Nonlinearity: Log-Transforming the Predictor

2017-04-11

Read in Data

In this set of notes, you will learn another method of dealing with nonlinearity. Specifically, we will look at transforming the predictor using a nonlinear transformation. The data we will use in this set of notes, *mnSchools.csv*, contains institutional data for several Minnesota colleges and universities collected in 2011. The variables are:

- `name`: College/university name
- `gradRate`: Six-year graduation rate, as a percentage
- `public`: Sector (1 = public college/university, 0 = private college/university)
- `sat`: Estimated median composite SAT score
- `tuition`: Amount of tuition and required fees covering a full academic year for a typical student, in U.S. dollars

These source of these data is: http://www.collegeresults.org. Using these data, we will examine if (and how) academic "quality" of the student-body (measured by SAT score) is related to institutiional graduation rates.

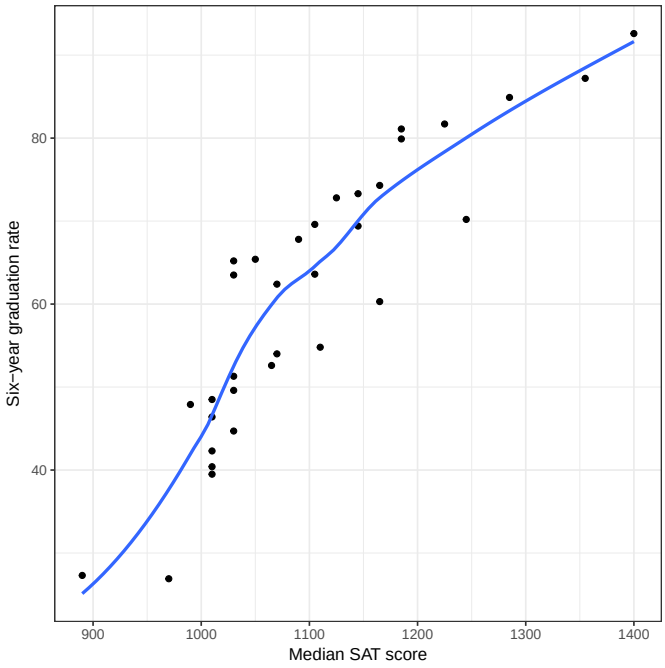
```
mn = read.csv(file = "~/Google Drive/Documents/epsy-8251/data/mnSchools.csv")
head(mn)
```

```
  id                              name gradRate public  sat tuition
1  1                  Augsburg College     65.2      0 1030   39294
2  3           Bethany Lutheran College     52.6      0 1065   30480
3  4 Bethel University, Saint Paul, MN     73.3      0 1145   39400
4  5                  Carleton College     92.6      0 1400   54265
5  6         College of Saint Benedict     81.1      0 1185   43198
6  7     Concordia College at Moorhead     69.4      0 1145   36590
```

```
# Load libraries
library(dplyr)
library(ggplot2)
library(sm)
```

Using the Natural Logarithm as a Transformation of the Predictor

Recall that the scatterplot of SAT scores and graduation rates suggested that the relationship between these variables was curvilinear.



To model this nonlinearity, we fitted a model that included a polynomial effect (quadratic). Another method of modeling nonlinearity is to transform the predictor (or outcome) using a nonlinear transformation. One commonly used nonlinear transformation is the logarithm. The logarithm is an inverse function of an exponent. The logarithm of a number (32 in our example) is the exponent to which the base (2 in our example) must be raised to produce that number. In other words,

$$\log_2(32) \longrightarrow 2^x = 32 \longrightarrow x = 5$$

Thus,

$$\log_2(32) = 5$$

To compute a logarithm using R, we use the `log()` function. We also specify the argument `base=`, since logarithms are unique to a particular base. For example, to compute the mathematical expression $\log_2(32)$, we use

```
log(32, base = 2)
```

```
[1] 5
```

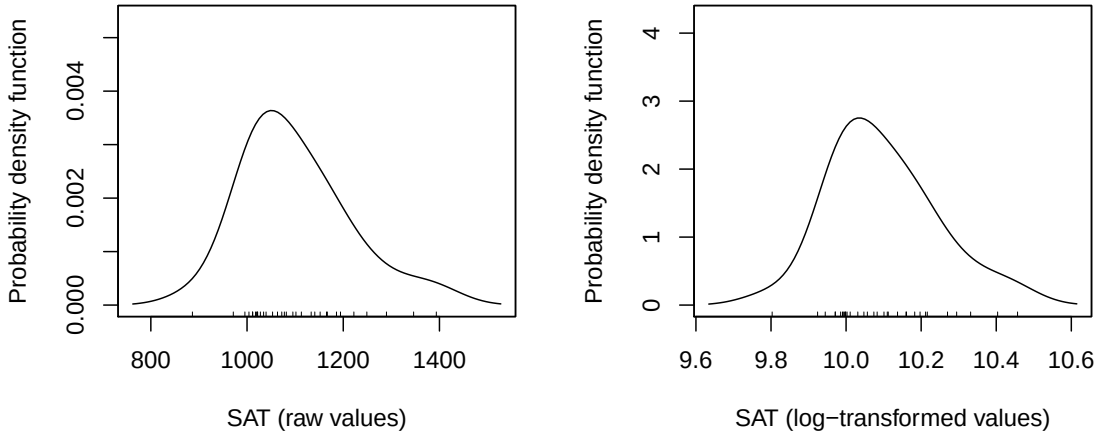
Log-Transforming Variables

For our purposes, we need to log-transform each value in a particular variable. Here, we will log-transform the SAT variable (using base-2).

```
mn = mn %>% mutate(L2sat = log(sat, base = 2))
head(mn)
```

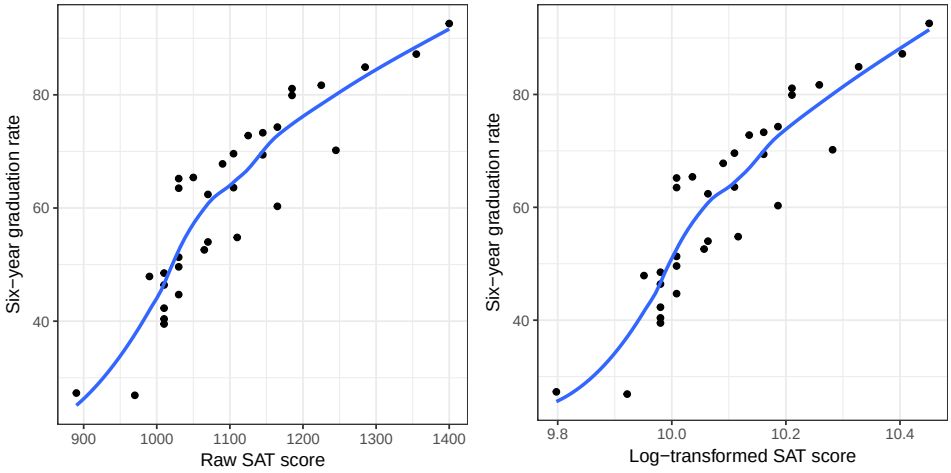
```
  id                              name gradRate public  sat tuition
1  1                  Augsburg College     65.2      0 1030   39294
2  3           Bethany Lutheran College     52.6      0 1065   30480
3  4 Bethel University, Saint Paul, MN     73.3      0 1145   39400
4  5                  Carleton College     92.6      0 1400   54265
5  6         College of Saint Benedict     81.1      0 1185   43198
6  7     Concordia College at Moorhead     69.4      0 1145   36590
     L2sat
1 10.00843
2 10.05664
3 10.16113
4 10.45121
5 10.21067
6 10.16113
```

How does this log-transformed variable compare to the original SAT predictor. We can examine the density plot of both the original and log-transformed variables to answer this.



- Comparing the shapes of the two variables, we see that the original variable was right-skewed. The log-transformed variable is also right-skewed, although it is LESS right-skewed than the original.
- The scale is quite different between the two variables (one is, after all, log-transformed). This has greatly affected the variation. After log-transforming, the variation is much smaller.

What happens when we use the log-transformed variable in a scatterplot with graduation rates?



The relationship between graduation rate and the log-transformed SAT scores is MORE linear than the relationship between graduation rates and the untransformed SAT scores. By transforming the variable using a nonlinear transformation (log) we have "linearized" the relationship with graduation rates. As such, we can fit a linear model to predict graduation rates using the Log-transformed SAT scores as a predictor.

Fitting the Regression Model

To fit the model, we use the `lm()` function and input the log-transformed SAT scores as the predictor.

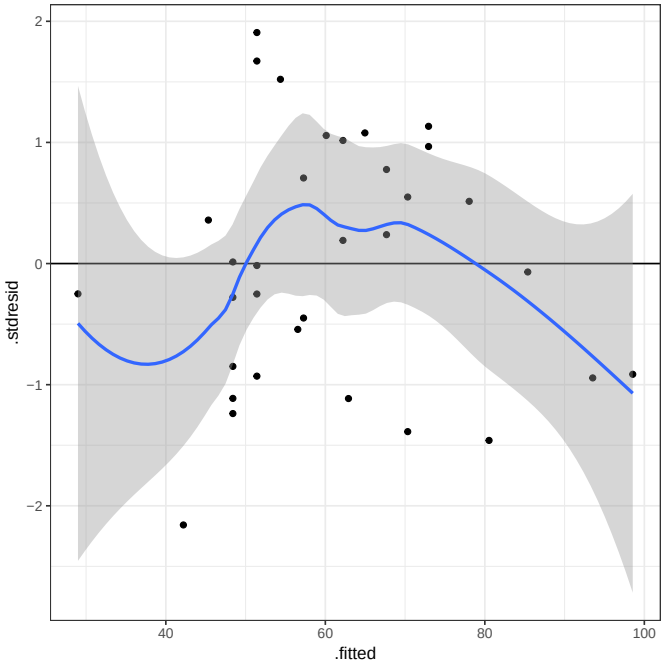
```
lm.1 = lm(gradRate ~ 1 + L2sat, data = mn)
```

Examine the Assumption of Linearity

Before examining the coefficients, we can scrutinize the residuals to see whether the log-transformation helped us meet the assumption of linearity.

```
# Obtain residuals
out = fortify(lm.1)

# Check linearity assumptions
ggplot(data = out, aes(x = .fitted, y = .stdresid)) +
  geom_point() +
  geom_hline(yintercept = 0) +
  geom_smooth() +
  theme_bw()
```



The assumption looks reasonably met as the horizontal line of $y = 0$ is encompassed in the confidence envelope of the loess smoother.

Interpret the Regression Results

We can now look at the `summary()` output and interpret the output.

```
summary(lm.1)
```

```
Call:
lm(formula = gradRate ~ 1 + L2sat, data = mn)

Residuals:
     Min       1Q   Median       3Q      Max
-15.3006  -6.1058  -0.1169   5.6295  13.7831

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -1013.872     93.098  -10.89 4.02e-12 ***
L2sat         106.439      9.219   11.55 9.30e-13 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 7.386 on 31 degrees of freedom
Multiple R-squared:  0.8113,    Adjusted R-squared:  0.8053
F-statistic: 133.3 on 1 and 31 DF,  p-value: 9.296e-13
```

Examining the model-level output, we see that differences in $\log_2(\text{SAT})$ explain 81.13% of the variation in graduation rates. This is statistically significant, $F(1,\ 31) = 133.3$, $p < .001$. Since differences in $\log_2(\text{SAT})$ imply that there are differences in the raw SAT scores, we would typically just say that "differences in SAT scores explain 81.13% of the variation in graduation rates."

Moving to the coefficient-level output, we can write the fitted equation as,

$$\hat{\text{Graduation}}\text{ Rate} = -1013.87 + 106.44\Big[\log_2(\text{SAT})\Big]$$

We can interpret the coefficients as we always do, recognizing that these interpretation are based on the log-transformed predictor.

- The intercept value of -1013.87 is the predicted average graduation rate for all colleges/universities with a $\log_2(\text{SAT})$ value of 0.
- The slope value of 106.44 indicates that each one-unit difference in $\log_2(\text{SAT})$ is associated with a 106.44-unit difference in graduation rate, on average.

Better Interpretations: Back-transforming

While these interpretations are technically correct, it is more helpful to your readers (and more conventional) to interpret any regression results in the metric of SAT scores rather than log-transformed SAT scores. This means we have to back-transform the interpretations. To back-transform a logarithm, we use its inverse function; exponentiation.

We interpreted the intercept as, "the predicted average graduation rate for all colleges/universities with a $\log_2(\text{SAT})$ value of 0". To interpret this using the SAT metric, we have to understand what $\log_2(\text{SAT}) = 0$ is.

$$\log_2(\text{SAT}) = 0 \longrightarrow 2^0 = \text{SAT}$$

In this computation, SAT = 1. Thus, rather than using the log-transformed interpretation, we can, instead, interpret the intercept as,

- The predicted average graduation rate for all colleges/universities with a median SAT score of 1 is -1013.87.

Since there are no colleges/universities in our data that have a SAT score of 1, this is extrapolation.

What about the slope? Our interpretation was that "each one-unit difference in $\log_2(\text{SAT})$ is associated with a 106.44-unit difference in graduation rate, on average." Working with the same ideas of back-transformation, we need to understand what a one-unit difference in $\log_2(\text{SAT})$ means. Consider four values of $\log_2(\text{SAT})$ that are each one-unit apart:

$$\log_2(\text{SAT}) = 1\,\log_2(\text{SAT}) = 2\,\log_2(\text{SAT}) = 3\,\log_2(\text{SAT}) = 4$$

If we back-transform each of these, then we can see how the four values of the raw SAT variable would differ.

$$\text{SAT} = 2^1 = 2\text{SAT} = 2^2 = 4\text{SAT} = 2^3 = 8\text{SAT} = 2^4 = 16$$

When $\log_2(\text{SAT})$ is increased by one-unit, the raw SAT scores is doubled. We can use this in our interpretation of slope:

- A doubling of median SAT score is associated with a 106.44-unit difference in graduation rate, on average.

The technical language for doubling is a "two-fold difference". So we would conventionally interpret this as:

- Each two-fold difference in median SAT score is associated with a 106.44-unit difference in graduation rate, on average.

To understand this further, consider a specific school, say Augsburg. Their median SAT score is 1030, and their log-transformed SAT score is 10.00843. Using the fitted regression equation (which employs the log-transformed SAT),

```
-1013.872 + 106.439 * 10.00843
```

```
[1] 51.41528
```

Augsburg's predicted graduation rate would be 51.4. If we increase the L2sat score by 1 to 11.00843 (which is equivalent to a raw SAT score of 2060; double 1030), their predicted graduation rate is,

```
-1013.872 + 106.439 * 11.00843
```

```
[1] 157.8543
```

This is an increase of 106.439.

Alternative Method of Fitting the Model

Rather that create the log-transformed SAT score in the data, we can use the `log()` function on SAT directly in the `lm()` computation.

```
lm.1 = lm(gradRate ~ 1 + log(sat, base = 2), data = mn)
summary(lm.1)
```

```

Call:
lm(formula = gradRate ~ 1 + log(sat, base = 2), data = mn)

Residuals:
     Min       1Q   Median       3Q      Max 
-15.3006  -6.1058  -0.1169   5.6295  13.7831 

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)    
(Intercept)        -1013.872     93.098  -10.89 4.02e-12 ***
log(sat, base = 2)   106.439      9.219   11.55 9.30e-13 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 7.386 on 31 degrees of freedom
Multiple R-squared:  0.8113,    Adjusted R-squared:  0.8053 
F-statistic: 133.3 on 1 and 31 DF,  p-value: 9.296e-13
```

Using this method of fitting the model will be useful as we plot the fitted model.

Plotting the Fitted Model

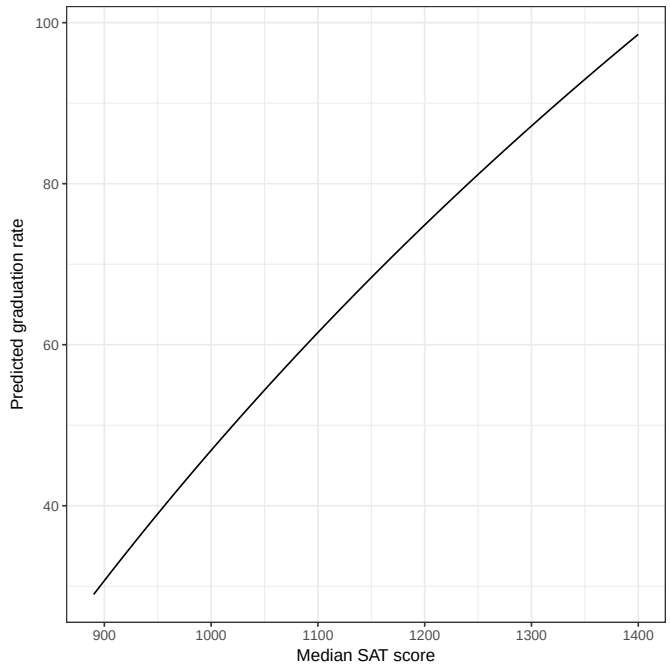
To aid interpretation of the effect of SAT on graduation rate, we can plot the fitted model. If we used the method of fitting in which we used `log()` directly in the `lm()` function, we only need to set up a sequance of SAT values, predict graduation rates using the fitted model, and finally connect these values using a line.

```
# Set up data
plotData = expand.grid(
    sat = seq(from = 890, to = 1400, by = 10)
    )

# Predict
plotData$yhat = predict(lm.1, newdata = plotData)

# Examine data
# head(plotData)

# Plot
ggplot(data = plotData, aes(x = sat, y = yhat)) +
    geom_line() +
    theme_bw() +
  xlab("Median SAT score") +
  ylab("Predicted graduation rate")
```



The plot shows the slight curvilinearity in the effect of SAT on graduation rates.

Different Base Values in the Logarithm

The base value we used in the `log()` function was base-2. Using a base value of 2 was an arbitrary choice. We can use any base value we want. For example, what happens if we use base-10.

```
mn$L10sat = log(mn$sat, base = 10)
head(mn)
```

```
  id                              name gradRate public  sat tuition
1  1                  Augsburg College     65.2      0 1030   39294
2  3           Bethany Lutheran College     52.6      0 1065   30480
3  4 Bethel University, Saint Paul, MN     73.3      0 1145   39400
4  5                  Carleton College     92.6      0 1400   54265
5  6         College of Saint Benedict     81.1      0 1185   43198
6  7     Concordia College at Moorhead     69.4      0 1145   36590
     L2sat   L10sat
1 10.00843 3.012837
2 10.05664 3.027350
3 10.16113 3.058805
4 10.45121 3.146128
5 10.21067 3.073718
6 10.16113 3.058805
```

Comparing the logarithms of SAT using base-10 to those using base-2 we see that the base-10 logarithms are smaller. This is because now we are using the base of 10 in our exponent (rather than 2). For example, for Augsburg,

$$10^{3.012837} = 1030$$

If we fit a model using the base-10 lagarithm,

```
lm.2 = lm(gradRate ~ 1 + log(sat, base = 10), data = mn)
summary(lm.2)
```

```

Call:
lm(formula = gradRate ~ 1 + log(sat, base = 10), data = mn)

Residuals:
     Min       1Q   Median       3Q      Max
-15.3006  -6.1058  -0.1169   5.6295  13.7831

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)         -1013.87      93.10  -10.89 4.02e-12 ***
log(sat, base = 10)   353.58      30.62   11.55 9.30e-13 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 7.386 on 31 degrees of freedom
Multiple R-squared:  0.8113,    Adjusted R-squared:  0.8053
F-statistic: 133.3 on 1 and 31 DF,  p-value: 9.296e-13
```

Examining the model-level output, we see that differences in $\log_{10}(\text{SAT})$ explain 81.13% of the variation in graduation rates. Or simply, that differences in SAT scores explain 81.13% of the variation in graduation rates. This is statistically significant, $F(1,\ 31) = 133.3$, $p < .001$. These model-level results are the same as when we used the base-2 logarithm.

The fitted equation is,

$$\hat{\text{Graduation}}\ \text{Rate} = -1013.87 + 353.58\Big[\log_{10}(\text{SAT})\Big]$$

We can interpret the coefficients using the base-10 logarithm of SAT scores as:

- The intercept value of -1013.87 is the predicted average graduation rate for all colleges/universities with a $\log_{10}(\text{SAT})$ value of 0.
- The slope value of 353.58 indicates that each one-unit difference in $\log_{10}(\text{SAT})$ is associated with a 353.58-unit difference in graduation rate, on average.

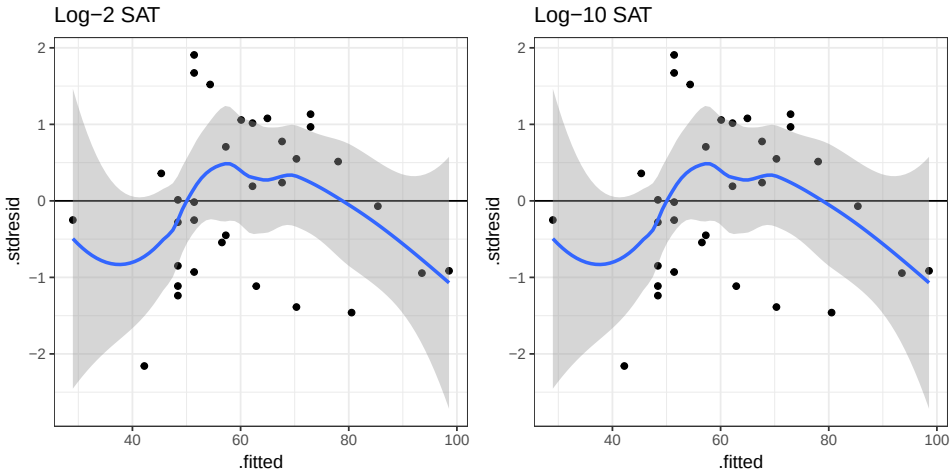
Better yet, we can back-transform the interpretations so that we are using SAT scores rather than $\log_{10}(\text{SAT})$ scores.

- The predicted average graduation rate for all colleges/universities with a median SAT score of 1 is -1013.87.
- Each *ten-fold* difference in median SAT score is associated with a 353.58-unit difference in graduation rate, on average.

To further think about the effect of SAT, if Augsburg improved its median SAT score ten-fold (i.e., going from a median SAT score of 1030 to a median SAT score of 10,300) we would predict its graduation rate to go up by 353.58.

The model-level information is all the same. Furthermore, the intercepts (and SE and p-value) was the same across both models. The slope coefficients and SEs were different in the two models, but the t-value and p-value for the effect of SAT was identical for both base-2 and base-10. The only real difference in using base-10 vs. base-2 in the logarithm is in the interpretation of the SAT effect.

What if we look at the residual fit?



The residuals fit EXACTLY the same. Why is this? Let's again use Augsburg as an example. Using the fitted model that employed the base-2 logarithm, we found that Augsburg's predicted graduation rate was,

$$\begin{aligned}\widehat{\text{Graduation Rate}} &= -1013.87 + 106.44\Big[\log_2(1030)\Big] \\ &= -1013.87 + 106.44\Big[10.00843\Big] \\ &= 51.42\end{aligned}$$

Using the model that employed the base-10 lpogarithm, augsburg's predicted graduation rate would be

$$\begin{aligned}\widehat{\text{Graduation Rate}} &= -1013.87 + 353.58\Big[\log_{10}(1030)\Big] \\ &= -1013.87 + 353.58\Big[3.012837\Big] \\ &= 51.42\end{aligned}$$

Augsburg's predicted graduation rate is exactly the same in the two models. This implies that Augsburg's residual would also be the same in the two models. This is true for every college. Because of this, increasing (or decreasing) the base used in the logarithm does not help improve the fit of the model. The fit is exactly the same no matter which base you choose. The only thing that changes when you choose a different base is the interpretation of the slope. You should choose the base to facilitate interpretation. For example, does it make more sense to talk about a *two-fold* difference in the predictor? A *five-fold* difference in the predictor? A *ten-fold* difference in the predictor?

Base-e Logarithm: The Natural Logarithm

In our example, neither of the bases we examined is satisfactory in terms of talking about the effect of SAT. Two-fold differences in SAT are very unlikely, to say anything of ten-fold differences. One base that is commonly used for log-transformations is base-e. e is a mathematical constant (Euler's number) that is approximately equal to 2.71828. We can obtain this by using the `exp()` function in R. This function takes e to some exponanent that is given as the argument. So to obtain the approximation of e we use

```
exp(1)
```

```
[1] 2.718282
```

The logarithm (base-e) for a number, referred to as the *natural logarithm*, can be obtained using the `log()` function with the argument `base=exp(1)`. However, this base is so commonly used that it is the default value for the `base=` argument. So, if we use the `log()` function without defining the `base=` argument, it will automatically use base-e. For example, the natural logarithm of Augsburg's SAT score of 1030 can be computed as

```
log(1030)
```

```
[1] 6.937314
```

If we took $e^{6.937314}$ we would obtain 1030. The natural logarithm even has its own mathematical notation; ln. For example, we would mathematically express the natural logarithm of 1030 as

$$\ln(1030) = 6.937314.$$

Using the Natural Logarithm in a Regression Model

Below we regress graduation rates on the log-transformed SAT scores, using the natural logarithm.

```
lm.3 = lm(gradRate ~ 1 + log(sat), data = mn)
summary(lm.3)
```

```
Call:
lm(formula = gradRate ~ 1 + log(sat), data = mn)

Residuals:
     Min       1Q   Median       3Q      Max
-15.3006  -6.1058  -0.1169   5.6295  13.7831

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -1013.9       93.1  -10.89 4.02e-12 ***
log(sat)       153.6       13.3   11.55 9.30e-13 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 7.386 on 31 degrees of freedom
Multiple R-squared:  0.8113,    Adjusted R-squared:  0.8053
F-statistic: 133.3 on 1 and 31 DF,  p-value: 9.296e-13
```

As with any base, using base-e results in the same model-level information ($R^2 = .811$, $F(1,\ 31) = 133.3$, $p < .001$). The intercept has the same coefficient ($\hat{\beta}_0 = -1013.9$), SE, t-value, and p-value as the intercept from the models using base-2 and base-10 log-transformations of SAT. (This is, again, because $2^0 = 10^0 = e^0 = 1$.) And, although the coefficient and SE for the effect of SAT is again different (a one-unit change in the three different log-scales doess not correspond to the same amount of change in raw SAT for the three models), the t-value and level of statistical significance ($t(31) = 11.55$, $p < .001$) for this effect, are the same as when we used base-2 and base-10.

So how can we interpret the model's coefficients? The intercept can be interpreted exactly the same as in the previous models in which we used base-2 or base-10; namely that the predicted average graduation rate for colleges/univeristies with a median SAT score of one is -1013.9. Interpreting the slope, we could say that an e-fold difference in median SAT score is associated with a 153.6-unit difference in graduation rates, on average. This, although correct, is again, unsatisfactory. If a two-fold difference is unlikely, a 2.71-fold difference is also unlikely.

Interpretation Using Percentage Change

Consider three schools, each having a SAT score that differs by 1%; say these schools have median SAT scores of 1000, 1010, 1020.1. Using the fitted equation, we can compute the predicted graduation rate for each of these hypothetical schools:

$$\widehat{\text{Graduation Rate}} = -1013.9 + 153.6\Big[\ln(\text{SAT})\Big]$$

The SAT scores and predicted graduation rates for these schools are given below:

```
     sat gradRate
1 1000.0 46.87784
2 1010.0 48.40581
3 1020.1 49.93378
```

The difference between each subsequent predicted graduation rate is 1.52797.

```
48.40581 - 46.87784
```

```
[1] 1.52797
```

```
49.93378 - 48.40581
```

```
[1] 1.52797
```

In other words, for schools that have a median SAT score that differs by 1%, their predicted graduation rate differs by 1.52797, on average.

Mathematical Explanation

To understand how we can directly compute this of this difference, consider the predicted values for two x-values that differ by one-percent, if we use symbolic notation:

$$\hat{y}_1 = \hat{\beta}_0 + \hat{\beta}_1 \left[\ln(x)\right]$$
$$\hat{y}_2 = \hat{\beta}_0 + \hat{\beta}_1 \left[\ln(1.01x)\right]$$

The difference in their predicted values is:

$$\begin{aligned}
\hat{y}_2 - \hat{y}_1 &= \hat{\beta}_0 + \hat{\beta}_1 \left[\ln(1.01x)\right] - \left(\hat{\beta}_0 + \hat{\beta}_1 \left[\ln(x)\right]\right) \\
&= \hat{\beta}_0 + \hat{\beta}_1 \left[\ln(1.01x)\right] - \hat{\beta}_0 - \hat{\beta}_1 \left[\ln(x)\right] \\
&= \hat{\beta}_1 \left[\ln(1.01x)\right] - \hat{\beta}_1 \left[\ln(x)\right] \\
&= \hat{\beta}_1 \left[\ln(1.01x) - \ln(x)\right] \\
&= \hat{\beta}_1 \left[\ln(\frac{1.01x}{1x})\right]
\end{aligned}$$

If we substitute in any value for x, we can now directly compute this constant difference. Note that a convenient value for x is 1. Then this reduces to:

$$\hat{\beta}_1 \left[\ln(1.01)\right]$$

So now, we can interpret this as: a one-percent difference in x is associated with a $\hat{\beta}_1 \left[\ln(1.01)\right]$-unit difference in Y, on average.

In our model, we can compute this difference using the fitted coefficient $\hat{\beta}_1 = 153.6$ as

$$153.6 \left[\ln(1.01)\right] = 1.528371$$

The same computation using R is

```
153.6 * log(1.01)
```

```
[1] 1.528371
```

This gives you the constant difference exactly. So you can interpret the effect of SAT as, each 1% difference in SAT score is associated with a difference in graduation rates of 1.53, on average.

Approximate Interpretaton

We can get an approximate estimate for this value by using the mathematical shortcut of

$$\frac{\hat{\beta}_1}{100}$$

Using our fitted results, we could approximate the effect as,

$$\frac{153.6}{100} = 1.536$$

We could then interpret the effect of SAT by saying a 1% difference in median SAT score is associated with a 1.53-unit difference in predicted graduation rate, on average.

Including Covariates

We can also include covariates in the model. Below we examine the nonlinear effect of SAT on graduation controlling for differences in sector.

```
lm.4 = lm(gradRate ~ 1 + public + log(sat), data = mn)
summary(lm.4)
```

```
Call:
lm(formula = gradRate ~ 1 + public + log(sat), data = mn)

Residuals:
     Min       1Q   Median       3Q      Max
-18.7978  -3.4414  -0.5165   4.8757  10.7387

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -958.496     81.457 -11.767 9.10e-13 ***
public        -8.501      2.444  -3.479  0.00156 **
log(sat)     146.016     11.616  12.570 1.74e-13 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.338 on 30 degrees of freedom
Multiple R-squared:  0.8656,    Adjusted R-squared:  0.8566
F-statistic: 96.58 on 2 and 30 DF,  p-value: 8.466e-14
```

The model explains 86.5% of the variation in graduation rates, $F(2,\ 30) = 96.58$, $p < .001$. Interpreting each of the coefficients using the raw SAT scores:

- The intercept value of -1013.87 is the predicted average graduation rate for all public colleges/universities with a median SAT score of 1 (extrapolation).
- There is a statistically significant effect of sector after controlling for differences in median SAT score ($p = .002$). Public schools have a predicted graduation rate that is 8.5-units lower, on average, than private schools controlling for differences in median SAT scores.
- There is a statistically significant effect of SAT after controlling for differences in sector ($p < .001$). A 1% difference in median SAT score is associated with a 1.46-unit difference in predicted graduation rate, on average, after controlling for differences in sector.

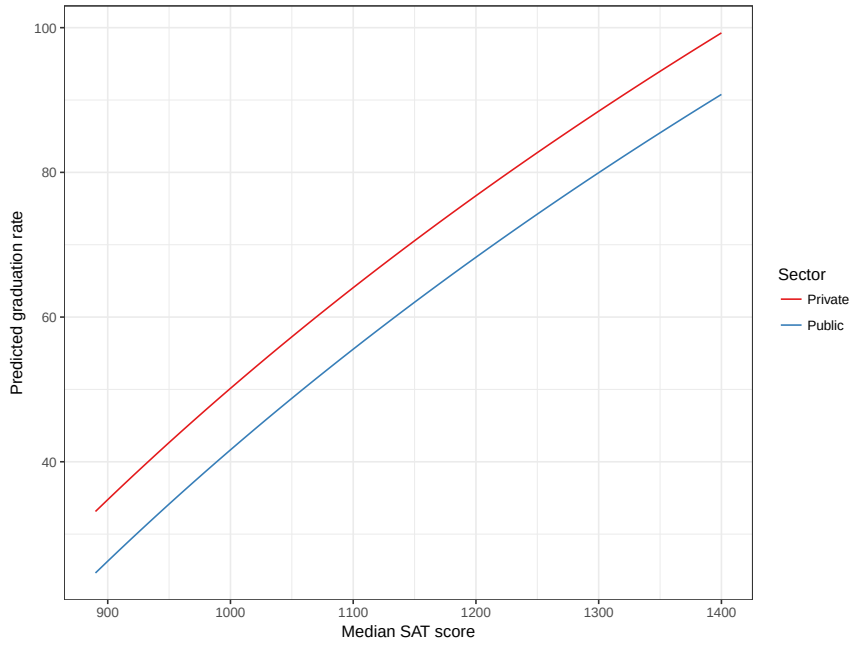
To further help interpret these effects, we can plot the fitted model.

```
# Set up data
plotData = expand.grid(
  sat = seq(from = 890, to = 1400, by = 10),
  public = c(0, 1)
)

# Predict
plotData$yhat = predict(lm.4, newdata = plotData)

# Examine data
# head(plotData)

# Plot
ggplot(data = plotData, aes(x = sat, y = yhat, color = factor(public))) +
  geom_line() +
  theme_bw() +
  xlab("Median SAT score") +
  ylab("Predicted graduation rate") +
  scale_color_brewer(name = "Sector", palette = "Set1", labels = c("Private", "Public"))
```

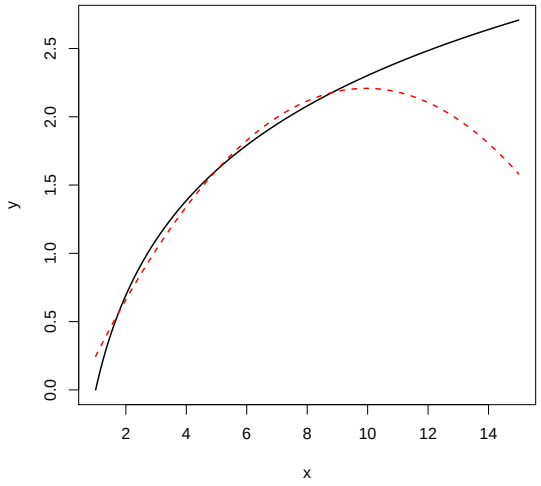


The plot shows the nonlinear, positive effect of SAT on graduation rate for both public and private schools. For schools with lower median SAT scores, there is a larger effect on graduation rates than for schools with higher median SAT scores (for both private and public schools). The plot also shows the controlled effect of sector. For schools with the same median SAT score, private schools have a higher predicted graduation rate than public schools, on average.

Polynomial Effects vs. Log-Transformations

The inclusion of polynomial effects and the use of a log-transformation was to model the nonlinearity observed in the relationship between SAT scores and graduation rates. Both methods were successful in this endeavor. While either method could be used in practice to model nonlinearity, there are some considerations when making the choice of which may be more appropriate for a given modeling situation.

The first consideration is one of theory. The plot below shows the mathematical function for a log-transformed X-value (solid, black line) and for a quadratic polynomial of X (dashed, red line).



Both functions are nonlinear, however then polynomial function changes direction. For low values of X, the function has a large positive effect. This effect diminishes as X gets bigger, and around $X = 9$ the effect is zero. For larger values of X, the effect is actually negative. For the logarithmic function, the effect is always positive, but it diminishes as X gets larger . Theoretically, these are very different ideas, and if substantive literature suggests one or the other, you should probably acknowledge that in the underlying statistical model that is fitted.

Empirically, the two functions are very similar especially within certain ranges of X. For example, although the predictions from these models would be quite different for really high values of X, if we only had data from the range of 2 to 8 ($2 \leq X \leq 8$) both functions would produce similar residuals. It might then be prudent to think about Occam's Razor—if two competing models produce similar predictions, adopt the simpler model. Between these two functions, the log-transformed model is simpler; it has one fewer predictor. The mathematical models make this clear:

$$\textbf{Polynomial}: \; Y_i = \beta_0 + \beta_1(X_i) + \beta_2(X_i^2) + \epsilon_i$$

$$\textbf{Log-Transform}: \; Y_i = \beta_0 + \beta_1\Big[\ln(X_i)\Big] + \epsilon_i$$

The quadratic polynomial model has two effects: a linear effect of X and a quadratic effect of X (remember it is an interaction model), while the model using the log-transformed predictor only has a single effect. If there is no theory to guide your model's functional form, and the residuals from the polynomial and log-transformed models seem to fit equally well, then the log-transformed model saves you a degree of freedom, and probably should be adopted.