Análisis Estadístico de Datos Climáticos

Regresión lineal simple (Wilks, cap. 6.2)

Von Storch and Zwiers (Cap. 8)

2015

Regresión

La regresión, en general, se utiliza habitualmente para estimar modelos paramétricos de la relación entre variables en una escala continua, ya sea para vincular variables aleatorias (ej., ancho de un anillo de árbol con la temperatura), o una variable aleatoria con uno o más factores externos no aleatorios (ej. modelar una tendencia con un polinomio).

Se puede utilizar para la *predicción* cuando las variables a relacionar no son simultáneas.

Regresión lineal simple

- Estimación de los parámetros
- Distribución de los residuos
- Tabla ANOVA
- Bondad del ajuste
- Análisis de los residuos
- Distribución muestral de coeficientes de la regresión
- Intervalos de confianza de la “predicción”

Regresión lineal simple

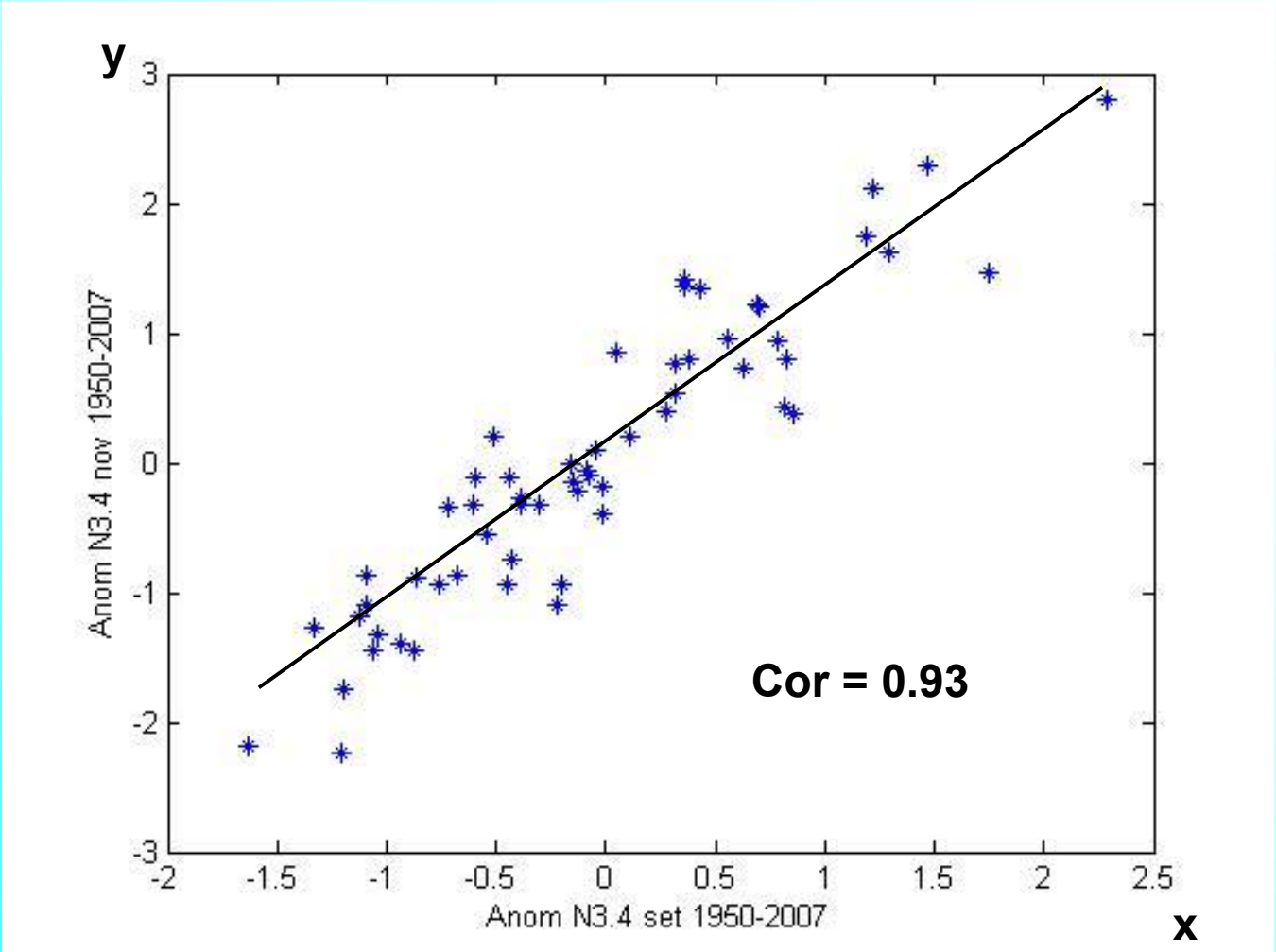
Dados los pares de valores: $(x_1, y_1), (x_2, y_2)\ldots.(x_n, y_n)$

se busca ajustar una recta de ecuación

$$\hat{y} = \hat{a} + \hat{b}\, x$$

(El ajuste puede ser más o menos bueno, según el caso.)

n=58



Regresión lineal simple

x variable independiente o "predictor"

y variable dependiente o "predictando"

(Las palabras "predictor" y "predictando" vienen del lenguaje de la estadística y su uso no se debe confundir con el que le damos p. ej., en el pronóstico meteorológico o climático, aunque puede darse ese caso también.)

No se debe suponer que necesariamente existe una relación de causalidad entre ambas variables.

$$\hat{y} = \hat{a} + \hat{b}\, x$$

$\hat{a}$ y $\hat{b}$ son parámetros a estimar

Regresión lineal simple

Es importante aclarar que, en este contexto, cuando decimos que ajustamos un modelo lineal, nos estamos refiriendo a linealidad en los parámetros (a y b).

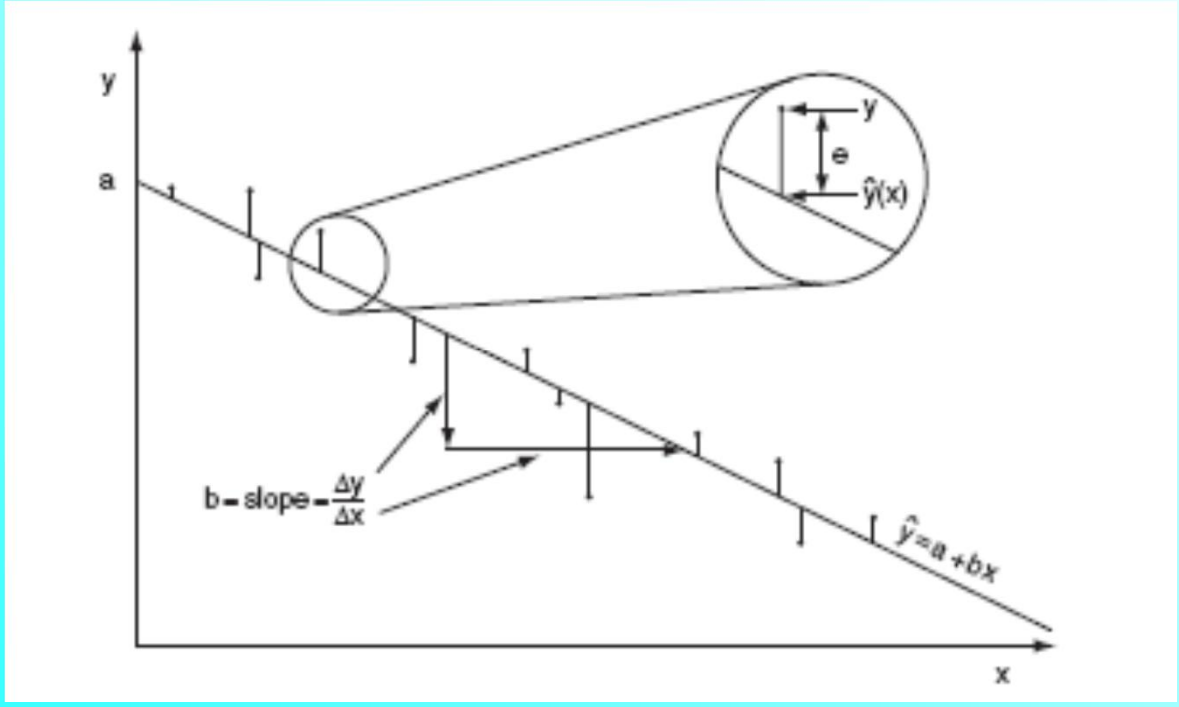
El valor de la más alta potencia de la variable predictora en el modelo se llama orden del modelo.

El modelo $\hat{y} = \hat{a} + \hat{b}\,x$ **es lineal (en a y b), de orden 1**

El modelo $\hat{y} = \hat{a} + \hat{b}x + \hat{c}x^2$ **es lineal (en a, b y c), de orden 2**

Estimación de los parámetros

El criterio más habitual para estimar los parámetros es el método de mínimos cuadrados.



$$e_i = y_i - \hat{y}(x_i)$$

$$y_i = \hat{a} + \hat{b}x_i + e_i$$

Los e_i son los residuos o errores

Se busca minimizar $\sum_{i=1}^{n} e_i^2 = \sum_{1}^{n} (y_i - \hat{a} - \hat{b}x_i)^2 = SSE$

(suma de errores cuadráticos)

Estimación de los parámetros

Se plantea la anulación de las derivadas parciales

respecto de $\hat{a}$ y $\hat{b}$ **obteniéndose las soluciones:**

$$\hat{b} = \frac{\sum_{1}^{n} x_i y_i - n\,\bar{x}\,\bar{y}}{\sum_{1}^{n} x_i^2 - n\,\bar{x}^2}$$

$$\hat{a} = \bar{y} - \hat{b}\,\bar{x}$$

$\bar{x}$ e $\bar{y}$ **son los promedios de los** x_i **e** y_i

La recta de ajuste pasa por el punto $(\bar{x}, \bar{y})$ **que es el centro de gravedad de la nube de puntos**

En el ejemplo:

$$\hat{b} = 1.26$$

$$\hat{a} = 0.09$$

Estimación de los parámetros

Se cumple que: $\hat{b} = \dfrac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} = \hat{r}_{xy}\, s_y / s_x$

siendo $\hat{r}_{xy}$ **la correlación de Pearson entre las series x e y,**

y s_x **y** s_y **las respectivas desviaciones estándar muestrales**

Como caso particular, si las series son estandarizadas, las

S_x **y** S_y **valen 1 y** $\hat{b} = \hat{r}_{xy}$

En Matlab:

```
A=[ones(58,1) n34set5007'];
Y=n34nov5007';
ab=A\Y % resuelve el problema de mínimos cuadrados
    (matrix left division)
ab =
    0.0894
    1.2563
```

También se obtiene el mismo resultado con:

```
ab=polyfit(n34set5007, n34nov5007,1);
```

que ajusta un polinomio de grado 1.

ATENCIÓN: Existe asimetría entre x e y (si se invierten, no se obtiene la misma recta!!)

Distribución de los residuos

Supondremos que los residuos (o errores) e_i son independientes e idénticamente distribuidos (iid) con media 0 y varianza σ (igual para todos los e_i).

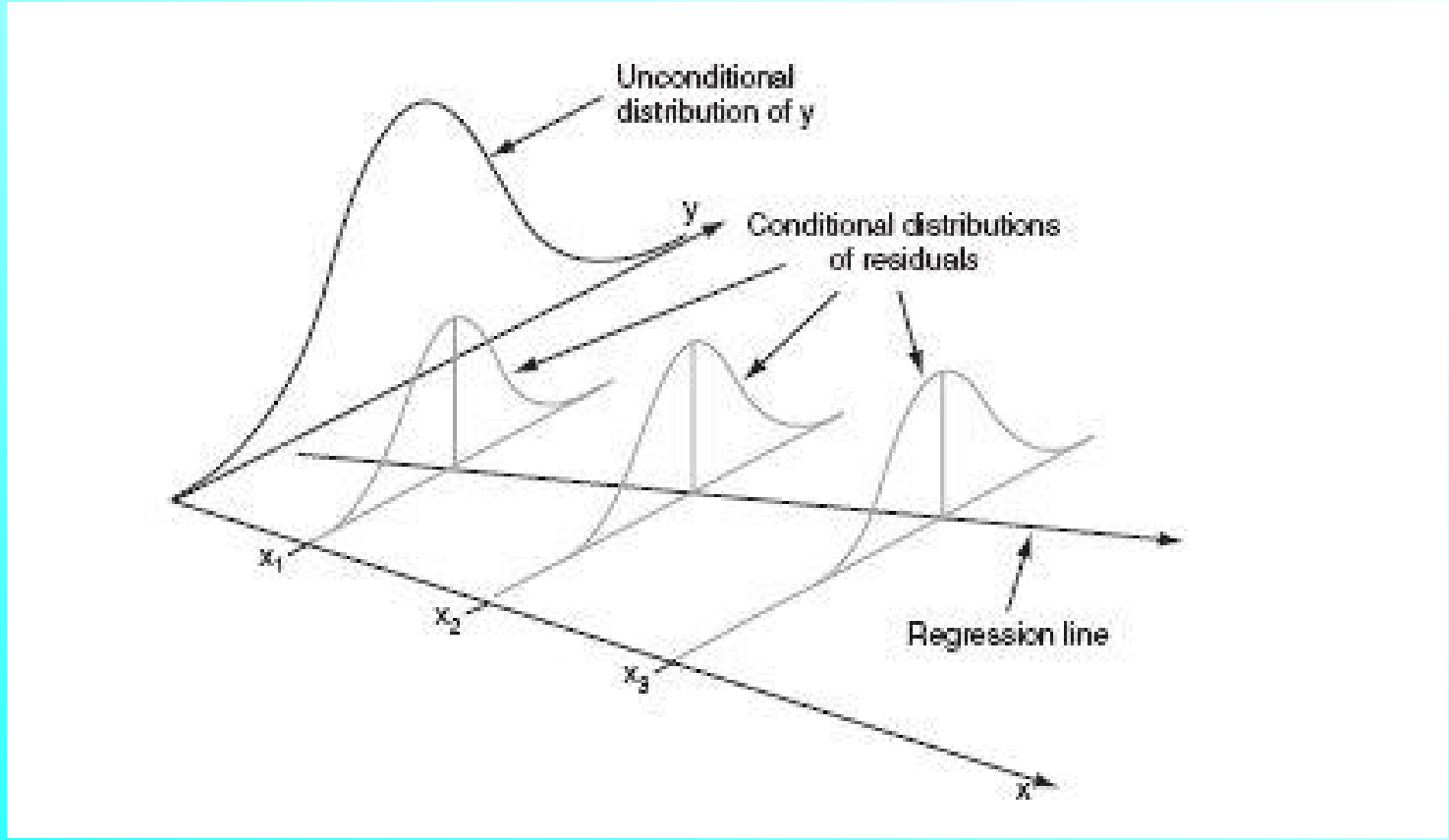
Además se suele suponer que los residuos siguen una distribución gausiana.

En general, cuantas más hipótesis se hagan, más ricas serán las conclusiones estadísticas que podremos extraer, pero más limitada será la aplicabilidad del modelo.

Cuanto mayor sea el tamaño de la muestra (n), más se atenúa la eventual no gausianidad.

Distribución de los residuos

Las suposiciones anteriores implican que la distribución de residuos condicionada a x constante, no depende de x.



Distribución de los residuos

Estimación de la varianza de los residuos

$$\hat{s}_e^2 = \frac{1}{n-2}\sum_1^n e_i^2$$

(es n-2 porque se estiman 2 parámetros, a y b.)

En el ejemplo: $\hat{s}_e^2 = 0.18$

Interesa analizar cuánto de la variación en los datos es *explicado* por la regresión.
Para eso se definen varias sumas de cuadrados:

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

suma de cuadrados total (o respecto de la media)

$$SSR = \sum_{1}^{n} [\hat{y}_i - \bar{y}]^2 = b^2 \left[\sum_{1}^{n} x_i^2 - n\,\bar{x}^2 \right]$$

suma de cuadrados dada por la regresión (es bueno que se acerque a SST)

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

suma de cuadrados de los residuos

Se cumple: SST = SSR + SSE

En el ejemplo anterior:
SST = 72.47 (ºC)²
SSR = 62.49 (ºC)²
SSE = 9.98 (ºC)²

Tabla ANOVA

(ANOVA = Análisis de varianza)

	Grados de libertad	Suma de cuadrados	Media cuadrática	
Total	n- 1	SST		
Regresión	1	SSR	MSR=SSR/1	F=MSR/MSE
Residuos	n-2	SSE	$MSE=s_e^2$	

En cada caso, los grados de libertad indican cuantos valores independientes de los valores y_1, y_2, …., y_n son necesarios para calcular la suma de cuadrados correspondiente.

Esta tabla es válida sólo para el caso con 2 parámetros.

Tabla ANOVA

Para el ejemplo:

	Grados de libertad	Suma de cuadrados	Media cuadrática	
Total	57	SST=72.47		
Regresión	1	SSR=62.49	MSR=SSR/1= 62.49	F=MSR/MSE = 347.2
Residuos	56	SSE=9.98	MSE=s_e^2 =0.18	

Bondad del ajuste

Hay 3 indicadores usuales para la bondad de ajuste:

1) $MSE = \hat{s}_e^2$ **(da un valor promedio de la exactitud del ajuste; lo ideal sería MSE=0)**

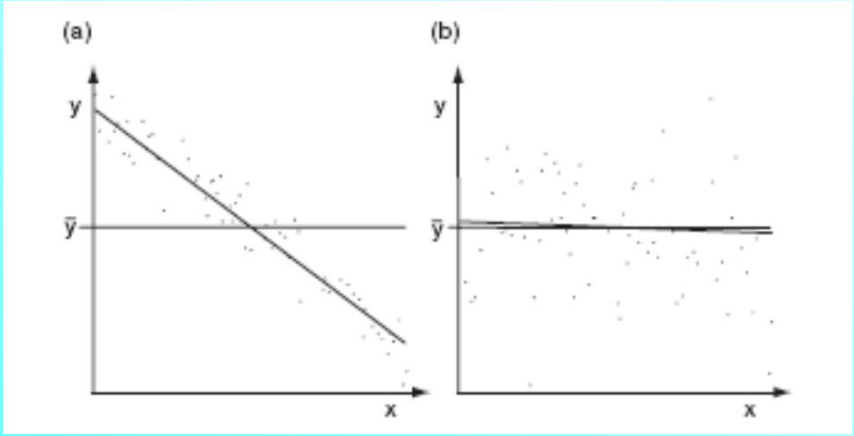
2) **Coeficiente de determinación:** $R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$
en el peor caso vale 0,
en el mejor, vale 1.

Sólo para el ajuste de una recta, se cumple que $R^2 = r_{xy}^2$;
en el ejemplo, $R^2 = 0.86$

3) **El estadístico F=SSR/MSE (es mayor cuanto mejor es el ajuste)**

Bondad del ajuste

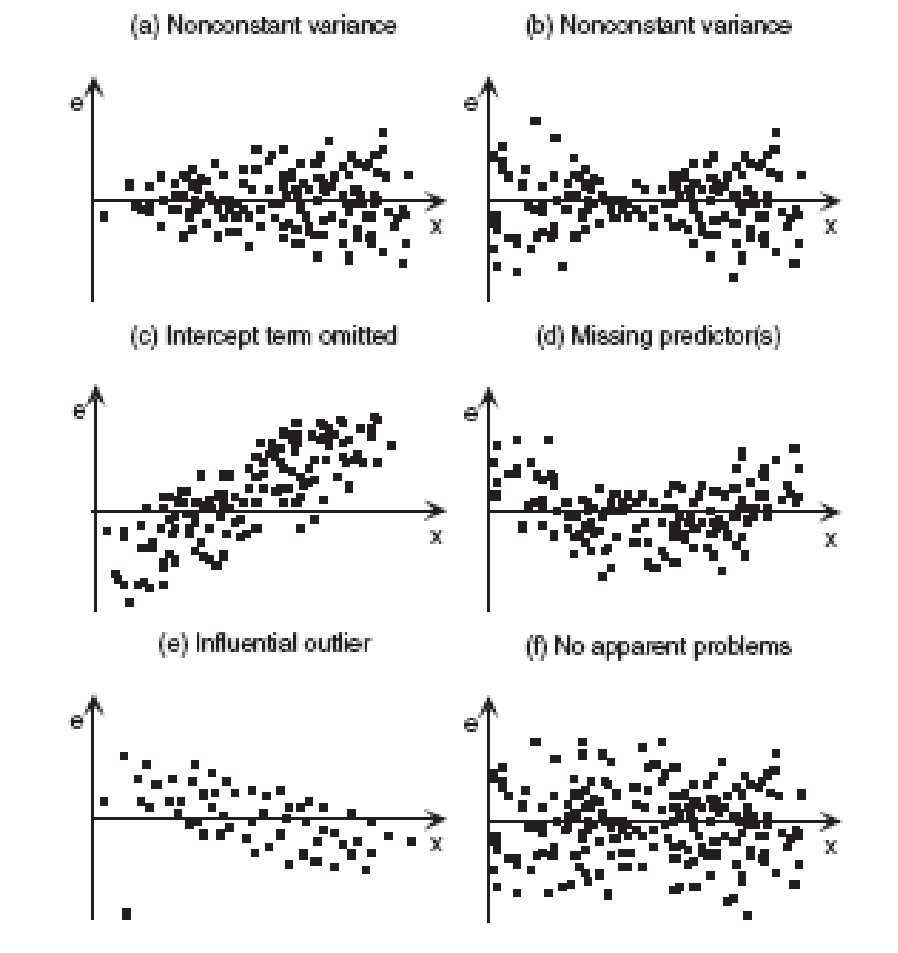
En general, cuanto más cercano a 0 esté el coeficiente angular b, menos información aporta la regresión lineal o, de otra forma, más débil es la relación entre x e y.



$\hat{b} = \hat{r}_{xy} s_y / s_x$ **=> b=0 si** $r_{xy} = 0$**, o si** $s_y = 0$

Análisis de los residuos

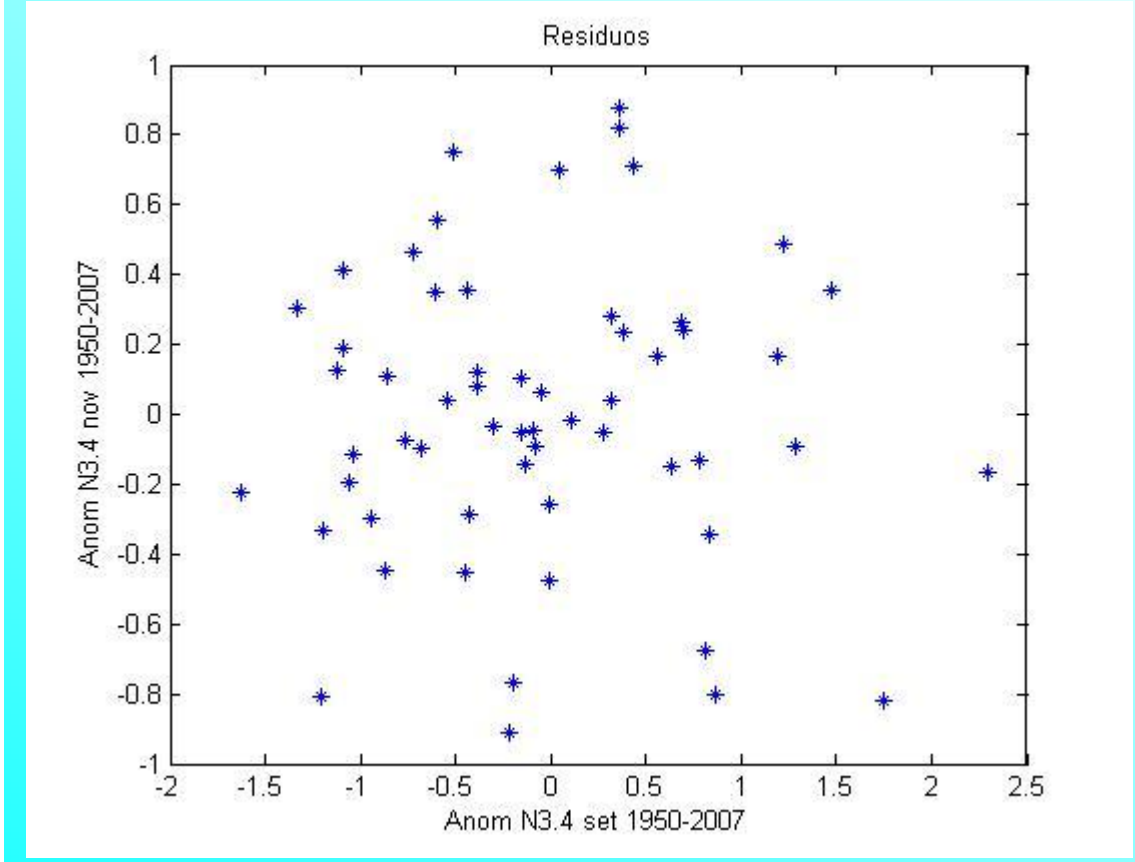
Es muy conveniente hacer una inspección visual de los residuos.



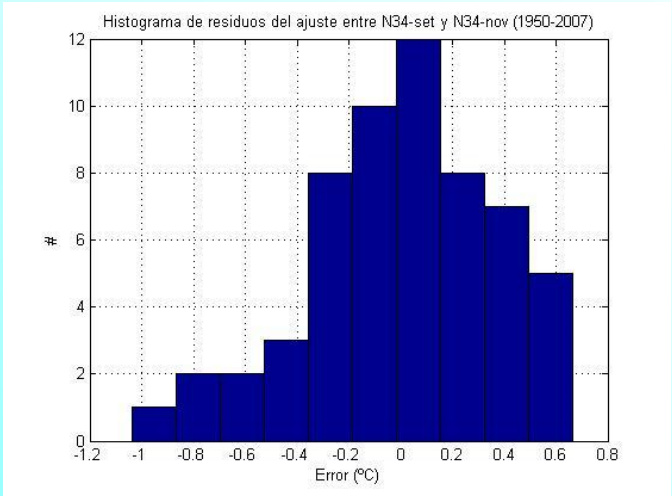
Adicionalmente se puede hacer un test para evaluar la hipótesis nula de que los residuos siguen una distribución gausiana

Análisis de los residuos

(para el ejemplo)



El ajuste parece razonablemente bueno



Un test χ2 no rechaza la Hipótesis de normalidad al 5%

Distribución muestral de coeficientes de la regresión

Los estimadores de a y b son insesgados y, en las hipótesis hechas, sus distribuciones son gaussianas, siendo sus desviaciones estándar respectivas:

$$\sigma_a = \sigma_e \left[\frac{\sum_1^n x_i^2}{n \sum_1^n (x_i - \bar{x})^2} \right]^{1/2}$$

y

$$\sigma_b = \frac{\sigma_e}{\left[\sum_1^n (x_i - \bar{x})^2 \right]^{1/2}}$$

Sin embargo, como $\hat{s}_e$ es una estimación, hay que usar la distribución t de Student con n-2 grados de libertad.

Distribución muestral de coeficientes de la regresión

Por ejemplo, para hacer una prueba en que la hipótesis nula sea H0: b = 0, contra la hipótesis H1: b ≠ 0 ,

se tiene que el estadístico $$t = \frac{\hat{b} - 0}{\left(\hat{s}_e / \sqrt{\sum_{1}^{n} (x_i - \bar{x})^2} \right)}$$

en la hipótesis nula sigue una distribución t de Student con n-2 grados de libertad

En nuestro ejemplo, obtenemos: t = 18.7 , que, con 56 grados de libertad, es muy significativa (a menos del 0.1%), por lo que se rechaza la hipótesis nula.

Es decir que un intervalo de confianza,del 99.9% no contiene al valor b=0.

No hay que olvidar que los datos pueden no ser independientes

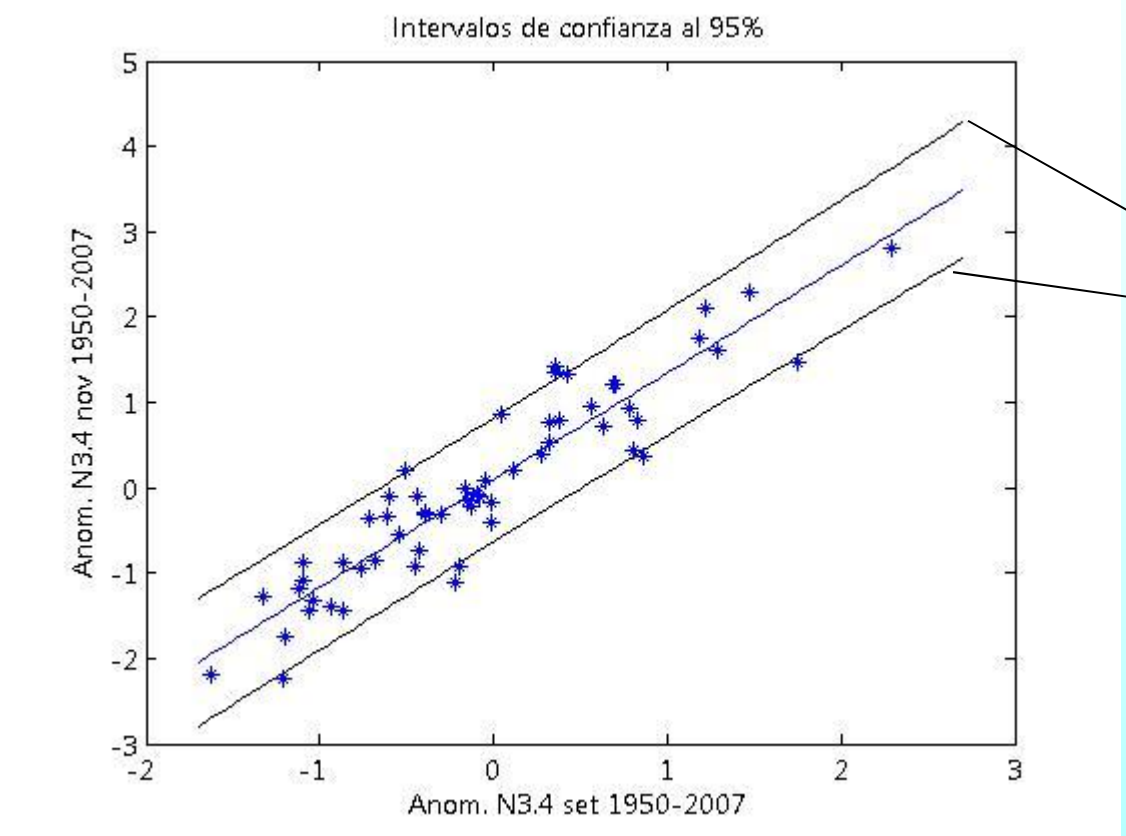
Intervalos de confianza de la “predicción”

Puede interesar hallar intervalos de confianza para $\hat{y}(x_0)$ siendo x_0 un valor cualquiera, independiente de los utilizados para construir el modelo.

Se obtiene:

$$s^2_{\hat{y}} = s^2_e \left[1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_1^n (x_i - \bar{x})^2} \right]$$

Intervalos de confianza de la “predicción”



No son rectas!

$$\hat{y}(x_0) \pm t_{\left(\frac{1+p}{2}\right)} \hat{s}_e \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_1^n (x_i - \bar{x})^2}}$$