Sesión 8

Curso: R Aplicado a los Proyectos de Investigación

Percy Soto-Becerra, M.D., M.Sc(c)

InkaStats Data Science Solutions | Medical Branch

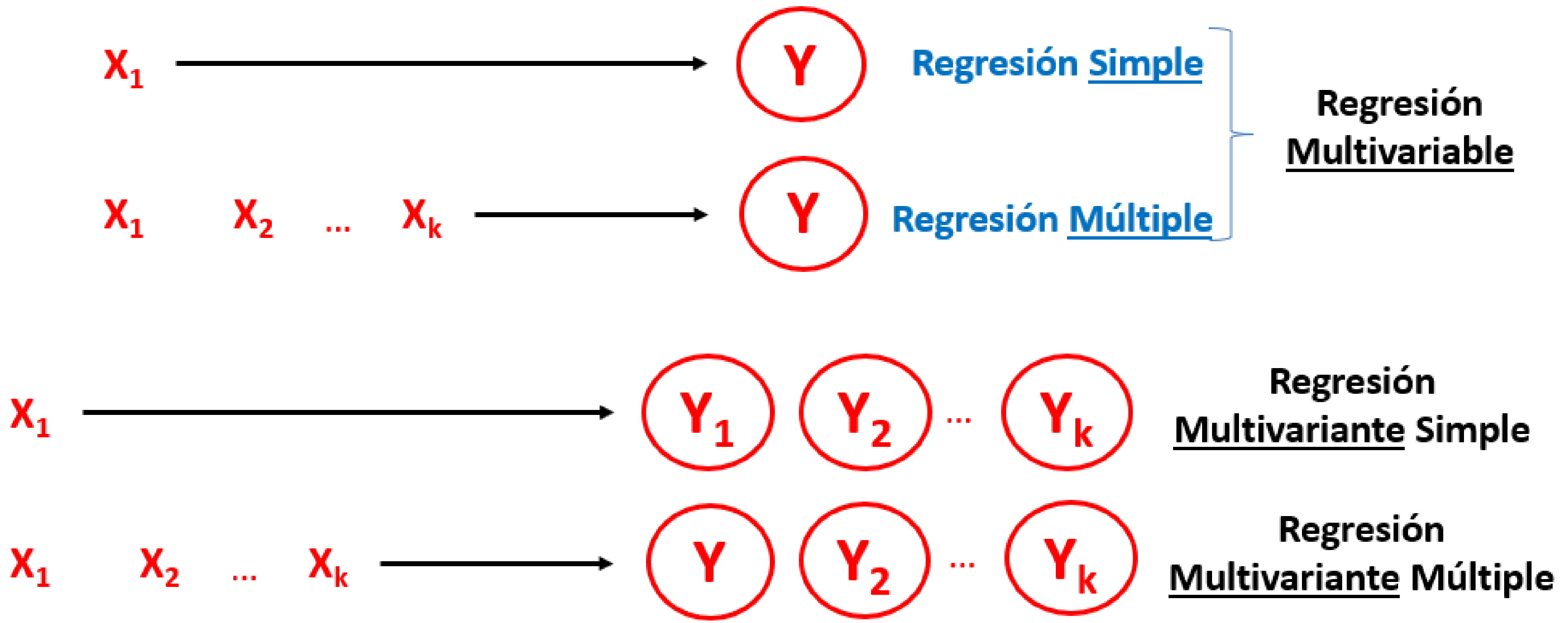
2022-10-19

https://github.com/psotob91

Agenda

1. **Introducción al modelado de regresión**
2. Modelo de Regresión Lineal
3. El Modelo Lineal Generalizado
4. La regresión (log) Poisson
5. Tablas de regresión reproducibles con {gtsummary}

- Conjunto de técnicas estadísticas para estimar la relación entre variables.



Modelos de regresión multivariable

- Modelan variable dependiente en función a variable independiente.
- Desenlace define el tipo de regression.

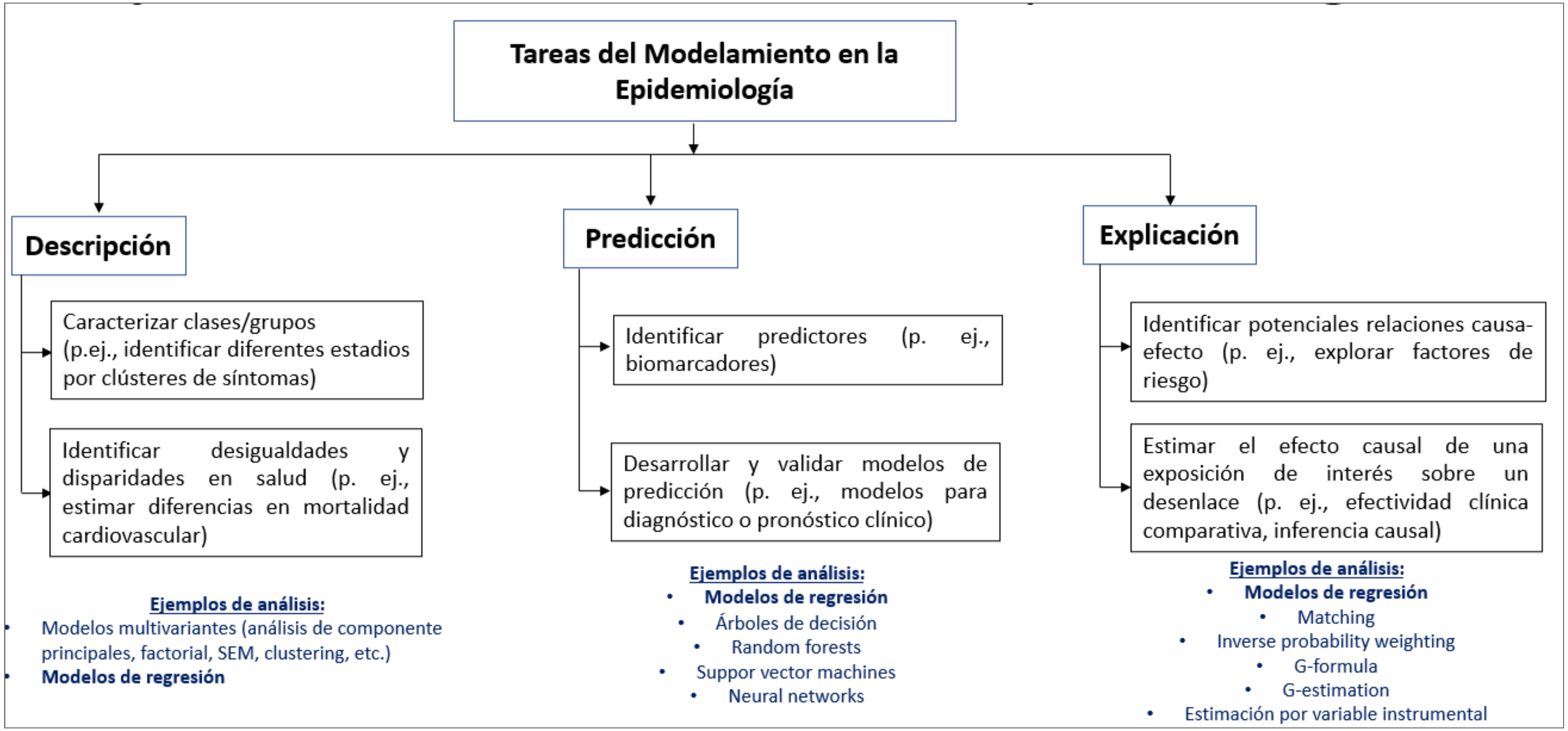
$$X \longrightarrow Y$$

Y ~ continua	→ Regresión lineal
Y ~ Bernoulli (dicotómica)	→ Regresión logística
Y ~ Poisson (conteo)	→ Regresión de Poisson
Y ~ Weibull (tiempo a evento)	→ Regresión de Cox

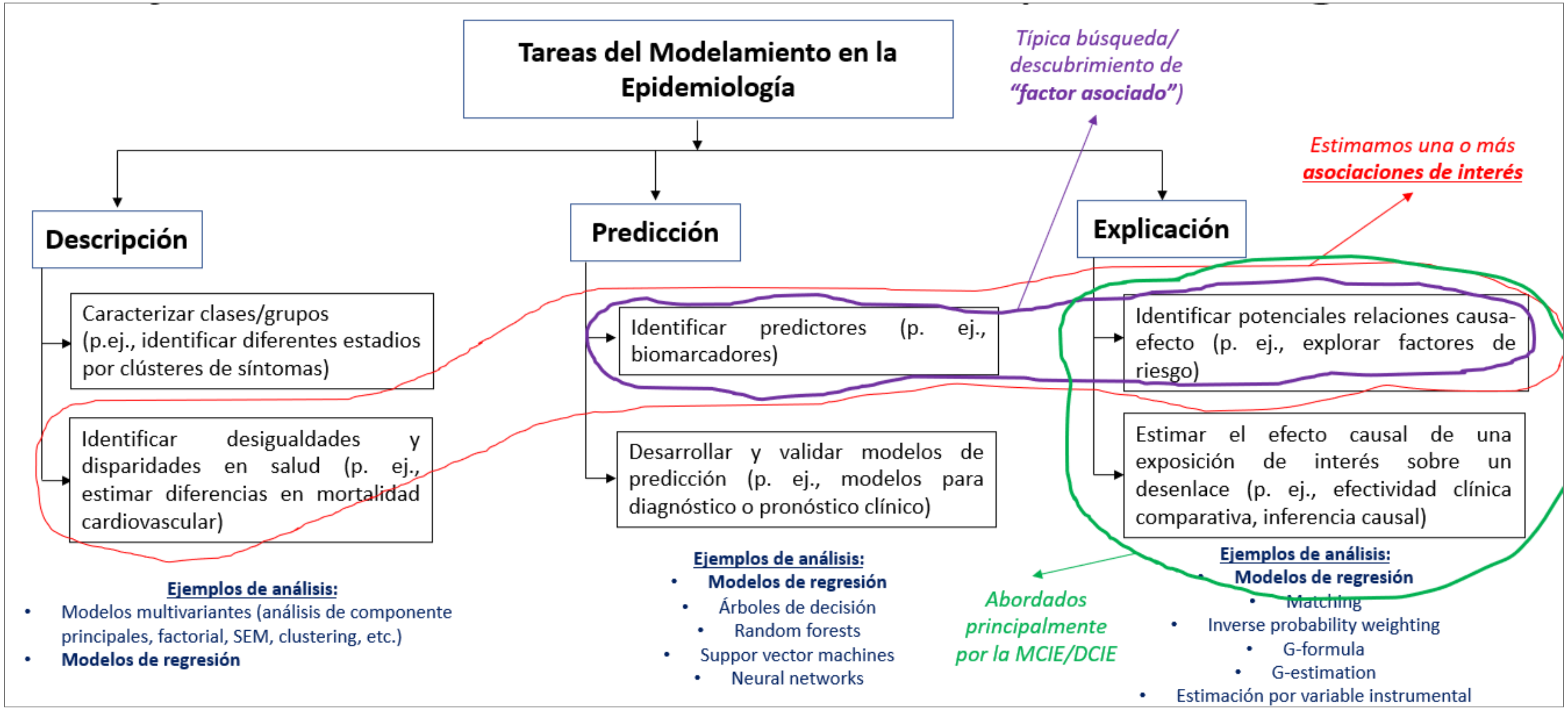
¿Para qué usamos los modelos de regresión?

- Según STRATOS podemos usar regresión para 3 propósitos diferentes:
  - Descripción*
  - Predicción
  - Explicación

Propósitos del modelamiento



Propósitos del modelamiento (cont.)



¿Para qué usamos los modelos de regresión? (cont.)

- Este curso se centrará solamente en algunas aplicaciones.
- Regresión para descripción:
  - “Factores asociados..:” No necesariamente importa que los factores sean causales.
  - Evaluación de la magnitud de desigualdades, magnitud de brechas, etc.

¿Para qué usamos los modelos de regresión? (cont.)

- Regresión para explicación:
    - “Efecto / Efectividad / Impacto”: Busca estimar efectos causales.
    - Explorar potenciales factores causales… (puede clasificarse dentro de descripción)
- Regresión para predicción:
    - Factores pronóstico o predictores de…“: Identifican predictores de interés que luego alimenten mdelos predictivos.
    - Modelos de predicción: Predicción para diagnóstico y pronóstico.

¿Para qué usamos los modelos de regresión? (cont.)

- No abordaremos modelos de regresión para desarrollar modelos o reglas de predicción clínica.
- Tampoco para métodos de inferencia causal robusta.

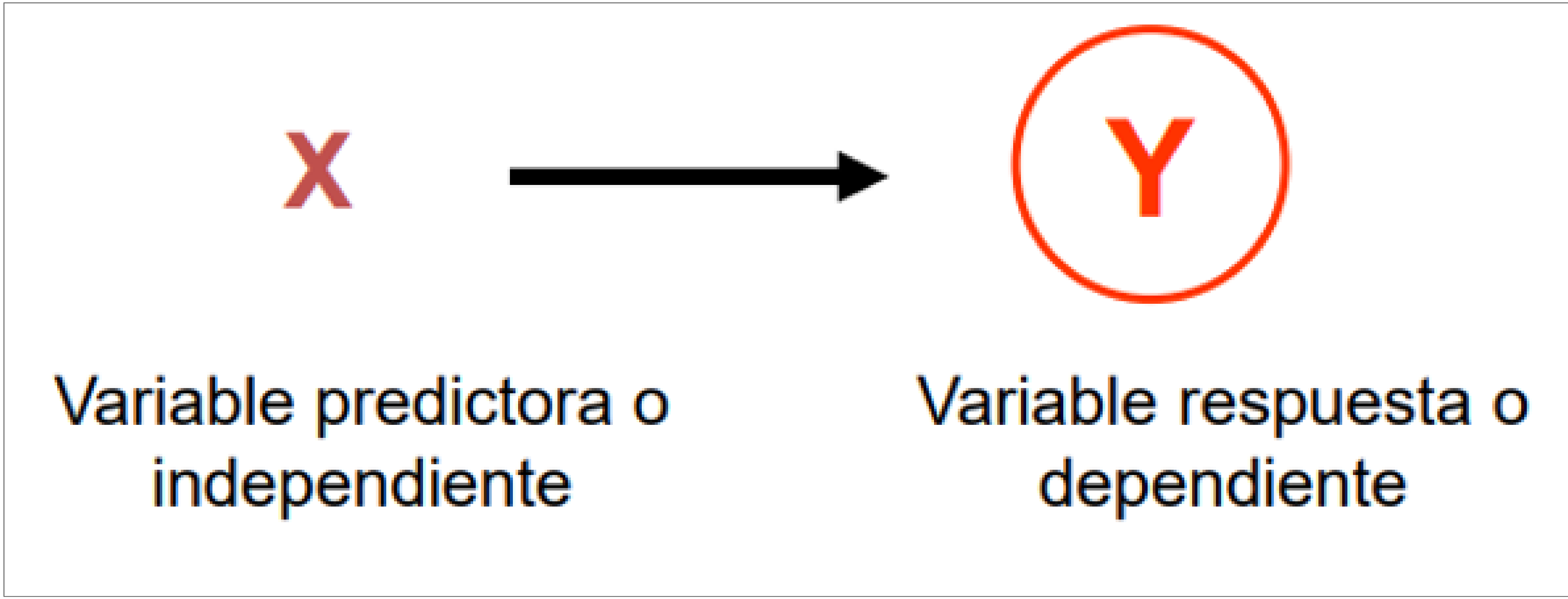
Clasificación inspirado en: Miguel A. Hernán, John Hsu & Brian Healy (2019) A Second Chance to Get Causal Inference Right: A Classification of Data Science Tasks, CHANCE, 32:1, 42-49, DOI: 10.1080/09332480.2019.1579578

Agenda

1. Introducción al modelado de regresión
2. **Modelo de Regresión Lineal**
3. El Modelo Lineal Generalizado
4. La regresión (log) Poisson
5. Tablas de regresión reproducibles con {gtsummary}

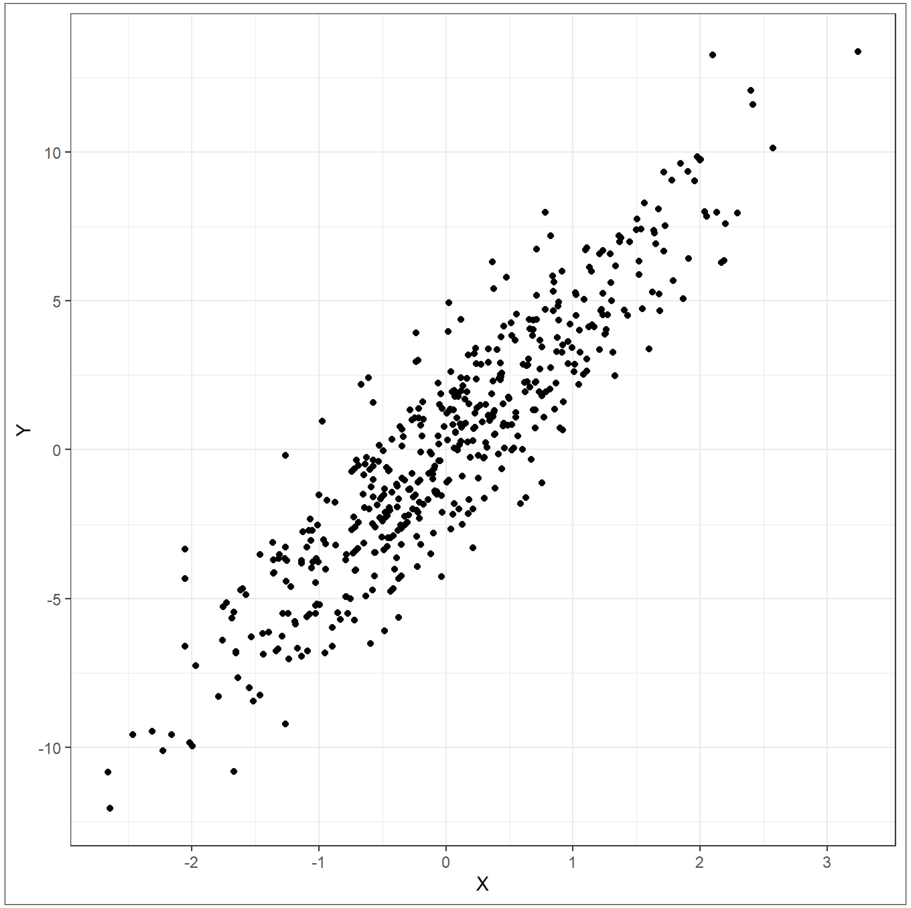
Regresión Lineal

- Método estadístico que modela la `relación` entre una `variable continua (dependiente)` y otras `variables (independientes)`.



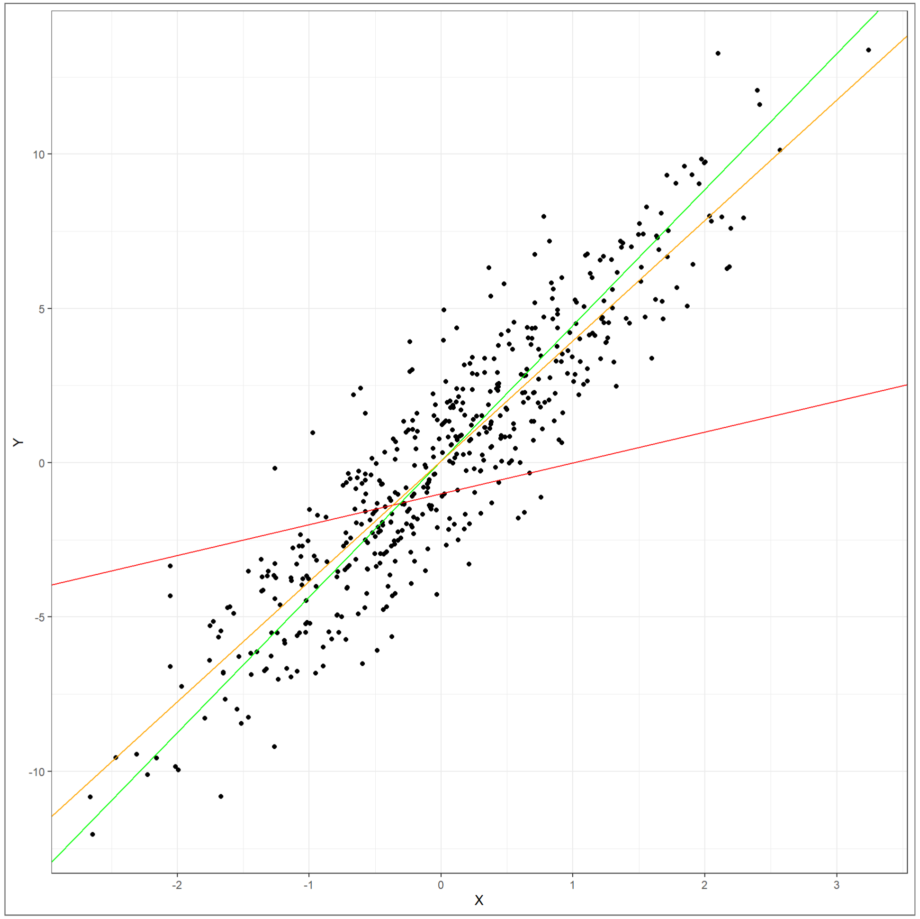
Relación entre dos variables

- Y es `variable resultado` (*outcome*), respuesta o dependiente.
- X es una `variable explicativa`, predictora o regresora.
- En la figura, a mayor valor de X, mayor valor de Y.



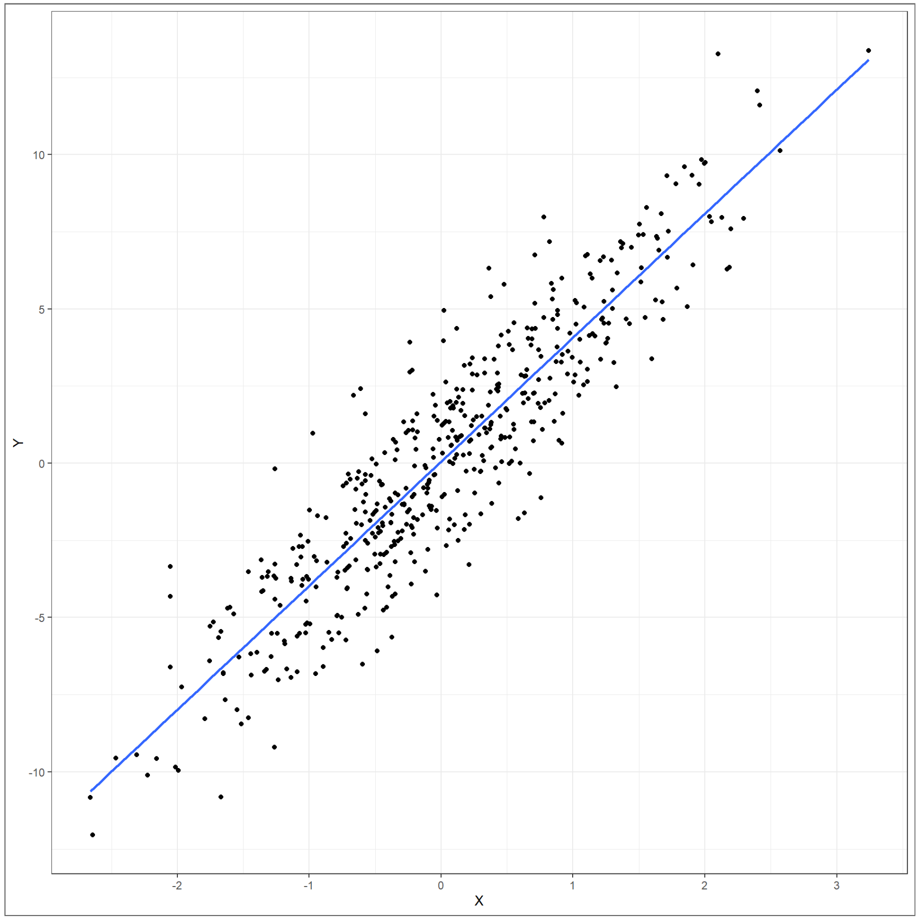
¿Cómo podemos resumir la relación entre ambas variables?

- Podemos tratar de dibujar una `línea recta` que `resuma` la relación.
- Existen `infinitas rectas posibles` que podríamos trazar: ¿Cuál elegir?



¿Cómo podemos resumir la relación entre ambas variables? (cont.)

- Una opción sería elegir una `recta` que pase por el `valor más representativo` del y_i en cada valor fijo de x_1.
  - Una `recta` que `conecte` los `promedios condicionados` en x_1



Anatomía de la RLS

- Entonces, la `recta que conecta los promedios` de y_i `condiciondos` en x_{1i} se puede expresar mediante la siguiente `combinación lineal`:

  $\beta_0 + \beta_1 x_{1i}$

- **Componente Sistemático:** Formalmente hablando, para cada observación i en la población, podemos `relacionar` el `valor esperado` (promedio) $E[y_i]$ de y_i (también llamado μ_i) con la `variable explicativa` x_{1i} mediante la siguiente `ecuación lineal`:

  $E[Y|X_1 = x_{1i}] = E[y_i] = \mu_i = \beta_0 + \beta_1 x_{1i}$

- Donde:
  - y_i son `variables aleatorias` independientes e idénticamente distribuidas (`i.i.d`)
  - x_1 es una variable cuyas valores son fijos y conocidos: $x_1 i$:
    - Se asume se `miden sin error`.
    - `No importa` su `distribución`.
  - β_0 y β_1 son `parámetros desconocidos` de una superpoblación infinita.
    - Llamados `coeficientes de regresión` y son una `medida de asociación`.
    - Es lo que `queremos estimar` con los datos de la muestra!

Anatomía de la RLS (cont.)

- Notar que el `componente sistemático` solo `relaciona` el `promedio condicionado` de y_i con las `variables explicativas`, NO con los valores individuales.
  - Esta es una manera de obtener una medida que resuma las relaciones individuales en una sola medida.
- **Componente aleatorio:** Para poder relacionar completamente los valores individuales se agrega un término de error ϵ, el cual se obtiene de restar el valor observado y_i con el valor esperado de este (μ_i):

$$\epsilon_i = y_i - \mu_i$$

- El problema es que el término de error ϵ_i no puede predecirse ni estimarse con los datos, se considera que es el componente no explicado por estos.
  - Para lidiar con este, se asume que su comportamiento puede predecirse a nivel probabilístico: Se asume una distribución de este.
  - El error ϵ_i hereda la distribución de probabilidad de y_i.
- Por lo tanto, el valor individual de cada y_i puede ser denotado por la siguiente expresión:

$$y_i = \beta_0 + \beta_1 x_{1i} + \epsilon_i$$

- Para hacer inferencia estadística, a menudo se asume lo siguiente:

$$y_i \sim N(\beta_0 + \beta_1 x_{1i}, \sigma^2)$$
$$\epsilon_i \sim N(0, \sigma^2)$$
Regresión Lineal Normal

Algunas notas sobre normalidad

- No es necesario que ϵ_i o y_i sigan una distribución normal para que los coeficientes de regresión β puedan estimarse de manera puntual.
- Sin embargo, para estimar el `valor p` o los `intervalos de confianza` mediante `inferencia clásica` sí se necesita asumir una distribución conocida. El modelo de regresión lineal normal asume normalidad de estos.
  - Asimismo, el modelo es robusto a desviaciones leves/moderadas de la normalidad cuando se cumple el TLC (número de observaciones grande).
- Otros enfoques para inferencia flexibilizan este supuesto: p. ej., bootstrap, varianza robusta, modelo lineal generalizado que asume otras distribuciones, etc.

Estimación de ecuación de regresión

- En la práctica no conocemos los valores de los parámetros, así que los estimamos de nuestros datos.

Notación para la ecuación de regresión

	Parámetro poblacional	Estadístico muestral
Intercepto y de la ecuación de regresión	β_0	b_0
Pendiente de la ecuación de regresión	β_1	b_1
Ecuación de la recta de regresión	$y = \beta_0 + \beta_1 x$	$\hat{y} = b_0 + b_1 x$

Triola 10ma Ed.

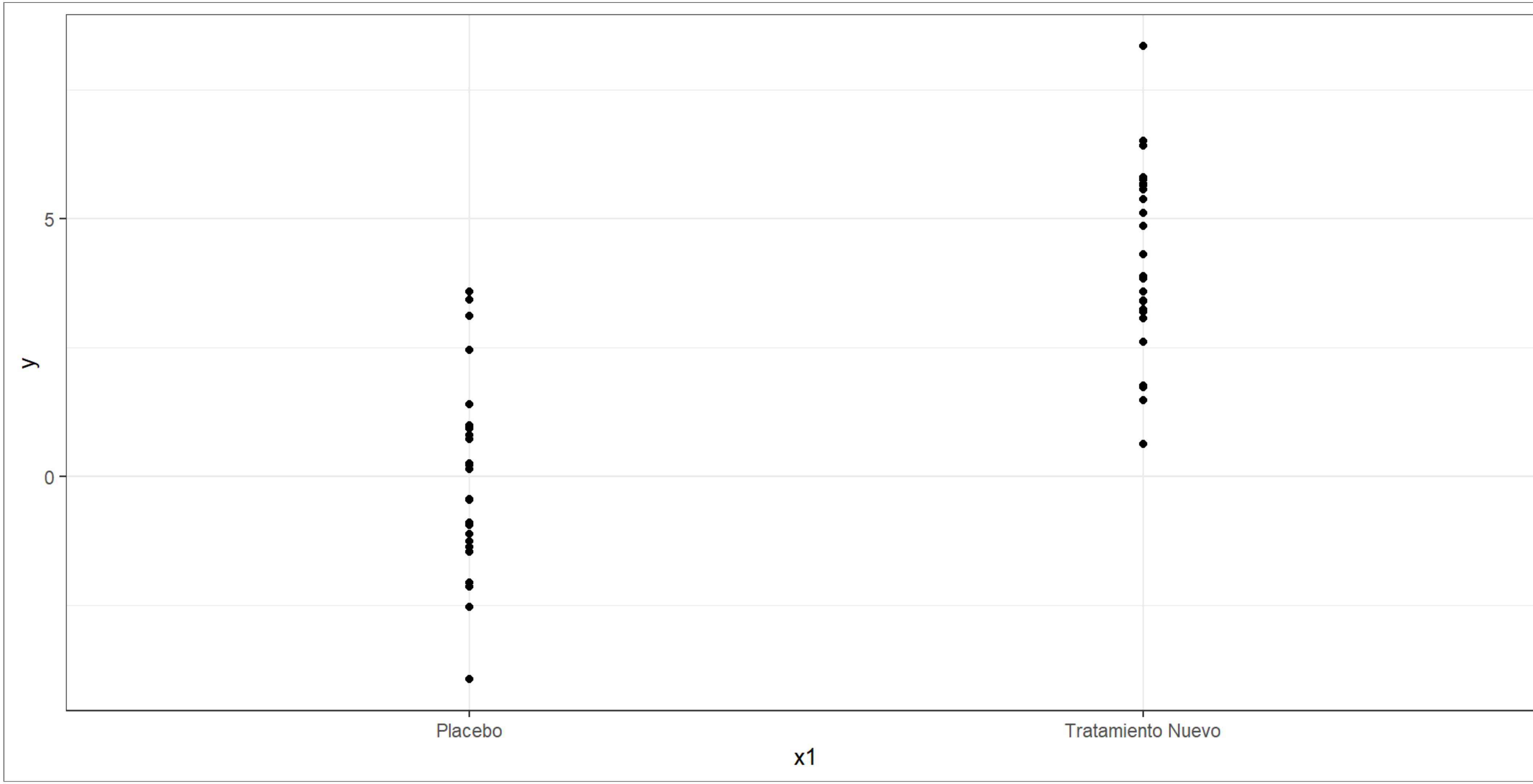
¿Cómo estimamos la ecuación lineal que mejor ajusta a los datos observados?

- Usamos métodos numéricos:
  - Método de Mínimos Cuadrados Ordinarios
  - Método de Máxima Verosimilitud
- Ambos métodos son equivalentes para el caso de la regresión lineal normal.

Regresión Lineal Simple sobre variable explicativa categórica

- Las variables categóricas no son continuas, en cambio son discretas y asumen unos cuantos valores.
- ¿Cómo estimar una medida de asociación cuando la variable explicativa es categórica?

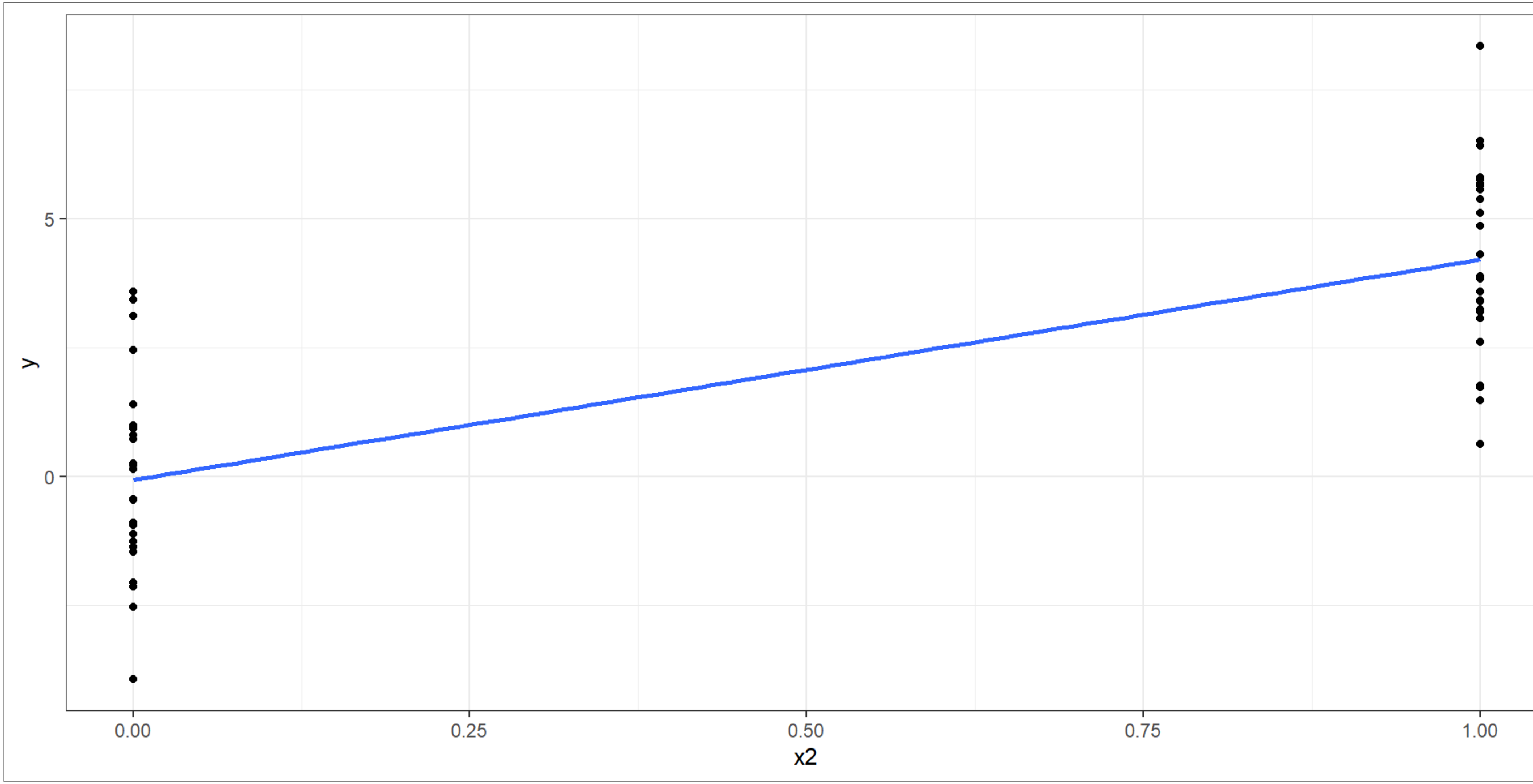
Regresión Lineal Simple sobre variable explicativa categórica



Regresión Lineal Simple sobre variable explicativa categórica (cont.)

- Si la variable es binaria, una forma de abordar el análisis es asignando a una categoría el valor de 1 y a otra el valor de 0.
  - Entonces, asumiremos que la variable categórica es numérica para los efectos de todo cálculo.
  - Sin embargo, la interpretación se centrará en la comparación de categorías.

Regresión Lineal Simple sobre variable explicativa categórica (cont.)



Regresión Lineal Simple en R

- Se usa la función `lm()` de R base. Sin embargo, la salida de esta no es muy informativa:

```r
lm(y ~ x1, data = datos)
```

```

Call:
lm(formula = y ~ x1, data = datos)

Coefficients:
        (Intercept)  x1Tratamiento Nuevo
           -0.06666              4.27094
```

- El modelo puede guardarse para realizar más operaciones sobre este. Por ejemplo, mejorar la salida:

```r
mod <- lm(y ~ x1, data = datos)
summary(mod)
```

```

Call:
lm(formula = y ~ x1, data = datos)

Residuals:
    Min      1Q  Median      3Q     Max 
-3.8666 -1.1168 -0.3487  1.3100  4.1336 

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)    
(Intercept)         -0.06666    0.37316  -0.179    0.859    
x1Tratamiento Nuevo  4.27094    0.52773   8.093 1.59e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.866 on 48 degrees of freedom
```

Interpretación de salida de RLS

Covariable numérica | Covariable categórica

- Usamos la función lm():

```
mod <- lm(y_peso_final ~ x3_peso_inicial, data = datos2)
summary(mod)
```

```

Call:
lm(formula = y_peso_final ~ x3_peso_inicial, data = datos2)

Residuals:
     Min       1Q   Median       3Q      Max 
-10.0568  -4.7717  -0.8704   5.1824  10.4953 

Coefficients:
                Estimate Std. Error t value Pr(>|t|)    
(Intercept)      -5.4317     2.6574  -2.044   0.0412 *  
x3_peso_inicial   1.3447     0.1766   7.615  6.1e-14 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.535 on 998 degrees of freedom
```

- El modelo estimado sería el siguiente:

$$y\_pesofinal = -5.4317 + 1.3447 * x3\_pesoinicial + \epsilon_i$$
$$\epsilon_i \sim Normal(0, 5.535^2)$$

- Usando el paquete `{broom}` y su función `tidy()` podemos obtener también los intervalos de confianza:

```
library(broom)
mod %>%
  tidy(conf.int = TRUE)
```

```
# A tibble: 2 × 7
  term            estimate std.error statistic  p.value conf.low conf.high
  <chr>              <dbl>     <dbl>     <dbl>    <dbl>    <dbl>     <dbl>
1 (Intercept)        -5.43     2.66      -2.04 4.12e- 2  -10.6      -0.217
2 x3_peso_inicial     1.34     0.177      7.62 6.10e-14    0.998     1.69
```

- Interpretación:
  - β_0 o `intercepto`: Este viene a ser el valor promedio de y cuando todos los valores de x son 0. En este caso, cuando el peso inicial es cero kg. ¿Esto es posible?, por tal motivo, no se suele interpretar este valor.
  - β_1 o coeficiente de regresión de `x3_peso_inicial`: Por `cada 1 kg adicional` de peso inicial, el `valor promedio` del peso final aumenta 1.43 kg (IC95% 1.00 a 1.69; $p < 0.001$).
- Usamos la función lm():

```r
mod <- lm(y_peso_final ~ x1_tto, data = datos2)
summary(mod)
```

```

Call:
lm(formula = y_peso_final ~ x1_tto, data = datos2)

Residuals:
    Min      1Q  Median      3Q     Max 
-7.7043 -1.6644 -0.0095  1.5849  8.5658 

Coefficients:
                        Estimate Std. Error t value Pr(>|t|)    
(Intercept)              19.8771     0.1112  178.67   <2e-16 ***
x1_ttoTratamiento Nuevo -10.2325     0.1573  -65.04   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.488 on 998 degrees of freedom
```

- Usando tidy de broom:

```r
mod %>%
  tidy(conf.int = TRUE)
```

```
# A tibble: 2 × 7
  term                    estimate std.error statistic p.value conf.low conf.h…¹
  <chr>                      <dbl>     <dbl>     <dbl>   <dbl>    <dbl>    <dbl>
1 (Intercept)                 19.9     0.111     179.        0     19.7    20.1 
2 x1_ttoTratamiento Nuevo    -10.2     0.157     -65.0       0    -10.5    -9.92
# … with abbreviated variable name ¹conf.high
```

- Interpretación:
  - β_0 (Intercept): A menudo no se interpreta. Es el valor promedio de y_i cuando los valores de x son cero. En este caso, cuando el tratamien es cero (placebo). ¿Esto es posible?, sí es posible pero no es de ayuda para modelos explicativos, por lo que no se interpreta.
  - $\beta 1$ x1Tratamiento Nuevo: El promedio de peso final en quienes recibieron el tratamiento nuevo fue 10.23 kg menor que el de quienes recibieron placebo (Dif. medias = -10.23; IC95% -10.54 a -9.92; $p < 0.001$).

Regresión Lineal Múltiple

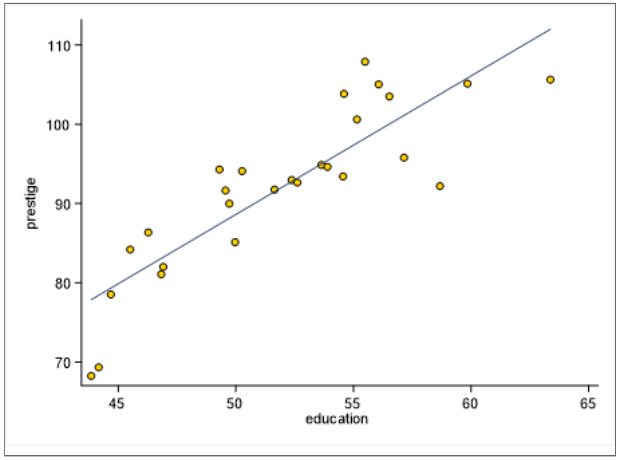
- Generaliza la RLS permitiendo evaluar la relación de varias covariables explicativas x sobre y_i.
- Para p variables explicativas, el modelo puede expresarse como:

Componente sistemático:
$E[Y|X_1 = x_{1i}, \ldots, X_p = x_{pi}] = E[y_i] = \mu_i = \beta_0 + \beta_1 x_{1i} + \ldots + \beta_p x_{pi}$
Componente aleatoria:
$y_i \sim N(\beta_0 + \beta_1 x_{1i} + \ldots + \beta_p x_{pi}, \sigma^2)$
$\epsilon_i \sim N(0, \sigma^2)$

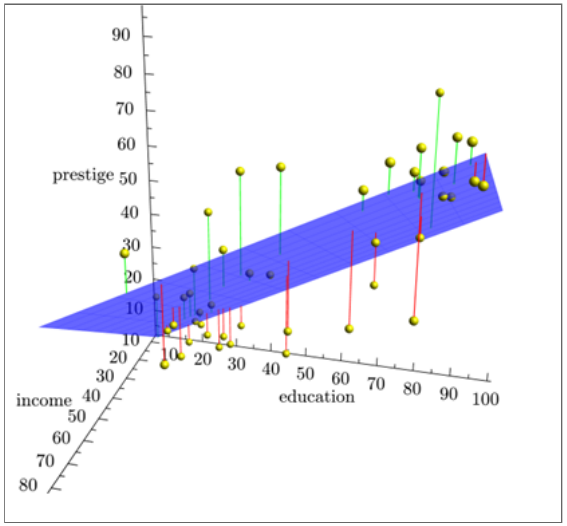
Regresión Lineal en gráficos

RLS | **RLM con 2 X** | **RLM con 3 o más X**

- La ecuación de la RLS representa una línea recta.



- La ecuación de la RLM con dos variables explicativas ya no representa una línea recta, sino un plano recto.



- Genera un hiperplano recto.
- No podemos imaginarnos una imagen de esto, pero sí podemos analizarlo a nivel estadístico.

- Algebra lineal proporciona herramientas para lidiar con esto usando matrices.

RLM en R

- Usamos la función lm():

```r
mod <- lm(y_peso_final ~ x1_tto + x3_peso_inicial, data = datos2)
summary(mod)
```

```

Call:
lm(formula = y_peso_final ~ x1_tto + x3_peso_inicial, data = datos2)

Residuals:
    Min      1Q  Median      3Q     Max 
-5.5598 -1.4213  0.1343  1.0768  5.4482 

Coefficients:
                         Estimate Std. Error t value Pr(>|t|)    
(Intercept)              -0.94719    0.99689   -0.95    0.342    
x1_ttoTratamiento Nuevo -10.25530    0.13111  -78.22   <2e-16 ***
x3_peso_inicial           1.38755    0.06614   20.98   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- El modelo estimado sería el siguiente:

$$y\_pesofinal = -0.94719 - 10.25530 * x1ttoTratamientoNuevo + 1.3875 * x3\_pesoinicial + \epsilon_i$$
$$\epsilon_i \sim Normal(0, 2.073^2)$$

- Usando el paquete `{broom}` y su función `tidy()` podemos obtener también los intervalos de confianza:

```r
library(broom)
mod %>%
  tidy(conf.int = TRUE)
```

```
# A tibble: 3 × 7
  term                    estimate std.error statistic  p.value conf.low conf.…¹
  <chr>                      <dbl>     <dbl>     <dbl>    <dbl>    <dbl>   <dbl>
1 (Intercept)               -0.947     0.997    -0.950 3.42e- 1    -2.90    1.01
```

```
2 x1_ttoTratamiento Nuevo  -10.3      0.131    -78.2   0          -10.5   -10.0
3 x3_peso_inicial            1.39     0.0661    21.0   3.10e-81     1.26    1.52
# … with abbreviated variable name ¹conf.high
```

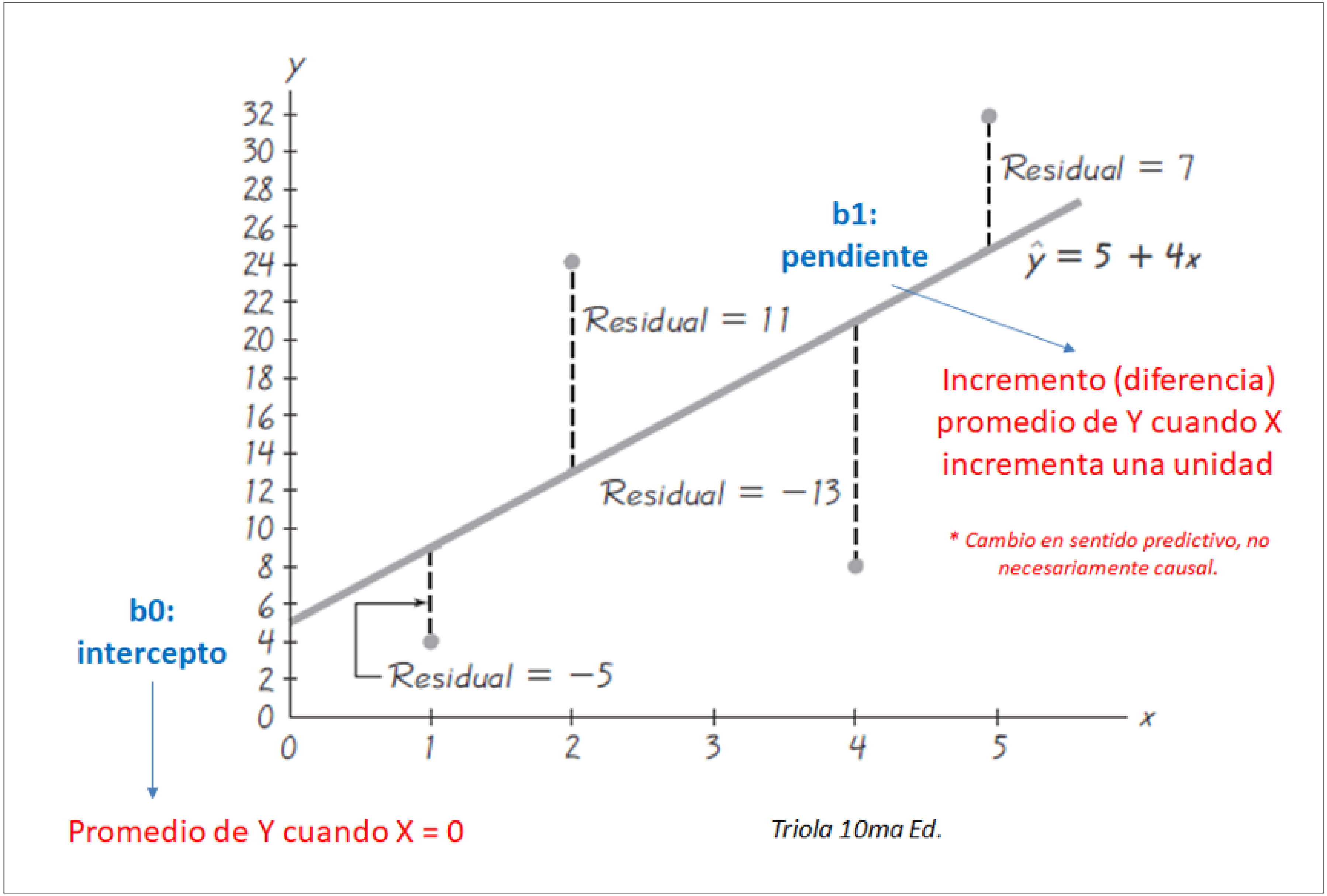
- Interpretación:
  - β_0 o `intercepto`: Este viene a ser el valor promedio de y cuando todos los valores de x son 0. En este caso, cuando el peso inicial es cero kg y cuando el tratamiento es placebo. ¿Esto es posible?, por tal motivo, no se suele interpretar este valor.
  - β_2 o coeficiente de regresión de `x1_ttoTratamiento Nuevo`: El promedio de peso final en quienes recibieron el tratamiento nuevo fue 10.26 kg menor que el de quienes recibieron placebo, luego de ajustar por peso inicial (Dif. medias = -10.26; IC95% -10.51 a -9.99; p < 0.001).
  - β_1 o coeficiente de regresión de `x3_peso_inicial`: Por `cada 1 kg adicional` de peso inicial, el `valor promedio` del peso final aumenta 1.39 kg, luego de ajustar por tatamiento recibido (IC95% 1.26 a 1.52; p < 0.001).

:::

Errores y residuos

- Los `errores` (ϵ_i) son medidas de la población a la que no tenemos acceso.
  - Sin embargo, varios supuestos de la regresión involucran a los errores inaccesibles por el investigador.
- Los `residuos` (e_i) son el análogo a los `errores` pero obtenidos de la `muestra observada`.
- Podemos usar los `residuos` para `evaluar` algunos `supuestos` sobre los `errores`.

Residuos gráficamente



Supuestos de la regresión lineal normal

Supuestos del modelo - Linealidad

- Independencia de observaciones
- Homocedasticidad de los errores ϵ_i
- Normalidad de los errores ϵ_i o de y_i.
- No problemas con la regresión:
  - Puntos influyentes.
  - (Multi) colinealidad: Solo cuando es un problema, no siempre lo es.

Supuestos adicionales que suelen acompañar a la regresión lineal normal

Supuestos si queremos generalizar a una población finita bien definida

- La muestra es representativa de la población.
  - Ideal para alcanzar esto es mediante muestreo probabilístico: representatividad estadística.
  - Cuando no lo tenemos, solo podemos generalizar a una población que sabemos que existe pero no podemos definir. ¿Qué tan relevante puede ser esto?
  - Otros consideran que, bajo ciertas condiciones, se puede alcanar una representativadad teórica.

Supuestos adicionales que suelen acompañar a la regresión lineal normal

Supuestos si queremos hacer inferencias causales

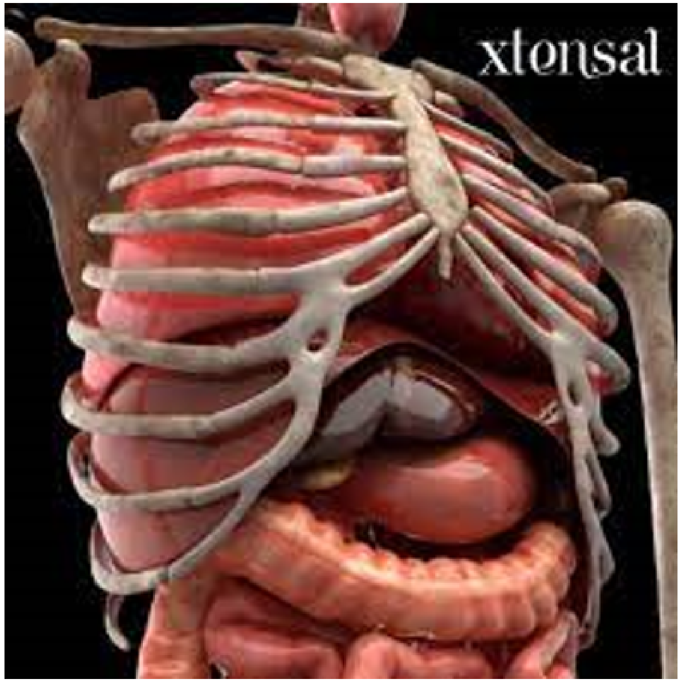
- Hay asignación aleatoria
  - Ideal para alcanzar esto es mediante experimento aleatorizado.
  - Cuando no lo tenemos, tenemos que poder asumir que se puede emular la asignación aleatoria de alguna manera:
    - El ajuste de regresión por confusores es una manera de pensar en esto.

Algunas notas sobre los errores y residuos

- En realidad, los supuestos de los modelos lineales son sobre el comportamiento probabilístico de y_i.
- Sin embargo, la idea de la existencia de los `errores` y de sus valores observados en la muestra, `residuos` resulta útil para evaluar supuestos.
  - Permiten reducir un problema de muchas dimensiones a solo 1 o 2 dimensiones.
  - Son como las placas radiográficas para el diagnóstico de los modelos.

Algunas notas sobre los errores y residuos

La **ecuación de regresión múltiple** y sus complejas relaciones multidimensionales que **no podemos ver directamente**



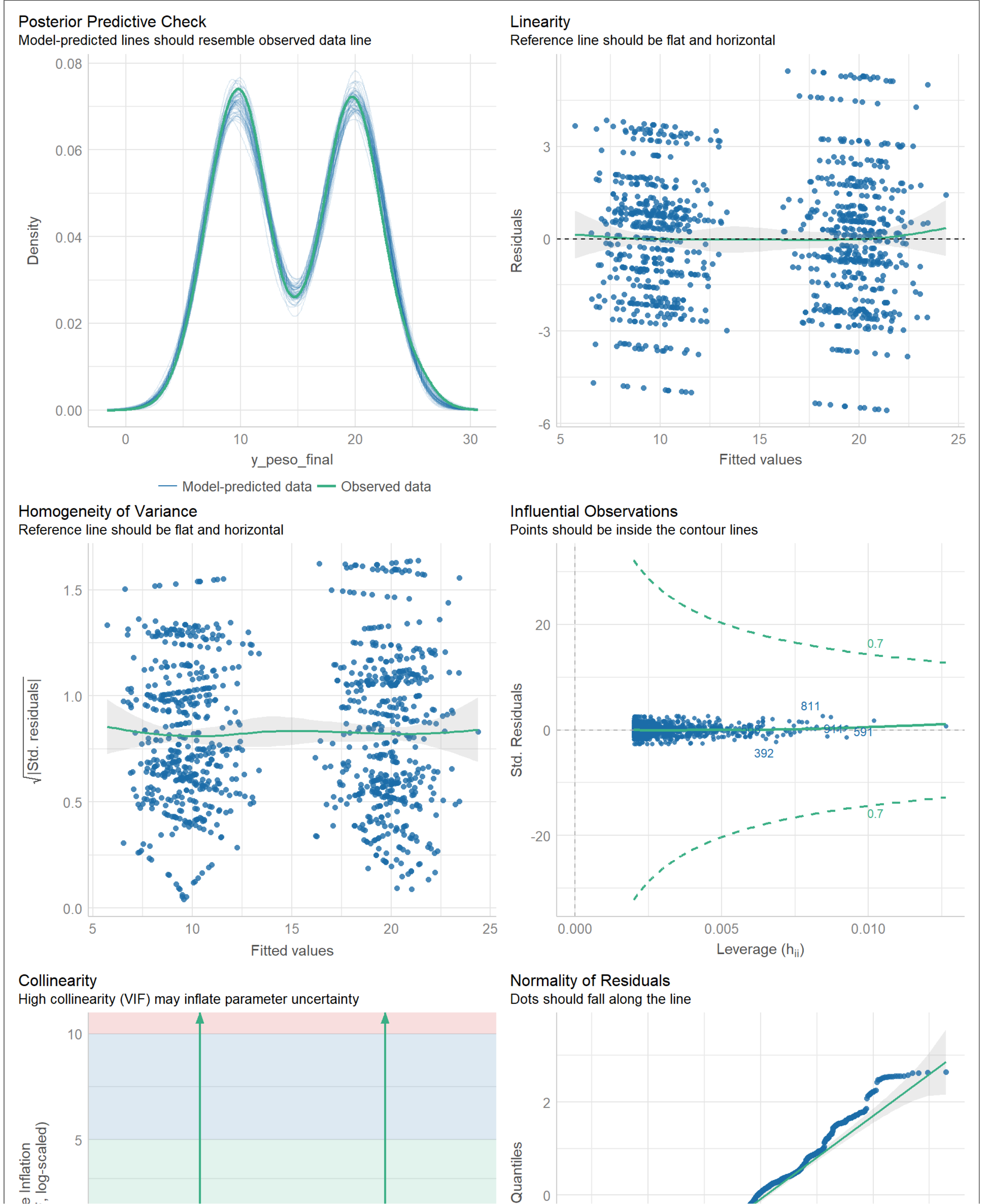
La "radiografía" de la ecuación de regresión para **diagnosticar problemas** en su planteamiento: Gráfico de residuales y otras "pruebas" diagnósticas, que **sí podemos ver directamente!**

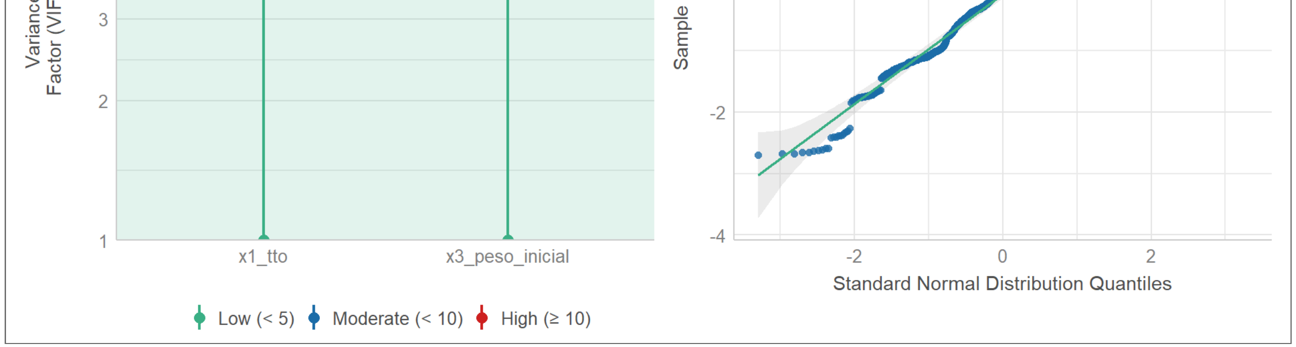
¿Cómo evaluar los supuestos de la regresión lineal?

- Se usan los residuos para explorar el comportamiento de los y_i o los errores ϵ.
- Preferiblemente usar gráficos de residuos.
  - Pruebas de hipótesis que usan residuos tienen los mismos problemas que discutimos en clases anteriores.
  - Podríamos usarlas para complementar análisis cuando los tamaños de muestra no son ni muy pequeños ni muy grandes.
- La función `check_model` del paquete `{performance}` genera un panel de gráficos muy útil para evalur estos supuestos.
- Podemos complentar el análisis de supuestos con funciones del paquete `{car}`.

Panel general | Lin. det. | Homo. det. | P. inf. det.

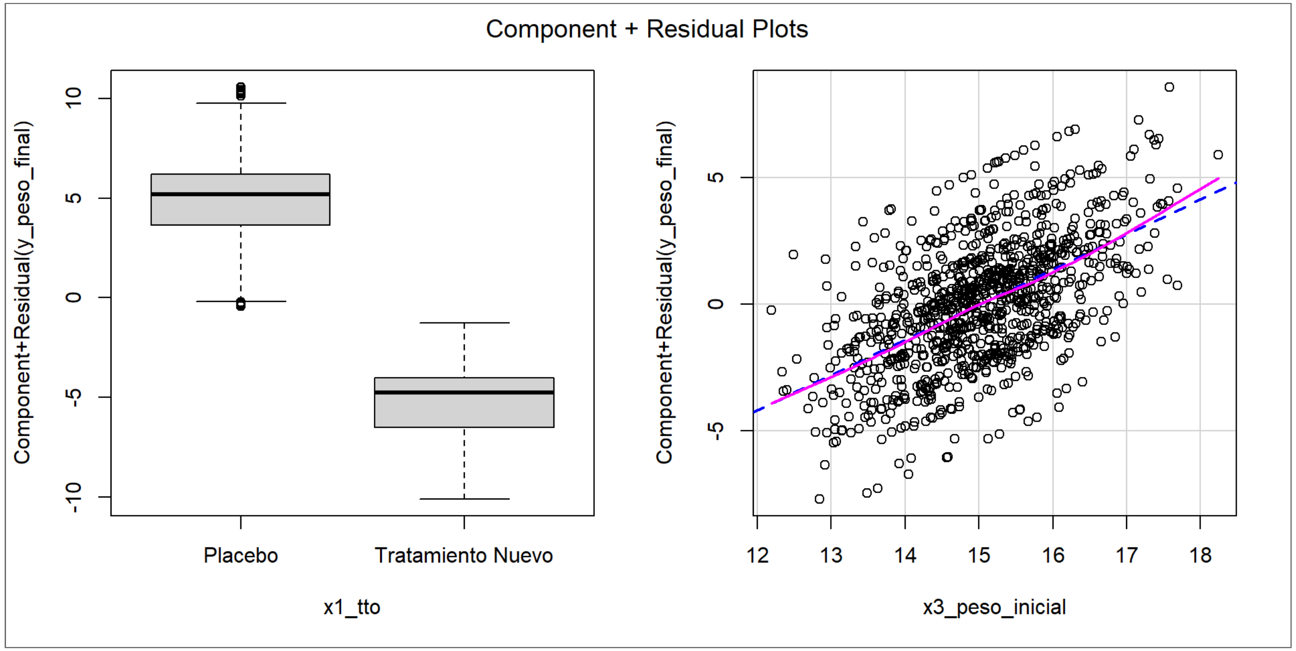
```
library(performance)
check_model(mod)
```


Posterior Predictive Check
Model-predicted lines should resemble observed data line
Density
y_peso_final
Model-predicted data
Observed data
Linearity
Reference line should be flat and horizontal
Residuals
Fitted values
Homogeneity of Variance
Reference line should be flat and horizontal
√|Std. residuals|
Fitted values
Influential Observations
Points should be inside the contour lines
Std. Residuals
Leverage (h_{ii})
0.7
811
392
591
Collinearity
High collinearity (VIF) may inflate parameter uncertainty
Normality of Residuals
Dots should fall along the line
Quantiles



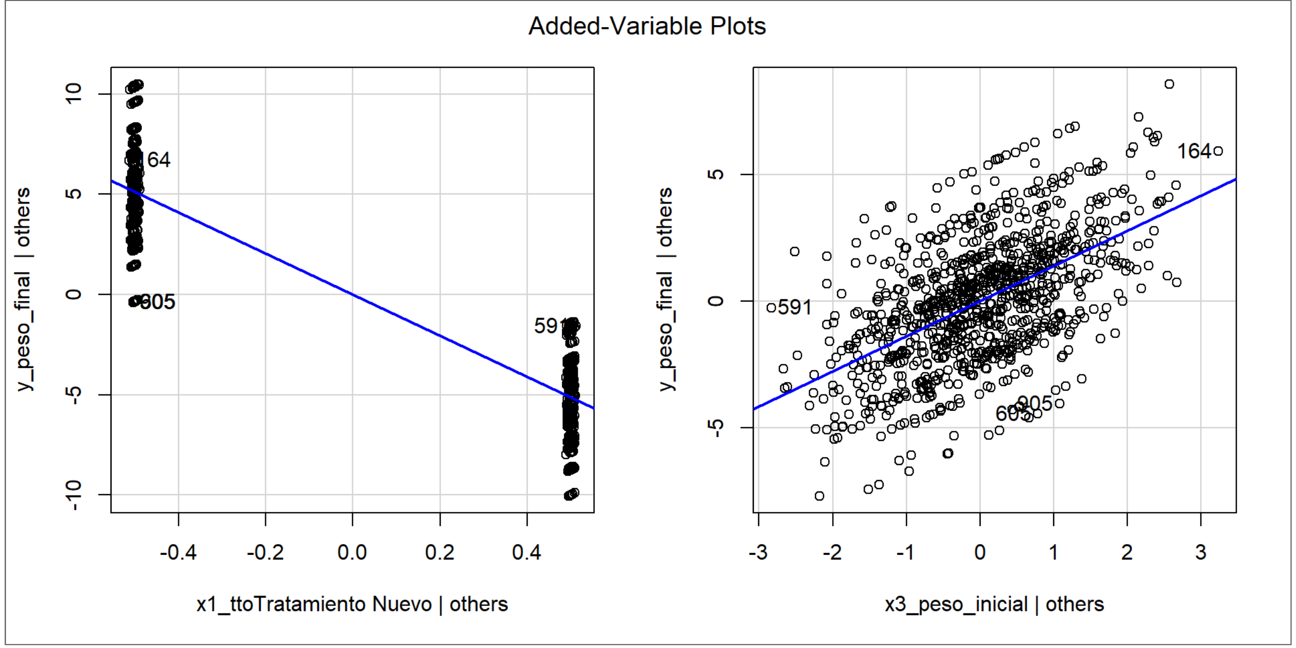
- Podemos usar gráficos de residuos parciales + Componente:

```r
library(car)
crPlots(mod)
```



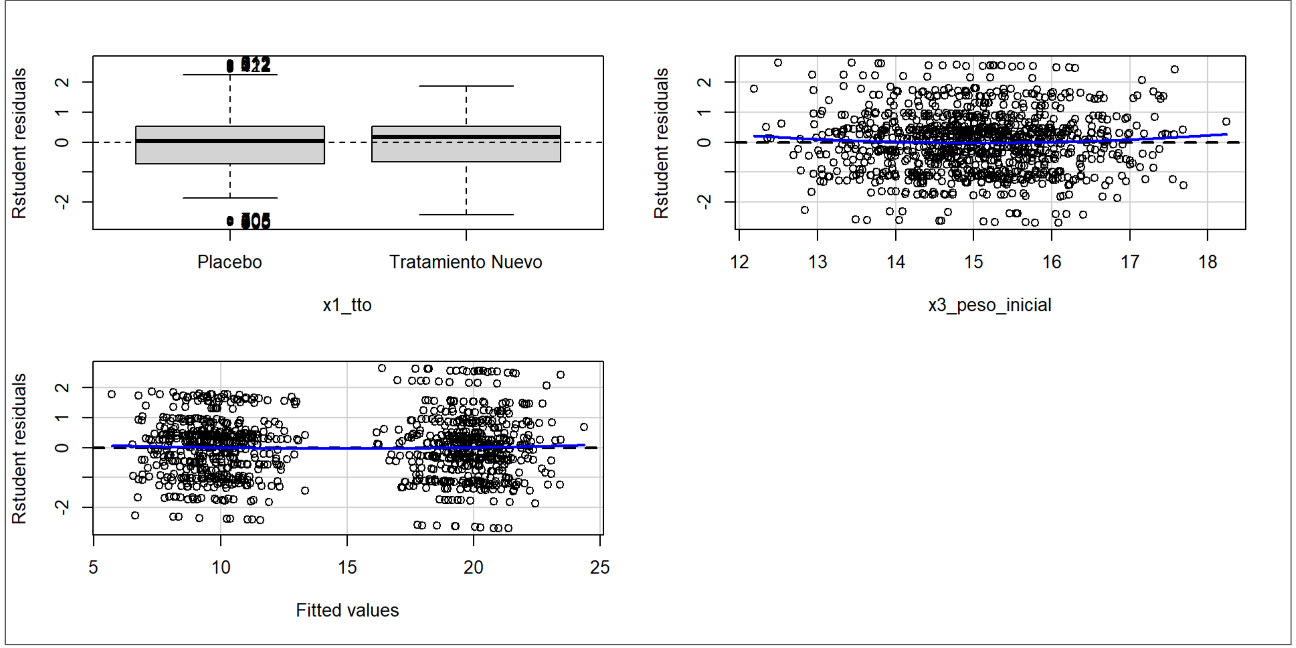
- También podemos usar gráficos de variable agregada

```r
avPlots(mod)
```



- Se puede evaluar si la homocedasticidad es consistente según cada variable predictora.
- Si no lo es, se puede optar por modelar esta heterogeneidad de varianzas.
- Se sugiere usar `residuos estudentizados`.

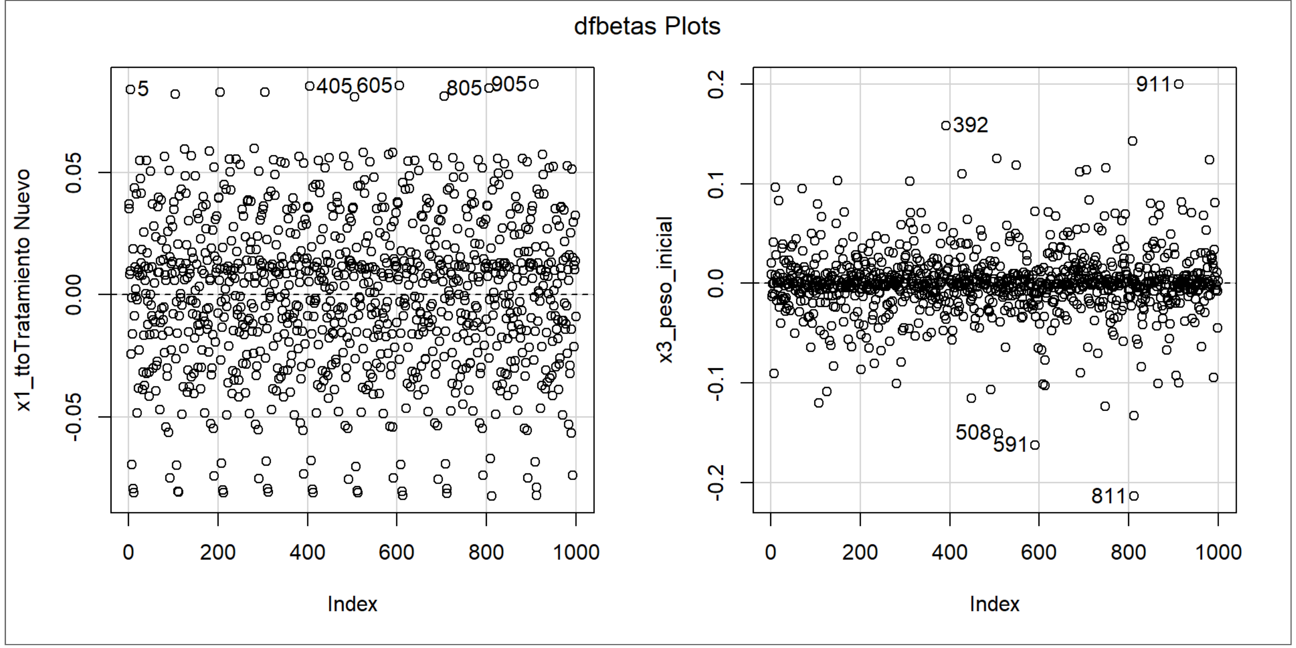
```
residualPlots(mod, type = "rstudent")
```




```
                 Test stat Pr(>|Test stat|)
x1_tto
x3_peso_inicial     1.2180           0.2235
Tukey test          0.5429           0.5872
```

- En el caso de modelos explicativos, importa determinar si hay un impacto en los coeficientes de regresion.
- Los `dfbetas` pueden ser útiles para evaluar esto:

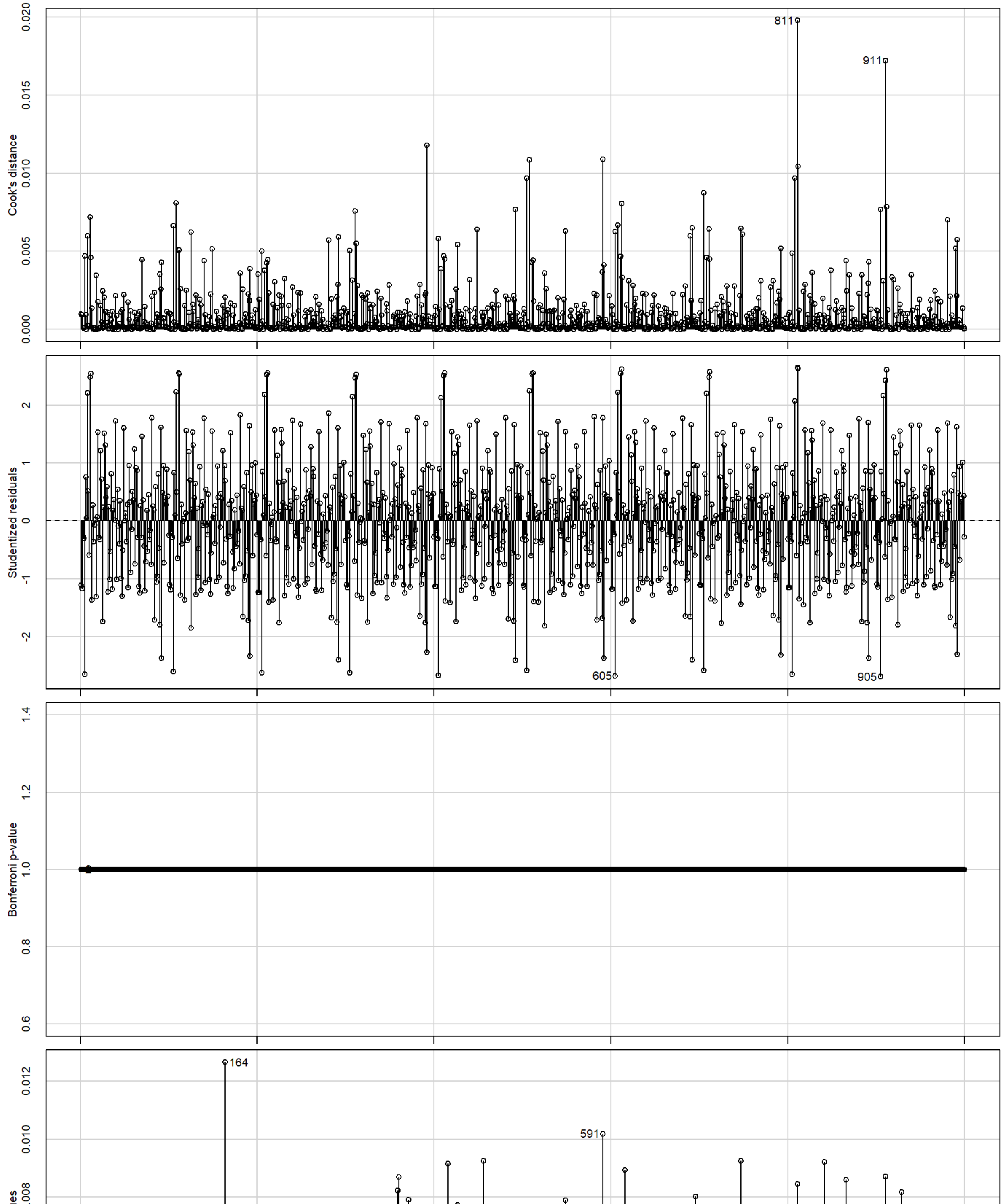
```
dfbetasPlots(model = mod, id.n = 5)
```

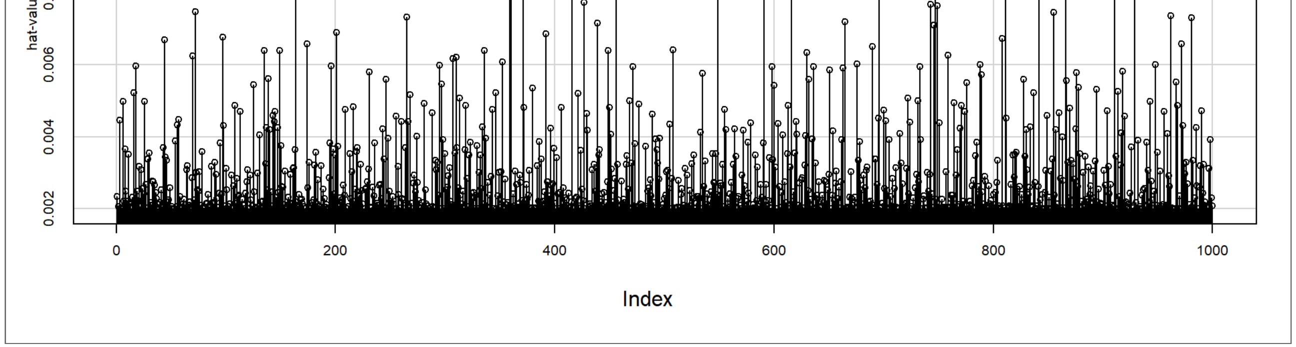


- Otras medidas también pueden evaluarse:

```
influenceIndexPlot(model = mod, id.n = 5)
```

Diagnostic Plots




0.002
0.004
0.006
0
200
400
600
800
1000
Index

¿Cómo flexibilizar supuestos?

No linealidad | Heterocedasticidad | No normalidad

- El supuesto de linealidad es sobre los coeficientes de regresión β, no sobre las covariables.
- Las variables X deben estar en una forma apropiada para que el supuesto se cumpla.
- Es bien difícil que exista linealidad en la realida, pero puede ocurrir en raras y excepcionales ocasiones.
  - Sobre todo cuando la variable está acotada en valores donde la linealidad es plausible.
- Se sugiere asumir no linealidad y pre-planear un modelamiento no lineal.
- Entre los métodos que pueden usarse, tenemos:
  - `Splines`: Bastante usado y sugerido en bioestadística. Útil para ajustar por variables continuas.
  - `Modelamiento Multivariablede polinomios fraccionales`. También usado y recomendado en literatura biomédica. Útil para modelar forma como objetivo principal.
  - `Polinomios`. Menos flexible, puede ser útil si se conoce bien la relación o se busca mejorar ajuste.
  - `Modelos aditivos generalizados`. Útil si se buscar modelar la relación. Complejos y requieren muchos datos.
- Veamos un ejemplo de modelamiento continuo con `splines`:

> ⓘ
>
> **Evite categorizar la variable continua**
>
> - Categorizar es muy malo: se pierde información y se corre el riesgo de sesgar resultados.
> - Si se quiere ajustar por variables continuas, use Splines o Polinomios fraccionales. No requiere interpretar sus resultados, pero si ajsutar bien!
> - Si se quiere evaluar la relación de la variable continua, planee un método estadístico para modelar la forma sin asumir linealidad.
>   - Presuponga que la relación no es lineal.
>   - Modelo y responda su pregunta. Si la relación es lineal, el modelo más complejo revelerá una línea recta.

- No homogeneidad de varianzas
- Podemos usar una estimación robusta de la varianza.
- Los paquetes `{sanwich}` y `{lmtest}` proporcionan funciones útiles para esto.
- Es bien difícil de creer que existe homogeneidad de varianzas en la vida real (salvo muy raras y excepcionales ocasiones).
  - Se sugiere planear el proyecto asumiendo que no hay homocedasticidad y usar inferencia robusta de manera pre-planeada.

```
library(lmtest)
library(sandwich)
coeftest(mod, vcov = vcovHC) %>%
  tidy(conf.int = TRUE)
```

```
# A tibble: 3 × 7
  term                    estimate std.error statistic  p.value conf.low conf.…¹
  <chr>                      <dbl>     <dbl>     <dbl>    <dbl>    <dbl>   <dbl>
1 (Intercept)               -0.947    1.05      -0.906 3.65e- 1    -3.00    1.10
2 x1_ttoTratamiento Nuevo  -10.3      0.131    -78.1   0          -10.5   -10.0
3 x3_peso_inicial            1.39     0.0692    20.0   2.45e-75     1.25    1.52
# … with abbreviated variable name ¹conf.high
```

- Si distribución es normal (cosa que no podemos saber con certeza), podemos dejar de preocuparnos por este supuesto.
- Si se cumple TLC, podemos dejar de preocuparnos por este supuesto.
- Si no se cumple TLC o hay dudas razonables, podemos optar por alguna de las siguientes alternativas:
  - Transformar Y para normalizar (p. ej., logaritmo)
  - Usar varianza robusta
  - Estimar varianza con bootstrapping u otro método de remuestreo.

Agenda

1. Introducción al modelado de regresión
2. Modelo de Regresión Lineal
3. **El Modelo Lineal Generalizado**
4. La regresión (log) Poisson
5. Tablas de regresión reproducibles con {gtsummary}

Modelo Lineal Generalizado

- Modelo lineal que permite modelar desenlaces de varios tipos.
- Generaliza el modelo de regresión lineal.
- Permite que Y_i siga otras distribuciones.

Modelo Lineal Generalizado: Anatomía

Componente sistemático:
$g(E(Y|x_{1i},...,x_{pi})) = g(E(Y_i)) = \eta_i = \beta_0 + \beta_1 x_{1i} + ... + \beta_p x_{ip}$

- $g()$ es la `función de enlace`.
- η_i es el `predictor linear`.

Modelo Lineal Generalizado: Anatomía (cont.)

Componente aleatorio:
$Y_i \sim Distribucion\ de\ la\ Familia\ Exponencial$

- Familia exponencial:

Variable respuesta	Distribución de FE	Función de enlace canónica $g()$	Otras funciones de enlace comunes
Binaria	Bernoulli (Binomial con n = 1)	$logit()$	$log()$
Conteo	Binomial (con n > 1)	$logit()$	$log()$
	Poisson	$log()$	
	Binomial negativo	$log(\mu + k)$	$log()$
Continua positiva	Gamma	$\frac{1}{\mu}$	
	Gausiana inversa		

Agenda

1. Introducción al modelado de regresión
2. Modelo de Regresión Lineal
3. El Modelo Lineal Generalizado
4. **La regresión (log) Poisson**
5. Tablas de regresión reproducibles con {gtsummary}

Regresión de Poisson

- Caso específico de Modelo Lineal Generalizado. Veamos el caso en el que usamos la función de enlace canónica para la distribución de Poisson: `log()`.
- **Componente sistemático:**

  $$log(E(y_i)) = \eta_i$$

- **Función de enlace:**

  $$\eta_i = \beta_0 + \beta_1 x_{1i} + \ldots + \beta_p x_{ip}$$

Regresión de Poisson (cont.)

- **Componente aleatorio:**

  $y_i \sim Poisson(\eta_i)$
  o, equivalentemente,
  $y_i \sim Poisson(\beta_0 + \beta_1 x_{1i} + \ldots + \beta_p x_{ip})$

¿Por qué usar log?

- Si usamos la función identidad de la regresión lineal, el modelo quedaría planetado de esta manera:

  $$E(y_i) = \beta_0 + \beta_1 x_{1i} + \ldots + \beta_p x_{ip}$$

- Entonces, el modelo predecirá valores fuera del rango natural de la variable y_i:
  - y_i es de conteo (discreto), pero se obtendrían predichos con decimales (continuo).
  - y_i es positivo siempre, pero se podrían ontener predichos negativos.

¿Por qué no asumir normalidad de y_i?

- Porque la distribución de y_i no es normal, es una variable de conteo.
- El principal problema de esto, es que al ser Poisson, la media = varianza, por lo que a mayor valor de la media, la varianza aumentará, lo que implica que y_i es heterocedástica.
  - El modelo normal necesita homocedasticidad, caso contrario, tiene que corregirse de alguna manera.
  - Poisson no necesita esto, su modelo es heterocedastico por naturaleza, lo que hace más eficiente la estimación: si el modelo es válido, los intervalos de confianza serán más precisos.

La regresión de Poisson retorna razón de medias

- La regresión de Poisson permite retornar directamente `razón de medias` (RM).
- Los coeficientes de regresión β del modelo son $log(RM)$, por lo tanto, podemos exponenciarlos para obtener los OR:

$$\beta = log(RM)$$

entonces

$$e^{\beta} = RM$$

Casos aplicado

- Identificar factores asociados a que el niño tenga alergia.

Caso | Interpretación | Supuestos | Evaluación de Supuestos

- Factores asociados al número de visitas médicas anuales.

```
md_visit <- import("rwm5yr.dta") %>%
  characterize()
```

- Especificación del modelo

```
mod <- glm(numcig ~ female + age + edlevel + outwork + hhninc + year,
           family = poisson(link = "log"),
           data = md_visit)
summary(mod)
```

```

Call:
glm(formula = numcig ~ female + age + edlevel + outwork + hhninc + 
    year, family = poisson(link = "log"), data = md_visit)

Deviance Residuals: 
    Min       1Q   Median       3Q      Max  
-3.2312  -1.0334  -0.1454   0.5515   4.1289  

Coefficients:
                    Estimate Std. Error z value Pr(>|z|)    
(Intercept)         8.998893   7.556099   1.191 0.233675    
female              0.220202   0.012252  17.973  < 2e-16 ***
age                 0.016771   0.000498  33.675  < 2e-16 ***
edlevelGrad School -0.108043   0.038168  -2.831 0.004644 ** 
edlevelHS grad      0.111976   0.033475   3.345 0.000823 ***
```

- Presentación con intervalos de confianza y exponenciada (OR):

```r
library(broom)
mod %>%
  tidy(conf.int = TRUE, exponentiate = TRUE)
```

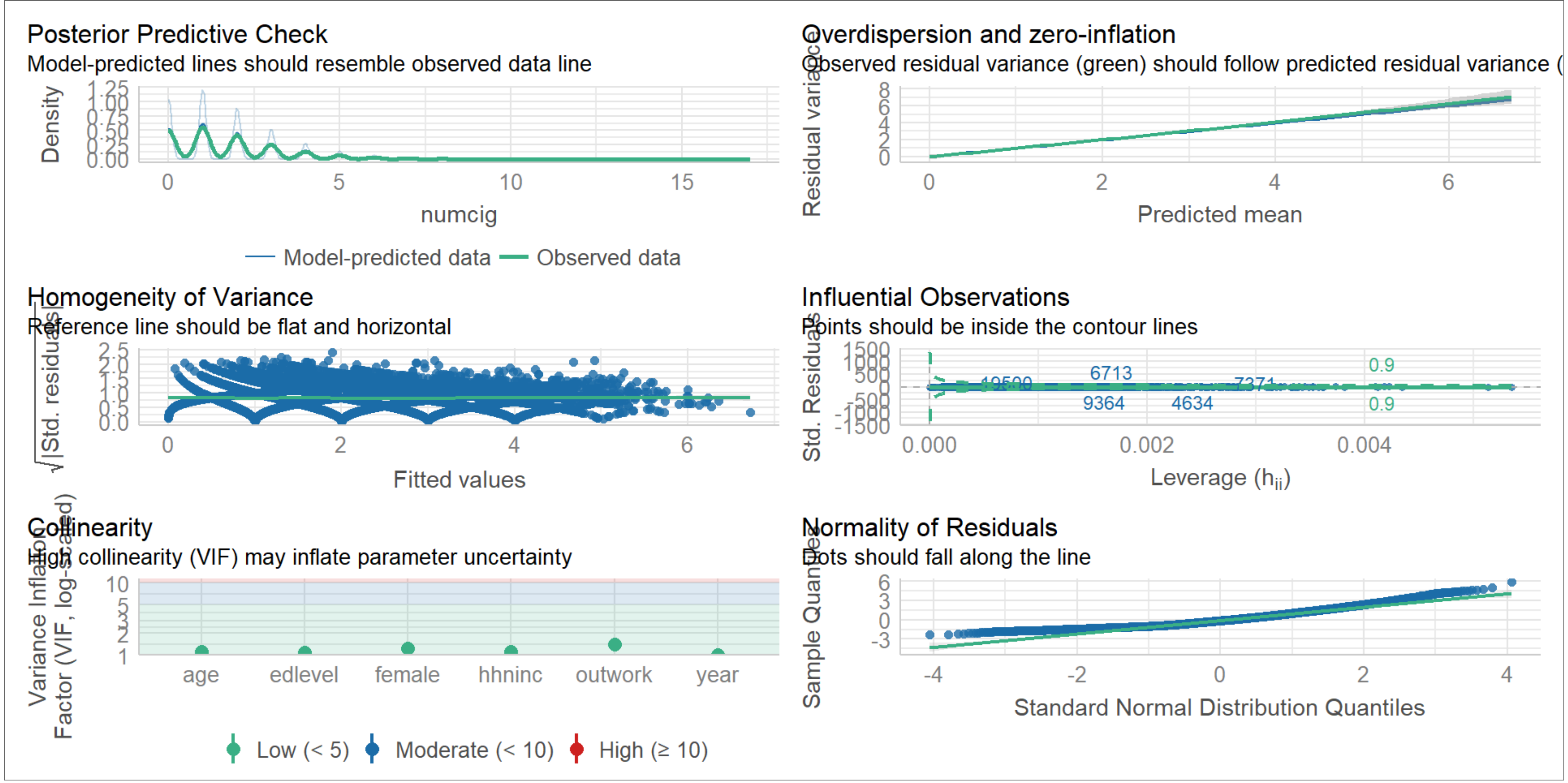
```
# A tibble: 9 × 7
  term               estimate std.error statistic   p.value conf.low conf.high
  <chr>                 <dbl>     <dbl>     <dbl>     <dbl>    <dbl>     <dbl>
1 (Intercept)        8094.     7.56          1.19 2.34e-  1  0.00300  2.19e+10
2 female                1.25   0.0123       18.0  3.19e- 72  1.22     1.28e+ 0
3 age                   1.02   0.000498     33.7  1.33e-248  1.02     1.02e+ 0
4 edlevelGrad School    0.898  0.0382       -2.83 4.64e-  3  0.833    9.67e- 1
5 edlevelHS grad        1.12   0.0335        3.35 8.23e-  4  1.05     1.19e+ 0
6 edlevelNot HS grad    1.21   0.0234        8.24 1.69e- 16  1.16     1.27e+ 0
7 outwork               1.17   0.0130       12.0  2.85e- 33  1.14     1.20e+ 0
8 hhninc                0.735  0.00496     -62.1  0          0.728    7.42e- 1
9 year                  0.996  0.00381      -1.15 2.52e-  1  0.988    1.00e+ 0
```

```
+ `female`: El número medio de visitas anuales al médico en mujeres fue 20% veces más el de los hombres (RM = 1.25;
IC95% 1.22 a 1.28; p < 0.001)

+ `age`: Por cada incremento de la edad en un año, el número medio de visitas anuales al médico se incrementa en 1%
(RM = 1.017; IC95% 1.016 a 1.018; p < 0.001).
```

- Linealidad del $log(y_i)$ respecto a la combinación lineal de predictores.
- Observaciones son independientes.
- Y_i sigue distribución de Poisson.
- No problemas de regresión:
  - No puntos influyentes
  - No colinealidad: Solo cuando esta es un problema.
- Supuestos específicos si se busca generalizar a poblaciones conocidas, hacer inferencias causales o ambas.

```r
library(performance)
check_model(mod)
```



Posterior Predictive Check
Model-predicted lines should resemble observed data line
Density
numcig
Model-predicted data
Observed data
Overdispersion and zero-inflation
Observed residual variance (green) should follow predicted residual variance (
Residual variance
Predicted mean
Homogeneity of Variance
Reference line should be flat and horizontal
√|Std. residuals|
Fitted values
Influential Observations
Points should be inside the contour lines
Std. Residuals
Leverage (h_{ii})
6713
9364
4634
7371
0.9
0.9
Collinearity
High collinearity (VIF) may inflate parameter uncertainty
Variance Inflation Factor (VIF, log-scaled)
age
edlevel
female
hhninc
outwork
year
Low (< 5)
Moderate (< 10)
High (≥ 10)
Normality of Residuals
Dots should fall along the line
Sample Quantiles
Standard Normal Distribution Quantiles

Agenda

1. Introducción al modelado de regresión
2. Modelo de Regresión Lineal
3. El Modelo Lineal Generalizado
4. La regresión (log) Poisson
5. **Tablas de regresión reproducibles con {gtsummary}**

Tablas de regresión lineal reproducible

- Podemos usar la librería {gtsummary} para esto.
- Veamos un ejemplo.

```
datos <- import("hb.dta") %>%
  characterize()
```

- Podemos reportar la tabla de regreion multivarible de la siguiente manera:
  - Primero realizamos el modelo:

```
mod <- lm(hb ~ age + sex, data = datos)
mod %>%
  tidy(conf.int = TRUE)
```

```
# A tibble: 3 × 7
  term        estimate std.error statistic   p.value conf.low conf.high
  <chr>          <dbl>     <dbl>     <dbl>     <dbl>    <dbl>     <dbl>
1 (Intercept)  11.0     0.0505       218.  0           10.9      11.1
2 age           0.0110  0.000341      32.3 5.81e-218   0.0104    0.0117
3 sex          -0.474   0.0258       -18.3 8.44e- 74  -0.524    -0.423
```

Tablas de regresión lineal reproducible

- Se puede crear una tabla de regresión multivariable con la función `tbl_regression()` de `{gtsummary}`:

```r
tabla_multi <- mod %>%
  tbl_regression()

tabla_multi
```

Characteristic	Beta	95% CI[1]	p-value
age	0.01	0.01, 0.01	<0.001
sex	-0.47	-0.52, -0.42	<0.001

[1] CI = Confidence Interval

Tablas de regresión lineal reproducible

- Podemos hacer la tabla de regresiones bivariada con la función `tbl_uvregression()` de `{gtsummary}`:

```r
tabla_univ <- datos %>%
  select(age, sex, hb) %>%
  tbl_uvregression(
    method = lm,
    y = hb
  )

tabla_univ
```

Characteristic	N	Beta	95% CI[1]	p-value
age	10,000	0.01	0.01, 0.01	<0.001
sex	10,000	-0.47	-0.53, -0.42	<0.001

[1] CI = Confidence Interval

Tablas de regresión lineal reproducible

- Luego, podemos fusionar ambas tablas en una sola con la función `tbl_merge()`:

```
tabla_final <- tbl_merge(list(tabla_univ, tabla_multi), tab_spanner = c("Modelos crudos", "Modelo ajustado"))

tabla_final
```

	Modelos crudos				Modelo ajustado		
Characteristic	**N**	**Beta**	**95% CI[1]**	**p-value**	**Beta**	**95% CI[1]**	**p-value**
age	10,000	0.01	0.01, 0.01	<0.001	0.01	0.01, 0.01	<0.001
sex	10,000	-0.47	-0.53, -0.42	<0.001	-0.47	-0.52, -0.42	<0.001

[1] CI = Confidence Interval

Tablas de regresión lineal reproducible

- Podemos exportarlo a MS Word para post-procesamiento y reporte:

```
tabla_final %>%
  as_flex_table() %>%
  save_as_docx(path = "Tabla_Final.docx")
```

¡Gracias!
¿Preguntas?

https://github.com/psotob91
percys1991@gmail.com