Multple Linear Regression

María Isabel Chuya

Multple Linear Regression

La regresión lineal múltiple es una generalización de la regresión lineal simple porque permite evaluar las relaciones lineales entre una variable explicativa (cuantitativa o cualitativa) y varias variables de respuesta.

- Cargar librerías a utilizar

```
library(ggplot2)
library(forcats)
library(performance)
library(visreg)
library(ggstatsplot)
```

```
You can cite this package as:
     Patil, I. (2021). Visualizations with statistical details: The 'ggstatsplot' approach.
     Journal of Open Source Software, 6(61), 3167, doi:10.21105/joss.03167
```

```
library(car)
```

```
Loading required package: carData
```

```
library(carData)
```

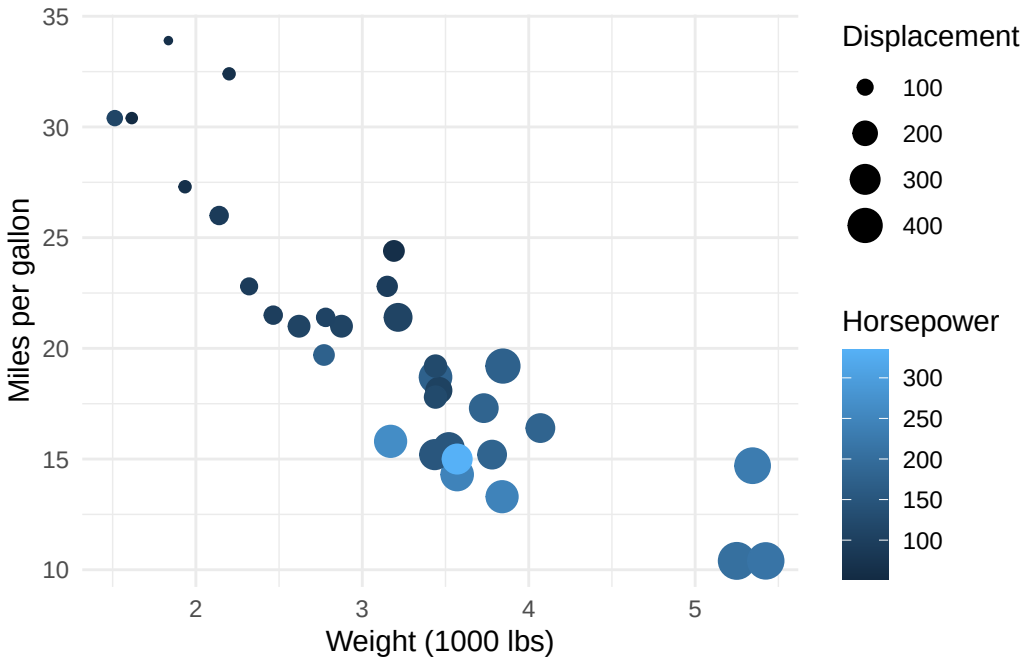
- Se cargan llos datos del conjunto “mtcars”
- El comando “head()” muestra por defecto las primeras 6 filas del conjunto de datos especificado.

```r
# Cargamos el conjunto de datos mtcars
data <- mtcars
head(data)
```

```
                   mpg cyl disp  hp drat    wt  qsec vs am gear carb
Mazda RX4         21.0   6  160 110 3.90 2.620 16.46  0  1    4    4
Mazda RX4 Wag     21.0   6  160 110 3.90 2.875 17.02  0  1    4    4
Datsun 710        22.8   4  108  93 3.85 2.320 18.61  1  1    4    1
Hornet 4 Drive    21.4   6  258 110 3.08 3.215 19.44  1  0    3    1
Hornet Sportabout 18.7   8  360 175 3.15 3.440 17.02  0  0    3    2
Valiant           18.1   6  225 105 2.76 3.460 20.22  1  0    3    1
```

El peso (wt), los caballos de fuerza (hp) y el desplazamiento (disp) de un automóvil están relacionados con su consumo de combustible (mpg). El volumen combinado de aire barrido (o desplazado) que resulta del movimiento hacia arriba y hacia abajo de los pistones en los cilindros se conoce como cilindrada del motor. Por lo general, cuanto más alto es el cilindro, más potente es el motor.

```r
ggplot(data) +
  aes(x = wt, y = mpg, colour = hp, size = disp) +
  geom_point() +
  scale_color_gradient() +
  labs(
    y = "Miles per gallon",
    x = "Weight (1000 lbs)",
    color = "Horsepower",
    size = "Displacement"
  ) +
  theme_minimal()
```



La relación estimada entre la variable dependiente y una variable explicativa es ajustada, es decir, libre de los efectos lineales de las otras variables explicativas, según la regresión lineal múltiple.

Interpretaciones de los coeficientes ˆ

El método de mínimos cuadrados da como resultado una estimación ajustada de los coeficientes. El término ajustado significa **después de tener en cuenta los efectos lineales** de las otras variables independientes en la variable dependiente, pero también en la variable predictora.

```r
model2 <- lm(mpg ~ wt + hp + disp, data = data)
summary(model2)
```

```
Call:
lm(formula = mpg ~ wt + hp + disp, data = data)

Residuals:
   Min     1Q Median     3Q    Max
-3.891 -1.640 -0.172  1.061  5.861
```

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 37.105505   2.110815  17.579  < 2e-16 ***
wt          -3.800891   1.066191  -3.565  0.00133 **
hp          -0.031157   0.011436  -2.724  0.01097 *
disp        -0.000937   0.010350  -0.091  0.92851
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.639 on 28 degrees of freedom
Multiple R-squared:  0.8268,    Adjusted R-squared:  0.8083
F-statistic: 44.57 on 3 and 28 DF,  p-value: 8.65e-11
```

La interpretación de las variables cualitativas independientes es ligeramente diferente en el sentido de que cuantifica el efecto de un nivel en comparación con el nivel de referencia, siendo todo lo demás igual. Por lo tanto, compara los diferentes grupos (formados por los diferentes niveles de la variable categórica) en términos de la variable dependiente (es por eso que la regresión lineal se puede ver como una extensión de la prueba t y el ANOVA).

```r
# Grabando dat$vs
data$vs <- as.character(data$vs)
data$vs <- fct_recode(data$vs,
  "V-shaped" = "0",
  "Straight" = "1"
)

model3 <- lm(mpg ~ wt + vs, data = data)
summary(model3)
```

```
Call:
lm(formula = mpg ~ wt + vs, data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-3.7071 -2.4415 -0.3129  1.4319  6.0156

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  33.0042     2.3554  14.012 1.92e-14 ***
wt           -4.4428     0.6134  -7.243 5.63e-08 ***
```

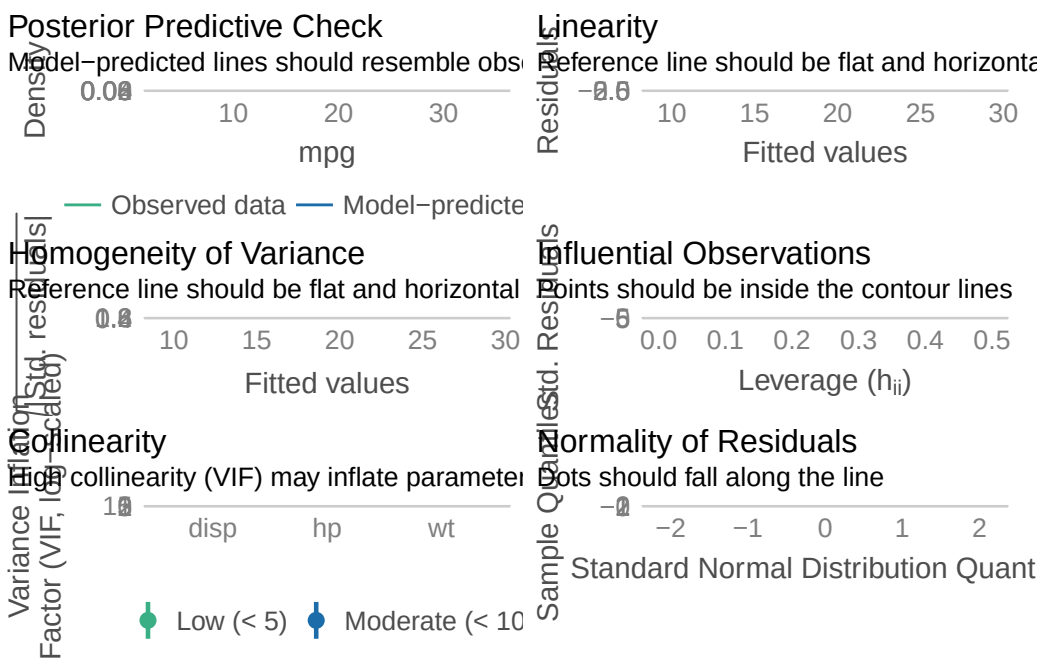
```
vsStraight    3.1544     1.1907   2.649   0.0129 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.78 on 29 degrees of freedom
Multiple R-squared:  0.801, Adjusted R-squared:  0.7873
F-statistic: 58.36 on 2 and 29 DF,  p-value: 6.818e-11
```

Se concluye que:

- Para un motor en forma de V y para un aumento de una unidad en el peso (es decir, un aumento de 1000 libras), el número de millas/galón disminuye, en promedio, en 4,44 (p-valor $< 0{,}001$).
- La distancia recorrida con un galón de combustible aumenta, en promedio, en 3,15 millas **cuando el motor es recto en comparación con un motor en forma de V**, para un peso constante (p-valor = 0,013).
- (Para completar, pero debe interpretarse solo cuando tenga sentido: para un peso = 0 y un motor en forma de V, podemos esperar que el coche tenga, en promedio, un consumo de combustible de 33 millas/galón (p-valor $< 0{,}001$).)

Condiciones de aplicación

```
check_model(model2)
```




```
summary(model2)
```

```
Call:
lm(formula = mpg ~ wt + hp + disp, data = data)

Residuals:
   Min     1Q Median     3Q    Max
-3.891 -1.640 -0.172  1.061  5.861

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 37.105505   2.110815  17.579  < 2e-16 ***
wt          -3.800891   1.066191  -3.565  0.00133 **
hp          -0.031157   0.011436  -2.724  0.01097 *
disp        -0.000937   0.010350  -0.091  0.92851
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.639 on 28 degrees of freedom
Multiple R-squared:  0.8268,    Adjusted R-squared:  0.8083
F-statistic: 44.57 on 3 and 28 DF,  p-value: 8.65e-11
```

Para la ilustración, comenzamos con un modelo con todas las variables en el conjunto de datos como variables independientes:

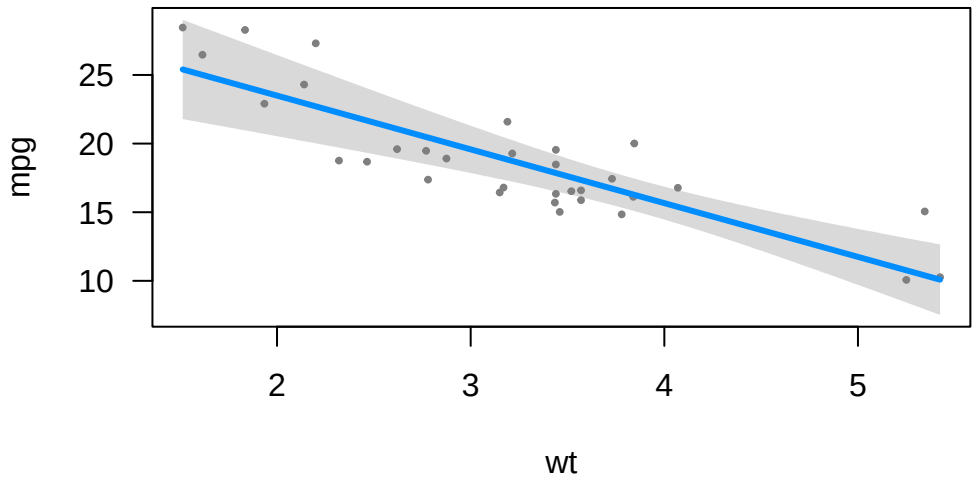
```r
## vs has already been transformed into factor
## so only am is transformed here

## Recoding dat$vs
library(forcats)
data$am <- as.character(data$am)
data$am <- fct_recode(data$am,
  "Automatic" = "0",
  "Manual" = "1"
)

model4 <- lm(mpg ~ ., data = data)
model4 <- step(model4, trace = FALSE)
summary(model4)
```

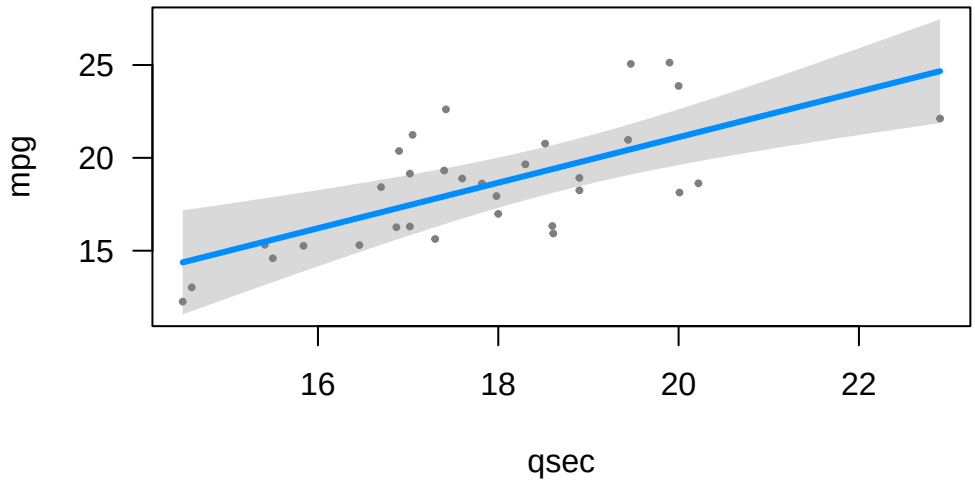
```
Call:
lm(formula = mpg ~ wt + qsec + am, data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-3.4811 -1.5555 -0.7257  1.4110  4.6610

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   9.6178     6.9596   1.382 0.177915
wt           -3.9165     0.7112  -5.507 6.95e-06 ***
qsec          1.2259     0.2887   4.247 0.000216 ***
amManual      2.9358     1.4109   2.081 0.046716 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.459 on 28 degrees of freedom
Multiple R-squared:  0.8497,    Adjusted R-squared:  0.8336
F-statistic: 52.75 on 3 and 28 DF,  p-value: 1.21e-11
```

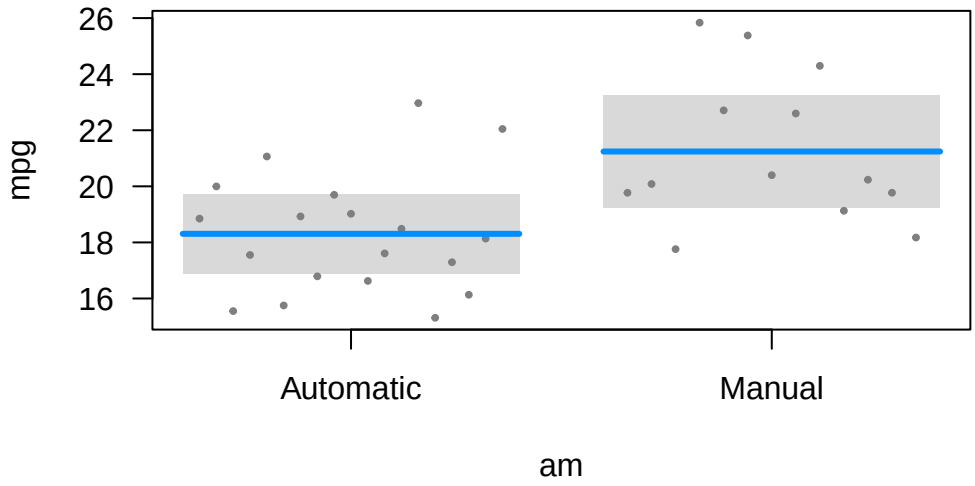
Visualizaciones

1. El comando Visreg () que ilustra las relaciones entre las variables dependientes e independientes en diferentes gráficos (uno para cada variable independiente a menos que especifique qué relación desea ilustrar):

```
visreg(model4)
```

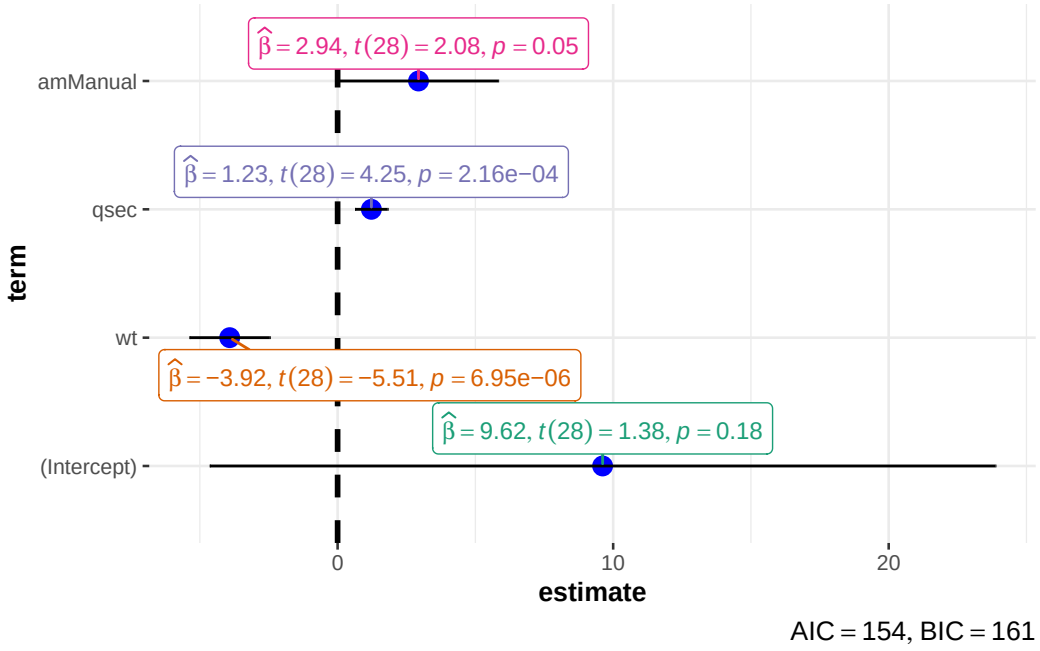






2. ggcoefstats () que ilustra los resultados en una sola parcela:

```
ggcoefstats(model4)
```



Predicciones

La regresión lineal también se utiliza muy a menudo con **fines predictivos**. Los intervalos de confianza y predicción para **nuevos datos** se pueden calcular con el predict() función.

```
# confidence interval for new data
predict(model4,
  new = data.frame(wt = 3, qsec = 18, am = "Manual"),
  interval = "confidence",
  level = .95
)
```

```
      fit      lwr    upr
1 22.87005 21.09811 24.642
```

```
# prediction interval for new data
predict(model4,
  new = data.frame(wt = 3, qsec = 18, am = "Manual"),
```

```
    interval = "prediction",
    level = .95
  )
```

```
       fit      lwr      upr
1 22.87005 17.53074 28.20937
```

Según nuestro modelo, se espera que este coche conduzca 22,87 millas con un galón.

La diferencia entre el intervalo de confianza y el intervalo de predicción es que:

- un intervalo **de confianza** da el valor predicho para la **media** de Y para una nueva observación, mientras que
- un intervalo **de predicción** da el valor predicho para un **individuo** Y para una nueva observación.

El intervalo de predicción es más amplio que el intervalo de confianza para tener en cuenta la **incertidumbre adicional debido a la predicción de una respuesta individual**, y no la media, para un valor dado de X.

Pruebas de Hipótesis LIneales

```
linearHypothesis(model4, c("wt = 0", "qsec = 0"))
```

```
Linear hypothesis test

Hypothesis:
wt = 0
qsec = 0

Model 1: restricted model
Model 2: mpg ~ wt + qsec + am

  Res.Df    RSS Df Sum of Sq      F   Pr(>F)
1     30 720.90
2     28 169.29  2    551.61 45.618 1.55e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Rechazamos la hipótesis nula y concluimos que al menos uno de 1 y 2 es diferente de 0 (pag-valor = 1.55e-09).

Efecto general de las variables categóricas

Cuando las variables independientes son categóricas con k categorías, la tabla de regresión proporciona k valores:

```
model5 <- lm(mpg ~ vs + am + as.factor(cyl), data = data)
summary(model5)
```

```
Call:
lm(formula = mpg ~ vs + am + as.factor(cyl), data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-6.2821 -1.4402  0.0391  1.8845  6.2179

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)       22.809      2.928   7.789 2.24e-08 ***
vsStraight         1.708      2.235   0.764  0.45135
amManual           3.165      1.528   2.071  0.04805 *
as.factor(cyl)6   -5.399      1.837  -2.938  0.00668 **
as.factor(cyl)8   -8.161      2.892  -2.822  0.00884 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.097 on 27 degrees of freedom
Multiple R-squared:  0.7701,    Adjusted R-squared:  0.736
F-statistic: 22.61 on 4 and 27 DF,  p-value: 2.741e-08
```

```
Anova(model5)
```

```
Anova Table (Type II tests)

Response: mpg
                Sum Sq Df F value  Pr(>F)
vs               5.601  1  0.5841 0.45135
am              41.122  1  4.2886 0.04805 *
as.factor(cyl)  94.591  2  4.9324 0.01493 *
Residuals      258.895 27
```

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

Interacción

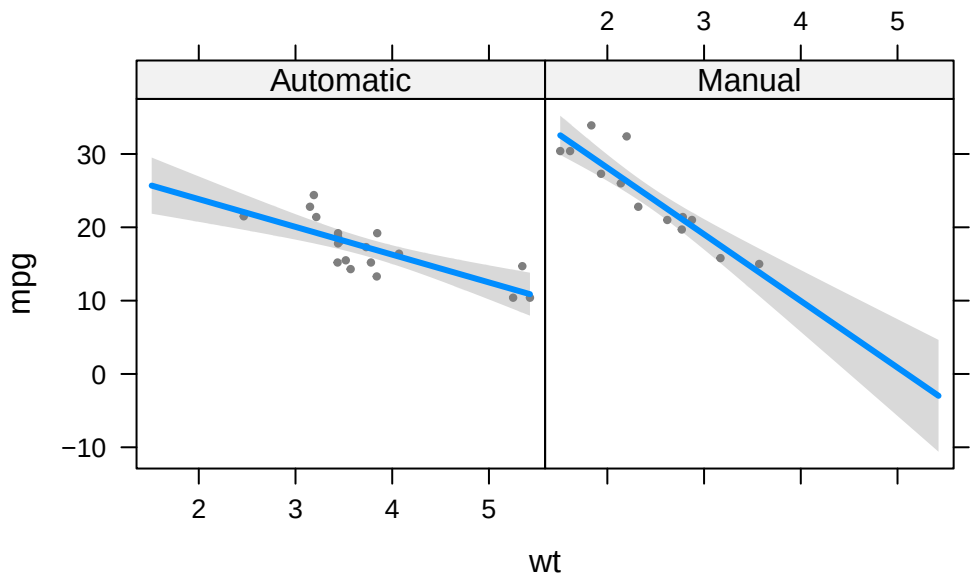
Existe un efecto **de interacción** entre los factores A y B **si el efecto del factor A sobre la respuesta depende del nivel que tome el factor** B.

```r
model6 <- lm(mpg ~ wt + am + wt:am, data = data)

# Or in a shorter way:
model6 <- lm(mpg ~ wt * am, data = data)

summary(model6)
```

```
Call:
lm(formula = mpg ~ wt * am, data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-3.6004 -1.5446 -0.5325  0.9012  6.0909

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  31.4161     3.0201  10.402 4.00e-11 ***
wt           -3.7859     0.7856  -4.819 4.55e-05 ***
amManual     14.8784     4.2640   3.489  0.00162 **
wt:amManual  -5.2984     1.4447  -3.667  0.00102 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.591 on 28 degrees of freedom
Multiple R-squared:  0.833, Adjusted R-squared:  0.8151
F-statistic: 46.57 on 3 and 28 DF,  p-value: 5.209e-11
```

La forma más fácil de manejar la interacción es visualizar la relación para cada nivel de la variable categórica:

```r
visreg(model6, "wt", by = "am")
```



Vemos que la relación entre el peso y las millas/galón es más fuerte (la pendiente es más pronunciada) para los coches con transmisión manual en comparación con los coches con transmisión automática.