Three Special Topics

Interaction Terms

Standardized Regression

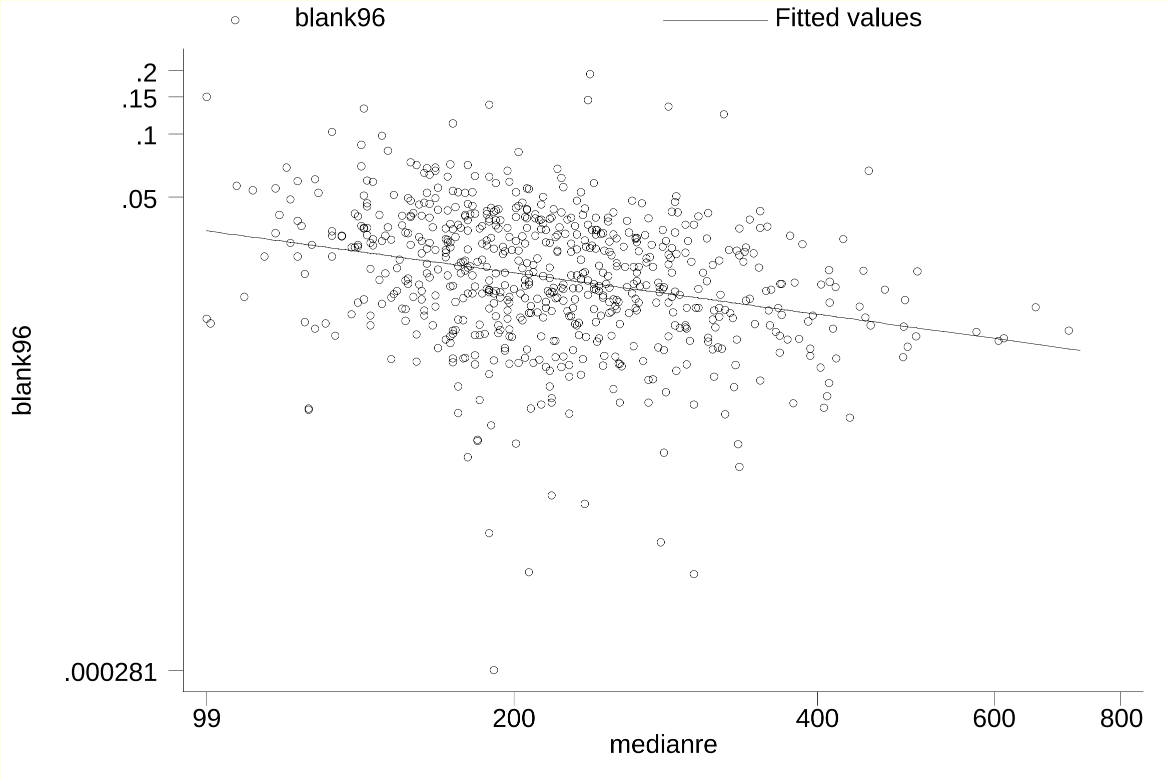
Measurement Error

Interaction Terms

What Happens When Different Models Apply in Different Situations?

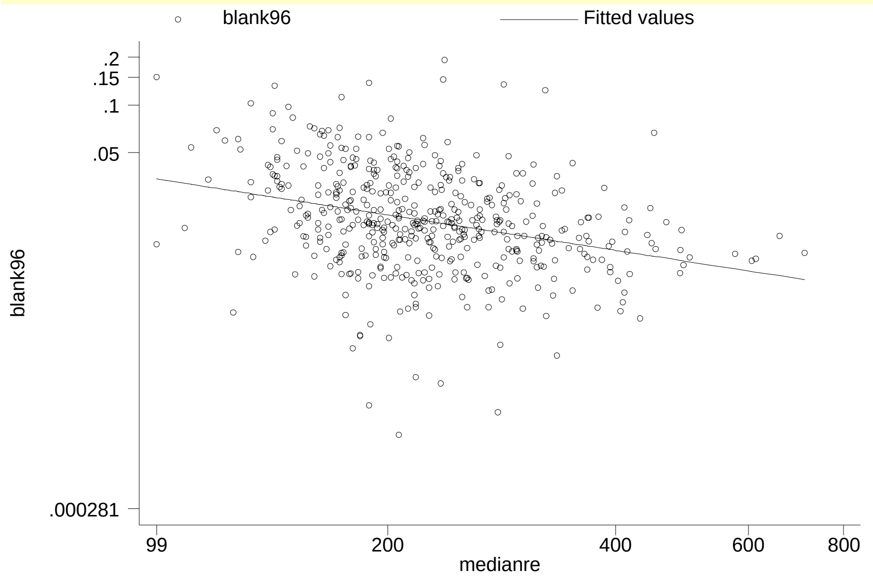
Regression of Blank Ballots (1996) on Median Rent (1990)

	Coeff.	s.e.
Intercept	-0.36	0.48
Slope	-0.65	0.088
N	663	
R^2	.077	



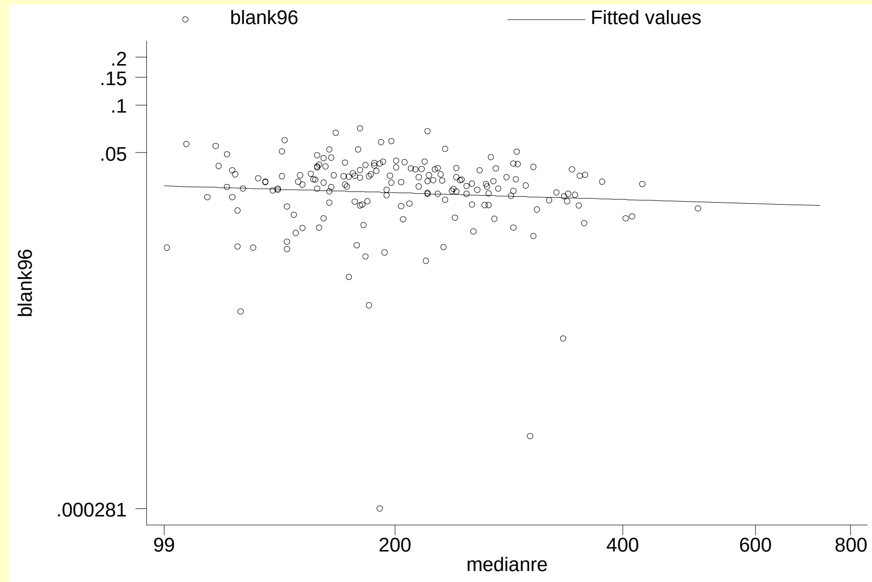
Regression of Blank Ballots (1996) on Median Rent (1990), By Ballot Type

Scanning



	Coeff.	s.e.
Intercept	0.040	0.48
Slope	-0.74	0.088
N	491	
R^2	.10	

Electronic



	Coeff.	s.e.
Intercept	-2.83	0.82
Slope	-0.14	0.15
N	172	
R^2	.005	

What to do?

- Run two separate regressions
  - Advantage: conceptually simple
  - Disadvantage: hypothesis testing cumbersome
- Interaction terms
  - Advantage: hypothesis testing facilitated
  - Disadvantage: conceptually complex

Interaction terms in the voting machine example

- Define S_c = 1 if the county uses optical scanning, 0 otherwise
- Run this regression:

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

Note that if $S_c = 0$ (i.e., electronic county), we have

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

If $S_c = 1$ (i.e., scanned county), we have

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

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

Doing this in *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
------------------------------------------------------------------------------
```

Interaction terms generally

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

Rewriting,

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

Standardized Regression

Comparing (Standardized) Apples
with (Standardized) Oranges

Which "matters" more in determining vote outcomes, popularity or the economy?

```
. 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
------------------------------------------------------------------------------
```

Solutions I

- Normalize into percentages
  - Take logs of everything
  - Advantage: elegant
  - Disadvantages:
    - Not always appropriate transform
    - Zero, negative numbers

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

Calculate $\partial y/\partial x_1$ and $\partial y/\partial x_2$ and rearrange terms:

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

Solutions II

- Transform the variables into unit deviates (I.e., mean 0, s.d. 1)
  - Subtract each variable from its mean and divide by its standard deviation, I.e.:

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

Doing this in *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                        .
------------------------------------------------------------------------------
```

Measurement Error

What Happens When You Can’t Measure Things Perfectly?

Suppose we measure *x* with error?

Instead of observing x, we observe $x' = x + e$

(e is random with mean $\bar{e}$ and variance v_e)

$\therefore$ instead of doing the regression

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

we do the regression

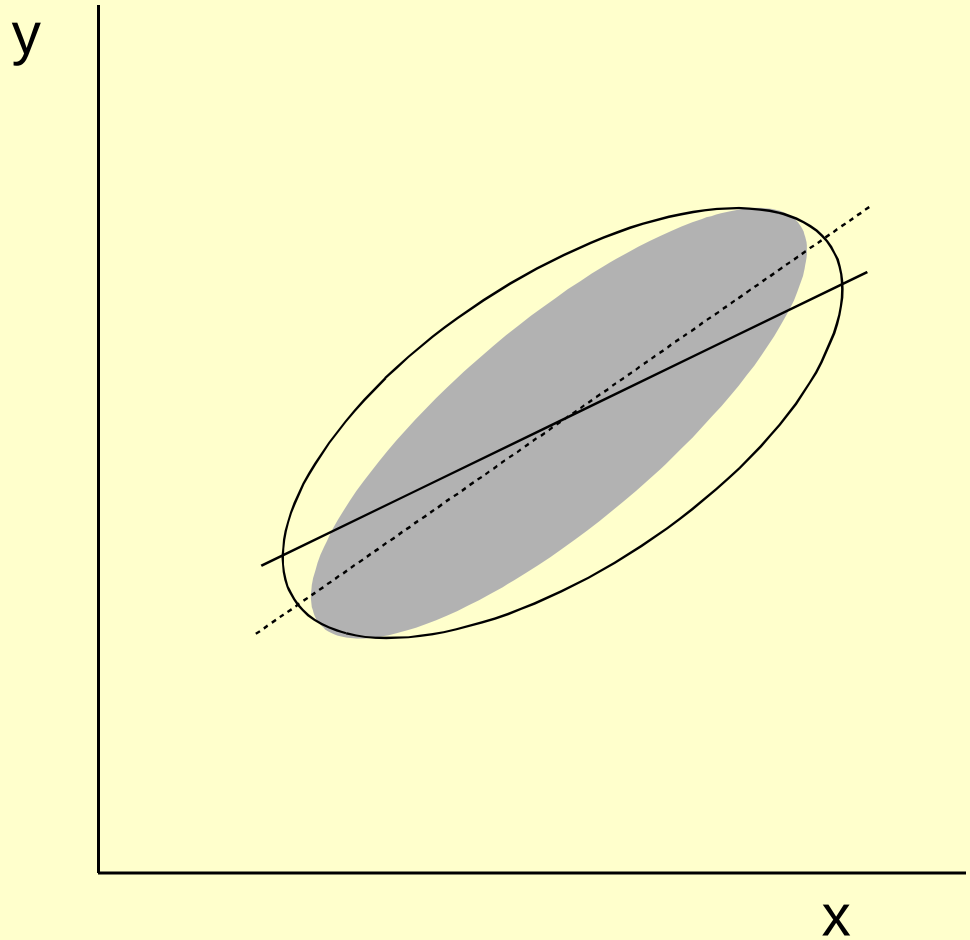
$y = \alpha + \beta' x' + \varepsilon.$

What is the relationship between β and β'?

Answer

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

Errors in Independent Variables: The Picture



Suppose we measure *y* with error

Instead of observing y, we observe $y' = y + e$

(e is random with mean $\bar{e}$ and variance v_e)

$\therefore$ instead of doing the regression

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

we do the regression

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

What is the relationship between β and β'?

The answer

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

But…

- Standard errors and s.e.r. inflated
- R^2 deflated

Errors in Dependent Variables: The Picture