# Modelos estadísticos

Técnicas estadísticas avanzadas para la conservación de la biodiversidad - Universidad de Huelva

David García Callejas

01/2021

# Regresión estadística

Hasta ahora:

- Sabemos cómo cuantificar una muestra o una población

# Regresión estadística

Hasta ahora:

- Sabemos cómo cuantificar una muestra o una población
- Sabemos los fundamentos del diseño experimental

# Regresión estadística

Hasta ahora:

- Sabemos cómo cuantificar una muestra o una población
- Sabemos los fundamentos del diseño experimental
- Sabemos cómo comparar dos muestras

# Regresión estadística

Hasta ahora:

- Sabemos cómo cuantificar una muestra o una población
- Sabemos los fundamentos del diseño experimental
- Sabemos cómo comparar dos muestras

- Pero aun queda todo un mundo de preguntas que podemos resolver:

# Regresión estadística

Hasta ahora:

- Sabemos cómo cuantificar una muestra o una población
- Sabemos los fundamentos del diseño experimental
- Sabemos cómo comparar dos muestras

- Pero aun queda todo un mundo de preguntas que podemos resolver:
  - ¿Cómo afecta una variable independiente a una respuesta?

# Regresión estadística

Hasta ahora:

- Sabemos cómo cuantificar una muestra o una población
- Sabemos los fundamentos del diseño experimental
- Sabemos cómo comparar dos muestras

- Pero aun queda todo un mundo de preguntas que podemos resolver:
  - ¿Cómo afecta una variable independiente a una respuesta?
  - ¿Podemos predecir una variable en función de otras?

# Regresión estadística

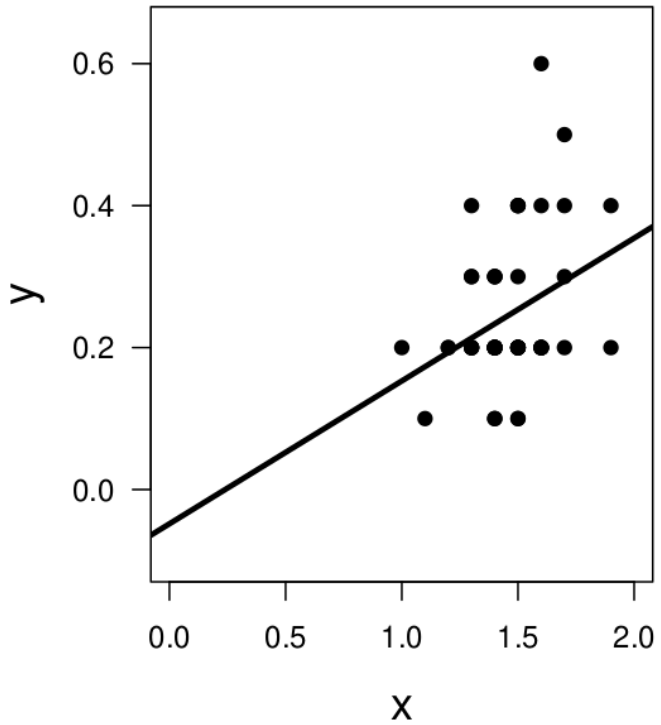
Hasta ahora:

- Sabemos cómo cuantificar una muestra o una población
- Sabemos los fundamentos del diseño experimental
- Sabemos cómo comparar dos muestras

- Pero aun queda todo un mundo de preguntas que podemos resolver:
  - ¿Cómo afecta una variable independiente a una respuesta?
  - ¿Podemos predecir una variable en función de otras?
  - ¿Qué ocurre cuando tenemos más de dos tratamientos en una población?

# Regresión estadística

$$y_i = a + bx_i + \varepsilon_i$$
$$\varepsilon_i \sim N\left(0, \sigma^2\right)$$




**Data**

$y$ = response variable

$x$ = predictor

**Parameters**

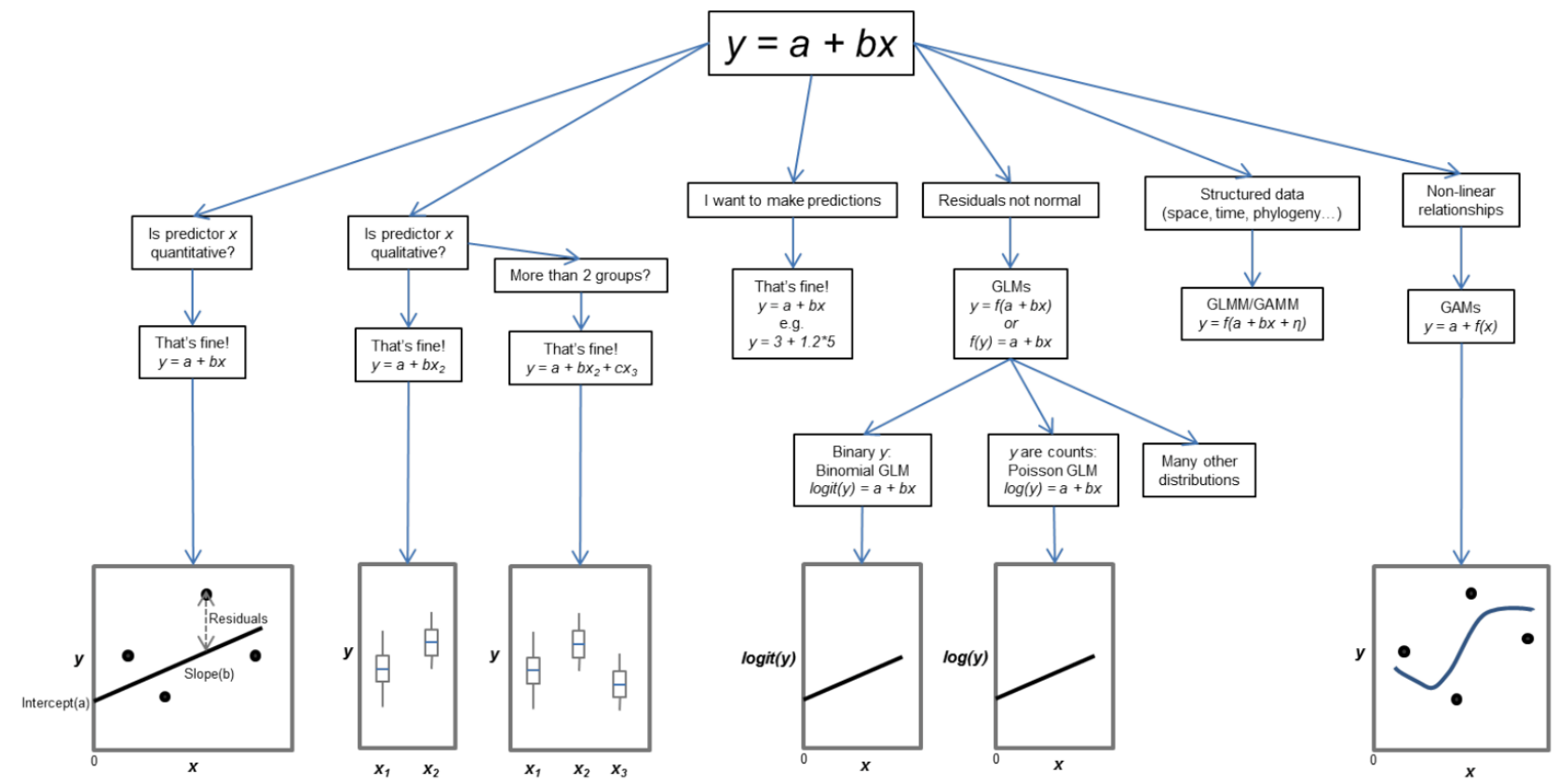
$a$ = intercept

$b$ = slope

$\sigma$ = residual variation

$\varepsilon$ = residuals

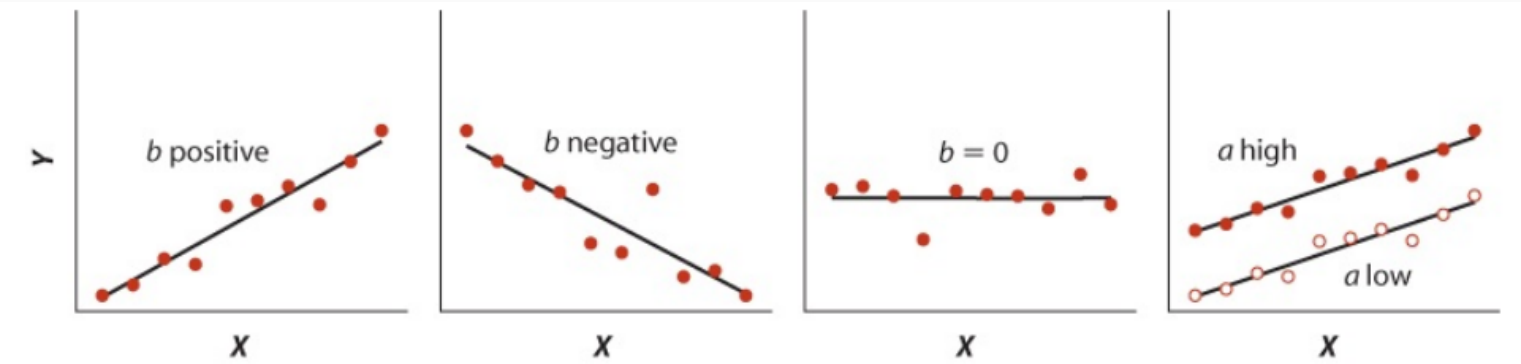
# Regresión estadística



# Regresión estadística

Regresión lineal:

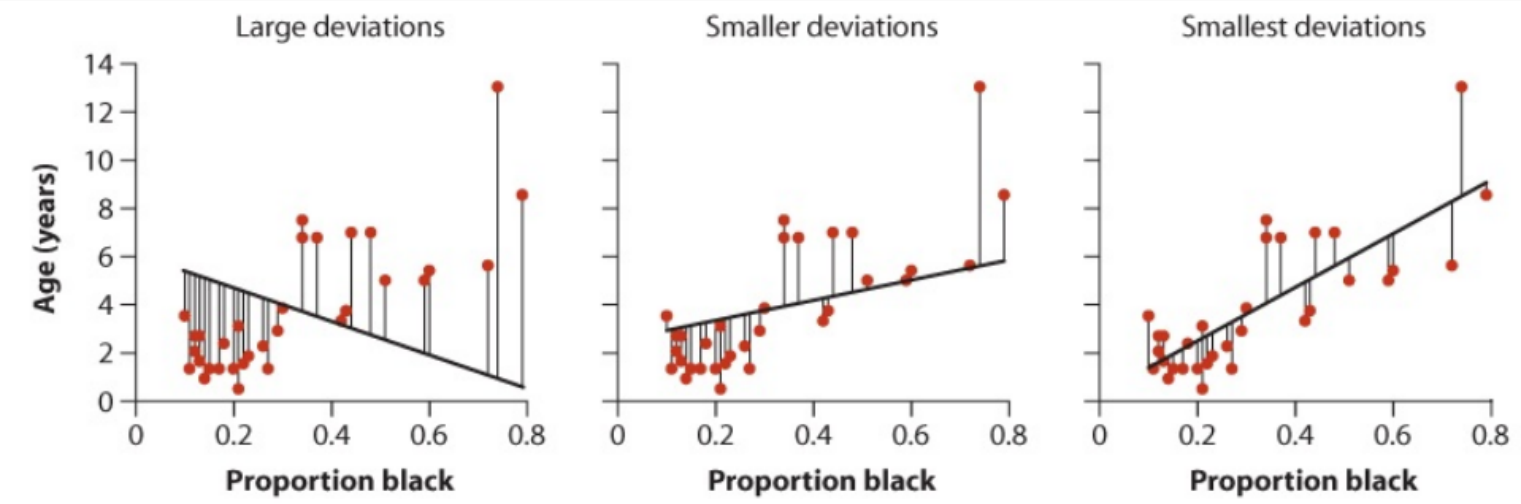
- Relación lineal entre las variables



# Regresión estadística

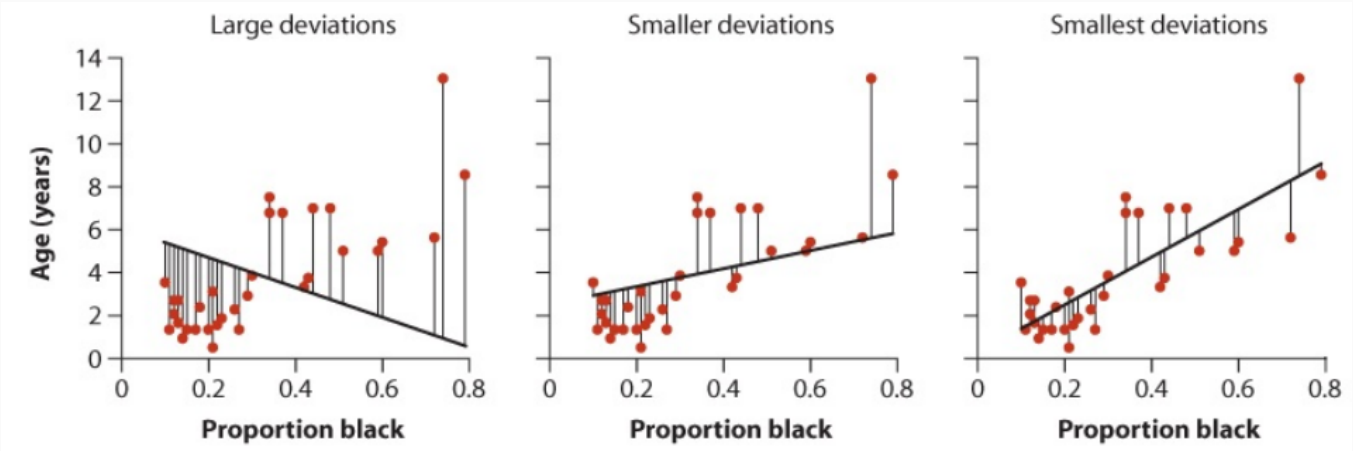
Regresión lineal:

- Minimiza el *error residual*



# Regresión estadística

- ¿Cómo calcular la recta con menor error residual? **Método de mínimos cuadrados**




$$b = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - x)^2} \tag{1}$$

$$a = \bar{y} - b\bar{x} \tag{2}$$

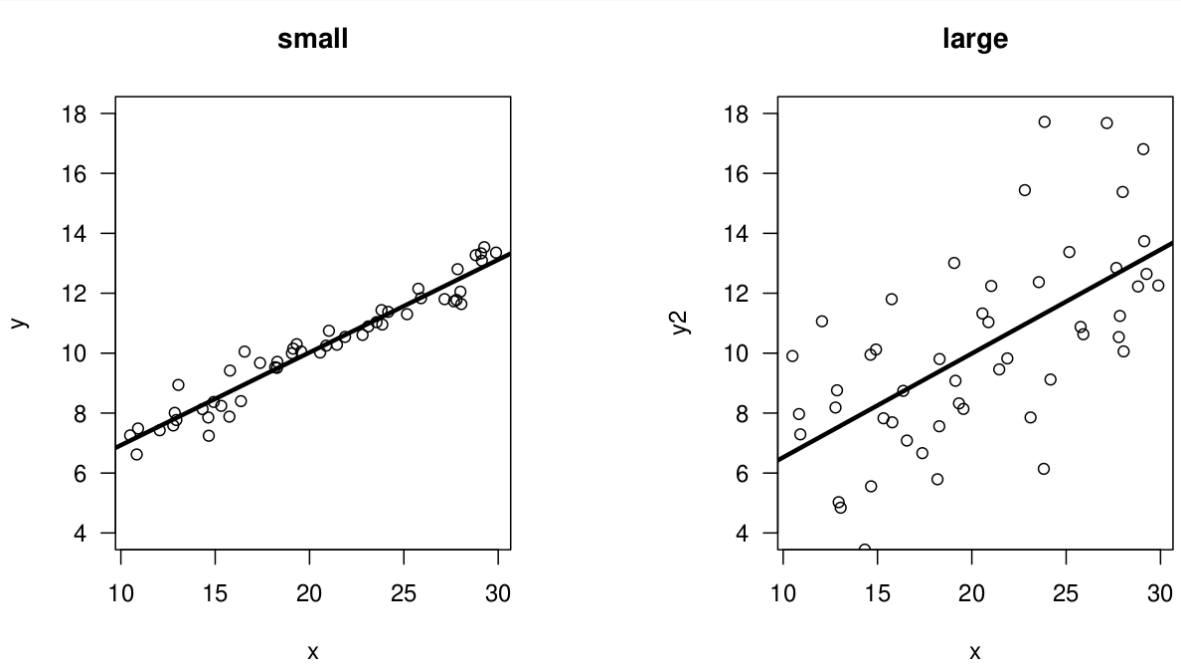
# Regresión estadística

- Residuos: diferencia entre valor observado y predicho
- Recuerda:

$$y_i = a + bx_i + \varepsilon_i$$
$$\varepsilon_i \sim N(0, \sigma^2)$$

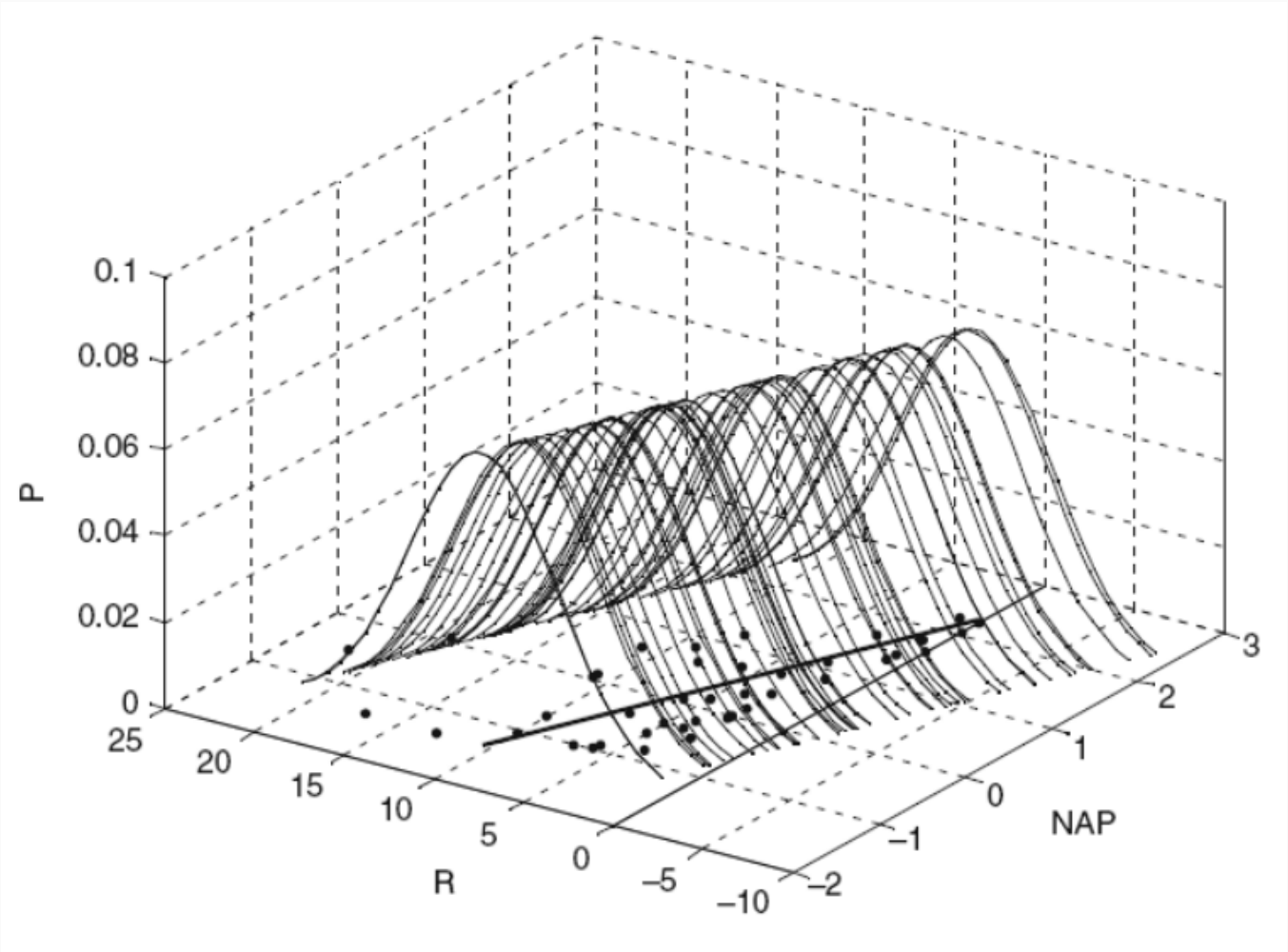
# Regresión estadística

- Residuos: diferencia entre valor observado y predicho



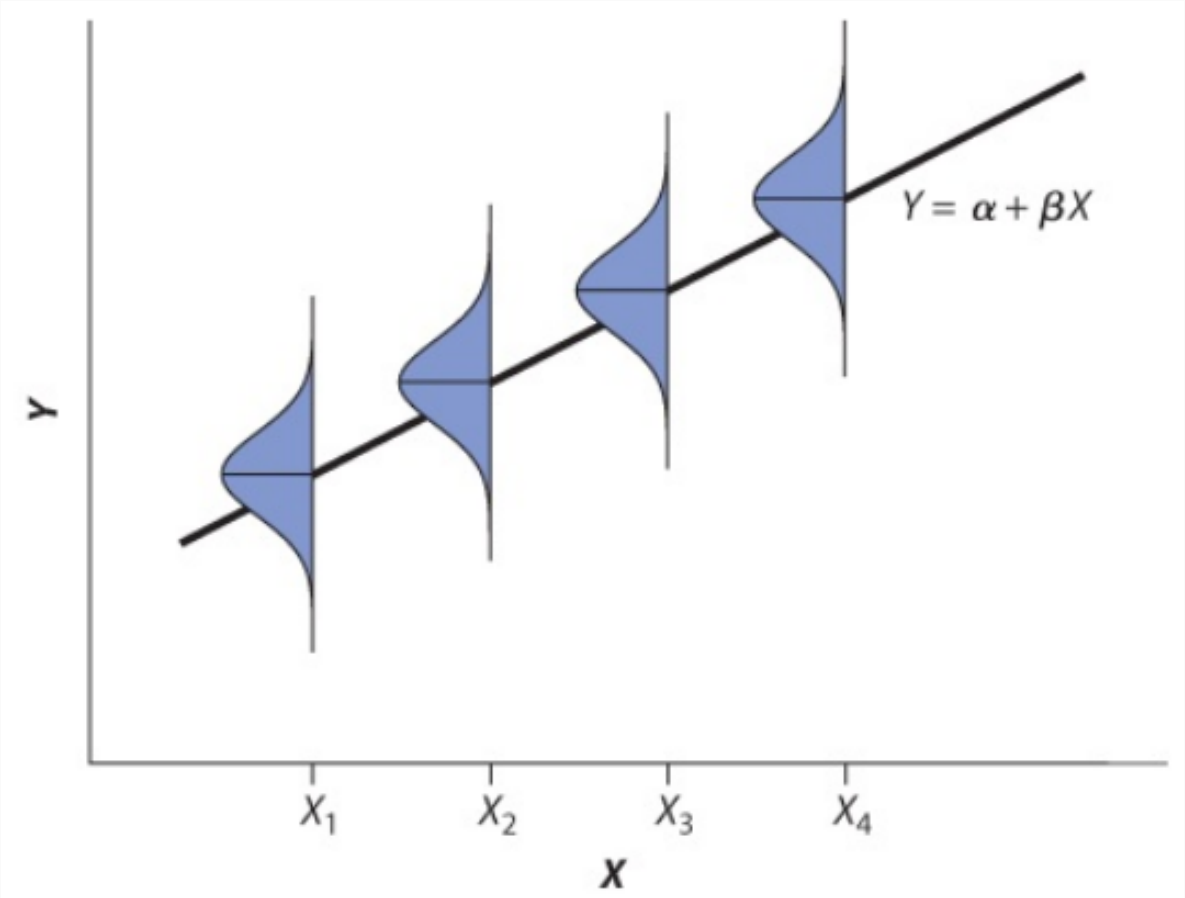
# Regresión estadística

- Para que la estimación sea correcta, la distribución de residuos debe ser normal
- y la varianza debe ser homogénea



# Regresión estadística

- **Again**: Para que la estimación sea correcta, la distribución de residuos debe ser normal y la varianza residual, homogénea



# Regresión estadística

*Importante*: Esto no implica que la variable respuesta, o la variable independiente, deban tener una distribución normal!

# Regresión estadística

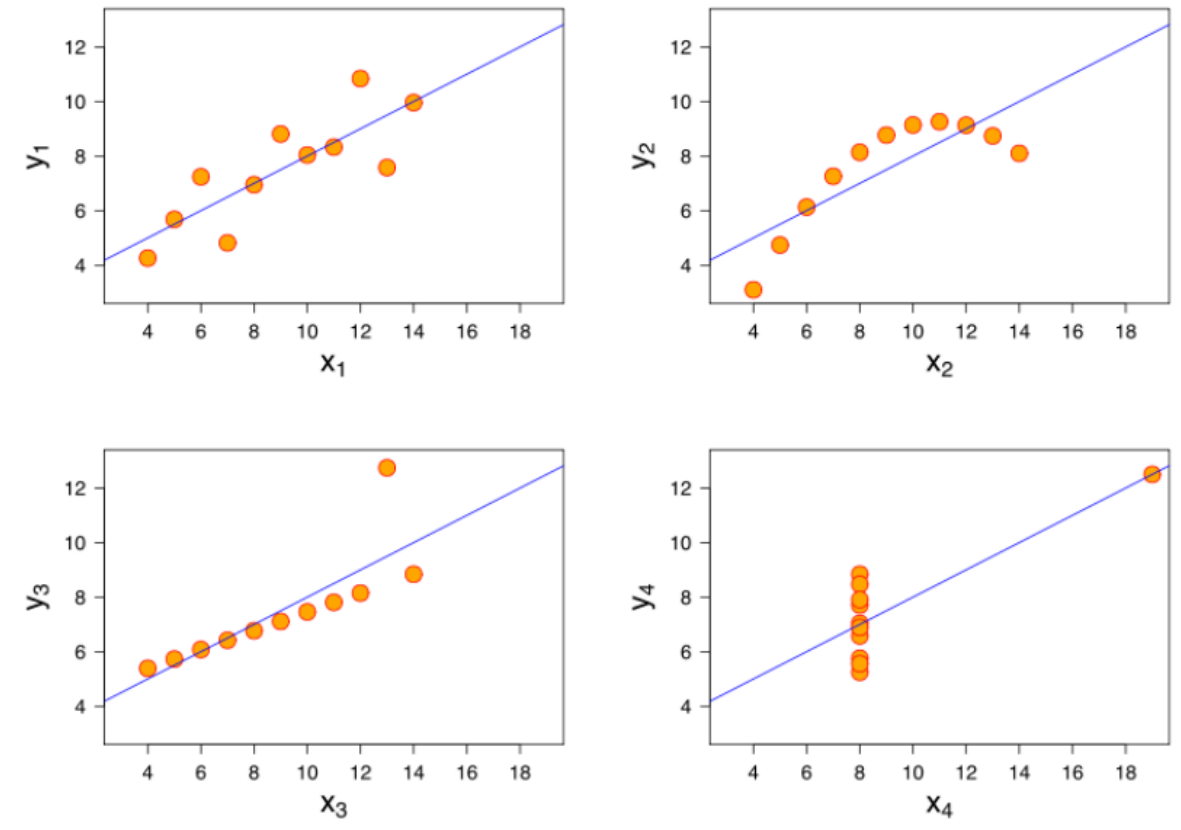
¿Podemos predecir la altura de un árbol a partir de su *dbh*?

```
trees <- read.csv(here::here("datasets","trees.csv"))
head(trees)
```

```
##   site   dbh height    sex dead
## 1    4 29.68   36.1   male    0
## 2    5 33.29   42.3   male    0
## 3    2 28.03   41.9 female    0
## 4    5 39.86   46.5 female    0
## 5    1 47.94   43.9 female    0
## 6    1 10.82   26.2   male    0
```

**Siempre**

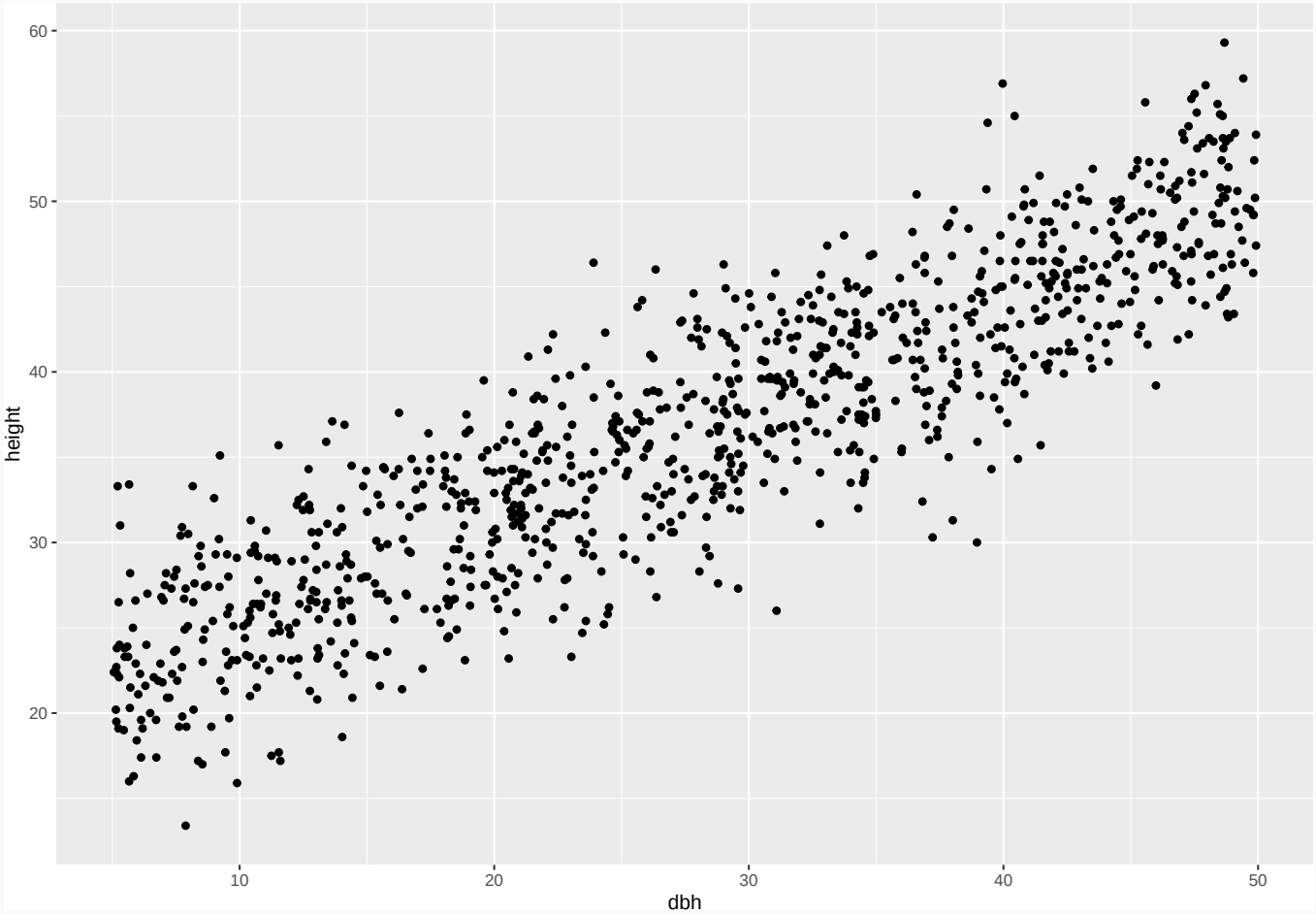
Visualiza los datos como primer paso



# Regresión estadística

¿Hay outliers en los datos?

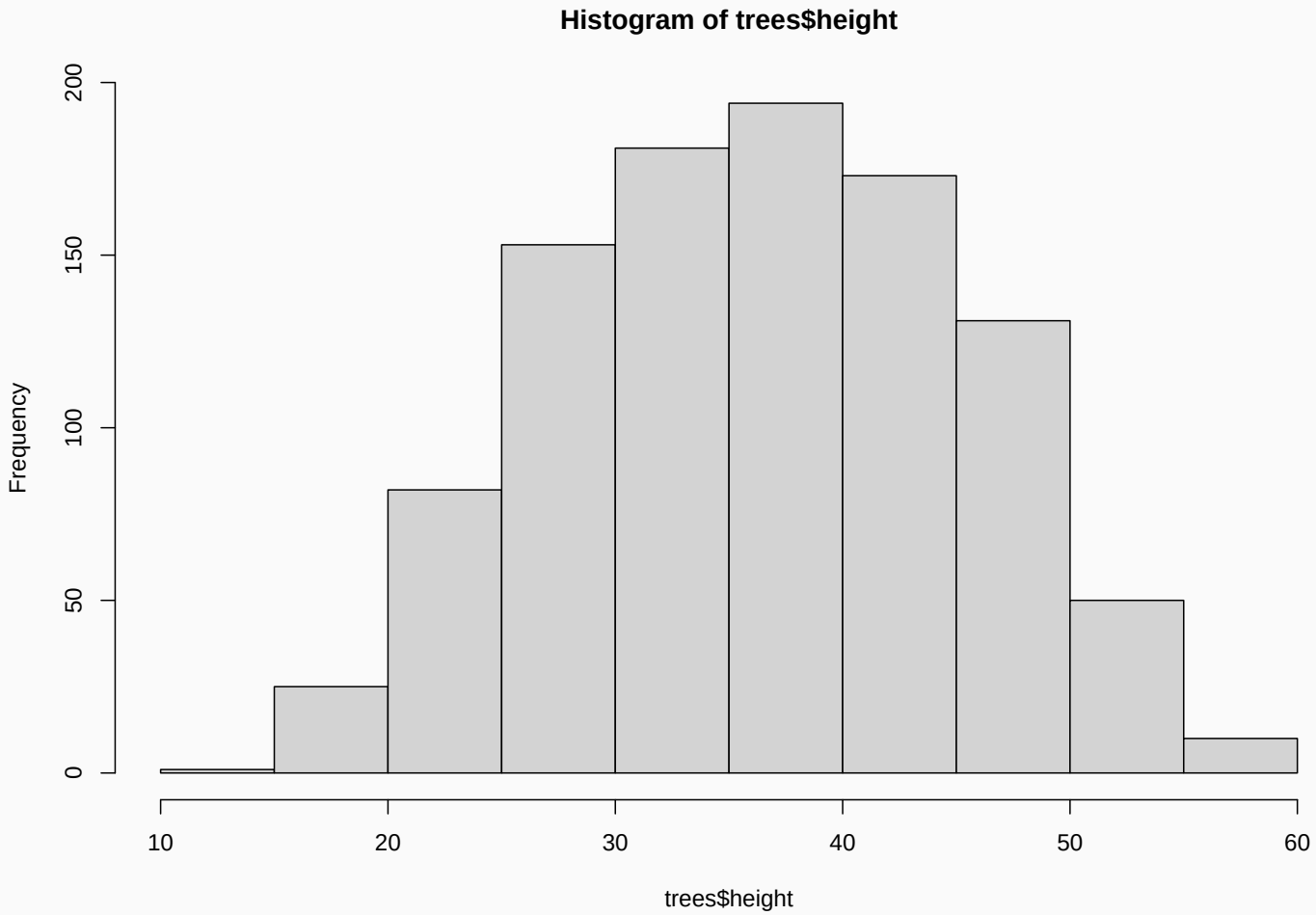
```
ggplot(trees, aes(dbh, height)) +
geom_point()
```



# Regresión estadística

¿Cómo están distribuidas las variables independientes y respuesta?

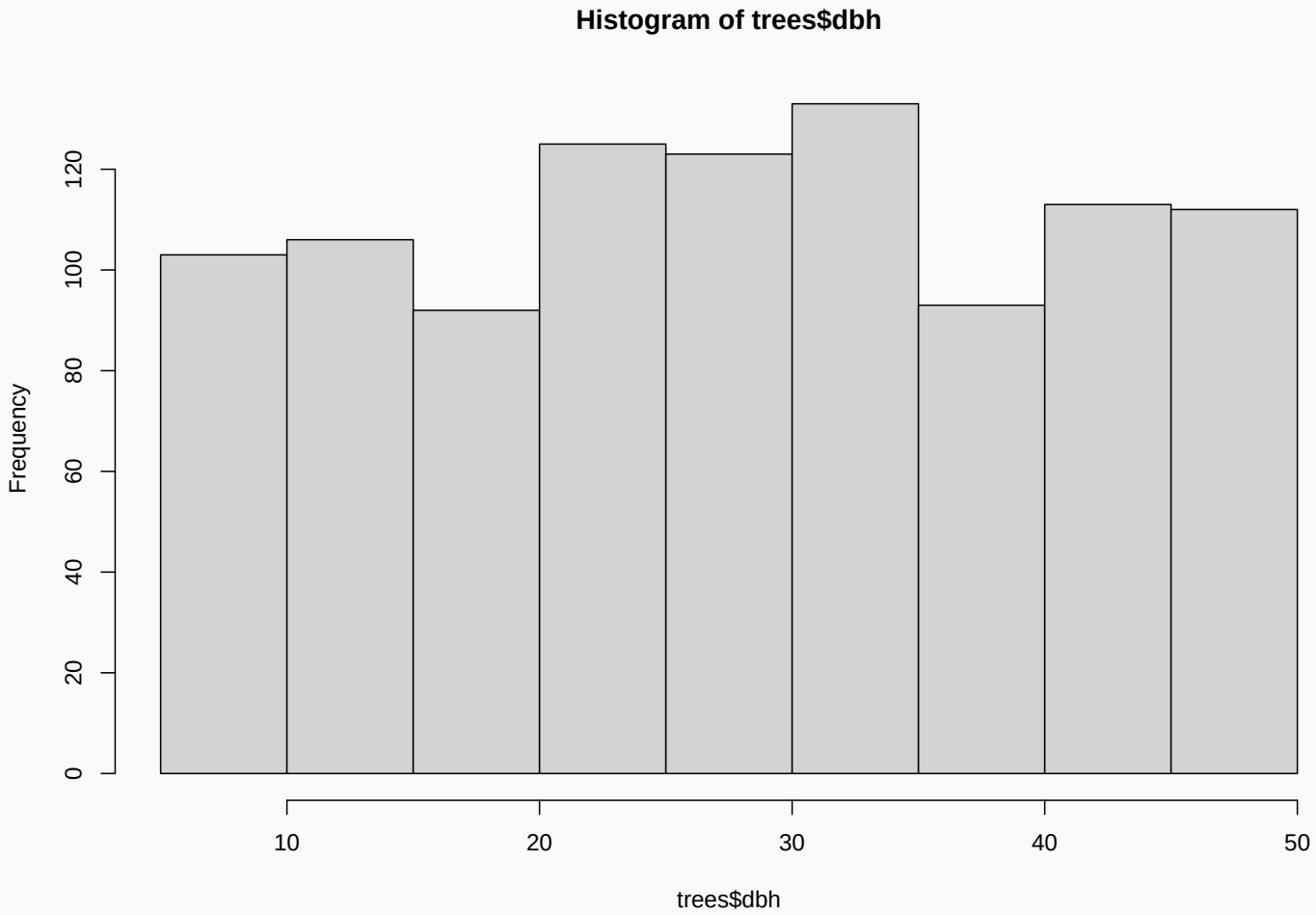
```
hist(trees$height)
```



# Regresión estadística

¿Cómo están distribuidas las variables independientes y respuesta?

```
hist(trees$dbh)
```



# Regresión estadística

Después del análisis exploratorio, si no hay nada raro, ajustamos el modelo:

```
m1 <- lm(height ~ dbh, data = trees)
```

que se corresponde con:

$$height_i = a + b \cdot DBH_i + \varepsilon_i$$
$$\varepsilon_i \sim N(0, \sigma^2)$$

# Regresión estadística

¿Y ahora?

```
m1
```

```
##
## Call:
## lm(formula = height ~ dbh, data = trees)
##
## Coefficients:
## (Intercept)          dbh
##     19.3392       0.6157
```

# Regresión estadística

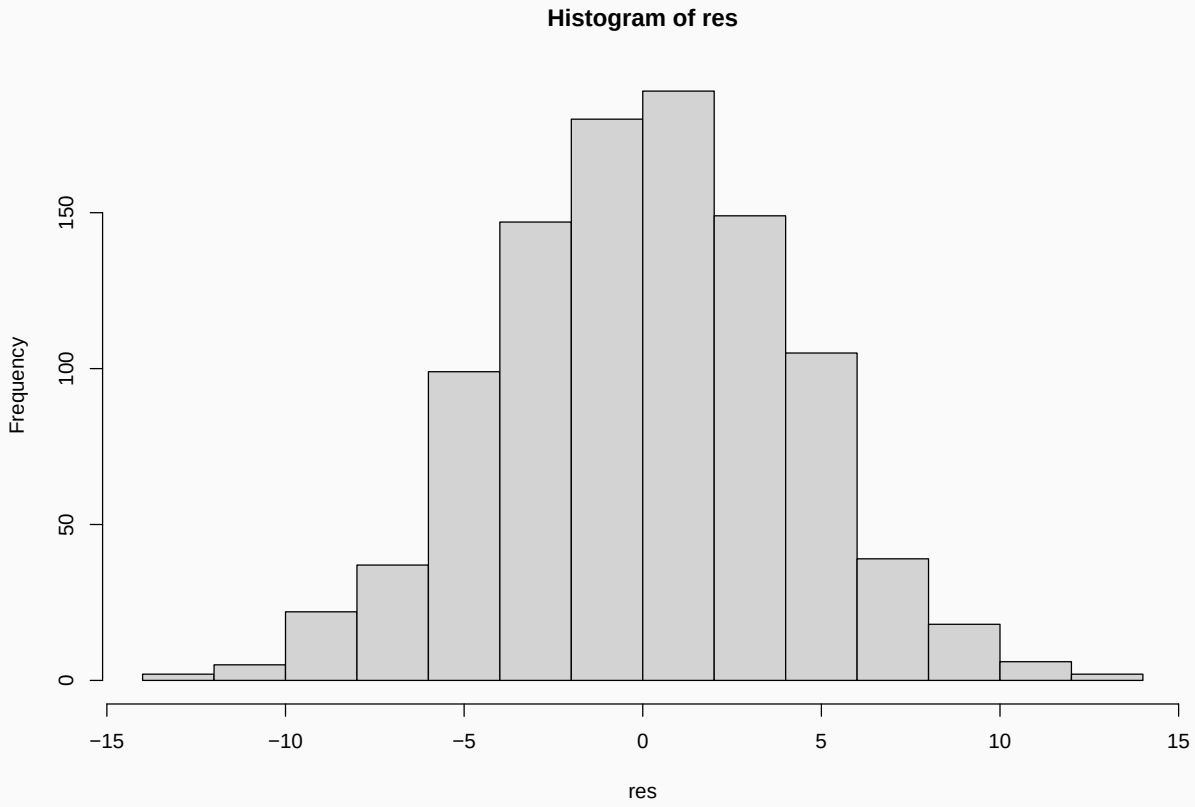
```
summary(m1)
```

```
##
## Call:
## lm(formula = height ~ dbh, data = trees)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -13.3270  -2.8978   0.1057   2.7924  12.9511
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 19.33920    0.31064   62.26   <2e-16 ***
## dbh          0.61570    0.01013   60.79   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.093 on 998 degrees of freedom
## Multiple R-squared:  0.7874, Adjusted R-squared:  0.7871
```

# Regresión estadística

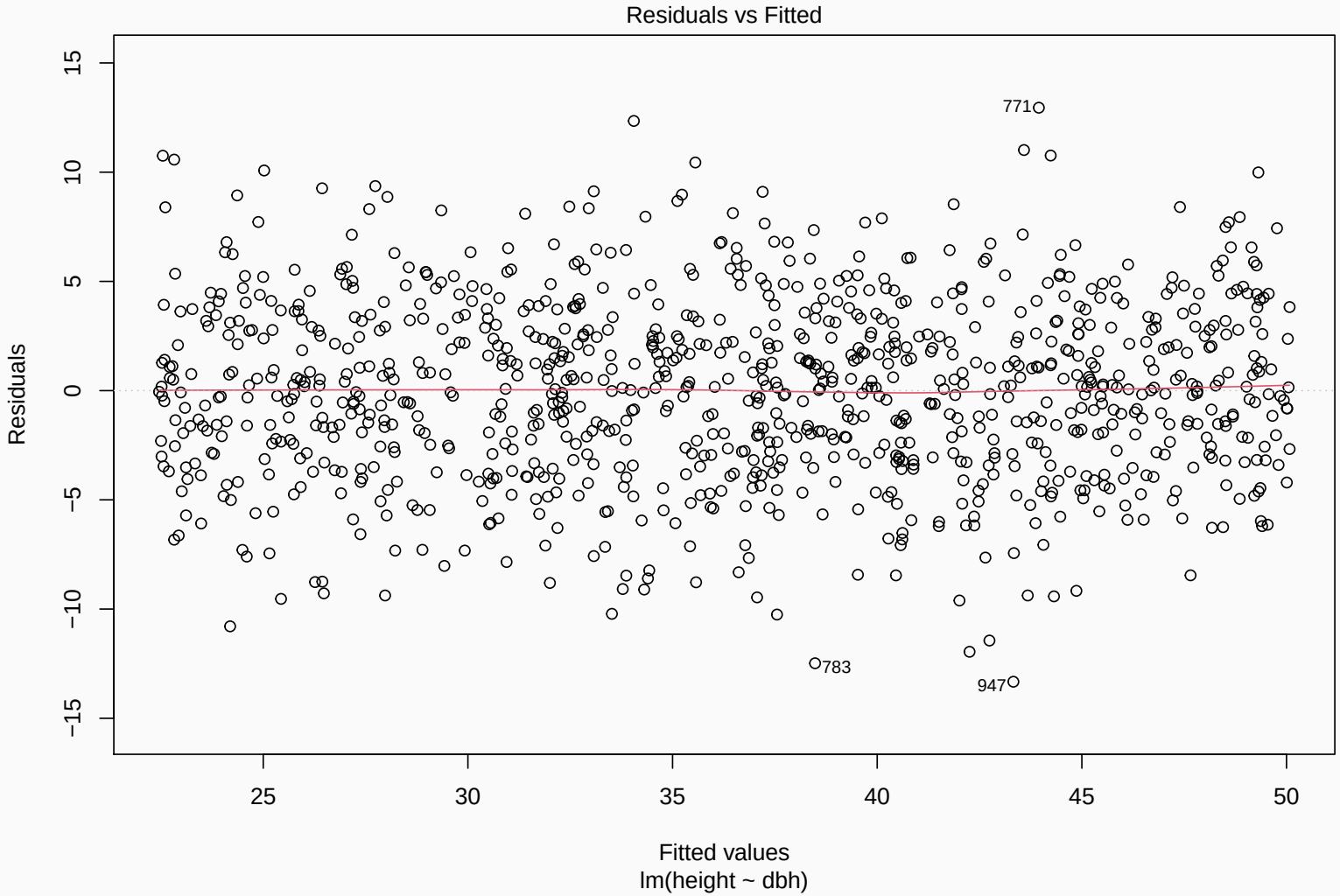
Antes de interpretar el resultado, comprobamos que los residuos se ajustan a una distribución normal

```
res <- resid(m1)
hist(res)
```



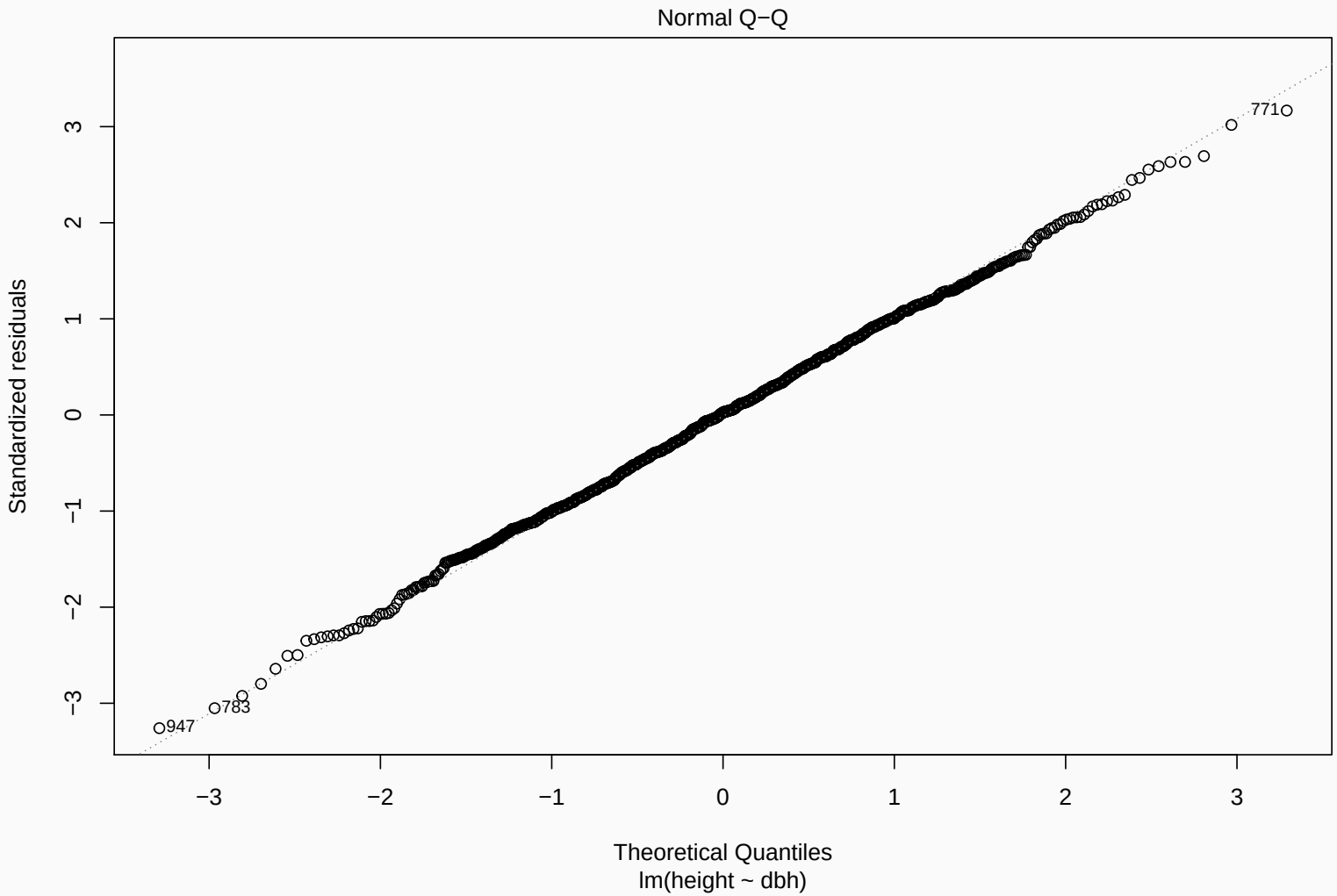
# Regresión estadística

```
plot(m1)
```



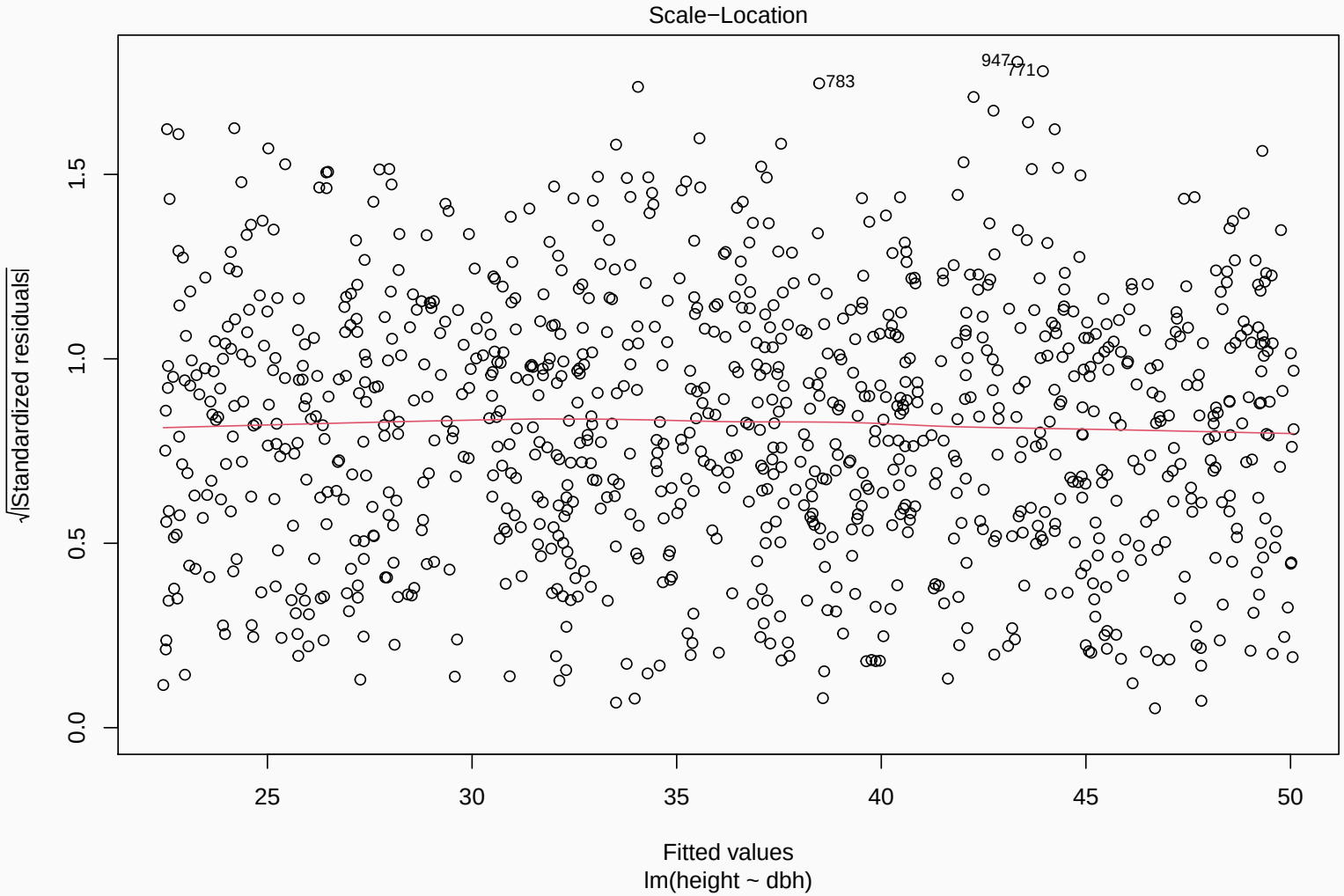
# Regresión estadística

```
plot(m1)
```



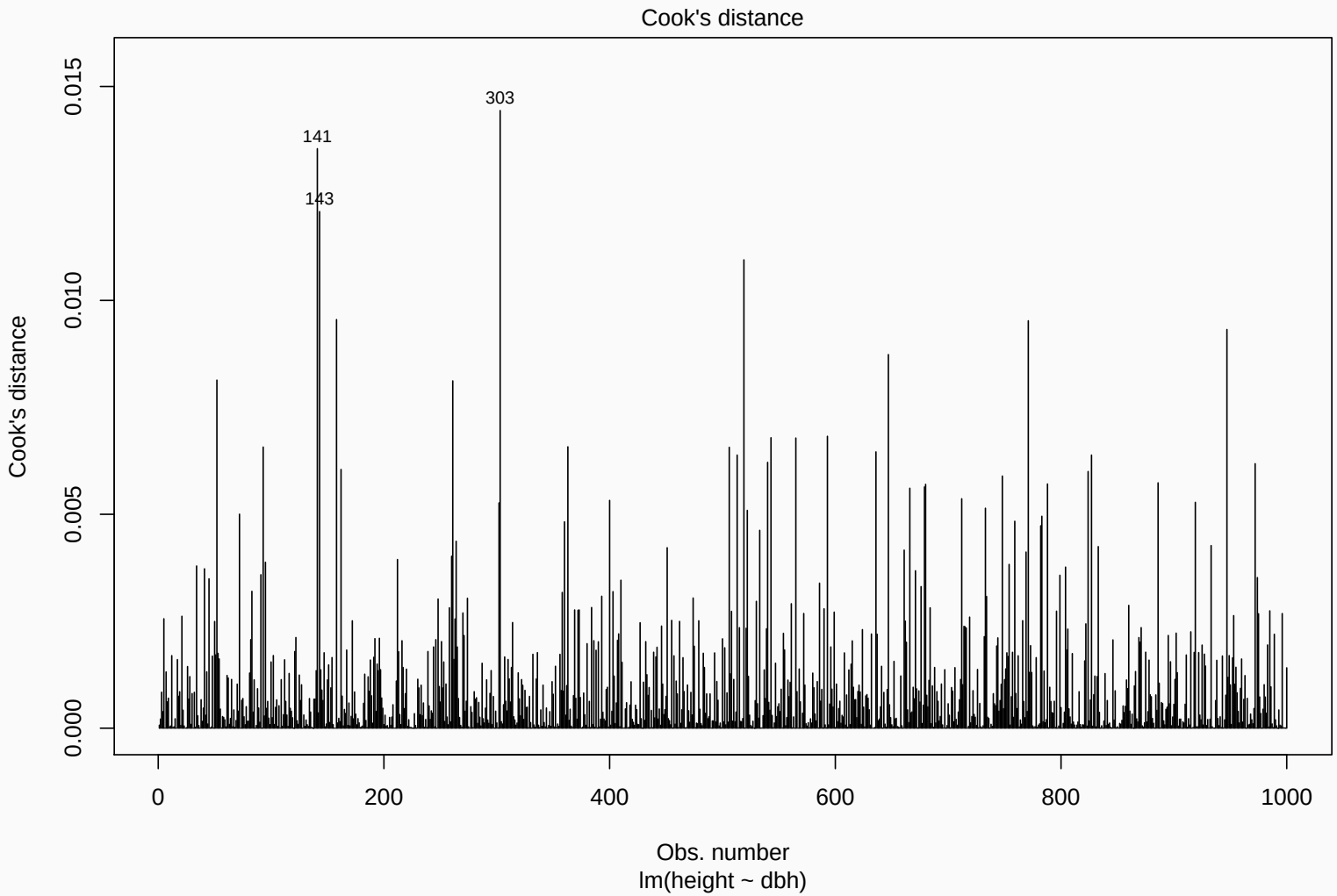
# Regresión estadística

```
plot(m1)
```



# Regresión estadística

```
plot(m1)
```



# Regresión estadística

Una vez comprobamos que el modelo ajusta bien, interpretamos los resultados

```
summary(m1)
```

```
##
## Call:
## lm(formula = height ~ dbh, data = trees)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -13.3270  -2.8978   0.1057   2.7924  12.9511
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 19.33920    0.31064   62.26   <2e-16 ***
## dbh          0.61570    0.01013   60.79   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

# Regresión estadística

```
library(broom)
tidy(m1)
```

```
## # A tibble: 2 x 5
##   term        estimate std.error statistic p.value
##   <chr>          <dbl>     <dbl>     <dbl>   <dbl>
## 1 (Intercept)   19.3      0.311       62.3       0
## 2 dbh            0.616    0.0101      60.8       0
```

Cada coeficiente tiene un valor estimado, el error asociado a ese valor (recordad el error estándar asociado a una muestra), y un p-valor.

Podemos recuperar los coeficientes directamente con

```
coef(m1)
```

```
## (Intercept)         dbh
##  19.3391968   0.6157036
```

# Regresión estadística

Nuestro modelo es:

$$height_i = 19.3392 + 0.6 \cdot dbh_i$$

# Regresión estadística

Y los intervalos de confianza (al 95%) para los coeficientes son

```
confint(m1)
```

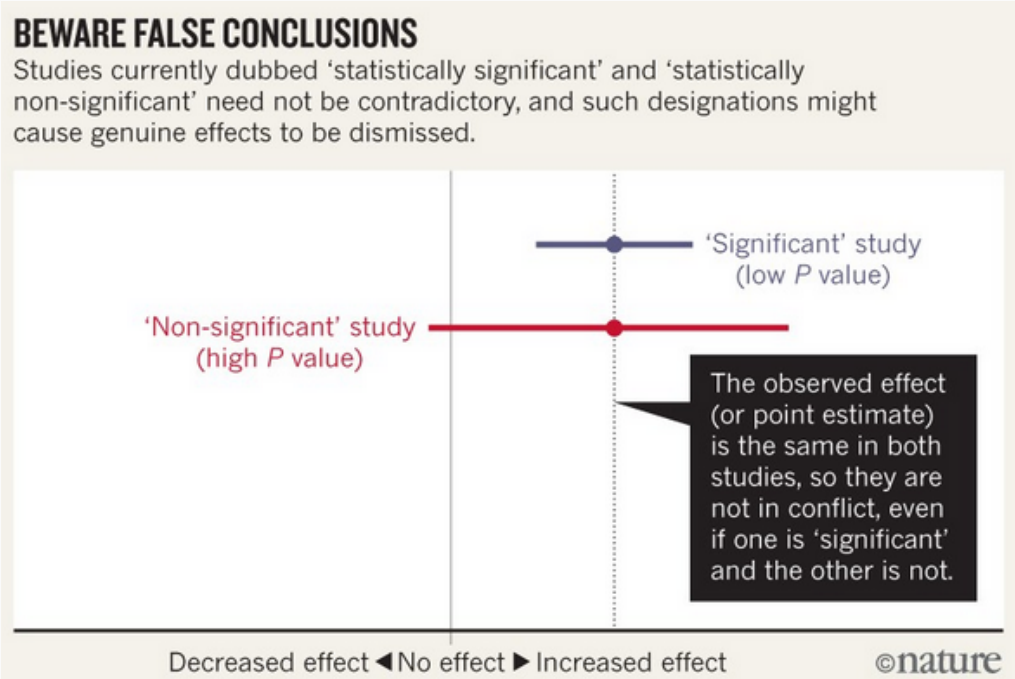
```
##                  2.5 %    97.5 %
## (Intercept) 18.7296053 19.948788
## dbh          0.5958282  0.635579
```

Recordad que un intervalo al 95% es, aproximadamente, $\mu \pm 2\sigma$. La desviación típica asociada a cada coeficiente es su error estándar. Así pues, para la pendiente de la recta (el coeficiente asociado a la DBH), $0.61 \pm 2 \cdot 0.01$ nos da los valores del intervalo.

# Regresión estadística

¿Cómo interpretar el p-valor asociado a un coeficiente?

Generalmente, se dice que si $p < 0.05$, la variable independiente tiene una relación significativa (diferente de cero) con la respuesta. Esto no es necesariamente así. Ya sabemos que 0.05 es un valor arbitrario, y que las relaciones entre variables no son dicotómicas.




Ver: https://doi.org/10.1038/d41586-019-00857-9

# Regresión estadística

¿Cómo interpretar el p-valor asociado a un coeficiente?

Es mucho más informativo comunicar el efecto asociado a una variable (e.g. aumentar una unidad de DBH implica aumentar en 0.6 unidades la altura de un árbol) y su incertidumbre asociada (su intervalo de confianza o su error estándar).

# Regresión estadística

We found a ~~significant~~ positive relationship between tree DBH and height ~~($p < 0.05$)~~ ($b = 0.61$, $SE = 0.01$)

# Regresión estadística

El último parámetro de interés es el "coeficiente de determinación", $R^2$. Nos informa de cómo de bueno es el ajuste de nuestro modelo. Literalmente, nos dice qué proporción de la varianza en los datos viene explicada por nuestro modelo. En nuestro caso,

```
summary(m1)$adj.r.squared
```

```
## [1] 0.7871477
```

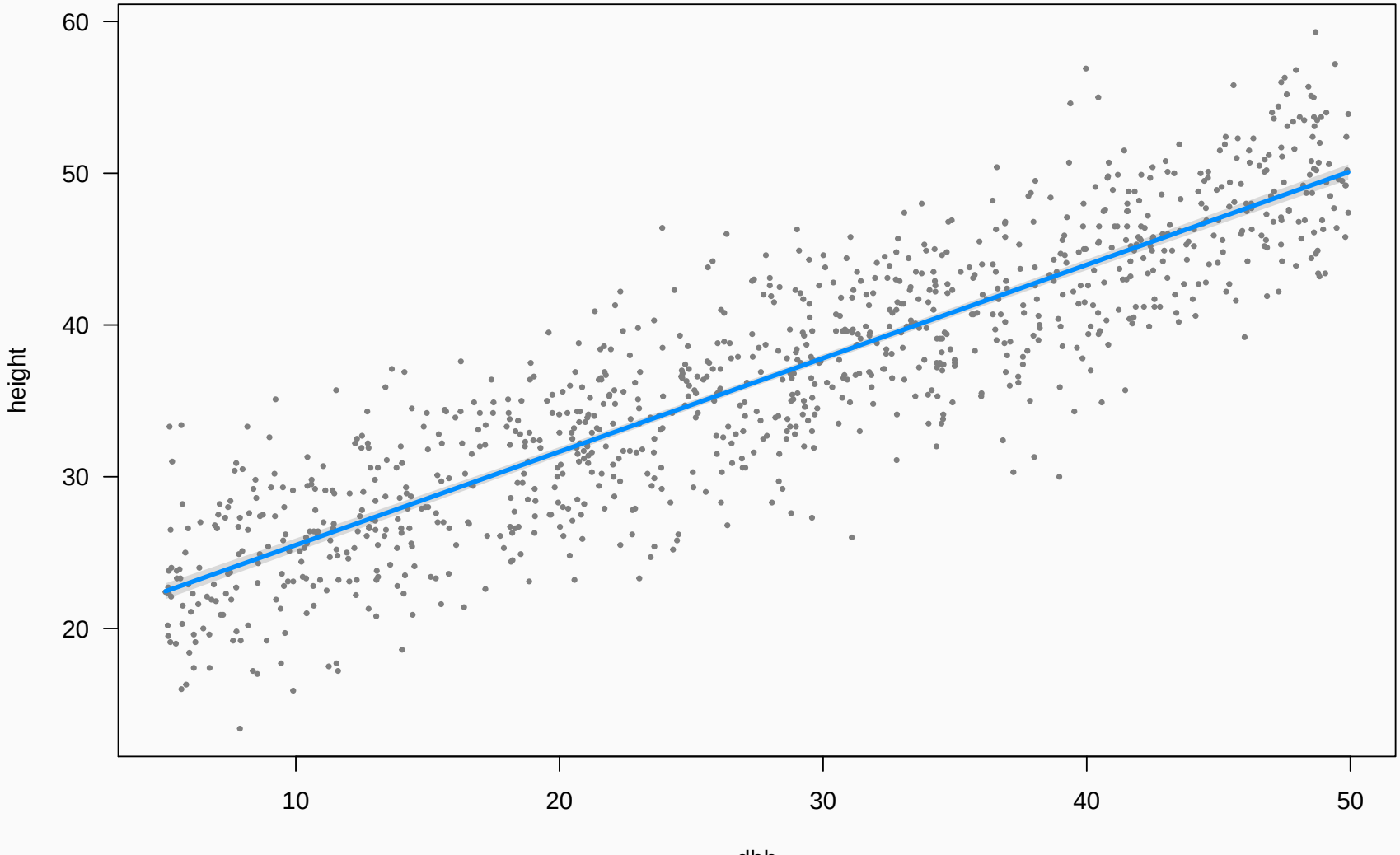
La variación en altura entre los árboles de nuestra muestra viene explicada en un 79% por su variación en DBH. Existe un 21% de variación en altura que responde potencialmente a otros factores, sean estocásticos, errores muestrales, o ecológicos.

- nota: un $R^2$ de 0.79 es *realmente alto* para los estándares de estudios en ecología...

# Regresión estadística

Visualización del modelo:

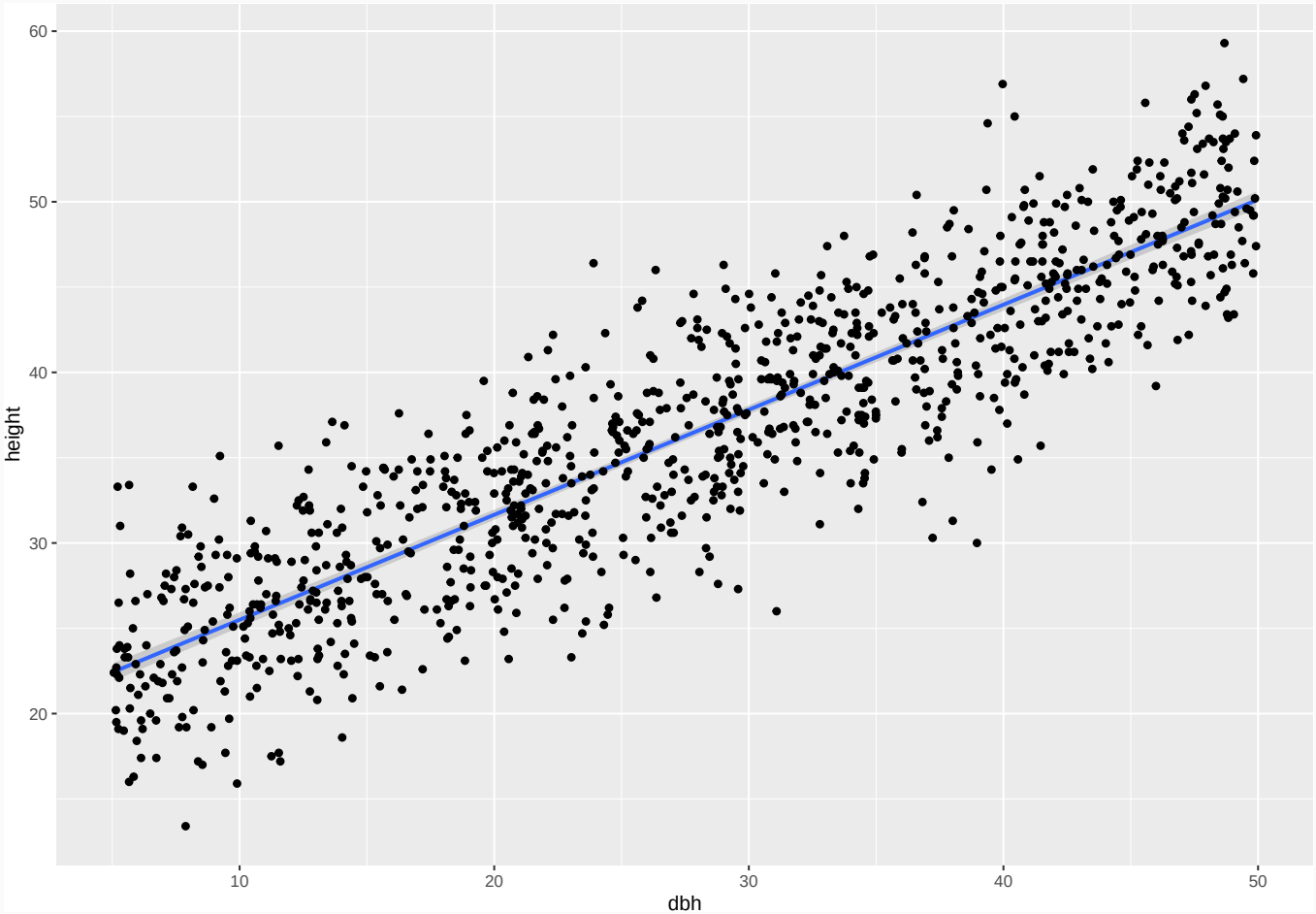
```
library(visreg)
visreg(m1)
```



# Regresión estadística

Visualización del modelo:

```
ggplot(trees, aes(x = dbh, y = height)) +
  geom_smooth(method = "lm") +
  geom_point()
```



# Regresión estadística

¿Podemos predecir la altura de un árbol nuevo, en función de su DBH?

$$height_i = 19.3392 + 0.6 \cdot dbh_i$$

```
new.dbh <-data.frame(dbh =c(12))
predict(m1, new.dbh,se.fit =TRUE)
```

```
## $fit
##        1
## 26.72764
##
## $se.fit
## [1] 0.2064598
##
## $df
## [1] 998
##
## $residual.scale
## [1] 4.092629
```

# Regresión estadística

¿Podemos predecir la altura de un árbol nuevo, en función de su DBH?

$$height_i = 19.3392 + 0.6 \cdot dbh_i$$

```
predict(m1, new.dbh,interval ="confidence")
```

```
##        fit      lwr      upr
## 1 26.72764 26.32249 27.13279
```

```
predict(m1, new.dbh,interval ="prediction")
```

```
##        fit      lwr    upr
## 1 26.72764 18.68628 34.769
```

# Regresión estadística

Estos intervalos nos ayudan a entender los dos tipos de predicciones asociadas a un modelo de regresión:

- Predecir el valor medio de la variable respuesta para un valor determinado de la variable predictora
- Predecir el valor concreto de la variable respuesta para un valor determinado de la variable predictora

En nuestro ejemplo, esto se traduce en dos cuestiones:

- Predecir la altura *media* de los árboles con una DBH determinada
- Predecir la altura *de un individuo* con una DBH determinada

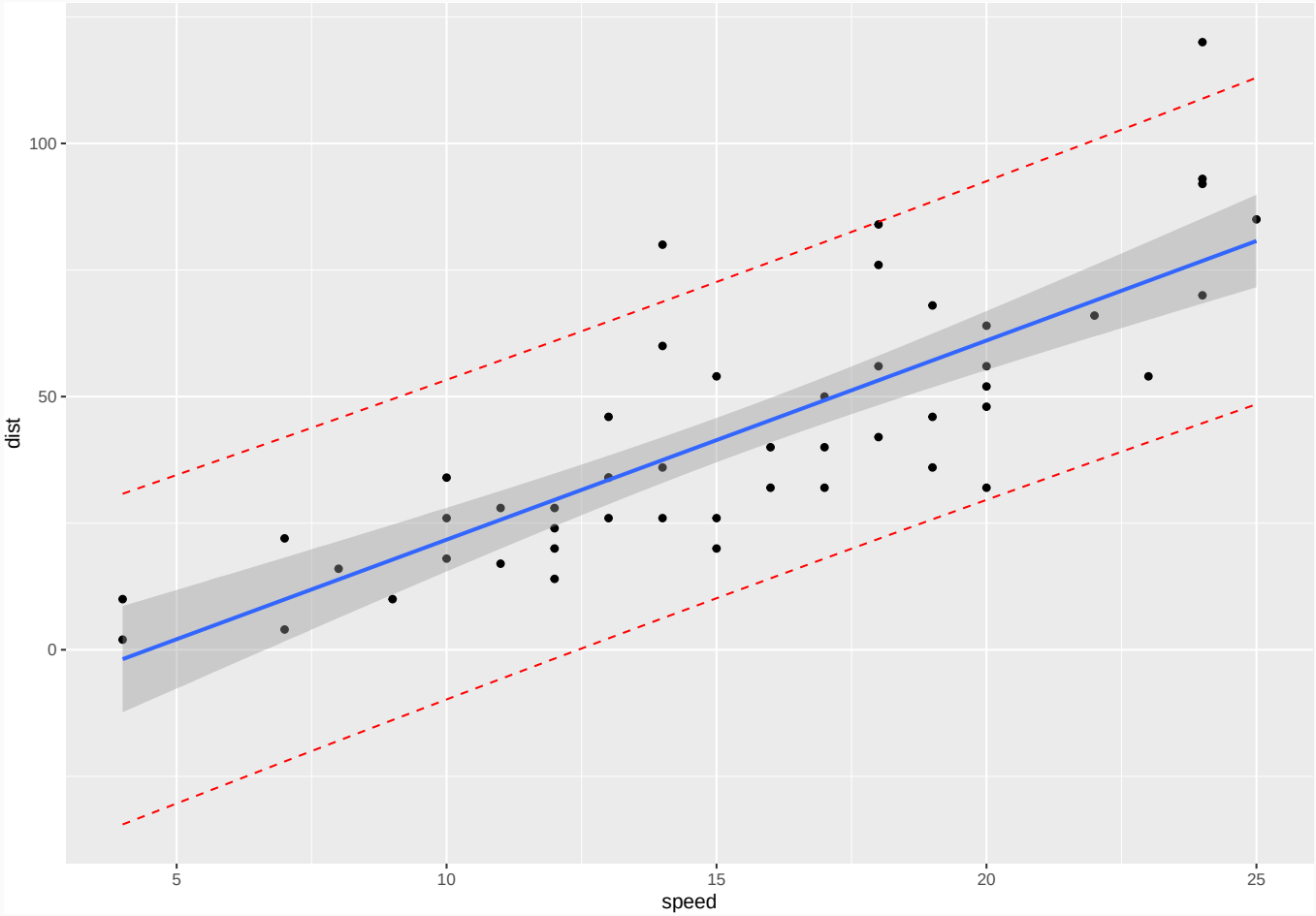
Aunque el valor predicho será el mismo en ambos casos, la incertidumbre asociada a las predicciones es diferente. La primera predicción tiene asociado un intervalo de confianza, la segunda predicción tiene asociado un intervalo de predicción.

# Regresión estadística

Ejemplo con otros datos (en los datos de árboles el intervalo de confianza es muy pequeño)

```
# 0. datos y modelo
data("cars", package = "datasets")
model <- lm(dist ~ speed, data = cars)
# predicciones
pred.int <- predict(model, interval = "prediction")
mydata <- cbind(cars, pred.int)
# visualizar recta de regresión e intervalos
library("ggplot2")
p <- ggplot(mydata, aes(speed, dist)) +
  geom_point() +
  stat_smooth(method = lm) +
  # intervalos de predicción
  geom_line(aes(y = lwr), color = "red", linetype = "dashed") +
  geom_line(aes(y = upr), color = "red", linetype = "dashed")
```

# Regresión estadística



# Regresión estadística

Pasos en la elaboración de modelos estadísticos

- Análisis exploratorio y visualización

# Regresión estadística

Pasos en la elaboración de modelos estadísticos

- Análisis exploratorio y visualización
- Ajustar modelo

# Regresión estadística

Pasos en la elaboración de modelos estadísticos

- Análisis exploratorio y visualización
- Ajustar modelo
- Comprobar residuos

# Regresión estadística

Pasos en la elaboración de modelos estadísticos

- Análisis exploratorio y visualización
- Ajustar modelo
- Comprobar residuos
- Visualizar modelo

# Regresión estadística

Pasos en la elaboración de modelos estadísticos

- Análisis exploratorio y visualización
- Ajustar modelo
- Comprobar residuos
- Visualizar modelo
- Interpretar resultados

# Regresión estadística

Pasos en la elaboración de modelos estadísticos

- Análisis exploratorio y visualización
- Ajustar modelo
- Comprobar residuos
- Visualizar modelo
- Interpretar resultados
- Predicción

# Regresión estadística

Otros casos:

- variable independiente categórica

# Regresión estadística

Otros casos:

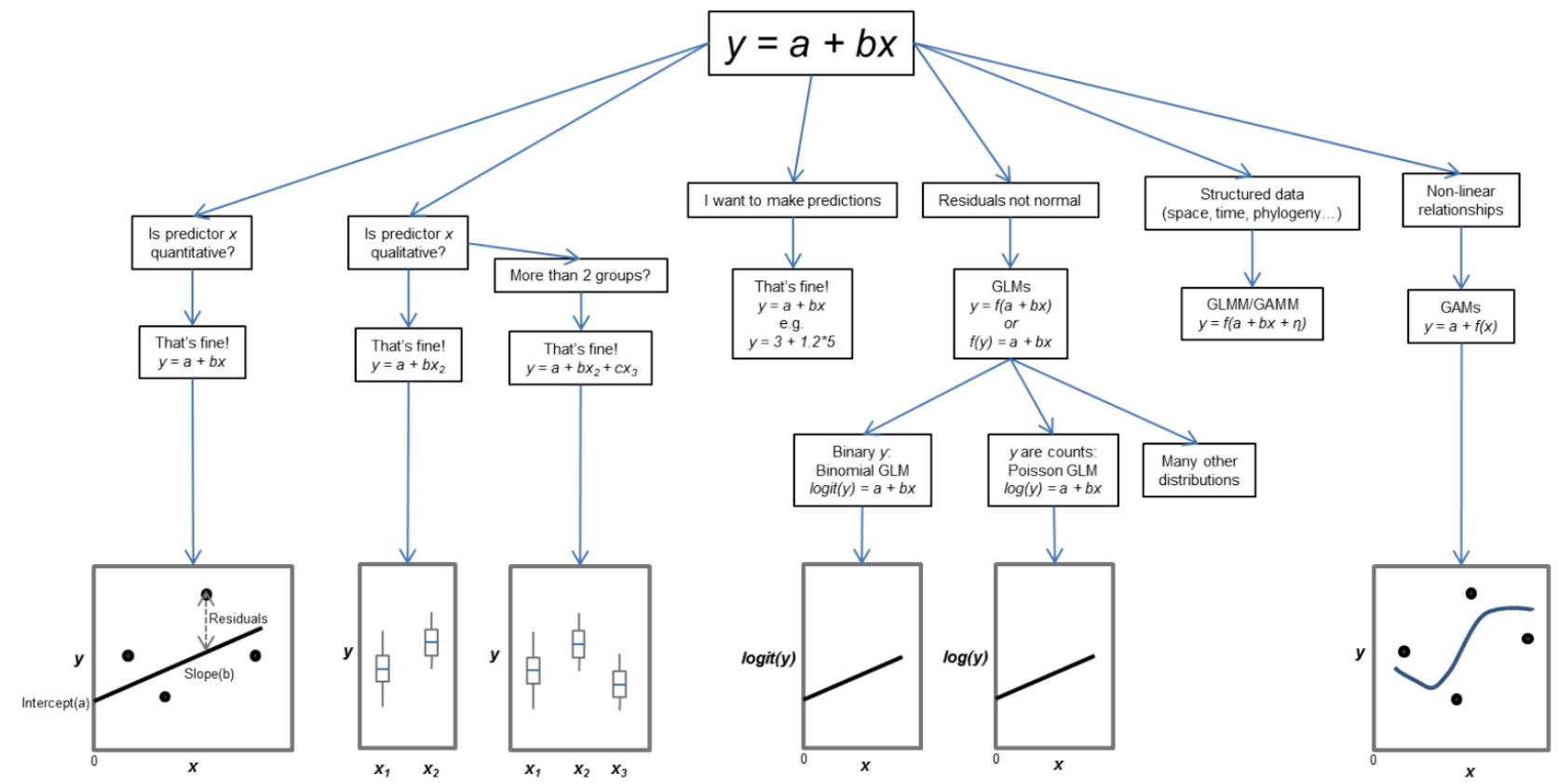
- variable independiente categórica
- múltiples variables independientes

# Regresión estadística

Otros casos:

- variable independiente categórica
- múltiples variables independientes
- datos más complejos: residuos no normales, variable respuesta discreta...

# Regresión estadística



# Regresión estadística

- Un predictor categórico: ¿Varía la altura de los árboles en función del sexo?

# Regresión estadística

- Un predictor categórico con varias categorías: ¿Varía la altura de los árboles en función del lugar de muestreo?

# Regresión estadística

- Combinando predictores categóricos y numéricos: ¿Varía la altura de los árboles en función de su DBH y el lugar de muestreo?

# Regresión estadística

- Interacciones entre predictores: ¿Varía la relación altura-DBH entre lugares de muestreo?

# Regresión estadística

https://bookdown.org/speegled/foundations-of-statistics/

- marco general (UHU, WS p671)

# Regresión estadística

https://bookdown.org/speegled/foundations-of-statistics/

- marco general (UHU, WS p671)
- residuos (WS p689)

# Regresión estadística

https://bookdown.org/speegled/foundations-of-statistics/

- marco general (UHU, WS p671)
- residuos (WS p689)
- distribucion normal

# Regresión estadística

https://bookdown.org/speegled/foundations-of-statistics/

- marco general (UHU, WS p671)
- residuos (WS p689)
- distribucion normal
- datos paco trees, primer lm

# Regresión estadística

https://bookdown.org/speegled/foundations-of-statistics/

- marco general (UHU, WS p671)
- residuos (WS p689)
- distribucion normal
- datos paco trees, primer lm
- visualización

# Regresión estadística

https://bookdown.org/speegled/foundations-of-statistics/

- marco general (UHU, WS p671)
- residuos (WS p689)
- distribucion normal
- datos paco trees, primer lm
- visualización
- ajuste

# Regresión estadística

https://bookdown.org/speegled/foundations-of-statistics/

- marco general (UHU, WS p671)
- residuos (WS p689)
- distribucion normal
- datos paco trees, primer lm
- visualización
- ajuste
- interpretación y comunicación de resultados (incluyendo effect sizes)

# Regresión estadística

https://bookdown.org/speegled/foundations-of-statistics/

- marco general (UHU, WS p671)
- residuos (WS p689)
- distribucion normal
- datos paco trees, primer lm
- visualización
- ajuste
- interpretación y comunicación de resultados (incluyendo effect sizes)
- validación (asunciones)

# Regresión estadística

https://bookdown.org/speegled/foundations-of-statistics/

- marco general (UHU, WS p671)
- residuos (WS p689)
- distribucion normal
- datos paco trees, primer lm
- visualización
- ajuste
- interpretación y comunicación de resultados (incluyendo effect sizes)
- validación (asunciones)
- predicción

# Regresión estadística

https://bookdown.org/speegled/foundations-of-statistics/

- marco general (UHU, WS p671)
- residuos (WS p689)
- distribucion normal
- datos paco trees, primer lm
- visualización
- ajuste
- interpretación y comunicación de resultados (incluyendo effect sizes)
- validación (asunciones)
- predicción
- tipos de variables independientes (continuas y categóricas)

# Regresión estadística

https://bookdown.org/speegled/foundations-of-statistics/

- marco general (UHU, WS p671)
- residuos (WS p689)
- distribucion normal
- datos paco trees, primer lm
- visualización
- ajuste
- interpretación y comunicación de resultados (incluyendo effect sizes)
- validación (asunciones)
- predicción
- tipos de variables independientes (continuas y categóricas)
- selección de modelos (R2)

# Generalized linear models

- distribuciones continuas y discretas

# Generalized linear models

- distribuciones continuas y discretas
- likelihood (WS p814)

# Generalized linear models

- distribuciones continuas y discretas
- likelihood (WS p814)
- esquema general: distribución de residuos, fórmula, función de enlace

# Generalized linear models

- distribuciones continuas y discretas
- likelihood (WS p814)
- esquema general: distribución de residuos, fórmula, función de enlace
- regresión logística (WS p701)

# Generalized linear models

- distribuciones continuas y discretas
- likelihood (WS p814)
- esquema general: distribución de residuos, fórmula, función de enlace
- regresión logística (WS p701)
- regresión de conteos (poisson, negbin)

# Generalized linear models

- distribuciones continuas y discretas
- likelihood (WS p814)
- esquema general: distribución de residuos, fórmula, función de enlace
- regresión logística (WS p701)
- regresión de conteos (poisson, negbin)
- selección de modelos (AIC)