Introduction to fractional outcome regression models using the `fracreg` and `betareg` commands

Miguel Dorta

Staff Statistician
StataCorp LP

Aguascalientes, Mexico

Introduction to fractional outcome regression models using the `fracreg` and `betareg` commands

Miguel Dorta

Staff Statistician
StataCorp LP

Aguascalientes, Mexico

Outline

- Introduction
- **fracreg** – Fractional response regression
  - Concepts
  - Example
- **betareg** – Beta regression
  - Concepts
  - Example
- Conclusion
- Questions

Introduction

- From version 14, Stata includes the **`fracreg`** and **`betareg`** commands for fractional outcome regressions.
- Continuous dependent variables (y) in [0,1] or (0,1).
- We want to fit a regression for the mean of y conditional on x: $E(y|x)$.
- Some case studies where fractional regression has been applied.
  - 401(k) retirement plan participation rates (Papke and Wooldridge, 1996).
  - Test pass rates for exams on students (Papke and Wooldridge, 2008).
  - Gini index values for the prices of art (Castellani et al., 2012).
  - Probability of a defendant's guilt and the verdict (Smithson et al., 2007).

Introduction

- Why do we need regression methods for dependent variables in [0,1] or (0,1)?
  - Avoid model misspecification and dubious statistical validity.
  - If we simply use **`regress`**, predictions could fall outside those intervals.
  - **`fracreg`** and **`betareg`** captures particular non linear relationships, especially when the outcome variable is near 0 or 1.
- Dependent variables in that range:
  - Fractions
  - Proportions
  - Rates
  - Indices
  - Probabilities

`fracreg` – Fractional response regression – Concepts

`fracreg` – Fractional response regression – Concepts

- We have a continuous dependent variable y in [0,1], and a vector of independent variables (x).
- We want to fit a regression for the mean of y conditional on x: $E(y|x)$.
- Because y is in [0,1], we want to restrict that $E(y|x)$ is also in [0,1].
- **`fracreg`** accomplishes that by using the following models:
  - probit: $E(y|x) = \Phi(x\beta)$
  - heteroskedastic probit: $E(y|x) = \Phi\left(x\beta/exp(z\gamma)\right)$
  - logit: $E(y|x) = exp(x\beta)/(1 + exp(x\beta))$

`fracreg` – Fractional response regression – Concepts

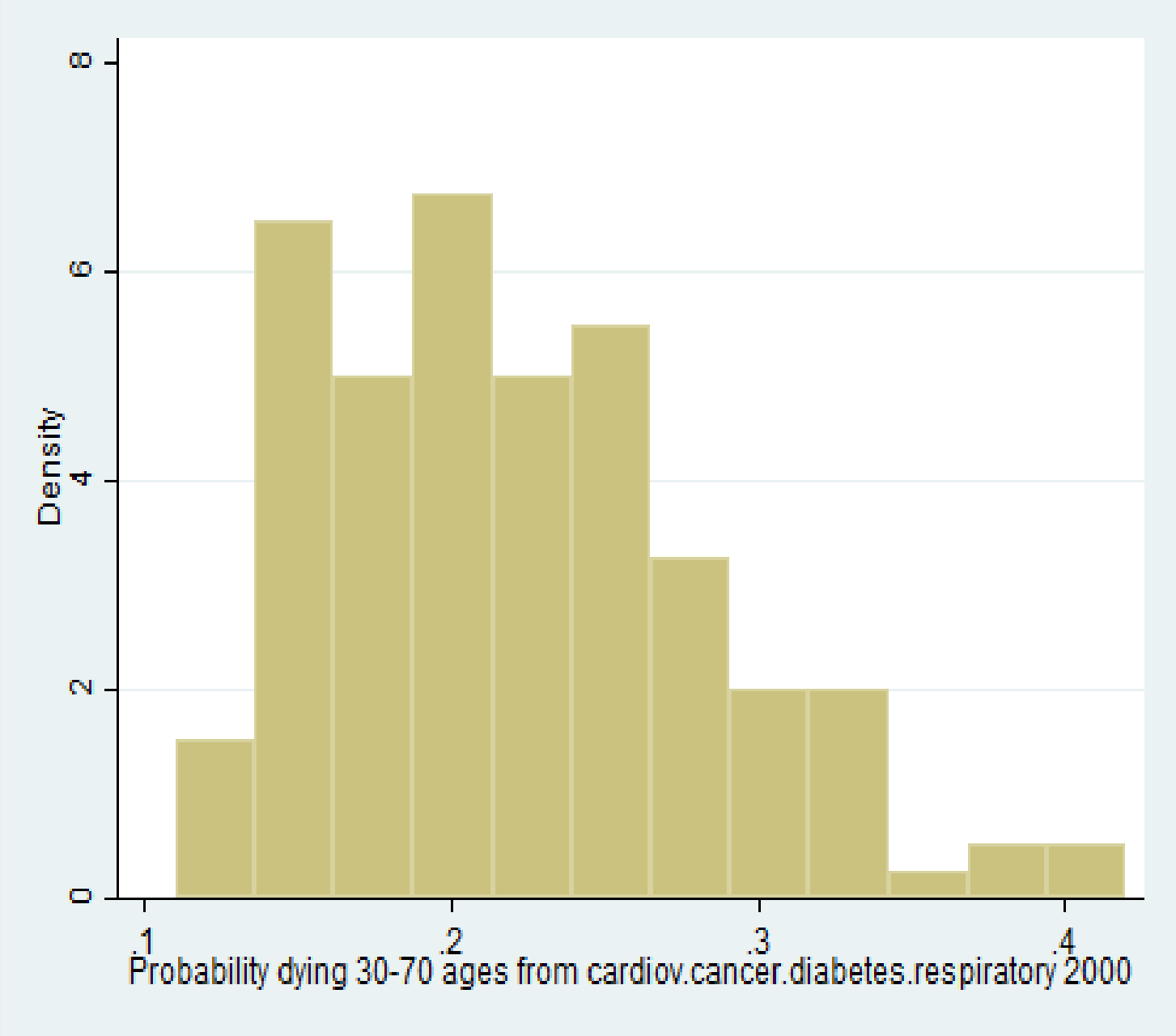
- `fracreg` implements quasilikelihood estimators.
  - No need to know the true distribution to obtain consistent parameter estimates.
  - We need the correct specification of the conditional mean.
  - `fracreg` computes robust standard errors by default.

An example with `fracreg`

An example with `fracreg`

- We are fitting a model for the conditional mean of the probability of dying between ages 30 and 70 from four important diseases (**prdying**) on a set of independent variables.
- Data on 155 countries (including Mexico) for year 2000.
- Independent variables:
  - **idwtotal**: Total population using improved drinking-water sources (tens of percentage points).
  - **pctexph**: Total expenditure per capita on health at average exchange rate (thousands of US$).
  - **gniperc**: Gross national income per capita (PPP thousands of US$).
  - **uvradiation**: Exposure to solar ultraviolet (UV) radiation (thousands of J/m2).
- Source: Global Health Observatory (GHO) data repository of the World Health Organization. http://www.who.int/gho/database/en/

An example with `fracreg`



An example with `fracreg`

```
. fracreg logit prdying idwtotal pctexph gniperc uvradiation, nolog

Fractional logistic regression                  Number of obs     =        155
                                                Wald chi2(4)      =      74.91
                                                Prob > chi2       =     0.0000
Log pseudolikelihood = -81.014058               Pseudo R2         =     0.0094

------------------------------------------------------------------------------
             |               Robust
     prdying |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    idwtotal |  -.0475306   .0174399    -2.73   0.006    -.0817122   -.013349
     pctexph |  -.2998815   .0759262    -3.95   0.000    -.4486941   -.1510689
     gniperc |   -.003473   .0032611    -1.06   0.287    -.0098647    .0029187
uvradiation  |  -.1367411   .0244849    -5.58   0.000    -.1847306   -.0887515
       _cons |  -.1831707   .2114469    -0.87   0.386    -.5975989    .2312576
------------------------------------------------------------------------------
```

An example with `fracreg`

```
. margins, dydx(*)
Average marginal effects                        Number of obs     =        155
Model VCE    : Robust
Expression   : Conditional mean of prdying, predict()
dy/dx w.r.t. : idwtotal pctexph gniperc uvradiation
```

	Delta-method					
	dy/dx	Std. Err.	z	P>|z|	[95% Conf.	Interval]
idwtotal	-.0080946	.0029576	-2.74	0.006	-.0138914	-.0022977
pctexph	-.0510706	.0128047	-3.99	0.000	-.0761673	-.0259739
gniperc	-.0005915	.0005565	-1.06	0.288	-.0016822	.0004992
uvradiation	-.0232874	.0041145	-5.66	0.000	-.0313517	-.015223

An example with `fracreg`

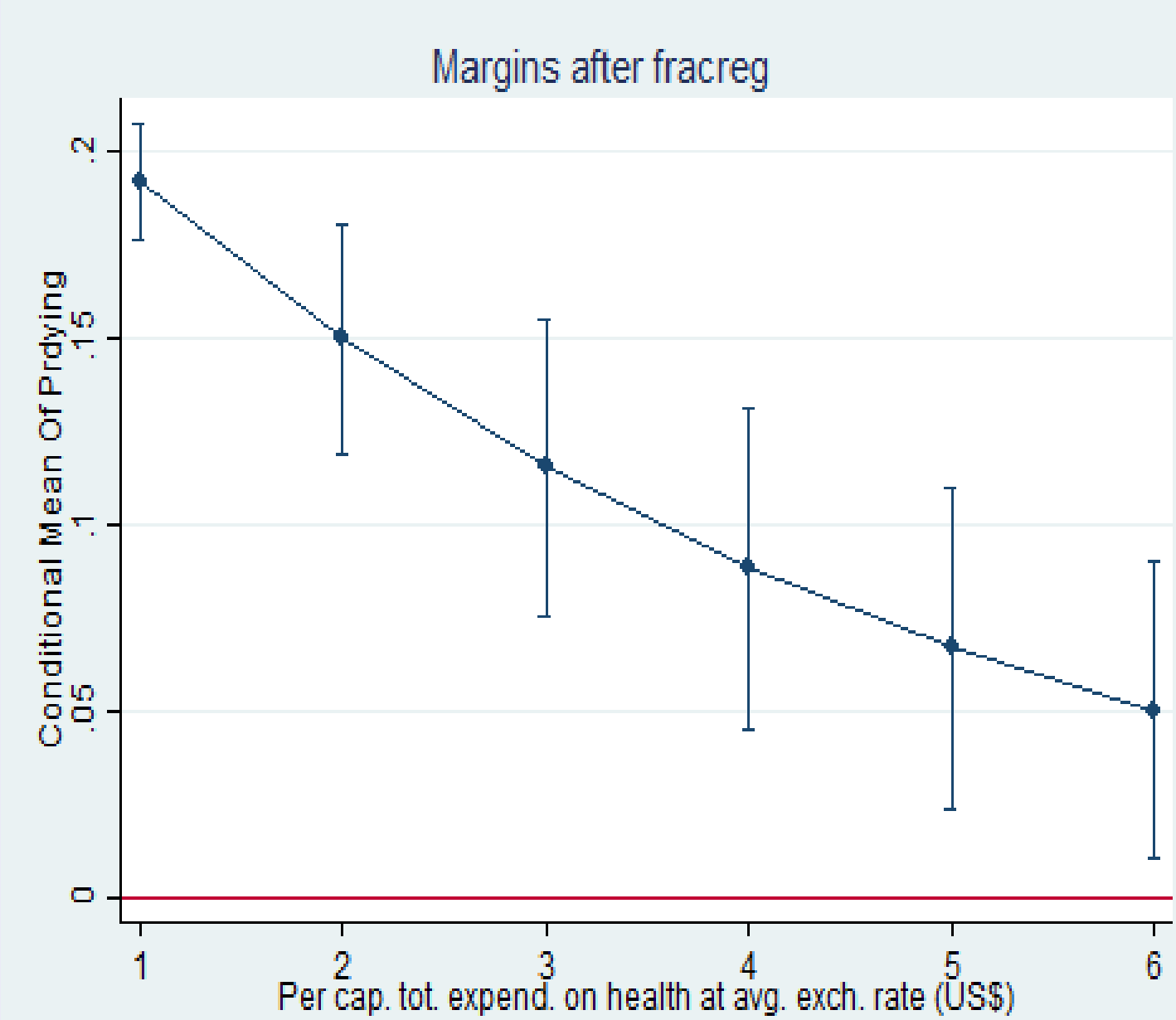
```
. margins, at(pctexph=(1(1)6)) noatlegend
Predictive margins                                Number of obs     =        155
Model VCE    : Robust
Expression   : Conditional mean of prdying, predict()
```

	Delta-method				
	Margin	Std. Err.	z	P>|z|	[95% Conf. Interval]
_at					
1	.1920181	.0078442	24.48	0.000	.1766437 .2073925
2	.1498362	.0157019	9.54	0.000	.1190611 .1806113
3	.1155769	.0202613	5.70	0.000	.0758654 .1552884
4	.0883243	.0220353	4.01	0.000	.045136 .1315126
5	.0670025	.0218357	3.07	0.002	.0242054 .1097997
6	.0505373	.0203957	2.48	0.013	.0105624 .0905122

```
. marginsplot, yline(0) title("Margins after fracreg")
  Variables that uniquely identify margins: pctexph
```

An example with `fracreg`

```
. marginsplot, yline(0) title("Margins after fracreg")
```



An example with `fracreg`

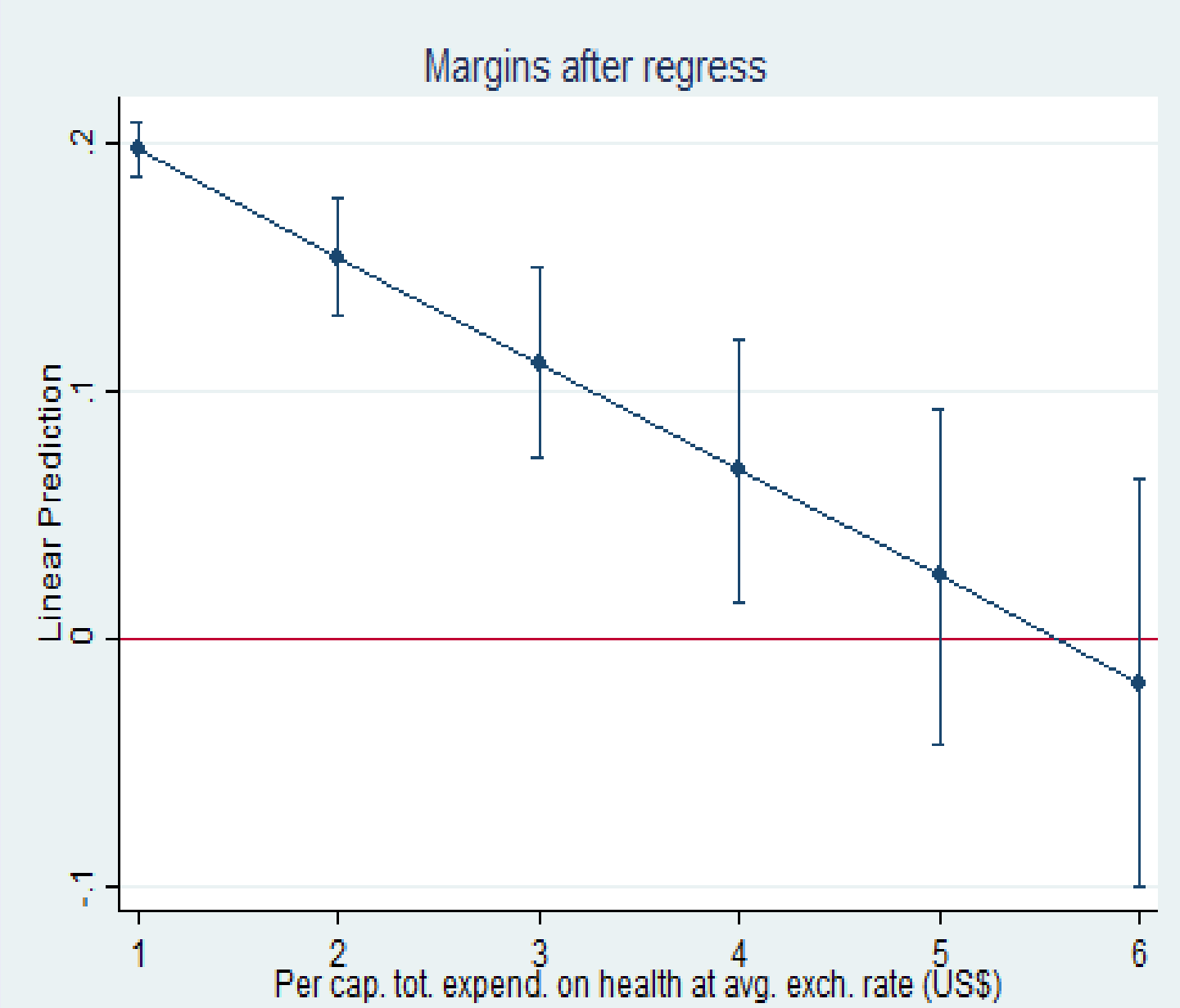
```
. qui regress prdying idwtotal pctexph gniperc uvradiation
. margins, at(pctexph=(1(1)6)) noatlegend
Predictive margins                                Number of obs     =        155
Model VCE    : OLS
Expression   : Linear prediction, predict()
```

	Delta-method					
	Margin	Std. Err.	t	P>|t|	[95% Conf.	Interval]
_at						
1	.1972132	.005612	35.14	0.000	.1861245	.208302
2	.1542242	.0121964	12.65	0.000	.1301253	.178323
3	.1112351	.0194517	5.72	0.000	.0728004	.1496699
4	.0682461	.0268393	2.54	0.012	.0152141	.121278
5	.025257	.0342738	0.74	0.462	-.0424647	.0929787
6	-.017732	.04173	-0.42	0.672	-.1001866	.0647225

```
. marginsplot, yline(0) title("Margins after regress")
  Variables that uniquely identify margins: pctexph
```

An example with `fracreg`

```
. marginsplot, yline(0) title("Margins after regress")
```



An example with `fracreg`

```
 . qui fracreg logit prdying idwtotal pctexph gniperc uvradiation
. estimates store flogit
. qui fracreg probit prdying idwtotal pctexph gniperc uvradiation
. estimates store fprobit
. qui fracreg probit prdying idwtotal pctexph gniperc uvradiation, ///
>      het(gniperc)
. estimates store fprobhet
. estimate stat flogit fprobit fprobhet
Akaike´s information criterion and Bayesian information criterion
```

Model	Obs	ll(null)	ll(model)	df	AIC	BIC
flogit	155	-81.78449	-81.01406	5	172.0281	187.2452
fprobit	155	-81.78449	-81.03322	5	172.0664	187.2836
fprobhet	155	-81.44187	-80.92097	6	173.8419	192.1025

Note: N=Obs used in calculating BIC; see [R] BIC note.

`betareg` – Beta regression – Concepts

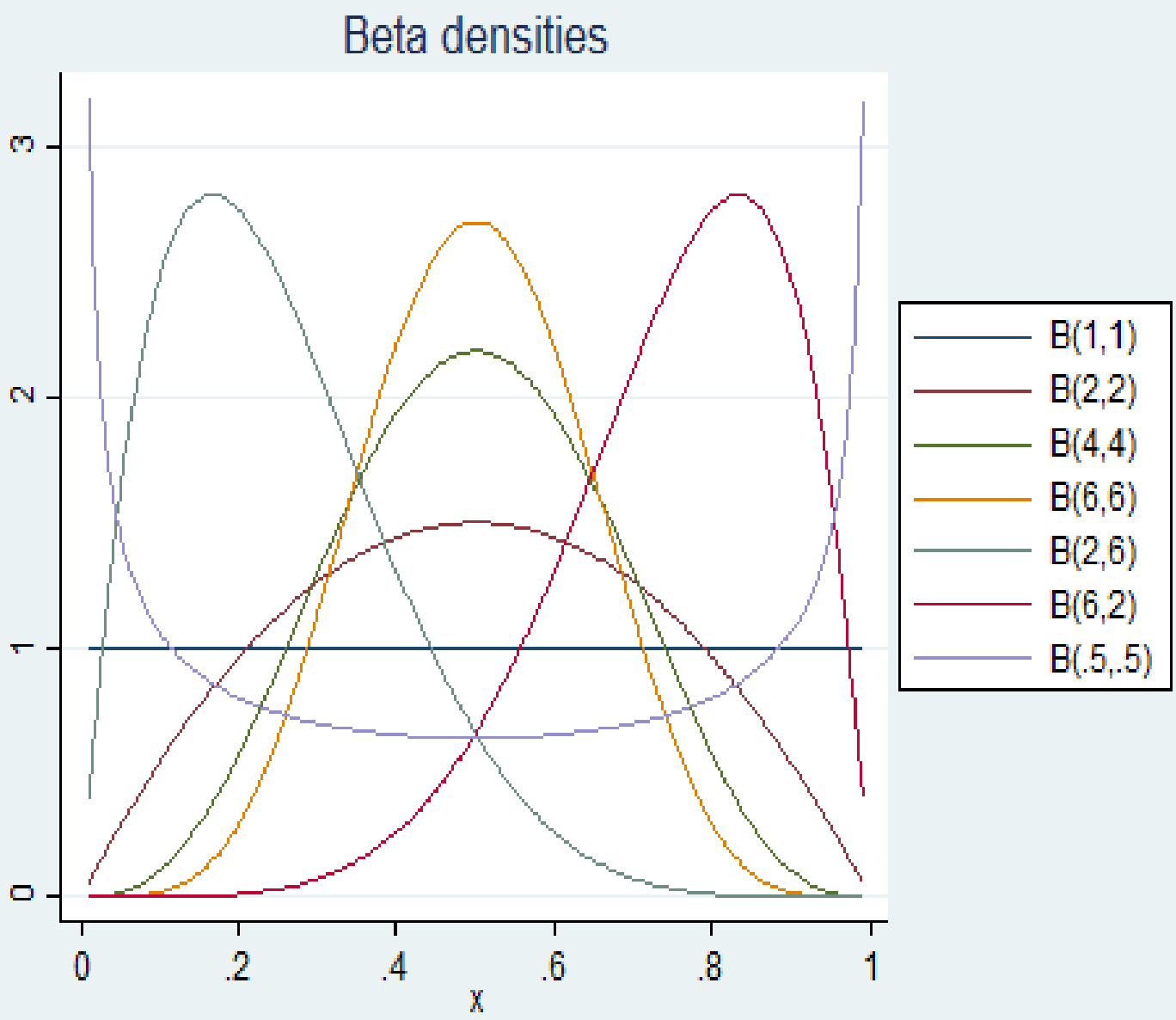
`betareg` – Beta regression – Concepts

- We have a continuous dependent variable y in (0,1), and a vector of independent variables (x).
- We need to fit a model for the mean of y conditional on x:

$$E(y/x) = \mu_x$$

- μ_x follows a Beta distribution; and therefore, μ_x must be in (0,1).
- **`betareg`** implements maximum likelihood estimators.
- The Beta distribution covers a wide spectrum of density shapes.

betareg – Beta regression – Concepts



`betareg` – Beta regression – Concepts

- **`betareg`** uses links functions $g(\mu_x) = x\beta$ so that $\mu_x = g^{-1}(x\beta)$ is in (0,1)
- By default, **`betareg`** works with the logit link

$$ln[\mu_x/(1-\mu_x)] = x\beta$$

$$\Rightarrow \mu_x = exp(x\beta)/(1+exp(x\beta))$$

- Link functions available:
  - logit: $g(\mu_x) = ln[\mu_x/(1-\mu_x)]$
  - probit: $g(\mu_x) = \Phi^{-1}(\mu_x)$
  - cloglog: $g(\mu_x) = ln[-ln(1-\mu_x]$
  - loglog: $g(\mu_x) = -ln[-ln(\mu_x)]$

`betareg` – Beta regression – Concepts

- The conditional variance of the beta distribution is

$$Var(y/x) = \mu_x(1-\mu_x)/(1+\psi_x)$$

- The parameter ψ_x rescales the conditional variance. We may use scale-link functions to restrict that $\psi_x > 0$:

$$h(\psi_x) = x\gamma$$

- Scale-link functions available:
  - log: $h(\psi_x) = ln(\psi_x)$ (default)
  - root: $h(\psi_x) = \sqrt{\psi_x}$
  - identity: $h(\psi_x) = \psi_x$

An example with `betareg`

An example with `betareg`

- Now, we are going to use **`betareg`** for fitting the previous model: the conditional mean of **prdying** on the same set of independent variables.
- Data on 155 countries (including Mexico) for year 2000.
- Independent variables:
  - **idwtotal**: Total population using improved drinking-water sources (tens of percentage points).
  - **pctexph**: Total expenditure per capita on health at average exchange rate (thousands of US$).
  - **gniperc**: Gross national income per capita (PPP thousands of US$).
  - **uvradiation**: Exposure to solar ultraviolet (UV) radiation (thousands of J/m2).
- Source: Global Health Observatory (GHO) data repository of the World Health Organization. http://www.who.int/gho/database/en/

An example with **betareg**

```
. betareg prdying idwtotal pctexph gniperc uvradiation, ///
>     nolog link(cloglog)
Beta regression                                 Number of obs     =        155
                                                LR chi2(4)        =      98.72
                                                Prob > chi2       =     0.0000

Link function  :  g(u) = log(-log(1-u))         [Comp. log-log]
Slink function :  g(u) = log(u)                 [Log]
Log likelihood =  266.78962
```

prdying	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
prdying						
idwtotal	-.0434229	.0136457	-3.18	0.001	-.0701681	-.0166778
pctexph	-.2896986	.0472833	-6.13	0.000	-.3823722	-.1970249
gniperc	-.002445	.0031457	-0.78	0.437	-.0086106	.0037205
uvradiation	-.1258703	.0177798	-7.08	0.000	-.160718	-.0910225
_cons	-.4028858	.1553157	-2.59	0.009	-.7072989	-.0984727
scale						
_cons	4.478092	.1131977	39.56	0.000	4.256229	4.699956

An example with **betareg**

```
. margins, dydx(*)
Average marginal effects                        Number of obs     =        155
Model VCE    : OIM
Expression   : Conditional mean of prdying, predict()
dy/dx w.r.t. : idwtotal pctexph gniperc uvradiation
```

	Delta-method					
	dy/dx	Std. Err.	z	P>|z|	[95% Conf.	Interval]
idwtotal	-.0083901	.0026356	-3.18	0.001	-.0135556	-.0032245
pctexph	-.0559747	.009124	-6.13	0.000	-.0738574	-.038092
gniperc	-.0004724	.0006078	-0.78	0.437	-.0016637	.0007189
uvradiation	-.0243203	.0034275	-7.10	0.000	-.0310382	-.0176024

An example with **betareg**

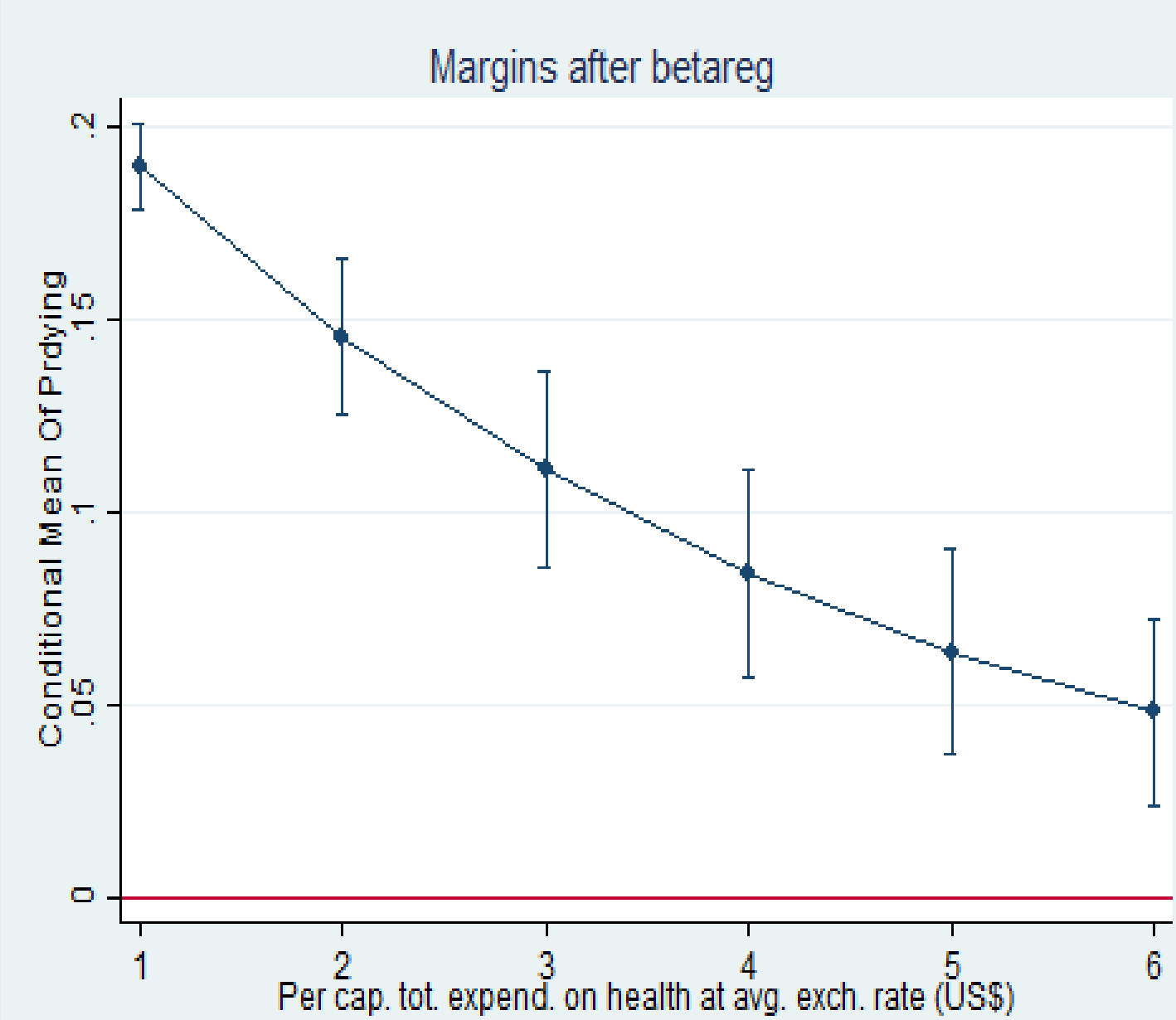
```
. margins, at(pctexph=(1(1)6)) noatlegend
Predictive margins                                Number of obs     =        155
Model VCE    : OIM
Expression   : Conditional mean of prdying, predict()
```

	Delta-method					
	Margin	Std. Err.	z	P>|z|	[95% Conf.	Interval]
_at						
1	.1896114	.0056458	33.58	0.000	.1785457	.200677
2	.1456751	.0103518	14.07	0.000	.1253859	.1659643
3	.1112062	.0129212	8.61	0.000	.0858811	.1365312
4	.0844814	.0137514	6.14	0.000	.0575292	.1114337
5	.0639434	.0134357	4.76	0.000	.0376098	.0902769
6	.048264	.0124457	3.88	0.000	.023871	.072657

```
. marginsplot, yline(0) title("Margins after betareg")
  Variables that uniquely identify margins: pctexph
```

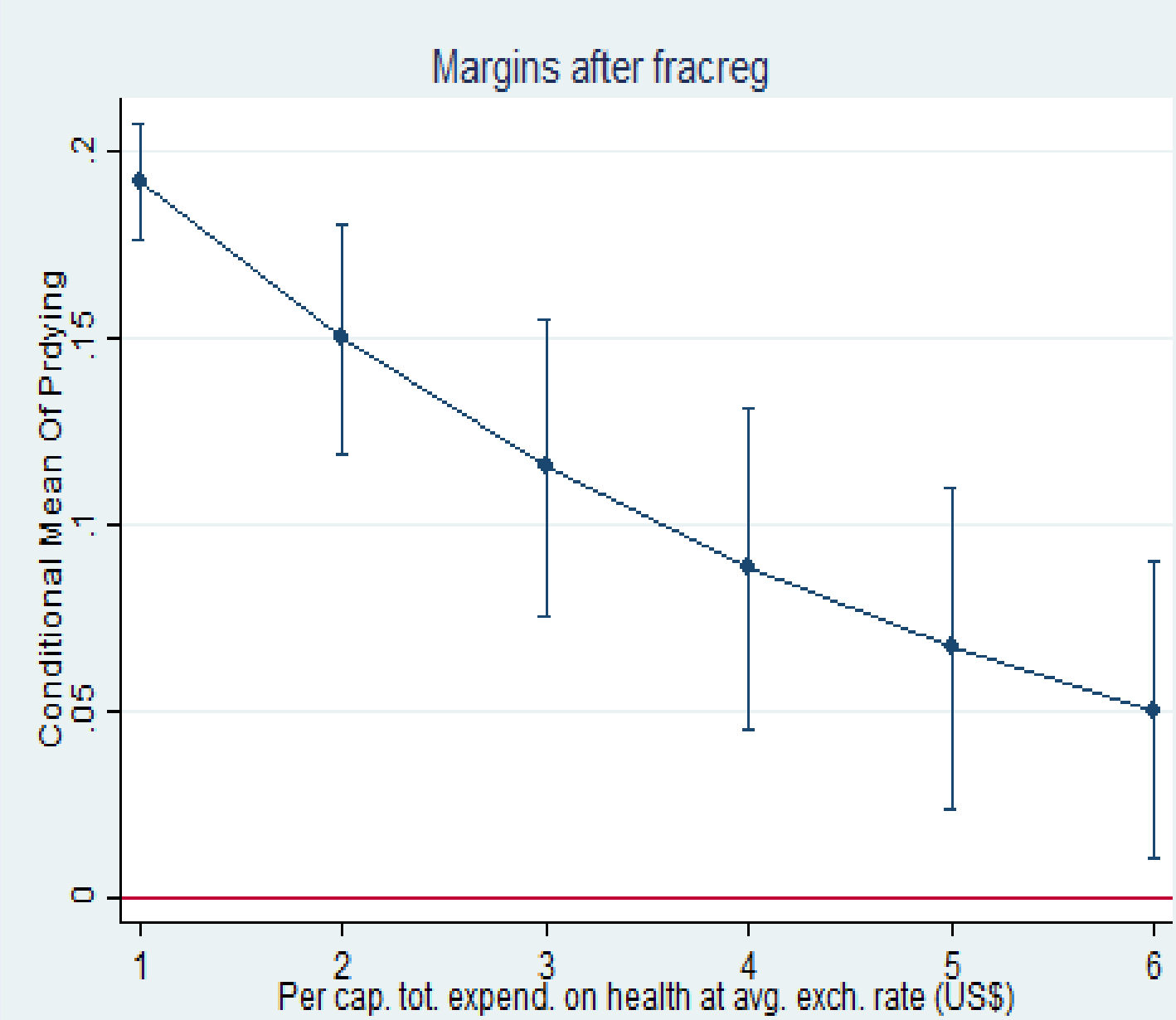
An example with **betareg**

```
. marginsplot, yline(0) title("Margins after betareg")
```



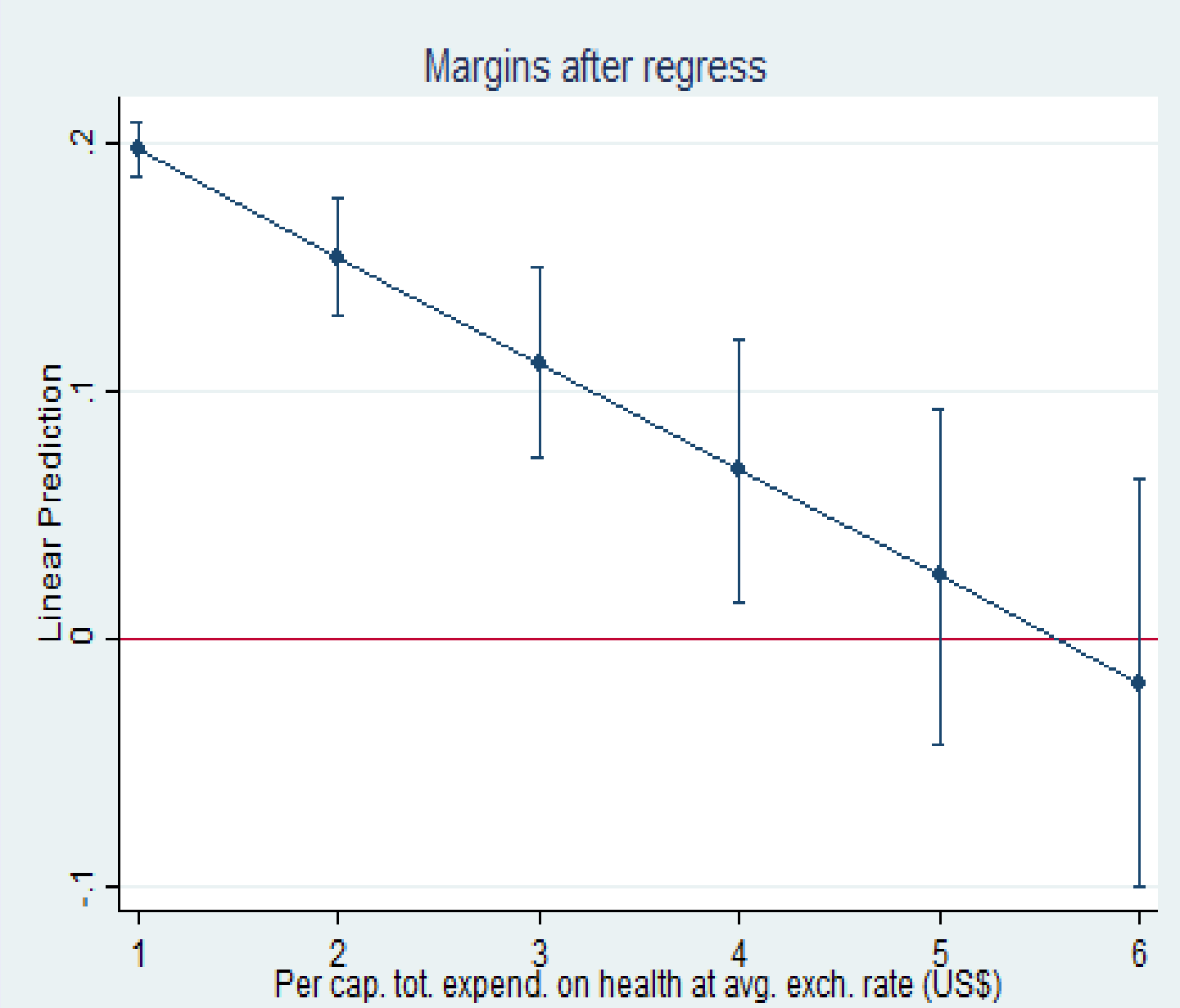
An example with **betareg**

```
. marginsplot, yline(0) title("Margins after fracreg")
```



An example with **betareg**

```
. marginsplot, yline(0) title("Margins after regress")
```



An example with **betareg**

```
 . qui betareg prdying idwtotal pctexph gniperc uvradiation
. estimates store blogit
. qui betareg prdying idwtotal pctexph gniperc uvradiation, ///
>     link(probit)
. estimates store bprobit
. qui betareg prdying idwtotal pctexph gniperc uvradiation, ///
>     link(cloglog)
. estimates store bcloglog
. qui betareg prdying idwtotal pctexph gniperc uvradiation, ///
>     link(loglog)
. estimates store bloglog
. estimate stat blogit bprobit bcloglog bloglog
Akaike´s information criterion and Bayesian information criterion
```

Model	Obs	ll(null)	ll(model)	df	AIC	BIC
blogit	155	217.431	265.7818	6	-519.5636	-501.303
bprobit	155	217.431	264.3145	6	-516.6291	-498.3685
bcloglog	155	217.431	266.7896	6	-521.5792	-503.3187
bloglog	155	217.431	262.1897	6	-512.3793	-494.1188

Note: N=Obs used in calculating BIC; see [R] BIC note.

Conclusion

- From version 14, Stata includes the **fracreg** and **betareg** regression commands for dependent variables in [0,1] and (0,1) respectively.
- Models specified and fitted with these commands are more appropriate than using **regress** when the dependent variables are in [0,1] or (0,1).
- **fracreg** and **betareg** guarantee that predictions will be in the correct intervals.
- **fracreg** computes quasilikelihood estimators based on probit or logit. Simpler but less flexible likelihood specification.
- **betareg** computes maximum likelihood estimators based on the beta distribution. Complex but likelihood specification adaptable to a wide spectrum of density shapes.
- The original coefficients are not very useful; and so, the **margins** command becomes an important tool for interpreting results after models fitted with **fracreg** or **betareg**.

It was a pleasure! Thank you!

- Any questions?