Log-Transforming the Predictor

2020-02-06

Preparation

In this set of notes, you will learn another method of dealing with nonlinearity. Specifically, we will look at log-transforming the predictor in a linear model. To do so, we will use the *mn-schools.csv* dataset (see the data codebook) to examine if (and how) academic "quality" of the student-body (measured by SAT score) is related to institutiional graduation rate.

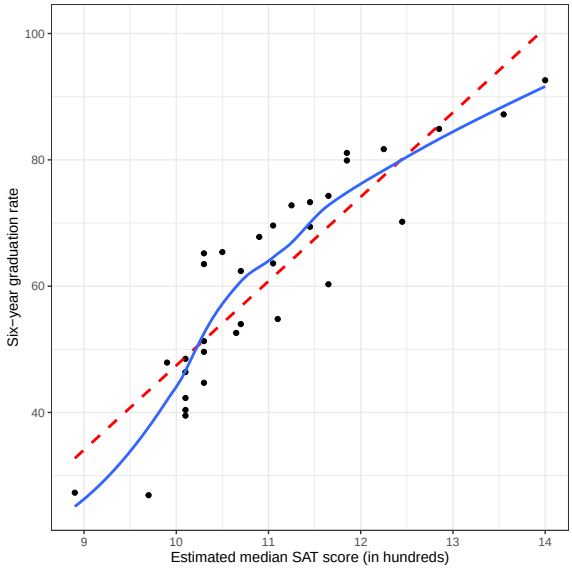
```
# Load libraries
library(broom)
library(corrr)
library(educate) #Need version 0.1.0.1
library(patchwork)
library(tidyverse)

# Read in data
mn = read_csv(file = "~/Documents/github/epsy-8252/data/mn-schools.csv")
```

Relationship between Graduation Rate and SAT Scores

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

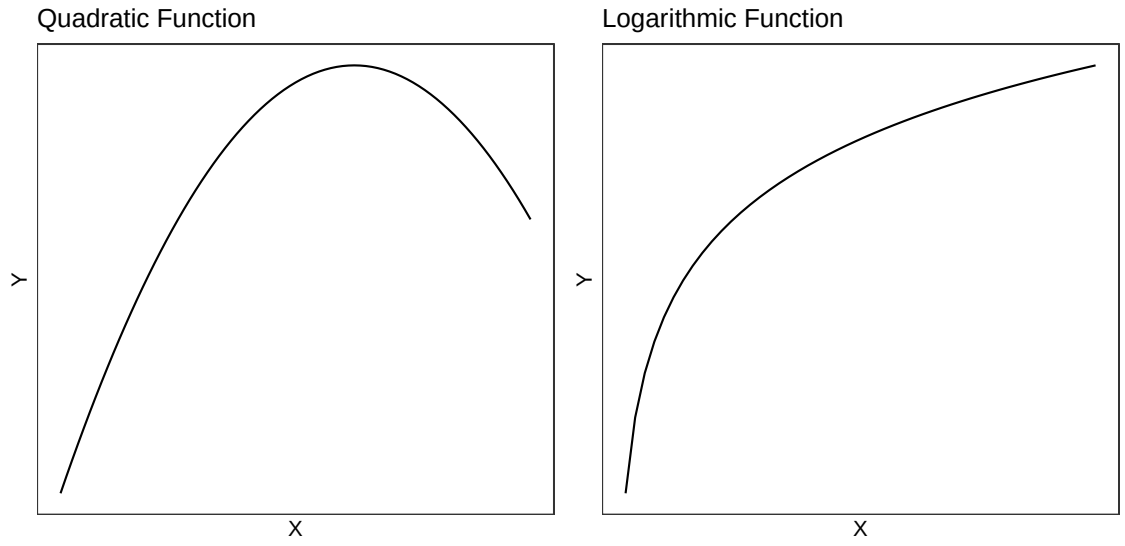
Figure 1
Six-Year Graduation Rate as a Function of Median SAT Score



Note. The loess smoother (blue, dashed line) and fitted regression line (red, dashed line) are also displayed.

One way to model this nonlinearity was to fit 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. Below is a comparison of the quadratic function to the logarithmic function.

Figure 2
Quadratic and Logarithmic Functions



The quadratic function shows continuous and diminishing growth followed by continuous and increasing loss (parabola; the function changes direction), while the logarithmic function models continuous, albeit diminishing, growth (the function does not change direction).

Quick Refresher on Logarithms

The logarithm is an inverse function of an exponent. Consider this example,

$$\log_2(32)$$

The logarithm of 32 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
```

There is also a shortcut function to use base-2.

```
log2(32)
```

```
[1] 5
```

Log-Transforming Variables

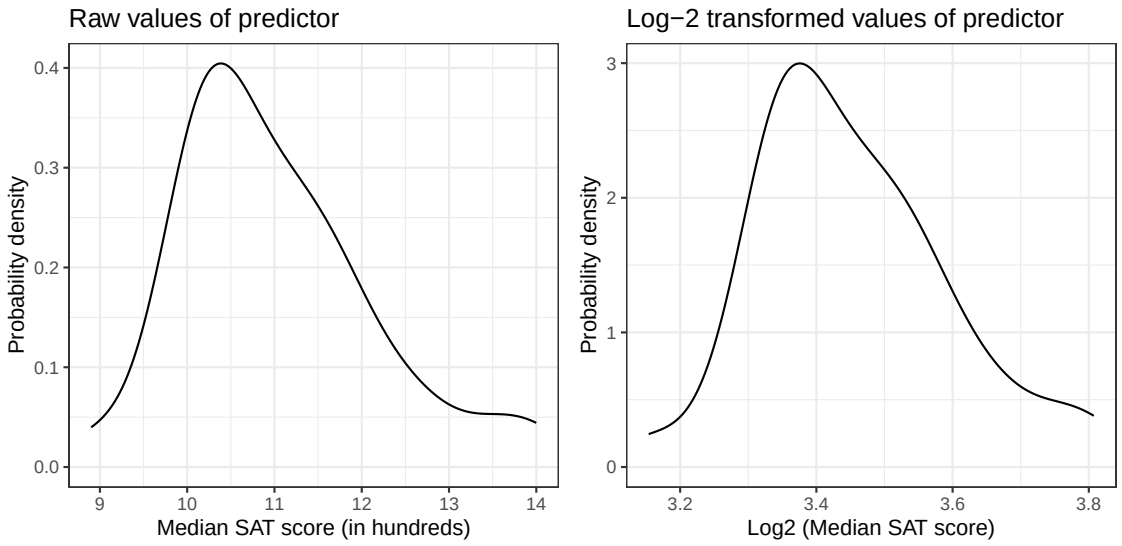
For our purposes, we need to log-transform each value in a particular variable. Here, we will log-transform the SAT predictor (using base-2).

```
# Create log2 transformed SAT scores
mn = mn %>%
  mutate(
    L2sat = log(sat, base = 2)
    )

# View data
head(mn)
```

```
# A tibble: 6 x 7
     id name                              grad public   sat tuition L2sat
  <dbl> <chr>                            <dbl>  <dbl> <dbl>   <dbl> <dbl>
1     1 Augsburg College                  65.2      0  10.3    39.3  3.36
2     3 Bethany Lutheran College          52.6      0  10.6    30.5  3.41
3     4 Bethel University, Saint Paul, MN 73.3      0  11.4    39.4  3.52
4     5 Carleton College                  92.6      0  14      54.3  3.81
5     6 College of Saint Benedict         81.1      0  11.8    43.2  3.57
6     7 Concordia College at Moorhead     69.4      0  11.4    36.6  3.52
```

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



- Comparing the shapes of the two distributions, we see that the original median SAT 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 center (mean) and the variation. After log-transforming, the center and variation are both much smaller.

Logarithmic transformations change the shape, center, and variation of a distribution!

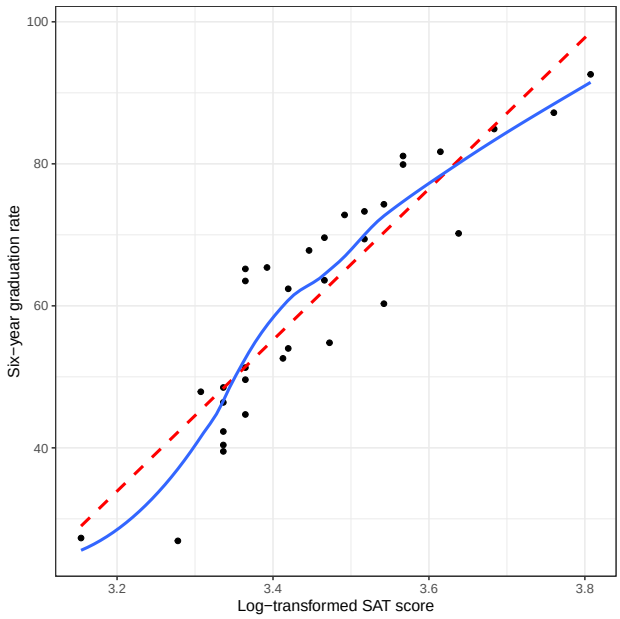
Relationship between Graduation Rate and Log-Transformed SAT Scores

What happens when we examine the relationship between graduation rates and the log-transformed SAT scores?

```
ggplot(data = mn, aes(x = L2sat, y = grad)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE, color = "red", linetype = "dashed") +
  geom_smooth(se = FALSE) +
  theme_bw() +
  xlab("Log-transformed SAT score") +
  ylab("Six-year graduation rate")
```

Figure 3

Scatterplot of Six-Year Graduation Rate versus Log-Transformed Median SAT Score (Base-2)



Note. The loess smoother (blue, solid line) and fitted regression line (red, dashed line) are also displayed.

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. (The loess smoother and fitted regression line are more closely aligned in this plot than in the plot of the untransformed data.) 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

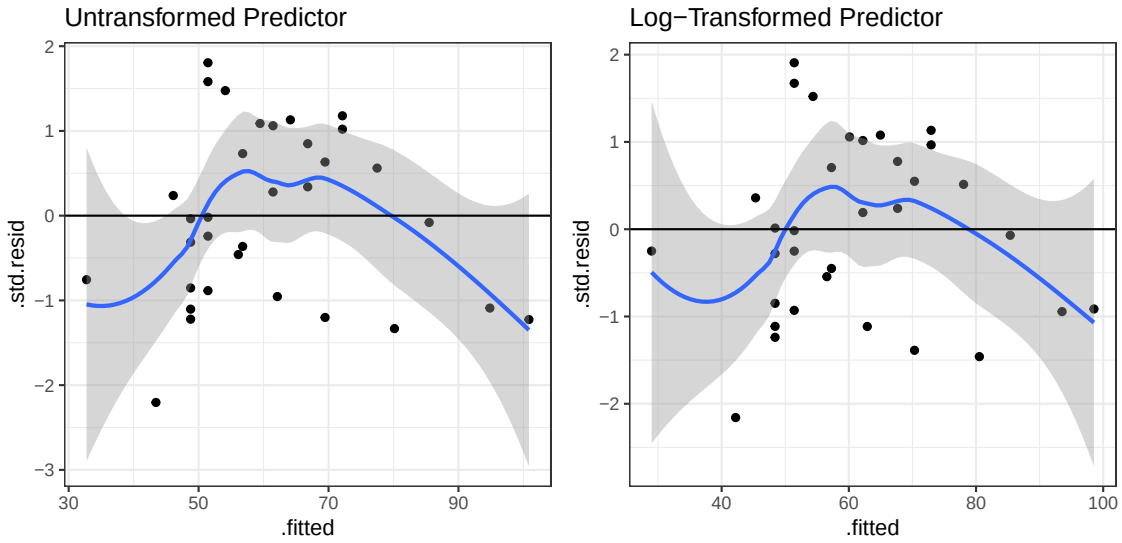
To fit the model, we use the `lm()` function and input the log-transformed SAT scores as the predictor.

```
# Fit regression model
lm.1 = lm(grad ~ 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. We will examine the standardized residuals versus the fitted values for models with both the untransformed and transformed predictor.

Figure 4
Residual Plots for Two Competing Models



Note. Standardized residuals versus the fitted values for models with both the untransformed (left) and log-transformed (right; base-2) predictor. In both plots the loess smoother (solid, blue line) is displayed along with the 95% confidence envelope (grey shaded area). The reference line of Y=0 is also displayed.

While the smoother shows the same overall pattern, the line $Y = 0$ is completely encompassed in the confidence envelope for the model using the log-transformed predictor. In the case of the model with the untransformed predictor, the $Y = 0$ reference line is not completely encompassed in the confidence envelope. Because of this, we are more satisfied with the tenability of the assumption that the average residual is 0 at each fitted value for the log-transformed model than with the untransformed model.

When we are trying to decide whether to use a transformed predictor/outcome in our model, it is examination of the residuals that will help us make this decision, not a statistical test.

Interpret the Regression Results

We can now look at the regression output and interpret the results.

```
# Model-level output
glance(lm.1)
```

```
# A tibble: 1 x 11
  r.squared adj.r.squared sigma statistic  p.value    df logLik   AIC   BIC
      <dbl>         <dbl> <dbl>     <dbl>    <dbl> <int>  <dbl> <dbl> <dbl>
1     0.811         0.805  7.39      133. 9.30e-13     2  -112.  230.  234.
# ... with 2 more variables: deviance <dbl>, df.residual <int>
```

Examining the model-level output, we see that differences in $\log_2(\text{SAT})$ explain 81.13% of the variation in graduation rates. Based on the inferential results, $F(1,\ 31) = 133.3$, $p < .001$, the model seems to explain variation in the outcome. 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,

```
# Coefficient-level output
tidy(lm.1)
```

```
# A tibble: 2 x 5
  term        estimate std.error statistic  p.value
  <chr>          <dbl>     <dbl>     <dbl>    <dbl>
1 (Intercept)    -307.     31.9      -9.62 7.94e-11
2 L2sat           106.      9.22     11.5  9.30e-13
```

We can write the fitted equation as,

$$\widehat{\text{Graduation Rate}}_i = -306.7 + 106.4\Big[\log_2(\text{SAT}_i)\Big]$$

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

- The intercept value of -306.7 is the predicted average graduation rate for all colleges/universities with a $\log_2(\text{SAT})$ value of 0.
- The slope value of 106.4 indicates that each one-unit difference in $\log_2(\text{SAT})$ is associated with a 106.4-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 raw metric of the variable rather than log-transformed metric. 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 raw metric of our SAT attribute, we have to understand what $\log_2(\text{SAT}) = 0$ is.

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

In this computation, $\text{SAT} = 1$. 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 value of 1 (which since this is measures in hundreds corresponds to a median SAT of 100) is -306.7. Since there are no colleges/universities in our data that have a median SAT value 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 106.4-unit difference in graduation rate, on average." Working with the same idea 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 value is doubled. We can use this in our interpretation of slope:

> A doubling of the SAT value is associated with a 106.4-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 SAT value is associated with a 106.4-unit difference in graduation rate, on average.

To understand this further, consider a specific school, say Augsburg. Their measurement on the raw SAT variable is 10.3, and their log-transformed SAT score is 3.36. Using the fitted regression equation (which employs the log-transformed SAT),

```
-306.7 + 106.4 * 3.36
```

```
[1] 50.804
```

Augsburg's predicted graduation rate would be 50.8. If we increase the `L2sat` score by 1 to 4.36 (which is equivalent to a raw SAT measurement of 20.6; double 10.3), their predicted graduation rate is,

```
-306.7 + 106.4 * 4.36
```

```
[1] 157.204
```

This is an increase of 106.4.

Alternative Method of Fitting the Model

Rather that create the log-transformed SAT score as a new column in the data and then using this column in the model, we can use the `log()` function directly on the SAT predictor in the `lm()` computation.

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

# Model-level output
glance(lm.1)
```

```
# A tibble: 1 x 11
  r.squared adj.r.squared sigma statistic  p.value    df logLik   AIC   BIC
      <dbl>         <dbl> <dbl>     <dbl>    <dbl> <int>  <dbl> <dbl> <dbl>
1     0.811         0.805  7.39      133. 9.30e-13     2  -112.  230.  234.
# ... with 2 more variables: deviance <dbl>, df.residual <int>
```

```
# Coefficient-level output
tidy(lm.1)
```

```
# A tibble: 2 x 5
  term               estimate std.error statistic  p.value
  <chr>                 <dbl>     <dbl>     <dbl>    <dbl>
1 (Intercept)           -307.     31.9      -9.62 7.94e-11
2 log(sat, base = 2)     106.      9.22     11.5  9.30e-13
```

Plotting the Fitted Curve

To aid interpretation of the effect of median SAT score on graduation rate, we can plot the fitted curve. Recall that our fitted equation was:

$$\widehat{\text{Graduation Rate}}_i = -306.7 + 106.4\left[\log_2(\text{SAT}_i)\right]$$

This relates the log-transformed SAT scores to graduation rates. We need to determine the relationship between raw SAT scores and graduation rates in order to appropriately plot the fitted curve. To do so we will express the fitted equation more generally:

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1\left[\log_2(X_i)\right]$$

Now we will work through the algebra to relate $\hat{Y}_i$ to X_i. Becuase there is a log of X we need to exponentiate both sides of the equation to get back to the raw X.

$$2^{\hat{Y}_i} = 2^{\hat{\beta}_0 + \hat{\beta}_1 \left[\log_2(X_i) \right]}$$

Then we can use the algebraic rules of exponents and logarithms to re-express this equation.

$$\begin{aligned}
2^{\hat{Y}_i} &= 2^{\hat{\beta}_0} \times 2^{\hat{\beta}_1 \left[\log_2(X_i) \right]} \\
2^{\hat{Y}_i} &= 2^{\hat{\beta}_0} \times \left[2^{\log_2(X_i)} \right]^{\hat{\beta}_1} \\
2^{\hat{Y}_i} &= 2^{\hat{\beta}_0} \times X_i^{\hat{\beta}_1}
\end{aligned}$$

Now the only problem is that the left-hand side is $2^{\hat{Y}_i}$ and not Y_i. To fix this we take the logarithm (base-2) of both sides of the equation.

$$\begin{aligned}
\log_2 \left(2^{\hat{Y}_i} \right) &= \log_2 \left(2^{\hat{\beta}_0} \times X_i^{\hat{\beta}_1} \right) \\
\hat{Y}_i &= \log_2 \left(2^{\hat{\beta}_0} \times X_i^{\hat{\beta}_1} \right)
\end{aligned}$$

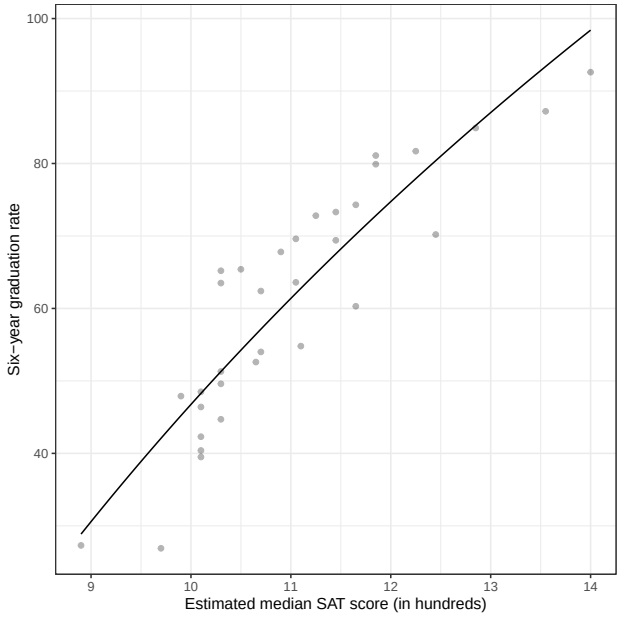
We now have an expression of the curve that relates the outcome to the predictor. In our example, the fitted equation can be re-expressed as:

$$\widehat{\text{Graduation Rate}}_i = \log_2 \left(2^{-306.7} \times \text{SAT}_i^{106.4} \right)$$

We can implement this into the `stat_function()` layer of our `ggplot()` syntax.

```
# Plot
ggplot(data = mn, aes(x = sat, y = grad)) +
    geom_point(alpha = 0.3) +
    stat_function(
      fun = function(x) {log(2^-306.7 * x^106.4 , base = 2)}
      ) +
    theme_bw() +
  xlab("Estimated median SAT score (in hundreds)") +
  ylab("Six-year graduation rate")
```

Figure 5
Plot of Predicted Graduation Rate as a Function of Median SAT Score



Note. The non-linearity in the plot indicates that there is a diminishing positive effect of SAT on graduation rates. We can also use the expression given in the fitted equation. The figure (not shown) would be identical.

```
# Plot
ggplot(data = mn, aes(x = sat, y = grad)) +
    geom_point(alpha = 0.3) +
    stat_function(
      fun = function(x) {-306.7 + 106.4 * log(x, base = 2)}
      ) +
    theme_bw() +
  xlab("Estimated median SAT score (in hundreds)") +
  ylab("Six-year graduation rate")
```

Different Base Values in the Logarithm

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

```
mn = mn %>%
  mutate(
    L10sat = log(mn$sat, base = 10)
  )

# Examine data
head(mn)
```

```
# A tibble: 6 x 8
```

```
     id name                              grad public   sat tuition L2sat L10sat
  <dbl> <chr>                            <dbl>  <dbl> <dbl>   <dbl> <dbl>  <dbl>
1     1 Augsburg College                  65.2      0  10.3    39.3  3.36   1.01
2     3 Bethany Lutheran College          52.6      0  10.6    30.5  3.41   1.03
3     4 Bethel University, Saint Paul, ~  73.3      0  11.4    39.4  3.52   1.06
4     5 Carleton College                  92.6      0  14      54.3  3.81   1.15
5     6 College of Saint Benedict         81.1      0  11.8    43.2  3.57   1.07
6     7 Concordia College at Moorhead     69.4      0  11.4    36.6  3.52   1.06
```

Comparing the logarithms of the SAT attribute 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,

$$10^{1.013} = 10.3$$

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

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

# Model-level output
glance(lm.2)
```

```
# A tibble: 1 x 11
  r.squared adj.r.squared sigma statistic  p.value    df logLik   AIC   BIC
      <dbl>         <dbl> <dbl>     <dbl>    <dbl> <int>  <dbl> <dbl> <dbl>
1     0.811         0.805  7.39      133. 9.30e-13     2  -112.  230.  234.
# ... with 2 more variables: deviance <dbl>, df.residual <int>
```

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, $F(1,\ 31) = 133.3$, $p < .001$. These model-level results are the same as when we used the base-2 logarithm.

```
# Coefficient-level output
tidy(lm.2)
```

```
# A tibble: 2 x 5
  term                estimate std.error statistic  p.value
  <chr>                  <dbl>     <dbl>     <dbl>    <dbl>
1 (Intercept)            -307.      31.9     -9.62 7.94e-11
2 log(sat, base = 10)     354.      30.6     11.5  9.30e-13
```

The fitted equation is,

$$\widehat{\text{Graduation Rate}}_i = -306.7 + 353.6\Big[\log_{10}(\text{SAT}_i)\Big]$$

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

- The intercept value of -306.7 is the predicted average graduation rate for all colleges/universities with a $\log_{10}(\text{SAT})$ value of 0.
- The slope value of 353.6 indicates that each one-unit difference in $\log_{10}(\text{SAT})$ is associated with a 353.6-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 SAT value of 1 (median SAT score = 100) is -306.7.
- Each *ten-fold* difference in SAT is associated with a 353.6-unit difference in graduation rate, on average.

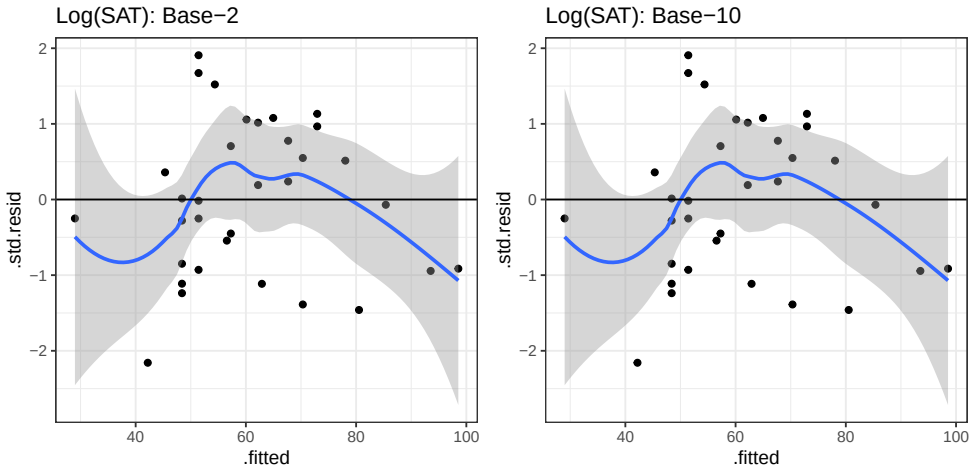
To further think about the effect of SAT, if Augsburg improved its median SAT score ten-fold (i.e., going from a SAT value of 10.3 to a value of 103) we would predict its graduation rate to go up by 353.6 percentage points!

Comparing the Output from the Two Bases

The model-level information is all the same. Furthermore, the intercepts (and SE and p-value) was the same across both models. The slope estimates and SEs were different in the two models, reflecting the change in scale and interpretation. However, the inferential results (t-value and p-value) are the same for both transformations.

What if we look at the residual fit?

Figure 6
Residual Plots for Two Models with Different Logarithmic Bases



Note. Standardized residuals versus the fitted values for the models fitted with the log-2 predictor (left) and the log-10 predictor (right). In both plots the loess smoother (solid, blue line) is displayed along with the 95% confidence envelope (grey shaded area). The reference line of Y=0 is also displayed.

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,

$$\begin{aligned}\widehat{\text{Graduation Rate}} &= -306.7 + 106.4\Big[\log_2(10.3)\Big] \\ &= -306.7 + 106.4\Big[3.36\Big] \\ &= 50.8\end{aligned}$$

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

$$\begin{aligned}\widehat{\text{Graduation Rate}} &= -306.7 + 353.6\Big[\log_{10}(10.3)\Big] \\ &= -306.7 + 353.6\Big[1.01\Big] \\ &= 50.8\end{aligned}$$

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?

The Natural Logarithm: Base-e

In our example, neither of the bases we examined is satisfactory in terms of talking about the effect of median SAT score. Two-fold differences in median SAT scores are very unlikely, to say anything of ten-fold differences. One base that is commonly used for log-transformations because it offers a reasonable interpretation 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 exponent 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(10.3)
```

```
[1] 2.332144
```

If we took $e^{2.332}$ we would obtain 10.3. The natural logarithm even has its own mathematical notation; ln. For example, we would mathematically express the natural logarithm of 10.3 as

$$\ln(10.3) = 2.332.$$

Using the Natural Logarithm in a Regression Model

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

```r
# Fit model
lm.3 = lm(grad ~ 1 + log(sat), data = mn)

# Model-level output
glance(lm.3)
```

```
# A tibble: 1 x 11
  r.squared adj.r.squared sigma statistic  p.value    df logLik   AIC   BIC
      <dbl>         <dbl> <dbl>     <dbl>    <dbl> <int>  <dbl> <dbl> <dbl>
1     0.811         0.805  7.39      133. 9.30e-13     2  -112.  230.  234.
# ... with 2 more variables: deviance <dbl>, df.residual <int>
```

As with any base, using base-e results in the same model-level information ($R^2 = .811$, $F(1,\ 31) = 133.3$, $p < .001$).

```r
# Coefficient-level output
tidy(lm.3)
```

```
# A tibble: 2 x 5
  term        estimate std.error statistic  p.value
  <chr>          <dbl>     <dbl>     <dbl>    <dbl>
1 (Intercept)    -307.      31.9     -9.62 7.94e-11
2 log(sat)        154.      13.3     11.5  9.30e-13
```

The fitted equation is,

$$\widehat{\text{Graduation Rate}}_i = -306.7 + 153.6\Big[\ln(\text{SAT}_i)\Big]$$

The intercept has the same coefficient ($\hat{\beta}_0 = -306.7$), 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 does not correspond to the same amount of change in raw SAT for the three models), the inferential results, $t(31) = 11.55$, $p < .001$, 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/universities with a SAT value of one (median SAT score of 100) is -306.7.
- Interpreting the slope, we could say that an e-fold difference in SAT value is associated with a 153.6-unit difference in graduation rates, on average.

Interpretation Using Percentage Change

Consider three schools, each having a median SAT values that differs by 1%; say these schools have median SAT values of 10, 10.1, 10.201. Using the fitted equation, we can compute the predicted graduation rate for each of these hypothetical schools:

$$\widehat{\text{Graduation Rate}} = -306.7 + 153.6\Big[\ln(\text{SAT})\Big]$$

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

Table 1
Median SAT Values and Graduation Rates for Three Hypothetical Schools that have Median SAT Values that Differ by One Percent.

SAT	Predicted Graduation Rate
10.000	46.8778
10.100	48.4058
10.201	49.9338

The difference between each subsequent predicted graduation rate is 1.53.

```
48.4058 - 46.8778
```

```
[1] 1.528
```

```
49.9338 - 48.4058
```

```
[1] 1.528
```

In other words, schools that have a SAT value that differ by 1%, have predicted graduation rates that differ by 1.53, on average.

Mathematical Explanation

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

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

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 $\hat{\beta}_1\,[\ln(1.01)]$-unit difference in Y, on average.

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

$$153.6\,[\ln(1.01)] = 1.528371$$

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 Interpretation

We can get an approximate estimate for the size of the effect by using the mathematical shortcut of

$$\text{Effect} \approx \frac{\hat{\beta}_1}{100}$$

Using our fitted results, we could approximate the size of 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.

```
# Fit model
lm.4 = lm(grad ~ 1 + public + log(sat), data = mn)

# Model-level output
glance(lm.4)
```

```
# A tibble: 1 x 11
  r.squared adj.r.squared sigma statistic  p.value    df logLik   AIC   BIC
      <dbl>         <dbl> <dbl>     <dbl>    <dbl> <int>  <dbl> <dbl> <dbl>
1     0.866         0.857  6.34      96.6 8.47e-14     3  -106.  220.  226.
# ... with 2 more variables: deviance <dbl>, df.residual <int>
```

The model explains 86.5% of the variation in graduation rates, $F(2,\ 30) = 96.58$, $p < .001$.

```
# Coefficient-level output
tidy(lm.4)
```

```
# A tibble: 3 x 5
  term        estimate std.error statistic  p.value
  <chr>          <dbl>     <dbl>     <dbl>    <dbl>
1 (Intercept)  -286.       28.0     -10.2  2.73e-11
2 public         -8.50      2.44     -3.48 1.56e- 3
3 log(sat)      146.       11.6      12.6  1.74e-13
```

The fitted equation is,

$$\widehat{\text{Graduation Rate}}_i = -286.1 - 8.5(\text{Public}_i) + 146.0\Big[\ln(\text{SAT}_i)\Big]$$

Interpreting each of the coefficients using the raw metric of median SAT value:

- The intercept value of -286.1 is the predicted average graduation rate for all public colleges/universities with a SAT value of 1 (extrapolation).
- Public schools have a predicted graduation rate that is 8.5-percentage points lower, on average, than private schools controlling for differences in median SAT scores ($p = .002$).
- Each 1% difference in median SAT value is associated with a 1.46-percentage point difference in predicted graduation rate, on average, after controlling for differences in sector ($p < .001$).

Plot of the Model Results

To further help interpret these effects, we can plot the fitted curves resulting from this fitted equation. Now we will have two curves one for public schools and one for private schools.

Private

$$\begin{aligned}\widehat{\text{Graduation Rate}}_i &= -286.1 - 8.5(0) + 146.0\Big[\ln(\text{SAT}_i)\Big] \\ &= -286.1 + 146.0\Big[\ln(\text{SAT}_i)\Big]\end{aligned}$$

Public

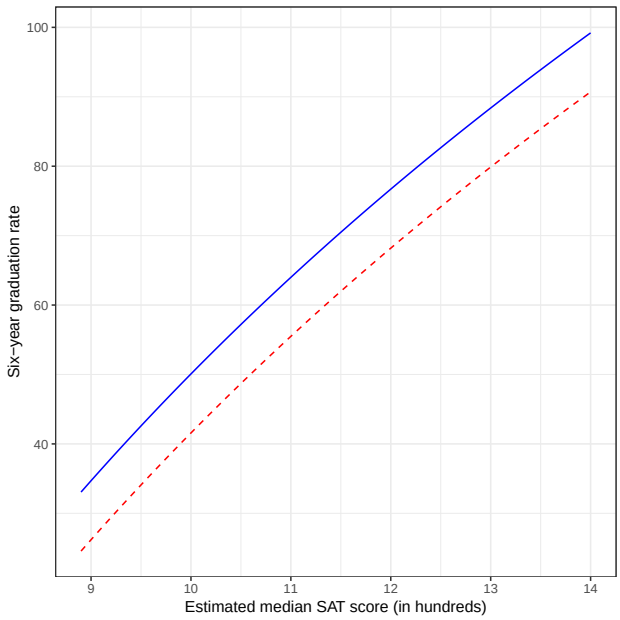
$$\begin{aligned}\widehat{\text{Graduation Rate}}_i &= -286.1 - 8.5(1) + 146.0\Big[\ln(\text{SAT}_i)\Big] \\ &= -294.6 + 146.0\Big[\ln(\text{SAT}_i)\Big]\end{aligned}$$

We can input each of these fitted equations into a `stat_function()` layer of our `ggplot()` syntax.

```r
# Plot
ggplot(data = mn, aes(x = sat, y = grad)) +
    geom_point(alpha = 0) +
    stat_function(
      fun = function(x) {-286.1 + 146.0*log(x)},
      color = "blue"
      ) +
  stat_function(
    fun = function(x) {-294.6 + 146.0*log(x)},
    color = "red",
    linetype = "dashed"
    ) +
    theme_bw() +
  xlab("Estimated median SAT score (in hundreds)") +
  ylab("Six-year graduation rate")
```

Figure 7

Plot of Predicted Graduation Rates as a Function of Median SAT Score for Public (Blue, Solid Line) and Private (Red, Dashed Line) Institutions



The plot shows the nonlinear, diminishing positive effect of median SAT score 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.

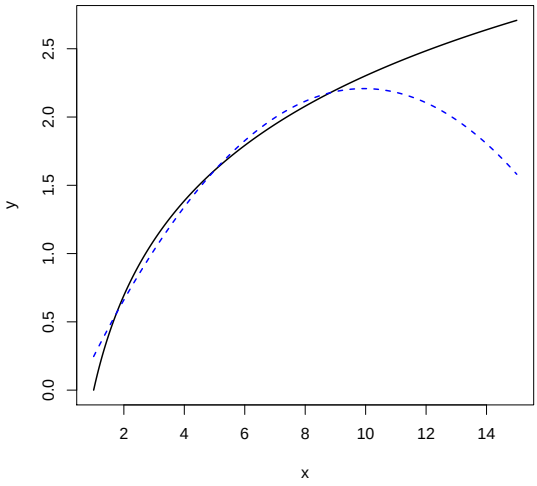
Polynomial Effects vs. Log-Transformations

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).

Figure 8
Comparison of Quadratic (Blue, Dashed Line) and Logarithmic (Black, Solid Line) Functions of X



Both functions are nonlinear, however the 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. (Functions that constantly increase, or constantly decrease, are referred to as *monotonic functions*.) 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. In this case, the residuals would likely not suggest better fit for either of the two models. In this case, it might 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 predictor compared to the two predictors in the quadratic model. The mathematical models make this clear:

$$\text{Polynomial}: \; Y_i = \beta_0 + \beta_1(X_i) + \beta_2(X_i^2) + \epsilon_i$$
$$\text{Log-Transform}: \; Y_i = \beta_0 + \beta_1\Big[\ln(X_i)\Big] + \epsilon_i$$

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.