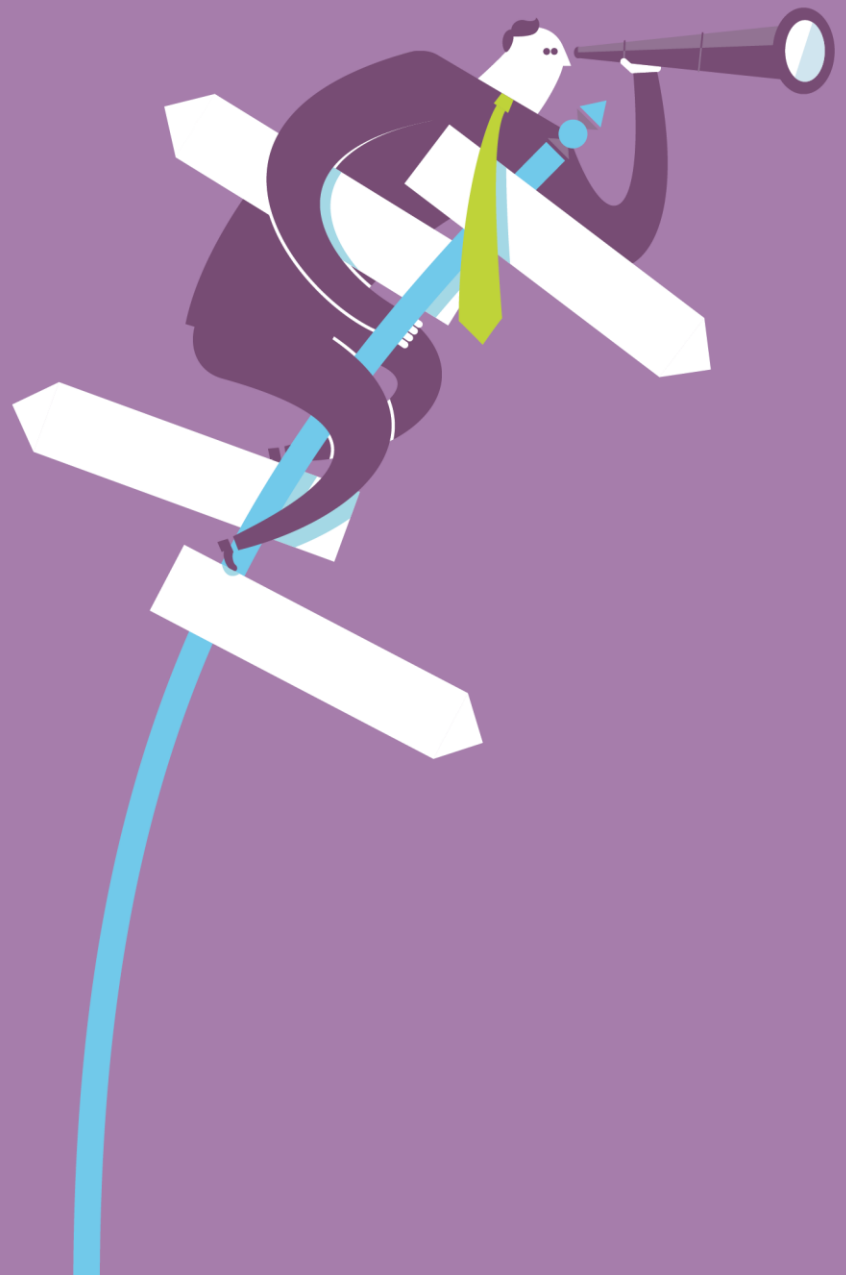


Tema 13: Inferencias en modelos de regresión múltiple y predicción.

Modelo de regresión múltiple: ejemplo.

Se realizó un experimento para determinar si el peso de un animal puede predecirse después de un periodo dado, con base en el peso inicial del animal y en la cantidad de alimento consumida por este. Se registraron los siguientes datos, medidos en kilogramos:

$$\hat{Y} = b_0 + b_1 X_1 + b_2 X_2$$

Peso Final Y	Peso inicial X_1	Alimentos consumidos X_2
95	42	272
77	33	226
80	33	259
100	45	292
97	39	311
70	36	183
50	32	173
80	41	236
92	40	230
94	38	235

Una vez estimado el modelo de regresión, se obtienen los siguientes resultados:

Estadísticas de la regresión	
Coeficiente de correlación múltiple	0.9012
Coeficiente de determinación R^2	0.8121
R^2 ajustado	0.7584
Error típico	7.5797
Observaciones	10

	Coeficientes	Error típico	Estadístico t	Probabilidad	Inferior 95%	Superior 95%
Intersección	-22.1377	22.2507	-0.9949	0.3529	-74.7523	30.4769
Variable X 1	1.4420	0.7297	1.9760	0.0887	-0.2836	3.1675
Variable X 2	0.2110	0.0724	2.9152	0.0225	0.0398	0.3821

De esta regresión puede apreciarse que el modelo de regresión estimado es:

$$\hat{Y} = -22.1377 + 1.4420\ X_1 + 0.2110X_2$$

La relación entre Y (peso final) y X_1 (peso inicial) se describe por b_1= 1.4420. De este número puede decirse que en este modelo, por cada unidad adicional de peso inicial, el peso final se incrementa en 1.4420 en promedio, manteniendo constante la X_2 (alimentos consumidos).

Supóngase que se desea predecir el peso promedio final cuando X1(el peso inicial), es de 35 kilogramos y X2 (los alimentos consumidos), es igual a 280 kilogramos.

Si se utiliza la ecuación de regresión múltiple, obtenida anteriormente:

$$\hat{Y} = -22.1377 + 1.4420\ X_1 + 0.2110X_2$$

Con $X_1 = 35$ y $X_2=280$, se tiene:

$$\hat{Y} = -22.1377 + 1.4420\ (35) + 0.2110(280)$$

Y así:

$$\hat{Y} = 87.55$$

Sin embargo la regresión no termina allí …

Recordemos que: la evaluación del modelo se puede hacer en tres formas:

Analizando estas tres vertientes, junto con el análisis de los supuestos de los errores (residuos), podremos tener un buen modelo de regresión: homocedasticidad, media cero, distribución normal, no multicolinealidad.

Error estándar de la estimación

En regresión múltiple, el error estándar de la estimación se define como sigue:

$$S_\varepsilon = \sqrt{\frac{SCE}{n-k-1}} = \sqrt{CME}$$

En donde:

n = número de observaciones

k = número de variables independientes en la función de regresión

SCE = suma de cuadrados del error

CME = cuadrado medio del error

El número de observaciones es n=10 y el error estándar de la estimación se determina con:

$$S_{\varepsilon} = \sqrt{\frac{402.1607}{10-2-1}} = \sqrt{\frac{402.1607}{7}} = \sqrt{57.4515} = 7.5797$$

En este caso, el error estándar del modelo de regresión es de 7.58.

Coeficiente de determinación

El coeficiente de determinación es dado por:

$$R^2 = \frac{\text{Suma de cuadrados de regresión}}{\text{Suma de cuadrados totales}}$$

Y representa la razón de la variación de la respuesta Y explicada por su relación con las X. Para el ejemplo anterior se tiene que el coeficiente de determinación es:

$$R^2 = \frac{\text{Suma de cuadrados de regresión}}{\text{Suma de cuadrados totales}} = \frac{1738.3393}{2140.5000} = 0.8121$$

En el contexto de este problema podemos decir que el 81.21% de la variación en el peso final se explica por X_1 (peso inicial) y X_2 (alimentos consumidos). En la práctica, $0 \leq R^2 \leq 1$, y el valor de R^2 debe interpretarse en relación con los extremos, 0 y 1.

Significancia de los coeficientes de regresión

	Coeficientes	Error típico	Estadístico t	Probabilidad	Inferior 95%	Superior 95%
Intercepción	-22.1377	22.2507	-0.9949	0.3529	-74.7523	30.4769
Variable X 1	1.4420	0.7297	1.9760	0.0887	-0.2836	3.1675
Variable X 2	0.2110	0.0724	2.9152	0.0225	0.0398	0.3821

Para evaluar la significancia de los coeficientes de la regresión, se hacen algunas pruebas de hipótesis respecto a los coeficientes.

Pruebas de hipótesis.

$H_0 : \beta_1 = \beta_2 = \ldots \beta_k = 0$ (Las variables independientes no afectan a Y)

En oposición a:

$H_a : \beta_i \neq 0$ (Al menos una variable X afecta a Y)

Para **evaluar la hipótesis** se hace uso del estadístico de prueba:

$$t_{calculada} = \frac{b_i - \beta_i}{S_{b_i}}$$

Regla de decisión

$$t_{calculada} = \frac{b_i - \beta_i}{S_{b_i}} = \frac{1.4420 - 0}{0.7297} = 1.9760$$

Rechazar H_0 si $|t_{calculada}| = 1.9760$ es mayor que $t_{teórica}$.

En donde:

$$t_{teórica} = t_{\alpha/2}(n-k-1) = t_{0.05/2}(7) = t_{0.025}(7) = 2.365$$

En donde el valor de $t_{teórica}$ se obtiene de la tabla de distribución de t.

Puesto que $|t_{calculada}| = 1.9760$ es *menor* que $t_{teórica} = 2.365$, **no** se rechaza H_0. (Esto es, **no** existe evidencia de que el peso inicial X_1 afecte el peso final Y, o bien, la variable peso inicial X_1 no tienen efecto significativo en el peso final Y).

Intervalo de confianza

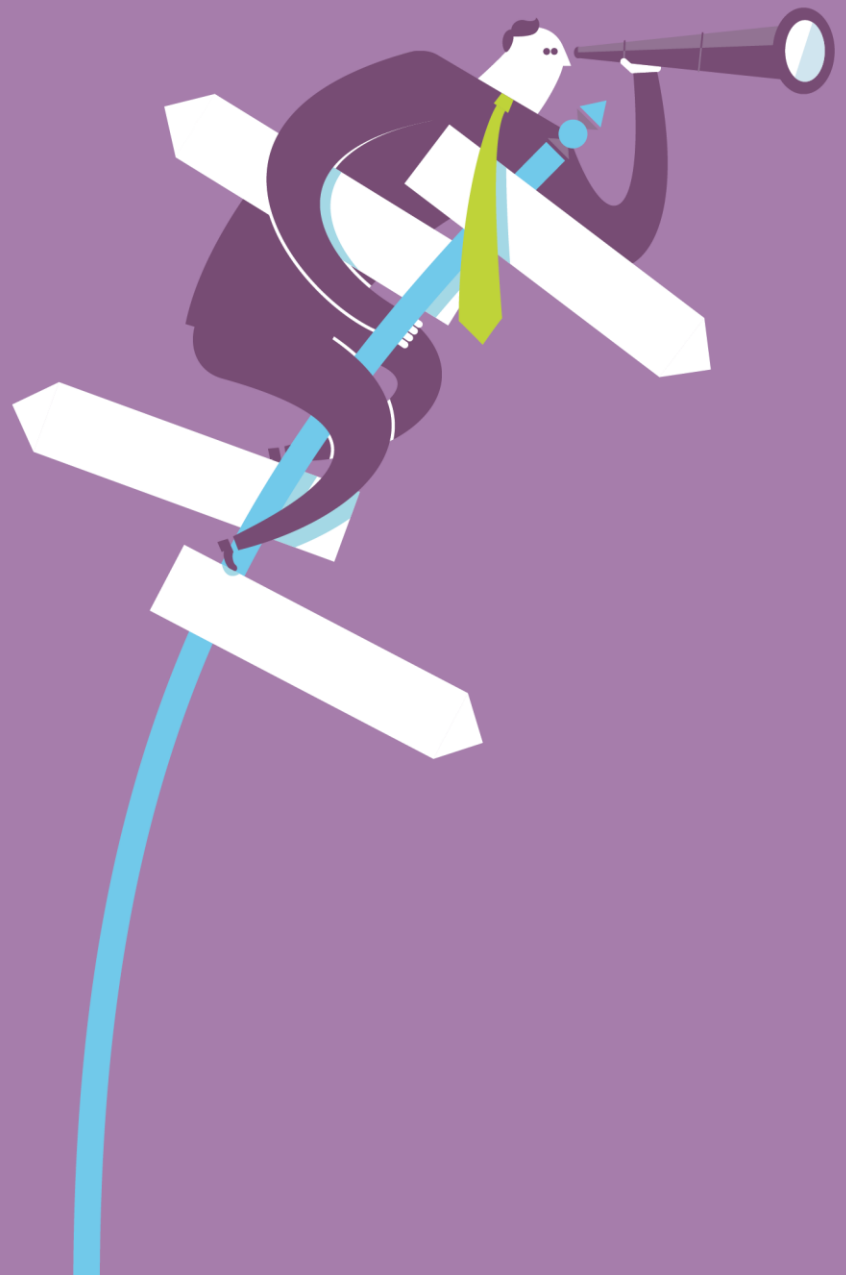
En el análisis de regresión múltiple, un intervalo de confianza para una pendiente de la población se puede estimar a partir de la siguiente expresión:

$$b_1 \pm t_{\alpha/2}(n - k - 1)\, S_{b_i}$$

Para el presente ejemplo:

	Coeficientes	Error típico	Estadístico t	Probabilidad	Inferior 95%	Superior 95%
Intercepción	-22.1377	22.2507	-0.9949	0.3529	-74.7523	30.4769
Variable X 1	1.4420	0.7297	1.9760	0.0887	-0.2836	3.1675
Variable X 2	0.2110	0.0724	2.9152	0.0225	0.0398	0.3821

Entonces, con un 95% de confianza, se tiene que el verdadero valor β_1 se encuentra en el intervalo (-0.2837, 3.1677). Desde el punto de vista de la prueba de hipótesis, puesto que este intervalo de confianza contiene al cero, se concluye que el coeficiente de correlación β_1 no tiene efecto significativo.



Tema 14: Transformaciones de modelos de regresión no lineales.

Transformaciones de modelos de regresión no lineales

Aunque el **modelo de regresión lineal simple** supone una línea recta entre Y y X, en general, un modelo lineal de regresión se refiere al grado u orden en las β (por ejemplo, β^2 no está presente), las variables de predicción (las X) pueden tomar varias formas y la metodología de regresión lineal sigue siendo apropiada.

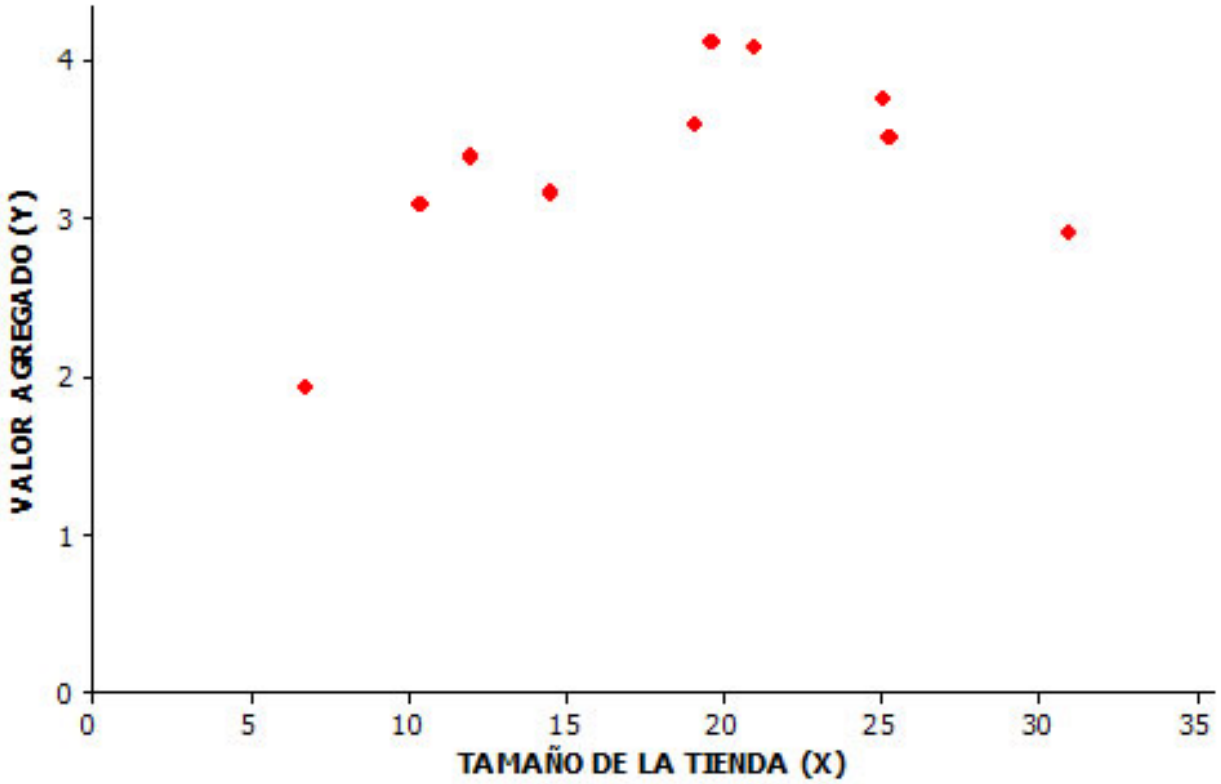
Los modelos de regresión pueden ser usados para modelar relaciones complejas entre Y y X (o muchas X) o para modelar una relación de línea recta entre la variable Y y alguna función (transformación) de X.

Ejemplo

En un estudio de variables que afectan la productividad en el negocio de abarrotes al menudeo, W. S. Good usa el valor agregado por hora de trabajo para medirla. Él define al valor agregado como el “excedente [dinero generado por el negocio] disponible para pagar mano de obra, muebles accesorios y equipo”. Los datos de acuerdo con la relación el valor agregado por hora de trabajo Y y el tamaño X de la tienda de abarrotes descrita en el artículo de Good para diez tiendas de abarrotes ficticias se muestran enseguida.

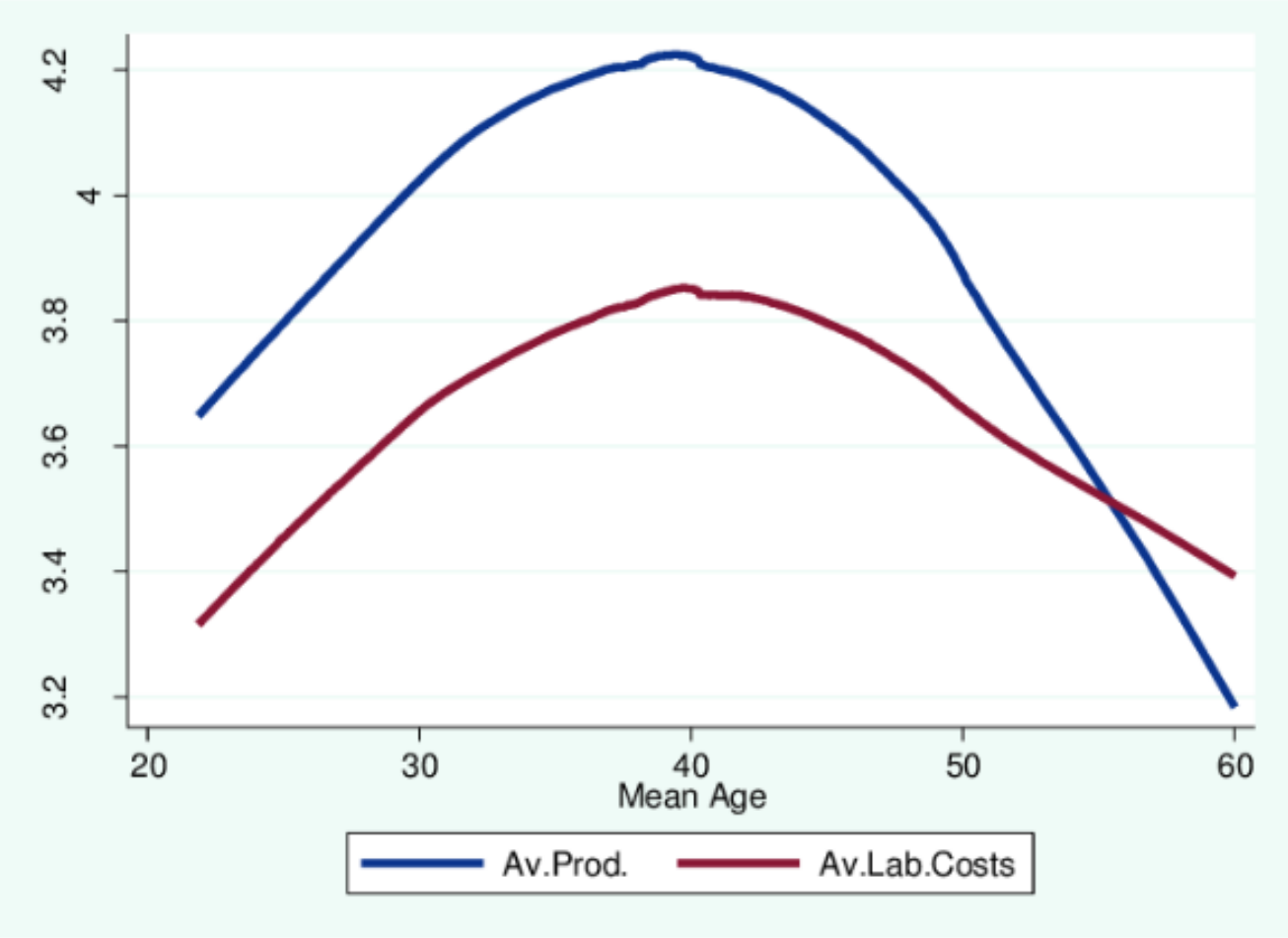
Datos en relación con el tamaño de tienda y el valor agregado		
Tienda	Valor agregado por hora de trabajo Y	Tamaño de la tienda (miles de pies cuadrados) X
1	4.08	21.0
2	3.40	12.0
3	3.51	25.2
4	3.09	10.4
5	2.92	30.9
6	1.94	6.8
7	4.11	19.6
8	3.16	14.5
9	3.75	25.0
10	3.60	19.1

Se puede investigar la relación entre Y y X por inspección de la gráfica de los datos. La gráfica hace pensar que la productividad Y se incrementa con el tamaño de la tienda X hasta que se alcanza un punto óptimo.



Arriba de cierto tamaño, la productividad tiende a disminuir. La relación parece curvilínea y un modelo cuadrático podría ser apropiado.

Otro ejemplo: productividad y costos laborales vs edad



A medida que aumenta la edad, la productividad laboral tiende a ser menor. Si estamos analizando la productividad laboral, lo más adecuado es utilizar un modelo cuadrático.

Con el primer ejemplo, se usó Minitab para ajustar un modelo cuadrático a los datos y graficar la curva de predicción cuadrática junto con los puntos de datos graficados. Se prueba la hipótesis para ver la idoneidad del modelo a ajustar.

La ecuación de regresión es:

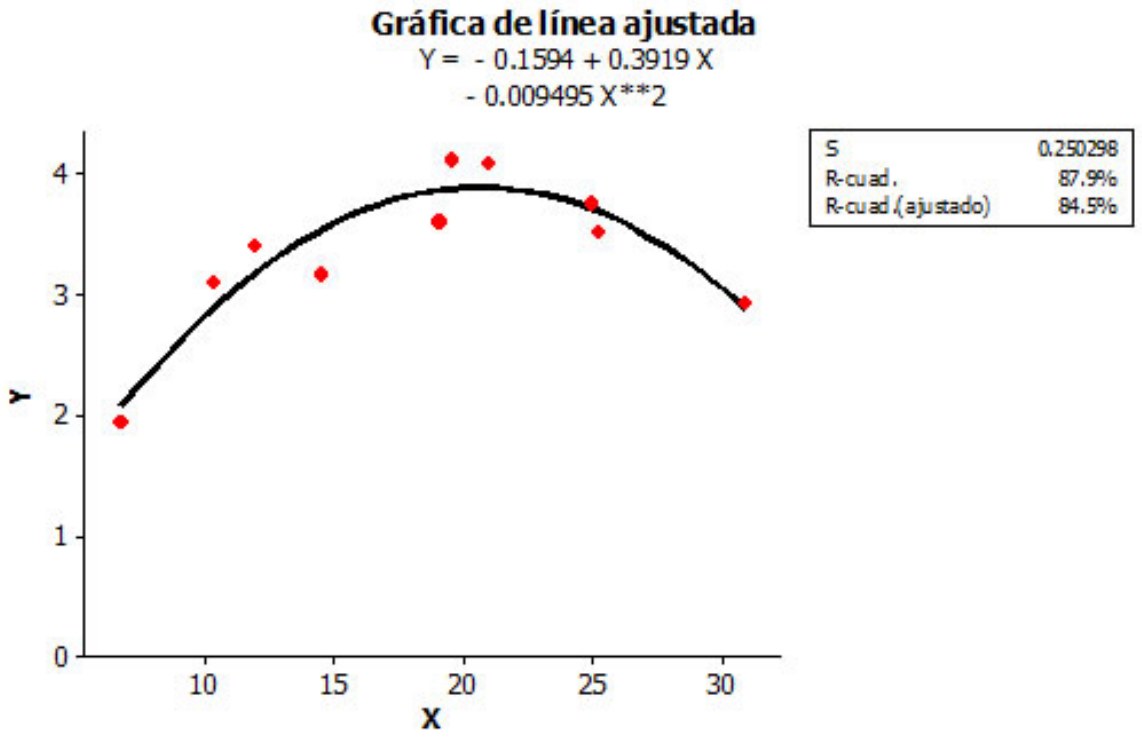
Y = - 0.159 + 0.392 X - 0.00949 X cuadrada

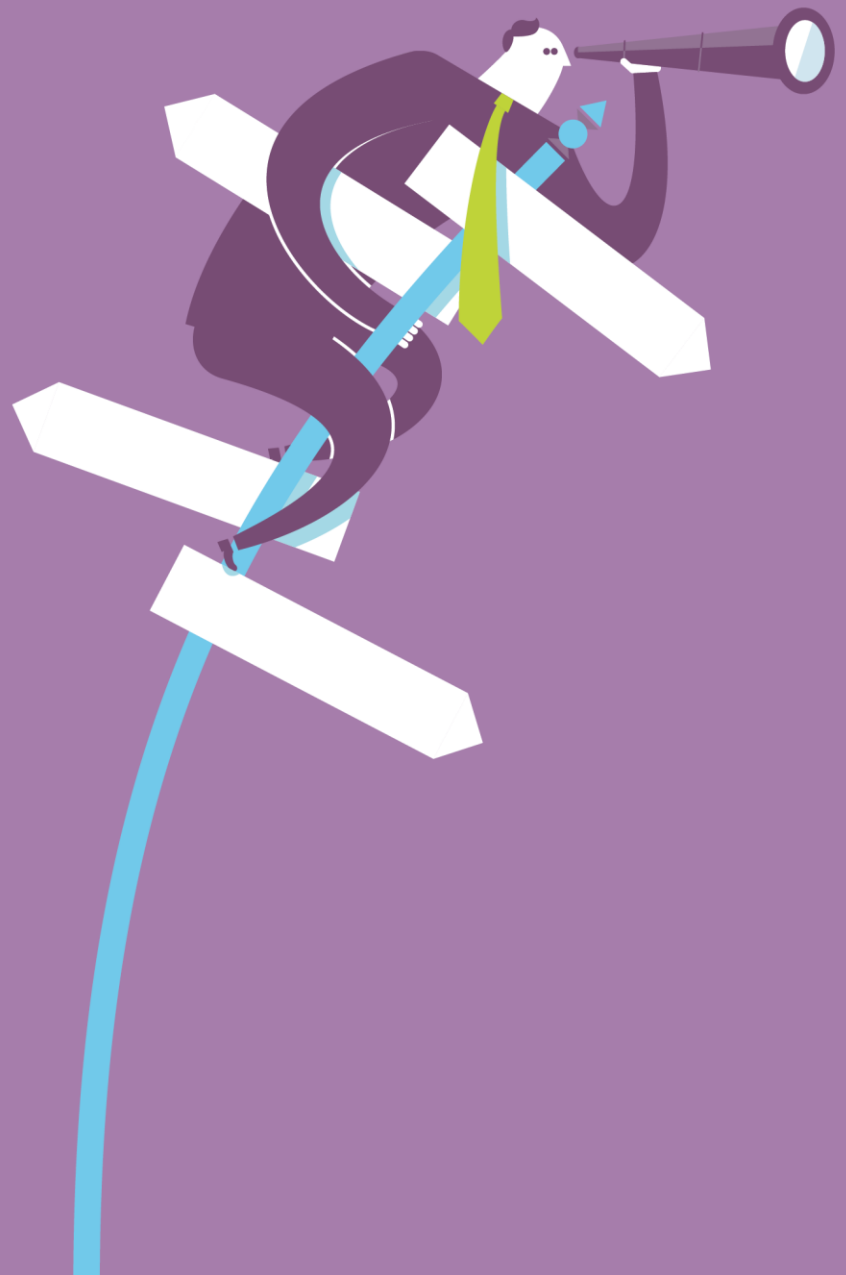
Predictor	Coef	Coef. de EE	T	P
Constante	-0.1594	0.5006	-0.32	0.760
X	0.39193	0.05801	6.76	0.000
X_cuad	-0.009495	0.001535	-6.19	0.000

De la impresión mostrada anteriormente, puede observarse que la ecuación de regresión es:

$$\hat{Y} = -0.1594 + 0.3919\ X - 0.009495\ X^2$$

La gráfica de la ecuación se presenta enseguida:





Tema 15: Multicolinealidad, diagnósticos de regresión y análisis de residuales.

Multicolinealidad

Según Hanke, y Wichern (2010) la *multicolinealidad* es la situación en la cual las variables independientes de una ecuación de regresión múltiple están sumamente intercorrelacionadas. Es decir, existe una relación lineal entre dos o más variables independientes.

En palabras más simples, la multicolinealidad se refiere a que dos o más variables pueden contener información repetida.

¿Cómo detectar la multicolinealidad?

El impacto que tiene la multicolinealidad en las variables se puede medir mediante un indicador que se conoce como **factor de inflación de la varianza** o VIF por sus siglas en inglés (variance inflation factor)

$$VIF_i = \frac{1}{1 - R_i^2}$$

En donde R_i^2 es el coeficiente de determinación de la regresión, indica qué tanto la variable independiente está relacionada con el resto de variables independientes.

Si el valor de VIF es cercano a 1 se puede interpretar que la multicolinealidad no es un gran problema para la variable en cuestión.

Caso contrario, un valor de VIF alto implica que existe información redundante en las variables independientes y que, por lo tanto, existen problemas de multicolinealidad.

Tratamiento de la multicolinealidad

La multicolinealidad es un problema de la muestra y, por tanto, no tiene solución simple, ya que estamos pidiendo a los datos más información de la que contienen.

Las dos únicas soluciones son:

1. Eliminar regresores, reduciendo el número de parametros a estimar

2. Cambiar por otra variable que explique mejor la relación con la variable dependiente.

Ejemplo en Stata

```
. reg  y x1 x2, beta

      Source |       SS       df       MS              Number of obs =     100
-------------+------------------------------           F(  2,    97) =    3.24
       Model |  55.7446181     2   27.872309            Prob > F      =  0.0436
    Residual |  835.255433    97  8.61088075            R-squared     =  0.0626
-------------+------------------------------           Adj R-squared =  0.0432
       Total |  891.000051    99  9.00000051            Root MSE      =  2.9344

------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      t    P>|t|                     Beta
-------------+----------------------------------------------------------------
          x1 |   .0153846   .1889008     0.08   0.935                 .025641
          x2 |   .1353847   .1889008     0.72   0.475                .2256411
       _cons |   10.49231   .6655404    15.77   0.000                       .
------------------------------------------------------------------------------
```

$$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$$

$$y = 10.49 + 0.015 X_1 + 0.135 X_2 + \varepsilon$$

Análisis de correlación entre las variables

```
. corr y x1 x2, means

(obs=100)

    Variable |       Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------
           y |         12            3    4.899272   18.91652
          x1 |         10            5   -1.098596   23.10749
          x2 |         10            5   -.0284863   23.72392


             |        y       x1       x2
-------------+---------------------------
           y |   1.0000
          x1 |   0.2400   1.0000
          x2 |   0.2500   0.9500   1.0000
```

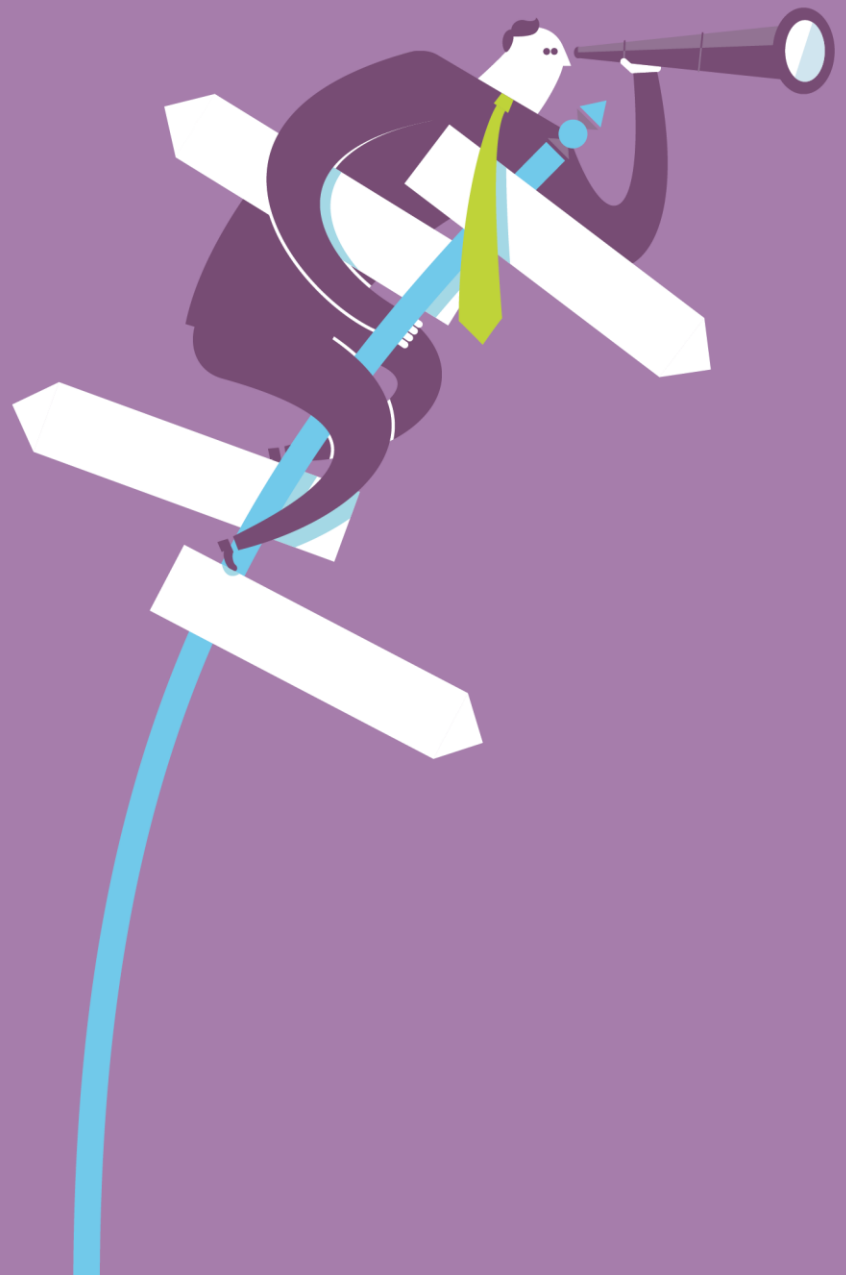
X1 y X2 están fuertemente correlacionadas (r=0.95).

Estos son indicadores de que la multicolinealidad podría ser un problema en estos datos.

Prueba VIF

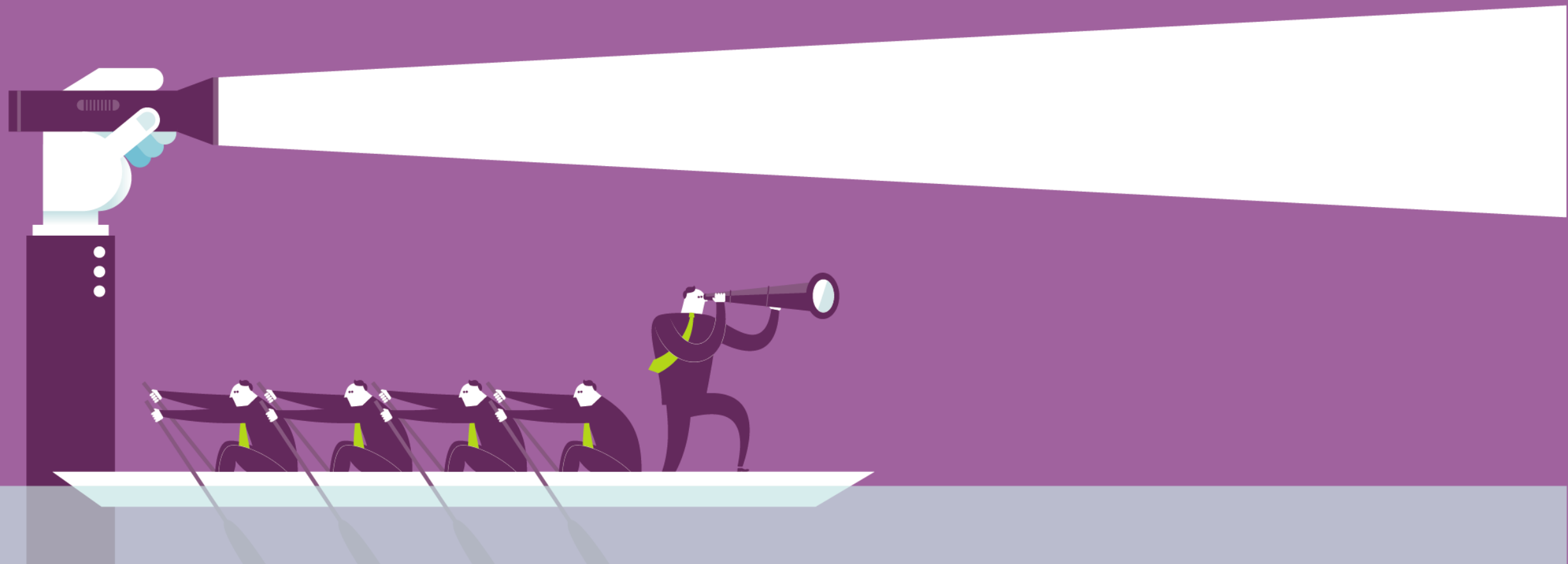
```
. vif

    Variable |       VIF       1/VIF
-------------+----------------------
          x1 |     10.26    0.097500
          x2 |     10.26    0.097500
-------------+----------------------
    Mean VIF |     10.26
```

X1 y X2 están muy altamente correlacionados (r = .95) y los VIF superan nuestra "regla de tener valor VIF cercano a 1" en este caso es de 10. Por lo tanto, en este modelo hay presencia de multicolinealidad.



Tema extra: Modelo de regresión logística (modelos logit).

Regresión logística



- Hasta ahora, cuando se hablaba de regresión, la variable dependiente (Y) era **continua**: precio, número de ventas, producción, gastos, salarios.
- **¿Qué pasa si ahora Y es binaria?** si se acepta o no una solicitud para una hipoteca; si el crédito de un cliente fue aprobado o no; si el cliente dejo nuestros servicios o sigue afiliado.
- Supongamos que la variable dependiente **Y representa la ocurrencia o no de un suceso,** por ejemplo:
  - El crédito de un cliente fue aprobado o no.
  - Las ventas de un promotor se concluyó o no.
  - Un cliente sigue con la empresa o abandono.

» Podemos decir que la variable dependiente **Y toma valor 1 si ocurre el suceso, y valor 0 si no ocurre el suceso.**

» La regresión logística se usa para **predecir una clase**, es decir, una **probabilidad**.

» La regresión logística puede predecir un resultado **binario**.


YES
NO

- La regresión logística se usa para **predecir una clase**, es decir, una **probabilidad**.
- La regresión logística puede predecir un resultado **binario**.
- Como ejemplo, **podría ser predecir si una persona compra un seguro o no**; en función de muchos atributos.
- La regresión logística es de la forma 0/1. y = 0 si no se compra el seguro, y = 1 si se compra.

En muchas aplicaciones de la regresión, **la variable dependiente asume sólo dos valores,** por ejemplo, en un banco suele necesitarse una ecuación de regresión estimada para predecir si a una persona se le aprobará su solicitud de tarjeta de crédito.

- Con la regresión logística, dado un conjunto particular de valores de las variables independientes elegidas, se **estima la probabilidad** de que el banco apruebe la solicitud de tarjeta de **crédito**.
- Un modelo de regresión logística difiere del modelo de regresión lineal de dos maneras:
  1. En primer lugar, la **regresión logística solo acepta entradas dicotómicas (binarias) como una variable dependiente** (es decir, un vector de 0 y 1).
  2. En segundo lugar, el resultado se mide mediante una función de enlace probabilístico llamada sigmoide debido a su forma de S.

Por otra parte, nos interesa estudiar la relación entre una o más variables independientes o explicativas: x_1, x_2, …, x_p y la variable y.

El modelo logístico **establece la siguiente relación entre la probabilidad de que ocurra el suceso**, dado que el individuo presenta los valores x_1, x_2, …, x_p.

donde denotamos con Pr (Y = 1| x_1, x_2, …, x_p) a la probabilidad condicional del suceso.

$$\Pr(Y=1\,|\,x_1,x_2,\cdots,x_p)=\frac{1}{1+\exp(-\alpha-\beta_1x_1-\beta_2x_2-\cdots-\beta_px_p)}.$$

Ejemplo: regresión logística para medir intención de voto

Weather

Today: *Variably cloudy, showers. High 53, Low 41.*
Monday: *Variably cloudy, warmer. High 62, Low 45.*
Details, Page **B2.**

The Washington Post

Inside: **Book World, TV Week, The Post Magazine, Comics**
Today's Contents on Page A2

$1.50

122ND YEAR, NO. 15 ... ...

SUNDAY, DECEMBER 20, 1998

Clinton Impeached

House Approves Articles Charging Perjury, Obstruction

Livingston Quits As Designated House Speaker

By ERIC PIANIN
Washington Post Staff Writer

Fearing that a controversy over his sexual past would undercut his power and tear apart his family, Rep. Bob Livingston (R-La.) yesterday told an astounded House he will not assume the speakership he claimed last month but would instead resign from Congress next year.

Livingston made his unexpected announcement during the impeachment debate on the House floor after pointedly

Mostly Partisan Vote Shifts Drama to Senate

By PETER BAKER *and* JULIET EILPERIN
Washington Post Staff Writers

The House of Representatives impeached the president of the United States yesterday for only the second time in American history, charging William Jefferson Clinton with "high crimes and misdemeanors" for lying under oath and obstructing justice to cover up an Oval Office affair with a young intern.

At 1:25 p.m. on a day of constitutional drama and personal trauma, the Republican-led House voted 228 to 206 largely along

Consideremos el voto de juicio político del Senado a Bill Clinton en 1999 como un ejemplo

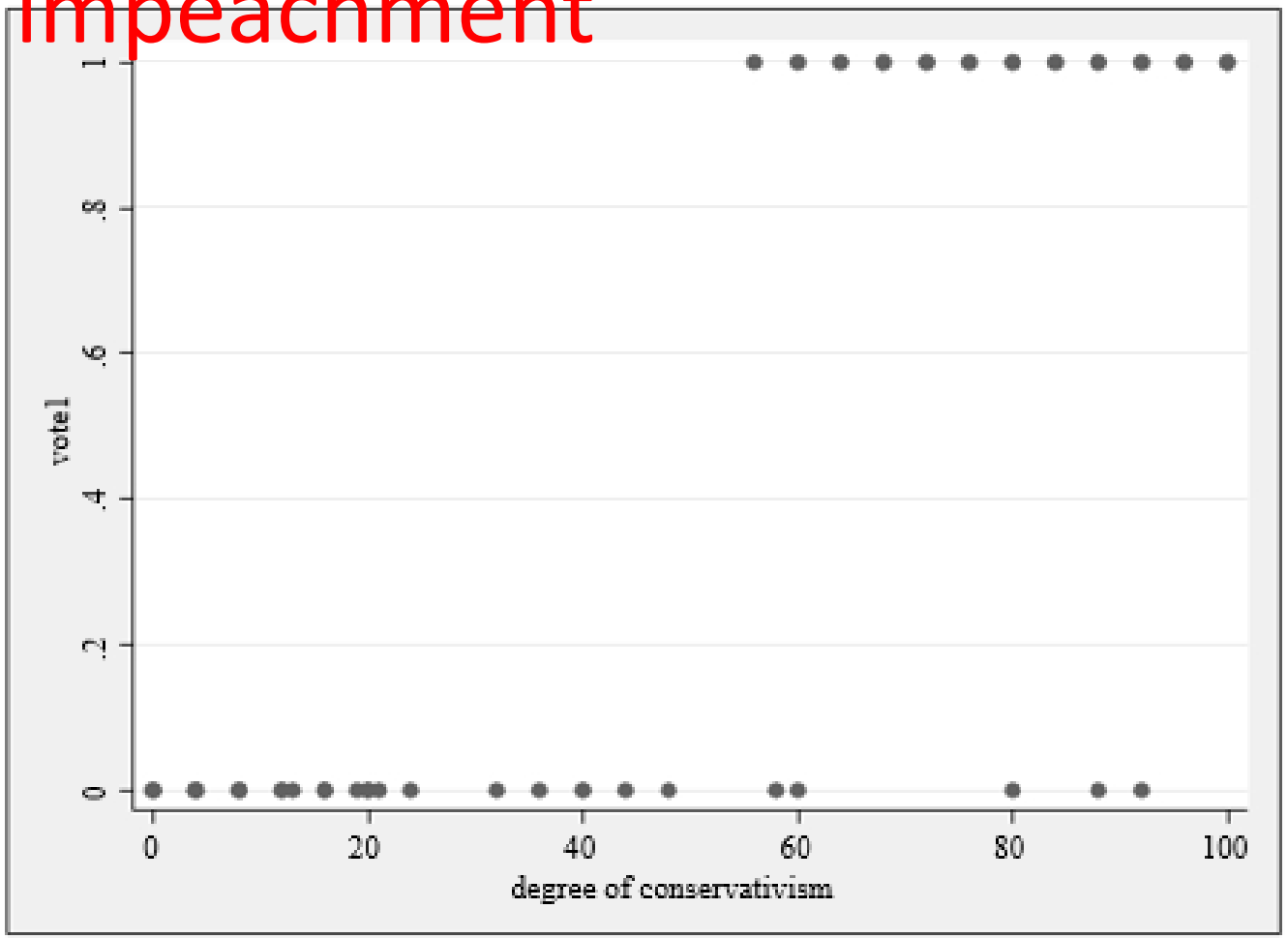
Ejemplo: regresión logística para medir intención de voto

Prediciendo el voto de apoyo

- **Variable dependiente:** 'culpable' (1) o 'no culpable' (0).
- **Variable predictora:** grado de conservadurismo ideológico del senador ('conservadurismo').
- Esa escala va de 0-100, 100 es más conservador.
- Emitido por la 'American Conservative Union' (http://conservative.org/).
- Basado en los registros de votación del senado.

Los datos sugieren que el nivel de conservadurismo predice el voto

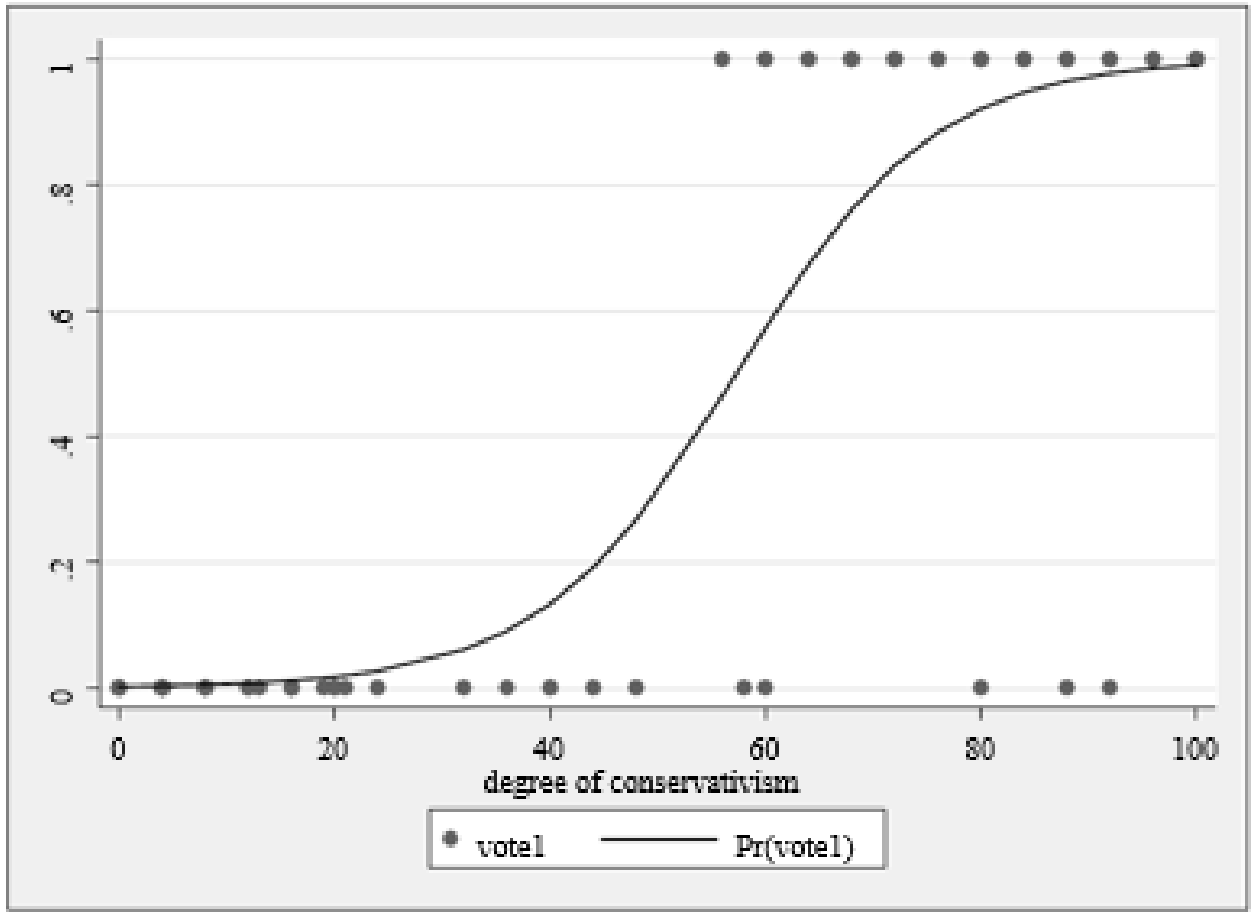
Regresión para el voto de impeachment



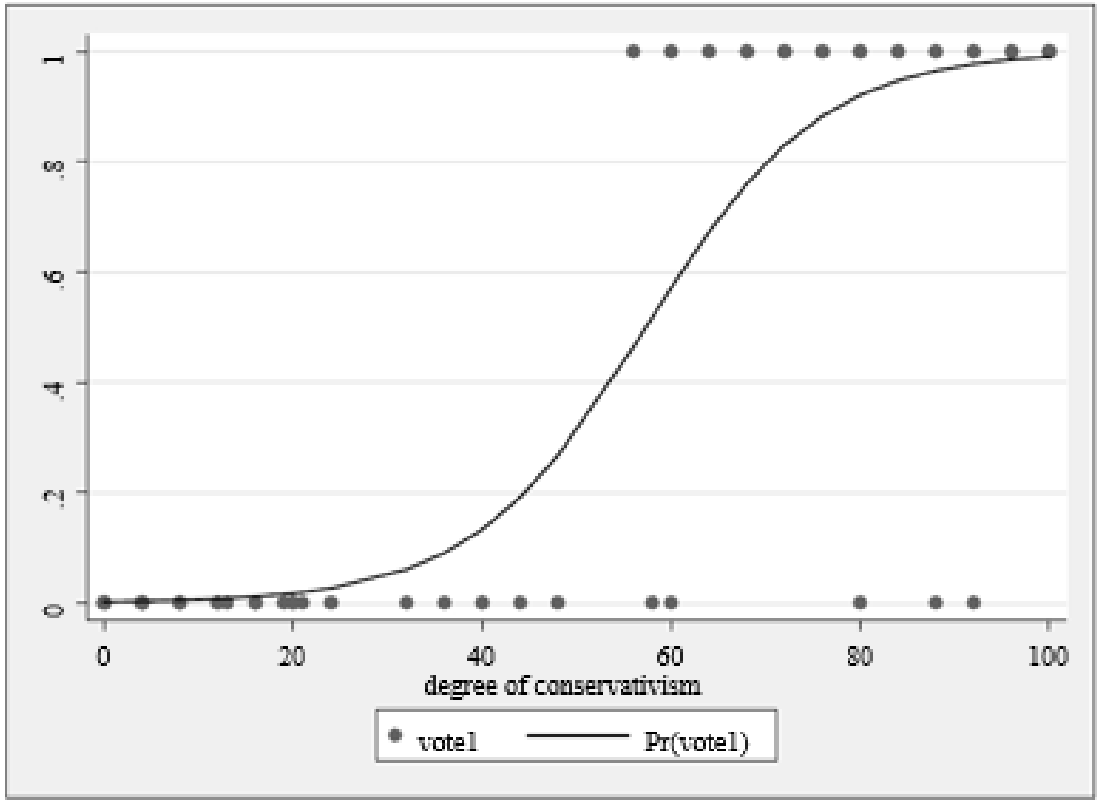
Queremos un método que corrija las fallas de la regresión

Predicción probabilística ideal

Voto



Un modelo de regresión logística es una forma de predecir opciones binarias...



- También conocido como 'regresión logística' (Daniel McFadden, Berkeley Econ Nobel Laureate 2000).
- Se usa cuando la variable dependiente es binaria:
  - Comprar / no comprar (modelos de elección de compra).
  - Izquierda / detenida (modelos de deserción, política).
  - Falló / no falló (mantenimiento predictivo).
  - Abandono / no abandono.

» La regresión logística es una técnica estadística muy útil para sacar perfiles y sobre todo para identificar las causas de los fenómenos, algo importantísimo si queremos incidir o intervenir sobre la realidad social.

Ejemplo de la regresión logística

a. Estudiamos los factores que influyen en la compra de un seguro médico.
b. Variable dependiente: si una persona tiene o no seguro médico (0 o 1).
c. Variables independientes: jubilado, edad, buen estado de salud, ingresos del hogar, años de educación, casado, hispano.

Seguro medico	Categorías	Frecuencia
Sí	1	39%
No	0	61%

Resultados de la regresión logística

Seguro medico	Coeficientes regresión logística
Jubilado	0.19*
Edad	-0.01
Estado de salud	0.31*
Ingreso	0.002
Educación	0.11*
Casado	0.57*
Hispano	-0.81*

* Indica significancia al nivel del 95%.

Interpretación de coeficientes: las personas jubiladas (en comparación con las personas no jubiladas), las personas con buen estado de salud, ingresos más altos, educación superior, casados, tienen más probabilidades de tener seguro médico. Los hispanos tienen menos probabilidades de tener seguro médico.

Efectos marginales

Seguro medico	Efectos marginales regresión logística
Jubilado	0.04*
Edad	-0.003
Estado de salud	0.07*
Ingreso	0.0005*
Educación	0.02*
Casado	0.12*
Hispano	-0.16*

Interpretación de los efectos marginales: las personas jubiladas tienen un 4% más de probabilidades de tener seguro (en comparación con las que no están jubiladas). Por cada año adicional de educación, las personas tienen un 2% más de probabilidades de tener un seguro. Los hispanos tienen un 16% menos de probabilidades de tener seguro que los no hispanos.