

PRACTICANDO LAS REGRESIONES LINEALES

Empezamos con los modelos y vamos a trabajar con las regresiones lineales.

Para ir familiarizándote te propongo dos ejercicios.

Una regresión lineal simple

Una regresión línea múltiple

Sigue los vídeos tech Zen porque será simplemente cambiar algunas variables.

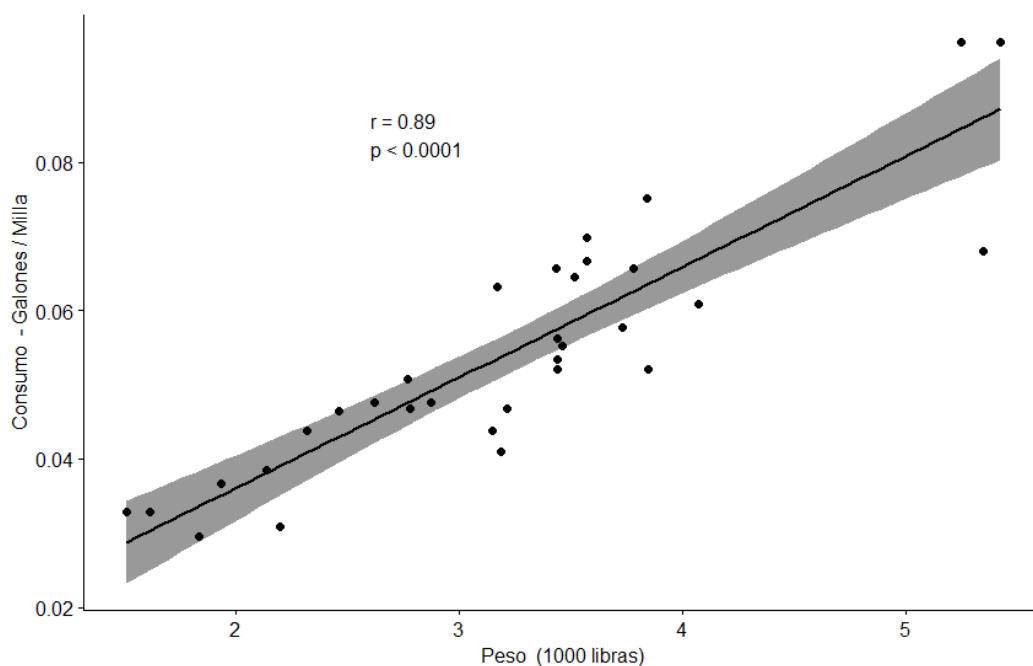
Te escribo el ejemplo de una regresión lineal simple para que te facilite el trabajo:

Queremos encontrar un modelo que permita obtener el consumo del coche solo sabiendo el peso. Este modelo lo podemos encontrar utilizando datos de consumo y peso de distintos coches.

Crear un gráfico de dispersión

consumo ↕	peso ↕
0.04761905	2.620
0.04761905	2.875
0.04385965	2.320
0.04672897	3.215
0.05347594	3.440
0.05524862	3.460
0.06993007	3.570
0.04098361	3.190
0.04385965	3.150
0.05208333	3.440
0.05617978	3.440
0.06097561	4.070
0.05780347	3.730
0.06578947	3.780
0.09615385	5.250
0.09615385	5.424
0.06802721	5.345
0.03086420	2.200
0.03280474	1.615

PRACTICANDO LAS REGRESIONES LINEALES



El significado del p-valor

- a. H_1 : los coeficientes son significativamente diferentes de cero
- b. H_0 : los coeficientes son igual a 0

Establece el nivel de significación o el margen de error

- c. 5%

Elige una prueba o test estadístico

- d. Regresión Lineal Simple

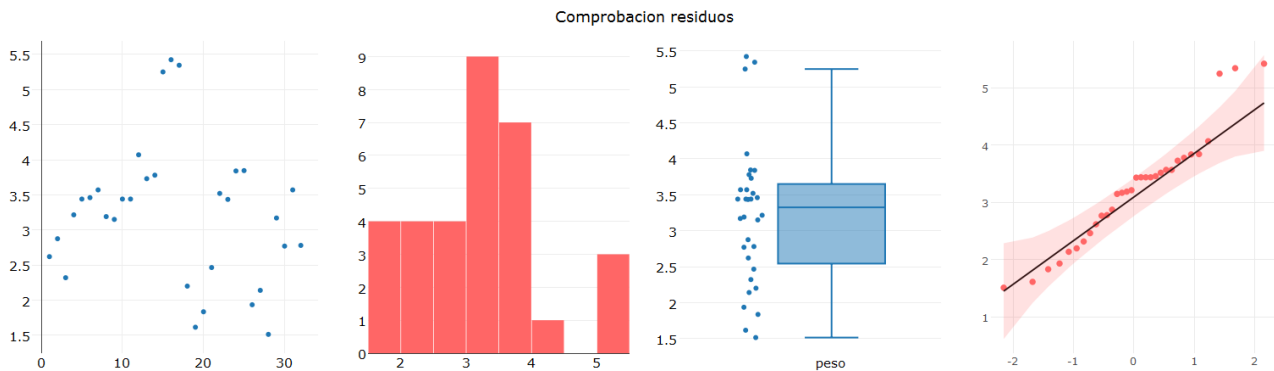
Calcula el p-valor y el coeficiente de correlación:

- e. Intercept o "a" = 0.006169 con un p-valor = 0.199 >> Debería ser 0 pero siempre se deja
- f. Pendiente o "b" = 0.014938 con un p-valor = 9.57e-12 >> significativo
- g. El modelo es:

$$\text{Consumo} = 0.014938 * \text{Peso} + 0.006169$$

- h. Los residuos son normales y NO presentan autocorrelación (gráficos).

PRACTICANDO LAS REGRESIONES LINEALES



Los gráficos reflejan la normalidad de los residuos y no presentan formas raras en el diagrama de dispersión.

Conclusión:

El modelo parece bueno aunque el p-valor de la b sea menos que 0.05. Normalmente el “intercept” se deja a no ser que estés muy seguro que ninguna otra variable tiene influencia con el consumo.

Este es el resultado de esta recta de regresión. Lo más importante está en la parte de coeficientes. Puedes ver el valor (Estimate) y el p-valor.

```
call:
lm(formula = consumo ~ peso, data = varEstudio)

Residuals:
    Min       1Q   Median       3Q      Max
-0.0179837 -0.0044301  0.0009413  0.0045354  0.0116583

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.006169   0.004695   1.314   0.199
peso         0.014938   0.001398  10.685 9.57e-12 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.007616 on 30 degrees of freedom
Multiple R-squared:  0.7919,    Adjusted R-squared:  0.785
F-statistic: 114.2 on 1 and 30 DF,  p-value: 9.566e-12
```

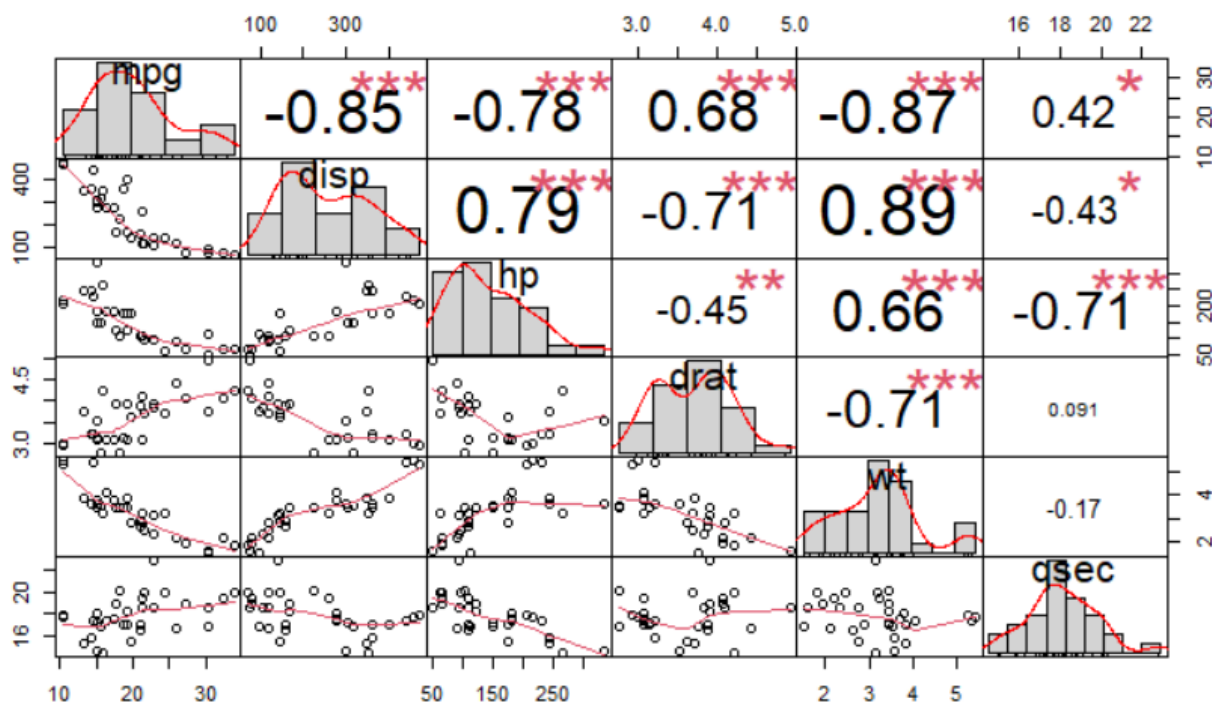
¡Bien! Ahora es tu turno.

REGRESIÓN LINEAL SIMPLE

Vamos a cambiar la variable de entrada a ver si conseguimos un buen modelo de regresión lineal.

Calcula el matrixplot de todas las variables:

Copia el matrixplot



Escoge una de las variables que no sea el peso que según el matrixplot y el coeficiente de correlación está más relacionada con el consumo (que no sea el peso)

Calcula el modelo y copia los resultados aquí:

Copia los resultados con la variables que has escogido

PRACTICANDO LAS REGRESIONES LINEALES

Los resultados son:

- La ecuación del modelo
- Los resultados de la tabla de coeficientes y del modelo
- La comprobación de los residuos

MPG Y DISP: Coeficiente de correlación -0.85

MPG: variable respuesta

Disp: Variable de salida

FORMULA

$\text{mpg} = a \cdot \text{DISP} + b + \text{error}$

MPG: $0.0001166 \cdot \text{disp} + 0.0273265$

Call:

`lm(formula = mpg ~ disp, data = df)`

Residuals:

Min	1Q	Median	3Q	Max
-0.021881	-0.003825	0.001061	0.005191	0.015194

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	0.0273265	0.0030013	9.105	3.88e-10 ***
disp	0.0001166	0.0000115	10.139	3.32e-11 ***

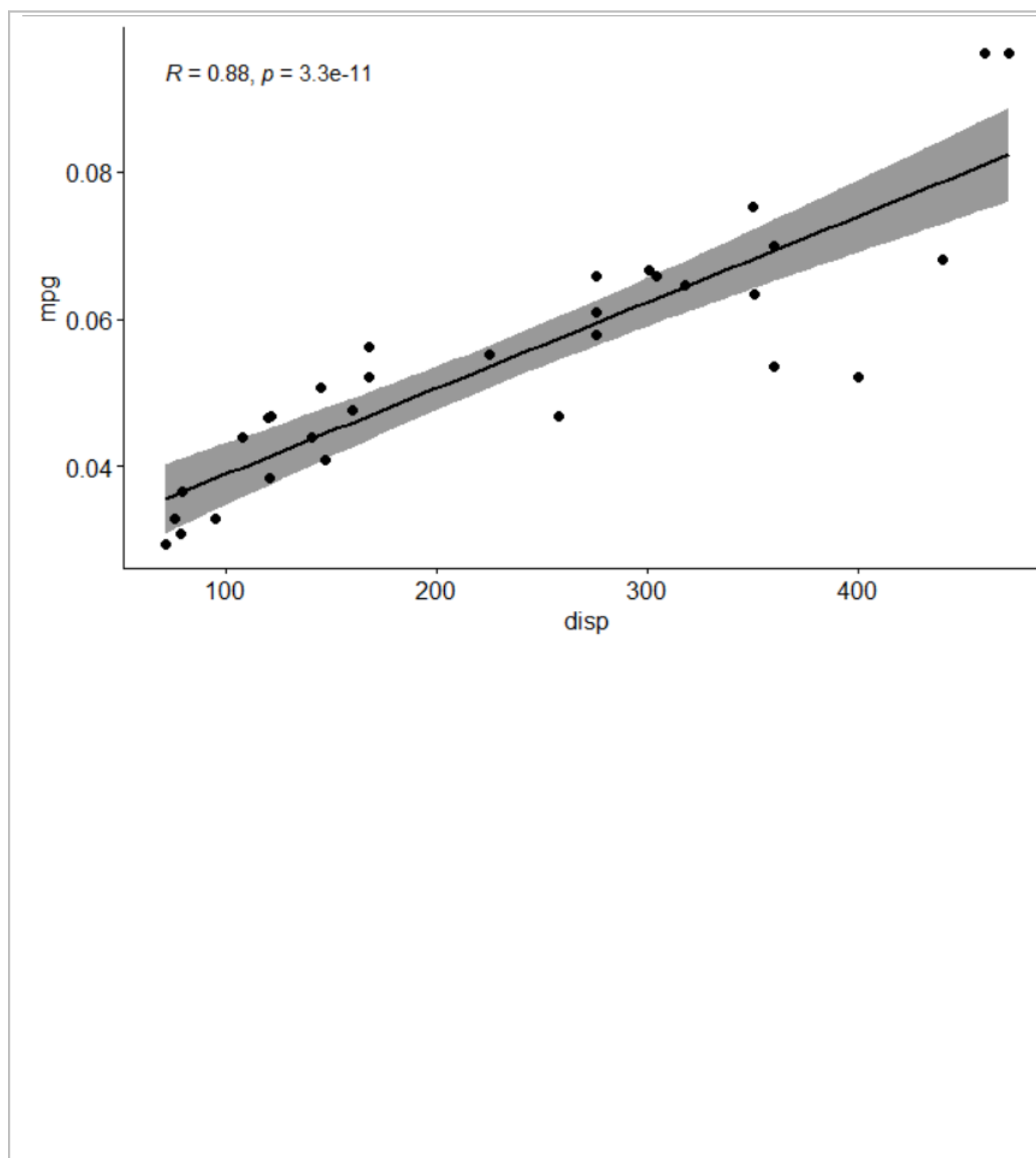
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.007936 on 30 degrees of freedom

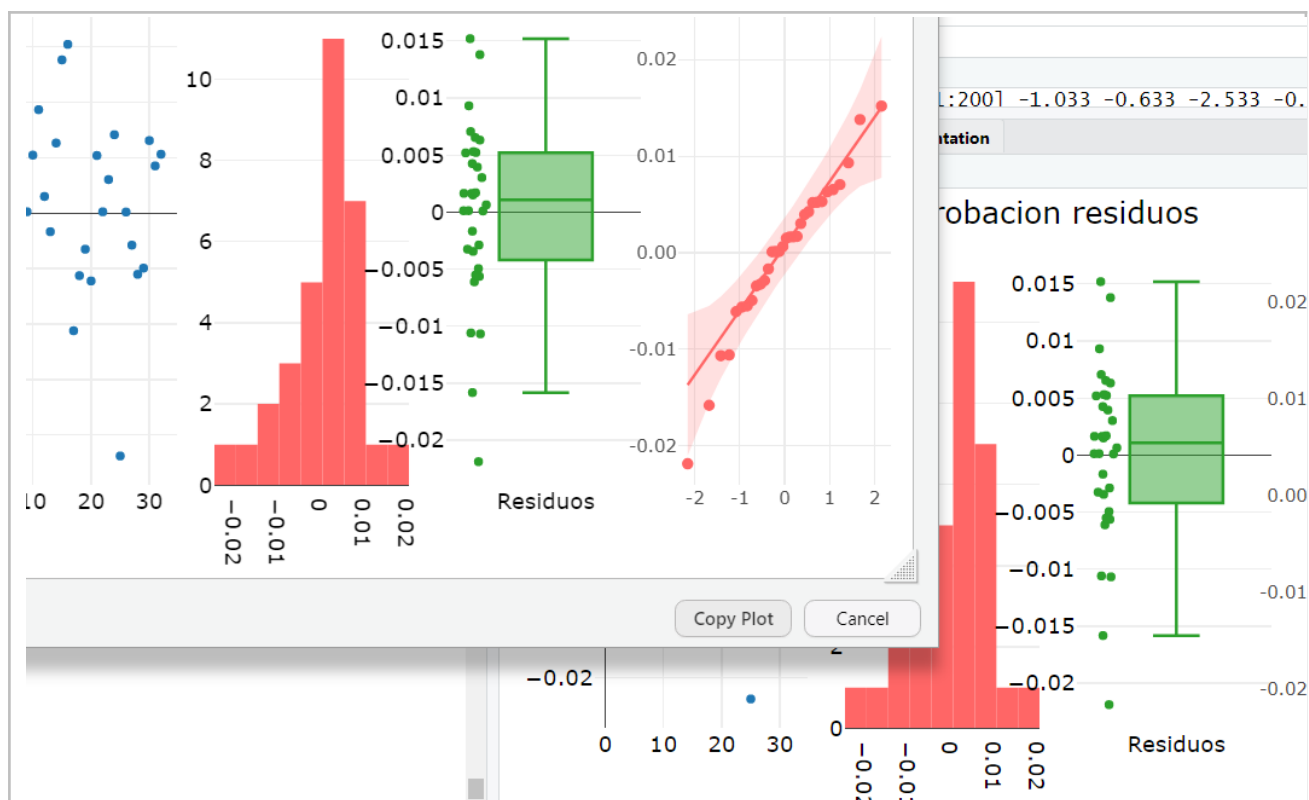
Multiple R-squared: 0.7741, Adjusted R-squared: 0.7666

F-statistic: 102.8 on 1 and 30 DF, p-value: 3.316e-11

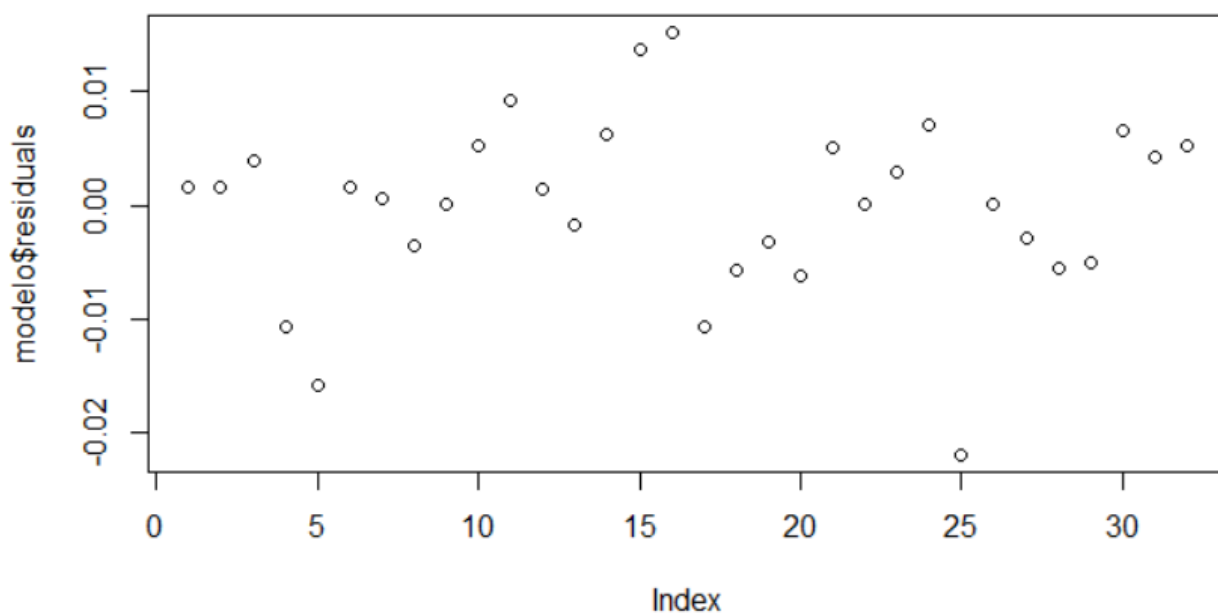
PRACTICANDO LAS REGRESIONES LINEALES



PRACTICANDO LAS REGRESIONES LINEALES



RESIDUOS



PRACTICANDO LAS REGRESIONES LINEALES

Shapiro-Wilk normality test

data: residuos

W = 0.96228, p-value = 0.3167

Residuos con distribución normal

¿Qué puedes interpretar según los resultados del modelo?

Describe lo que ves en los resultados.

El significado del p-valor

- i. H1: los coeficientes son significativamente diferentes de cero
- j. H0: los coeficientes son igual a 0

Establece el nivel de significación o el margen de error

- k. 5%

Elige una prueba o test estadístico

- l. Regresión Lineal Simple

Calcula el p-valor y el coeficiente de correlación:

- m. Intercept o "a" = 0.0273265 con un p-valor = 9.105 3.88e-10 ***>>
- n. Pendiente o "b" = 0.0001166 con un p-valor = 10.139 3.32e-11 *** >> significativo
- o. El modelo es:

Consumo = $0.0001166 \cdot \text{disp} + 0.0273265$

- p. Los residuos son normales y NO presentan autocorrelación (gráficos).

PRACTICANDO LAS REGRESIONES LINEALES

Las variables MPG y Disp están relacionadas de forma inversa a menor DISP mayor consumo, con coeficiente de correlación de -0.85. Se aplica modelo de regresión lineal obteniendo la formula **Consumo = $0.0001166 \cdot \text{disp} + 0.0273265$** con ambos coeficientes significativos

El modelo parece bueno, aunque hay algún valor que es atípico y se sale de los valores esperados, Aplicando el test de Sahpiro obtenemos que los residuos tienen una distribución normal.

REGRESIÓN LINEAL MÚLTIPLE

Calcula dos modelos lineales múltiples:

Modelo 1: Consumo = $a_1 \cdot \text{Peso} + a_2 \cdot \text{Caballos} + a_3 \cdot \text{Cilindrada} + b$

Modelo 2: Consumo = $a_1 \cdot \text{Peso} + a_2 \cdot \text{Caballos} + b$

Copia los resultados del modelo 1

Modelo 1: Consumo = $a_1 \cdot \text{Peso} + a_2 \cdot \text{Caballos} + a_3 \cdot \text{Cilindrada} + b$

Residuals:

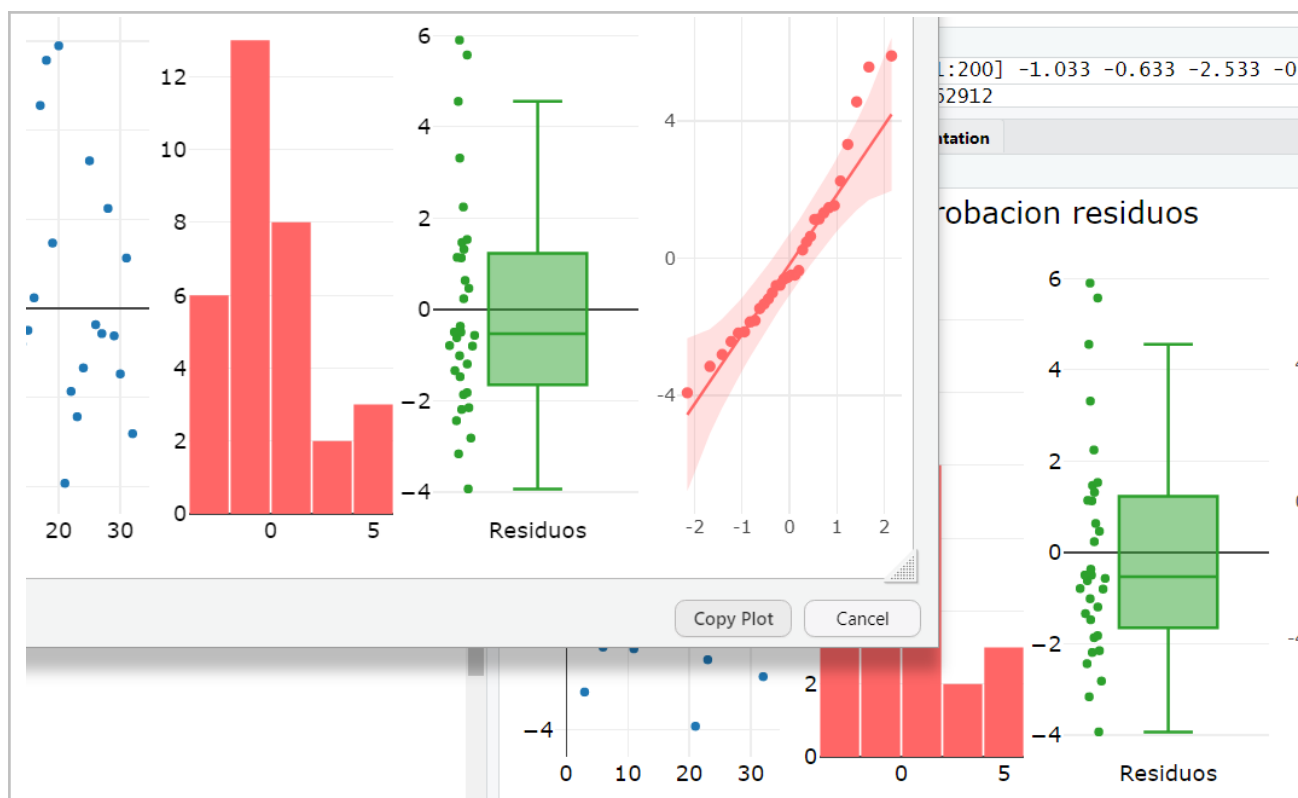
Min	1Q	Median	3Q	Max
-3.9290	-1.5598	-0.5311	1.1850	5.8986

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	38.75179	1.78686	21.687	< 2e-16 ***
peso	-3.16697	0.74058	-4.276	0.000199 ***
Caballos	-0.01804	0.01188	-1.519	0.140015
cilindrada	-0.94162	0.55092	-1.709	0.098480 .

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

PRACTICANDO LAS REGRESIONES LINEALES



Shapiro-Wilk normality test

data: residuos

W = 0.93455, p-value = 0.05252

Realizando el modelo de regresión lineal multiple vemos que el coeficiente de la variable caballos y de cilindrada no son significativos y haciendo el test de shapiro test vemos que los residuos no tienen distribución normal

Copia los resultados del modelo 2

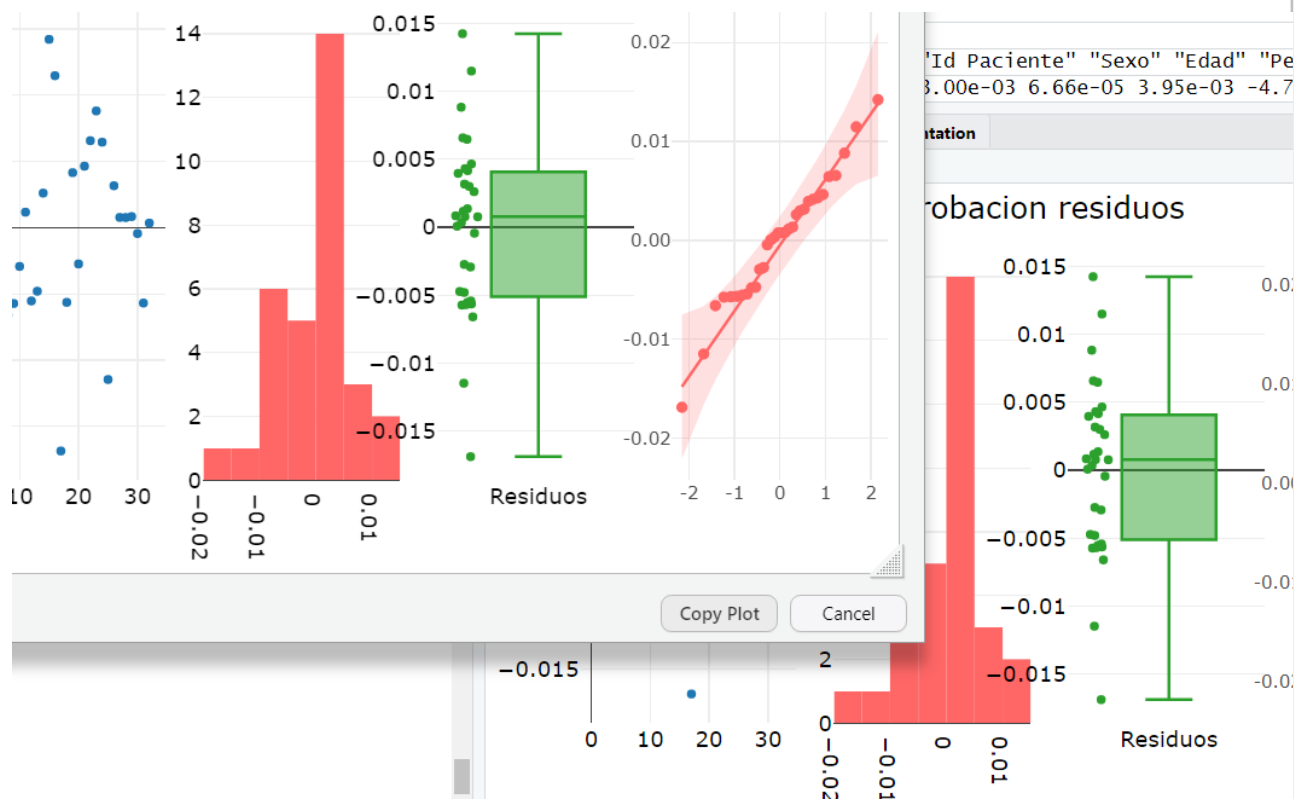
Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	6.305e-03	4.094e-03	1.540	0.13435
peso	1.149e-02	1.620e-03	7.089	8.45e-08 ***
Caballos	7.479e-05	2.312e-05	3.235	0.00303 **

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

PRACTICANDO LAS REGRESIONES LINEALES

En el modelo de regresión lineal tanto el peso como los caballos son significativos por lo que tiene sentido usarlos en el modelo



Shapiro-Wilk normality test

data: residuos

W = 0.97539, p-value = 0.6589

El test de Shapiro muestra que los residuos tienen una distribución normal

Modelo 2: Consumo = $1.149e-02$ *Peso + $7.479e-05$ *Caballos + $6.305e-03$

PRACTICANDO LAS REGRESIONES LINEALES

Compara los dos modelos con el BIC o AIC y elige uno de ellos.

¿Qué puedes interpretar de este análisis?

Yo elegiría el modelo 2 Modelo 2: Consumo = $1.149e-02$ *Peso + $7.479e-05$ *Caballos + $6.305e-03$

Ya que los residuos tienen una distribución normal y los coeficientes son significativos. Llama la atención el valor tan bajo del Intercepto y de los coeficientes que acompañan a las variables, por lo que modifican poco el valor de peso y caballos.

¡Tus primeras experiencias con modelos!

La siguiente es flexibilizar un poquito el modelo lineal $\odot$ y relajar las condiciones de uso.

Estás cerca de los modelos lineales generalizados