

# Modelos estadísticos

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

---

David García Callejas

01/2021

Hasta ahora:

- Sabemos cómo cuantificar una muestra o una población

Hasta ahora:

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

Hasta ahora:

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

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:

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?

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?

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?



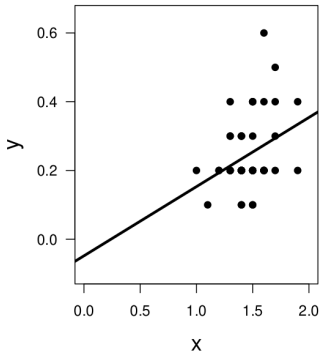
Hasta ahora:

- 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?
- Respuesta:  $y = a + bx$

# Regresión estadística

$$y_i = a + bx_i + \varepsilon_i$$

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



## Data

$y$  = response variable

$x$  = predictor

## Parameters

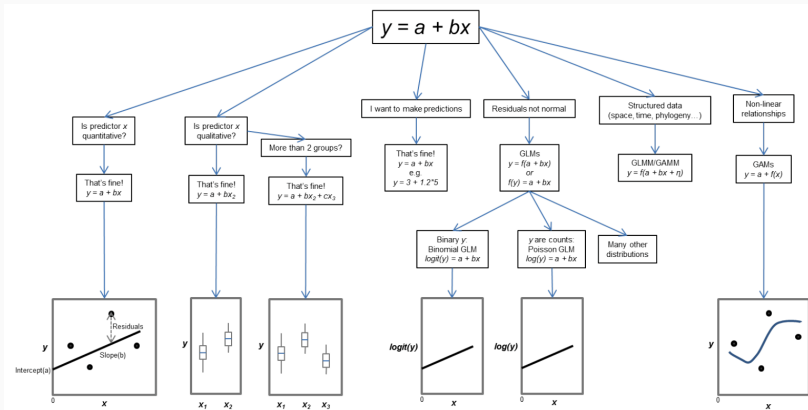
$a$  = intercept

$b$  = slope

$\sigma$  = residual variation

$\varepsilon$  = residuals

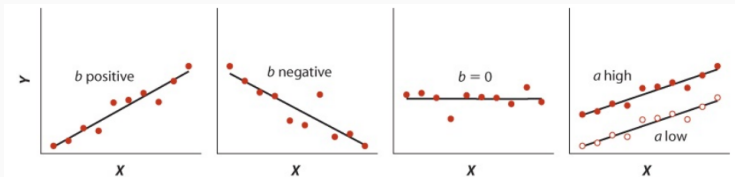
# Regresión estadística



# Regresión estadística

Regresión lineal:

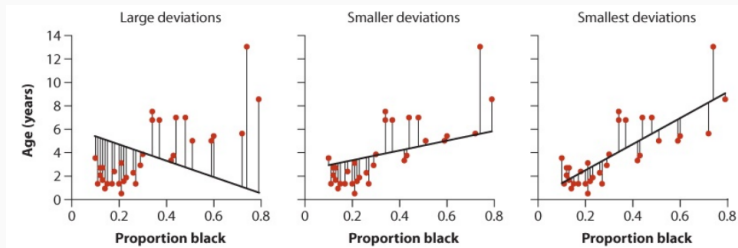
- Relación lineal entre las variables



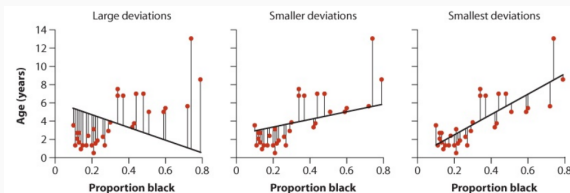
# Regresión estadística

Regresión lineal:

- Minimiza el *error residual*



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



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

$$a = \bar{y} - b\bar{x} \quad (2)$$

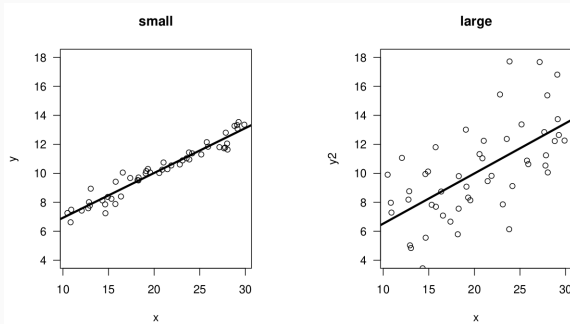
- Residuos: diferencia entre valor observado y predicho
- Recuerda:

$$y_i = a + bx_i + \varepsilon_i$$

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

# Regresión estadística

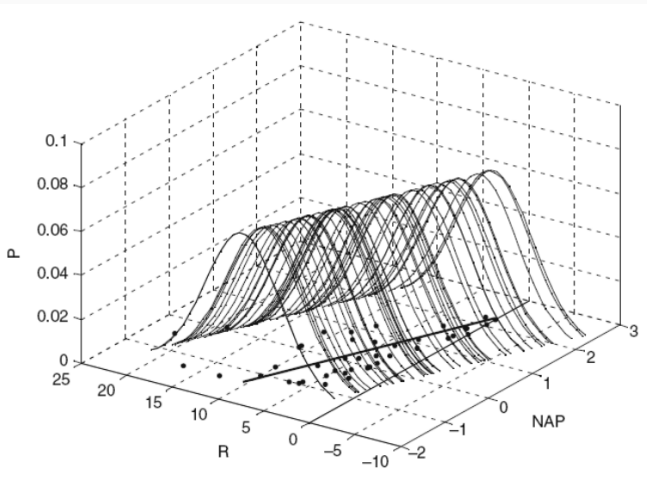
- Residuos: diferencia entre valor observado y predicho





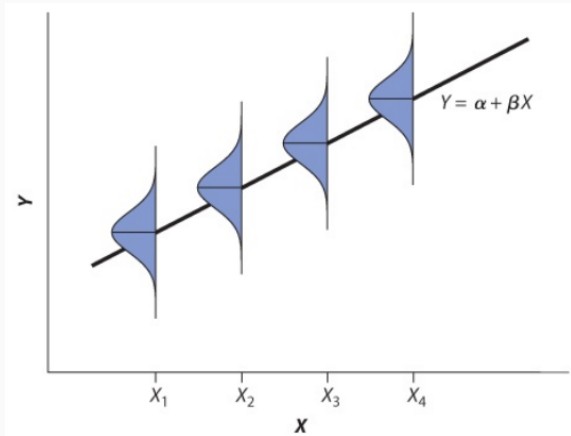
# Regresión estadística

- 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



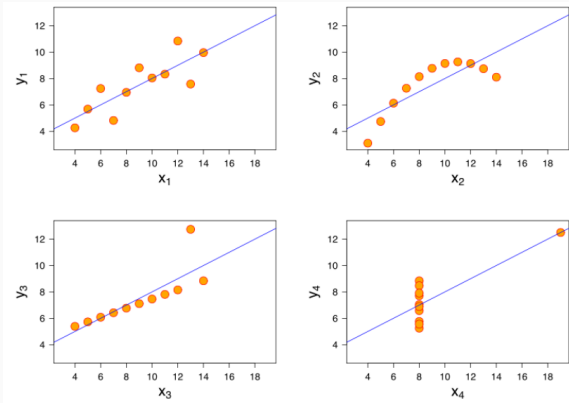
*Importante:* Esto no implica que la variable respuesta, o la variable independiente, deban tener una distribución normal!

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

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

**Siempre**

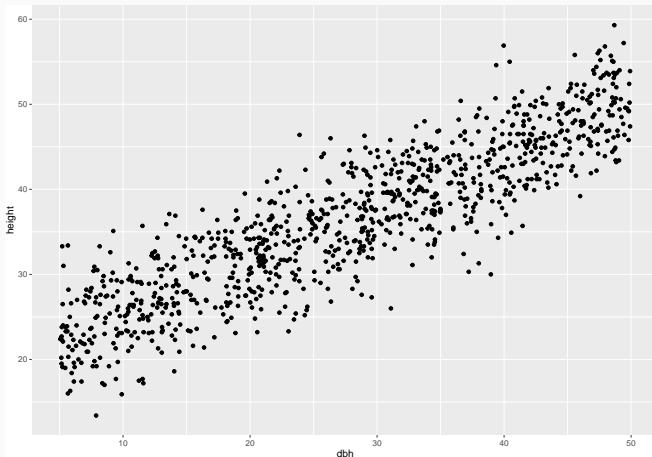
Visualiza los datos como primer paso



# Regresión estadística

¿Hay outliers en los datos?

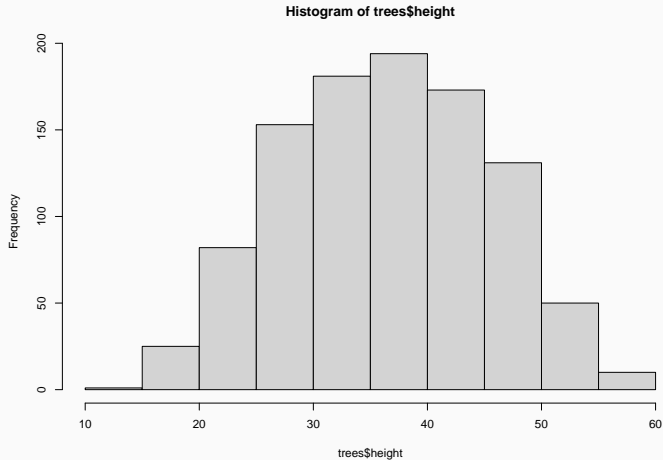
```
ggplot(trees, aes(dbh, height)) +  
  geom_point()
```



# Regresión estadística

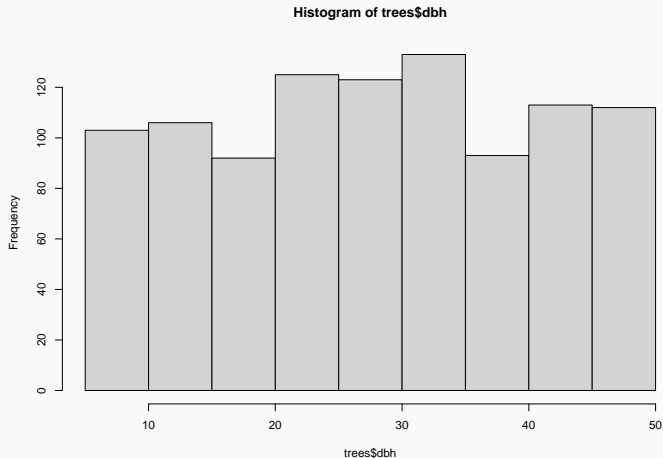
¿Cómo están distribuidas las variables independientes y respuesta?

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



¿Cómo están distribuidas las variables independientes y respuesta?

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





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

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

que se corresponde con:

$$height_i = a + b \cdot DBH_i + \varepsilon_i$$

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

¿Y ahora?

```
m1
```

```
##
```

```
## Call:
```

```
## lm(formula = height ~ dbh, data = trees)
```

```
##
```

```
## Coefficients:
```

```
## (Intercept)          dbh
```

```
##      19.3392      0.6157
```

# Regresión estadística

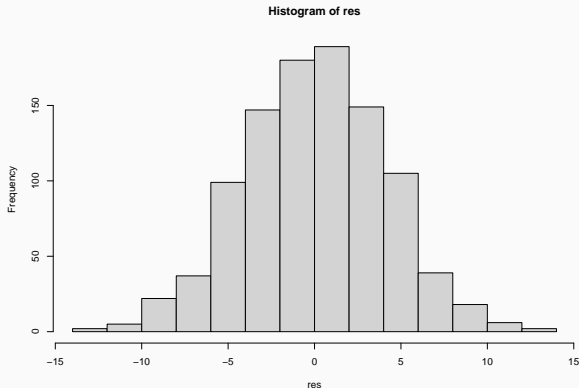
```
summary(m1)
```

```
##  
## 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

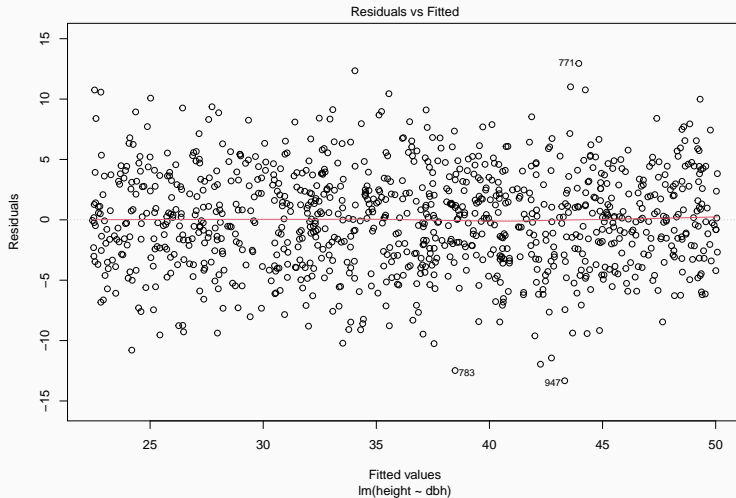
Antes de interpretar el resultado, comprobamos que los residuos se ajustan a una distribución normal

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

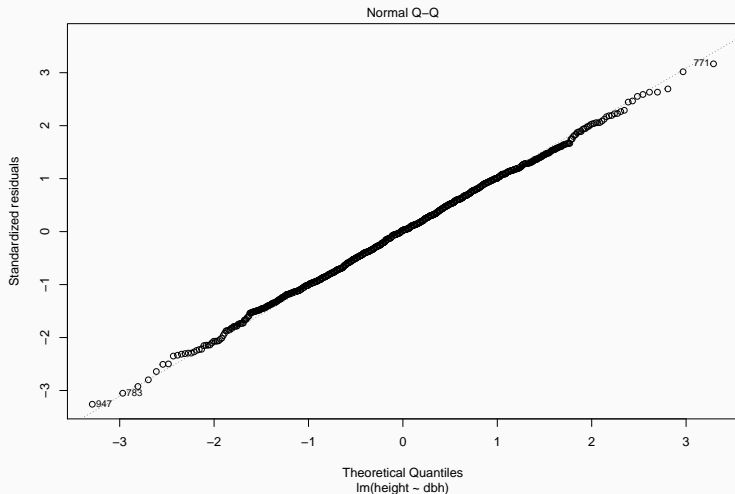


# Regresión estadística

```
plot(m1)
```

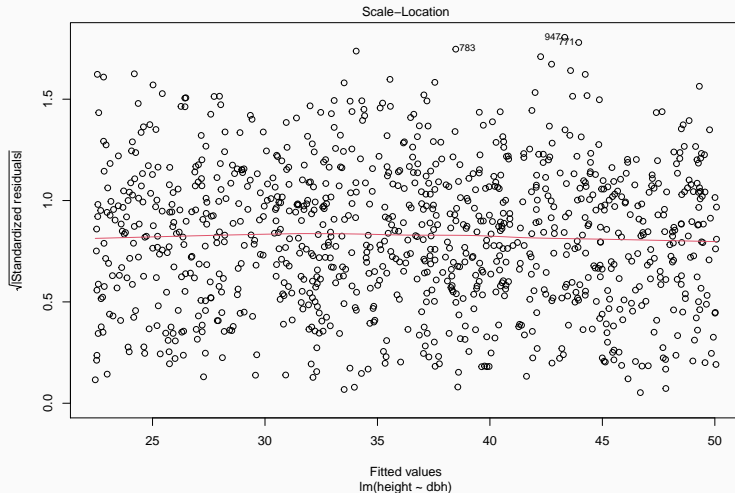


```
plot(m1)
```



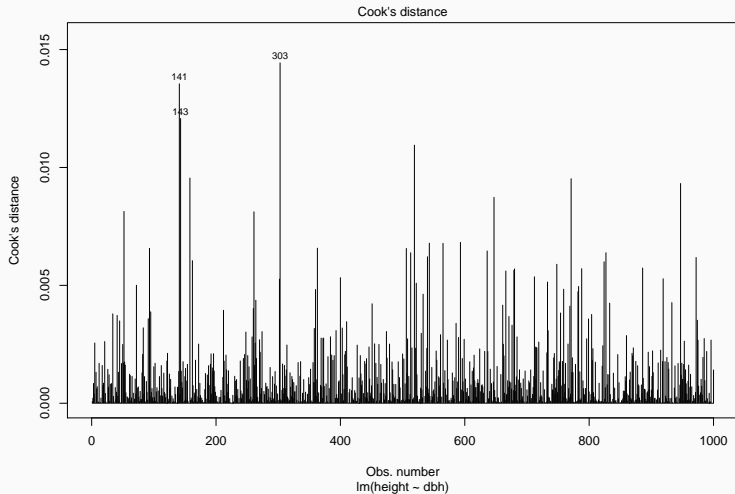
# Regresión estadística

```
plot(m1)
```



# Regresión estadística

```
plot(m1)
```





Una vez comprobamos que el modelo ajusta bien

# Regresión estadística

- Caso básico:
  - variable respuesta continua
  - Una variable independiente continua

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

- marco general (UHU, WS p671)

# Regresión estadística

- Caso básico:
  - variable respuesta continua
  - Una variable independiente continua

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

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

# Regresión estadística

- Caso básico:
  - variable respuesta continua
  - Una variable independiente continua

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

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

# Regresión estadística

- Caso básico:
  - variable respuesta continua
  - Una variable independiente continua

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

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

# Regresión estadística

- Caso básico:
  - variable respuesta continua
  - Una variable independiente continua

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

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

# Regresión estadística

- Caso básico:
  - variable respuesta continua
  - Una variable independiente continua

<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

- Caso básico:
  - variable respuesta continua
  - Una variable independiente continua

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

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



# Regresión estadística

- Caso básico:
  - variable respuesta continua
  - Una variable independiente continua

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

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

# Regresión estadística

- Caso básico:
  - variable respuesta continua
  - Una variable independiente continua

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

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

# Regresión estadística

- Caso básico:
  - variable respuesta continua
  - Una variable independiente continua

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

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

- Caso básico:
  - variable respuesta continua
  - Una variable independiente continua

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

- marco general (UHU, WS p671)
- residuos (WS p689)
- distribución 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 ( $R^2$ )

- 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)

- 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)