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

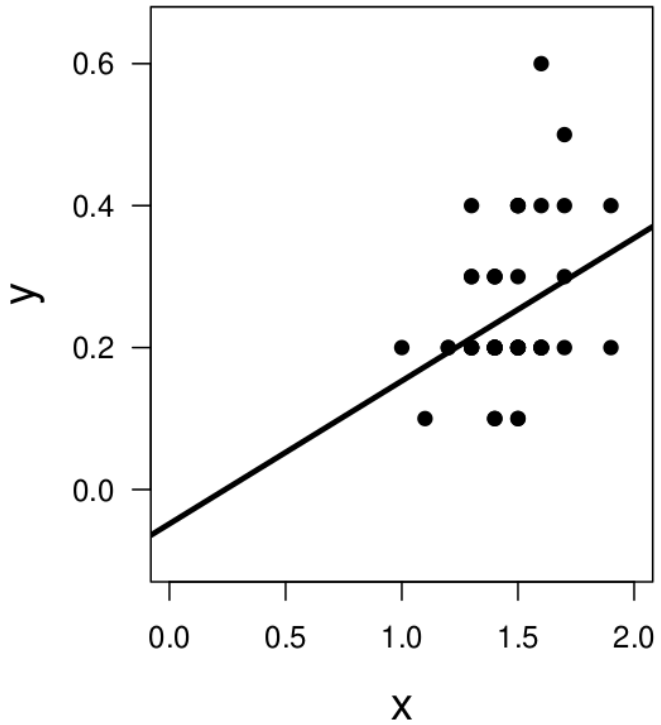
Hasta ahora:

- 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

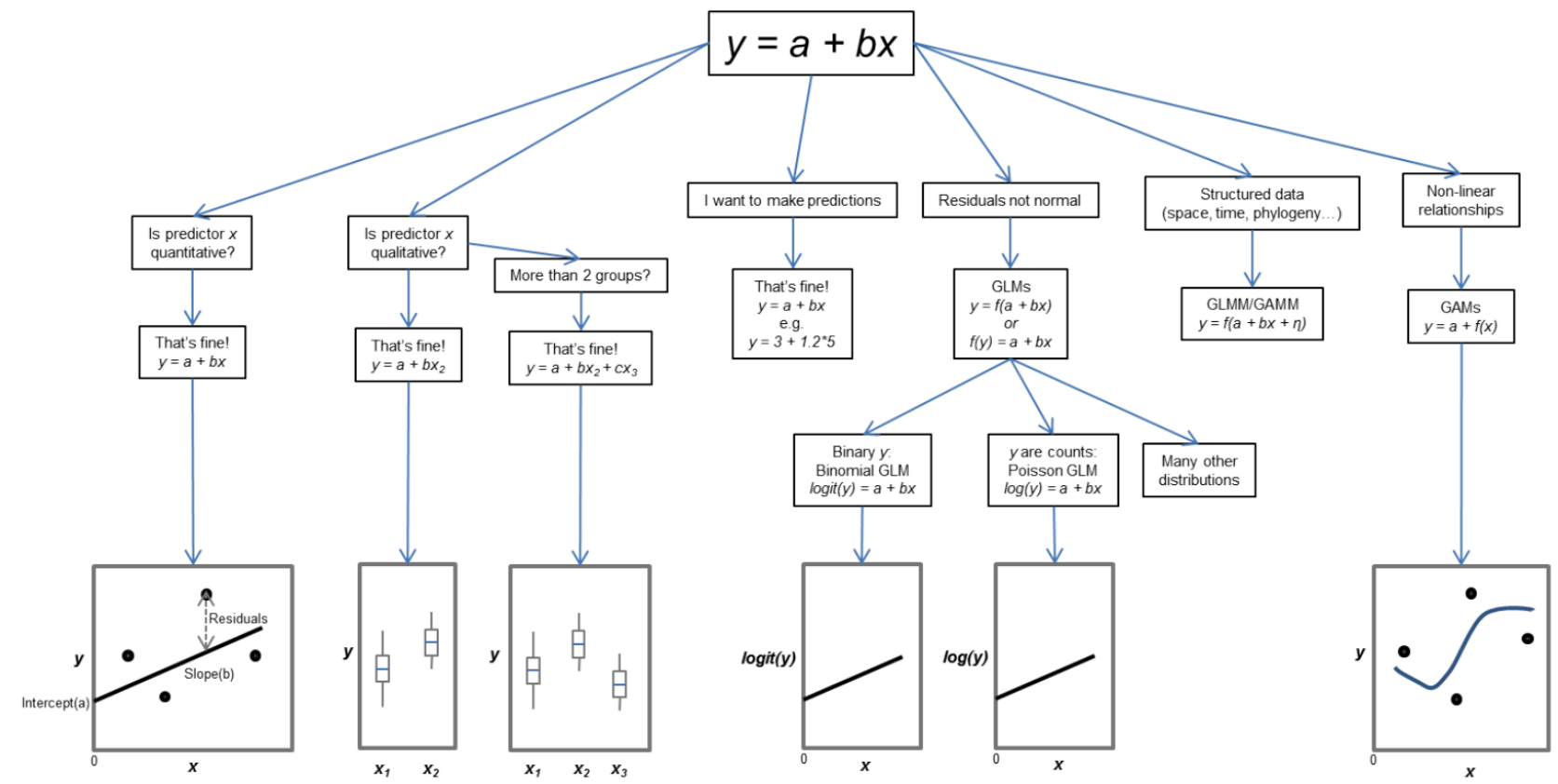
a = intercept

b = slope

σ = residual variation

ε = residuals

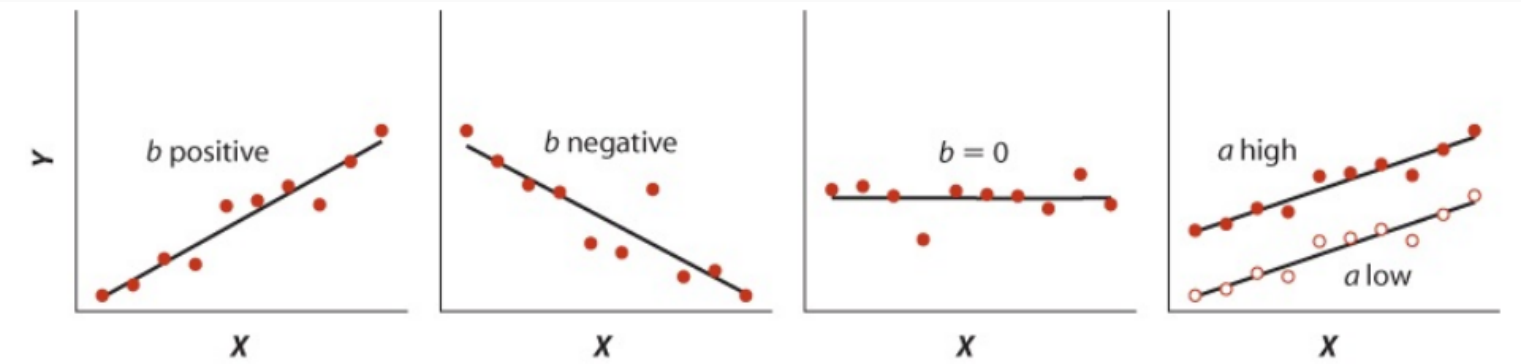
Regresión estadística



Regresión estadística

Regresión lineal:

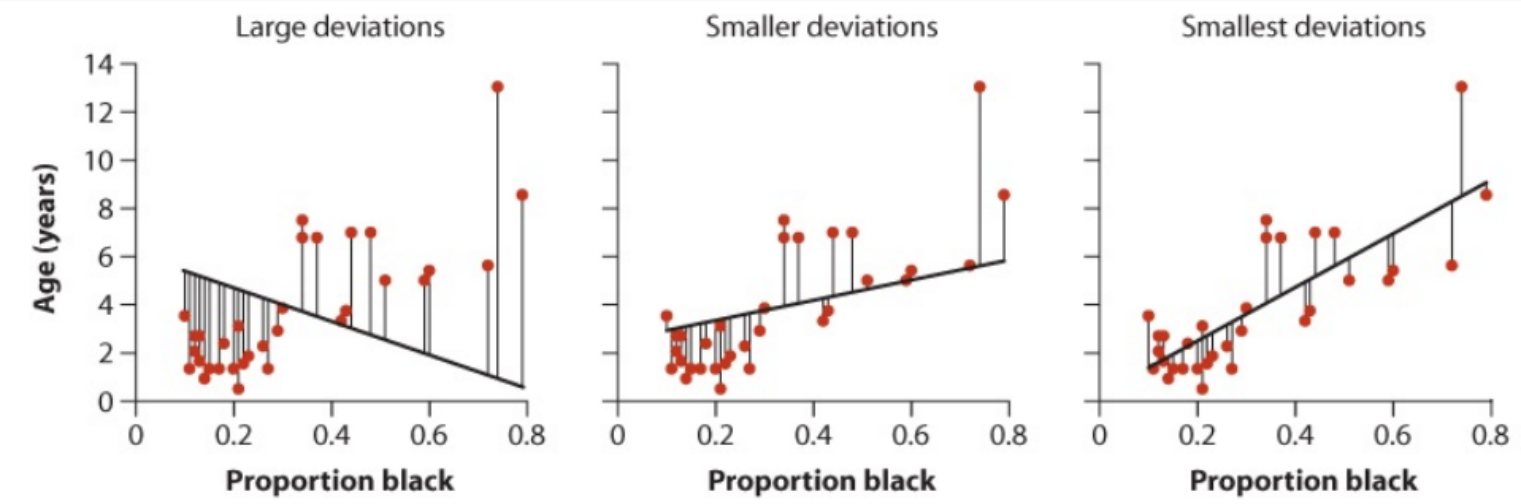
- Relación lineal entre las variables



Regresión estadística

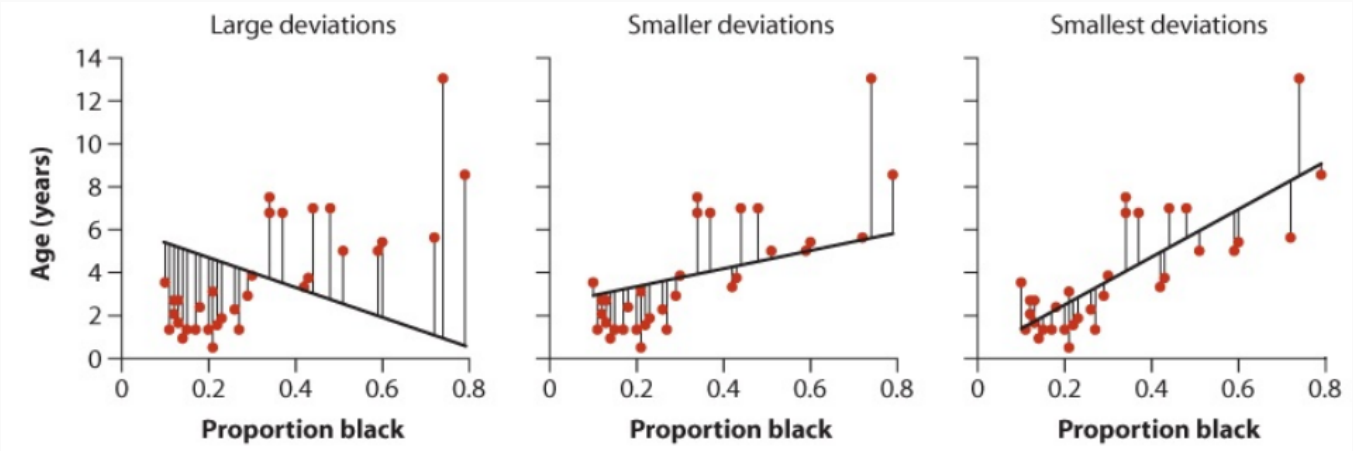
Regresión lineal:

- Minimiza el *error residual*



Regresión estadística

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



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

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

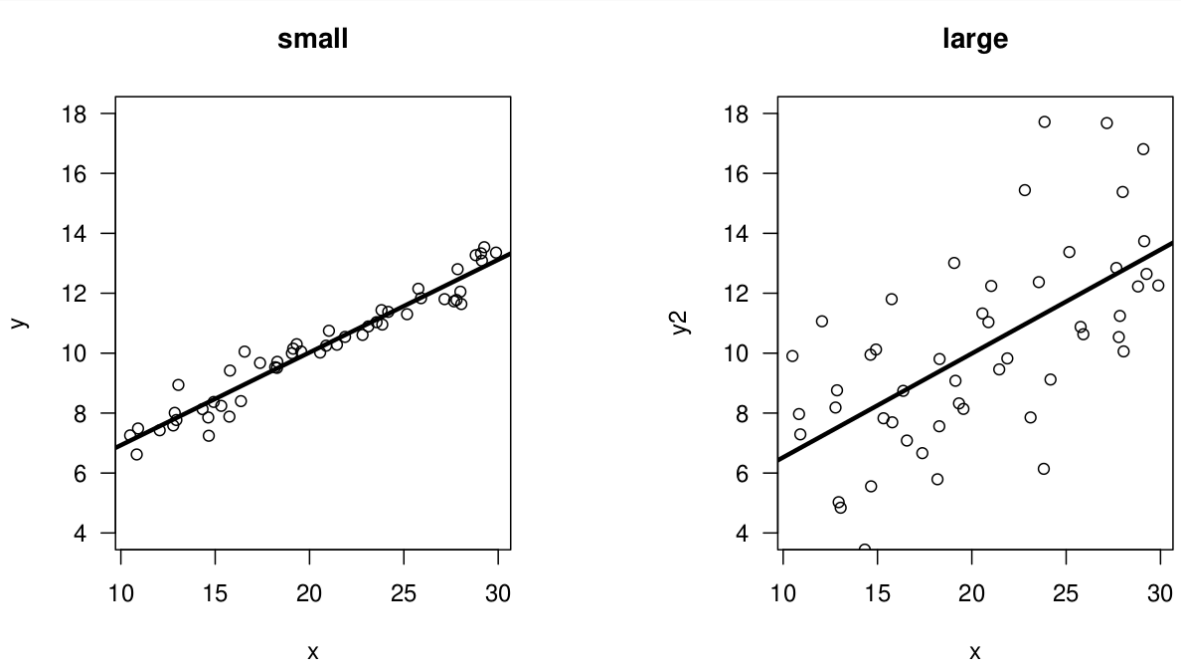
Regresión estadística

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

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

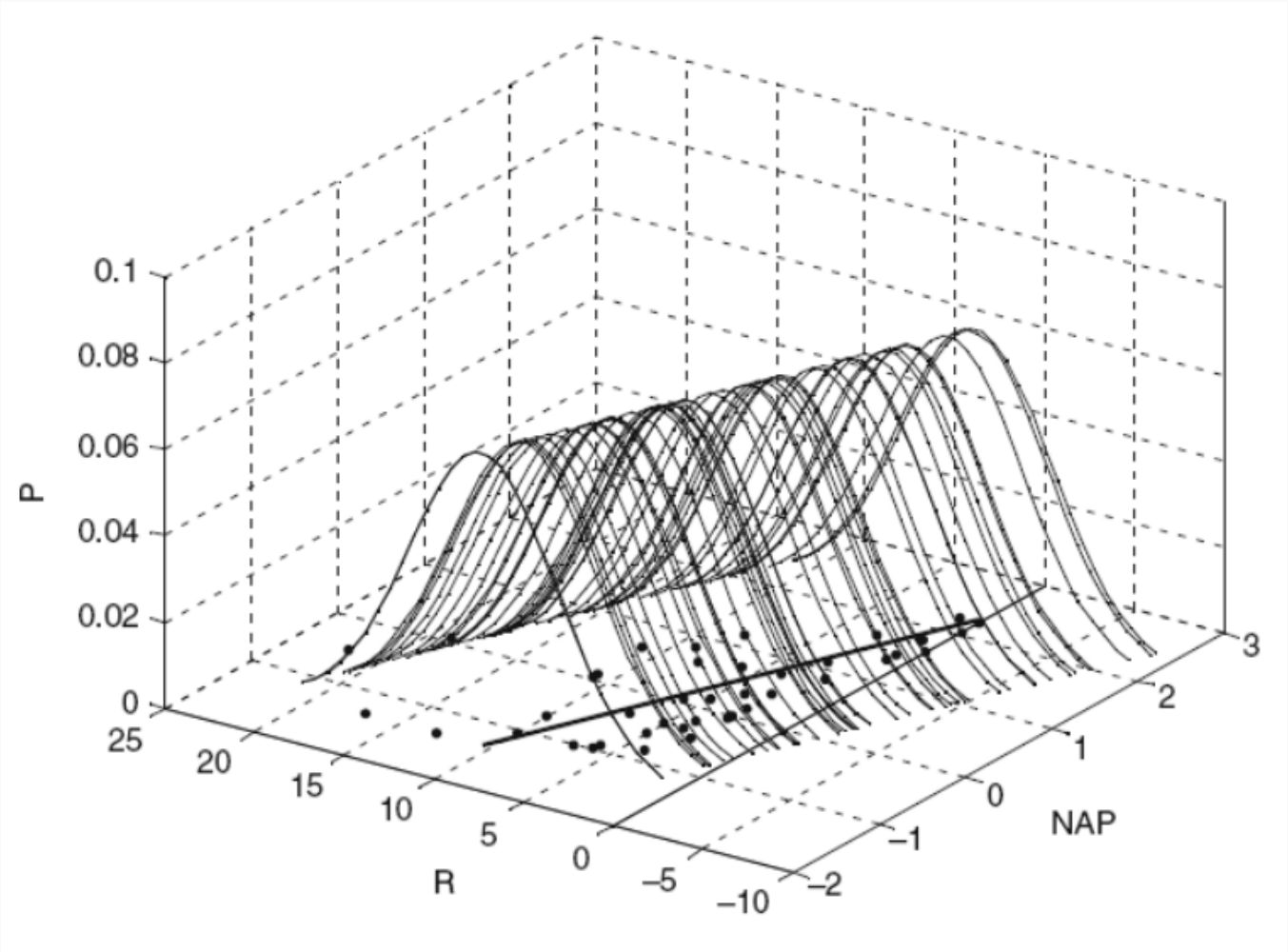
Regresión estadística

- Residuos: diferencia entre valor observado y predicho



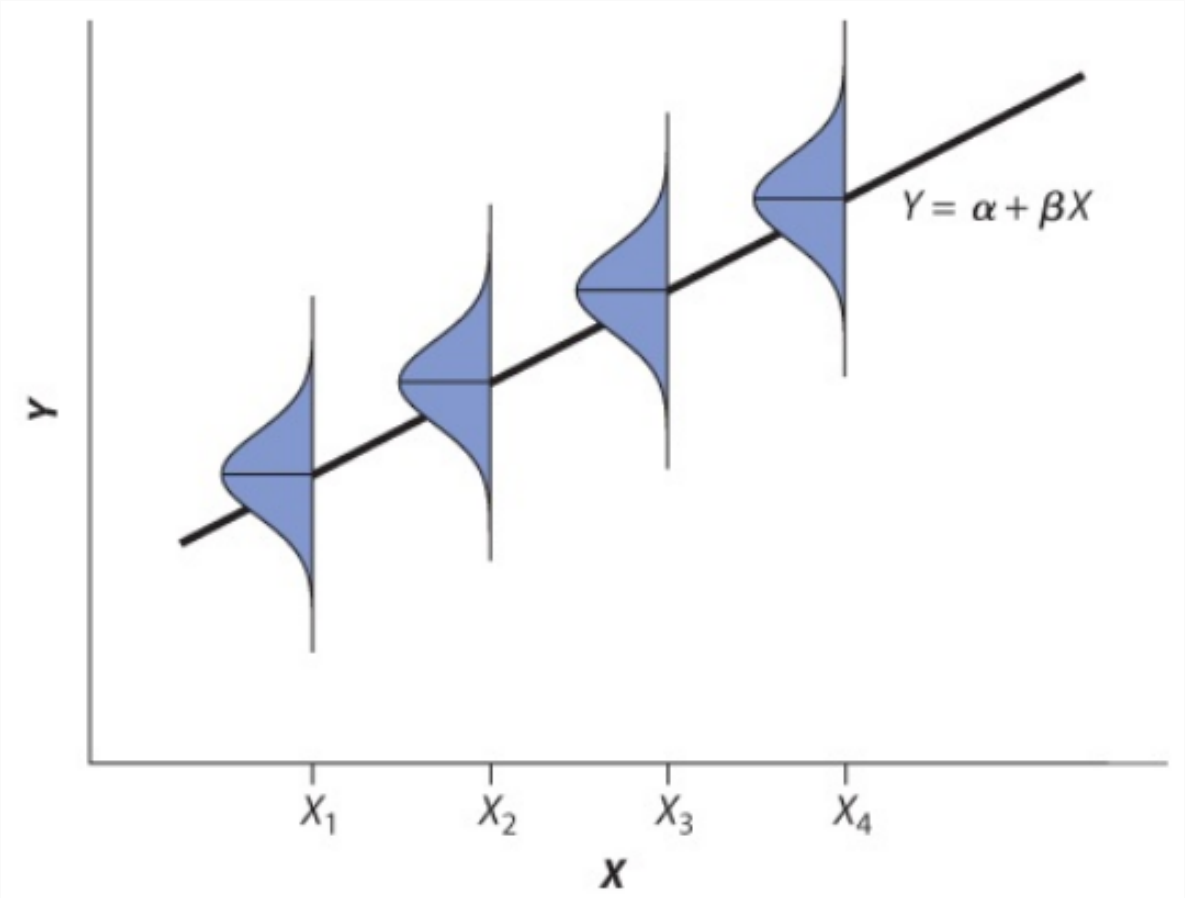
Regresión estadística

- 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

Asunciones de la regresión lineal:

- variable respuesta: normal

Regresión estadística

Asunciones de la regresión lineal:

- variable respuesta: normal
- distribución de residuos: normal

Regresión estadística

Asunciones de la regresión lineal:

- variable respuesta: normal
- distribución de residuos: normal
- varianza residual: homogénea

Regresión estadística

Asunciones de la regresión lineal:

- variable respuesta: normal
- distribución de residuos: normal
- varianza residual: homogénea
- observaciones independientes entre sí

Regresión estadística

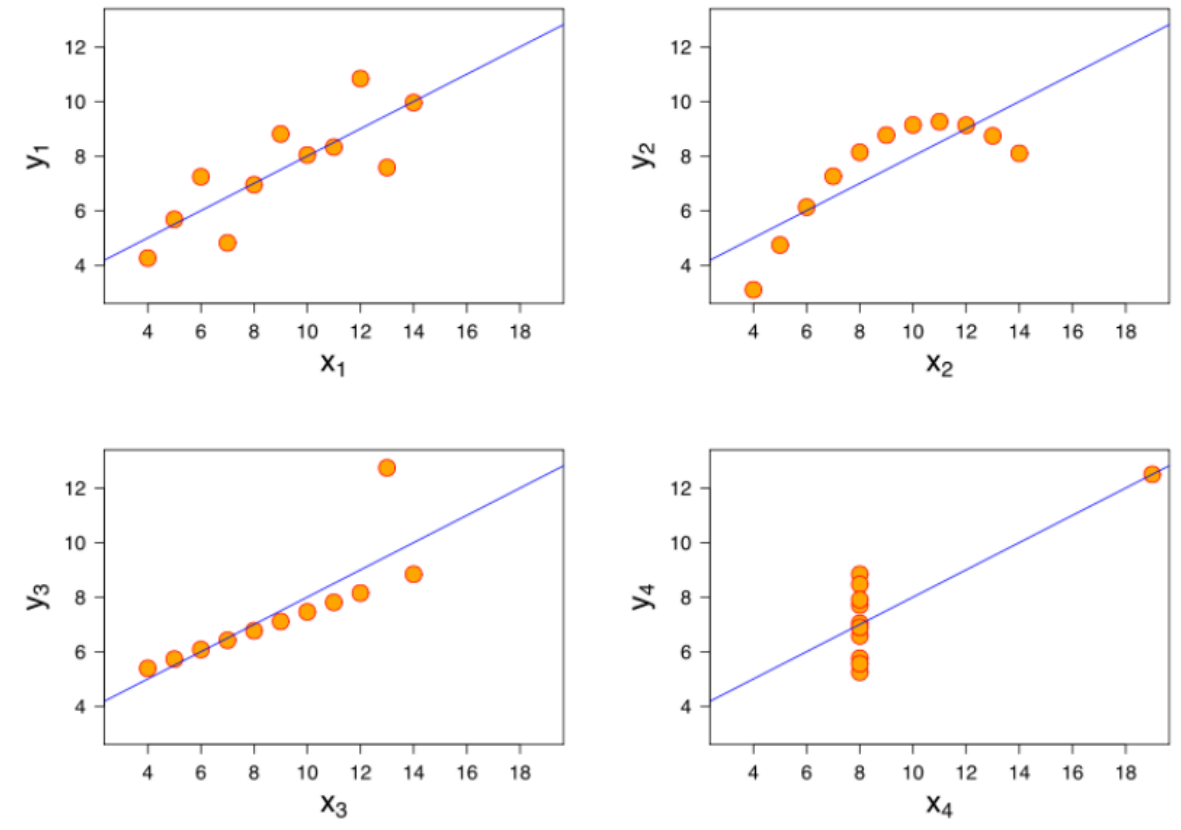
¿Podemos predecir la altura de un árbol a partir de su *dbh*?

```
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

- Visualizando el cuarteto de Anscombe

```
aq <- datasets::anscombe
head(aq)
```

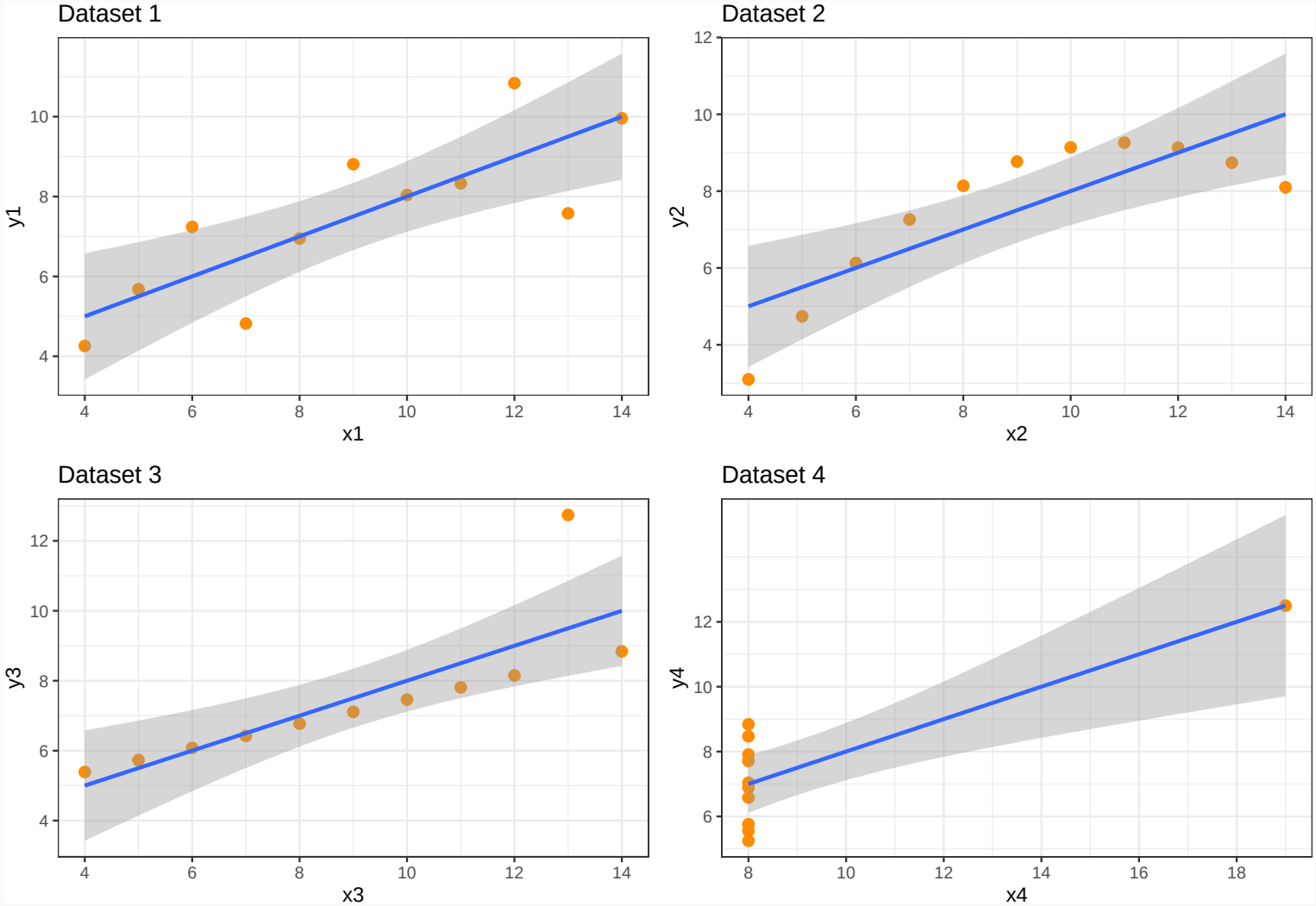
```
##   x1 x2 x3 x4   y1   y2    y3   y4
## 1 10 10 10  8 8.04 9.14  7.46 6.58
## 2  8  8  8  8 6.95 8.14  6.77 5.76
## 3 13 13 13  8 7.58 8.74 12.74 7.71
## 4  9  9  9  8 8.81 8.77  7.11 8.84
## 5 11 11 11  8 8.33 9.26  7.81 8.47
## 6 14 14 14  8 9.96 8.10  8.84 7.04
```

Regresión estadística

```
p1 <- ggplot(aq,aes(x1, y1)) +
  geom_point(color = "darkorange", size = 2.5) +
  scale_x_continuous(breaks = seq(0,20,2)) +
  scale_y_continuous(breaks = seq(0,12,2)) +
  # expand_limits(x = 0, y = 0) +
  labs(x = "x1", y = "y1",
       title = "Dataset 1" ) +
  geom_smooth(method = "lm") +
  theme_bw()
```

Regresión estadística

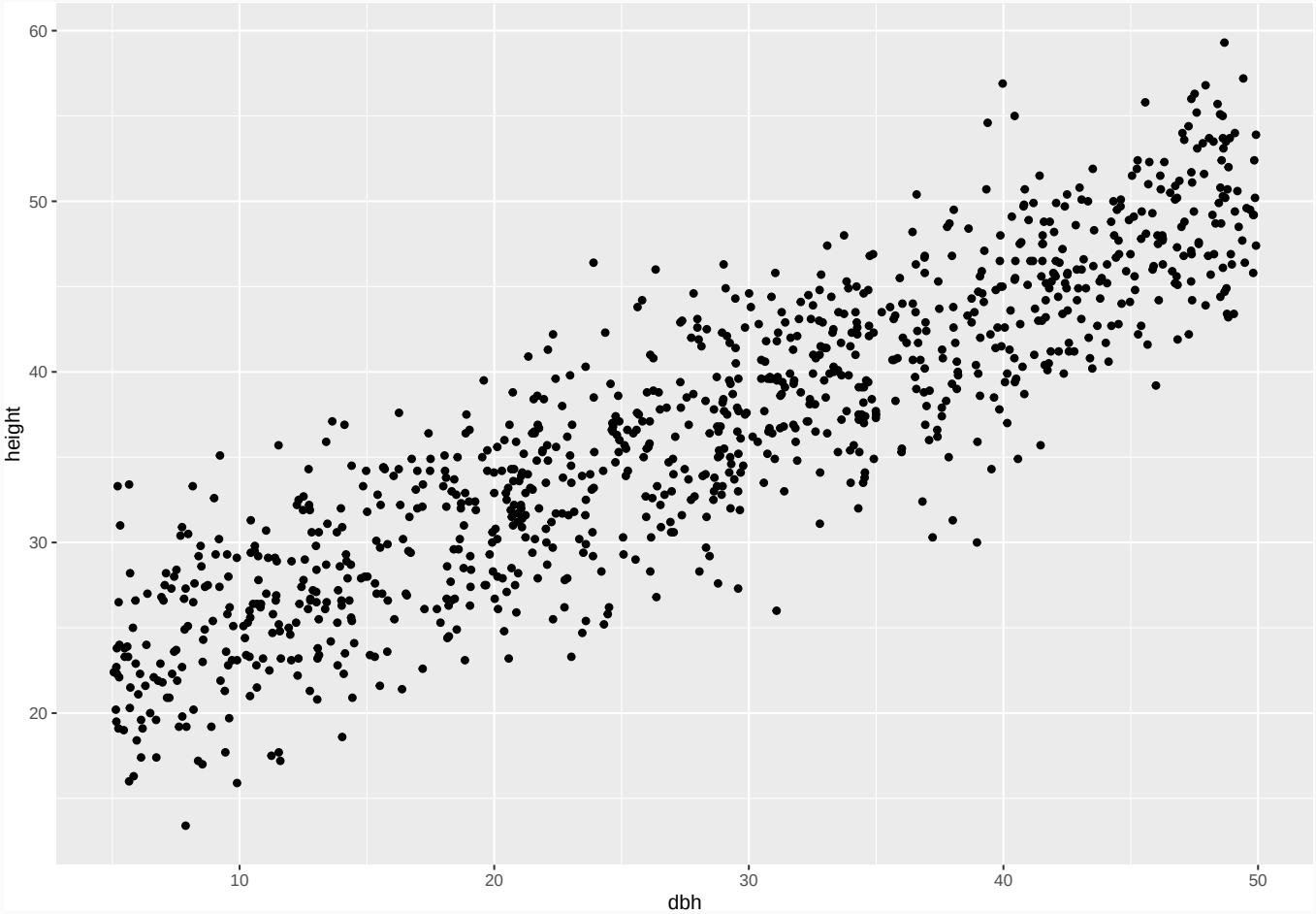
```
library(patchwork)
wrap_plots(p1,p2,p3,p4)
```



Regresión estadística

- Volvemos a nuestros datos de árboles: ¿Hay outliers en los datos?

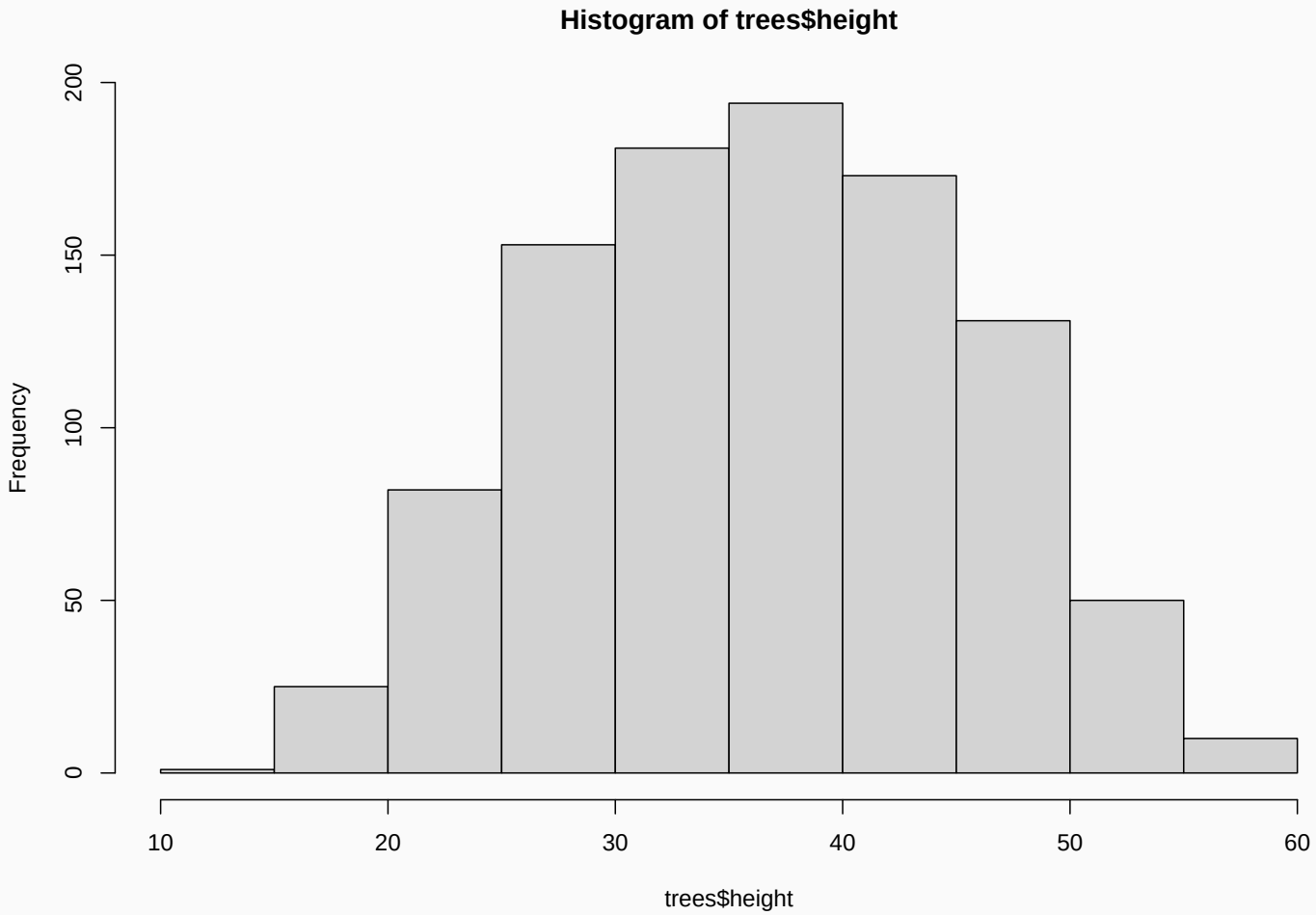
```
ggplot(trees, aes(dbh, height)) +
geom_point()
```



Regresión estadística

¿Cómo están distribuidas las variables independientes y respuesta?

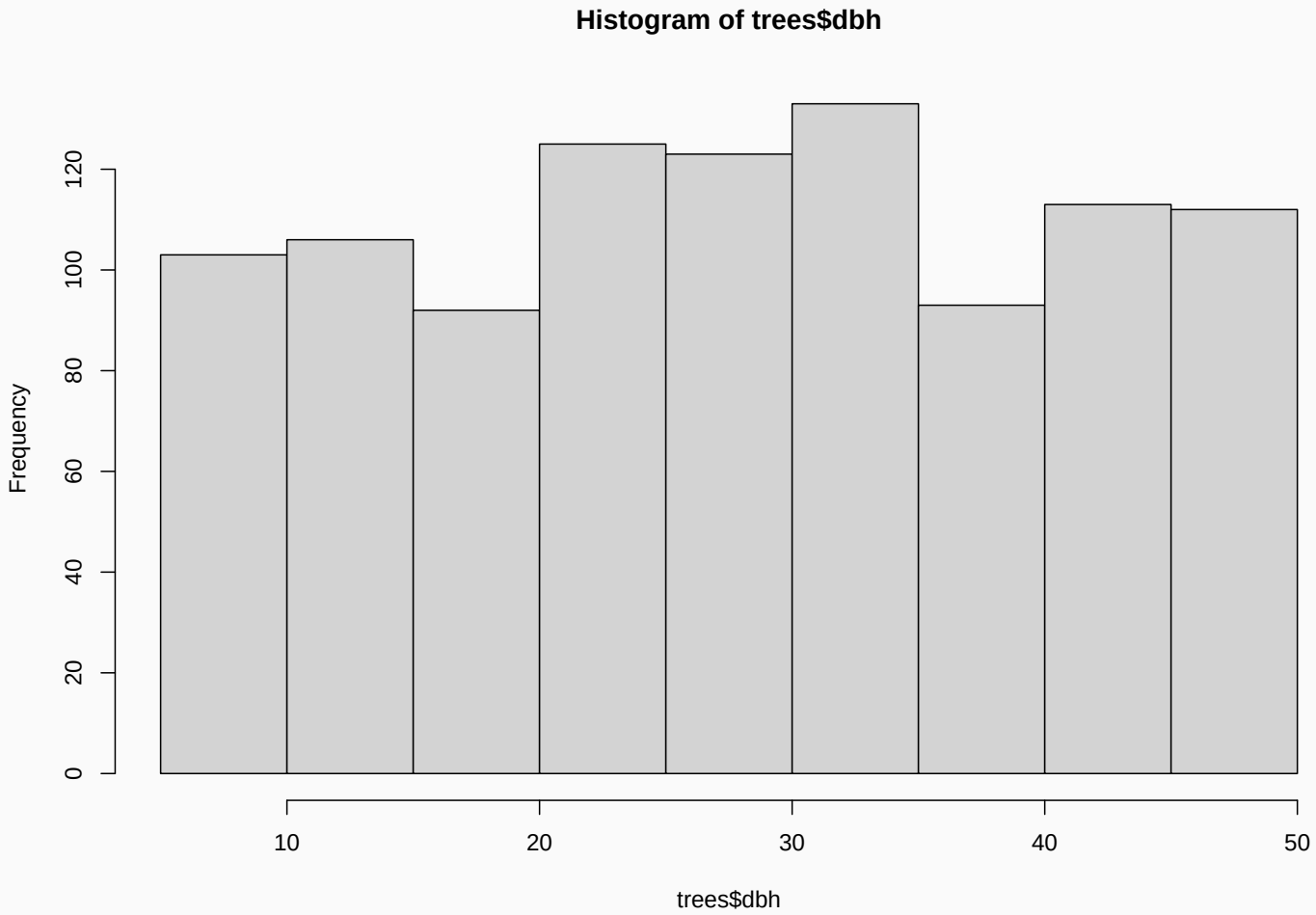
```
hist(trees$height)
```



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

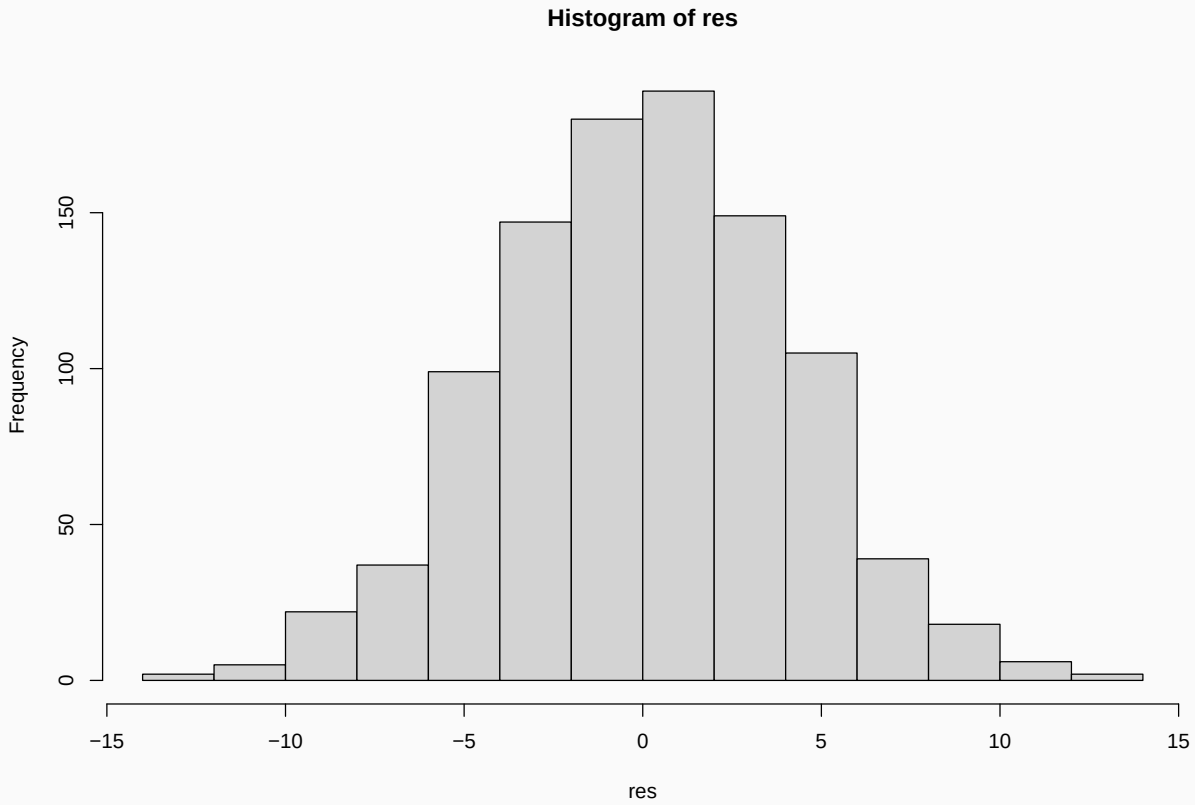
```
summary(m1)
```

```
##
## 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

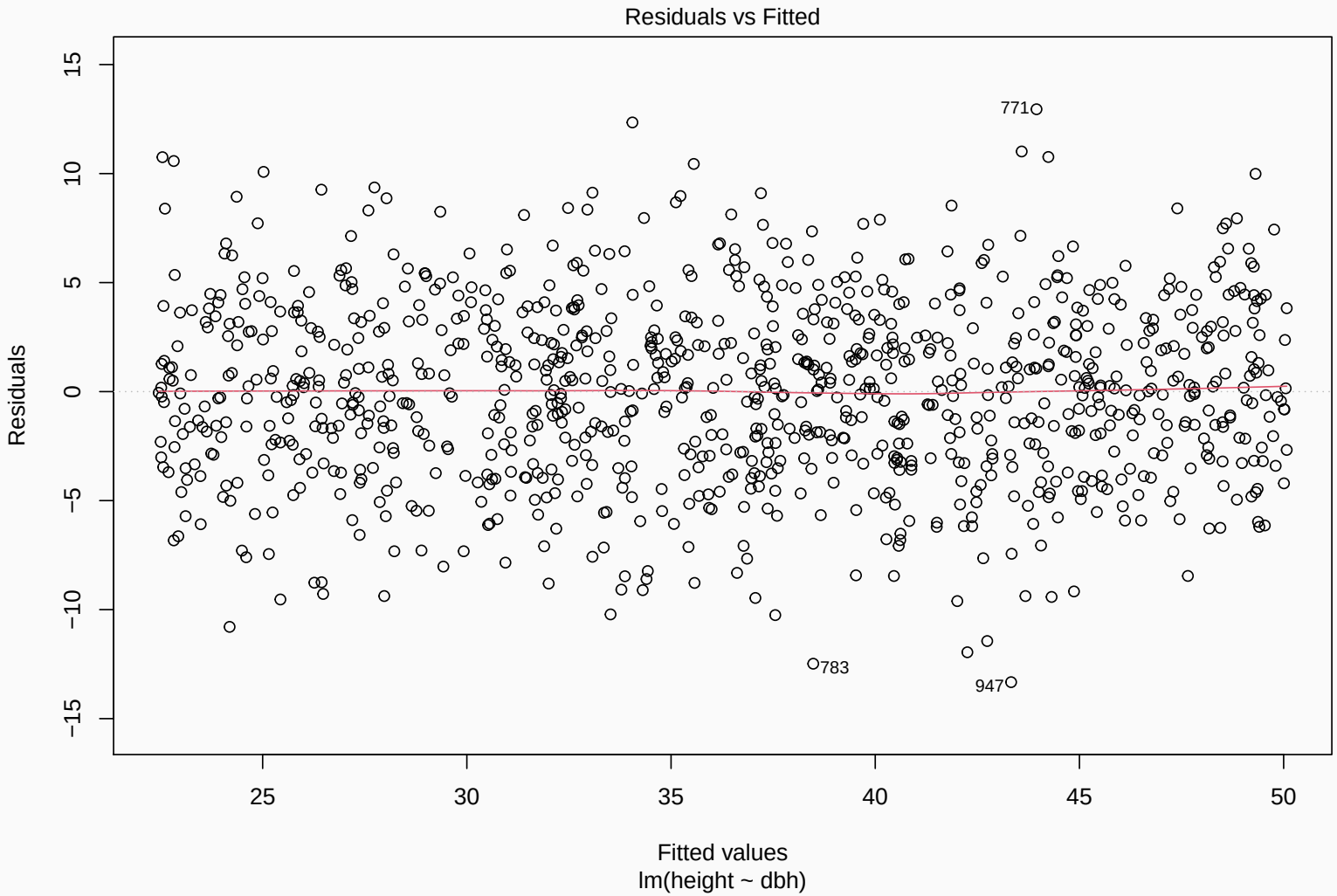
Antes de interpretar el resultado, comprobamos que los residuos se ajustan a una distribución normal

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



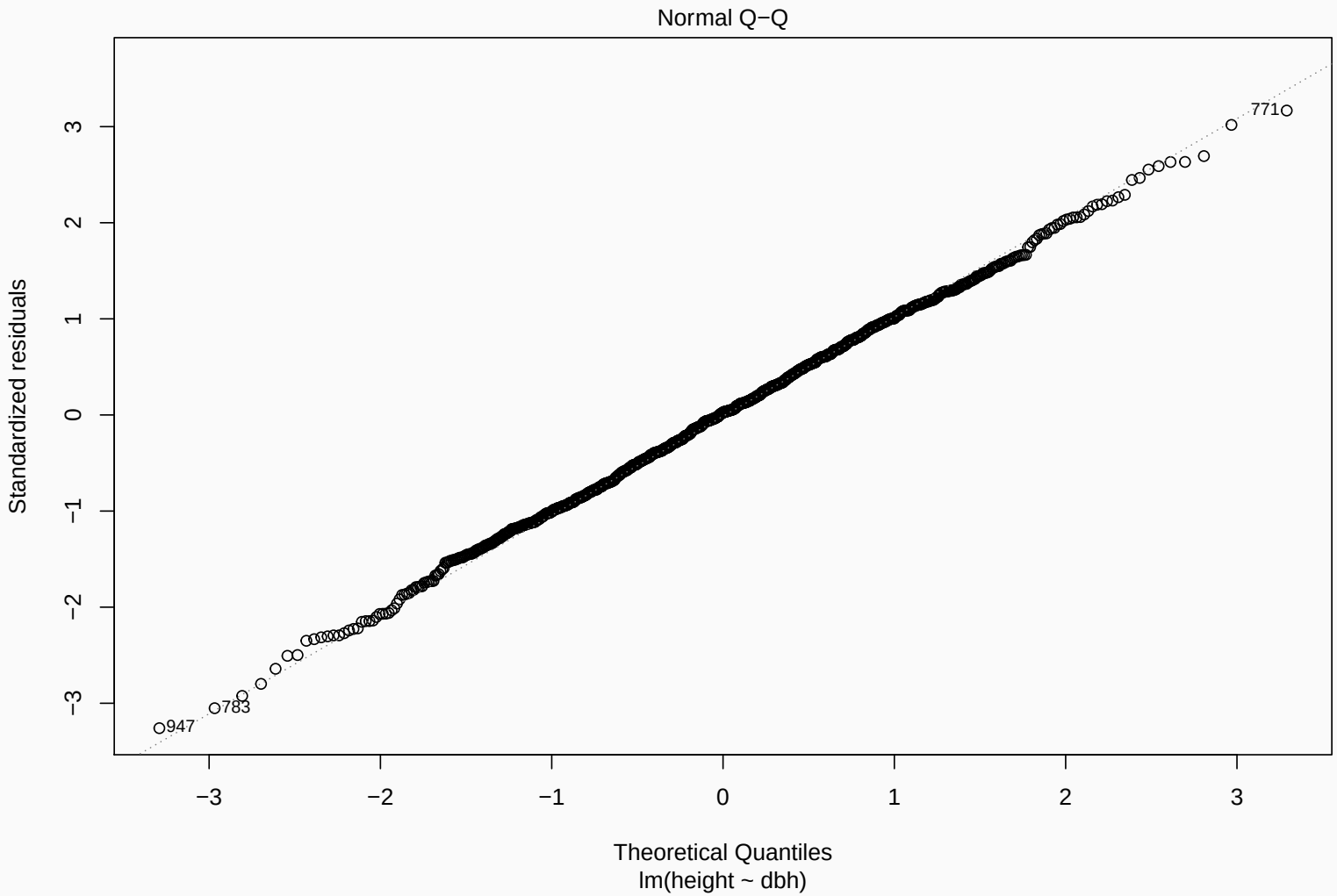
Regresión estadística

```
plot(m1)
```



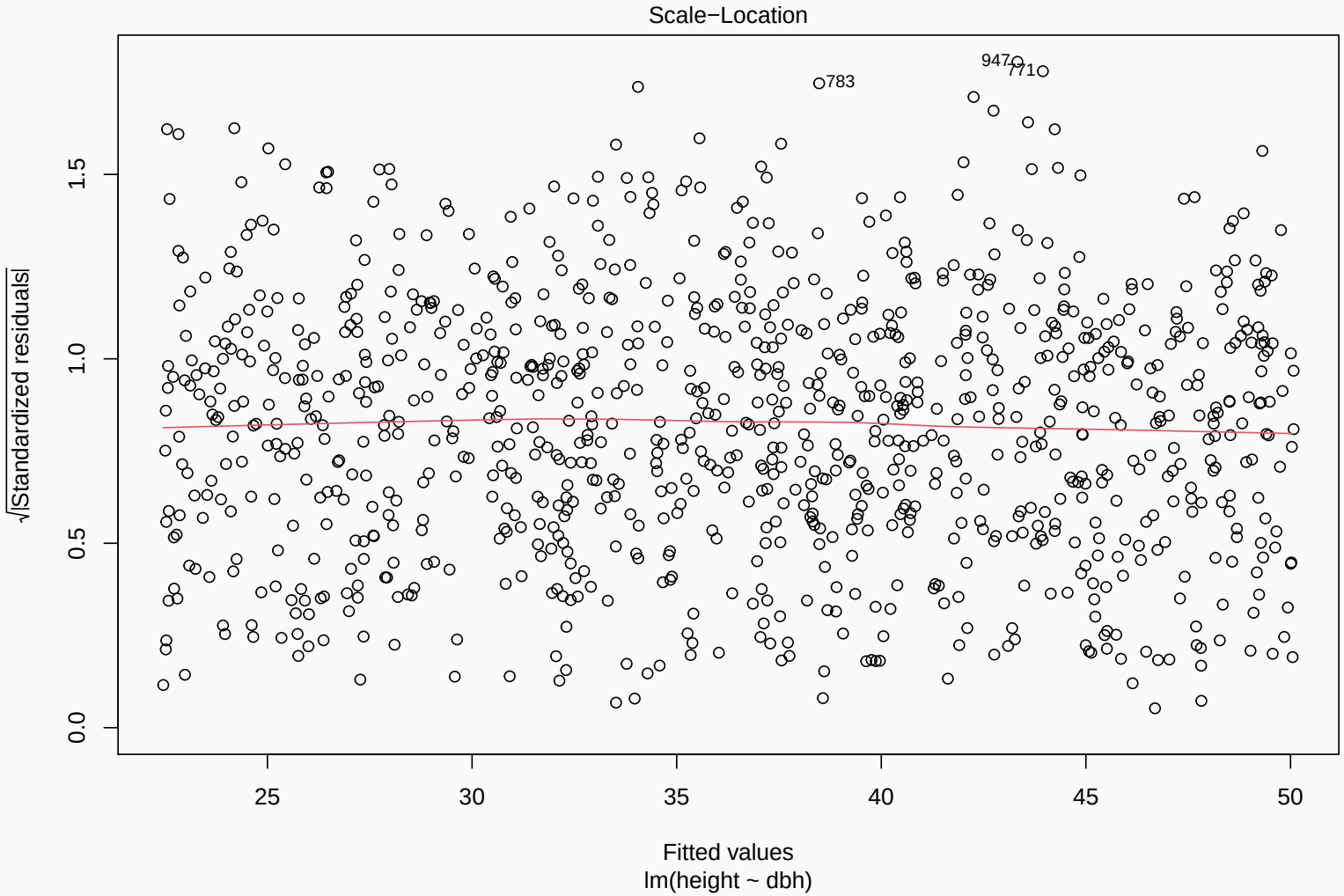
Regresión estadística

```
plot(m1)
```



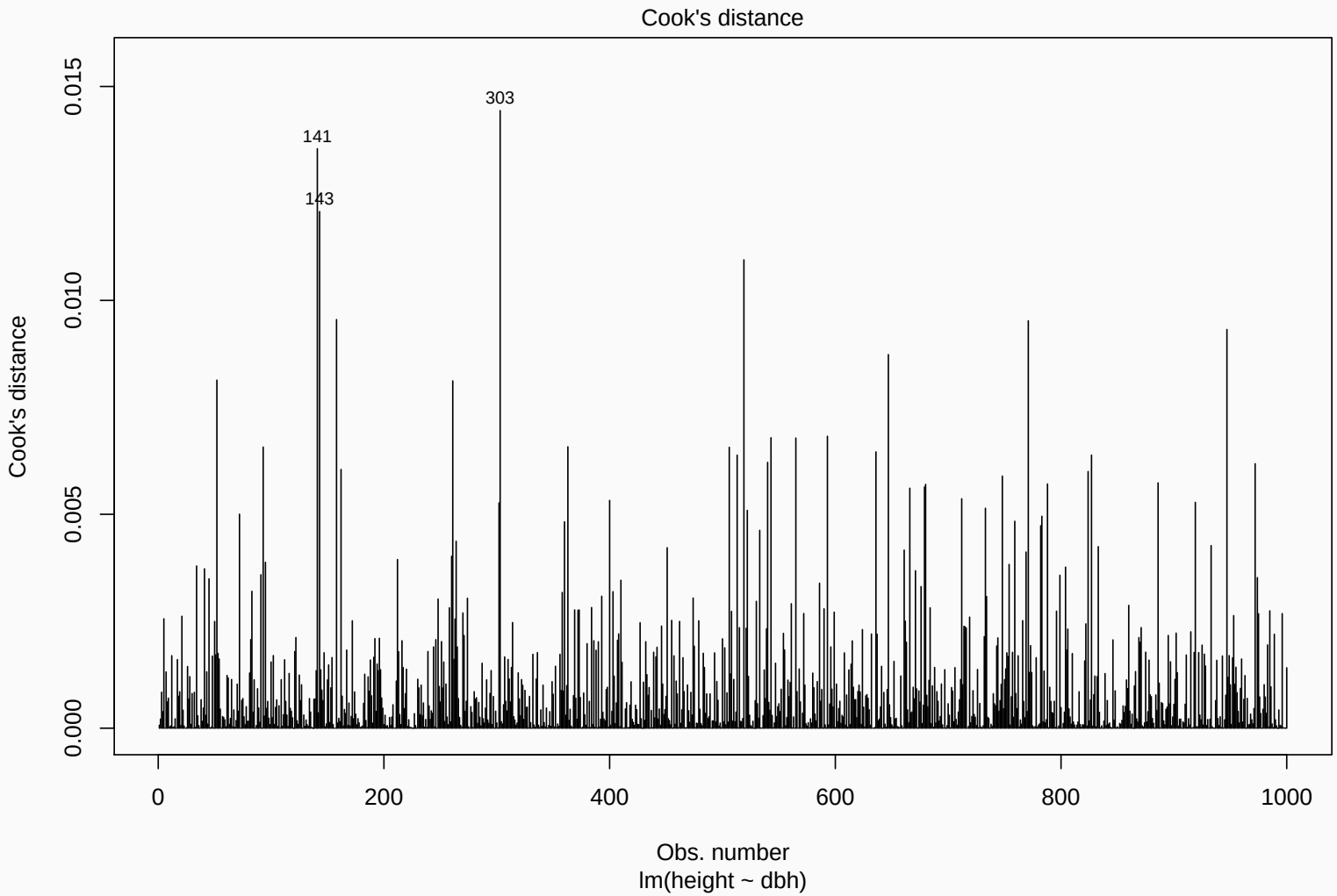
Regresión estadística

```
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)
```

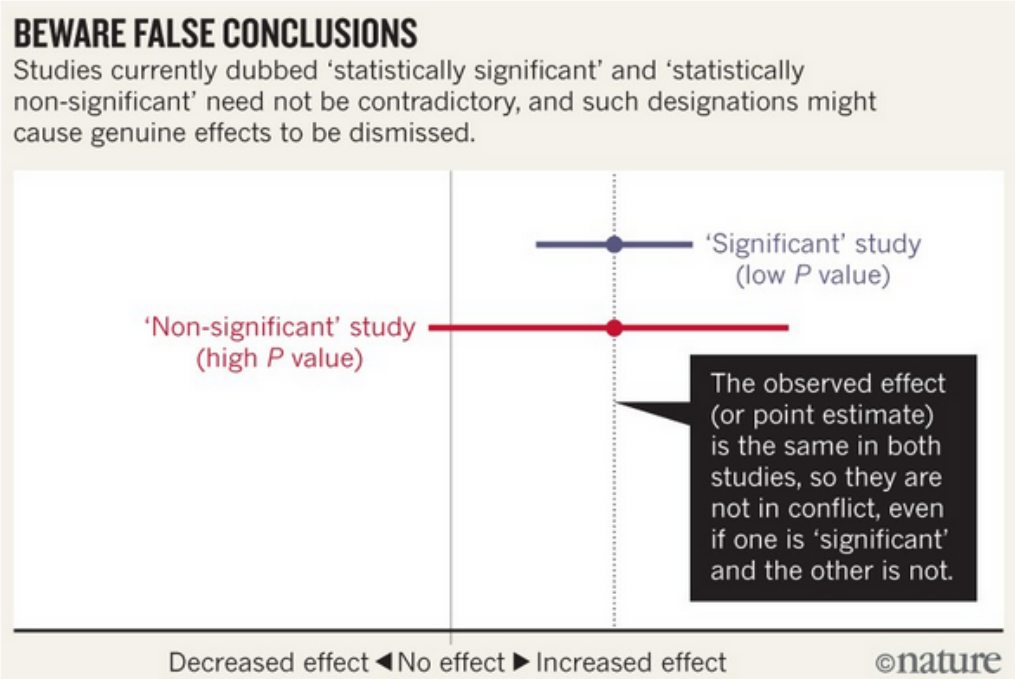
```
##                  2.5 %    97.5 %
## (Intercept) 18.7296053 19.948788
## dbh          0.5958282  0.635579
```

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
```

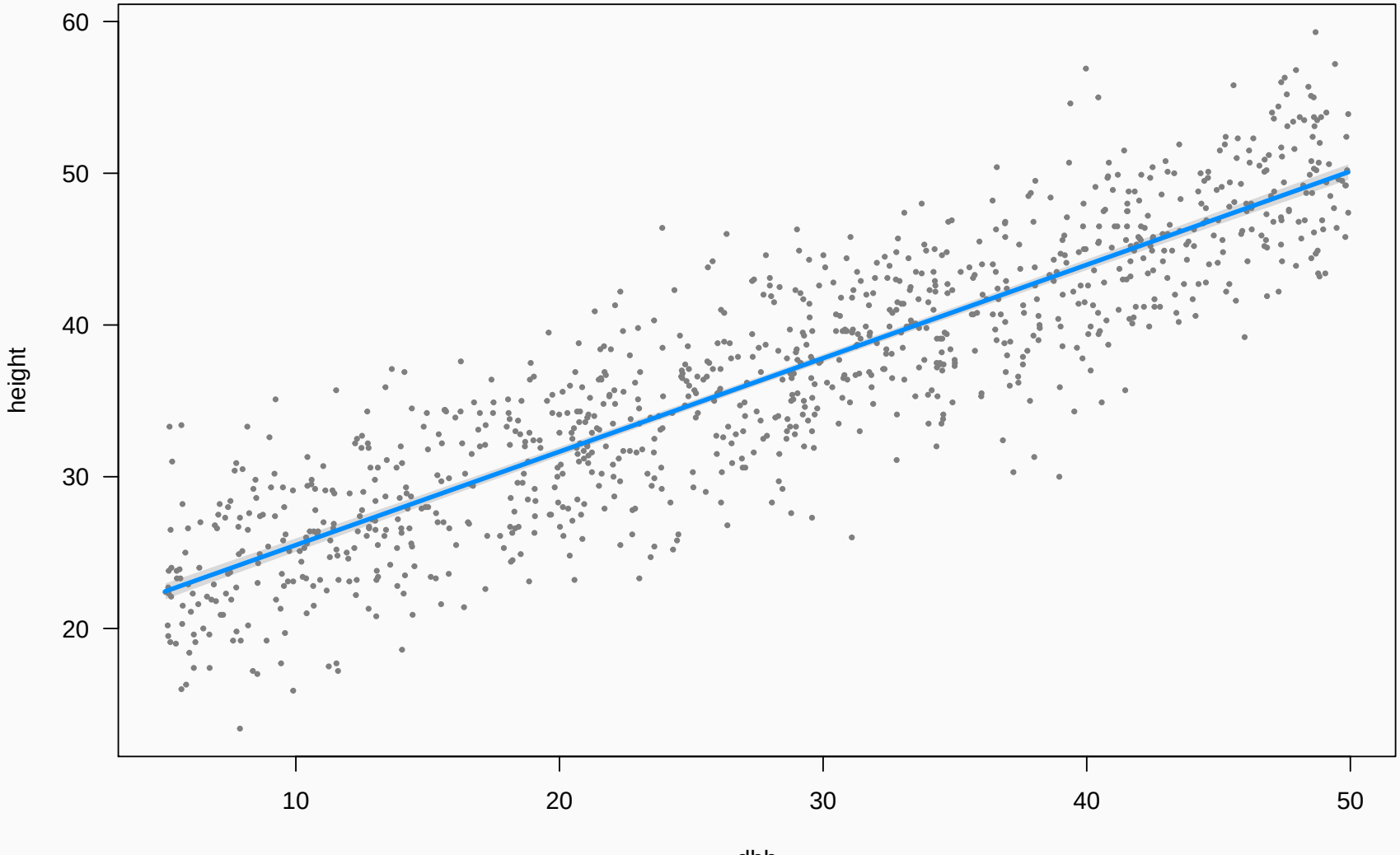
La variación en altura entre los árboles de nuestra muestra viene explicada en un 79% por su variación en DBH. Existe un 21% de variación en altura que responde potencialmente a otros factores, sean estocásticos, errores muestrales, o ecológicos.

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

Regresión estadística

Visualización del modelo:

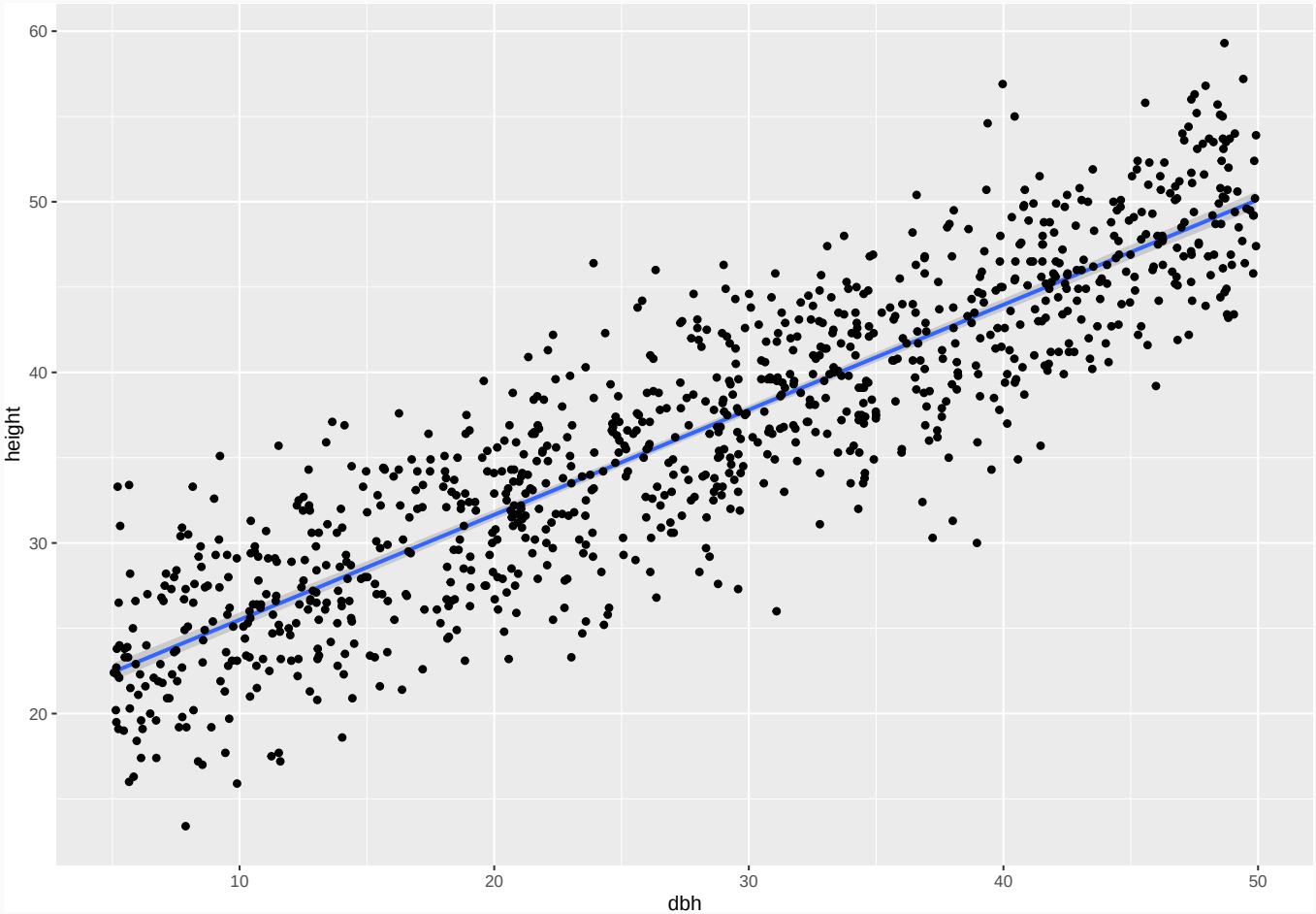
```
library(visreg)
visreg(m1)
```



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

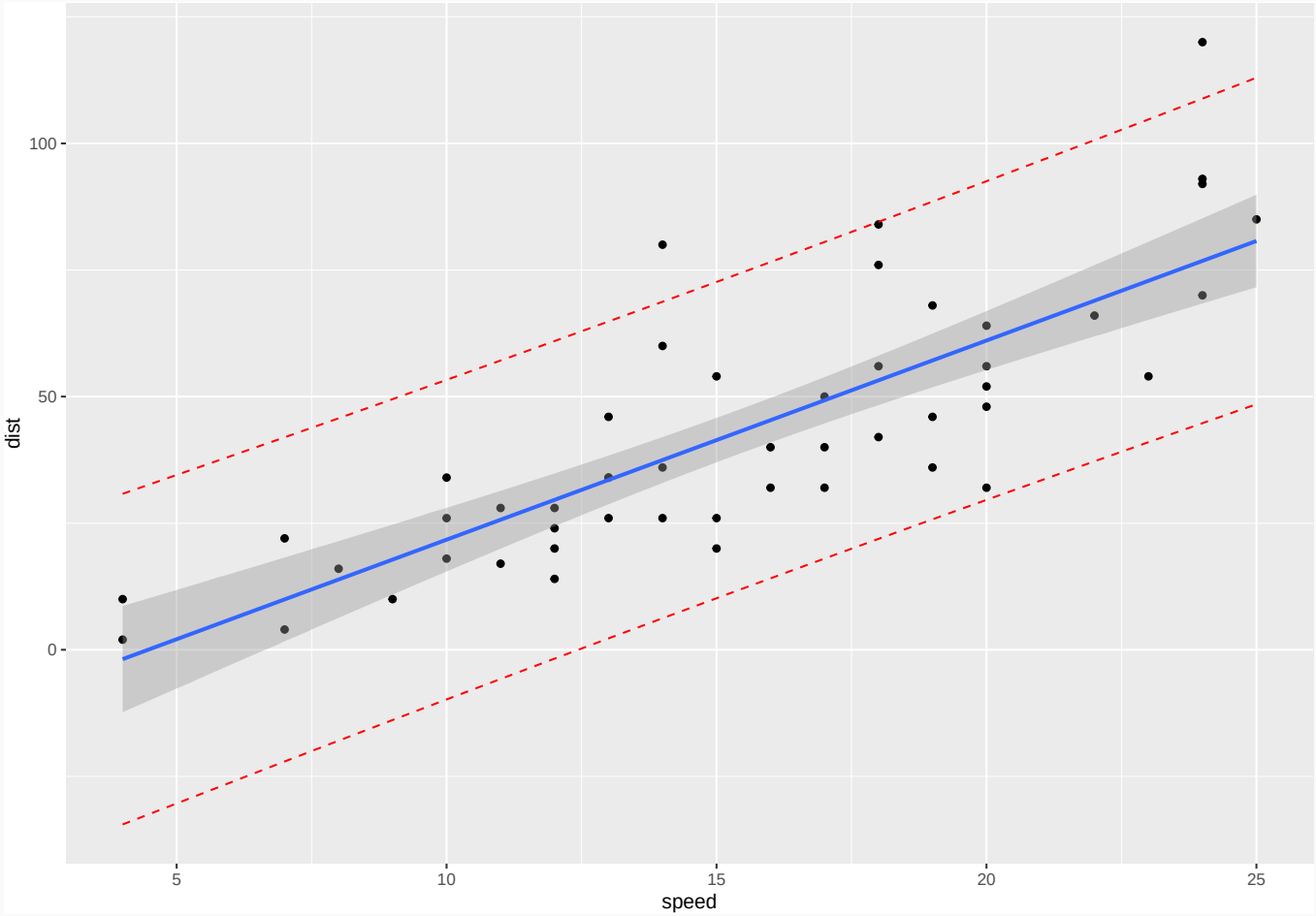
Aunque el valor predicho será el mismo en ambos casos, la incertidumbre asociada a las predicciones es diferente. La primera predicción tiene asociado un intervalo de confianza, la segunda predicción tiene asociado un intervalo de predicción.

Regresión estadística

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

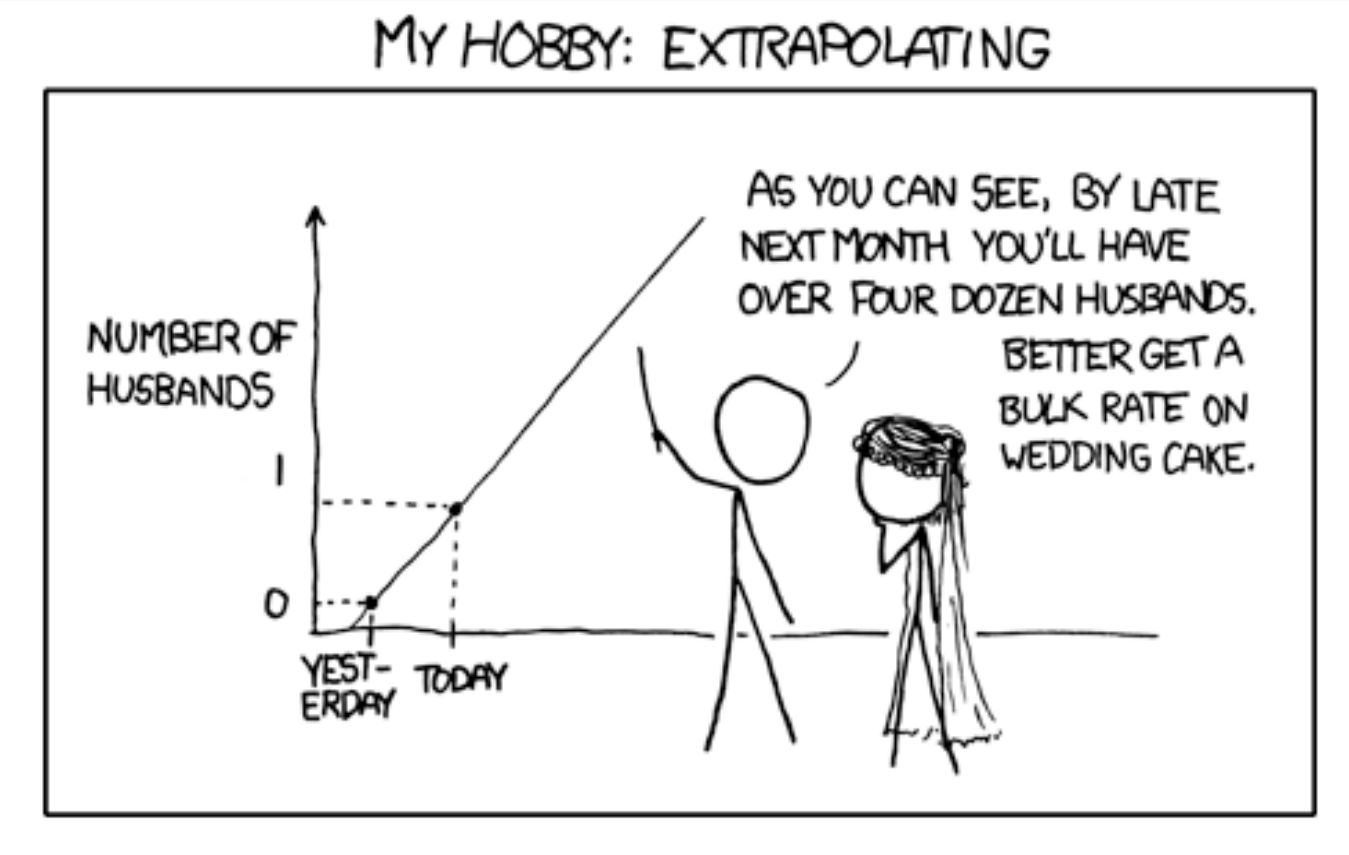
```
# 0. datos y modelo
data("cars", package = "datasets")
model <- lm(dist ~ speed, data = cars)
# predicciones
pred.int <- predict(model, interval = "prediction")
mydata <- cbind(cars, pred.int)
# visualizar recta de regresión e intervalos
library("ggplot2")
p <- ggplot(mydata, aes(speed, dist)) +
  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

Los modelos estadísticos permiten predecir datos que están fuera del rango de las observaciones... cuidado con ello...



Recordatorio

- En general, esperamos que la *variable respuesta* tenga una distribución normal. Pero esto no siempre es necesario: es mucho más importante la distribución de los residuos, que como hemos visto nos ayudan a calcular el error asociado al modelo y los intervalos de confianza.

Regresión estadística

Ejemplo con datos simulados

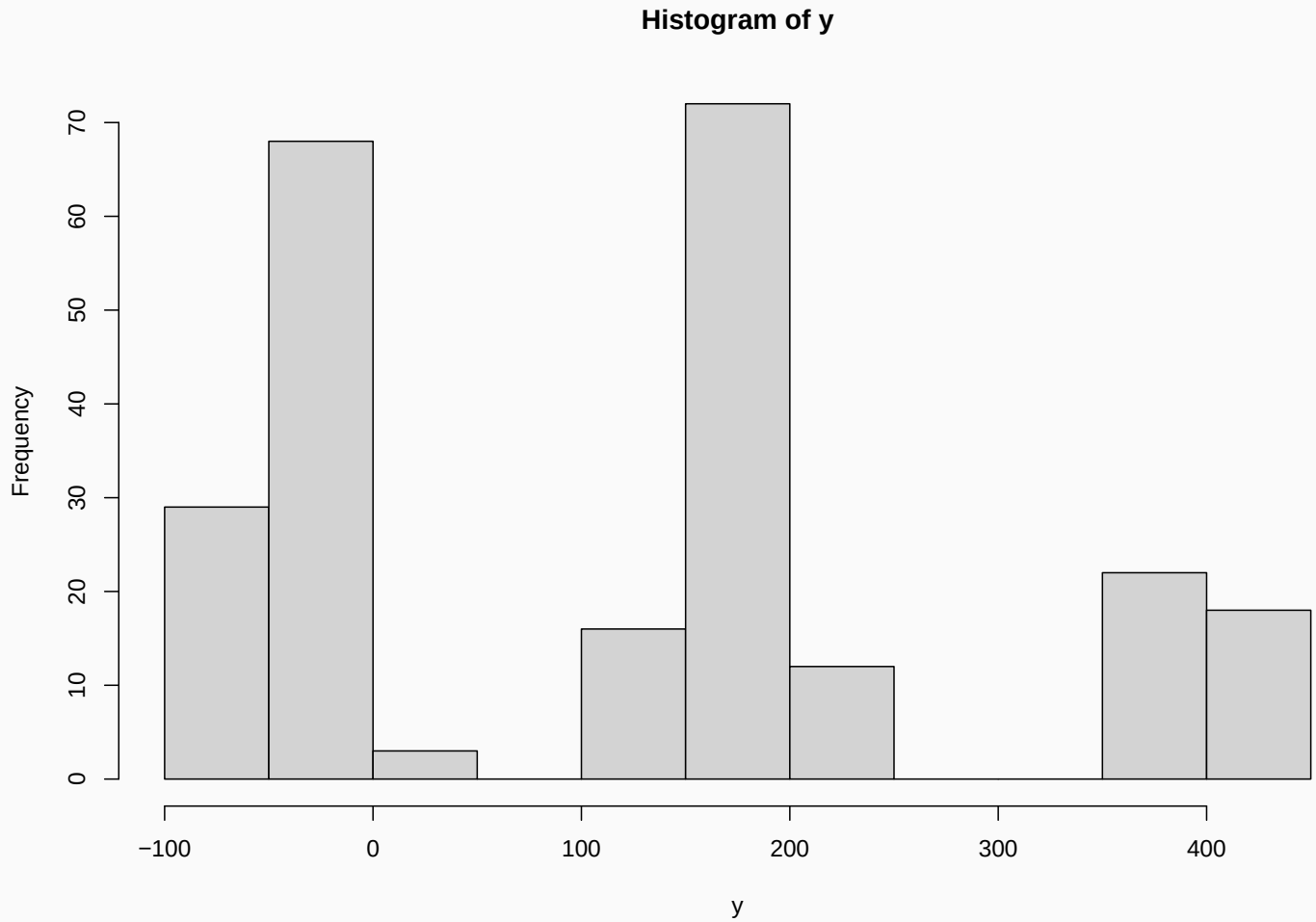
```
Ns <- c(100,100,40)
ms <- c(-10,44,100)

x <- pmap(list(Ns,ms), function(n,m){
  rnorm(n, mean = m, sd = 5)}) %>%
  flatten() %>%
  as_vector()

beta <- 4
err <- 10
y <- rnorm(sum(Ns), beta * x, sd = err)
df <- data.frame(x = x, y = y)
```

Ejemplo con datos simulados

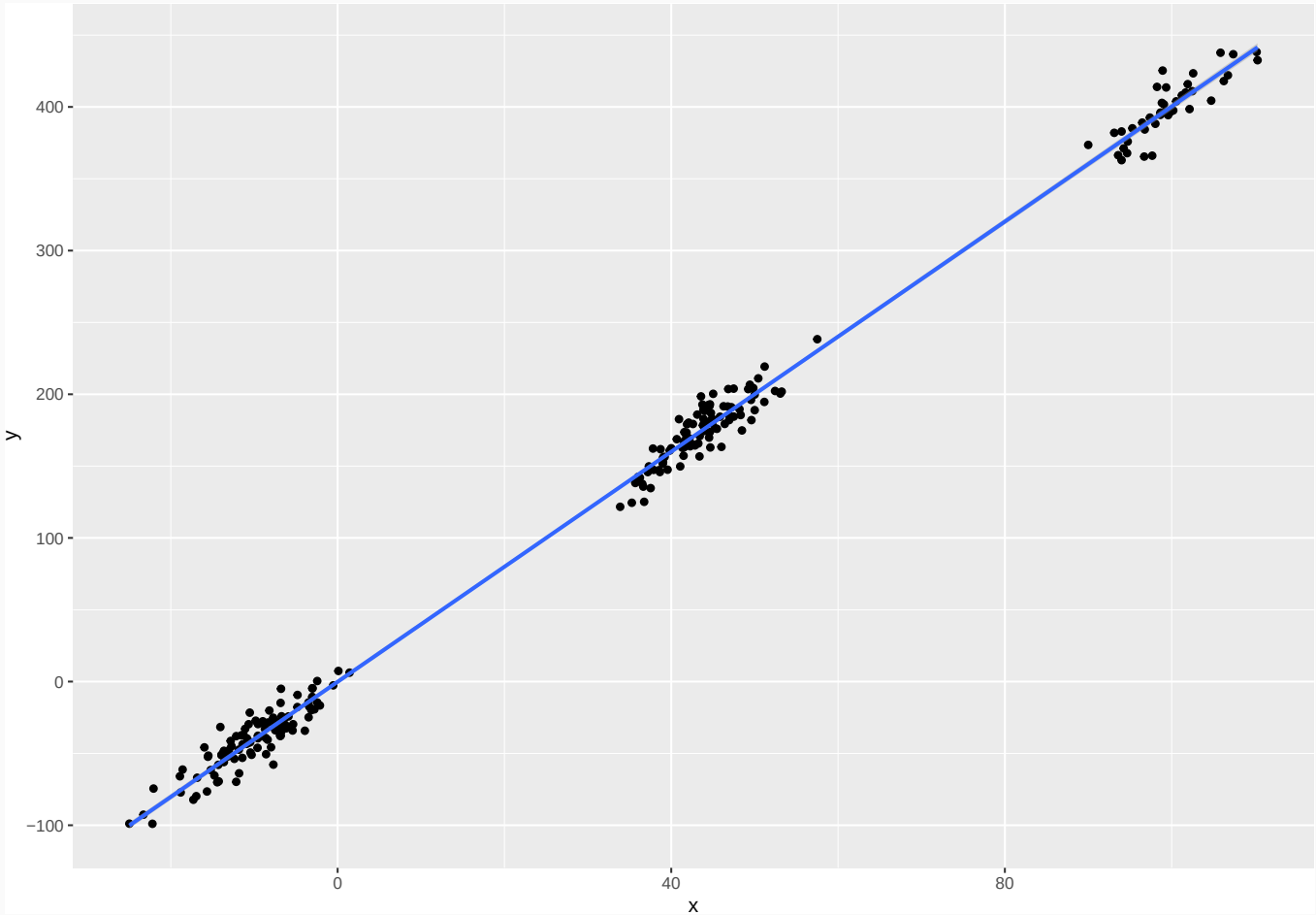
```
hist(y)
```



Regresión estadística

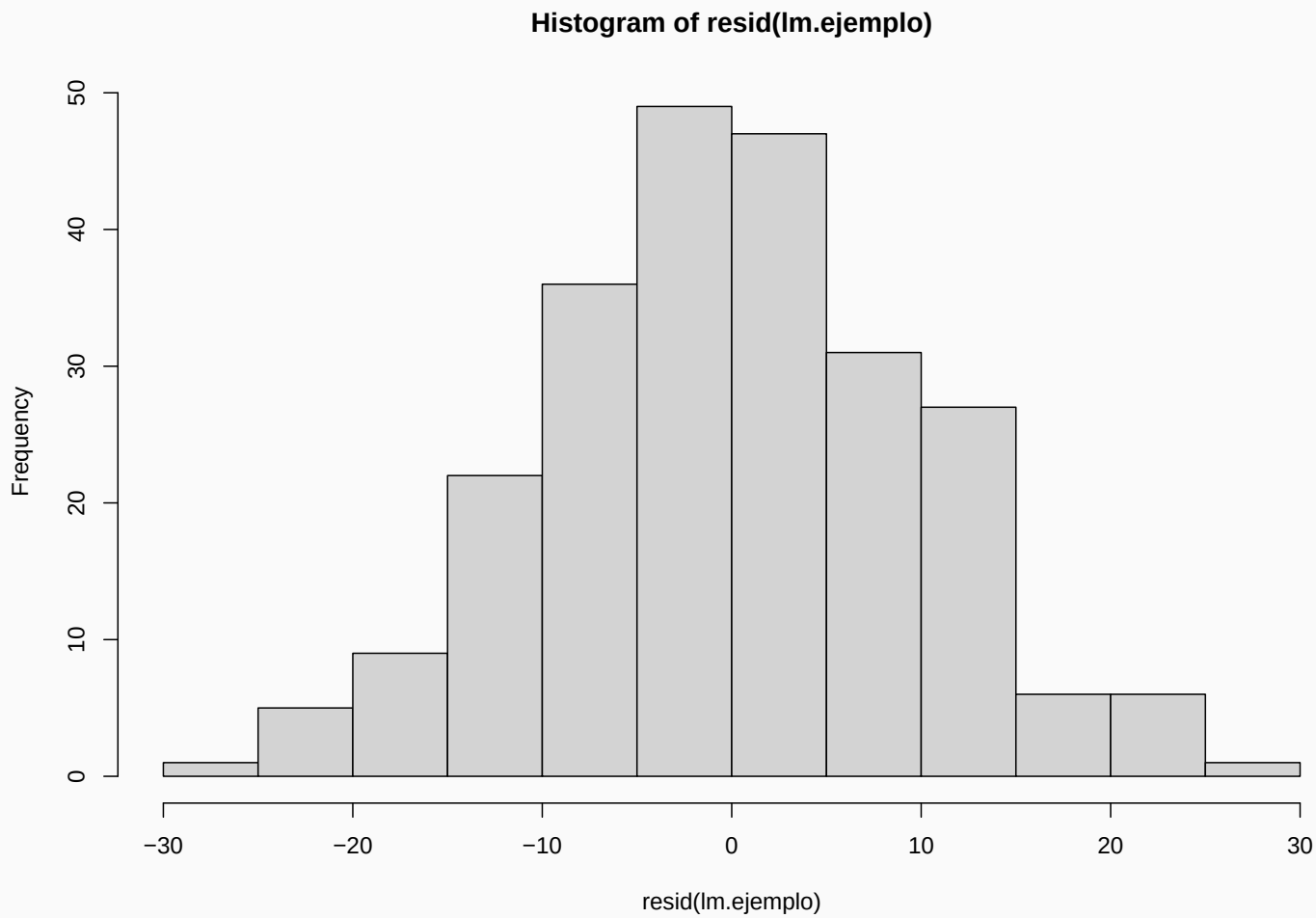
Ejemplo con datos simulados

```
ggplot(df, aes(x = x, y = y)) +
  geom_point() +
  geom_smooth(method = "lm")
```



Regresión estadística

```
lm.ejemplo <- lm(y~x, data = df)
hist(resid(lm.ejemplo))
```



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:

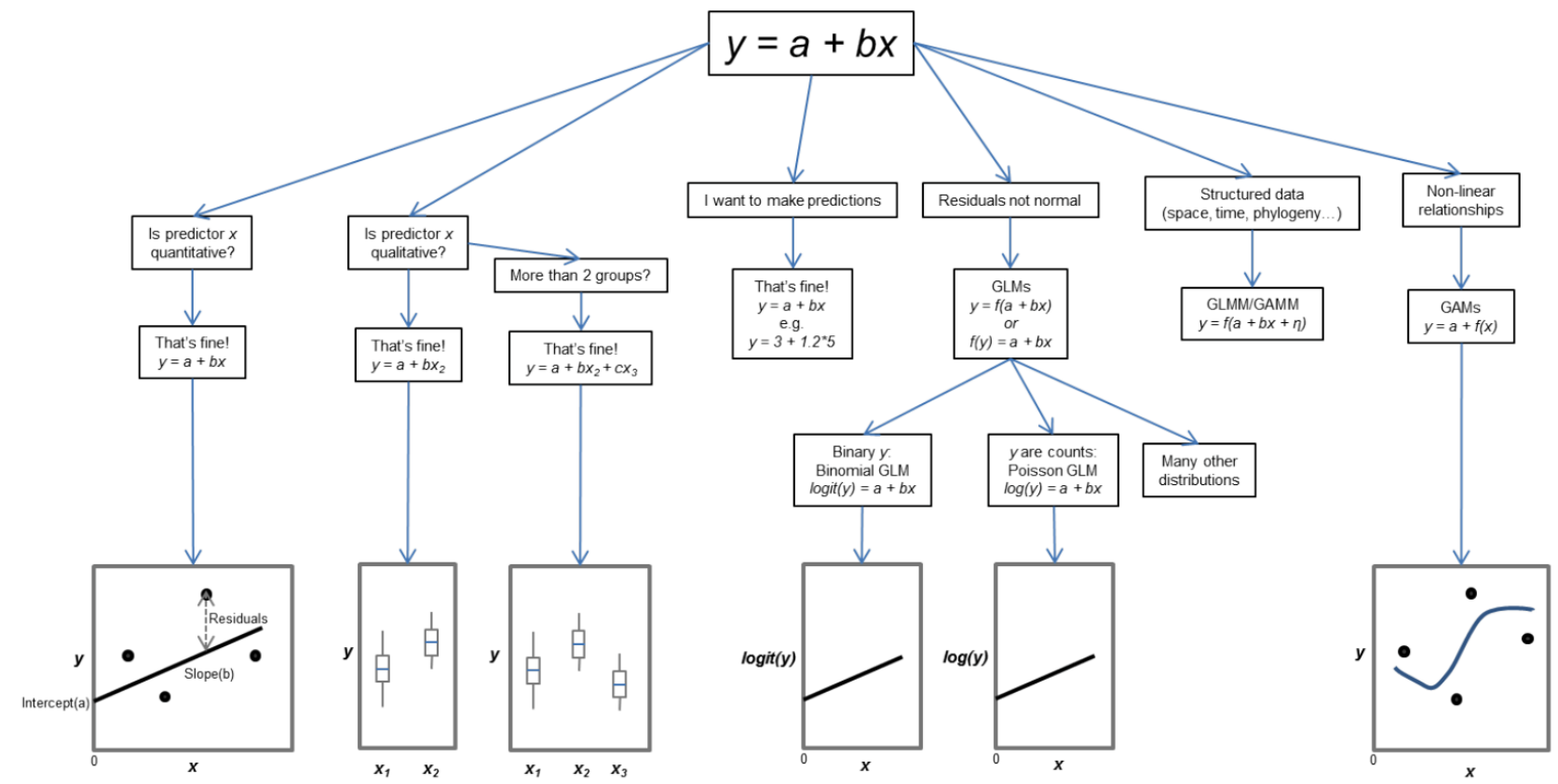
- variable independiente categórica
- múltiples variables independientes

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?

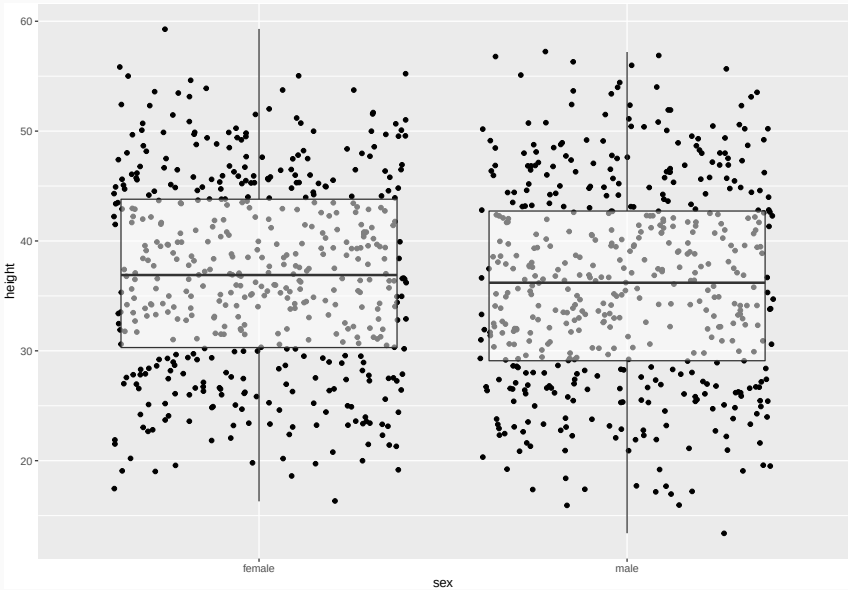
```
head(trees)
```

```
##   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
```

Regresión estadística

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

```
ggplot(trees, aes(x = sex, y = height)) +
  geom_point(position = position_jitter()) +
  geom_boxplot(alpha = 0.5)
```



Regresión estadística

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

```
m2 <- lm(height ~ sex, data = trees)
```

que se corresponde con

$$height_i = a + b_{male} + \varepsilon_i$$
$$\varepsilon_i \sim N(0, \sigma^2)$$

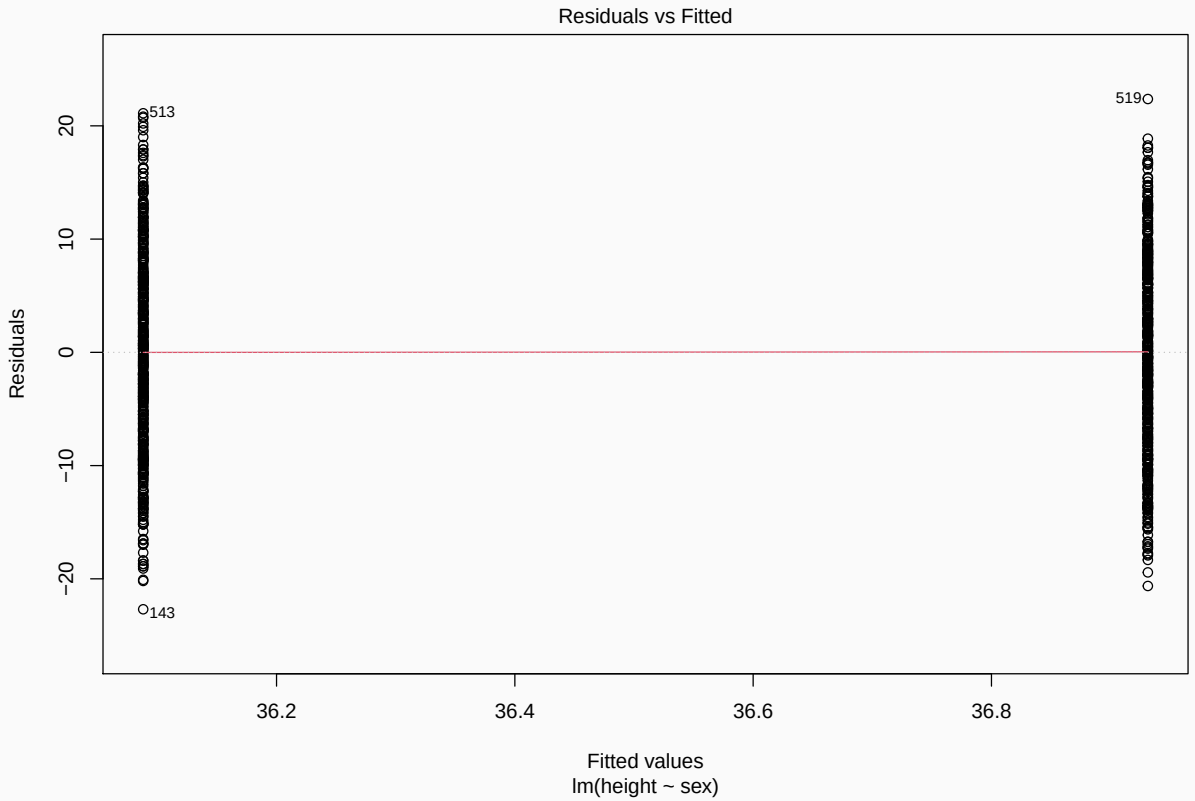
Regresión estadística

```
summary(m2)
```

```
##
## Call:
## lm(formula = height ~ sex, data = trees)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -22.6881  -6.7881  -0.0097   6.7261  22.3687
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  36.9312     0.3981  92.778   <2e-16 ***
## sexmale      -0.8432     0.5607  -1.504    0.133
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.865 on 998 degrees of freedom
## Multiple R-squared:  0.002261,   Adjusted R-squared:  0.001261
```

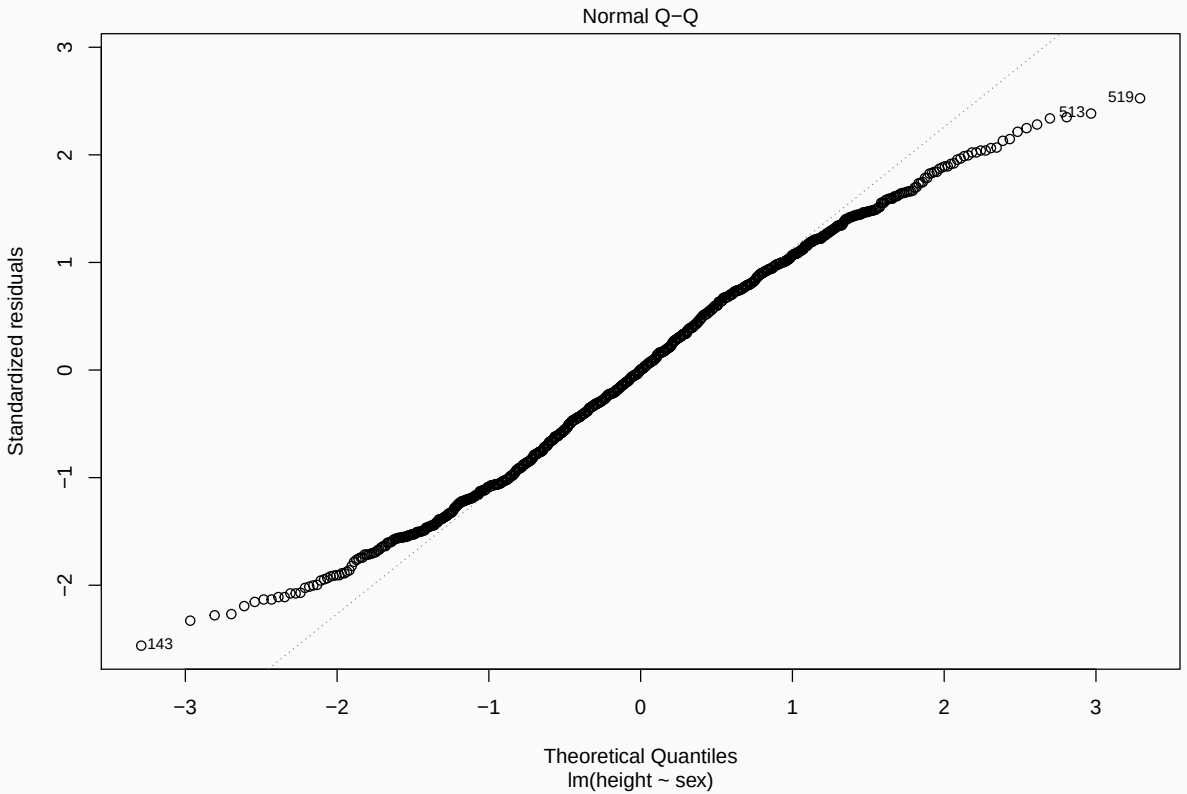
Regresión estadística

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



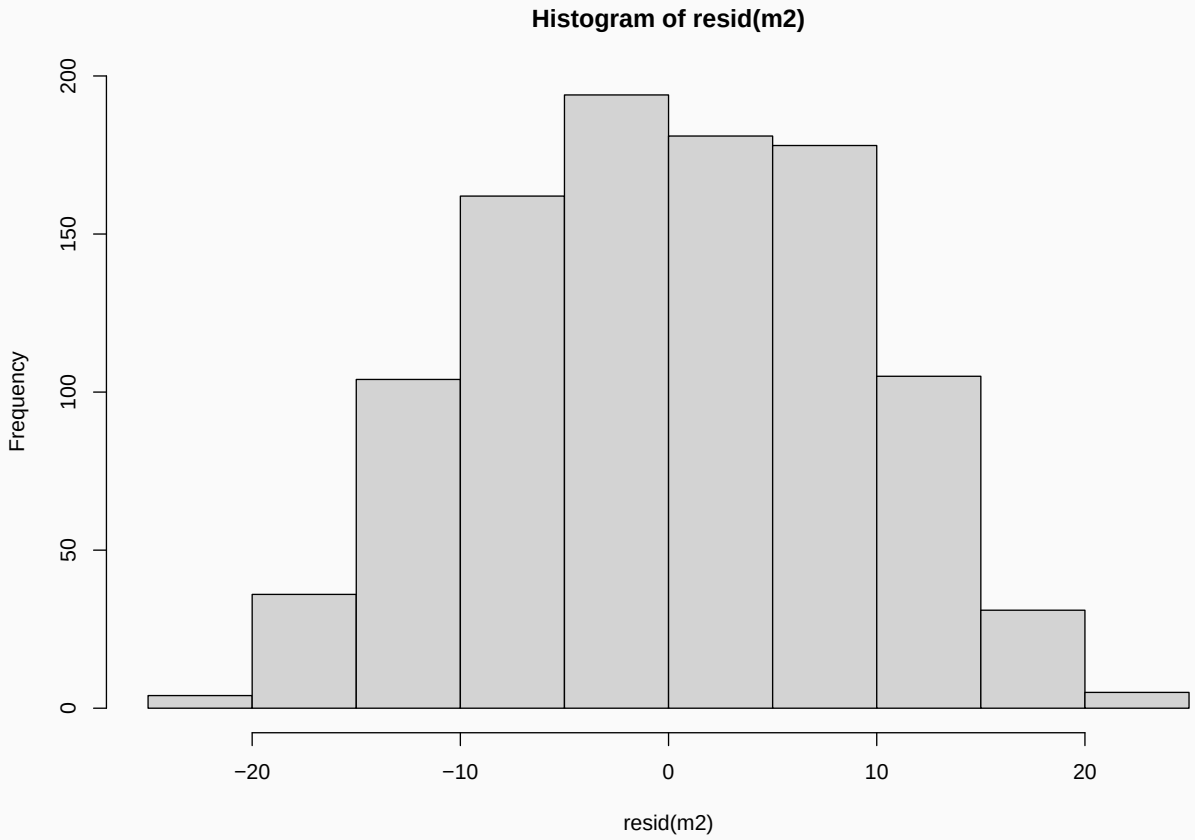
Regresión estadística

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



Regresión estadística

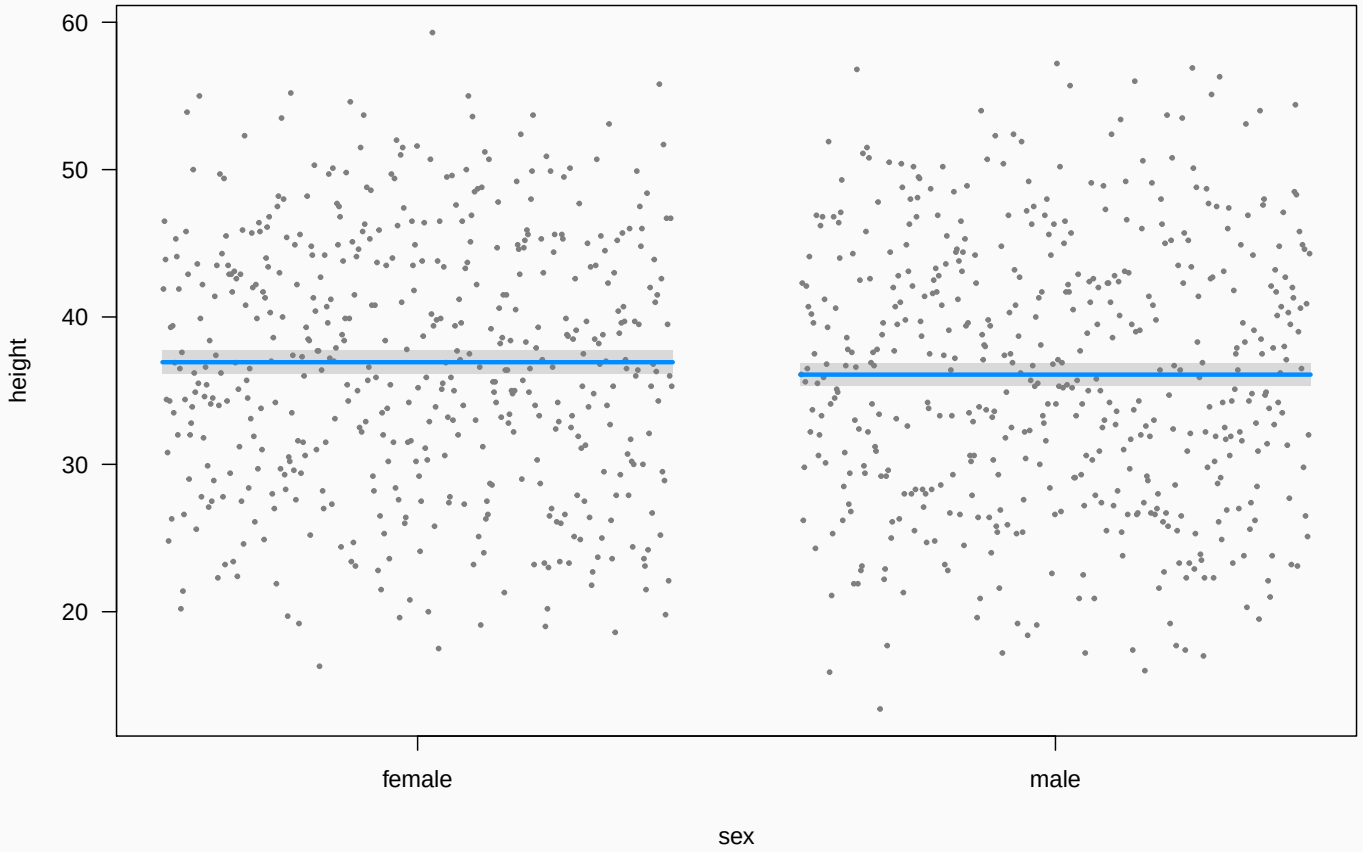
- Un predictor categórico: ¿Varía la altura de los árboles en función del sexo?
  - Comprobar residuos



Regresión estadística

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

```
visreg(m2)
```



Regresión estadística

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

```
confint(m2)
```

```
##                 2.5 %     97.5 %
## (Intercept) 36.150120 37.7123803
## sexmale     -1.943447  0.2571379
```

```
summary(m2)$adj.r.squared
```

```
## [1] 0.001260919
```

Regresión estadística

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

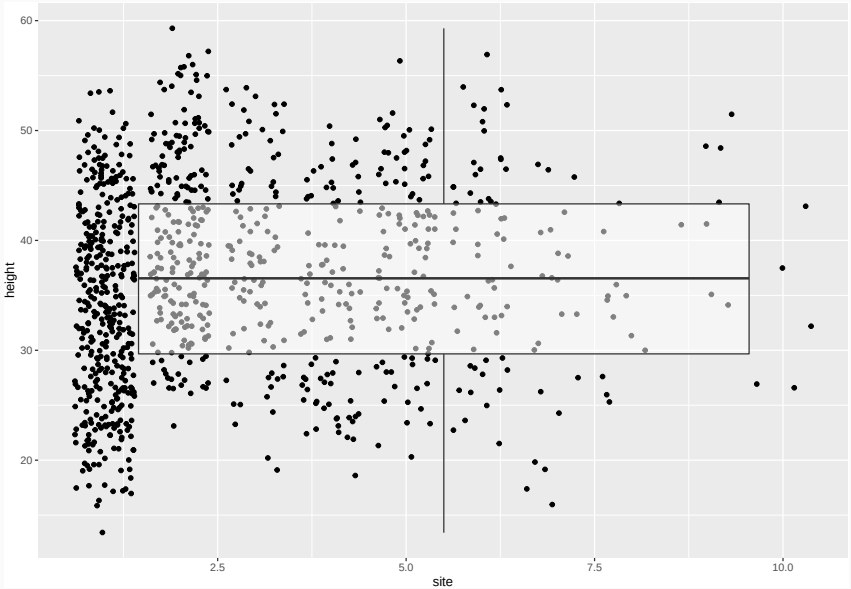
```
head(trees)
```

```
##   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
```

Regresión estadística

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

```
ggplot(trees, aes(x = site, y = height)) +
  geom_point(position = position_jitter()) +
  geom_boxplot(alpha = 0.5)
```



- ¿Qué ocurre?

Regresión estadística

```
str(trees)
```

```
## 'data.frame':    1000 obs. of  5 variables:
##  $ site  : int  4 5 2 5 1 1 2 2 2 1 ...
##  $ dbh   : num  29.7 33.3 28 39.9 47.9 ...
##  $ height: num  36.1 42.3 41.9 46.5 43.9 26.2 29.8 35.6 42.1 36.5 ...
##  $ sex   : chr  "male" "male" "female" "female" ...
##  $ dead  : int  0 0 0 0 0 0 0 0 0 0 ...
```

- "Site" es una variable numérica... debería ser categórica!

Regresión estadística

```
str(trees)
```

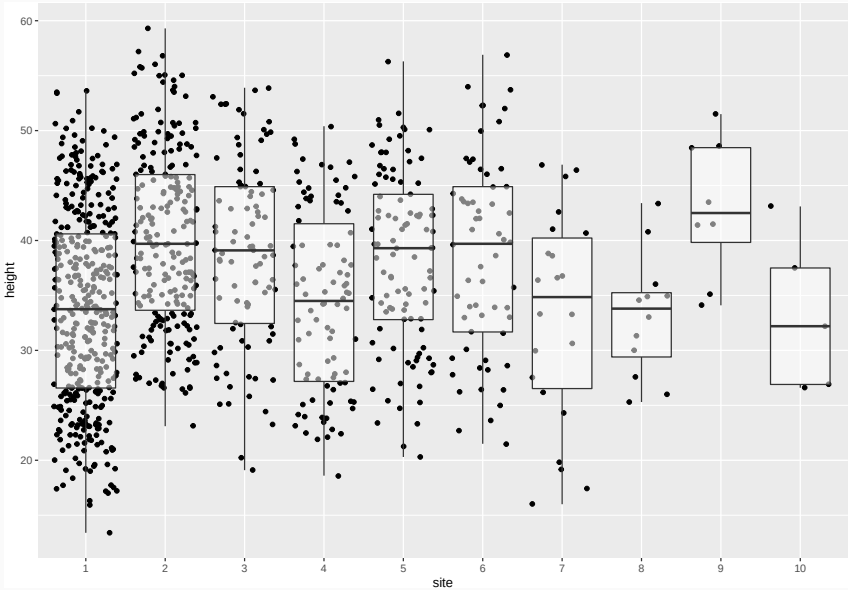
```
## 'data.frame':    1000 obs. of  5 variables:
##  $ site  : int  4 5 2 5 1 1 2 2 2 1 ...
##  $ dbh   : num  29.7 33.3 28 39.9 47.9 ...
##  $ height: num  36.1 42.3 41.9 46.5 43.9 26.2 29.8 35.6 42.1 36.5 ..
##  $ sex   : chr  "male" "male" "female" "female" ...
##  $ dead  : int  0 0 0 0 0 0 0 0 0 0 ...
```

- "Site" es una variable numérica... debería ser categórica!

```
trees$site <- as.factor(trees$site)
```

Regresión estadística

```
ggplot(trees, aes(x = site, y = height)) +
  geom_point(position = position_jitter()) +
  geom_boxplot(alpha = 0.5)
```



Regresión estadística

- Un predictor categórico: ¿Varía la altura de los árboles en función del lugar de muestreo?
  - Ajustar modelo

```
m3 <- lm(height ~ site, data = trees)
```

que se corresponde con

$$height_i = a + b_{site2} + c_{site3} + d_{site4} + ... + \varepsilon_i$$
$$\varepsilon_i \sim N(0, \sigma^2)$$

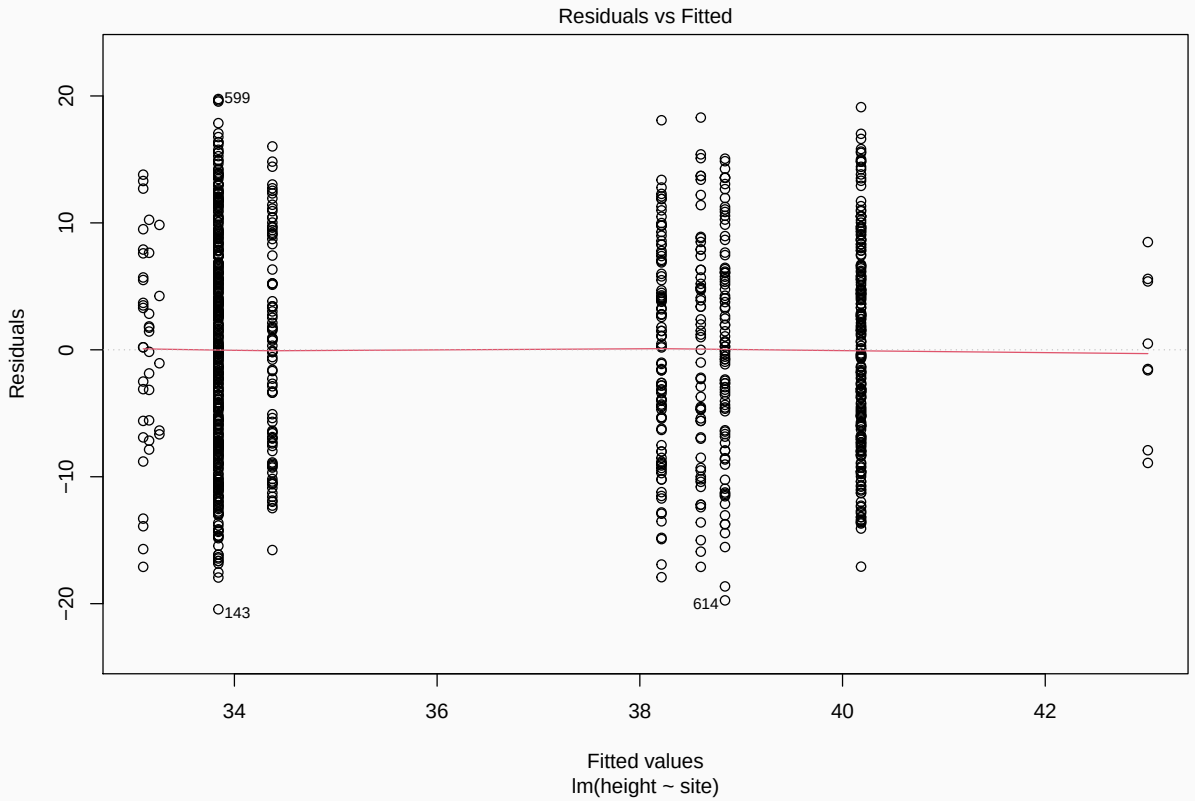
Regresión estadística

```
summary(m3)
```

```
##
## Call:
## lm(formula = height ~ site, data = trees)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -20.4416  -6.9004   0.0379   6.3051  19.7584
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  33.8416     0.4266  79.329  < 2e-16 ***
## site2         6.3411     0.7126   8.899  < 2e-16 ***
## site3         4.9991     0.9828   5.086 4.36e-07 ***
## site4         0.5329     0.9872   0.540  0.58949
## site5         4.3723     0.9425   4.639 3.97e-06 ***
## site6         4.7601     1.1709   4.065 5.18e-05 ***
## site7        -0.7416     1.8506  -0.401  0.68871
## site8        -0.6832     2.4753  -0.276  0.78258
## site9         9.1709     3.0165   3.040  0.00243 **
## site10       -0.5816     3.8013  -0.153  0.87843
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.446 on 990 degrees of freedom
## Multiple R-squared:  0.1016, Adjusted R-squared:  0.09344
## F-statistic: 12.44 on 9 and 990 DF,  p-value: < 2.2e-16
```

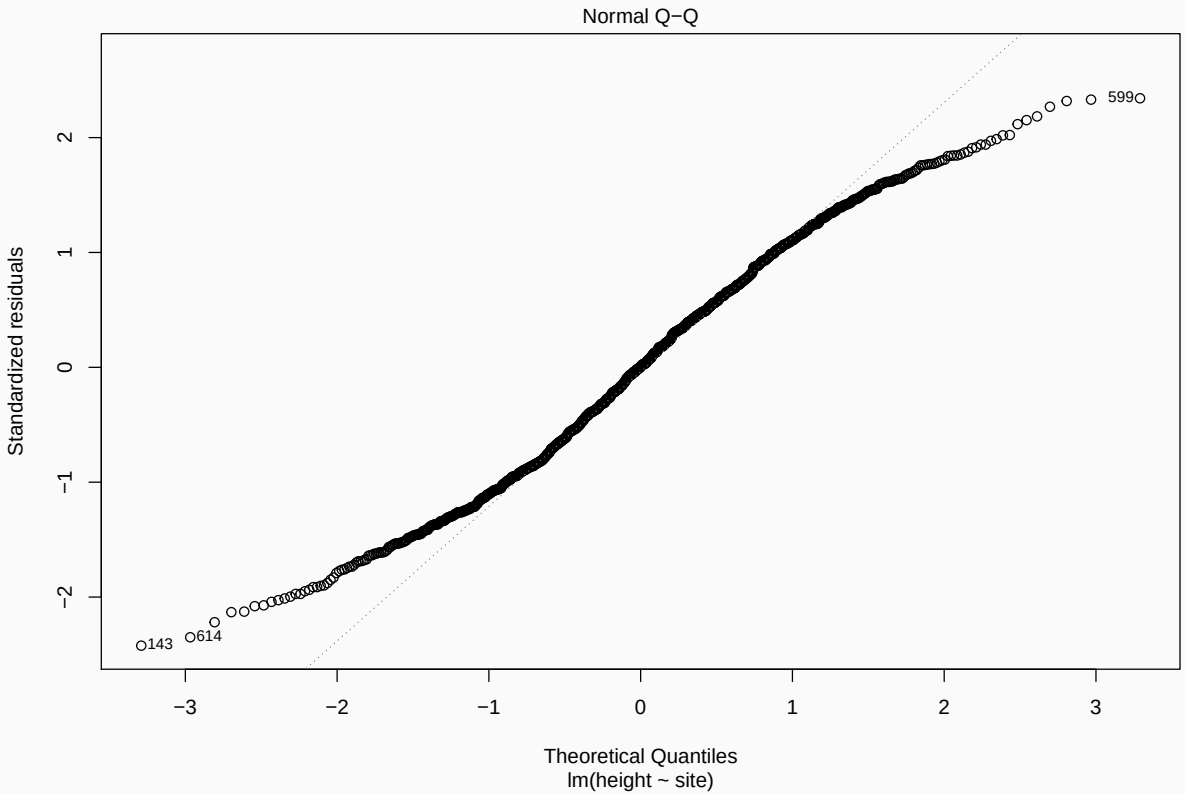
Regresión estadística

- Un predictor categórico: ¿Varía la altura de los árboles en función del lugar de muestreo?
  - Comprobar residuos



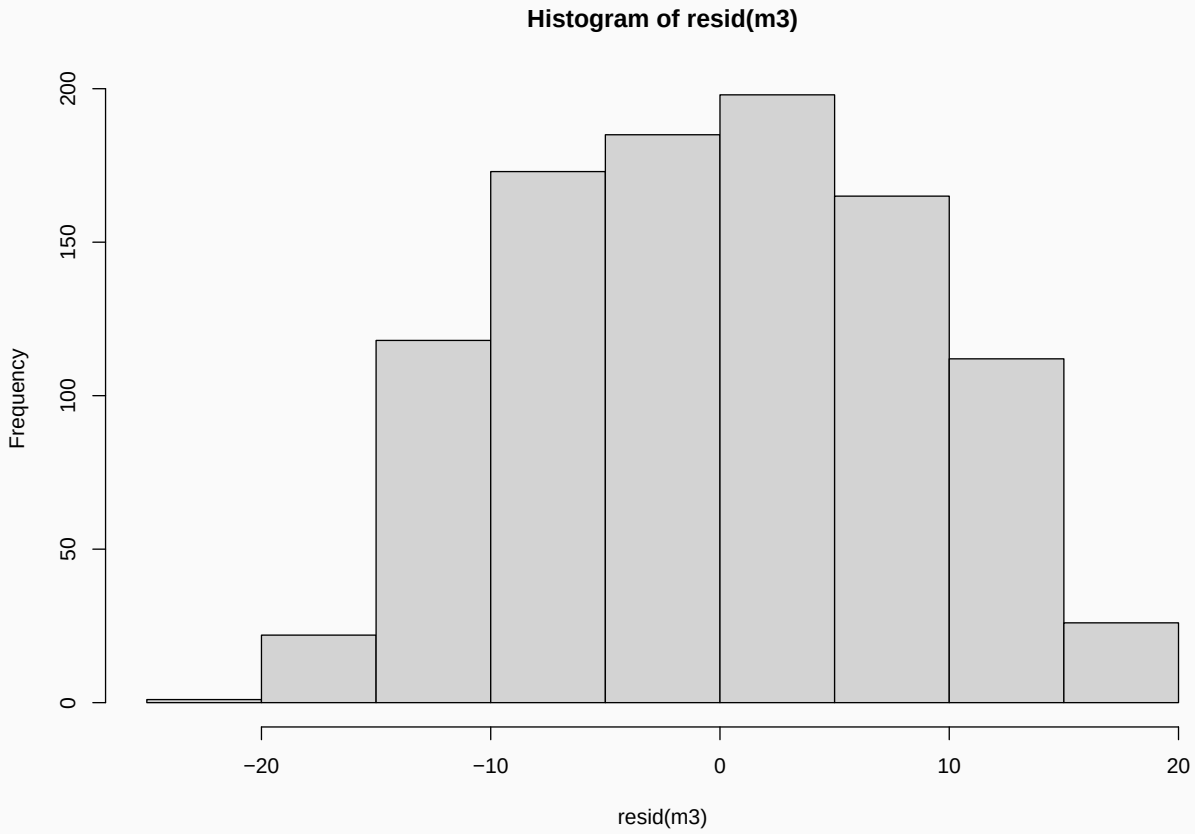
Regresión estadística

- Un predictor categórico: ¿Varía la altura de los árboles en función del lugar de muestreo?
  - Comprobar residuos



Regresión estadística

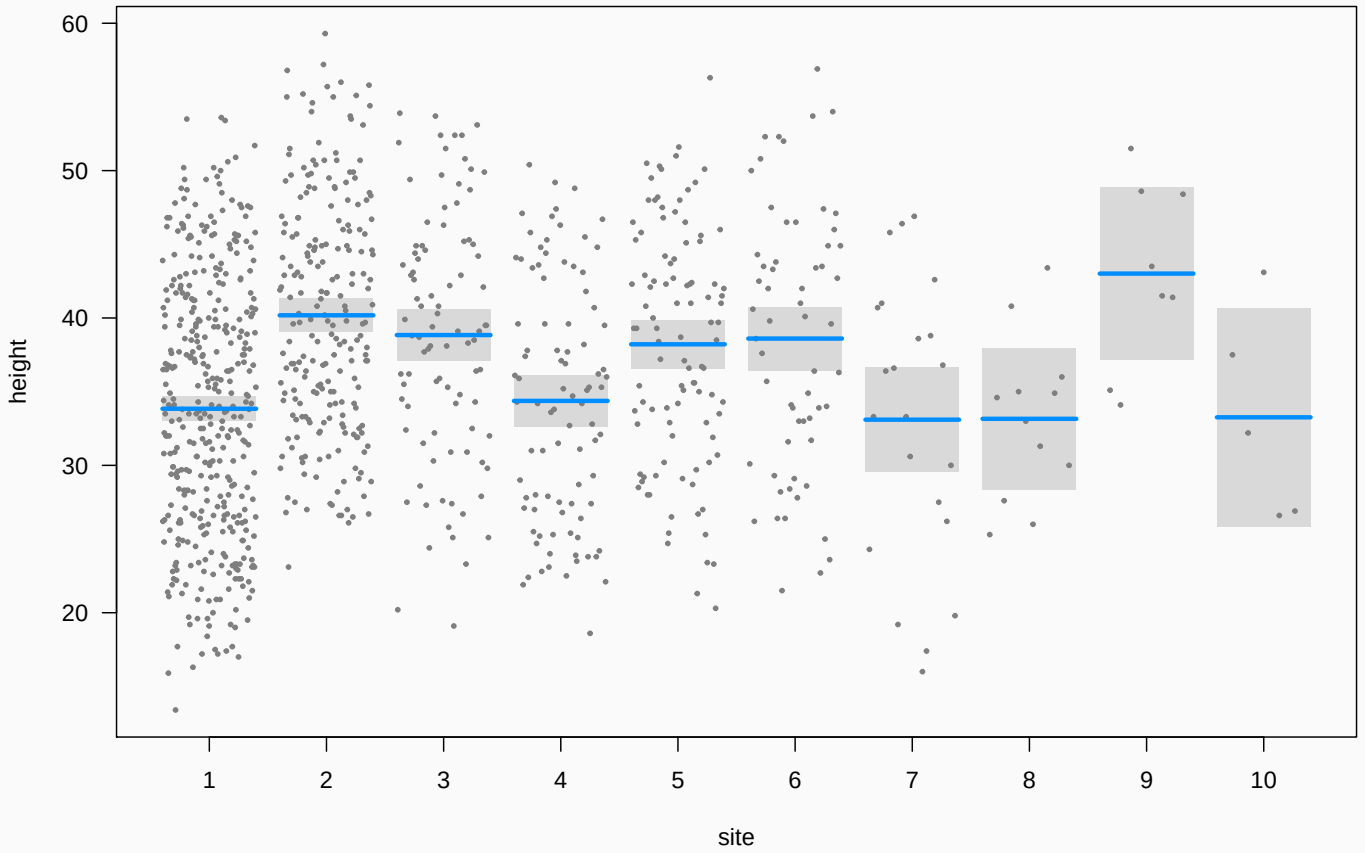
- Un predictor categórico: ¿Varía la altura de los árboles en función del lugar de muestreo?
  - Comprobar residuos



Regresión estadística

- Un predictor categórico: ¿Varía la altura de los árboles en función del lugar de muestreo?
  - Visualizar modelo

```
visreg(m3)
```



Regresión estadística

- Un predictor categórico: ¿Varía la altura de los árboles en función del lugar de muestreo?
    - Visualizar modelo

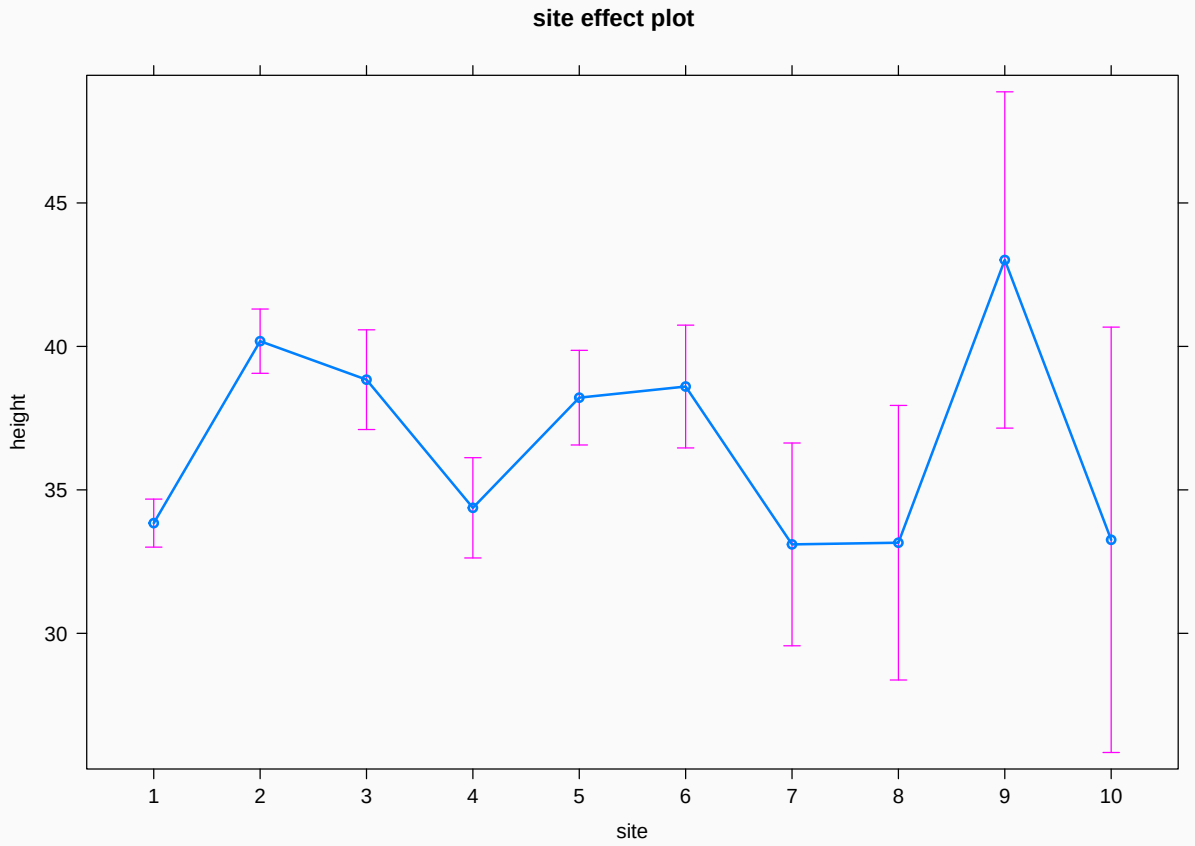
```
library(effects)
allEffects(m3)
```

```
##  model: height ~ site
##
##  site effect
## site
##        1        2        3        4        5        6        7        8
## 33.84158 40.18265 38.84066 34.37444 38.21386 38.60167 33.10000 33.15833
##        9       10
## 43.01250 33.26000
```

Regresión estadística

- Un predictor categórico: ¿Varía la altura de los árboles en función del lugar de muestreo?
  - Visualizar modelo

```
library(effects)
plot(allEffects(m3))
```



Regresión estadística

- Un predictor categórico: ¿Varía la altura de los árboles en función del lugar de muestreo?
    - Interpretar resultados

```
confint(m3)
```

```
##                 2.5 %    97.5 %
## (Intercept) 33.004441 34.678723
## site2        4.942777  7.739357
## site3        3.070436  6.927719
## site4       -1.404455  2.470181
## site5        2.522750  6.221810
## site6        2.462391  7.057779
## site7       -4.373093  2.889930
## site8       -5.540590  4.174094
## site9        3.251437 15.090399
## site10      -8.041067  6.877904
```

```
summary(m3)$adj.r.squared
```

```
## [1] 0.09343655
```

Regresión estadística

Recordemos las asunciones de los modelos lineales:

- variable respuesta: normal
- distribución de residuos: normal
- varianza residual: homogénea
- observaciones independientes entre sí

En este caso, las observaciones entre lugares de muestreo claramente *no son* independientes entre sí (en el modelo anterior hemos observado un efecto del lugar de muestreo sobre la altura de los árboles). Por otro lado, ya sabemos que la DBH es una variable importante, así que podemos pensar en un modelo que incluya ambos factores.

Regresión estadística

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

```
m4 <- lm(height ~ site + dbh, data = trees)
```

que se corresponde con

$$height_i = a + b_{site2} + c_{site3} + d_{site4} + ... + k \cdot DBH_i + \varepsilon_i$$
$$\varepsilon_i \sim N(0, \sigma^2)$$

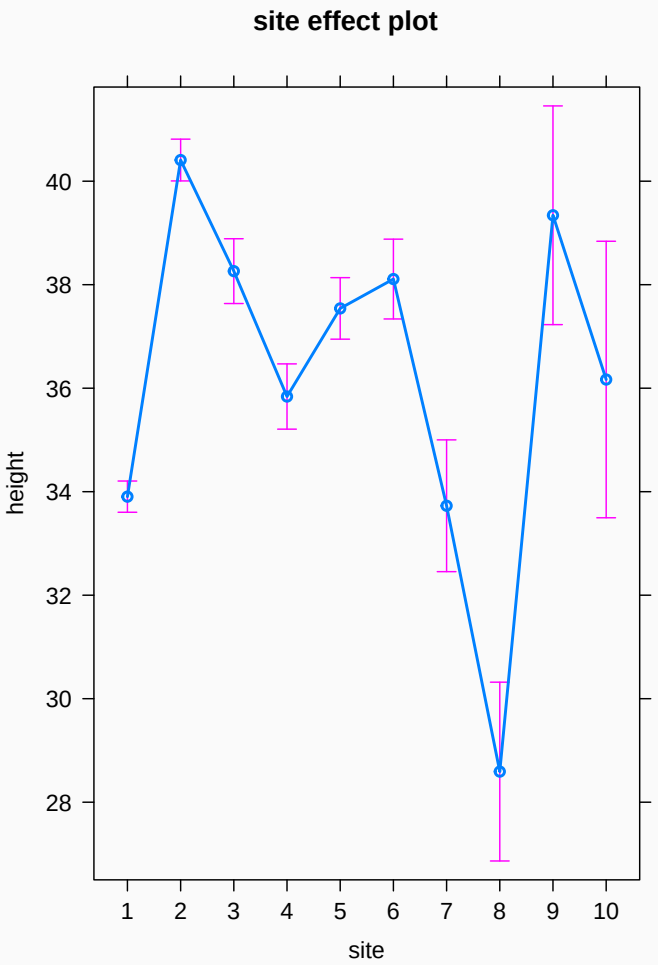
Regresión estadística

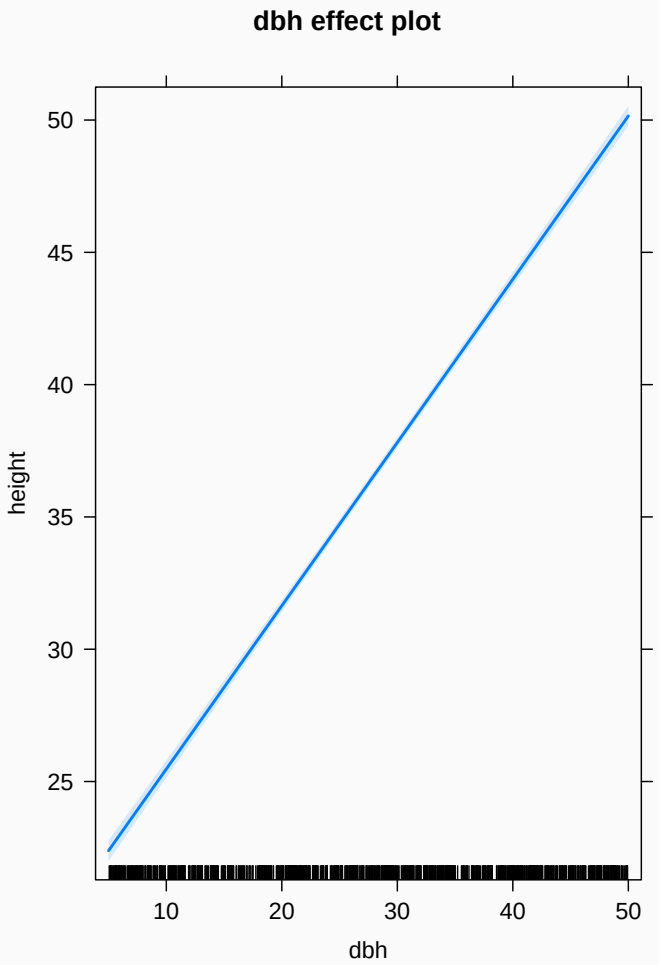
```
summary(m4)
```

```
##
## Call:
## lm(formula = height ~ site + dbh, data = trees)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10.1130  -1.9885   0.0582   2.0314  11.3320
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 16.699037   0.260565  64.088  < 2e-16 ***
## site2        6.504303   0.256730  25.335  < 2e-16 ***
## site3        4.357457   0.354181  12.303  < 2e-16 ***
## site4        1.934650   0.356102   5.433 6.98e-08 ***
## site5        3.637432   0.339688  10.708  < 2e-16 ***
## site6        4.204511   0.421906   9.966  < 2e-16 ***
## site7       -0.176193   0.666772  -0.264   0.7916
## site8       -5.312648   0.893603  -5.945 3.82e-09 ***
## site9        5.437049   1.087766   4.998 6.84e-07 ***
## site10       2.263338   1.369986   1.652   0.0988 .
## dbh          0.617075   0.007574  81.473  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.043 on 989 degrees of freedom
## Multiple R-squared:  0.8835, Adjusted R-squared:  0.8823
## F-statistic:   750 on 10 and 989 DF,  p-value: < 2.2e-16
```

Regresión estadística

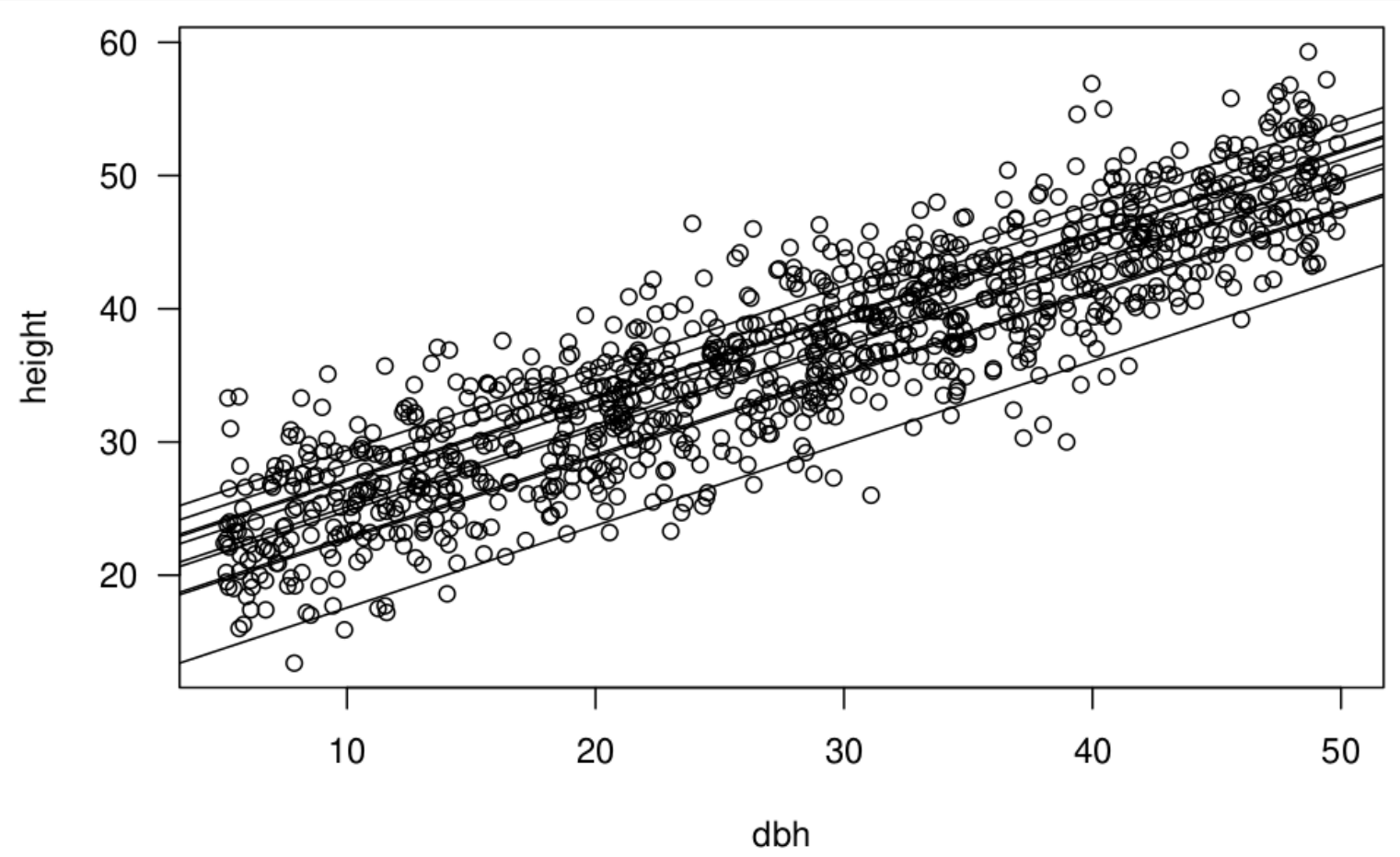
```
plot(allEffects(m4))
```





Regresión estadística

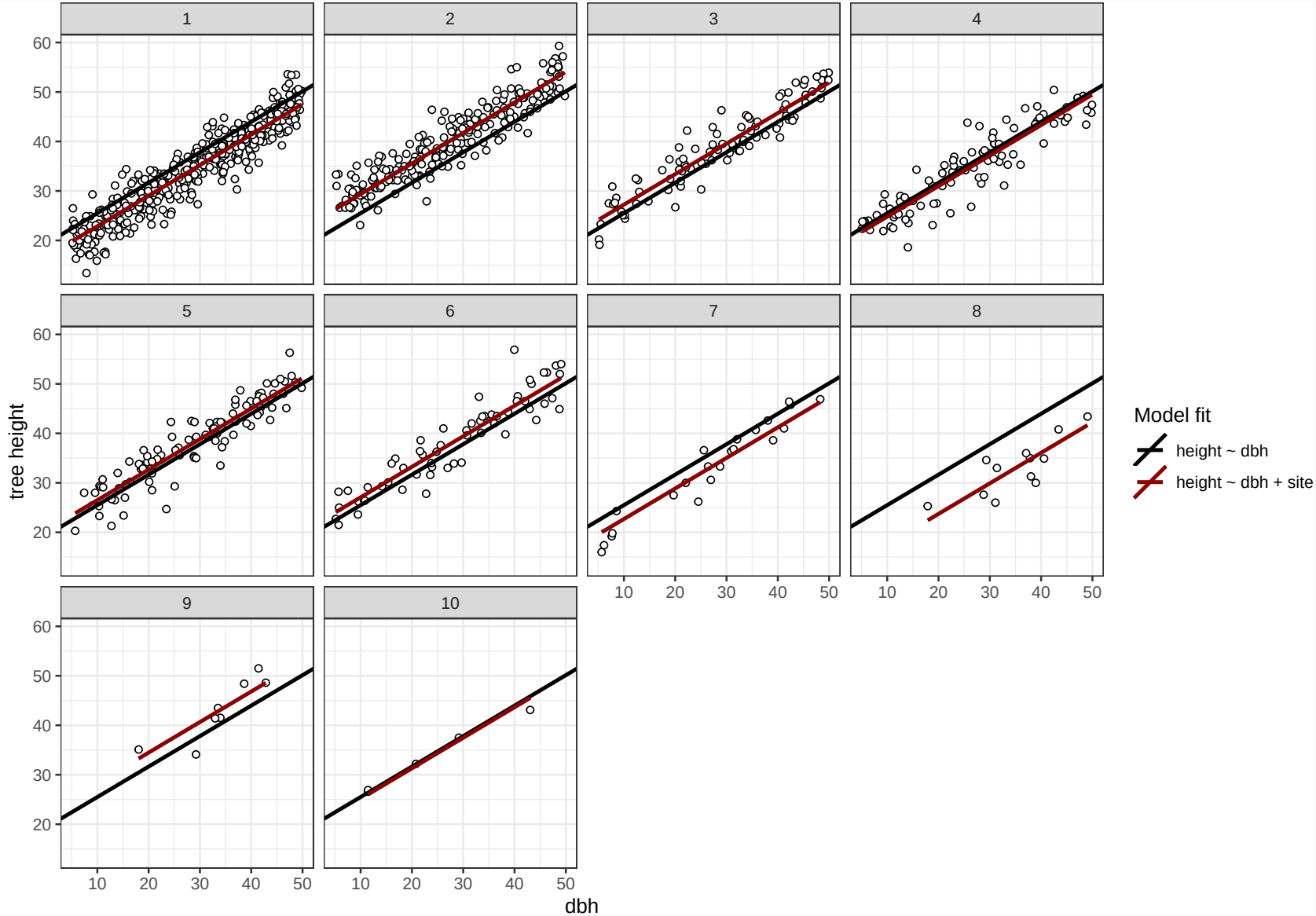
- Hemos ajustado un modelo con diferentes puntos de corte (uno por cada *site*) y una sola pendiente (DBH)



Regresión estadística

```
trees$fitted.m4 <- fitted(m4)
ggplot(trees ,aes(x=dbh,y=height))+
  geom_point(show.legend = FALSE,shape = 21, fill = "white") +
  facet_wrap(~site)+
  geom_abline(aes(intercept = coef(m1)[[1]],
                  slope = coef(m1)[[2]],
                  color = 'black'),size = 1) +
  geom_line(aes(x=dbh,y=fitted.m4, color = "darkred"),size=1)+
  scale_color_identity(labels=c("height ~ dbh",
                                "height ~ dbh + site"),
                       guide="legend")+
  labs(x = "dbh",y = "tree height",color="Model fit") +
  theme_bw()
```

Regresión estadística



Regresión estadística

- Hemos ajustado un modelo con diferentes puntos de corte (uno por cada *site*) y una sola pendiente (DBH)
- ¿varía la relación altura-DBH (= la pendiente) entre diferentes lugares de muestreo?

Regresión estadística

- Interacciones entre predictores: ¿Varía la relación altura-DBH en función del lugar de muestreo?

```
m5 <- lm(height ~ site * dbh, data = trees)
```

que se corresponde con

$$height_i = a + b_{site2} + c_{site3} + d_{site4} + ...+ \tag{3}$$
$$k \cdot DBH + l \cdot DBH_{i,site2} + m \cdot DBH_{i,site3} + ... + \varepsilon_i$$
$$\varepsilon_i \sim N(0, \sigma^2)$$

Regresión estadística

```
summary(m5)
```

```
##
## Call:
## lm(formula = height ~ site * dbh, data = trees)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10.1017  -1.9839   0.0645   2.0486  11.1789
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 16.359437   0.360054  45.436  < 2e-16 ***
## site2        7.684781   0.609657  12.605  < 2e-16 ***
## site3        4.518568   0.867008   5.212 2.28e-07 ***
## site4        2.769336   0.813259   3.405 0.000688 ***
## site5        3.917607   0.870983   4.498 7.68e-06 ***
## site6        4.155161   1.009379   4.117 4.17e-05 ***
## site7       -2.306799   1.551303  -1.487 0.137334
## site8       -2.616095   4.090671  -0.640 0.522630
## site9        2.621560   5.073794   0.517 0.605492
## site10       4.662340   2.991072   1.559 0.119378
## dbh          0.629299   0.011722  53.685  < 2e-16 ***
## site2:dbh   -0.042784   0.020033  -2.136 0.032950 *
## site3:dbh   -0.006031   0.027640  -0.218 0.827312
## site4:dbh   -0.031633   0.028225  -1.121 0.262677
## site5:dbh   -0.010173   0.027887  -0.365 0.715334
## site6:dbh    0.001337   0.032109   0.042 0.966797
## site7:dbh    0.079728   0.052056   1.532 0.125951
## site8:dbh   -0.079027   0.113386  -0.697 0.485984
## site9:dbh    0.081035   0.146649   0.553 0.580679
## site10:dbh  -0.101107   0.114520  -0.883 0.377522
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

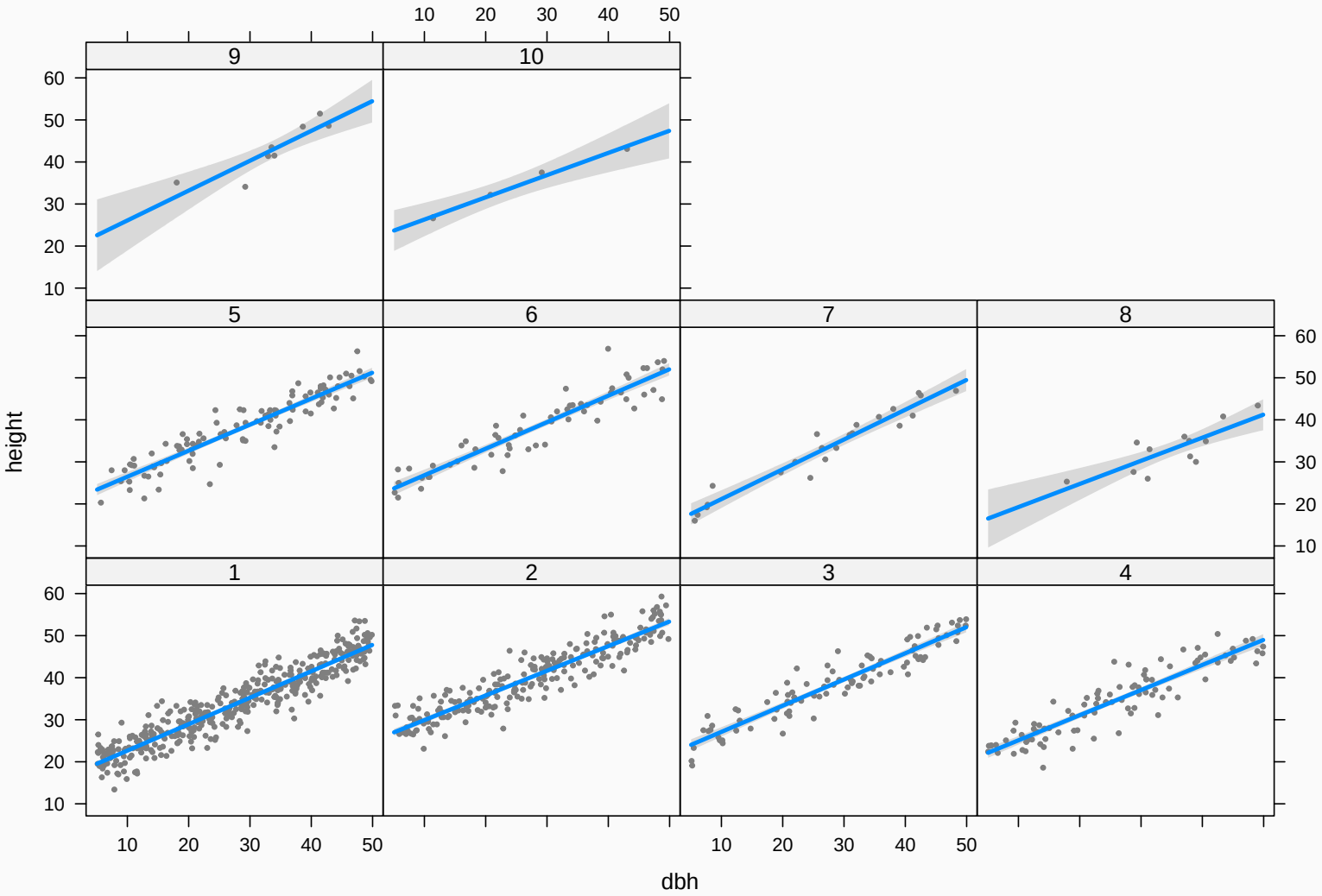
Regresión estadística

Interpretar coeficientes en modelos con interacciones sigue la misma filosofía que hemos visto en modelos más sencillos:

- El punto de corte representa la categoría base: para DBH 0, en en site 1, la altura esperada es 16.3594372 (lo cual tiene poco sentido biológico). Igualmente, el efecto asociado a dbh nos da la pendiente de la recta para el site 1 (una unidad de DBH aumenta la altura en 0.6292993)
- Los efectos asociados a sites (site2,site3,etc) nos dan la diferencia en altura, para DBH $= 0$, con respecto al site 1 (categoría base). Por ejemplo, los árboles de DBH $= 0$ en el site 2 tienen una altura esperada de 16.3594372 + 7.6847807.
- Los efectos compuestos, por ejemplo site2:dbh, nos dicen la diferencia en el efecto de la DBH con respecto al efecto base para cada site. Por ejemplo, el efecto de aumentar una unidad de DBH en el site 2 es de 0.6292993 + -0.0427843.

Regresión estadística

```
visreg(m5, xvar = "dbh", by = "site")
```



Regresión estadística

Otros puntos a tener en cuenta:

- Las variables independientes no deben estar correlacionadas entre sí. Cuando ajustamos modelos con múltiples variables numéricas, es importante comprobar previamente la correlación entre ellas (ver sección *test de hipótesis*) y eliminar las variables correlacionadas.

Regresión estadística

Otros puntos a tener en cuenta:

- Las variables independientes no deben estar correlacionadas entre sí. Cuando ajustamos modelos con múltiples variables numéricas, es importante comprobar previamente la correlación entre ellas (ver sección *test de hipótesis*) y eliminar las variables correlacionadas.
- Un modelo lineal en el que la variable independiente es categórica con varios factores es equivalente a una ANOVA (ver, por ejemplo, aquí)

Regresión estadística

Otros puntos a tener en cuenta:

- Las variables independientes no deben estar correlacionadas entre sí. Cuando ajustamos modelos con múltiples variables numéricas, es importante comprobar previamente la correlación entre ellas (ver sección *test de hipótesis*) y eliminar las variables correlacionadas.
- Un modelo lineal en el que la variable independiente es categórica con varios factores es equivalente a una ANOVA (ver, por ejemplo, aquí)
- A veces puede ser importante centrar y/o estandarizar las variables independientes. Ver, por ejemplo, discusión aquí

Resumen de modelos lineales

- permiten predecir una variable respuesta en función de variables independientes

Resumen de modelos lineales

- permiten predecir una variable respuesta en función de variables independientes
- modelan *relaciones lineales*

Resumen de modelos lineales

- permiten predecir una variable respuesta en función de variables independientes
- modelan *relaciones lineales*
- son la base para muchas de las técnicas estadísticas que trabajaréis

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

Recetario de R

- Ajustar un modelo: `modelo <- lm(respuesta ~ predictores, data = datos)`
- comprobar residuos: `plot(modelo)`, `resid(modelo)`
- visualizar modelo: `visreg(modelo)`, `allEffects(modelo)`
- coeficientes: `coef(modelo)`, `tidy(modelo)`
- intervalos de confianza: `confint(modelo)`
- coeficiente de determinación: `summary(modelo)$adj.r.squared`

Regresión estadística

Otros recursos:

- https://bookdown.org/speegled/foundations-of-statistics/
- https://bookdown.org/egarpor/PM-UC3M/