

DIPLOMADO MÉTODOS

$y = g(x)$

Secant Lines

$$f'(x) = \lim_{h \rightarrow 0} \frac{f(x+h) - f(x)}{h}$$
$$f'(x) = \lim_{h \rightarrow 0} \frac{(x+h)^2 - x^2}{h}$$
$$= \lim_{h \rightarrow 0} \frac{x^2 + 2xh + h^2 - x^2}{h}$$
$$= \lim_{h \rightarrow 0} \frac{2xh + h^2}{h}$$
$$= \lim_{h \rightarrow 0} h(2x + h)$$
$$= 1$$
$$f(x) = \dots$$
$$f(a) = \dots$$
$$f(a) = \dots$$

REGRESIÓN LINEAL: DE LO SIMPLE A LO MÚLTIPLE

MÓDULO 5

REGRESIÓN
LINEAL MÚLTIPLE

PONENTE: DRA.
ANA ESCOTO

10 DE JUNIO

MODELOS LINEALES II - REGRESIÓN LINEAL MULTIPLE



Repaso regression lineal simple



Supuestos.



Estimación e
interpretación de coeficientes.



Análisis de varianza.



Bondad de ajuste.



Limitantes.



Extensiones de la regresión lineal

LA LÍNEA DE MÍNIMOS CUADRADOS ORDINARIOS

- La idea de mínimos cuadrados. Para cada observación, hallar la distancia vertical de cada punto de la gráfica de dispersión de una línea de regresión. La línea de regresión de mínimos cuadrados hace que la suma de los cuadrados de estas distancias tan pequeñas como sea posible.

TERMINOS (WOOLRIDGE)

TABLA 2.1

Terminología en la regresión simple

<i>y</i>	<i>x</i>
Variable dependiente	Variable independiente
Variable explicada	Variable explicativa
Variable de respuesta	Variable de control
Variable predicha	Variable predictora
Regresando	Regresor

- Para encontrar la pendiente y la intersección y de la mejor línea de ajuste, se usa la fórmula:

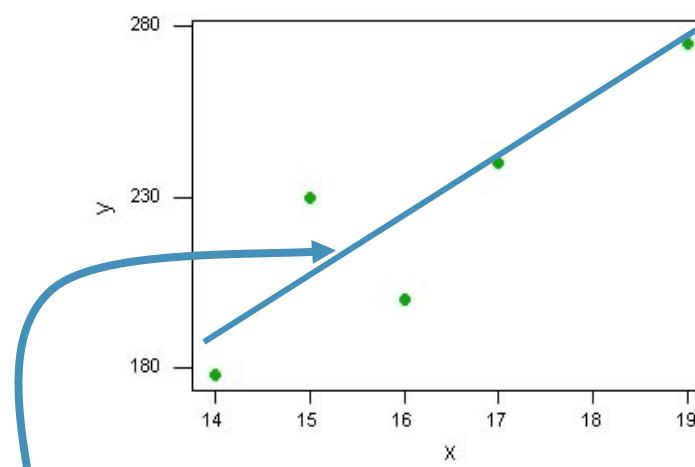
$$\hat{\beta}_1 = r \frac{s_x}{s_y}$$

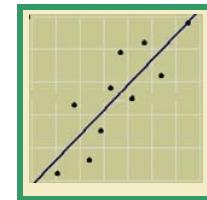
$$\hat{\beta}_0 = \bar{y} + \hat{\beta}_1 \bar{x}$$

La línea de mínimos

cuadrados es

$$y = \hat{\beta}_0 + \hat{\beta}_1 x$$





- **Modelo determinístico:** $y = \beta_0 + \beta_1 x$

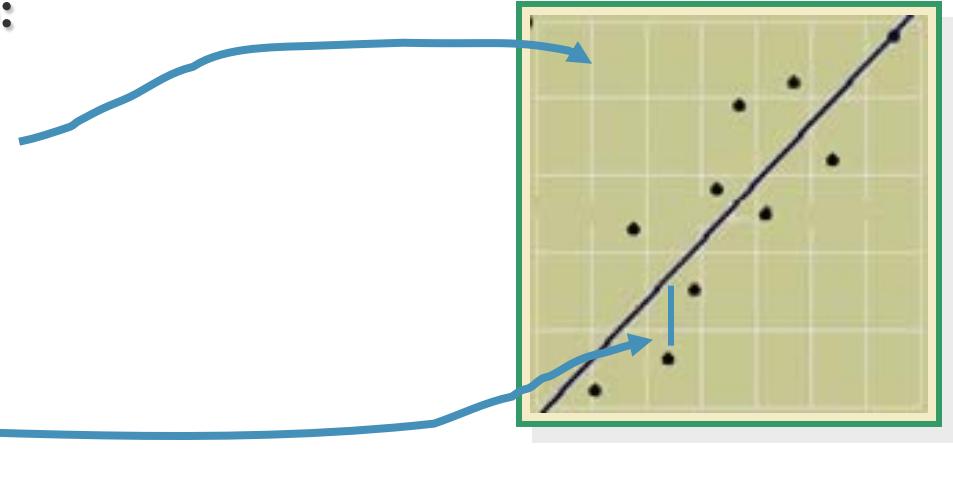
- **Modelo probabilístico:**

$y = \text{El modelo determinístico} + \text{error aleatorio}$

$$y = \beta_0 + \beta_1 x + \varepsilon$$

$$\textcolor{blue}{y} = \beta_0 + \beta_1 x + u.$$

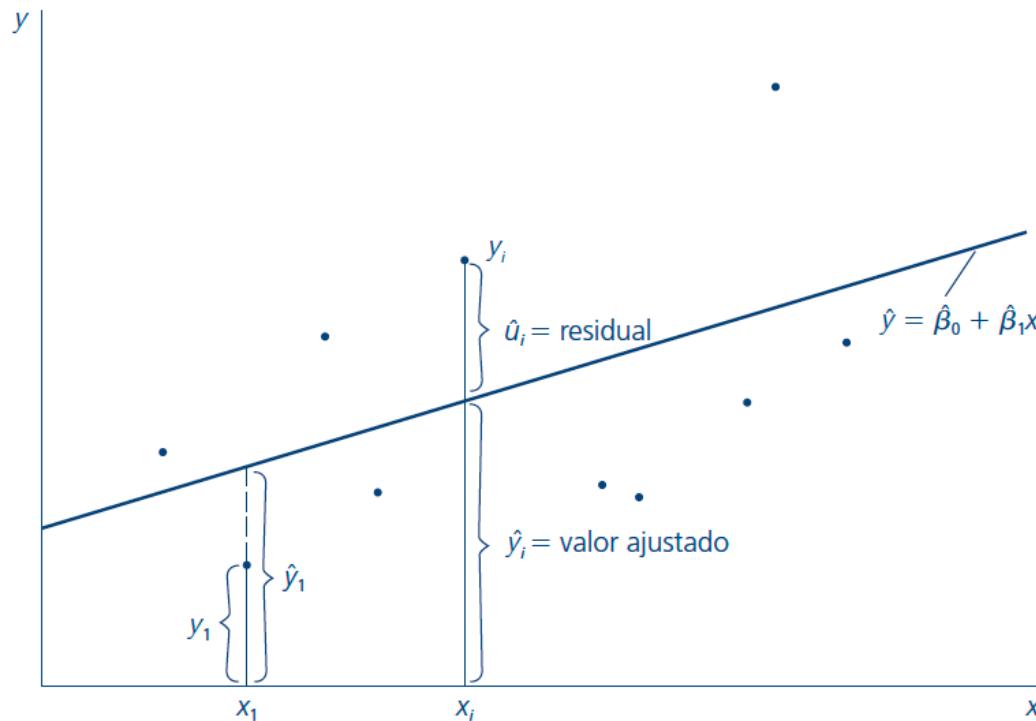
- Dado que las mediciones bivariadas que observamos generalmente no caen exactamente en una línea recta, elegimos utilizar:
- **Modelo probabilístico:**
 - $y = \alpha + \beta x + \varepsilon$
 - $E(y) = \alpha + \beta x$



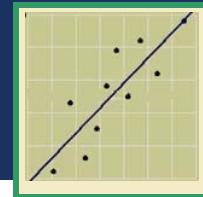
**Los puntos se desvían de la
línea de medias por una
cantidad ε
donde ε tiene una media 0 y
una varianza constante σ^2 .**

FIGURA 2.4

Valores ajustados y residuales.



EL ERROR ALEATORIO



- La línea de medias que viene de , $E(y) = \alpha + \beta x$, describes el valor promedio de cualquier valor para y y para cualquier valor fijo de x .
- La población de mediciones se genera cuando "y" se desvía de la línea de población por " ε ". Estimamos " α " y " β " usando información muestral.

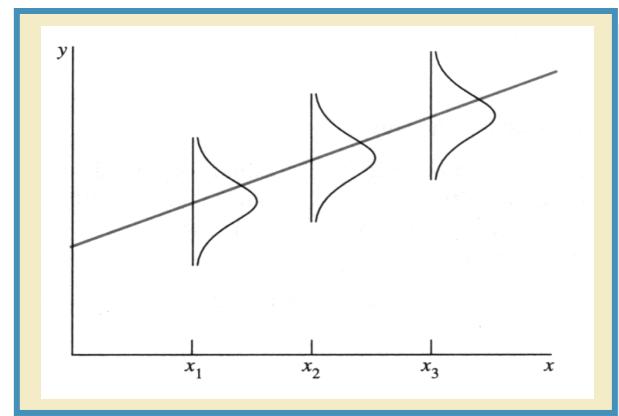
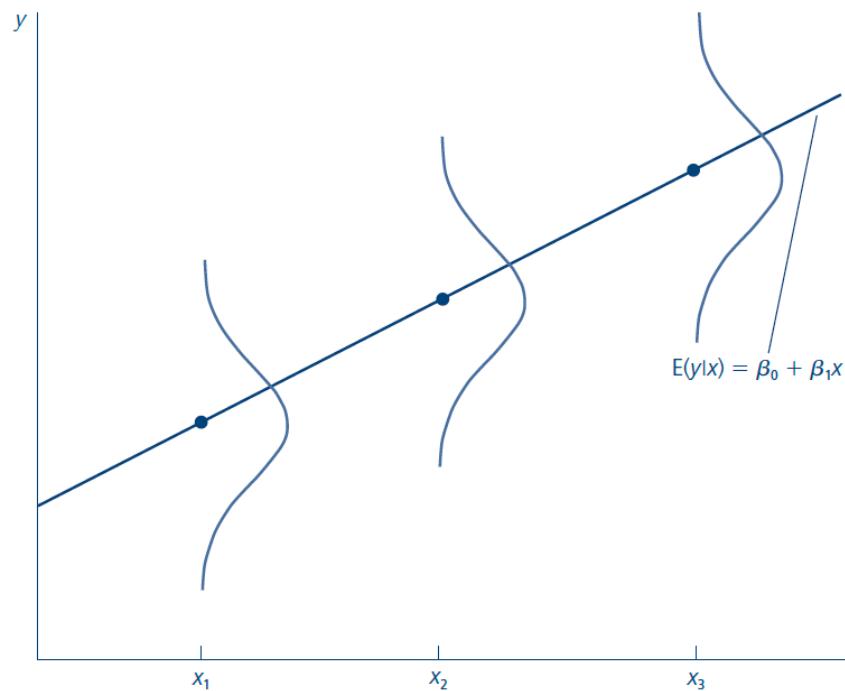
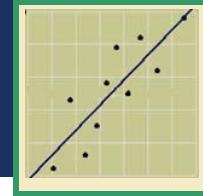


FIGURA 2.1

$E(y|x)$ como función lineal de x .

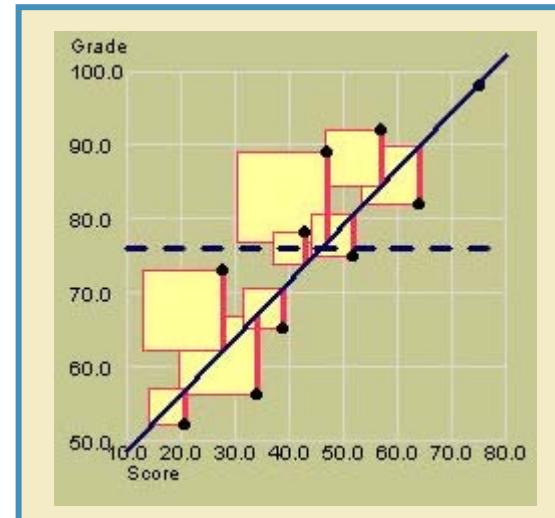


EL MÉTODO DE MÍNIMOS CUADRADOS



- La ecuación de la línea de mejor ajuste se calcula utilizando un conjunto de n pares (x_i, y_i) .

•Elegimos nuestras estimaciones "a" y "b" para estimar " α " y " β " de modo que se minimicen las distancias verticales de los puntos desde la línea.

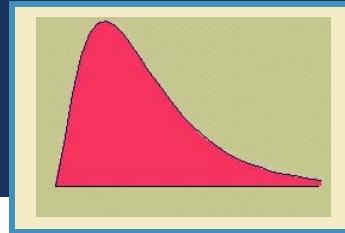


$$\hat{y} = a + bx$$

- Hay que escoger la que minimice los errores al cuadrado

$$\sum (y - \hat{y})^2 = \sum (y - a - bx)^2$$

EL TEST F



- Se puede probar la utilidad general del modelo usando una prueba F. Si el modelo es útil, la media de los errores de la regresión será grande en comparación con la variación inexplicable, la media de los cuadrados los errores.

To test H_0 : model is useful in predicting y

$$\text{Test Statistic } F = \frac{M\ SR}{M\ SE}$$

Reject H_0 if $F > F_\alpha$ with 1 and $n - 2$ df.

Esta prueba es
exactamente
equivalente a la
prueba t, con $t^2 = F$.

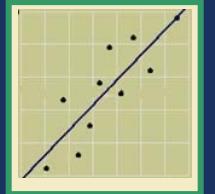
MEDIR LA FUERZA DE LA RELACIÓN

- Si la variable independiente x es útil para predecir y, querrá saber qué tan bien se ajusta el modelo.
- La fuerza de la relación entre x e y puede medirse usando:

$$\text{Correlation coefficient : } r = \frac{S_{xy}}{\sqrt{S_{xx} S_{yy}}}$$

$$\text{Coefficient of determination : } r^2 = \frac{S_{xy}^2}{S_{xx} S_{yy}} = \frac{\text{SSR}}{\text{TotalSS}}$$

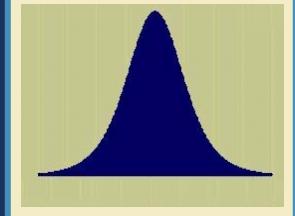
PROBANDO LA UTILIDAD DEL MODELO



- La primera pregunta a hacer es si la variable independiente x sirve en la predicción de y.
- Si no lo es, entonces el valor de y no cambia, independientemente del valor de x. Esto implica que la pendiente de la recta, b, es cero.

$$H_0 : \beta = 0 \quad \text{versus} \quad H_a : \beta \neq 0$$

PROBANDO LA UTILIDAD DEL MODELO



- El estadístico de prueba es función de β , nuestra mejor estimación de β . Usando el error estándar como la mejor estimación de la variación aleatoria σ^2 , obtenemos un estadístico t .

$$\text{Teststatistic } t = \frac{b - 0}{\sqrt{\frac{MSE}{S_{xx}}}} \text{ which has a } t \text{ distribution}$$

with $df = n - 2$ or a confidence interval: $b \pm t_{\alpha/2} \sqrt{\frac{MSE}{S_{xx}}}$

Ojo
El error
estándar
cambia para
 β_0 o α

ES DECIR

$$t = \frac{\widehat{\beta}_0 - 0}{SE(\widehat{\beta}_0)}$$

$$t = \frac{\widehat{\beta}_1 - 0}{SE(\widehat{\beta}_1)}$$

MEDIR LA FUERZA DE LA RELACIÓN

Como Total SS = SSR + SSE, r^2 mide

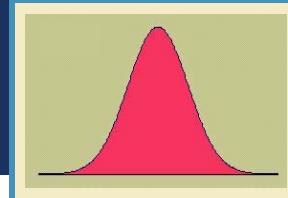
- la proporción de la variación total en las respuestas que se puede explicar usando la variable independiente x en el modelo.
- El porcentaje de reducción de la variación total usando la ecuación de regresión en vez de usar la media de la muestra y-bar para estimar y.

$$r^2 = \frac{\text{SSR}}{\text{TotalSS}}$$

INTERPRETACIÓN DE UNA "REGRESIÓN SIGNIFICATIVA"

- Incluso si no se rechaza la hipótesis nula de que la pendiente de la recta es igual a 0, no significa necesariamente que y e x no tengan una relación.
- **Tipo II error**- declarando falsamente que la pendiente es 0 y que x e y no están relacionados.
- Puede suceder que y e x estén perfectamente relacionados de una manera **no lineal**.

CHECANDO LOS SUPUESTOS DE LA REGRESIÓN



•Recuerde que los resultados de un análisis de regresión sólo son válidos cuando se han satisfecho los supuestos necesarios.

1. La relación entre x y y es lineal, dada por $y = \alpha + \beta x + \varepsilon$.
2. Los términos de error aleatorio ε son independientes y, para cualquier valor de x , tienen una distribución normal con media 0 y varianza σ^2 .

Supuesto RLS.1 (Lineal en los parámetros)

En el modelo poblacional, la variable dependiente, y , está relacionada con la variable independiente, x , y con el error (o perturbación), u , de la manera siguiente:

$$y = \beta_0 + \beta_1 x + u,$$

donde β_0 y β_1 son los parámetros poblacionales correspondientes al intercepto y a la pendiente, respectivamente.

Supuesto RLS.2 (Muestreo aleatorio)

Se tiene una muestra aleatoria de tamaño n , $\{(x_i, y_i): i = 1, 2, \dots, n\}$, que sigue el modelo poblacional del supuesto RLS.1.

Supuesto RLS.3 (Variación muestral en la variable explicativa)

Los valores muestrales de x , a saber, $\{x_i, i = 1, \dots, n\}$, no son todos iguales.

Supuesto RLS.4 (Media condicional cero)

Dado cualquier valor de la variable explicativa, el valor esperado de error u es cero. En otras palabras,

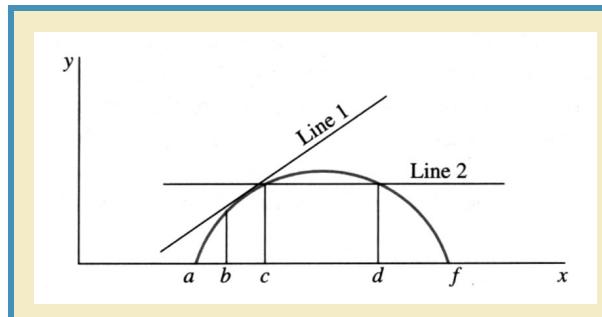
$$\text{E}(u|x) = 0.$$

Supuesto RLS.5 (Homocedasticidad)

Para cualquier valor de la variable explicativa, el error u tienen la misma varianza. En otras palabras,

$$\text{Var}(u|x) = \sigma^2.$$

PRECAUCIONES



- Es posible que haya ajustado el modelo equivocado
 - 1. Extrapolación
 - Se predice valores de y fuera del rango de los datos ajustados.
 - 2. Causalidad
 - No concluya que x causa y . Puede haber una variable desconocida en el trabajo! “cofounding”

Resumen de las formas funcionales en las que se emplean logaritmos

Modelo	Variable dependiente	Variable independiente	Interpretación de β_1
Nivel-nivel	y	x	$\Delta y = \beta_1 \Delta x$
Nivel-log	y	$\log(x)$	$\Delta y = (\beta_1 / 100) \% \Delta x$
Log-nivel	$\log(y)$	x	$\% \Delta y = (100 \beta_1) \Delta x$
Log-log	$\log(y)$	$\log(x)$	$\% \Delta y = \beta_1 \% \Delta x$

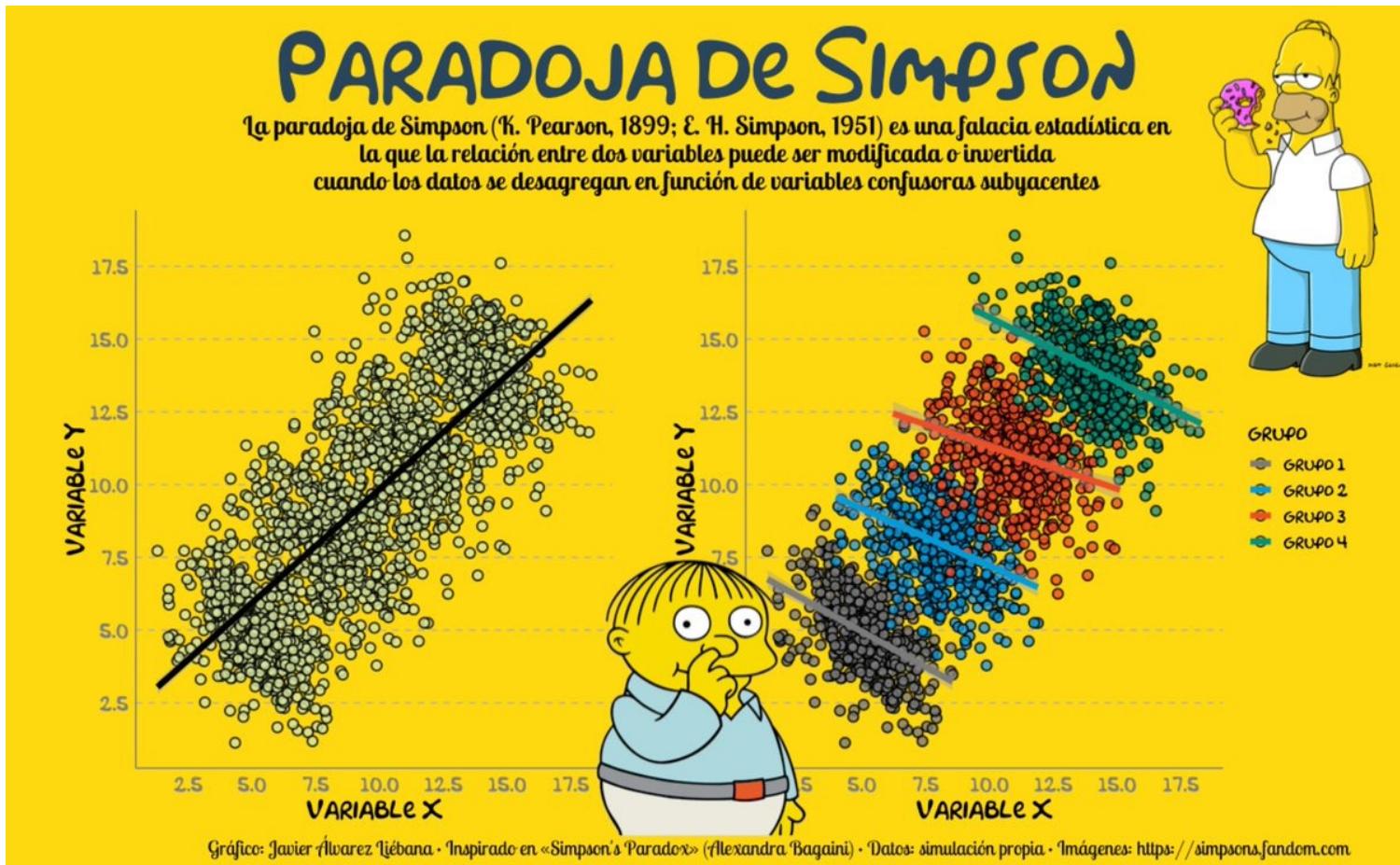
MÁS DE UNA VARIABLE EXPLICATIVA



REGRESIÓN MÚLTIPLE

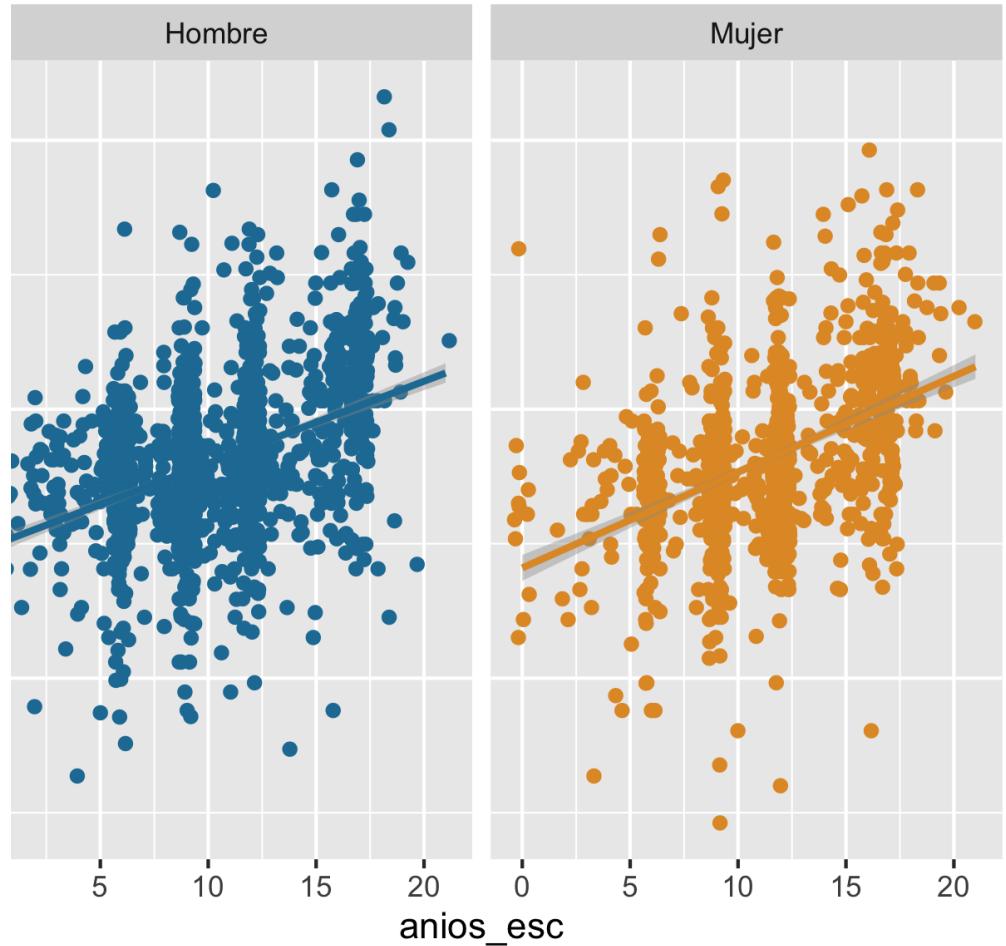
- La regresión múltiple **es una técnica estadística** que se puede utilizar para analizar la relación entre una sola variable dependiente y varias variables independientes.
- El **objetivo** del análisis de regresión múltiple es utilizar las variables independientes cuyos valores se conocen para predecir el valor del valor dependiente único.
- Se **pondera cada valor predictor**, y los **pesos indican su contribución relativa** a la predicción general.
- En 1908, Pearson usó por primera vez “regresión múltiple”
- Así que estamos aprendiendo una técnica de más de 100 años

AÑADIENDO UNA VARIABLE CATEGÓRICA



UNA VARIABLE CATEGÓRICA HACE GRUPOS

- Y podemos analizar la regresión lineal a través de esos grupos.
- El caso más sencillo será pensar que tenemos dos grupos.

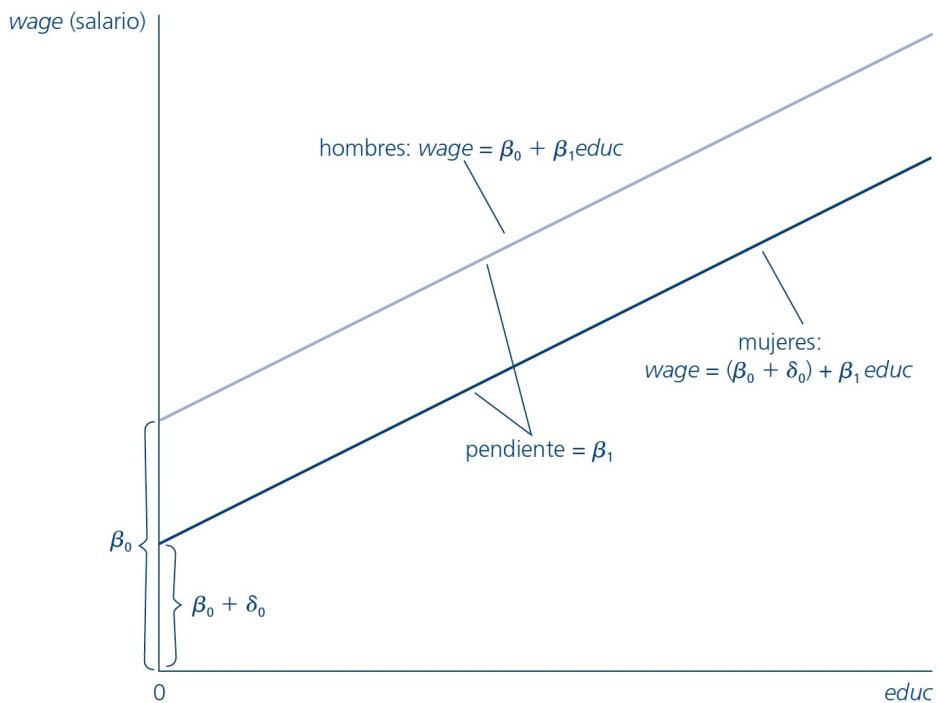


¿Entonces es como tener dos modelos?

AÑADIENDO UNA VARIABLE CATEGÓRICA

FIGURA 7.1

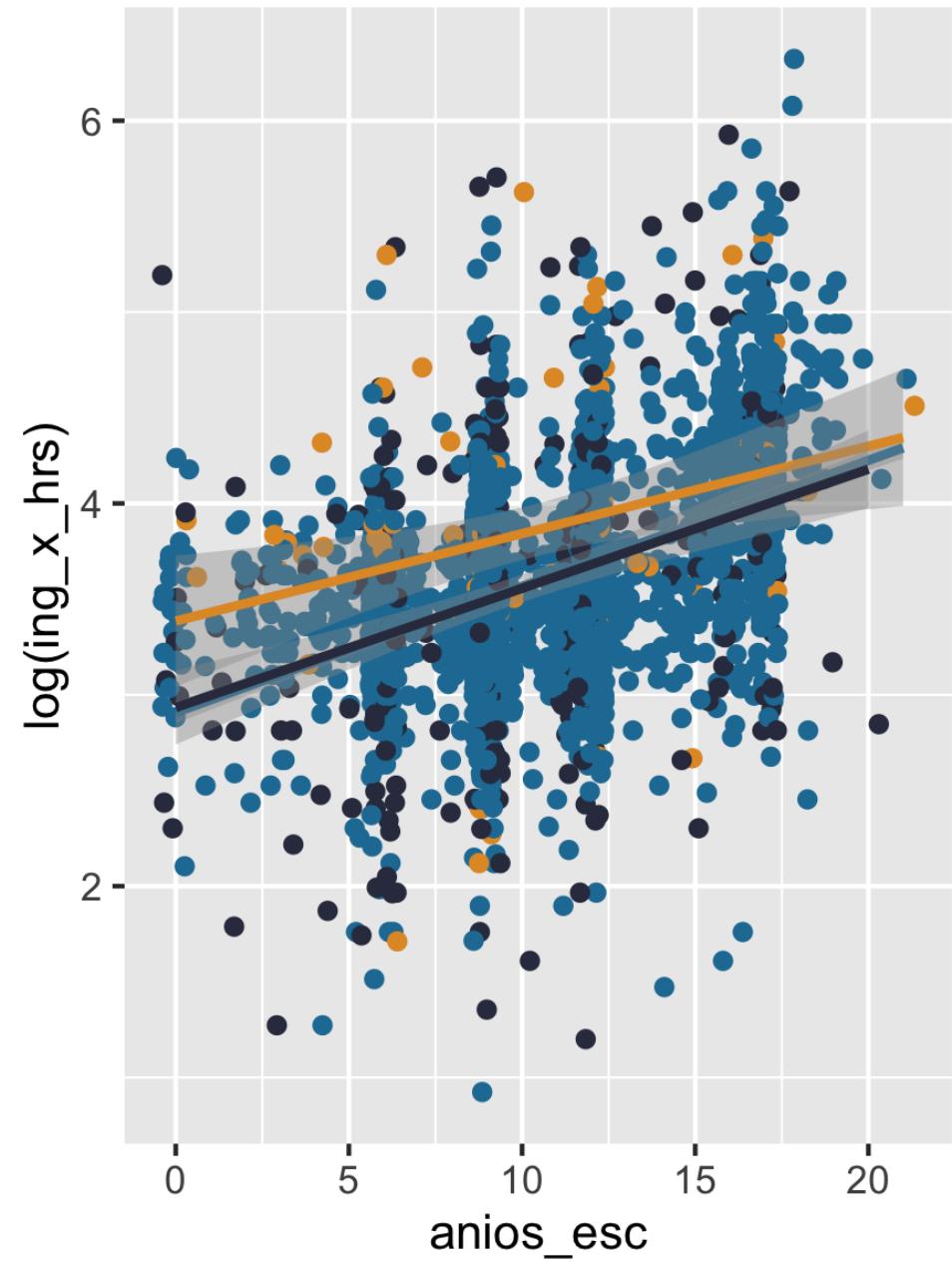
Gráfica de $wage = \beta_0 + \delta_0 female + \beta_1 educ$ en la que $\delta_0 < 0$.



- Una variable categórica nos divide a la población en grupos
- Una relación puede variar a lo largo de los grupos.
- Por el momento sólo veremos cambios en el intercepto

$$y = \beta_0 + \beta_1 x_1 + \delta_2 x_2 + \epsilon$$

- Sea x_2 , una variable “*dummy*”. Es decir es categórica pero la codificamos como un número (no es realmente numérica). Puede tomar valores de 0 y 1.
- Sea $x_2=1$ cuando es mujer, $x_2=0$ cuando es varón
- Cuando $x_2=1$ el modelo se escribe
 - $y = \beta_0 + \beta_1 x_1 + \delta_2 * 1 + \epsilon$
 - $y = (\beta_0 + \delta_2) + \beta_1 x_1 + \epsilon$
- Cuando $x_2=2$ el modelo se escribe
 - $y = \beta_0 + \beta_1 x_1 + \delta_2 * 0 + \epsilon$
 - $y = \beta_0 + \beta_1 x_1 + \epsilon$
- Si restamos ambas ecuaciones: vemos que δ_2 es el cambio de 0 a 1 de x_2



as_label(pos_ocu)

- Trabajadores subordinados y remunerados
- Empleadores
- Trabajadores por cuenta propia

LA CATEGORÍA DE REFERENCIA

- El valor de nuestra variable categórica de x2 se comparaba con los varones, δ_2 es la diferencia de las remuneraciones contra las mujeres.
- Cuando tenemos más de una categoría, se deben introducir **k -1 dummies** (donde k es número de categorías)
- En este caso el modelo quedaría así:
- $y = \beta_0 + \beta_1 x_1 + \delta_2 \text{empleadores} + \delta_3 \text{cuentapropia} + \epsilon$
- Cuando no son empleadores ni cuenta propias, las variables dummies valen cero y esa será la categoría de referencia

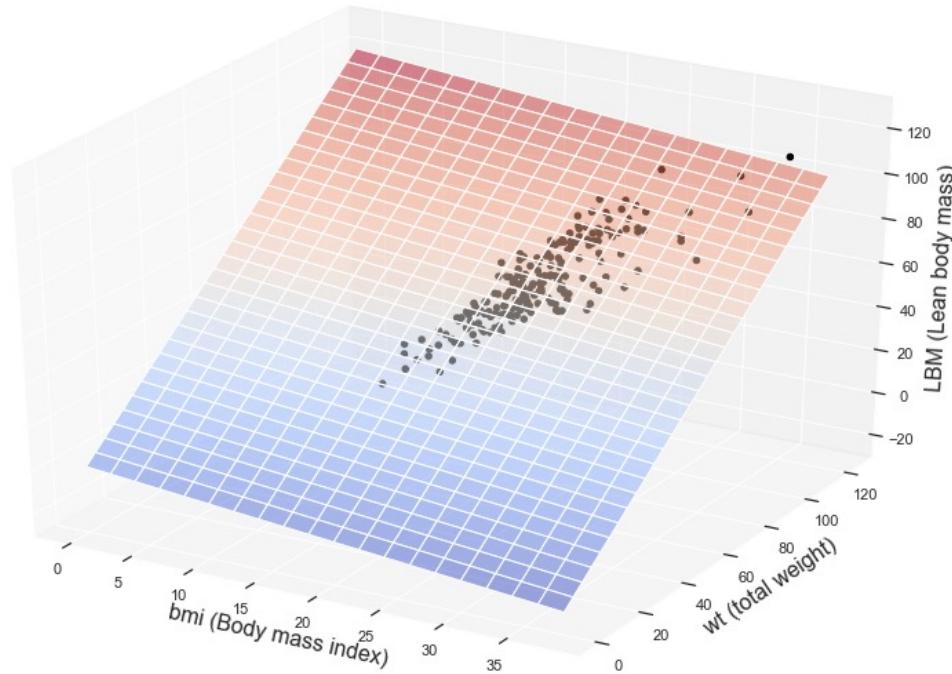
¿MEJORAN EL MODELO LA INCLUSIÓN DE VARIABLES DUMMIES?

- Al separar el análisis en grupos, esperamos modelar una relación que se repite en grupos con diferencias en los interceptos.
- Podemos hacer el análisis por variable verificando una prueba t para cada estimador δ
- Podemos también hacer una prueba F de ajuste, comparando los ajustes de los modelos (sin las dummies) y con las dummies.
- Estadístico de Chow.

$$F = \frac{[SRC_P - (SRC_1 + SRC_2)]}{SRC_1 + SRC_2} \cdot \frac{[n - 2(k + 1)]}{k + 1},$$

Asume homocedasticidad

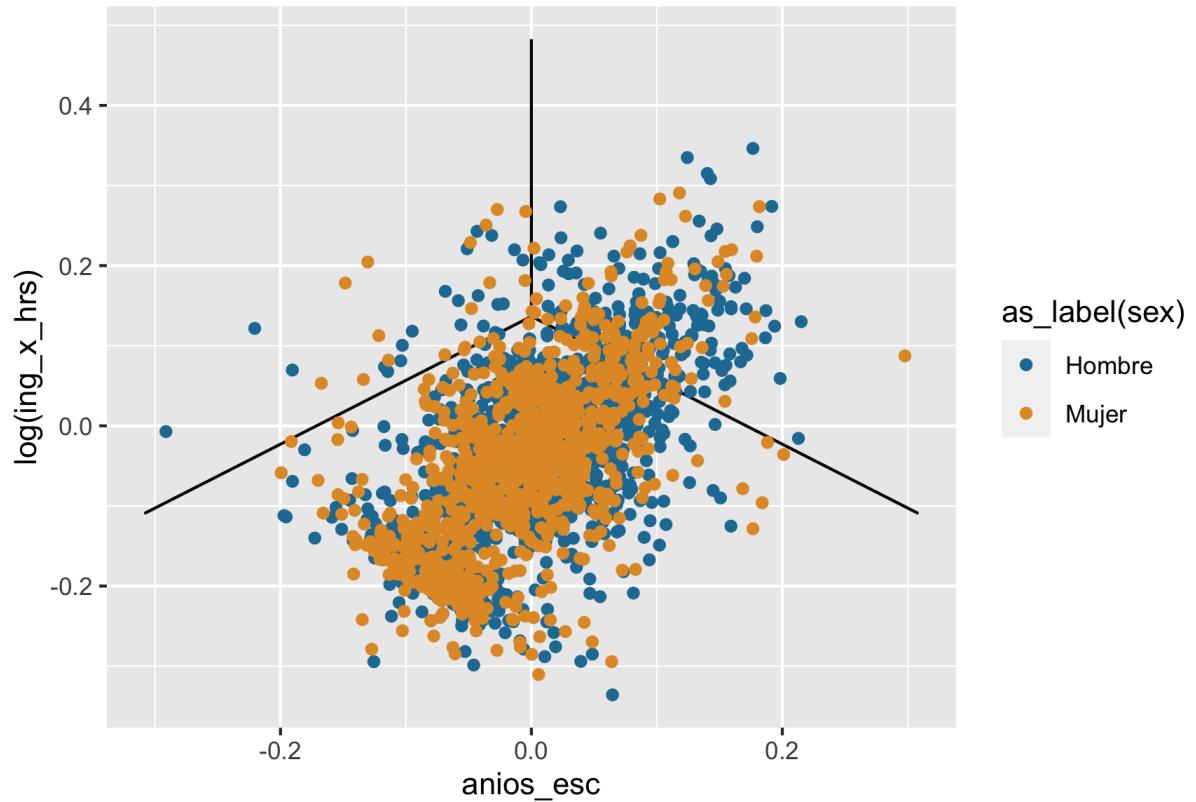
REGRESIÓN LINEAL CON DOS VARIABLES CUANTITATIVAS EXPLICATIVAS



Geométricamente, $\beta_0 + \beta_1X_1 + \beta_2x_2$ describe un plano

REGRESIÓN LINEAL CON DOS VARIABLES CUANTITATIVAS EXPLICATIVAS

- con nuestros datos



cada observación tiene una triple coordenada

INTERPRETACIÓN DE LA REGRESIÓN MÚLTIPLE

$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \dots + \hat{\beta}_k x_k$, de modo que

$\Delta \hat{y} = \Delta \hat{\beta}_1 x_1 + \Delta \hat{\beta}_2 x_2 + \dots + \Delta \hat{\beta}_k x_k$,

y si mantenemos x_2, \dots, x_k constantes, implica que

$\Delta \hat{y} = \Delta \hat{\beta}_1 x_1$, es decir, cada β tiene

una interpretación *ceteris paribus*

POR LO TANTO

- Por cada unidad que cambia x_1 , y cambia otra cantidad, pero lo demás se mantiene constante.
- Esto también sucede con las variables categóricas, pero con un cambio de 0 a 1.

AJUSTE DEL MODELO

Call:

```
lm(formula = log(ing_x_hrs) ~ anios_esc + sex + eda, data = mydata)
```

Residuals:

Min	1Q	Median	3Q	Max
-2.63819	-0.28954	-0.00834	0.29308	2.25535

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	2.5997575	0.0447779	58.059	< 2e-16 ***
anios_esc	0.0718597	0.0026085	27.548	< 2e-16 ***
sexMujer	-0.0680613	0.0200914	-3.388	0.000715 ***
eda	0.0073802	0.0007421	9.945	< 2e-16 ***

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5194 on 2854 degrees of freedom
Multiple R-squared: 0.2119, Adjusted R-squared: 0.2111
F-statistic: 255.9 on 3 and 2854 DF, p-value: < 2.2e-16

Ajuste individual

Ajuste global

AJUSTE DEL MODELO

```
> broom::tidy(model)
# A tibble: 4 × 5
  term      estimate std.error statistic   p.value
  <chr>     <dbl>    <dbl>     <dbl>     <dbl>
1 (Intercept) 2.60     0.0448     58.1     0
2 anios_esc    0.0719   0.00261    27.5  2.41e-148
3 sexMujer    -0.0681   0.0201    -3.39  7.15e- 4
4 eda         0.00738  0.000742    9.94  6.23e- 23
```

Ajuste individual

```
> broom::glance(model)
# A tibble: 1 × 12
  r.squared adj.r.squared sigma statistic   p.value      df logLik     AIC     BIC deviance df.residual
  <dbl>        <dbl> <dbl>     <dbl>     <dbl>     <dbl> <dbl>     <dbl>     <dbl>     <int>
1 0.212       0.211  0.519     256.  4.80e-147     3 -2181. 4372. 4402.     770.    2854
```

... with 1 more variable: nobs <int>

$$R^2 = 1 - (\text{SRC}/n)/(\text{STC}/n), \quad \text{Ajuste global}$$

SRC es la suma de los residuales cuadrados

STC es la suma total de cuadrados

$$\bar{R}^2 = 1 - [\text{SRC}/(n - k - 1)]/[\text{STC}/(n - 1)] \\ = 1 - \hat{\sigma}^2/[\text{STC}/(n - 1)],$$

PRUEBAS DE HIPÓTESIS ASOCIADAS

Individual

- $H_0: \beta_i = 0$
- $H_a: \beta_i \neq 0$

$$t = \frac{\text{estimador} - \text{valor hipotético}}{\text{error estándar}}$$

- Es una prueba bilateral
- ¿Por qué comparar contra cero?

Global

- Más de un parámetro es igual a cero.
- En el caso del ajuste, se comparan las varianzas residuales del modelo restringido (sin parámetros), contra el modelo con todos los parámetros

```
> model<-lm(log(ing_x_hrs)~ anios_esc + sex+ eda , data=mydata)
> model0<-lm(log(ing_x_hrs)~ 1 , data=mydata)
> anova(model0, model)
Analysis of Variance Table
```

	Model 1: log(ing_x_hrs) ~ 1	Model 2: log(ing_x_hrs) ~ anios_esc + sex + eda				
	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	2857	976.90				
2	2854	769.85	3	207.05	255.86 < 2.2e-16 ***	

						Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

SUPUESTOS DEL MODELO

$$y|\mathbf{x} \sim \text{Normal}(\beta_0 + \beta_1x_1 + \beta_2x_2 + \dots + \beta_kx_k, \sigma^2),$$

Supuesto RLM.1 (Lineal en los parámetros)

El modelo poblacional puede expresarse como

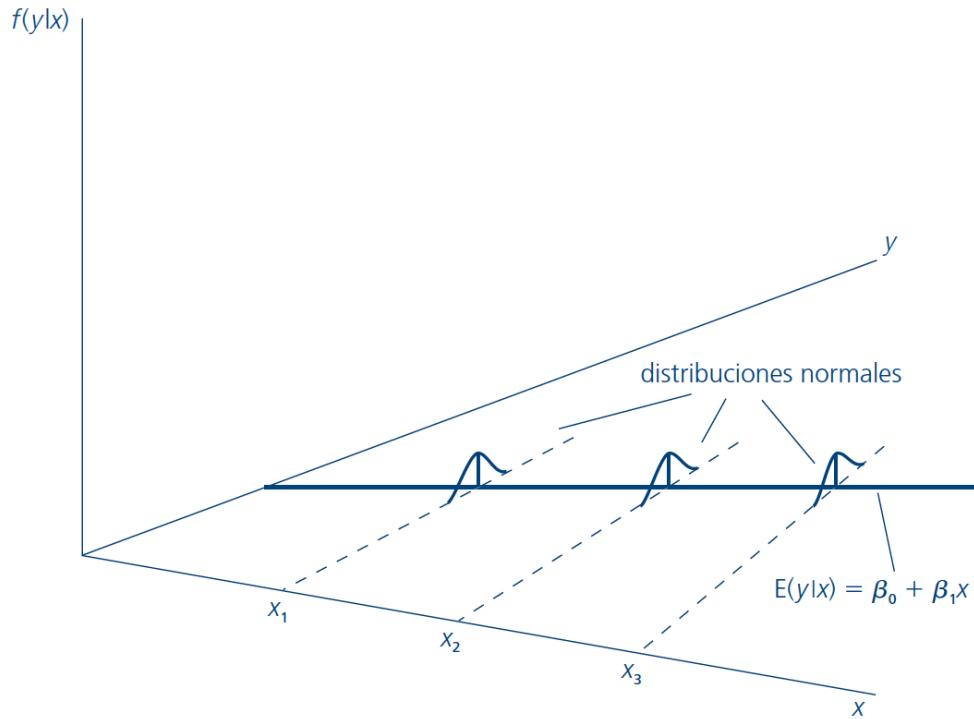
$$y = \beta_0 + \beta_1x_1 + \beta_2x_2 + \dots + \beta_kx_k + u,$$

donde $\beta_0, \beta_1, \dots, \beta_k$ son los parámetros (constantes) desconocidos de interés y u es el error aleatorio o término de perturbación no observable.

¿CÓMO SE VE?

FIGURA 4.1

Distribución normal homocedástica con una sola variable explicativa



LINEALIDAD EN LOS PARÁMETROS

- Como es evidente en el nombre de regresión **lineal** múltiple, se supone que la relación entre variables es lineal.
- En la práctica, esta suposición prácticamente nunca se puede confirmar.
- Sin embargo, como regla, es prudente revisar siempre **el diagrama de dispersión bivariado de las variables de interés**.
- **Si no hay linealidad**, y hay relaciones curvilíneas en las relaciones, uno puede considerar transformar las variables o permitir explícitamente el componente no lineal (esto lo veremos más adelante)

SUPUESTOS DEL MODELO

Supuesto RLM.2 (Muestreo aleatorio)

Se tiene un muestreo aleatorio con n observaciones, $\{(x_{i1}, x_{i2}, \dots, x_{ik}, y_i): i = 1, 2, \dots, n\}$, de acuerdo con el modelo poblacional del supuesto RLM.1.



Supuesto RLM.3 (No hay colinealidad perfecta)

En la muestra (y por tanto en la población), ninguna de las variables independientes es constante y no hay relaciones *lineales exactas* entre las variables independientes.

Supuesto RLM.4 (Media condicional cero)

El error u tiene un valor esperado de cero dados cualesquiera valores de las variables independientes.
En otras palabras,

$$E(u|x_1, x_2, \dots, x_k) = 0.$$

Supuesto RLM.5 (Homocedasticidad)

El error u tiene la misma varianza dado cualquier valor de las variables explicativas. En otras palabras,

$$\text{Var}(u|x_1, \dots, x_k) = \sigma^2.$$

EL ÚLTIMO SUUESTO

Supuesto RLM.6 (Normalidad)

El error poblacional u es *independiente* de las variables explicativas x_1, x_2, \dots, x_k y está distribuido normalmente, con media cero y varianza σ^2 : $u \sim \text{Normal}(0, \sigma^2)$.

EL PROBLEMA DE LA MULTICOLINEALIDAD: EL CASO DE LA MULTICOLINEALIDAD PERFECTA

Consideremos el modelo

$$Y_t = \beta_1 + \beta_2 X_{t2} + \beta_3 X_{t3} + u_t,$$

donde $X_{t2} - X_{t3} = 1$.

Entonces, sin más que sustituir $X_{t2} = 1 + X_{t3}$ en el modelo original:

$$\begin{aligned} Y_t &= \beta_1 + \beta_2 \cdot (1 + X_{t3}) + \beta_3 X_{t3} + u_t \\ &= (\beta_1 + \beta_2) + (\beta_2 + \beta_3) \cdot X_{t3} + u_t, \end{aligned}$$

obtenemos que las combinaciones lineales estimables de los parámetros originales son:

$$\beta_1 + \beta_2, \quad \beta_2 + \beta_3.$$

Por ello no introducimos todas las dummies

EL PROBLEMA DE LA MULTICOLINEALIDAD: EL CASO DE LA MULTICOLINEALIDAD PERFECTA

- La **multicolinealidad exacta o perfecta** hace referencia a la existencia de una relación lineal exacta entre dos o más variables independientes
- Por lo general los paquetes estadísticos no computarán el modelo con esas variables (R) o botarán uno (STATA)

```
> mydata$sex1<-mydata$sex==1  
> mydata$sex2<-mydata$sex==2  
>  
> lm(log(ing_x_hrs)~ anios_esc + sex1 + sex2 , data=mydata)
```

Call:

```
lm(formula = log(ing_x_hrs) ~ anios_esc + sex1 + sex2, data = mydata)
```

Coefficients:

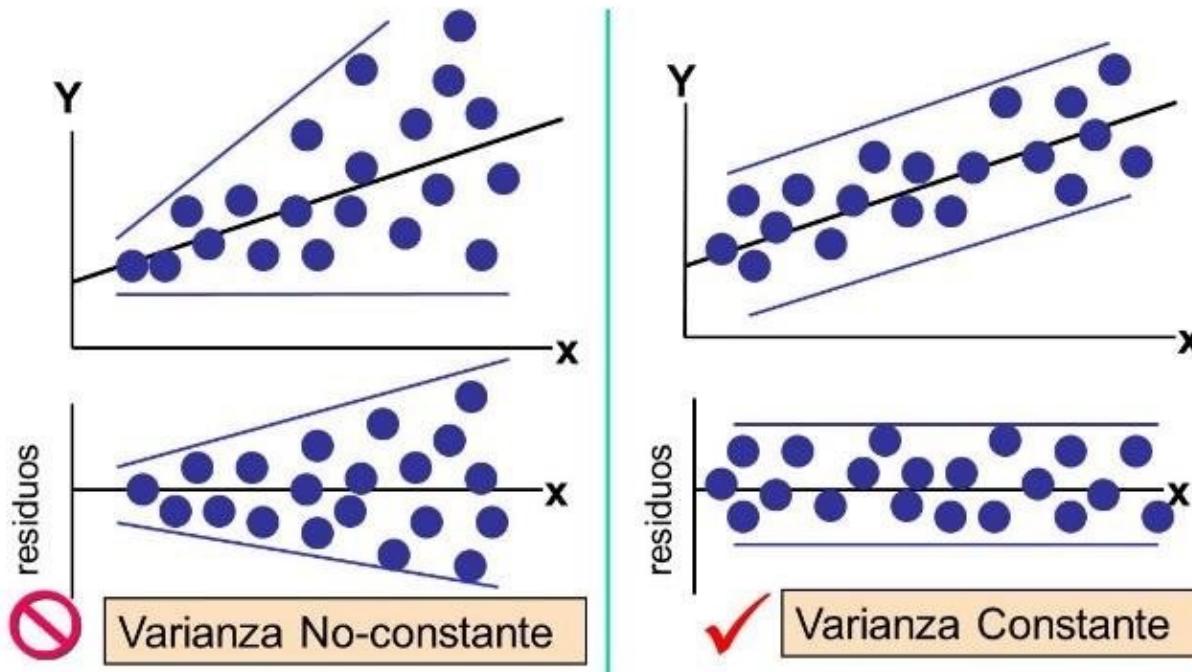
(Intercept)	anios_esc	sex1TRUE	sex2TRUE
2.93011	0.06442	NA	NA

EL PROBLEMA DE LA MULTICOLINEALIDAD: EL CASO DE LA MULTICOLINEALIDAD APROXIMADA

- La **multicolinealidad aproximada** hace referencia a la existencia de una relación lineal aproximada entre dos o más variables independientes
- Esto trae los siguientes problemas:
 - Las varianzas de los estimadores son muy grandes.
 - Al efectuar contrastes de significación individual no se rechaza la hipótesis nula, mientras que al realizar contrastes conjuntos sí. (pruebas t vs pruebas F)
 - Los coeficientes estimados serán muy sensibles ante pequeños cambios en los datos.
 - Un coeficiente de determinación elevado.

LA HETEROCEDASTICIDAD

Análisis de la Homoscedasticidad



LA HETEROCEDASTICIDAD

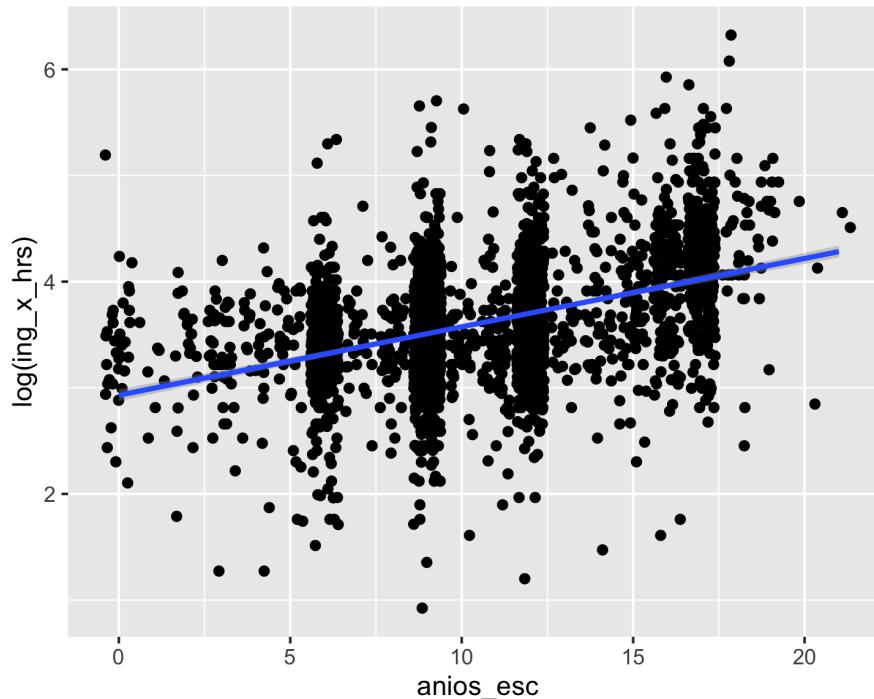
- El problema de una varianza no constante tiene que ver también con los errores estándar y las pruebas de especificación. No obstante hay varias formas de solucionarlo
- Se identifican con varias pruebas estadísticas: Brusch-Pagan o estadístico de White o de manera gráfica

EXTENSIONES DE LA REGRESIÓN LINEAL

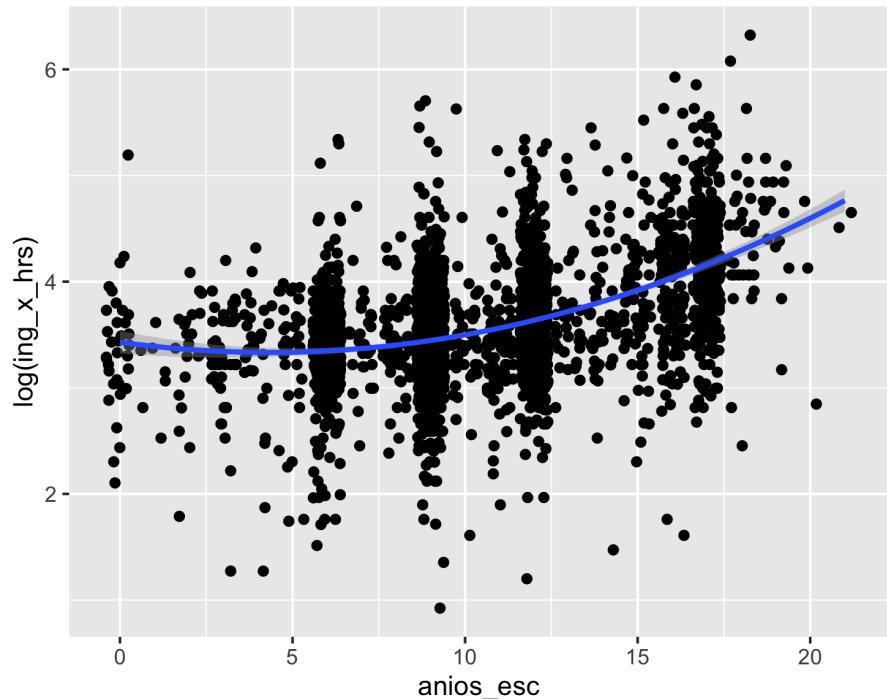
- Transformaciones logarítmicas
- Modelos de más de primer orden.
- Interacciones
- Elasticidad

¿CUÁL SE AJUSTA MEJOR?

Lineal



Cuadrático



RESPUESTA

```
> anova(model, model2)
Analysis of Variance Table

Model 1: log(ing_x_hrs) ~ anios_esc + sex + eda
Model 2: log(ing_x_hrs) ~ anios_esc + I(anios_esc^2) + sex + eda
Res.Df   RSS Df Sum of Sq      F    Pr(>F)
1     2854 769.85
2     2853 749.58  1     20.269 77.147 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

PRÁCTICA

- [RStudio Cloud](#)
- [curso \(aniuxa.github.io\)](#)

TODO SE PUEDE COMPRENDER COMO UNA REGRESIÓN LINEAL

Common statistical tests are linear models

Last updated: 02 April, 2019

See worked examples and more details at the accompanying notebook: <https://lindeloev.github.io/tests-as-linear>

Common name	Built-in function in R	Equivalent linear model in R	Exact?	The linear model in words	Icon	
Simple regression: $\text{Im}(y \sim 1 + x)$	y is independent of x P: One-sample t-test N: Wilcoxon signed-rank	<code>t.test(y) wilcox.test(y)</code>	<code>lm(y ~ 1) lm(signed_rank(y) ~ 1)</code>	✓ <code>for N > 14</code>	One number (intercept, i.e., the mean) predicts y . - (Same, but it predicts the <i>signed rank</i> of y .)	
	P: Paired-sample t-test N: Wilcoxon matched pairs	<code>t.test(y1, y2, paired=TRUE) wilcox.test(y1, y2, paired=TRUE)</code>	<code>lm(y2 - y1 ~ 1) lm(signed_rank(y2 - y1) ~ 1)</code>	✓ <code>for N > 14</code>	One intercept predicts the pairwise y₂-y₁ differences. - (Same, but it predicts the <i>signed rank</i> of y₂-y₁ .)	
	y ~ continuous x P: Pearson correlation N: Spearman correlation	<code>cor.test(x, y, method='Pearson') cor.test(x, y, method='Spearman')</code>	<code>lm(y ~ 1 + x) lm(rank(y) ~ 1 + rank(x))</code>	✓ <code>for N > 10</code>	One intercept plus x multiplied by a number (slope) predicts y . - (Same, but with <i>ranked x</i> and y)	
	y ~ discrete x P: Two-sample t-test P: Welch's t-test N: Mann-Whitney U	<code>t.test(y1, y2, var.equal=TRUE) t.test(y1, y2, var.equal=FALSE) wilcox.test(y1, y2)</code>	<code>lm(y ~ 1 + G₁)^a gls(y ~ 1 + G₂, weights=...)^b lm(signed_rank(y) ~ 1 + G₂)^a</code>	✓ ✓ <code>for N > 11</code>	An intercept for group 1 (plus a difference if group 2) predicts y . - (Same, but with one variance <i>per group</i> instead of one common.) - (Same, but it predicts the <i>signed rank</i> of y .)	
Multiple regression: $\text{Im}(y \sim 1 + x_1 + x_2 + \dots)$	P: One-way ANOVA N: Kruskal-Wallis	<code>aov(y ~ group) kruskal.test(y ~ group)</code>	<code>lm(y ~ 1 + G₂ + G₃ + ... + G_N)^a lm(rank(y) ~ 1 + G₂ + G₃ + ... + G_N)^a</code>	✓ <code>for N > 11</code>	An intercept for group 1 (plus a difference if group ≠ 1) predicts y . - (Same, but it predicts the <i>rank</i> of y .)	
	P: One-way ANCOVA	<code>aov(y ~ group + x)</code>	<code>lm(y ~ 1 + G₂ + G₃ + ... + G_N + x)^a</code>	✓	- (Same, but plus a slope on x .) Note: this is discrete AND continuous. ANCOVAs are ANOVAs with a continuous x .	
	P: Two-way ANOVA	<code>aov(y ~ group * sex)</code>	<code>lm(y ~ 1 + G₂ + G₃ + ... + G_N + S₂ + S₃ + ... + S_K + G₂*S₂+G₃*S₃+...+G_N*S_K)</code>	✓	Interaction term: changing sex changes the y - group parameters. Note: $G_{2..N}$ is an <i>indicator (0 or 1)</i> for each non-intercept levels of the group variable. Similarly for $S_{2..K}$ for sex . The first line (with G₁) is main effect of group , the second (with S₁ for sex and the third is the group × sex interaction. For two levels (e.g. male/female), line 2 would just be " S₁ " and line 3 would be S₁ multiplied with each G_i .	[Coming]
	Counts ~ discrete x N: Chi-square test	<code>chisq.test(groupXsex_table)</code>	Equivalent log-linear model <code>glm(y ~ 1 + G₂ + G₃ + ... + G_N + S₂ + S₃ + ... + S_K + G₂*S₂+G₃*S₃+...+G_N*S_K, family=...)^a</code>	✓	Interaction term: (Same as Two-way ANOVA.) Note: Run <code>glm</code> using the following arguments: <code>glm(modele, family=poisson())</code> As linear-model, the Chi-square test is $\log(y) = \log(N) + \log(a) + \log(\beta) + \log(a\beta)$ where a and β are proportions. See more info in the accompanying notebook.	Same as Two-way ANOVA
N: Goodness of fit	<code>chisq.test(y)</code>	<code>glm(y ~ 1 + G₂ + G₃ + ... + G_N, family=...)^a</code>	✓	(Same as One-way ANOVA and see Chi-Square note.)	1W-ANOVA	

List of common parametric (P) non-parametric (N) tests and equivalent linear models. The notation $y \sim 1 + x$ is R shorthand for $y = 1 \cdot b + a \cdot x$ which most of us learned in school. Models in similar colors are highly similar, but really, notice how similar they *all* are across colors! For non-parametric models, the linear models are reasonable approximations for non-small sample sizes (see "Exact" column and click links to see simulations). Other less accurate approximations exist, e.g., Wilcoxon for the sign test and Goodness-of-fit for the binomial test. The signed rank function is `signed_rank = function(x) sign(x) * rank(abs(x))`. The variables **G_i** and **S_j** are "dummy coded" indicator variables (either 0 or 1) exploiting the fact that when $\Delta x = 1$ between categories the difference equals the slope. Subscripts (e.g., G_2 or y) indicate different columns in data. `Im` requires long-format data for all non-continuous models. All of this is exposed in greater detail and worked examples at <https://lindeloev.github.io/tests-as-linear>.

^a See the note to the two-way ANOVA for explanation of the notation.

^b Same model, but with one variance per group: `gls(value ~ 1 + G2, weights = varIdent(form = ~1|group), method="ML")`.



Jonas Kristoffer Lindeløv
<https://lindeloev.net>