Log-Transforming Predictors

Andrew Zieffler

April 18, 2016

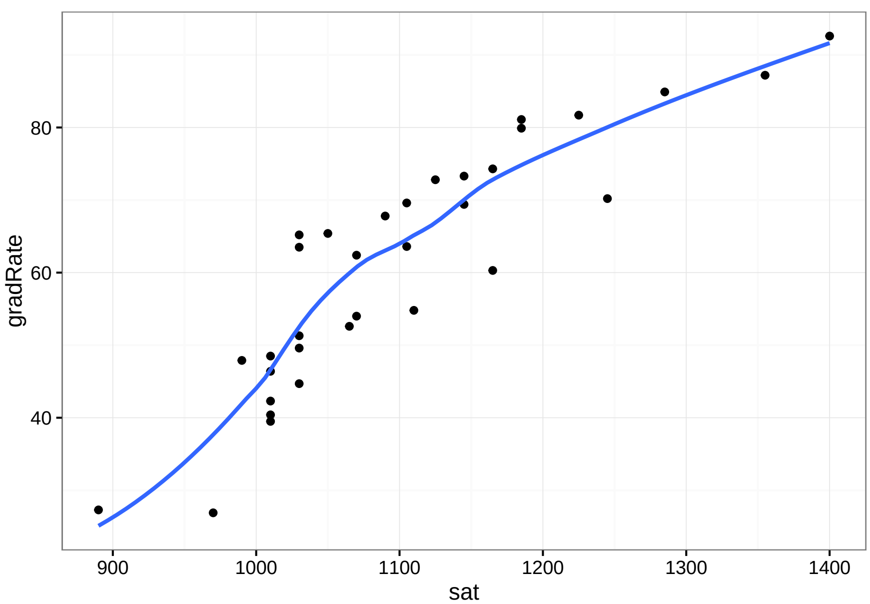
Read in Data and Load Libraries

```
mn = read.csv(file = "/Users/andrewz/Documents/EPsy-8262/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(ggplot2)
library(sm)
```

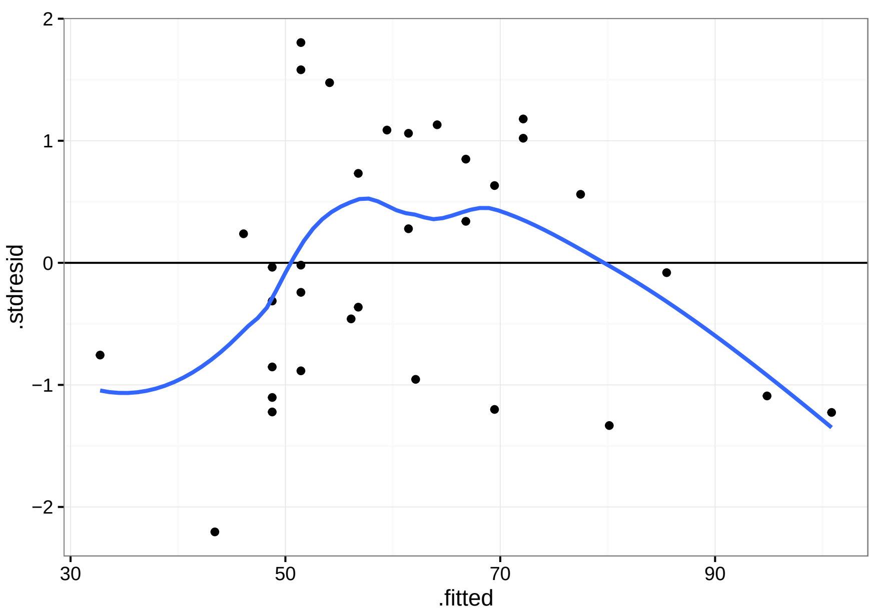
Examine Relationship between Graduation Rate and SAT Scores



The loess line suggests that the relationship between SAT scores and graduation rate is non-linear. A one-unit change in SAT scores does have the same effect on graduation rates. . . for low SAT scores, a one-unit difference in SAT is associated with a larger change in graduation rates than the same one-unit change for higher SAT values.

Sometimes this non-linear relationship is easier to see in the residual plots.

```r
lm.1 = lm(gradRate ~ sat, data = mn)
out = fortify(lm.1)

ggplot(data = out, aes(x = .fitted, y = .stdresid)) +
  geom_point() +
  geom_hline(yintercept = 0) +
  geom_smooth(se = FALSE) +
  theme_bw()
```



The scatterplot of the standardized residuals versus the fitted values suggest that the assumption of linearity is likely violated. There is systematic over-estimation for low fitted values, systematic under-estimation for moderate fitted values, and systematic over-estimation for high fitted values.

Create log base-2 predictor

This function is consistent with a logarithmic relationship. To model this type of function we will transform the predictor using a base-2 logarithm.

```r
mn$L2sat = log(mn$sat, base = 2)
head(mn)
```

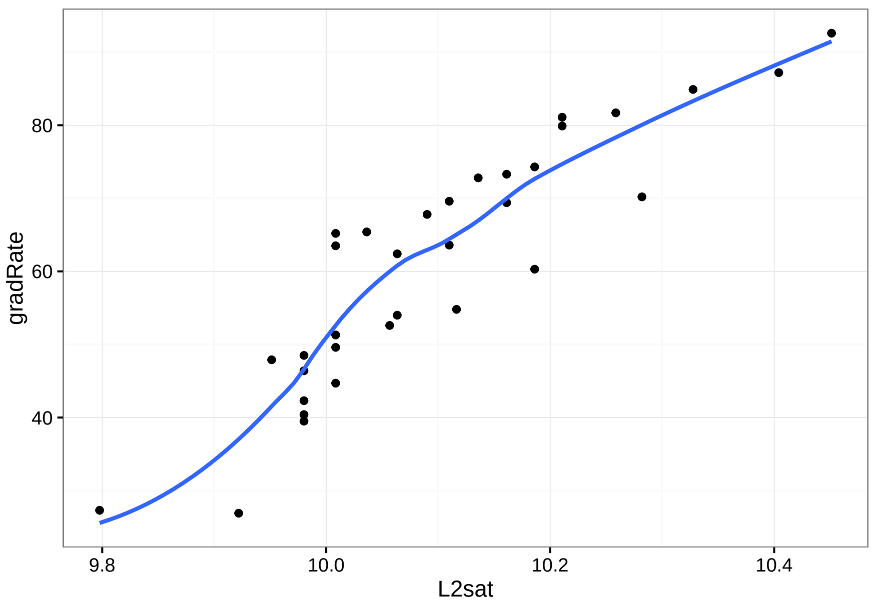
```
##   id                              name gradRate public  sat tuition
## 1  1                  Augsburg College     65.2      0 1030   39294
## 2  3          Bethany Lutheran College     52.6      0 1065   30480
## 3  4 Bethel University, Saint Paul, MN     73.3      0 1145   39400
## 4  5                  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
```

The log base-2 predictor, `L2sat`, is the result of taking $2^{\text{L2sat}} = \text{sat}$. So for Augsburg, $2^{10.00843} = 1030$.

Examining the relationship between the `L2sat` predictor and graduation rates, we see that these variables have a linear relationship.

```
ggplot(data = mn, aes(x = L2sat, y = gradRate)) +
    geom_point() +
    geom_smooth(se = FALSE) +
    theme_bw()
```



Fitting and Interpreting the Log-Transformed Model

```
lm.2 = lm(gradRate ~ L2sat, data = mn)
summary(lm.2)
```

```
##
## Call:
## lm(formula = gradRate ~ 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
```

Note the model-level summary: Differences in log-2 SAT scores, which is the same thing as differences in SAT scores, explains 81.13% of the variation in graduation rates. This is statistically reliable, $F(1,31) = 133.3$, $p < 0.001$.

The fitted equation is

$$\text{grad}\hat{\text{R}}\text{ate} = -1013.872 + 106.439(\text{L2sat})$$

To interpret the coefficients:

- $\hat{\beta}_0 = -1013.872$. This is the average estimated graduation rate when `L2sat` is equal to 0. Equivalently, when `L2sat` $= 0$, SAT $= 2^0 = 1$. The average estimated graduation rate for all school that have an SAT score of 1 is -1013.872.
- $\hat{\beta}_1 = 106.439$. A one-unit difference in `L2sat` is associated with a 106.4% difference in graduation rate, on average. A one-unit difference in `L2sat` is equivalent to a two-fold difference in SAT (e.g., SAT of 200 to an SAT of 400). Thus we interpret the slope here as a two-fold difference in SAT is associated with a 106.4% difference in graduation rate, on average.

Plot the Results of the Log-Transformed Model

Set up a sequence of x-values…in this case `L2sat`, and predict using the fitted model.

```r
# Set up data
plotData = expand.grid(
    L2sat = seq(from = 9.80, to = 10.5, by = 0.1)
    )

# Predict
plotData$yhat = predict(lm.2, newdata = plotData)

# Examine data
head(plotData)
```

```
##   L2sat     yhat
## 1   9.8 29.23190
## 2   9.9 39.87582
## 3  10.0 50.51974
## 4  10.1 61.16366
## 5  10.2 71.80758
## 6  10.3 82.45149
```

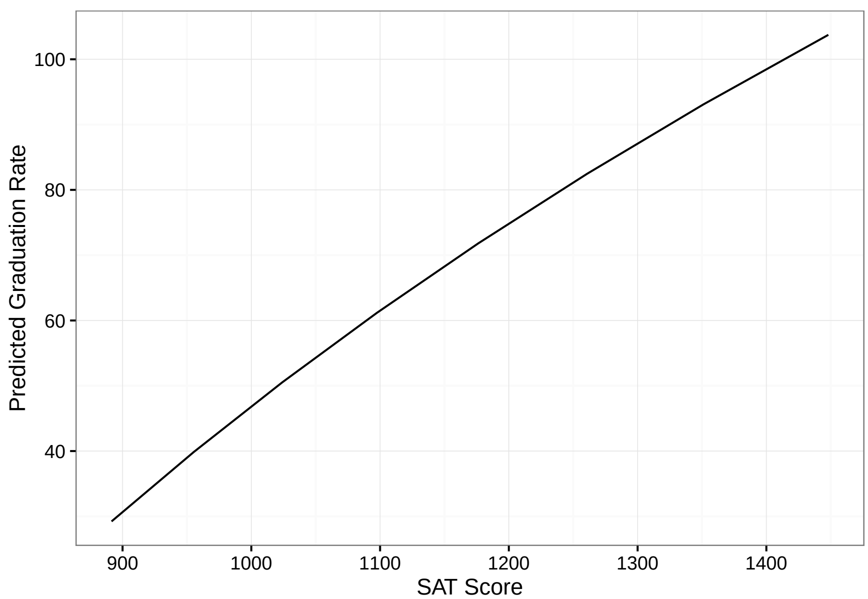
After predicting, we can back-transform the log-2 SAT scores to the original metric.

```r
# Back-transform any log terms
plotData$sat = 2 ^ plotData$L2sat

# Re-examine data
head(plotData)
```

```
##   L2sat     yhat       sat
## 1   9.8 29.23190  891.4438
## 2   9.9 39.87582  955.4258
## 3  10.0 50.51974 1024.0000
## 4  10.1 61.16366 1097.4960
## 5  10.2 71.80758 1176.2671
## 6  10.3 82.45149 1260.6919
```

Now we can plot the back-transformed SAT scores versus the fitted values.

```r
ggplot(data = plotData, aes(x = sat, y = yhat)) +
    geom_line() +
    theme_bw() +
  xlab("SAT Score") +
  ylab("Predicted Graduation Rate")
```



This will display the non-linearity between SAT scores and graduation rates that we observed in the original data.

Changing the Base

Let's see what would happen if we had used the base-10 logarithm of SAT score rather than the base-2 logarithm.

```r
# Create the base-10 logarihm of SAT scores
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
```

The log base-10 predictor, `L10sat`, is the result of taking $10^{\text{L2sat}} = \text{sat}$. So for Augsburg, $10^{3.012837} = 1030$.

Now we will fit the regression model.

```r
lm.3 = lm(gradRate ~ L10sat, data = mn)
summary(lm.3)
```

```
##
## Call:
## lm(formula = gradRate ~ L10sat, 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 ***
## L10sat        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
```

Note the model-level summary: Differences in log-10 SAT scores, which is the same thing as differences in SAT scores, explains 81.13% of the variation in graduation rates. This is statistically reliable, $F(1, 31) = 133.3$, $p < 0.001$. These are the exact same results we obtained when we use the base-2 logarithm.

The fitted equation is

$$\text{gra}\hat{\text{dR}}\text{ate} = -1013.872 + 353.58(\text{L10sat})$$

To interpret the coefficients:

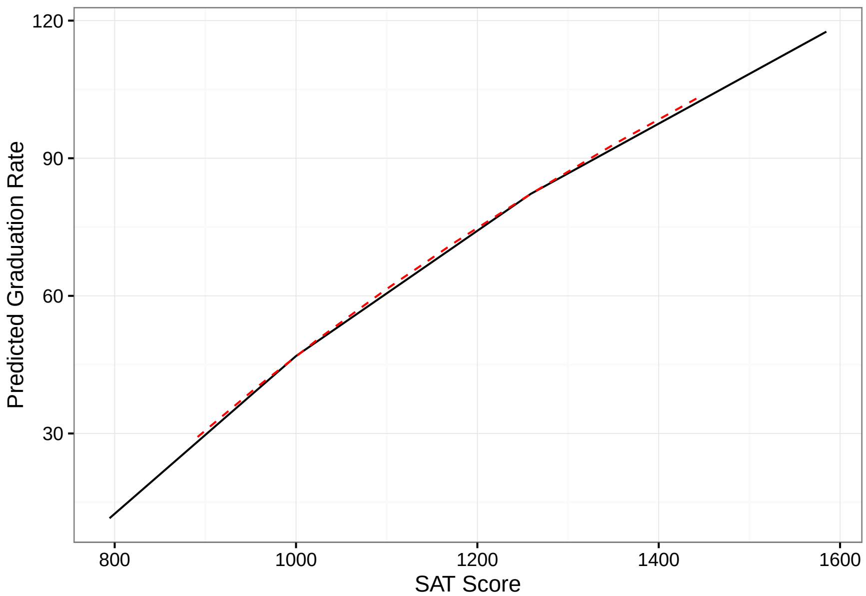
- $\hat{\beta}_0 = -1013.872$. This is the average estimated graduation rate when `L10sat` is equal to 0. Equivalently, when `L10sat` $= 0$, SAT $= 2^0 = 1$. The average estimated graduation rate for all school that have an SAT score of 1 is -1013.872.
- $\hat{\beta}_1 = 353.58$. A one-unit difference in `L10sat` is associated with a 353.6% difference in graduation rate, on average. A one-unit difference in `L10sat` is equivalent to a *ten-fold* difference in SAT (e.g., SAT of 200 to an SAT of 2000). Thus we interpret the slope here as a ten-fold difference in SAT is associated with a 353.6% difference in graduation rate, on average.

Here we plot both log models.

```
# Set up data
plotData2 = expand.grid(
    L10sat = seq(from = 2.9, to = 3.2, by = 0.1)
    )

# Predict
plotData2$yhat = predict(lm.3, newdata = plotData2)

# After predicting, back-transform any log terms
plotData2$sat = 10 ^ plotData2$L10sat

# Examine data
head(plotData)
```

```
##   L2sat     yhat       sat
## 1   9.8 29.23190  891.4438
## 2   9.9 39.87582  955.4258
## 3  10.0 50.51974 1024.0000
## 4  10.1 61.16366 1097.4960
## 5  10.2 71.80758 1176.2671
## 6  10.3 82.45149 1260.6919
```

```
# Plot
ggplot(data = plotData2, aes(x = sat, y = yhat)) +
    geom_line() +
  geom_line(data = plotData, linetype = "dashed", color = "red") +
    theme_bw() +
  xlab("SAT Score") +
  ylab("Predicted Graduation Rate")
```



Here the base-10 log model is shown as a black, solid line. The base-2 log model is shown as a red, dashed line. The lines are on top of each other because the changing the base does not change the relationship; they are the same model. (Note the difference in range over the SAT scores is just a function of the choices I made in `seq()`.)

We can also see this by examining the fitted values and residuals from the two models:

```
# Base-2 residuals
head(fortify(lm.2)[-c(3:5)])
```

```
##   gradRate    L2sat  .fitted    .resid  .stdresid
## 1     65.2 10.00843 51.41687 13.783127  1.9072880
## 2     52.6 10.05664 56.54821 -3.948210 -0.5435624
## 3     73.3 10.16113 67.67048  5.629516  0.7764559
## 4     92.6 10.45121 98.54628 -5.946282 -0.9142367
## 5     81.1 10.21067 72.94342  8.156576  1.1330068
## 6     69.4 10.16113 67.67048  1.729516  0.2385450
```

```
# Base-10 residuals
head(fortify(lm.3)[-c(3:5)])
```

```
##   gradRate   L10sat  .fitted    .resid  .stdresid
## 1     65.2 3.012837 51.41687 13.783127  1.9072880
## 2     52.6 3.027350 56.54821 -3.948210 -0.5435624
## 3     73.3 3.058805 67.67048  5.629516  0.7764559
## 4     92.6 3.146128 98.54628 -5.946282 -0.9142367
## 5     81.1 3.073718 72.94342  8.156576  1.1330068
## 6     69.4 3.058805 67.67048  1.729516  0.2385450
```

Both the fitted values and residuals are identical between the two models. This indicates that (1) the estimated conditional means of Y will be the same regardless of the base chosen, and (2) the model–data fit is the same, regardless of the base chosen.

In general, regardless of the base you choose, the model is the same. The same amount of variation will be explained. The fit based on the residuals will be the same. The only difference will be in the interpretation of the slope coefficient. In our two models, the interpretation was for a *two-fold* difference in SAT sccores or for a *ten-fold* difference in SAT scores. **The base should be chosen to facilitate interpretation.**