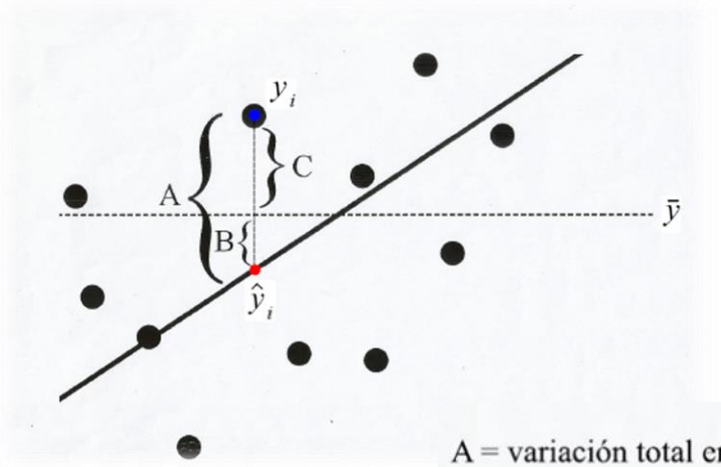
$$T = 2\pi\sqrt{\frac{l}{g}} \rightarrow T^2 = \frac{4\pi^2}{g} l$$

$y(x)$ = x, b = $\frac{4\pi^2}{g}$

Con los datos experimentales se obtiene T en función de l .
Se ajustan los datos T^2 vs l a una relación lineal

$$y(x) = a + bx$$

Debería ser cero



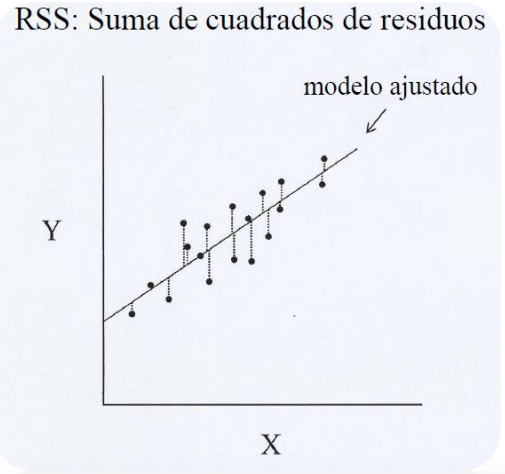
A = variación total en y
B = variación explicada en y
C = residuo no explicado en y

Supongamos tener una variable y que la medimos varias veces.

Tendremos y_i valores

Vamos a poder calcular un valor medio de la misma $\bar{y}$

Si generalizamos, supongamos que existe (o queremos probar) un modelo $\hat{y}$ que explique los resultados



Definamos algunos conceptos

A. Suma residual de los cuadrados.
B. Suma explicada de los cuadrados.
C. Suma total de los cuadrados

$$RSS = \Sigma(y_i - \hat{y}_i)^2 = SSE$$

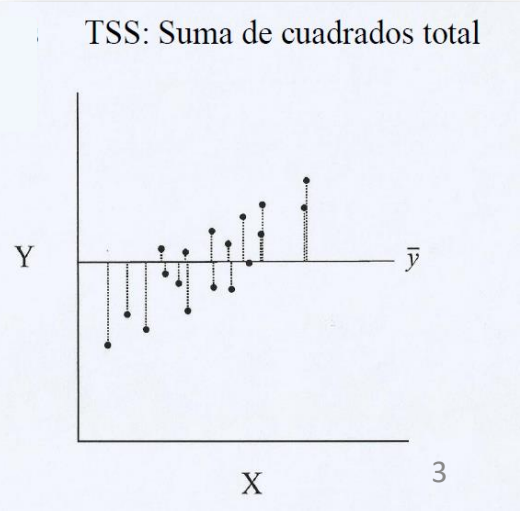
$$ESS = \Sigma(\hat{y}_i - \bar{y})^2 = SSR$$

$$TSS = \Sigma(y_i - \bar{y})^2 = SST$$

$$TSS = ESS + RSS$$

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

Vale para regresiones lineales o, en general, para variables con errores gaussianos y parámetros independientes (siempre que los residuos sean también independientes)



y_i son los valores medidos que se distribuyen normalmente alrededor de los valores del modelo $\hat{y}_i$

La probabilidad de haber obtenido c/u de esos valores es

$$P_i = \frac{1}{\sqrt{2\pi}\sigma_i} exp\left(-\frac{(y_i - \hat{y}_i)^2}{2\sigma_i^2}\right)$$

La probabilidad de realizar el conjunto observado de medidas de los N valores de y_i es el producto de la probabilidad de cada observación.

$$P = P_1 P_2 P_3 P_4 \ldots . P_N$$

$$P = \prod_i P_i = \prod_i^N \frac{1}{\sqrt{2\pi}\sigma_i} exp\left(-\frac{1}{2}\sum_i^N \frac{(y_i - \hat{y}_i)^2}{\sigma_i^2}\right)$$

Si se usa el principio de máxima probabilidad ("lo que se observó es lo más probable")

P máximo

$$\chi^2 = \sum_i^N \frac{(y_i - \hat{y}_i)^2}{\sigma_i^2} \quad \text{mínimo}$$

$$RSS = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \quad \text{mínimo}$$

Si la dispersión σ es constante

Repasemos las distintas clases de Estadísticas

Estadística descriptiva

Considera únicamente de las propiedades de los datos observados y no se basa en la suposición de que los datos provienen de una población más grande.

Estadística inferencial

- ✓ Utiliza el análisis de datos para inferir las propiedades de una distribución de probabilidad subyacente.
- ✓ Infiere propiedades de una población, por ejemplo, probando hipótesis y derivando estimaciones.
- ✓ Se supone que el conjunto de datos observados se extrae de una población más grande.

Test χ^2

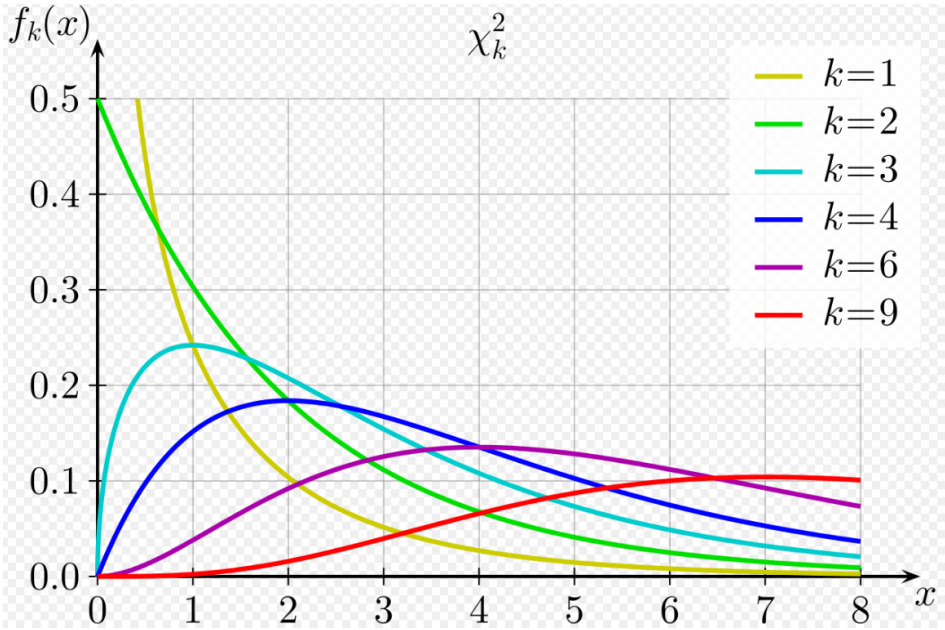
- Es una prueba de hipótesis estadística que es válida para realizar cuando el estadístico de prueba es χ^2 distribuido bajo la hipótesis nula.
- La prueba χ^2 se utiliza para determinar si existe una diferencia estadísticamente significativa entre las frecuencias esperadas y las frecuencias observadas en una o más categorías de una tabla de contingencia.

- En teoría de probabilidad y estadística, la distribución χ^2 con k grados de libertad es la distribución de una suma de los cuadrados de k variables aleatorias normales estándar independientes.

$$\chi^2 = \sum_i^N \frac{(y_i - \hat{y}_i)^2}{\sigma_i^2}$$

- La distribución χ^2 es un caso especial de la distribución gamma y es una de las distribuciones de probabilidad más utilizadas en la estadística inferencial, especialmente en la prueba de hipótesis y en la construcción de intervalos de confianza.

La distribución χ^2 se utiliza en las pruebas de χ^2 comunes para determinar la bondad de ajuste de una distribución observada a una teórica.



La función de densidad de probabilidad de la distribución χ^2 es

$$f(x;\, k) = \begin{cases} \dfrac{x^{\frac{k}{2}-1} e^{-\frac{x}{2}}}{2^{\frac{k}{2}} \Gamma\left(\frac{k}{2}\right)}, & x > 0; \\ 0, & x \leq 0 \end{cases}$$

Función gama

La integral de esta función debe ser 1

Prueba de bondad del ajuste χ^2

- Es una prueba no paramétrica que se utiliza para averiguar si el valor observado (medido) es significativamente diferente del valor esperado.

Si el valor de χ^2 es mayor que un valor mínimo prefijado se rechaza la Hipótesis nula y se asume que existe una diferencia significativa entre la frecuencia observada y medida.

Chi cuadrado reducido $\chi_\nu^2 = \frac{\chi^2}{\nu}$

Grados de libertad

$\nu = n - m$

Número de parámetros del modelo

Número de observaciones (mediciones)

Si el modelo es lineal m = 2

Hipótesis

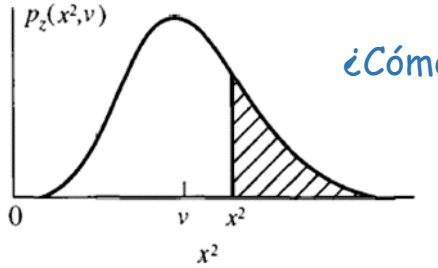
Ho = Supone que no hay diferencia significativa entre el valor observado y el esperado.

Ha = Asume que existe una diferencia significativa entre el valor observado y el esperado.

Supongamos que se hace un ajuste de 20 mediciones experimentales a una función lineal

$\nu = n - m = 20 - 2 = 18$

Se calcula χ^2 . Supongamos que $\chi^2 = 8.54$ ➡ $\chi^2_\nu = 8.54 / 18 = 0{,}474$



¿Cómo calculo la probabilidad P ? Uso la Tabla

Se busca esa valor en la tabla para los grados de libertad

TABLE C.4

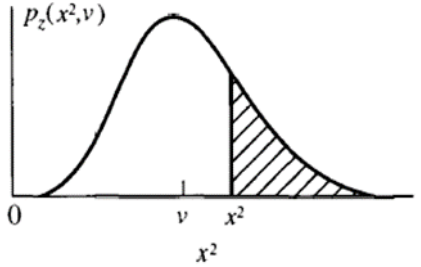
χ^2 distribution. Values of the reduced chi-square $\chi^2_\nu = \chi^2/\nu$ corresponding to the probability $P_\chi(\chi^2; \nu)$ of exceeding χ^2 versus the number of degrees of freedom ν

	P							
ν	0.99	0.98	0.95	0.90	0.80	0.70	0.60	0.50
1	0.00016	0.00063	0.00393	0.0158	0.0642	0.148	0.275	0.455
2	0.0100	0.0202	0.0515	0.105	0.223	0.357	0.511	0.693
3	0.0383	0.0617	0.117	0.195	0.335	0.475	0.623	0.789
4	0.0742	0.107	0.178	0.266	0.412	0.549	0.688	0.839
5	0.111	0.150	0.229	0.322	0.469	0.600	0.731	0.870
6	0.145	0.189	0.273	0.367	0.512	0.638	0.762	0.891
7	0.177	0.223	0.310	0.405	0.546	0.667	0.785	0.907
8	0.206	0.254	0.342	0.436	0.574	0.691	0.803	0.918
9	0.232	0.281	0.369	0.463	0.598	0.710	0.817	0.927
10	0.256	0.306	0.394	0.487	0.618	0.727	0.830	0.934
11	0.278	0.328	0.416	0.507	0.635	0.741	0.840	0.940
12	0.298	0.348				0.753	0.848	0.945
13	0.316	0.367				0.764	0.856	0.949
14	0.333	0.383				0.773	0.863	0.953
15	0.349	0.399	0.484	0.570	0.687	0.781	0.869	0.956
16	0.363	0.413	0.498	0.582	0.697	0.789	0.874	0.959
17	0.377	0.427	0.510	0.593	0.706	0.796	0.879	0.961
18	0.390	0.439	0.522	0.604	0.714	0.802	0.883	0.963
19	0.402	0.451	0.532	0.613	0.722	0.808	0.887	0.965
20	0.413	0.462	0.543	0.622	0.729	0.813	0.890	0.967
22	0.434	0.482	0.561	0.638	0.742	0.823	0.897	0.970

La probabilidad se ubica entre 98 % y 95 %

	P							
ν	0.40	0.30	0.20	0.10	0.05	0.02	0.01	0.001
1	0.708	1.074	1.642	2.706	3.841	5.412	6.635	10.827
2	0.916	1.204	1.609	2.303	2.996	3.912	4.605	6.908
3	0.982	1.222	1.547	2.084	2.605	3.279	3.780	5.423
4	1.011	1.220	1.497	1.945	2.372	2.917	3.319	4.617
5	1.026	1.213	1.458	1.847	2.214	2.678	3.017	4.102
6	1.035	1.205	1.426	1.774	2.099	2.506	2.802	3.743
7	1.040	1.198	1.400	1.717	2.010	2.375	2.639	3.475
8	1.044	1.191	1.379	1.670	1.938	2.271	2.511	3.266
9	1.046	1.184	1.360	1.632	1.880	2.187	2.407	3.097
10	1.047	1.178	1.344	1.599	1.831	2.116	2.321	2.959
11	1.048	1.173	1.330	1.570	1.789	2.056	2.248	2.842
12	1.049	1.168	1.318	1.546	1.752	2.004	2.185	2.742
13	1.049	1.163	1.307	1.524	1.720	1.959	2.130	2.656
14	1.049	1.159	1.296	1.505	1.692	1.919	2.082	2.580
15	1.049	1.155	1.287	1.487	1.666	1.884	2.039	2.513
16	1.049	1.151	1.279	1.471	1.644	1.852	2.000	2.453
17	1.048	1.148	1.271	1.457	1.623	1.823	1.965	2.399
18	1.048	1.145	1.264	1.444	1.604	1.797	1.934	2.351
19	1.048	1.142	1.258	1.432	1.586	1.773	1.905	2.307
20	1.048	1.139	1.252	1.421	1.571	1.751	1.878	2.266
22	1.047	1.134	1.241	1.401	1.542	1.712	1.831	2.194
24	1.046	1.129	1.231	1.383	1.517	1.678	1.791	2.132
26	1.045	1.125	1.223	1.368	1.496	1.648	1.755	2.079
28	1.045	1.121	1.215	1.354	1.476	1.622	1.724	2.032
30	1.044	1.118	1.208	1.342	1.459	1.599	1.696	1.990
32	1.043	1.115	1.202	1.331	1.444	1.578	1.671	1.953
34	1.042	1.112	1.196	1.321	1.429	1.559	1.649	1.919
36	1.042	1.109	1.191	1.311	1.417	1.541	1.628	1.888
38	1.041	1.106	1.186	1.303	1.405	1.525	1.610	1.861
40	1.041	1.104	1.182	1.295	1.394	1.511	1.592	1.835
42	1.040	1.102	1.178	1.288	1.384	1.497	1.576	1.812
44	1.039	1.100	1.174	1.281	1.375	1.485	1.562	1.790

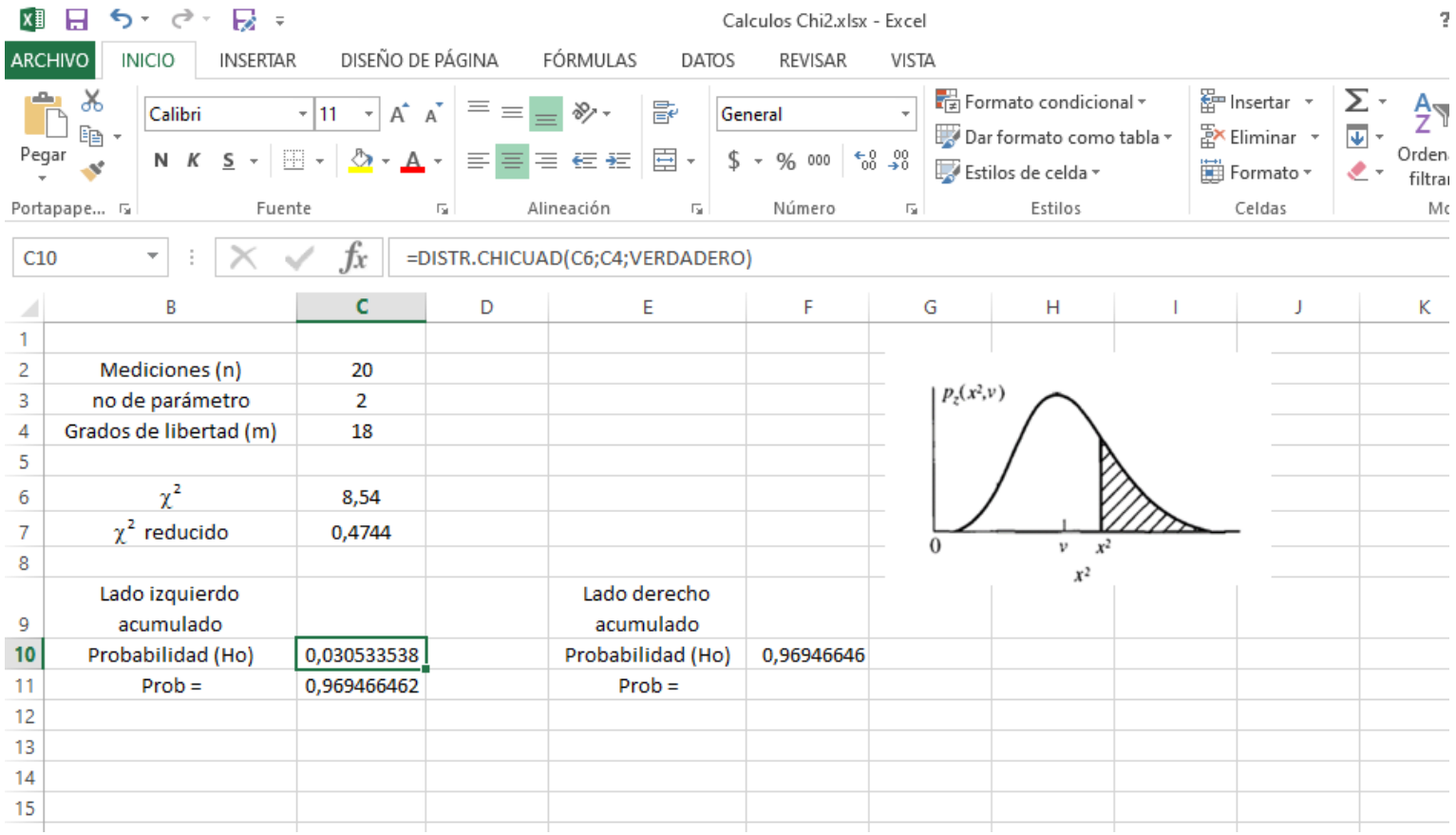
También se puede programar la función y obtener el valor exactos



Excel tiene estas funciones

- DISTR.CHICUAD(χ^2 ; ν ;VERDADERO) — Lado izquierdo acumulado
- DISTR.CHICUAD.CD(χ^2 ; ν) — Lado derecho acumulado

Da la integral bajo la curva para el χ^2 obtenido y los ν grados de libertad.



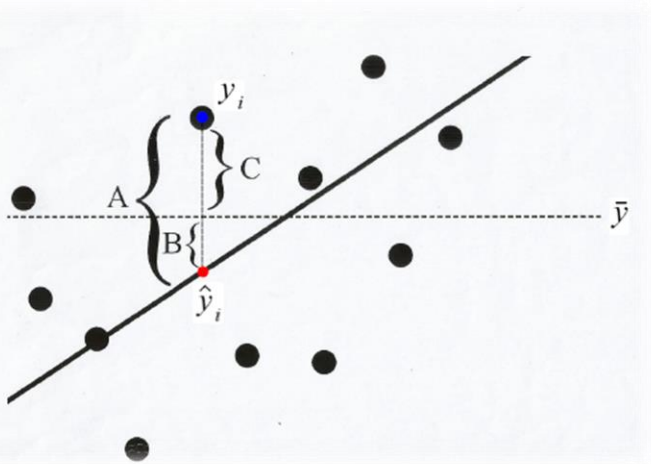
P ≈ 97 %

Repasando

Tendremos y_i valores

un valor medio de la misma $\bar{y}$

Si generalizamos, supongamos que existe (o que vamos a probar) un modelo $\hat{y}$ que explique los resultado.



A = variación total en y
B = variación explicada en y
C = residuo no explicado en y

A. Suma residual de los cuadrados. $RSS = \Sigma(y_i - \hat{y}_i)^2 = SSE$
B. Suma explicada de los cuadrados. $ESS = \Sigma(\hat{y}_i - \bar{y})^2 = SSR$
C. Suma total de los cuadrados $TSS = \Sigma(y_i - \bar{y})^2 = SST$

$$TSS = ESS + RSS$$

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

Coeficiente de correlación R^2

Es la relación entre Suma explicada de los cuadrados (*ESS*) y la suma total de los cuadrados (*TSS*)

$$TSS = \sum_{i=1}^{n}(y_i - \bar{y})^2 \qquad ESS = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 \qquad RSS = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

$$R^2 = \frac{\overset{ESS}{\overbrace{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}}}{\underset{TSS}{\underbrace{\sum_{i=1}^{n}(y_i - \bar{y})^2}}}$$

$$TSS = ESS + RSS \qquad RSS = \Sigma\left(y_i - \hat{y}_i\right)^2$$

$$R^2 = \frac{ESS}{TSS} = 1 - \frac{RSS}{TSS}$$

Minimizando *RSS* se obtiene el máximo R^2

¿ Que es el coeficiente $\bar{R}^2$ ajustado ?

El coeficiente de regresión R^2 indica la proporción de variación de la variable y que es debida una variación de las variables x

A veces se usa coeficiente de regresión R^2 ajustado que es una penalización de los modelos que tienen muchos parámetros

Partiendo de

$$R^2 = 1 - \frac{RSS}{TSS}$$

$$1 - R^2 = \frac{RSS}{TSS}$$

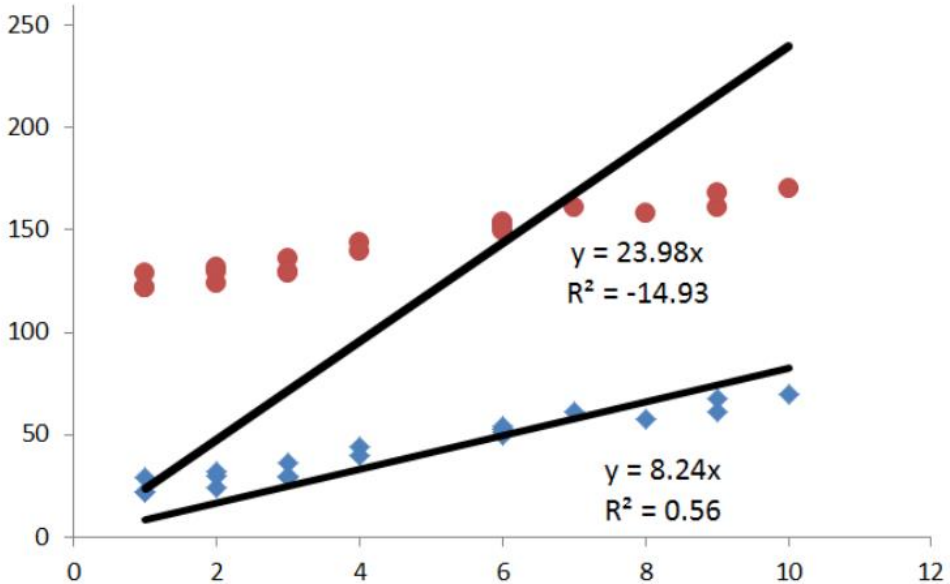
n es la cantidad de datos

m la cantidad de parámetros del modelo

$$1 - \bar{R}^2 = \frac{RSS/(n-m)}{TSS/(n-1)}$$

Se define $\bar{R}^2$ ajustado como

$$\bar{R}^2 = 1 - (1 - R^2)\frac{(n-1)}{(n-m)}$$



Un valor de R^2 negativo significa que el ajuste es peor que usar el valor medio para representar los datos

La forma mas usual de terminar con valor de R^2 negativo es forzar a la línea de regresión que pase por un punto específico.

El resultado es que la suma del error del cuadrado de la regresión es mayor que si se utilizara el valor medio. → $R^2 < 0$

Grados de libertad

Los grados de libertad indican la cantidad de valores independientes que pueden variar en un análisis manteniendo las restricciones existentes

Los grados de libertad son una combinación de cuantos datos hay y cuantos parámetros se necesita estimar, o sea, cuanta información independiente entra en la estimación de los parámetros.

Para tener estimaciones más precisas se requieren muchos grados de libertad

Los grados de libertad definen las distribuciones de probabilidad para las estadísticas de prueba de hipótesis.

Las prueba de hipótesis utilizan la distribución *t*, la distribución *F* y la distribución χ^2 para determinar la significancia estadística

$$TSS = \sum_{i=1}^{n}(y_i - \bar{y})^2$$

$$ESS = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$$

$$RSS = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

grados de libertad $\quad \nu = n - 1 \qquad \nu = m - 1 \qquad \nu = n - m$

Pruebas para validar regresiones

1. Que la variación explicada por el modelo no sea al azar (*F*-test).
2. Que los parámetros de la regresión sean significativamente distintos de 0 (es el *t*-test de cada parámetro).
3. Que el valor de R^2 sea cercano a 1 (test Chi-cuadrado)
4. Que el residuo tenga una distribución aleatoria

La estadística *F* nos dice si la dependencia es al azar o no

La estadística *t* nos dice si cada parámetro es significativo o no para el modelo

El coeficiente R^2 nos dice las proporción de residuo que no explica el modelo

El análisis del residuo, entre otras cosas, puede determinar si hay o no otras variables no consideradas en el modelo.

Hipótesis mas comunes de las regresiones

- ✓ Los errores (residuos) no varían con la variable independiente.
- ✓ Los residuo son independientes. Esto significa que el valor de un residuo no influye sobre el valor de otro.
- ✓ Los residuos siguen una distribución normal

¿ Qué nos indica un test de normalidad ?

- ✓ Si la variable tiene una distribución normal o gaussiana, se pueden usar estadísticas paramétricas que se basan esta suposición.
- ✓ Si una variable falla a la prueba de normalidad :
  - ✓ Se debe observar el histograma y la función de distribución normal
  - ✓ Se debe investigar si algún valor o grupo de valores causan esa no normalidad.
  - ✓ Si no hay valores atípicos, se puede intentar una transformación para que los datos sean normales.
  - ✓ Si no hay una transformación viable, se pueden usar métodos no paramétricos que no requieren normalidad.

Se puede verificar la significancia del modelo usando la estadística *F* de la siguiente forma :

$$ESS = \sum\left(\hat{y}_i - \bar{y}\right)^2$$

$$RSS = \sum\left(y_i - \hat{y}_i\right)^2$$

$$F = \frac{ESS/(m-1)}{RSS/(n-m)}$$

$$\nu = n - m$$

Grados de libertad

Cantidad de parámetros del modelo

Tamaño de la muestra

La distribución *F* (distribución de Fisher-Snedecor) es una distribución de probabilidad continua.
Aparece frecuentemente como la distribución nula de una prueba estadística, especialmente en el análisis de varianza.

Sea *X* una variable aleatoria continua y sean m,n números enteros.
Se dice que la variable aleatoria *X* tiene una distribución *F* con m y n grados de libertad y escribimos $X \sim F_{m,n}$ si su función de densidad está dada por

$$f(x) = \frac{\Gamma\left(\frac{m+n}{2}\right)}{\Gamma\left(\frac{m}{2}\right)\Gamma\left(\frac{n}{2}\right)}\left(\frac{m}{n}\right)^{\frac{m}{2}} \frac{x^{\frac{m-2}{2}}}{\left(1+\frac{mx}{n}\right)^{\frac{m+n}{2}}}$$

para $x > 0$.

¿ Como se interpretan los valores de F en el análisis de regresión ?

En general, en regresiones el *test F* compara el ajuste de diferentes modelos.
A diferencia del *test t*, que evalúa solo un coeficiente de regresión (parámetro del modelo) a la vez, el *test F* evalúa múltiples coeficiente simultáneamente.

En particular, el *test F* de significancia general es una forma específica del *test F*.
Comprar un modelo sin predictores con el modelo que se usa.

La hipótesis para el *test F* de significancia general son :

- ✓ Hipótesis nula : el ajuste del modelo de solo intercepción y el modelo son iguales.
- ✓ Hipótesis alternativa : el ajuste del modelo de solo intercepción es significativamente peor en comparación con el modelo.

Distribución *t*

La distribución *t* (de Student) es una distribución de probabilidad que surge del problema de estimar la media de una población normalmente distribuida cuando el tamaño de la muestra es pequeño y la desviación estándar poblacional es desconocida.

Se puede verificar si la hipótesis nula para un parámetro *b* usando la distribución *t* de la siguiente forma

$$t = \frac{b}{s_b}$$

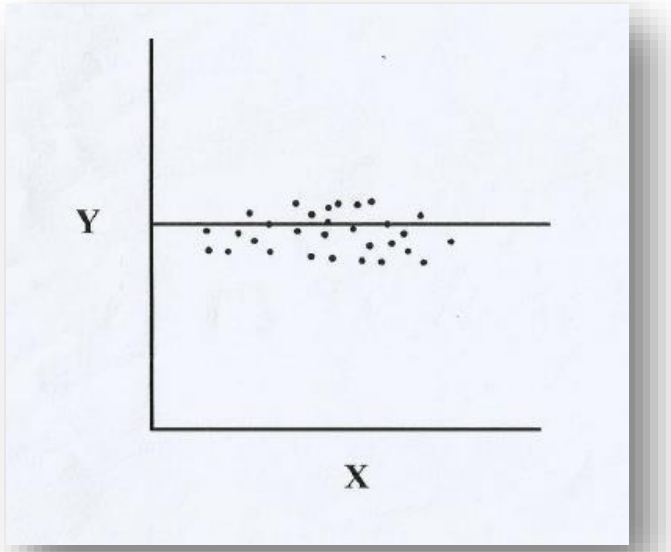
s_b es la desviación standard del parámetro *b*

$$s_b = \sqrt{\frac{RSS/n-m}{\sum x^2}}$$

$$RSS = \sum \left(y_i - \hat{y}_i\right)^2$$

$$\nu = n - m$$

- ν: Grados de libertad
- n: Tamaño de la muestra
- m: Cantidad de parámetros del modelo

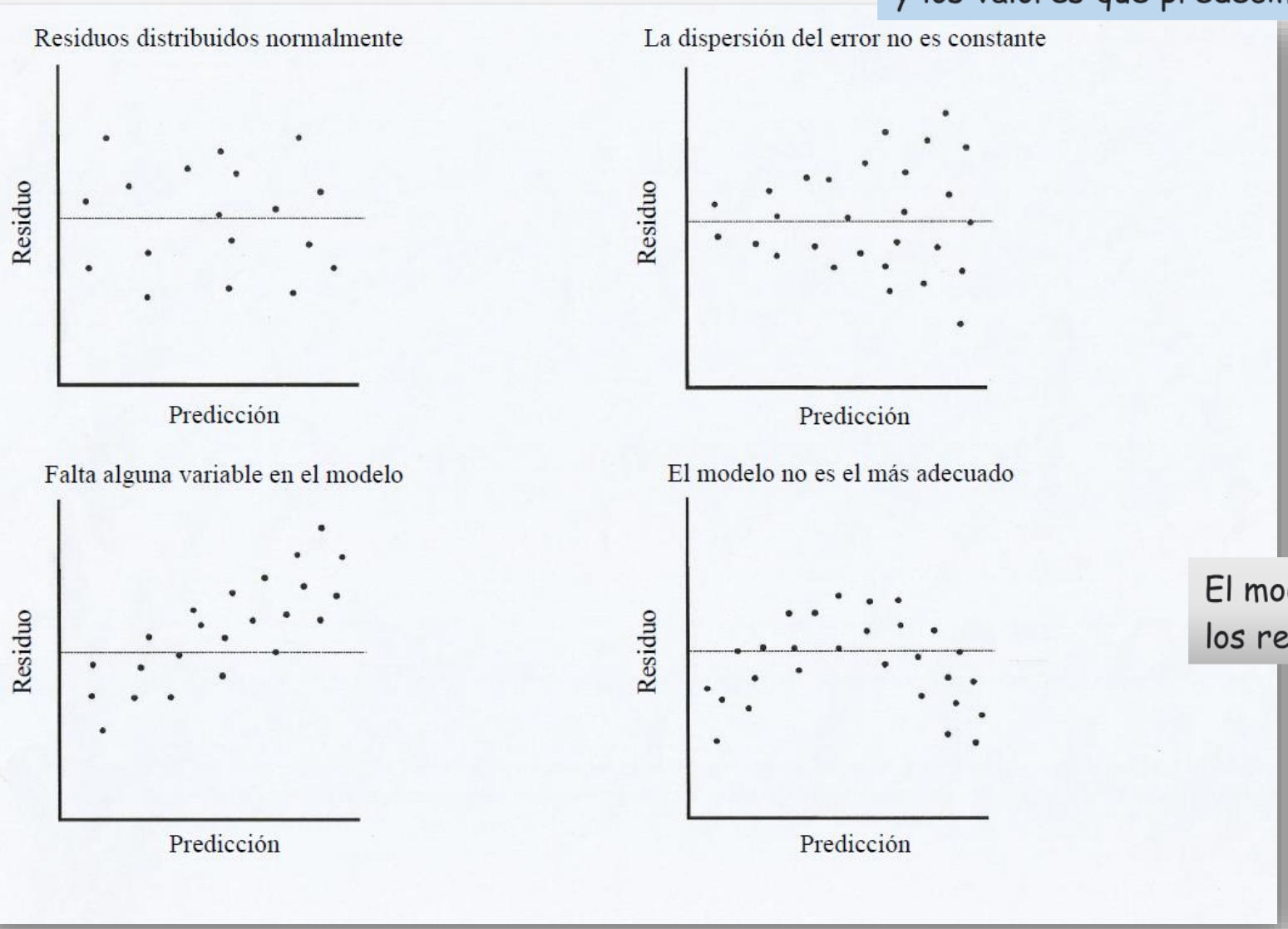


Cuando *b* = 0 no hay cambios en *y* al variar *x*

Distribución de residuos

En el contexto de la regresión lineal, llamamos residuos a las diferencias entre los valores de la variable dependiente observados y los valores que predecimos a partir de nuestra recta de regresión.

$$(y_i - \hat{y}_i)$$



El modelo de regresión lineal también supone que los residuos siguen una distribución normal

¿ Cómo se interpretan los valores *p* en una análisis de una regresión ?

El valor de *p* para cada variable del modelo prueba la hipótesis nula de que el coeficiente es igual a cero (sin efecto).

Un valor bajo ($p < 0.05$) indica que se puede rechazar la hipótesis nula.

Un predictor que tiene un valor bajo no es debido al azar y probablemente sea un parámetro significativo al modelo.

↳ Los cambios en el predictor están relacionados con cambios en la variable respuesta

Un valor de *p* mayor sugiere que los cambios en el predictor no están relacionados con cambios en la variable respuesta.

↳ Existe una alta probabilidad que la dependencia observada sea al azar.

Normalmente se utilizan los valores de *p* para determinar que términos mantener en el modelo de regresión.

Ejemplo de regresión lineal

Supongamos que queremos medir la velocidad de un cuerpo en MRU.

En forma directa, por ejemplo usando una pistola de velocidad, se obtiene:

$$v = v_0 \pm \Delta v$$

Precisión del instrumento si se mide una vez

En forma indirecta, midiendo la posición a dos tiempos distintos, y luego haciendo el cociente entre la distancia recorrida y el tiempo.

$$v = \frac{x_2 - x_1}{t_2 - t_1} \qquad \Delta v = \sqrt{\sum_{i=1}^{4} \left(\frac{\partial v}{\partial q_i}\right)^2 (\Delta q_i)^2} \qquad q_i = x_1, x_2, t_1, t_2$$

En ambos casos, podríamos medir *N* veces, y reportar la media de todas las velocidades obtenidas.

El error sería $\Delta v_C = \sqrt{(\Delta v_A)^2 + (\Delta v_B)^2}$

Es el nominal

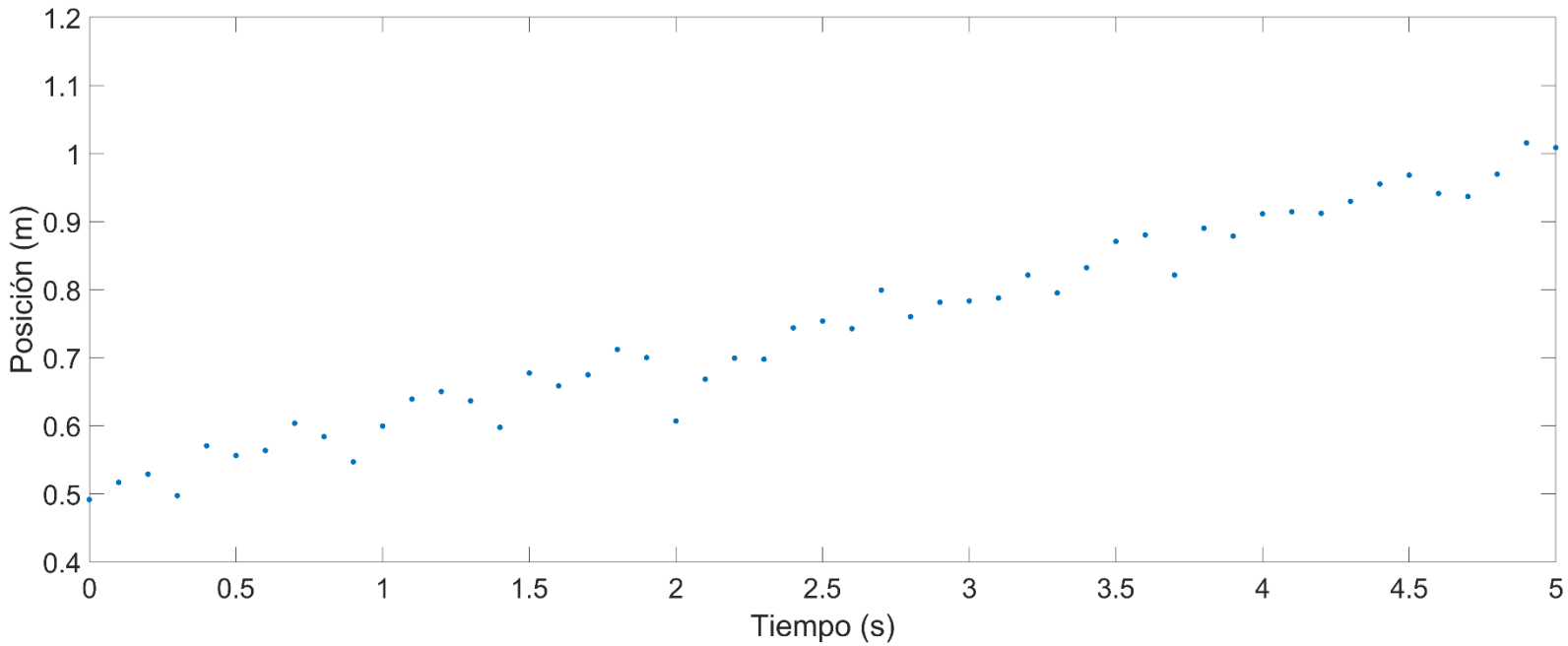
Es el estadístico

¿Por qué no medir directamente la posición *N* veces?

Si estamos en un MRU, la ecuación horaria resultante es:

$$x(t) = x_0 + vt$$

Supongamos que el cuerpo se mueve con *v* = 0.1 m/s, y medimos 10 veces por segundo.
Entonces podríamos obtener algo así:



¿ Porqué no daría un línea recta ?
¿ Hay algo mal en la ecuación horaria ?

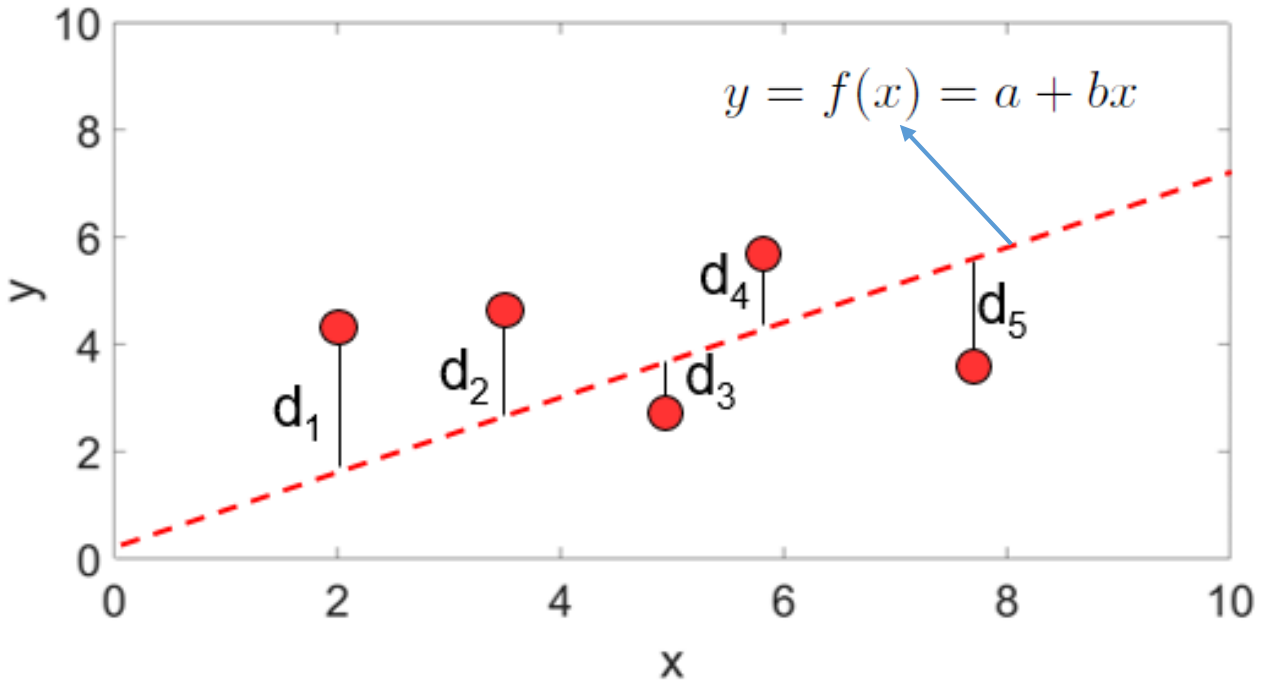
¿ Cómo podemos estimar el mejor valor para x_0 y v ?

¿ Qué quiere decir mejor ?

Lo que queremos, es que la recta estimada se ajuste bien a los datos

Esto es un ejemplo particular de un problema más general

N pares de puntos medidos (x_i, y_i)
y queremos elegir la recta con la menor distancia a los puntos, en algún sentido estadístico.

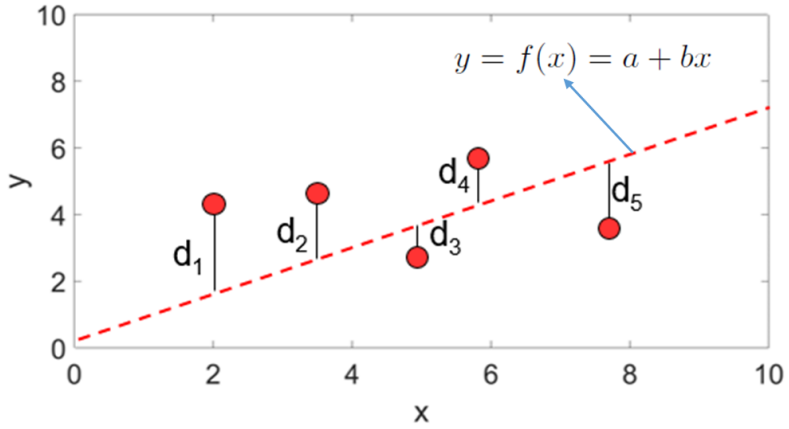


Para determinar a y b,
se aplica
el **método de los cuadrados mínimos**,
que minimiza suma de las distancias cuadráticas verticales de los puntos a la recta.

Queremos minimizar la suma:

$$d_1^2+d_2^2+d_3^2+d_4^2+d_5^2$$

Formalmente, queremos minimizar: $$\chi^2 = \sum_{i=1}^{N} [y_i - (a + bx_i)]^2$$



Los *N* puntos (x_i, y_i) están medidos, son constantes. Podemos pensar a χ^2 como una función de *a* y *b*. Para encontrar los valores de *a* y *b* que minimizan la expresión, derivamos e igualamos a cero:

$$\frac{\partial \chi^2}{\partial a} = 0 \qquad \frac{\partial \chi^2}{\partial b} = 0$$

Esto lleva a un par de ecuaciones lineales con dos incógnitas, denominadas ecuaciones normales.

Su solución es única y sencilla:

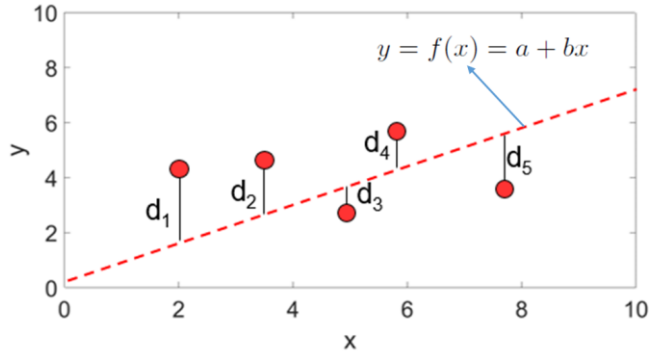
$$\chi^2 = \sum y_i^2 + b^2 \sum x_i^2 + N\, a^2 + 2ab \sum x_i - 2b \sum x_i y_i - 2a \sum y_i$$

$$\frac{\partial \chi^2}{\partial b} = 2b \sum x_i^2 + 2a \sum x_i - 2 \sum x_i y_i = 0$$

$$\frac{\partial \chi^2}{\partial a} = 2bN + 2b \sum x_i - 2 \sum y_i = 0$$

Vale si:
- el error en *x* es despreciable.
- el error en *y* es despreciable, o igual en todas las mediciones.

$$y = f(x) = a + bx$$



Si se resuelve

$$\frac{\partial \chi^2}{\partial b} = 2b\sum x_i^2 + 2a\sum x_i - 2\sum x_i y_i = 0$$

$$\frac{\partial \chi^2}{\partial a} = 2bN + 2b\sum x_i - 2\sum y_i = 0$$

$$a = \frac{\sum x_i^2 \sum y_i - \sum x_i \sum (x_i y_i)}{N\sum x_i^2 - (\sum x_i)^2}$$

$$b = \frac{N\sum x_i y_i - \sum x_i \sum y_i}{N\sum x_i^2 - (\sum x_i)^2}$$

Data reduction and Error Analysis
P.R.Bevington. D.K.Robinson

En un caso un poco más general, consideramos errores en las mediciones de y → Conocemos los σ_i

Nos interesa que la recta ajuste mejor a los puntos medidos con mayor precisión.

Esto se logra definiendo χ^2

$$\chi^2 = \sum_{i=1}^{N} \left[\frac{y_i - (a + bx_i)}{\sigma_i} \right]^2$$

en vez de $$\chi^2 = \sum_{i=1}^{N} [y_i - (a + bx_i)]^2$$

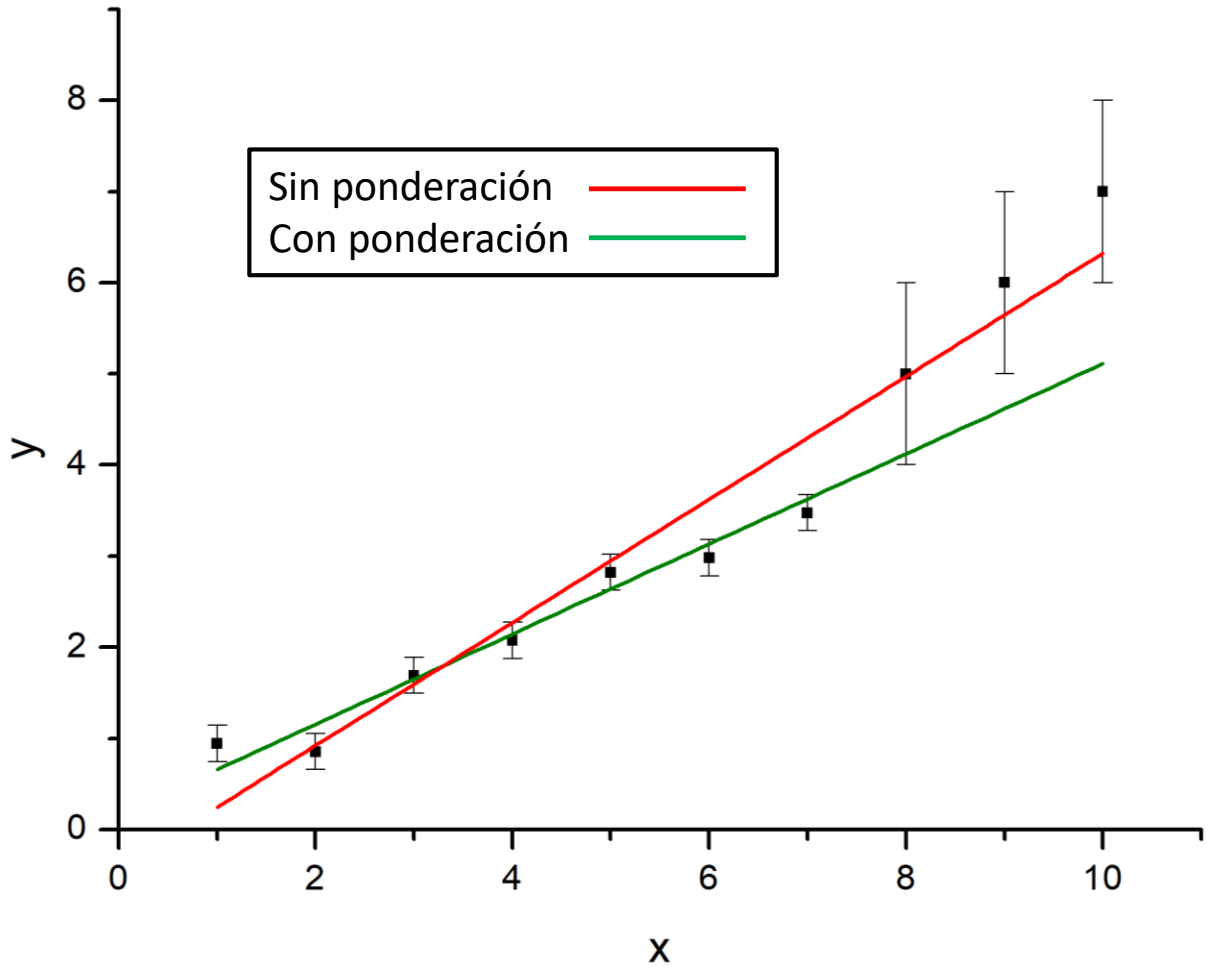
La solución de las ecuaciones normales queda:

$$\begin{cases} a = \frac{1}{\Delta}\left(\sum \frac{x_i^2}{\sigma_i^2} \sum \frac{y_i}{\sigma_i^2} - \sum \frac{x_i}{\sigma_i^2} \sum \frac{x_i y_i}{\sigma_i^2}\right) \\ b = \frac{1}{\Delta}\left(\sum \frac{1}{\sigma_i^2} \sum \frac{x_i y_i}{\sigma_i^2} - \sum \frac{x_i}{\sigma_i^2} \sum \frac{y_i}{\sigma_i^2}\right) \end{cases}$$

con $$\Delta = \sum \frac{1}{\sigma_i^2} \sum \frac{x_i^2}{\sigma_i^2} - \left(\sum \frac{x_i}{\sigma_i^2}\right)^2$$

Si los valores σ_i fuesen los mismos para todos los puntos y_i (por ejemplo en la medición de y_i solo se tiene el error instrumental) → se llega al mismo resultado para a y b como si no se computara el error

La ponderación del error mitiga la distorsión que pueden introducir las mediciones poco precisas.



En este ejemplo solo se incluye el error en la variable dependiente y, Δy .

Se está despreciando el error en x, Δx .

¿ Y si el error en *x* no es despreciable ?

Hay métodos generales para errores en ambas variables, pero son más complicados.

Lo habitual es elegir como *x* la variable con menor error, y despreciarlo.

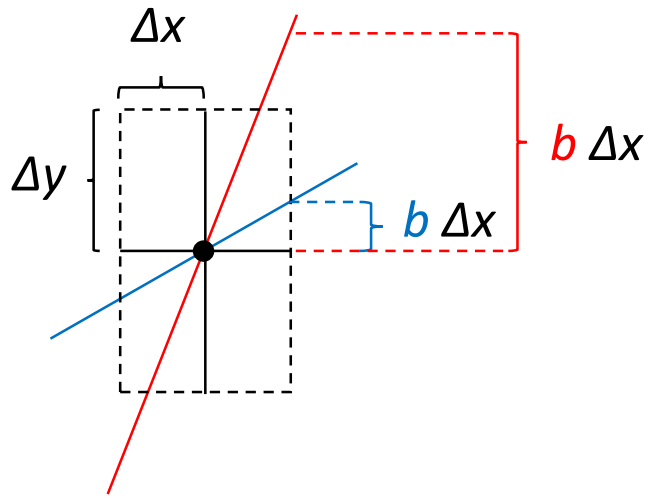
¿ Qué criterio se puede tomar ?

$$\Delta x \leq \Delta y$$

¿Y si *x* e *y* no son la misma magnitud?

$$\frac{\Delta x}{x} \leq \frac{\Delta y}{y}$$

Mejor ¿pero no depende de más nada?



Lo mejor es elegir *x* e *y*, aplicar cuadrados mínimos para estimar *b*, y chequear que se satisfaga: $b\,\Delta x \leq \Delta y$

Si no se cumple, hay que invertir los ejes, y aplicar el método para encontrar otros parámetros:

$$x = c + m\,y$$

Pero queremos $y = a + b\,x \quad \Rightarrow \quad b = \frac{1}{m} \quad a = -\frac{c}{m}$

¿ Qué errores tienen *a* y *b* ?

En su determinación están involucradas las *N* mediciones de (x_i, y_i), así que se estima propagando el error de todas ellas.

Suponiendo que las distintas mediciones no están correlacionadas

$$\sigma_z^2 = \sum_{i=1}^{N} \left[\sigma_i^2 \left(\frac{\partial z}{\partial y_i} \right)^2 \right] \qquad z = a, b$$

$$\frac{\partial a}{\partial y_j} = \frac{1}{\Delta}\left(\frac{1}{\sigma_j^2} \sum \frac{x_i^2}{\sigma_i^2} - \frac{x_j}{\sigma_j^2} \sum \frac{x_i}{\sigma_i^2} \right)$$

$$\frac{\partial b}{\partial y_j} = \frac{1}{\Delta}\left(\frac{x_j}{\sigma_j^2} \sum \frac{1}{\sigma_i^2} - \frac{1}{\sigma_j^2} \sum \frac{x_i}{\sigma_i^2} \right)$$

con $\quad \Delta = \sum \frac{1}{\sigma_i^2} \sum \frac{x_i^2}{\sigma_i^2} - \left(\sum \frac{x_i}{\sigma_i^2} \right)^2$

$$\sigma_a^2 \simeq \sum_{j=1}^{N} \frac{\sigma_j^2}{\Delta^2} \left[\frac{1}{\sigma_j^4} \left(\sum \frac{x_i^2}{\sigma_i^2} \right)^2 - \frac{2x_j}{\sigma_j^4} \sum \frac{x_i^2}{\sigma_i^2} \sum \frac{x_i}{\sigma_i^2} + \frac{x_j^2}{\sigma_j^4} \left(\sum \frac{x_i}{\sigma_i^2} \right)^2 \right]$$

$$= \frac{1}{\Delta^2} \left[\sum \frac{1}{\sigma_j^2} \left(\sum \frac{x_i^2}{\sigma_i^2} \right)^2 - 2 \sum \frac{x_j}{\sigma_j^2} \sum \frac{x_i^2}{\sigma_i^2} \sum \frac{x_i}{\sigma_i^2} + \sum \frac{x_j^2}{\sigma_j^2} \left(\sum \frac{x_i}{\sigma_i^2} \right)^2 \right]$$

$$= \frac{1}{\Delta^2} \left(\sum \frac{x_i^2}{\sigma_i^2} \right) \left[\sum \frac{1}{\sigma_j^2} \sum \frac{x_i^2}{\sigma_i^2} - \left(\sum \frac{x_i}{\sigma_i^2} \right)^2 \right]$$

$$\sigma_a^2 = \frac{1}{\Delta} \sum_{i=1}^{N} \frac{x_i^2}{\sigma_i^2}$$

$$\sigma_b^2 \simeq \sum_{j=1}^{N} \frac{\sigma_j^2}{\Delta^2}\left[\frac{x_j^2}{\sigma_j^4}\left(\sum\frac{1}{\sigma_i^2}\right)^2 - \frac{2x_j}{\sigma_j^4}\sum\frac{1}{\sigma_i^2}\sum\frac{x_i}{\sigma_i^2} + \frac{1}{\sigma_j^4}\left(\sum\frac{x_i}{\sigma_i^2}\right)^2\right]$$

$$= \frac{1}{\Delta^2}\left[\sum\frac{x_j^2}{\sigma_j^2}\left(\sum\frac{1}{\sigma_i^2}\right)^2 - 2\sum\frac{x_j}{\sigma_j^2}\sum\frac{1}{\sigma_i^2}\sum\frac{x_i}{\sigma_i^2} + \sum\frac{1}{\sigma_j^2}\left(\sum\frac{x_i}{\sigma_i^2}\right)^2\right]$$

$$= \frac{1}{\Delta^2}\left(\sum\frac{x_j^2}{\sigma_i^2}\right)\left[\sum\frac{1}{\sigma_j^2}\sum\frac{1}{\sigma_i^2} - \left(\sum\frac{x_i}{\sigma_i^2}\right)^2\right]$$

con $$\Delta = \sum\frac{1}{\sigma_i^2}\sum\frac{x_i^2}{\sigma_i^2} - \left(\sum\frac{x_i}{\sigma_i^2}\right)^2$$

$$\sigma_b^2 = \frac{1}{\Delta}\sum_{i=1}^{N}\frac{1}{\sigma_i^2}$$

Si las incerteza son iguales $\sigma_i = \sigma$ para todos los valores de *y*

$$\sigma_a^2 = \frac{\sigma^2}{\Delta'}\sum x_i^2$$

$$\sigma_b^2 = N\frac{\sigma^2}{\Delta'}$$

con $$\Delta' = N\sum x_i^2 - \left(\sum x_i\right)^2$$

Si realizamos una regresión donde no consideramos errores en las variable

$$\sigma_a^2 = \frac{\sigma^2}{\Delta'}\sum x_i^2$$

$$\sigma_b^2 = N\frac{\sigma^2}{\Delta'}$$

con

$$\Delta' = N\sum x_i^2 - \left(\sum x_i\right)^2$$

$$\sigma = \sqrt{\sum_{i=1}^{n}\frac{(y_i - \hat{y}_i)^2}{(n-m)}}$$

Importante!

Las estimaciones de *a* y *b* no son independientes !

Una fluctuación en *a* implica una en *b*, y viceversa.

Esto es fundamental a la hora de propagar los errores de *a* y *b* en cálculos que los involucran.

El concepto de cuadrados mínimos se puede aplicar a todo tipo de funciones, no solamente a rectas.

Hay una gran diferencia según si la función es lineal o no lineal en los parámetros que se quieren determinar.

Lineales: $a + b\,x + c\,x^2, \quad a\,\mathrm{sen}(x), \quad a\exp(x), \quad a\log(x)$

Todo lo que vimos sigue valiendo, las ecuaciones normales son diferentes pero siempre tienen solución, y es única.

No lineales: $a + a^2\,x, \quad \mathrm{sen}(a\,x), \quad \exp(a\,x), \quad \log(a\,x)$

Las ecuaciones normales no siempre tienen solución única, y en general no tienen solución cerrada.
La suma de los cuadrados se minimiza numéricamente usando algoritmos de optimización, y muchas veces necesita una solución inicial, de la que depende el resultado final.

A veces se pueden convertir un caso no lineal en otro lineal haciendo una sustitución de variable.

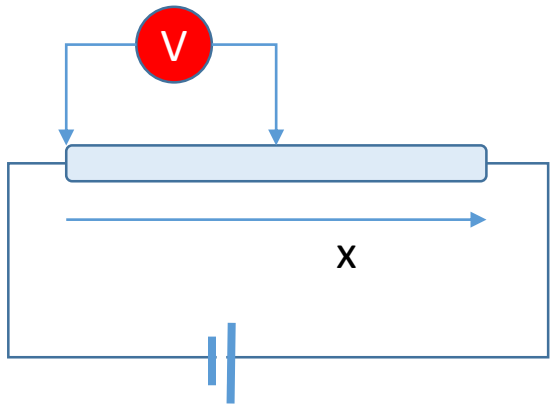
Ejemplo: ajustar N datos (x_i, y_i) con una función exponencial:

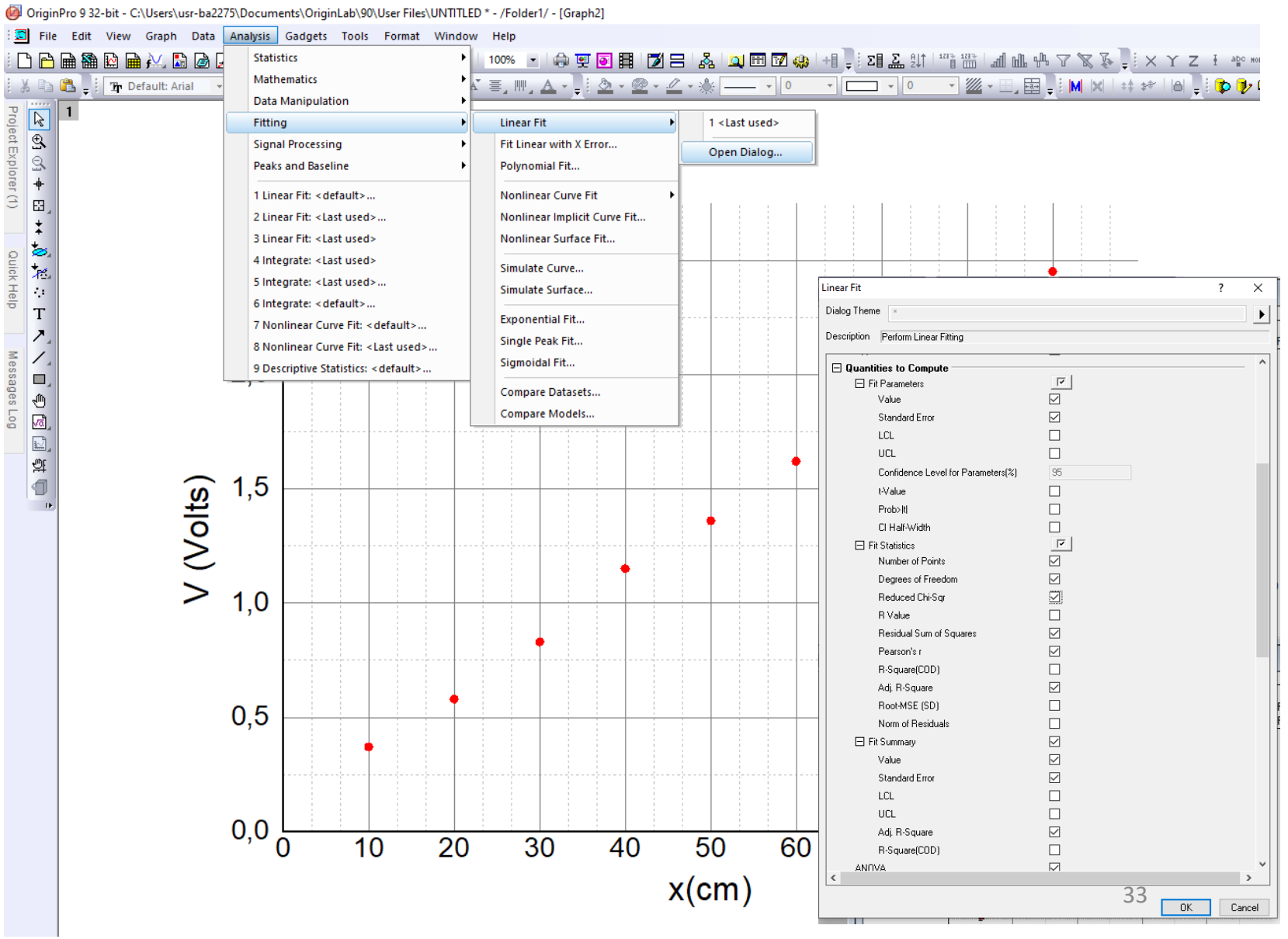
$$y = \exp(ax) \qquad u = \ln(y) \Rightarrow u = ax$$

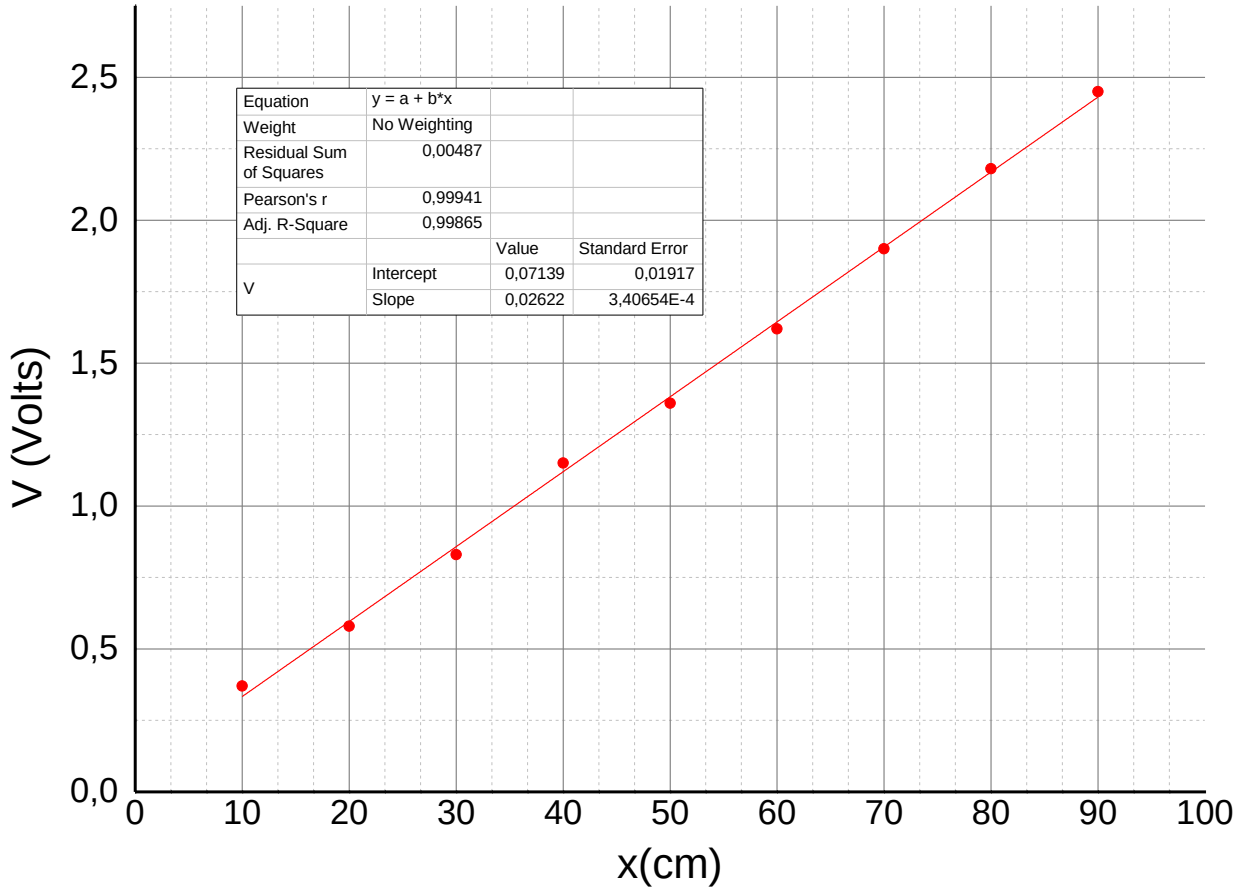
Ejemplo

Se mide la diferencia de potencial V a lo largo de un conductor de nickel/plata en 9 posiciones x

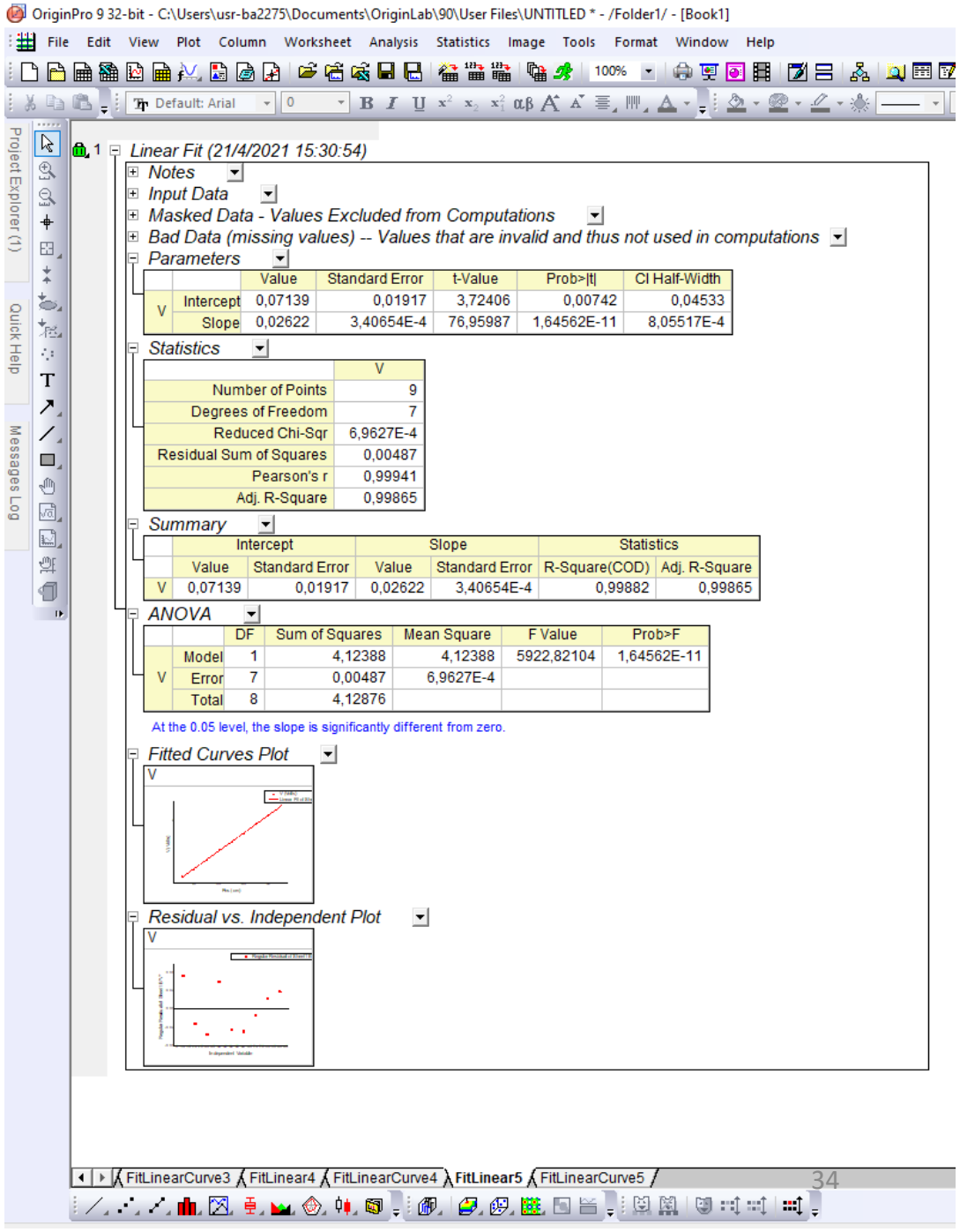
Point number	Postition x_i (cm)	Potential difference V_i (V)
1	10.0	0.37
2	20.0	0.58
3	30.0	0.83
4	40.0	1.15
5	50.0	1.36
6	60.0	1.62
7	70.0	1.90
8	80.0	2.18
9	90.0	2.45

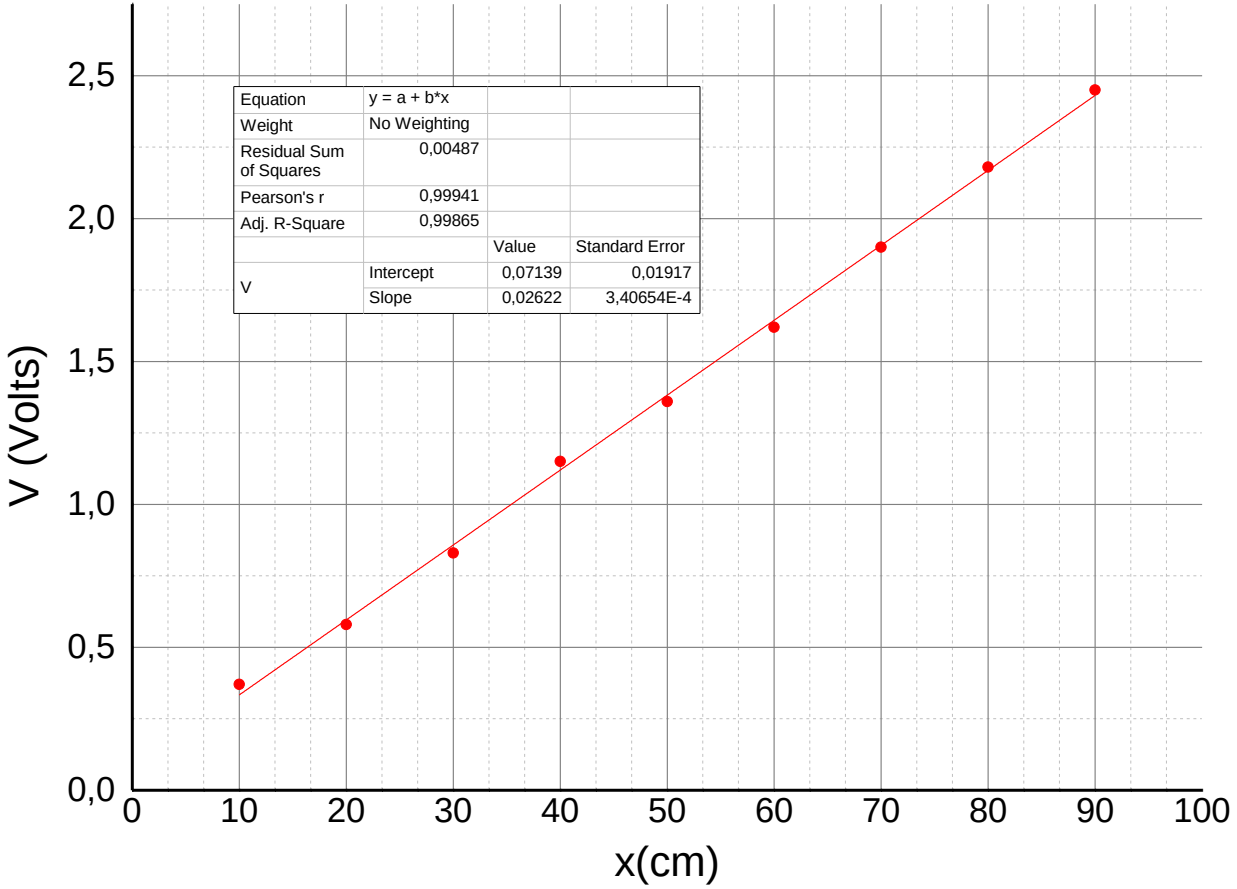






Un valor bajo ($p < 0.05$) indica que se puede rechazar la hipótesis nula





$$V = V_o + bx$$

Parameters

		Value	Standard Error	t-Value	Prob>|t|	CI Half-Width
V	Intercept	0,07139	0,01917	3,72406	0,00742	0,04533
	Slope	0,02622	3,40654E-4	76,95987	1,64562E-11	8,05517E-4

Statistics

	V
Number of Points	9
Degrees of Freedom	7
Reduced Chi-Sqr	6,9627E-4
Residual Sum of Squares	0,00487
Pearson's r	0,99941
Adj. R-Square	0,99865

$$\nu$$

$$\chi^2_\nu = \frac{\chi^2}{\nu}$$

$$RSS = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

Summary

	Intercept		Slope		Statistics	
	Value	Standard Error	Value	Standard Error	R-Square(COD)	Adj. R-Square
V	0,07139	0,01917	0,02622	3,40654E-4	0,99882	0,99865

ANOVA

		DF	Sum of Squares	Mean Square	F Value	Prob>F
V	Model	1	4,12388	4,12388	5922,82104	1,64562E-11
	Error	7	0,00487	6,9627E-4		
	Total	8	4,12876			

$$\chi^2 = \sum_i^N \frac{(y_i - \hat{y}_i)^2}{\sigma_i^2}$$

$$\Delta' = N\sum x_i^2 - \left(\sum x_i\right)^2$$

$$= (9 \times 28{,}500) - (450)^2 = 54{,}000$$

$$a = \frac{1}{\Delta'}\left(\sum x_i^2 \sum y_i - \sum x_i \sum x_i y_i\right)$$

$$= (28{,}500 \times 12.44 - 450.0 \times 779.30)/54{,}000$$

$$= 0.0714$$ → Concuerda con Origin

$$b = \frac{1}{\Delta'}\left(N\sum x_i y_i - \sum x_i \sum y_i\right)$$

$$= (9 \times 779.30 - 450.0 \times 12.44)/54{,}000$$

$$= 0.0262$$ → Concuerda con Origin

Point number	Postition x_i (cm)	Potential difference V_i (V)	x_i^2	$x_i V_i$
1	10.0	0.37	100	3.70
2	20.0	0.58	400	11.60
3	30.0	0.83	900	24.90
4	40.0	1.15	1,600	46.00
5	50.0	1.36	2,500	68.00
6	60.0	1.62	3,600	97.20
7	70.0	1.90	4,900	133.00
8	80.0	2.18	6,400	174.40
9	90.0	2.45	8,100	220.50
Sums	450.0	12.44	28,500	779.30
	$\sum x_i$	$\sum y_i$	$\sum x_i^2$	$\sum x_i y_i$

		Value	Standard Error
V	Intercept	0,07139	0,01917
	Slope	0,02622	3,40654E-4

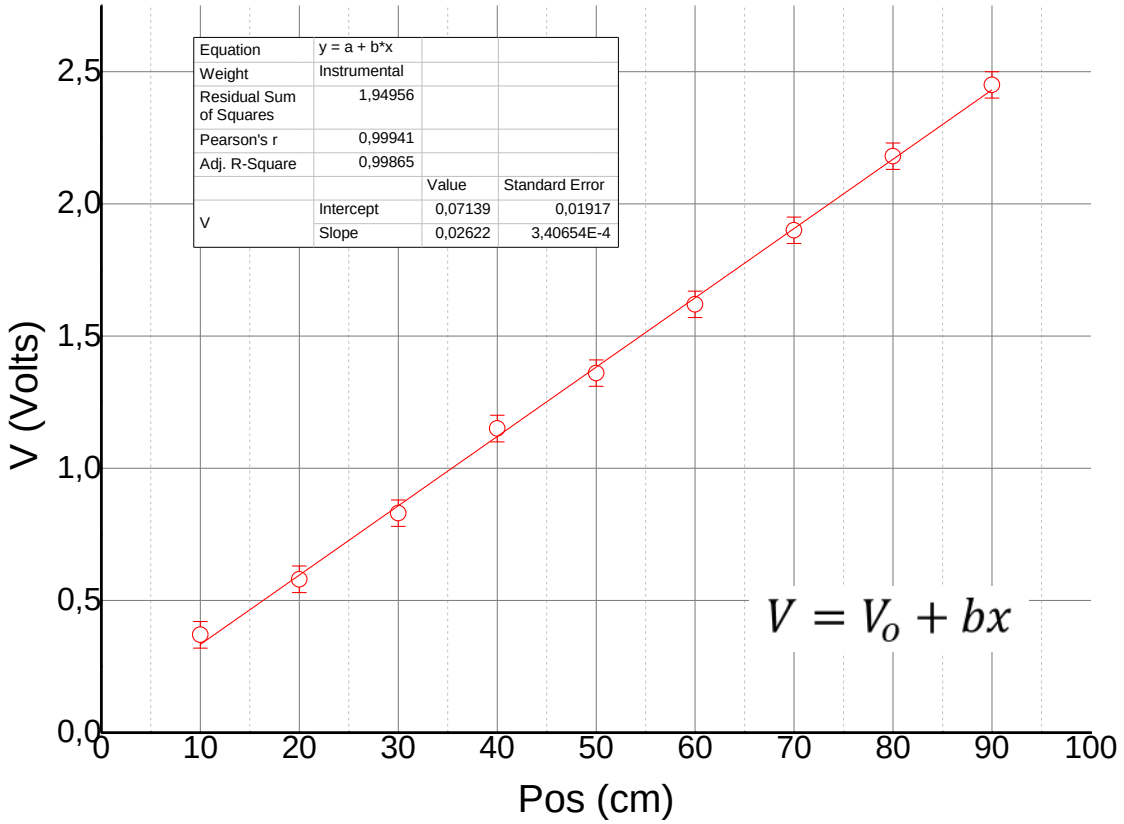
$$\sigma_a^2 = \frac{\sigma^2}{\Delta'}\sum x_i^2$$

$$\sigma_b^2 = N\frac{\sigma^2}{\Delta'}$$

Concuerda con Origin
Verificar

$$\sigma = \sqrt{\sum_{i=1}^{n} \frac{(y_i - \hat{y}_i)^2}{(n-m)}}$$

Supongamos que la diferencia de potencial V medida tiene un error (σ) de 0,05 Volts.
En ese caso en el Origin 9 graficamos con la barra de error en V y luego hacemos el ajuste lineal.
Obsérvese que valores de *a* y *b* no cambian.



Linear Fit (21/4/2021 15:30:17)

Notes

Input Data

Masked Data - Values Excluded from Computations

Bad Data (missing values) -- Values that are invalid and thus not used in computations

Parameters

		Value	Standard Error	t-Value	Prob>|t|	CI Half-Width
V	Intercept	0,07139	0,01917	3,72406	0,00742	0,04533
	Slope	0,02622	3,40654E-4	76,95987	1,64562E-11	8,05517E-4

Statistics

	V
Number of Points	9
Degrees of Freedom	7
Reduced Chi-Sqr	0,27851
Residual Sum of Squares	1,94956
Pearson's r	0,99941
Adj. R-Square	0,99865

cambian → $\chi^2 = \sum_i^N \frac{(y_i - \hat{y}_i)^2}{\sigma_i^2}$

Summary

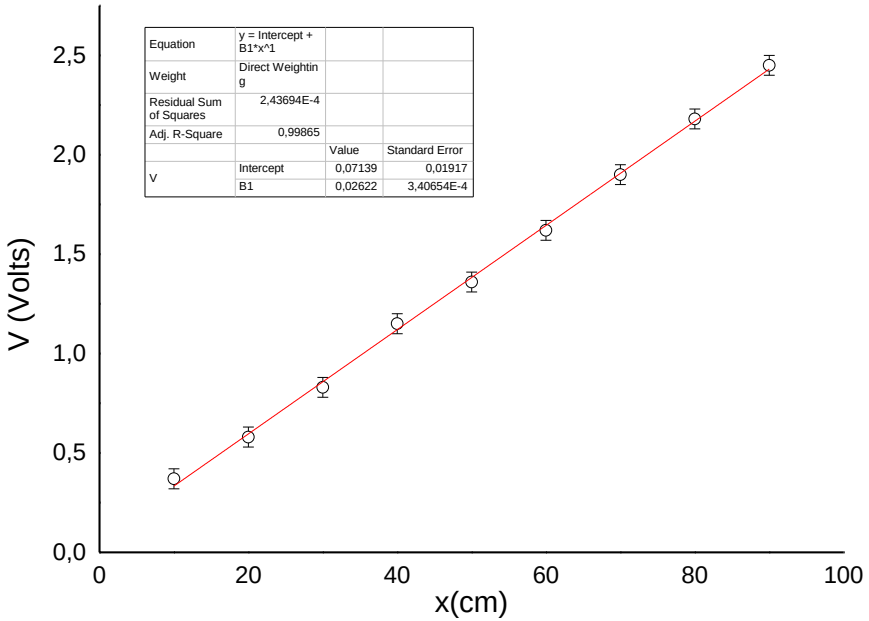
	Intercept		Slope		Statistics	
	Value	Standard Error	Value	Standard Error	R-Square(COD)	Adj. R-Square
V	0,07139	0,01917	0,02622	3,40654E-4	0,99882	0,99865

ANOVA

		DF	Sum of Squares	Mean Square	F Value	Prob>F
V	Model	1	1649,55267	1649,55267	5922,82104	1,64562E-11
	Error	7	1,94956	0,27851		
	Total	8	1651,50222			

cambian

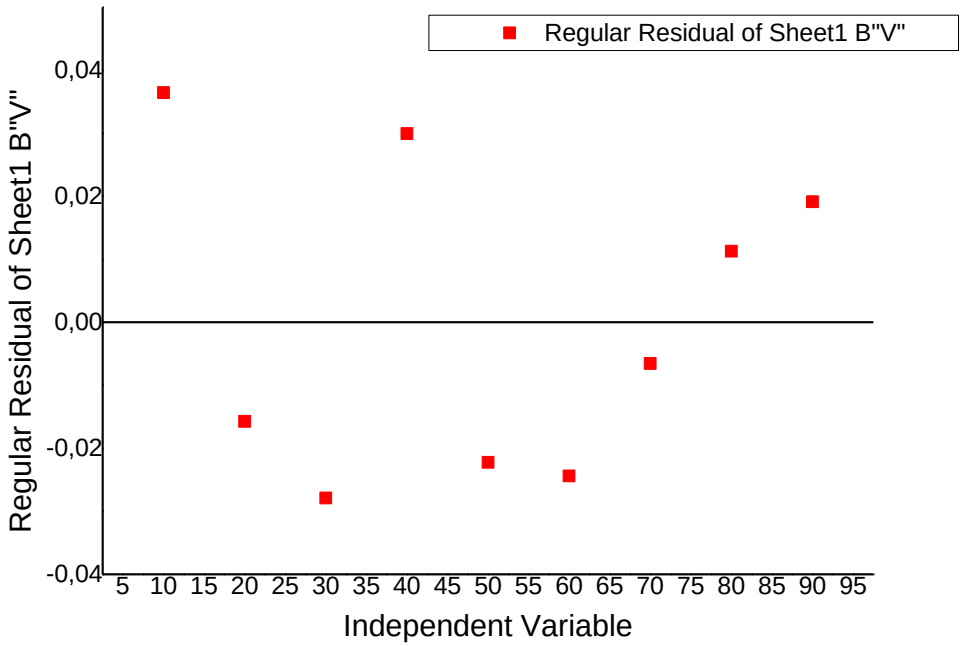
Generar errores diferentes σ_i y calcular para ver que sucede con el ajuste.



En el contexto de la regresión lineal, llamamos residuos a las diferencias entre los valores de la variable dependiente observados y los valores que predecimos a partir de nuestra recta de regresión.

$$(y_i - \hat{y}_i)$$

El modelo de regresión lineal también supone que los residuos siguen una distribución normal

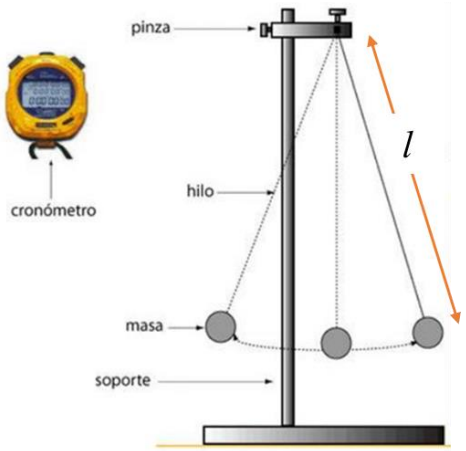


Experiencia

Materiales a utilizar

- Se usara el péndulo de la Clase 1 (donde se pueda cambiar la longitud del hilo fácilmente).

- ✓ Determine la longitud del hilo y su error.
- ✓ Asegurase de trabajar con oscilaciones de amplitud inicial baja (bajo ángulo inicial).
- ✓ Medir el período del péndulo con 7 longitudes de hilo diferentes.
- ✓ Filme la experiencia y analice los resultados de cada oscilación con el programa Tracker.
- ✓ Para calcular el periodo y su error mida por lo menos 50 oscilaciones.
- ✓ Calcule la aceleración de la gravedad a partir de la regresión lineal entre el cuadrado del periodo y la longitud del hilo.
- ✓ ¿Qué diferencia se observa en el valor de la aceleración de la gravedad si utilizo los valores de las variables considerando sus errores?



Experiencia de péndulo de longitud variable

modelo

$$T = 2\pi\sqrt{\frac{l}{g}} \quad \Rightarrow \quad T^2 = \frac{4\pi^2}{g} l$$

$y(x)$ x b

Con los datos experimentales se obtiene T en función de l.
Se ajustan los datos T^2 vs l a una relación lineal

$$y(x) = a + bx$$

Debería ser cero


¿ PREGUNTAS ?