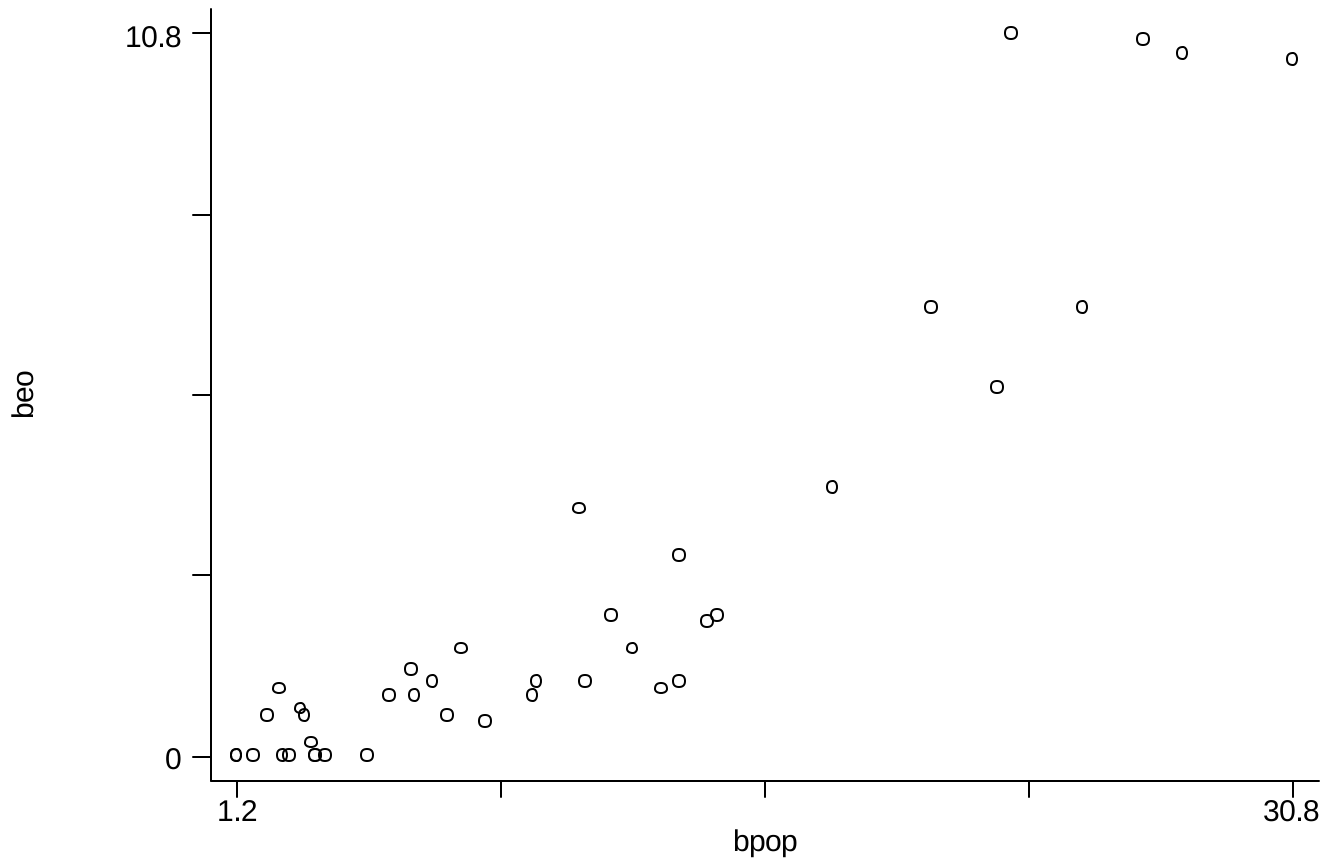
Regression Forced March

Regression quantifies how one variable can be described in terms of another

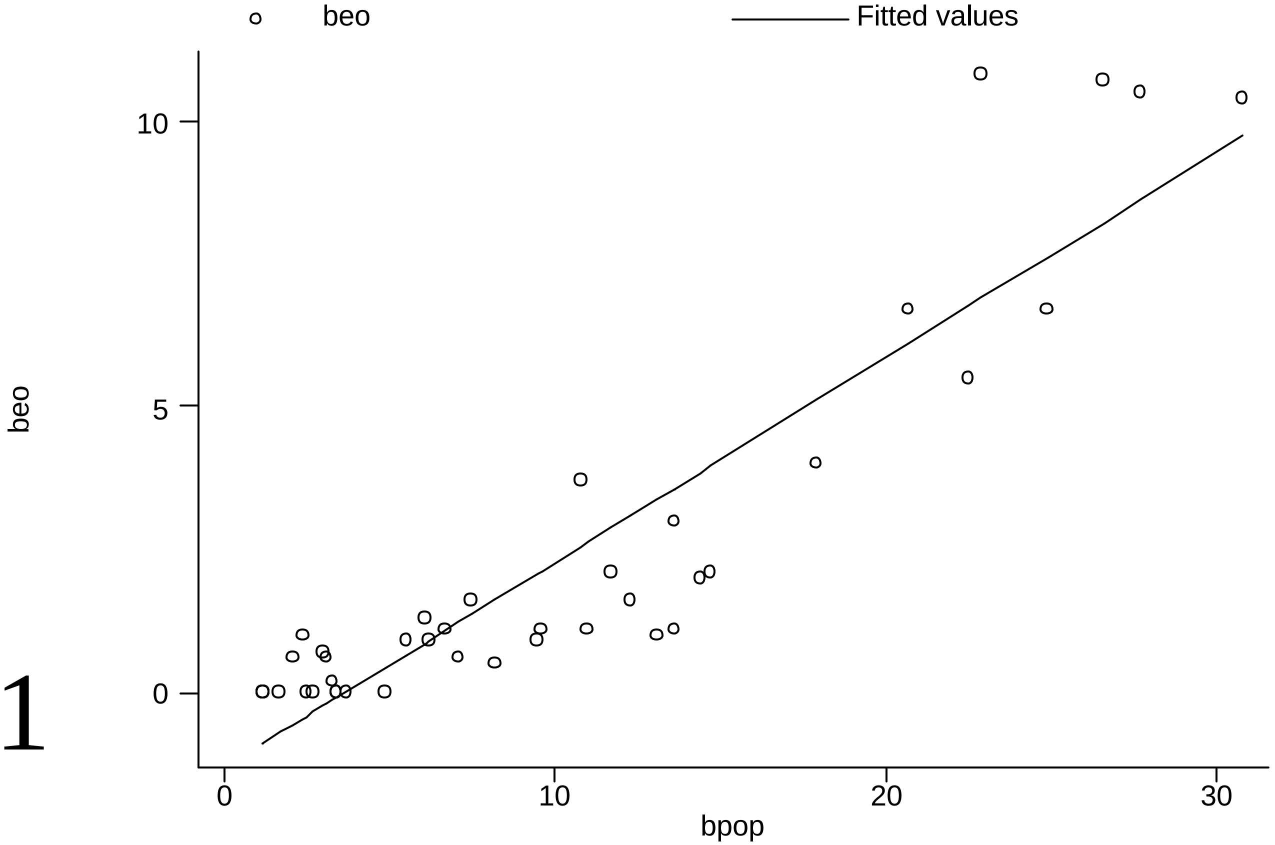
Black Elected Officials Example I



The Linear Relationship between Two Variables

$$Y_i = b_0 + b_1 X_i + e_i$$

The Linear Relationship between African American Population & Black Legislators



$$b_0 = -1.31$$

$$b_1 = 0.359$$

The Method of Least Squares

Pick b_0 and b_1 to minimize

$$\sum_{i=1}^{n} (e_i - \hat{Y}_i)^2 \quad \text{or}$$

$$\sum_{i=1}^{n} (e_i - b_0 - b_1 X_i)^2$$

Solve for $\dfrac{\partial \sum_{i=1}^{n} (e_i - b_0 - b_1 X_i)^2}{\partial b_1} = 0$

$$b_1 = \frac{\sum_{i=1}^{n} (\bar{Y} - Y_i)(\bar{X} - X_i)}{\sum_{i=1}^{n} (\bar{X} - X_i)^2} \quad \text{or}$$

$$\frac{\operatorname{cov}(X, Y)}{\operatorname{var}(X)}$$

Solve for $\frac{\partial \sum_{i=1}^{n} (e_i - b_0 - b_1 X_i)^2}{\partial b_0} = 0$

$$b_0 = \bar{Y} - b_1 \bar{X}$$

Note that if you rearrange.....

$$\bar{Y} = b_0 + b_1 \bar{X}$$

How to Think About Regression Results

- Deterministically
- Semi-deterministically
- Expectations

Determining Goodness of Fit I

- Coefficients
  - Standard error of a coefficient
  - *t*-statistic: *coeff./s.e.*

Determining Goodness of Fit II

- Standard error of the regression (Root mean square error in STATA and Freedman)

$$s.e.e. = \sqrt{\frac{\sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2}{d.f.}}$$

Determining Goodness of Fit III

- R-squared

$$r^2 = \frac{\sum_{i=1}^{n}(\hat{Y}_i - \overline{Y})^2}{\sum_{i=1}^{n}(Y_i - \overline{Y})^2} \quad \text{or}$$

$$\text{percent variance "explained"}$$

Return to Black Elected Officials Example

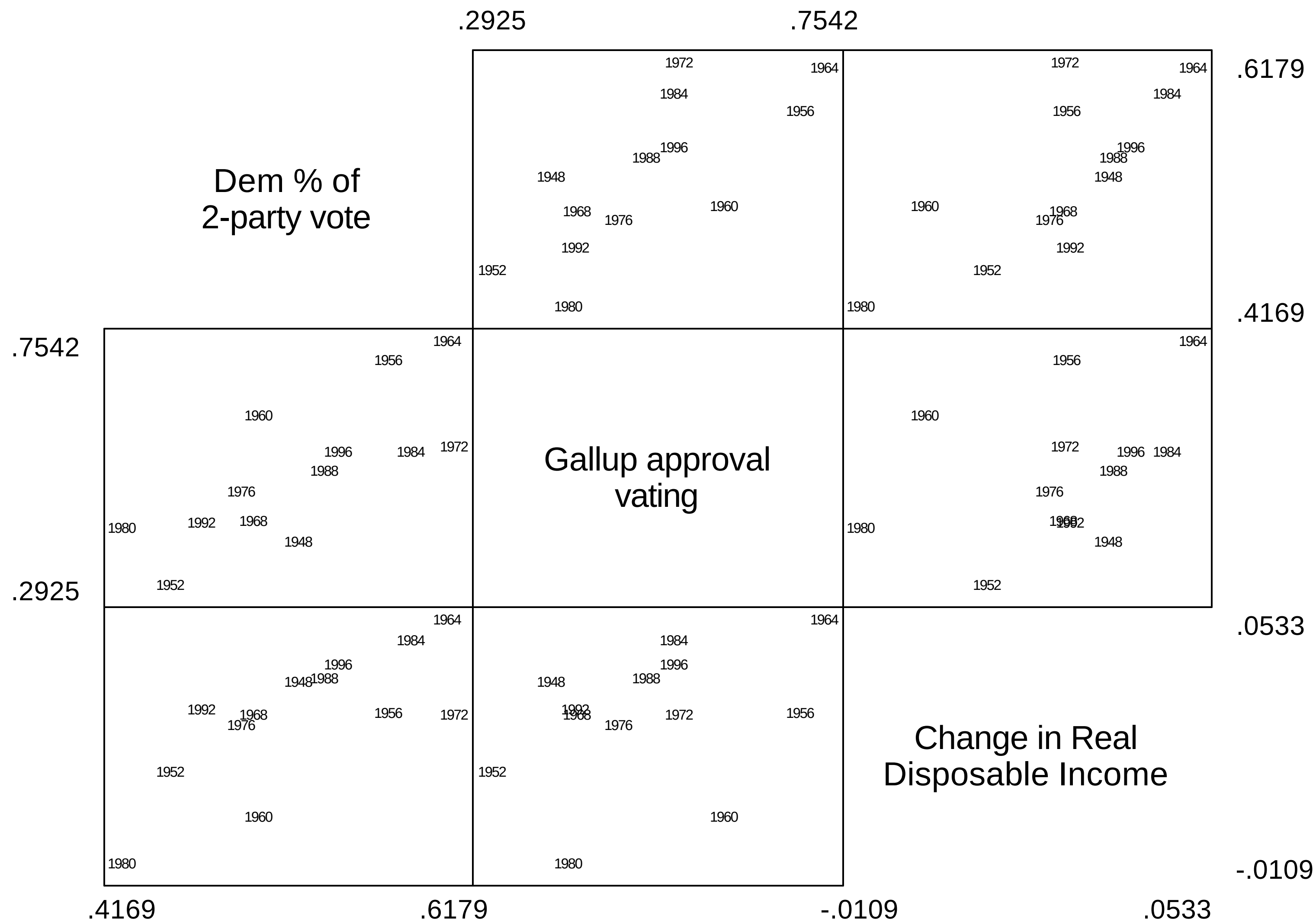
```
. reg beo bpop

      Source |       SS       df       MS              Number of obs =      41
-------------+------------------------------           F(  1,    39) =  202.56
       Model |   351.26542     1   351.26542           Prob > F      =  0.0000
    Residual |  67.6326195    39  1.73416973           R-squared     =  0.8385
-------------+------------------------------           Adj R-squared =  0.8344
       Total |  418.898039    40   10.472451           Root MSE      =  1.3169

------------------------------------------------------------------------------
         beo |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        bpop |   .3584751   .0251876    14.23   0.000     .3075284    .4094219
       _cons |  -1.314892   .3277508    -4.01   0.000    -1.977831   -.6519535
------------------------------------------------------------------------------
```

Multiple Regression

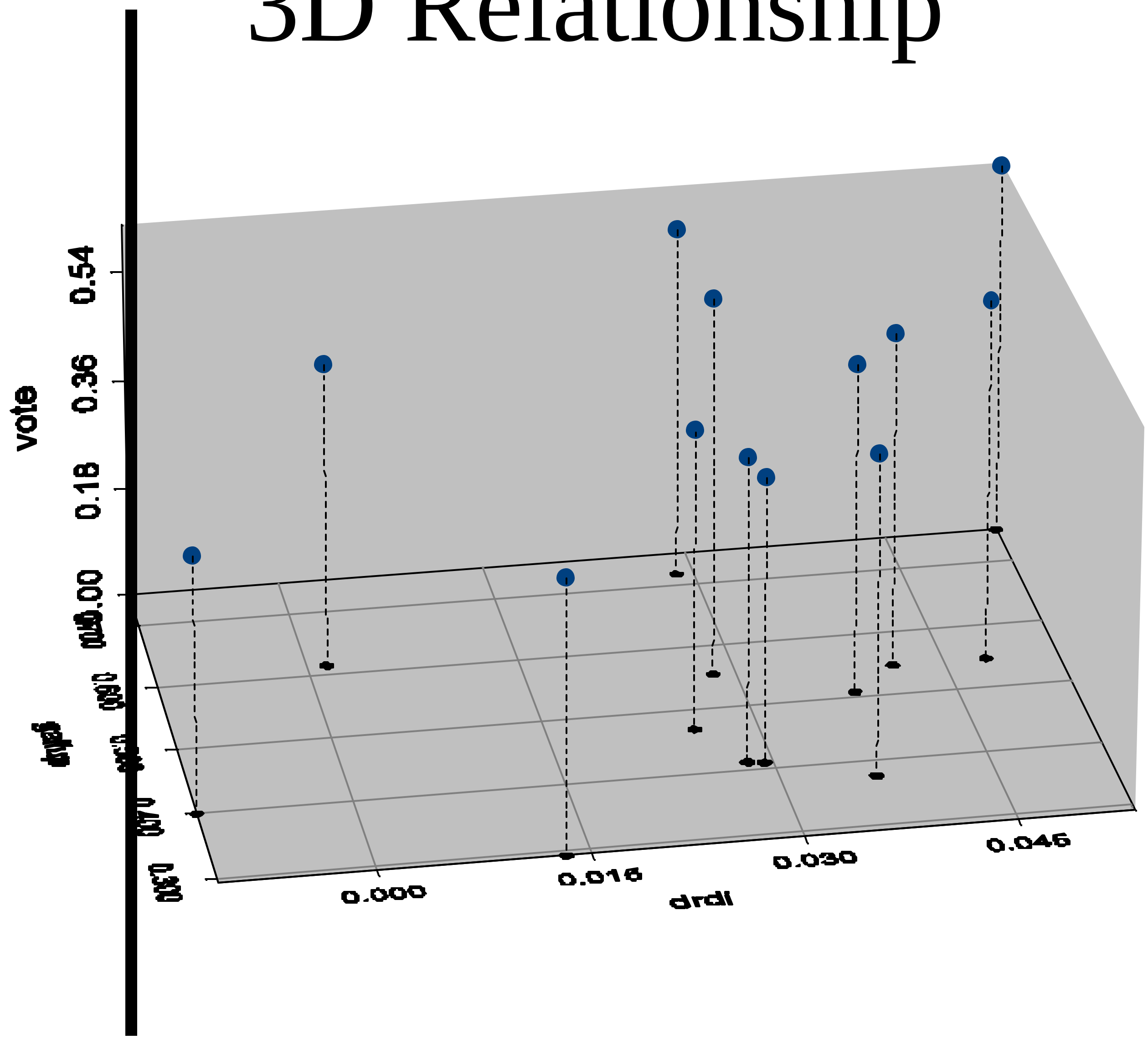
Why to Control

Smiley Face Example

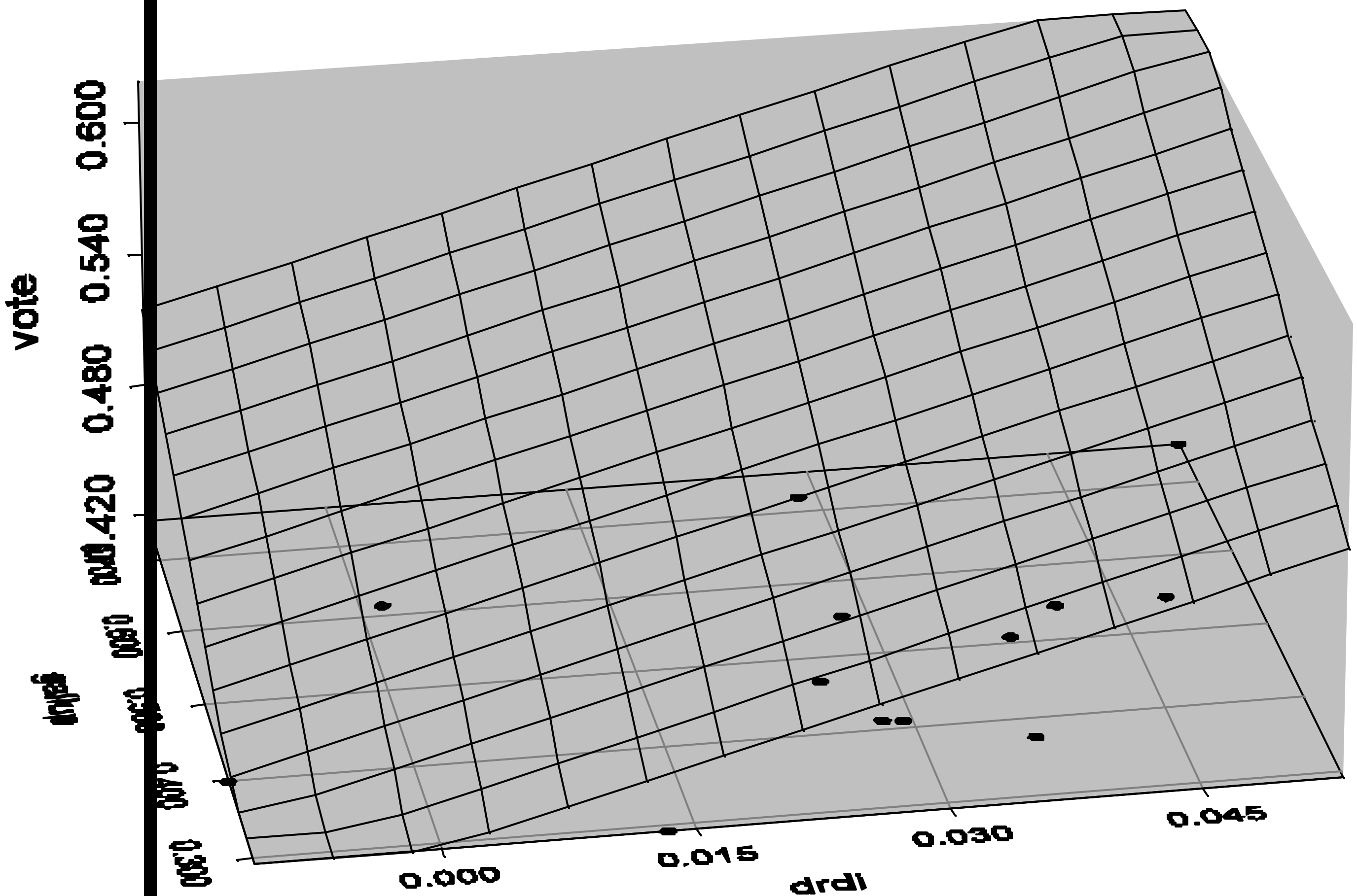
Presidential Vote Graphs



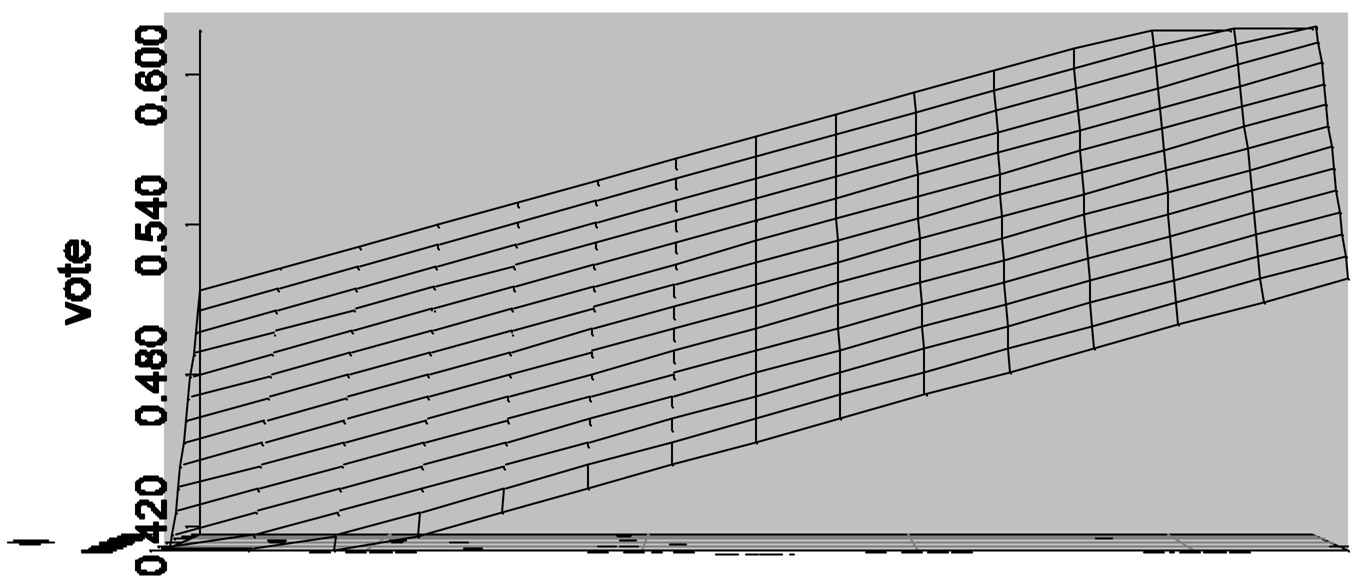
3D Relationship



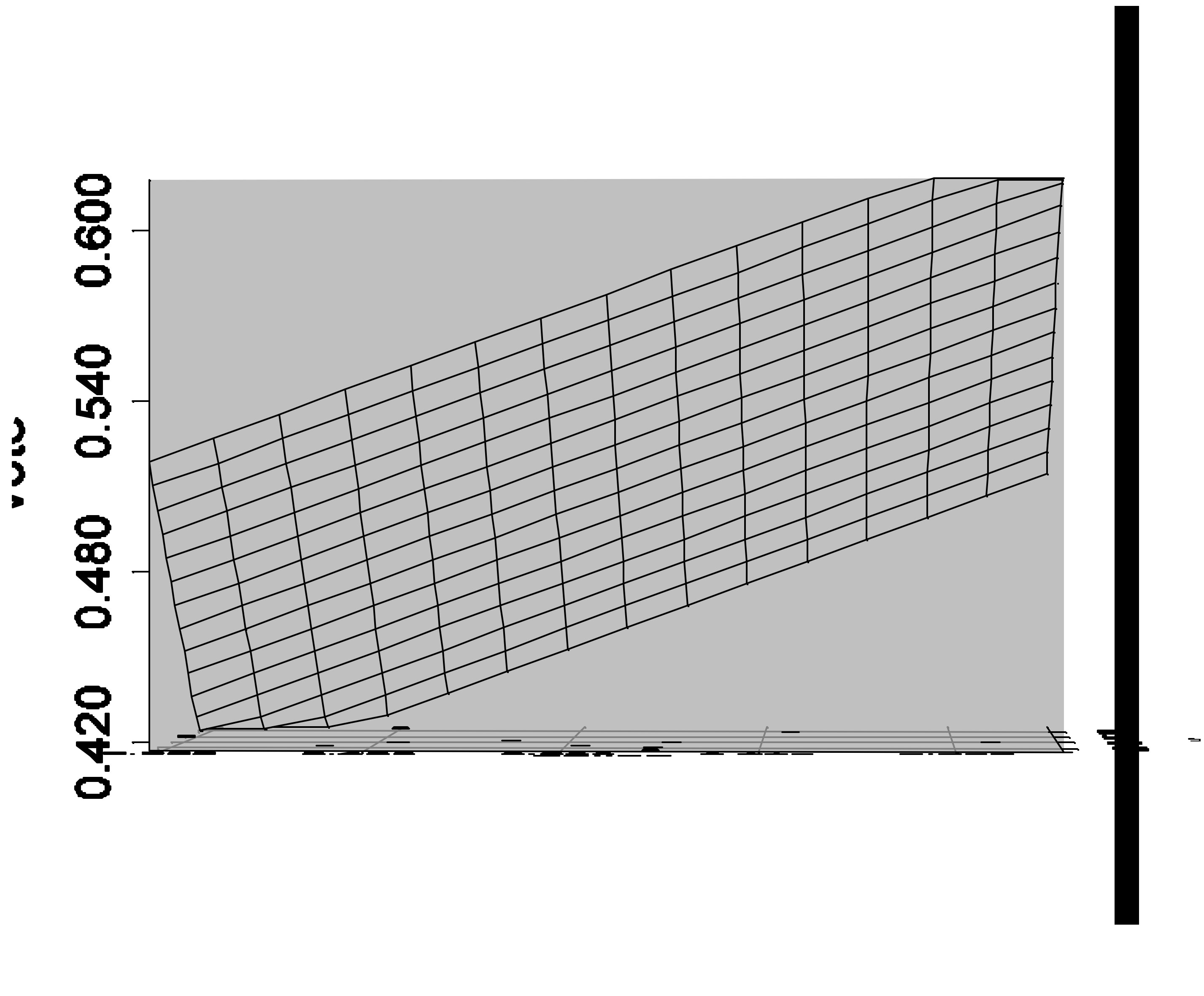
3D Linear Relationship



3D Relationship: drdi



3D Relationship: gallup



The Linear Relationship between Three Variables

$$Y_i = b_0 + b_1 X_{1,i} + b_2 X_{2,1} + e_i$$

The Intercept

$$b_0 = \bar{Y} - b_1\bar{X}_1 - b_2\bar{X}_2$$

Note that if you rearrange.....

$$\bar{Y} = b_0 + b_1\bar{X}_1 + b_2\bar{X}_2$$

The Slope Coefficients

$$\hat{b}_1 = \frac{\sum_{i=1}^{n}(\overline{Y} - Y_i)(\overline{X}_1 - X_{1,i})}{\sum_{i=1}^{n}(\overline{X}_1 - X_{1,i})^2} - \hat{b}_2 \frac{\sum_{i=1}^{n}(\overline{X}_1 - X_{1,i})(\overline{X}_2 - X_{2,i})}{\sum_{i=1}^{n}(\overline{X}_1 - X_{1,i})^2} \text{ and}$$

$$\hat{b}_2 = \frac{\sum_{i=1}^{n}(\overline{Y} - Y_i)(\overline{X}_2 - X_{1,i})}{\sum_{i=1}^{n}(\overline{X}_2 - X_{2,i})^2} - \hat{b}_1 \frac{\sum_{i=1}^{n}(\overline{X}_1 - X_{1,i})(\overline{X}_2 - X_{2,i})}{\sum_{i=1}^{n}(\overline{X}_2 - X_{2,i})^2}$$

The Slope Coefficients More Simply

$$\hat{b}_1 = \frac{\operatorname{cov}(X_1, Y)}{\operatorname{var}(X_1)} - \hat{b}_2 \frac{\operatorname{cov}(X_1, X_2)}{\operatorname{var}(X_1)} \text{ and}$$

$$\hat{b}_2 = \frac{\operatorname{cov}(X_2, Y)}{\operatorname{var}(X_2)} - \hat{b}_2 \frac{\operatorname{cov}(X_1, X_2)}{\operatorname{var}(X_2)}$$

What Difference Does This Make?

One Regression vs. r regressions

Separate regressions

	(1)	(2)	(3)
Intercept	0.346	0.451	0.342
Gallup	0.352		0.255
dRDI		2.644	1.909

Why did the Gallup Coefficient change from 0.352 to 0.255?

- Covariance matrix:
- **. corr vote gallup drdi,cov**
- `(obs=13)`

```
             |     vote   gallup     drdi
-------------+---------------------------
        vote |  .004056
      gallup |  .006394   .01816
        drdi |  .000845   .00092  .000319
```

The Calculations

$$\hat{b} = \frac{\text{cov}(gallup, vote)}{\text{var}(gallup)} = \frac{0.006394}{0.01816} = 0.352$$

$$\hat{b}_1 = \frac{\text{cov}(gallup, vote)}{\text{var}(gallup)} - \hat{b}_2 \frac{\text{cov}(gallup, drdi)}{\text{var}(gallup)}$$

$$= \frac{0.006394}{0.01816} - 1.909 \frac{0.00092}{0.01816}$$

$$= 0.352 - 0.097$$

$$= 0.255$$

The Output

```
. reg vote gallup drdi

      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]
-------------+----------------------------------------------------------------
      gallup |   .2554055     .07231     3.53   0.005     .0942888    .4165223
        drdi |   1.908849    .545243     3.50   0.006     .6939719    3.123726
       _cons |   .3422054   .0350065     9.78   0.000     .2642061    .4202047
------------------------------------------------------------------------------
```