

Analítica de datos aplicada a estudios sobre desarrollo

Giovany Babativa, PhD

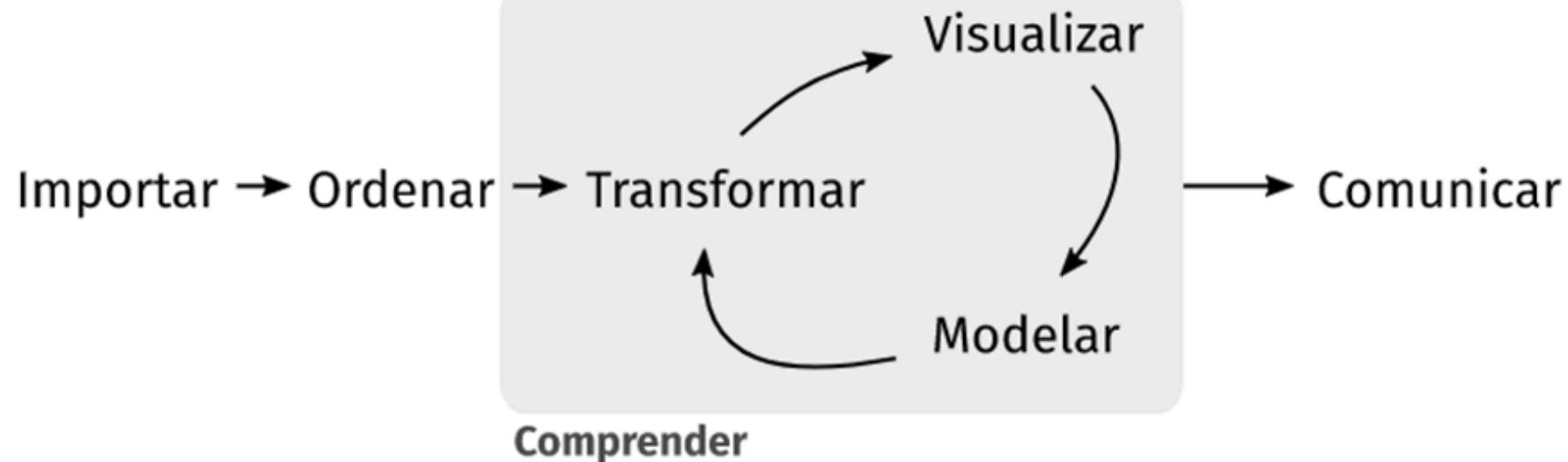
Sobre Mi

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

Proceso de analítica

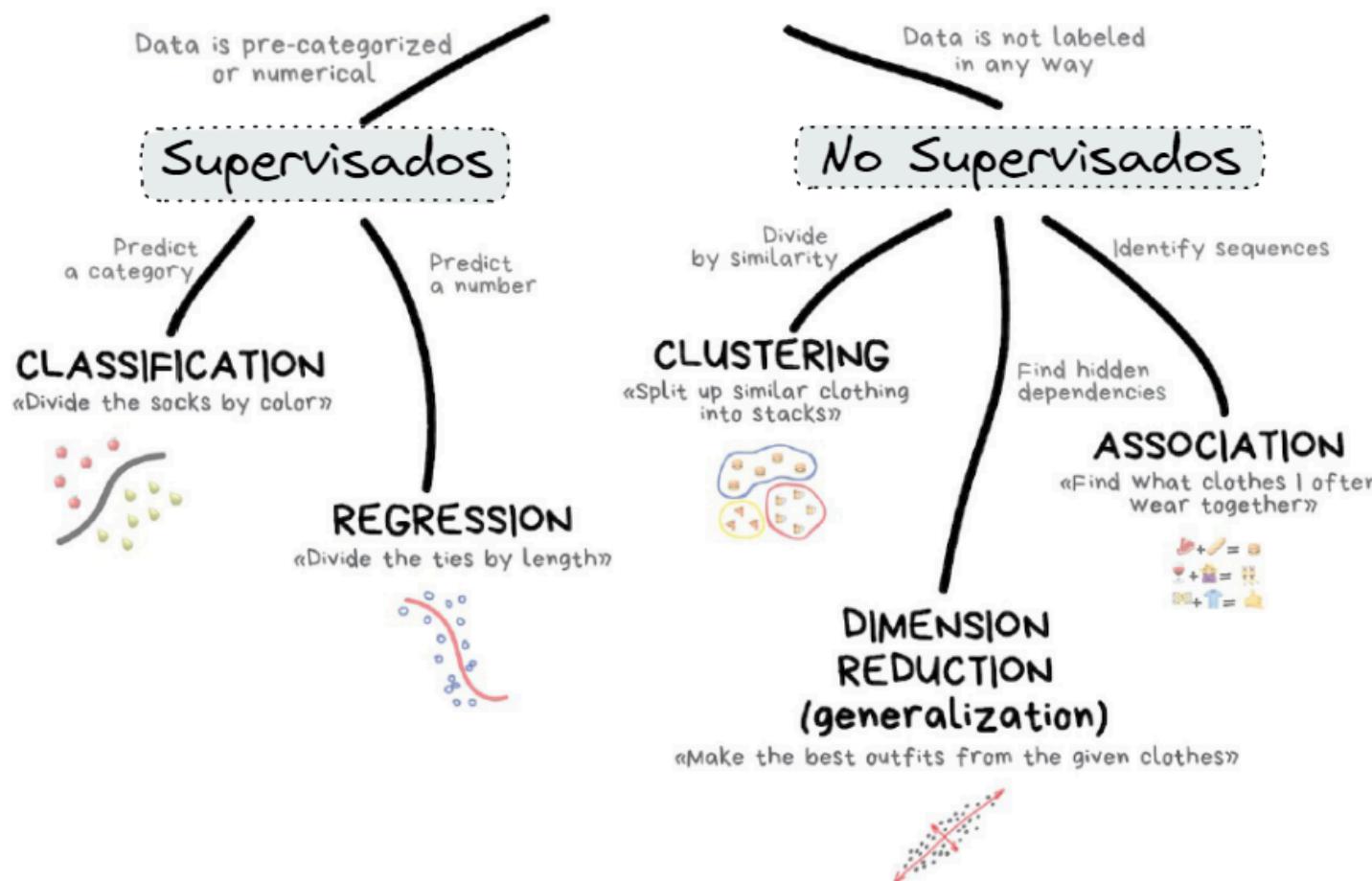


Wickham, H. y otros (2023)

MÉTODOS MULTIVARIANTES

Modelos de analítica

Modelos Multivariantes

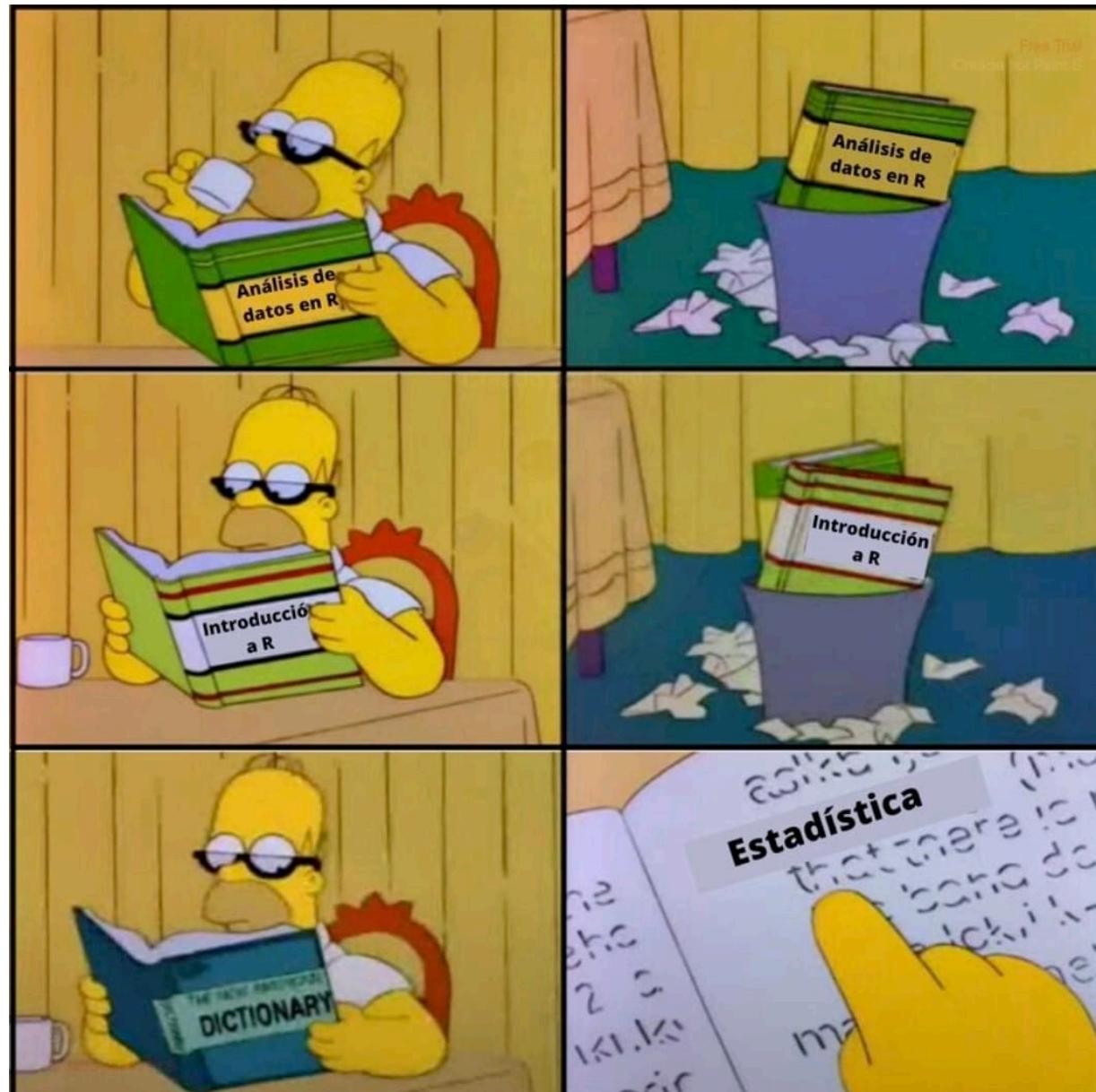


Fuente: Machine Learning for Everyone

Modelando datos

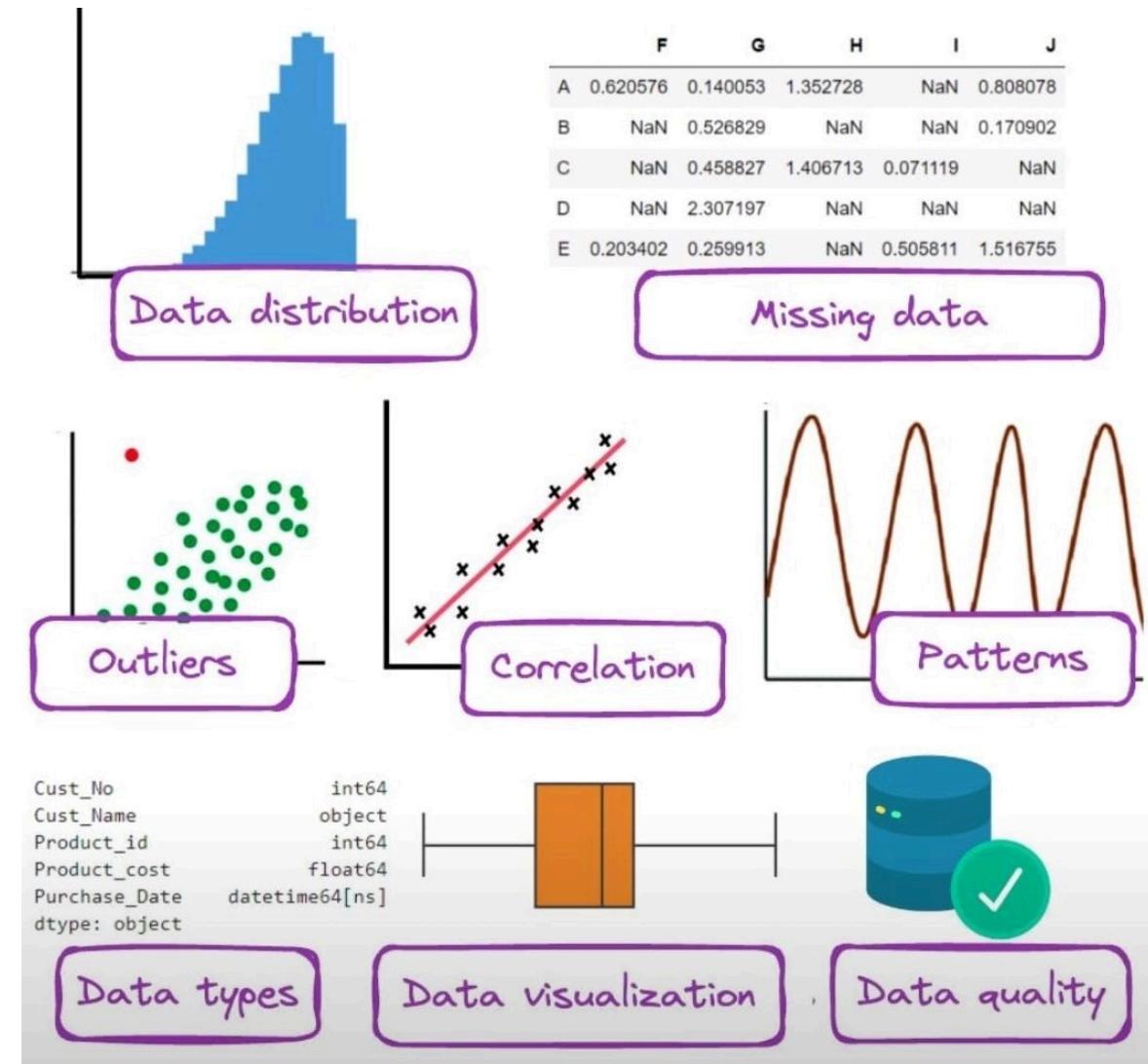


Análisis de datos usando R



Análisis exploratorio

1. Tipos de variables.
2. Visualizar los datos
3. Identificar relaciones
4. Datos atípicos
5. Datos faltantes



Análisis Exploratorio con DataExplorer

Puede realizar una exploración de datos rápido y eficiente, generando resúmenes visuales y estadísticos con muy poco código.

Usemos un conjunto de datos estándar del paquete `modelr`

```
1 library(pacman)
2 p_load(tidyverse, broom, modelr,
3         patchwork, performance, haven,
4         DataExplorer, skimr, corrplot, psych, gt, gtsummary)
5
6 datos <- heights
```

Explore los comandos `create_report()`, `glimpse()` y `skim()`.

```
1 create_report(datos)
2 glimpse(datos)
3 skim(datos)
```

El entorno tidyverse

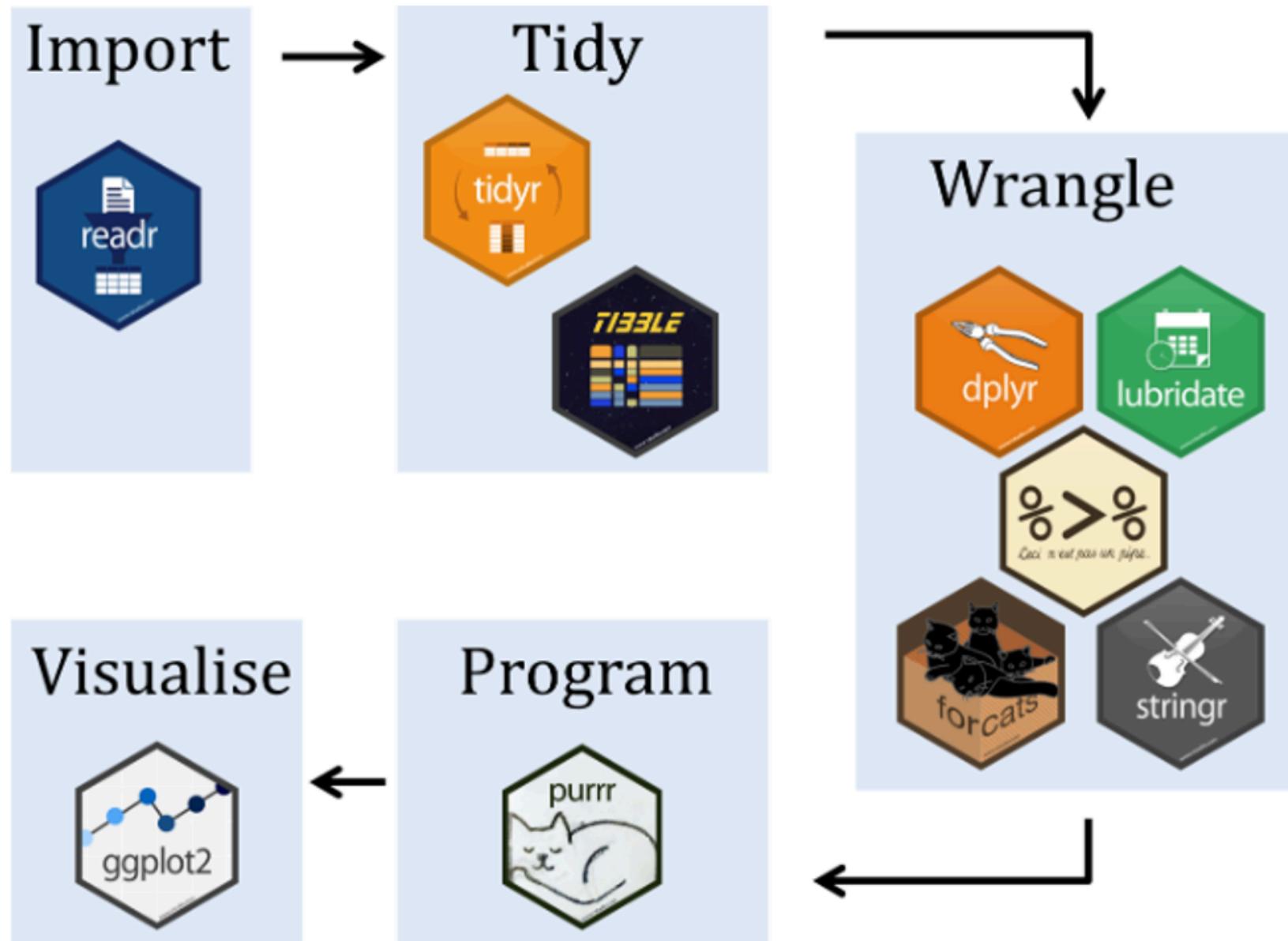


```
library(tidyverse)
```



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

Flujo de trabajo



La gramática de las gráficas

Requiere de al menos 3 elementos:

- Datos
- *aesthetics*: variables.
- geometría

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

Haga un diagrama de dispersión entre la estatura y el peso, el ingreso y la edad, el ingreso y los años de educación por sexo

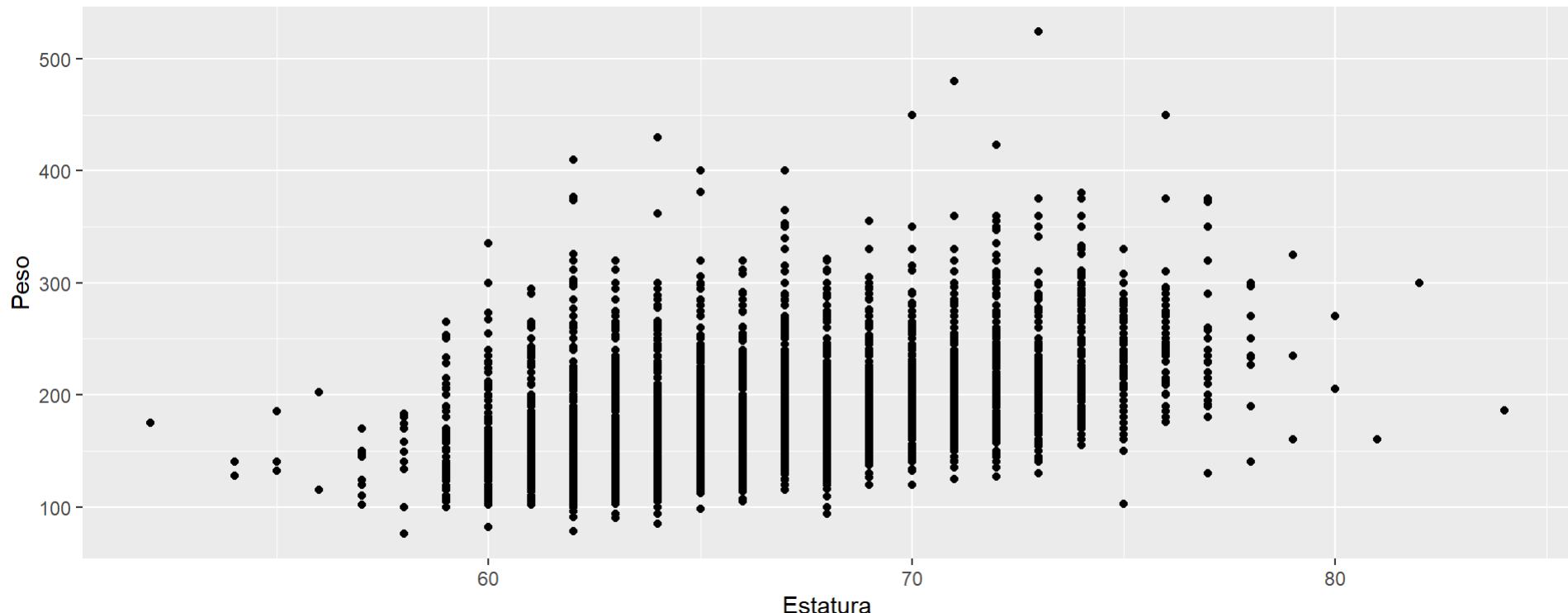
Diagramas de dispersión

Diagrama 1

Diagrama 2

Diagrama 3

```
1 ggplot(data = datos, aes(x = height, y = weight)) +  
2   geom_point() +  
3   labs(x = 'Estatura', y = 'Peso')
```



Correlación Gráfico

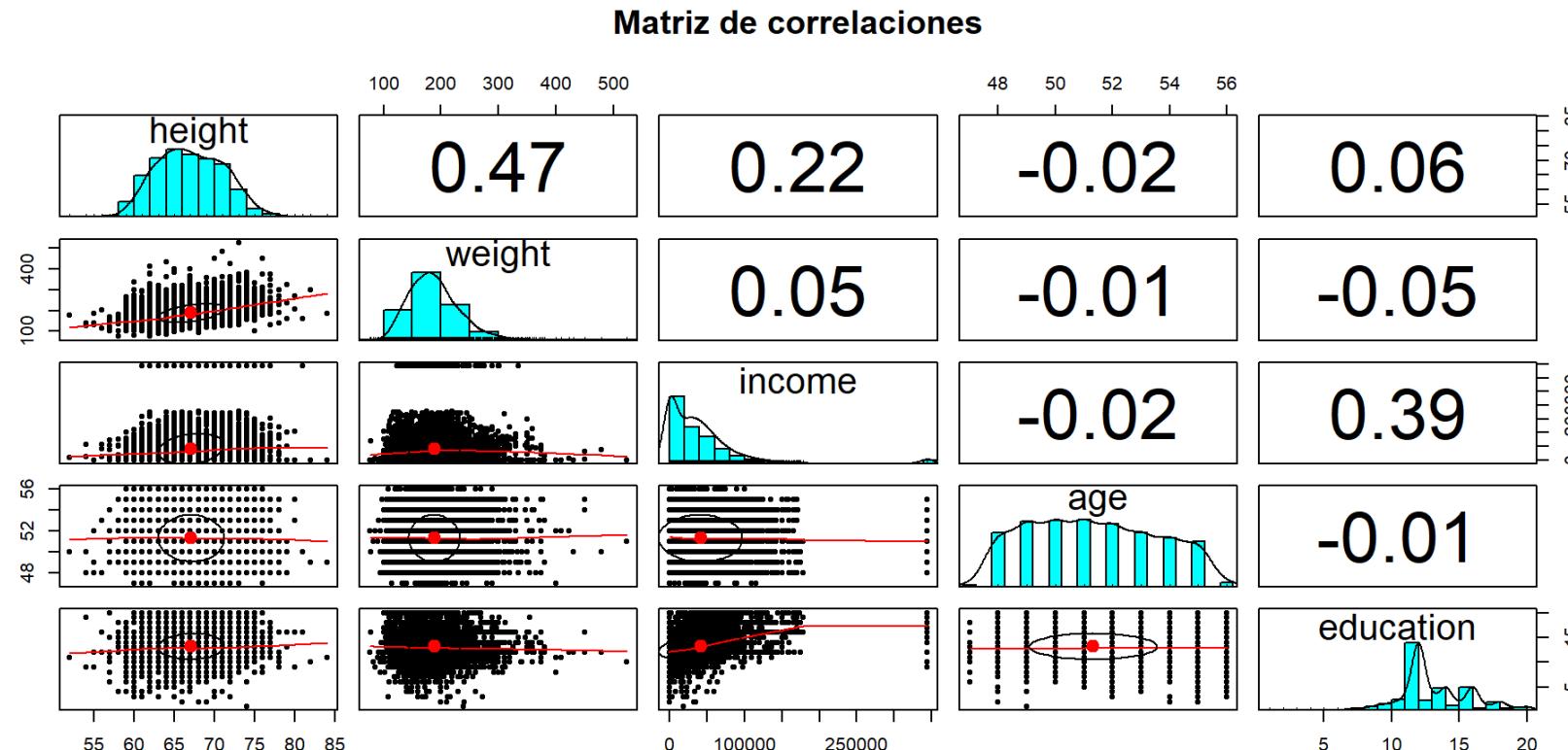
```
1 vars <- datos |> select(height, weight, income, age, education)
2
3 cor(vars, use = "complete")
```

	height	weight	income	age	education
height	1.0000000	0.46819603	0.21795196	-0.018027269	0.064838443
weight	0.46819603	1.00000000	0.05247293	-0.014656703	-0.045021628
income	0.21795196	0.05247293	1.00000000	-0.023703215	0.394288235
age	-0.01802727	-0.01465670	-0.02370322	1.000000000	-0.005998421
education	0.06483844	-0.04502163	0.39428823	-0.005998421	1.000000000

Matriz de correlación

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

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



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(haven)
2
3 dass <- read_sav("datos/DASS21.sav")
```

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

Punto 1 - Taller

1. Grafique el 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. Teniendo en cuenta que las variables sobre la participación en actividades no son cuantitativas, investigue y discuta sobre la forma en que podría identificarse alguna asociación con la depresión.

ANÁLISIS DE REGRESIÓN

Conceptos básicos

Denominar a $\mathbf{X} = (\mathbf{x}_1, \dots, \mathbf{x}_p)$ como variables explicativas o predictoras se debe a uno de dos propósitos:

- **Modelo para explicar las relaciones:** busca describir y cuantificar explícitamente la relación entre la variable de resultado y y un conjunto de **variables explicativas \mathbf{X}** , así como determinar la importancia de cualquier relación.
- **Modelo para la predicción:** busca predecir una variable de resultado y basado en la información contenida en un conjunto de **variables predictivas \mathbf{X}** . Acá no necesariamente importa comprender cómo se relacionan e interactúan todas las variables entre sí, sólo lograr buenas predicciones sobre y utilizando la información en \mathbf{X} .

¿Qué objetivo se persigue?

En cada caso indique si el objetivo del modelo debe ser explicativo o predictivo. Suponga que tenemos interés en identificar:

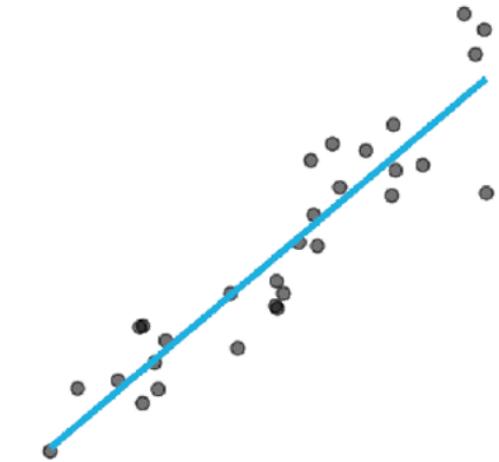
1. ¿Cuáles son los factores de riesgo (como el hábito de fumar, la edad, etc) que se asocian con el cáncer de pulmón?.
2. ¿Qué contenido que le gustaría ver con mayor probabilidad a un usuario de una plataforma digital?.
3. ¿Cuál es el efecto de la edad de la mujer, nivel educativo de la pareja, hábitos de la pareja (fuma, toma, etc), composición del hogar (con hijos/sin hijos) sobre la violencia doméstica?.

Especificación del modelo

...



Datos



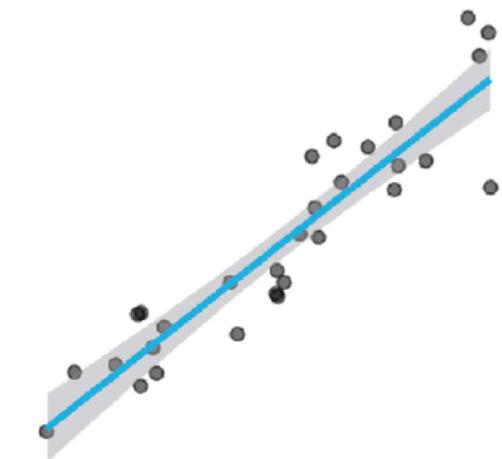
Función de enlace

¿Es un buen modelo?

...



Datos



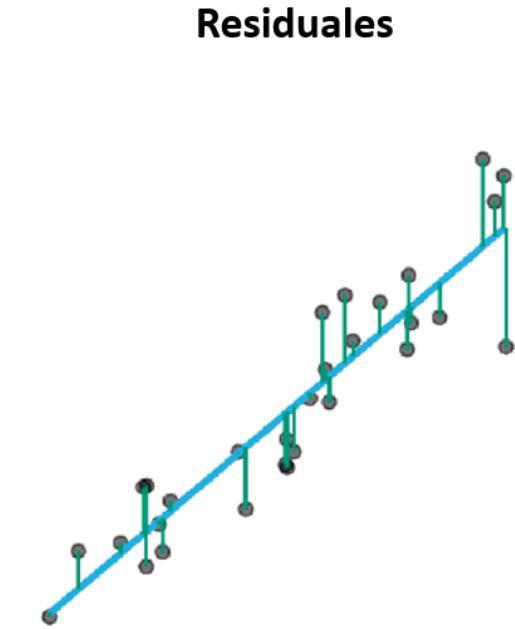
Función de enlace

Análisis de los residuales

...



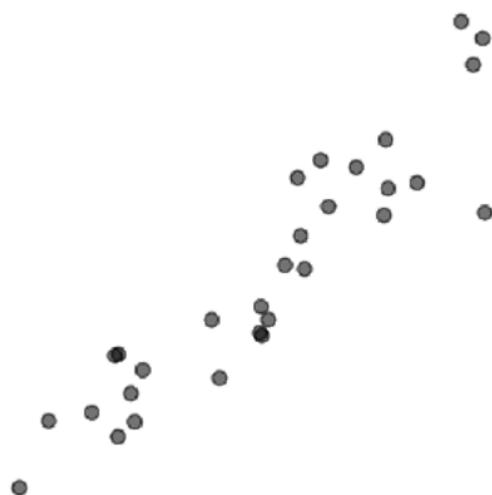
Datos



Función de enlace

Pronósticos

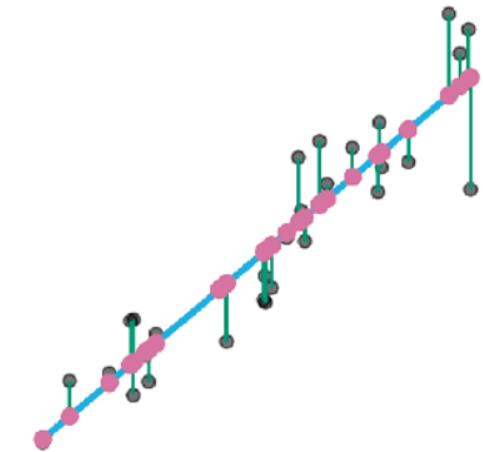
...



Datos



Predicciones



Función de enlace

Algoritmos

Algunos modelos son:

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

Formulación del modelo

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:

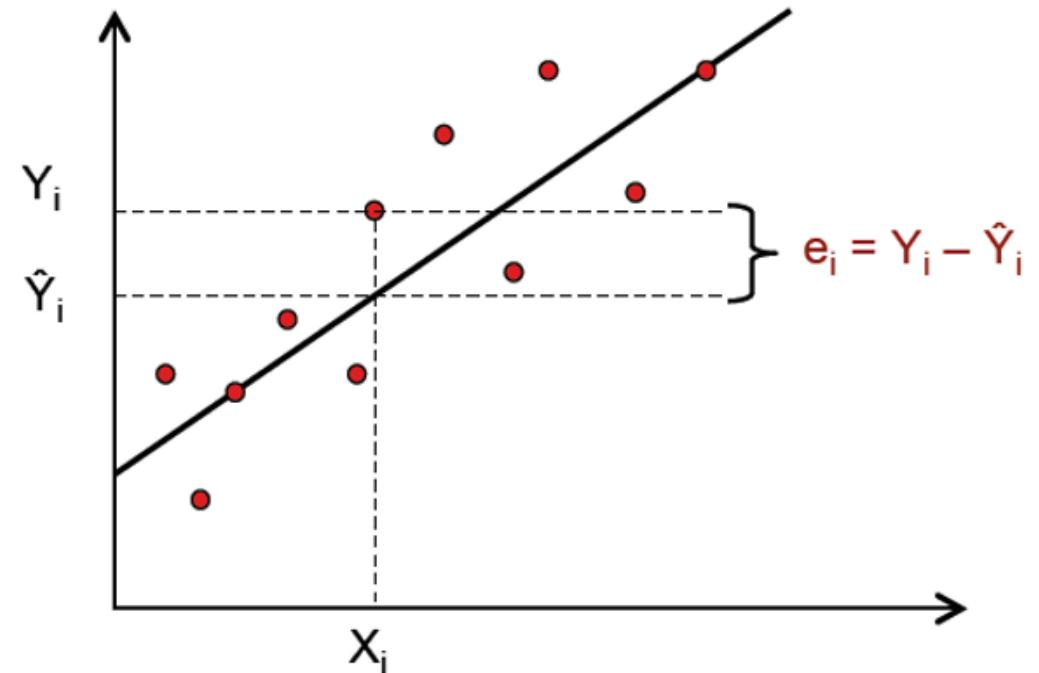
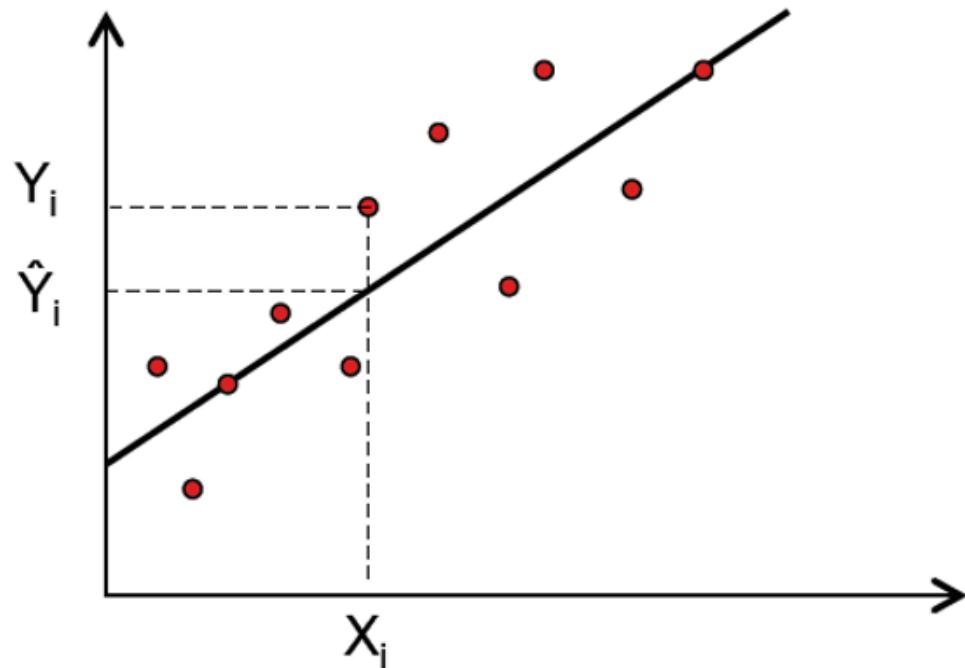
$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \varepsilon_i \text{ con } \varepsilon_i \stackrel{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma^2)$$

Supuestos del modelo

- **Linealidad:** $\mu = E(y|\mathbf{x}_i) = \mathbf{X}\beta$
- **Independencia:** $Cov(\varepsilon_i, \varepsilon_j) = 0$ para $i \neq j$
- **Homocedasticidad:** $V(\varepsilon_i|\mathbf{X}) = \sigma^2$
- **Normalidad:** $\varepsilon_i \stackrel{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma^2)$

Estimación de los parámetros

- Puntos: Valores observados
- Recta: Valores ajustados



Estimación de los parámetros

$$Y_i = \hat{Y}_i + (Y_i - \hat{Y}_i) = \hat{Y}_i + e_i$$

El objetivo entonces es minimizar

$$\sum_{i=1}^n e_i^2 = \sum_{i=1}^n (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^n (Y_i - [\beta_0 + \beta_1 X_i])^2$$

El procedimiento se conoce como Mínimos Cuadrados Ordinarios (MCO).

EJEMPLO

Variable explicativa cuantitativa

Considere nuevamente los datos del paquete `modelr`

```
1 modelo1 <- lm(income ~ education, data = datos)
```

Teniendo en cuenta que el ingreso está medido en dólares al año:

- Escriba la ecuación del modelo
- ¿Cuál debe ser la interpretación de los coeficientes de regresión?
- ¿Cómo debe interpretarse el p-value?

```
1 library(gtsummary)
2 tbl_regression(modelo1, intercept = TRUE)
```

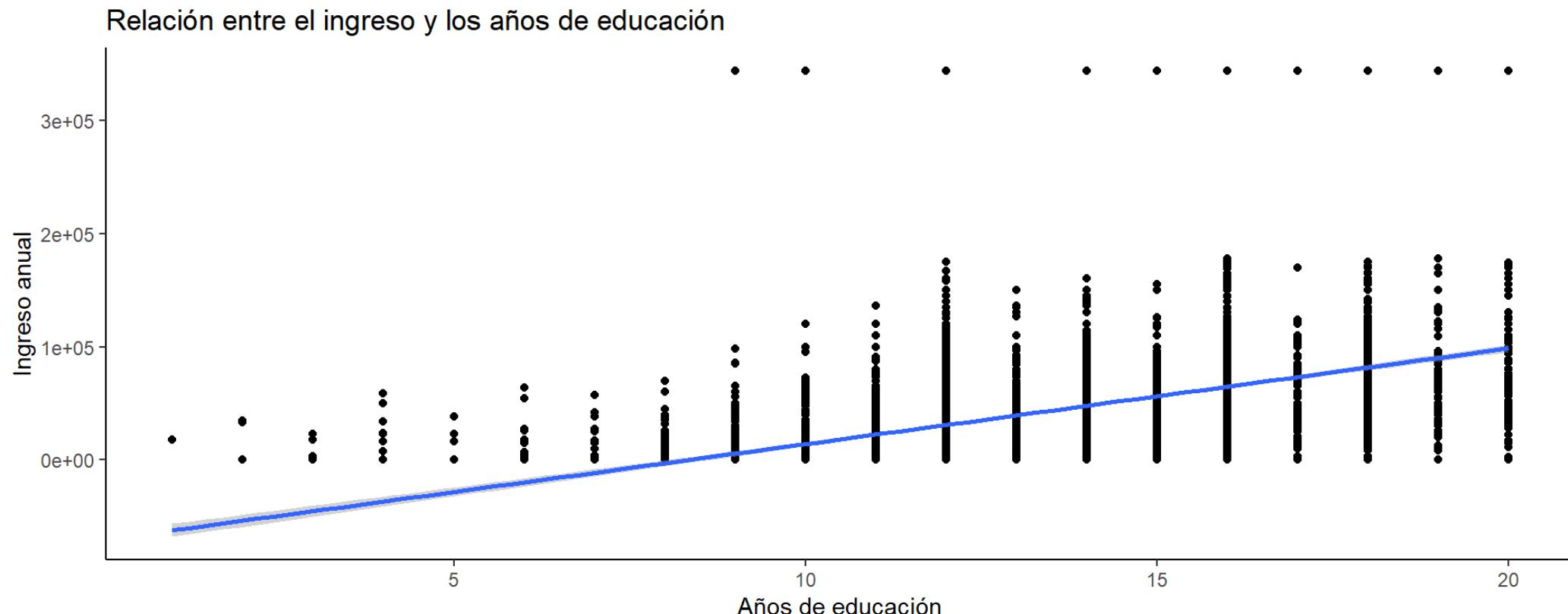
Characteristic	Beta	95% CI	p-value
(Intercept)	-70,571	-76,815, -64,327	<0.001
education	8,459	7,996, 8,923	<0.001

CI = Confidence Interval

Diapositivas disponibles en [GitHub](#).

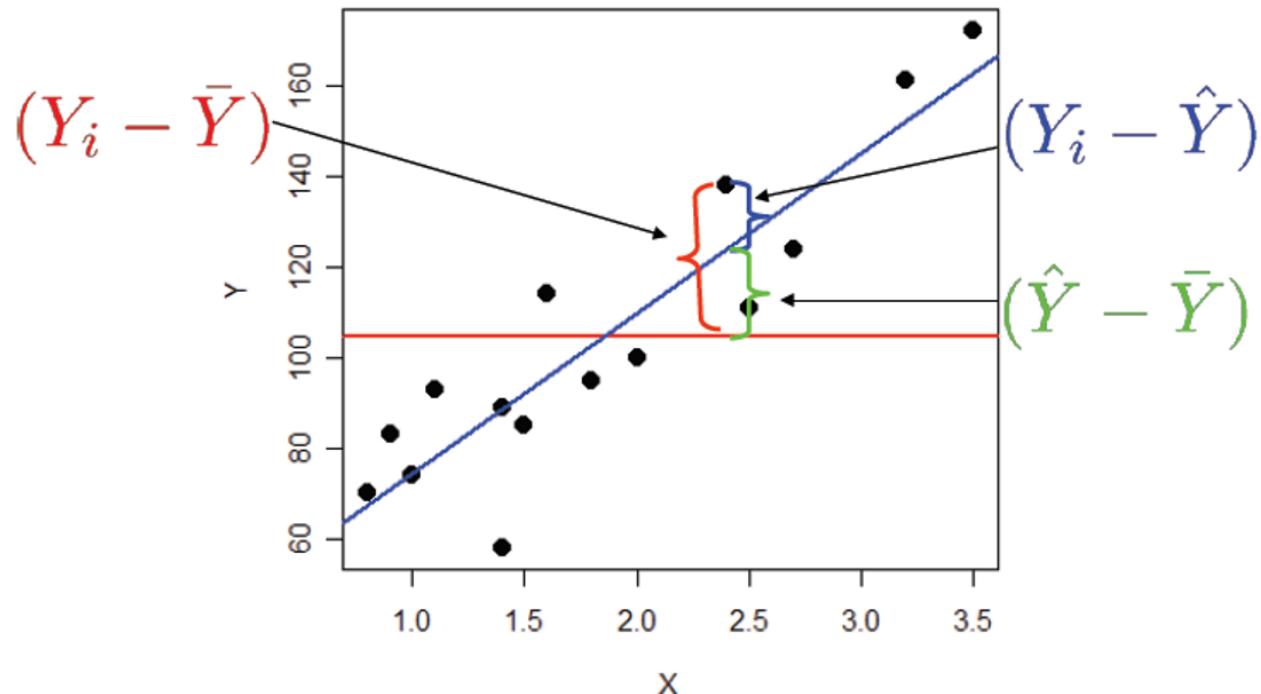
Visualización

```
1 ggplot(datos, aes(x = education, y = income)) +  
2   geom_point() +  
3   labs(x = "Años de educación", y = "Ingreso anual",  
4         title = "Relación entre el ingreso y los años de educación") +  
5   geom_smooth(method = "lm", formula = 'y ~ x') +  
6   theme_classic()
```



¿Es bueno el modelo?

Descomposición de la varianza



$$\sum_{i=1}^n (Y_i - \bar{Y}_i)^2 = \sum_{i=1}^n (Y_i - \hat{Y}_i)^2 + \sum_{i=1}^n e^2$$

Bondad del ajuste

El coeficiente de determinación es un indicador entre 0 y 1:

$$\sum_{i=1}^n (Y_i - \bar{Y}_i)^2 = \sum_{i=1}^n (Y_i - \hat{Y})^2 + \sum_{i=1}^n e^2$$

$$SCT = SCR + SCE$$

Se deduce que:

$$R^2 = \frac{SCR}{SCT} = 1 - \frac{SCE}{SCT}$$

El valor de R^2 está entre 0 y 1.

Bondad del ajuste

```
1 tab <- modelo1 |> glance()
2 gt::gt(tab)
```

r.squared	adj.r.squared	sigma	statistic	p.value	df	logLik	AIC	BIC	deviance	df.residual	nobs
0.1547128	0.1545919	51407.62	1280.111	1.313761e-257	1	-85815.3	171636.6	171657.1	1.848335e+13	6994	6996

- ¿cuál debe ser la interpretación del R^2 ?
- ¿cómo debe interpretarse el p-value?

```
1 tbl_regression(modelo1, intercept = TRUE) |>
2   add_glance_table(include = c(r.squared, p.value))
```

Characteristic	Beta	95% CI		p-value
		1	2	
(Intercept)	-70,571	-76,815,	-64,327	<0.001
education	8,459	7,996,	8,923	<0.001
R^2	0.155			
p-value		<0.001		

Variable explicativa categórica

Podemos preguntarnos si el ingreso depende del sexo de la persona.
Para ello es clave que la variable categórica sea de clase **factor**.

Revisión de la clase de la variable del sexo

```
1 class(datos$sex)  
[1] "factor"  
  
1 levels(datos$sex)  
[1] "male"    "female"  
  
1 datos$sex <- relevel(datos$sex, ref = "male")
```

El ajuste del modelo no cambia:

```
1 modelo2 <- lm(income ~ sex, data = datos)
```

Resultados

```
1 tbl_regression(modelo2, intercept = TRUE) |>  
2 add_glance_table(include = c(r.squared, p.value))
```

Characteristic	Beta	95% CI	p-value
(Intercept)	53,510	51,675, 55,345	<0.001
sex			
male	—	—	
female	-23,922	-26,481, -21,364	<0.001
R ²	0.046		
p-value	<0.001		

¹
CI = Confidence Interval

Punto 2 - Taller: DASS21

1. Ajuste los modelos de regresión simple para interpretar las relaciones entre:

- Depresión y Satisfacción con su trabajo
- Ansiedad y Satisfacción con su trabajo
- Estrés y Satisfacción con su trabajo

2. Interprete los coeficientes de los modelos ajustados y discuta la bondad del ajuste.

Punto 3 - Taller: DASS21

Ajuste un modelo de regresión simple que le permita identificar si el puntaje de depresión se relaciona con el sexo

1. Convierta la variable `sexo` en factor así: `dass$sexo <- as_factor(dass$sexo)`
2. Ajuste el modelo de regresión y presente los resultados.
3. Interprete los coeficientes y el valor p.

Variables explicativas mixtas

Ajuste un modelo para el ingreso en función de las variables de años de educación, sexo y estado civil.

$$Ingreso_i = f(Educa, Sexo, Est. Civil) + \varepsilon_i$$

Observe que las variables explicativas son cuantitativas y cualitativas. Verifique que la clase esté bien definida.

```
1 modelo3 <- lm(income ~ education + sex + marital, data = datos)
```

Escriba la ecuación del modelo e interprete los resultados.

Resultados

```
1 tbl_regression(modelo3, intercept = TRUE) |>  
2 add_glance_table(include = c(r.squared, p.value))
```

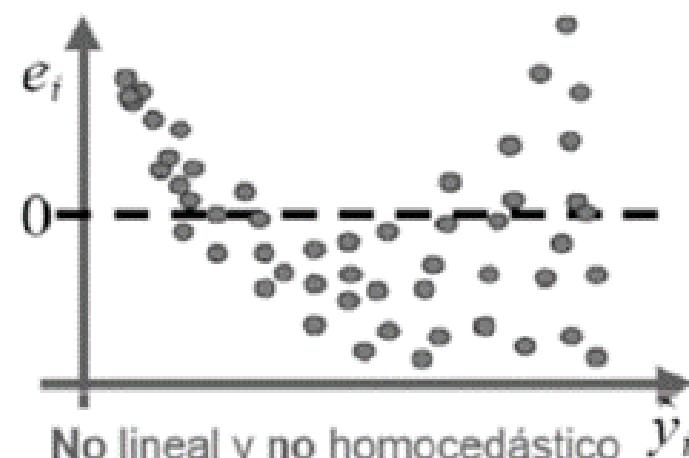
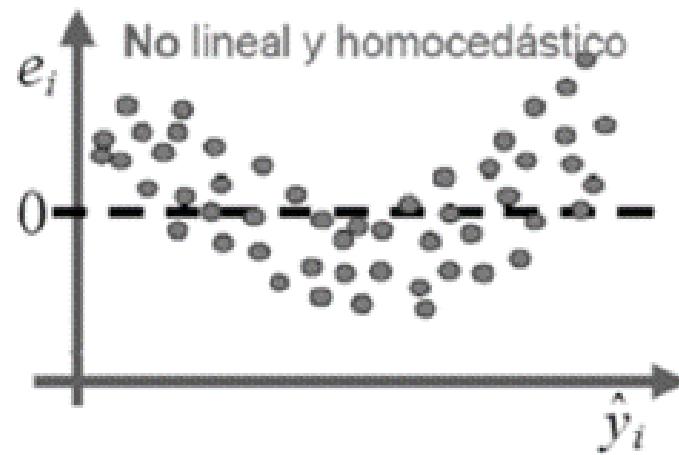
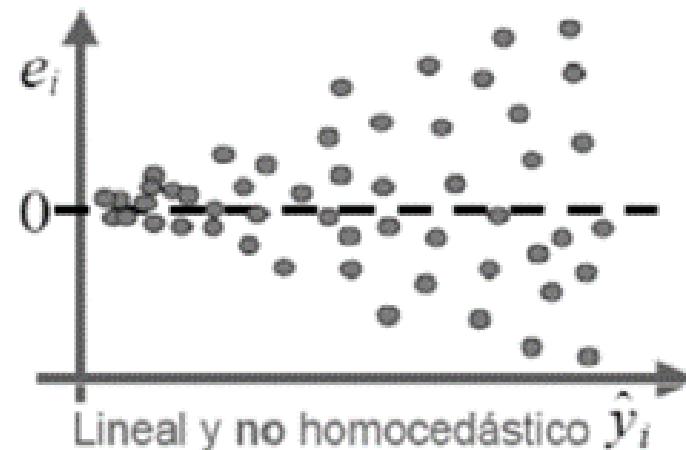
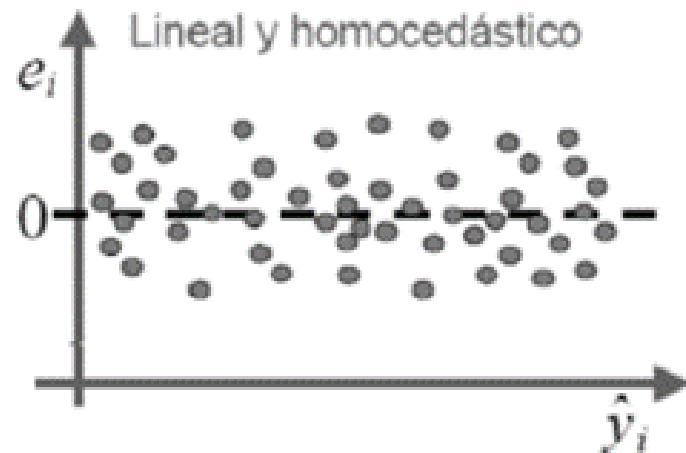
Characteristic	Beta	95% CI		p-value
		1	7	
(Intercept)	-67,047	-73,482,	-60,611	<0.001
education	8,305	7,854,	8,756	<0.001
sex				
male	—	—	—	
female	-26,580	-28,906,	-24,254	<0.001
marital				
single	—	—	—	
married	18,574	15,277,	21,872	<0.001
separated	2,500	-3,312,	8,311	0.4
divorced	7,864	4,080,	11,649	<0.001
widowed	9,967	1,784,	18,151	0.017
R ²	0.229			

1

CI = Confidence Interval
[Diapositivas disponibles en GitHub.](#)

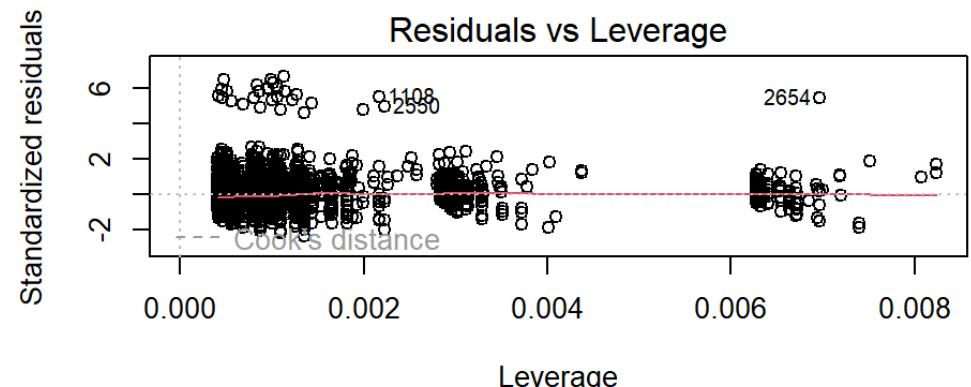
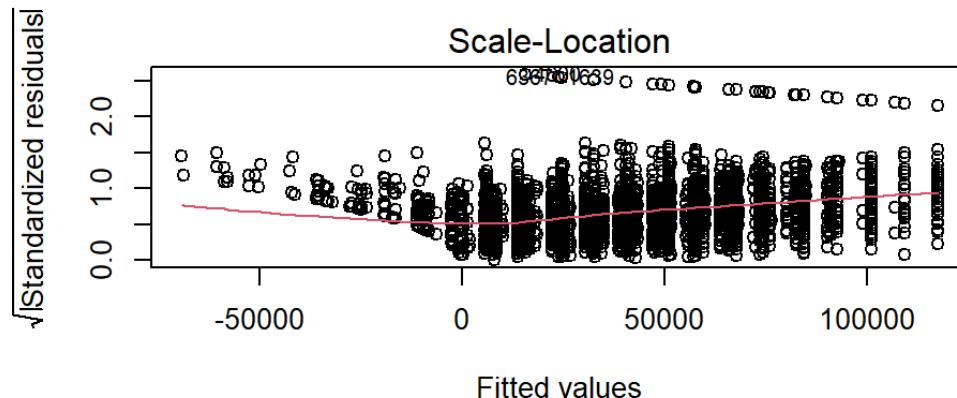
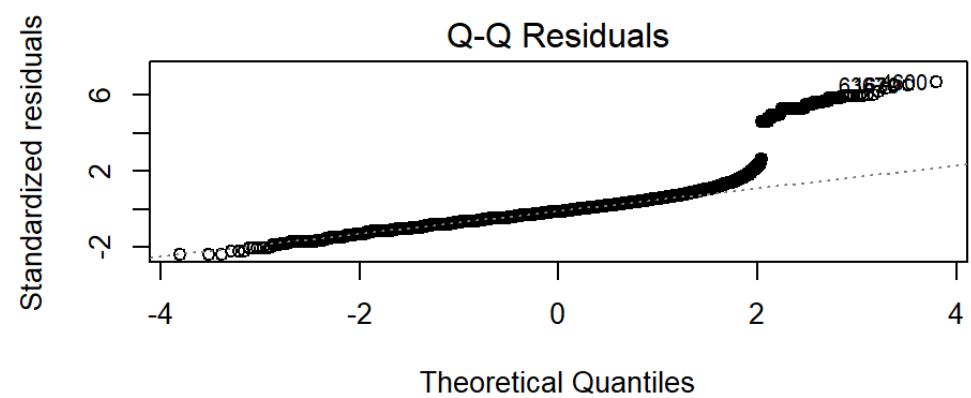
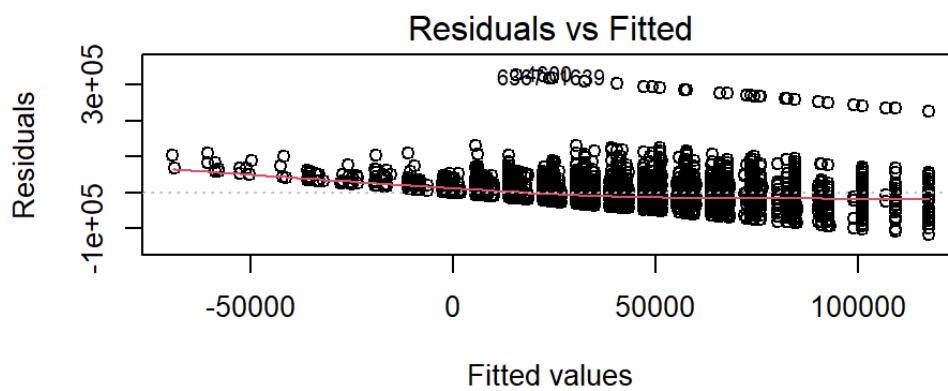
Análisis de los supuestos

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(modelo3)
```



Análisis de los supuestos

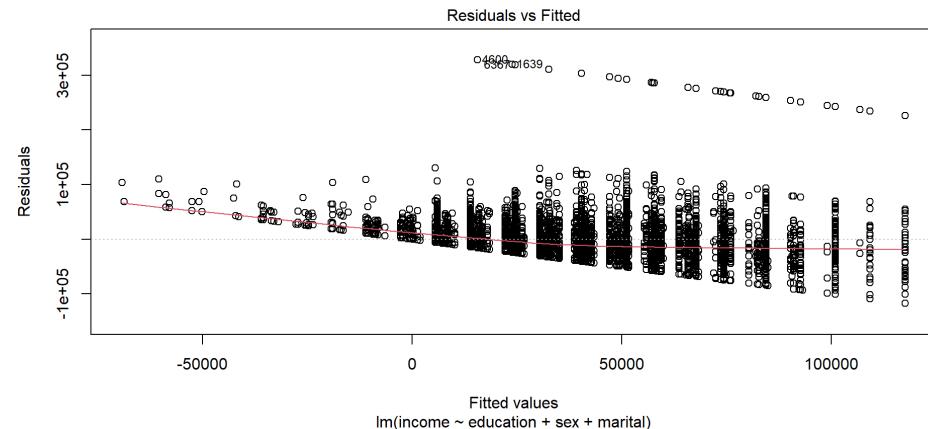
Linealidad

Homogeneidad

Normalidad

Influyentes

```
1 plot(modelo3, 1)
```



No se debe presentar un patrón, así que la línea roja debe estar aproximadamente de forma horizontal en cero.

Análisis de los supuestos

```
1 ajuste <- augment(modelo3)
2
3 ajuste |>
4   slice_max(.cooksdi, n=3)

# A tibble: 5 × 11
  .rownames income education sex marital .fitted .resid     .hat .sigma .cooksdi
  <chr>      <int>     <int> <fct> <fct>    <dbl>   <dbl>    <dbl>   <dbl>    <dbl>
1 2654       343830        16 male  widowed  75805. 2.68e5 0.00696 49015. 0.0300
2 1108       343830        20 fema... single  72478. 2.71e5 0.00217 49013. 0.00949
3 2550       343830        20 male  single   99058. 2.45e5 0.00222 49033. 0.00792
4 4973       343830        20 male  single   99058. 2.45e5 0.00222 49033. 0.00792
5 5060       343830        20 male  single   99058. 2.45e5 0.00222 49033. 0.00792
# i 1 more variable: .std.resid <dbl>
```

Revisar que no existan distancias mayores que 1.

Paquete performance

Revise los supuestos del modelo, verifique que VIF es menor que 10 (no hay multicolinealidad) y haga las pruebas de los supuestos

```
1 check_model (modelo3)
2 check_collinearity (modelo3)
3 check_outliers (modelo3)
4 check_autocorrelation (modelo3)
5 check_heteroscedasticity (modelo3)
6 check_normality (modelo3)
```

Punto 4 - Taller: DASS 21

1. Ajuste un modelo de regresión lineal múltiple con al menos 3 variables explicativas que resulten significativas para modelar el puntaje de depresión. Escriba la ecuación, interprete los coeficientes, revise los supuestos y concluya.
2. Ajuste un modelo de regresión lineal múltiple con al menos 3 variables explicativas que resulten significativas para modelar el puntaje de estrés. Escriba la ecuación, interprete los coeficientes, revise los supuestos y concluya.

TALLER

Reglas

- Se deben realizar los 4 ejercicios propuestos a lo largo de la sesión.
- El trabajo se debe realizar con los grupos que han conformado.
- Fecha de entrega: 09 de octubre.
- Si existen dos trabajos exactamente iguales, la nota será dividida entre el número de personas involucradas.

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/>
- 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>

