Single Variable Regression: Transformations and Functional Form

EC420 MSU Spring 2021

Justin Kirkpatrick
Last updated December 10, 2020

This lecture

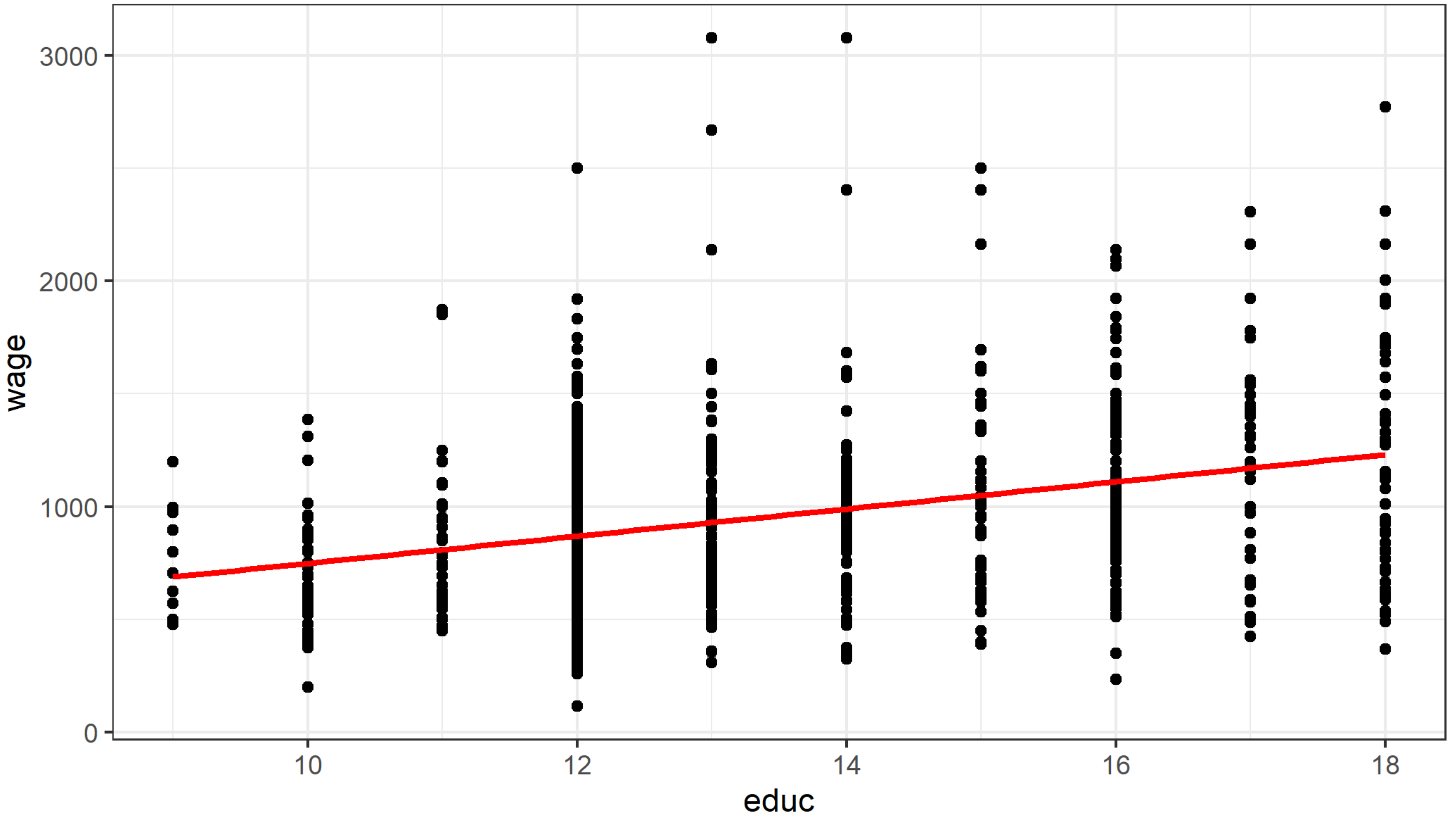
Goal:

1. Interpretation of regression coefficients
2. Re-scaling
3. Re-scaling Y
4. Non-linear functional forms
5. Intuition and uses of non-linear forms in economics
6. Regression in R

Interpretation

Last time, we discussed a single variable regression from Wooldridge `wage2` where Y is $wage$ and is $educ$:

$$wage = \beta_0 + \beta_1 educ + u$$



This resulted in a $\hat{\beta}_1 = 60.21$. How do we interpret this coefficient?

Interpretation

Let's start with our simple linear regression model:

where $wage$ and $educ$ are random variables

$$wage = \beta_0 + \beta_1 educ + u$$

Our PRF is:

$$E[wage|educ] = \beta_0 + \beta_1 educ$$

Interpretation

MICHIGAN STATE UNIVERSITY

Let's start with our simple linear regression model:

where $wage$ and $educ$ are random variables

$$wage = \beta_0 + \beta_1 educ + u$$

Our PRF is:

$$E[wage|educ] = \beta_0 + \beta_1 educ$$

- "One additional year of education is associated with a 60.21 increase in expected monthly earnings, all else held equal"
- Why "all else held equal"? Because we have assumed that $E[U| \quad] = 0$, so our estimate tells us how $E[Y]$ changes as *and not* U changes.
  - U is held at zero

Interpretation

Ceteris Paribus

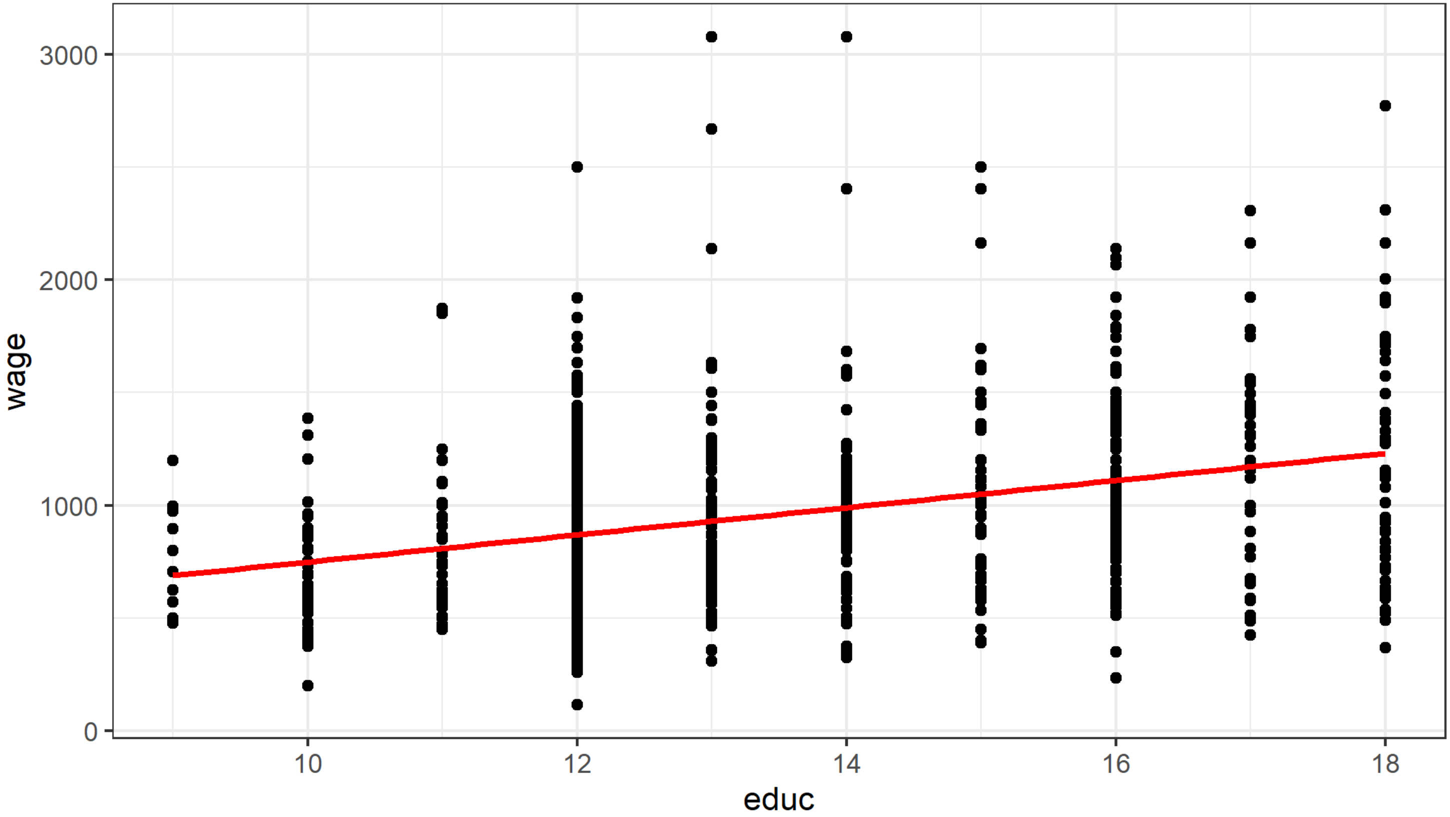
Latin for "all else held equal"

So $\hat{\beta}_1$ is

"the increase in the expectation of $wage$ associated with a 1-unit increase in $educ$, ceteris paribus"

The "all else held equal" part is very important.

Interpretation



- $\hat{\beta}_1$ is $\frac{wage}{educ}$
- $\hat{\beta}_1$ is the slope of the line
  - The line is $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$, the SRF

# Interpretation

```
myRegression = lm(wage ~ educ, data=wage2)
summary(myRegression)
```

```

Call:
lm(formula = wage ~ educ, data = wage2)

Residuals:
    Min      1Q  Median      3Q     Max 
-877.38 -268.63  -38.38  207.05 2148.26 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  146.952     77.715   1.891   0.0589 .  
educ          60.214      5.695  10.573   <2e-16    
---
Signif. codes:  0 '   ' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 382.3 on 933 degrees of freedom
Multiple R-squared:  0.107,	Adjusted R-squared:  0.106 
F-statistic: 111.8 on 1 and 933 DF,  p-value: < 2.2e-16
```

Rescaling Y and X

What happens if we re-scale the dependent variable, wage?

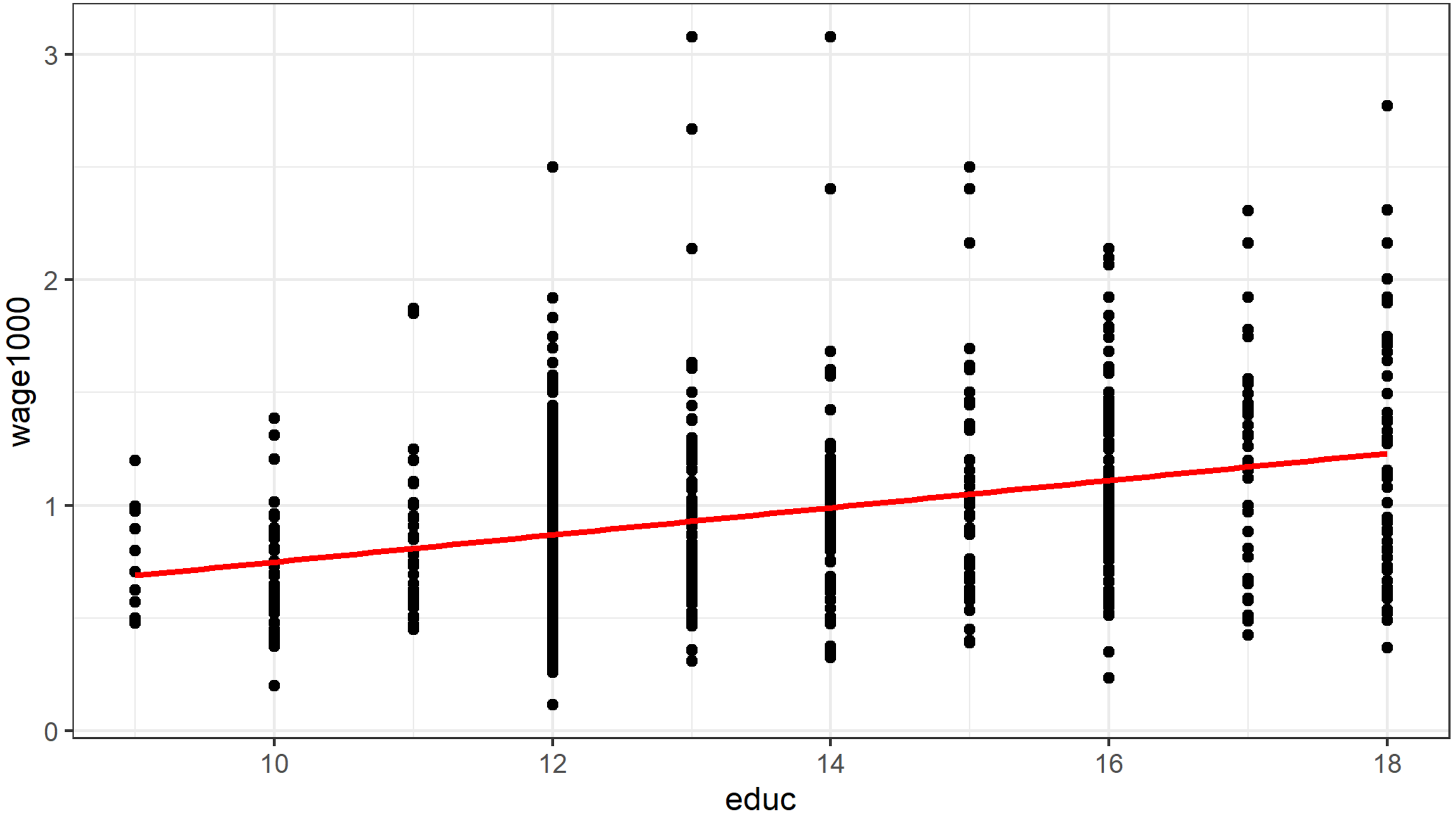
Maybe we have $wage$ in dollars, but want it in thousands of dollars

We hope that it still gives us the same relationship

Define $wage1000 = .001 \times wage$

- Any ideas what will happen to our coefficient?

Rescaling Y and X



Looks pretty similar, right? But the y-axis scale is very different.

Rescaling Y and X

A regression of:

$$wage1000 = \beta_0 + \beta_1 educ + u$$

```
Call:
lm(formula = wage1000 ~ educ, data = wage2)

Residuals:
     Min       1Q   Median       3Q      Max
-0.87738 -0.26863 -0.03838  0.20705  2.14826

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.146952   0.077715   1.891   0.0589 .
educ        0.060214   0.005695  10.573   <2e-16
---
Signif. codes:  0 '   ' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3823 on 933 degrees of freedom
Multiple R-squared:  0.107,    Adjusted R-squared:  0.106
F-statistic: 111.8 on 1 and 933 DF,  p-value: < 2.2e-16
```

Rescaling Y and X

$\hat{\beta}_1 = \mathbf{0.06}$ when we use $\boldsymbol{wage1000}$

$\hat{\beta}_1 = \mathbf{60.21}$ when we use $\boldsymbol{wage}$.

Rescaling Y and X

$\hat{\beta}_1 = 0.06$ when we use $wage1000$

$\hat{\beta}_1 = 60.21$ when we use $wage$.

Re-scaling the dependent variable, *wage*, results in an equal rescaling of the coefficient.

The relationship predicted by the *SRF* stays the same.

Rescaling Y and X

Now, let's re-scale the *independent* variable

- That's the "right hand side" variable, $educ$.
- Let's do education in months: $educ\quad onths = educ \times 12$

Rescaling Y and X

Now, let's re-scale the *independent* variable

- That's the "right hand side" variable, $educ$.
- Let's do education in months: $educ \quad onths = educ \times 12$
- Any predictions on what will result?

Rescaling Y and X

```
Call:
lm(formula = wage ~ educMonths, data = wage2)

Residuals:
    Min      1Q  Median      3Q     Max
-877.38 -268.63  -38.38  207.05 2148.26

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 146.9524    77.7150   1.891   0.0589 .
educMonths    5.0179     0.4746  10.573   <2e-16
---
Signif. codes:  0 '   ' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 382.3 on 933 degrees of freedom
Multiple R-squared:  0.107,    Adjusted R-squared:  0.106
F-statistic: 111.8 on 1 and 933 DF,  p-value: < 2.2e-16
```

What was the result?

Rescaling Y and X

Re-scaling the independent variable simply rescales the coefficient by the *inverse* amount:

- $12 \times educ \Rightarrow \hat{\beta}_1^{new} = \frac{\hat{\beta}_1}{12}$

Rescaling Y and X

Re-scaling the independent variable simply rescales the coefficient by the *inverse* amount:

- $12 \times educ \Rightarrow \hat{\beta}_1^{new} = \frac{\hat{\beta}_1}{12}$

Re-scaling the dependent variable simply rescales the coefficient on it by an equal amount:

- $\hat{\beta}_1^{new} = \hat{\beta}_1 \times .001$

Rescaling Y and X

Re-scaling the independent variable simply rescales the coefficient by the *inverse* amount:

- $12 \times educ \Rightarrow \hat{\beta}_1^{new} = \frac{\hat{\beta}_1}{12}$

Re-scaling the dependent variable simply rescales the coefficient on it by an equal amount:

- $\hat{\beta}_1^{new} = \hat{\beta}_1 \times .001$

The relationship always remains the same

Rescaling Y and X

Let's take a look at the R^2 of the original regression:

```

Call:
lm(formula = wage ~ educ, data = wage2)

Residuals:
    Min      1Q  Median      3Q     Max 
-877.38 -268.63  -38.38  207.05 2148.26 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  146.952     77.715   1.891   0.0589 .  
educ          60.214      5.695  10.573   <2e-16    
---
Signif. codes:  0 '   ' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 382.3 on 933 degrees of freedom
Multiple R-squared:  0.107,	Adjusted R-squared:  0.106 
F-statistic: 111.8 on 1 and 933 DF,  p-value: < 2.2e-16
```

Rescaling Y and X

Now, the re-scaled dependent variable:

```
Call:
lm(formula = wage1000 ~ educ, data = wage2)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.87738 -0.26863 -0.03838  0.20705  2.14826 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 0.146952   0.077715   1.891   0.0589 .  
educ        0.060214   0.005695  10.573   <2e-16 
---
Signif. codes:  0 '   ' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3823 on 933 degrees of freedom
Multiple R-squared:  0.107,	Adjusted R-squared:  0.106 
F-statistic: 111.8 on 1 and 933 DF,  p-value: < 2.2e-16
```

Rescaling Y and X

And the re-scaled independent variable:

```
Call:
lm(formula = wage ~ educMonths, data = wage2)

Residuals:
    Min      1Q  Median      3Q     Max 
-877.38 -268.63  -38.38  207.05 2148.26 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 146.9524    77.7150   1.891   0.0589 .  
educMonths    5.0179     0.4746  10.573   <2e-16 
---
Signif. codes:  0 '   ' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 382.3 on 933 degrees of freedom
Multiple R-squared:  0.107,	Adjusted R-squared:  0.106 
F-statistic: 111.8 on 1 and 933 DF,  p-value: < 2.2e-16
```

Rescaling Y and X

Heck, let's rescale both and look at the R^2

```
Call:
lm(formula = wage1000 ~ educMonths, data = wage2)

Residuals:
     Min       1Q   Median       3Q      Max
-0.87738 -0.26863 -0.03838  0.20705  2.14826

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.1469524  0.0777150   1.891   0.0589 .
educMonths  0.0050179  0.0004746  10.573   <2e-16
---
Signif. codes:  0 '   ' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3823 on 933 degrees of freedom
Multiple R-squared:  0.107,    Adjusted R-squared:  0.106
F-statistic: 111.8 on 1 and 933 DF,  p-value: < 2.2e-16
```

Rescaling Y and X

The R^2 is the same in every single one!

The "fraction of variance explained by the model" does not change.

Intuitively, you shouldn't be able to explain more variance simply by re-scaling a variable. The relationship that holds for wages and years of education must hold for 12 x years of education as well.

Rescaling Y and X

The R^2 is the same in every single one!

The "fraction of variance explained by the model" does not change.

Intuitively, you shouldn't be able to explain more variance simply by re-scaling a variable. The relationship that holds for wages and years of education must hold for 12 x years of education as well.

Since rescaling linearly doesn't matter, we can use a scale that is easiest to interpret and to read.

- $wage1000$ in thousands of dollars is a lot easier to look at than the larger number we get using $wage$.
- You often don't want to have very extreme numbers of decimal places (e.g. a coefficient of .00000051 will be a lot easier to talk about if it's in millions: 5.1)

Rescaling Y and X

Now that we've seen an example, can we derive this result from the definition of β_1 ?

$$\beta_1 = \frac{Cov(\quad, Y)}{Var(\quad)}$$

$$\beta_1^{rescaled} = \frac{Cov(a\quad, Y)}{Var(a\quad)}$$

Let's do this in class....

Non-linear Functional Forms

Non-linear functional forms

What do we mean by "non-linear" function?

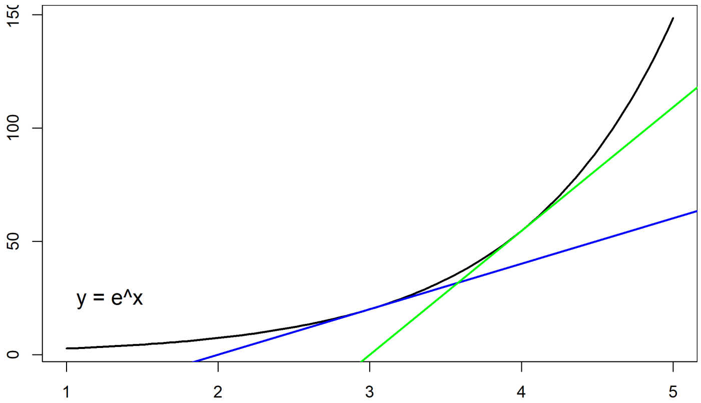
A function here is any mathematical operation or transformation that takes an input (usually called x) and returns an output (usually called y).

A non-linear function is any function where the graph is not a straight line.

- "Affine transformation" is the technical term for $y = ax + b$.
- "Non-affine transformation" is non-linear

Non-linear Another way of thinking about non-linear functions is that ` $\frac{y}{x}$ $dependsonthevalueofx$`

- The slope of the graph changes as x changes.
- The slope at x_1 (blue) is different than the slope at x_2 (green)



Non-linear functional forms

In the previous slide, we saw a non-linear function, the exponential function, e^x. If we wanted a model to use in a regression that includes an exponential function, we could use:

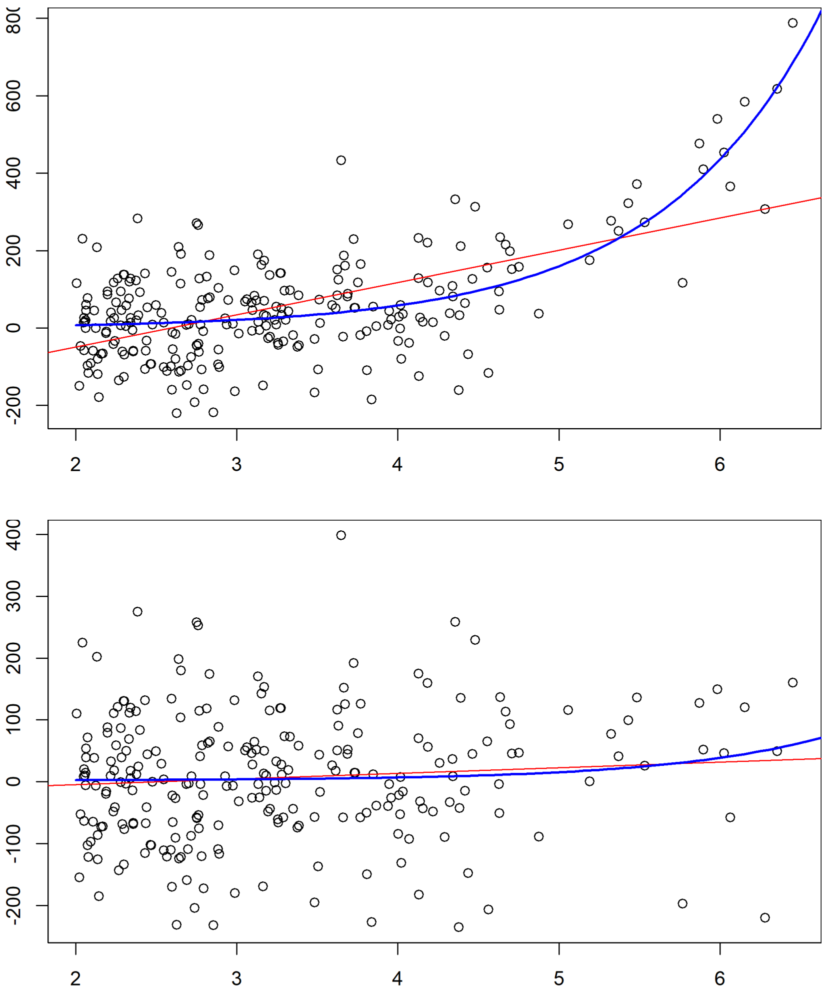
$$y_i = \beta_0 + \beta_1 e^{x_i} + u_i$$

Note that the value of x_i is exponentiated.

- So this model has a non-linear term.
- It lets y respond to changes in x more flexibly

Non-linear functional forms

- but imposes that relationship whether it is appropriate (top) or not (bottom).



Non-linear functional forms

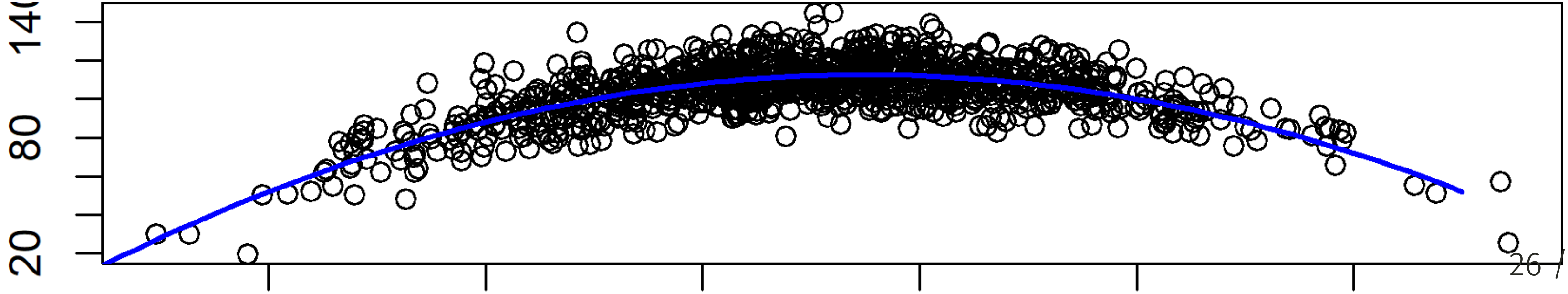
The most common non-linear transformation is the **polynomial**

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \cdots + u$$

For instance, plant growth rates over temperatures may be quadratic

- The *marginal effect* of an increase in temperature will be big and positive at lower temperatures.
- The *marginal effect* of an increase in temperature will be negative at very high temperatures.
- And somewhere in the middle, the *marginal effect* will be around zero.

The *marginal effect* is another way of saying "the change in y per change in x", or $\frac{dy}{dx}$.



Non-linear functional forms

If we have a polynomial relationship:

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + u$$

Then we can obtain the slope, $\frac{dy}{dx}$ as the derivative of the relationship:

$$\frac{\partial y}{\partial x} = \beta_1 + 2\beta_2 x$$

Non-linear functional forms

If we have a polynomial relationship:

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + u$$

Then we can obtain the slope, $\frac{dy}{dx}$ as the derivative of the relationship:

$$\frac{\partial y}{\partial x} = \beta_1 + 2\beta_2 x$$

If we propose a "higher order polynomial" relationship like:

$$y = \beta_0 + \beta 1 x + \beta_2 x^2 + \beta_3 x^3$$

Then we get a more complicated function for the slope at any x:

$$\frac{\partial y}{\partial x} = \beta_1 + 2\beta_2 x + 3\beta_3 x^2$$

Non-linear functional forms

There are other possible non-linear forms: $\overline{x}$, the natural log, log_{10}, the inverse hyperbolic sine...

Non-linear functional forms

There are other possible non-linear forms: $\overline{x}$, the natural log, log_{10}, the inverse hyperbolic sine…

Even though these specifications are non-linear transformations, the regression is still **linear-in-parameters**

That is, all of the transformations we have discussed are still in the category of "linear models" because they are linear in the parameters.

So, our PRF (population regression function) is still linear, even with one of these transformations.

Intuition and uses in economics

The quadratic specification, $y = \beta_0 + \beta_1 x + \beta_2 x^2$ is particularly useful anytime you have an effect of x on y that dissipates or declines with increasing values of x.

Quick question: if the *effect* of x on y **declines** as x increases, then is the slope *increasing* or *decreasing* as x gets larger?

Intuition and uses in economics

An example:

In many cases, the effect of household income on some behavior may change as income increases.

- A low-income person may spend more on food when income increases
- But a high-income person may not spend much more on food when their income increases
    - But of course, the high-income will spend more on food than the low-income person.

We see these declining effects in many economic situations, but we also see increasing effects.

- Installing solar panels
- Others?

Intuition and uses in economics

An example:

In many cases, the effect of household income on some behavior may change as income increases.

- A low-income person may spend more on food when income increases
- But a high-income person may not spend much more on food when their income increases
    - But of course, the high-income will spend more on food than the low-income person.

We see these declining effects in many economic situations, but we also see increasing effects.

- Installing solar panels
- Others?

The quadratic "specification" can capture these phenomon.

Intuition and uses in economics

MICHIGAN STATE UNIVERSITY

The natural log, $ln(x)$

The natural log is the most common transformation. It is particularly useful because of the following:

$$ln(1+x) \approx x \quad \text{when} \quad x \approx 0$$

Let's say $x^1 = x^0 + \quad x$.

$$ln(x^1) - ln(x^0) = ln \quad \frac{x^1}{x^0} \quad = ln \quad \frac{x^0 + \quad x}{x^0} \quad = ln \quad 1 + \frac{\quad x}{x^0} \quad \approx \frac{\quad x}{x^0}$$

- This is the percent change in x: $\frac{\quad x}{x}$
- $100 \times [ln(x^1) - ln(x^0)] \approx \qquad x$

Intuition and uses in economics

The natural log, $ln(x)$

Recall the formula for *elasticity*: $\frac{y}{x} = \frac{y}{x} \times \frac{x}{y}$

Intuition and uses in economics

The natural log, $ln(x)$

Recall the formula for *elasticity*: $\frac{y}{x} = \frac{y}{x} \times \frac{x}{y}$

And recall that, in a linear model ($y = \beta_0 + \beta_1 x$), this elasticity is **not** constant:

$$\frac{y}{x} \times \frac{x}{y} = \beta_1 \times \frac{x}{y} = \beta_1 \times \frac{x}{\beta_0 + \beta_1 x + u}$$

Intuition and uses in economics

But, when a model takes the form: $ln(y) = \beta_0 + \beta_1 ln(x)$

$$\frac{y}{x} \approx \frac{ln(y^1) - ln(y^0)}{ln(x^1) - ln(x^0)} = \frac{\beta_1[ln(x^1) - ln(x^0)]}{ln(x^1) - ln(x^0)} = \beta_1$$

The coefficient on a log-log model is the elasticity

$ln(y) = \beta_0 + \beta_1 ln(x)$ results in β_1 being the elasticity of y, or "percent change in y from a 1 percent change in x".

Econometrics is frequently about estimating that elasticity.

First, data

There is a very helpful packages called "wooldridge" that you can install with `install.packages('wooldridge')`. Then, we can use R's built-in "data" function to load `wage2`

```r
require(wooldridge)
data(wage2) # creates a wage2 object
print(wage2[1:5,])
```

```
  wage hours  IQ KWW educ exper tenure age married black south urban sibs
1  769    40  93  35   12    11      2  31       1     0     0     1    1
2  808    50 119  41   18    11     16  37       1     0     0     1    1
3  825    40 108  46   14    11      9  33       1     0     0     1    1
4  650    40  96  32   12    13      7  32       1     0     0     1    4
5  562    40  74  27   11    14      5  34       1     0     0     1   10
  brthord meduc feduc    lwage
1       2     8     8 6.645091
2      NA    14    14 6.694562
3       2    14    14 6.715384
4       3    12    12 6.476973
5       6     6    11 6.331502
```

Second, run the regression

We will use the `lm()` function. You will provide the regression formula and the name of the data to use.

The formula will be of the form *y* ~ *x*. You'll specify the data with `data = wage2`

```
MyRegression = lm(wage ~ educ, data=wage2)
print(MyRegression)
```

```
Call:
lm(formula = wage ~ educ, data = wage2)

Coefficients:
(Intercept)         educ
     146.95        60.21
```

Regression in R

Finally, we want a little more detail.

`MyRegression` is an R object. We can ask R to summarize it, and R will know to give us information about the regression:

Regression in R

```
summary(MyRegression)
```

```
Call:
lm(formula = wage ~ educ, data = wage2)

Residuals:
    Min      1Q  Median      3Q     Max
-877.38 -268.63  -38.38  207.05 2148.26

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  146.952     77.715   1.891   0.0589 .
educ          60.214      5.695  10.573   <2e-16
---
Signif. codes:  0 '   ' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 382.3 on 933 degrees of freedom
Multiple R-squared:  0.107,    Adjusted R-squared:  0.106
F-statistic: 111.8 on 1 and 933 DF,  p-value: < 2.2e-16
```