Proportions as dependent variable

Maarten L. Buis

Vrije Universiteit Amsterdam

Department of Social Research Methodology

http://home.fsw.vu.nl/m.buis

Outline

- Problems with using `regress` for proportions as dependent variable
- Methods for dealing with a single proportion
- Methods for dealing with multiple proportions
- Caveat: Ecological Fallacy

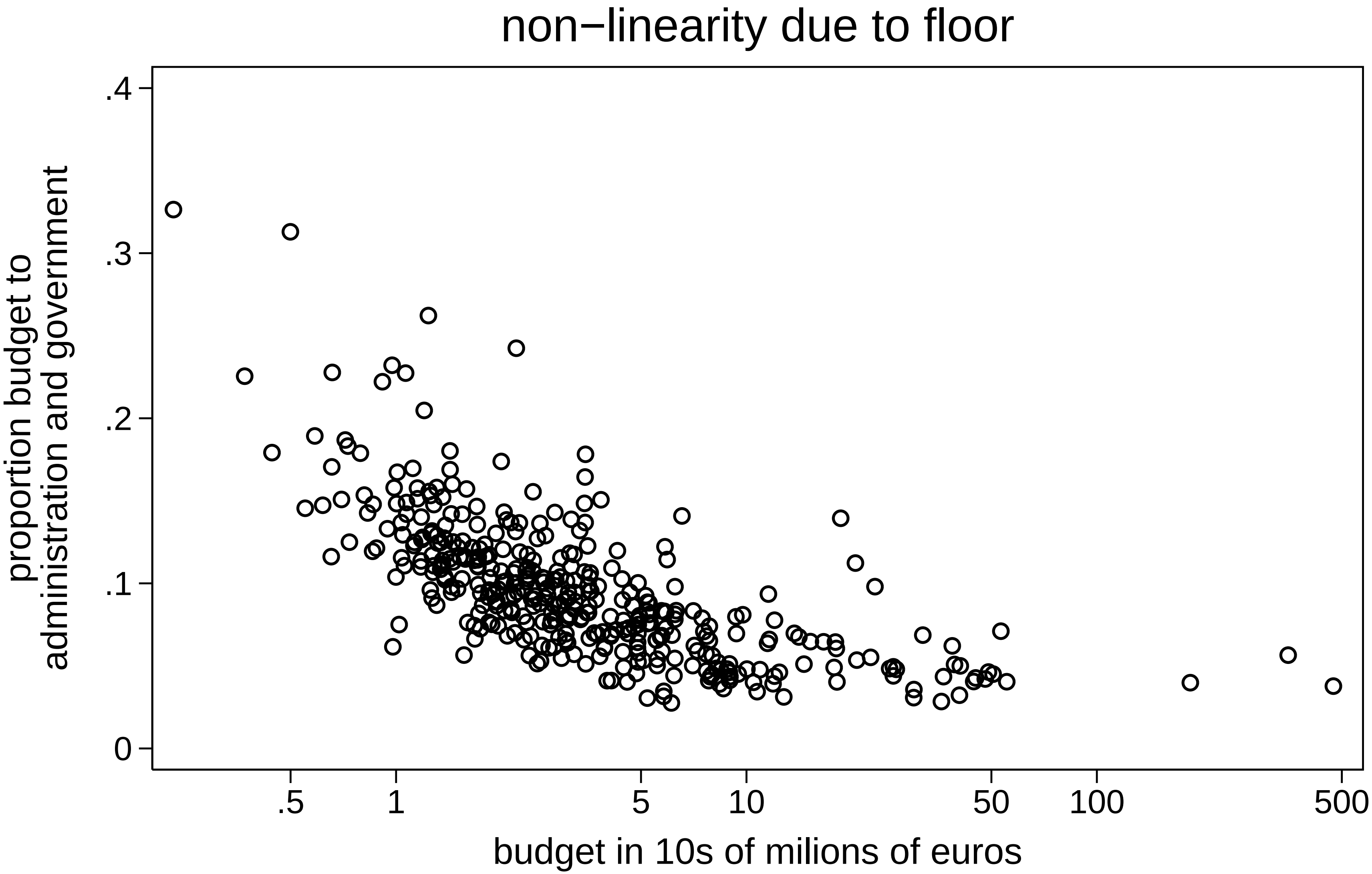
Example

- Explaining proportion of Dutch city budgets spent on administration and government with:
  - Size of budget (natural logarithm of budget in 10s of millions euros)
  - Average house price (in 100,000s of euros)
  - Population density (in 1000s of persons per square km)
  - Political orientation of city government (either no left parties in city government, left parties are a minority in city government, or left parties are a majority in city government)

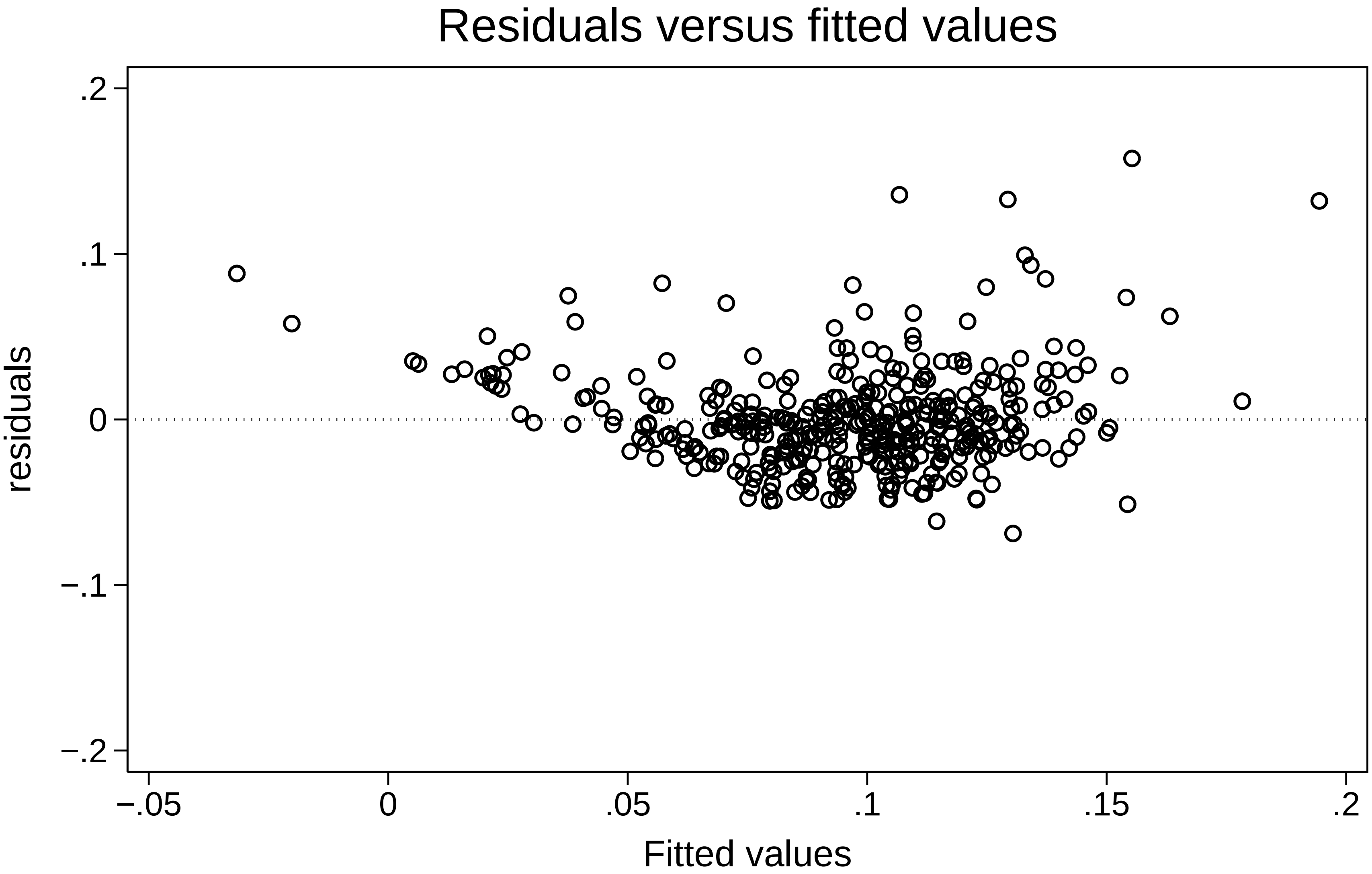
OLS results

	b	se
lntot	-0.030	(0.002)
houseval	0.013	(0.004)
popdens	0.008	(0.002)
noleft	-0.001	(0.005)
minorityleft	-0.007	(0.004)
constant	0.109	(0.008)
R^2	0.499	

Non linear effects due to floor



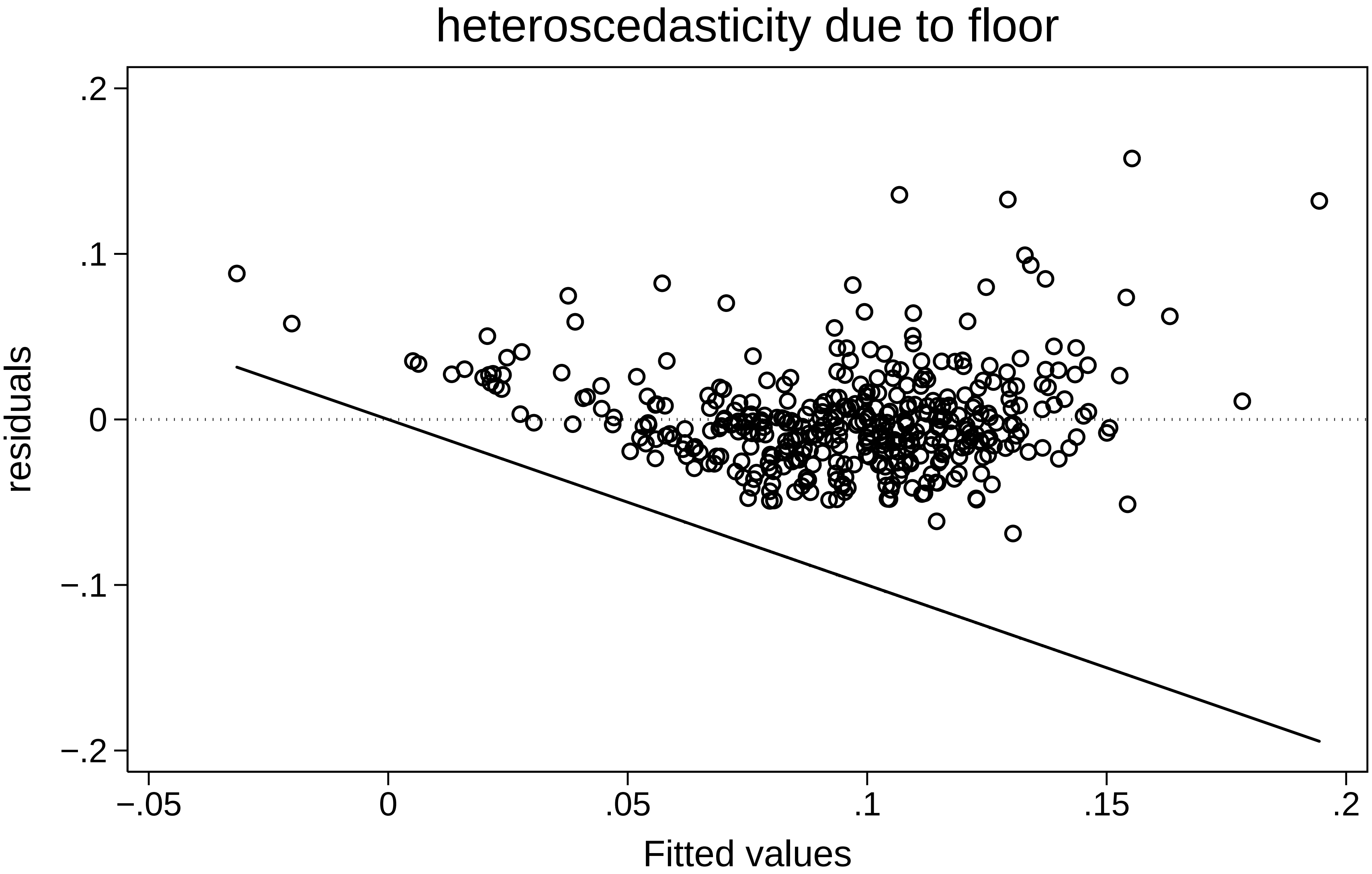
Residuals versus fitted values



Floor

$$
\begin{aligned}
observed &= fitted + residual \\
observed &\geq 0 \; (and \;\; \leq \; 1) \\
fitted + residual &\geq 0 \\
residual &\geq -fitted
\end{aligned}
$$

Residuals versus fitted values



Problems with `regress`

- Impossible predictions.
- Non-normal errors.
- Heteroscedasticity.
- Non-linear effects.

Outline

- Problems with using `regress` for proportions as dependent variable
- Methods for dealing with a single proportion
- Methods for dealing with multiple proportions
- Caveat: Ecological Fallacy

A solution: `betafit`

- Assumes that the proportion follows a beta distribution.
- The beta distribution is bounded between 0 and 1 (but does not include either 0 or 1).
- The beta distribution models heteroscedasticity in such a way that the variance is largest when the average proportion is near 0.5.

Two parameterizations

- the conventional parametrization with two shape parameters (α and β)
  - Corresponds to the formulas of the beta distribution in textbooks.
  - Does not correspond to conventions of Generalized Linear Models where one models how the mean of the distribution of the dependent variable changes as the explanatory variables change.
- the alternative parametrization with one location and one scale parameter (μ and ϕ)
  - Does not correspond to textbook formulas of the beta distribution but does correspond to the GLM convention.

Two parameterizations

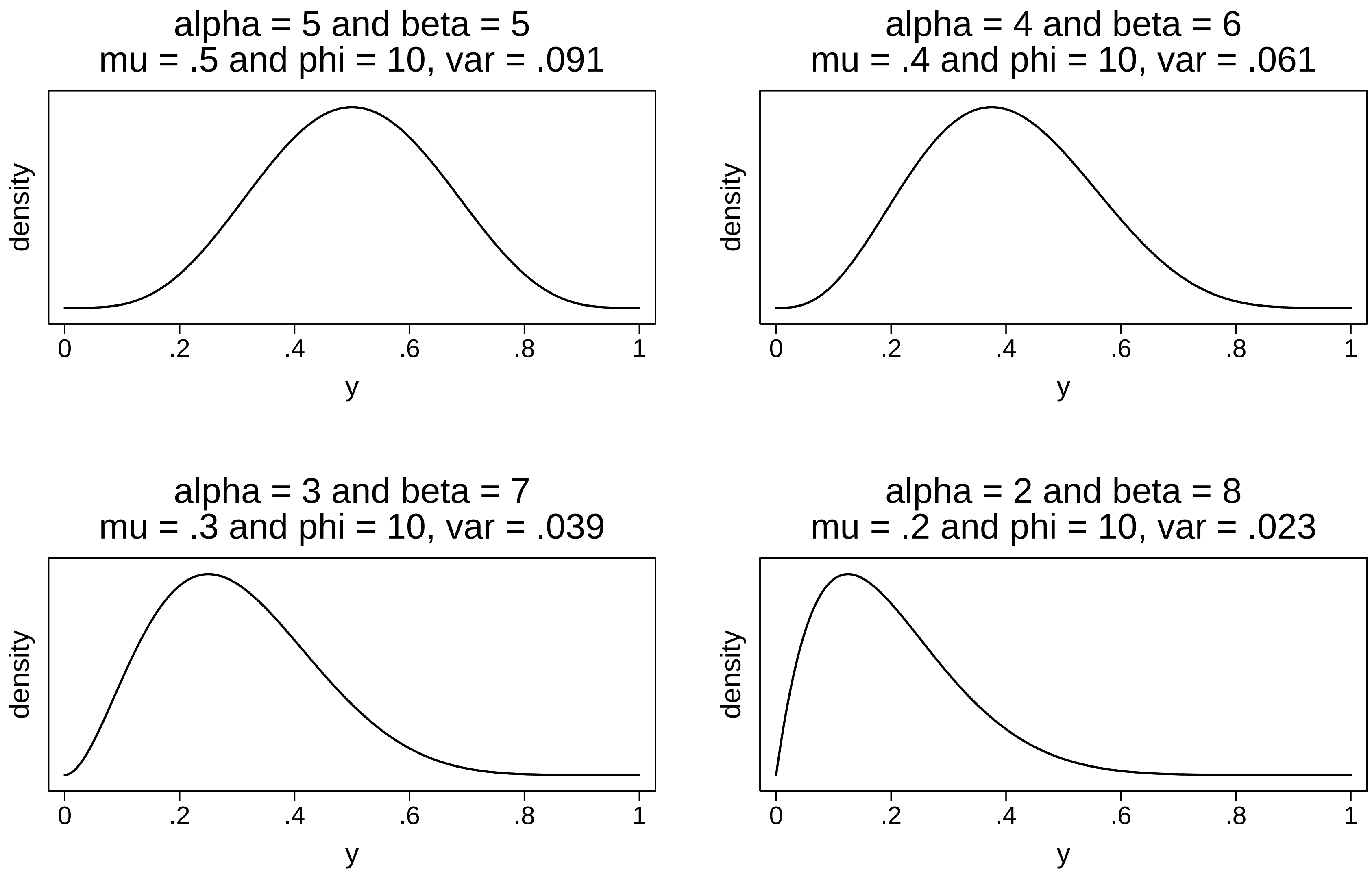
- conventional parametrization

$$\begin{aligned} f(y|\alpha,\beta) &\propto y^{\alpha-1}(y-1)^{\beta-1} \\ E(y) &= \frac{\alpha}{\alpha+\beta} \\ Var(y) &= \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)} \end{aligned}$$

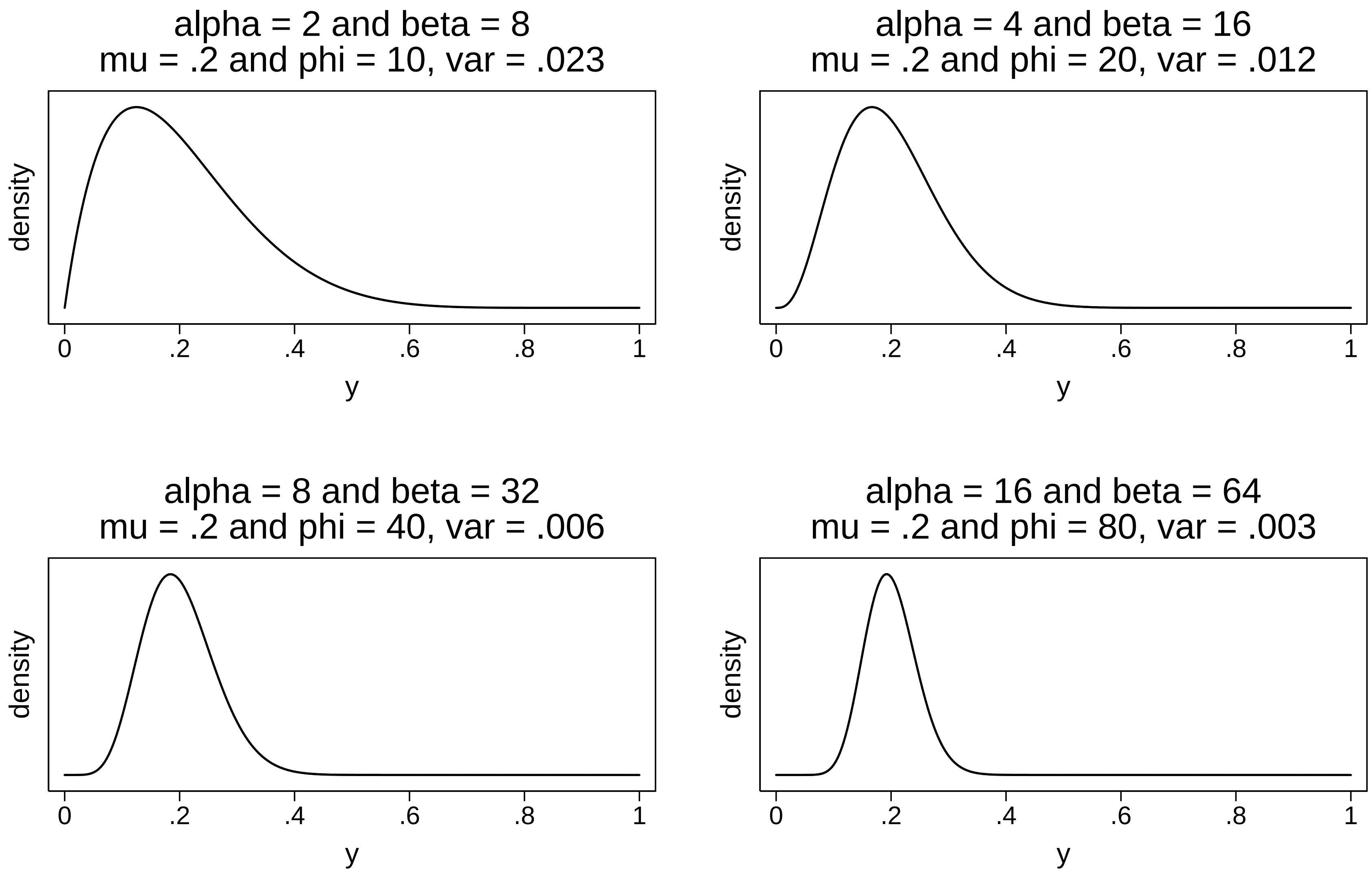
- alternative parametrization

$$\begin{aligned} f(y|\mu,\phi) &\propto y^{\mu\phi-1}(y-1)^{(1-\mu)\phi-1} \\ E(y) &= \mu \\ Var(y) &= \mu(1-\mu)\frac{1}{1+\phi} \end{aligned}$$

different μ fixed ϕ



different ϕ fixed μ



Modeling the mean

- We allow different cities to have different μs depending on their values of the explanatory variables.
- $\mu_i = f(b_0 + b_1 x_{1i} + b_2 x_{2i} \cdots)$
- The logistic transformation is used to ensure μ_i remains between 0 and 1.
- $\mu_i = \frac{e^{b_0 + b_1 x_{1i} + b_2 x_{2i} \cdots}}{1 + e^{b_0 + b_1 x_{1i} + b_2 x_{2i} \cdots}}$
- which is the same as:
- $\ln(\frac{\mu}{1-\mu}) = b_0 + b_1 x_{1i} + b_2 x_{2i} \cdots$

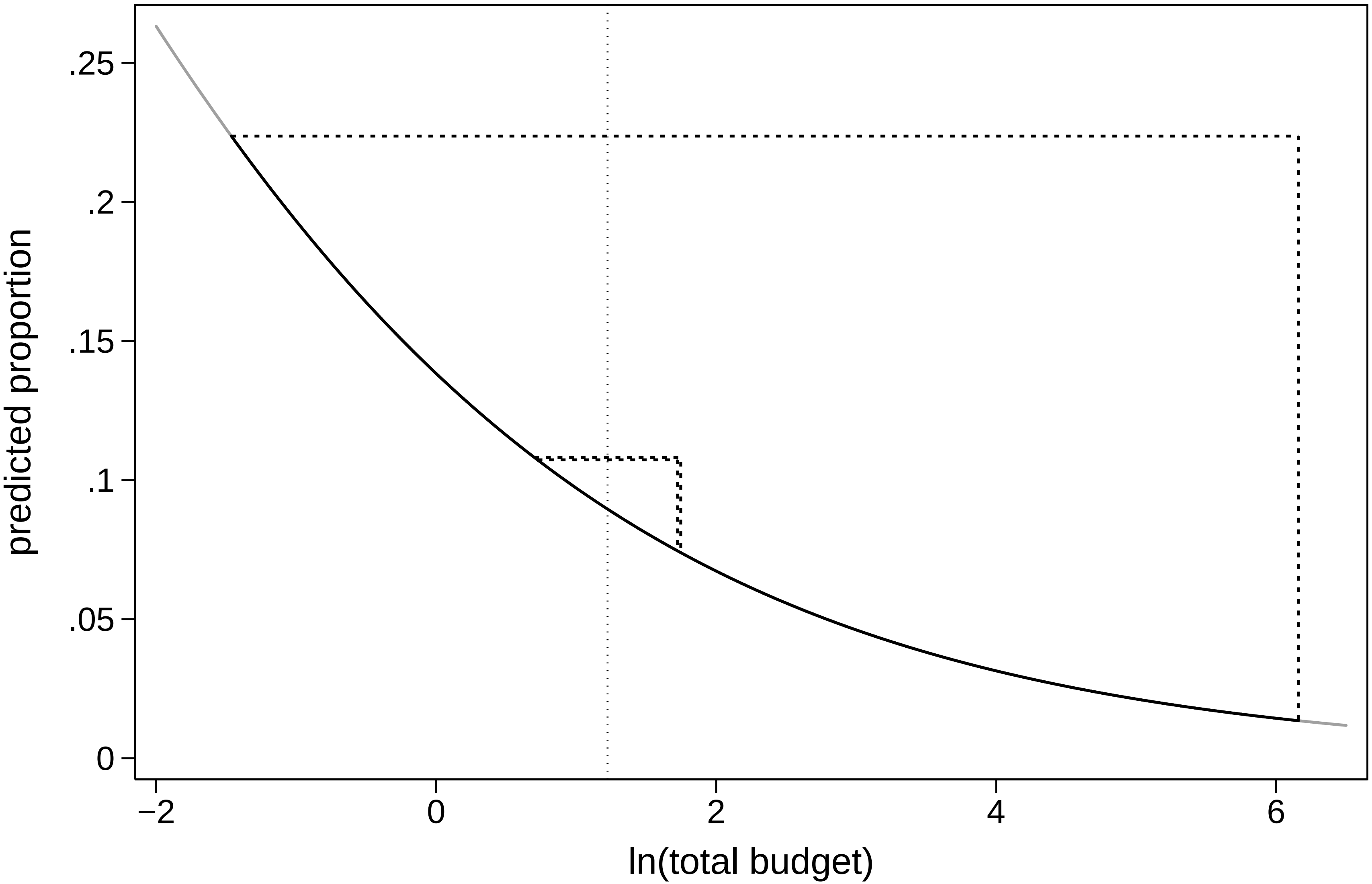
output of `betafit`

```
. betafit gov, mu(lntot houseval popdens noleft minorityleft ) nolog

ML fit of beta (mu, phi)          Number of obs =        394
                                  Wald chi2(5)  =     473.19
Log likelihood =  887.97456       Prob > chi2   =     0.0000

------------------------------------------------------------
             |   Coef.     se       z    P>|z|  [   95% CI   ]
-------------+----------------------------------------------
       lntot |  -.3999  .0227  -17.58  0.000  -.4445  -.3553
    houseval |   .1138  .0385    2.96  0.003   .0384   .1892
     popdens |   .0830  .0216    3.85  0.000   .0408   .1253
      noleft |   .0185  .0445    0.42  0.677  -.0686   .1057
minorityleft |  -.0080  .0450   -0.18  0.859  -.0962   .0802
       _cons | -2.0545  .0707  -29.06  0.000 -2.1931 -1.9160
-------------+----------------------------------------------
     /ln_phi |  4.7968  .0715   67.13  0.000  4.6568  4.9368
-------------+----------------------------------------------
         phi |   121.1 8.6545                 105.3    139.3
------------------------------------------------------------
```

interpretation using `dbetafit`

```
. dbetafit , at(noleft 0 minorityleft 0)
------------------------------------------------------------------
discrete       |  Min --> Max          +-SD/2           +-1/2
change         |  coef.      se       coef.      se    coef.      se
--------------+---------------------------------------------------
         lntot |-.2116   .0122     -.0344     .002   -.033   .0019
      houseval | .0291   .0105      .0037    .0013   .0093   .0032
       popdens | .0447   .0133      .0063    .0016   .0068   .0018
        noleft | .0015   .0037
  minorityleft |-6.6e-04   .0037
------------------------------------------------------------------
```

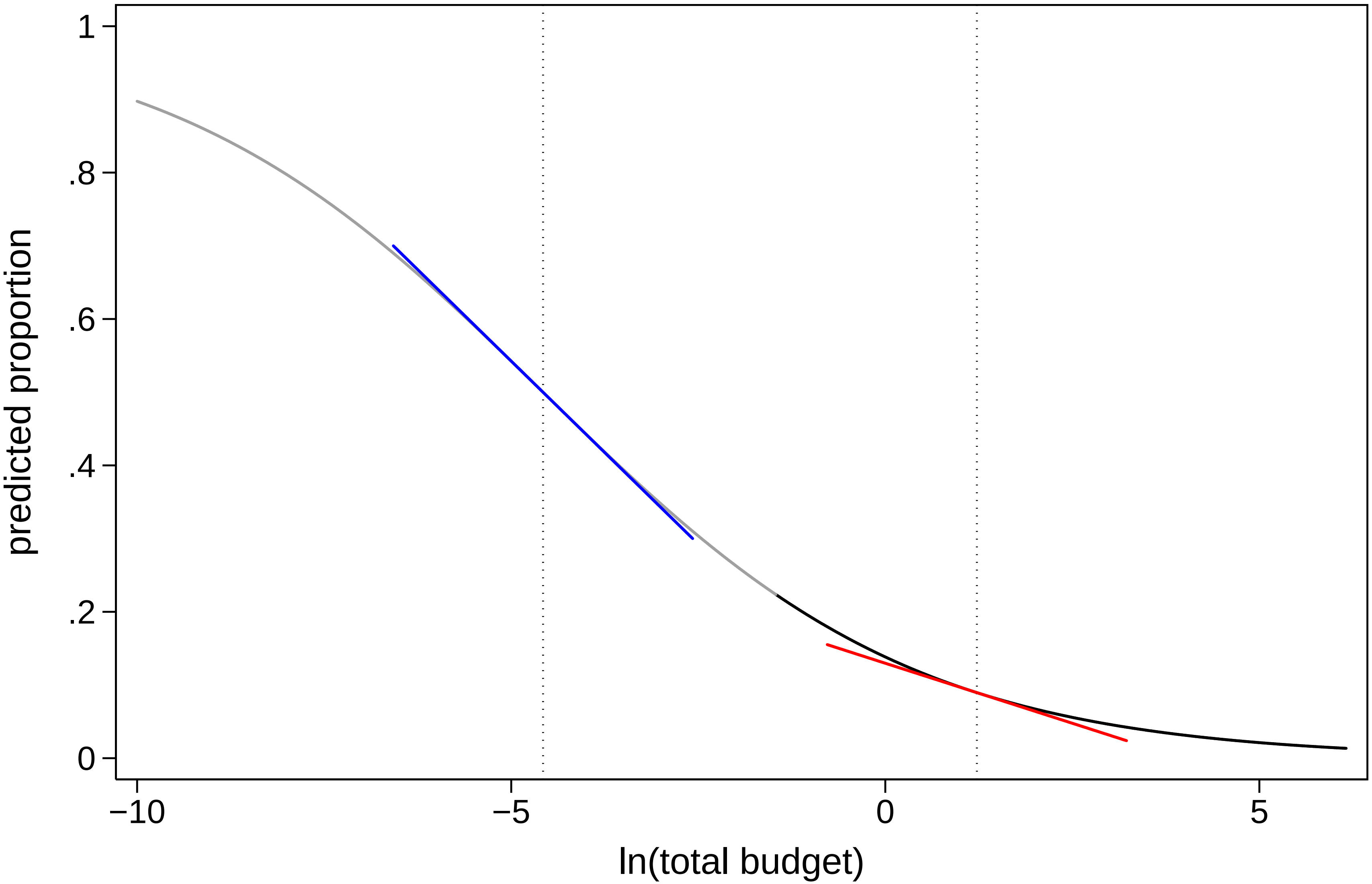
discrete changes in lntot



marginal effects

Marginal Effects	MFX at x coef.	MFX at x se	Max MFX coef.	Max MFX se
lntot	-.0328	.0019	-.1	.0057
houseval	.0093	.0032	.0284	.0096
popdens	.0068	.0018	.0208	.0054

marginal effects of lntot



Fractional logit

- Although the implied variance in `betafit` makes sense, it is still an assumption and some think it is too restrictive.
- The fractional logit has been proposed as an alternative by Papke and Wooldridge (1996).
- Fractional logit can handle proportions of exactly 0 or 1, unlike `betafit`.
- This model can be estimated by typing: `glm` *`varlist`*`, family(binomial) link(logit) robust`.
- Marginal effects like those from `dbetafit` can be obtained with `mfx, predict(mu)`.

Does it matter?

	OLS		betafit		glm	
	dy/dx	se	dy/sx	se	dy/dx	se
lntot	-.0296	.0027	-.0328	.0019	-.0330	.0026
houseval	.0135	.0051	.0093	.0032	.0105	.0036
popdens	.0078	.0019	.0068	.0018	.0071	.0018
noleft*	-.0010	.0056	.0015	.0037	.0008	.0046
minorityleft*	-.0065	.0047	-.0007	.0037	-.0019	.0042

* dy/dx is for discrete change of dummy variable from 0 to 1

Outline

- Problems with using `regress` for proportions as dependent variable
- Methods for dealing with a single proportion
- Methods for dealing with multiple proportions
- Caveat: Ecological Fallacy

Multiple proportions

Cities also spent money on other categories:

- Safety (which includes public health, fire department, and the police department)
- Education (mostly primary and secondary schools)
- recreation (which includes sport facilities and culture)
- social (which includes social work and some social security benefits)
- urbanplanning (which includes roads and houses)

Multiple proportions

- The proportions spent on each category should remain between 0 and 1, *and*
- the proportions should add up to 1.
- The proportions could be modeled with separate `betafit` models.
- This would ensure the first condition is met, *but*
- it would ignore the second condition.

A solution: `dirifit`

- Assumes that the proportions follow a Dirichlet distribution.
- The Dirichlet distribution is the multivariate generalization of the beta distribution.
- It ensures that the proportions remain between 0 and 1, *and* that they add up to 1.

Two parameterizations

- the conventional parametrization with one shape parameters for each proportion (α_1, α_2, . . ., α_k)
  - Corresponds to the formulas of the Dirichlet distribution in textbooks.
  - Does not correspond to conventions of Generalized Linear Models where one models how the mean of the distribution of the dependent variable changes as the explanatory variables change.
- the alternative parametrization with on location location parameter for each proportion and one scale parameter (μ_1, μ_2, . . ., μ_k, and ϕ)
  - Does not correspond to textbook formulas of the Dirichlet distribution but does correspond to the GLM convention.
  - One location parameter is redundant: $\mu_1 = 1 - (\mu_2 + \mu_3 + \ldots + \mu_k)$.

Modeling the mean

- We allow different cities to have different μ_js depending on their values of the explanatory variables.
- The multinomial logistic transformation is used to ensure the μ_js remain between 0 and 1 and add up to 1.

output of dirifit

```
. dirifit gov-urban, mu(lntot houseval popdens noleft minorityleft ) nolog

-----------------------------------------------------------------
             |   Coef.     se        z     P>|z|    [    95% CI   ]
-------------+---------------------------------------------------
mu2          |
       lntot |   .1445   .0406    3.56   0.000    .0649    .2240
    houseval |  -.0518   .0718   -0.72   0.471   -.1924    .0889
     popdens |  -.0700   .0390   -1.79   0.073   -.1465    .0065
      noleft |   .0817   .0827    0.99   0.323   -.0805    .2439
minorityleft |   .1043   .0826    1.26   0.207   -.0577    .2662
       _cons |   .5274   .1318    4.00   0.000    .2690    .7858
-------------+---------------------------------------------------
mu3          |
       lntot |   .4123   .0423    9.74   0.000    .3293    .4952
<snip>
-------------+---------------------------------------------------
         phi |   45.01   1.407                     42.33    47.85
-----------------------------------------------------------------
mu2 = safety          mu4 = recreation  mu6 = urbanplanning
mu3 = education       mu5 = social      base outcome = governing
```

Marginal effects obtained with `ddirifit`

	governing	safety	education	recreation	social	urban planning
lntot	-.0320*	-.0314*	.0115*	-.0067*	.0265*	.0321*
houseval	.0132*	.0143*	-.0321*	.0065	-.0496*	.0477*
popdens	.0074*	.0009	-.0067	.0002	.0072	-.0090*
noleft†	.0006	.0161*	-.0266*	.0048	-.0168	.0219*
minorityleft†	-.0019	.0154	-.0164*	.0085	-.0105	.0049

† discrete change of dummy variable from 0 to 1

* significant at 5% level

Variance and covariance of y in `dirifit`

- The variance of y_i is $\mu_i(1-\mu_i)\frac{1}{1+\phi}$
- The covariance of y_i and y_j implicit in `dirifit` is $-\mu_i\mu_j\frac{1}{1+\phi}$
- It depends on the means in a similar fashion as the multinomial distribution, and on a precision parameter ϕ.
- Covariance is forced to be negative. This makes sense in that there is less room for other categories if the fraction in one category increases.

Variance Covariance structure too restrictive?

- Though the implied variances and covariances make sense, they do not have to be true.
- Alternatives have been proposed for cases where this structure is violated.
- For `dirifit` a multivariate normal model for logit transformed dependent variables has been proposed by Aitcheson (2003).

Variance Covariance structure too restrictive?

This model can be estimated by typing:

```
gen logity1 = logit(y1)
gen logity2 = logit(y2)
.
.
gen logityk = logit(yk)

mvreg logity1 - logityk = indepvars, corr
```

Outline

- Problems with using `regress` for proportions as dependent variable
- Methods for dealing with a single proportion
- Methods for dealing with multiple proportions
- Caveat: Ecological Fallacy

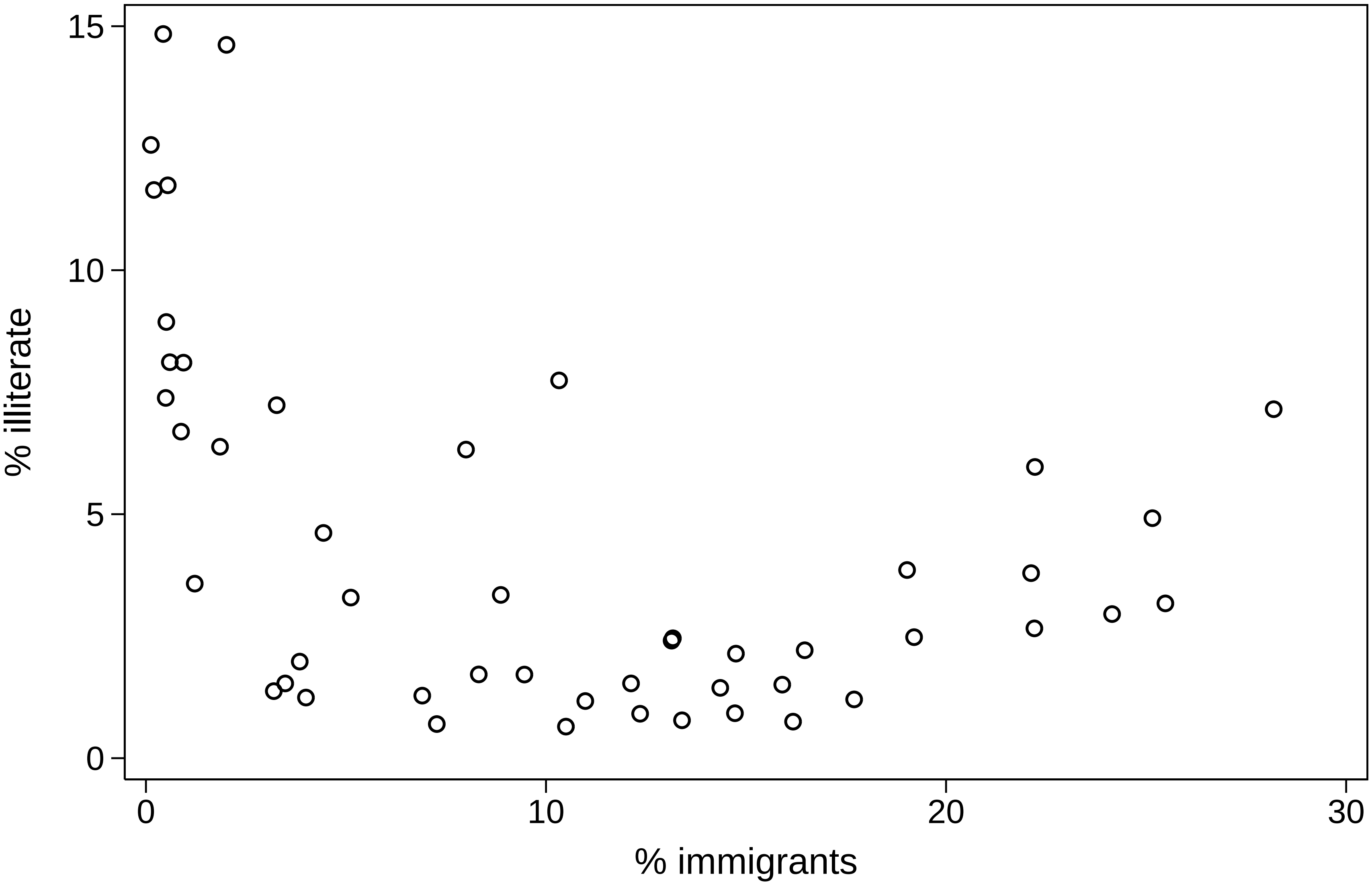
Ecological Fallacy

- Sometimes one wants to study behavior of individuals but one only has information on a aggregate level.
- This aggregate information is often in the form of proportions.
- One might be tempted to use the methods discussed previously to analyze this data.
- Example from Robinson (1950): Relationship between immigrant status and literacy in the 1930 US census.

Individual level analysis

	illiterate		
immigrant	literate	illiterate	Total
native born	96.72	3.28	100.00
foreign born	90.75	9.25	100.00
Total	95.87	4.13	100.00

State level analysis



Ecological Fallacy

- Aggregate level relationships can be completely different from individual level relationships.
- If it is remotely possible to use individual level data, do so!
- If that is not possible start reading up on Ecological Inference. A good place to start is Gary King (1997)
- `Ecol` package from Department of Political Science, Aarhus University, Denmark: `http://www.ps.au.dk/stata/`

Summary (1)

- The constraint that a proportion must remain between 0 and 1 causes problems with `regress`.
- `betafit` is one possible solution.
- Multiple proportions have the additional constraint that they must add up to 1.
- `dirifit` is one possible solution.

Summary (2)

- Both `betafit` and `dirifit` make assumptions about the variance (covariance) structure of the dependent variable that does make sense but that some find too restrictive.
- Fractional logit and multivariate regression have been proposed as alternatives.
- None of these techniques are appropriate for studying individual behavior from aggregate data.

References

Aitcheson, John. 2003. *The Statistical Analysis of Compositional Data.* Blackburn Press.

King, Gary. 1997. *A solution to the Ecological Inference Problem, Reconstructing Individual Behavior from Aggregate Data.* Princeton University Press.

Papke, Leslie E. and Jeffrey M. Wooldridge. 1996. "Econometric Methods for Fractional Response Variables with an Application to 401(k) Plan Participation Rates." *Journal of Applied Econometrics* 11(6):619–632.

Robinson, W.S. 1950. "Ecological Correlations and the Behavior of Individuals." *Amercian Sociological Review* 15(3):351–357.