Introducción al análisis de datos con R

Análisis de regresión lineal y logística con R

Manuel Mejías Leiva

Universidad de Valladolid | manuel.mejias@uva.es

5 - 9 junio de 2023

Primeros pasos: librerías, directorio de trabajo y datos

Antes de comenzar...

Primero, cargamos las librerías necesarias para el análisis:

```
library(tidyverse)
```

Segundo, definimos el directorio de trabajo en el que trabajaremos:

```
# setwd()
```

Tercero, importamos el fichero de datos que está en formato csv:

```
df <- read_csv("egd.csv")
```

Antes de comenzar...

```
glimpse(df)
```

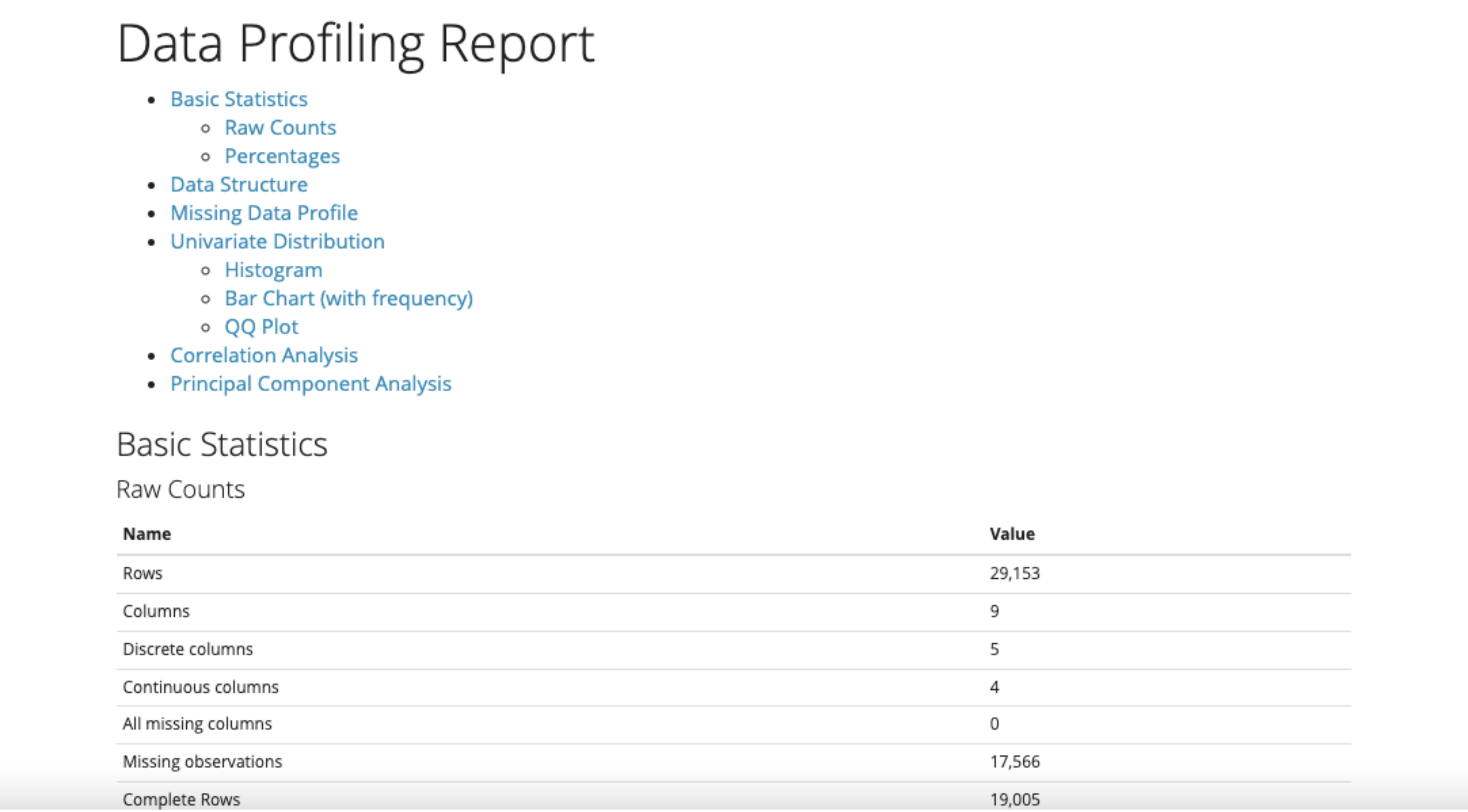
```
## Rows: 29,153
## Columns: 9
## $ id_centro             <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1…
## $ sexo                  <chr> "Chico", "Chica", "Chico", "Chica", "Chico", "Chi…
## $ trimestre_nacimiento <chr> "3 tr", "2 tr", "2 tr", "4 tr", "3 tr", "2 tr", "…
## $ anos_educ_infantil   <dbl> 5, 4, NA, 6, 4, 4, 6, 4, NA, 4, 4, 4, 6, 5, 4, 4,…
## $ isec                  <dbl> -0.37127, -0.65631, -0.24825, -1.06724, -0.65631,…
## $ estudios_madre        <chr> "Bachillerato", NA, NA, NA, NA, "Estudios obligat…
## $ repite_curso          <chr> "Repite", "No repite", "Repite", "Repite", "Repit…
## $ expectativas_educ     <chr> "Hasta terminar los estudios obligatorios (ESO)",…
## $ mates_score           <dbl> 384.3092, 451.8030, 419.8543, NA, 498.0551, 483.4…
```

Exploratory Data Analysis (EDA)

Exploratory Data Analysis (EDA)

- `DataExplorer` permite crear un resumen estadístico muy completo de las variables.

```
library(DataExplorer)
create_report(df)
```



Data Profiling Report

- Basic Statistics
  - Raw Counts
  - Percentages
- Data Structure
- Missing Data Profile
- Univariate Distribution
  - Histogram
  - Bar Chart (with frequency)
  - QQ Plot
- Correlation Analysis
- Principal Component Analysis

Basic Statistics

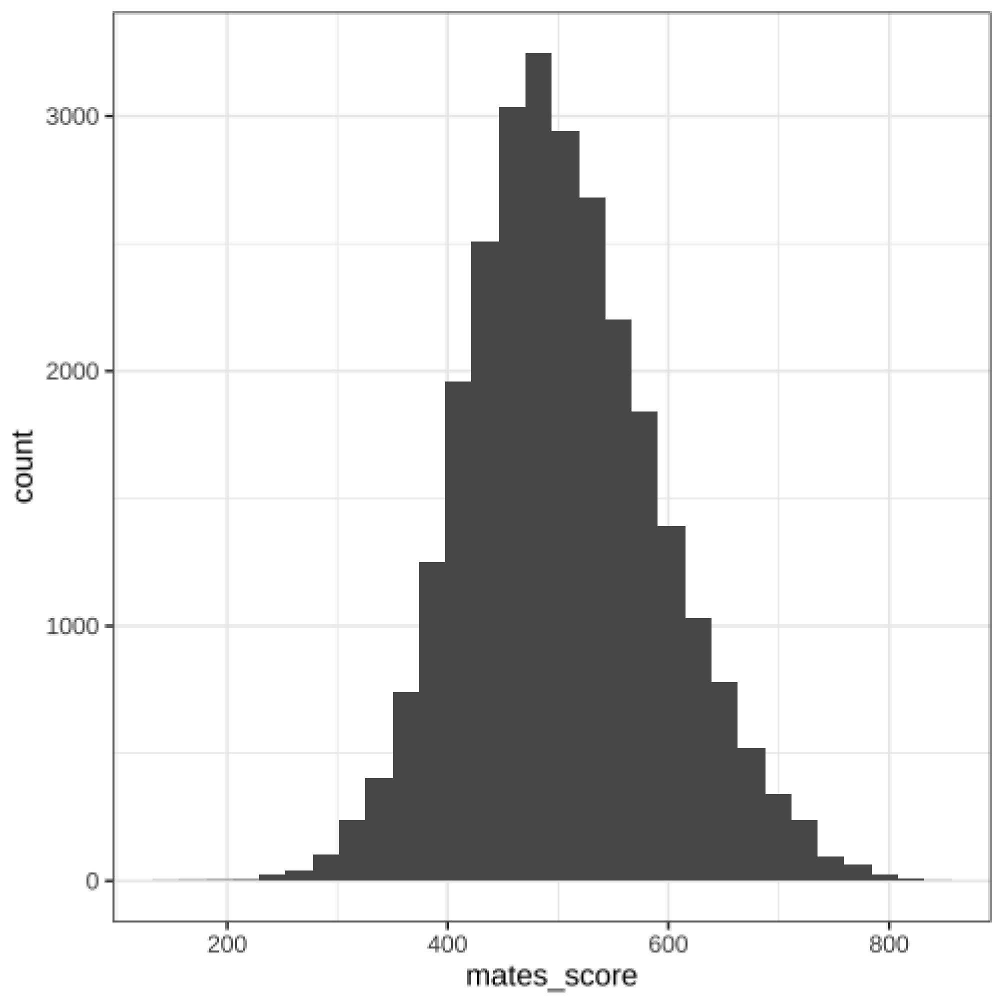
Raw Counts

Name	Value
Rows	29,153
Columns	9
Discrete columns	5
Continuous columns	4
All missing columns	0
Missing observations	17,566
Complete Rows	19,005

Exploratory Data Analysis (EDA)

- Histograma de la variable de respuesta

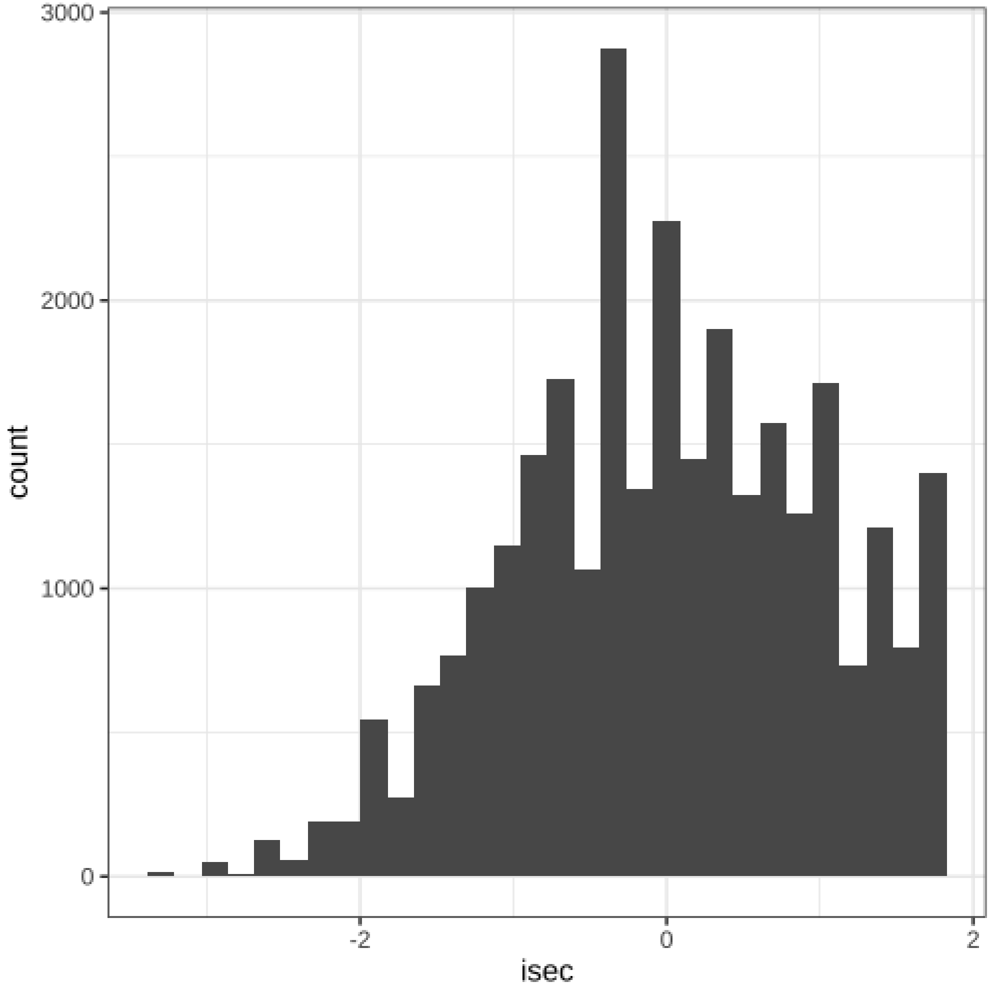
```
df %>%
  ggplot(aes(mates_score)) +
  geom_histogram() +
  theme_bw(base_size = 15)
```



Exploratory Data Analysis (EDA)

- Histograma de la variable predictora

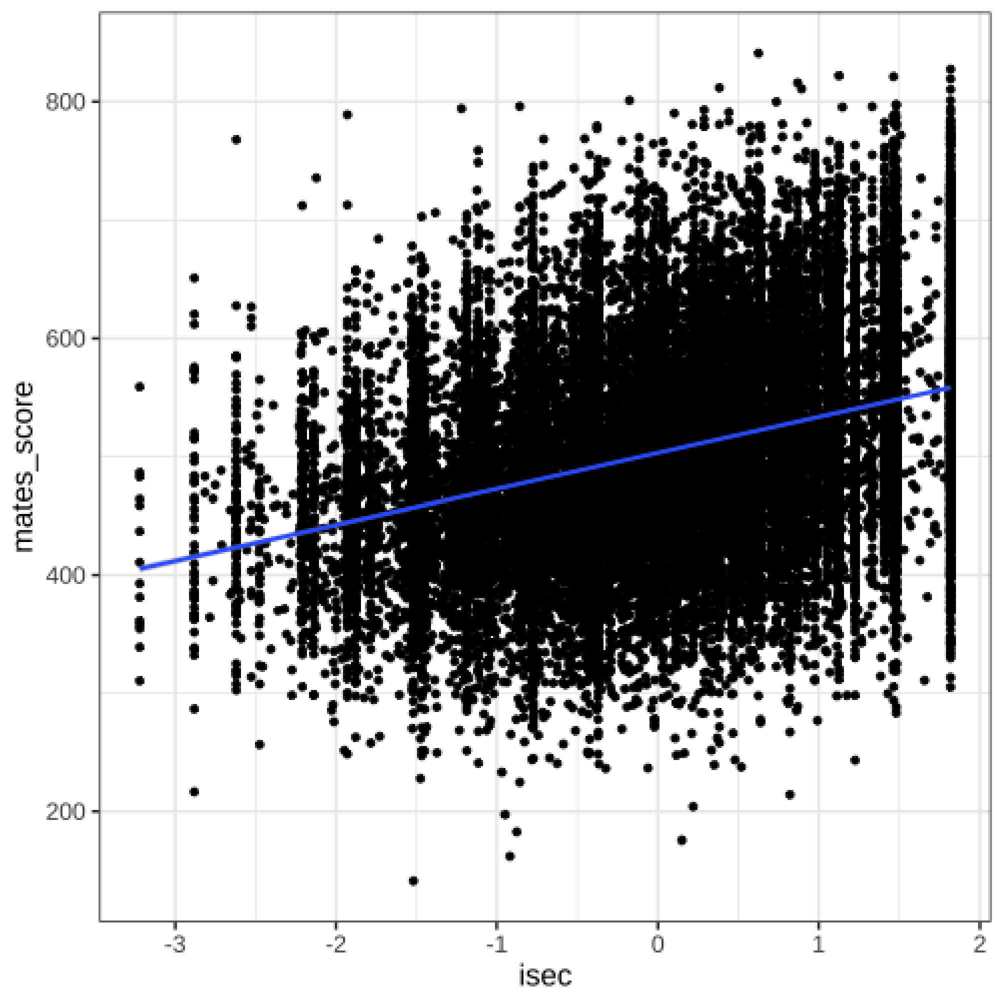
```
df %>%
  ggplot(aes(isec)) +
  geom_histogram() +
  theme_bw(base_size = 15)
```



Exploratory Data Analysis (EDA)

- Gráfico de dispersión: puntuación matemáticas e índice de estatus socioeconómico

```
df %>%
  ggplot(aes(x = isec, y = mates_score)) +
  geom_point() +
  geom_smooth(method = "lm") +
  theme_bw(base_size = 15)
```



Regresión lineal simple: lm()

Regresión lineal simple: lm()

- La regresión lineal se usa para **predecir el valor de una variable Y en función de una o más variables de predicción de entrada X**. Por consiguiente, nos sirve para responder preguntas como....
    - ¿Cuál será el precio de la gasolina mañana en España?
    - ¿Cuánto se gastarán las familias españolas estas navidades?
    - ¿Cuál es el número de votos de un partido "p" en las próximas elecciones generales?
- Objetivos
    - **PREDECIR** los valores que adoptará la variable dependiente (VD) a partir de valores conocidos del conjunto de variables independientes (VIs). Para ello, buscaremos la ecuación que mejor represente la asociación lineal existente entre las variables incluidas en el análisis.
    - **CUANTIFICAR** la relación de dependencia mediante el coeficiente de determinación, que informa de la proporción de varianza de la VD que queda explicada por la suma de VIs.
    - **DETERMINAR EL GRADO DE CONFIANZA** con que se puede afirmar que la relación observada en los datos muestras se da en la población.

Regresión lineal simple: lm()

- Regresión lineal con una sola variable numérica:

```
model1 <- lm(mates_score ~ isec, data = df)
summary(model1)
```

```
##
## Call:
## lm(formula = mates_score ~ isec, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -331.91  -56.10   -4.03   51.86  344.37
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 503.1402     0.4974 1011.62   <2e-16 ***
## isec         30.3868     0.4995   60.84   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 82.83 on 27742 degrees of freedom
##   (1409 observations deleted due to missingness)
## Multiple R-squared:  0.1177,    Adjusted R-squared:  0.1177
## F-statistic:  3701 on 1 and 27742 DF,  p-value: < 2.2e-16
```

Regresión lineal simple: lm()

La regresión lineal puede representarse formalmente de la siguiente manera:

```r
library(equatiomatic)

model1 %>%
  extract_eq(use_coef=FALSE, wrap = TRUE, terms_per_line=1)
```

$$
\begin{aligned}
\operatorname{mates\_score} &= \alpha + \\
&\quad \beta_{1}(\operatorname{isec}) + \\
&\quad \epsilon
\end{aligned}
$$

Regresión lineal simple: lm()

- usando `report()` para ayudarnos a interpretar el modelo

```
library(report)
report(model1)
```

```
## We fitted a linear model (estimated using OLS) to predict mates_score with isec
## (formula: mates_score ~ isec). The model explains a statistically significant
## and weak proportion of variance (R2 = 0.12, F(1, 27742) = 3701.02, p < .001,
## adj. R2 = 0.12). The model's intercept, corresponding to isec = 0, is at 503.14
## (95% CI [502.17, 504.12], t(27742) = 1011.62, p < .001). Within this model:
##
##   - The effect of isec is statistically significant and positive (beta = 30.39,
## 95% CI [29.41, 31.37], t(27742) = 60.84, p < .001; Std. beta = 0.34, 95% CI
## [0.33, 0.35])
##
## Standardized parameters were obtained by fitting the model on a standardized
## version of the dataset. 95% Confidence Intervals (CIs) and p-values were
## computed using a Wald t-distribution approximation.
```

Regresión lineal simple: lm()

¿Cómo podemos interpretar el modelo?

La fórmula del modelo 1 indica que estamos tratando de predecir la puntuación en matemáticas basada en el índice socioeconómico. Podemos interpretar los coeficientes del modelo de la siguiente manera:

- El **coeficiente** asociado a isec es de 30.3868. Esto indica que por cada aumento de una desviación estándar en el índice socioeconómico, se espera un aumento promedio de 30.3868 en la puntuación en matemáticas. Esto sugiere que hay una relación positiva entre el índice socioeconómico y la puntuación en matemáticas, donde los estudiantes con índices socioeconómicos más altos tienden a obtener mejores resultados en matemáticas en comparación con aquellos con índices socioeconómicos más bajos.

- El **valor p** asociado al coeficiente de isec también es muy pequeño (<2e-16), lo que indica que hay evidencia estadística sólida de una relación significativa entre el índice socioeconómico y la puntuación en matemáticas.

- En este caso, el **R-cuadrado** es 0.1177, lo que significa que aproximadamente el 11.77% de la variabilidad en la puntuación en matemáticas puede explicarse por el índice socioeconómico en este modelo. Esto indica que el índice socioeconómico es solo uno de los muchos factores que influyen en la puntuación en matemáticas, y hay otros factores que también deben tenerse en cuenta.

Regresión lineal simple: lm()

Parametros a considerar para interpretar el modelo:

- **Coeficiente de Determinación R2**: El coeficiente de determinación explica cuánta varianza de la variable dependiente y podemos explicar con nuestro modelo. Su valor puede oscilar entre 0 y 1, y cuanto mayor sea su valor, más preciso será el modelo de regresión.

- Los **coeficientes** indican la contribución de cada variable independiente al modelo de regresión. El valor del coeficiente indica que, en promedio, un incremento de una unidad en la variable Xi, produce un incremento de βi en la variable dependiente.

- La evaluación de la **significatividad de los coeficientes (βi)** comienza con la definición de hipótesis sobre los valores de los parámetros poblaciones:

  - Hipótesis nula: H0; Bi=0 (el valor de un determinado coeficiente en la población es 0)
  - Hipótesis alternativa: H1; Bi≠0 (el valor de un determinado coeficiente en la población es distinto de 0). Esta es la hipótesis que esperamos corroborar en nuestros análisis

- El **contraste de hipótesis** siempre se realiza a un nivel de significación que el investigador escoge. El mínimo más recurrente es **valor p=0.05**, que supone una probabilidad de acierto del 95 por ciento (o de 5 por ciento de equivocarse al rechazar la H0 cuando es cierta).

Regresión lineal simple: lm()

- Regresión lineal con una variable factor (o categórica):

```
model2 <- lm(mates_score ~ sexo, data = df)
summary(model2)
```

```
##
## Call:
## lm(formula = mates_score ~ sexo, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -359.29  -60.41   -6.62   56.34  340.16
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 500.8176     0.7669 653.080  < 2e-16 ***
## sexoChico     8.5529     1.0818   7.906 2.75e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 88.12 on 26541 degrees of freedom
##   (2610 observations deleted due to missingness)
## Multiple R-squared:  0.00235,	Adjusted R-squared:  0.002312
## F-statistic: 62.51 on 1 and 26541 DF,  p-value: 2.754e-15
```

Regresión lineal simple: lm()

- Regresión lineal con una variable factor (o categórica):

Usando `relevel()` para elegir la categoría de referencia de la variable factor:

```
df$sexo <- as.factor(df$sexo)
df$sexo <- relevel(df$sexo, ref = "Chico")
```

```
model2 <- lm(mates_score ~ sexo, data = df)
summary(model2)
```

```
##
## Call:
## lm(formula = mates_score ~ sexo, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -359.29  -60.41   -6.62   56.34  340.16
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  509.371      0.763 667.570  < 2e-16 ***
## sexoChica     -8.553      1.082  -7.906 2.75e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Regresión lineal simple: lm()

- usando `report()` para ayudarnos a interpretar el modelo

```
report(model2)
```

```
## We fitted a linear model (estimated using OLS) to predict mates_score with sexo
## (formula: mates_score ~ sexo). The model explains a statistically significant
## and very weak proportion of variance (R2 = 2.35e-03, F(1, 26541) = 62.51, p <
## .001, adj. R2 = 2.31e-03). The model's intercept, corresponding to sexo =
## Chico, is at 509.37 (95% CI [507.88, 510.87], t(26541) = 667.57, p < .001).
## Within this model:
##
##   - The effect of sexo [Chica] is statistically significant and negative (beta =
## -8.55, 95% CI [-10.67, -6.43], t(26541) = -7.91, p < .001; Std. beta = -0.10,
## 95% CI [-0.12, -0.07])
##
## Standardized parameters were obtained by fitting the model on a standardized
## version of the dataset. 95% Confidence Intervals (CIs) and p-values were
## computed using a Wald t-distribution approximation.
```

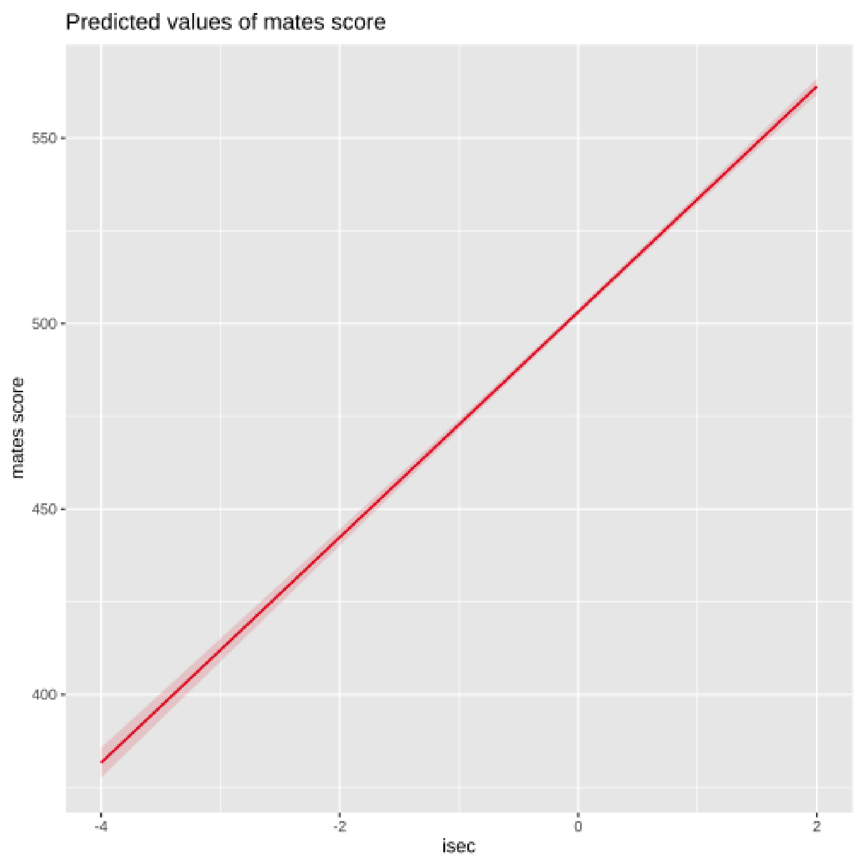
Visualizando los valores pronosticados: lm simple

Visualizando los valores pronosticados: lm simple

Una manera rápida de presentar los resultados de la regresión es representar gráficamente los coeficientes.

```r
library(sjPlot)

plot_model(model1, type="eff")
```

```
## $isec
```

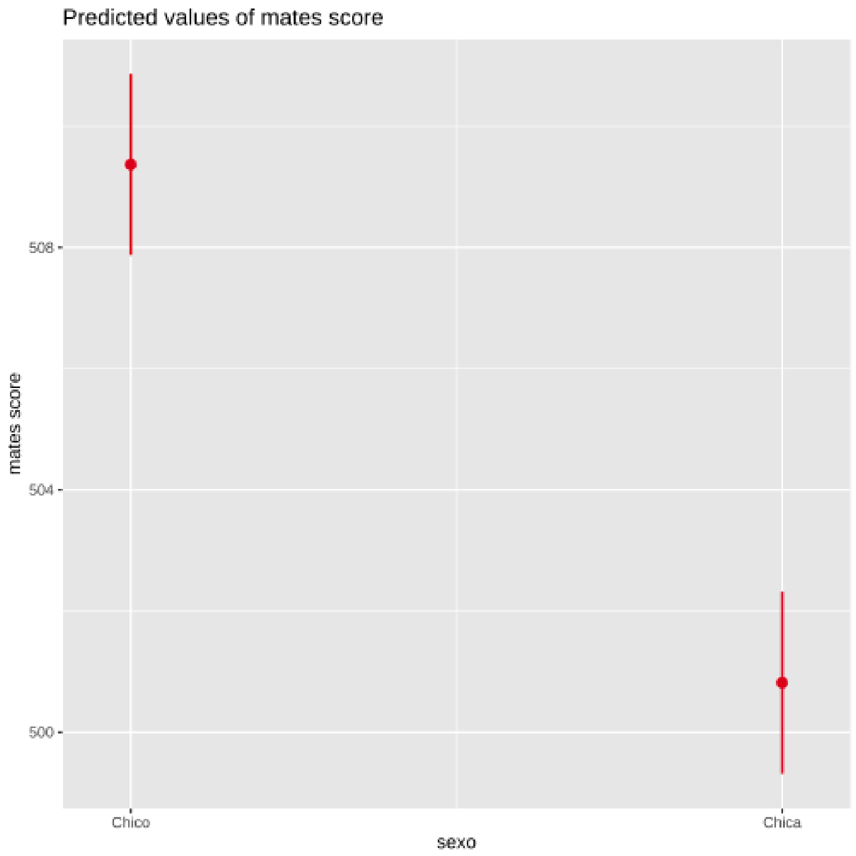


Visualizando los valores pronosticados: lm simple

Una manera rápida de presentar los resultados de la regresión es representar gráficamente los coeficientes.

```r
plot_model(model2, type="eff")
```

```
## $sexo
```



Regresión lineal multivariante

Regresión lineal multivariante

Llamamos regresión lineal múltiple (o multivariante) al análisis de regresión que incluye más de una variable independiente.

```
model3 <- lm(mates_score ~ sexo + anos_educ_infantil + isec, data = df)
summary(model3)
```

```
##
## Call:
## lm(formula = mates_score ~ sexo + anos_educ_infantil + isec,
##     data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -330.47  -55.90   -3.61   52.06  337.80
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)        491.0069     2.3282 210.898  < 2e-16 ***
## sexoChica           -8.7238     1.0241  -8.519  < 2e-16 ***
## anos_educ_infantil   3.9295     0.4828   8.139 4.17e-16 ***
## isec                28.8389     0.5209  55.364  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 82.65 on 26066 degrees of freedom
```

Regresión lineal multivariante

```
report(model3)
```

```
## We fitted a linear model (estimated using OLS) to predict mates_score with
## sexo, anos_educ_infantil and isec (formula: mates_score ~ sexo +
## anos_educ_infantil + isec). The model explains a statistically significant and
## weak proportion of variance (R2 = 0.12, F(3, 26066) = 1190.32, p < .001, adj.
## R2 = 0.12). The model's intercept, corresponding to sexo = Chico,
## anos_educ_infantil = 0 and isec = 0, is at 491.01 (95% CI [486.44, 495.57],
## t(26066) = 210.90, p < .001). Within this model:
##
##   - The effect of sexo [Chica] is statistically significant and negative (beta =
## -8.72, 95% CI [-10.73, -6.72], t(26066) = -8.52, p < .001; Std. beta = -0.10,
## 95% CI [-0.12, -0.08])
##   - The effect of anos educ infantil is statistically significant and positive
## (beta = 3.93, 95% CI [2.98, 4.88], t(26066) = 8.14, p < .001; Std. beta = 0.05,
## 95% CI [0.04, 0.06])
##   - The effect of isec is statistically significant and positive (beta = 28.84,
## 95% CI [27.82, 29.86], t(26066) = 55.36, p < .001; Std. beta = 0.33, 95% CI
## [0.32, 0.34])
##
## Standardized parameters were obtained by fitting the model on a standardized
## version of the dataset. 95% Confidence Intervals (CIs) and p-values were
## computed using a Wald t-distribution approximation.
```

Regresión lineal multivariante

Interpretando el modelo de regresión lineal multivariante

- En regresión lineal multivariante, los coeficientes de regresión representan el cambio medio en la VD para una unidad de cambio en la VI **mientras se mantienen constantes los otros predictores en el modelo**. Este control estadístico que proporciona la regresión es muy importante, porque aisla el papel de una variable de todas las otras del modelo.

```
summary(model3)
```

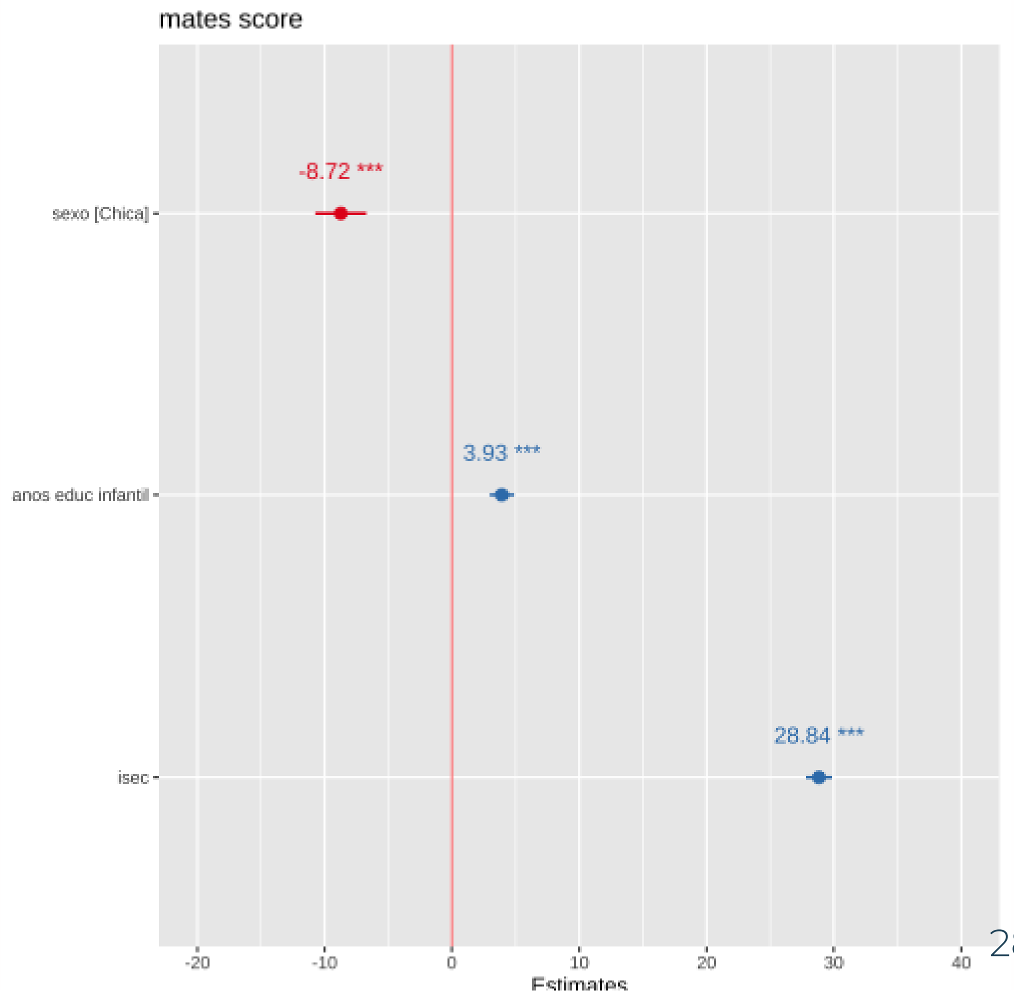
```
##
## Call:
## lm(formula = mates_score ~ sexo + anos_educ_infantil + isec,
##     data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -330.47  -55.90   -3.61   52.06  337.80
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)         491.0069     2.3282 210.898  < 2e-16 ***
## sexoChica            -8.7238     1.0241  -8.519  < 2e-16 ***
## anos_educ_infantil    3.9295     0.4828   8.139 4.17e-16 ***
## isec                 28.8389     0.5209  55.364  < 2e-16 ***
## ---
```

Visualizando los valores pronosticados: lm multivariante

Visualizando los valores pronosticados: lm multivariante

`plot_model` muestra los coeficientes asociados a cada variable (y sus categorías), y permite visualizar información como el grado de significatividad.

```
plot_model(model3,
           show.values = TRUE,
           vline.color = "red")
```



Errores estándar robustos

¿Qué pasa con los **errores estándar robustos o agrupados**? Hay *muchas* formas de obtenerlos en R. Sin embargo, mi forma preferida actualmente es utilizar el paquete `estimatr`.

```
library(estimatr)

model_robust <- lm_robust(mates_score ~ isec, data = df,
                          se_type = "HC1") #calcula los errores estandar robustos
summary(model_robust)
```

```
##
## Call:
## lm_robust(formula = mates_score ~ isec, data = df, se_type = "HC1")
##
## Standard error type:  HC1
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|) CI Lower CI Upper    DF
## (Intercept)   503.14     0.4963  1013.8        0    502.2   504.11 27742
## isec           30.39     0.5015    60.6        0     29.4    31.37 27742
##
## Multiple R-squared:  0.1177 ,    Adjusted R-squared:  0.1177
## F-statistic:  3672 on 1 and 27742 DF,  p-value: < 2.2e-16
```

Otros temas: términos de interacción

Podemos estar interesados en conocer el **efecto moderador** de una tercera variable en la relación entre el índice de estatus socioeconómico y la puntuación en matemáticas.

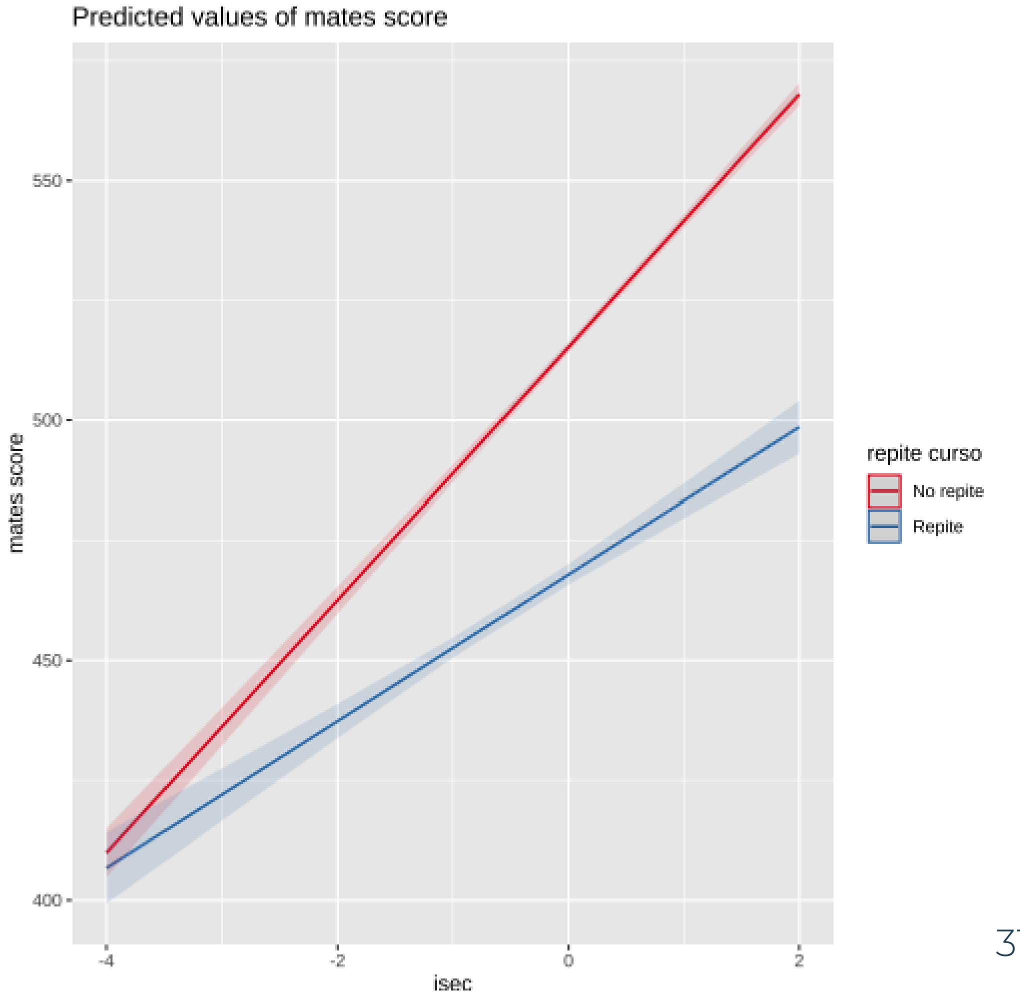
```
model_interaction <- lm(mates_score ~ isec*repite_curso, data = df)
summary(model_interaction)
```

```
##
## Call:
## lm(formula = mates_score ~ isec * repite_curso, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -343.40  -54.52   -3.24   51.50  324.47
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)              515.233      0.593 868.818   <2e-16 ***
## isec                      26.341      0.598  44.052   <2e-16 ***
## repite_cursoRepite       -47.308      1.257 -37.640   <2e-16 ***
## isec:repite_cursoRepite  -11.038      1.214  -9.091   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 80.8 on 26702 degrees of freedom
##   (2447 observations deleted due to missingness)
```

Visualizando los valores pronosticados: interacción

Es recomendable visualizar la interacción porque **facilita su interpretación**. Los coeficientes asociados a los términos de interacción son, por lo general, bastantes complejos de entender a simple vista.

```r
plot_model(model_interaction,
           type = "eff",
           terms = c("isec", "repite_curso"))
```



⚠️ Supuestos de la regresión lineal

El ajuste y análisis del modelo de regresión lineal se sustenta en varias suposiciones basicas. Debemos comprobar que estas hipótesis se cumplen, al menos aproximadamente:

- La relación entre las variables x e y es lineal (una recta)
- La varianza de los errores es constante (heterocesdasticidad)
- Los errores tienen distribucion normal
- Ausencia de multicolinealidad perfecta
- La media de los residuos es igual a cero
- Los errores son independientes

```r
library(performance)

# check_model(model1)
```

Regresión logística: glm()

Regresión logística: glm()

El análisis de regresión logística es una técnica para el análisis de **variables dependientes categóricas**, con dos categorías (dicotómicas) o más (polinómicas). Sirve para modelar la probabilidad de ocurrencia de un evento como función de otros factores, y responder preguntas como:

- ¿Qué factores explican la victoria/derrota de un candidato en unas elecciones?
- ¿Qué variables determinan que una persona fume?
- ¿Qué factores incrementan/disminuyen el riesgo de desempleo?
- ¿Cómo podemos explicar el abandono escolar?
- ¿Qué factores afectan a la probabilidad de tener un/otro hijo?

Regresión logística: glm()

El modelo de regresión lineal no es válido cuando la variable respuesta no es normal, por ejemplo: respuestas si/no, conteos, probabilidades, etc.

Al igual que la regresión lineal, la regresión logística busca:

- **Predecir/explicar** una VD a partir de una o mas VI.
- Medir el grado de relación de la VD con las VI.
- Comprobar su significatividad.

A diferencia de la regresión lineal:

- La función que vincula a las VI con la VD no es lineal, sino logística.
- Los coeficientes de regresión se estiman por el procedimiento de Máxima Verosimilitud, buscando maximizar la probabilidad de ocurrencia del evento que se analiza.

Regresión logística: glm()

Compartidos con la Regresión Lineal:

- Tamaño muestral elevado.
- Introducción de VI relevantes.
- Variables predictoras continuas o dicotómicas.
- Ausencia de colinealidad entre las VI
- Aditividad

Específicos:

- No-linealidad: La función de vinculación logit es no-lineal. Esto implica que el cambio en la VD producido por el incremento de una unidad en la VI depende del valor que dicha variable tenga. Es menos importante en los extremos de las VI, y mas importante en los valores centrales.

Regresión logística: glm()

- la variable dependiente en la regresión logística tiene que ser **factor**.

```
df <- df %>%
  mutate(repite_curso = if_else(repite_curso == "Repite", 1, 0), #1=repite;0=no repite
         repite_curso = as.factor(repite_curso),
         sexo = as.factor(sexo),
         estudios_madre = as.factor(estudios_madre)) #la consideramos como factor
```

- Podemos emplear la función `class()` para asegurarnos de que es factor:

```
class(df$repite_curso)
```

```
## [1] "factor"
```

- Establecemos las **categorías de referencia** para las variables con `relevel`. Esto se suele elegir de acuerdo con la literatura sobre el tema de estudio o a criterio del investigador/a.

```
df$sexo <- relevel(df$sexo, ref = "Chico")
df$estudios_madre <- relevel(df$estudios_madre, ref = "Universitarios superiores")
```

Regresión logística: glm()

- Definimos el modelo de regresión con la función `glm()`

```
model1_glm <- glm(repite_curso ~ sexo + estudios_madre,
                  data = df, family = binomial("logit"))
summary(model1_glm)
```

```
##
## Call:
## glm(formula = repite_curso ~ sexo + estudios_madre, family = binomial("logit"),
##     data = df)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.2492  -0.8554  -0.5675   1.1073   2.2570
##
## Coefficients:
##                                              Estimate Std. Error z value
## (Intercept)                                  -1.92792    0.05928 -32.524
## sexoChica                                    -0.43532    0.03336 -13.048
## estudios_madreBachillerato                    0.84970    0.06983  12.168
## estudios_madreEstudios obligatorios (ESO,EGB)  1.54629    0.06330  24.428
## estudios_madreSin estudios obligatorios       2.09511    0.07189  29.143
## estudios_madreTecnico FP grado medio          1.04771    0.09111  11.500
## estudios_madreTecnico superior FP             0.61852    0.08874   6.970
## estudios_madreUniversitarios medios          -0.10216    0.09126  -1.119
```

Regresión logística: glm()

La regresión logística puede representarse formalmente de la siguiente manera:

```
model1_glm %>%
  extract_eq(use_coef=FALSE, wrap = TRUE, terms_per_line=1)
```

$$
\begin{aligned}
\log\left[\frac { P(\operatorname{repite\_curso} = \operatorname{1}) }{ 1 - P(\operatorname{repite\_curso} = \operatorname{1}) } \right] &= \alpha + \\
&\quad \beta_{1}(\operatorname{sexo}_{\operatorname{Chica}}) + \\
&\quad \beta_{2}(\operatorname{estudios\_madre}_{\operatorname{Bachillerato}}) + \\
&\quad \beta_{3}(\operatorname{estudios\_madre}_{\operatorname{Estudios\ obligatorios\ (ESO,EGB)}}) + \\
&\quad \beta_{4}(\operatorname{estudios\_madre}_{\operatorname{Sin\ estudios\ obligatorios}}) + \\
&\quad \beta_{5}(\operatorname{estudios\_madre}_{\operatorname{Tecnico\ FP\ grado\ medio}}) + \\
&\quad \beta_{6}(\operatorname{estudios\_madre}_{\operatorname{Tecnico\ superior\ FP}}) + \\
&\quad \beta_{7}(\operatorname{estudios\_madre}_{\operatorname{Universitarios\ medios}})
\end{aligned}
$$

Regresión logística: glm()

```
report(model1_glm)
```

```
## We fitted a logistic model (estimated using ML) to predict repite_curso with
## sexo and estudios_madre (formula: repite_curso ~ sexo + estudios_madre). The
## model's explanatory power is weak (Tjur's R2 = 0.09). The model's intercept,
## corresponding to sexo = Chico and estudios_madre = Universitarios superiores,
## is at -1.93 (95% CI [-2.05, -1.81], p < .001). Within this model:
##
##   - The effect of sexo [Chica] is statistically significant and negative (beta =
## -0.44, 95% CI [-0.50, -0.37], p < .001; Std. beta = -0.44, 95% CI [-0.50,
## -0.37])
##   - The effect of estudios madre [Bachillerato] is statistically significant and
## positive (beta = 0.85, 95% CI [0.71, 0.99], p < .001; Std. beta = 0.85, 95% CI
## [0.71, 0.99])
##   - The effect of estudios madre [Estudios obligatorios (ESO,EGB)] is
## statistically significant and positive (beta = 1.55, 95% CI [1.42, 1.67], p <
## .001; Std. beta = 1.55, 95% CI [1.42, 1.67])
##   - The effect of estudios madre [Sin estudios obligatorios] is statistically
## significant and positive (beta = 2.10, 95% CI [1.96, 2.24], p < .001; Std. beta
## = 2.10, 95% CI [1.96, 2.24])
##   - The effect of estudios madre [Tecnico FP grado medio] is statistically
## significant and positive (beta = 1.05, 95% CI [0.87, 1.23], p < .001; Std. beta
## = 1.05, 95% CI [0.87, 1.23])
##   - The effect of estudios madre [Tecnico superior FP] is statistically
## significant and positive (beta = 0.62, 95% CI [0.44, 0.79], p < .001; Std. beta
```

Regresión logística: glm()

Los estimadores representan el logaritmo del cociente de probabilidades. Por ejemplo:

- el coeficiente para "sexoChica" es -0.43532, lo que significa que ser chica en lugar de chico se asocia con una disminución de la probabilidad de repetir el curso.

Esta interpretación de los coeficientes es muy poco intuitiva. Tenemos varias alternativas: expresar los coeficientes como **odds ratio**, calcular las **probabilidades predichas** o calcular los **efectos marginales**. Principalmente, veremos las dos últimas.

- Por ejemplo, los Odds Ratio se pueden calcular de la siguiente manera:

```
exp(cbind(OR = coef(model1_glm), confint(model1_glm)))
```

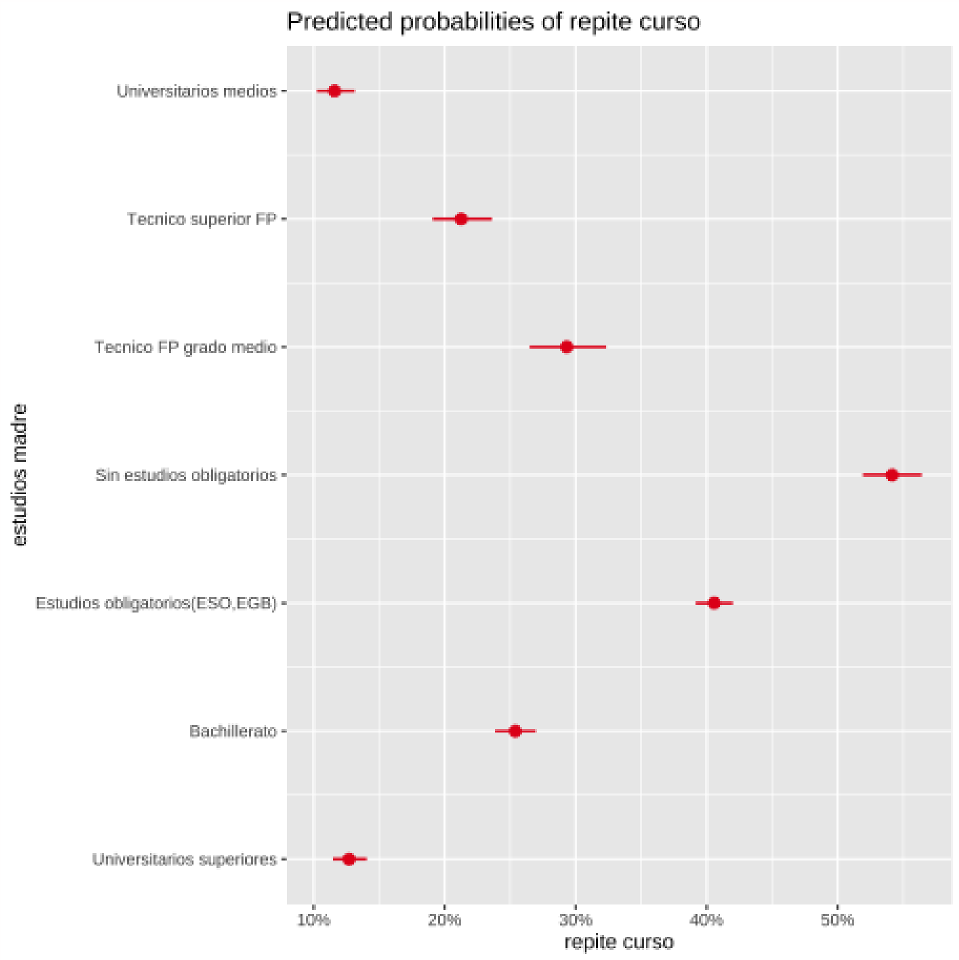
```
##                                                    OR     2.5 %    97.5 %
## (Intercept)                                 0.1454509 0.1292864 0.1631177
## sexoChica                                   0.6470582 0.6060600 0.6907427
## estudios_madreBachillerato                  2.3389372 2.0416606 2.6847104
## estudios_madreEstudios obligatorios (ESO,EGB) 4.6940380 4.1518785 5.3215987
## estudios_madreSin estudios obligatorios     8.1263051 7.0656390 9.3663143
## estudios_madreTecnico FP grado medio        2.8511237 2.3841113 3.4078994
## estudios_madreTecnico superior FP           1.8561876 1.5591997 2.2082127
## estudios_madreUniversitarios medios         0.9028839 0.7542749 1.0789089
```

Visualizando los valores pronosticados: glm

Visualizando los valores pronosticados: plot_model

- Calculamos con `plot_model` las **probabilidades predichas** de repetir curso según el nivel educativo de la madre

```
plot_model(model1_glm,
           type = "pred",
           terms = c("estudios_madre")) +
  coord_flip() # gira los ejes del gráfico
```



Visualizando los valores pronosticados: ggeffects

- `ggeffects` nos devuelve los valores en un dataframe que podemos combinar fácilmente con ggplot para la visualización de los resultados de una manera más estilizada:

- Cargamos la librería y llamamos a `ggpredict()`:

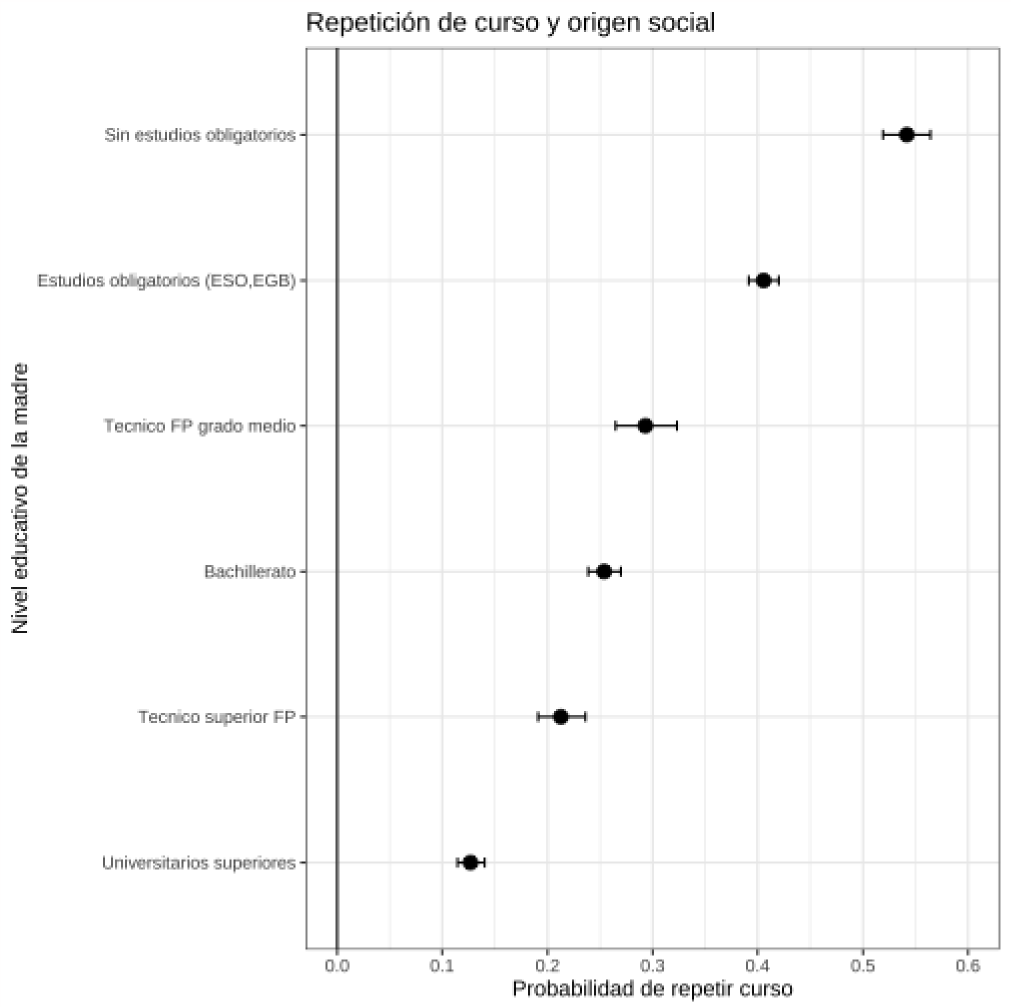
```
library(ggeffects)

ggdata1 <- ggpredict(model1_glm, terms = c("estudios_madre"))

head(ggdata1)
```

```
## # Predicted probabilities of repite_curso
##
## estudios_madre                 | Predicted |       95% CI
## ----------------------------------------------------------
## Universitarios superiores      |      0.13 | [0.11, 0.14]
## Bachillerato                   |      0.25 | [0.24, 0.27]
## Estudios obligatorios (ESO,EGB) |      0.41 | [0.39, 0.42]
## Sin estudios obligatorios      |      0.54 | [0.52, 0.56]
## Tecnico FP grado medio         |      0.29 | [0.26, 0.32]
## Tecnico superior FP            |      0.21 | [0.19, 0.24]
##
## Adjusted for:
## * sexo = Chico
```

Visualizando los valores pronosticados: ggeffects

- Mejoramos la visualización del gráfico mediante las funciones de `ggplot()`

```
ggdata1 %>%
  mutate(x = reorder(x, predicted)) %>%
  ggplot(aes(x = x, y = predicted)) +
  geom_point(position = position_dodge(width=0.3),
             size=3) +
  geom_errorbar(aes(ymin=conf.low, ymax=conf.high),
                width = 0.07,
                position = position_dodge(width=0.3))+
  geom_hline(yintercept = 0, col = "black") +
  scale_y_continuous(limits = c(0,0.6),
                     breaks = seq(0,0.6,by=0.1)) +
  coord_flip() +
  labs(title = "Repetición de curso y origen social",
       x = "Nivel educativo de la madre",
       y = "Probabilidad de repetir curso")+
  theme_bw()
```



Average Marginal Effects (AMEs)

- Los average marginal effects (AMEs) se utilizan para medir el impacto promedio de un cambio en una variable independiente sobre la variable dependiente, manteniendo todas las demás variables constantes.

- Una de las principales ventajas de los AMEs es que permiten comparar los efectos de diferentes variables independientes en una escala común.

- Cómo se interpreta: si el AME del nivel educativo de la madre "Sin estudios" es 0.38, esto significa que tener una madre con un nivel educativo "sin estudios" (en comparación con tener una madre con estudios "universitarios", manteniendo el resto de variables constante) se asocia, en promedio, con un aumento del 38% en el hecho de repetir curso.

```r
library(margins)

margins_summary(model1_glm, data = df)
```

```
##                                          factor     AME     SE
##                    estudios_madreBachillerato  0.1107 0.0086
##  estudios_madreEstudios obligatorios (ESO,EGB)  0.2497 0.0080
##        estudios_madreSin estudios obligatorios  0.3812 0.0118
##           estudios_madreTecnico FP grado medio  0.1459 0.0143
##              estudios_madreTecnico superior FP  0.0742 0.0113
##            estudios_madreUniversitarios medios -0.0093 0.0083
##                                      sexoChica -0.0765 0.0058
##    lower   upper
##   0.0939  0.1275
##   0.2340  0.2654
##   0.3580  0.4044
##   0.1180  0.1739
##   0.0520  0.0964
##  -0.0255  0.0069
##  -0.0879 -0.0651
```

Average Marginal Effects (AMEs)

- Añadimos los AME a un objeto que es un dataframe

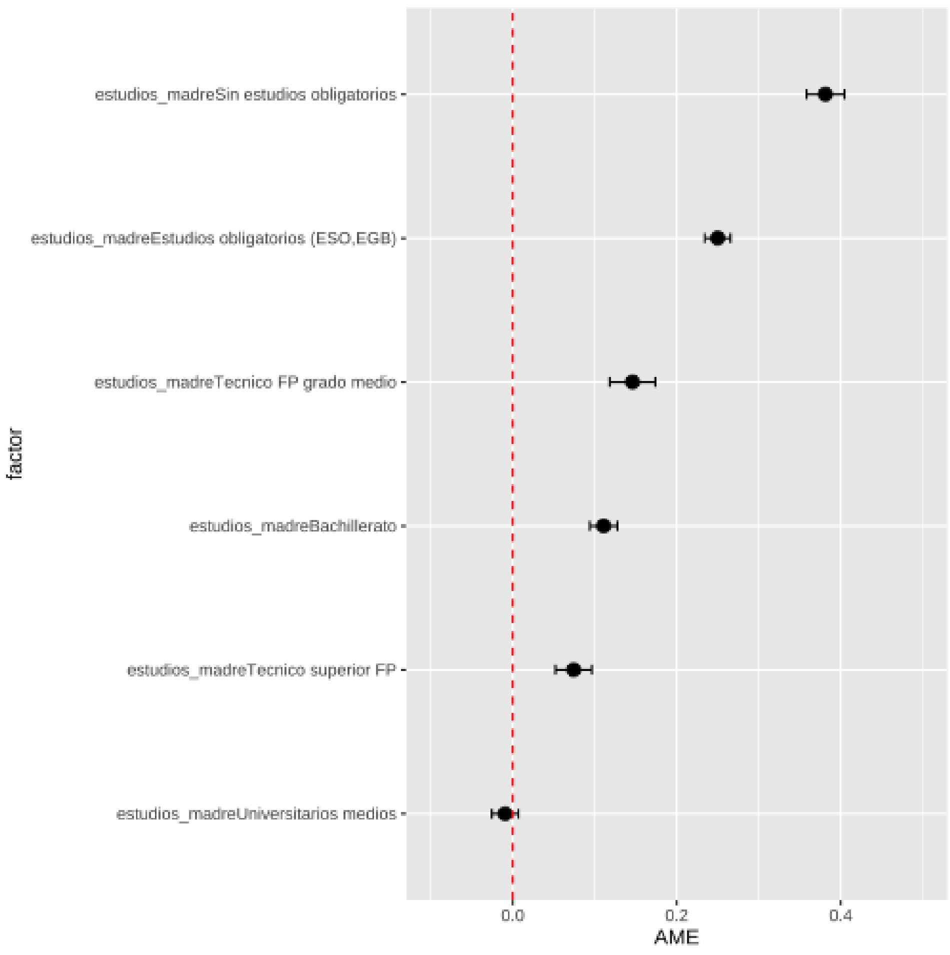
```
#añadimos los AME a un objeto que es un dataframe
ame <- margins_summary(model1_glm, data =  df, variables = "estudios_madre")
ame
```

```
##                                     factor     AME     SE       z      p
##                estudios_madreBachillerato  0.1107 0.0086 12.8932 0.0000
##  estudios_madreEstudios obligatorios (ESO,EGB)  0.2497 0.0080 31.1962 0.0000
##        estudios_madreSin estudios obligatorios  0.3812 0.0118 32.2135 0.0000
##           estudios_madreTecnico FP grado medio  0.1459 0.0143 10.2275 0.0000
##              estudios_madreTecnico superior FP  0.0742 0.0113  6.5548 0.0000
##            estudios_madreUniversitarios medios -0.0093 0.0083 -1.1271 0.2597
##    lower  upper
##   0.0939 0.1275
##   0.2340 0.2654
##   0.3580 0.4044
##   0.1180 0.1739
##   0.0520 0.0964
##  -0.0255 0.0069
```

Average Marginal Effects (AMEs)

- Los visualizamos con `ggplot`:

```
ame %>%
  mutate(factor = reorder(factor, AME)) %>%
  ggplot(aes(x = factor, y = AME)) +
  geom_point(size=3) +
  geom_errorbar(aes(ymin=lower, ymax=upper),
                width = 0.07)+
  geom_hline(yintercept = 0,
             col = "red",
             linetype = 2) +
  scale_y_continuous(limits = c(-0.1,0.5))+
  coord_flip()
```



Exportando los resultados de los modelos de regresión

Exportando los resultados en tablas

```
tab_model(model1) #librería sjPlot
```

	mates score		
Predictors	*Estimates*	*CI*	*p*
(Intercept)	503.14	502.17 – 504.12	**<0.001**
isec	30.39	29.41 – 31.37	**<0.001**
Observations	27744		
R^2 / R^2 adjusted	0.118 / 0.118		

Exportando los resultados en tablas

```
tab_model(model1,
          p.style = "stars") #añadimos asteriscos para marcar la significatividad de los valores
```

	mates score	
Predictors	*Estimates*	*CI*
(Intercept)	503.14 ***	502.17 – 504.12
isec	30.39 ***	29.41 – 31.37
Observations	27744	
R^2 / R^2 adjusted	0.118 / 0.118	

** p<0.05   ** p<0.01   *** p<0.001*

Exportando los resultados en tablas

```
tab_model(model1,model2, p.style = "stars", dv.labels = c("Modelo 1", "Modelo 2"))
```

	Modelo 1		**Modelo 2**	
Predictors	*Estimates*	*CI*	*Estimates*	*CI*
(Intercept)	503.14 ***	502.17 – 504.12	509.37 ***	507.88 – 510.87
isec	30.39 ***	29.41 – 31.37		
sexo [Chica]			-8.55 ***	-10.67 – -6.43
Observations	27744		26543	
R^2 / R^2 adjusted	0.118 / 0.118		0.002 / 0.002	
				* *p<0.05* ** *p<0.01* *** *p<0.001*

Exportando los resultados en tablas

- Podemos exportar la tabla de regresión a un archivo .doc:

```
tab_model(model1, p.style = "stars", dv.labels = c("Modelo 1"),
          file = "tabla_regresion.doc")
```

	Modelo 1	
Predictors	*Estimates*	*CI*
(Intercept)	503.14 ***	502.17 – 504.12
isec	30.39 ***	29.41 – 31.37
Observations	27744	
R^2 / R^2 adjusted	0.118 / 0.118	

* p<0.05   ** p<0.01   *** p<0.001

Recursos para seguir aprendiendo sobre regresiones en R

Más sobre regresiones 🔍

- **Regression and Other Stories**. Andrew Gelman, et al: https://avehtari.github.io/ROS-Examples/ 📚
- **Thinking Clearly with Data**. Ethan Bueno de Mesquita y Anthony Fowler: https://press.princeton.edu/books/hardcover/9780691214368/thinking-clearly-with-data 📚
- **El arte de la estadística. Cómo aprender de los datos**. David Spiegelhalter: https://capitanswing.com/libros/el-arte-de-la-estadistica/ 📚