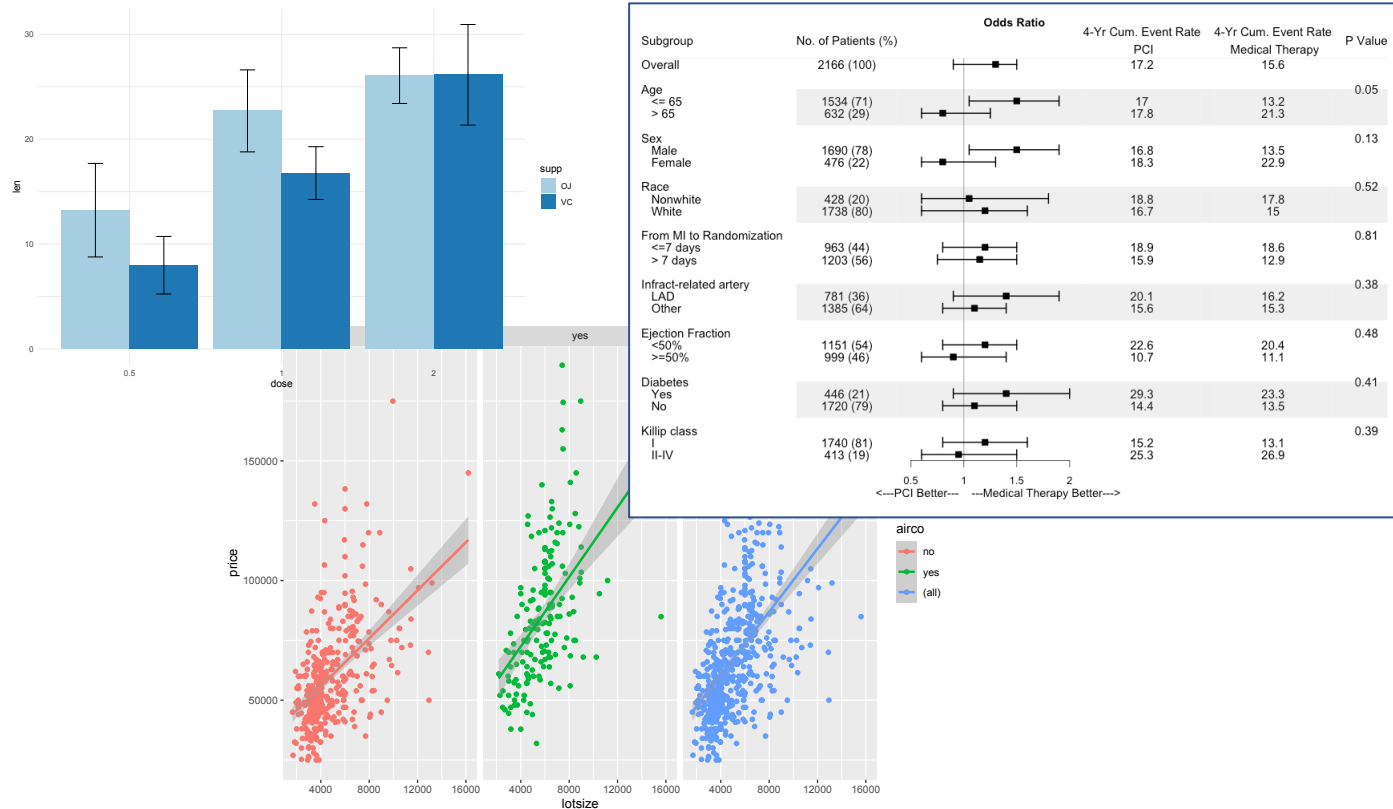


# Análisis Estadístico con



# 7. Modelos lineales, regresión múltiple

## Índice

- 7.1. Introducción
- 7.2. Búsqueda del mejor modelo
- 7.3. Regresión polinómica
- 7.4. Confusión, Interacción(modificador), mediación
- 7.5. Diagnóstico de los modelos

# 7. Modelos lineales, regresión múltiple

## 7.1. Introducción

- $Y = a + b_1X_1 + b_2X_2 + \dots + b_kX_k + E$
- $X_k$ =variables explicativas, predictores, variables independientes ...
- $Y$ = outcome, variable dependiente, ...
- Variable de interés (variable explicativa de estudio)

'age' a numeric vector, age in years.

'sex' a numeric vector code, 0: male, 1:female.

'height' a numeric vector, height (cm).

'weight' a numeric vector, weight (kg).

'bmp' a numeric vector, body mass (% of normal).

'fev1' a numeric vector, forced expiratory volume.

'rv' a numeric vector, residual volume.

'frc' a numeric vector, functional residual capacity.

'tlc' a numeric vector, total lung capacity.

'pemax' a numeric vector, maximum expiratory pressure.

	age	sex	height	weight	bmp	fev1	rv	frc	tlc	pemax
1	7	0	109	13.1	68	32	258	183	137	95
2	7	1	112	12.9	65	19	449	245	134	85
3	8	0	124	14.1	64	22	441	268	147	100
4	8	1	125	16.2	67	41	234	146	124	85
5	8	0	127	21.5	93	52	202	131	104	95
6	9	0	130	17.5	68	44	308	155	118	80
7	11	1	139	30.7	89	28	305	179	119	65
8	12	1	150	28.4	69	18	369	198	103	110
9	12	0	146	25.1	67	24	312	194	128	70
10	13	1	155	31.5	68	23	413	225	136	95
11	13	0	156	39.9	89	39	206	142	95	110
12	14	1	153	42.1	90	26	253	191	121	90
13	14	0	160	45.6	93	45	174	139	108	100
14	15	1	158	51.2	93	45	158	124	90	80
15	16	1	160	35.9	66	31	302	133	101	134
16	17	1	153	34.8	70	29	204	118	120	134
17	17	0	174	44.7	70	49	187	104	103	165
18	17	1	176	60.1	92	29	188	129	130	120
19	17	0	171	42.6	69	38	172	130	103	130
20	19	1	156	37.2	72	21	216	119	81	85
21	19	0	174	54.6	86	37	184	118	101	85
22	20	0	178	64.0	86	34	225	148	135	160
23	23	0	180	73.8	97	57	171	108	98	165
24	23	0	175	51.1	71	33	224	131	113	95
25	23	0	179	71.5	95	52	225	127	101	195

# 7. Modelos lineales, regresión múltiple

## 7.1. Introducción

OUTCOME: maximum expiratory pressure

`lm(pemax~age+sex+height+weight+bmp+fev1+rv+frc+tlc)`

`summary(lm(pemax~age+sex+height+weight+bmp+fev1+rv+frc+tlc))`

	age	sex	height	weight	bmp	fev1	rv	frc	tlc	pemax
1	7	0	109	13.1	68	32	258	183	137	95
2	7	1	112	12.9	65	19	449	245	134	85
3	8	0	124	14.1	64	22	441	268	147	100
4	8	1	125	16.2	67	41	234	146	124	85
5	8	0	127	21.5	93	52	202	131	104	95
6	9	0	130	17.5	68	44	308	155	118	80
7	11	1	139	30.7	89	28	305	179	119	65
8	12	1	150	28.4	69	18	369	198	103	110
9	12	0	146	25.1	67	24	312	194	128	70
10	13	1	155	31.5	68	23	413	225	136	95
11	13	0	156	39.9	89	39	206	142	95	110
12	14	1	153	42.1	90	26	253	191	121	90
13	14	0	160	45.6	93	45	174	139	108	100
14	15	1	158	51.2	93	45	158	124	90	80
15	16	1	160	35.9	66	31	302	133	101	134
16	17	1	153	34.8	70	29	204	118	120	134
17	17	0	174	44.7	70	49	187	104	103	165
18	17	1	176	60.1	92	29	188	129	130	120
19	17	0	171	42.6	69	38	172	130	103	130
20	19	1	156	37.2	72	21	216	119	81	85
21	19	0	174	54.6	86	37	184	118	101	85
22	20	0	178	64.0	86	34	225	148	135	160
23	23	0	180	73.8	97	57	171	108	98	165
24	23	0	175	51.1	71	33	224	131	113	95
25	23	0	179	71.5	95	52	225	127	101	195

```
Call:
lm(formula = pemax ~ age + sex + height + weight + bmp + fev1 +
    rv + frc + tlc)

Residuals:
    Min       1Q   Median       3Q      Max
-37.338 -11.532   1.081  13.386  33.405

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  176.0582   225.8912   0.779   0.448
age          -2.5420    4.8017  -0.529   0.604
sex          -3.7368   15.4598  -0.242   0.812
height       -0.4463    0.9034  -0.494   0.628
weight        2.9928    2.0080   1.490   0.157
bmp          -1.7449    1.1552  -1.510   0.152
fev1          1.0807    1.0809   1.000   0.333
rv           0.1970    0.1962   1.004   0.331
frc          -0.3084    0.4924  -0.626   0.540
tlc           0.1886    0.4997   0.377   0.711

Residual standard error: 25.47 on 15 degrees of freedom
Multiple R-squared:  0.6373,    Adjusted R-squared:  0.4197
F-statistic: 2.929 on 9 and 15 DF,  p-value: 0.03195
```

# 7. Modelos lineales, regresión múltiple

## 7.1. Introducción

```

> anova(lm(pemax~age+sex+height+weight+bmp+fev1+rv+frc+tlc))
Analysis of Variance Table

Response: pemax
      Df Sum Sq Mean Sq F value    Pr(>F)
age     1 10098.5  10098.5   15.5661 0.001296 **
sex     1   955.4    955.4    1.4727 0.243680
height  1   155.0    155.0    0.2389 0.632089
weight  1   632.3    632.3    0.9747 0.339170
bmp     1  2862.2   2862.2    4.4119 0.053010 .
fev1    1  1549.1   1549.1    2.3878 0.143120
rv      1   561.9    561.9    0.8662 0.366757
frc     1   194.6    194.6    0.2999 0.592007
tlc     1    92.4     92.4    0.1424 0.711160
Residuals 15  9731.2   648.7

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

- t.test= dicen algo sobre que ocurre si quitas una variable y dejas las otras
- F tests del ANOVA se hacen quitando variables desde el final hacia arriba

# 7. Modelos lineales, regresión múltiple

## 7.1. Introducción

```

> summary(lm(pemax~age))

Call:
lm(formula = pemax ~ age)

Residuals:
    Min       1Q   Median       3Q      Max
-48.666 -17.174   6.209  16.209  51.334

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   50.408     16.657   3.026  0.00601 **
age           4.055       1.088   3.726  0.00111 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 26.97 on 23 degrees of freedom
Multiple R-squared:  0.3764,    Adjusted R-squared:  0.3492
F-statistic: 13.88 on 1 and 23 DF,  p-value: 0.001109

> anova(lm(pemax~age))

Analysis of Variance Table

Response: pemax
      Df Sum Sq Mean Sq F value    Pr(>F)
age     1  10098  10098.5    13.88 0.001109 **
Residuals 23  16734    727.6
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

# 7. Modelos lineales, regresión múltiple

## 7.2. Búsqueda del mejor modelo

- `step()` → Basado en Akaike Information Criterion
- Backwards, ...
- `modelo <- lm(pemax ~ age + sex + height + weight + bmp + fev1 + rv + frc + tlc)`
- `step(modelo)`

```

Step:  AIC=160.66
pemax ~ weight + bmp + fev1 + rv

      Df Sum of Sq  RSS   AIC
<none>      10355 160.66
-  rv         1   1183.6 11538 161.36
-  bmp         1   3072.6 13427 165.15
-  fev1        1   3717.1 14072 166.33
-  weight      1  10930.2 21285 176.67

Call:
lm(formula = pemax ~ weight + bmp + fev1 + rv)

Coefficients:
(Intercept)      weight          bmp          fev1           rv
   63.9467      1.7489     -1.3772      1.5477      0.1257
  
```

# 7. Modelos lineales, regresión múltiple

## 7.2. Búsqueda del mejor modelo

- Backward manual permite una estructura lógica (criterio significación estadística, ajuste obligatorio...)

```
summary(lm(pemax~age+sex+height+weight+bmp+fev1+rv+frc+tlc))  
summary(lm(pemax~age+sex+height+weight+bmp+fev1+rv+frc))  
summary(lm(pemax~age+sex+height+weight+bmp+fev1+rv))  
summary(lm(pemax~age+sex+height+weight+bmp+fev1))  
summary(lm(pemax~age+sex+height+weight+bmp))  
summary(lm(pemax~age+height+weight+bmp))  
summary(lm(pemax~height+weight+bmp))  
summary(lm(pemax~weight+bmp))  
summary(lm(pemax~weight))  
summary(lm(pemax~age+weight+height))  
summary(lm(pemax~age+height))  
summary(lm(pemax~age))  
summary(lm(pemax~height))
```

Edad, peso y altura estarán muy correlacionadas dado que son niños y adolescentes.



# 7. Modelos lineales, regresión múltiple

## 7.3.Regresión polinómica

- Regresion no lineal entre X e Y
- Problemas de colinealidad pueden producir inestabilidades en el ajuste
- $Y = a + b_1X + b_2X^2 + \dots + b_kX^k + E$
- `summary(lm(pemax~height+I(height^2)))`

```

Call:
lm(formula = pemax ~ height + I(height^2))

Residuals:
    Min       1Q   Median       3Q      Max
-51.411 -14.932  -2.288   12.787   44.933

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  615.36248   240.95580    2.554   0.0181 *
height       -8.08324    3.32052   -2.434   0.0235 *
I(height^2)    0.03064    0.01126    2.721   0.0125 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 24.18 on 22 degrees of freedom
Multiple R-squared:  0.5205,    Adjusted R-squared:  0.4769
F-statistic: 11.94 on 2 and 22 DF,  p-value: 0.0003081
  
```

# 7. Modelos lineales, regresión múltiple

## 7.4. Confusión, Interacción, Mediación

- **Confusión:**
  - Distorsión en el efecto estimado de una variable explicativa sobre la variable respuesta debido a una interposición de otra covariable, denominada factor de confusión, cuyo efecto se confunde o se mezcla con el verdadero efecto de la variable explicativa de interés.
  - La distorsión puede ser grande y dar lugar a sobreestimación o infraestimación del efecto, incluido cambiar la dirección del efecto de estudio.
  - Condiciones:
    - 1. Confusor linealmente relacionado con variable explicativa.
    - 2. Confusor asociado con la variable respuesta E independientemente de su asociación con la variable explicativa.
    - 3. Confusor no debe ser un paso intermedio en la relación de la variable explicativa con la variable respuesta Y.
    - En la práctica se comparan la **estimación cruda y ajustada**

# 7. Modelos lineales, regresión múltiple

## 7.4. Confusión, Interacción, Mediación

- **Interacción o efecto modificador:**
  - Cambio en la magnitud de la asociación entre la variable explicativa de interés y la variable respuesta a diferentes niveles de otra variable.
  - Ajuste del modelo con el término de interacción o estratificando los análisis

Ejemplo

```

> summary(lm(pemax~age+as.factor(sex)+age*as.factor(sex)))

Call:
lm(formula = pemax ~ age + as.factor(sex) + age * as.factor(sex))

Residuals:
    Min       1Q   Median       3Q      Max
-54.901 -12.447   5.069  15.099  45.099

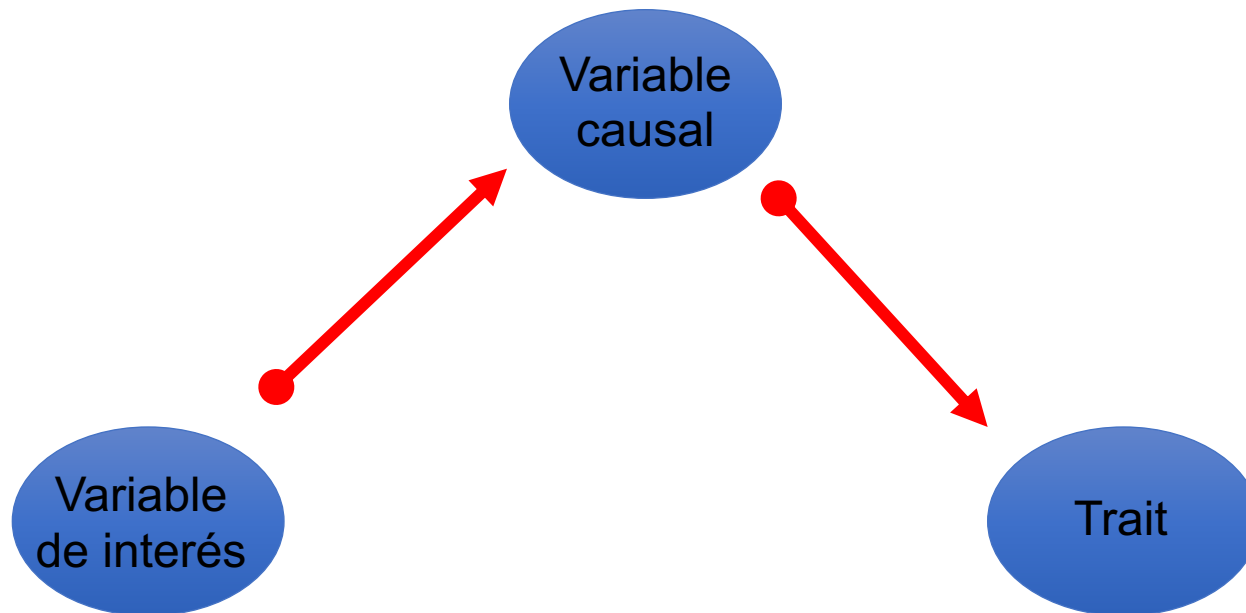
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    54.185     20.806   2.604  0.01656 *
age             4.162      1.281   3.249  0.00384 **
as.factor(sex)1  5.683     37.968  0.150  0.88243
age:as.factor(sex)1 -1.313     2.602  -0.505  0.61911
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 27.25 on 21 degrees of freedom
Multiple R-squared:  0.419,    Adjusted R-squared:  0.336
F-statistic: 5.048 on 3 and 21 DF,  p-value: 0.008655
  
```

# 7. Modelos lineales, regresión múltiple

## 7.4. Confusión, Interacción, Mediación

- **Mediación:**



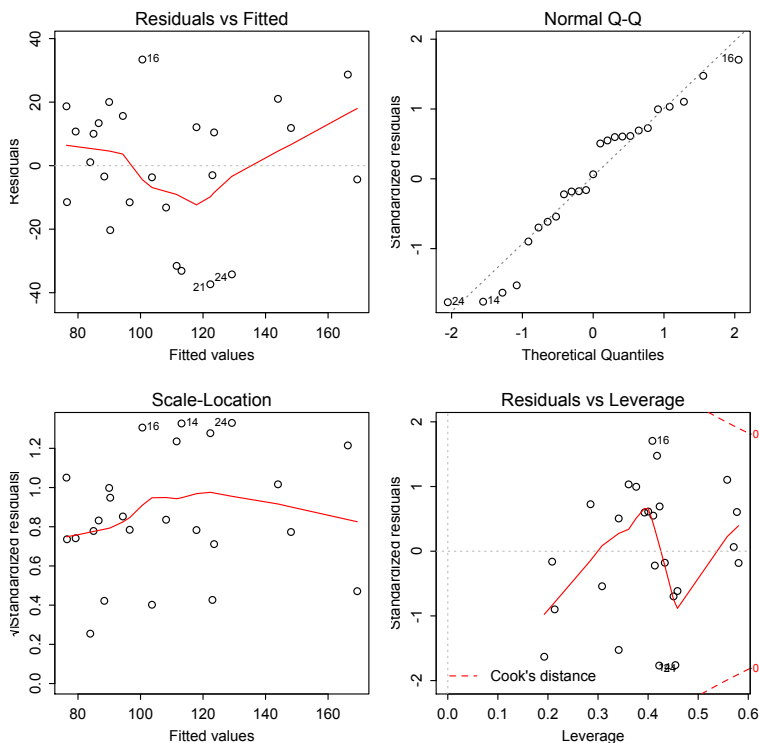
# 7. Modelos lineales, regresión múltiple

## 7.5. Diagnóstico de los modelos

```

modelo<-lm(pemax~age+sex+height+weight+bmp+fev1+rv+frc+tlc)
par(mfrow=c(2,2),mex=0.6)
plot(modelo)

```



Distancia de Cook = medida de la influencia de cada observación en los coeficientes de la regresión

# 7. Modelos lineales, regresión múltiple

## Otras cosas

- ANCOVA: analisis de la covarianza
- Linealidad sobre grupos
- Variables explicativas politómicas (en ejercicios)
- Variables Dummy

# EJERCICIOS

## “ejercicios.7 .modelos.lineales.regresion.múltiple.R”