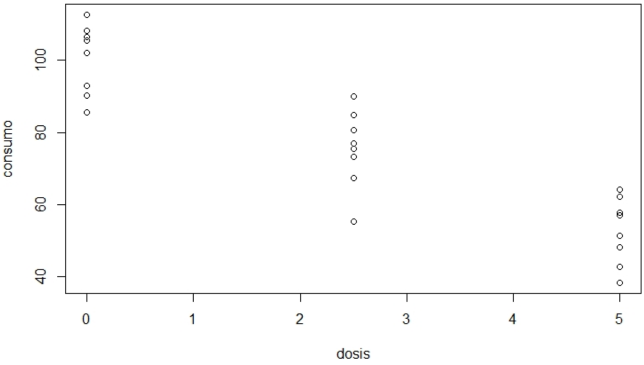
Regresión Lineal Simple

Departamento de Estadística e I.O.
Universitat de València

Regresión Lineal Simple

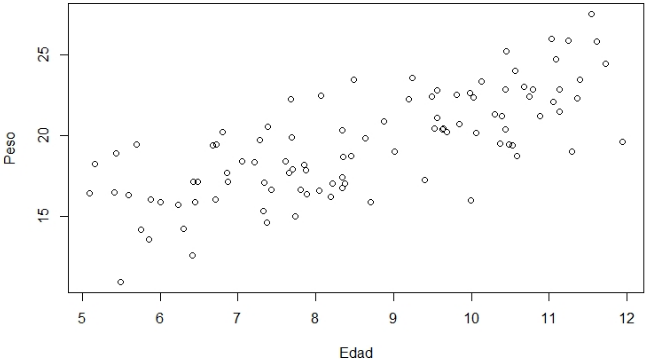
Un ejemplo: Efecto de las anfetaminas

- Efecto de las anfetaminas para inhibir el apetito. Se observa el consumo de comida en 24 ratones (gr) expuestos a dosis (mg/Kg) de anfetamina.



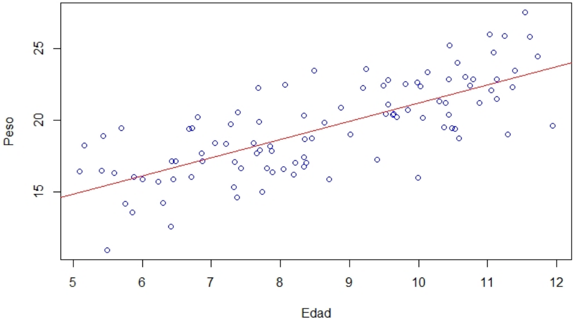
Otro ejemplo: Relación edad-peso

- En un estudio sobre el consumo de verduras en niños, se estudiaron a 100 niños observando su edad, talla, peso, sexo y consumo de verduras.



Recta de mínimos cuadrados

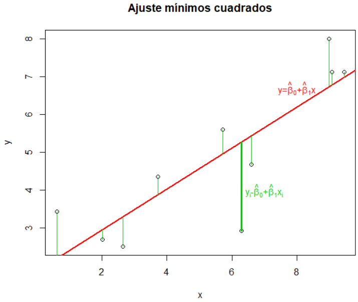
- Buscamos explicar Y, variable respuesta, a partir de X
- Relación más sencilla: $Y = f(X) + \varepsilon$ con $f(X) = \beta_0 + \beta_1\ X$
- ¿Qué recta es la más cercana a la nube de puntos en el diagrama de dispersión?



Distancia de una recta a una nube de puntos

- Distancia de un punto (x_i, y_i) a una recta $y = \hat{\beta}_0 + \hat{\beta}_1 \, x$
  - Debe ser distancia vertical: $|y_i - \hat{y}_i|$ con $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 \, x_i$
  - Definimos el residuo del par (x_i, y_i): $e_i = y_i - \hat{y}_i$
- Distancia mínimos cuadrados:

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



Estimadores mínimos cuadrados

- La recta de mínimos cuadrados es la que minimiza la distancia de mínimos cuadrados.
- Derivando e igualando a cero:

$$\frac{\partial}{\partial \beta_0}\left(\sum_{i=1}^{n}(y_i-(\beta_0+\beta_1 x_i))^2\right) = -2\sum_{i=1}^{n}(y_i-(\beta_0+\beta_1 x_i)) = 0$$

$$\frac{\partial}{\partial \beta_1}\left(\sum_{i=1}^{n}(y_i-(\beta_0+\beta_1 x_i))^2\right) = -2\sum_{i=1}^{n}(y_i-(\beta_0+\beta_1 x_i))x_i = 0$$

- El mínimo se obtiene para:

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x} \text{ y } \hat{\beta}_1 = \frac{SS_{xy}}{SS_x}$$

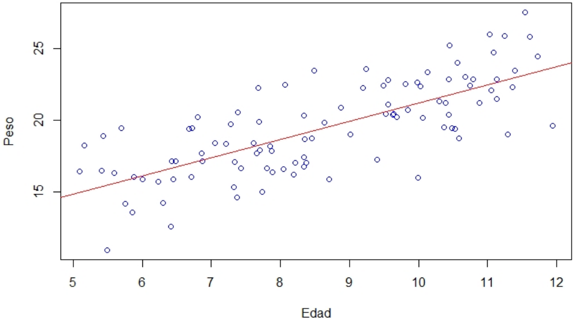
Con $SS_x = \sum_i (x_i-\bar{x})^2$ y $SS_{xy} = \sum_i (x_i-\bar{x})(y_i-\bar{y})$

Ejemplo relación edad-peso (continuación)

- En R, la recta de mínimos cuadrados se obtiene:

```
regre <- lm(Peso ~ Edad, data=Pesos)
coefficients(regre)
```

- Obtenemos $\hat{\beta}_0 = 8{,}48$ y $\hat{\beta}_1 = 1{,}27$: $\widehat{\text{Peso}} = 8{,}48 + 1{,}27\ \text{Edad}$

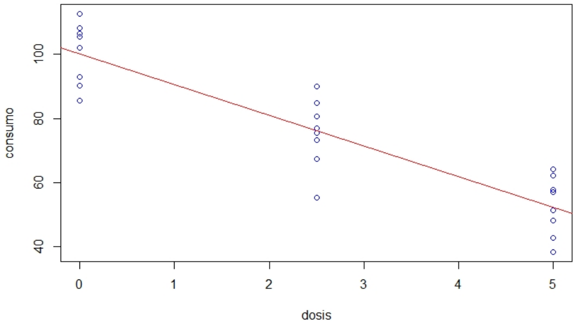


Ejemplo efecto anfetaminas (continuación)

- Obtenemos y dibujamos la recta de mínimos cuadrados:

```
reg <- lm(consumo  dosis, data=anfetas)
with(anfetas,plot(dosis,consumo,col='BLUE'))
abline(coef=coefficients(reg),col='RED')
```

- La recta es: $\text{consumo} = 100{,}10 - 9{,}53 \text{ dosis}$



Algunas propiedades

- Interpretación de los coeficientes de la recta
  - El intercepto es el valor predicho, $\hat{y}$, para $x = 0$
  - La pendiente es el incremento en Y por incremento unitario en X
  - El signo de la pendiente indica el sentido (directo o inverso) de la asociación. Si no hay relación, la pendiente vale cero.
  - La recta pasa por $(\bar{x}, \bar{y})$
- Propiedades de los residuos
  - Los residuos tienen media cero

$$e_i = y_i - \hat{y}_i = y_i - \hat{\beta}_0 - \hat{\beta}_1\, x_i = (y_i - \bar{y}) - \hat{\beta}_1\, (x_i - \bar{x})$$

$$\sum_i e_i = \sum_i (y_i - \bar{y}) - \hat{\beta}_1\, \sum_i (x_i - \bar{x}) = 0$$

- Definimos la varianza residual como:

$$s^2_{y|x} = \frac{1}{n-2} \sum_{i=1}^{n}\, e_i^2 = \frac{1}{n-2} SS_{y|x}$$

Coeficiente de Determinación

- La recta de regresión será útil para predecir o estimar si conocer X disminuye nuestra incertidumbre sobre Y
- El coeficiente de determinación es una medida de la disminución de la incertidumbre
  - Es la proporción de variabilidad explicada por la regresión
    - Medimos la variabilidad de Y mediante su varianza s_y^2
    - Conociendo X, la variabilidad de Y es la de los residuos $s_{y|x}^2$.
    - La disminución de variabilidad es $s_y^2 - s_{y|x}^2$, y en proporción

$$R^2 = 1 - \frac{SS_{y|x}}{SS_y}$$

- Ejercicio: Calcular R^2 para los datos de anfetaminas y de la relación edad/peso

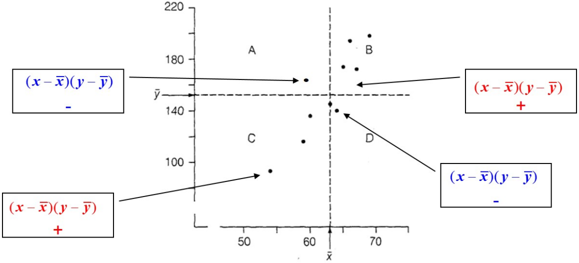
Regresión Lineal Simple

Covarianza

- La covarianza es una medida de la relación lineal entre dos variables numéricas

$$s_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \frac{1}{n-1}SS_{xy}$$

(Notad que ya apareció al obtener $\hat{\beta}_1$)



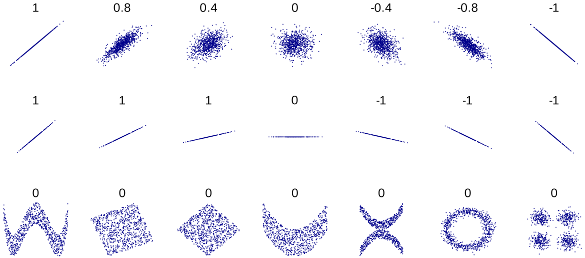
- Tiene unidades, por lo que es difícil de interpretar

Correlación lineal de Pearson

- El coeficiente de correlación lineal de Pearson es

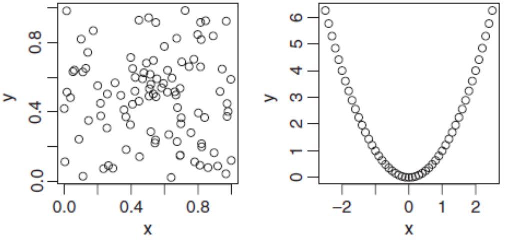
$$r = \frac{s_{xy}}{s_x s_y} = \frac{SS_{xy}}{\sqrt{SS_x SS_y}}$$

- Es una medida relativa. Toma valores entre -1 y 1. Cuanto más cerca de 1 esté $|r|$, mayor relación lineal.



Solamente relaciones lineales

- El coeficiente de correlación solo sirve para medir las relaciones lineales entre variables numéricas
- Las siguientes gráficas tienen un coeficiente de correlación 0



(a) (b)

- En la gráfica de la izquierda, X e Y son independientes. A la derecha, $Y = X^2$, una relación cuadrática perfecta.

Cálculo de la correlación

- En R, el coeficiente de correlación se obtiene:

```
with(Pesos,cor(Edad,Peso))
```

- Podemos contrastar si en la población puede ser 0.

```
with(Pesos,cor.test(Edad,Peso))
```

- Y entre todos los pares de variables numéricas

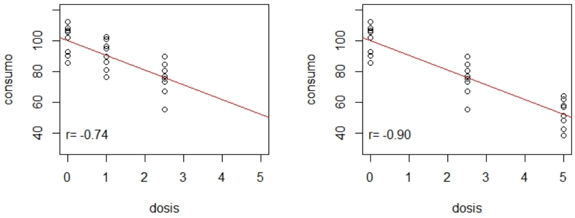
```
cor(subset(Pesos, select = -Sexo))
```

- La matriz de correlación es simétrica con 1's en la diagonal

	Edad	Verduras	Peso	Altura
Edad	1.0000000	0.5401488	0.7382047	0.9670731
Verduras	0.5401488	1.0000000	0.2397046	0.5069650
Peso	0.7382047	0.2397046	1.0000000	0.7023951
Altura	0.9670731	0.5069650	0.7023951	1.0000000

Cálculo de la correlación (continuación)

- Solo se puede estimar la correlación si X e Y son variables. No todas las columnas de un banco de datos son variables.
  - En el ejemplo de las anfetaminas, el consumo es variable pero la dosis no lo es, pues la han fijado los investigadores.
  - Si hubieran decidido observar dosis 0, 1 y 2.5 (en lugar de 0, 2.5 y 5) la varianza de X cambiaría artificialmente y la correlación también



- Además está pensado para variables continuas.

Regresión y correlación

- La correlación es simétrica, la regresión no
  - En regresión es muy importante especificar la variable respuesta
    - Ejercicio: Con los datos de los pesos de niños, calculad la recta de regresión considerando la edad como variable respuesta (dependiente) y comparadla con la obtenida anteriormente
- La pendiente de la recta es proporcional a la correlación

$$\hat{\beta}_1 = \frac{SS_{xy}}{SS_x} = \frac{SS_{xy}}{\sqrt{SS_x \times SS_y}} \frac{\sqrt{SS_y}}{\sqrt{SS_x}} = r \times \frac{\sqrt{SS_y}}{\sqrt{SS_x}}$$

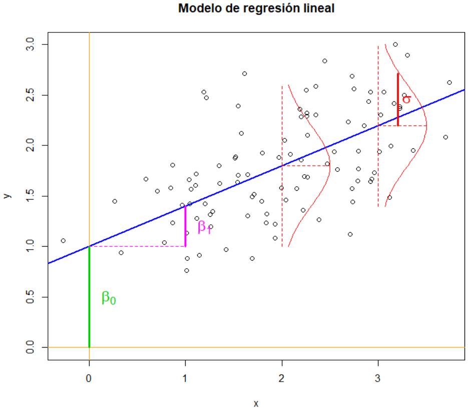
  - La pendiente de la recta es 0 si y solo si la correlación es 0
  - Los signos de la pendiente, la correlación y la covarianza coinciden
- El coeficiente de determinación es el cuadrado de él de correlación: $R^2 = r^2$

Regresión Lineal Simple

Modelo de regresión lineal simple

- Modelo de regresión lineal simple normal homocedástico

$$Y = \beta_0 + \beta_1 X + \epsilon \quad \text{siendo} \quad \epsilon \sim N(0, \sigma^2)$$



Estimación de los parámetros

- Es un modelo estocástico, permite utilizar herramientas estadísticas para medir la incertidumbre
- Si los datos son una muestra aleatoria de la población:
  - Los estimadores máximo verosímiles de los coeficientes coinciden con los estimadores mínimos cuadrados

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \text{ y } \hat{\beta}_1 = \frac{SS_{xy}}{SS_x}$$

  - El estimador máximo verosímil de σ^2 es la varianza de los residuos, $\sum_{i=1}^{n} e_i^2/n$, pero se utiliza más el estimador insesgado:

$$\hat{\sigma}^2 = \frac{1}{n-2}\sum_{i=1}^{n} e_i^2 = \frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \frac{1}{n-2} SS_{y|x}$$

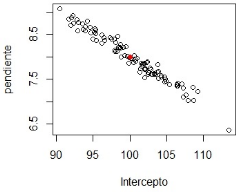
- Estimador insesgado: su esperanza coincide con el valor del parámetro que se desea estimar
- Estimador máxima verosimilitud: valor del parámetro que maximiza la función de verosimilitud

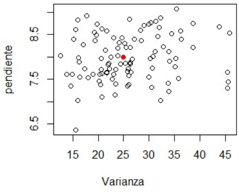
Distribuciones en el muestreo

- Bajo esas condiciones, las distribuciones en el muestreo son tales que:

$$\frac{\hat{\beta}_0 - \beta_0}{\hat{\sigma}\sqrt{(1+\bar{x}^2/s_x^2)/n}} \sim t_{n-2}, \quad \frac{\hat{\beta}_1 - \beta_1}{\hat{\sigma}/(\sqrt{n}s_x)} \sim t_{n-2} \text{ y } \frac{(n-2)\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{n-2}$$

- La siguiente gráfica muestra los estimadores de 100 muestras de tamaño $n = 20$ generadas con $\beta_0 = 100$, $\beta_1 = 8$, y $\sigma^2 = 25$





Propiedades de los estimadores

- De las distribuciones en el muestreo anteriores se deduce:
  - Los tres estimadores son insesgados:
    $E(\hat{\beta}_0) = \beta_0$; $E(\hat{\beta}_1) = \beta_1$ y $E(\hat{\sigma}) = \sigma$
  - Los errores estándar de $\hat{\beta}_0$ y $\hat{\beta}_1$ son:

$$SE(\hat{\beta}_0) = \frac{\hat{\sigma}}{\sqrt{n}}\sqrt{1 + \bar{x}^2/s_x^2}, \quad SE(\hat{\beta}_1) = \frac{\hat{\sigma}}{\sqrt{n}s_x}$$

  - $\hat{\beta}_0$ y $\hat{\beta}_1$ son estimadores consistentes (la varianza disminuye con el tamaño de la muestra). Además, al ser estimadores máximo verosímiles también son asintóticamente eficientes (varianza del estimador es mínima).
- Los intervalos de confianza al 95 %, aproximadamente,

$$IC95(\beta_0) = \hat{\beta}_0 \pm 2\, SE(\hat{\beta}_0), \quad IC95(\beta_1) = \hat{\beta}_1 \pm 2\, SE(\hat{\beta}_1)$$

(Notad que esos IC95 serían exactos si se utiliza el cuantil 0.975 de t_{n-2} en lugar de 2)

```
regre <- lm(Peso  Edad, data=Pesos)
confint(regre,level=0.95)
```

Contrastes de hipótesis sobre los parámetros de la recta

- Contraste de linealidad
  - Hipótesis nula $\beta_1 = 0$ frente a una alternativa bilateral
  - Bajo la hipótesis nula, el estadístico de contraste sigue una distribución t_{n-2}

$$t = \frac{\hat{\beta}_1}{\hat{\sigma}/(\sqrt{n}s_x)} \sim t_{n-2}$$

- Contraste sobre correlación 0
  - La hipótesis nula $\rho = 0$ es equivalente a $\beta_1 = 0$
- Contraste sobre el intercepto de la recta
  - Hipótesis nula $\beta_0 = 0$ frente a una alternativa bilateral
  - Bajo la hipótesis nula, el estadístico de contraste sigue una distribución t_{n-2}

$$t = \frac{\hat{\beta}_0}{\hat{\sigma}\sqrt{(1+\bar{x}^2/s_x^2)/n}} \sim t_{n-2}$$

Regresión lineal en R: Peso

- En el ejemplo del peso, observad las salidas de R a los siguientes comandos:

```
regreA <- lm(Peso ~ Altura, data=Pesos)
summary(regreA)
confint(regreA)
```

  ¿Cómo de bueno es el modelo? ¿Dirías que Altura explica algo el Peso?¿Sería razonable forzar $b_0 = 0$?

- Si quisiéramos hacer ésto último:

```
regreA2 <- lm(Peso ~ Altura-1, data=Pesos)
summary(regreA2)
confint(regreA2)
```

Regresión lineal en R: Anfetaminas

- En el ejemplo de las anfetaminas, observad las salidas de R a los siguientes comandos:

```
reg <- lm(consumo ~ dosis, data=anfetas)
summary(reg)
confint(reg)
```

  ¿Cómo de bueno es el modelo? ¿Dirías que Dosis explica algo del Consumo?¿Sería razonable forzar $b_0 = 0$?

- Si quisiéramos forzar que la recta pase por el origen:

```
reg.2 <- lm(consumo ~ dosis -1, data=anfetas)
summary(reg.2)
confint(reg.2)
```

Precisión de la Estimación de la recta de regresión

- El valor de la recta en un punto genérico x_0 es $E(Y|x_0) = \beta_0 + \beta_1 \ x_0$
  - Su estimador puntual es $\hat{\beta}_0 + \hat{\beta}_1 \ x_0$
  - Es insesgado y con error estándar

$$SE(E(Y|x_0)) = \frac{\hat{\sigma}}{\sqrt{n}} \sqrt{1 + \left(\frac{x_0 - \bar{x}}{s_x}\right)^2}$$

  - Su distribución en el muestreo es tal que

$$\frac{(\hat{\beta}_0 + \hat{\beta}_1 \ x_0) - (\beta_0 + \beta_1 \ x_0)}{SE(E(Y|x_0))} \sim t_{n-2}$$

- El intervalo de confianza al 95 % es:

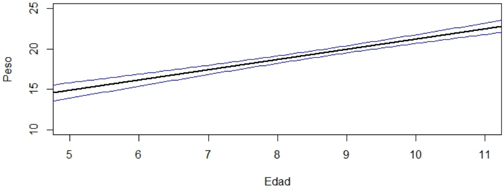
$$(\hat{\beta}_0 + \hat{\beta}_1 x_0) \pm t_{n-2,0,975} SE(E(Y|x_0))$$

Bandas de confianza

- La longitud de los intervalos de confianza sobre la recta depende del punto x_0 que se desee utilizar
  - Por ejemplo, con los datos de Pesos, los intervalos para edades 5, 8 y 10 años se obtienen en R

```
regre <- lm(Peso ~ Edad, data=Pesos)
edades <- data.frame(list(Edad = c(5,8,10)))
predict(regre, newdata = edades, interval = 'confidence')
```

- Si se dibujan esos intervalos forman unas bandas que no son paralelas a la recta de regresión

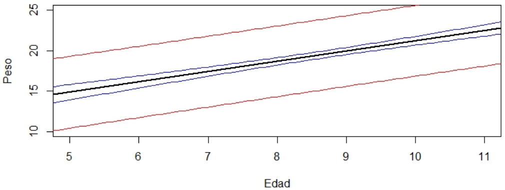


Predicción

- Uno de los principales objetivos de la regresión es la predicción
  - La predicción puntual para un nuevo individuo con $X = x_0$ es el valor de la recta en ese punto $\hat{\beta}_0 + \hat{\beta}_1 \; x_0$
  - Error estándar de predicción, mayor que él de la recta,

$$SE(\text{Pred}|x_0) = \hat{\sigma} + SE(E(Y|x_0)) = \frac{\hat{\sigma}}{\sqrt{n}} \sqrt{\mathbf{n} + 1 + \left(\frac{x_0 - \bar{x}}{s_x}\right)^2}$$

- No debe hacerse fuera del rango de x en la base de datos (extrapolación)
- Las bandas de predicción son más amplias que las bandas de confianza
  - En R: predict(regre, newdata = edades, interval = 'prediction')



Regresión Lineal Simple

Uso de predictores dicotómicos

- Cuando x es una variable de tipo éxito/fracaso, querríamos que el coeficiente b_1 estimara el cambio al pasar de una categoría (categoria de referencia) a la otra (categoría de efecto).
- Para ello debemos,
  - Tener la variable dicotómica identificada como factor en la base de datos:

    ```
    regre2 <- lm(Peso~ Sexo, data=Pesos)
    ```

  - o bien especificarlo en la sintaxis del modelo:

    ```
    regre2 <- lm(Peso~ factor(Sexo), data=Pesos)
    ```

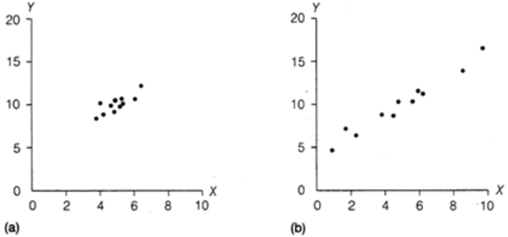
- R mostrará el coeficiente de la categoría con etiqueta mayor u orden alfabético posterior. Para facilitar la interpretación, se identifica en la salida la categoria de efecto.

Variabilidad de la variable explicativa

- Notar que la varianza de b_1 es inversamente proporcional a la varianza de x:

$$SE(\hat{\beta}_1) = \frac{\hat{\sigma}}{\sqrt{n}S_x}$$

- Cuanto más ‘dispersa’ es la x, más ‘precisa ’es la estimación de la pendiente.



(a) SS_X más pequeño (b) SS_X más grande

- En el diseño, se debe intentar que x tenga la máxima varianza.

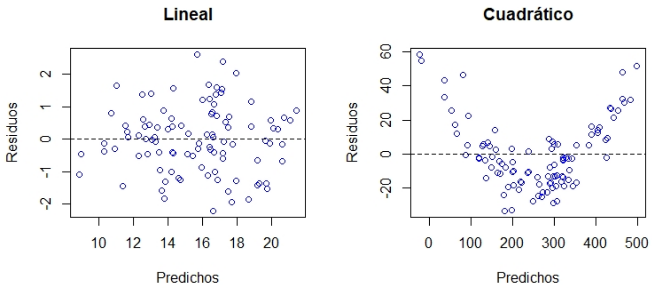
Regresión Lineal Simple

Condiciones de validez del modelo

- Obtención de los datos: Deben ser muestra aleatoria
  - La variable respuesta, dada la recta, se distribuye de manera independiente
    - Relaciones temporales y/o espaciales no son válidas
  - Los datos $\{x_i, y_i\}$ deben ser independientes y representativos de la población objetivo
- Hipótesis del modelo lineal
  - Linealidad
    - La relación $Y = f(X)$ debe ser una recta: $Y = \beta_0 + \beta_1 X$
  - Homocedasticidad
    - La variabilidad alrededor de la recta es constante, no puede depender de X
  - Normalidad
    - La distribución condicional $Y|X$ debe ser Normal

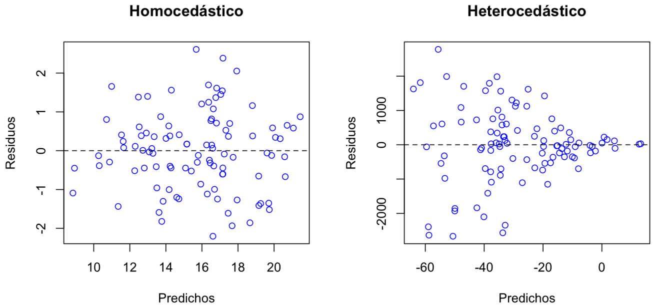
Análisis de residuos: linealidad

- Gráfica de valores predichos frente a residuos, muy útil
  - Linealidad y no linealidad, un ejemplo



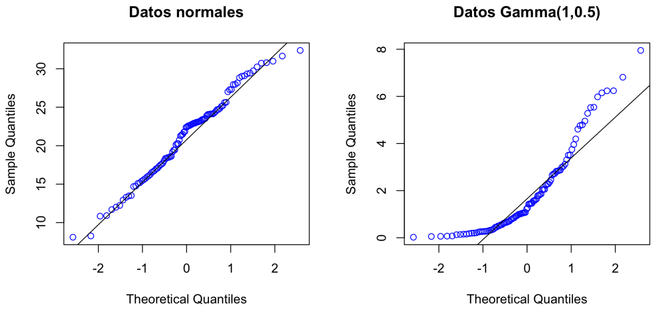
Análisis de residuos: homocedasticidad

- Gráfica de valores predichos frente a residuos, muy útil
  - Homocedasticidad y heterocedasticidad, un ejemplo



Análisis de residuos: normalidad

- Gráfica de normalidad de los residuos

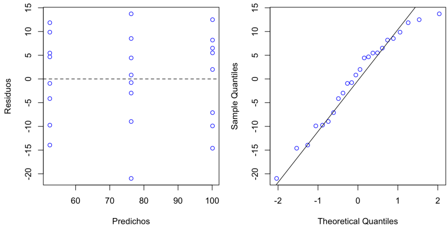


- Pueden utilizarse contrastes de normalidad, pero:
  - El modelo lineal es robusto frente a no normalidad

Análisis de residuos en R: Un ejemplo

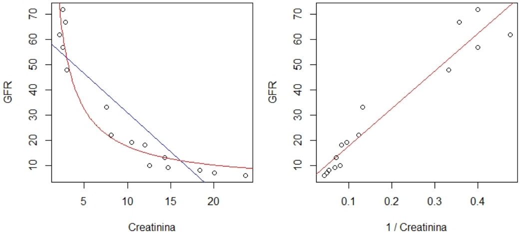
- Ambas gráficas, con datos anfetaminas:

```
reg <- lm(consumo ~ dosis, data=anfetas)
par.tipicos <- par()
par(mfcol=c(1,2),mar=c(4,4,1,1))
plot(reg$fitted.values,reg$residuals, col='BLUE',
        xlab = 'Predichos',ylab = 'Residuos')
abline(h=0,lty=2)
qqnorm(reg$residuals, col='BLUE', main='')
qqline(reg$residuals)
```



Uso de transformaciones

- La falta de linealidad y/o homocedasticidad se puede solucionar utilizando transformaciones de las variables
  - Un ejemplo: Los datos GFR.csv relacionan la creatinina con la tasa de filtración glomedular

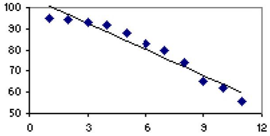


- Si heterocedasticidad hay que transformar Y. Transformar X no lo soluciona
- Cuidado con la interpretación de los efectos!
- Si no funcionan las transformaciones: modelos más flexibles

Observaciones aberrantes

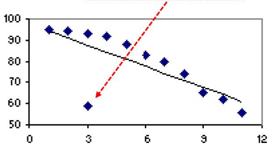
- Las observaciones aberrantes, outliers, son datos que se distancian del resto de valores observados
  - Para una sola variable, los detectamos con el diagrama de cajas
  - En regresión, una descripción de los residuos es útil

SIN OUTLIER



Ecuación de regresión: $\hat{y}_i = 104.78 - 4.10\, x_i$

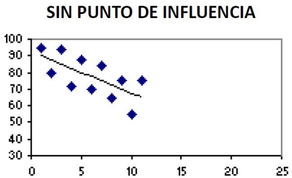
CON OUTLIER



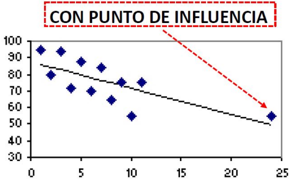
Ecuación de regresión: $\hat{y}_i = 97.51 - 3.32\, x_i$

Observaciones influyentes

- Las observaciones influyentes son datos que, si se suprimieran, cambiaría bastante el resultado
  - Es importante detectar estas observaciones, por si fueran erróneas, pues su exclusión puede cambiar los resultados del estudio
  - Estudiaremos más adelante métodos para detectarlas



Ecuación de regresión: $\hat{y}_i = 92.54 - 2.5\,x_i$



Ecuación de regresión: $\hat{y}_i = 87.59 - 1.6\,x_i$

Regresión Lineal Simple

Regresión KNN

- La regresión KNN (K vecinos más próximos) es un método no paramétrico para estimar la función $f(X)$
  - Dado un valor x_0, $N_k(x_0)$ es el conjunto de índices de los k (x_i) más cercanos a x_0. Entonces:

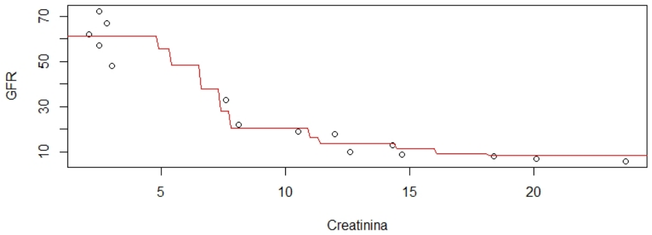
$$\hat{y}_0 = \frac{1}{k} \sum_{i \in N_k(x_0)} y_i$$

  - Es conveniente probar varios valores de k y evitar sobreestimación
  - Permite visualizar cómo debería ser la función $f(X)$
  - No modeliza el error, por lo que no permite calcular bandas de confianza

Regresión KNN en R

- Un ejemplo: utilizando los datos GFR

```
plot(GFR$Creatinina,GFR$GFR, xlab='Creatinina', ylab='GFR')
xx <- seq(1,25,0.1)
creat <- data.frame(list(Creatinina = xx))
reg <- FNN::knn.reg(GFR$Creatinina, creat, y=GFR$GFR, k=5)
lines(xx,reg$pred,col='red')
```



Regresión loess

- La regresión loess es un método de suavizado basado en regresiones locales ponderadas
  - Utiliza los k vecinos más próximos, como KNN
  - En lugar de calcular la media aritmética, ajusta una regresión ponderada a los vecinos
  - Los pesos se calculan con la función tricúbica
    - Se calculan las distancias de x_0 a los k vecinos más próximos: $D_1, \dots, D_k$
    - Estas se dividen por el máximo $d_i = D_i / \max(D_1, \dots, D_k)$
    - El vecino x_i se pondera con $w(d_i)$, siendo:

$$w(d_i) = (1 - |d_i|^3)^3 \times I_{(|d_i|<1)}$$

  - También hay que probar varios valores de k

Regresión loess en R

- Un ejemplo: utilizando los datos GFR

```
reg <- loess(GFR ~ Creatinina, data=GFR)
xx <- seq(1,25,0.1)
creat <- data.frame(list(Creatinina = xx))
plot(GFR$Creatinina,GFR$GFR, xlab='Creatinina', ylab='GFR')
prediccion <- predict(reg, newdata = creat, se = T)
lines(xx[!is.na(prediccion$fit)],
       prediccion$fit[!is.na(prediccion$fit)],col='RED')
```