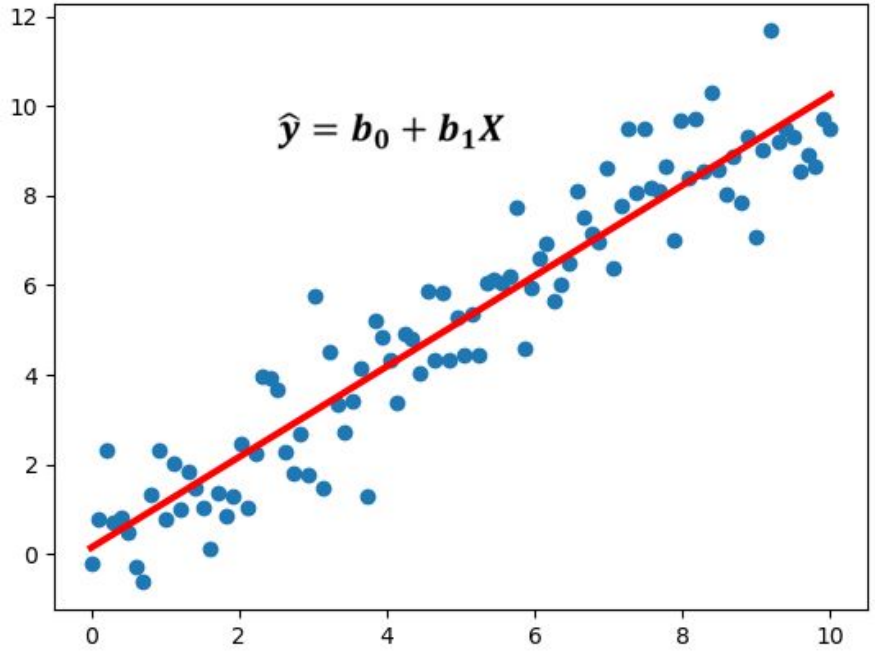
Modelos de regresión

Estadística para el análisis político 2

¿Qué son los modelos de regresión? (I)

Definición general: un modelo matemático que busca determinar la relación entre una variable dependiente (Y), con respecto a una o varias variables independientes (X).

Nos ayuda a *explicar* o *predecir* el comportamiento de una variable dependiente en términos de causalidad. Y para ello, desde la ciencia política, planteamos un marco teórico, que se pone a prueba con el modelo de regresión.



¿Qué son los modelos de regresión? (II)

“Correlación no es igual a causalidad” o la importancia de *controlar* la relación entre variables e identificar explicaciones alternativas

En un estudio, se encontró que las personas más altas de un salón de clase también presentaron las notas más altas. ¿Esto implica que la altura determina la desempeño académico de los estudiantes? Tal vez no sea la altura, sino la edad de las personas del salón. Y, para demostrar que la altura impacta en las notas, habría que explorar esta relación *manteniendo la edad de todas las personas como constante.*

De esto consta el control de variables, de mantener constantes (controlar) posibles explicaciones alternas o elementos exógenos que podrían afectar nuestra fenómeno de estudio.

Modelo de regresión lineal (I)

$$Y = B0 + B1 * X1 + \cdots Bn * Xn + e$$

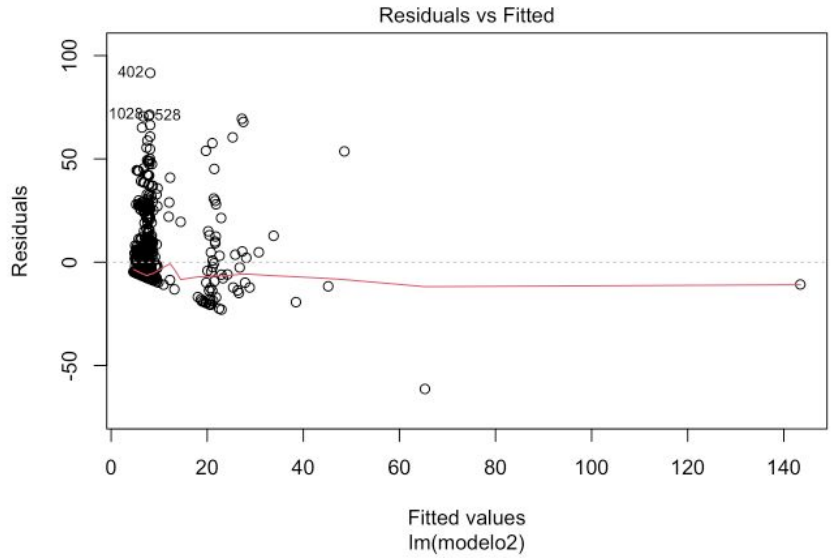
- Y: la variable dependiente
- B0: el intercepto (por donde pasa la recta, el valor que toma Y cuando X o las Xs es igual a 0, etc.)
- B1: pendiente de la variable o el impacto que se tiene sobre la variable dependiente el aumento de un punto en la independiente.
- E: error residual o todo aquello que no puede ser explicar por el modelo

*La variable dependiente es una numérica continua no acotada. Pero en la práctica hay más flexibilidad con este tema.

Modelo de regresión lineal (II): linealidad

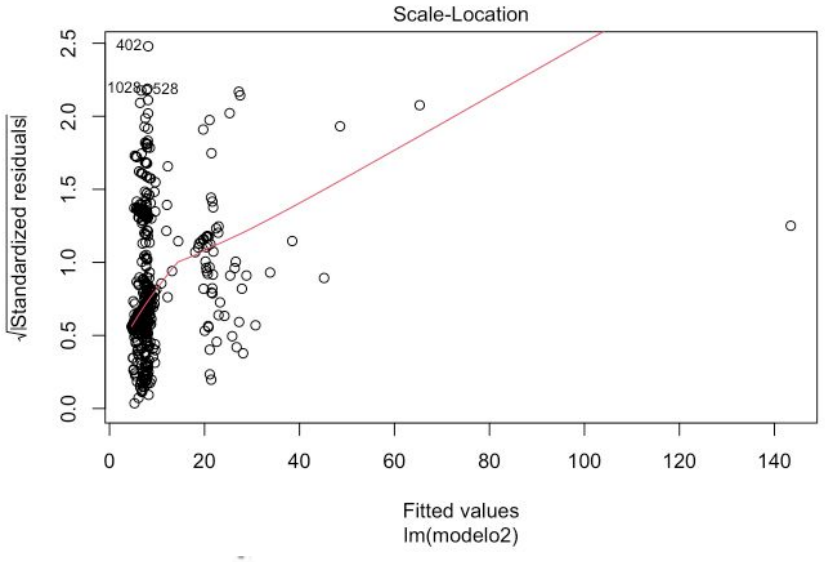
Se asume una *relación lineal entre la Y y las Xs*. Esperamos una línea roja que tienda a ser horizontal y un promedio de residuos (diferencia entre el valor esperado y observado de la data) cercano a 0. Más pequeño significa que está más pegado a la recta.

```
# linea roja debe tender a horizontal
plot(reg2, 1)
```



Modelo de regresión lineal (III): homocedasticidad

```
# linea roja debe tender a horizontal
plot(reg2, 3)
```



Esperamos que *las varianzas (dispersión) de los errores de estimación son constantes en el modelo de regresión*. Es decir, no existe un patrón o sesgo en en estos errores (un modelo heterocedástico).

Queremos una línea que sea horizontal y un valor de la prueba de Breusch-Pagan mayor a 0.05

```
library(lmtest)
```

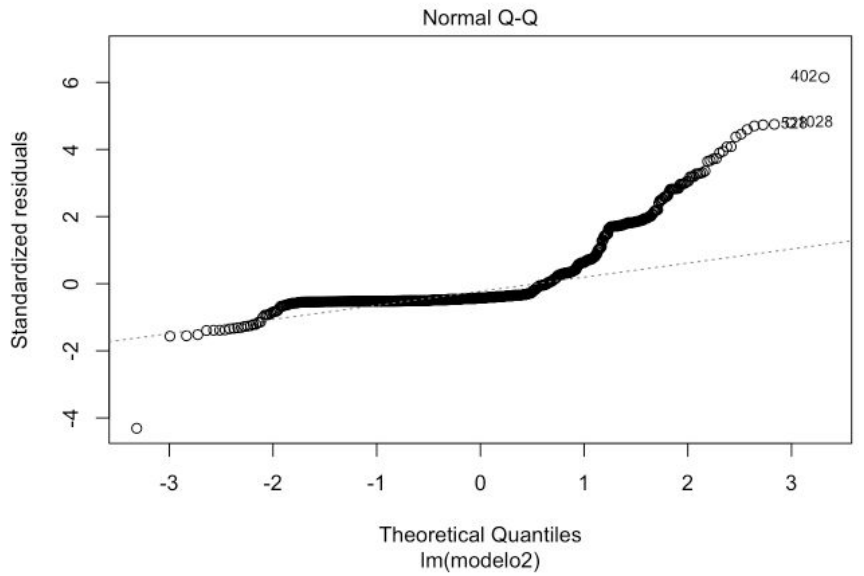
```
# null: modelo homocedastico
bptest(reg3)
```

Modelo de regresión lineal (IV): normalidad de residuos

Los residuos (la distancia entre el valor esperado y el observado de la variable) deben distribuirse de manera normal.

Esperamos que los puntos se acerquen lo más posible a la línea del gráfico y un p-valor de la prueba de Shapiro-Wilk mayor a 0.05.

```
# puntos cerca a la diagonal
plot(reg2, 2)
```




```
shapiro.test(reg3$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  reg3$residuals
## W = 0.9989, p-value = 0.9776
```

Modelo de regresión lineal (V): no multicolinealidad

No queremos tener una fuerte correlación entre las variables independientes. En la práctica, estaríamos midiendo lo mismo dos veces.

Queremos una el *Factor de Inflación de la Varianza,* o prueba VIF, con valores menores a 5.

```
library(DescTools)
VIF(reg2) # > 5 es problematico
```

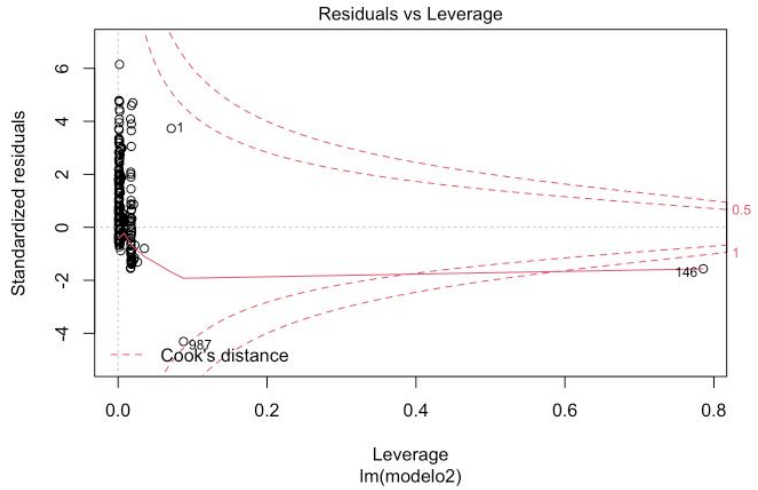
```
##          pctopo   consejocomunal poblacioncienmil
##        1.000465         1.035584         1.035113
```

Modelo de regresión lineal (VI): cuidado con los valores influyentes

Casos en nuestra data que no siguen el patrón general del resto de casos.

El gráfico para identificar influyentes es *Residual vs. Leverage*. En este gráfico, los patrones no son relevantes. Debemos buscar los casos identificados en el gráfico por encima de 1.

```
plot(reg2, 5)
```

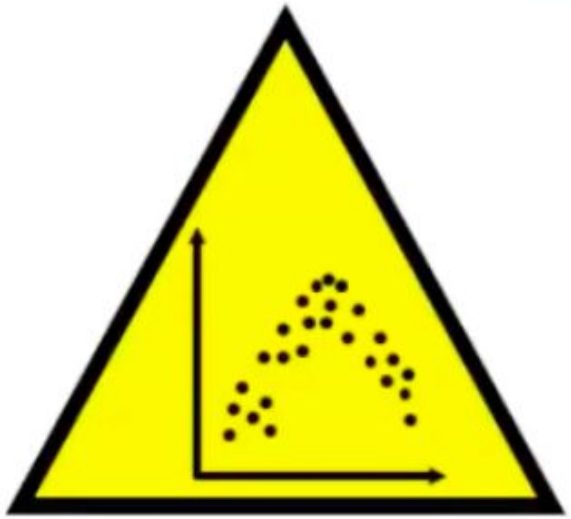


Asi podemos recuperar los casos influyentes:

```
checkReg2=as.data.frame(influence.measures(reg2)$is.inf)
head(checkReg2)
```

Normalmente le prestamos atencion al indice de Cook y a los valores predecidos (los *hat* values):

```
checkReg2[checkReg2$cook.d & checkReg2$hat,]
```



CUIDADO

Relaciones no lineales

CUIDADO

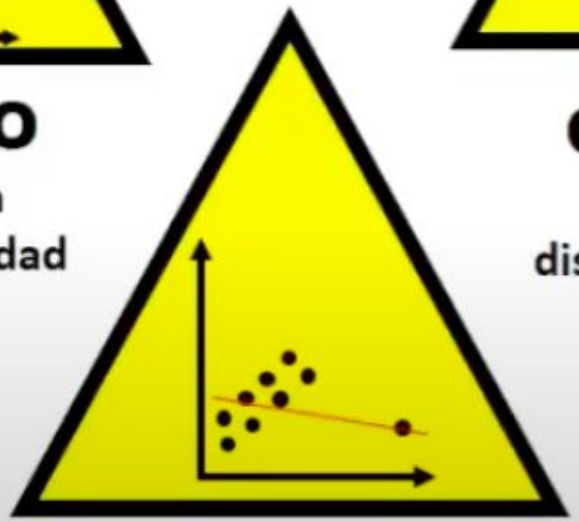
Residuos con Heterocedasticidad

CUIDADO

Residuos sin distribución normal

CUIDADO

Multicolinealidad



CUIDADO

Presencia de valores influyentes

Modelo de regresión lineal (VII): interpretación

El aumento de un punto en la variable independiente eleva (o disminuye) en (el valor del coeficiente) la variable dependiente. Por ejemplo, "el aumento de un año de edad aumenta el salario de la persona en 158.2 soles, manteniendo todos los demás regresores constantes". Ojo con las escalas de medición de las variables.

¿Cómo se comparan modelos? Miramos el *valor de significancia de nuestra prueba F.* Si es menor a 0.05, significa que nuestro modelo es válido. Luego vemos el R2, como medida de ajuste o bondad del modelo de regresión; es decir, *qué tanta varianza es explicada*. Nos quedamos con el modelo que tenga un mayor nivel de explicación de varianza.

```
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -32736  -3965  -1214   2458  46474
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)     -1.027e+04  2.960e+03  -3.469 0.000571 ***
## salario_inicial  1.927e+00  4.437e-02  43.435  < 2e-16 ***
## antiguedad       1.732e+02  3.468e+01   4.995 8.32e-07 ***
## experiencia     -2.251e+01  3.339e+00  -6.742 4.59e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7586 on 470 degrees of freedom
## Multiple R-squared:  0.8039, Adjusted R-squared:  0.8026
## F-statistic: 642.2 on 3 and 470 DF,  p-value: < 2.2e-16
```

Regresión logística binomial (I)

- https://peopleanalytics-regression-book.org/bin-log-reg.html

Objetivo: modelar la probabilidad de ocurrencia de un evento en función a ciertas variables independientes. No podemos plantearlo en términos lineales, ya que nuestra variable dependiente tiene solo dos categorías, pero podemos expresar la relación asimétrica (de dependencia) *en términos lineales*.

1 □ Presencia/ocurrencia del evento
0 □ Ausencia del evento

Conceptos importantes:

Probabilidad: Grado de incertidumbre de que un evento pueda ocurrir. Va de 0 a 1. Es el número de veces que ocurre un evento divido por la cantidad total de eventos posibles.

Odds: Es la probabilidad de que suceda un evento dividido por la probabilidad de que no suceda. Oscilan entre 0 e infinito y se pueden calcular para la ocurrencia del evento como para la no ocurrencia del evento. Ejemplo: la probabilidad de ganar una apuesta en un partido de futbol es 1.5 veces más probable que perder.

Odds ratio: La razón entre dos *odds.* Permite comparar los odds de un evento en dos grupos. Va de 0 a infinito. Los modelos de regresión logística están basados en probabilidades entre dos variables.

Regresión logística binomial (II)

- Como saben, lo que se está modelando es a un *lograritmo del odds.* ¿Qué significa eso? Es difícil de interpretar, pero el punto es que necesitamos valores numéricos que nos den una referencia del impacto que se tiene sobre nuestra variable dependiente.

- Como los *odds* son difíciles de interpretar, sacar un exponencial nos va a decir muy poco. Miremos los *efectos marginales*. Estos valores expresan cómo cambia la probabilidad de ocurrencia de un evento (variable dependiente) frente a un cambio en la(s) variable(s) independiente(s).

- En promedio, el aumento del salario de la persona en un dólar disminuye su probabilidad de apoyar la democracia en .15 (o 15%).

```
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -0.21104    0.07201  -2.931  0.00338 **
## sexmale     -0.24934    0.10845  -2.299  0.02150 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1933.5  on 1420  degrees of freedom
## Residual deviance: 1928.2  on 1419  degrees of freedom
## AIC: 1932.2
##
## Number of Fisher Scoring iterations: 4
```

```r
# interpracion usando marginal effects:
library(margins)
#
(model = margins(rlog3))
```

```
## Average marginal effects
```

```
## glm(formula = volunteer ~ ., family = binomial, data = vars3)
```

```
##  neuroticism extraversion sexfemale
##      0.00152      0.01585   0.05623
```

Regresión Poisson

$$Y = EXP(B0 + B1 * X1 + \cdots Bn * Xn) + e$$

$$\log(Y) = B0 + B1 * X1 + \cdots Bn * Xn$$

Usamos la regresión Poisson con una variable dependiente no negativa y entera que represente *conteos* en espacio o tiempo. Una variable cuantitativa discreta. Por ejemplo: número de veces que una persona ha sido acosada, número de veces que una persona ha sido arrestada, número de veces

Objetivo: predecir o explicar fenómenos que ocurren en un intervalo de tiempo específico.

Requisitos:

- Independencia de las observaciones: los casos no se encuentran relacionados entre sí.
- La media es igual a su varianza.
- Linealidad: el logaritmo de la media de los datos (log(λ)) debe ser una función lineal de los datos.

- Veamos si se cumple el requisito de igualdad de medias y de varianzas. Si no se cumple, utilizamos un modelo *quasipoisson* (subdispersión y sobredispersión) o una *binomial negativa* (sobredispersión).