

# Analítica de datos aplicada a estudios sobre desarrollo

Giovany Babativa, PhD

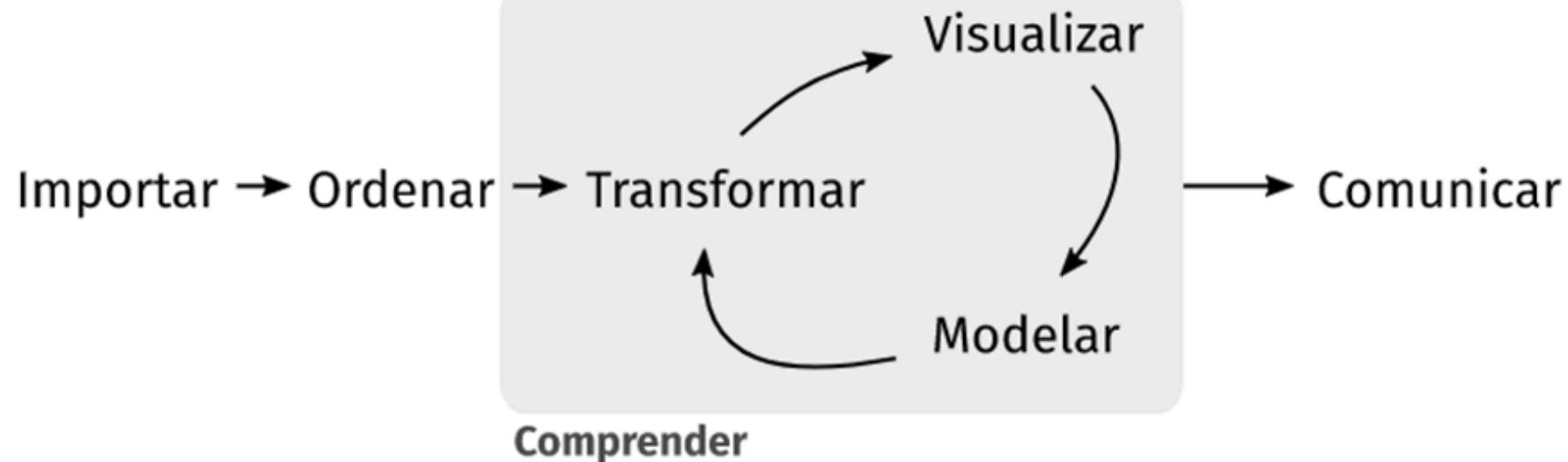
# Sobre Mi

Soy PhD en Estadística, más de 15 años de experiencia en el sector académico, actual director de analítica en CNC, miembro del comité de expertos en pobreza en el DANE y consultor experto de la División de Estadística de la CEPAL. Ex-decano de la Facultad de Estadística USTA, ex-director de operaciones en el ICFES, más de 40 evaluaciones de impacto o de resultados...

Puedes encontrarme en:

-  [Google scholar](#)
-  [GitHub. <https://github.com/jgbabativam>](https://github.com/jgbabativam)
-  [linkedin](#)
-  [j.babativamarquez@uniandes.edu.co](mailto:j.babativamarquez@uniandes.edu.co)

# Proceso de analítica



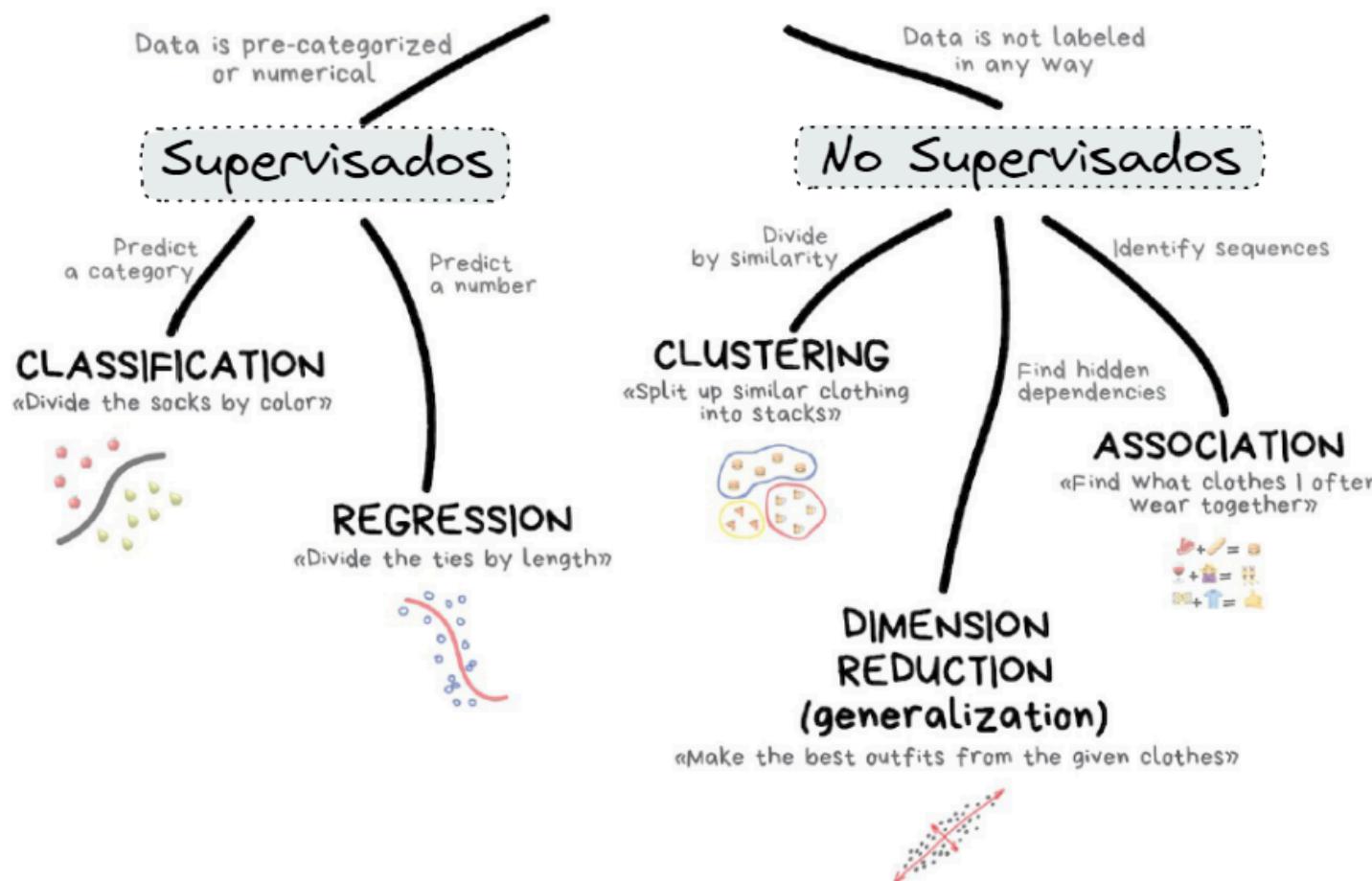
Programar

Wickham, H. y otros (2023)

# MÉTODOS MULTIVARIANTES

# Modelos de analítica

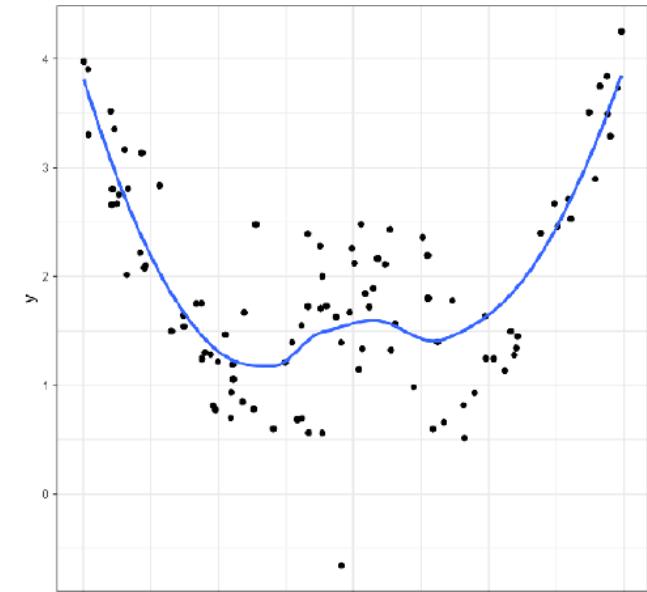
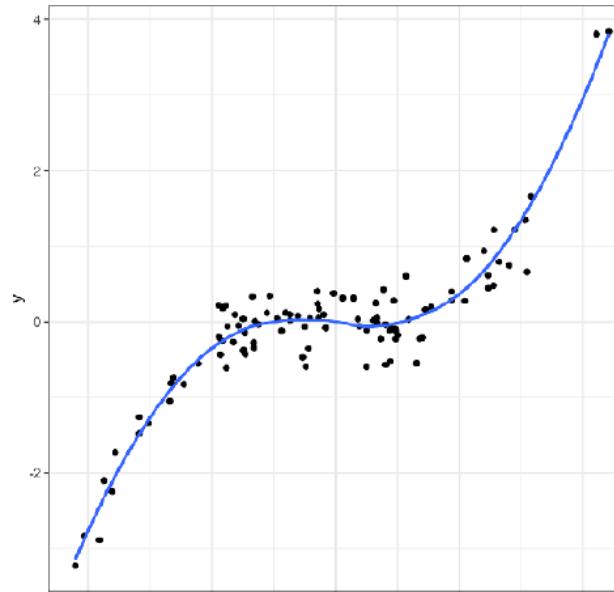
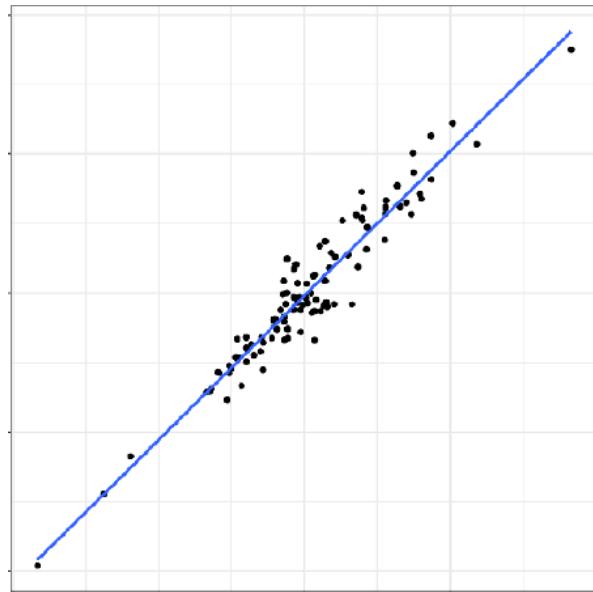
## Modelos Multivariantes



Fuente: Machine Learning for Everyone

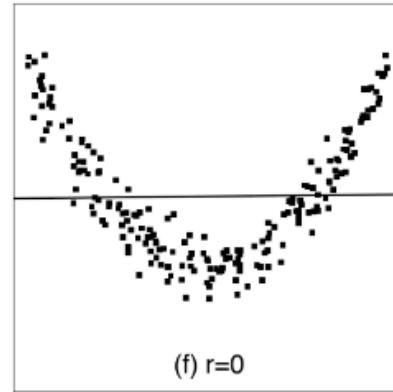
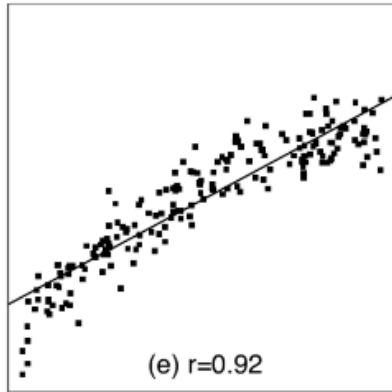
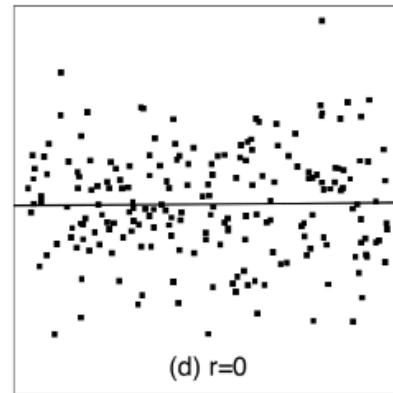
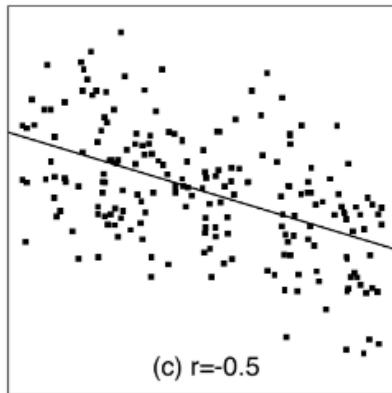
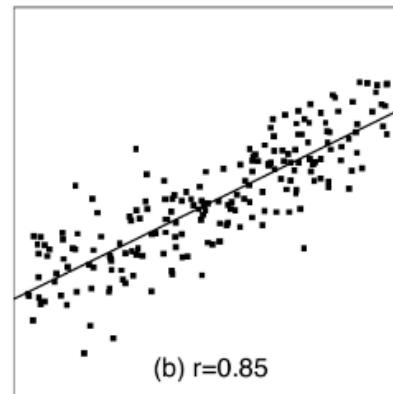
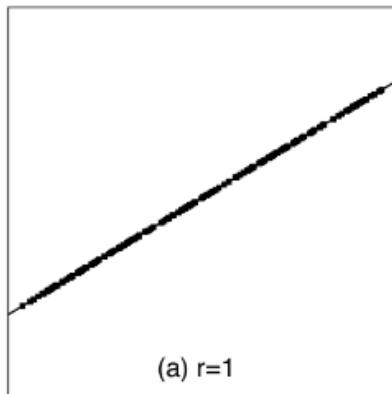
# Modelando datos

Primero explore los datos para identificar el tipo de relación: lineal o no lineal.



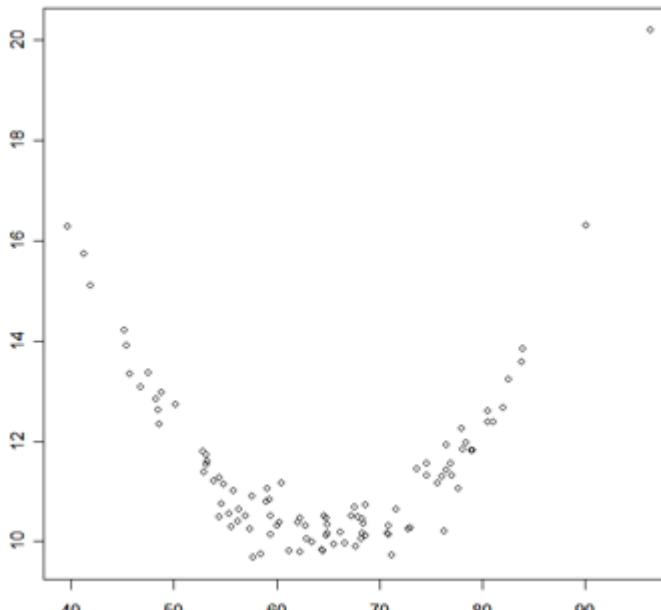
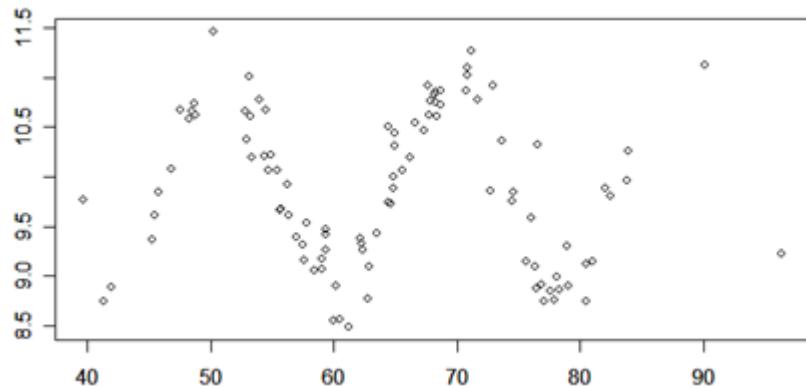
Correlación de Pearson, gráficos de dispersión simples o matricial.

# Correlación lineal



# Otros tipos de asociación

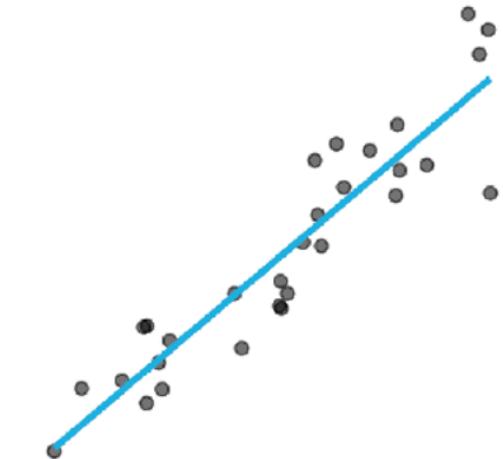
Una Correlación de **CERO** no debe interpretarse como “No existe asociación”, solo permite concluir que no hay asociación **lineal** pero puede existir una relación de otro tipo. Por ejemplo, Salario Vs. Experiencia. Otro aspecto a considerar es la presencia de datos atípicos (“raros”) que puedan ser influyentes.



# Especificación del modelo



**Datos**

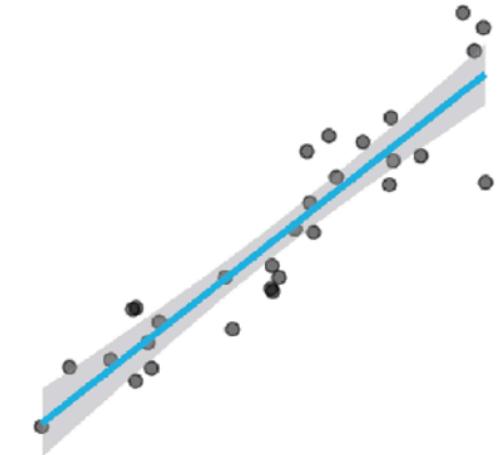


**Función de enlace**

# ¿Es un buen modelo?

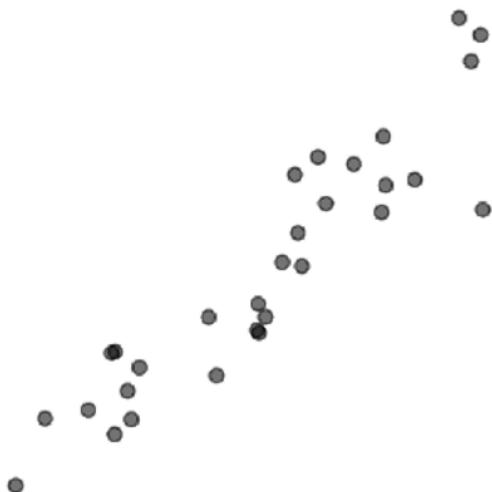


Datos

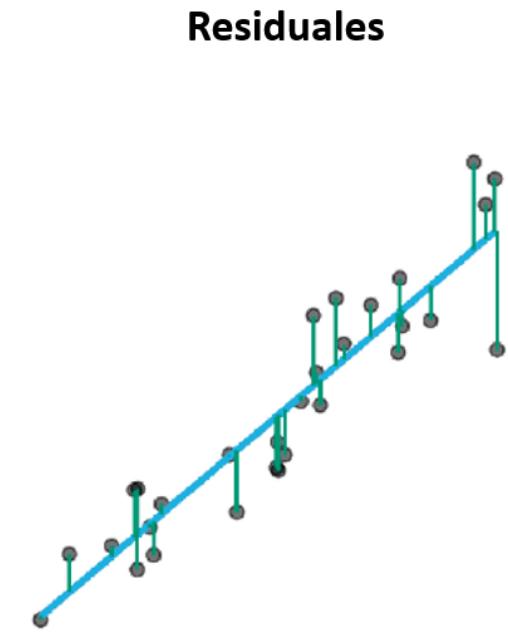


Función de enlace

# Análisis de los residuales



Datos



Función de enlace

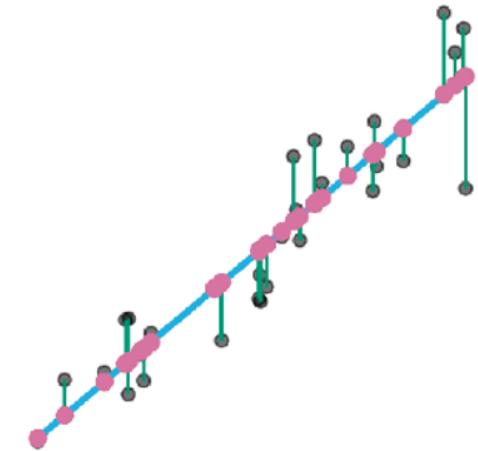
# Pronósticos



Datos



Predicciones



Función de enlace

# Algoritmos

Algunos modelos son:

- Lineales: `lm()`.
- Generalizados: `glm()`.
- Bayesianos: `stan_glm()`
- Penalizados: `glmnet()`
- ML: `tidymodels`

# Modelo Lineal

Sea  $\mathcal{D} = \{(y_i, \mathbf{x}_i) : i = 1, \dots, n\}$ , con  $y_i$  la  $i$ -ésima respuesta medida en una escala continua;  $\mathbf{x}_i = (x_{i1}, \dots, x_{ip})^t \in \mathbb{R}^p$  es el vector de variables predictoras; y  $n$  ( $\gg p$ ) es el tamaño de la muestra. El modelo lineal se especifica así:

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \varepsilon_i \text{ con } \varepsilon_i \stackrel{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma^2)$$

# Aprender un lenguaje de analítica



Arte de Allison Horst

Diapositivas disponibles en [GitHub](#).

# El entorno tidyverse

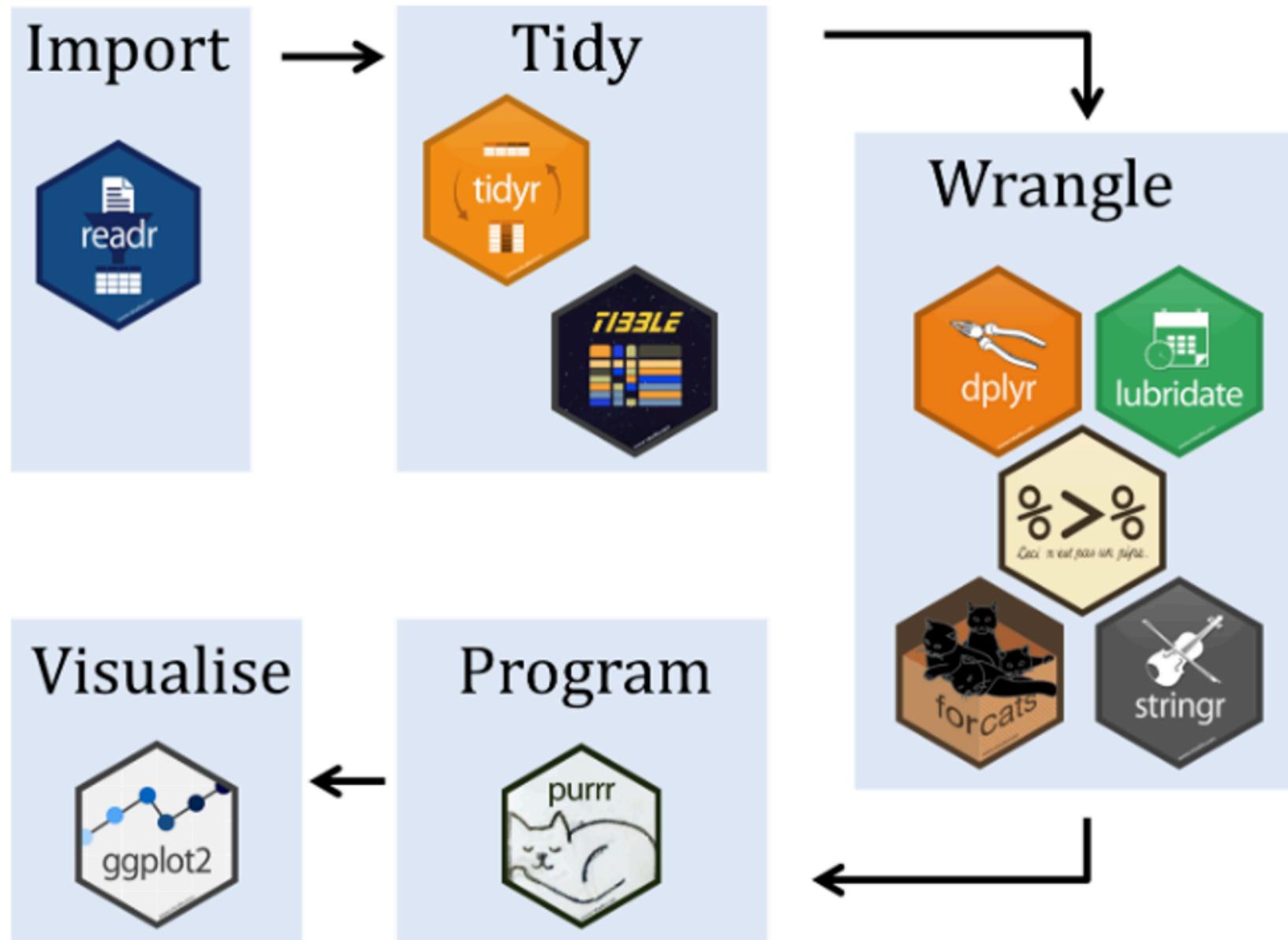


```
library(tidyverse)
```



```
library(readr)
library(dplyr)
library(tidyr)
library(ggplot2)
library(purrr)
library(tibble)
library(stringr)
library(forcats)
```

# Flujo de trabajo



# ESTUDIO DE CASO

- El instrumento del DASS 21 permite construir una escala de Depresión, Ansiedad y Estrés (DASS-21). Investigue más sobre su construcción y propiedades psicométricas. Una versión del instrumento puede ser consultada [aquí](#)
- Explore el conjunto de datos `DASS21.sav` el cual contiene los resultados para una muestra de 800 personas de Colombia realizada en el año 2022.

```
1 library(pacman)
2 p_load(tidyverse, broom, modelr, haven, labelled, performance,
3         skimr, corrplot, psych, gt, gtsummary, pROC)
4
5 url <- "https://github.com/jgbabativam/AnaDatos/raw/main/datos/DASS21.sav"
6 dass <- read_sav(url)
```

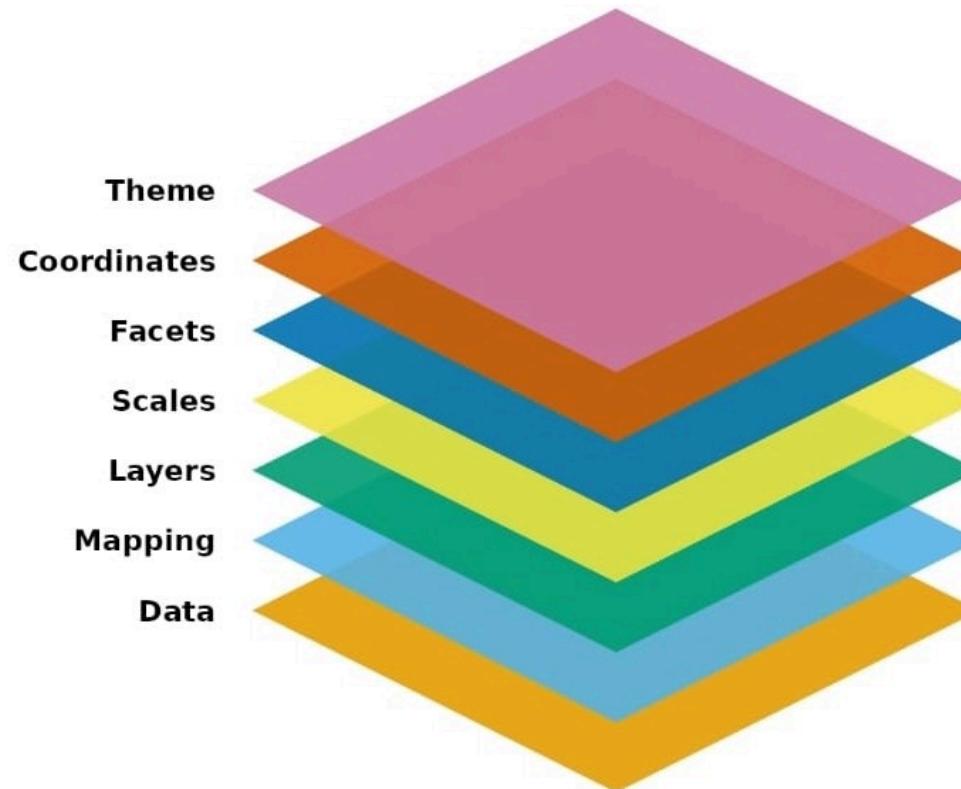
Puede usar `lapply(dass, function(x) attributes(x)$label)` para ver las etiquetas de las preguntas.

1. Realice un diagrama de dispersión y calcule la correlación entre las variables cuantitativas de nivel de depresión, estrés y ansiedad.
2. ¿Considera que el grado de asociación se diferencia entre hombres y mujeres?, haga los gráficos de dispersión segmentados por sexo
3. Realice los análisis que le permitan concluir sobre la asociación entre la depresión y la satisfacción con la vivienda, trabajo, amigos, vecinos y el barrio.
4. Ajuste el modelo de regresión usando las variables del numeral anterior y agregue la variable del sexo.

# La gramática de las gráficas

Requiere de al menos 3 elementos: datos, variables (aes), geometria.

```
1 ggplot(data = datos, aes(x = ___, y = ___)) +  
2   geom_point()
```



# Exploración visual de datos



Arte de Allison Horst

Diapositivas disponibles en [GitHub](#).

# Solución ejercicios 1 y 2

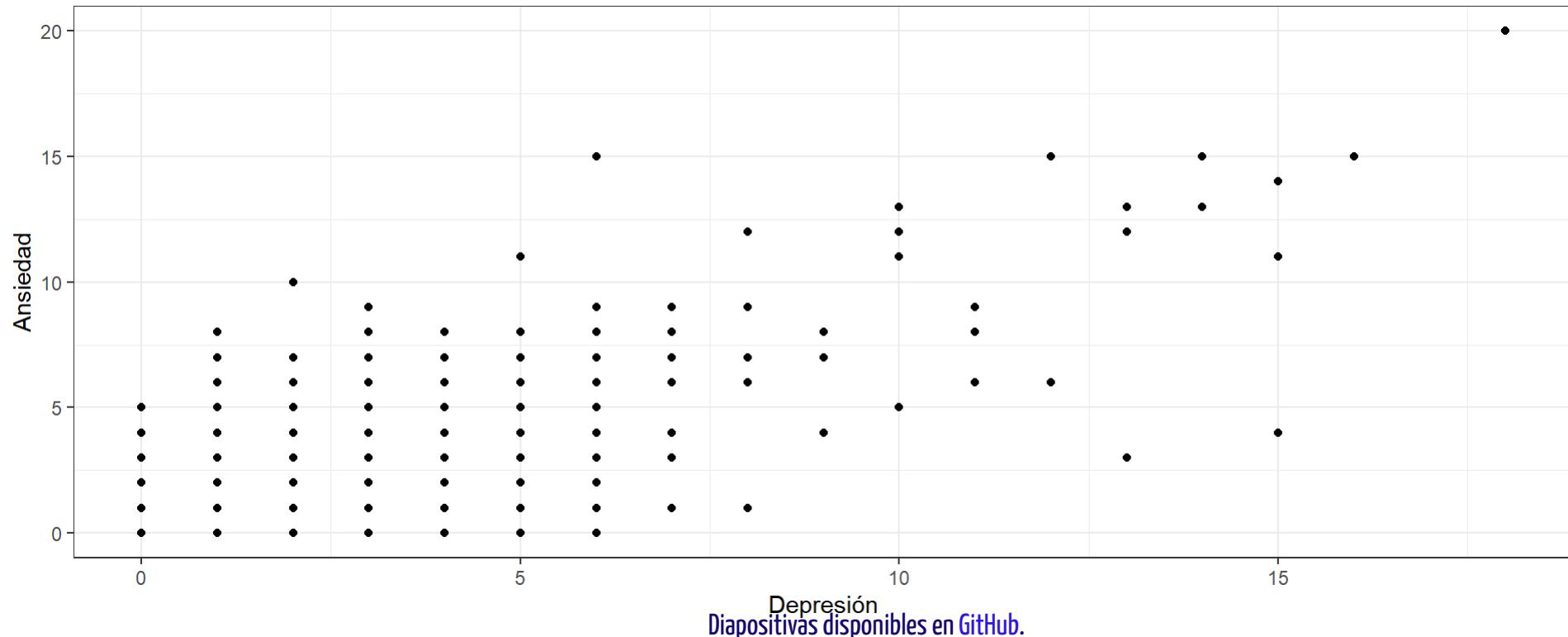
Diagrama 1

Diagrama 2

Diagrama 3

Diagrama 4

```
1 dass |>
2   ggplot(aes(x = DEPRESION, y = ANSIEDAD)) +
3     geom_point() +
4     labs(x = 'Depresión', y = 'Ansiedad') +
5     theme_bw()
```



## Correlación Gráfico

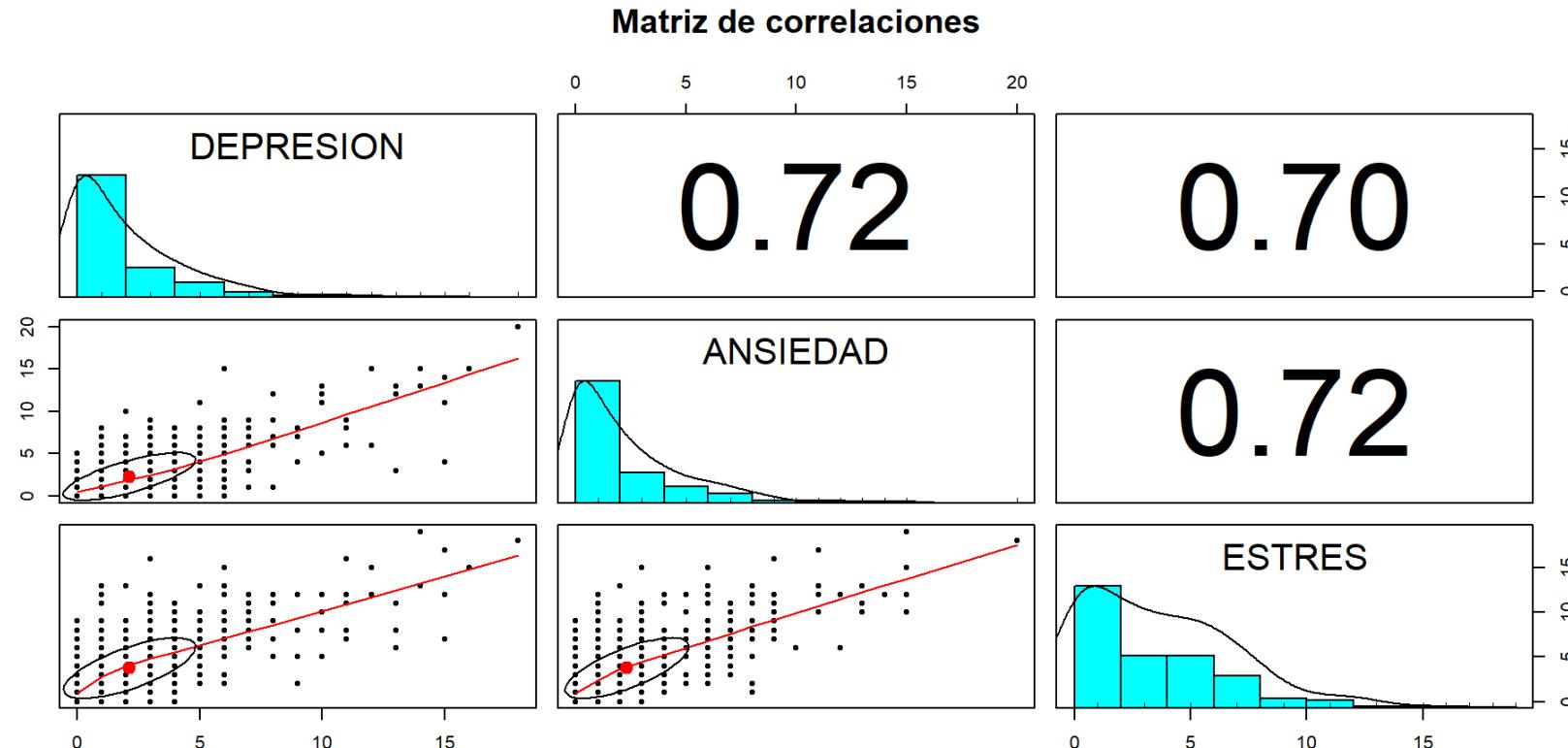
```
1 vars <- dass |> select(DEPRESION, ANSIEDAD, ESTRES)
2
3 cor(vars, use = "complete")
```

	DEPRESION	ANSIEDAD	ESTRES
DEPRESION	1.0000000	0.7237418	0.7007826
ANSIEDAD	0.7237418	1.0000000	0.7220293
ESTRES	0.7007826	0.7220293	1.0000000

# Matriz de correlación

Puede ver la correlación en forma de matriz usando `pairs.panels()` del paquete `psych`

```
1 pairs.panels(vars, main="Matriz de correlaciones")
```



# Punto 3 - Taller: DASS21

Realice los análisis que le permitan concluir sobre la asociación entre la depresión y la satisfacción con la vivienda, trabajo, amigos, vecinos y el barrio.

```
1 vars <- dass |> select(DEPRESION, P1_1, P1_2, P1_3, P1_4, P1_5)  
2  
3 round(cor(vars, use = "complete"), 3)
```

	DEPRESION	P1_1	P1_2	P1_3	P1_4	P1_5
DEPRESION	1.000	-0.107	-0.161	-0.100	-0.096	-0.097
P1_1	-0.107	1.000	0.430	0.413	0.357	0.429
P1_2	-0.161	0.430	1.000	0.601	0.480	0.556
P1_3	-0.100	0.413	0.601	1.000	0.592	0.590
P1_4	-0.096	0.357	0.480	0.592	1.000	0.622
P1_5	-0.097	0.429	0.556	0.590	0.622	1.000

# Punto 4 - Taller: DASS21

Ajuste del modelo de regresión adicionando la variable del sexo.

1. Revise la escala de las variables que se usarán en el modelo.
2. Convierta la variable `sexo` en factor así: `dass$sexo <- as_factor(dass$sexo)`
3. Ajuste el modelo de regresión y presente los resultados.
4. Interprete los coeficientes y el valor p.
5. Analice la validez de los resultados.

# Preparación de los datos

```
1 dass$sexo <- as_factor(dass$sexo)
2 vars <- dass |> select(DEPRESION, P1_1, P1_2, P1_3, P1_4, P1_5, sexo)
```

Use las funciones `glimpse(vars)` y `skim(vars)` para inspeccionar el conjunto de los datos.

# Ajuste del modelo

El modelo es de tipo lineal, así que se usa la función `lm()`

```
1 modelo <- lm(DEPRESION ~ P1_1 + P1_2 + P1_3 + P1_4 + P1_5 + sexo,  
2                   data = dass)
```

# Resultados

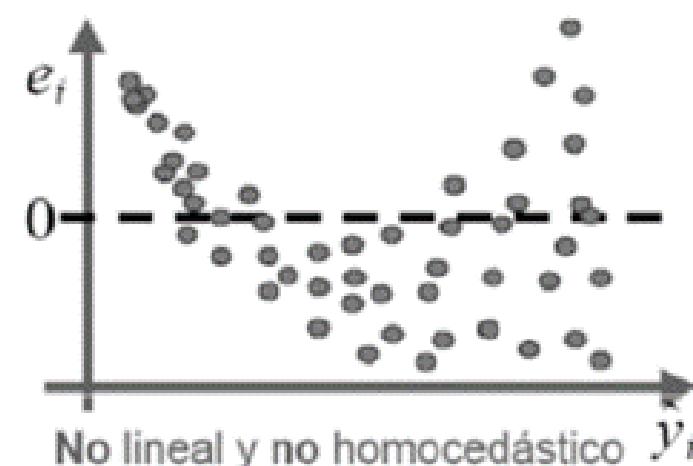
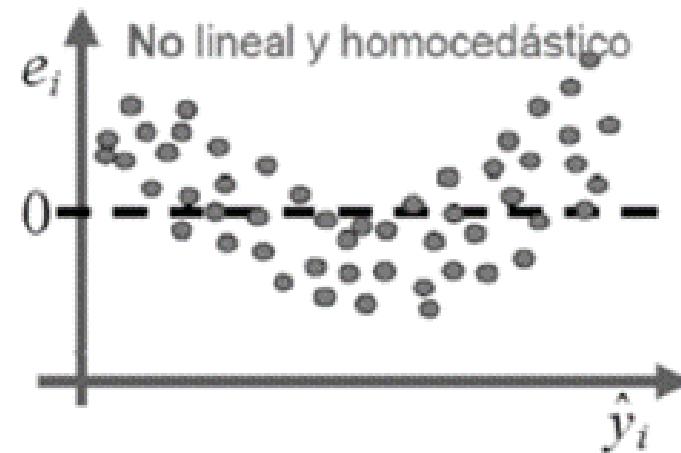
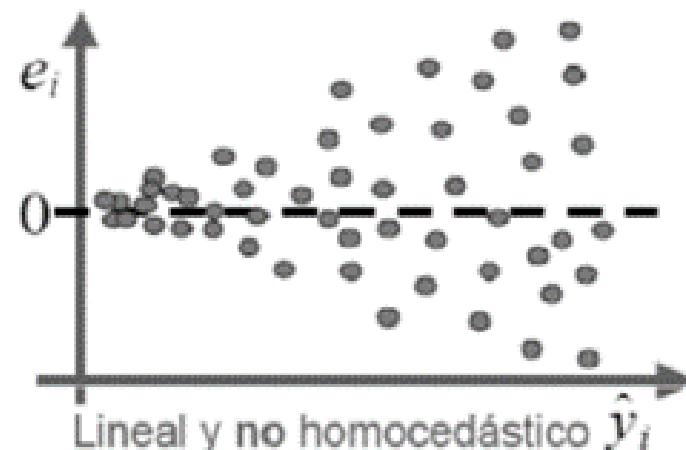
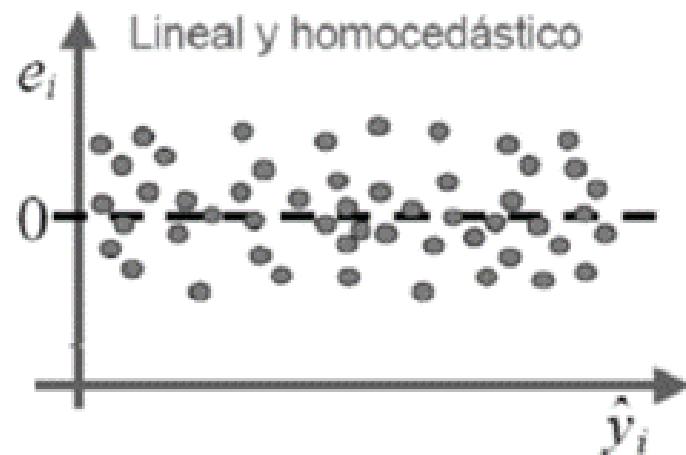
```
1 tbl_regression(modelo, intercept = TRUE) |>  
2 add_glance_table(include = c(r.squared, p.value))
```

Characteristic	Beta	95% CI		p-value
		<sup>1</sup>		
(Intercept)	4.0	3.0, 4.9		<0.001
Satisfacción con La vivienda	-0.09	-0.24, 0.06		0.3
Satisfacción con Su trabajo	-0.36	-0.59, -0.13		0.002
Satisfacción con Sus amigos	0.03	-0.22, 0.28		0.8
Satisfacción con Sus vecinos	-0.06	-0.31, 0.18		0.6
Satisfacción con Su barrio	0.02	-0.23, 0.28		0.9
B. Sexo:				
Hombre	—	—		
Mujer	0.06	-0.34, 0.47		0.8
R <sup>2</sup>	0.028			
p-value		<0.001		

<sup>1</sup>  
CI = Confidence Interval

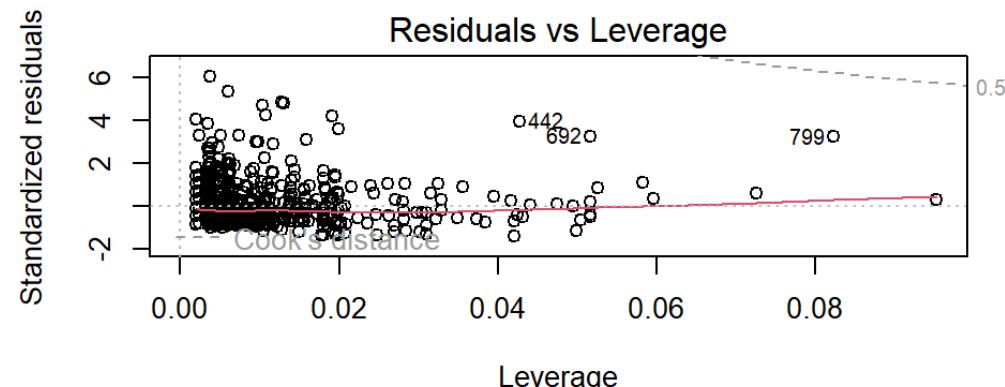
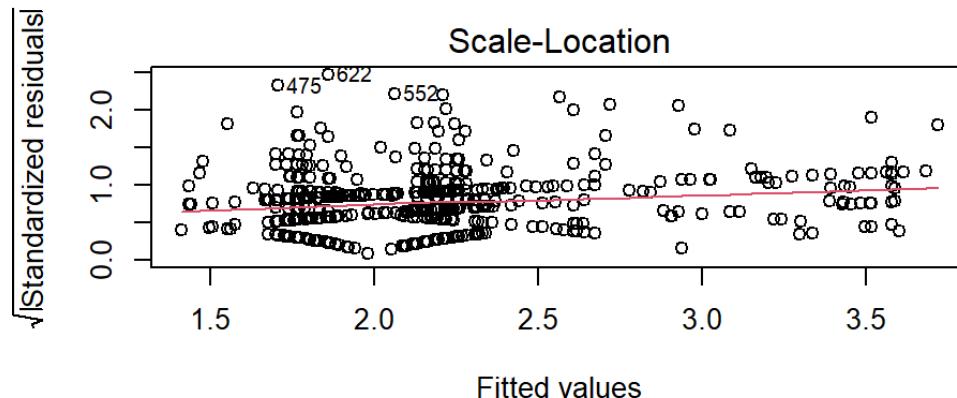
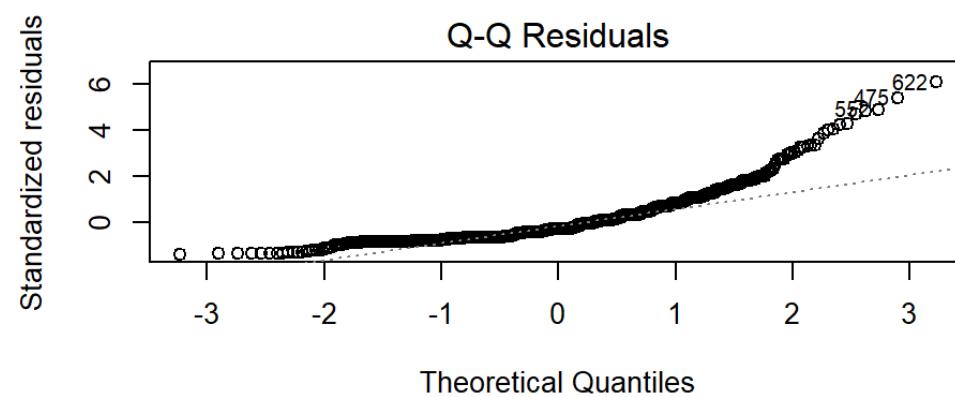
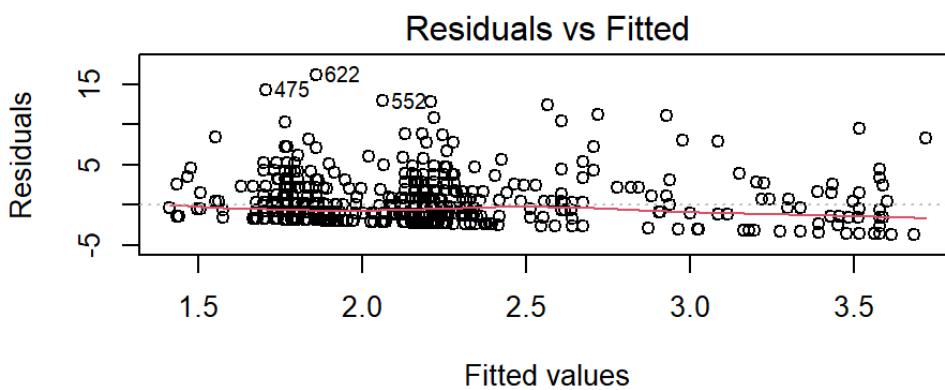
# Análisis de los supuestos

Que no se cumplan los supuestos puede afectar varios aspectos:  
sesgos, problemas de pronóstico, error de contraste.



# Análisis de los supuestos

```
1 par(mfrow = c(2, 2))  
2 plot(modelo)
```

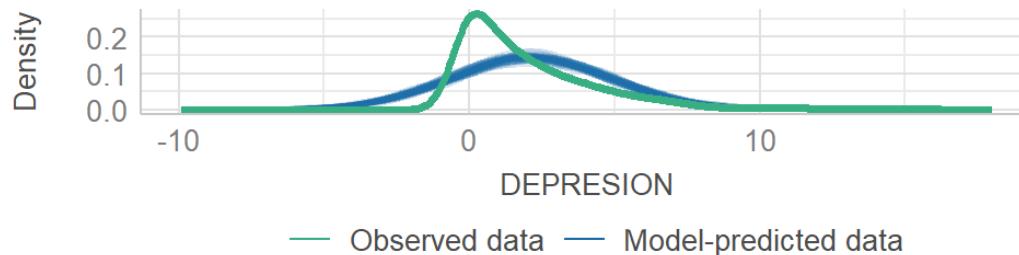


# Análisis de los supuestos

```
1 check_model(modelo)
```

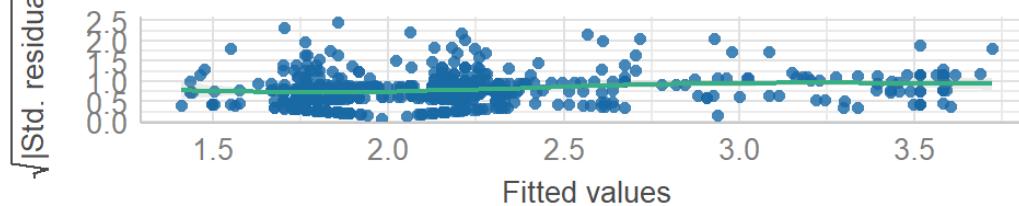
## Posterior Predictive Check

Model-predicted lines should resemble observed data line



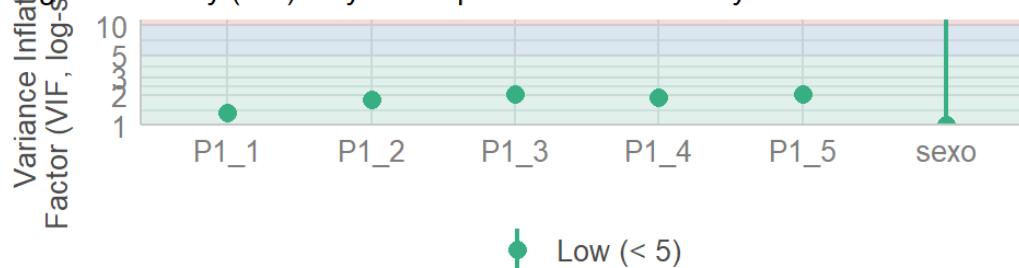
## Homogeneity of Variance

Reference line should be flat and horizontal



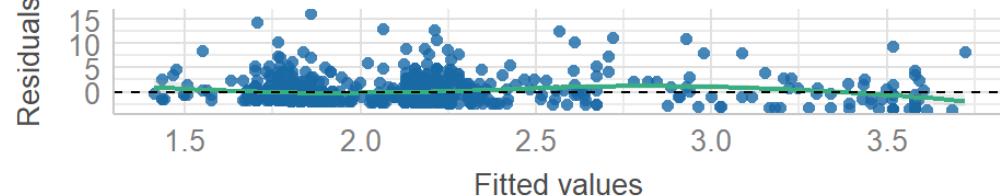
## Collinearity

High collinearity (VIF) may inflate parameter uncertainty



## Linearity

Reference line should be flat and horizontal



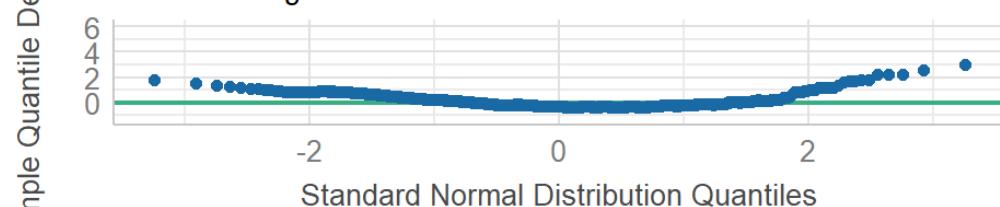
## Influential Observations

Points should be inside the contour lines



## Normality of Residuals

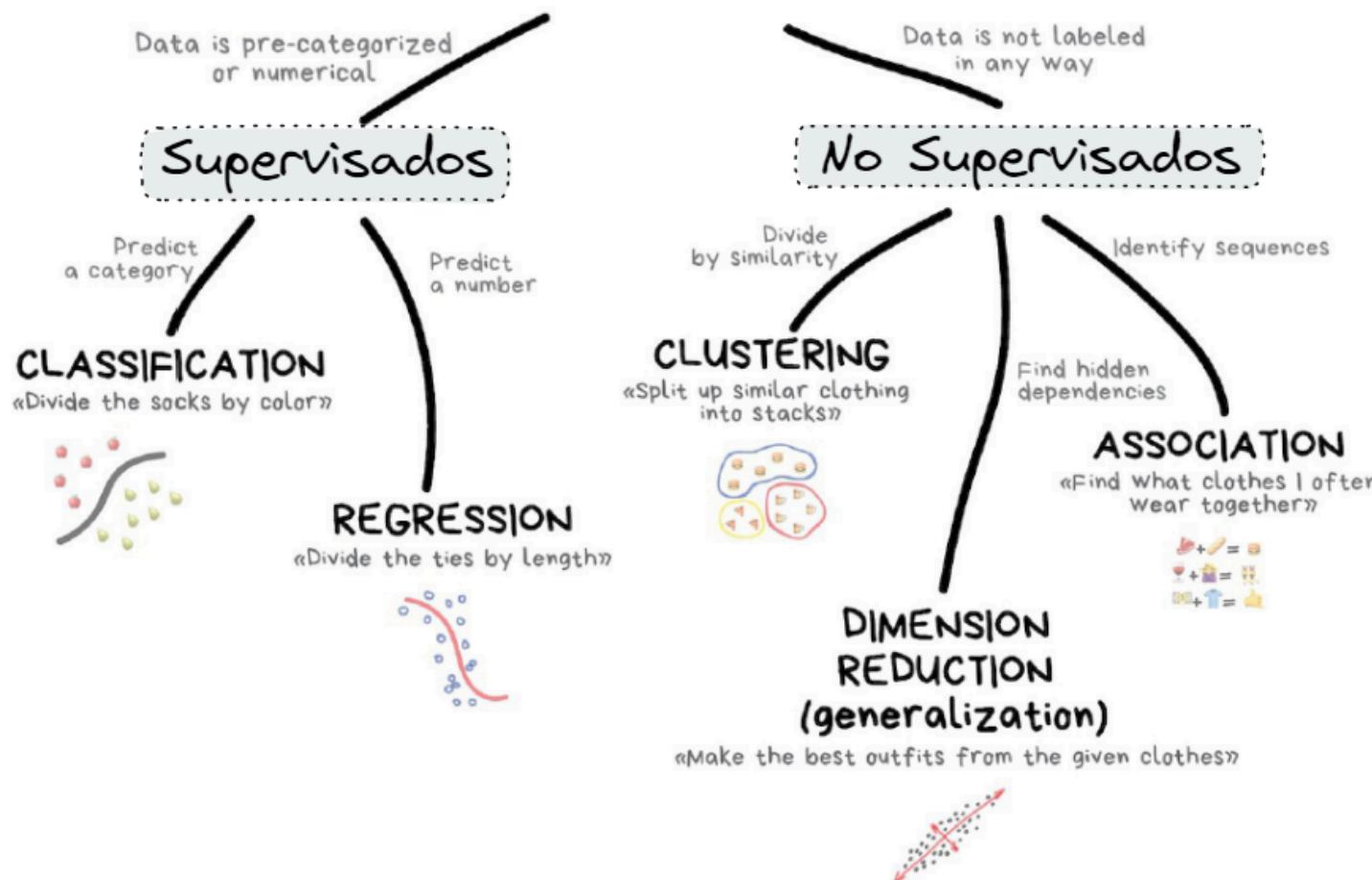
Dots should fall along the line



# REGRESIÓN NO LINEAL

# Modelos de analítica

## Modelos Multivariantes



Fuente: Machine Learning for Everyone

# MODELOS SUPERVISADOS DE CLASIFICACIÓN

# Modelos de respuesta discreta

Se usan cuando la variable dependiente es de tipo discreto, nominal o multinomial.

## Variable dependiente binaria

- Aprobar/no aprobar el examen / gestión del pte.
- Ser o no víctima de violencia doméstica.
- Tener o no tener una enfermedad.
- Estar o no en estado de pobreza.

En este caso:

$y_i \in \{0, 1\}$ , en donde 1 representa aprobar y 0 no aprobar.

# Modelos de respuesta discreta

Se usan cuando la variable dependiente es de tipo discreto, nominal o multinomial.

## Variable dependiente nominal

- Estado de ocupación: empleo formal, empleo informal, desempleo, inactivo, etc.
- Clasificación étnica: Sin étnia, Afrocolombiano, Indígena, Raizal, Palenquero.
- Tipo de cliente: Élite, Plata, Bronce, Masivo.

En este caso:

$y_i \in \{A, B, C, D, \dots\}$ , en donde cada letra representa una categoría nominal que es excluyente de las demás.

# Modelos de respuesta discreta

Se usan cuando la variable dependiente es de tipo discreto, nominal o multinomial.

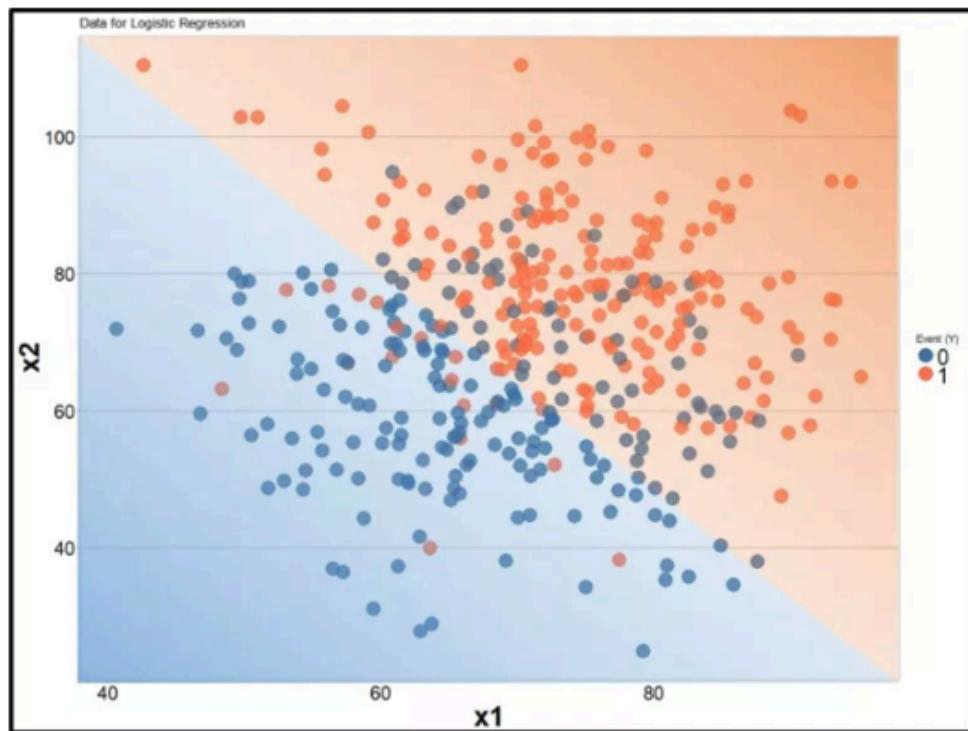
## Variable dependiente ordinal

- La clasificación del nivel de estrés en la escala: Sin estrés, con estrés leve, estrés moderado, estrés severo y estrés extremadamente severo.
- Clasificación de pobreza: Con pobreza extrema, en estado de pobreza, no pobre.

En este caso:

$y_i \in \{1, 2, 3, \dots\}$ , en cada valor está en una escala ordinal.

# Regla de decisión



- Necesitamos una regla para clasificar a cada observación.
- La imagen presenta una regla dadas dos variables independientes.
- La función de clasificación puede ser lineal o no lineal.

Imagen de "Elements of Statistical Learning" de Hastie and Tibshirani

# Regla de decisión

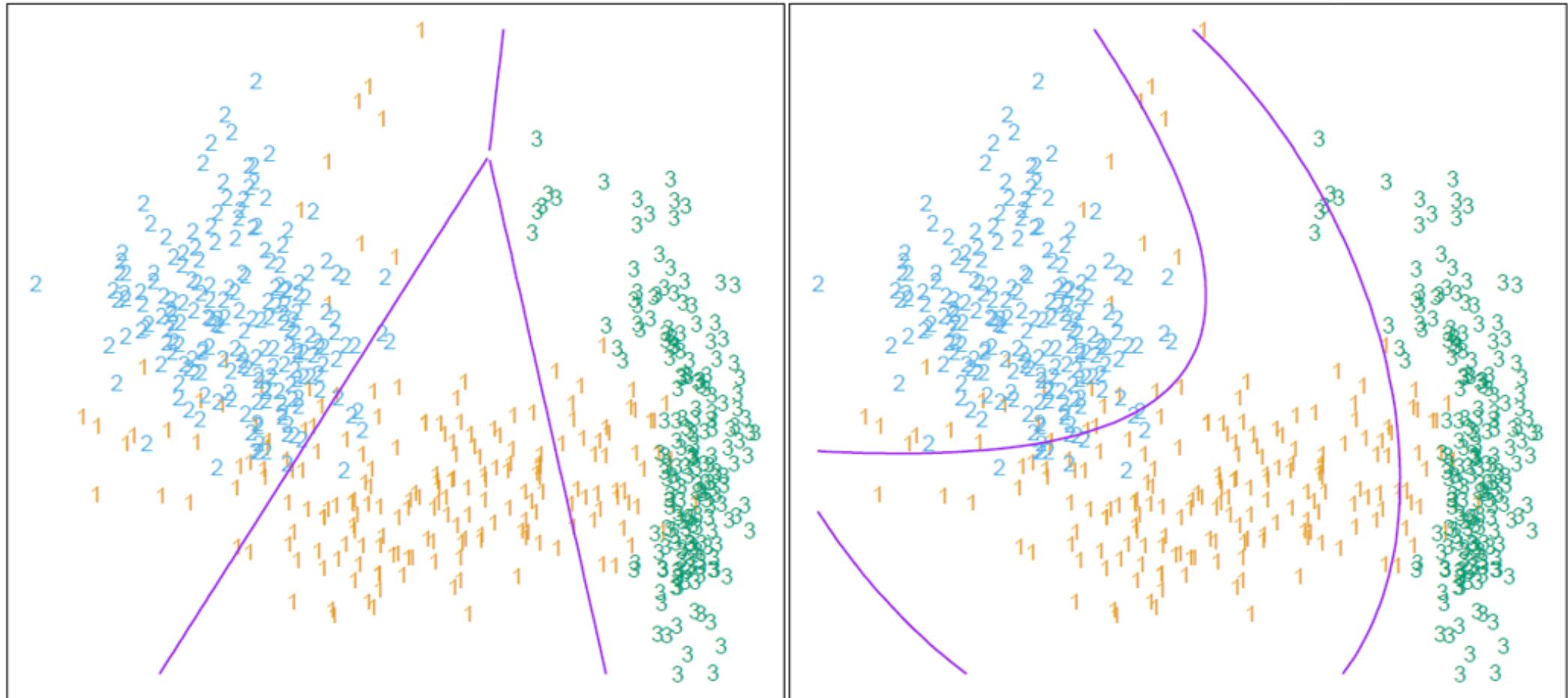


Imagen de “Elements of Statistical Learning” de Hastie and Tibshirani

# MODELO DE REGRESIÓN LOGÍSTICA

Suponga que puede clasificar a todas las observaciones en 2 grupos/clases (1: Éxito, 0: Fracaso). La función logística, permite calcular la probabilidad de cada suceso.

$$\pi(x) = P(Y = 1 | \mathbf{X} = \mathbf{x}) = \frac{\exp\{\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k\}}{1 + \exp\{\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k\}}$$

$$1 - \pi(x) = P(Y = 0 | \mathbf{X} = \mathbf{x})$$

¿Por qué no se debería aplicar regresión lineal en este caso?

# Riesgo (Odds)

$$Odds = \frac{\pi(x)}{1 - \pi(x)}$$

Suponga que en el evento de desertar del colegio, se encontró que la probabilidad de deserción cuando el niño tiene una madre sin estudios es  $\pi(x) = 0.71$ . Así que

$$Odds_A = \frac{0.71}{1 - 0.71} = 2.45$$

Por cada niño que no deserta, hay casi 3 que si lo hacen cuando la madre no tienen ningún nivel de estudios.

# Razón de Riesgos (Odds-Ratio - OR)

$$OR = \frac{Odds_A}{Odds_B}$$

Si la probabilidad de deserción cuando la madre tiene un nivel educativo de bachillerato es  $\pi(x) = 0.18$ . Calcule el OR entre el grupo con madre sin estudios y el grupo con madre con nivel de bachiller.

$$OR = \frac{Odds_{\text{Madre sin estudios}}}{Odds_{\text{Madre bachiller}}} = \frac{\frac{0.71}{1-0.71}}{\frac{0.18}{1-0.18}} = \frac{2.45}{0.22} = 11.2$$

Es 11 veces más probable que un niño cuya madre no tiene estudios abandone sus estudios en comparación con los niños cuyas madres

# Modelo de regresión logística

Cuando la probabilidad se calcula con una f.d.p. logística, se puede mostrar que:

$$\frac{\pi(x)}{1 - \pi(x)} = e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k}$$

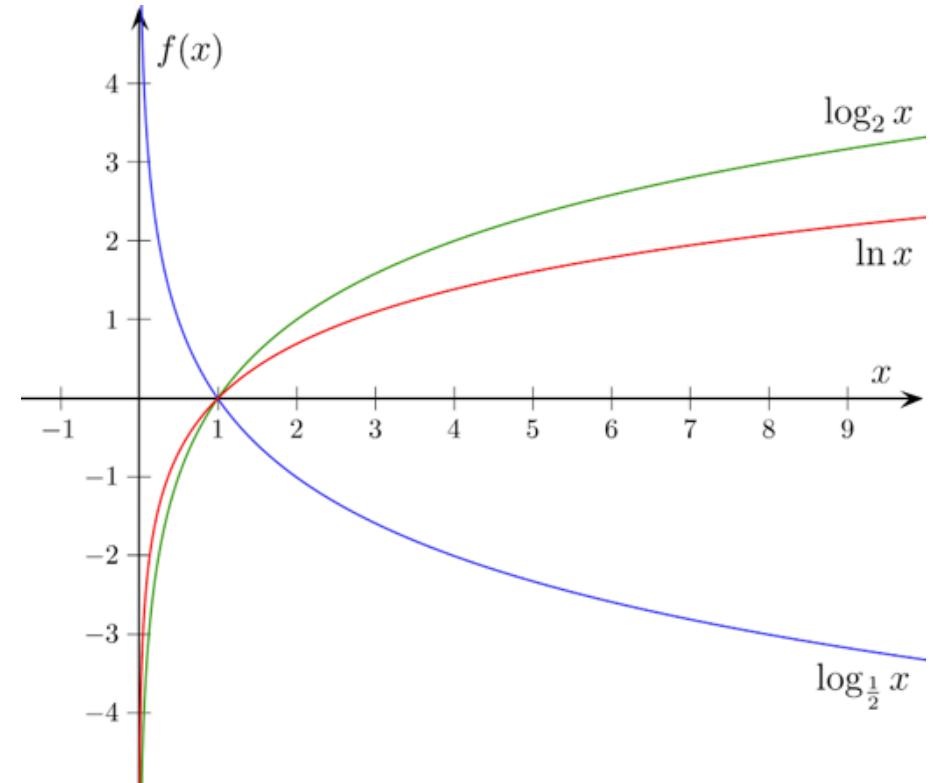
De esta forma el log-odds, conocido como el logit, conlleva a una relación lineal fácil de manejar y de interpretar:

$$\ln\left(\frac{\pi(x)}{1 - \pi(x)}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k$$

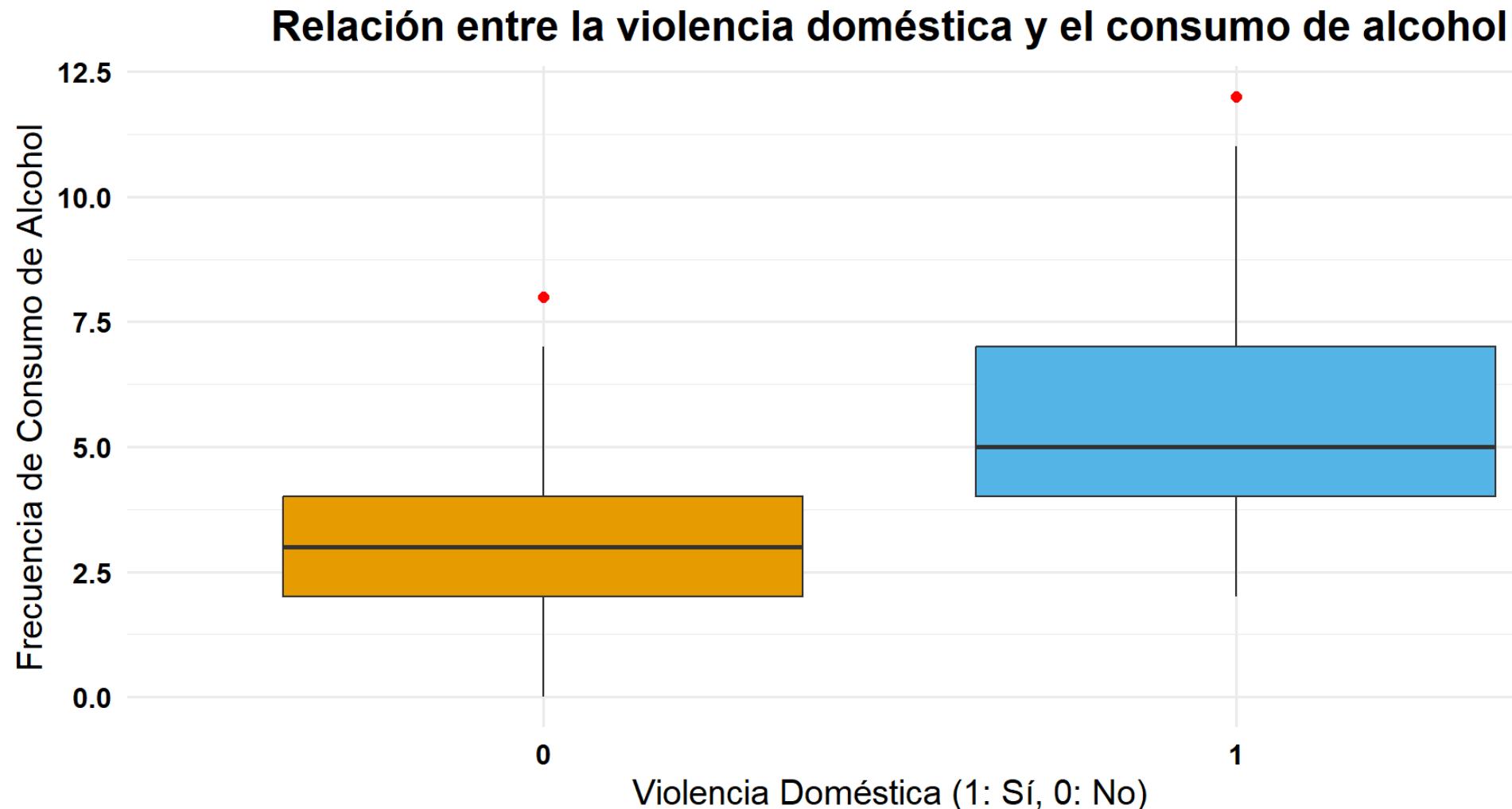
# Aspectos técnicos

La estimación de los parámetros debe hacerse por métodos iterativos debido a que la función de pérdida no es convexa.

- Newton – Raphson
- Descenso del gradiente
- Fisher
- Híbrido



# Análisis exploratorio



# Prueba Chi-cuadrado

Permite identificar la asociación entre dos variables cualitativas.

- Fumar está asociado con el cáncer de pulmón
- El género está asociado con la preferencia por tipos de música
- El nivel educativo está asociado con la afiliación política
- El tipo de alimentación está asociado con la presencia de enfermedades cardíacas
- La deserción estudiantil está asociada con el nivel educativo de la madre/padre.
- Empleabilidad tras un programa de formación: Empleado después de la formación (Sí/No) vs Edad, nivel educativo, género, años de experiencia laboral...

# Prueba Chi-cuadrado

$H_0$  : Las variables son independientes

$H_1$  : Las variables están relacionadas.

Si A y B son eventos independientes entonces se cumple que:

$$\mathbb{P}(A \cap B) = \mathbb{P}(A) * \mathbb{P}(B)$$

Si el valor  $p$  es inferior a un nivel  $\alpha$  se dice que hay suficiente evidencia estadística para rechazar  $H_0$ .

# Desempeño del modelo: Matriz de confusión

Tabla de clasificación<sup>a</sup>

Observado		Pronosticado		Corrección de porcentaje
		Cáncer		
Paso 1	Cáncer	No	Si	
		217	76	74,1
	No	29	558	95,1
		Porcentaje global		88,1

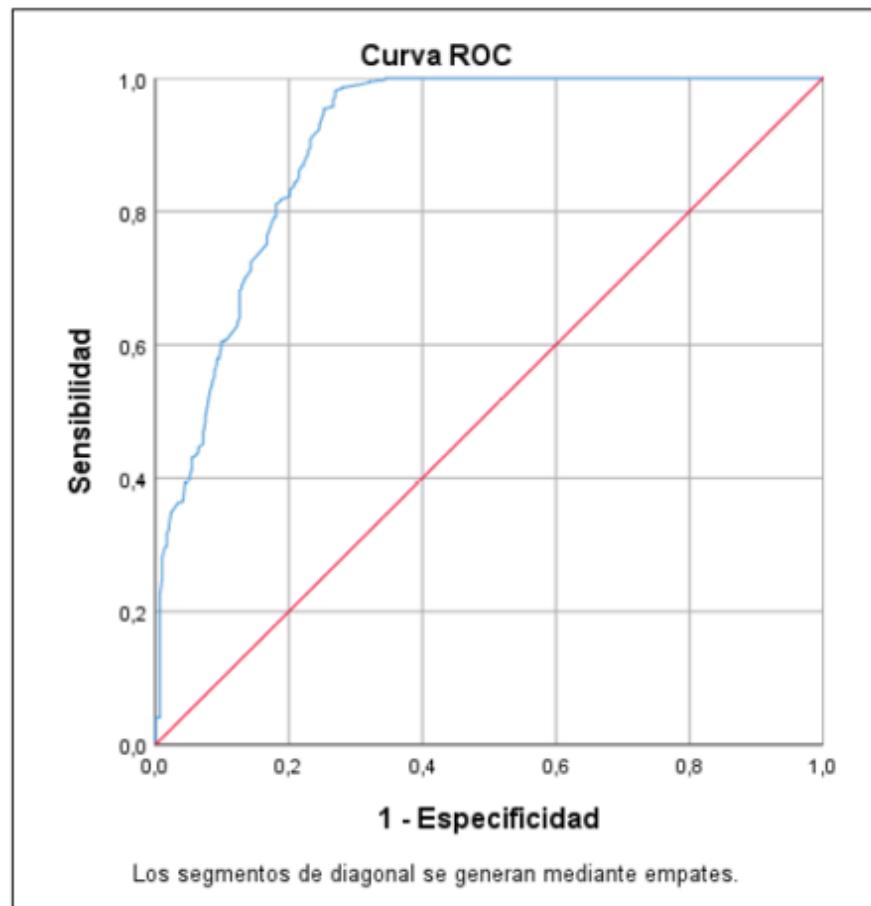
a. El valor de corte es ,500

**Sensibilidad:** Porcentaje de verdaderos positivos.

**Especificidad:** Porcentaje de verdaderos negativos.

# Desempeño del modelo: Curva ROC

Evalúa la capacidad predictiva del modelo. Se eligen varios puntos de corte y se representa gráficamente la capacidad del modelo para realizar una correcta clasificación



# Ejemplo

De acuerdo con la escala DASS, una persona con un puntaje de 8 o más puede ser considerada con problemas de depresión, ansiedad o estrés.

Ajuste un modelo que le permita identificar los factores de riesgo y los factores de protección contra la ansiedad, para ello utilice las variables del sexo, satisfacción y participación que se encuentran dentro del instrumento aplicado.

# Transformación de datos



Arte de Allison Horst

Diapositivas disponibles en [GitHub](#).

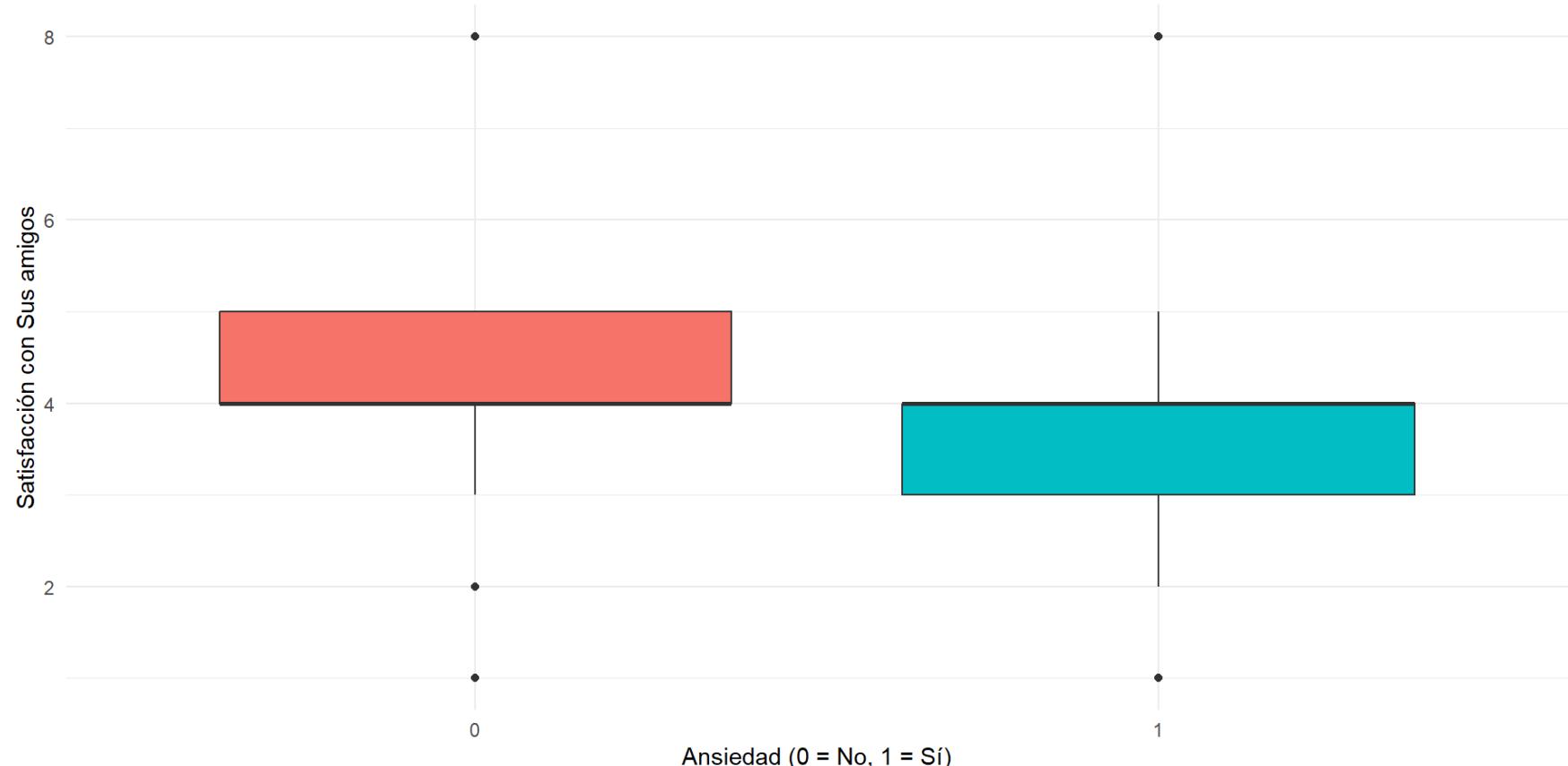
# Factores de riesgo asociados a la ansiedad

```
1 prepara <- dass |>
2   mutate(ansi = ifelse(ANSIEDAD >= 8, 1, 0)) |>
3   mutate(sexo = as_factor(sexo),
4         across(starts_with("P2"), ~relevel(as_factor(.), ref = "NO")))
```

Explore el conjunto de datos e identifique las diferencias frente a los datos iniciales.

# Análisis exploratorio: Box-plot

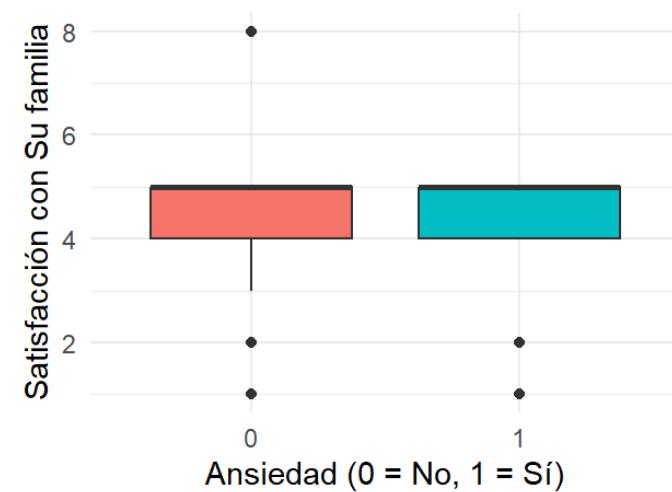
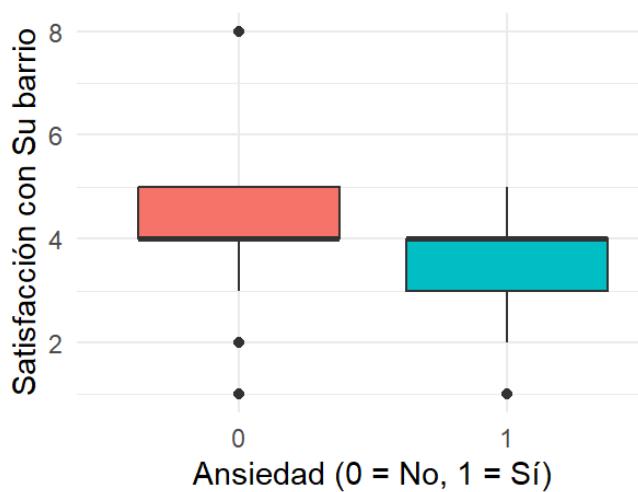
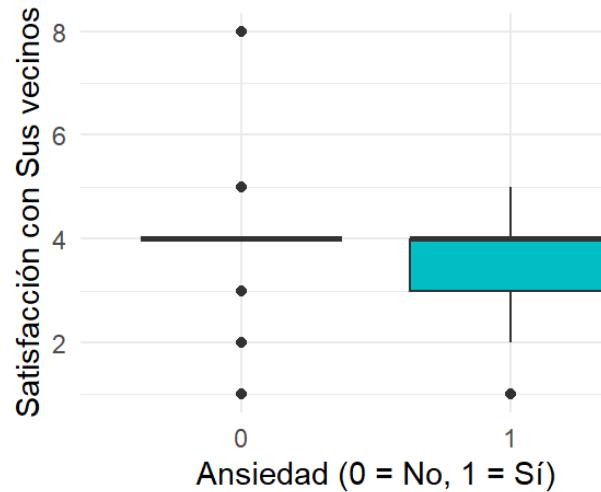
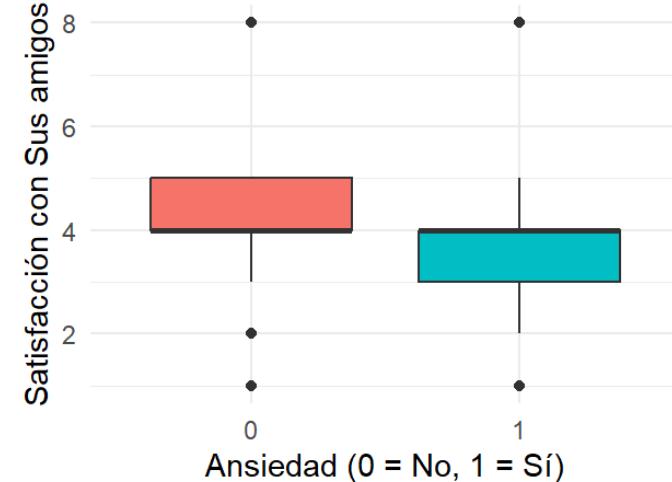
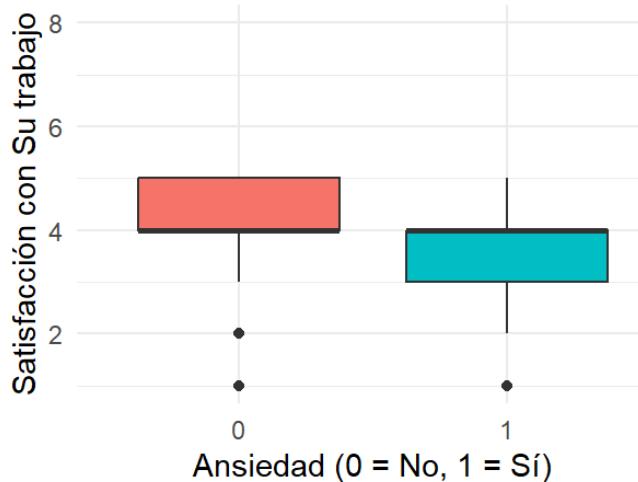
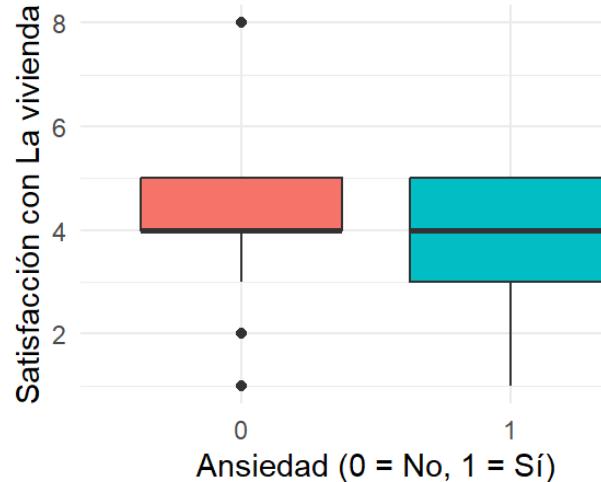
```
1 ggplot(prepara, aes(x = as.factor(ansi), y = P1_3, fill = as.factor(ansi))) +
2   geom_boxplot() +
3   labs(x = "Ansiedad (0 = No, 1 = Sí)", y = label_attribute(prepara$P1_3)) +
4   ylim(1, 8) +
5   theme_minimal() +
6   theme(legend.position = "none")
```



# Análisis exploratorio: Box-plot

```
1 grafica <- function(variable) {  
2   label_y <- attr(prepara[[deparse(substitute(variable))]], "label")  
3   ggplot(prepara, aes(x = as.factor(ansi), y = {{variable}}), fill = as.factor(ansi))  
4     geom_boxplot() +  
5     labs(x = "Ansiedad (0 = No, 1 = Sí)", y = label_y) +  
6     ylim(1, 8) +  
7     theme_minimal() +  
8     theme(legend.position = "none")  
9 }  
10  
11 p1 <- grafica(P1_1); p2 <- grafica(P1_2); p3 <- grafica(P1_3);  
12 p4 <- grafica(P1_4); p5 <- grafica(P1_5); p6 <- grafica(P1_6);  
13  
14 (p1 | p2 | p3) / (p4 | p5 | p6)
```

# Análisis exploratorio: Box-plot



# Análisis exploratorio: Prueba Chi-Cuadrado.

Esta prueba contrasta la hipótesis nula de independencia frente a la alternativa de asociación.

```
1 chisq.test(prepara$ansi, prepara$sexo)
```

```
Pearson's Chi-squared test with Yates' continuity correction

data: prepara$ansi and prepara$sexo
X-squared = 0.31065, df = 1, p-value = 0.5773
```

```
1 attributes(dass$P2_7)$label
```

```
[1] "Ha participado en Prácticas restaurativas para superar situaciones difíciles"
```

```
1 chisq.test(prepara$ansi, prepara$P2_7)
```

```
Pearson's Chi-squared test with Yates' continuity correction

data: prepara$ansi and prepara$P2_7
X-squared = 7.854, df = 1, p-value = 0.005071
```

# Ajuste del modelo

```
1 modelo_logit <- glm(ansi ~ sexo +  
2                               P1_1 + P1_2 + P1_3 + P1_4 + P1_5 + P1_6 +  
3                               P2_1 + P2_2 + P2_3 + P2_4 + P2_5 + P2_6 + P2_7 + P2_8,  
4                         data = prepara,  
5                         family = binomial)
```

La función `glm()` permite ajustar un modelo lineal generalizado. Se usa `binomial` para modelos de regresión logística binaria donde la función de enlace es de tipo logit, cuando la variable dependiente son recuentos se usa `poisson` y si la variable dependiente es positiva continua entonces se puede usar `gamma`.

# Resultados

```
1 summary(modelo_logit)
```

Call:

```
glm(formula = ansi ~ sexo + P1_1 + P1_2 + P1_3 + P1_4 + P1_5 +  
    P1_6 + P2_1 + P2_2 + P2_3 + P2_4 + P2_5 + P2_6 + P2_7 + P2_8,  
    family = binomial, data = prepara)
```

Coefficients:

	Estimate	Std. Error	z value	Pr(> z )
(Intercept)	-1.17846	0.72108	-1.634	0.102196
sexoMujer	0.31186	0.37322	0.836	0.403382
P1_1	-0.26461	0.14629	-1.809	0.070480 .
P1_2	-0.27997	0.19843	-1.411	0.158253
P1_3	-0.12968	0.20598	-0.630	0.528993
P1_4	-0.37452	0.23396	-1.601	0.109425
P1_5	-0.02688	0.22994	-0.117	0.906931
P1_6	0.55290	0.22809	2.424	0.015350 *
P2_1SÍ	-0.08366	0.48328	-0.173	0.862560
P2_2SÍ	-0.30776	0.48578	-0.634	0.526382

Note que hay múltiples variables que no son significativas en el modelo.

# Modelo reducido

Para llevar a cabo un proceso de eliminación automática de variables que no son relevantes se puede usar

```
1 step(modelo_logit)
```

```
Start: AIC=337.74
ansi ~ sexo + P1_1 + P1_2 + P1_3 + P1_4 + P1_5 + P1_6 + P2_1 +
      P2_2 + P2_3 + P2_4 + P2_5 + P2_6 + P2_7 + P2_8
```

	Df	Deviance	AIC
- P2_6	1	305.75	335.75
- P1_5	1	305.76	335.76
- P2_1	1	305.77	335.77
- P2_3	1	305.84	335.84
- P2_2	1	306.14	336.14
- P1_3	1	306.14	336.14
- P2_4	1	306.22	336.22
- P2_5	1	306.40	336.40
- sexo	1	306.47	336.47
- P1_2	1	307.65	337.65
<none>		305.74	337.74
- P1_4	1	308.39	338.39
- P1_1	1	309.30	339.30

```
Call: glm(formula = ansi ~ P1_1 + P1_2 + P1_4 + P1_6 + P2_7 + P2_8,
          family = binomial, data = prepara) Dispositivas disponibles en GitHub.
```

Coefficients:

(Intercept)	P1_1	P1_2	P1_4	P1_6	P2_7SÍ
-1.3812	-0.2577	-0.3010	-0.4337	0.4938	1.5991
P2_8SÍ					
-0.9809					

Degrees of Freedom: 799 Total (i.e. Null); 793 Residual

Null Deviance: 340.8

Residual Deviance: 308.9 AIC: 322.9

# Modelo final

```
1 var_label(prepara$P2_7) <- var_label(dass$P2_7)
2 var_label(prepara$P2_8) <- var_label(dass$P2_8)
3
4 modelo_final <- glm(ansi ~ P1_1 + P1_2 + P1_4 + P1_6 + P2_7 + P2_8,
5                      data = prepara,
6                      family = binomial)
```

# Resultados del modelo reducido

```
1 tbl_regression(modelo_final, exponentiate = TRUE) |>  
2 bold_labels()
```

Characteristic	OR <sup>1</sup>	95% CI <sup>1</sup>	p-value
<b>Satisfacción con La vivienda</b>	0.77	0.58, 1.01	0.070
<b>Satisfacción con Su trabajo</b>	0.74	0.51, 1.09	0.12
<b>Satisfacción con Sus vecinos</b>	0.65	0.44, 0.96	0.031
<b>Satisfacción con Su familia</b>	1.64	1.12, 2.42	0.011
<b>Ha participado en Prácticas restaurativas para superar situaciones difíciles</b>			
NO	—	—	
SÍ	4.95	2.17, 11.4	<0.001
<b>Ha participado en Entrenamiento para fortalecer habilidades socio-emocionales</b>			
NO	—	—	
SÍ	0.37	0.16, 0.85	0.021

<sup>1</sup>  
OR = Odds Ratio, CI = Confidence Interval

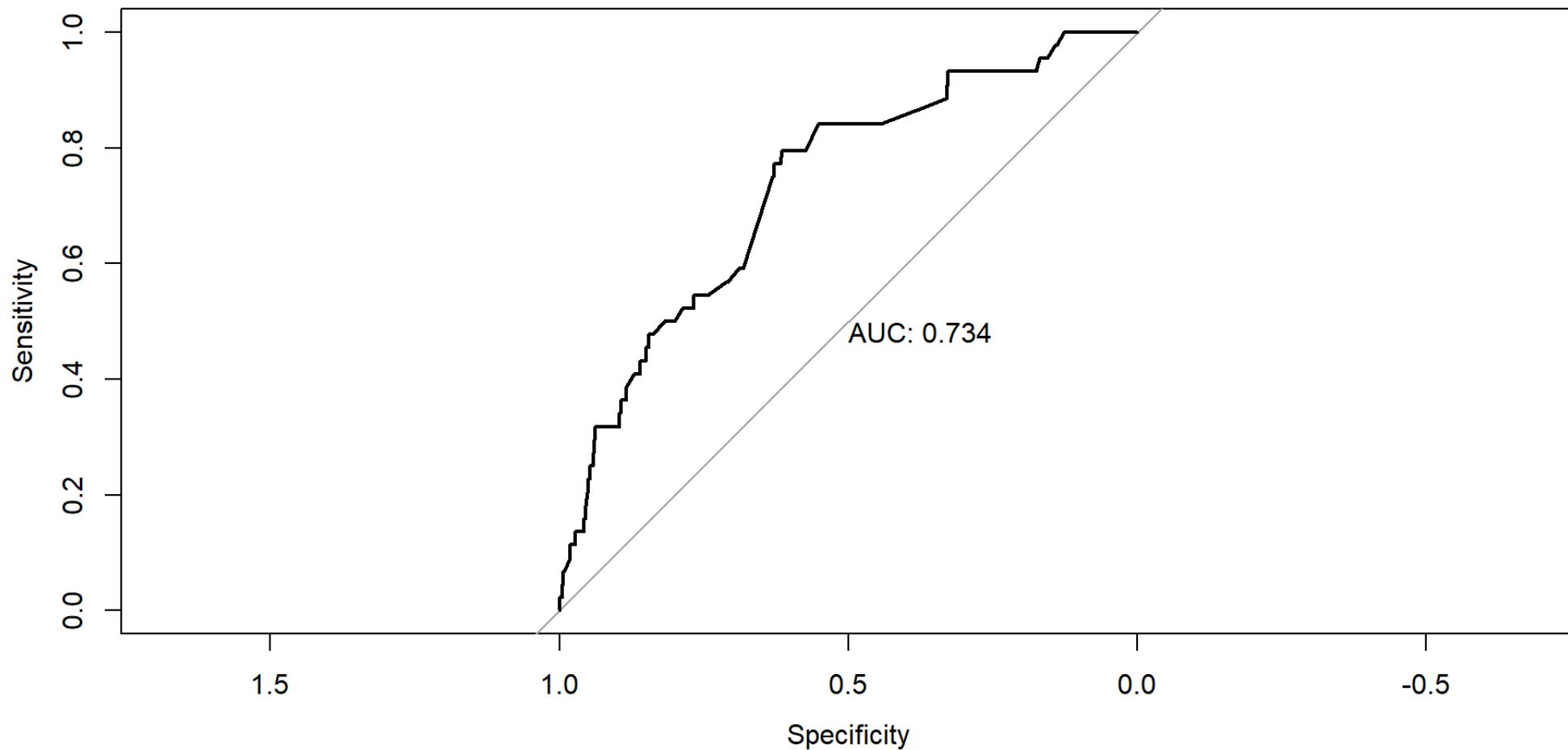
```
1 summary(modelo_final)
```

$$\ln\left(\frac{\pi}{1 - \pi}\right) = -1.38 - 0.26S.Viv - 0.30S.Trab - 0.43S.Vec \\ + 0.49S.Flia + 1.6P.Restau - 0.98Hab.Soc$$

¿Cómo calcularía la probabilidad de que una persona con las siguientes características sufra de ansiedad? Sat.Viv = 2, Sat.Trab = 1, Sat.Vec = 2, Sat.Flia = 3, Parti.Pract.Restau = Si, Entrenamiento SocEmo = No. ¿Qué se esperaría si la persona cambia a un trabajo que en verdad le haga muy feliz (5)?

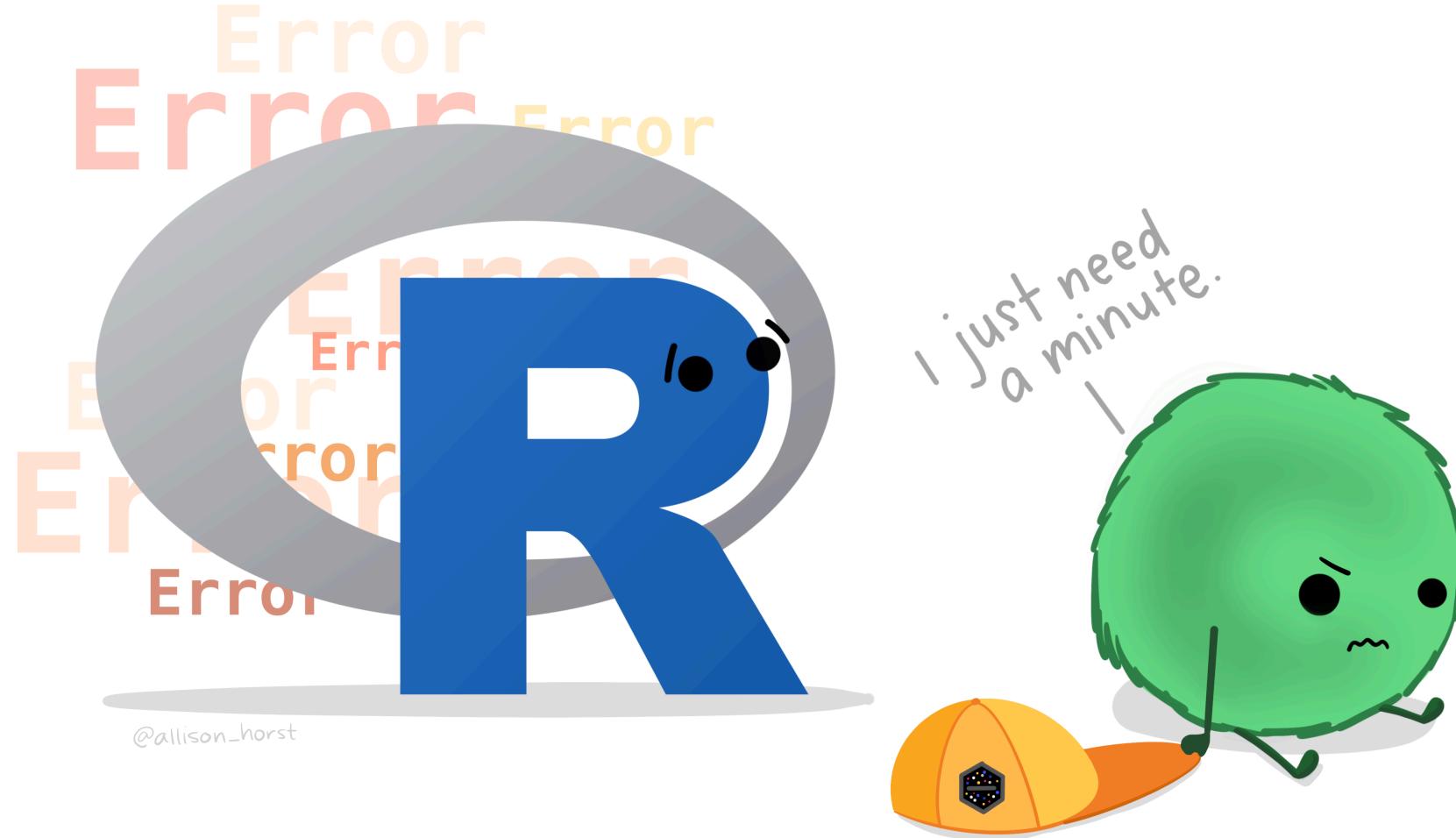
# Curva ROC

```
1 predicciones <- predict(modelo_final, prepara, type = "response")
2 roc_curve <- roc(prepara$ansi, predicciones, plot = TRUE, print.auc = TRUE)
```



# BONUS: IA y Programación

# La etapa de la frustración



Arte de Allison Horst

Diapositivas disponibles en [GitHub](#).

# La IA como herramienta

Estamos en un mundo de constante evolución, ¿la IA nos va a reemplazar?

- Enviar una carta en papel por correo
- Pedir un domicilio por teléfono
- Solicitar un taxi por teléfono
- Orientarse en una ruta con un mapa de papel

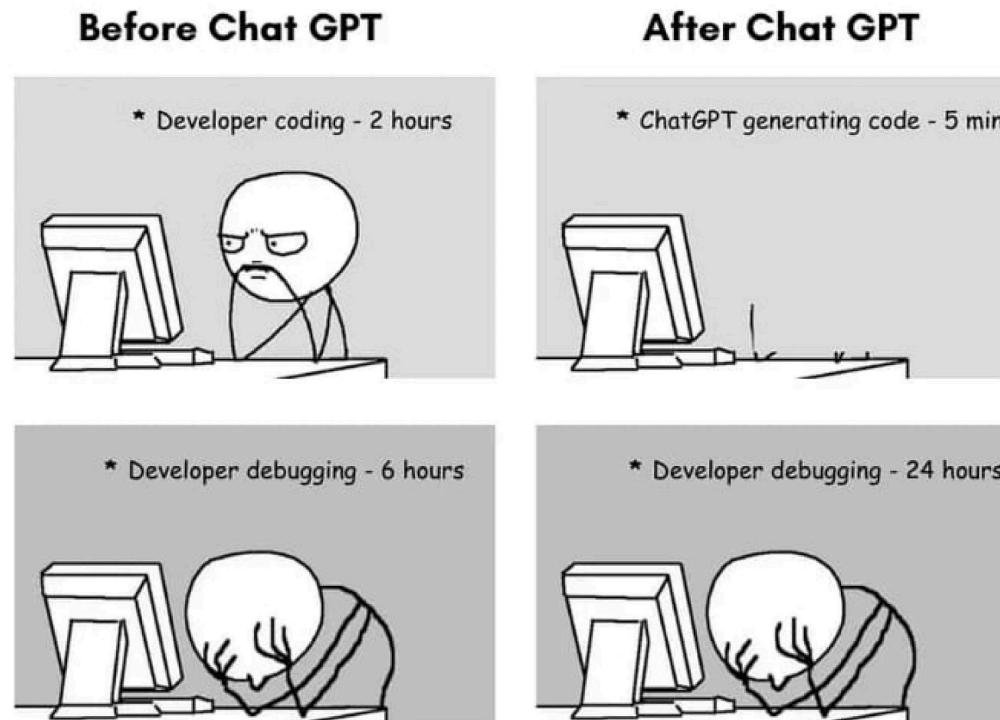


Imagen de Caracol Radio

@tiangolo

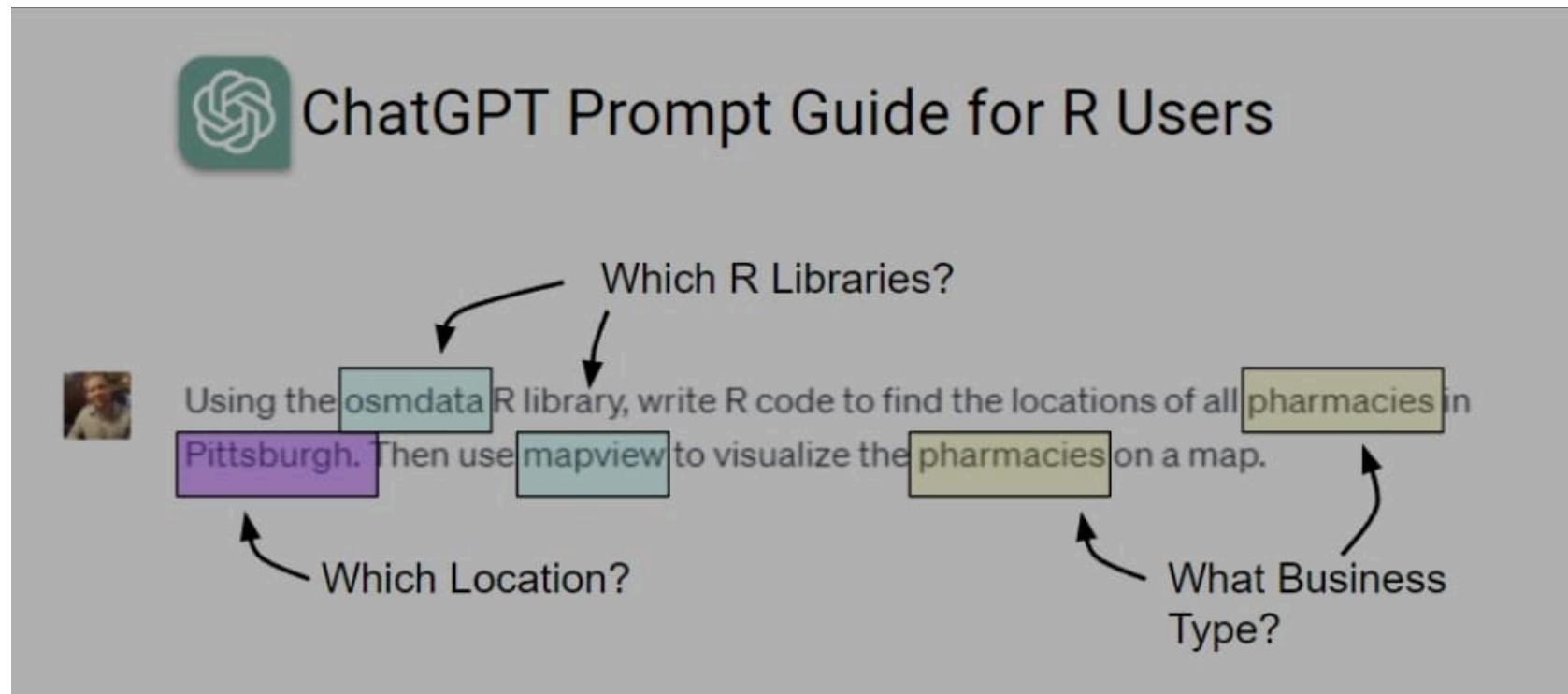
# GPT y Modelos LLMs

El GPT (Generative Pre-trained Transformer) es un modelo de LLM (Large Language Model). Mientras que GPT-3 usaba 175 MM de parámetros usando modelos soportados en texto, GPT-4 usa 100 BN de parámetros usando modelos soportados en texto e imágenes.



# Instrucciones

No pretenda que todo ocurra en un solo paso, a veces se obtienen mejores resultados precisando un *prompt* en cada paso.



@mdancho84

# Herramientas

- <https://rtutor.ai/>
- <https://www.codeconvert.ai/r-to-python-converter>
- ChatGPT
- Copilot
- Gemini
- ...

# Ejercicios

Use un asistente de IA para realizar algunos análisis sobre el conjunto de datos DASS:

- Haga un prompt que solicite crear un gráfico de correlación entre las variables que empiezan con el prefijo “P1\_”, pida que use el paquete *corrplot* de R, solicite que se vea elegante y se presente los valores de las correlaciones.
- Solicite asistencia para generar la gráfica de dispersión del punto 1 del taller.

Haga el Debug de los códigos anteriores, es decir, encontrar errores que pueden impedir que los códigos funcionen.

# GRACIAS!

# Referencias

- Çetinkaya-Rundel, M. and Hardin, J. (2021) Introduction to modern statistics. Sections of Regression modeling: 7, 8, 9 y 10. Disponible aquí: <https://openintro-ims.netlify.app/>
- Hastie, T., Tibshirani, R., Friedman, J. H., & Friedman, J. H. (2009). The elements of statistical learning: data mining, inference, and prediction (Vol. 2, pp. 1-758). New York: springer.
- Ismay, C., & Kim, A.Y. (2019). Statistical Inference via Data Science: A ModernDive into R and the Tidyverse (1st ed.). Chapman and Hall/CRC. <https://doi.org/10.1201/9780367409913>
- Thompson, J. (2019). Tidy Data Science with the tidyverse and tidymodels. <https://tidyds-2021.wjakethompson.com>

