

BIOMETRÍA II

CLASE 3

SUPUESTOS DE LOS MODELOS LINEALES

Adriana Pérez
Depto de Ecología, Genética y Evolución
FCEN, UBA

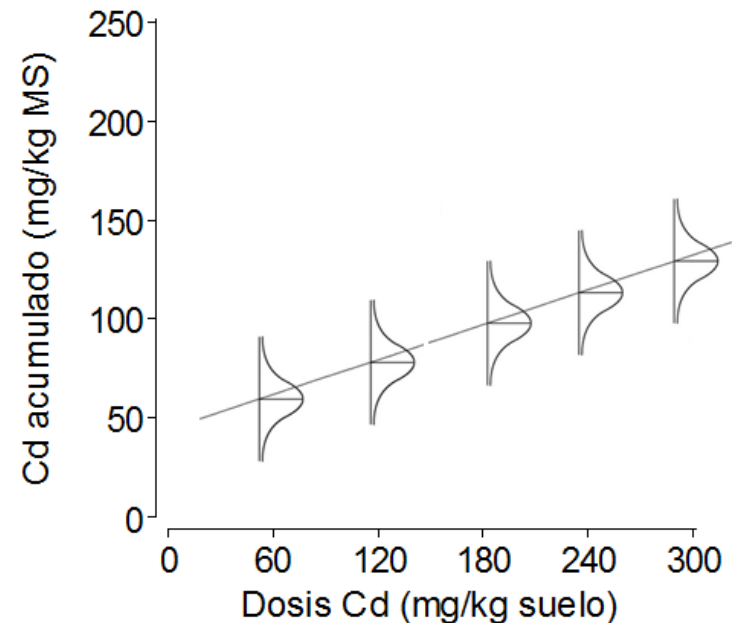
Restauración con césped de suelos contaminados con cadmio



2

- Interesa estudiar la capacidad detoxificadora del césped *Eremochloa ophiuroides* en suelos contaminados con Cd
- A 20 macetas con césped se les asignará una de 5 dosis de Cd diferente (60, 120, 180, 240 y 300 mg Cd kg⁻¹); 4 macetas por dosis
- Luego de 36 días en invernadero se medirá el Cd acumulado por la planta (expresado como mg Cd kg⁻¹ materia seca)
- Se sospecha una relación lineal

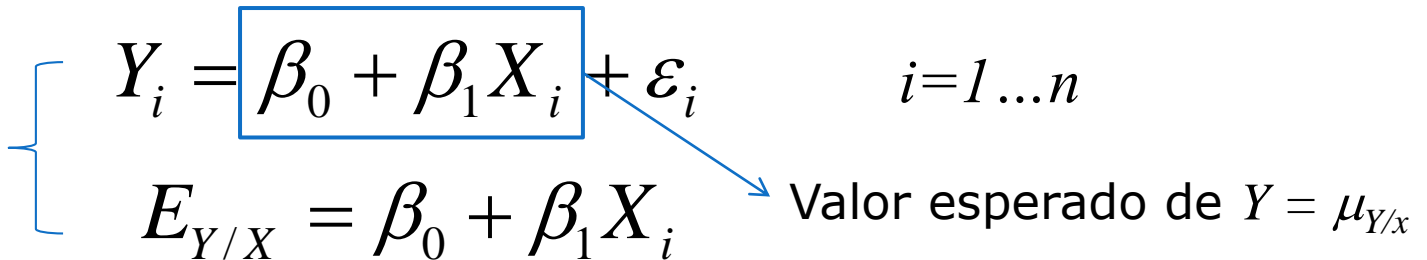
- UE
- VR (Y)
- VE (X)
- Réplicas
- Modelo



Modelo de regresión lineal simple

3

equivalentes $\left\{ \begin{array}{l} Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \\ E_{Y/X} = \beta_0 + \beta_1 X_i \end{array} \right. \quad i=1 \dots n$



- Y_i es la i -ésima observación de la variable dependiente Y
- X_i es el i -ésimo valor de la variable predictora X
- β_0 y β_1 son los **parámetros** ordenada al origen y pendiente (o coeficiente de regresión)
 - Si el alcance del modelo incluye a $X=0$, β_0 es el valor esperado de Y cuando $X=0$
 - β_1 indica el cambio esperado en Y por cada aumento unitario de X
- ε_i es el error aleatorio, variación de Y no explicada por X_i

$$\varepsilon_i \sim NID(0, \sigma^2)$$

Dosis Cd
(mg Cd/kg)

Concentración Cd
(mg Cd/kg MS)

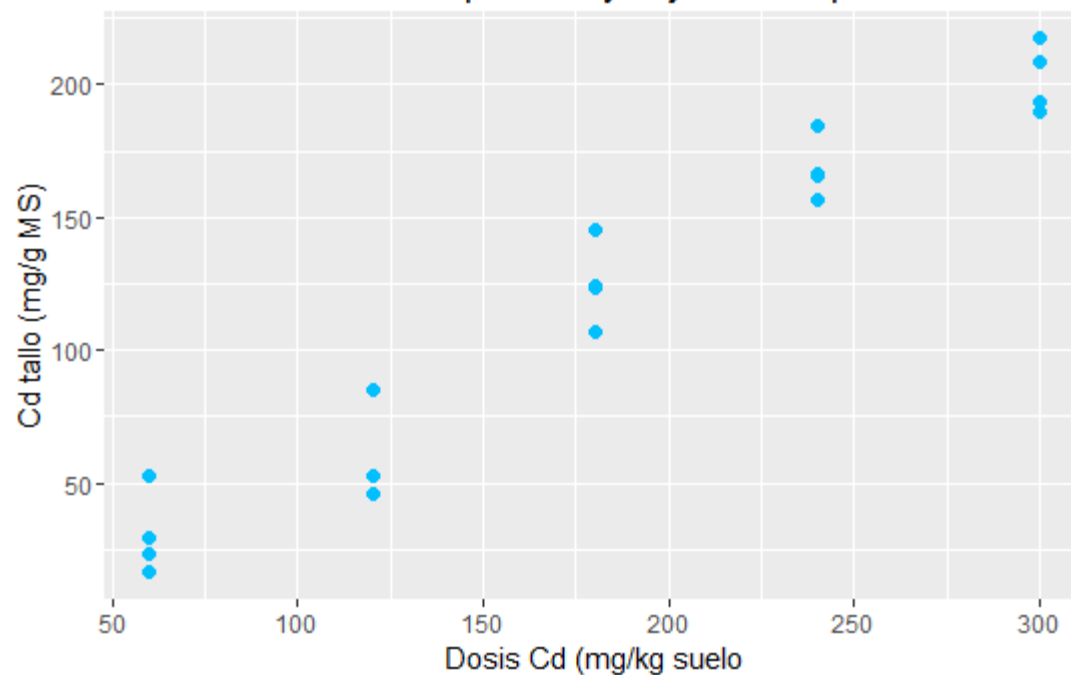
Tallo y hojas

Raíz

```
> summary(cadmio)
```

dosis_cd	cd_tallo	cd_raiz
Min. : 60	Min. : 16.20	Min. : 104.6
1st Qu.:120	1st Qu.: 52.65	1st Qu.: 245.1
Median :180	Median :123.70	Median : 465.0
Mean :180	Mean :117.02	Mean : 502.8
3rd Qu.:240	3rd Qu.:171.18	3rd Qu.: 664.6
Max. :300	Max. :217.70	Max. :1067.1

Absorción de Cd por tallo y hojas de *E.ophiuroides*



```
modelo1<-lm(cd_tallo~dosis_cd, data=cadmio)
```

Estimación de los parámetros del modelo

5

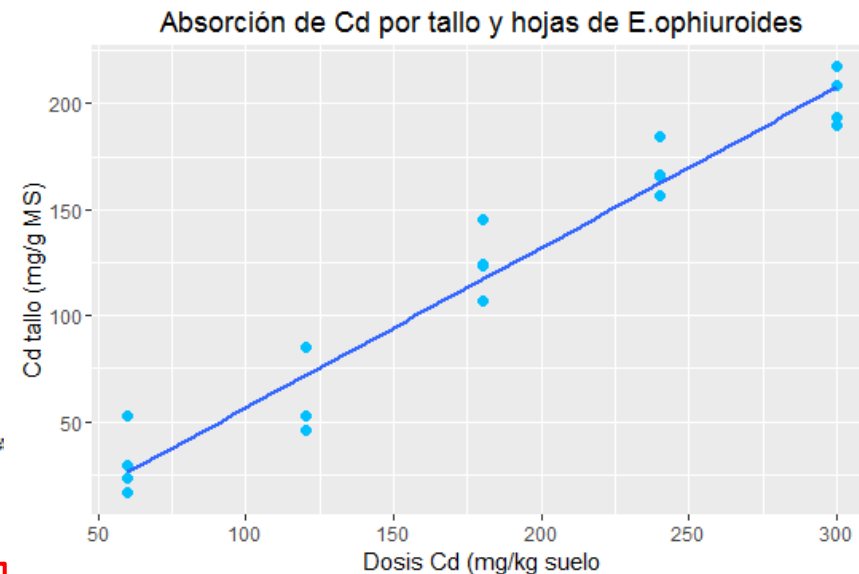
```
> summary(model1)
```

```
Call:
lm(formula = cadmio$cd_tallo ~ cadmio$dosis_cd)

Residuals:
    Min       1Q   Median       3Q      Max
-25.970 -11.220   1.730   7.655  28.680

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -19.03000    8.35547  -2.278   0.0352 *
cadmio$dosis_cd  0.75583    0.04199  18.001 5.88e-13 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 15.93 on 18 degrees of freedom
Multiple R-squared:  0.9474,    Adjusted R-squared:  0.9445
F-statistic:  324 on 1 and 18 DF,  p-value: 5.882e-13
```



$$\hat{y} = -19,03 + 0,76x$$

$$Cd \text{ tallo} = -19,03 + 0,76 \cdot Cd \text{ suelo}$$

Si los errores son independientes y su distribución es normal, los estimadores por mínimos cuadrados son los estimadores por máxima verosimilitud

Calculando varianzas

Grados de libertad: Piezas de información independiente = n – cantidad de parámetros que se debieron estimar previamente. Dividir por GL en vez de por n asegura que el estimador de σ^2 sea insesgado

6

	dosis	Cd	Cd	tallo	Predichos	Residuos
1		60		23.2	26.32	-3.12
2		60		16.2	26.32	-10.12
3		60		52.7	26.32	26.38
4		60		29.1	26.32	2.78
5		120		52.5	71.67	-19.17
6		120		45.7	71.67	-25.97
7		120		52.9	71.67	-18.77
8		120		84.9	71.67	13.23
9		180		123.5	117.02	6.48
10		180		106.9	117.02	-10.12
11		180		123.9	117.02	6.88
12		180		145.7	117.02	28.68
13		240		166.8	162.37	4.43
14		240		165.9	162.37	3.53
15		240		184.3	162.37	21.93
16		240		157.0	162.37	-5.37
17		300		208.4	207.72	0.68
18		300		189.9	207.72	-17.82
19		300		217.7	207.72	9.98
20		300		193.2	207.72	-14.52

$$S^2_Y = \frac{\sum_{i=1}^n (y_i - \bar{y})^2}{n-1} = \frac{86834.53}{19} = 4570.23 (mg/g MS)^2$$

$$S^2_e = S^2_{Y/X} = CM_{error} = \frac{\sum_{i=1}^n e_i^2}{n-2} = \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{n-2} =$$

$$= \frac{\sum_{i=1}^n (y_i - (\hat{\beta}_0 - \hat{\beta}_1 x_i))^2}{n-2} = \frac{4569.63}{18} = 253,87 (mg/g MS)^2$$

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

$$e_i = y_i - \hat{y}_i$$

Residuos
estandarizados

$$RE = \frac{e_i}{\sqrt{S_e^2}}$$

Variación total de VR =
variación explicada por el modelo + Variación no explicada (error o residual)

Supuestos del modelo

7

- X medida sin error, no es V.A., sus valores son determinados por el investigador

Muchas veces no se cumple; pero es grave sólo si la magnitud del error de X es grande en relación a la magnitud de X (ie, > 10%). En ese caso, Modelo II

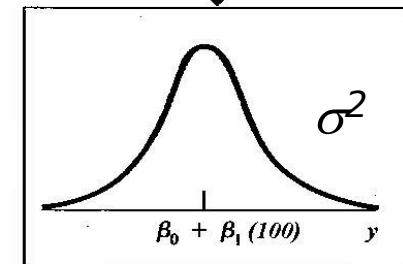
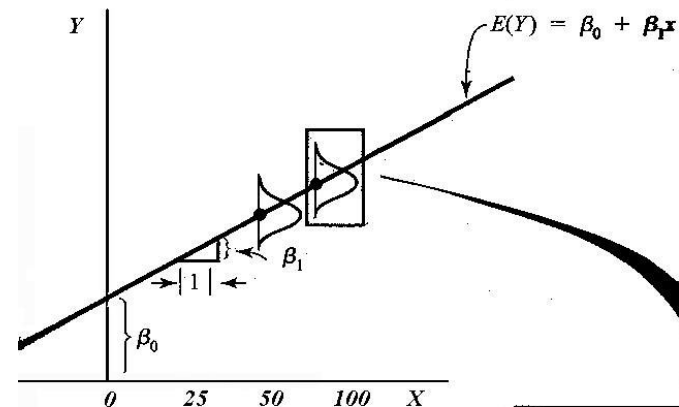
- **Independencia** entre las observaciones (no se cumple para medidas repetidas o datos estructurados espacialmente). No debe existir correlación entre observaciones. $cov(i;j) = 0$ para $i \neq j$.

Supuestos del modelo

No es necesarios para
estimar β_0 y β_1 pero sí para
hacer inferencia

8

- Para cada valor de X existe una subpoblación de Y
 - La media de cada una de estas subpoblaciones es
 $E_{Y/X} = \beta_0 + \beta_1 X_i$ (linealidad)
 - La distribución de cada subpoblación es normal $Y_{i/X} \approx NID(\mu_{Y/X}, \sigma^2)$
 - las varianzas de las subpoblaciones son iguales, es decir que el modelo asume una varianza constante σ^2 , sin importar el nivel de X $\text{Var}[Y/X] = \sigma^2$



Estos supuestos se pueden resumir en: $\varepsilon_i \approx NID(0, \sigma^2)$

Los residuos constituyen el insumo básico
para estudiar los supuestos del modelo

Linealidad y homocedasticidad

9

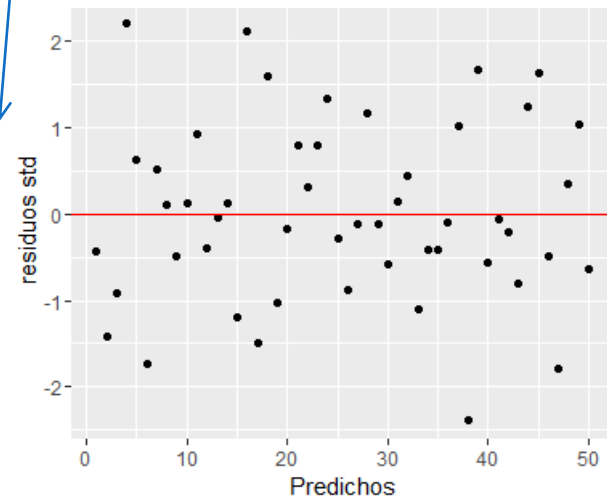
Gráficamente: **gráfico de dispersión de residuos vs predichos**

- Determinar si el modelo lineal está bien especificado (los residuos deberían distribuirse aleatoriamente, sin patrones)
- Determinar si la variabilidad es constante (homocedasticidad)
- Detectar **outliers** o **datos atípicos en Y** (con residuo muy grande)

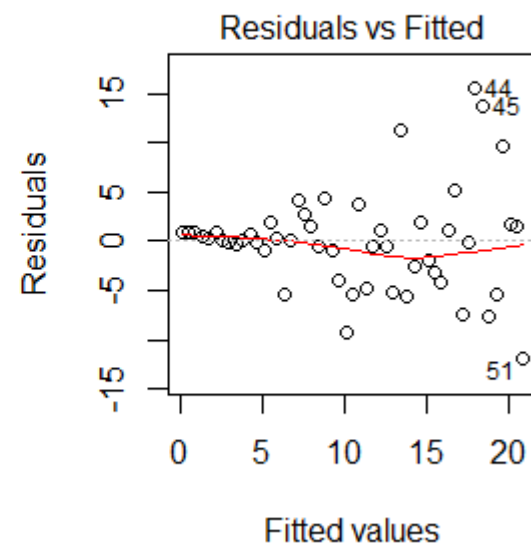
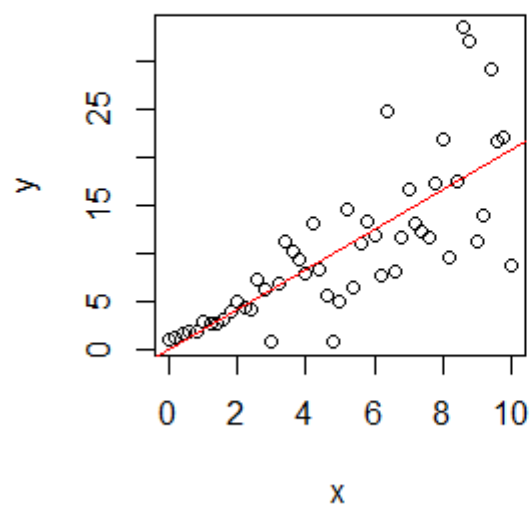
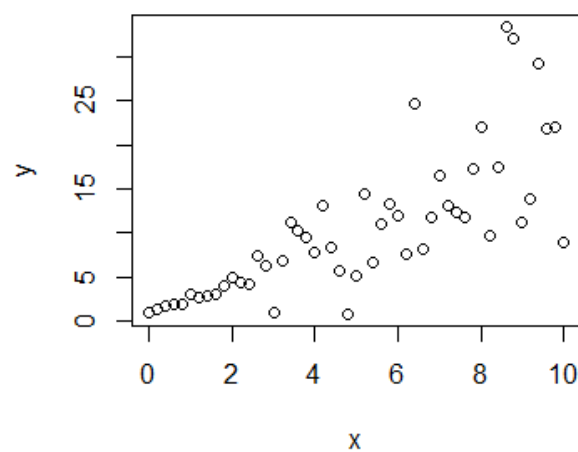
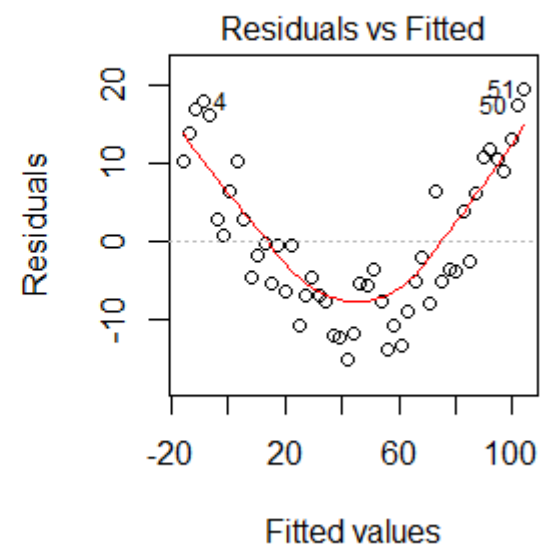
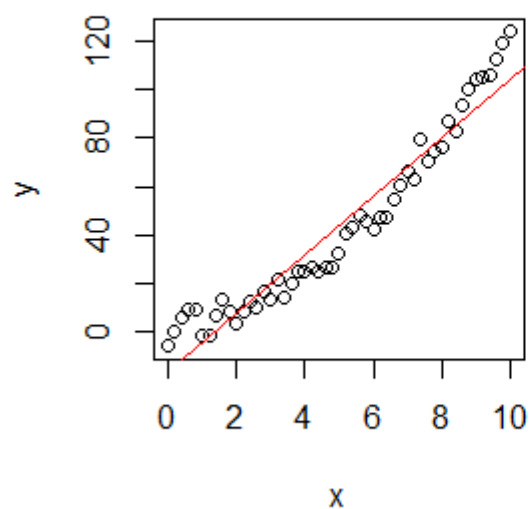
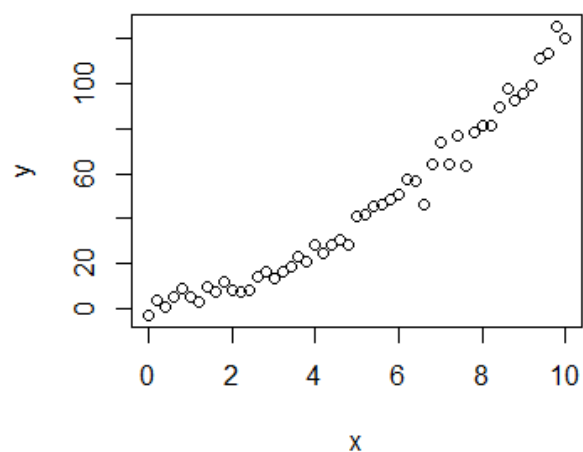
Se espera encontrar una distribución al azar (sin patrones) y con variabilidad constante

Conviene utilizar residuos estandarizados, ya que permite detectar outliers ($RE > 2$ o $RE < -2$)

Gráfico de dispersión de residuos vs predichos



El valor predicho para cada observación es la respuesta obtenida a partir de la ecuación estimada



Homocedasticidad

11

- Analíticamente: [Prueba de Levene](#)
- Es un análisis de la varianza de un factor utilizando como VR el valor absoluto de los residuos
- H_0 : todas las varianzas poblacionales son iguales
- Solo se puede efectuar cuando existen réplicas para cada nivel de X (raro en estudios observacionales)

```
library(car)  
levene.test(modelo2, data=cadmio)
```

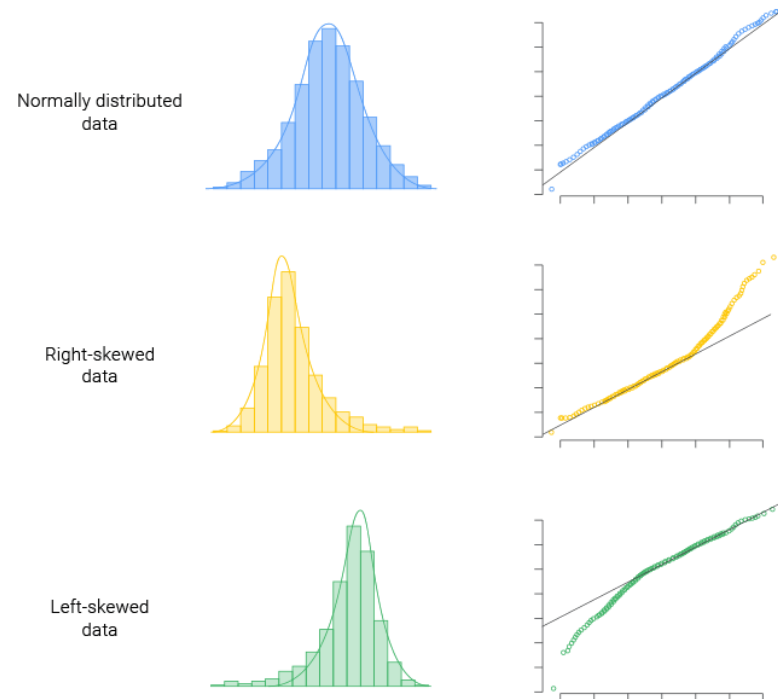
Normalidad

12

□ Gráficamente: **QQ plot**

Es un gráfico de dispersión de los percentiles (quantiles) de las observaciones vs los percentiles (cuantiles) de una distribución normal con media y DE estimados a partir de la muestra

La normalidad se estudia utilizando los residuos del modelo



Normalidad

13

- Analíticamente: Prueba de Shapiro-Wilk

$H_0: \varepsilon_i \approx normal$

El estadístico W^* de la prueba de Shapiro-Wilk oscila entre 0 y 1. Cuanto más cercano a 1 mayor evidencia de normalidad. Básicamente, mide cuan cerca de una recta está la curva que describen los puntos graficados en el QQ-plot

Los errores del modelo no son observables; para probar el supuesto se utilizan sus correlatos empíricos, los residuos

Causas del incumplimiento de los supuestos

14

- la presencia de outliers puede generar heterocedasticidad
- Si la distribución de la variable no es normal (lognormal, gamma, etc) puede detectarse tanto falta de normalidad como de homocedasticidad
- La falta de linealidad implica que la relación de la VR con la VE no es lineal. Puede solucionarse agregando más términos al modelo (cuadrático, cúbico, interacciones, etc) o tratando a las VE cuantitativas como cualitativas

Consecuencias del incumplimiento de los supuestos

15

Heterocedasticidad

Las estimaciones de los parámetros son insesgadas y consistentes, pero los errores estándares de los estimadores no \Rightarrow la inferencia (Pruebas de hipótesis e IC) no es confiable; provoca un aumento en la probabilidad de cometer error tipo I; el nivel de confianza no es el propuesto

- Los efectos son más graves si el diseño es desbalanceado
- Más grave si una varianza es mucho mayor que el resto
- Menos grave si una varianza es mucho menor que el resto

No normalidad

Menos grave. Si el apartamiento de la normalidad no es severo y no hay heterocedasticidad, las estimaciones e inferencia son razonables

No linealidad

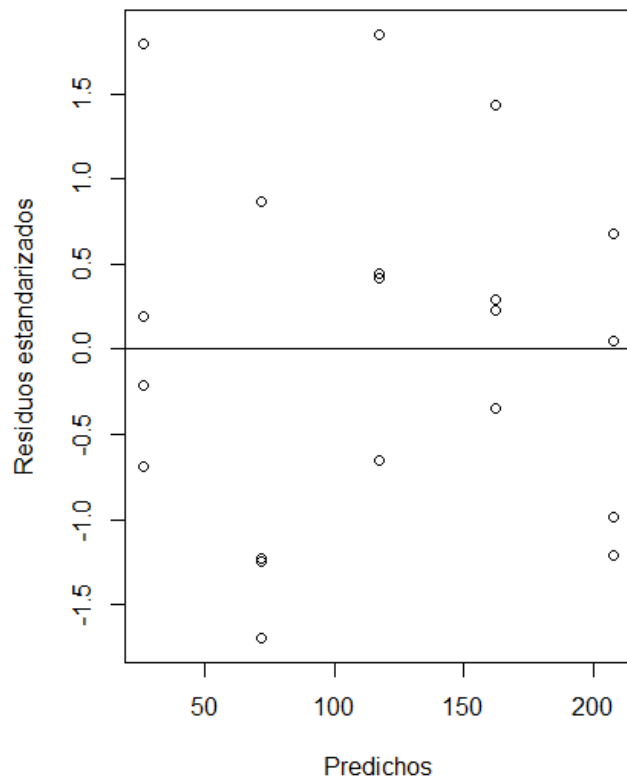
Los coeficientes de regresión no miden la verdadera relación con la VE

¿Cómo corregimos la heterocedasticidad o la falta de normalidad?

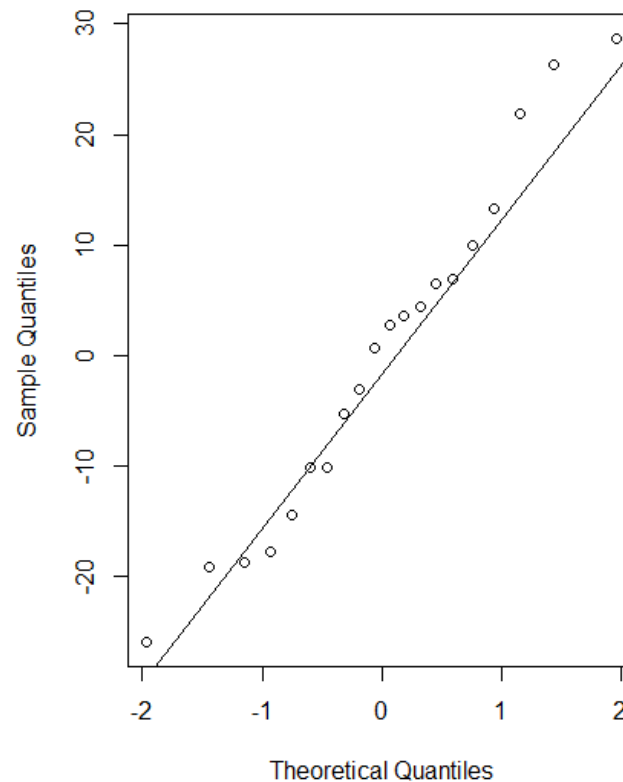
16

1. Aplicando modelos que permitan modelar la heterocedasticidad
2. Aplicando modelos que permitan otra distribución de probabilidad de la VR (**modelos lineales generalizados**)
3. Aplicando transformaciones monotónicas a los datos (i.e. aplicando logaritmo), pero que implican cambiar la escala de la VR
4. Regresión ponderada (para heterocedasticidad): menor peso a los datos con mayor dispersión
5. Métodos robustos

Gráfico de dispersión de RE vs PRED



Normal Q-Q Plot



Levene's Test for Homogeneity of Variance (center = median)

	Df	F value	Pr(>F)
group	4	0.0782	0.9878
	15		

```
> shapiro.test(e)
```

Shapiro-Wilk normality test

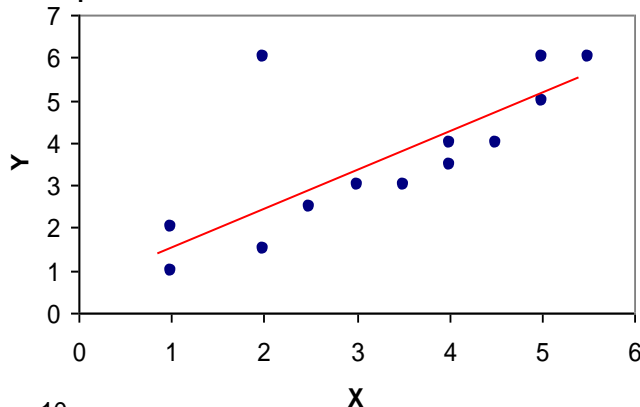
```
data: e
W = 0.9676, p-value = 0.7035
```

Observaciones atípicas y observaciones influyentes

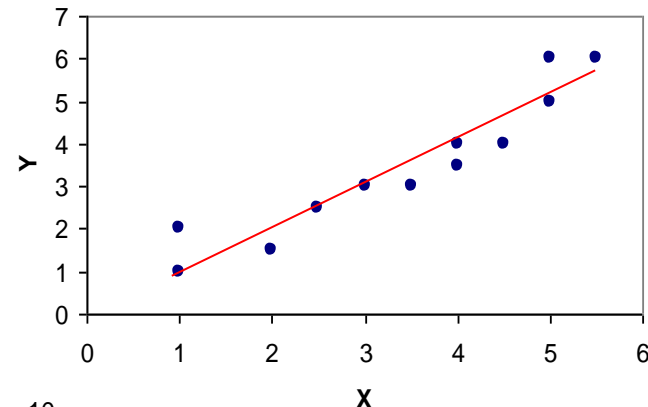
18

- **Atípicas (outliers en Y):** Observaciones con un patrón distinto al resto de los datos, que producen un residuo grande
- **Influyentes (con alta palanca):** Observaciones cuyo valor de X se encuentra alejado del promedio y que tienen mucho peso en las estimaciones de los parámetros. Al ser eliminadas pueden provocar cambios sustanciales en las estimaciones

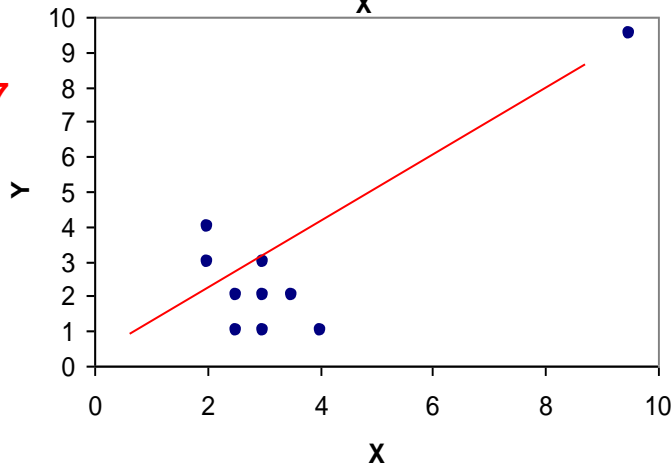
$R^2=0.52$



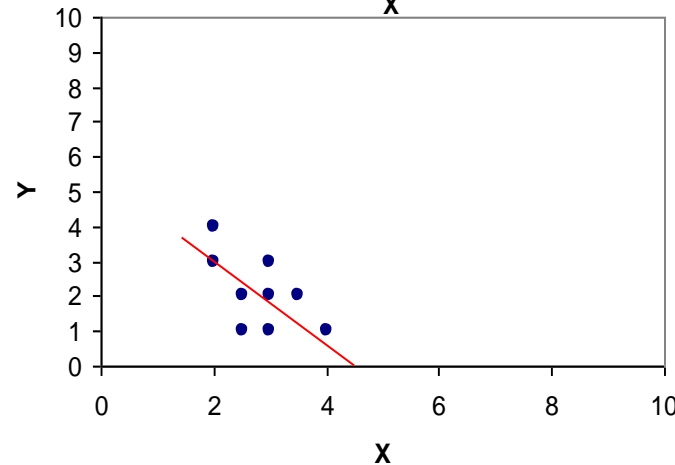
$R^2=0.89$



$R^2=0.67$



$R^2=0.36$



Cómo detectar observaciones atípicas

19

□ Residuos estandarizados

- ▣ Permite detectar outliers en Y
- ▣ Se identifican valores con $RE < -2$ o > 2

$$RE = \frac{e_i}{\sqrt{S_e^2}}$$

□ Residuos studentizados

- ▣ Permite detectar outliers en Y
- ▣ Se calculan como:
- ▣ Se identifican valores con $RS < -2$ o > 2
- ▣ h_{ii} es el **Leverage** o palanca

$$RS = \frac{e_i}{\sqrt{S_e^2 (1 - h_{ii})}}$$

Cómo detectar observaciones influyentes

20

□ Leverage o palanca h_{ii}

- Es una medida que mide cuán lejos cae el valor de X_i de la media muestral de las X (outlier en X)
- Mide, de alguna manera, cuánto es el aporte de la observación i -ésima a la varianza muestral de las X
- Puede tomar valores entre $1/n$ y 1
- Valores altos (alta palanca) indican que esa observación contribuye más en la predicción de Y , es decir que fuerza a la recta a pasar por el valor observado de Y
- Se consideran outliers en X las observaciones con $h_{ii} > 2k/n$, donde k es la cantidad de variables predictoras del modelo

$$h_{ii} = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{k=1}^n (x_k - \bar{x})^2}$$

Cómo detectar observaciones influyentes

21

□ Distancia de Cook

- Mide el efecto global de una observación sobre las estimaciones de los parámetros del modelo y sobre los valores predichos
- Grandes valores indican observaciones cuya eliminación tiene gran influencia sobre las estimaciones y sobre los valores predichos (dato influyente)

$$D_{Cook} = \left(\frac{e_i}{\sqrt{CMerror(1 - h_{ii})}} \right)^2 \left(\frac{h_{ii}}{1 - h_{ii}} \right) \left(\frac{1}{p} \right)$$
$$D_{Cook} = \frac{\sum (\hat{y}_j - \hat{y}_{j(i)})^2}{pCMerror}$$

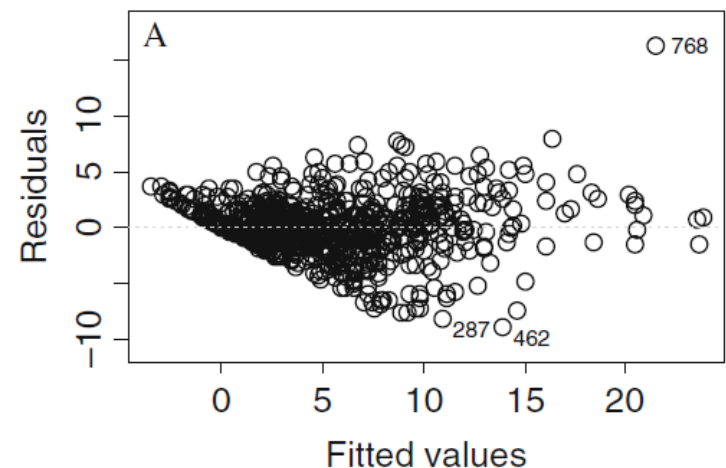
- Se consideran influyentes las observaciones con $D > 1$

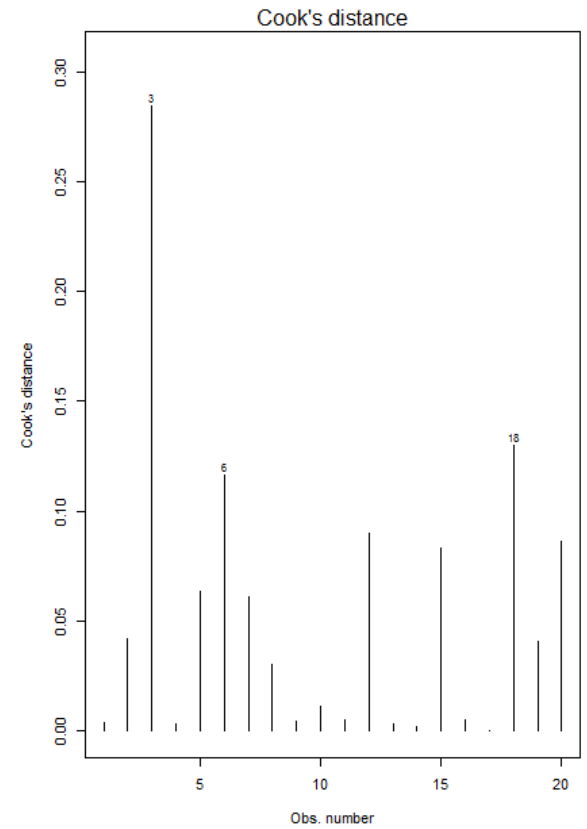
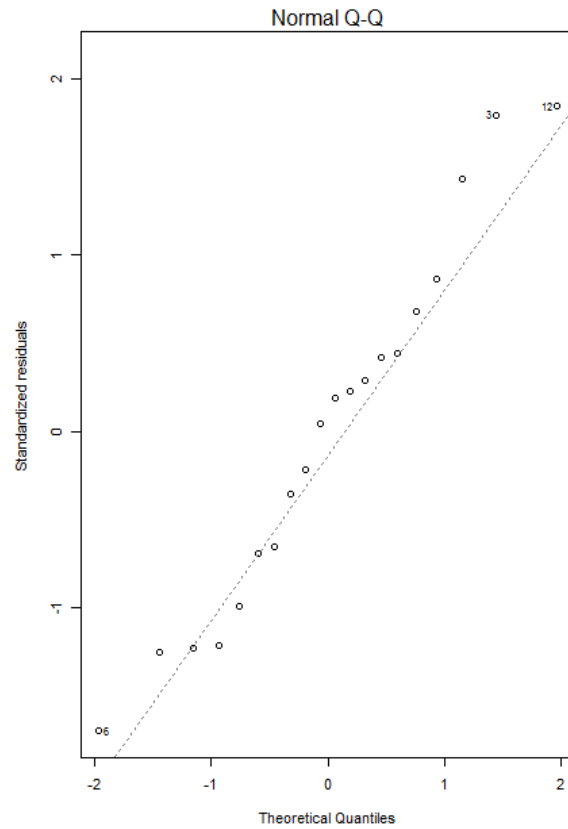
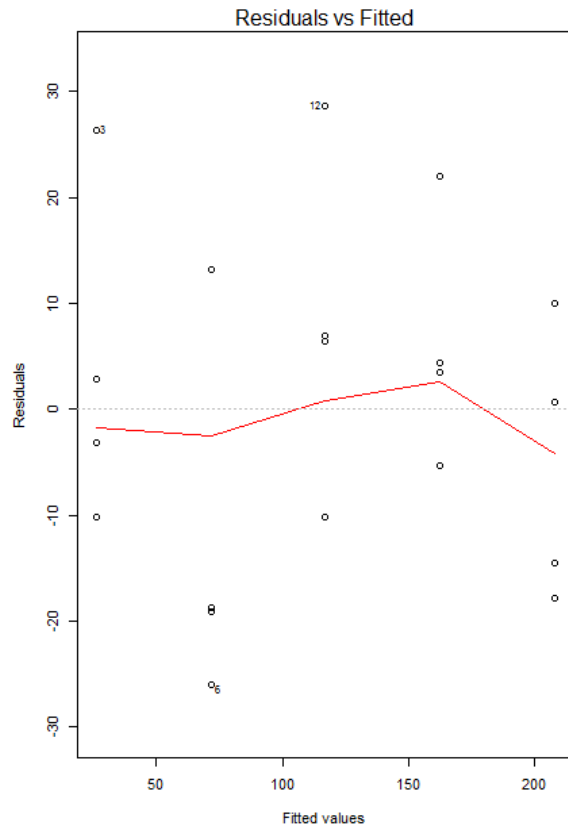
VR limitada:

Datos censurados o truncados

22

- Truncamiento: cuando por el proceso de recopilación de los datos solo se obtiene datos de un subconjunto de una población de interés más grande.
 - Por ejemplo: VR: Glucosa en plasma, pero solo participaron individuos con Glu > 110
- Censura: cuando todos los valores de un cierto rango se transforman (o se informan como) un solo valor.
 - Por ejemplo: VR: tiempo de respuesta (hasta 60 seg)
- Se detectan patrones en los residuos
- Exigen un modelado especial





No se detectan violaciones a los supuestos -> podemos hacer inferencia

```
> summary(model1)
```

Call:

```
lm(formula = cadmio$cd_tallo ~ cadmio$dosis_cd)
```

Residuals:

Min	1Q	Median	3Q	Max
-25.970	-11.220	1.730	7.655	28.680

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	-19.03000	8.35547	-2.278	0.0352	*
cadmio\$dosis_cd	0.75583	0.04199	18.001	5.88e-13	***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 15.93 on 18 degrees of freedom

Multiple R-squared: 0.9474, Adjusted R-squared: 0.9445

F-statistic: 324 on 1 and 18 DF, p-value: 5.882e-13

Ojo: diferencias
"significativas" no quiere
decir "importantes", sino
"poco probables de
obtener sólo por azar"

Alternativamente puede construirse un IC para β_1
y determinar si cero pertenece o no a dicho intervalo

$$\hat{\beta}_1 \pm t_{n-2; 1-\alpha/2} \sqrt{\frac{s^2_{error}}{\sum (x_i - \bar{x})^2}}$$

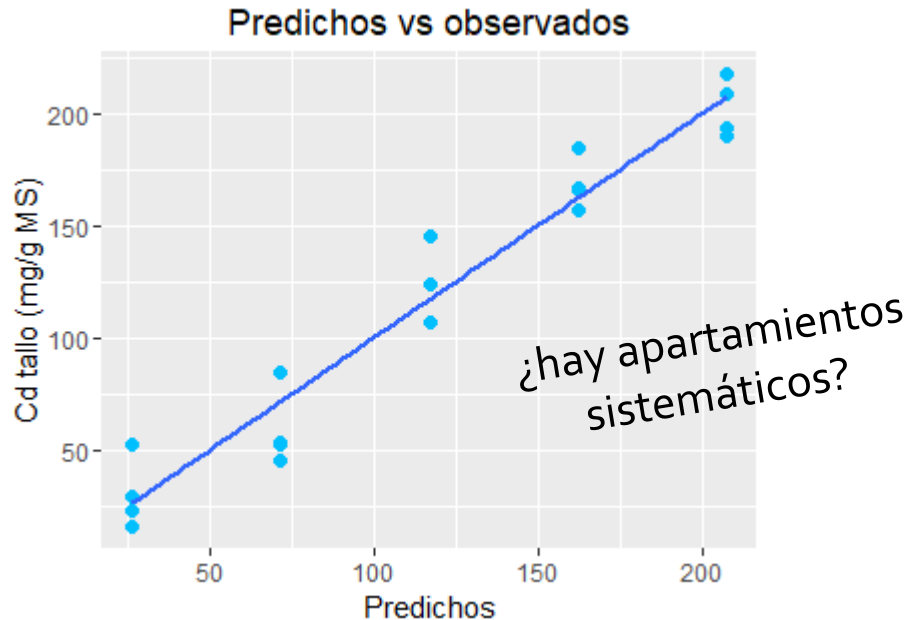
```
> round(confint(modelo1), 2) (en paquete ISwR)
```

	2.5 %	97.5 %
(Intercept)	-36.58	-1.48
cadmio\$dosis_cd	0.67	0.84

β_1 mide la magnitud
del efecto

Validación del modelo

25



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

`cor(pre, cd_tallo)`
0.9733321

- Coeficiente de determinación R^2 para evaluar cuánto de la variabilidad de Y está explicada por el modelo
- Correlación entre predichos y observados para evaluar la capacidad predictiva del modelo. Pero está sobrevaluada!
- La validación cruzada es un conjunto de métodos para medir el desempeño de un modelo evaluando su capacidad para predecir **un nuevo conjunto de datos** (próximamente)

Importancia de visualizar los datos

Intervalos de confianza para las predicciones

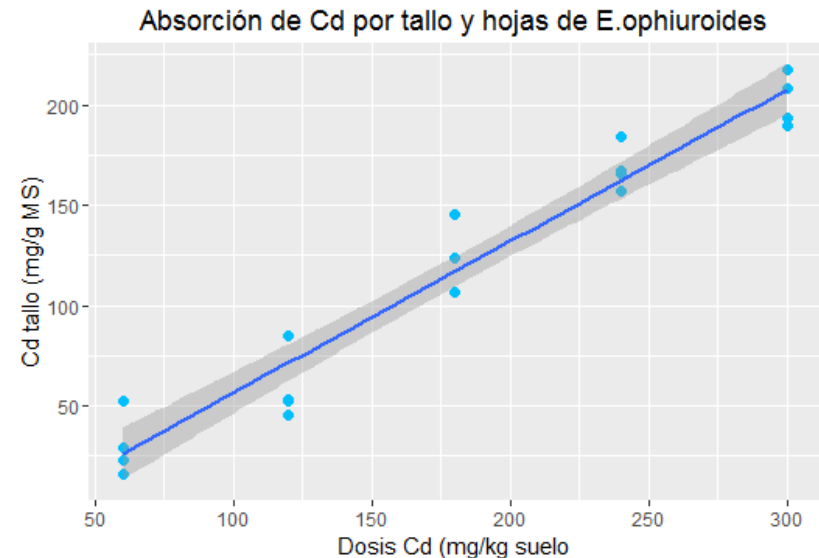
26

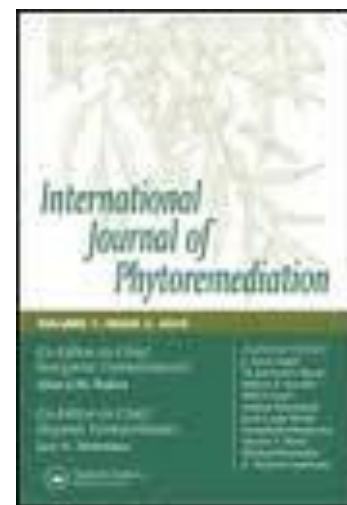
- Dos aplicaciones de los modelos de regresión: explicación y predicción
- Una vez **estimados los parámetros y validado el modelo**, es posible realizar **predicciones** acerca del valor que tomaría la VR para una unidad extramuestral.
- Se pueden construir **intervalos de confianza** sobre dichos valores
- Los pronósticos son válidos en el rango estudiado

IC para $\mu_{Y/X}$

$$\hat{y}_0 \pm t_{n-2; 1-\alpha/2} \sqrt{S_e^2 \left[\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum (x_i - \bar{x})^2} \right]}$$

Penaliza el
alejamiento
del centro





PHYSIOLOGICAL RESPONSES AND TOLERANCE THRESHOLD TO CADMIUM CONTAMINATION IN *EREMOCHLOA OPHIUROIDES*

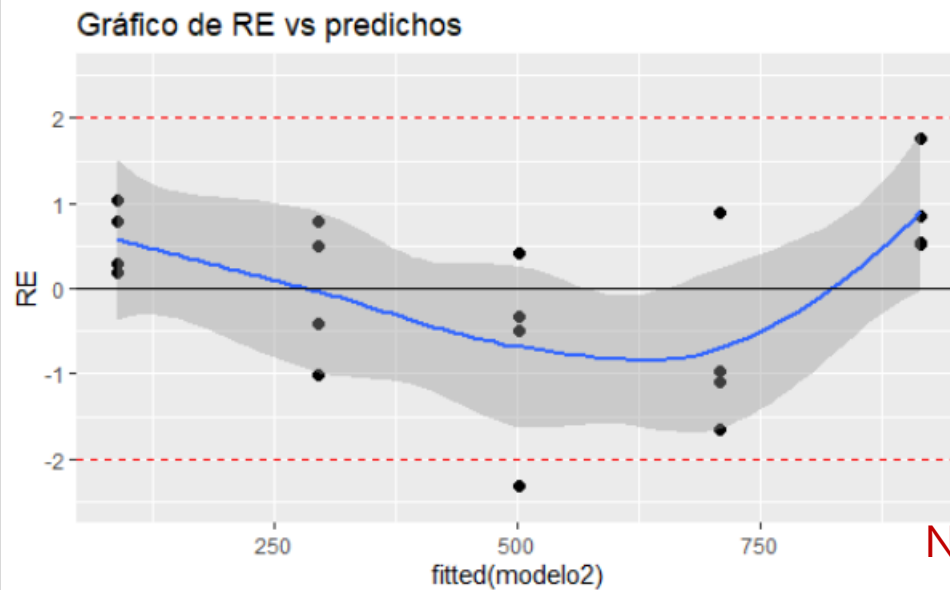
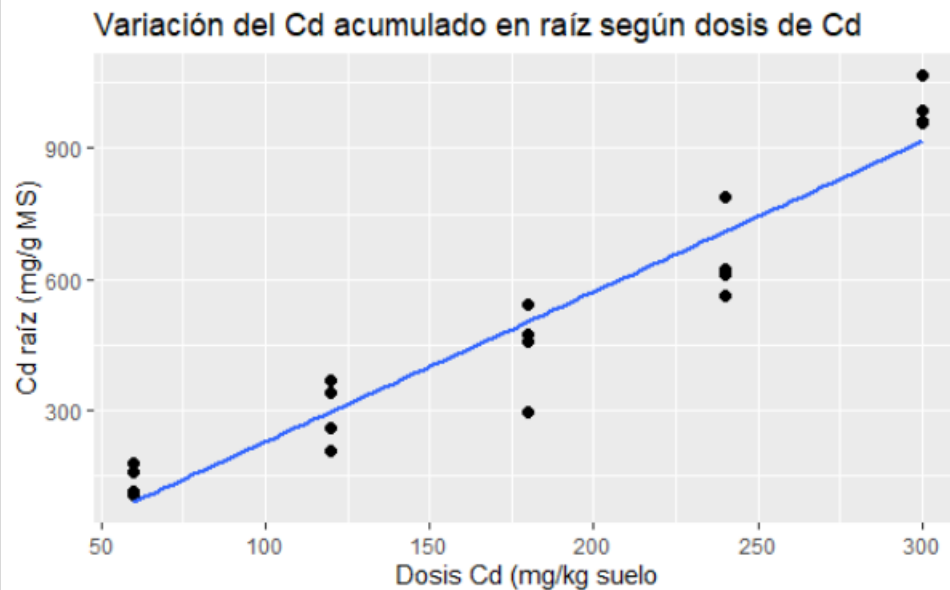
Yiming Liu, Kai Wang, Peixian Xu, and Zhaolong Wang

School of Agriculture and Biology, Shanghai Jiaotong University, Shanghai, China

Table 2 The capacity of Cd phytoextraction of centipedegrass under various Cd concentration treatments

Treatments (mg Cd/kg)	Biomass (g/pot)		Cd concentration (mg/kg DW)		Cd accumulation (mg/pot)			Phytoextraction (%)
	Shoot	Root	Shoot	Root	Shoot	Root	Total	
0	58.9 a	5.1 a	—	—	—	—	—	—
60	58.5a	5.2 a	30.3 e	104.6 e	1.8 bc	0.5 d	2.3	0.60
120	57.6 ab	4.7 a	59.0 d	258.4 d	3.4 b	1.2 cd	4.6	0.57
180	57.9 a	4.4 a	135.1 c	457.5 c	7.8 a	2.0 bc	9.8	0.87
240	53.9 ab	4.3 a	168.5 b	612.7 b	9.1 a	2.6 ab	11.7	0.76
300	46.4 b	3.4 b	202.3 a	988.9 a	9.4 a	3.4 a	12.8	0.63

Dosis Cd (mg Cd/kg)	Concentración Cd (mg Cd/kg MS)	
	Tallo y hojas	Raíz
60	23,2	104,6
	16,2	156,0
	52,7	114,9
	29,1	176,9
120	52,5	258,4
	45,7	340,9
	52,9	205,3
	84,9	366,8
180	123,5	457,5
	106,9	540,8
	123,9	472,5
	145,7	294,3
240	166,8	612,7
	165,9	789,6
	184,3	622,9
	157,0	562,6
300	208,4	988,9
	189,9	1067,1
	217,7	959,6
	193,2	962,9



Residual standard error: 92.65 on 18 degrees of freedom
Multiple R-squared: 0.9171, Adjusted R-squared: 0.9125

Regresión polinomial

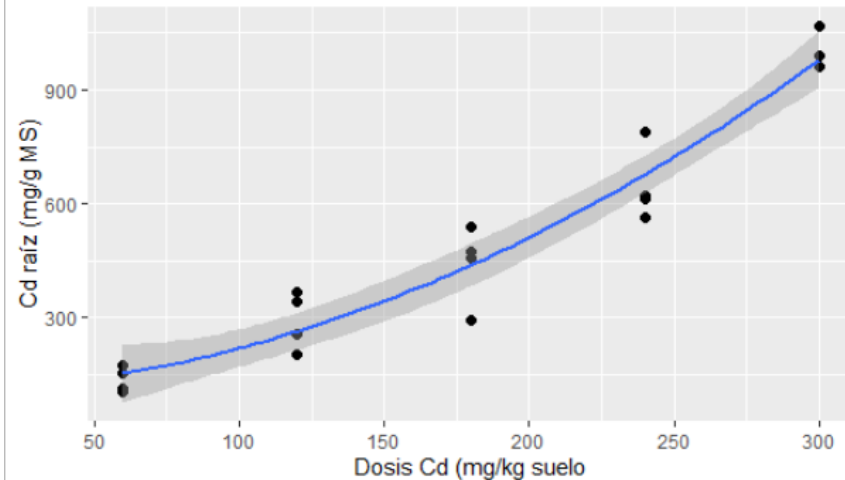
29

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \dots + \beta_p x_i^p + \varepsilon_i$$

- El modelo incluye términos de potencias sucesivas de la VE cuantitativa X
- Es un caso particular de **regresión múltiple**: las distintas potencias de X actúan como distintas v. explicatorias
- p es el **grado** del polinomio (**máxima potencia**)
- Si $p = 1$ entonces la regresión polinomial se reduce a regresión lineal simple
- El grado máximo al que se puede ajustar un conjunto de n datos es $n-1$. Si se desea hacer inferencia, $n-2$
- Y como siempre:

$$\varepsilon_i \approx NID(0, \sigma^2)$$

Variación del Cd acumulado en raíz según dosis de Cd



$$E(y_i) = \beta_0 + \beta_1 \text{dosis}_i + \beta_2 \text{dosis}_i^2$$

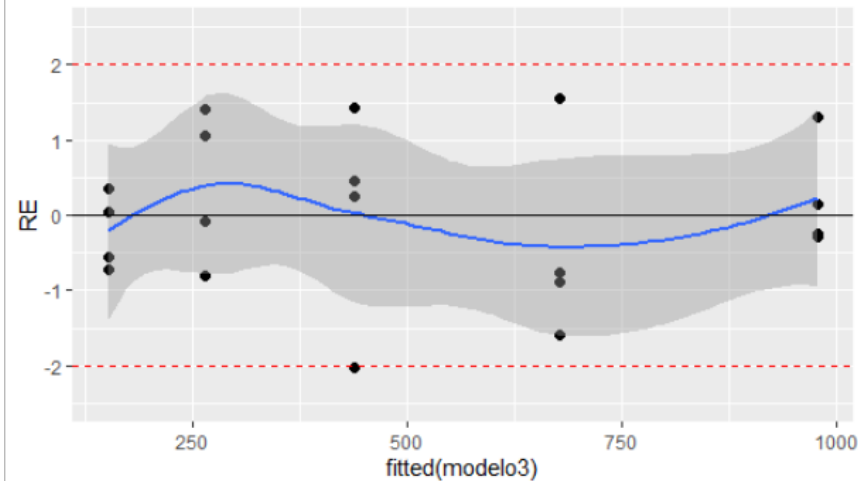
```
modelo3 <- lm(cd_raiz ~ dosis_cd
+ I(dosis_cd^2), bd)
```

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	1.042e+02	8.160e+01	1.277	0.21891
dosis_cd	2.802e-01	1.036e+00	0.270	0.79009
dosis_cd_cuad	8.792e-03	2.824e-03	3.113	0.00633 **

 Residual standard error: 76.09 on 17 degrees of freedom
 Multiple R-squared: 0.9472, Adjusted R-squared: 0.941
 F-statistic: 152.5 on 2 and 17 DF, p-value: 1.39e-11

Gráfico de RE vs predichos



Bibliografía

31

Quinn, G., & Keough, M. (2002). Cap 5: Correlation and regression. In *Experimental Design and Data Analysis for Biologists*. Cambridge: Cambridge University Press.

- 5.3.8 Assumptions of regression analysis
- 5.3.9 Regression diagnostics
- 5.3.10 Diagnostic graphics
(pp 92 -98)