

## CURVILINEAR REGRESSION MODELS

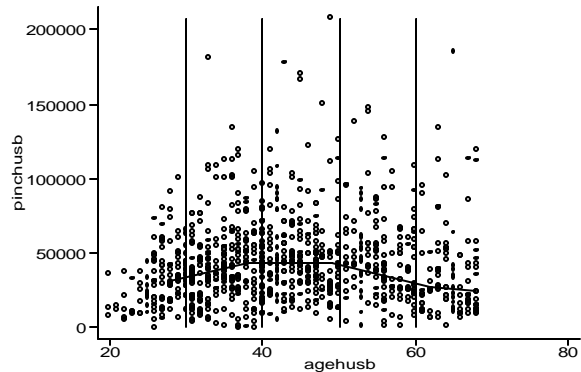
The regression models we have been discussing are *linear* models. We now know a lot about how to handle linear regression. In many cases, however, the relationships we want to study are *curvilinear*, which we discover on the basis of examination of our data through scatterplots and other diagnostics or from our theory about the relationship.

To see whether there is curvilinearity, one method is to use graphs to look at whether the relationship between two variables is linear - by examining the relationship within *bands* defined along the predictor variable.

To illustrate, I will use a dataset on 974 spousal pairs and look at the personal income of the husbands. Here, I restrict consideration to men 20-69 years of age, so that teenagers and men who were in the retirement are omitted. I also omit all men whose personal income (`pinchusb`) is not greater than zero.

```
. use spouses2.dta
. keep if agehusb<69 & agehusb>=20 & pinchusb>0
      (113 observations deleted)
. graph pinchus agehusb, connect(m) bands(5) xlabel ylabel xline(30,40,50,60)
. graph pinchus edhusb, connect(m) bands(5) xlabel ylabel xline(5,10,15)
```

What is graphed here is the *medians within the bands*, connected by a line. The graph to the right shows that personal income first goes up with age, and then declines. The second graph, not shown) shows that the relationship with education is not simply linear. There are also a lot of high outliers.



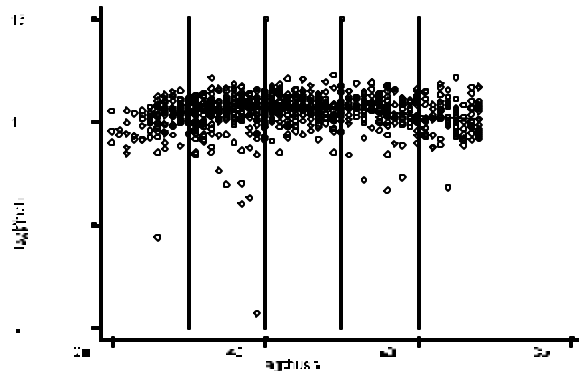
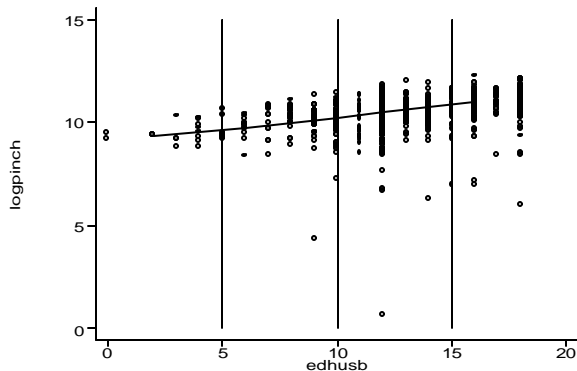
We can use `ladder` to look for a transformation that works. But *none* does. This is because `ladder` is usually *not informative* when samples sizes are large. The reason: the test we use is whether the transformed variable fits a normal distribution well. The larger the sample, the more sensitive is the test to deviations from normality -- and so it gives us P values that are about zero.

One way to handle these data would be to fit curves -- and we'll return to that approach later. BUT curvilinear regression is difficult to carry out. Therefore, we try to *trick* our model into being linear.

One of the easiest ways is one we have already discussed - that of using transformations that make variables more normal in distribution and that reduce skewness and heteroscedasticity in relationships.

One transformation we can use is to take the log of personal income -- attempting to get rid of those unusual high residuals.

```
. gen logpinch = log(pinch)
. graph logpinch edhusb, connect(m) bands(5) xlabel ylabel xline(5,10,15)
. graph logpinch agehusb, connect(m) bands(5) xlabel ylabel xline(30,40,50,60)
```



Next, it is worth thinking about interpretation of coefficients when we do regression using variables that are in log terms. We have talked about transforming X, Y or both. Today, I want to go over the *interpretation* of the resulting coefficients in each of those circumstances.

LOG TRANSFORMATIONS

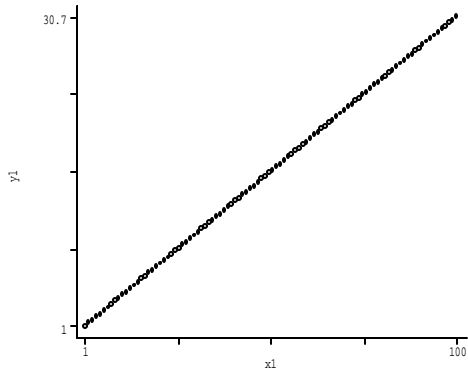
I: Untransformed:  $Y = \beta_0 + \beta_1 X + \epsilon$ ,

Our interpretation of the estimated coefficient,  $b_1$ , is that a change of 1 unit in X is associated with a change of  $b_1$  units in Y.

$$\Delta Y = b_1 \Delta X$$

where  $\Delta$  is the mathematical symbol for "change in".

The example that we use has  $X_1 = 1, \dots, 100$ .  $Y_1 = .7 + .3 * X_1$ . When we plot  $Y_1$  against  $X_1$ , we see the straight line. When we do the regression, we get our results back.



```
. set obs 100
. gen x1=_n
. gen y1 = .7 + .3 * x1
```

```
. regress y1 x1
```

Source	SS	df	MS	Number of obs =	100
Model	7499.25001	1	7499.25001	F( 1, 98) =	.
Residual	1.8190e-11	98	1.8561e-13	Prob > F	= 0.0000
Total	7499.25001	99	75.7500001	R-squared	= 1.0000
				Adj R-squared	= 1.0000
				Root MSE	= 4.3e-07

y1	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
x1	.3	1.49e-09	.	0.000	.3 .3
_cons	.7	8.68e-08	.	0.000	.6999998 .7000002

As always, when  $x_1$  changes by 1,  $y_1$  changes by  $\beta_1$ :

```
. display y1[13] - y1[12] .29999971
```

```
. display y1[50] - y1[49]          .30000019
. display y1[70] - y1[69]          .30000114
```

II: Y is in regular units, X is log units:  $Y_2 = \beta_0 + \beta_1 \ln X_2 + \epsilon$

Our interpretation of the estimated coefficient,  $\beta_1$ , is that a change of 1% in X is associated with a change of  $.01 \beta_1$  in Y.

$$\Delta Y = \beta_1 (\Delta X / X)$$

WHY?  $Y + \Delta Y = \beta_0 + \beta_1 \ln[(1.01) X] = \beta_0 + \beta_1 \ln(X) + \beta_1 \ln(1.01)$

$$\Delta Y = \beta_1 \ln(1.01) \quad \text{but } \ln(1.01) \approx .01$$

so the change in Y is about  $.01 \beta_1$

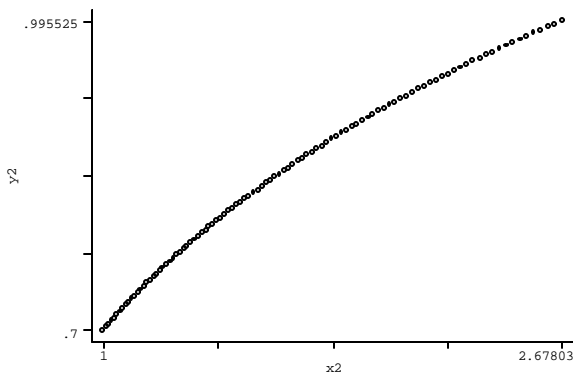
To see this:

```
. gen x2 = 1
. replace x2 = 1.01 * x2[_n-1] in 2/100
. gen lnx2 = ln(x2)
. gen y2 = .7 + .3 * lnx2
```

```
. regress y2 lnx2
```

Source	SS	df	MS	Number of obs =	100
Model	.742493907	1	.742493907	F( 1, 98) =	.
	3.0087e-14	98	3.0701e-16	Prob > F =	0.0000 Residual
Total	.742493907	99	.007499938	R-squared =	1.0000
				Adj R-squared =	1.0000
				Root MSE =	1.8e-08

y2	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
lnx2	.3	6.10e-09	.	0.000	.3 .3
_cons	.7	3.48e-09	.	0.000	.7 .7



What we've done by using the log of X2 instead of the original variable is transform this graph into a straight line from its curvilinear form.

```
. display y2[13]-y2[12]          .00298512
. display y2[50]-y2[49]          .00298512
```

```
. display y2[70]-y2[69]          .00298512
```

In all cases,  $x_2$  changed by 1%, and  $y_2$  then changes by about  $.01*b_1$ .

III. Y is in log units, X is in regular units  $\ln Y = \beta_0 + \beta_1 X + \epsilon$ ,  
suppose  $X$  is 1. Then

$$\ln Y_2 = \beta_0 + \beta_1 (X_1 + 1) \quad \text{and} \quad \ln Y_2 - \ln Y_1 = \beta_1$$

The ratio  $Y_2 / Y_1$  then is give by e raised to the power  $\beta_1$

but if  $\beta_1$  is small, then  $\exp(\beta_1)$  is approximately  $1 + \beta_1$

Conclusion: if X changes by 1, then Y changes by  $100 \beta_1$  %

```
. gen x3 = x1
. gen lny3 = .7 + .3 * x3
. gen y3 = exp(lny3)
```

What we've done, then is to trick the curvilinear graph of  $y_3$  against  $x_3$  into becoming the linear relationship between  $y$  and  $\ln(x)$ . The ratio of each  $y_3$  to the previous one is  $100 \beta_1$  % (approximately!)

```
. display y3[13]/y3[12]          1.3498585
. display y3[50]/y3[49]          1.3498591
. display y3[70]/y3[69]          1.3498604
```

More precisely, the ratio is  $e^{.3}$ , which can be found in STATA as:

```
. display exp(.3)                1.3498588
```

IV. Both X and Y are in log units  $\ln Y = \beta_0 + \beta_1 \ln X + \epsilon$ ,

$$\ln Y_2 = \beta_0 + \beta_1 \ln[(1.01)X] = \beta_0 + \beta_1 \ln X + \beta_1 \ln(1.01)$$

$$\ln Y_2 - \ln Y_1 = \ln(Y_2 / Y_1) = \beta_1 \ln(1.01) \quad \text{or} \quad Y \text{ changes by about } \beta_1 \%$$

A frequently used example is the demand function in economics

$$Q_i = A P_i^B$$

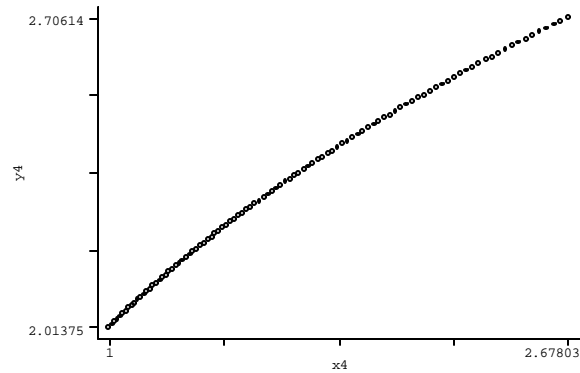
$\ln Q_i = \ln A + B \ln P_i$  where  $P$  = price and  $Q$  = quantity

If price increases by 1% quantity increases by  $B\%$  ---  $B$  is the *elasticity* of the given good.

What sign do you expect on  $B$ ?

Returning to our examples:

```
. gen x4 = x2
. gen lnx4 = ln(x4)
. gen lny4 = .7 + .3 * lnx4
. gen y4 = exp(lny4)
```



```
. display y4[13]/y4[12]          1.0029895
. display y4[50]/y4[49]          1.0029896
. display y4[70]/y4[69]          1.0029895
```

#### SUMMARY OF INTERPRETATION WHEN VARIABLES ARE IN LOG TERMS:

**X**

! If X is in its original scale, the coefficient of X represents the *absolute change* in the dependent variable expected when X increases by 1 unit (models 1 and 3)

! If the predictor is, instead, log X, then the interpretation of the coefficient of log X is that the *absolute change* in the dependent variable when X increases by 1% is .01 times the coefficient (models 2 and 4)

**Y:**

! If Y is in its original scale, the results above hold for models 1 and 2.

! If lnY is used as the dependent variable, then the results above hold for the predicted change in **lnY**. But for **Y** itself:

a. if X is in the original scale, then Y changes by  $100 \cdot b_1\%$  when X increases by 1 (i.e.  $b_1 = .3$  means Y increases 30% for every 1 unit change in X)

b. if lnX is used, the Y changes by  $b_1\%$  when X increases by 1% (i.e.  $b_1 = .3$  means Y increases by .3% for every 1% increase in X)

USE OF  $X^2$  TERM TO DETECT CURVILINEARITY

Economists frequently add this term to regressions to detect curvilinearity. We will look at several different relationships where there really are curves and see if regression leads to significant coefficients on those squared terms.

```
. regress pinchu edhus edh2 agehu ageh2
```

Source	SS	df	MS	Number of obs =	857
Model	1.0564e+11	4	2.6410e+10	F( 4, 852) =	72.61
Residual	3.0989e+11	852	363719811	Prob > F =	0.0000
				R-squared =	0.2542
				Adj R-squared =	0.2507
Total	4.1553e+11	856	485433546	Root MSE =	19071

pinchusb	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
edhusb	-877.0916	972.7439	-0.902	0.367	-2786.347 1032.164
edh2	160.9159	39.7684	4.046	0.000	82.8604 238.9714
agehusb	2246.841	405.1229	5.546	0.000	1451.685 3041.997
ageh2	-23.32859	4.413251	-5.286	0.000	-31.99071 -14.66647
_cons	-37172.4	10515.77	-3.535	0.000	-57812.25 -16532.55

Here the interpretation of the coefficients CANNOT be done singly. The coefficients of the age variable TOGETHER tell us that the relationship to age is *curvilinear*.

## USE OF DUMMY VARIABLES TO DETECT CURVILINEARITY

For completeness, I note that we could divide our age variable into categories. I will use 4 categories representing quartiles of agehusb – this is similar to the approach we used in an earlier example where we divided years of education in 4 categories (<HS, HS, Somecl, College+), but there we used “natural” categories whereas here I’m using quartiles.

```
. sum agehusb, detail
```

Percentiles		Smallest		
1%	22	20		
5%	26	20		
10%	29	20	Obs	861
25%	34	21	Sum of Wgt.	861
50%	42		Mean	43.8676
		Largest	Std. Dev.	12.23001
75%	54	68		
90%	62	68	Variance	149.5731
95%	65	68	Skewness	.2720456
99%	68	68	Kurtosis	2.059909

```
. gen young =0
. replace young=1 if ageh<=34
(227 real changes made)

. gen ymid =0
. replace ymid=1 if ageh>34 & ageh<=42
(208 real changes made)

. gen omid = 0
. replace omid=1 if ageh>42 & ageh<=54
(222 real changes made)

. gen old=0
. replace old=1 if ageh>54
(204 real changes made). regress logpinch ymid omid old
```

Source	SS	df	MS	Number of obs =	861
Model	22.9401091	3	7.64670302	F( 3, 857) =	10.36
Residual	632.543137	857	.738090008	Prob > F	= 0.0000
				R-squared	= 0.0350
				Adj R-squared	= 0.0316
Total	655.483246	860	.762189821	Root MSE	= .85912

logpinch	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
ymid	.2687645	.0824622	3.259	0.001	.106913	.4306159
omid	.4041256	.0810939	4.983	0.000	.2449596	.5632916
old	.0626667	.082883	0.756	0.450	-.1000107	.2253441
_cons	10.24392	.0570219	179.649	0.000	10.13201	10.35584