Análisis e Interpretación de Datos

MÁSTER UNIVERSITARIO EN ANÁLISIS Y VISUALIZACIÓN DE DATOS MASIVOS / VISUAL ANALYTICS AND BIG DATA

Miller Janny Ariza Garzón

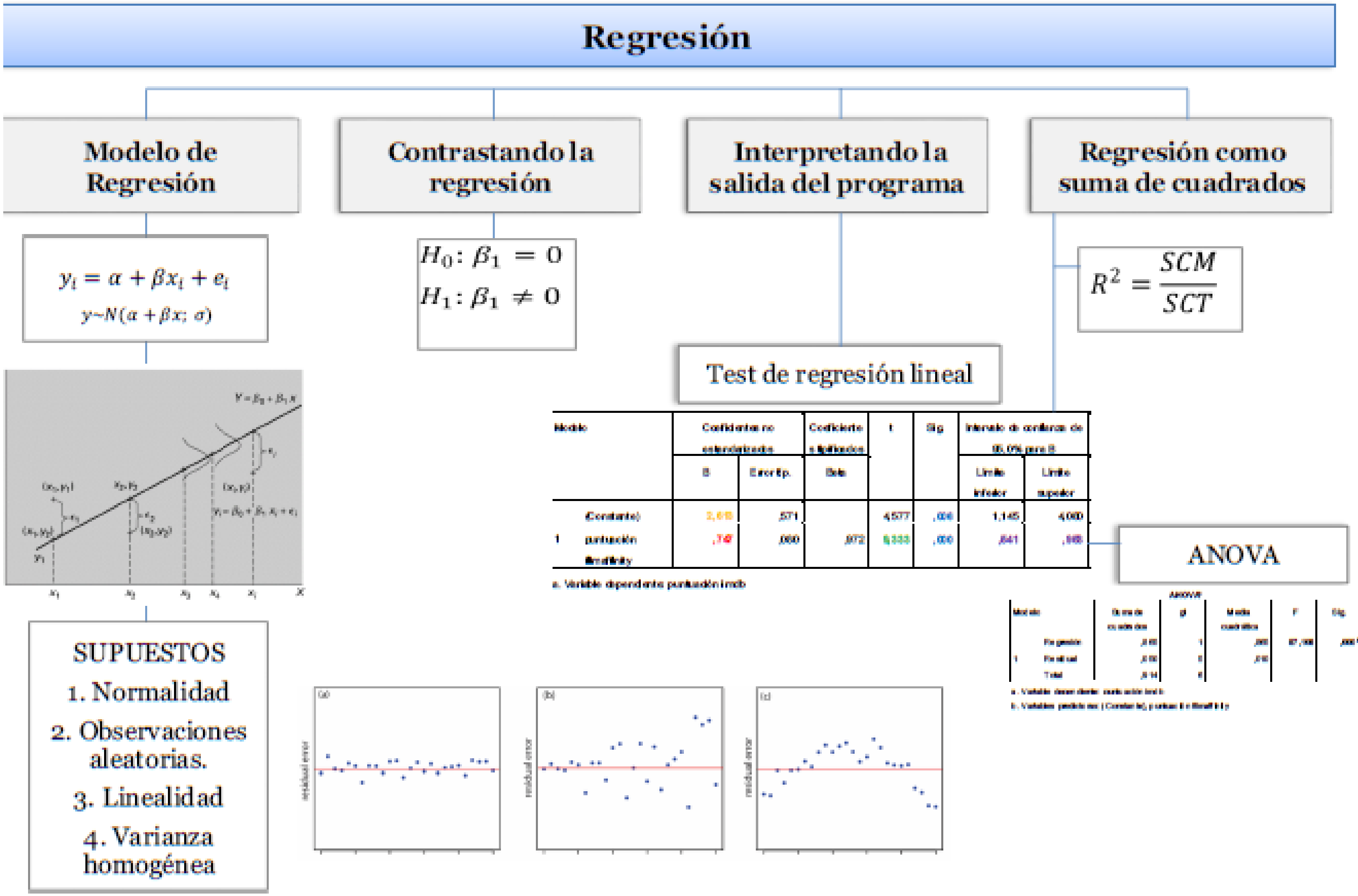
Tema 9. Regresión

LA UNIVERSIDAD EN INTERNET

Universidad Internacional de La Rioja

Tabla de contenido

Tabla de contenido



Modelación

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_k X_{ik} + \varepsilon_i$$

Se deben desarrollar varios pasos:

1. Especificación
2. Estimación (parámetros)
3. Análisis inferencial (**pruebas t's** y F)
4. Análisis inferencial (pruebas t's y **F**)
5. Evaluación de bondad de ajuste (R2 y s)
6. Validación de supuestos
7. Interpretación de estimaciones
8. Predicción o Pronóstico

Modelación

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_k X_{ik} + \varepsilon_i$$

1. Especificación (Problema-soporte teórico)
2. Estimación modelo lineal simple:

M.C.O

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

$$\sum (Y - \beta_0 + \beta_1 X_i)^2 = \sum {\varepsilon_i}^2$$

$$\frac{\partial \sum {\varepsilon_i}^2}{\partial \beta_0} = ?$$

$$\frac{\partial \sum {\varepsilon_i}^2}{\partial \beta_1} = ?$$

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$$

$$\hat{\beta}_1 = \frac{\sum[(X - \bar{X})(Y - \bar{Y})]}{\sum (X - \bar{X})^2}$$

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \frac{Cov\,(x, y)}{Var(x)}$$

Modelación

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_k X_{ik} + \varepsilon_i$$

1. Especificación (Problema-soporte teórico)
2. Estimación modelo lineal múltiple:

$$\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon}\ ; \quad \boldsymbol{\varepsilon} \sim \boldsymbol{N}(\boldsymbol{0}, \boldsymbol{\sigma}^2 \boldsymbol{I})$$

$$\hat{\beta} = (X'X)^{-1}\mathrm{X}'\mathrm{Y}$$

$$Var(\hat{\beta}) = s^2 (X'X)^{-1}$$

$$s^2 = \frac{\varepsilon'\varepsilon}{n-(k+1)}$$

Modelación

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_k X_{ik} + \varepsilon_i$$

3. Análisis inferencial (**pruebas t's** y F)

$H_0\!: \beta_1 = 0$ $H_I\!: \beta_1 \neq 0$

$H_0\!: \beta_1 = 0$ $H_I\!: \beta_1 \neq 0$

$H_0\!: \beta_i = 0$ $H_I\!: \beta_i \neq 0$

$$t_c = \frac{\hat{\beta}_i - \beta_{i_{(H_0)}}}{\delta_{\hat{\beta}_i}} \sim t_{n-(k+1)}$$

k: Número de variables

Al rechazar la hipótesis nula, se considera la variable X_i relevante individualmente para explicar el comportamiento de Y

Modelación

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_k X_{ik} + \varepsilon_i$$

4. Análisis inferencial (pruebas t's y **F**)

Tabla ANOVA				
Fuente de variación	**Grados de libertad**	**Suma de cuadrados**	**Varianza**	**"F" calculado**
Modelo	k	SCM	SCM/k	$\frac{SCM/k}{SCR/(n-(k+1))}$
Residuos	$n-(k+1)$	SCR	$SCR/(n-(k+1))$	
Total	$n-1$	SCT		

$$H_0\colon \beta_1 = \beta_2 = \cdots \beta_k = 0$$

Al rechazar la hipótesis nula se concluye que **al menos una** de las variables explicativas es relevante para explicar Y

Modelación

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_k X_{ik} + \varepsilon_i$$

5. Evaluación de bondad de ajuste (R2 y s)

Tabla ANOVA				
Fuente de variación	**Grados de libertad**	**Suma de cuadrados**	**Varianza**	**"F" calculado**
Modelo	k	SCM	SCM/k	$\frac{SCM/k}{SCR/(n-(k+1))}$
Residuos	$n\text{-}(k+1)$	SCR	$SCR/(n\text{-}(k+1))$	
Total	$n\text{-}1$	SCT		

$R^2 = \frac{SCM}{SCT}$ porcentaje de variación de la variable dependiente que es explicado por el comportamiento de las variables independientes.

$s = \sqrt{CMR} = \sqrt{\frac{SCR}{n-(k+1)}}$ Promedio de error al pronosticar Y

Modelación

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_k X_{ik} + \varepsilon_i$$

6. Validación de supuestos. (Gráficos-residuos- y Pruebas de hipótesis)

- Relación lineal
- Linealidad de los parámetros.
- El valor medio de la perturbación es igual a cero $E(\varepsilon_i) = 0$.
- Homocedasticidad $V(\varepsilon_i) = \sigma^2$
- No autocorrelación $Cov(\varepsilon_i, \varepsilon_j) = 0$
- No multicolinealidad excesiva (Regresión múltiple)
- El error se distribuye como una normal $\varepsilon_i \sim N$

$$\boldsymbol{\varepsilon \sim N(0, \sigma^2 I)}$$

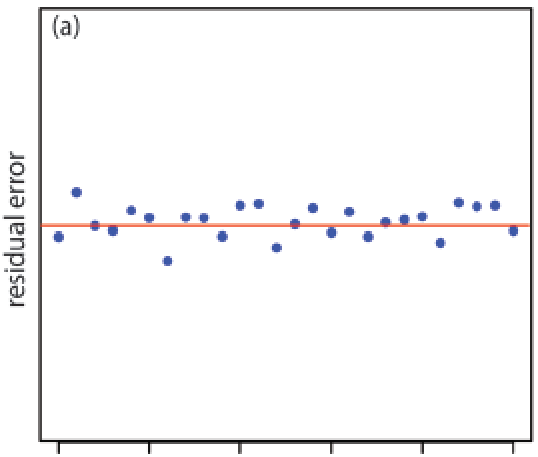
Modelación

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_k X_{ik} + \varepsilon_i$$

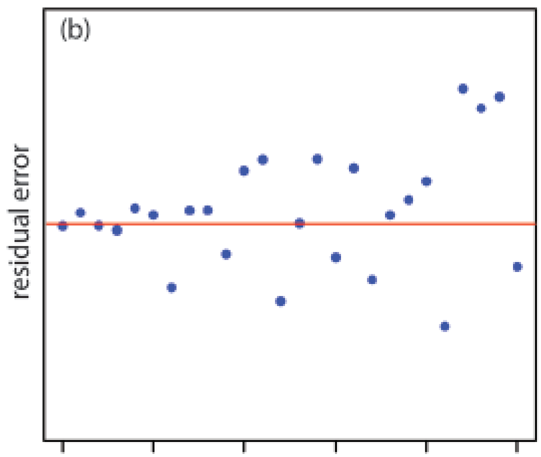
6. Validación de supuestos. (Gráficos-residuos- y Pruebas de hipótesis)

aleatoriedad



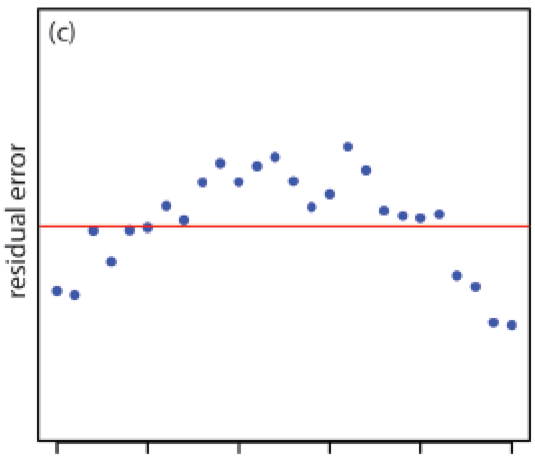
- Deseable.
- Errores unos por encima y otros por debajo.
- No varía su magnitud.

heterocedasticidad



- Varianza no es constante → heterocedasticidad

no linealidad



- Falta de linealidad.

Modelación

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_k X_{ik} + \varepsilon_i$$

7. Interpretación de las estimaciones. Evaluar magnitudes y signos

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{i1}$$

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{i1} + \hat{\beta}_2 X_{i2} + \cdots + \hat{\beta}_k X_{ik}$$

$\hat{\beta}_j$: Promedio de variación de Y por cada unidad que varíe X_j, bajo ceteris paribus. ($j = 1, \ldots k$)

8. Predicción o Pronóstico. Reemplazar en la ecuación diferentes valores de X's para saber en promedio como se comporta Y.

Ejemplo

¿Es posible explicar las Y: puntuaciones de IMDB en función de las X: puntuaciones de FilmAffinity de las películas nominadas al Óscar (2015)? (Evaluación Inferencial)

IMDB : Internet Movie Database. Base de datos en línea, de las más grandes del mundo. Almacena información relacionada con películas. Se basa en votos de los usuarios registrados del sitio web. Solamente son tomados en consideración los largometrajes que tengan más de 25 000 calificaciones de usuarios.

FilmAffinity: es un sitio web español dedicado al cine. Una de las mejores webs en la categoría de entretenimiento.

	Filmaffinity	IMBD
Bidrman	7.2	7.9
Boyhood	7.4	8.1
El gran Hotel Budapest	7.2	8.1
El Francotirador	6.3	7.4
La Teoría del Todo	7.1	7.8
Whiplash	7.9	8.6
Selma	6.7	7.6

Ejemplo

Coeficientes[a]

Modelo		Coeficientes no estandarizados		Coeficientes tipificados	t	Sig.	Intervalo de confianza de 95,0% para B	
		B	Error típ.	Beta			Límite inferior	Límite superior
1	(Constante)	2,613	,571		4,577	,006	1,145	4,080
	puntuación filmaffinity	,747	,080	,972	9,333	,000	,541	,953

a. Variable dependiente: puntuación imdb

- Naranja y rojo: estimación de los coeficientes. 0.747 es la pendiente → positiva indica similar tendencia.
- Verde: Estadístico t → en este caso tiene un valor muy elevado. Nos ayuda decidir sobre el rechazo o no de H_0. Se rechaza H_0.
- Azul: P-valores de constante y pendiente.
  - Son muy significativos (< 0.01).
  - Cuanto más pequeño más significativa es la variable
  - El segundo es especialmente importante por ser el de la pendiente.
- Morado: Intervalo de confianza para la pendiente.
  - No contiene el valor 0 → la pendiente es distinta de 0.

Ejemplo

ANOVA[a]

Modelo		Suma de cuadrados	gl	Media cuadrática	F	Sig.
1	Regresión	,865	1	,865	**87,108**	**,000b**
	Residual	,050	5	,010		
	Total	,914	6			

a. Variable dependiente: puntuación imdb

b. Variables predictoras: (Constante), puntuación filmaffinity

Prueba de dependencia, prueba conjunta, prueba derivada del análisis de varianza (ANOVA).

- Como $P - Value < \alpha$, existe dependencia, la variable puntaje IMDB es significativa para explicar el comportamiento del puntaje de filmaffinity.

- $R^2 = \frac{SCM}{SCT} = \frac{0.865}{0.914} = 0.9463$; 94.6% de la variabilidad del puntaje IMDB es explicado por el puntaje de filmaffinity
- $s = error\ típico\ del\ modelo = \sqrt{CMR} = \sqrt{0.010} = 0.1$, en promedio, al pronosticar el puntaje IMDB, se comete un error de 0.1.

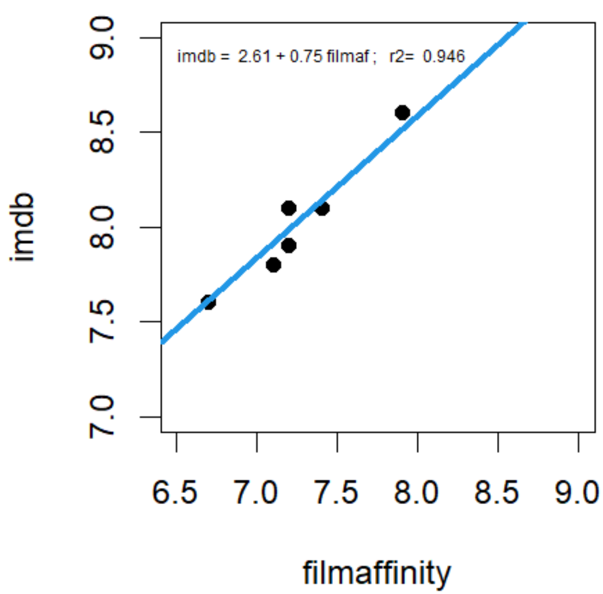
Ejemplo

In R

```
imdb=c(7.9,8.1,8.1,7.4,7.8,8.6,7.6)
filmaffinity=c(7.2,7.4,7.2,6.3,7.1,7.9,6.7)

plot(filmaffinity, imdb, pch = 16,ylim=c(7,9),xlim=c(6.5,9))
abline(lm(imdb~ filmaffinity), col = 4, lwd = 3)
coef <- round(coef(lm(imdb ~ filmaffinity)), 2)
text(7.4, 8.9,  paste("imdb = ", coef[1], "+", coef[2], "filmaf",";  "," r2= ",
                    format(summary(lm(imdb~ filmaffinity))$r.squared, digits = 3)), cex=0.5)

summary(lm(imdb ~ filmaffinity))
```



Coeficientes

Est. t's

P-values

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  2.61255    0.57083   4.577 0.005965 ** 
filmaffinity 0.74723    0.08006   9.333 0.000238 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.09963 on 5 degrees of freedom
Multiple R-squared:  0.9457,	Adjusted R-squared:  0.9349 
F-statistic: 87.11 on 1 and 5 DF,  p-value: 0.0002378
```

Est. F

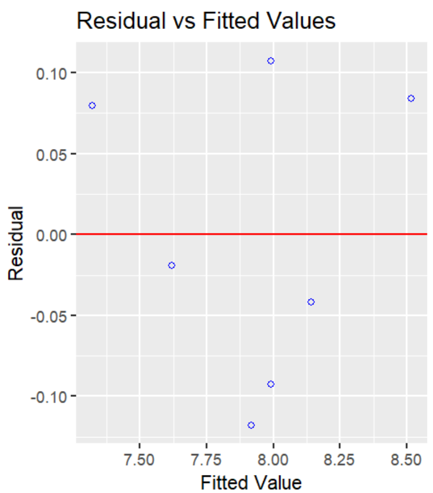
R^2

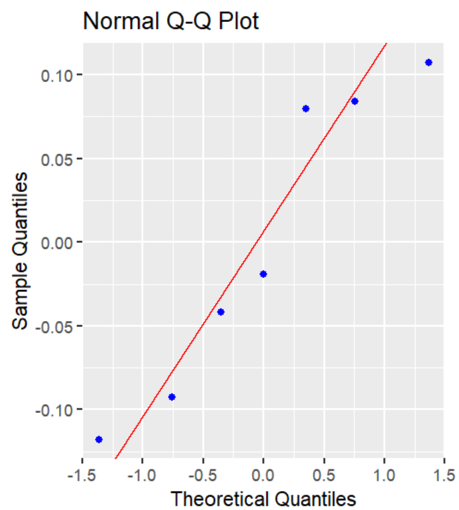
s

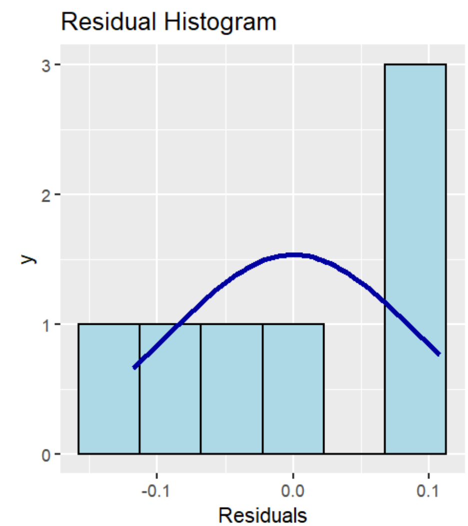
P-value F

Ejemplo

library("olsrr")
ols_plot_resid_fit(lm(imdb ~ filmaffinity))
ols_plot_resid_qq(lm(imdb ~ filmaffinity))
ols_plot_resid_hist(lm(imdb ~ filmaffinity))







err=lm(imdb ~ filmaffinity)$residuals
ks.test(err,"pnorm")

```
        One-sample Kolmogorov-Smirnov test

data:  err
D = 0.45724, p-value = 0.07421
alternative hypothesis: two-sided
```

Ejemplo

Predicción:

```
mod1=lm(imdb ~ filmaffinity)
nuevo <- data.frame(filmaffinity=8)
predict(object=mod1, newdata=nuevo, interval="prediction", level=0.95)
```

```
       fit      lwr      upr
1 8.590406 8.261485 8.919327
```

Próxima sesión

- Tema 10: Análisis de componentes principales.

LA UNIVERSIDAD
EN INTERNET