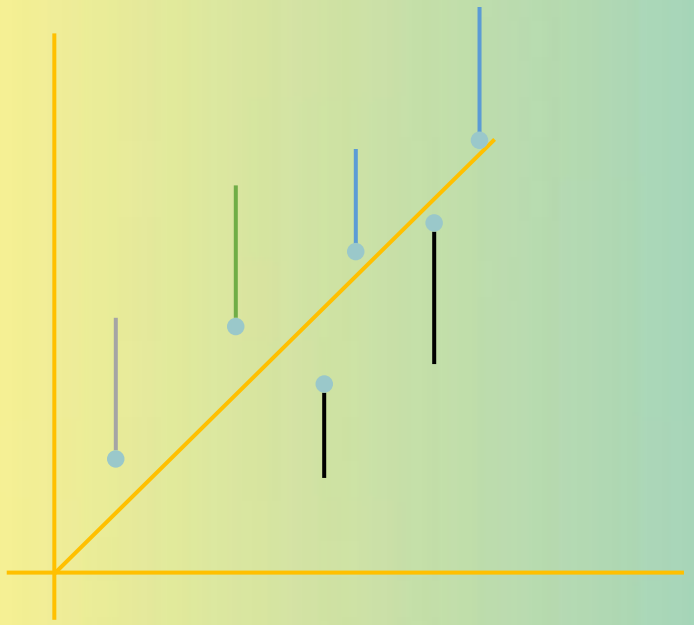
How to choose the right type of regression method

Bishwajit Ghose



Type of dependent variable

Numerical

Discrete

Number/ count of something that cannot be fractioned. e.g. days in a week.

Poisson/ Negative binomial

Continuous

Number/ count of something that cannot be fractioned. e.g. age, height, weight, BMI.

Ordinary least squares (OLS), Quantile regression.

Categorical

Nominal

No natural order. e.g. type of sex, nationality.

Binary (2 categories)

Multinomial (>2 categories)

Ordinal

Possible to assign order. e.g. degree of life satisfaction.

Ordered logistic (>2 categories)

Type of regression

Numerical ➡ Discrete

- Poisson regression
  - Dependent variable:
    household size (houssiz)
  - Independent variables:
    race (categorical)
    region (categorical)
    rural (categorical)

Incidence rate ratio

```
. webuse nhanes2

. poisson houssiz  i.race i.region i.rural, irr nolog cformat(%9.2f) baselevels

Poisson regression                              Number of obs     =     10,351
                                                LR chi2(6)        =     229.90
                                                Prob > chi2       =     0.0000
Log likelihood =  -19160.45                     Pseudo R2         =     0.0060
```

houssiz	IRR	Std. Err.	z	P>|z|	[95% Conf.	Interval]
race						
White	1.00	(base)				
Black	1.22	0.02	10.79	0.000	1.18	1.27
Other	1.45	0.05	10.22	0.000	1.35	1.56
region						
NE	1.00	(base)				
MW	0.95	0.02	-2.94	0.003	0.92	0.98
S	0.96	0.02	-2.30	0.022	0.93	0.99
W	0.94	0.02	-3.34	0.001	0.91	0.98
rural						
0	1.00	(base)				
1	1.10	0.01	7.63	0.000	1.07	1.13
_cons	2.86	0.04	79.22	0.000	2.79	2.94

Numerical ➡ Continuous

- OLS regression
  - Dependent variable:
    diastolic bp (bpdiast)
  - Independent variables:
    age (continuous)
    sex (categorical)
    bmi (continuous)
    race (categorical)

```
. webuse nhanes2

. reg bpsystol age i.sex bmi i.race, cformat(%9.2f)
```

Source	SS	df	MS		
Model	1767281.35	5	353456.27	Number of obs	= 10,351
Residual	3867388.67	10,345	373.841341	F(5, 10345)	= 945.47
				Prob > F	= 0.0000
				R-squared	= 0.3136
				Adj R-squared	= 0.3133
Total	5634670.03	10,350	544.412563	Root MSE	= 19.335

bpsystol	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
age	0.59	0.01	52.75	0.000	0.57	0.61
sex						
Female	-4.07	0.38	-10.70	0.000	-4.82	-3.33
bmi	1.29	0.04	32.79	0.000	1.22	1.37
race						
Black	2.71	0.62	4.35	0.000	1.49	3.94
Other	2.06	1.38	1.49	0.136	-0.65	4.78
_cons	71.52	1.09	65.70	0.000	69.38	73.65

Numerical ➡ Continuous

- Quantile regression
  - Dependent variable:
    bmi
  - Independent variables:
    age (continuous)
    sex (categorical)
    race (categorical)
    rural (categorical)

```
. webuse nhanes2

. sqreg bmi age i.sex i.race i.rural, quantile(.25 .75) cformat(%9.2f)
(fitting base model)

Bootstrap replications (20)
----+--- 1 ---+--- 2 ---+--- 3 ---+--- 4 ---+--- 5
.....................

Simultaneous quantile regression                Number of obs =     10,351
  bootstrap(20) SEs                             .25 Pseudo R2 =     0.0372
                                                .75 Pseudo R2 =     0.0258
```

bmi	Coef.	Bootstrap Std. Err.	t	P>|t|	[95% Conf.	Interval]
q25						
age	0.05	0.00	17.95	0.000	0.04	0.05
sex						
Female	-1.06	0.09	-11.29	0.000	-1.25	-0.88
race						
Black	0.36	0.19	1.88	0.060	-0.02	0.74
Other	-1.11	0.24	-4.64	0.000	-1.57	-0.64
1.rural	0.15	0.10	1.50	0.133	-0.05	0.35
_cons	20.56	0.16	126.74	0.000	20.24	20.88
q75						
age	0.06	0.00	19.13	0.000	0.05	0.07
sex						
Female	0.18	0.14	1.31	0.189	-0.09	0.44
race						
Black	2.16	0.28	7.85	0.000	1.62	2.70
Other	-1.52	0.37	-4.15	0.000	-2.23	-0.80
1.rural	0.75	0.13	5.72	0.000	0.49	1.01
_cons	24.47	0.14	170.88	0.000	24.19	24.75

Categorical ➡ Nominal

- Binary logistic regression
  - Dependent variable:
    diabetes
  - Independent variables:
    age (continuous)
    sex (categorical)
    bmi (continuous)
    rural (categorical)

Odds ratio

```
. webuse nhanes2

. logit diabetes  age i.sex bmi i.rural, or nolog cformat(%9.2f) baselevels

Logistic regression                             Number of obs     =     10,349
                                                LR chi2(4)        =     419.59
                                                Prob > chi2       =     0.0000
Log likelihood = -1789.9636                     Pseudo R2         =     0.1049

------------------------------------------------------------------------------
    diabetes | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |       1.06       0.00    15.10   0.000        1.05        1.07
             |
         sex |
        Male |       1.00  (base)
      Female |       1.07       0.10     0.70   0.484        0.89        1.29
             |
         bmi |       1.08       0.01     8.96   0.000        1.06        1.09
             |
       rural |
           0 |       1.00  (base)
           1 |       0.95       0.09    -0.52   0.606        0.79        1.15
             |
       _cons |       0.00       0.00   -24.89   0.000        0.00        0.00
------------------------------------------------------------------------------
```

Categorical ➡ Nominal

- Mutinomial logistic regression
  - Dependent variable:
    presidential candidate (candidat)
  - Independent variables:
    inc (categorical)

Relative risk ratio

```
. sysuse voter.dta

. mlogit candidat i.inc, rrr nolog cformat(%9.2f) baselevels

Multinomial logistic regression                 Number of obs     =         15
                                                LR chi2(8)        =       0.00
                                                Prob > chi2       =     1.0000
Log likelihood = -16.479184                     Pseudo R2         =     0.0000
```

candidat	RRR	Std. Err.	z	P>|z|	[95% Conf.	Interval]
Clinton	(base outcome)					
Bush						
inc						
<$15k	1.00	(base)				
$15-30k	1.00	2.00	0.00	1.000	0.02	50.40
$30-50k	1.00	2.00	0.00	1.000	0.02	50.40
$50-75k	1.00	2.00	0.00	1.000	0.02	50.40
$75k+	1.00	2.00	0.00	1.000	0.02	50.40
_cons	1.00	1.41	0.00	1.000	0.06	15.99
Perot						
inc						
<$15k	1.00	(base)				
$15-30k	1.00	2.00	0.00	1.000	0.02	50.40
$30-50k	1.00	2.00	0.00	1.000	0.02	50.40
$50-75k	1.00	2.00	0.00	1.000	0.02	50.40
$75k+	1.00	2.00	0.00	1.000	0.02	50.40
_cons	1.00	1.41	0.00	1.000	0.06	15.99

Categorical ➡ Ordinal

Ordered logistic regression

- Dependent variable:
  health status (hlthstat)
- Independent variables:
  age (continuous)
  sex (categorical)
  race (categorical)
  rural (categorical)

```
. webuse nhanes2

. ologit hlthstat age i.sex i.race i.rural, or nolog cformat(%9.2f) baselevels

Ordered logistic regression                     Number of obs     =     10,349
                                                LR chi2(5)        =   1787.79
                                                Prob > chi2       =    0.0000
Log likelihood = -14976.969                     Pseudo R2         =    0.0563
```

hlthstat	Odds Ratio	Std. Err.	z	P>|z|	[95% Conf.	Interval]
age	1.04	0.00	37.51	0.000	1.04	1.04
sex						
Male	1.00	(base)				
Female	1.14	0.04	3.70	0.000	1.06	1.22
race						
White	1.00	(base)				
Black	2.70	0.16	16.69	0.000	2.40	3.03
Other	1.36	0.17	2.45	0.014	1.06	1.74
rural						
0	1.00	(base)				
1	1.46	0.05	9.93	0.000	1.35	1.57
/cut1	0.90	0.06			0.78	1.01
/cut2	2.16	0.06			2.04	2.28
/cut3	3.59	0.07			3.46	3.72
/cut4	5.07	0.08			4.92	5.21
/cut5	9.16	0.28			8.62	9.70