

# Type Of Variables

## • CATEGORICAL/QUALITATIVE

- Ordinal (Health Status)
- Nominal (Fav.Color)
- Binary

## • NUMERICAL/QUANTITATIVE

- Continuous (Temperature)
- Discrete (Shoe Number)

## • OTHER TYPE

- Data Time
- Locations

ID TRAVEL	DESTIN	TRANSP	HOTEL STAR	HOTEL ROOM	HOTEL COST
1	MAD	TRAIN	3 Good	302	35,32€
2	ROME	PLANE	2 Bad	150	30,02€
3	LON	CAR	4 Very Good	390	70€
4	MAD	TRAIN	5 Excellent	302	105,32€
ID	NOM	NOM	ORD	NOMINAL	CONTINUOUS

## Preprocessing

### 1 - Data Selection

#### 1.1 - Instance And Variable Selection

### 2 - Outliers And Missing Detection

### 3 - Outliers And Missing Treatment

### 4 - Data Transformation

#### 4.1 - Units

#### 4.2 - Language

#### 4.3 - Enshort Modalities

#### 4.4 - Spelling Mistakes

### 5 - Variables Creation

### 6 - Feature Selection

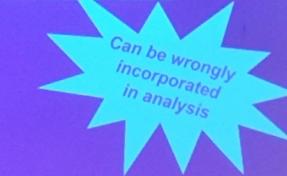
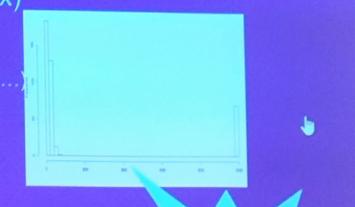
## Missing data (empty cells in data matrix)

- Random missing
  - non problematic
  - casual
  - follow same distribution as present data
  - inputation is easy: mean, 0
- Non random missing: absence is informative
  - come from some particular part of population
  - probably correspond to special values
  - difficult to induce from the present data
  - inputation is much difficult
  - very critical
  - very dangerous to ignore those individuals
  - asking religion in israel (muslims do not answer)
  - Asking age to a lady over 45
  - Frequency of observations (microbio tests in water)
- Non applicable value (non-random, structural)
  - salary of a non-working person
  - number of pregnancies of a man
  - number of cigarettes of a non-smoker person
  - age of menopause

©K. Gibert

## Empty cells in data matrix

- Representation:
  - \*, ?, "", depending on software
  - numerical variables: sometimes codified (0, 99999, -1...)
  - categorical variables: special modality (Ns/Nc, ...)
- Standardize missing representation
- Causes of missing data:
  - voluntary hidden (religion in israel) (always non-random)
  - data non-provided
  - data non-achievable
    - technical limitations (example anemometers IKE hurricane)
    - accessibility (no privileges, sensitive information)
  - data lost
  - data forced to missing (as a result of correction)



- Identification:
  - Numerical indicators (stdev...)

## Outlier

- Rare observation (presumed out of range)
- Multivariate vs univariate outlier

BIAS

Types of outliers:

- Mistake (Transcription Error or Measurement Error)
  - A person 560 years old
  - FIRST VERIFY If possible correct.  
If not, substitute by missing
- Informative point
  - A single informative point of a missing part of the population
  - Complete the sample when impossible, restrict scope of analysis
- Extreme value of the population
  - Very old person, 99 years old
  - Keep
- Value of another population
  - One swedish in the middle of a cannibal tribe, measuring height
  - Treat apart. CLEARLY REPORT ABOUT IT
- Missing code
  - Substitute by missing or inpute

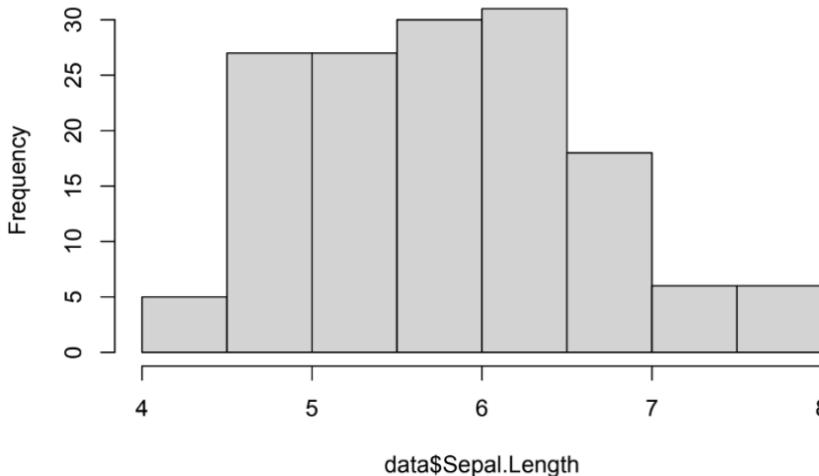
Care with suppressions

©K. Gibert

## ##NUMERICAL

```
hist(data$Sepal.Length)
```

Histogram of data\$Sepal.Length



```
data<-iris  
iris
```

```
summary(data$Sepal.Length)
```

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
##	4.300	5.100	5.800	5.843	6.400	7.900

- **Mínimo (Min.)**: Valor más pequeño (4.3 cm).
- **Primer cuartil (1st Qu.)**: Valor que separa el 25% inferior de los datos (5.1 cm).
- **Mediana (Median)**: Valor central (5.8 cm).
- **Media (Mean)**: Promedio (5.84 cm).
- **Tercer cuartil (3rd Qu.)**: Valor que separa el 75% inferior (6.4 cm).
- **Máximo (Max.)**: Valor más grande (7.9 cm).

Para variables numéricas:

### 1. Distribución:

- Si la **media ≈ mediana** (ejemplo: 5.84 vs 5.8), la distribución es aproximadamente simétrica.
- Si **media > mediana**, hay sesgo a la derecha (y viceversa).

### 2. Dispersion:

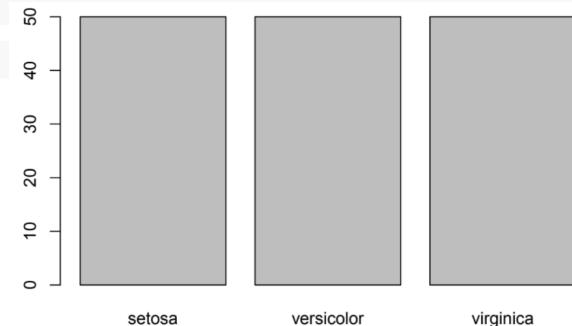
- **Rango intercuartílico (IQR)**:  $3rd\ Qu. - 1st\ Qu.$  ( $6.4 - 5.1 = 1.3$  cm). Un IQR grande indica alta variabilidad.
- **Rango total**:  $Max. - Min.$  ( $7.9 - 4.3 = 3.6$  cm).

### 3. Valores extremos:

- Si el mínimo o máximo están muy alejados de los cuartiles, podrían ser outliers (no es el caso en **iris**).

## ##CATEGORICAL

```
barplot(table(data$Species))
```

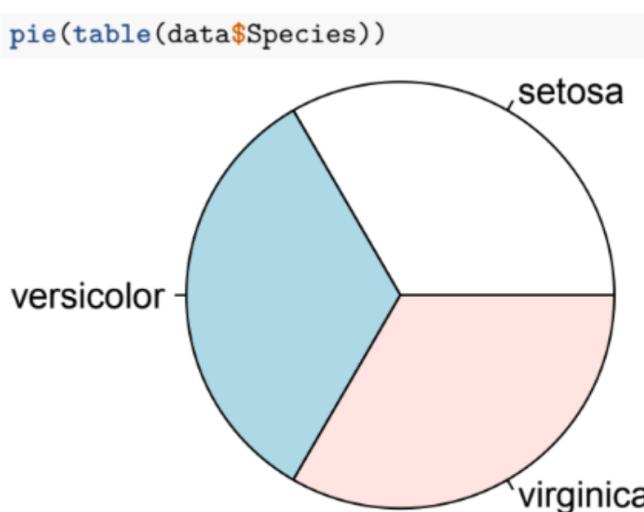


- Si **mediana** en el centro de la caja entonces la distribución es **simétrica**.
- Si **mediana** corta la caja en dos lados desiguales se tiene:
  - **Asimetría positiva o segada a la derecha** si la parte más larga de la caja es la parte superior a la mediana.
  - **Asimetría negativa o sesgada a la izquierda** si la parte más larga es la inferior a la mediana.

```

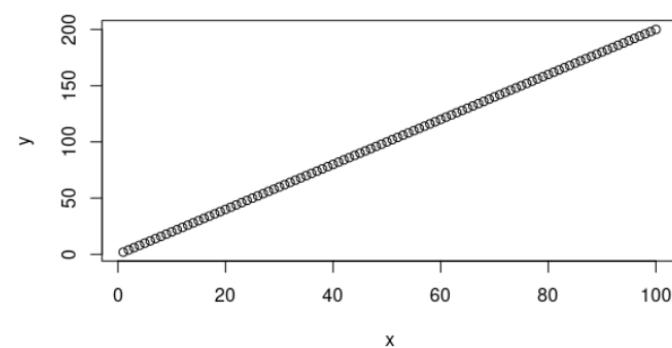
class(data[,1])
for ( i in 1:5){
  if(is.numeric(data[,i])){
    hist(data[,i])
    boxplot(data[,i])
    table(data[,i])
    summary(data[,i])
  } else{
    table(data[,i])
    barplot(table(data[,i]))
    pie(table(data[,i]))
  }
}

```



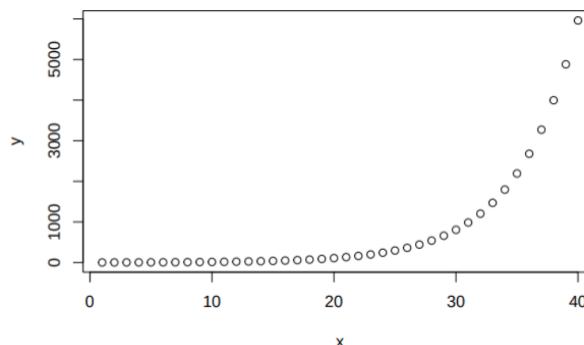
## ## Linear Model

**plot(x, y)**



a. Gráfico de Dispersion (`plot(x, y)`)

- **Propósito:** Visualizar la relación entre dos variables.
- **Interpretación:**
  - **Tendencia lineal:** Si los puntos siguen una línea recta, sugiere que un modelo lineal es adecuado.
  - **Dispersión aleatoria:** Indica ruido o falta de relación lineal.



a. `lm()` : Ajustar un Modelo Lineal

- **Qué hace:** Calcula los coeficientes (intercepto y pendientes) de un modelo lineal mediante mínimos cuadrados.

```

model <- lm(y ~ x, data = df) # Modelo lineal simple

```

Copy

- **Parámetros:**
  - **Fórmula (`y ~ x`):** Especifica la variable dependiente (`y`) y las independientes (`x`).
  - **data :** Dataframe que contiene las variables.

```

# Normal distributed error
e <- rnorm(100, 0, 5)

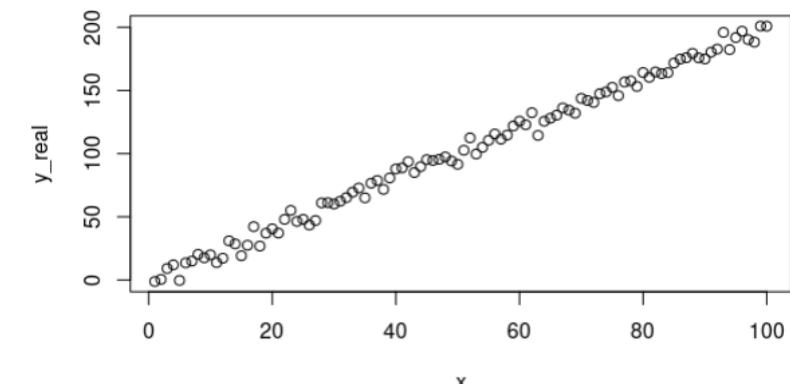
```

```

# Real data from a sample
y_real <- y + e

```

**plot(x, y\_real)**



`df <- data.frame(x, y_real)`

**model\_1 <- lm(y\_real~x, data=df)**

## summary(model\_1)

```
lm(formula = y_real ~ x, data = df)
```

Residuals:

Min	1Q	Median	3Q	Max
-11.6221	-3.0825	0.6418	3.3548	9.7599

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	-0.2927	0.9653	-0.303	0.762
x	2.0054	0.0166	120.841	<2e-16 ***
---				

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

Residual standard error: 4.79 on 98 degrees of freedom

Multiple R-squared: 0.9933, Adjusted R-squared: 0.9933

F-statistic: 1.46e+04 on 1 and 98 DF, p-value: < 2.2e-16

### a. Coeficientes (Coefficients):

- **Estimate:** Valor estimado del intercepto ( $\beta_0$ ) y pendiente ( $\beta_1$ ).
- **Std. Error:** Incertidumbre del estimador (error estándar).
- **t value:** Estadístico t para probar  $H_0 : \beta = 0$ .
- **Pr(>|t|):** p-valor. Si  $<0.05$ , el coeficiente es significativo.

### b. Bondad de Ajuste:

- **Residual standard error:** Desviación estándar de los residuos. Valores bajos indican mejor ajuste.
- **R<sup>2</sup>:** Proporción de varianza explicada (ej: 0.72 = 72% explicado).
- **R<sup>2</sup> ajustado:** Penaliza por variables innecesarias. Útil para comparar modelos.

### c. Interpretación:

- Por cada año adicional de educación, el prestigio aumenta 5.36 unidades ( $p < 0.001$ ).
- El intercepto no tiene interpretación práctica aquí (educación no puede ser 0).

$r_{XY}^2 = R^2$  = coeficiente de determinación

$0 \leq R^2 \leq 1 \rightarrow \begin{cases} \text{más cerca del 1} \rightarrow \text{más capacidad predictiva y poca variabilidad} \\ \text{más cerca del 0} \rightarrow \text{menos capacidad predictiva y mucha variabilidad} \end{cases}$

### ✓ Interpretación de cada columna:

#### • Estimate:

- **Intercepto (3.001)** → Predicción de YB cuando XB = 0.
- **Pendiente (0.500)** → Por cada unidad que aumenta XB, YB aumenta **en promedio** 0.500 unidades.

#### • Std. Error:

- Mide la precisión de la estimación.
- **Cuanto menor sea el error estándar, más confiable es la estimación.**

#### • t value:

- Se calcula como **Estimate / Std. Error**.
- **Valores altos indican que el coeficiente es estadísticamente significativo.**

#### • Pr(>|t|) (p-valor):

- $p < 0.05$  → La variable es significativa en el modelo.
- $p > 0.05$  → No hay suficiente evidencia para decir que la variable afecta YB.
- En este caso, **XB es muy significativa (p = 0.00218)**.

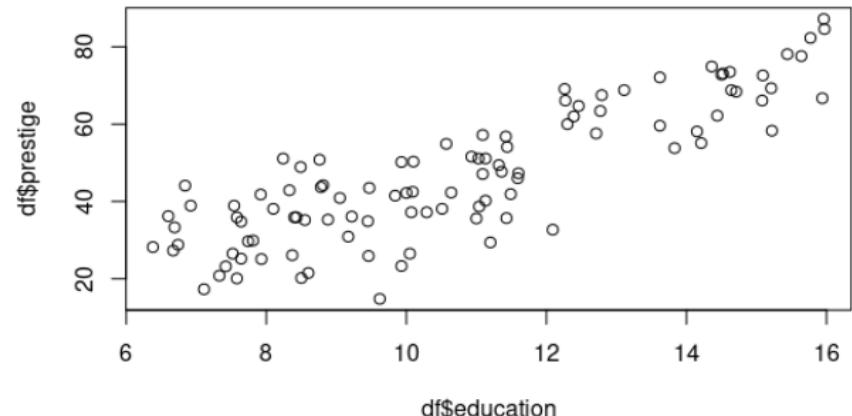
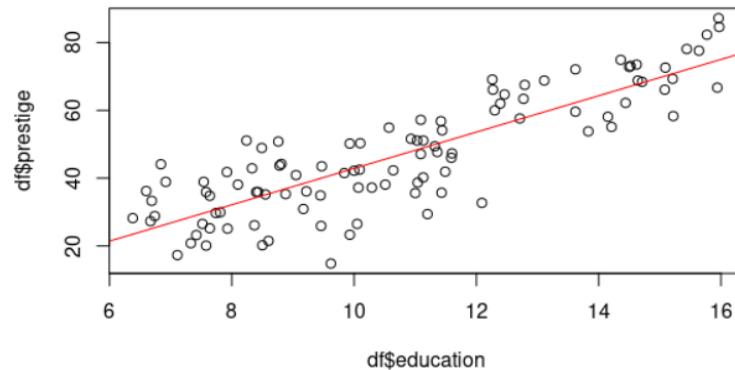
### ✓ Conclusión:

El coeficiente XB es **significativo** ( $p < 0.05$ ), lo que significa que hay una relación entre YB y XB.

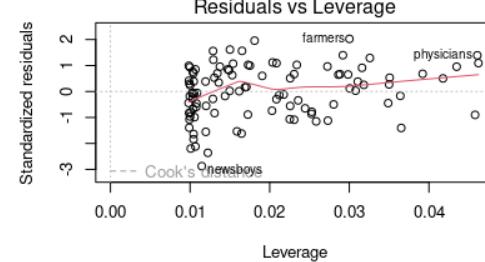
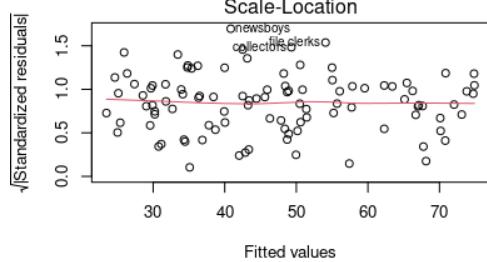
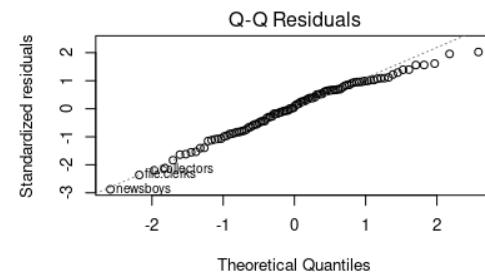
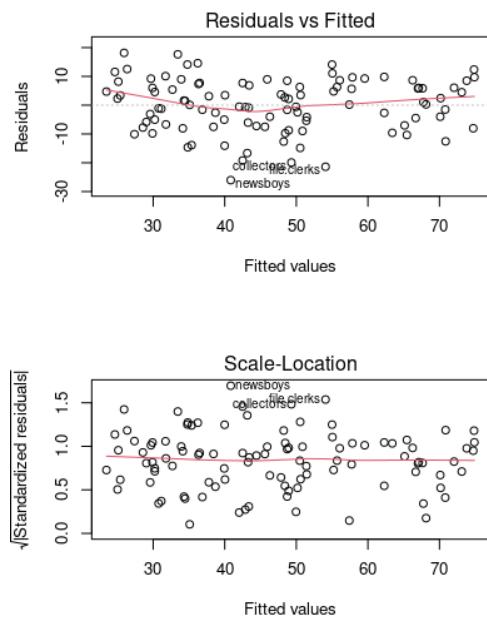
### 4. Validación de Supuestos

1. **Linealidad:** Verificada con gráficos de dispersión y residuos vs ajustados.
2. **Normalidad:** Evaluada con Q-Q plot.
3. **Homocedasticidad:** Confirmada si los residuos tienen varianza constante.
4. **Independencia:** Asumida si los datos no tienen estructura temporal o espacial.

```
plot(df$education, df$prestige)
abline(a = m1$coefficients[1], b = m1$coefficients[2], col = "red")
```

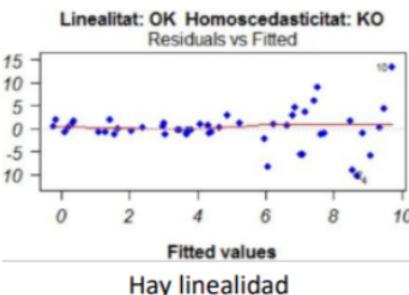


**plot(m1)**

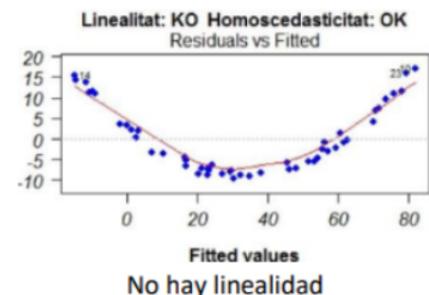


- **En Residuos vs Ajustados:**

- Residuos distribuidos **aleatoriamente** alrededor de cero, sin patrones (ej: forma de embudo, U, o nube curvada).
- Línea roja suavizada cercana a la horizontal (no curva ascendente/descendente).



Hay linealidad



No hay linealidad

We have moreless homocedasticity (some education levels have larger variance), higher values have higher residuals (breaking a bit normality). There are not many too influencial observations.

## 2. Normalidad

Gráfico clave:

- **Q-Q Plot de residuos:** `plot(modelo, which = 2)` (segunda gráfica al usar `plot(modelo)`).

Qué buscar:

- Los puntos deben seguir la **línea diagonal teórica**.
- Desviaciones menores en las colas son aceptables.

Problemas comunes:

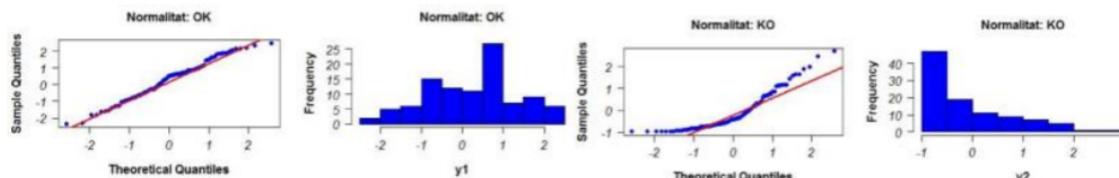
- **Puntos fuera de la línea en las colas:** Colas pesadas (residuos más extremos de lo normal).

- **Puntos en S o curva:** Asimetría (residuos sesgados).

Solución:

- Transformar la variable respuesta (ej: `log(y)`).

- **Normalidad:** en el caso de un gráfico que se mantengan los datos sobre la recta `qqnorm` y en el caso de un histograma que forme una campana.



## 3. Homocedasticidad

Gráficos clave:

- **Residuos vs Valores Ajustados:** `plot(modelo, which = 1)`.
- **Scale-Location:** `plot(modelo, which = 3)` (tercera gráfica al usar `plot(modelo)`).

Qué buscar:

- **Residuos vs Ajustados:** Dispersión constante (misma amplitud vertical en todo el rango de valores ajustados).
- **Scale-Location:** Línea roja suavizada **horizontal** (sin inclinación).

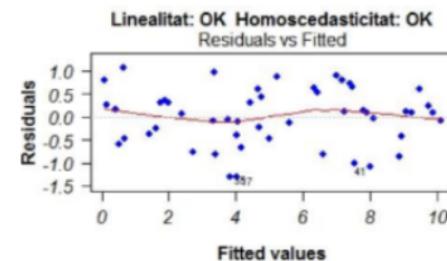
Problemas comunes:

- **Forma de embudo:** La varianza aumenta con los valores ajustados (heterocedasticidad).

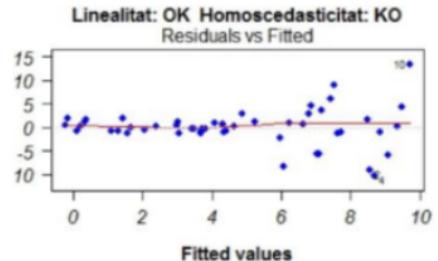
Solución:

- Transformar la respuesta (ej: `sqrt(y)`).

- **Homoscedasticidad:** misma  $\sigma^2$  para cualquier caso/dato, cuando miramos el gráfico vemos que todos los datos tienen el mismo error.



Hay homoscedasticidad



Hay heteroscedasticidad (no hay homoscedasticidad)

## 4. Independencia

Gráfico clave:

- **Residuos vs Orden de Observación:** `plot(residuos, type = "b")` (no generado automáticamente por `plot(modelo)`).

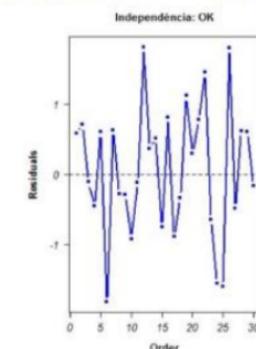
Qué buscar:

- **Aleatoriedad:** Sin patrones (ej: tendencias ascendentes/descendentes, ciclos).
- **Si hay estructura temporal/espacial:**
  - Ejemplo: Residuos positivos seguidos de negativos en secuencia (autocorrelación).

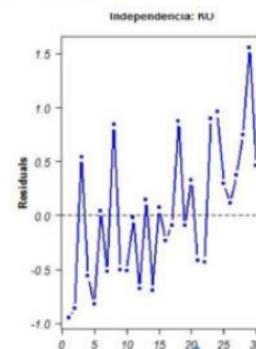
Problemas comunes:

- **Autocorrelación:** Patrón en serie (ej: datos temporales no independientes).

- **Independencia:** Detecta si existe o no dependencia entre los datos.



Hay independencia, porque no observamos ningún patrón en el gráfico



No hay independencia, porque observamos un patrón en el gráfico

We add now a third regressor: **women**.

```
m3 <- lm(prestige~education+income+women, data=df)
```

```
summary(m3)
```

Call:

```
lm(formula = prestige ~ education + income + women, data = df)
```

Residuals:

Min	1Q	Median	3Q	Max
-19.8246	-5.3332	-0.1364	5.1587	17.5045

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	-6.7943342	3.2390886	-2.098	0.0385 *
education	4.1866373	0.3887013	10.771	< 2e-16 ***
income	0.0013136	0.0002778	4.729	7.58e-06 ***
women	-0.0089052	0.0304071	-0.293	0.7702

---

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

Residual standard error: 7.846 on 98 degrees of freedom

Multiple R-squared: 0.7982, Adjusted R-squared: 0.792

F-statistic: 129.2 on 3 and 98 DF, p-value: < 2.2e-16

We see now that women is not significative, that is, it does not explain the variable Prestige. We see that with the addition of this variable we do not improve the residual standard error. The explained variability (R<sup>2</sup>) always increases with the addition of variables. To solve this, we check R<sup>2</sup> adj, which takes into consideration the number of parameters and penalises the addition. We see that the addition of variable women worsens R<sup>2</sup> adj.

### 1. `residualPlot(ma)`

Genera un gráfico de residuos contra los valores ajustados para evaluar la linealidad y la homocedasticidad.

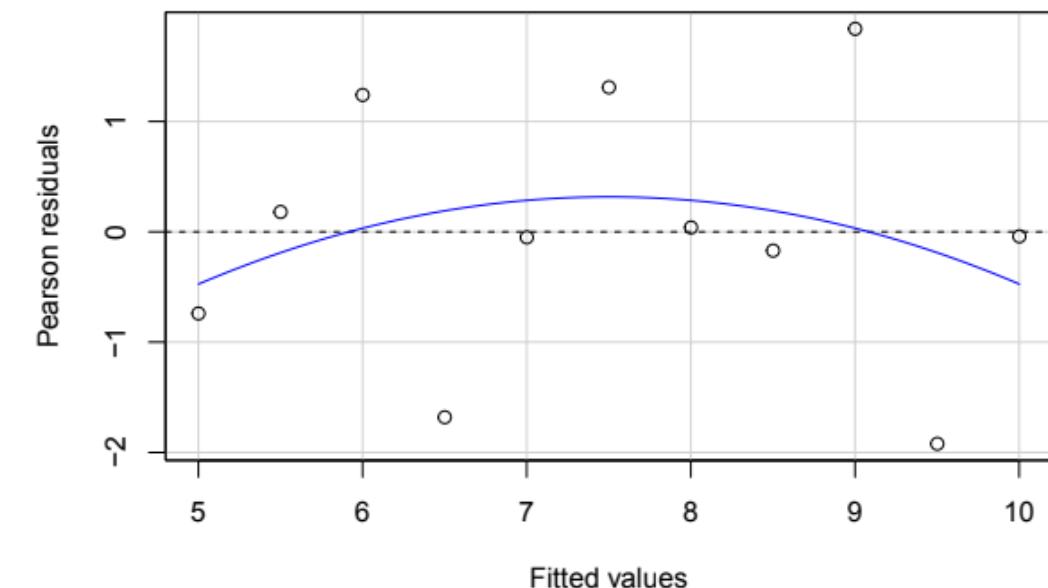
#### ✓ Qué ver en el gráfico:

- Si los residuos están aleatoriamente distribuidos alrededor de 0 → El modelo es adecuado.
- Si se observa un patrón curvo → Puede indicar que la relación no es lineal y que una transformación o un modelo polinómico podría ser mejor.
- Si la dispersión de los residuos aumenta con los valores ajustados → Puede haber heterocedasticidad (varianza no constante), lo que significa que los errores no son uniformes a lo largo del rango de los datos.

#### 💡 Ejemplo de interpretación:

- Un patrón en forma de "U" o "V" sugiere que falta un término cuadrático en el modelo.
- Un abanico (residuos dispersos en los valores grandes) indica heterocedasticidad.

### `residualPlot(ma)`



## 2. marginalModelPlots(ma)

Genera gráficos que comparan la relación observada en los datos con la relación modelada por la regresión.

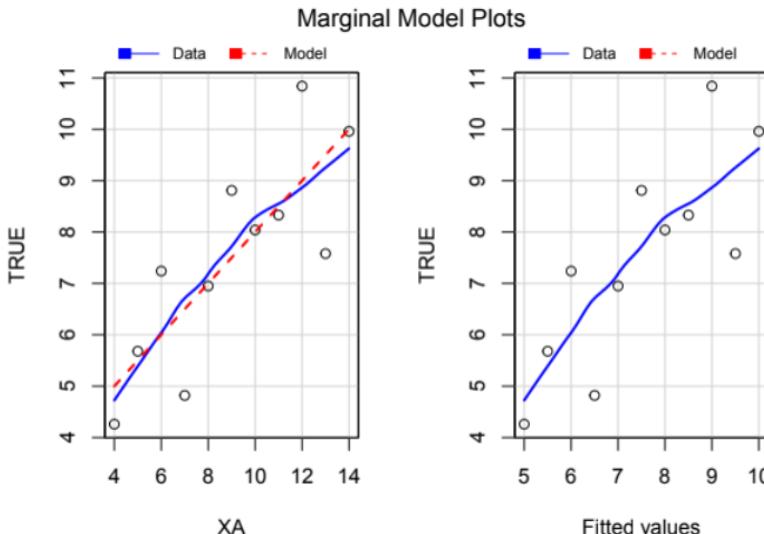
### Cómo interpretar:

- Cada gráfico muestra una variable predictoría (`XA`) en el eje X y el valor de `YA` en el eje Y.
- Se dibujan dos curvas:
  1. La línea de datos reales (valores observados).
  2. La línea ajustada por el modelo (predicciones del modelo de regresión).
- Si las dos líneas son similares, significa que el modelo está capturando correctamente la relación entre las variables.
- Si las líneas difieren mucho, el modelo puede estar mal especificado.

### Ejemplo de interpretación:

- Si los datos siguen una curva y el modelo es una línea recta, es probable que necesites una regresión no lineal o un término polinómico.

```
marginalModelPlots(ma)
```



late the AIC and BIC value from the linear model we just estimated:

```
AIC(ma) #AIC
```

```
## [1] 39.68137
```

```
AIC(ma, k = log(nrow(anscombe))) #BIC
```

```
## [1] 40.87506
```

## a) Forward Selection (Selección Hacia Adelante)

- Se parte de un **modelo nulo** (`YA ~ 1`), que solo tiene el intercepto.
- Se agregan variables una por una, evaluando si mejoran AIC/BIC.
- Se detiene cuando agregar más variables no mejora AIC/BIC.
- El **modelo inicial** (`YA ~ 1`) tiene un **AIC de 16.55**.
- Agregar `XA` reduce el AIC a **6.46**, lo que signifICA que es una mejora significativa.

Let's run the stepwise regression algorithm to determine the best linear model:

```
ma_0 <- lm(YA ~ 1, data=anscombe) # Null model  
step(ma_0, ~XA, direction="forward", data=anscombe)
```

```
## Start: AIC=16.55  
## YA ~ 1  
##  
##           Df Sum of Sq   RSS   AIC  
## + XA     1    27.51 13.763  6.4647  
## <none>          41.273 16.5454  
##  
## Step: AIC=6.46  
## YA ~ XA  
##  
## Call:  
## lm(formula = YA ~ XA, data = anscombe)  
##  
## Coefficients:  
## (Intercept)          XA  
##       3.0001        0.5001
```

## Selección de Modelo

```
AIC(ma), AIC(ma, k = log(nrow(anscombe)))
```

El **Akaike Information Criterion (AIC)** y **Bayesian Information Criterion (BIC)** evalúan qué tan bueno es el modelo.

- Menor AIC/BIC → Mejor modelo.
- Mayor AIC/BIC → Peor modelo.

✓ Uso Si se comparan varios modelos, el que tenga el menor AIC/BIC es preferible.

## b) Backward Selection (Selección Hacia Atrás)

- Se parte del **modelo completo** (con todas las variables).
- Se eliminan variables una por una y se evalúa AIC/BIC.
- Se detiene cuando quitar más variables empeora el modelo.

```
step(ma, direction = "backward", data=anscombe)
```

```
## Start:  AIC=6.46
## YA ~ XA
##
##          Df Sum of Sq   RSS   AIC
## <none>            13.763  6.4647
## - XA     1    27.51 41.273 16.5454
##
## Call:
## lm(formula = YA ~ XA)
##
## Coefficients:
## (Intercept)      XA
##           3.0001    0.5001
```

## ¿Qué hace `acf(ma$residuals)`?

`acf()` calcula la **función de autocorrelación (Autocorrelation Function, ACF)** de los residuos del modelo `ma`.

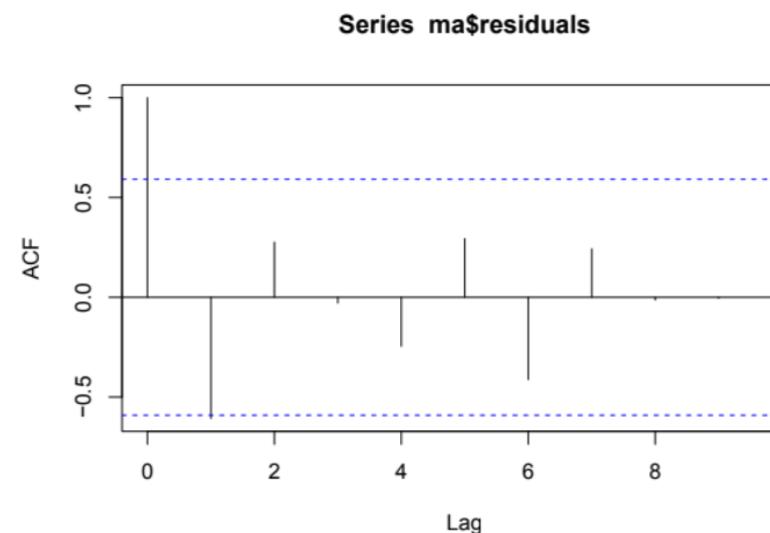
⚠️ **Autocorrelación:** Ocurre cuando los errores de la regresión (residuos) están correlacionados entre sí. Esto **viola el supuesto de independencia de los residuos**, lo que puede hacer que los resultados del modelo no sean confiables.

### Cómo interpretar el gráfico de `acf(ma$residuals)`:

1. **Eje X:** Representa los **lags** (retrasos).
  - Lag 1: Comparación entre un residuo y el residuo anterior.
  - Lag 2: Comparación entre un residuo y el residuo de dos observaciones atrás, etc.
2. **Eje Y:** Muestra la **correlación** entre los residuos en cada lag.
3. **Líneas azules (límites de confianza):**
  - Si los valores de autocorrelación (barras negras) quedan dentro de las **bandas azules**, no hay evidencia de autocorrelación.
  - Si una barra **sale fuera de las bandas azules**, hay autocorrelación significativa en ese lag.

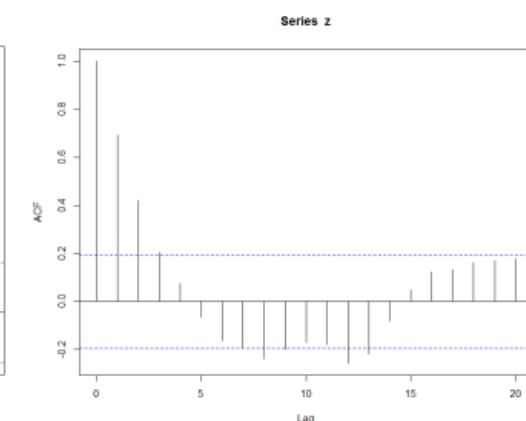
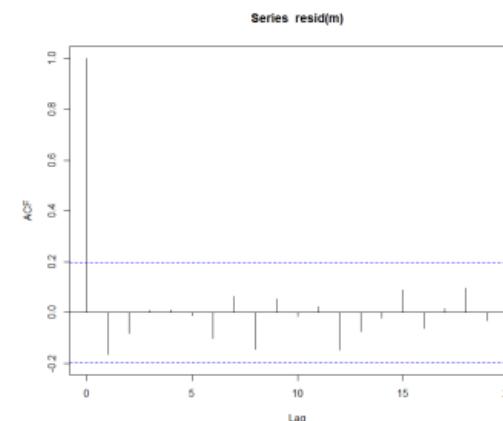
Autocorrelation / Independence:

```
acf(ma$residuals)
```



No temporal dependency

Temporal dependency



### Ejemplo de interpretación:

- Si **todas las barras están dentro de las bandas azules** → No hay autocorrelación → El modelo cumple el supuesto de independencia.
- Si **hay barras que sobresalen** → Hay autocorrelación → Esto sugiere que el modelo no ha capturado toda la estructura de los datos.

### 3. Diagnóstico de multicolinealidad

```
vif(m1) # Variance Inflation Factor (VIF)
```

Copiar Editar

Detecta si hay correlación entre las variables predictoras.

- Valores de VIF mayores a 5 o 10 indican alta colinealidad y pueden afectar la estabilidad del modelo.
- Aquí, los valores de VIF están por debajo de 3, lo que indica que no hay problemas graves de multicolinealidad.

### 4. Transformación de la variable dependiente

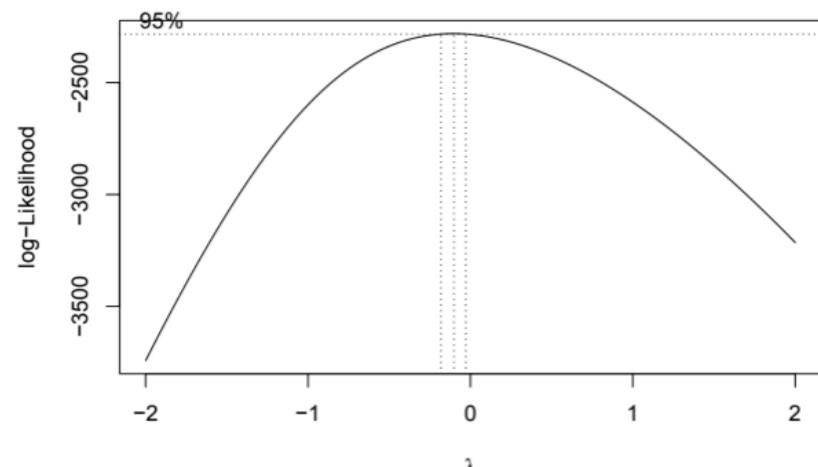
```
boxcox(price ~ mileage + tax + mpg + age, data = df)
```

Copiar Editar

La función `boxcox()` sugiere si es necesaria una transformación de `price`.

- Si  $\lambda = 0$ , se recomienda usar `log(price)`.

```
# Target variable transformation?  
boxcox(price~mileage+tax+mpg+age,data=df)
```



# Transformations to my regressors?

```
boxTidwell(log(price)~mileage+tax+mpg+age,data=df[!df$mout=="YesMOut",])
```

```
##          MLE of lambda Score Statistic (t)  Pr(>|t|)  
## mileage      0.13269                  1.0375 0.2997598  
## tax          0.46073                 -3.2720 0.0011062 **  
## mpg          -2.25574                  8.9625 < 2.2e-16 ***  
## age          1.41541                 -3.5963 0.0003393 ***  
## ---  
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1  
##  
## iterations =  11  
##  
## Score test for null hypothesis that all lambdas = 1:  
## F = 21.964, df = 4 and 955, Pr(>F) = < 2.2e-16  
# Power transformations of the predictors in a linear model
```

Evalúa si las variables explicativas deberían transformarse.

- Si  $p < 0.05$ , la variable no tiene una relación lineal y necesita transformación.

Lambda = 0 --- log

Lambda = 0.5 --- sqrt

Lambda = 1 ---- no transformation

Lambda = 2 --- Poly

1. `avPlots(m2, id=list(method=cooks.distance(m2), n=5))`

¿Qué hace?

- Genera gráficos de valores añadidos (Added Variable Plots, AV Plots).
- Muestran cómo una variable afecta `log(price)` después de considerar las demás variables.
- Destaca los 5 puntos más influyentes usando `Cook's Distance`.

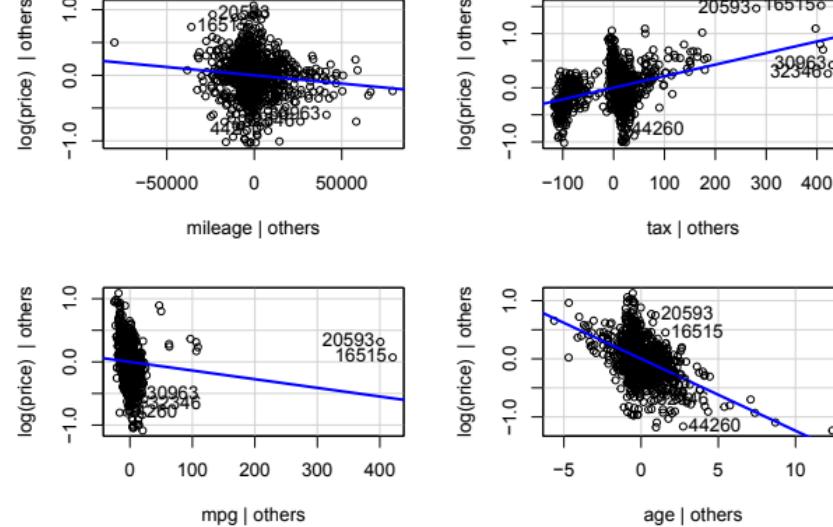
Cómo interpretarlo:

- 1 Si la relación es lineal (recta diagonal bien ajustada) → La variable se comporta bien en el modelo.
- 2 Si hay curvatura → Podría necesitar una transformación (ej. `log(variable)`, `poly(variable, 2)`).
- 3 Si hay puntos muy alejados de la recta → Son observaciones influyentes que podrían afectar el modelo.

Acción recomendada:

- Si hay puntos extremos, puedes analizarlos con `influencePlot(m2)`.
- Si ves curvatura, podrías probar una transformación.

## Added-Variable Plots



2. `crPlots(m2, id=list(method=cooks.distance(m2), n=5))`

💡 ¿Qué hace?

- Genera gráficos de residuos parciales (Component + Residual Plots, CR Plots).
- Muestra la relación entre cada variable predictora y `log(price)`, pero sin considerar otras variables.

✓ Cómo interpretarlo:

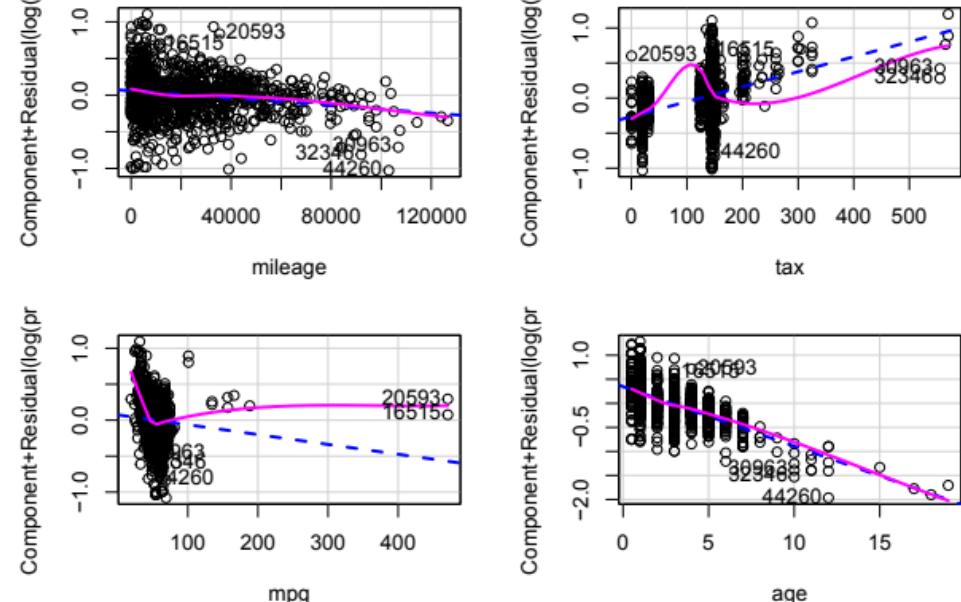
- Si la relación es lineal → La variable es adecuada en un modelo lineal.
- Si hay curvatura → Podría ser necesario usar `log(variable)`, `sqrt(variable)`, o `poly(variable, 2)`.
- Si hay valores extremos destacados por Cook's Distance → Pueden estar afectando mucho el modelo y quizás deban eliminarse o analizarse.

💡 Acción recomendada:

- Si la relación no es lineal, prueba una transformación (`log`, `sqrt`, `poly`).
- Si hay valores atípicos, revisa con `influencePlot(m2)`.

```
crPlots(m2, id=list(method=cooks.distance(m2), n=5))
```

## Component + Residual Plots



💡 ¿Qué hace Anova(m3) ?

La función `Anova()` del paquete `car` realiza un **test ANOVA tipo II**, que evalúa el **efecto neto de cada variable** en el modelo de regresión `m3`.

- A diferencia de `summary(m3)`, que evalúa cada coeficiente individualmente, `Anova(m3)` mide la contribución global de cada variable al modelo.

```
# Validation and effects consideration:  
Anova(m3) #Net effect test
```

```
## Anova Table (Type II tests)  
##  
## Response: log(price)  
##              Sum Sq Df F value Pr(>F)  
## log(mileage)   0.184  1  2.5864 0.1081  
## sqrt(tax)      5.040  1 70.6516 <2e-16 ***  
## poly(mpg, 2) 17.777  2 124.6003 <2e-16 ***  
## age            27.036  1 379.0041 <2e-16 ***  
## Residuals     68.338 958  
## ---  
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

## 1. ¿Cómo interpretar la tabla de Anova(m3) ?

Variable	Sum Sq	Df	F value	Pr(>F)	¿Significativa?
log(mileage)	0.184	1	2.59	0.1081	✗ No
sqrt(tax)	5.040	1	70.65	<2e-16	✓ Sí (** muy significativa)
poly(mpg, 2)	17.777	2	124.60	<2e-16	✓ Sí (** muy significativa)
age	27.036	1	379.00	<2e-16	✓ Sí (** muy significativa)
Residuals	68.338	958	-	-	-

# anova(model\_no\_type, model\_final) # Falla, type té NAs!!

## 2. Explicación de cada columna

### 1 Sum Sq (Suma de cuadrados)

- Mide cuánta variabilidad del `log(price)` es explicada por cada variable.
- Mayor valor = Mayor impacto en el precio.

### 2 Df (Grados de libertad)

- Muestra cuántos términos se usaron en cada variable.
- `poly(mpg, 2)` tiene **Df = 2** porque se usó un polinomio de grado 2.

### 3 F value

- Indica cuán fuerte es el efecto de la variable en el modelo.
- Valores altos = Variable importante.
- `age` tiene el **mayor impacto en el precio** ( $F = 379.00$ ).

### 4 Pr(>F) (p-valor)

- Si  $p < 0.05 \rightarrow$  La variable es significativa.
- Si  $p > 0.05 \rightarrow$  No es significativa y podría eliminarse.

## 4. ¿Qué hacer después?

- Si una variable tiene  $p > 0.05$ , se puede considerar eliminarla y volver a evaluar el modelo (`summary(m4)`).
- Si una variable tiene una F muy alta, confirma que tiene un impacto fuerte en el precio.
- El polinomio `poly(mpg, 2)` es significativo, por lo que el efecto de `mpg` no es lineal.

## 3. ¿Cómo interpretar los resultados?

### ✓ Variables SIGNIFICATIVAS ( $p < 0.05$ )

- `sqrt(tax)` ( $p < 2e-16$ ) : El impuesto tiene un fuerte efecto sobre el precio.
- `poly(mpg, 2)` ( $p < 2e-16$ ) : La eficiencia de combustible afecta el precio, de manera no lineal.
- `age` ( $p < 2e-16$ ) : La antigüedad del auto es la variable más importante.

### ✗ Variable NO significativa ( $p > 0.05$ )

- `log(mileage)` ( $p = 0.1081$ ):
  - No tiene un impacto fuerte en el precio.
  - Podría eliminarse del modelo (`m4 <- update(m3, . ~ . - log(mileage))` y ver si  $R^2$  mejora.

## 1. ¿Cómo interpretar influencePlot(m5) ?

Este gráfico muestra tres cosas importantes para cada observación:

### 1 Cook's Distance (y-axis)

- Mide cuánto cambiaría el modelo si eliminamos una observación.
- Si `Cook's Distance` > 1, el punto tiene demasiada influencia.
- Si `Cook's Distance` es pequeño, el punto no afecta mucho el modelo.

### 2 Leverage (x-axis)

- Mide cuánto se aleja una observación del resto de los datos.
- Valores altos indican que la observación está en una región donde no hay muchos datos.

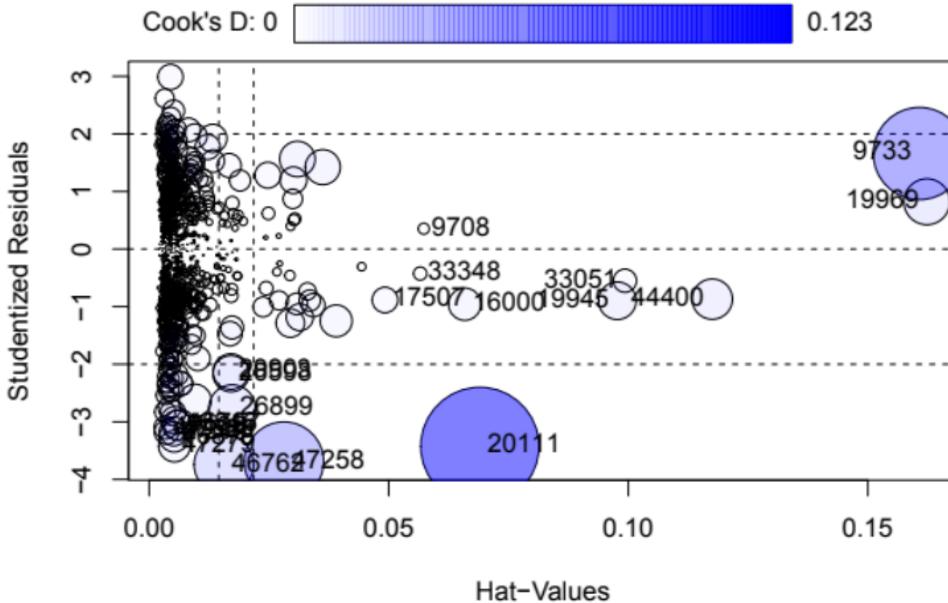
### 3 Residual Studentizado (punto en el gráfico)

- Mide cuánto se desvía un punto del patrón esperado en el modelo.
- Si un punto tiene un residual grande y un leverage alto, podría ser un outlier problemático.

### Ejemplo de interpretación:

- Si un punto está en la parte superior del gráfico (`Cook's Distance` grande) → Es un punto altamente influyente y podría afectar el modelo.
- Si un punto está a la derecha (`Leverage` alto) → Es un punto con información única, pero no necesariamente problemático.
- Si un punto tiene ambas cosas (alto `Cook's Distance` y alto `Leverage`) → Es un dato problemático y podría valer la pena eliminarlo o analizarlo más a fondo.

```
influencePlot(m5, id=list(n=10))
```



📌 1. `with(df, plotMeans(prestige, type, error.bars = "sd"))`

📌 ¿Qué hace?

- Muestra las medias de `prestige` para cada categoría de `type`.
- Incluye barras de error con desviación estándar (`sd`).

✓ Cómo interpretarlo:

- Si las barras de error se superponen mucho → No hay mucha diferencia entre los grupos de `type`.
- Si hay diferencias claras entre medias → `type` tiene un impacto fuerte en `prestige`.

📌 Acción recomendada:

- Si las diferencias no son claras, revisar `Anova(model)` para confirmar si `type` es significativo.

📌 2. `emmip(model, ~type, CIs=T)`

📌 ¿Qué hace?

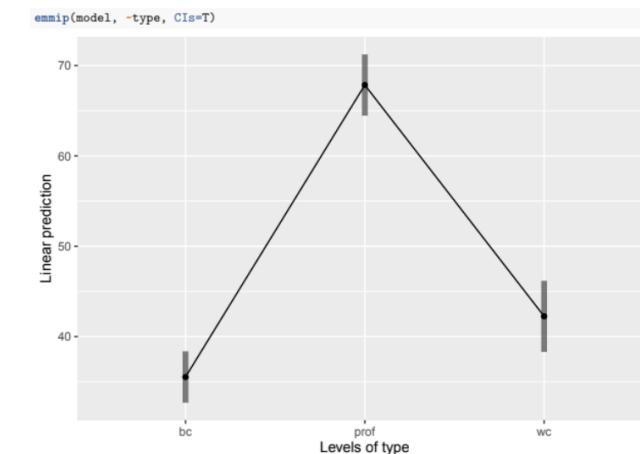
- Muestra la media marginal estimada (Estimated Marginal Means) de `prestige` para cada categoría de `type`.
- Incluye intervalos de confianza (`CIs=T`).

✓ Cómo interpretarlo:

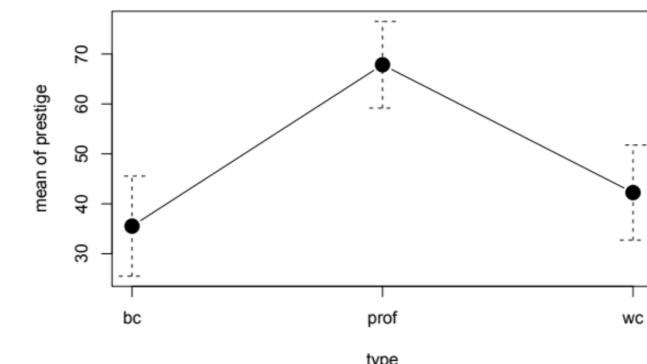
- Si los intervalos de confianza son estrechos, la predicción de `prestige` es precisa.
- Si los intervalos de confianza no se superponen, `type` tiene un efecto fuerte sobre `prestige`.
- Si los intervalos se superponen, `type` podría no ser tan importante.

📌 Acción recomendada:

- Si `type` no muestra diferencias claras, reconsiderar su inclusión en el modelo.



Plot of Means



## DFBetas

The most direct approach to assessing influence is to assess how the regression coefficients change if outliers are omitted from the model. We can use DFBetas\_ij(). Use dfbetas(model) in R.

```
head(dfbetas(model_final3))  
  
## (Intercept) education log(income) typeprof  
  
## gov.administrators -1.390206e-02 -0.024635407 0.0216433448 0.029981713  
## general.managers 1.418582e-01 0.150406078 -0.1857267486 -0.122448927  
## accountants 9.715269e-05 -0.002022608 0.0006121935 0.002769585  
## purchasing.officers -4.351735e-03 0.062199298 -0.0175433476 -0.069930235  
## chemists 2.359544e-02 0.048338908 -0.0390045426 0.013123389  
## physicists -1.770880e-02 0.062381406 -0.0050452428 -0.029866497  
  
## typewc  
## gov.administrators 0.017650150  
## general.managers -0.109932134  
## accountants 0.001401876  
## purchasing.officers -0.043058491  
## chemists -0.034492189  
## physicists -0.042675674
```

### 💡 ¿Qué hace?

- Mide cuánto cambia cada coeficiente ( $\beta$ ) si se elimina una observación.
- Se calcula para cada observación y cada variable en el modelo.

### ✓ Cómo interpretarlo:

Observación	Intercepto	Education	log(income)	typeprof	typewc
gov.administrators	-0.0139	-0.0246	0.0216	0.0299	0.0176
general.managers	0.1418	0.1504	-0.1857	-0.1224	-0.1099
accountants	0.0001	-0.0020	0.0006	0.0028	0.0014
physicists	-0.0177	0.0624	-0.0050	-0.0299	-0.0427

### 💡 Reglas de interpretación:

- 1 Si DFBeta está cerca de 0, la observación no influye mucho.
- 2 Si DFBeta es alto (ej.  $> 0.2$  o  $< -0.2$ ), indica que la observación cambia mucho el coeficiente del modelo.
- 3 Valores grandes en DFBeta sugieren que la observación podría ser atípica y afectar los resultados.

### 💡 Acción recomendada:

- Si hay valores extremos, revisarlos con influencePlot(model\_final3).
- Si eliminarlos mejora la estabilidad del modelo (summary(model\_final3)), considerar excluirlos.

## 2. dfbetasPlots(model\_final3) : Visualización de DFBetas

r

Copiar Editar

```
dfbetasPlots(model_final3)
```

### 💡 ¿Qué hace?

- Grafica cómo cambia cada coeficiente al remover cada observación.
- Resalta puntos que afectan más el modelo.

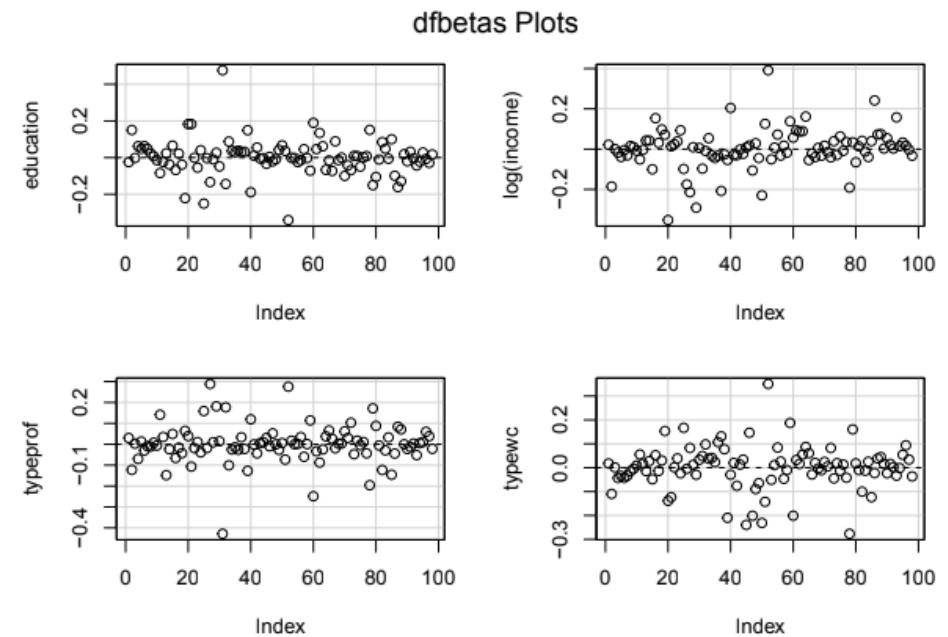
### ✓ Cómo interpretarlo:

- 1 Si la mayoría de los valores están dentro de (-0.2, 0.2), no hay problemas.
- 2 Si hay puntos fuera de (-0.2, 0.2), esas observaciones afectan mucho el modelo.
- 3 Si hay patrones en la gráfica, podría haber colinealidad o errores sistemáticos.

### 💡 Acción recomendada:

- Revisar observaciones fuera del rango (-0.2, 0.2) con influencePlot(model\_final3).
- Si un punto afecta solo una variable, podríamos hacer una transformación en esa variable.

```
dfbetasPlots(model_final3)
```



## Cook's D

To overcome the problem of having a 2D object we have Cook's Dthat presents a single summary measure for each observation. Use `cooks.distance(model)` in R.

```
head(cooks.distance(model_final3))

## gov.administrators    general.managers      accountants purchasing.officers
## 4.722351e-04        1.087006e-02       2.733529e-06     1.230362e-03
## chemists            physicists
## 4.790623e-03        2.368733e-03
```

### 💡 ¿Qué hace?

- Mide cuánto cambia el modelo completo si se elimina una observación.
- Valores altos indican observaciones muy influyentes.

### ✓ Cómo interpretarlo:

Observación	Cook's Distance
gov.administrators	0.00047
general.managers	<b>0.01087</b>
accountants	0.0000027
purchasing.officers	0.00123
chemists	0.00479
physicists	0.00237

### 💡 Reglas generales:

- Si `Cook's Distance > 1`, la observación tiene demasiada influencia en el modelo y debería revisarse.
- Si `Cook's Distance` es bajo ( $< 0.1$ ), el punto no afecta mucho.
- Si hay un solo punto con valor muy alto, el modelo puede depender demasiado de esa observación.

### 💡 Acción recomendada:

- Si `Cook's Distance` es alto, revisar con `influencePlot(model_final3)`.
- Si la eliminación del punto mejora el modelo () `summary(model_final3)`, considerar excluirlo.

```
matplot(dfbetas(model_final3), type = "l",
        col=2:7, lwd=2, xlim = c(0, 100), ylim = c(-1.5, 0.6))
lines(sqrt(cooks.distance(model_final3)), col=1, lwd=3)
abline(h = 2/sqrt(dim(df)[1]), lty=3, lwd=1, col=5)
abline(h = -2/sqrt(dim(df)[1]), lty=3, lwd=1, col=5)
abline(h = sqrt(4/(dim(df)[1]-length(names(coef(model_final3))))),
        lty=3, lwd=1, col=6)
llegenda <- c("Cook d", names(coef(model_final3)), "DFBETA Cut-off", "Ch-H Cut-off")
# legend(locator(n=1), legend=llegenda,
#         col=1:length(llegenda), lty=c(1,2,2,2,3,3), lwd=c(3,2,2,2,1,1))
legend(x = 60, y = -0.4, legend=llegenda,
       col=1:length(llegenda), lty=c(1,2,2,2,3,3), lwd=c(3,2,2,2,1,1))
```

## 4. Relación entre DFBetas y Cook's Distance

r Copiar Editar

```
matplot(dfbetas(model_final3), type = "l", col=2:7, lwd=2)
lines(sqrt(cooks.distance(model_final3)), col=1, lwd=3)
```

### 💡 ¿Qué hace?

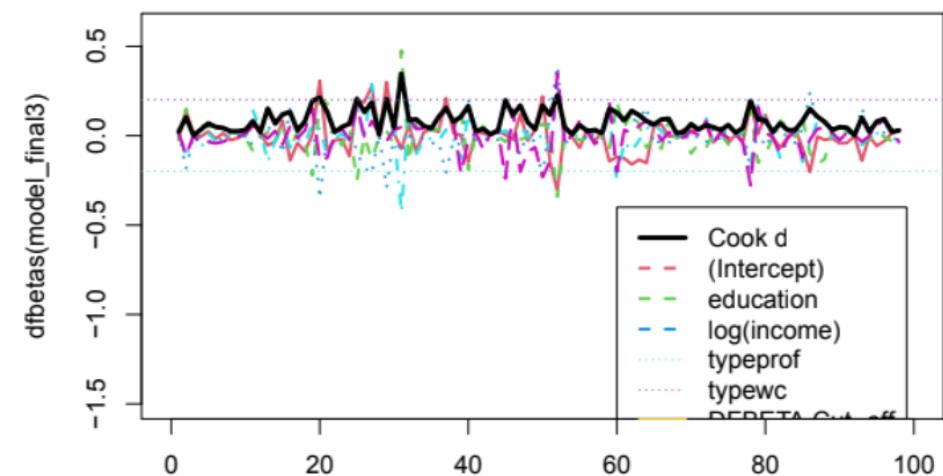
- Grafica DFBetas y `Cook's Distance` juntos para comparar sus efectos.
- Muestra si las observaciones con `DFBeta` alto también tienen `Cook's Distance` alto.

### ✓ Cómo interpretarlo:

- Si los valores altos de `Cook's Distance` coinciden con valores extremos de `DFBeta`, esos puntos afectan el modelo en gran medida.
- Si un punto tiene `Cook's Distance` alto pero `DFBeta` bajo, afecta la predicción pero no los coeficientes.
- Si hay muchos valores extremos, podría haber problemas con la selección de variables o valores atípicos.

### 💡 Acción recomendada:

- Eliminar observaciones con `Cook's Distance > 1` y `DFBeta` extremos y verificar si `summary(model_final3)` mejora.
- Si hay demasiados valores extremos, considerar una transformación de la variable dependiente (`log(prestige)`).



## DFFits

One can argue that if the final objective is rather predictive than explicative, one can use the difference in the fitted values rather than in the beta parameters. DFFits are related to Cook's distance and combine studentized residuals and leverages. Use `dffits(model)` in R.

```
head(dffits(model_final3))
```

```
##   gov.administrators    general.managers      accountants purchasing.officers
##       0.048341859        -0.232237528       0.003677053       -0.078037005
##   chemists                  physicists
##       0.154456285        0.108372435
```

### 💡 ¿Qué hace?

- Mide cuánto cambia la predicción del modelo si se elimina una observación.
- Es similar a `Cook's Distance`, pero enfocado en la predicción en lugar de los coeficientes.

### ✓ Cómo interpretarlo:

Observación	DFFits
gov.administrators	0.048
general.managers	-0.232
accountants	0.0036
purchasing.officers	-0.078
chemists	0.154
physicists	0.108

### 💡 Reglas generales:

- Si `DFFits > 2 * sqrt(p/n)`, el punto afecta mucho la predicción.
- Valores negativos indican que eliminar ese punto haría que la predicción aumente.

### 💡 Acción recomendada:

- Si `DFFits` es alto, revisar si eliminarlo mejora `summary(model_final3)`.
- Comparar con `Cook's Distance` para ver si el punto afecta tanto predicción como coeficientes.



## 4. Evaluar la multicolinealidad

Si las variables explicativas están correlacionadas, el modelo puede volverse inestable.

### 💡 Cómo se verifica:

- Se usa el **VIF (Factor de Inflación de la Varianza)**.
- $VIF < 5$ : No hay problema.
  - $VIF > 10$ : Fuerte multicolinealidad.

## 1. Hipótesis en Regresión Múltiple

### Hipótesis sobre los coeficientes del modelo

Para un modelo de regresión múltiple:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_p X_p + \epsilon$$

Debes contrastar si las variables explicativas  $X_i$  tienen un efecto significativo sobre la variable dependiente  $Y$ .

- **Hipótesis nula ( $H_0$ )**:  $\beta_i = 0$  (La variable  $X_i$  no tiene efecto sobre  $Y$ )
- **Hipótesis alternativa ( $H_1$ )**:  $\beta_i \neq 0$  (La variable  $X_i$  sí tiene efecto sobre  $Y$ )

### 💡 Cómo se verifica:

Se usa el **test t** para cada coeficiente:

$$t = \frac{\hat{\beta}_i}{SE(\hat{\beta}_i)}$$

Si el valor p es menor que un umbral (generalmente 0.05), rechazamos  $H_0$  y concluimos que  $X_i$  es significativo.

### Hipótesis global sobre la regresión

Esta prueba verifica si el conjunto de variables explicativas ayuda a predecir  $Y$ .

- **Hipótesis nula ( $H_0$ )**:  $\beta_1 = \beta_2 = \dots = \beta_p = 0$  (ninguna variable explicativa afecta  $Y$ )
- **Hipótesis alternativa ( $H_1$ )**: Al menos un  $\beta_i \neq 0$  (hay efecto de al menos una variable)

### 💡 Cómo se verifica:

Se usa el **test F** en la tabla de ANOVA. Si el valor p < 0.05, rechazamos  $H_0$  y concluimos que hay un efecto significativo en el modelo.

### Hipótesis sobre la multicolinealidad

Si las variables explicativas están muy correlacionadas, los coeficientes pueden volverse inestables.

- **Hipótesis nula ( $H_0$ )**: No hay multicolinealidad (las variables son independientes)
- **Hipótesis alternativa ( $H_1$ )**: Hay multicolinealidad

### 💡 Cómo se verifica:

Se usa el **factor de inflación de la varianza (VIF)**. Si  $VIF > 10$ , hay multicolinealidad fuerte.

## 2. Hipótesis en ANOVA

En ANOVA (Análisis de Varianza), queremos comparar medias entre varios grupos.

Para un ANOVA de un solo factor con  $k$  grupos:

- **Hipótesis nula ( $H_0$ ):**  $\mu_1 = \mu_2 = \dots = \mu_k$  (todas las medias son iguales)
- **Hipótesis alternativa ( $H_1$ ):** Al menos una media es diferente

💡 **Cómo se verifica:** Se usa el **test F**:

Si el valor  $p < 0.05$ , rechazamos  $H_0$  y concluimos que al menos un grupo es diferente.

### Hipótesis en ANOVA de dos factores

Para un modelo con dos factores  $A$  y  $B$ , tenemos tres hipótesis:

#### 1. Efecto del factor A:

- $H_0 : \mu_{A1} = \mu_{A2} = \dots$  (Las medias de  $A$  son iguales)
- $H_1 : \text{Al menos una media de } A \text{ es diferente}$

#### 2. Efecto del factor B:

- $H_0 : \mu_{B1} = \mu_{B2} = \dots$  (Las medias de  $B$  son iguales)
- $H_1 : \text{Al menos una media de } B \text{ es diferente}$

#### 3. Interacción entre A y B:

- $H_0 : \text{No hay interacción entre } A \text{ y } B$  (el efecto de un factor es el mismo en todos los niveles del otro)
- $H_1 : \text{Sí hay interacción}$

💡 **Cómo se verifica:** Se usa un **ANOVA factorial**, con valores  $p$  para cada factor y su interacción.

## 3. Hipótesis en ANCOVA

El **Análisis de Covarianza (ANCOVA)** combina ANOVA y regresión.

- **Hipótesis nula ( $H_0$ ):** No hay diferencia entre los grupos después de ajustar por la covariante.
- **Hipótesis alternativa ( $H_1$ ):** Sí hay diferencias entre los grupos, incluso después de controlar la covariante.

💡 **Cómo se verifica:** Se usa un **test F** en la tabla de ANCOVA.

## Cómo verificar si un modelo es válido

Después de ajustar un modelo, es fundamental validarlo. Aquí están los pasos clave:

### 1. Verificar supuestos del modelo

Para que los resultados sean confiables, los modelos deben cumplir ciertos supuestos:

- **Linealidad:** La relación entre  $X$  y  $Y$  debe ser lineal.
- **Independencia:** Los errores deben ser independientes.
- **Normalidad:** Los errores deben seguir una distribución normal.
- **Homocedasticidad:** La varianza de los errores debe ser constante.

💡 **Cómo se verifica:**

- **Gráfico de residuos vs predicciones:** Debe verse un patrón aleatorio.
- **Prueba de normalidad de Shapiro-Wilk:** Si  $p > 0.05$ , los errores son normales.
- **Prueba de Breusch-Pagan:** Si  $p > 0.05$ , no hay heterocedasticidad.

### 2. Evaluar el ajuste del modelo

- **$R^2$  y  $R^2$  ajustado:** Miden qué porcentaje de la variabilidad en  $Y$  explica el modelo.
  - **$R^2$  alto:** Buen ajuste, pero cuidado con sobreajuste.
  - **$R^2$  ajustado:** Penaliza modelos con muchas variables innecesarias.
- **Error estándar de la estimación (SEE):** Mide el error promedio en la predicción.

💡 **Cómo se verifica:** Se usa la tabla de regresión, observando  $R^2$  y el SEE.

### 3. Identificar valores atípicos e influyentes

- **Outliers:** Observaciones que no siguen el patrón general.
- **Puntos influyentes:** Datos que cambian drásticamente el modelo.

💡 **Cómo se verifica:**

- **Gráfico de residuos:** Busca puntos alejados del patrón.
- **Distancia de Cook:** Si  $D_i > 1$ , el punto es influyente.