UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Departament d'Organització d'Empreses

Modelos de regresión

Jose M Sallan

17 de julio de 2018

Contenidos

Regresión lineal

El análisis de regresión consiste en analizar las relaciones entre:

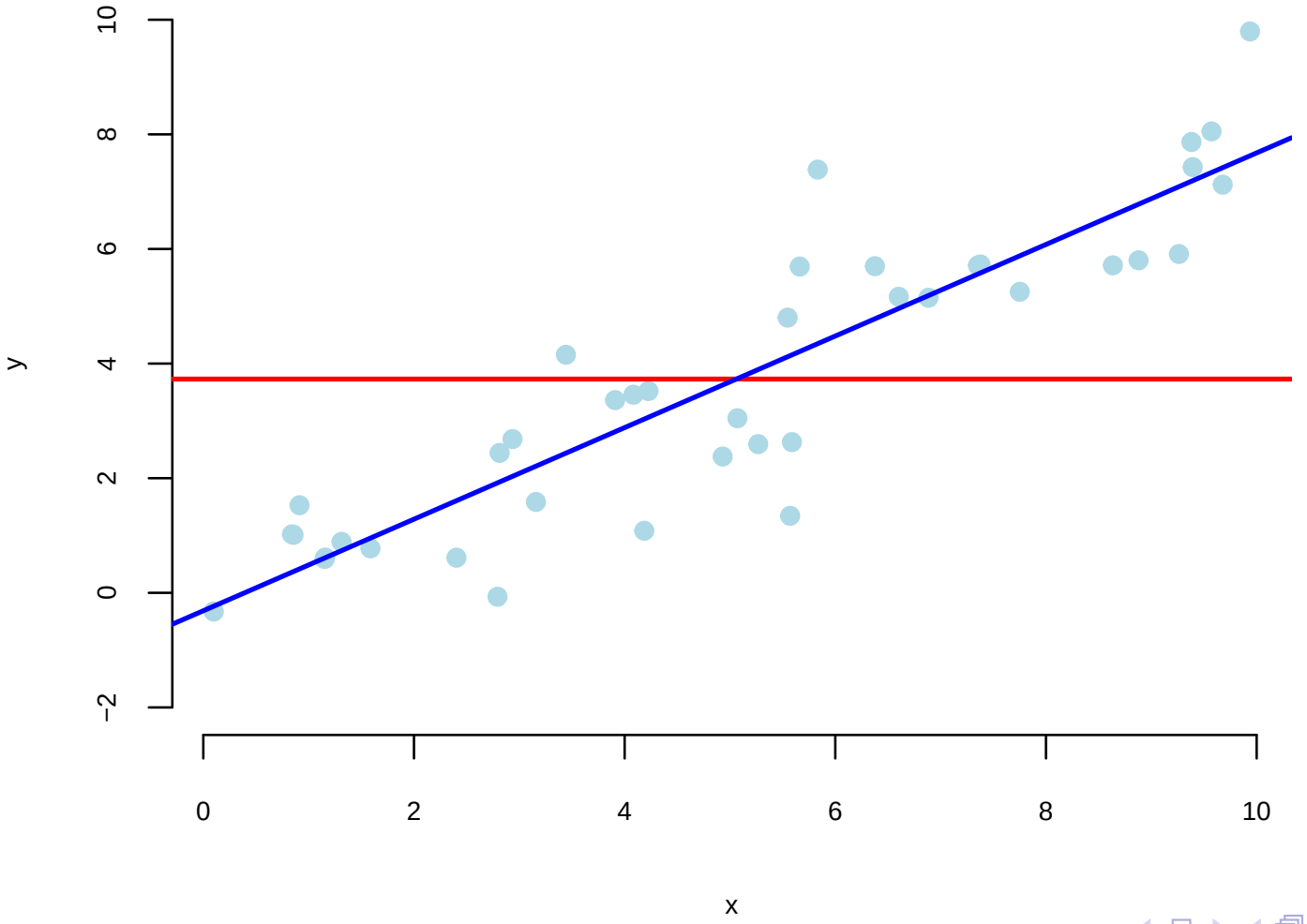
- **una variable dependiente (o endogéna, respuesta o criterio)** y
- un conjunto de p variables **independentes (exógenas, predictoras)** x_j

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \varepsilon_i \tag{1}$$

Un ejemplo univariante

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Departament d'Organització d'Empreses

Estimando con la media (rojo) o con una recta de regresión (azul):



Ejemplo multivariante: mtcars 1

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Departament d'Organització d'Empreses

`mtcars` incluye información sobre consumo de gasolina en millas por galón `mpg` y diez aspectos de diseño y desempeño para 32 automóviles (modelos de 1973-1974).

```
> head(mtcars)
```

	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

Ejemplo multivariante: mtcars 1

Resultados de la regresión lineal

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Departament d'Organització d'Empreses

Regresión de `mpg` frente a potencia `hp` peso `wt` y aceleración `qsec`

```
> mtcars01 <- lm(mpg ~ hp + wt + qsec, data=mtcars)
> summary(mtcars01)

Call:
lm(formula = mpg ~ hp + wt + qsec, data = mtcars)

Residuals:
    Min      1Q  Median      3Q     Max
-3.8591 -1.6418 -0.4636  1.1940  5.6092

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 27.61053    8.41993   3.279  0.00278 **
hp          -0.01782    0.01498  -1.190  0.24418
wt          -4.35880    0.75270  -5.791 3.22e-06 ***
qsec         0.51083    0.43922   1.163  0.25463
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.578 on 28 degrees of freedom
Multiple R-squared:  0.8348,        Adjusted R-squared:  0.8171
F-statistic: 47.15 on 3 and 28 DF,  p-value: 4.506e-11
```

Significación global del modelo

- **Coeficiente de determinación** R^2: fracción de la variablidad de y explicada por el modelo de regresión (modelos con R^2 cercano a 1 tienen buen ajuste).
- R^2 **ajustado**: coeficiente ajustado por el número de predictores p.
- **F-statistic**: test de si el modelo de regresión explica y mejor que la media $\bar{y}$.

Significación de los coeficientes de regresión

Nos interesa saber si hay relación entre y y las x_j, más que el valor del coeficiente.

- **Hipótesis nula** para cada variable: *el coeficiente de regresión es cero* (no hay relación entre x_j e y).
- Podemos descartar la hipótesis nula si el p-valor es lo bastante pequeño:

.	$p < 0{,}1$
*	$p < 0{,}05$
**	$p < 0{,}01$
***	$p < 0{,}001$

El p-valor es la probabilidad de obtener un valor igual o más extremo que el obtenido si la hipótesis nula es cierta.

Interpretación del modelo mtcars 1


```
> mtcars01 <- lm(mpg ~ hp + wt + qsec, data=mtcars)
> summary(mtcars01)

Call:
lm(formula = mpg ~ hp + wt + qsec, data = mtcars)

Residuals:
    Min      1Q  Median      3Q     Max
-3.8591 -1.6418 -0.4636  1.1940  5.6092

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 27.61053    8.41993   3.279  0.00278 **
hp          -0.01782    0.01498  -1.190  0.24418
wt          -4.35880    0.75270  -5.791 3.22e-06 ***
qsec         0.51083    0.43922   1.163  0.25463
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.578 on 28 degrees of freedom
Multiple R-squared:  0.8348,        Adjusted R-squared:  0.8171
F-statistic: 47.15 on 3 and 28 DF,  p-value: 4.506e-11
```

Interpretación del modelo mtcars 1

Ajuste global del modelo:

- El estadístico F muestra que la regresión explica mejor el modelo que la media.
- Coeficiente de determinación elevado $R^2_{aj.} = 0,82$.

Coeficientes de regresión:

- Expresan la relación entre la variable dependiente y cada una de las independientes.
- En este modelo, sólo son significativos el término independiente y el coeficiente de regresión de `wt`.
- Cuanto mayor es el peso `wt`, mayor es el consumo (menor es `mpg`).

Variables categóricas

En una regresión lineal podemos introducir como variables dependientes **variables categóricas**, que representan niveles de un determinado factor:

- género (masculino, femenino)
- sector indutrial (energía, farmacia, TIC, ...)

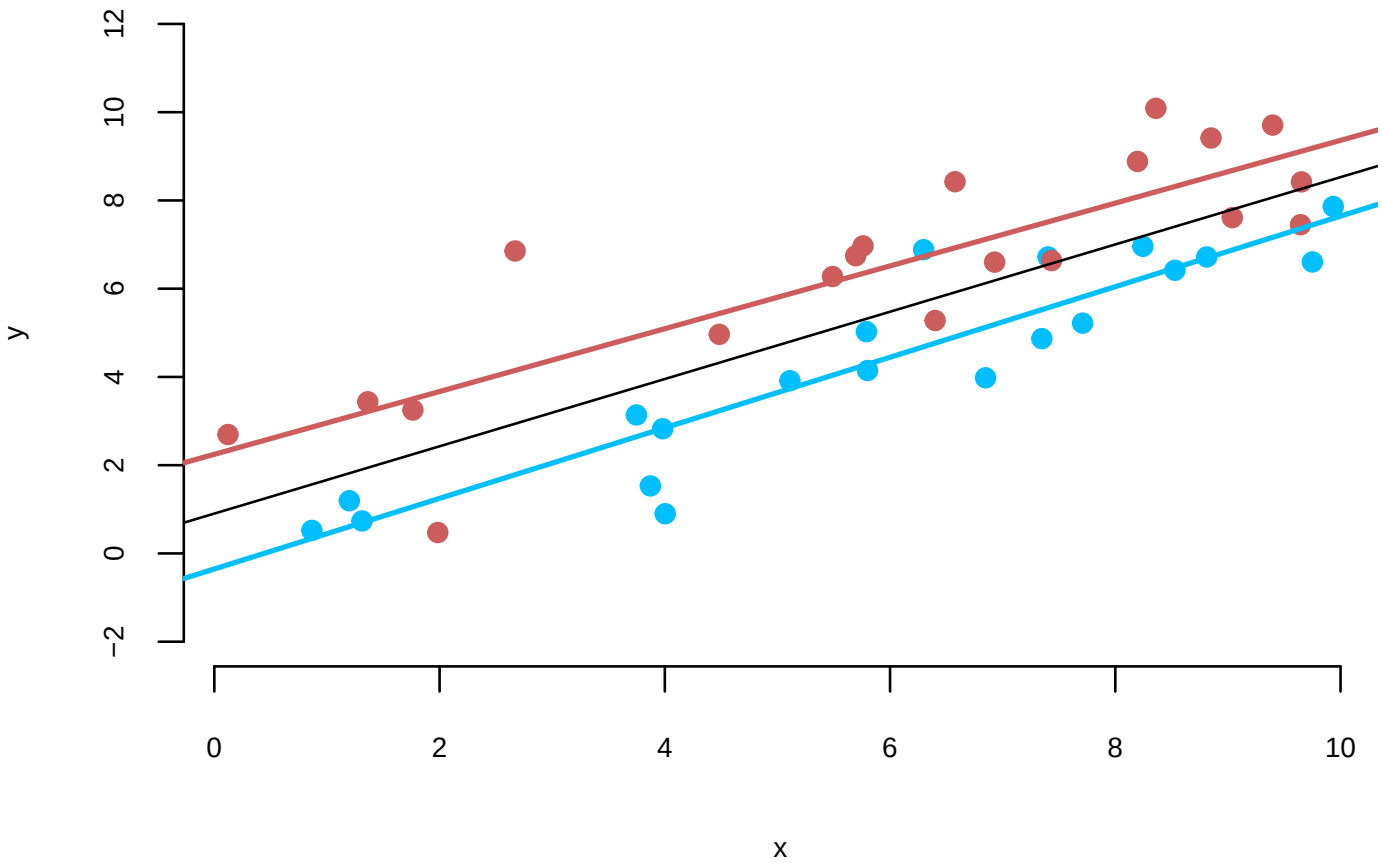
Una variable categórica de k niveles se puede representar usando $k-1$ variables **dummy** (binarias).

Sector	d_1	d_2	d_3
Energía (base level)	0	0	0
Farmacia	1	0	0
TIC	0	1	0
Construcción	0	0	1

Si la variable categórica se codifica como factor, **R** genera las variables dummy automáticamente. Los coeficientes de regresión de las variables dummy representan la diferencia de valor de la variable dependiente entre el nivel definido por la dummy y en nivel base.

Variable binaria

En este caso, y depende de x y del género de los individuos (azul: hombres, rojo: mujeres).



Variable binaria

¿Qué poder explicativo añade la variable binaria?

```
> summary(mod.dummy)

Call:
lm(formula = y.dummy ~ x.dummy)

Residuals:
    Min      1Q  Median      3Q     Max
-3.0556 -1.0599 -0.2609  1.2634  3.9147

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.90376    0.57665   1.567    0.125
x.dummy      0.76241    0.08804   8.660 1.59e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.573 on 38 degrees of freedom
Multiple R-squared:  0.6637,        Adjusted R-squared:  0.6549
F-statistic: 74.99 on 1 and 38 DF,  p-value: 1.587e-10
```

Variable binaria

¿Qué poder explicativo añade la variable binaria?

```
> summary(mod.dummy4)

Call:
lm(formula = y.dummy ~ x.dummy + dummy)

Residuals:
    Min      1Q  Median      3Q     Max
-3.0232 -0.6226  0.1100  0.6344  2.8411

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.07296    0.46384  -0.157    0.876
x.dummy      0.75183    0.06565  11.452 1.01e-13 ***
dummywomen   2.07853    0.37098   5.603 2.17e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.173 on 37 degrees of freedom
Multiple R-squared:  0.8181,        Adjusted R-squared:  0.8082
F-statistic: 83.18 on 2 and 37 DF,  p-value: 2.035e-14
```

Variable binaria

Comparando los dos modelos:

```
> anova(mod.dummy, mod.dummy4)

Analysis of Variance Table

Model 1: y.dummy ~ x.dummy
Model 2: y.dummy ~ x.dummy + dummy
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1     38 94.046
2     37 50.879  1    43.167 31.392 2.167e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

El modelo con la variable binaria añade poder explicativo.

Variable binaria


```
> library(stargazer)
> stargazer(mod.dummy, mod.dummy4, type = "text")

===============================================================
                              Dependent variable:
                    ---------------------------------------
                                     y.dummy
                              (1)                  (2)
---------------------------------------------------------------
x.dummy                     0.762***             0.752***
                            (0.088)              (0.066)

dummywomen                                       2.079***
                                                 (0.371)

Constant                     0.904               -0.073
                            (0.577)              (0.464)

---------------------------------------------------------------
Observations                  40                   40
R2                           0.664                0.818
Adjusted R2                  0.655                0.808
Residual Std. Error    1.573 (df = 38)      1.173 (df = 37)
F Statistic         74.994*** (df = 1; 38) 83.182*** (df = 2; 37)
===============================================================
Note:                                *p<0.1; **p<0.05; ***p<0.01
```

Ejemplo: mtcars 2

Definimos un modelo para evaluar la influencia del número de marchas `gears`:

```
> levels(as.factor(mtcars$gear))

[1] "3" "4" "5"

> mtcars02 <- lm(mpg ~ wt + hp + factor(gear), data=mtcars)
> coef(summary(mtcars02))

                Estimate Std. Error    t value     Pr(>|t|)
(Intercept)   34.87245123 2.58015801 13.5156262 1.558098e-13
wt            -3.23852439 0.87781636 -3.6892960 1.000770e-03
hp            -0.03497069 0.01260201 -2.7750090 9.897557e-03
factor(gear)4  1.26489784 1.34083819  0.9433635 3.538604e-01
factor(gear)5  1.87355541 1.86661986  1.0037156 3.244269e-01
```

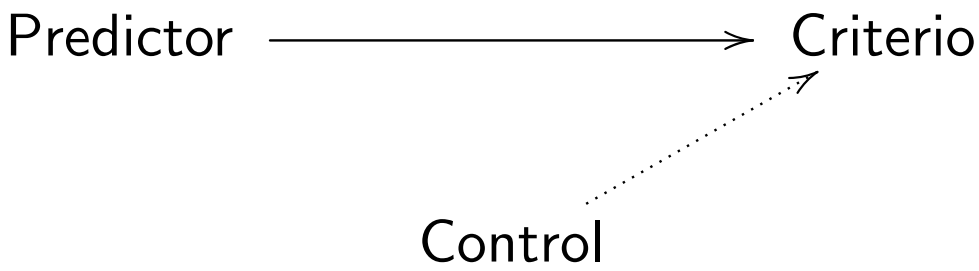
En este modelo, no se aprecia que `gears` influya en el consumo de combustible.

Regresión jerárquica

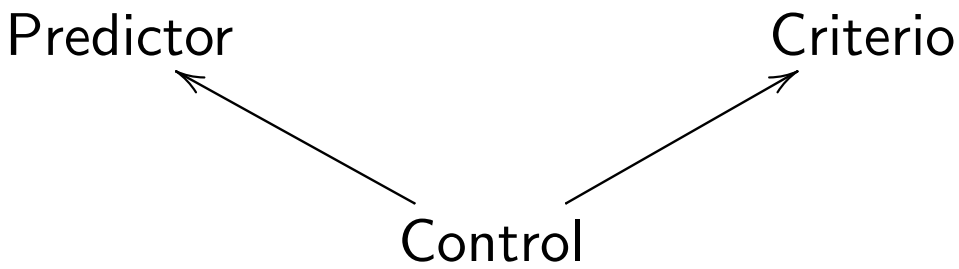
Cuando estudiamos la relación entre una variable **predictora** y una variable **criterio**, queremos estar seguros de controlar fuentes de variabilidad común (variables de **control**).

Correlación espúrea

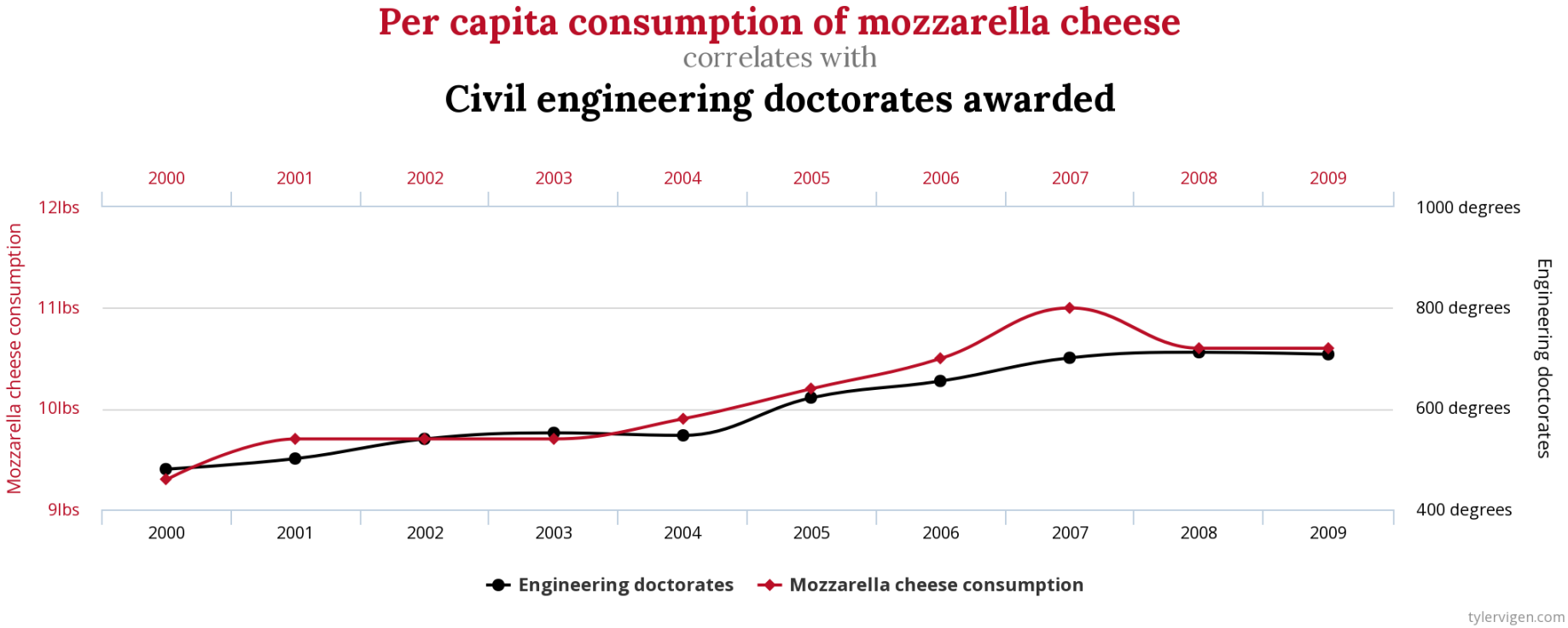
Relación no espúrea:



Relación espúrea:



Correlación espúrea



Regresión jerárquica

En la regresión jerárquica las variables se entran en dos pasos:

- Primer modelo: regresión de la variable predictora con las **variables de control**.
- Segundo modelo: regresión de la variable predictora con las **variables de control y de criterio**.

Regresión jerárquica

Para tener seguridad de que no hay relación espúrea:

- El segundo modelo ha de tener más poder explicativo que el primero.
- Los coeficientes de regresión de las variables de control no han de experimentar cambios significativos cuando se introducen las variables predictoras.

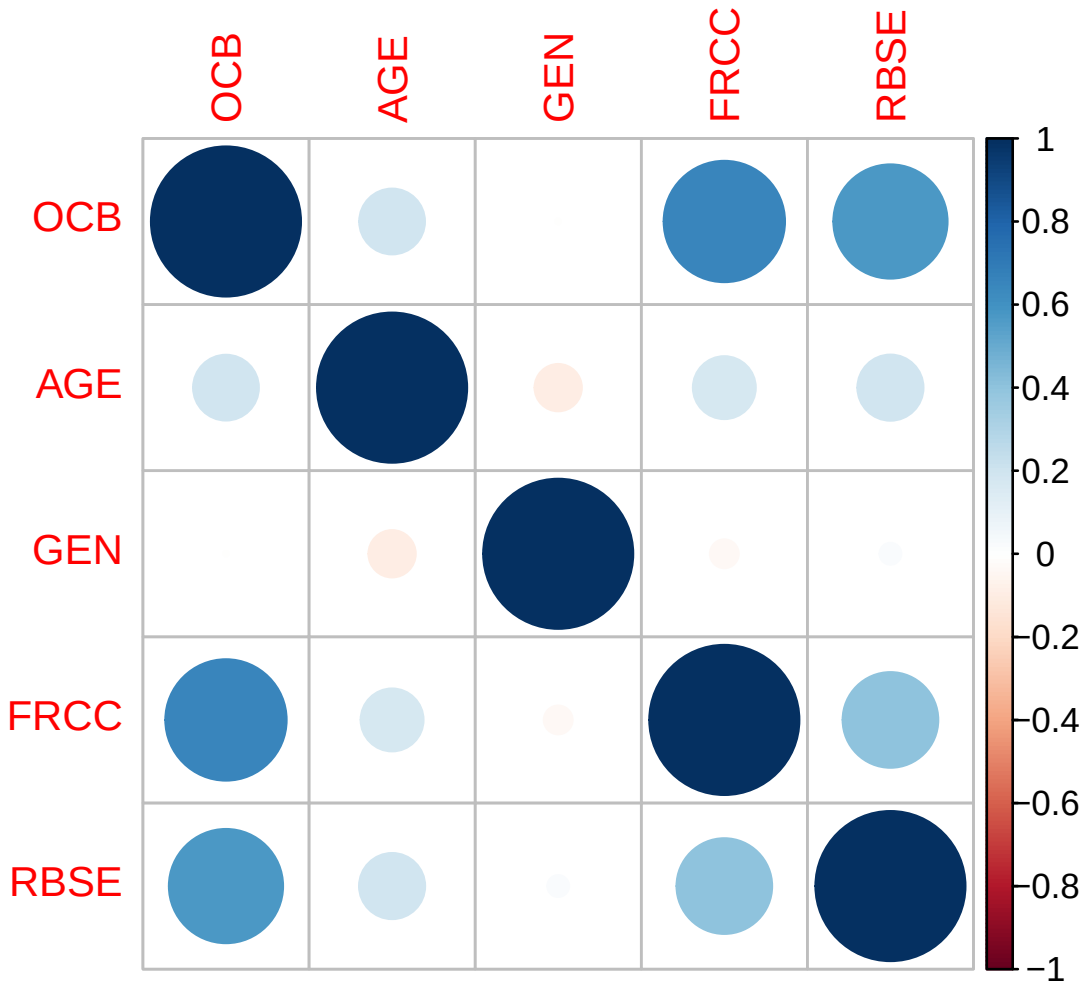
Ejemplo OCB

Datos con 602 observaciones:

- Criterio: comportamiento ciudadano (OCB).
- Control: edad (AGE) y género (GEN).
- Predictor: Responsabilidad de cambio constructivo (FRCC), eficacia percibida en el rol (RBSE).

Ejemplo OCB

Correlograma de variables

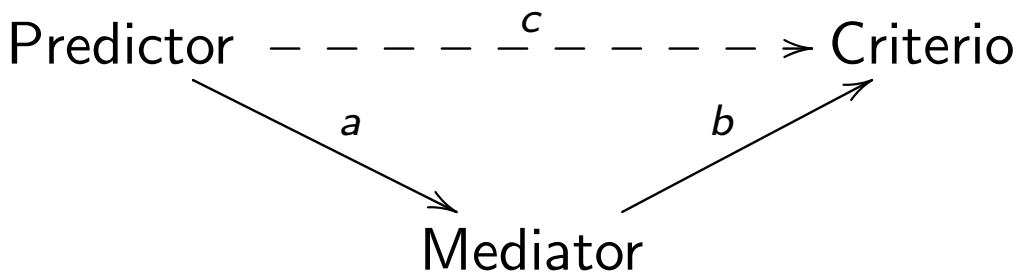


Ejemplo OCB

	Dependent variable:	
	OCB	
	(1)	(2)
AGE	0.015***	0.003
	(0.003)	(0.002)
factor(GEN)1	0.028	0.020
	(0.063)	(0.044)
FRCC		0.128***
		(0.008)
RBSE		0.390***
		(0.032)
Constant	3.151***	0.287*
	(0.141)	(0.151)
Observations	602	602
R2	0.036	0.544
Adjusted R2	0.033	0.541
Residual Std. Error	0.772 (df = 599)	0.531 (df = 597)
F Statistic	11.331*** (df = 2; 599)	178.267*** (df = 4; 597)
Note:		*p<0.1; **p<0.05; ***p<0.01

Mediación

Un **mediador** es una variable que actúa como intermediaria entre la relación entre **predictor** y **criterio**:



Mediation

Criterios de Baron y Kenny

Existe mediación si:

- variaciones de nivel de la variable independiente explican variaciones de la candidata a mediadora (camino a).
- variaciones en el mediador explican variaciones de la variable dependiente (camino b).
- cuando los caminos a y b están controlados, deja de ser significativa una relación entre variable dependiente e independiente (camino c).

Simulando una relación mediada


```
> set.seed(3333)
> pred <- rnorm(100, 2, 1)
> med <- 3 + 2*pred + rnorm(100, sd=0.3)
> cri <- 2 + med + rnorm(100, sd=0.2)
> bk01 <- lm(med ~ pred)
> bk02 <- lm(cri ~ pred)
> bk03 <- lm(cri ~ pred + med)
```

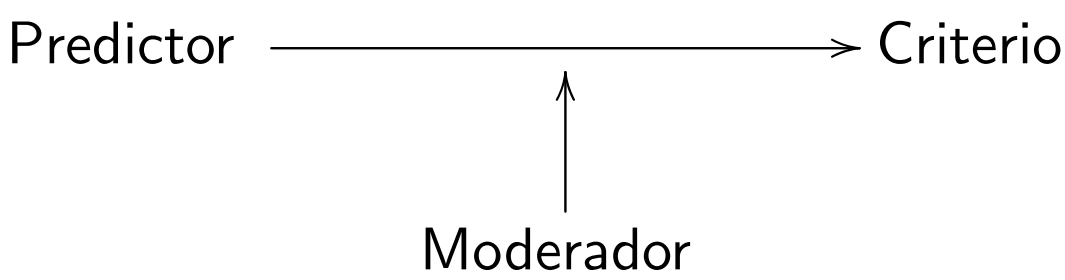
Simulando una relación mediada

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Departament d'Organització d'Empreses

```
> coef(summary(bk01))

            Estimate Std. Error  t value     Pr(>|t|)
(Intercept) 3.018608 0.07659320 39.41092 6.544253e-62
pred        2.006344 0.03358632 59.73696 6.803052e-79

> coef(summary(bk02))

            Estimate Std. Error  t value     Pr(>|t|)
(Intercept) 5.010529 0.09142991 54.80186 2.493842e-75
pred        2.006993 0.04009225 50.05938 1.310783e-71

> coef(summary(bk03))

              Estimate Std. Error    t value     Pr(>|t|)
(Intercept) 2.10701046 0.22339750  9.4316652 2.264074e-15
pred        0.07714424 0.14597351  0.5284811 5.983726e-01
med         0.96187320 0.07177705 13.4008470 8.487747e-24
```

Moderación

Un **moderador** es una variable que afecta la dirección o intensidad de la relación entre las variables **predictora** y **criterio**.



Hay moderación cuando el coeficiente del **término de interacción** (producto de predictor y moderador) es significativo.

Ejemplo: mtcars 3

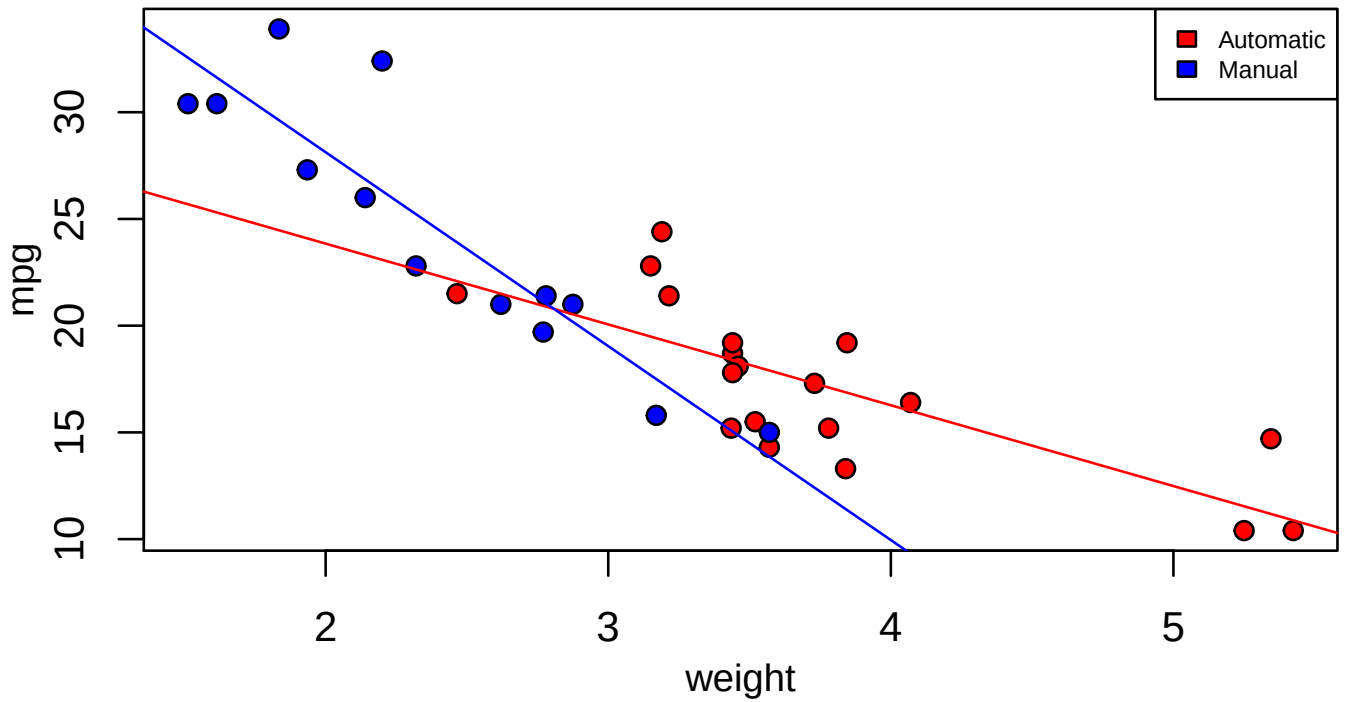
UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Departament d'Organització d'Empreses

¿La relación entre consumo de combustible `mpg` y peso `wt` depende del tipo de transimisión `am` (0 automática, 1 manual)?

```
> summary(lm(mpg ~ am*wt, mtcars))

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

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 ***
am           14.8784     4.2640   3.489  0.00162 **
wt           -3.7859     0.7856  -4.819 4.55e-05 ***
am:wt        -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
```

Ejemplo: mtcars 3

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Departament d'Organització d'Empreses



Las millas por galón decrecen más lentamente con el peso cuando se usa transmisión automática.

Residuos

Los **residuos** son la diferencia entre el valor real y la predicción del modelo: $e_i = y_i - \hat{y}_i$.

El método más habitual de obtener los estimadores del modelo de regresión es minimizando:

$$\sum (y_i - \hat{y}_i)^2$$

Mínimos cuadrados ordinarios *(ordinary least squares OLS)*.

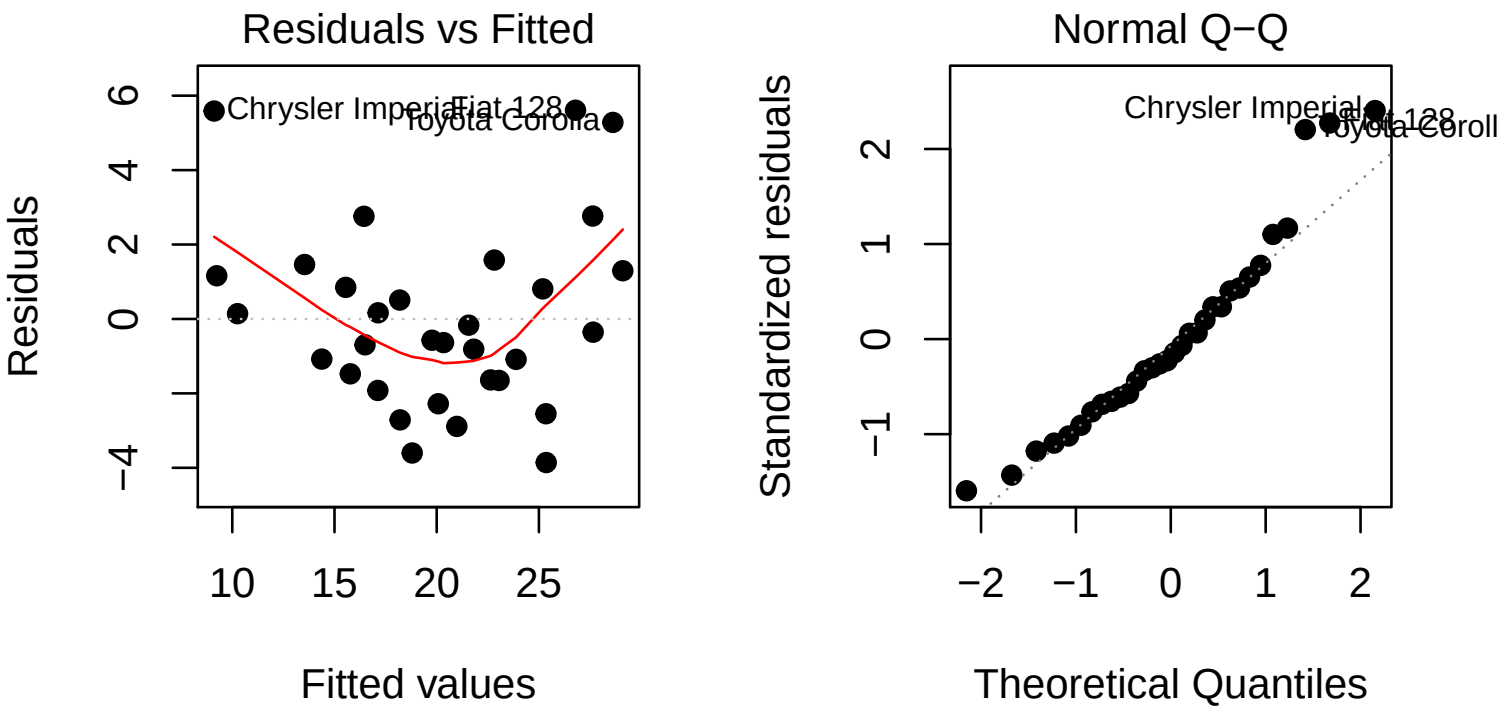
Diagnóstico de residuos

Para que los estimadores obtenidos por mínimos cuadrados sean insesgados y de máxima verosimilitud los residuos:

- han de tener media a cero y varianza constante para $\hat{y}_i$.
- han de seguir una distribución normal.
- han de ser independientes del tiempo (si aplica).

Análisis de residuos en mtcars 1

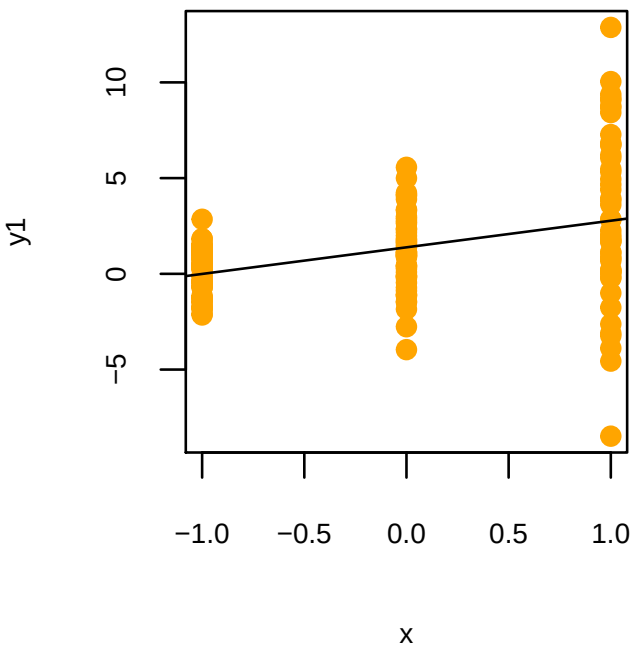
```
> par(mfrow=c(1,2), pty="s")
> plot(mtcars01, which=1, pch=19)
> plot(mtcars01, which=2, pch=19)
```



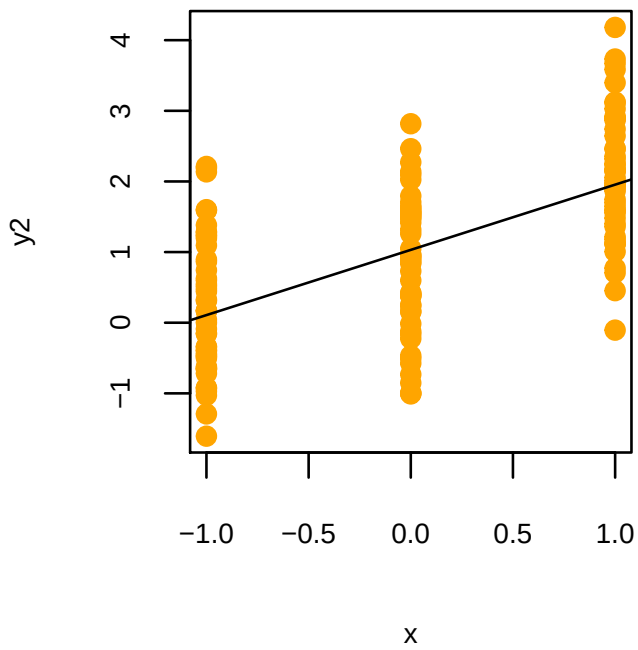
Varianza de los residuos

Los residuos son **homocedásticos** si su varianza es independiente de $\hat{y}_i$, y **heterocedásticos** en caso contrario.

heterocedástico



homocedástico



Varianza de los residuos

Comparativa de los coeficientes de x de los modelos 1 (heterocedástico) y 2 (homocedástico). El coeficiente de regresión de población es 1.

```
> coef(summary(mod.y1))
```

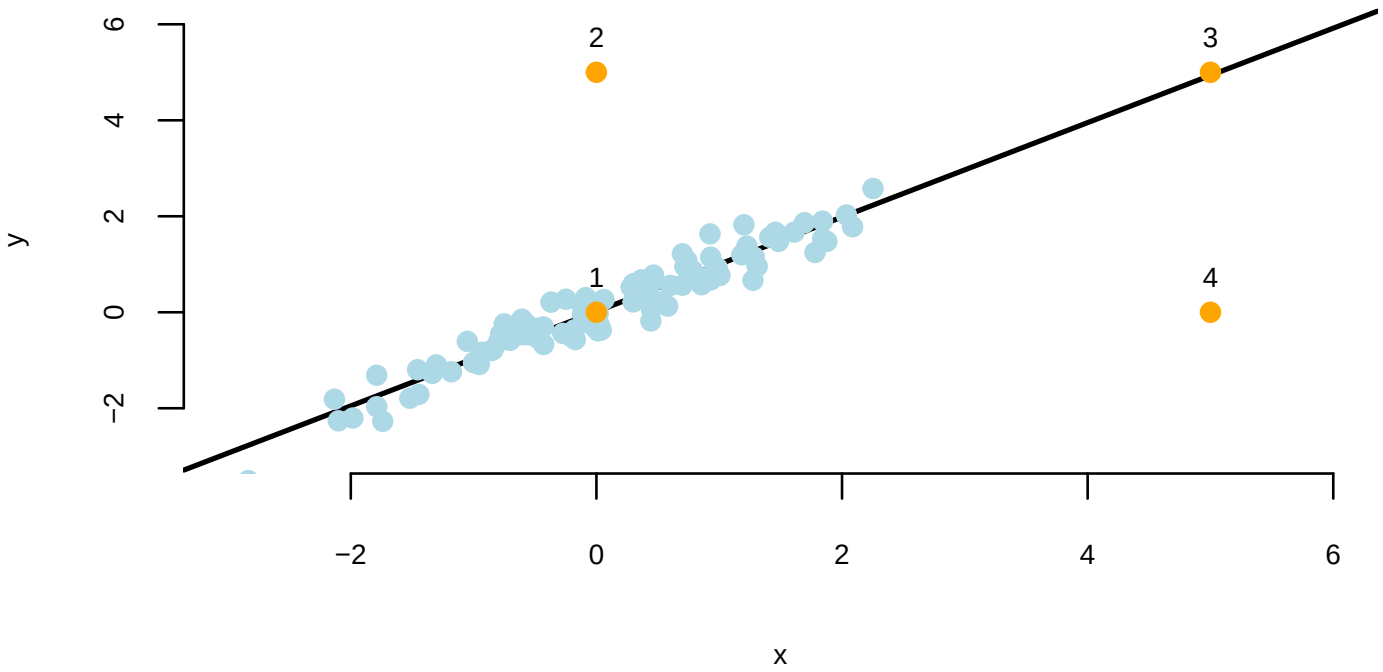
```
            Estimate Std. Error  t value     Pr(>|t|)
(Intercept) 1.385123  0.2333131 5.936757 1.984273e-08
x           1.390449  0.2857490 4.865981 2.885719e-06
```

```
> coef(summary(mod.y2))
```

```
             Estimate Std. Error   t value     Pr(>|t|)
(Intercept) 1.0312575 0.07809141 13.205774 8.432673e-27
x           0.9233003 0.09564205  9.653707 2.100486e-17
```

Para evitar la heterocedasticidad pueden hacerse transformaciones de las variables (logaritmo, raíz cuadrada).

Valores atípicos



El **punto 1** no es atípico **(outlier)**, el resto sí.
El **punto 2** es de bajo apalancamiento y baja influencia.
El **punto 3** es de alto apalancamiento y baja influencia.
El **punto 4** es de alto apalancamiento y alta influencia.

Diagnósticos de influencia

Puede accederse a una colección de los diagnósticos de influencia disponibles en **R** escribiendo en la consola `?influence.measures`

- `cooks.distance` La distancia de Cook de la observación i es la variación estandardizada de las predicciones $\hat{y}_i$ cuando se elimina la observación i.
- `dfbeta` / `dfbetas` es el cambio en los coeficientes de regresión no estandarizados y estandarizados cuando se elimina la observación i.

Distancias de Cook en mtcars 1


```
> plot(mtcars01, which=4)
```

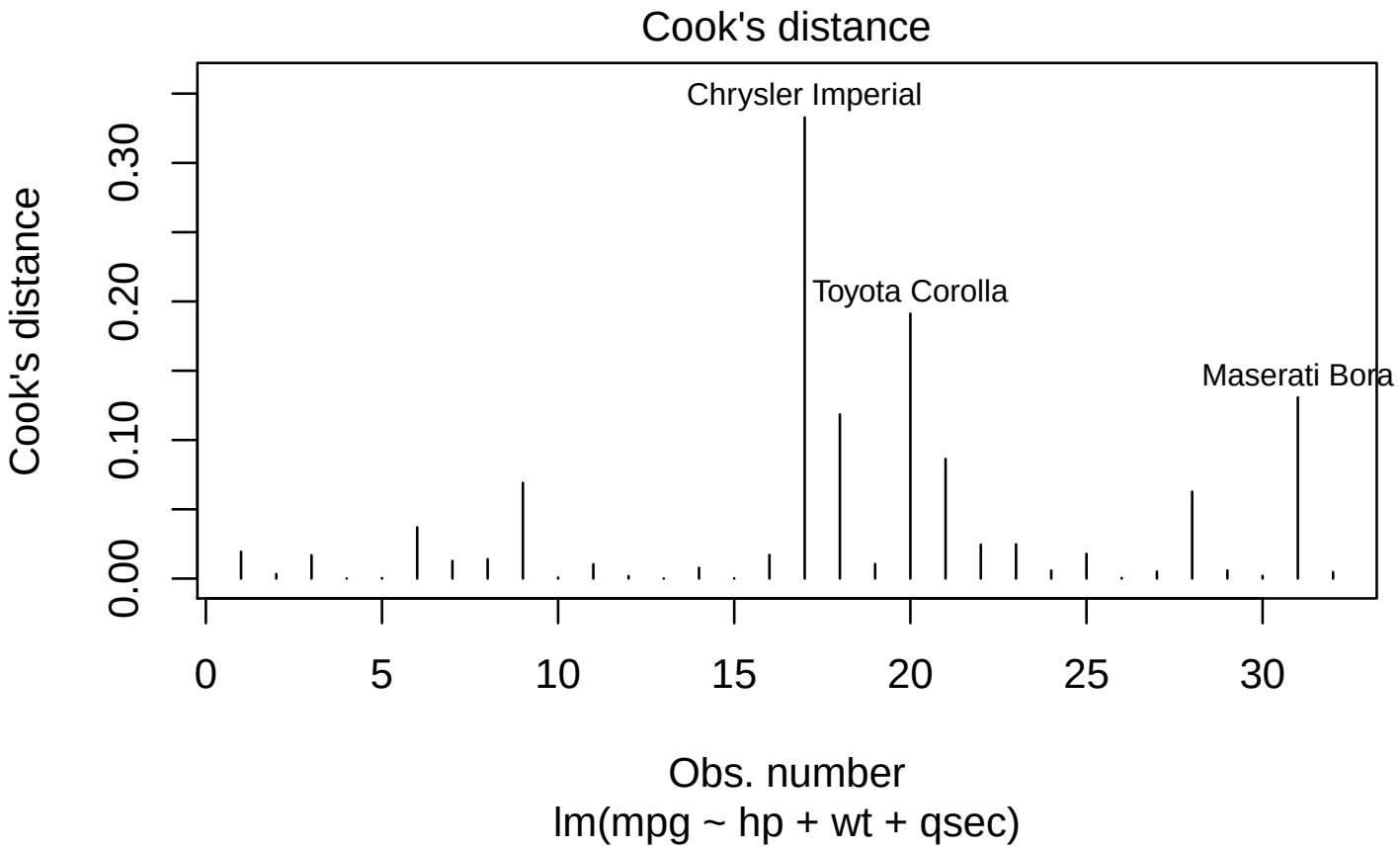
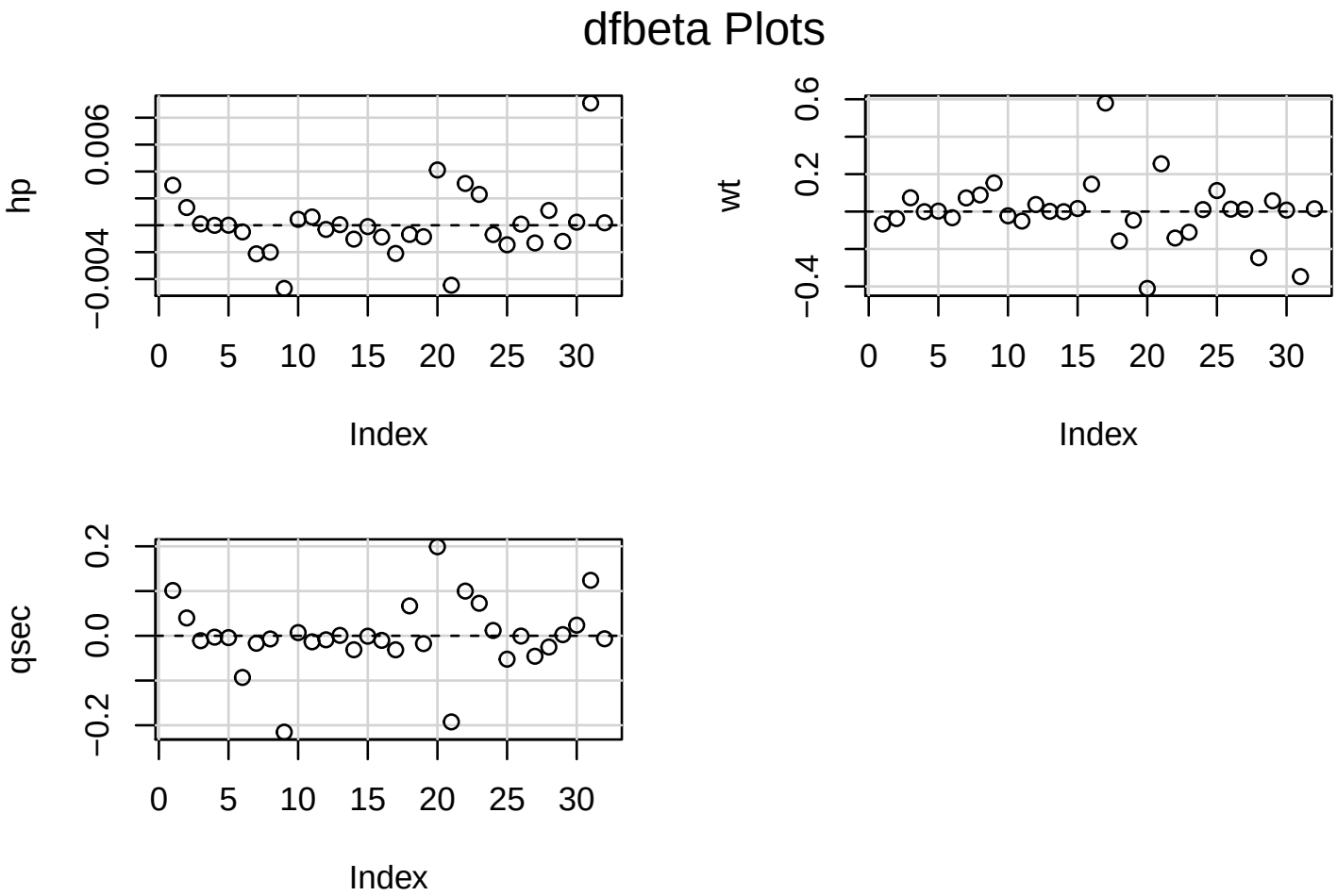


Gráfico dfbeta en mtcars 1


```
> library(car)
> dfbetaPlots(mtcars01)
```



Multicolinealidad

Hay **multicolinealidad** si algunas de las variables independientes x_j están muy correlacionadas entre sí.

Algunos efectos de añadir variables independientes correlacionadas con las ya añadidas:

- El coeficiente de determinación ajustado $R^2_{aj.}$ disminuye.
- La varianza de los coeficientes de regresión aumenta.
- Uno o varios coeficientes de regresión dejan de ser significativos.

Diagnósticos de multicolinealidad

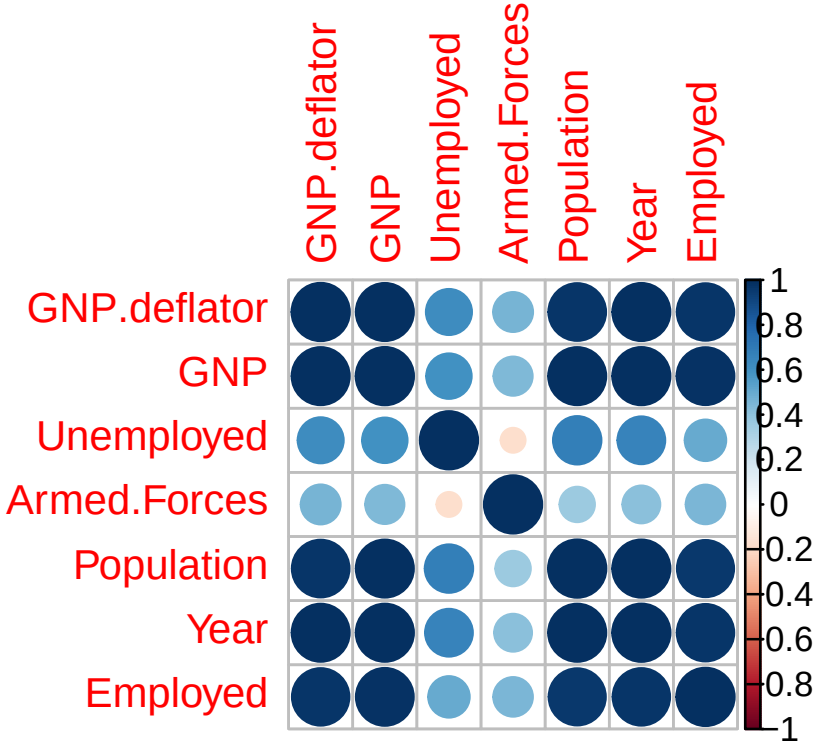
Se obtienen diagnósticos de multicolinealidad (MC) mediante R_j^2, el coeficiente de determinación de la regresión de x_j respecto del resto de variables independientes.

- La **tolerancia** de x_j es $1 - R_j^2$ (valores altos si no hay multicolinealidad).
- El **factor de inflación de la varianza** es el inverso de la tolerancia (bajo si no hay multicolinealidad).

Ejemplo: Longley

`longley` es un conjunto de datos macroeconómicos:

```
> library(corrplot)
> corrplot(cor(longley), method="circle")
```



Cálculo de los FIV en longley

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Departament d'Organització d'Empreses

Evaluamos dos modelos:

```
> longley01 <- lm(Employed ~ ., data=longley)
> library(car)
> vif(longley01)
```

```
GNP.deflator          GNP   Unemployed Armed.Forces   Population         Year
   135.53244   1788.51348     33.61889      3.58893    399.15102   758.98060
```

```
> longley02 <- lm(Employed ~ GNP + Armed.Forces, data=longley)
> vif(longley02)
```

```
        GNP Armed.Forces
   1.248916     1.248916
```

Comparativa de modelos en Longley

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Departament d'Organització d'Empreses

```
> coef(summary(longley01))
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-3.482259e+03	8.904204e+02	-3.9108029	0.0035604037
GNP.deflator	1.506187e-02	8.491493e-02	0.1773760	0.8631408328
GNP	-3.581918e-02	3.349101e-02	-1.0695163	0.3126810611
Unemployed	-2.020230e-02	4.883997e-03	-4.1364274	0.0025350917
Armed.Forces	-1.033227e-02	2.142742e-03	-4.8219853	0.0009443668
Population	-5.110411e-02	2.260732e-01	-0.2260511	0.8262117958
Year	1.829151e+00	4.554785e-01	4.0158898	0.0030368033

```
> coef(summary(longley02))
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	51.683468731	0.804340962	64.2556716	1.164460e-17
GNP	0.034393472	0.001965575	17.4979215	2.036369e-10
Armed.Forces	0.001147955	0.002807338	0.4089122	6.892606e-01

Conclusiones

Ejemplo de aplicación de regresión lineal:

- Wassmer, U., & Meschi, P.-X. (2011). The effect of code-sharing alliance formations and terminations on firm value: The role of co-specialization and scope extension. *Journal of Air Transport Management*, 17(5), 305–308.

Resultados Wassmer et al. (2011)

Table 4
Results for alliance terminations. [a, b]

Variables	Hypothesizedsign	Model 1 (controls only)	Model 2	Model 3
Alliance co-specialization (A) [c]	(−)		−0.491*	−0.457*
Alliance extension (B)	(−)		−0.316	−0.230
CAR(formation)$_{-1,+1}$ (C)			0.433	0.395†
(A)×(C) [d]	(−)			0.033**
Focal firm size [c]		3.355	3.595	1.988
Focal firm debt		0.320†	0.357*	0.371†
Focal firm ROS		8.355	5.085	2.315
Alliance experience [c]		−1.114	−1.839	−1.234
Focal firm internationalization		5.275†	5.571†	4.161*
Relative partner size		0.055*	0.059*	0.042†
Relative market size		0.002	0.001	0.001
Cultural distance		1.182	1.030	0.431
Alliance duration		0.236	0.280	0.033
R^2		0.150*	0.196**	0.385**
VIF test		From 1.13 to 1.56	From 1.16 to 1.58	From 1.16 to 1.58
Breusch–Pagan test for heteroskedasticity		26.14***	48.99***	6.61*

$^{\dagger}p < 0.1$*$p < 0.05$**$p < 0.01$***$p < 0.001$.
[a] Model intercepts are not reported.
[b] As some sample firms engaged in more than one alliance, we adjusted the standard errors of the regression coefficients using the robust estimates of the standard errors, clustered by firms.
[c] Logarithmic transformation.
[d] Variables are centered for interaction terms.

Interacción en Wassmer et al. (2011)

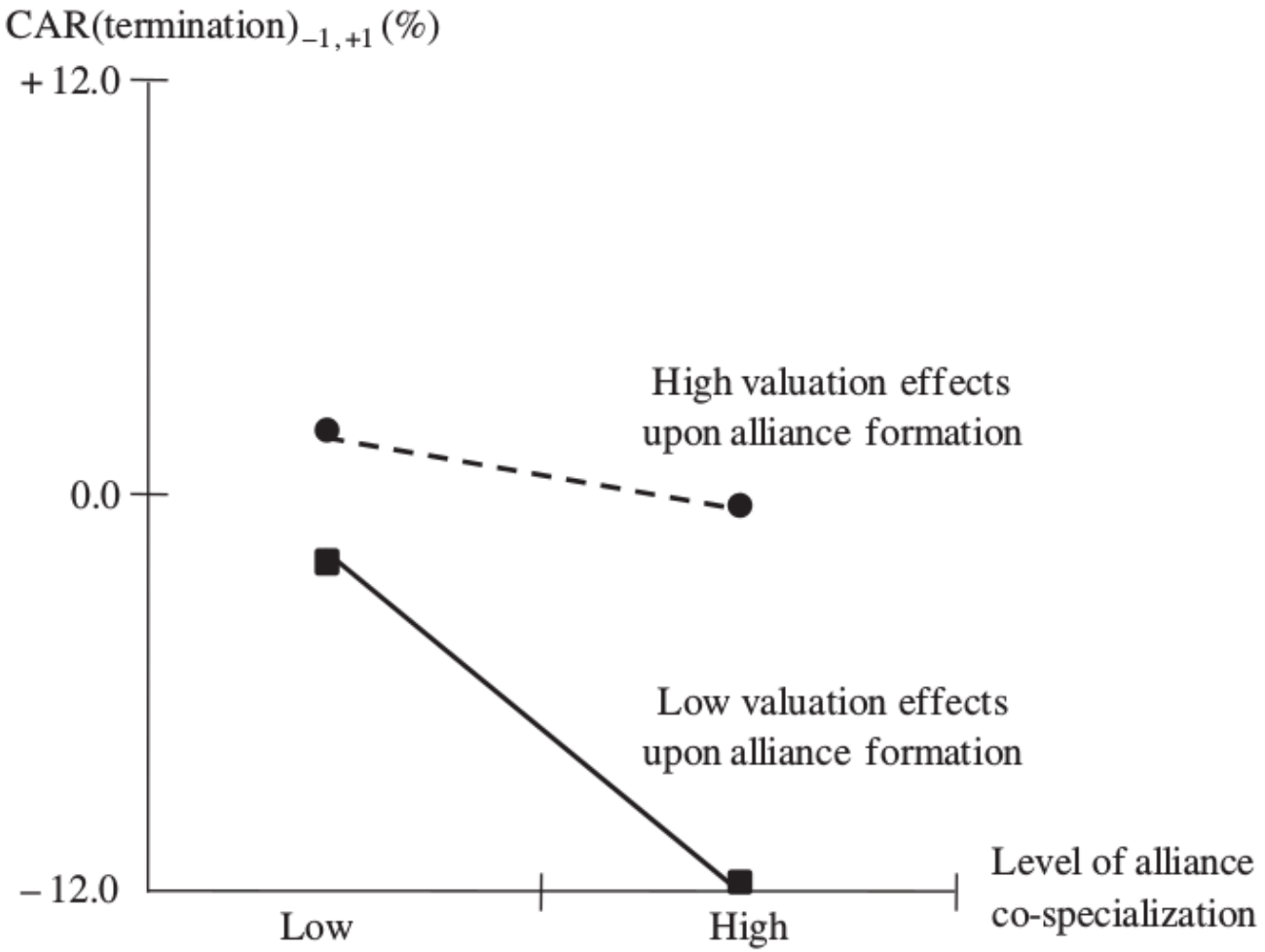


Fig. 1. Interaction effect of CAR(formation)$_{-1,+1}$ and the level of alliance co-specialization.

Conclusiones

- Los modelos de regresión lineal permiten evaluar la existencia de relaciones entre una variable **dependiente** y variables **independientes**
- Pueden usarse **variables categóricas** como variables independientes. Si son variables dependientes deben usarse otros modelos (regresión logística)
- Puede evaluarse el impacto de **variables de control** usando **regresión jerárquica**
- También pueden estimarse modelos de **mediación** y **moderación**

Información y contacto

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Departament d'Organització d'Empreses

Correo: jose.maria.sallan@upc.edu

Blog: `http://josemsallan.blogspot.com/`

Documentación: `https://github.com/jmsallan/quantitative`