Tres temas especiales

Términos de interacción

Regresión estandarizada

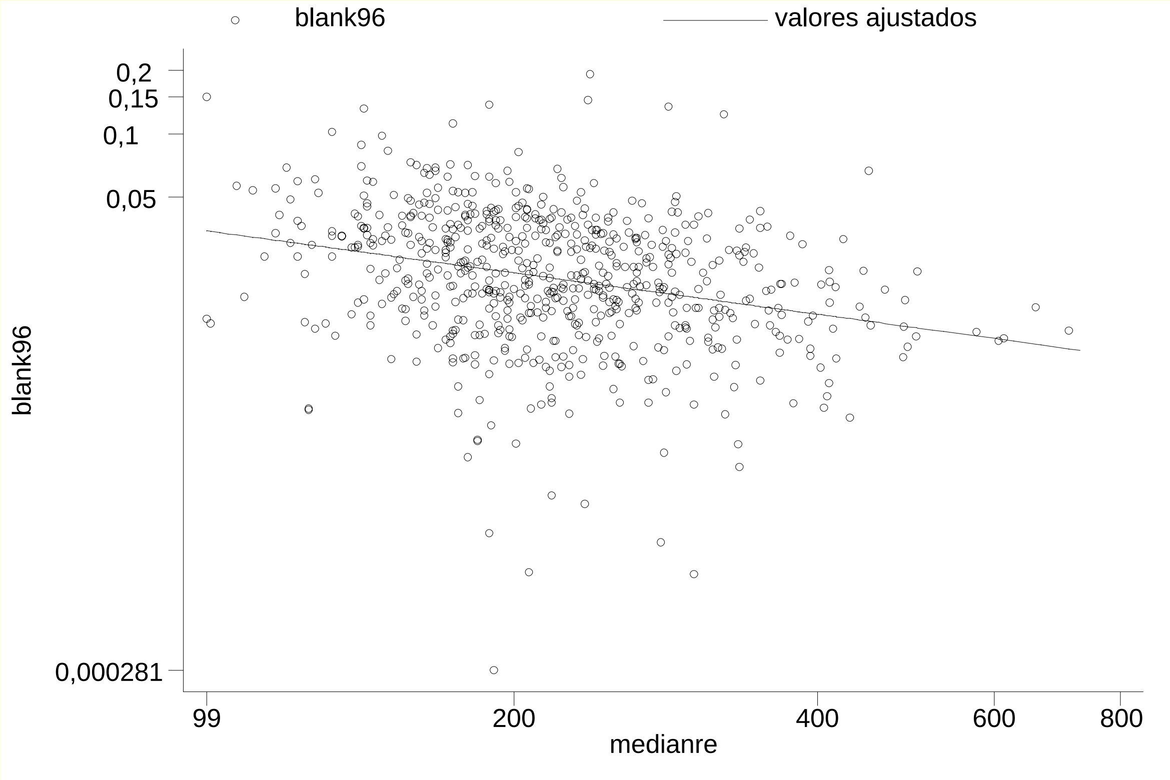
Error de medida

Terminos de interacción

¿Qué ocurre cuando en situaciones diferentes se aplican modelos distintos?

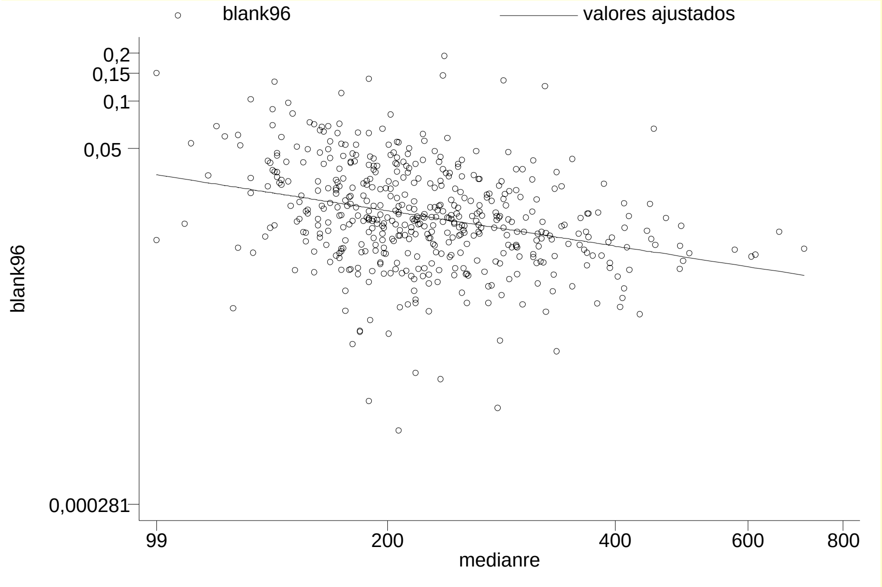
Regresión del voto en blanco (1996) sobre la renta media (1990)

	Coef.	s.e.
Constante	-0,36	0,48
Pendiente	-0,65	0,088
N	663	
R^2	0,077	



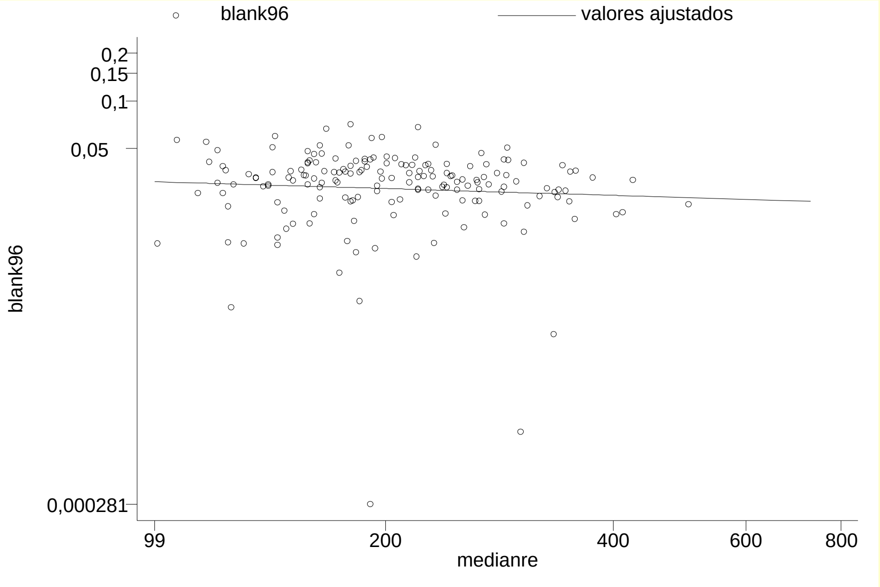
Regresión del voto en blanco (1996) sobre la renta media (1990), por sistema de voto

Escáner



	Coeff.	s.e.
Constante	0,040	0,48
Pendiente	-0,74	0,088
N	491	
R^2	0,10	

Electrónico



	Coeff.	s.e.
Constante	-2,83	0,82
Pendiente	-0,14	0,15
N	172	
R^2	0,005	

¿Qué hacer?

- Ejecute dos regresiones separadas
  - Ventaja: sencillo conceptualmente
  - Desventaja: prueba de hipótesis lenta y difícil
- Términos de interacción
  - Ventaja: facilidad en la prueba de hipótesis
  - Desventaja: conceptualmente complejo

Términos de interacción en el ejemplo de las máquinas para votar

- Defina S_c = 1 si el condado utiliza escáner óptico, 0 en caso contrario.

- Ejecute la regresión:

$$blankpct_c = \beta_0 + \beta_1 \times rent_c + \beta_2 \times S_c + \beta_3 \times S_c \times rent_c + \varepsilon_c$$

$$\text{Vea que si } S_c = 0 \text{ (i.e., condado electrónico), tenemos}$$

$$blankpct_c = \beta_0 + \beta_1 \times rent_c + \varepsilon_c$$

$$\text{Si } S_c = 1 \text{ (i.e., condado con escáner), tenemos}$$

$$blankpct_c = \beta_0 + \beta_1 \times rent_c + \beta_2 + \beta_3 \times rent_c + \varepsilon_c \text{ o}$$

$$blankpct_c = (\beta_0 + \beta_2) + (\beta_1 + \beta_3) \times rent_c + \varepsilon_c$$

Haciéndolo en *STATA*:

```
. gen scan=ve96_cod==“5”
. gen s=scan
. gen scanrent=scan*rent
. reg blank rent scan scanrent

      Source |       SS       df       MS              Number of obs =     663
-------------+------------------------------           F(  3,   659) =   32.71
       Model |   48.597571     3  16.1991903           Prob > F      =  0.0000
    Residual |   326.36878   659  .495248527           R-squared     =  0.1296
-------------+------------------------------           Adj R-squared =  0.1256
       Total |  374.966351   662  .566414427           Root MSE      =  .70374

------------------------------------------------------------------------------
       blank |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        rent |    -.14256   .1686598    -0.85   0.398    -.4737353    .1886153
        scan |   2.866141   1.051092     2.73   0.007     .8022468    4.930035
    scanrent |  -.6017711    .196494    -3.06   0.002     -.987601   -.2159413
       _cons |  -2.826596   .8975356    -3.15   0.002     -4.58897   -1.064222
------------------------------------------------------------------------------
```

Términos de interacción en general

$$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + \varepsilon$$

Replanteamiento,

$$y = \beta_0 + (\beta_1 + \beta_3 X_2) X_1 + \beta_2 X_2 + \varepsilon$$

Regresión estándar

Comparar manzanas (estándar)
con naranjas (estándar)

¿Qué “importa” más al determinar el resultado de los votos, la popularidad o la economía?

```
. reg vote drdi gallup

      Source |       SS       df       MS              Number of obs =      13
-------------+------------------------------           F(  2,    10) =   20.01
       Model |  .038942217     2  .019471109           Prob > F      =  0.0003
    Residual |  .009732889    10  .000973289           R-squared     =  0.8000
-------------+------------------------------           Adj R-squared =  0.7601
       Total |  .048675106    12  .004056259           Root MSE      =   .0312

------------------------------------------------------------------------------
        vote |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        drdi |   1.908849    .545243     3.50   0.006     .6939719    3.123726
      gallup |   .2554055     .07231     3.53   0.005     .0942888    .4165223
       _cons |   .3422054   .0350065     9.78   0.000     .2642061    .4202047
------------------------------------------------------------------------------
```

Soluciones I

- Normalizar en porcentajes
  - Tomar *logs* de todo
  - Ventaja: elegante
  - Desventaja:
    - No siempre transformación apropiada
    - Cero, números negativos

$$\ln(y) = \beta_0 + \beta_1 \ln(x_1) + \beta_2 \ln(x_2) + \varepsilon$$

Calcular $\partial y/\partial x_1$ y $\partial y/\partial x_2$ y cambiar términos :

$$\beta_1 = \frac{\partial y / y}{\partial x_1 / x_1}, \beta_2 = \frac{\partial y / y}{\partial x_2 / x_2}$$

Soluciones II

- Transformar las variables en desviaciones unitarias (I.e., media 0, desviación estándar 1)
  - Restar cada variable de su media y dividir por su desviación estándar, I.e.:

$$z_{i,j} = \frac{(Z_{i,j} - \bar{Z}_i)}{s_{Z_i}}$$

Haciéndolo en *STATA*:

```
. reg vote drdi gallup,beta

      Source |       SS       df       MS              Number of obs =      13
-------------+------------------------------           F(  2,    10) =   20.01
       Model |  .038942217     2  .019471109           Prob > F      =  0.0003
    Residual |  .009732889    10  .000973289           R-squared     =  0.8000
-------------+------------------------------           Adj R-squared =  0.7601
       Total |  .048675106    12  .004056259           Root MSE      =   .0312

------------------------------------------------------------------------------
        vote |      Coef.   Std. Err.      t    P>|t|                     Beta
-------------+----------------------------------------------------------------
        drdi |   1.908849    .545243     3.50   0.006                  .535644
      gallup |   .2554055     .07231     3.53   0.005                 .5404141
       _cons |   .3422054   .0350065     9.78   0.000                        .
------------------------------------------------------------------------------
```

Error de medida

¿Qué ocurre cuando no puede medirse todo perfectamente?

Supongamos que medimos x con un error

En lugar de observar x, observamos $x' = x + e$

(e es aleatorio con media $\bar{e}$ y varianza v_e)

$\therefore$ en lugar de realizar la regresión

$y = \alpha + \beta x + \varepsilon,$

realizamos la regresión

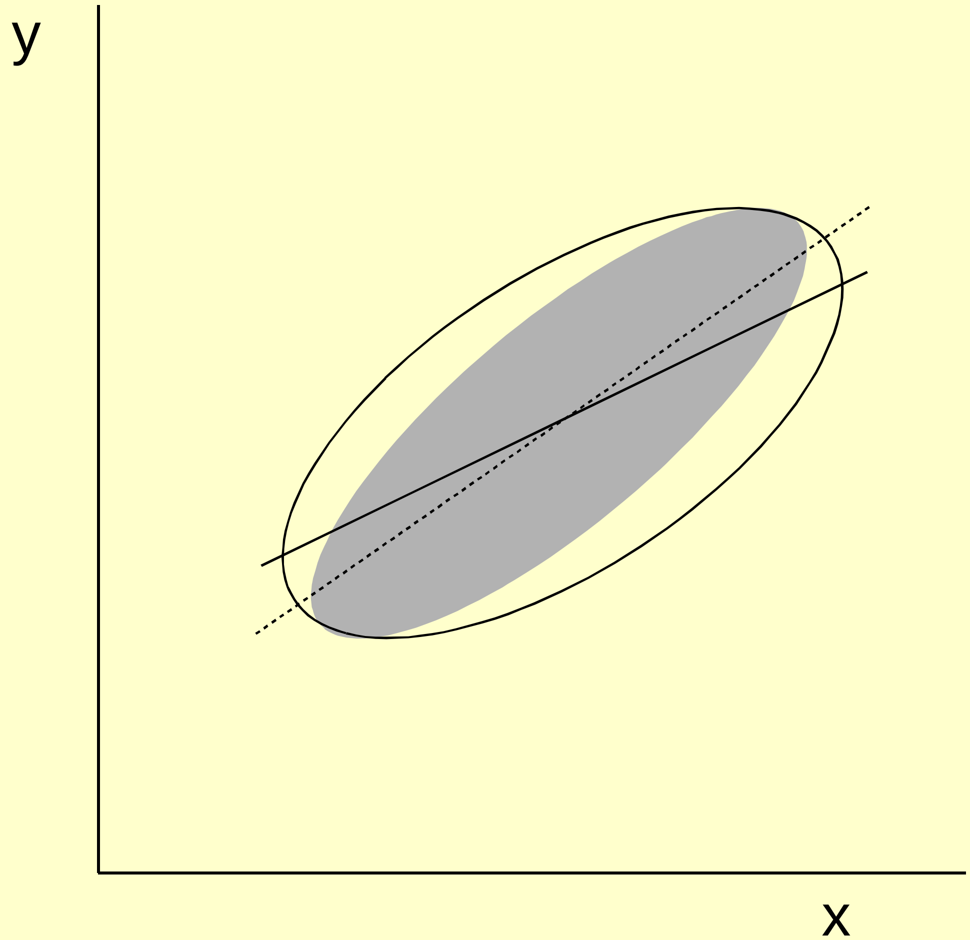
$y = \alpha + \beta' x' + \varepsilon.$

¿Cuál es la relación entre β y β'?

Respuesta

$$\beta' = \frac{\mathrm{cov}(x, y)}{\mathrm{var}(x) + \mathrm{var}(e)}$$

Errores en variables independientes: el cuadro



Supongamos que medimos *y* con un error

En lugar de observar y, observamos $y' = y + e$

(e es aleatorio con media $\bar{e}$ y varianza v_e)

$\therefore$ en lugar de realizar la regresión

$y = \alpha + \beta x + \varepsilon,$

realizamos la regresión

$y' = \alpha + \beta' x + \varepsilon.$

¿Cuál es la relación entre β y β'?

La respuesta

$$\beta' = \frac{\mathrm{cov}(x, y)}{\mathrm{var}(x)} = \beta$$

Pero…

- s.e. y s.e.r. resultan amplificados
- R^2 comprimido

Errores en variables dependientes: el cuadro