

Regresión Linear Múltiple

Nicolás Jadán

La regresión lineal múltiple es una técnica estadística que amplía la regresión lineal simple para evaluar relaciones lineales entre una variable de respuesta (cuantitativa) y múltiples variables explicativas (cuantitativas o cualitativas).

A continuación, se cargan diversas librerías en R para facilitar el análisis de datos.

```
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(equatiomatic)
library(car)
```

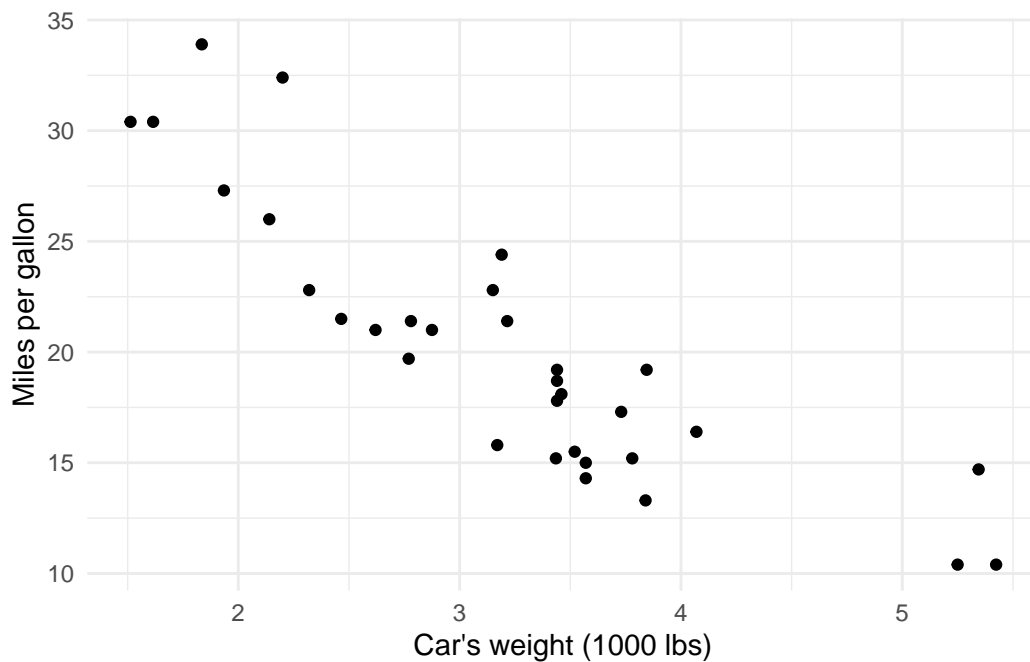
Loading required package: carData

```
library(carData)
```

A continuación, procedemos a evaluar si existe una relación lineal entre la distancia recorrida con un galón de combustible (mpg) y el peso de los automóviles (wt) utilizando el conjunto de datos “mtcars”:

```
# Cargamos el conjunto de datos
data <- mtcars
```

```
ggplot(data, aes(x = wt, y = mpg)) +
  geom_point() +
  labs(
    y = "Miles per gallon",
    x = "Car's weight (1000 lbs)"
  ) +
  theme_minimal()
```



El diagrama de dispersión muestra una relación negativa entre la distancia recorrida de un galón de combustible y el peso de un auto.

Para realizar una regresión lineal en R, usamos la función `lm()` (que significa modelo lineal).

```
model <- lm(mpg ~ wt, data = data)
summary(model)
```

Call:

```
lm(formula = mpg ~ wt, data = data)
```

Residuals:

Min	1Q	Median	3Q	Max
-4.5432	-2.3647	-0.1252	1.4096	6.8727

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	37.2851	1.8776	19.858	< 2e-16 ***
wt	-5.3445	0.5591	-9.559	1.29e-10 ***

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

Residual standard error: 3.046 on 30 degrees of freedom

Multiple R-squared: 0.7528, Adjusted R-squared: 0.7446

F-statistic: 91.38 on 1 and 30 DF, p-value: 1.294e-10

El valor $p = 1,29e-10 < 0,05$, por lo que rechazamos la hipótesis nula en el nivel de significancia =5%. Por lo tanto, concluimos que existe una relación significativa entre el peso de un automóvil y su consumo de combustible.

Condiciones de Aplicación

Existen tres condiciones fundamentales que se deben tener en cuenta al aplicar el análisis de regresión lineal:

1. Independencia: Las observaciones deben ser independientes entre sí. Esto significa que los datos deben provenir de diferentes individuos o unidades experimentales para que se consideren independientes. Si los datos se recopilan en diferentes momentos para los mismos individuos, es probable que no cumplan con esta condición de independencia.
2. Normalidad de los residuos: Para muestras grandes, los intervalos de confianza y las pruebas de los coeficientes son aproximadamente válidos, independientemente de si los errores (residuos) siguen o no una distribución normal. Sin embargo, para muestras pequeñas, es importante que los residuos se distribuyan de manera aproximadamente normal. Esto puede verificarse visualmente mediante un gráfico QQ (cuantil-cuantil) o un histograma, o mediante pruebas estadísticas formales como la prueba de Shapiro-Wilk.
3. Homocedasticidad de los residuales: La varianza de los errores (residuales) debe ser constante en todos los niveles de las variables independientes. Es decir, la dispersión de los residuos no debe depender del valor predicho (valor ajustado) por el modelo. Esta condición puede evaluarse visualmente mediante gráficos de dispersión de los residuos estandarizados frente a los valores ajustados o mediante la prueba de Breusch-Pagan.

Al cumplir con estas tres condiciones, se garantiza que los resultados del análisis de regresión lineal sean válidos y confiables para realizar inferencias sobre las relaciones entre las variables involucradas.

Regresión Linear Múltiple

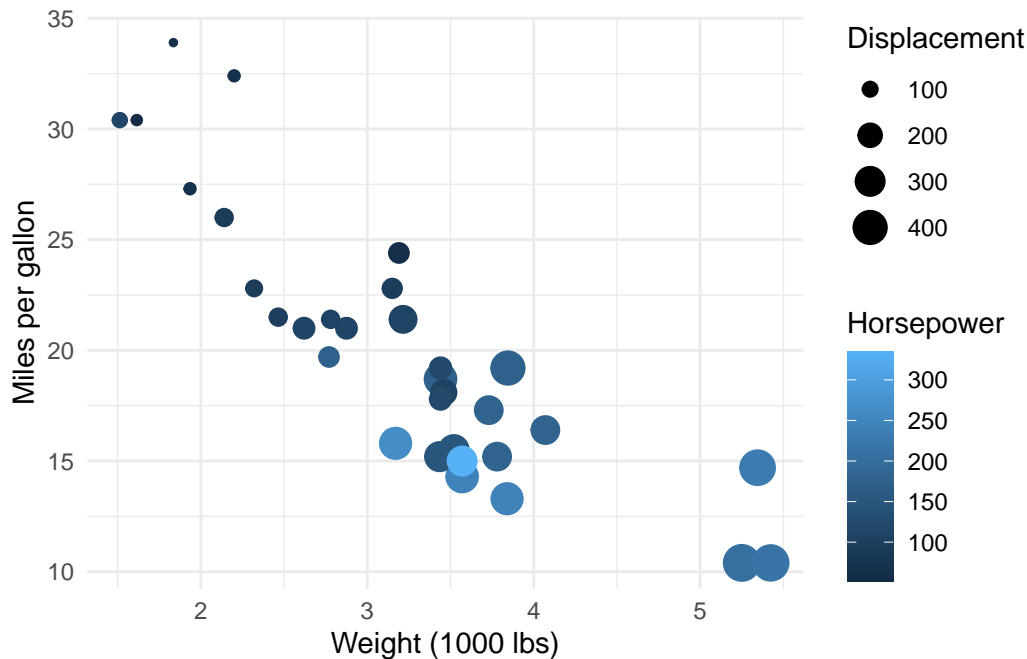
El comando `head()` muestra por defecto las primeras 6 filas del conjunto de datos especificado.

```
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

Se puede visualizar la relación entre el consumo de combustible de un automóvil (mpg) junto con su peso (wt), caballos de fuerza (hp) y desplazamiento (disp) (la cilindrada del motor es el volumen combinado de aire barrido (o desplazado) resultante del movimiento hacia arriba y hacia abajo de los pistones en los cilindros, generalmente cuanto más alto, más potente es el automóvil):

```
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()
```



Se observa que existen dos relaciones negativas: una entre el consumo de combustible (millas/galón) y la potencia (caballos de fuerza), y otra entre el consumo de combustible y el desplazamiento del motor. Los puntos más claros, que indican mayor potencia, tienden a estar más presentes en niveles bajos de millas por galón, al igual que los puntos más grandes, que representan mayor desplazamiento. Por lo tanto, para entender mejor la relación entre el consumo de combustible y el peso del automóvil, se realizan ajustes adicionales incorporando información sobre la potencia y el desplazamiento. De esta manera, podemos evaluar la relación directa entre el consumo de combustible y el peso, sin que el efecto indirecto de la potencia y el desplazamiento influya en la interpretación.

```
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

Las conclusiones obtenidas son las siguientes:

1. Se encontró una relación significativa y negativa entre el consumo de millas por galón y el peso de los automóviles, manteniendo constantes la potencia y el desplazamiento. Es decir, por cada aumento de una unidad en el peso (equivalente a 1000 libras), en promedio, el consumo de millas por galón disminuye en 3.8 unidades (valor de $p = 0.001$).
2. Se identificó una relación significativa y negativa entre el consumo de millas por galón y la potencia de los automóviles, considerando un nivel constante de peso y desplazamiento. Por cada aumento de una unidad en la potencia, en promedio, el consumo de millas por galón disminuye en 0.03 unidades (valor de $p = 0.011$).
3. No se encontró evidencia para rechazar la hipótesis de que no existe una relación entre el consumo de millas por galón y el desplazamiento, siempre que el peso y la potencia se mantengan constantes ($p\text{-valor} = 0.929 > 0.05$).
4. Para completar la interpretación, cuando el peso, la potencia y el desplazamiento son iguales a cero, se estima que un automóvil tendría un consumo de combustible promedio de 37.11 millas por galón (valor de $p < 0.001$).

En el siguiente análisis, se ilustra el modelo del consumo de combustible (mpg) considerando el peso (wt) y la forma del motor (vs), siendo esta última una variable categórica con dos niveles.

```
# 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

```
#check_model(model2)
```

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

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

```
## Recoding dat$vs
data$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

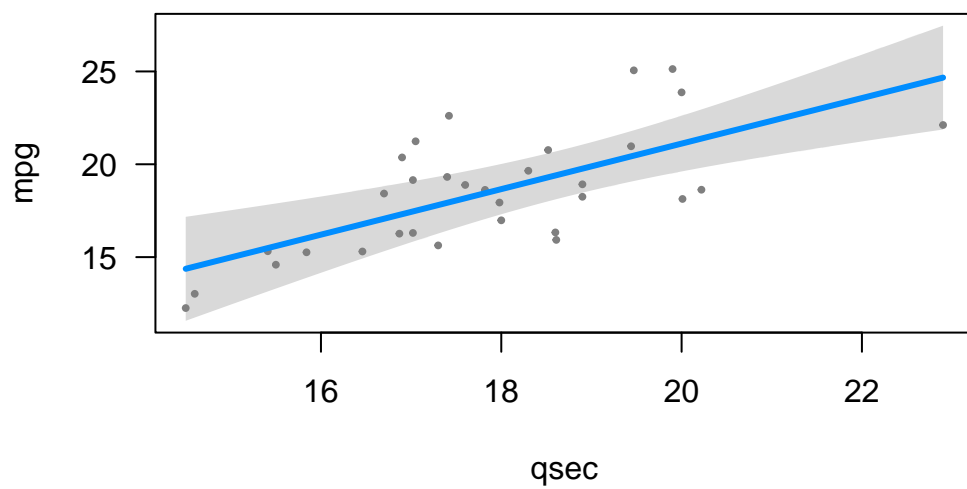
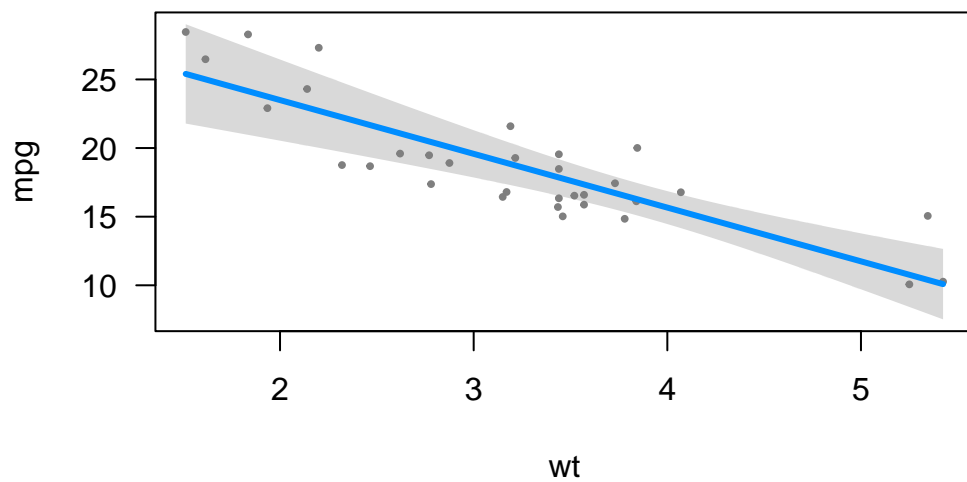
Multiple R-squared: 0.8497, Adjusted R-squared: 0.8336

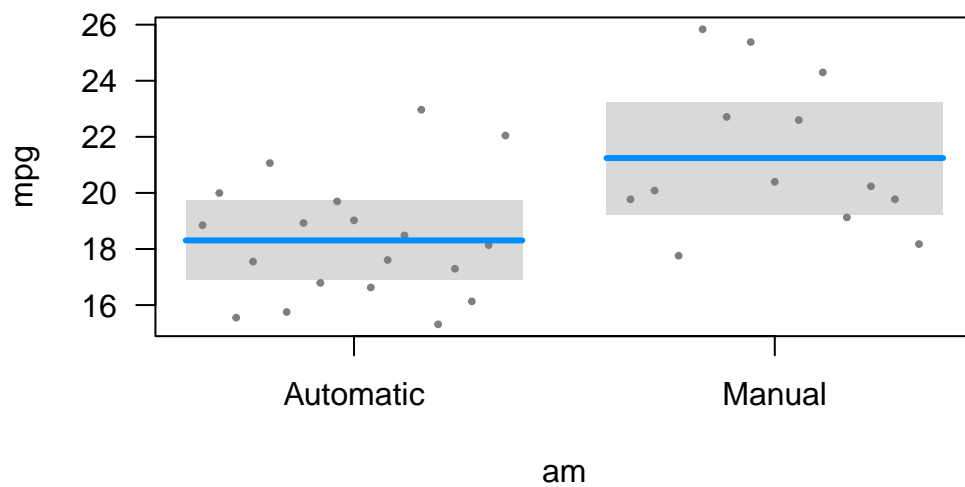
F-statistic: 52.75 on 3 and 28 DF, p-value: 1.21e-11

Hay muchas formas de visualizar los resultados de una regresión lineal. Los 2 más fáciles son:

1. 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):

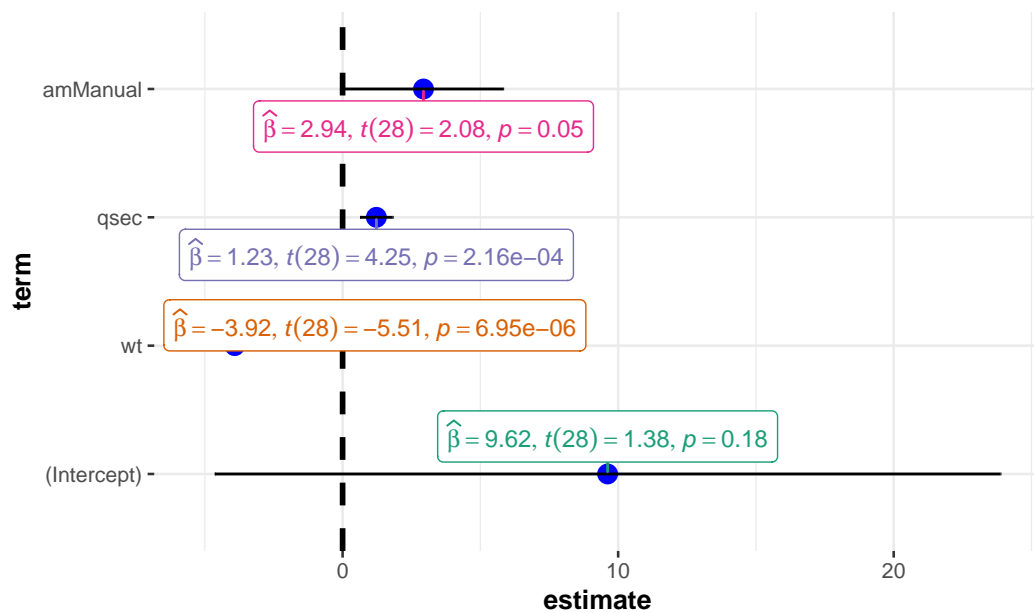
```
visreg(model4)
```



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

```
ggcoefstats(model4)
```



AIC = 154, BIC = 161

```
#extract_eq(model4,
# use_coefs = TRUE, # display coefficients
#wrap = TRUE, # multiple lines
#terms_per_line = 2
#)
```

Predicciones

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

```
# 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
```

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

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

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.

```

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

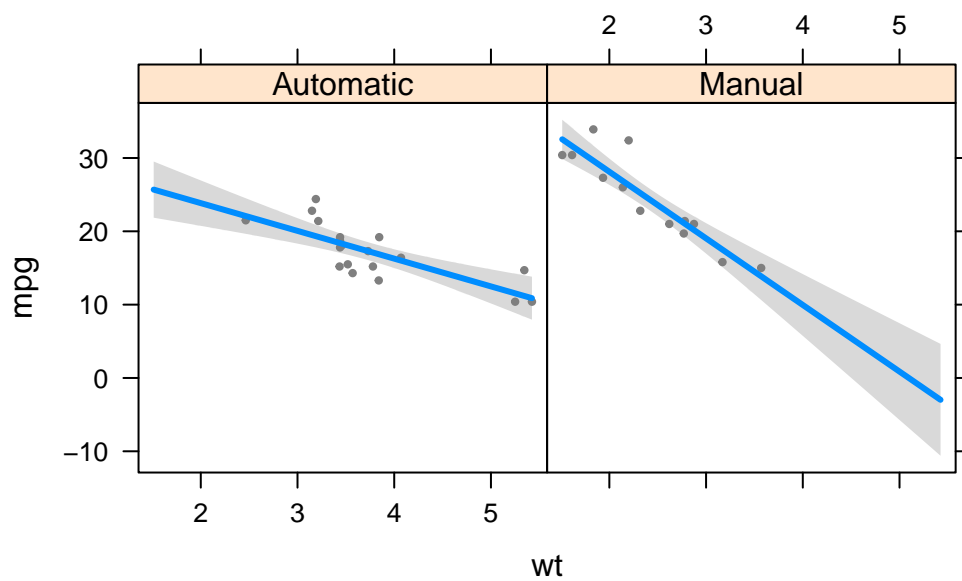
Residual standard error: 2.591 on 28 degrees of freedom

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:

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



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