BS2280 – Econometrics I
Homework 9: Nonlinear Models and Transformation of Variables I - Solution

1

Until now we have assumed that our regression model is linear in variables and parameters. Explain what this means.

A regression is linear in variables if the variables on the right-hand side of the equation are included exactly as defined, rather than as function. E.g.

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + ... + \beta_k X_{ki} + u_i$$

where X_2 could be age of a person measured in years and X_3 the work experience measured in year.

A regression is linear in parameters if all the coefficients (betas) are just multiplied with the variables on the right hand side of the equation. E.g.

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + ... + \beta_k X_{ki} + u_i$$

where $\beta_2 X_{2i}$ shows the impact of a specific level of X_2 on Y.

2

It has often been observed that there is a weak tendency for years of schooling to be inversely related to the number of siblings (brothers and sisters) of an individual. The regression shown below has been fitted on the hypothesis that the adverse effect is nonlinear. Z is defined as the reciprocal of the number of siblings, for individuals with at least one sibling. Sketch the regression relationship and provide an interpretation of the regression results.

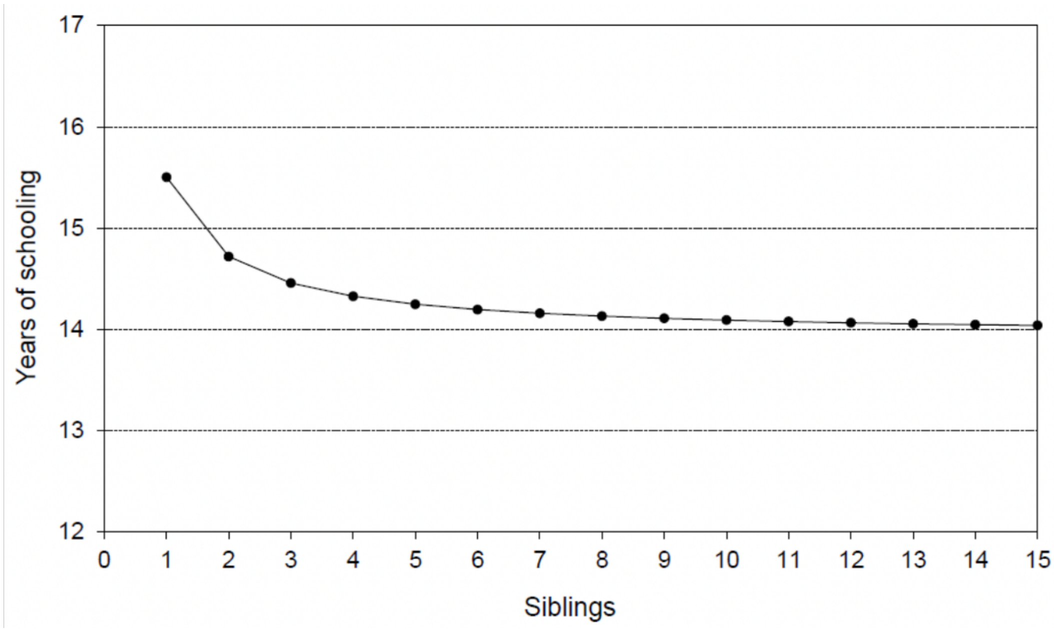
```
> EAWE21$Z <- 1 / EAWE21$SIBLINGS
> Sfit <- lm(S~Z,data=subset(EAWE21, SIBLINGS>0))
> summary(Sfit)

Call:
lm(formula = S ~ Z, data = subset(EAWE21, SIBLINGS > 0))

Residuals:
    Min      1Q  Median      3Q     Max
-7.5028 -2.4569  0.4972  1.8917  5.6738

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  13.9340     0.2710   51.41  < 2e-16 ***
Z             1.5688     0.4023    3.90  0.00011 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.727 on 471 degrees of freedom
Multiple R-squared:  0.03128,   Adjusted R-squared:  0.02923
F-statistic: 15.21 on 1 and 471 DF,  p-value: 0.0001102
```

Given the definition of Z, its coefficient does not have any straightforward interpretation. The best one can say is that the relationship predicts between 15 and 16 years of schooling for those with only one sibling and that the relationship approaches a lower limit of 14 years of schooling for those with many siblings. The relationship is not defined for those with no siblings.

We have to transform the results so that we get more meaningful results. Write down the equation where you substitute $\frac{1}{Siblings_i}$ for Z:

$$\hat{S}_i = 13.93 + 1.57\frac{1}{Siblings_i}$$

By substituting number of siblings into the equation, you will get the data to draw the nonlinear regression line diagram.



To calculate the marginal effect, you have to take the first derivative:

$$\frac{\partial S_i}{\partial Siblings_i} = -1.57 \frac{1}{Siblings_i^2}$$

(Make use of the power rule: Derivative of $\frac{1}{x}$ is $-\frac{1}{x^2}$)

Now you can calculate the marginal effect of number of siblings on educational attainment. As expected, the sign is negative, therefore having more siblings has an adverse effect on educational attainment.

3

The output below shows the result of regressing $LGWT04$, the logarithm of weight in 2004, measured in pounds, on $LGHEIGHT$, the logarithm of height, measured in inches. Provide an interpretation of the slope coefficient and evaluate the regression results.

```
> EAWE21$LGWT04 <- log(EAWE21$WEIGHT04)
> EAWE21$LGHEIGHT <- log(EAWE21$HEIGHT)
> LGWT04fit <- lm(LGWT04~LGHEIGHT,data=EAWE21)
> summary(LGWT04fit)

Call:
lm(formula = LGWT04 ~ LGHEIGHT, data = EAWE21)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.40323 -0.13720 -0.03225  0.10760  0.61840 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  -3.7883     0.6109  -6.201 1.18e-09 ***
LGHEIGHT      2.1064     0.1449  14.536  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1935 on 498 degrees of freedom
Multiple R-squared:  0.2979,	Adjusted R-squared:  0.2965 
F-statistic: 211.3 on 1 and 498 DF,  p-value: < 2.2e-16
```

We have to transform the results so that we get more meaningful results. Write down the equation:

$$\widehat{LGWT}04_i = -3.78 + 2.10LGHEIGHT_i$$

We have here a log-log model! Therefore the slope coefficient indicates that the elasticity of weight with respect to height was 2.10, the coefficient being highly significant statistically.
A 1% increase in height will lead to a 2.10% increase in weight. As should be expected, variation in height is responsible for some of the variation in weight, but other factors are evidently important. The intercept has no meaning.

4

The output below shows the result of regressing $LGWT04$, the logarithm of weight in 2004, measured in pounds, on $HEIGHT$, height measured in inches. Provide an interpretation of the slope coefficient and evaluate the regression results.

We have to transform the results so that we get more meaningful results. Write down the equation:

$$\widehat{LGWT}04_i = 2.98 + 0.03HEIGHT_i$$

```
> LGWT04fit2 <- lm(LGWT04~HEIGHT,data=EAWE21)
> summary(LGWT04fit2)

Call:
lm(formula = LGWT04 ~ HEIGHT, data = EAWE21)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.40150 -0.13819 -0.03537  0.10552  0.62194 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 2.988730   0.145213   20.58   <2e-16 ***
HEIGHT      0.030990   0.002137   14.50   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1936 on 498 degrees of freedom
Multiple R-squared:  0.2969,    Adjusted R-squared:  0.2955 
F-statistic: 210.3 on 1 and 498 DF,  p-value: < 2.2e-16
```

Here we have a log-level model, also known as a semi-log model! Since it is a semilogarithmic specification, the coefficient of HEIGHT indicates that an extra inch of height increases weight by 3.1 percent. The coefficient is highly significant statistically. Note that the intercept will be meaningless as a person cannot have a height of 0 inches!