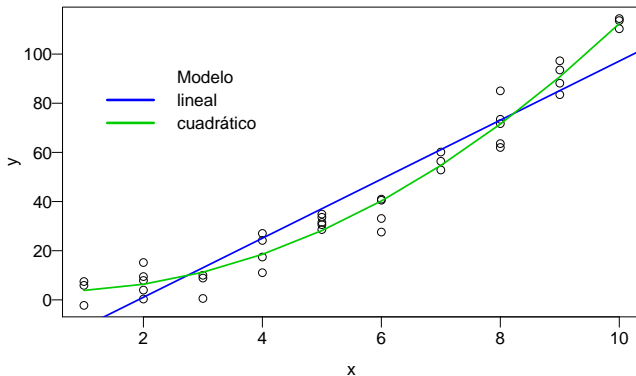


4 RLS – Validación



Contenido

1 Introducción

- Modelación Estadística

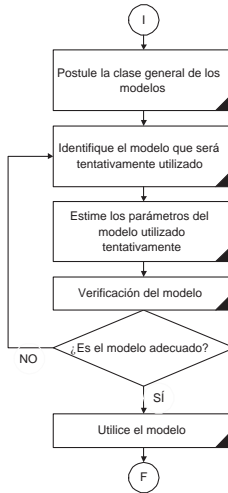
2 Modelo correcto

- Error Puro y Falta de Ajuste

3 Análisis de residuales

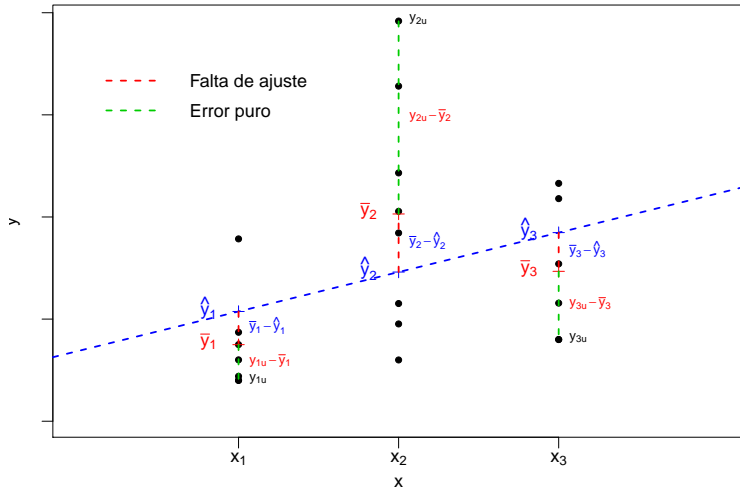
- Varianza constante
- Autocorrelación
- Normalidad
- Posibles remedios

Modelación Estadística¹



¹Box and Jenkins (1970, p19)

Modelo Correcto: Error Puro y Falta de Ajuste



Considere los datos:

$$y_{iu}; \quad u = 1, \dots, n_i; \quad i = 1, \dots, m$$

la u -ésima observación de la respuesta al nivel $X = x_i$. Hay en total $n = \sum_{i=1}^m n_i$ observaciones.

★ La *Suma de cuadrados debida al error puro*:

$$SC_{\text{Error Puro}} = SC_{EP} = \sum_{i=1}^m \sum_{u=1}^{n_i} (y_{iu} - \bar{y}_{i.})^2$$

con $\sum_{i=1}^m (n_i - 1) = n - m$ grados de libertad y donde $\bar{y}_{i.} = \frac{1}{n_i} \sum_{u=1}^{n_i} y_{iu}$ es la respuesta promedio al nivel $X = x_i$.

★ El *Cuadrado medio debido al error puro*:

$$CM_{EP} = SC_{EP} / (n - m)$$

es un estimador de σ^2 , independientemente del modelo.

Note que SC_{EP} es parte de la SC_{Resid} pues

$$\begin{aligned} (y_{iu} - \hat{y}_i) &= (y_{iu} - \bar{y}_{i.}) + (\bar{y}_{i.} - \hat{y}_i) \\ \text{Residual} &= \text{Error Puro} + \text{Falta de Ajuste} \end{aligned}$$

Error Puro y Falta de Ajuste

En la presencia de réplicas puras la *Suma de cuadrados de los residuales* se puede descomponer como

$$\begin{aligned} \sum_{i=1}^m \sum_{u=1}^{n_i} (y_{iu} - \hat{y}_i)^2 &= \sum_{i=1}^m \sum_{u=1}^{n_i} (y_{iu} - \bar{y}_{i\cdot})^2 + \sum_{i=1}^m n_i (\bar{y}_{i\cdot} - \hat{y}_i)^2 \\ \text{SC}_{\text{Residuales}} &= \text{SC}_{\text{Error Puro}} + \text{SC}_{\text{Falta de Ajuste}} \\ \text{grados de libertad} &= \\ (n - 2) &= (n - m) + (m - 2) \end{aligned}$$

Bajo los supuestos del modelo, $\text{CM}_{\text{EP}} = \text{SC}_{\text{EP}} / (n - m)$ y $\text{CM}_{\text{FA}} = \text{SC}_{\text{FA}} / m - 2$ son estimaciones independientes de σ^2 y su cociente sería aproximadamente 1. De hecho, bajo los supuestos del modelo:

$$\hat{F} = \frac{\text{CM}_{\text{FA}}}{\text{CM}_{\text{EP}}} \sim F_{(m-2, n-m)}$$

Entonces, si

$$\hat{F} > F_{(1-\alpha; m-2, n-m)} \quad \text{entonces} \quad \text{"El modelo no es correcto"}$$

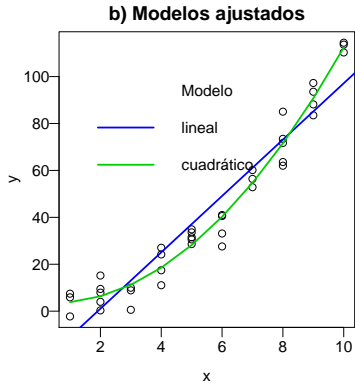
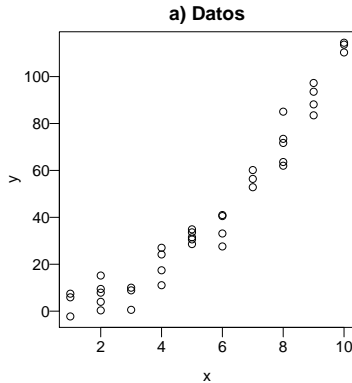
Ejemplo Simulado

Datos:

i	x_i	y_i	\bar{y}_i	\hat{y}_{1i}	\hat{y}_{2i}	i	x_i	y_i	\bar{y}_i	\hat{y}_{1i}	\hat{y}_{2i}
1	1	-2.25	3.71	-10.96	3.84	21	6	33.11	35.56	49.10	40.25
2	1	7.46	3.71	-10.96	3.84	22	6	27.60	35.56	49.10	40.25
3	1	5.90	3.71	-10.96	3.84	23	6	40.54	35.56	49.10	40.25
4	2	0.34	7.38	1.05	6.34	24	6	40.98	35.56	49.10	40.25
5	2	15.19	7.38	1.05	6.34	25	7	56.39	56.46	61.11	54.69
6	2	3.98	7.38	1.05	6.34	26	7	60.16	56.46	61.11	54.69
7	2	9.47	7.38	1.05	6.34	27	7	52.84	56.46	61.11	54.69
8	2	7.90	7.38	1.05	6.34	28	8	62.01	71.17	73.13	71.52
9	3	10.01	6.49	13.06	11.24	29	8	73.47	71.17	73.13	71.52
10	3	0.57	6.49	13.06	11.24	30	8	85.05	71.17	73.13	71.52
11	3	8.87	6.49	13.06	11.24	31	8	63.58	71.17	73.13	71.52
12	4	11.06	19.93	25.08	18.52	32	8	71.72	71.17	73.13	71.52
13	4	24.18	19.93	25.08	18.52	33	9	93.55	90.62	85.14	90.75
14	4	27.03	19.93	25.08	18.52	34	9	97.28	90.62	85.14	90.75
15	4	17.43	19.93	25.08	18.52	35	9	83.49	90.62	85.14	90.75
16	5	30.60	31.84	37.09	28.19	36	9	88.16	90.62	85.14	90.75
17	5	34.89	31.84	37.09	28.19	37	10	110.30	112.79	97.15	112.36
18	5	31.50	31.84	37.09	28.19	38	10	113.64	112.79	97.15	112.36
19	5	28.60	31.84	37.09	28.19	39	10	114.44	112.79	97.15	112.36
20	5	33.60	31.84	37.09	28.19						

donde x_i es el nivel del regresor; y_i la respuesta observada; \bar{y}_i la respuesta media observada; \hat{y}_{1i} la respuesta media ajustada por el modelo 1; y \hat{y}_{2i} la respuesta media ajustada por el modelo 2.

Ejemplo Simulado (cont.)



Ejemplo Simulado (cont.)

Ajuste modelo lineal

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	-22.975	3.652	-6.292	2.53e-07
x	12.013	0.594	20.224	< 2e-16

Residual standard error: 10.28 on 37 degrees of freedom

Multiple R-squared: 0.917, Adjusted R-squared: 0.9148

F-statistic: 409 on 1 and 37 DF, p-value: < 2.2e-16

Analysis of Variance Table

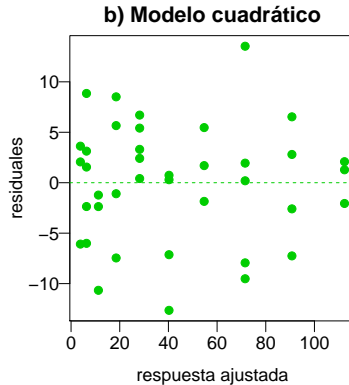
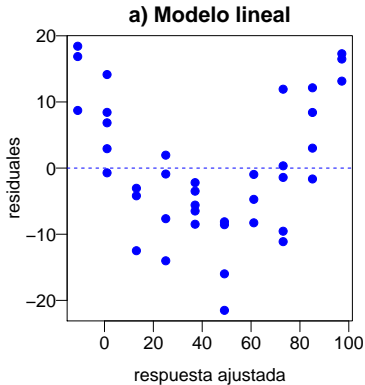
Response: y

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
x	1	43255	43255	409.02	< 2.2e-16
Residuals	37	3913	106		

SCRes	SCEP	SCFA	F	p
3.912829e+03	1.021897e+03	2.890932e+03	1.025507e+01	2.090028e-06

Ejemplo Simulado (cont.)

Análisis de Residuales



Ejemplo Simulado (cont.)

Ajuste modelo cuadrático

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	3.7195	3.7326	0.996	0.326
x	-1.0763	1.5490	-0.695	0.492
x2	1.1940	0.1378	8.665	2.47e-10

Residual standard error: 5.935 on 36 degrees of freedom

Multiple R-squared: 0.9731, Adjusted R-squared: 0.9716

F-statistic: 651.5 on 2 and 36 DF, p-value: < 2.2e-16

Analysis of Variance Table

Response: y

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
x	1	43255	43255	1227.983	< 2.2e-16
x2	1	2645	2645	75.084	2.467e-10
Residuals	36	1268	35		

SCRes	SCEP	SCFA	F	p
1268.0666396	1021.8971671	246.1694725	0.8732428	0.5390620

Error Puro y Falta de Ajuste

$$\eta_i = \mathbb{E}[y_i] = \mathbb{E}[y|x = x_i]$$

es el valor esperado de la respuesta y al nivel del regresor $x = x_i$. Y sea

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

el valor ajustado del modelo al mismo nivel.

$$\begin{aligned}\hat{e}_i = (y_i - \hat{y}_i) &= (y_i - \hat{y}_i) - \mathbb{E}[y_i - \hat{y}_i] + \mathbb{E}[y_i - \hat{y}_i] \\ &= \underbrace{[(y_i - \hat{y}_i) - (\eta_i - \mathbb{E}[\hat{y}_i])]}_{q_i} + \underbrace{(\eta_i - \mathbb{E}[\hat{y}_i])}_{b_i}\end{aligned}$$

donde b_i es el *sesgo al nivel* $x = x_i$.

- Si el modelo es correcto $\mathbb{E}[\hat{y}_i] = \eta_i \implies b_i = 0$
- Por otro lado, $\mathbb{E}[q_i] = 0$ *independientemente* del modelo.
- Se puede mostrar que los q_i son correlacionados y $\mathbb{E}[\sum q_i] = (n-2)\sigma^2$. De donde,

$$\mathbb{E}[s^2] = \mathbb{E}\left[\frac{1}{n-2} \sum (y_i - \hat{y}_i)^2\right] = \begin{cases} \sigma^2 & \text{si el modelo es correcto} \\ \sigma^2 + \frac{1}{n-2} \sum b_i^2 & \text{si el modelo **no** es correcto} \end{cases}$$

Análisis de Residuales

Mediante los residuales se intenta verificar si los supuestos del modelo se satisfacen.

Supuestos

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \quad i = 1, \dots, n$$

$$\epsilon \sim N(0, \sigma^2), \quad \text{v.a.i.i.d.}$$

$$\hat{e}_i = y_i - \hat{y}_i = y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)$$

- Los residuales son “errores observados” *si el modelo es correcto*. Pero por las ecuaciones normales se tiene dependencia entre ellos.
- ¿Sugieren los residuales que los supuestos no se satisfacen?
- Análisis más sofisticados: pruebas estadísticas formales.
- Definición de otros residuales.

Análisis de Residuales

Varianza constante

- Prueba de *Bartlett*, prueba de *Levene*.
- Prueba de *Breusch y Pagan*, prueba de *White*.
- ★ Gráficas de residuales: $(\hat{e}_i \text{ vs. } \hat{y}_i)$; $(\hat{e}_i \text{ vs. } i)$; $(\hat{e}_i \text{ vs. } x_{ji})$.
- $(\hat{e}_i \text{ vs. } y_i)$ no se grafican por estar correlacionados.

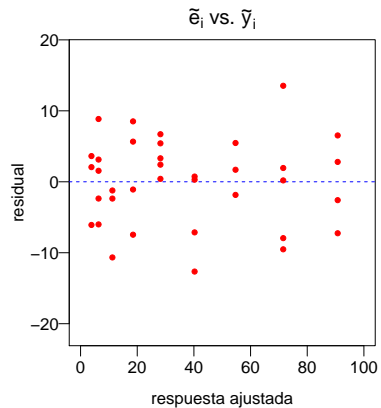
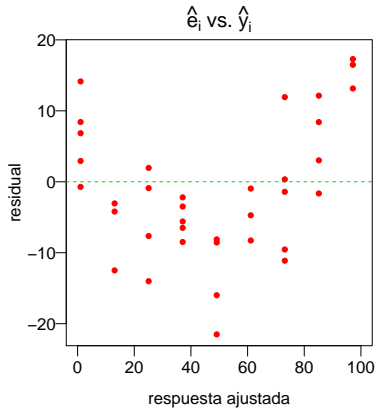
Correlación

- Significancia de la autocorrelación de residuales \hat{e}_t .
- Correlogramas.
- Prueba de *Durbin-Watson*.
- ★ Gráficas de residuales: $(\hat{e}_{i-1} \text{ vs. } \hat{e}_i)$.

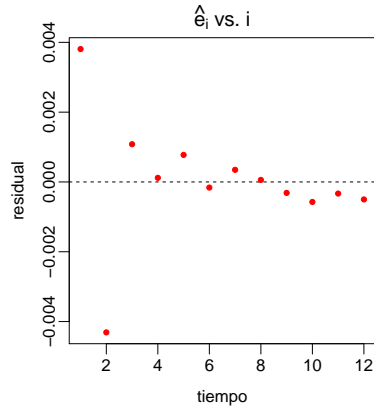
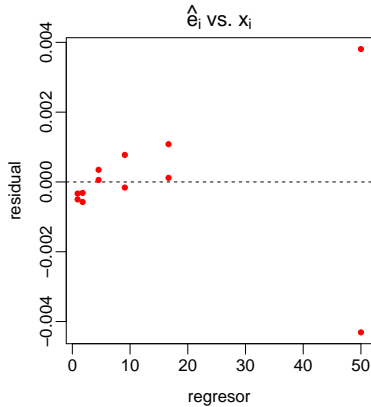
Normalidad

- Pruebas de Bondad de Ajuste: χ^2 ; *Kolmogorov-Smirnov*; *Anderson-Darling*; *Jarque-Bera*, *Cramér-Von Mises*, *Lilliefors*, etc.
- ★ Gráficas de residuales \hat{e} en *papel de probabilidad normal (cuantil-cuantil normal)*.

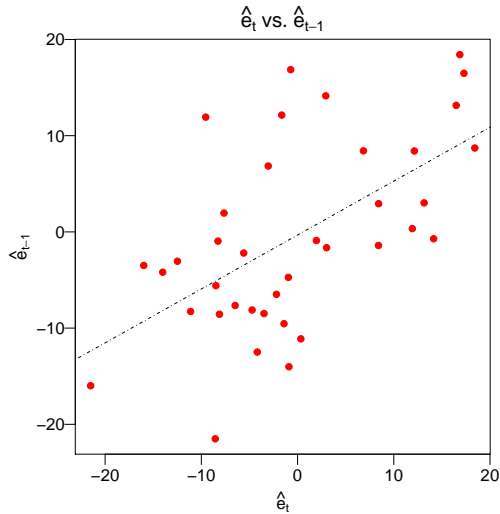
Varianza constante



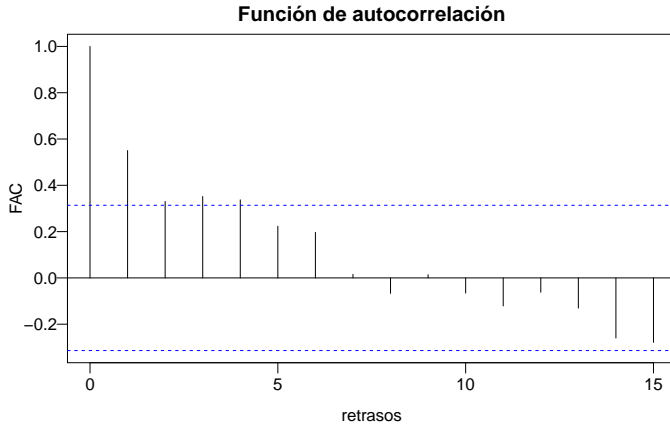
Varianza constante



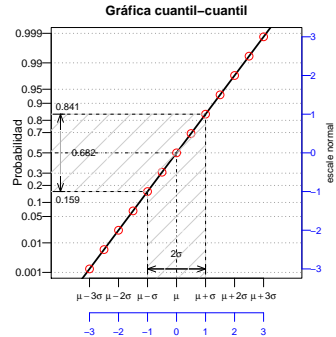
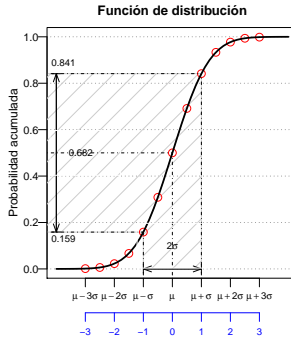
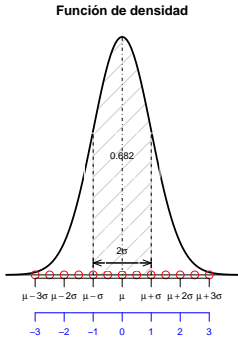
Autocorrelación



Autocorrelación



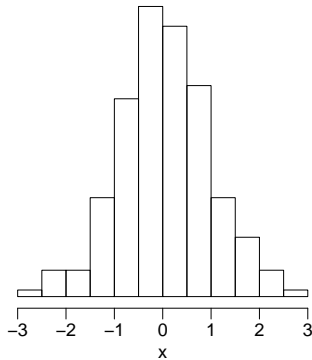
Gráfica cuantil-cuantil normal



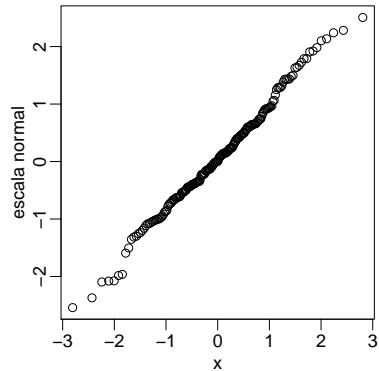
Normalidad

Ejemplo simulado: 200 observaciones normales

Histograma



Gráfica cuantil–cuantil

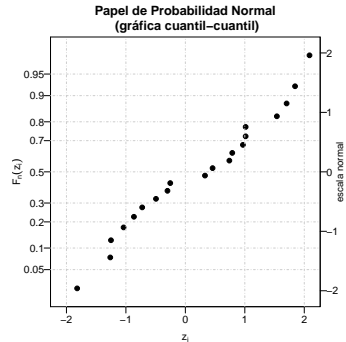


Normalidad

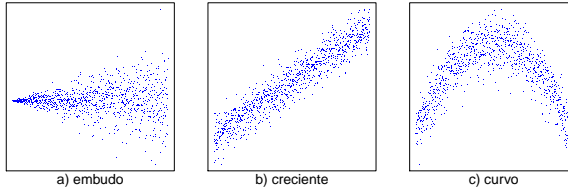
Ejemplo

- 1 Ordene la muestra $\{z_1, \dots, z_n\}$: $x_1 = z_{(1)}, \dots, x_n = z_{(n)}$
- 2 Calcule la **probabilidad empírica acumulada**: $y_i = F_n(x_i) = \frac{i-1/2}{n}$
- 3 Grafique (x_i, y_i) , $i = 1, \dots, n$ en **papel de probabilidad normal (cuantil-cuantil)**.

i	z_i	ord.	x_i	$F_n(x_i)$
1	1.703	9	-1.820	0.025
2	-0.865	17	-1.263	0.075
3	-0.301	11	-1.251	0.125
4	1.014	18	-1.039	0.175
5	2.088	2	-0.865	0.225
6	0.740	13	-0.726	0.275
7	-0.254	16	-0.494	0.325
8	0.968	3	-0.301	0.375
9	-1.820	7	-0.254	0.425
10	1.014	14	0.330	0.475
11	-1.251	12	0.458	0.525
12	0.458	6	0.740	0.575
13	-0.726	20	0.789	0.625
14	0.330	8	0.968	0.675
15	1.845	4	1.014	0.725
16	-0.494	10	1.014	0.775
17	-1.263	19	1.539	0.825
18	-1.039	1	1.703	0.875
19	1.539	15	1.845	0.925
20	0.789	5	2.088	0.975



Posibles remedios cuando residuales insatisfactorios²



Patrón:	Gráfica de \hat{e}_i versus:		
	Orden temporal	Respuesta ajustada \hat{y}_i	Valores x_{ji}
a) Embudo indicando varianza no constante	Uso de Mínimos Cuadrados Ponderados	Uso de Mínimos Cuadrados Ponderados o transformación de la y_i	Uso de Mínimos Cuadrados Ponderados o transformación de la y_i
b) Banda ascendente o descendente	Considere incluir un término lineal en el tiempo	Error en el análisis u omisión de β_0	Error en los cálculos. Efecto de primer orden de X_j no eliminado
c) Banda curva	Considere incluir términos lineal y cuadrático en el tiempo	Considere añadir términos extra al modelo o transformar la respuesta y_i	Considere añadir términos extra al modelo o transformar la respuesta y_i

²Draper and Smith (1998) p. 64