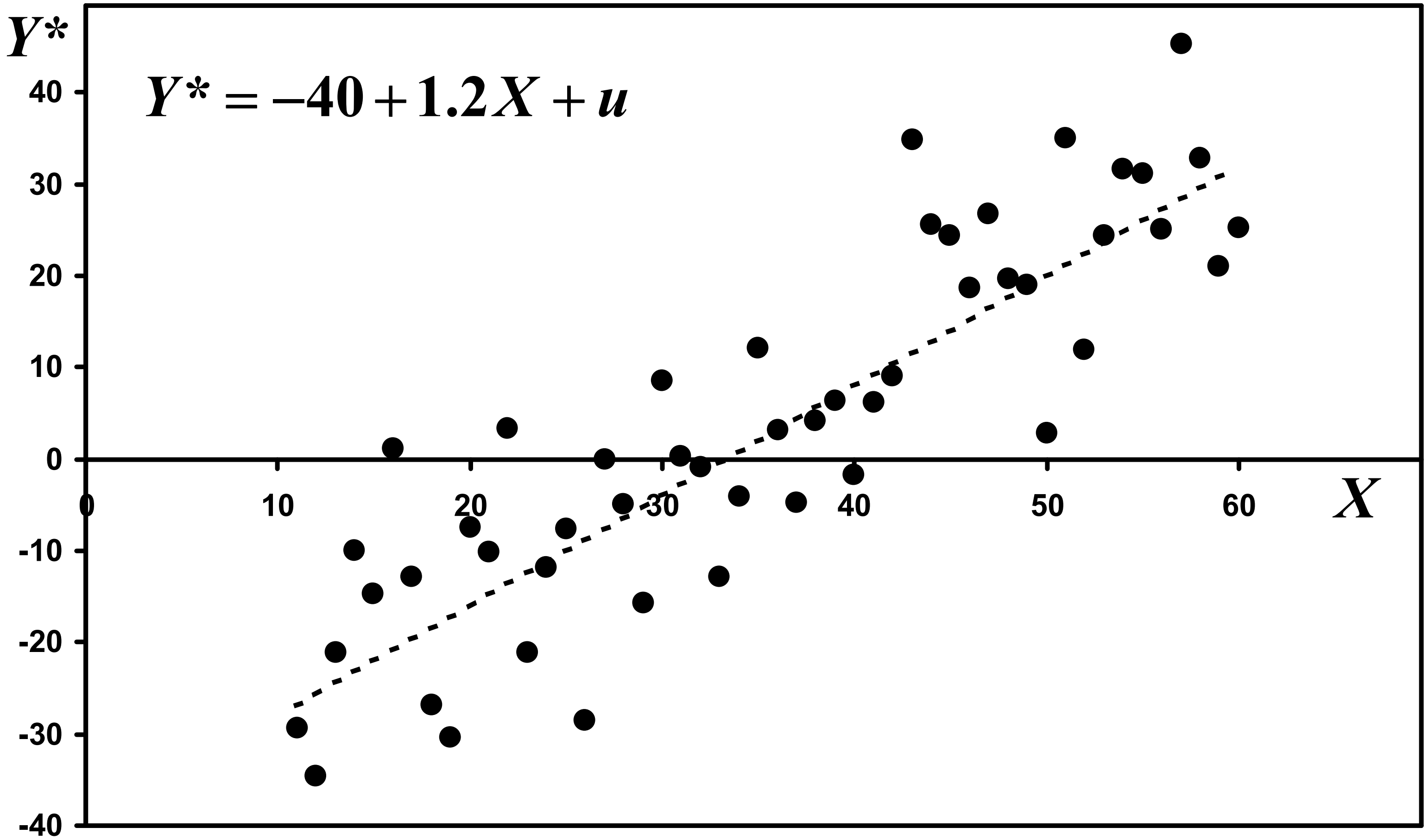
TOBIT ANALYSIS

$$Y^* = \beta_1 + \beta_2 X + u$$

Sometimes the dependent variable in a regression model is subject to a lower limit or an upper limit, or both. Suppose that in the absence of any constraints, *Y* is related to *X* by the model shown.

TOBIT ANALYSIS



For example, suppose that the true relationship is as shown, and that, given a set of random numbers for the disturbance term drawn from a normal distribution with mean 0 and standard deviation 10, we have a sample of observations as shown.

TOBIT ANALYSIS

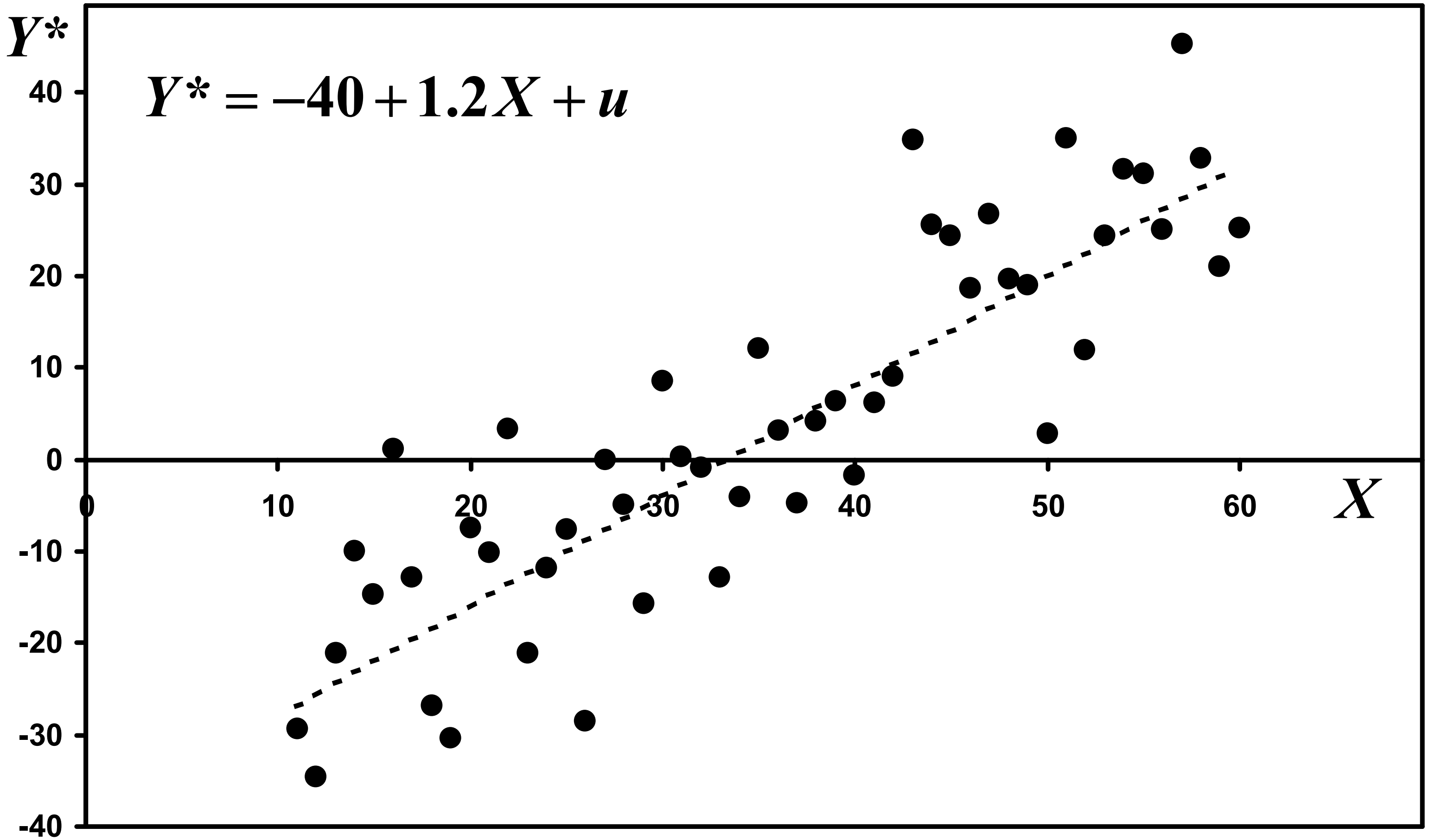
$$Y^* = \beta_1 + \beta_2 X + u$$

$$Y = Y^* \text{ if } Y^* > 0$$

$$Y = 0 \quad \text{if } Y^* \leq 0$$

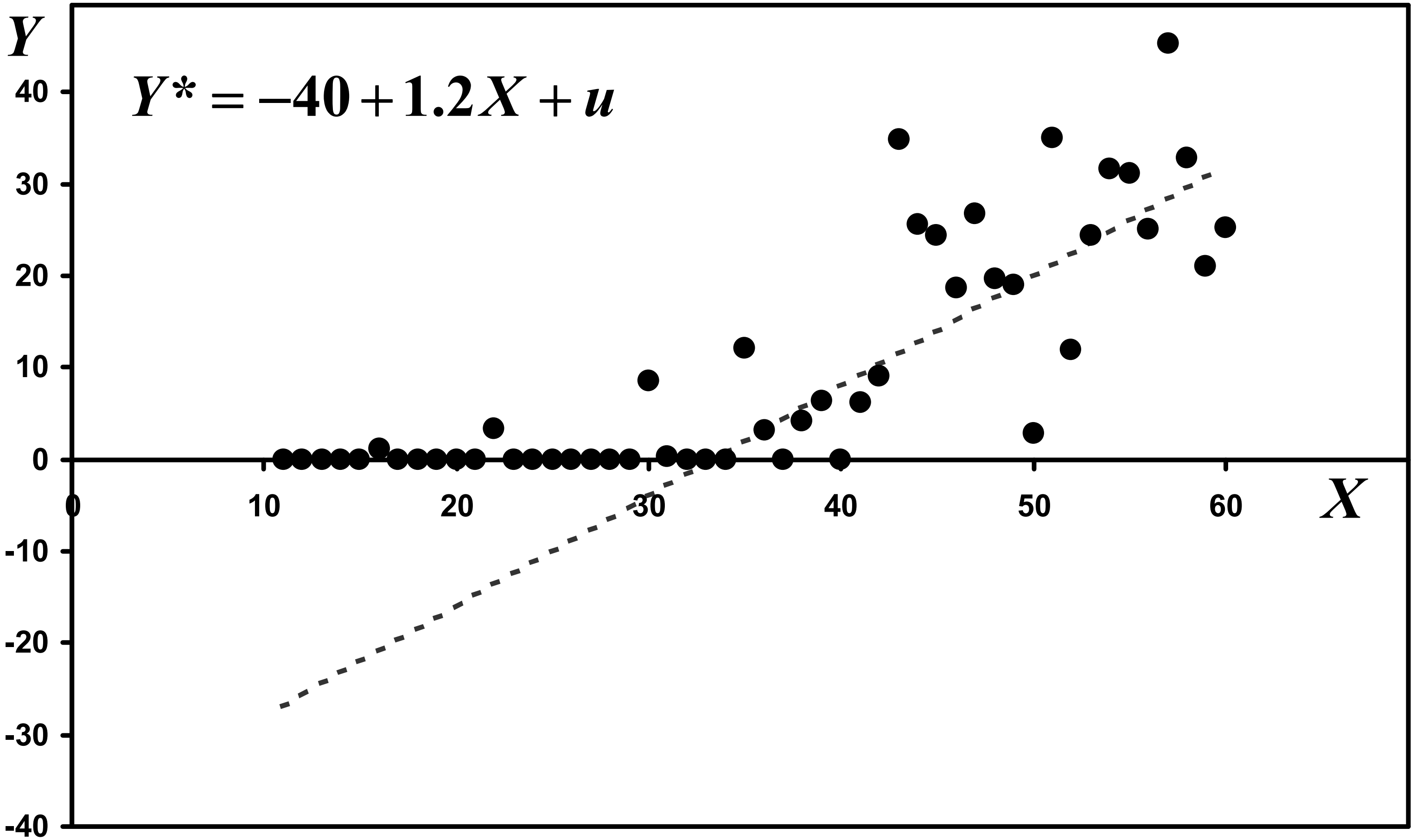
However, suppose that the dependent variable is subject to a lower bound, in this case 0. Then Y will be as given by the model if $Y^* > 0$, and it will be 0 if $Y^* = 0$ or if $Y^* < 0$.

TOBIT ANALYSIS



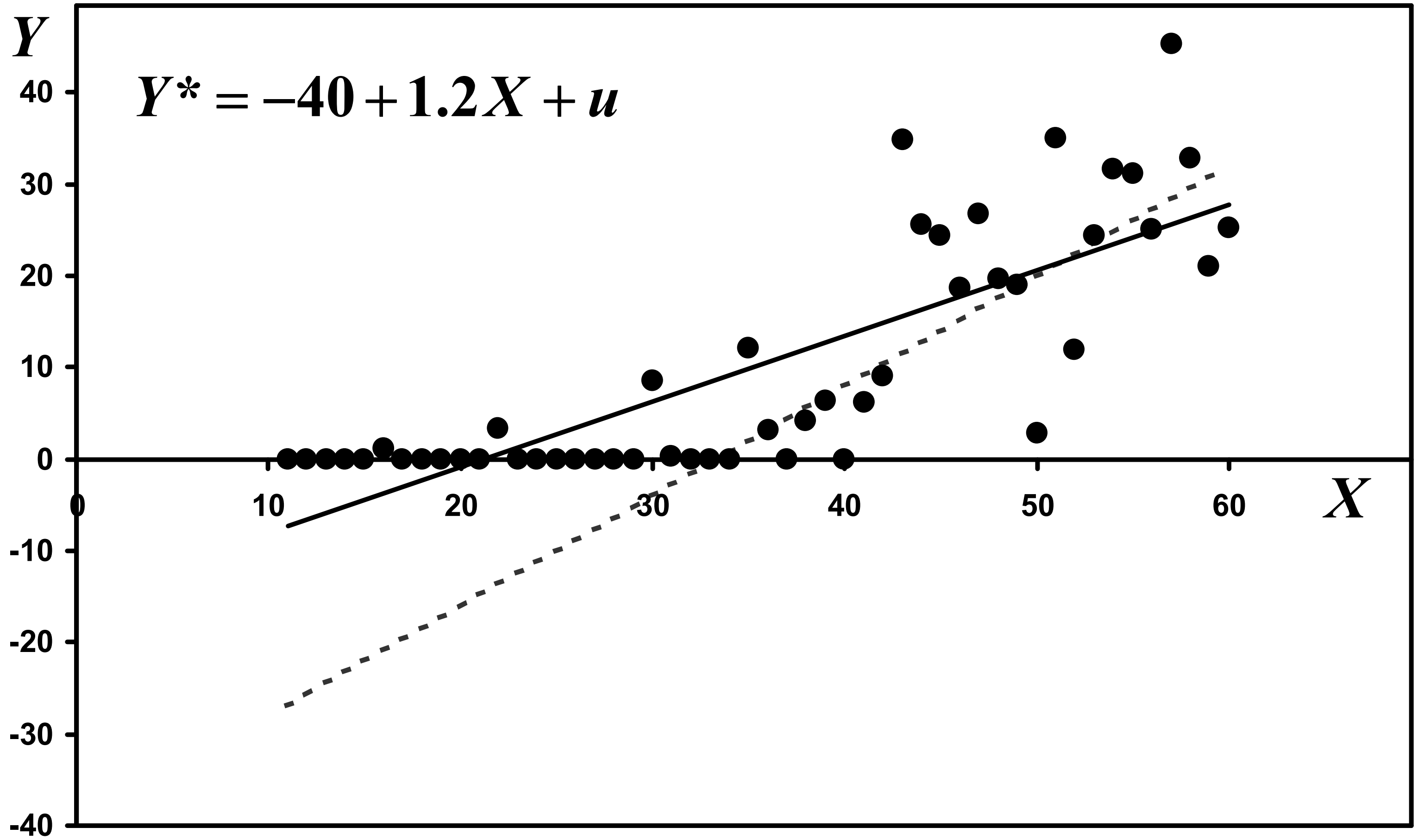
For example, suppose that we have a labor supply model with *Y* hours of labor supplied per week as a function of *X*, the wage that is offered. It is not possible to supply a negative number of hours.

TOBIT ANALYSIS



Those individuals with negative *Y** will simply not work. For them, the actual *Y* is 0.

TOBIT ANALYSIS



What would happen if we ran an OLS regression anyway? Obviously, in this case the slope coefficient would be biased downwards.

TOBIT ANALYSIS

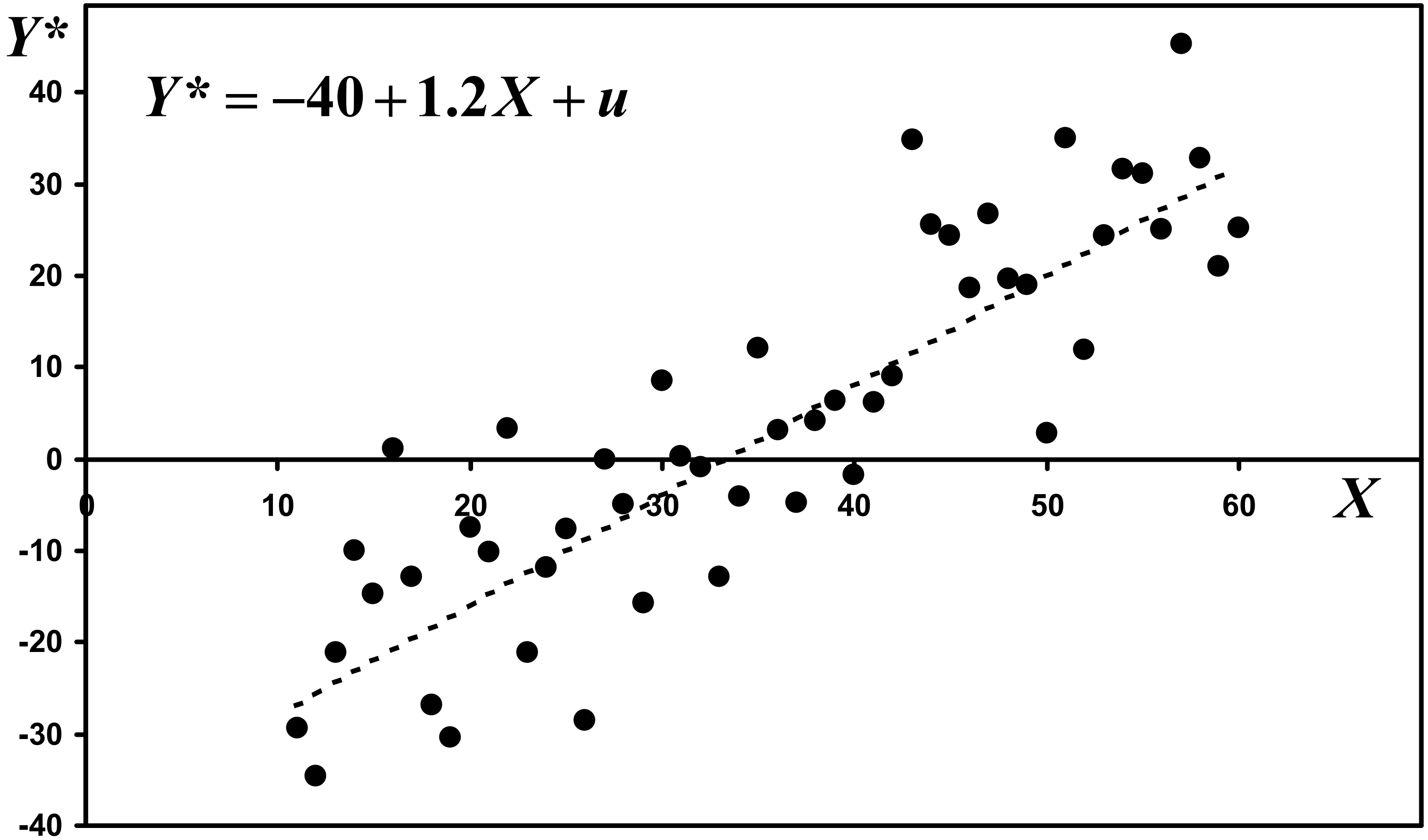
$$Y^* = \beta_1 + \beta_2 X + u$$

$$Y = Y^* \text{ if } Y^* > 0$$

$$\text{observation dropped if } Y^* \leq 0$$

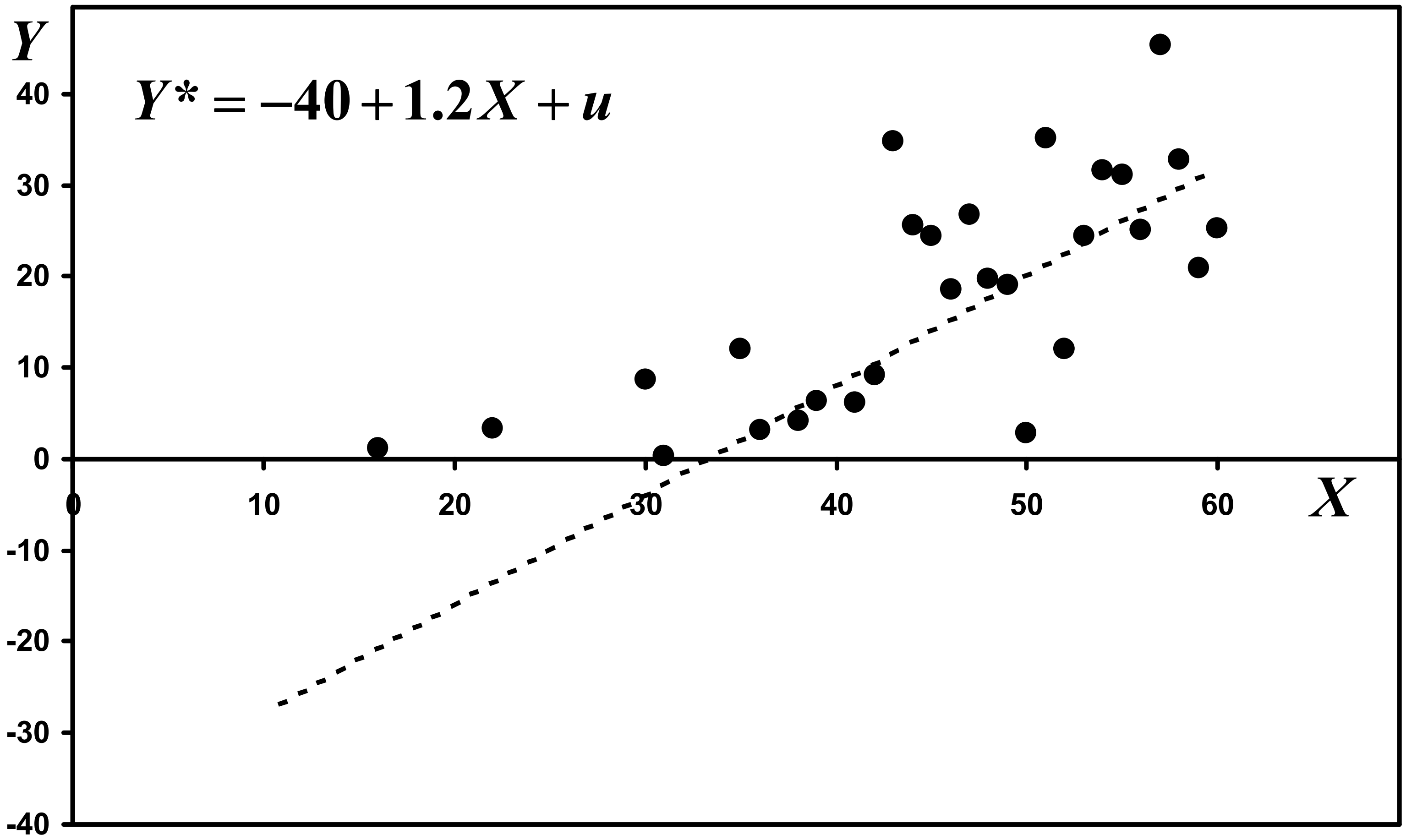
It would be natural to suppose that the problem could be avoided by dropping the constrained observations. Unfortunately, this does not work.

TOBIT ANALYSIS



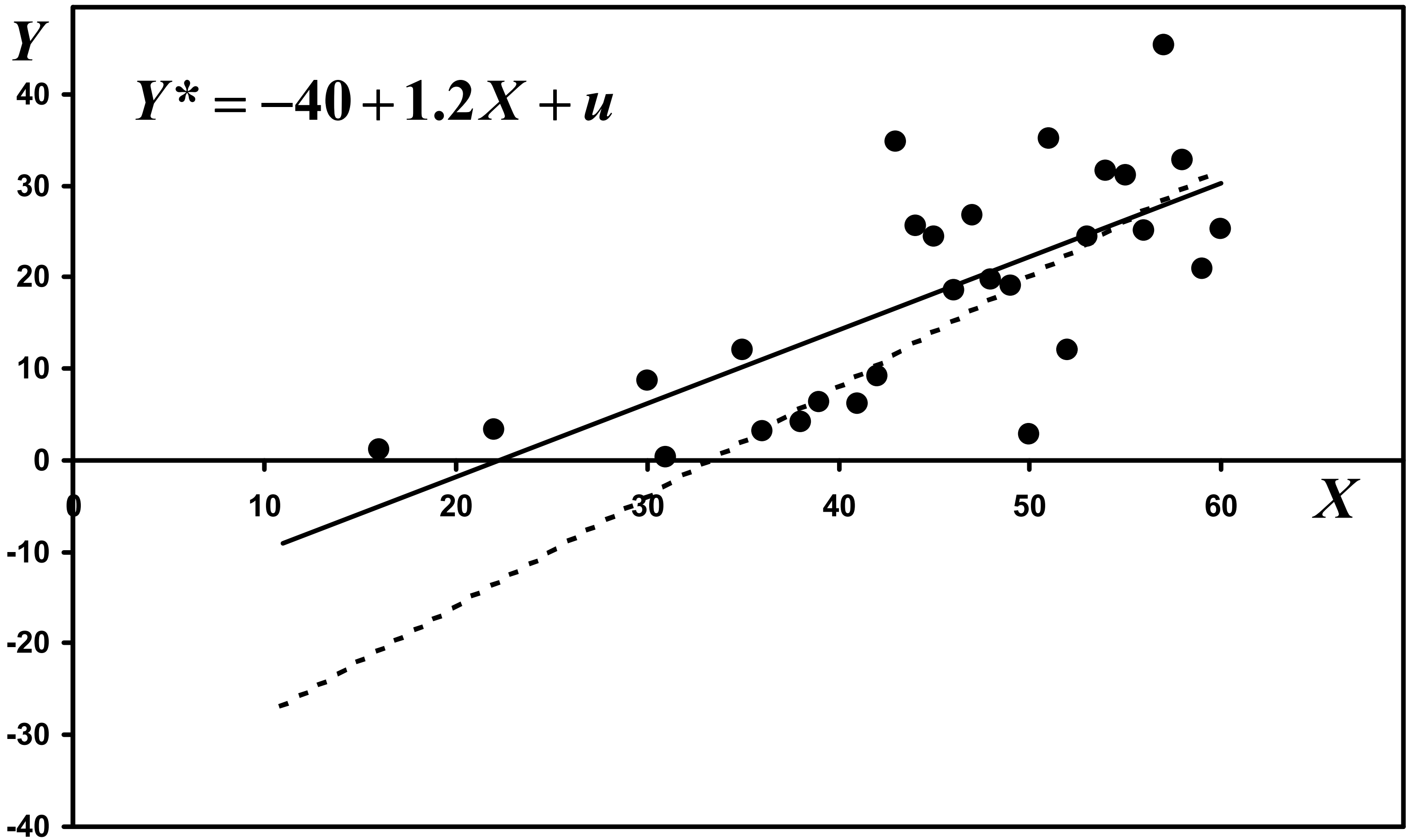
Here again is the sample as it would be if *Y* were not constrained.

TOBIT ANALYSIS



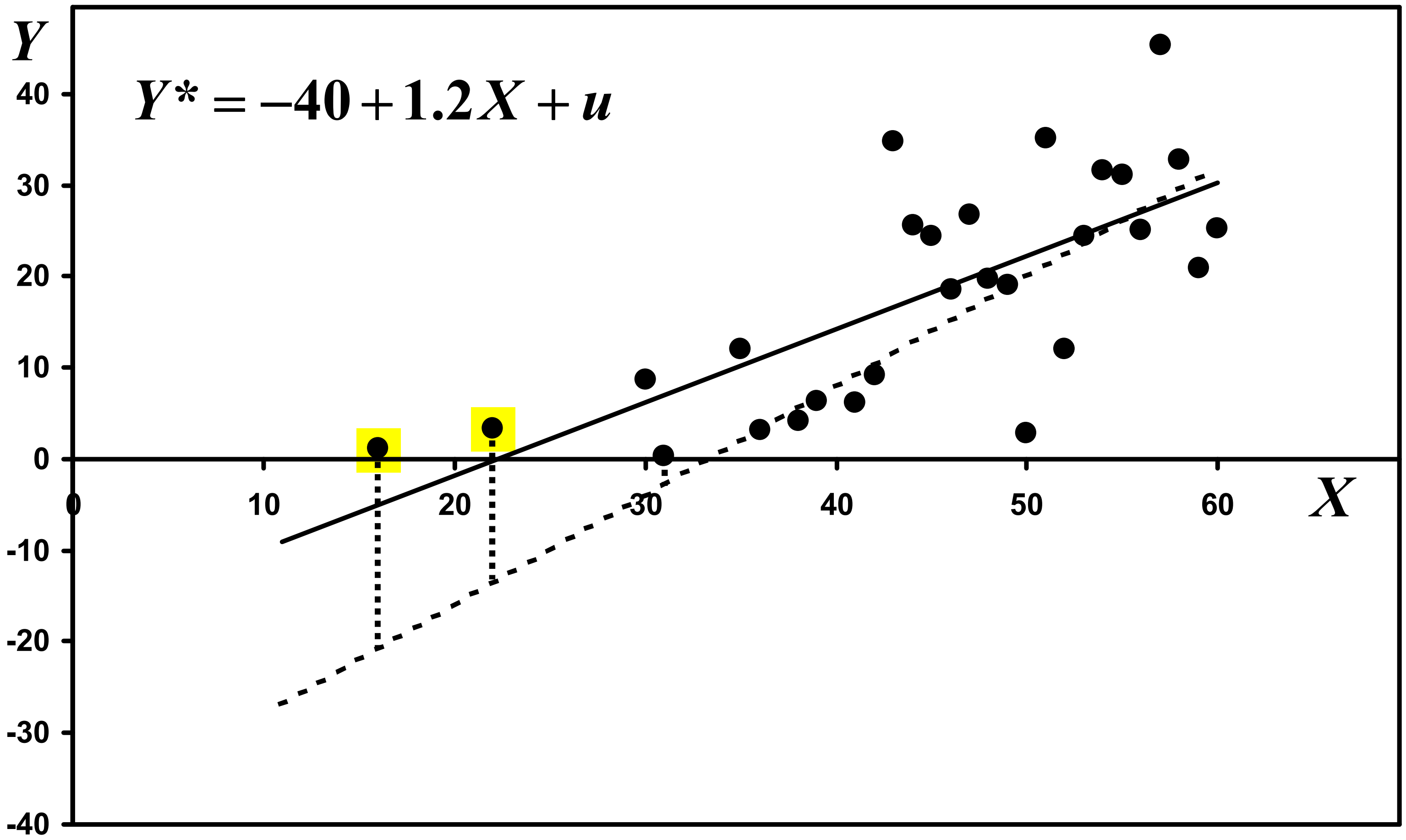
Here is the sample with the constrained observations dropped.

TOBIT ANALYSIS



An OLS regression again yields a downwards-biased estimate of the slope coefficient and an upwards-biased estimate of the intercept. We will investigate the reason for this.

TOBIT ANALYSIS



Look at the two observations highlighted. For such low values of *X*, most of the observations are constrained. The reason that these two observations appear in the sample is that their disturbance terms happen to be positive and large.

TOBIT ANALYSIS

$$Y^* = -40 + 1.2X + u$$

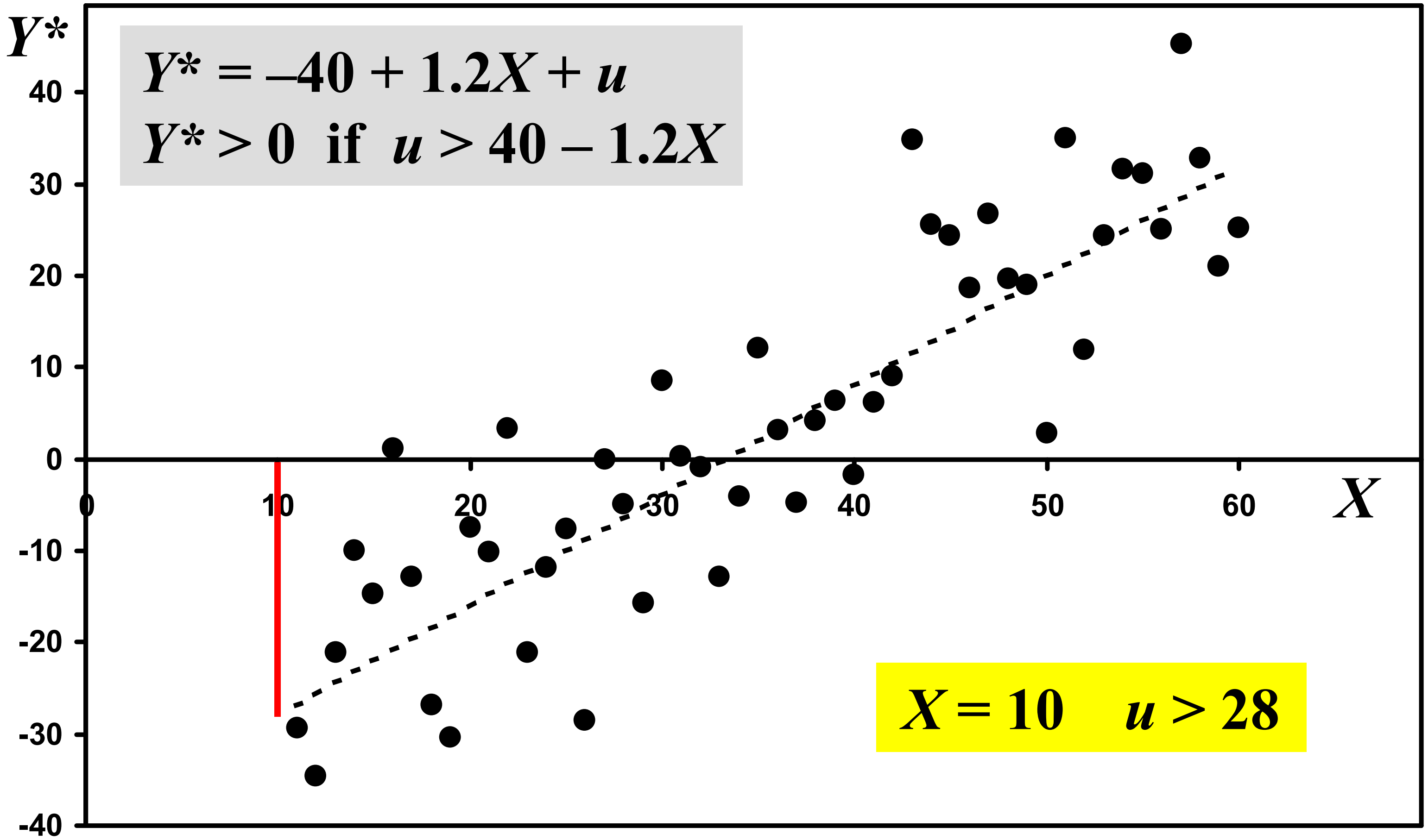
$$Y = Y^* \quad \text{if } Y^* > 0$$

$$\text{observation dropped if } Y^* \leq 0$$

$$Y^* > 0 \quad \text{if } u > 40 - 1.2X$$

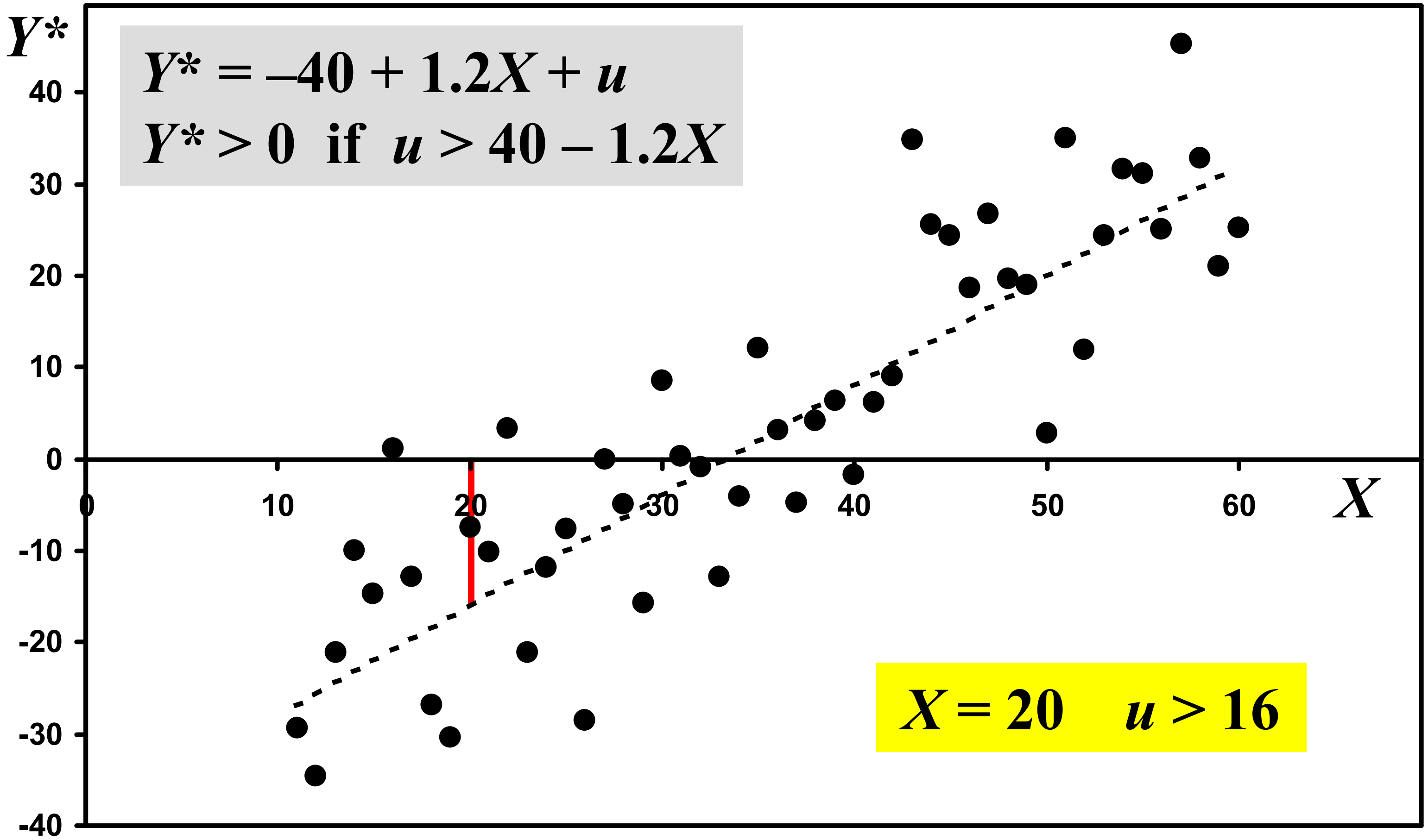
In general, for an observation to appear in the sample, *Y must be positive, and this requires that *u* > 40 – 1.2*X*.**

TOBIT ANALYSIS

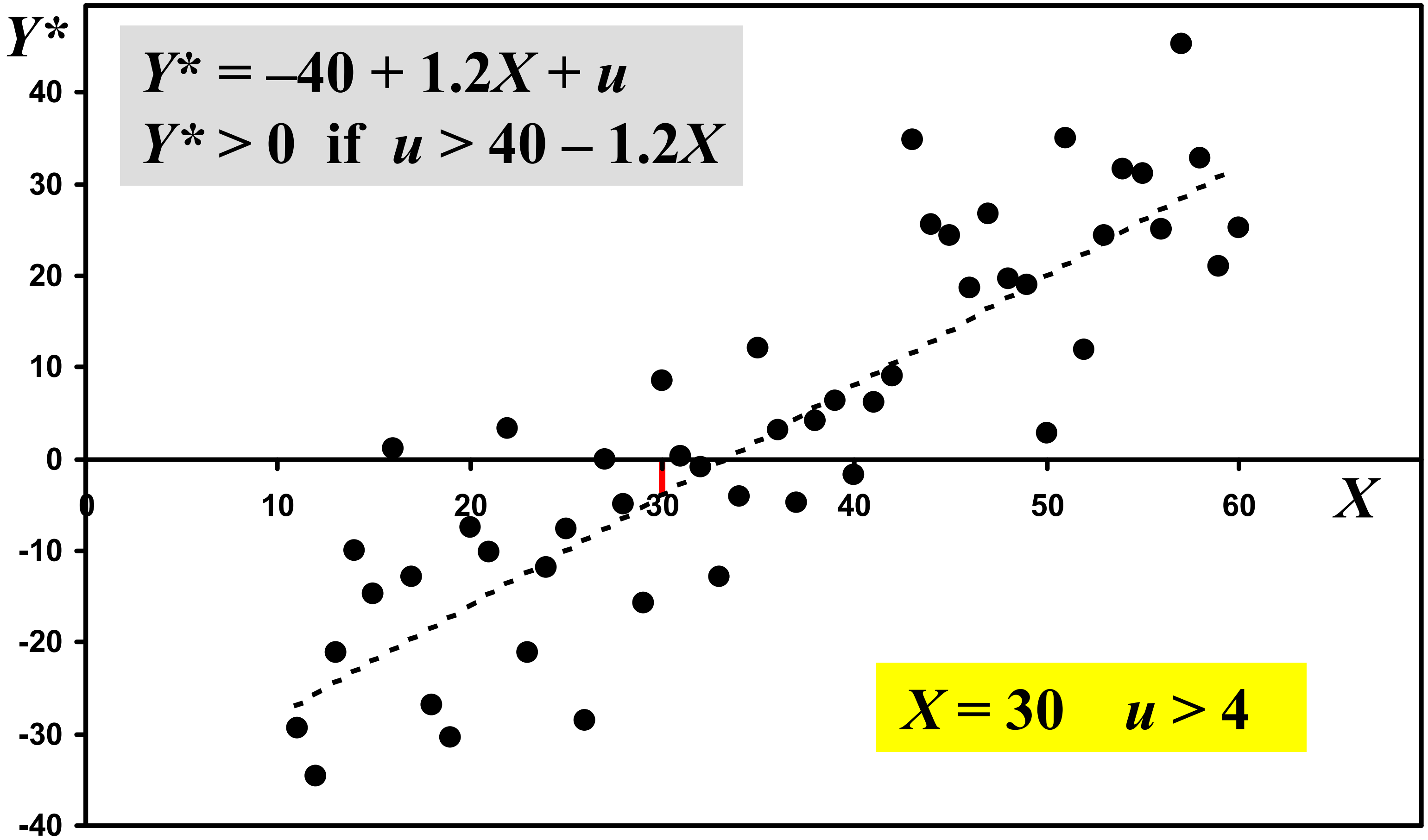


If *X* is equal to 10, *u* must be greater than 28 if the observation is to appear in the sample.

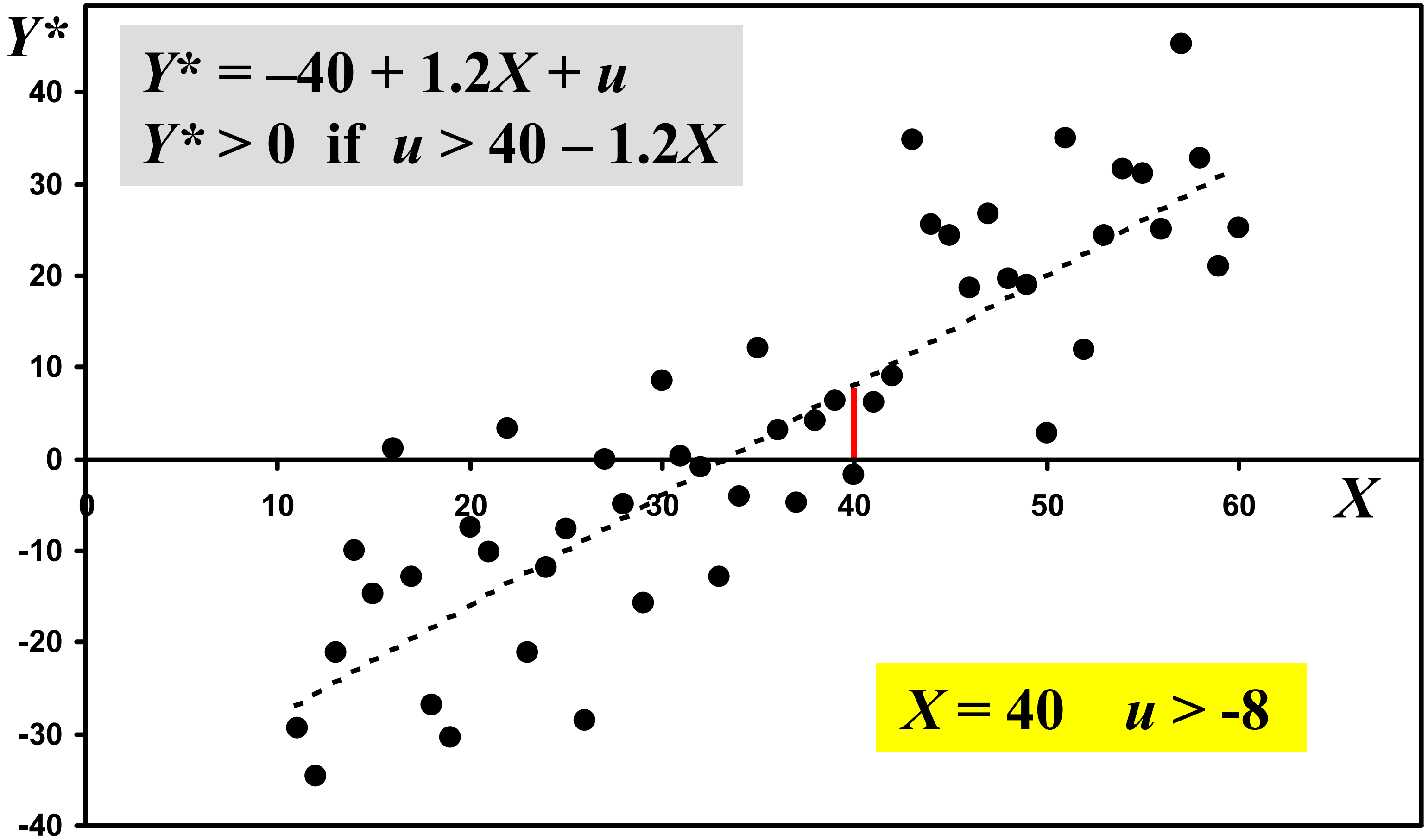
TOBIT ANALYSIS



If *X* is equal to 20, *u* must be greater than 16.

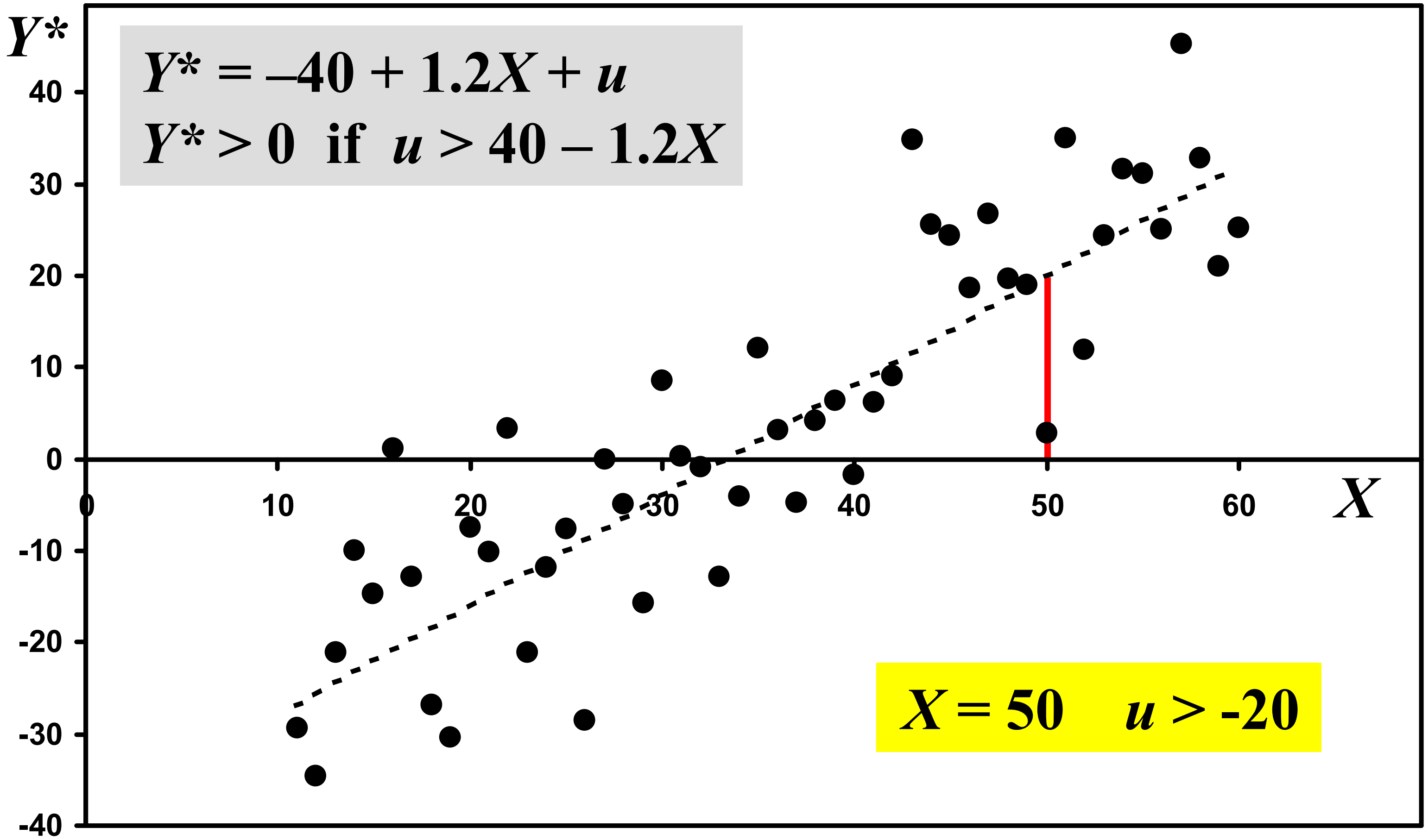


If X is equal to 30, u must be greater than 4.



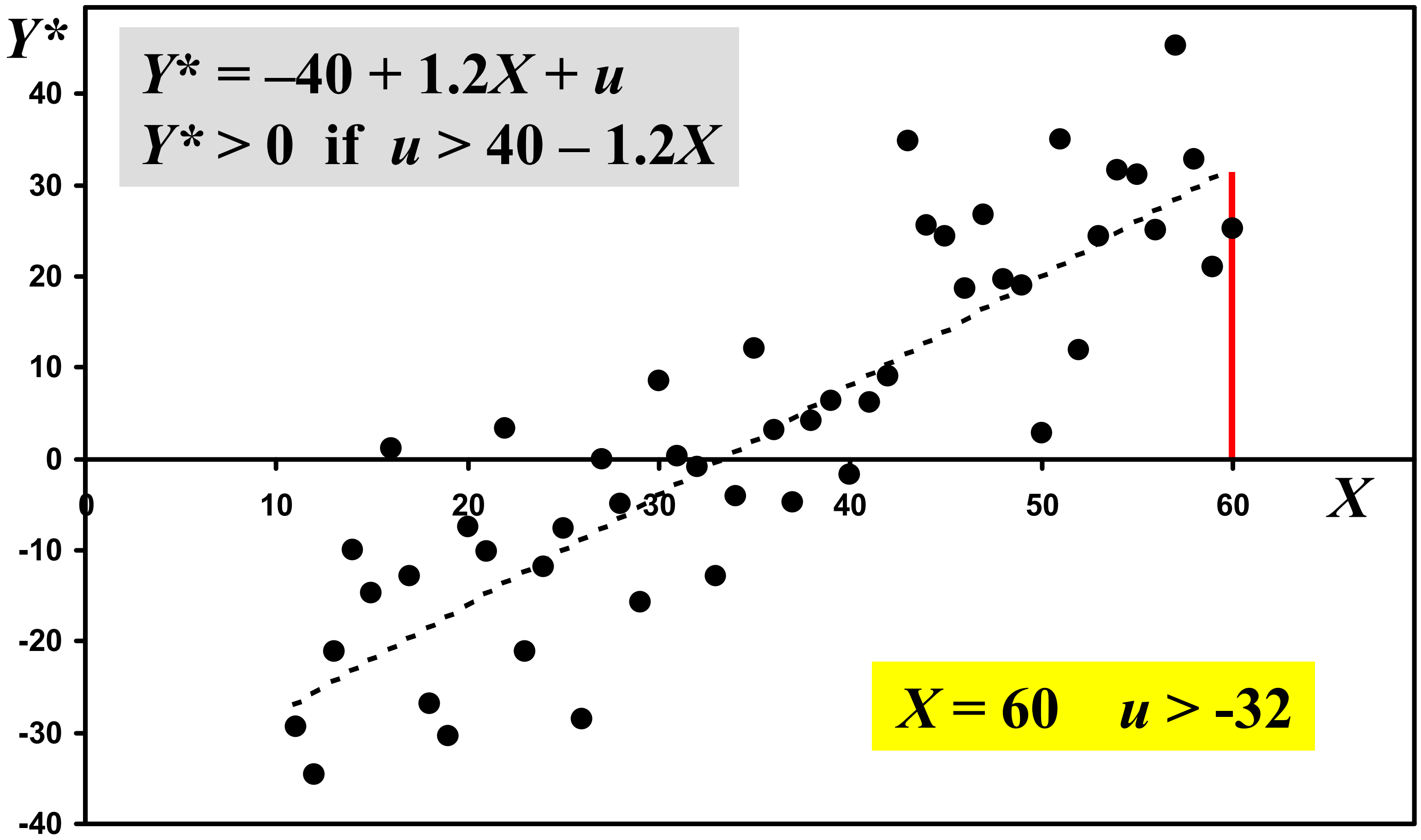
If X is equal to 40, the observation will appear in the sample for any positive value of u and even some negative ones. The condition is that u must be greater than –8.

TOBIT ANALYSIS

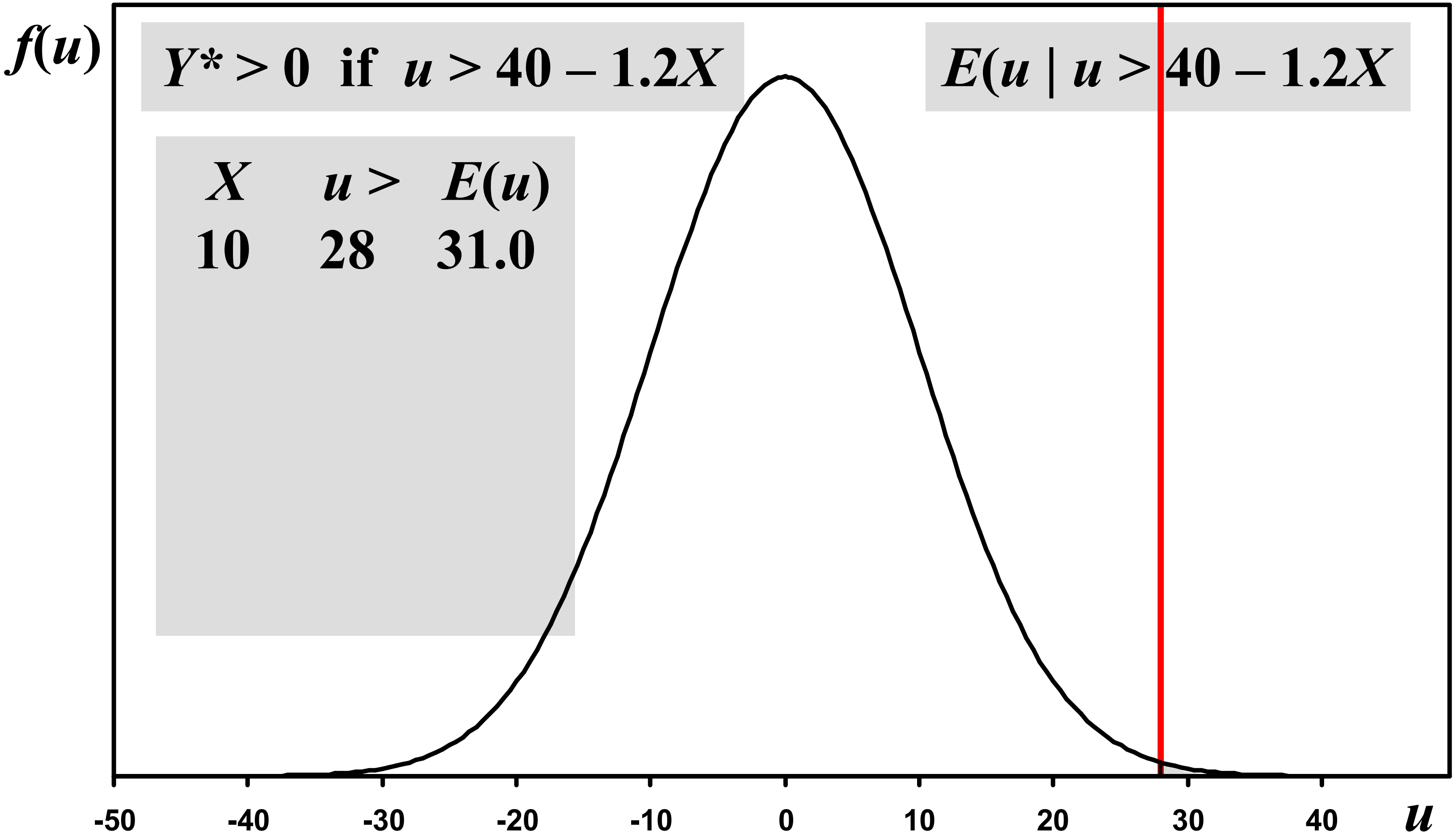


If X is equal to 50, u must be greater than –20.

TOBIT ANALYSIS

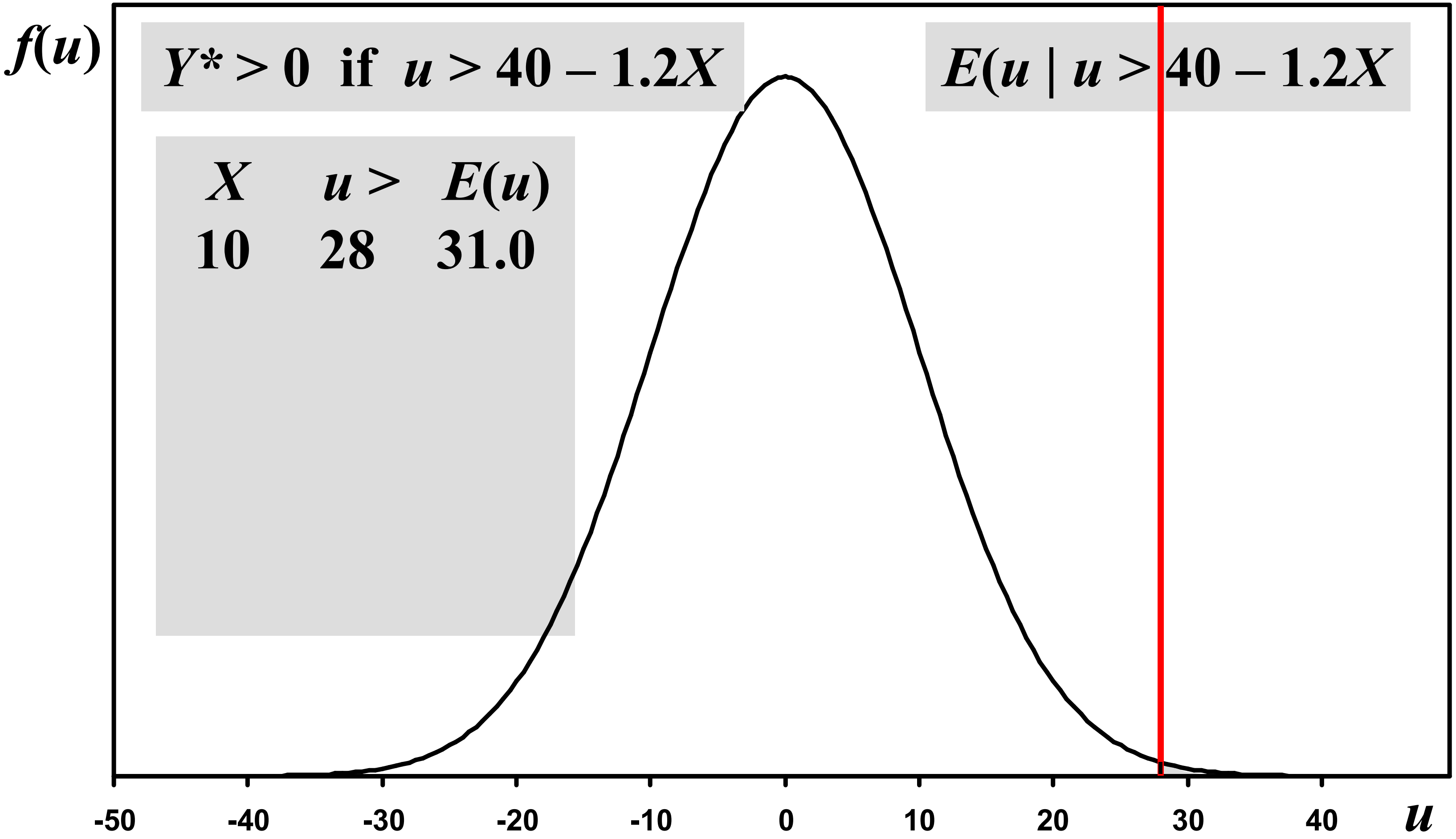


If *X* is equal to 60, *u* must be greater than –32. A value of less than –32 is very unlikely, so in this part of the sample virtually every observation will appear.



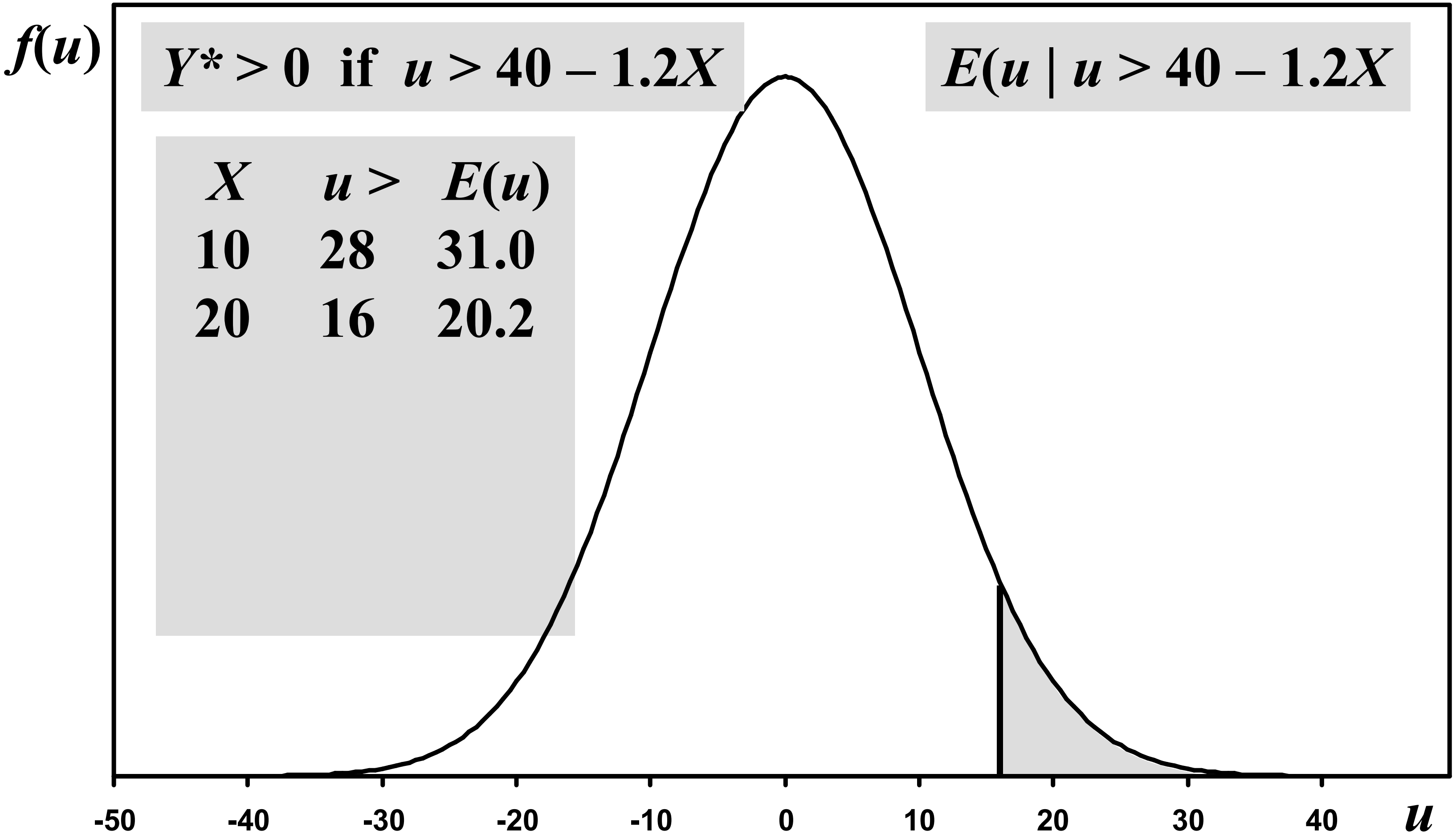
We will now show that, for observations that appear in the sample, there is a negative correlation between *X* and *u*.

TOBIT ANALYSIS

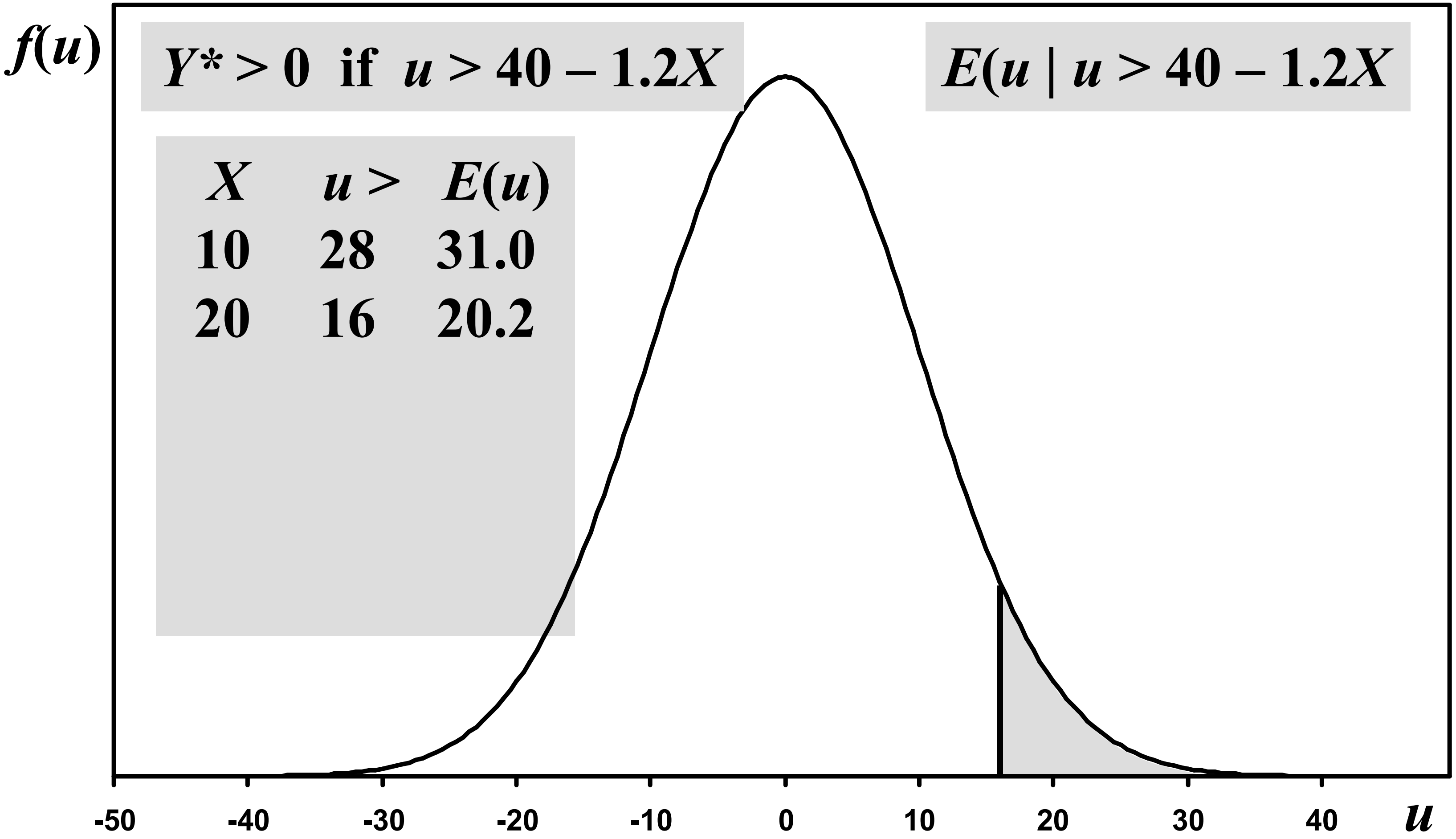


When *X* is equal to 10, *u* must be greater than 28. The expected value of *u* for observations appearing in the sample is its expected value in the tail to the right of the red line. It turns out to be 31.0.

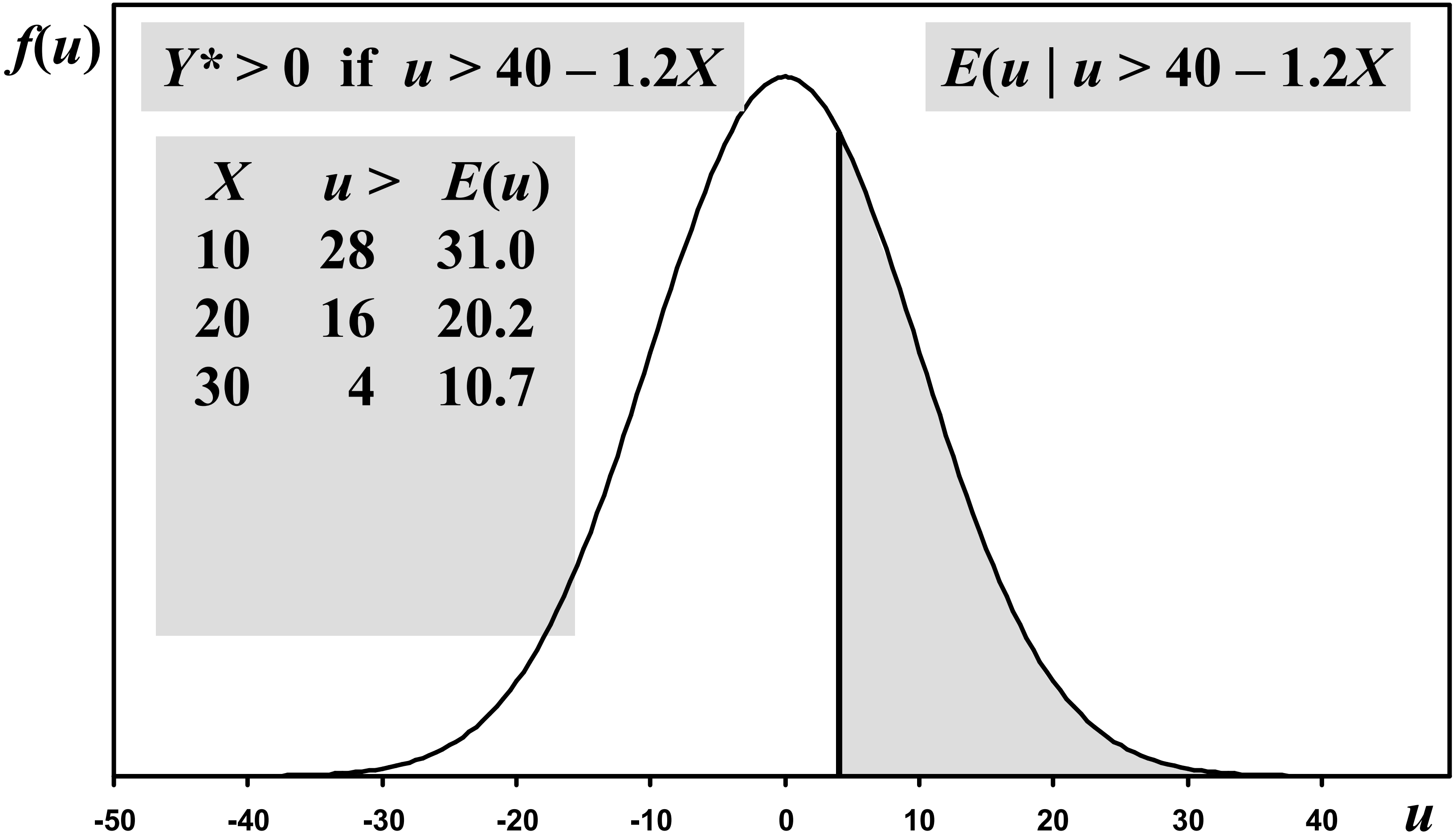
TOBIT ANALYSIS



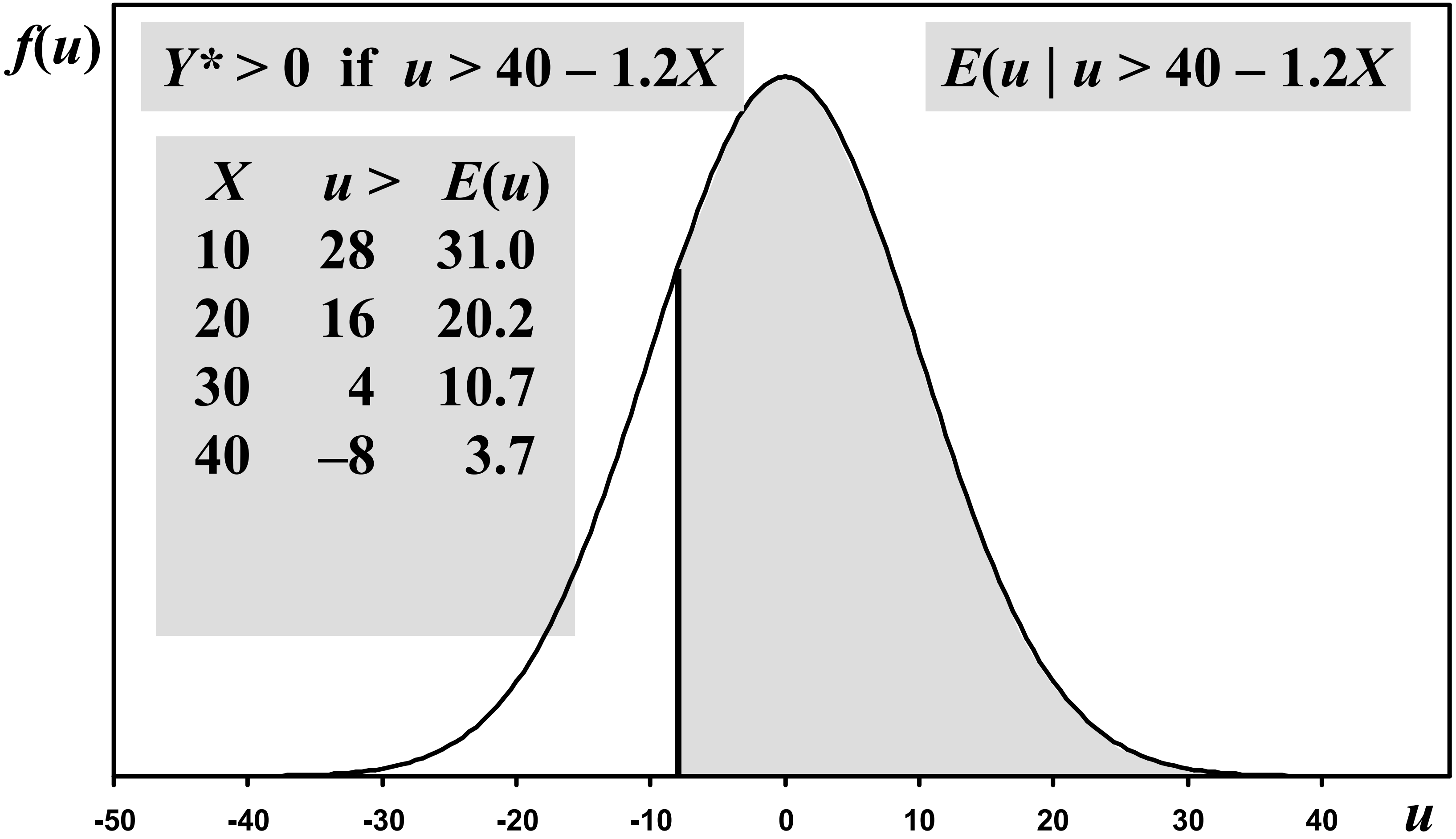
When *X* is equal to 20, *u* must be greater than 16. The expected value of *u*, for observations that appear in the sample, is the expected value in the shaded area. This is 20.2.



The rest of the distribution is irrelevant because an observation cannot appear in the sample if u < 16.

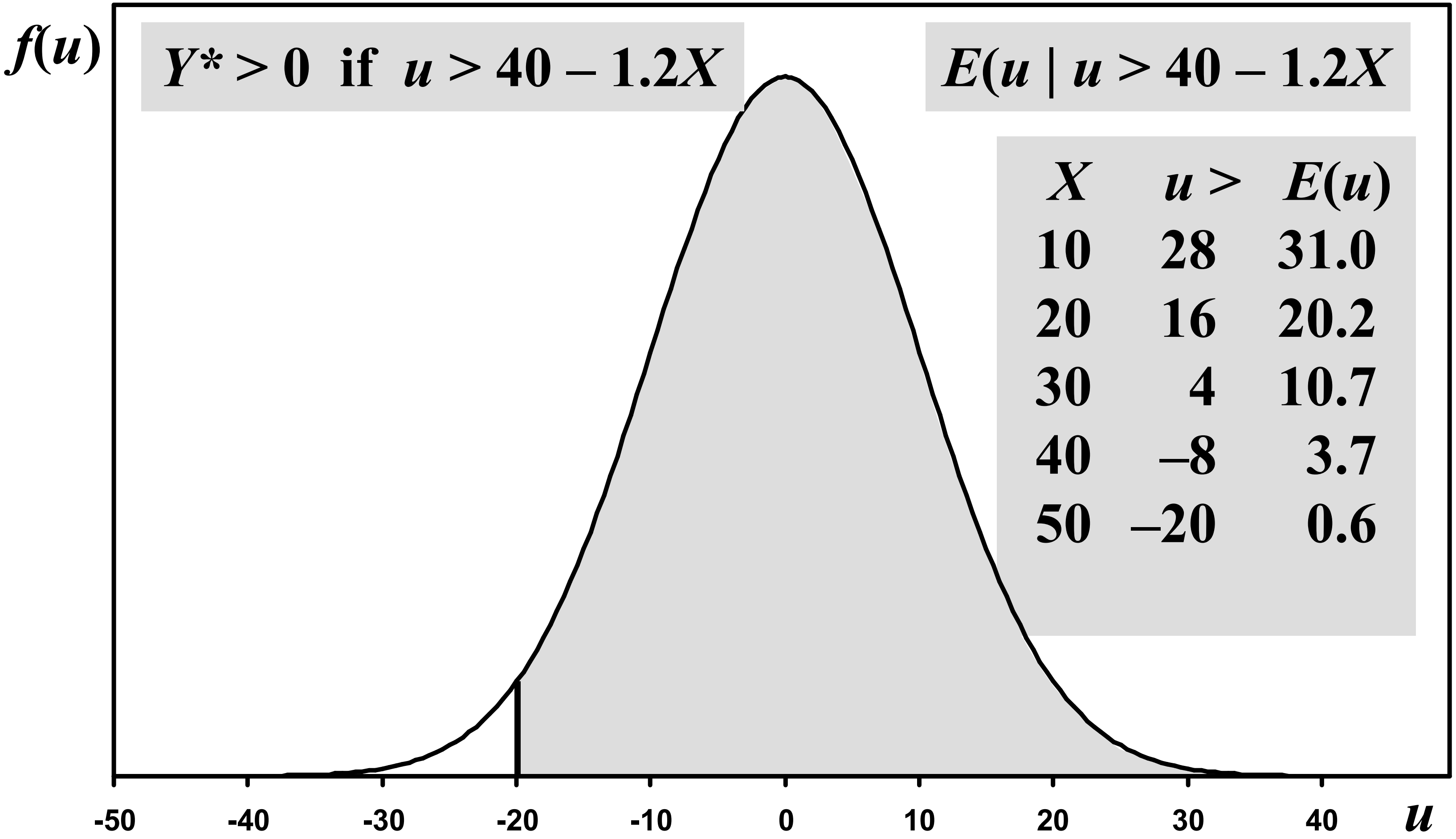


When *X* is equal to 30, *u* must be greater than 4. Its expected value, conditional on it being greater than 4, is 10.7.

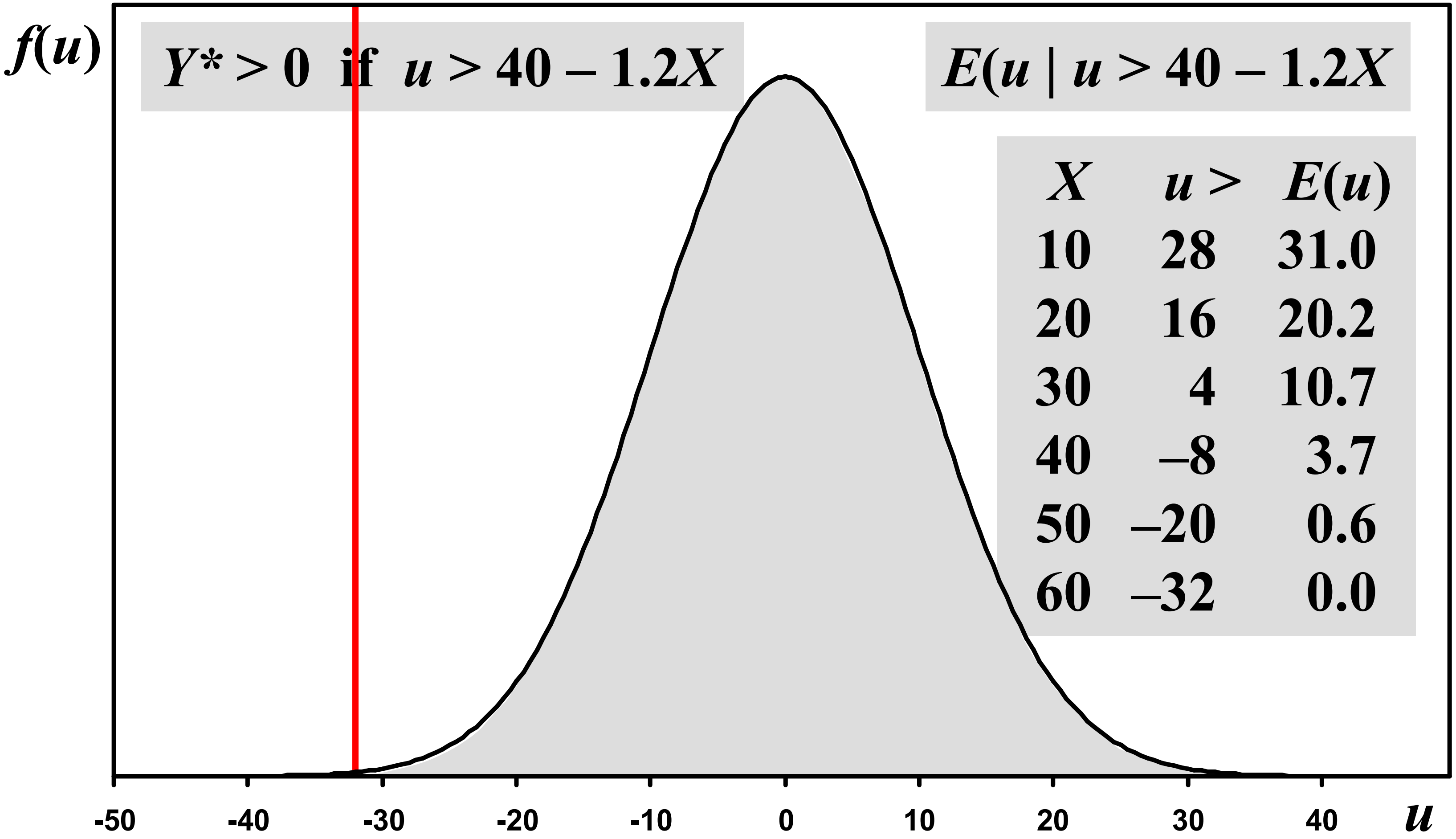


When *X* is equal to 40, *u* must be greater than –8. Its expected value, subject to this condition, is 3.7.

TOBIT ANALYSIS



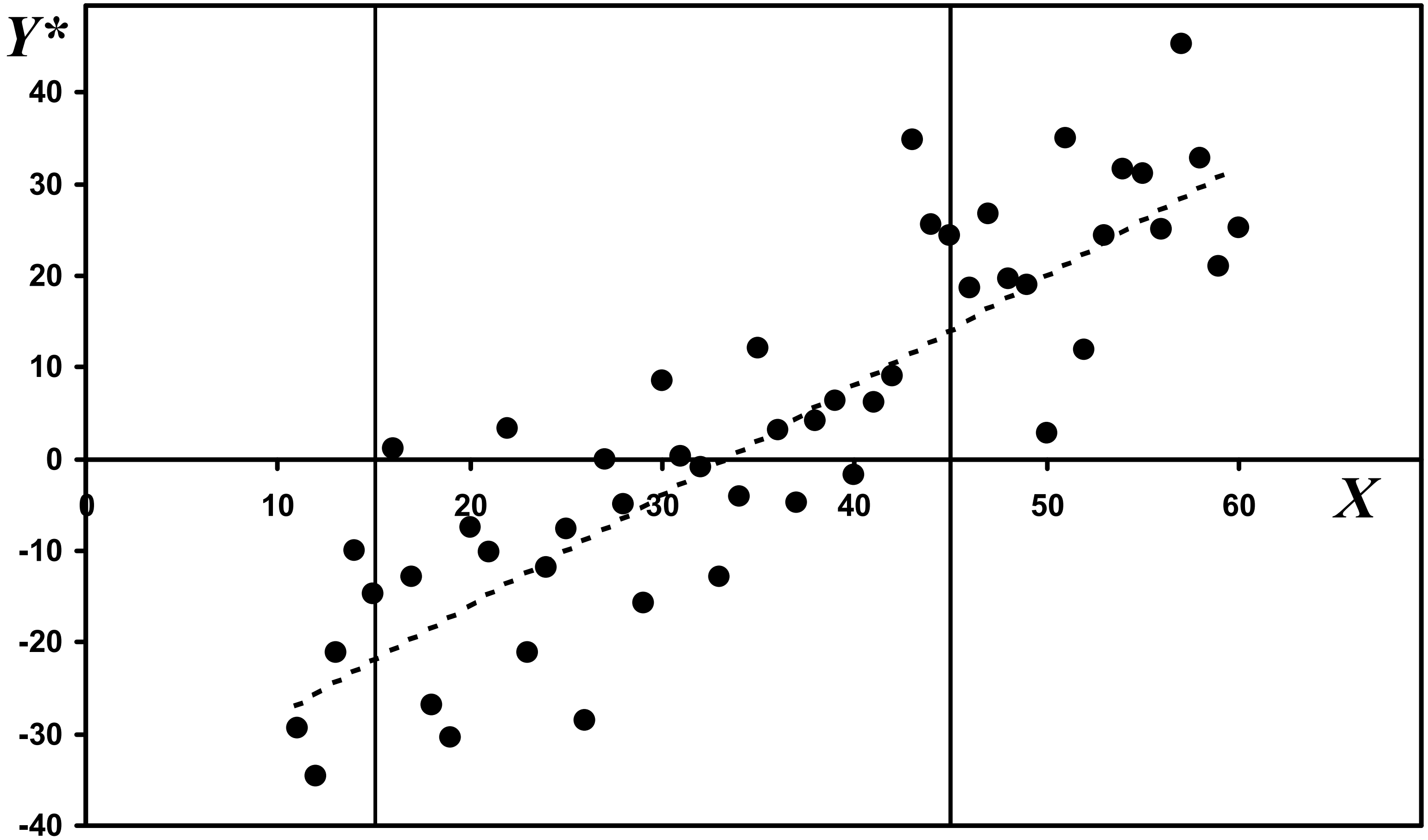
When *X* is equal to 50, *u* must be greater than –20. It will satisfy this condition nearly all the time and its conditional expected value, 0.6, is hardly any greater than its unconditional expected value, 0.



X	u >	E(u)
10	28	31.0
20	16	20.2
30	4	10.7
40	–8	3.7
50	–20	0.6
60	–32	0.0

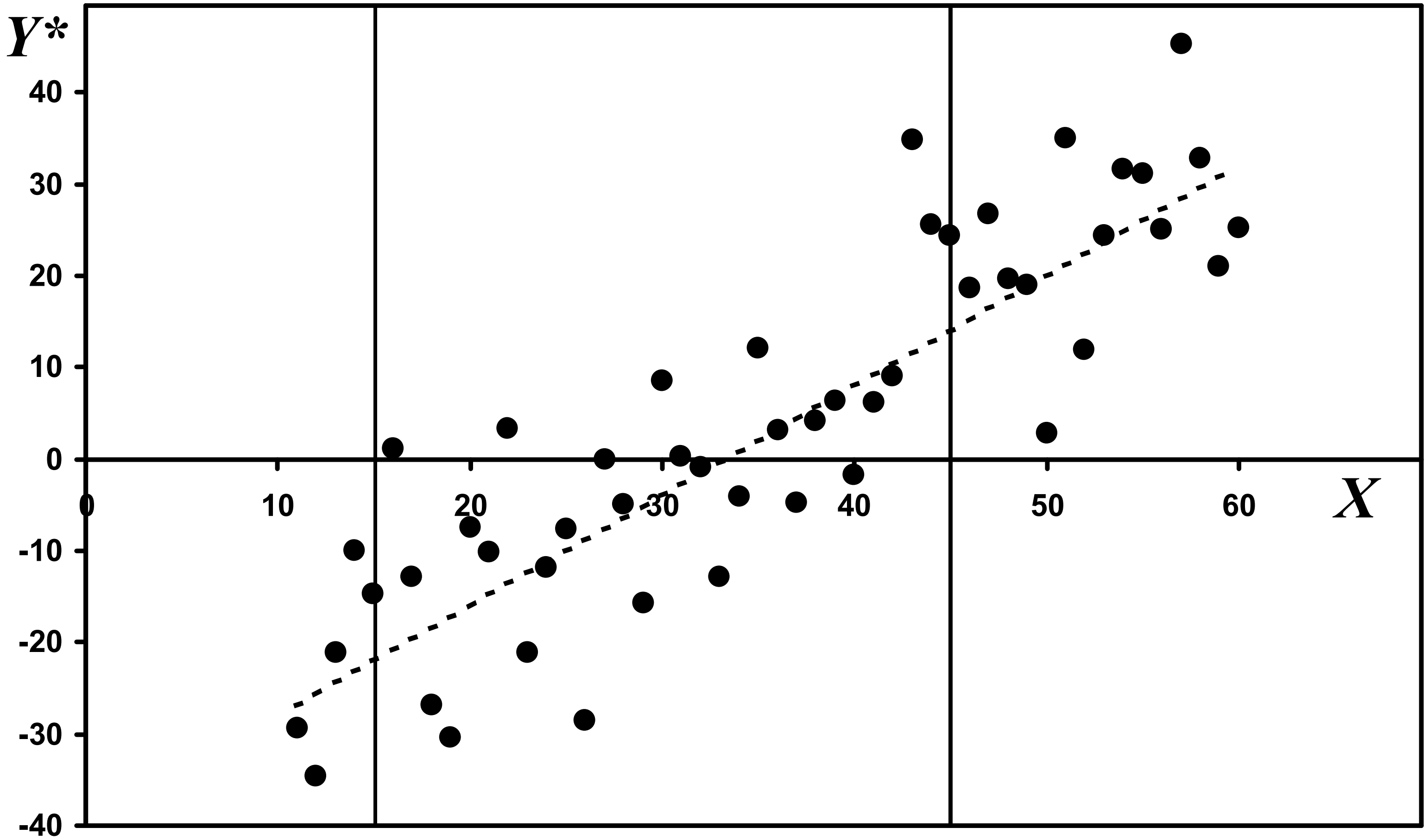
When *X* is 60 or higher, the condition will always be satisfied and the observation will always appear in the sample. For *X* > 60, *E*(*u*) is equal to its unconditional value of 0 and so there is no negative correlation between *X* and *u*.

TOBIT ANALYSIS



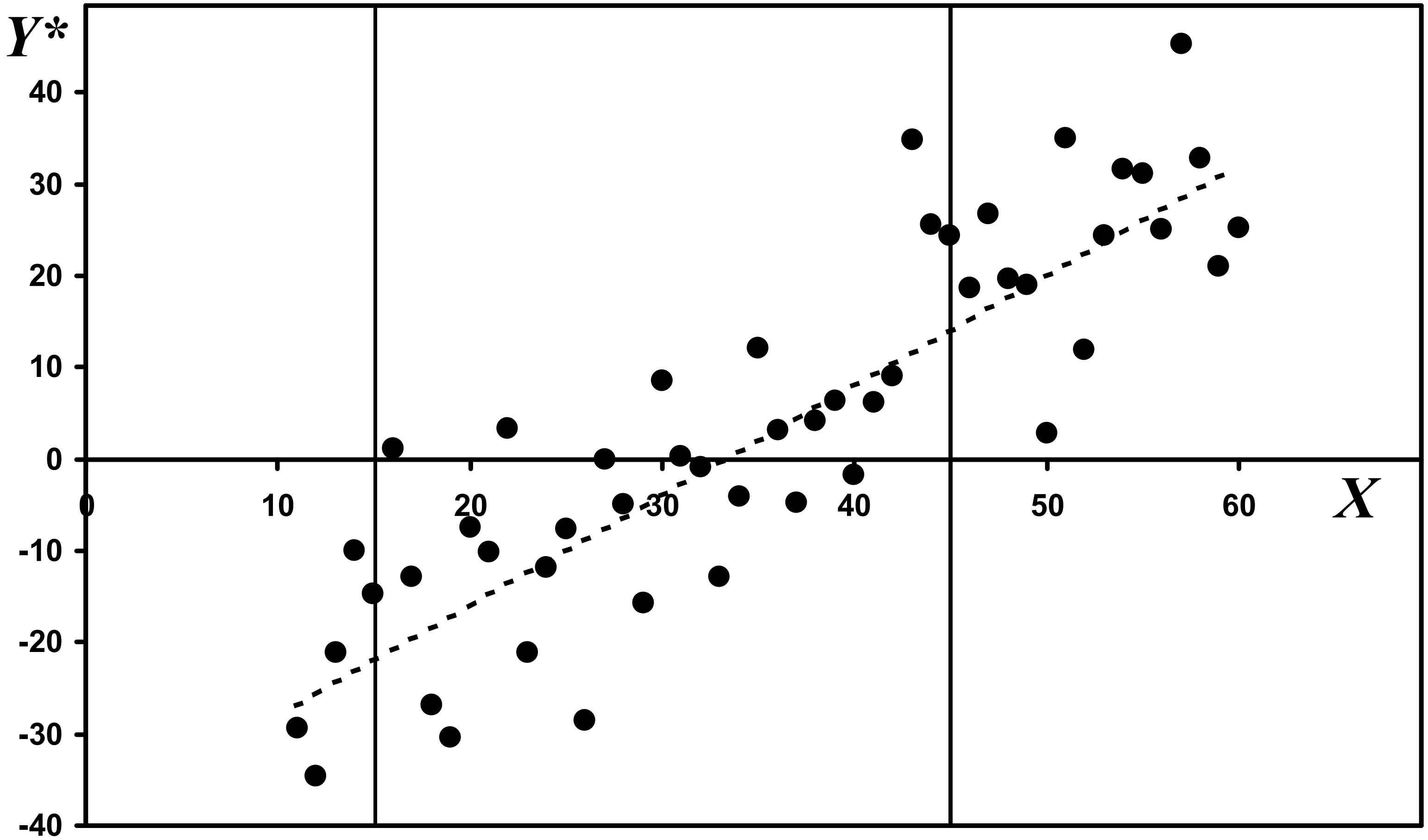
The range over which one observes a negative correlation between *X* and *u* is approximately 15 to 45. Below 15, an observation is almost certainly going to be constrained and so deleted from the sample.

TOBIT ANALYSIS



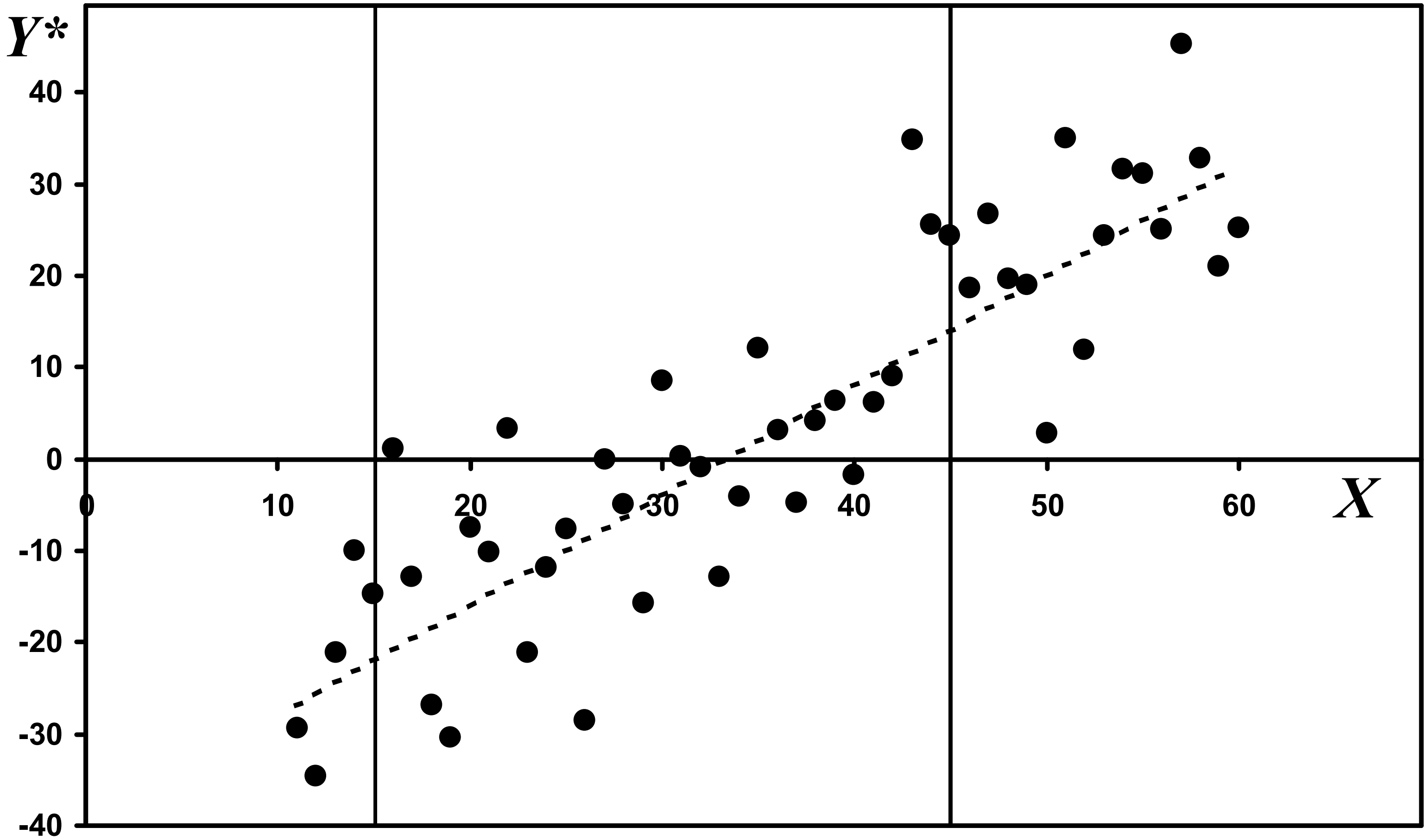
Above 45, almost all are going to appear in the sample, irrespective of the value of *u*.

TOBIT ANALYSIS



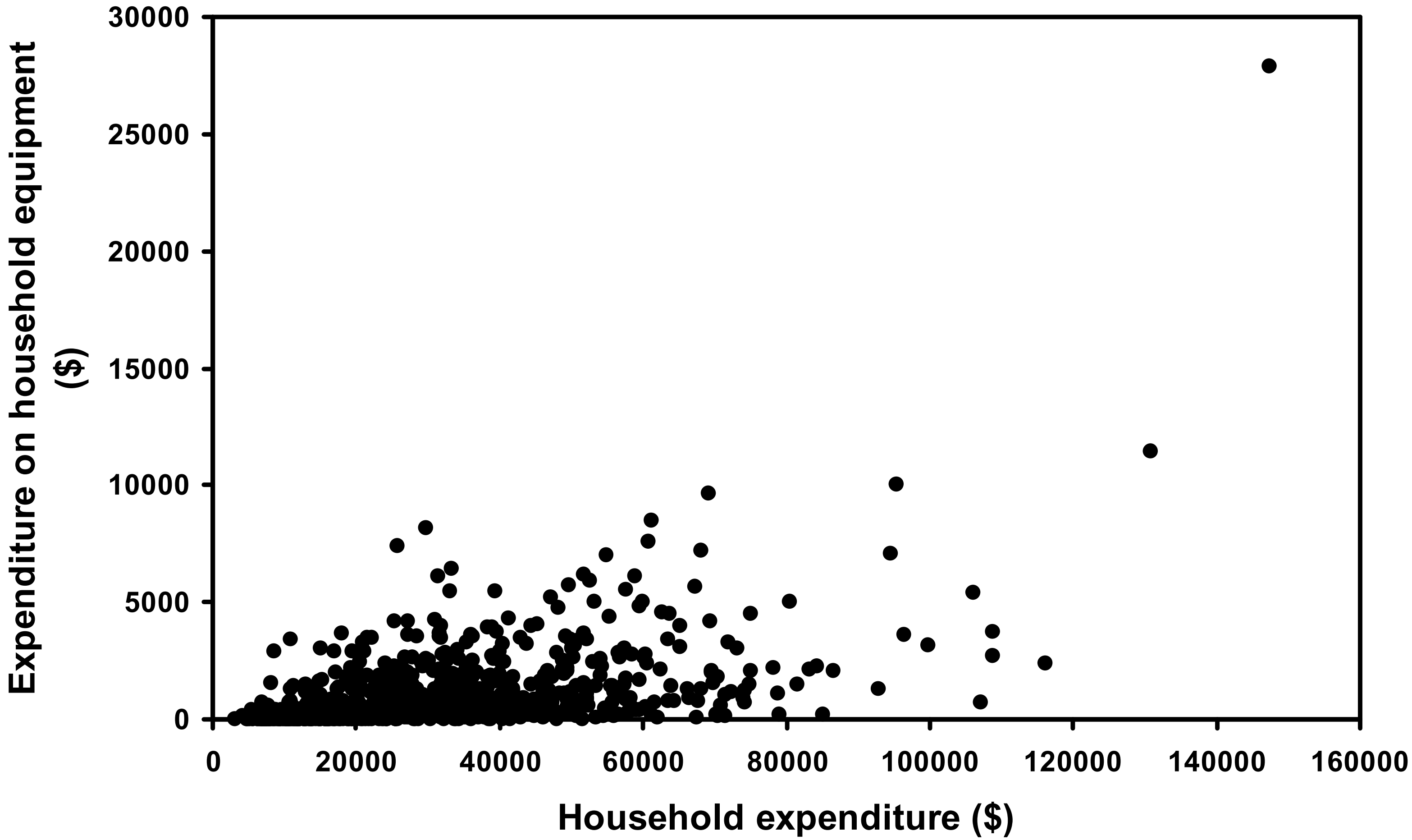
The solution to the problem is to have a hybrid model which effectively uses probit analysis to investigate why some observations have positive *Y while others do not, and then, for those with *Y** > 0, regression analysis to quantify the relationship.**

TOBIT ANALYSIS



The model is fitted using maximum likelihood estimation. We will not be concerned with the technicalities here.

TOBIT ANALYSIS



We will use the Consumer Expenditure Survey data set to illustrate the use of tobit analysis. The figure plots annual household expenditure on household equipment, *HEQ*, on total household expenditure, *EXP*, both measured in dollars.

TOBIT ANALYSIS

```
. tab HEQ if HEQ<10
```

HEQ	Freq.	Percent	Cum.
0	86	89.58	89.58
3	1	1.04	90.62
4	2	2.08	92.71
6	1	1.04	93.75
7	1	1.04	94.79
8	5	5.21	100.00
Total	96	100.00	

For 86 households, *HEQ* was 0. (The tabulation has been confined to small values of *HEQ*. We are only interested in finding out how many actually had *HEQ* = 0.)

TOBIT ANALYSIS

```
. reg HEQ EXP

  Source |       SS       df       MS                  Number of obs =     869
---------+------------------------------               F(  1,   867) =  353.91
   Model |  729289164      1   729289164               Prob > F      =  0.0000
Residual | 1.7866e+09    867  2060635.12               R-squared     =  0.2899
---------+------------------------------               Adj R-squared =  0.2891
   Total | 2.5159e+09    868  2898456.01               Root MSE      =  1435.5

------------------------------------------------------------------------------
     HEQ |      Coef.   Std. Err.       t     P>|t|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
     EXP |   .0471546   .0025065     18.813   0.000        .042235    .0520742
   _cons |  -397.2088   89.44449     -4.441   0.000      -572.7619   -221.6558
------------------------------------------------------------------------------
```

Here is a regression using all the observations. We anticipate that the coefficient of *EXP* is biased downwards.

TOBIT ANALYSIS

```
. reg HEQ EXP if HEQ>0

  Source |       SS       df       MS                  Number of obs =     783
---------+------------------------------               F(  1,   781) =  291.04
   Model |  656349265      1   656349265               Prob > F      =  0.0000
Residual | 1.7613e+09    781  2255219.19               R-squared     =  0.2715
---------+------------------------------               Adj R-squared =  0.2705
   Total | 2.4177e+09    782  3091656.59               Root MSE      =  1501.7

------------------------------------------------------------------------------
     HEQ |      Coef.   Std. Err.       t     P>|t|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
     EXP |   .0467672   .0027414     17.060   0.000       .0413859    .0521485
   _cons |  -350.1704   101.8034     -3.440   0.001      -550.0112   -150.3296
------------------------------------------------------------------------------
```

Here is an OLS regression with the constrained observations dropped. The estimate of the slope coefficient is almost the same, just a little lower.

TOBIT ANALYSIS

```
. tobit HEQ EXP, ll(0)

Tobit Estimates                                   Number of obs =    869
                                                  chi2(1)       = 315.41
                                                  Prob > chi2   = 0.0000
Log Likelihood = -6911.0175                       Pseudo R2     = 0.0223

------------------------------------------------------------------------------
     HEQ |      Coef.   Std. Err.       t     P>|t|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
     EXP |   .0520828   .0027023     19.273   0.000       .0467789    .0573866
   _cons |  -661.8156   97.95977     -6.756   0.000      -854.0813   -469.5499
---------+--------------------------------------------------------------------
     _se |   1521.896    38.6333               (Ancillary parameter)
------------------------------------------------------------------------------

Obs. summary:         86 left-censored observations at HEQ<=0
                     783 uncensored observations
```

Here is the tobit regression. The Stata command is '`tobit`', followed by the dependent variable and the explanatory variables, then a comma, then '`ll`' and in parentheses the lower limit.

TOBIT ANALYSIS

```
. tobit HEQ EXP, ll(0)

Tobit Estimates                                   Number of obs =    869
                                                  chi2(1)       = 315.41
                                                  Prob > chi2   = 0.0000
Log Likelihood = -6911.0175                       Pseudo R2     = 0.0223

------------------------------------------------------------------------------
     HEQ |      Coef.   Std. Err.       t     P>|t|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
     EXP |   .0520828   .0027023     19.273   0.000       .0467789    .0573866
   _cons |  -661.8156   97.95977     -6.756   0.000      -854.0813   -469.5499
---------+--------------------------------------------------------------------
     _se |   1521.896    38.6333             (Ancillary parameter)
------------------------------------------------------------------------------

Obs. summary:        86 left-censored observations at HEQ<=0
                    783 uncensored observations
```

If the dependent variable were constrained by an upper limit, we would use '`ul`' instead of '`ll`', with the upper limit in parentheses. The method can handle lower limits and upper limits simultaneously.

TOBIT ANALYSIS

```
. tobit HEQ EXP, ll(0)
------------------------------------------------------------------------------
         HEQ |      Coef.   Std. Err.      t     P>|t|       [95% Conf. Interval]
-------------+----------------------------------------------------------------
         EXP |   .0520828   .0027023    19.273   0.000      .0467789    .0573866
       _cons |  -661.8156   97.95977    -6.756   0.000     -854.0813   -469.5499
-------------+----------------------------------------------------------------
         _se |   1521.896    38.6333             (Ancillary parameter)
------------------------------------------------------------------------------
. reg HEQ EXP
------------------------------------------------------------------------------
         HEQ |      Coef.   Std. Err.      t     P>|t|       [95% Conf. Interval]
-------------+----------------------------------------------------------------
         EXP |   .0471546   .0025065    18.813   0.000       .042235    .0520742
       _cons |  -397.2088   89.44449    -4.441   0.000     -572.7619   -221.6558
------------------------------------------------------------------------------
. reg HEQ EXP if HEQ>0
------------------------------------------------------------------------------
         HEQ |      Coef.   Std. Err.      t     P>|t|       [95% Conf. Interval]
-------------+----------------------------------------------------------------
         EXP |   .0467672   .0027414    17.060   0.000      .0413859    .0521485
       _cons |  -350.1704   101.8034    -3.440   0.001     -550.0112   -150.3296
------------------------------------------------------------------------------
```

We see that the coefficient of *EXP* is indeed larger in the tobit analysis, confirming the downwards bias in the OLS estimates. In this case the difference is not very great. That is because only 10 percent of the observations were constrained.