Non-linearities in Simple Regression

1. Example: Monthly Earnings and Years of Education

In this tutorial, we will focus on an example that explores the relationship between total monthly earnings and years of education.

The code below downloads a CSV file that includes data from 1980 for 935 individuals on variables including their total monthly earnings (`MonthlyEarnings`) and a number of variables that could influence income, including years of education (`YearsEdu`) and assigns it to a dataset that we call `wages`.

```
wages <- read.csv("http://murraylax.org/datasets/wage2.csv");
```

We estimate the simple regression with the following call to `lm()` and store the output in an object we call `lmwages`:

```
lmwages <- lm(wages$MonthlyEarnings ~ wages$YearsEdu)
```

2. Log Function

It may not be appropriate that there is a *linear* relationship between years of education and monthly earnings. With a linear relationship, we assume that each year of education results in the same dollar increase in monthly earnings.

It may be more appropriate to suggest that each year of education leads to a similar *percentage* increase in monthly earnings. To estimate such a relationship, we estimate the following regression equation that includes the natural logarithm of the dependent variable (monthly earnings):

$$ln(y_i) = b_0 + b_1 x_i + e_i$$

where y_i denotes the *income* of individual i, $ln(y_i)$ is the natural logarithm of y_i, and x_i denotes the number of *years of education* of individual i.

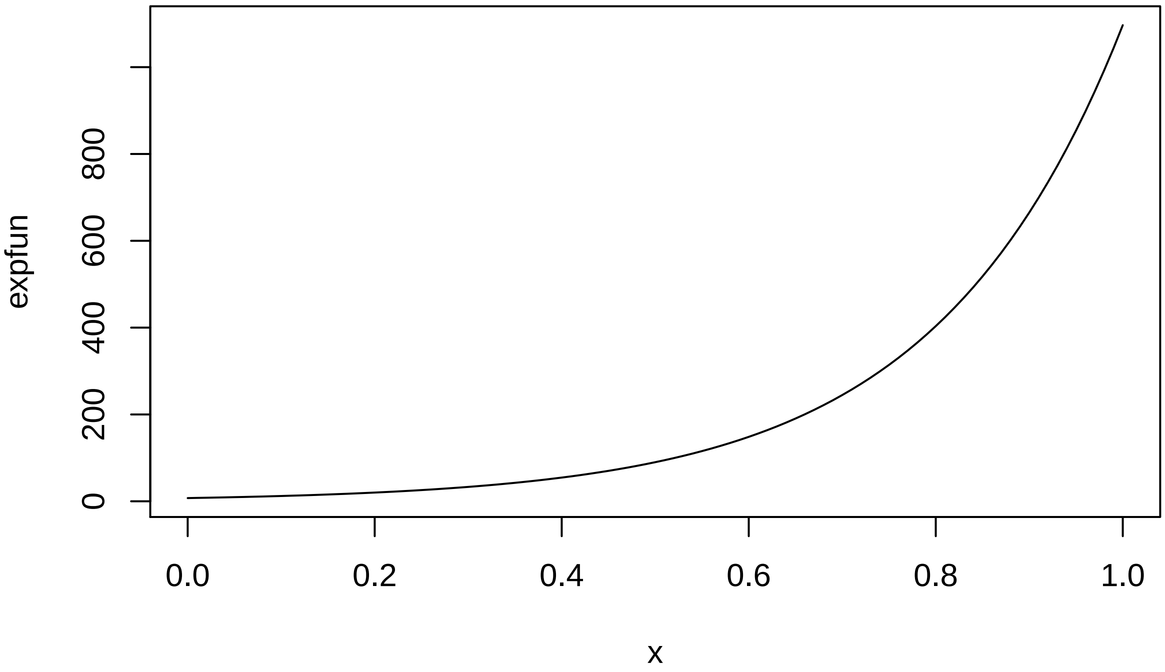
When we have a relationship of the form $ln(y) = b_0 + b_1 x$, this can be transformed to the exponential function, $y = \exp b_0 + b_1 x$. To get an idea what this function looks like, we can make up some numbers for b_0 and b_1 and plot the function.

In the line of code below we create a function called `expfun` and set it equal to the function, $f(x) = \exp 2 + 5x$.

```
expfun <- function(x) exp(2 + 5*x)
```

We can see a plot of this function with a simple call to `plot()`:

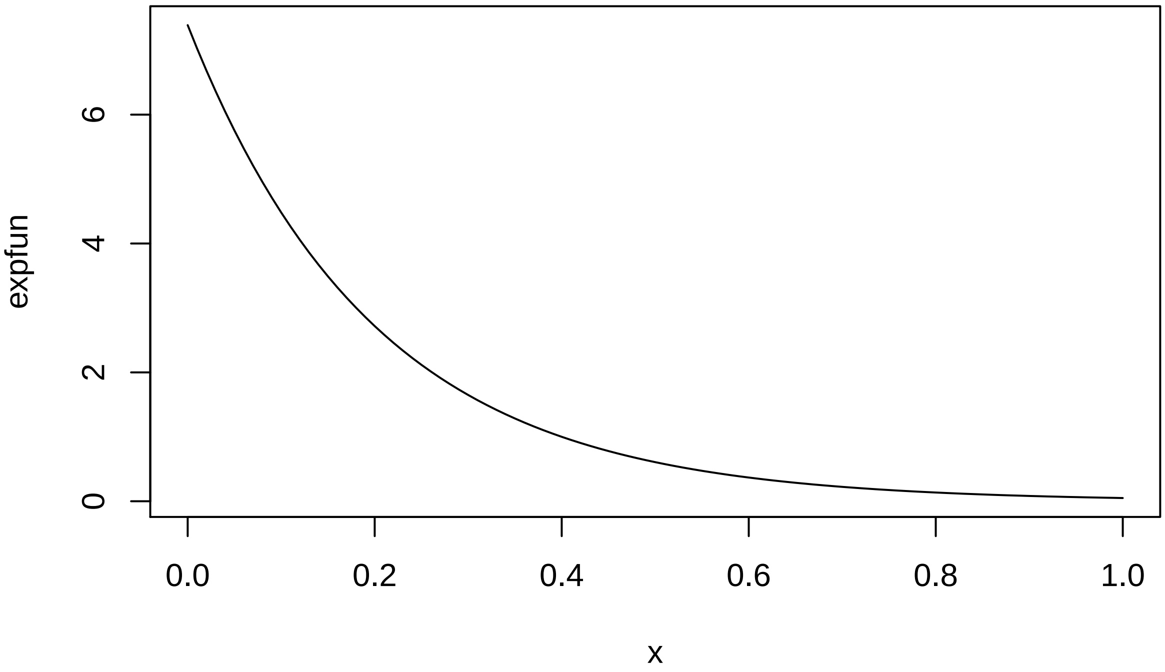
```
plot(expfun)
```



We can see that this kind of relationship implies that the outcome variable, y, increases and *an increasing rate* as x increases.

Let us look at what the function looks like if instead the coefficient for b_1 is negative. In the code below, we create the function `expfun`, but with a coefficient on x equal to -5 instead of $+5$, then plot it.

```
expfun <- function(x) exp(2 - 5*x)
plot(expfun)
```



Here we see this means that the outcome variable, y, decreases as x increases, and at a *decreasing* rate.

3. Regression with a log dependent variable

We estimate the regression equation with the log of monthly earnings as the outcome variable with the following call to `lm()` that assigns the output to an object that we call `loglmwages`:

```
loglmwages <- lm(log(wages$MonthlyEarnings) ~ wages$YearsEdu)
```

We can view the summary of the regression output with the following call to `summary()`:

```
summary(loglmwages)
```

```
##
## Call:
## lm(formula = log(wages$MonthlyEarnings) ~ wages$YearsEdu)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.94620 -0.24832  0.03507  0.27440  1.28106
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)    5.973062   0.081374   73.40   <2e-16 ***
## wages$YearsEdu 0.059839   0.005963   10.04   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4003 on 933 degrees of freedom
## Multiple R-squared:  0.09742,    Adjusted R-squared:  0.09645
## F-statistic: 100.7 on 1 and 933 DF,  p-value: < 2.2e-16
```

The coefficient years of education is equal to `0.0598`, which is how much the predicted value for $ln(monthly\ earnings)$ increases when educational attainment increases by one year. We can express this mathematically as,

$$b_1 = \frac{\Delta ln(\hat{y})}{\Delta x} = 0.0598$$

It turns out that this is a close approximation to the *percentage* increase in y from a one unit increase in x. That is,

$$\frac{\Delta ln(\hat{y})}{\Delta x} \approx \frac{\% \Delta \hat{y}}{\Delta x}.$$

Therefore, our regression predicts that a one additional year of education is associated with approximately a 6% higher monthly salary.

4. Log-Log Relationship

Let us instead consider the possibility for the following non-linear relationship between monthly earnings and educational attainment:

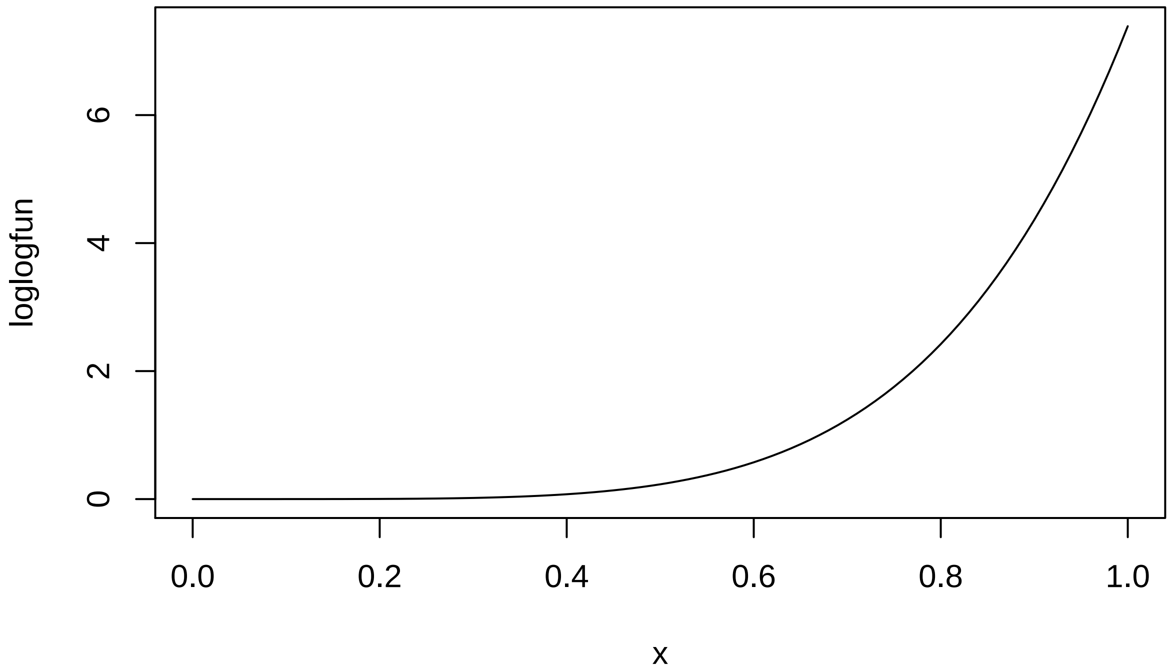
$$ln(y_i) = b_0 + b_1 ln(x_i) + \epsilon_i$$

Let us make up some numbers for b_0 and b_1 to visualize what such a function looks like. First let us solve for y_i by taking the exponential function of both sides of the equation. This yields the equivalent function:

$$y_i = \exp b_0 + b_1 ln(x_i) + \epsilon_i$$

In the code below, we make up a function with $b_0 = 2$ and $b_1 = 5$, call it `loglogfun` and plot the curve to see what it looks like:

```
loglogfun <- function(x) exp(2 + 5*log(x))
plot(loglogfun)
```



The function also predicts that y increases at an increasing rate with x, but an examination of the magnitude of the y axis labels reveals rate of increase in smaller.

We can estimate a log-log regression of monthly earnings on educational attainment with the following call to `lm()`, where the output is assigned to an object we call `lglglmwages`:

```
lglglmwages <- lm( log(wages$MonthlyEarnings) ~ log(wages$YearsEdu) )
```

We summarize the output with the following call to the `summary()` function:

```
summary(lglglmwages)
```

```
##
## Call:
## lm(formula = log(wages$MonthlyEarnings) ~ log(wages$YearsEdu))
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.94925 -0.24818  0.03866  0.27282  1.27167
##
```

```
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)          4.63932    0.21297   21.78   <2e-16 ***
## log(wages$YearsEdu)  0.82694    0.08215   10.07   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4002 on 933 degrees of freedom
## Multiple R-squared:  0.09796,    Adjusted R-squared:  0.09699
## F-statistic: 101.3 on 1 and 933 DF,  p-value: < 2.2e-16
```

The coefficient $b_1 = 0.8269$ is a measure of how much the *natural log* of earnings increases when the *natural log* of educational attainment increases by one unit, expressed mathematically as,

$$b_1 = \frac{\Delta ln(\hat{y})}{\Delta ln(x)} = 0.8269$$

It turns out that this is approximately equal to the predicted *percentage* increase in y when x increases by *one percent.* Mathematically,

$$b_1 = \frac{\Delta ln(\hat{y})}{\Delta ln(x)} = \frac{\%\Delta \hat{y}}{\%\Delta x} = 0.8269$$

That is, monthly earnings on average are 0.83% higher for each 1% increase in education attainment. In economics, we call a measure like this an **elasticity**.