Transformations

Andrew Zieffler

Educational Psychology

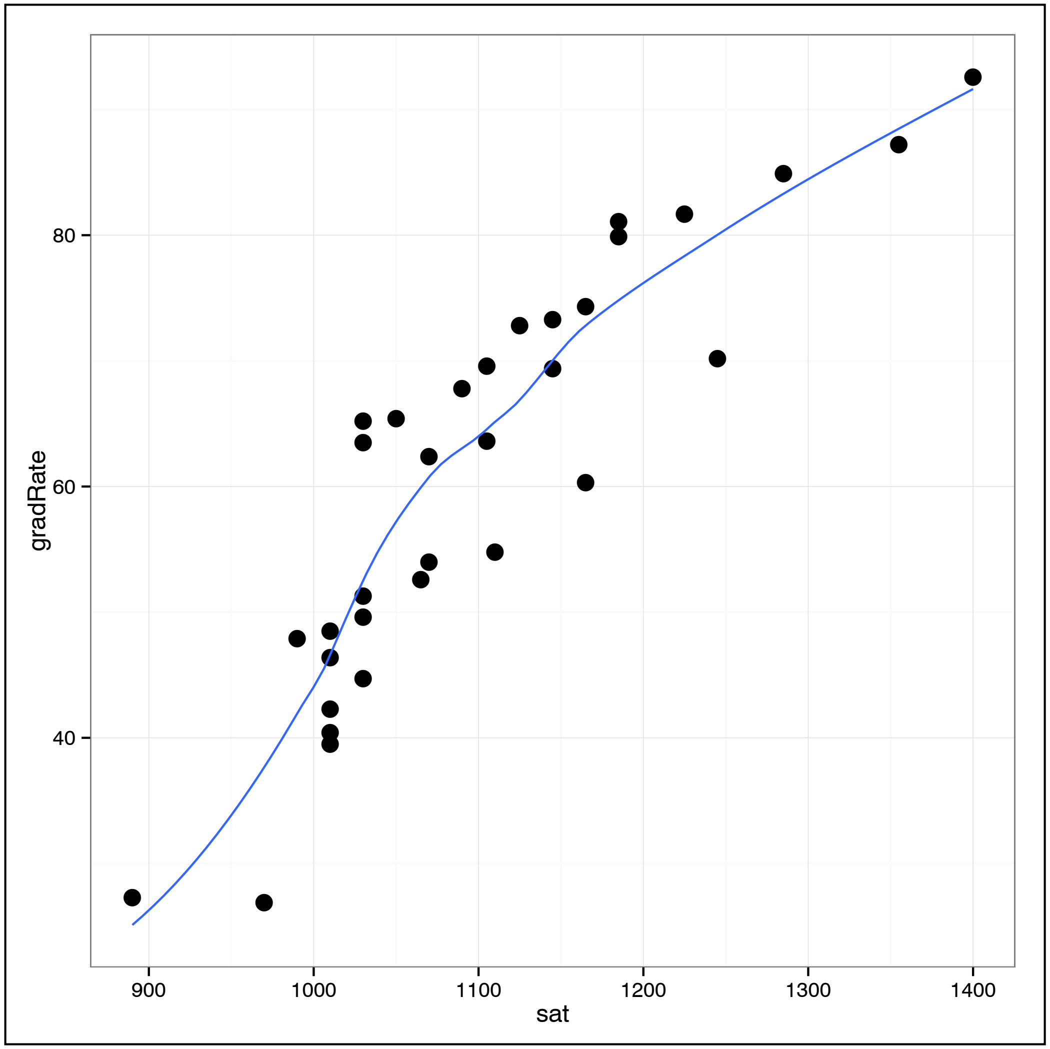
University of Minnesota

Driven to Discover℠

We will use the *mnSchools.csv* data.

```
> mn = read.csv(file = "~/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
```

RQ: Do SAT scores predict variation in graduation rates?



The relationship is non-linear.

This time, rather than fitting a polynomial model to account for the non-linearity, we will use a log-model.

Base-2 Logarithm of SAT Score

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

> head(mn)

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

Augsburg College

$$2^{10.00843} = 1030$$

$$2^{\text{L2sat}} = \text{sat}$$

Bethany Lutheran College

$$2^{10.05664} = 1065$$

College	L2sat	sat
A	8	
B	9	
C	10	
D	11	

Logarithms transform multiplicative differences into additive differences.

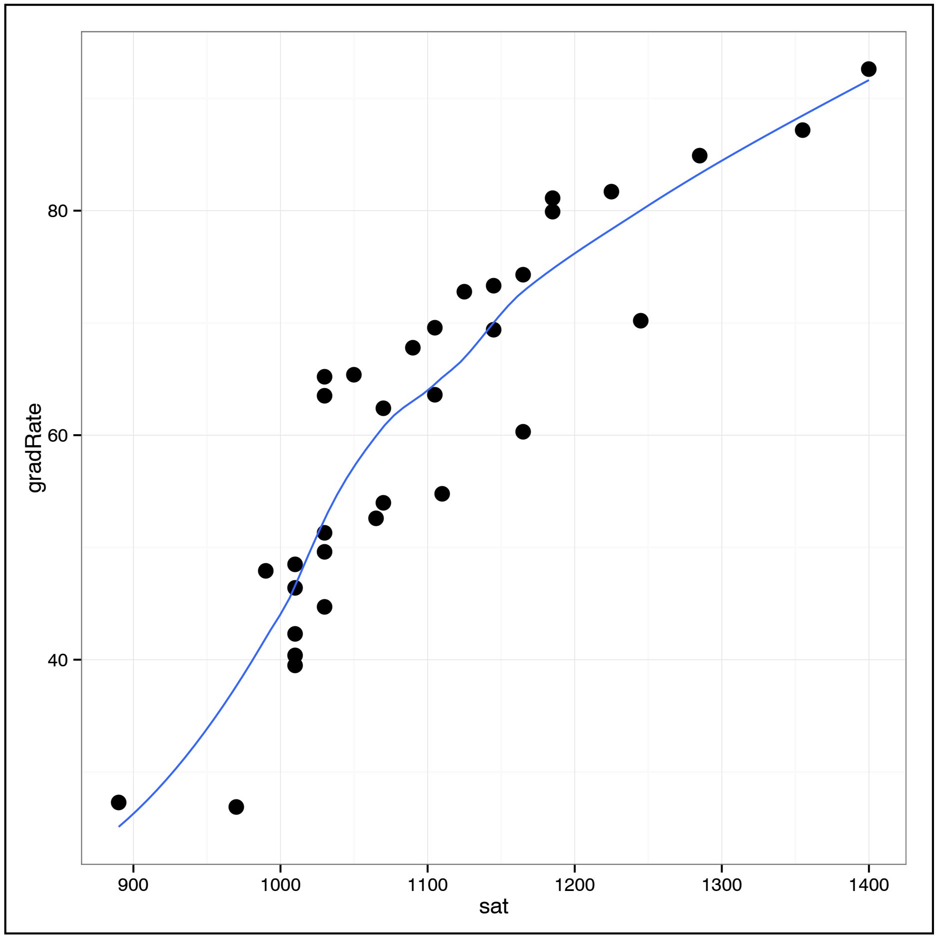
The flood is over and the ark has landed. "Go forth and multiply," Noah tells the animals.

A few months later, he decides to take a stroll and see how the animals are doing. Everywhere he looks he finds baby animals. Everyone is doing fine except for one pair of little snakes. "Please, Noah," say the snakes, "we need you to cut down some trees for us."

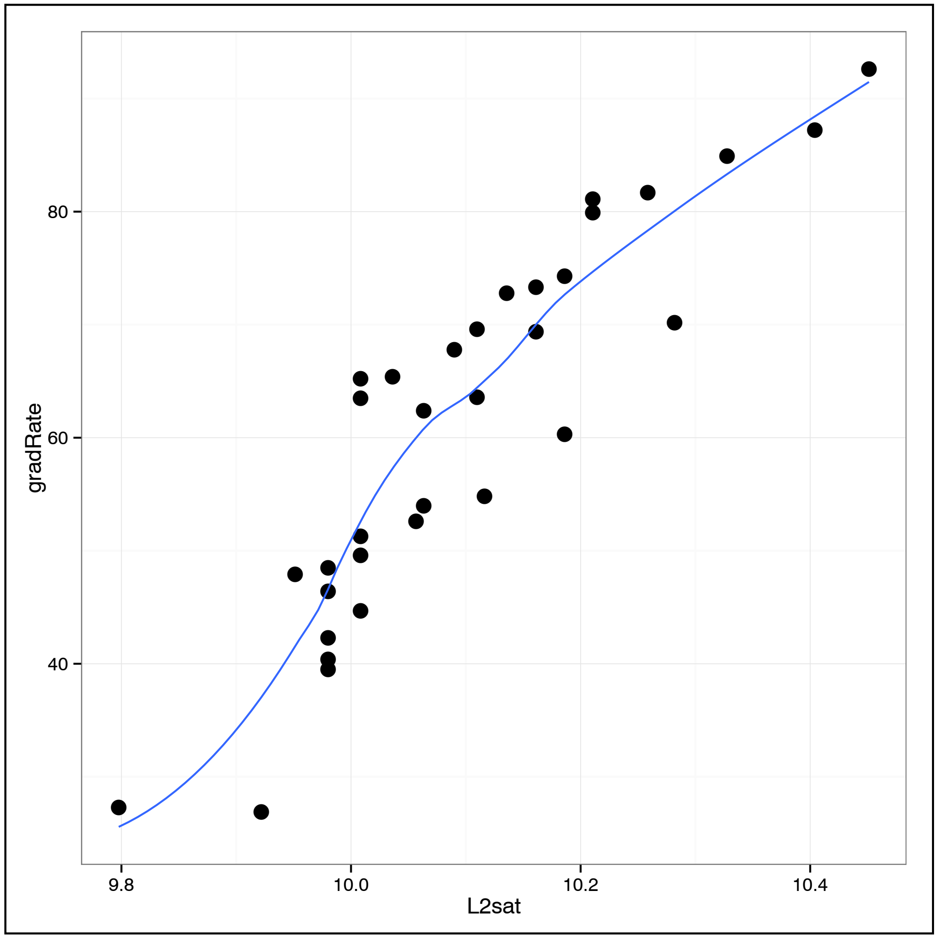
"No problem," says Noah. He cuts down a few trees and goes home scratching his head. A few weeks later he gets curious and come back to check on the snakes. They now have lots of little snakes and everyone is happy. "What happened?" he asks them.

"We are adders," the snakes explain. "So we need logs to multiply."

http://www.math.psu.edu/tseng/mathjoke1.html


gradRate
80
60
40
900
1000
1100
1200
1300
1400
sat

sat


gradRate
80
60
40
9.8
10.0
10.2
10.4
L2sat

L2sat

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

> summary(lm.1)

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
```

Differences in the log (base-2) median SAT scores explain roughly 81% of the variation in graduation rates, $F(1, 31) = 133.3$, $p < .001$.

…but differences in log(x) imply differences in x, so we would say…

Differences in median SAT scores explain roughly 81% of the variation in graduation rates, $F(1, 31) = 133.3$, $p < .001$.

```
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
```

The average graduation rate for all schools with a log (base-2) median SAT score of 0 is predicted to be –1014.

…but when log(x) = 0; x = 1…

The average graduation rate for all schools with a median SAT score of 1 is predicted to be –1014.

Each one-unit difference in the log (base-2) median SAT score is associated with a 106% difference in the predicted graduation rate.

…but a one-unit difference in log(x) is the same as a 2-times difference in x…

Each doubling (two-fold difference) of the median SAT score is associated with a difference of 106% in the predicted graduation rate.

The predicted `gradRate` for a school with a SAT score of 800:

$$\text{L2sat} = \log_2(800) = 9.64$$

$$\text{grad}\hat{\text{R}}\text{ate} = -1014 + 106(9.64) = 7.84$$

The predicted `gradRate` for a school with a SAT score of 1600:

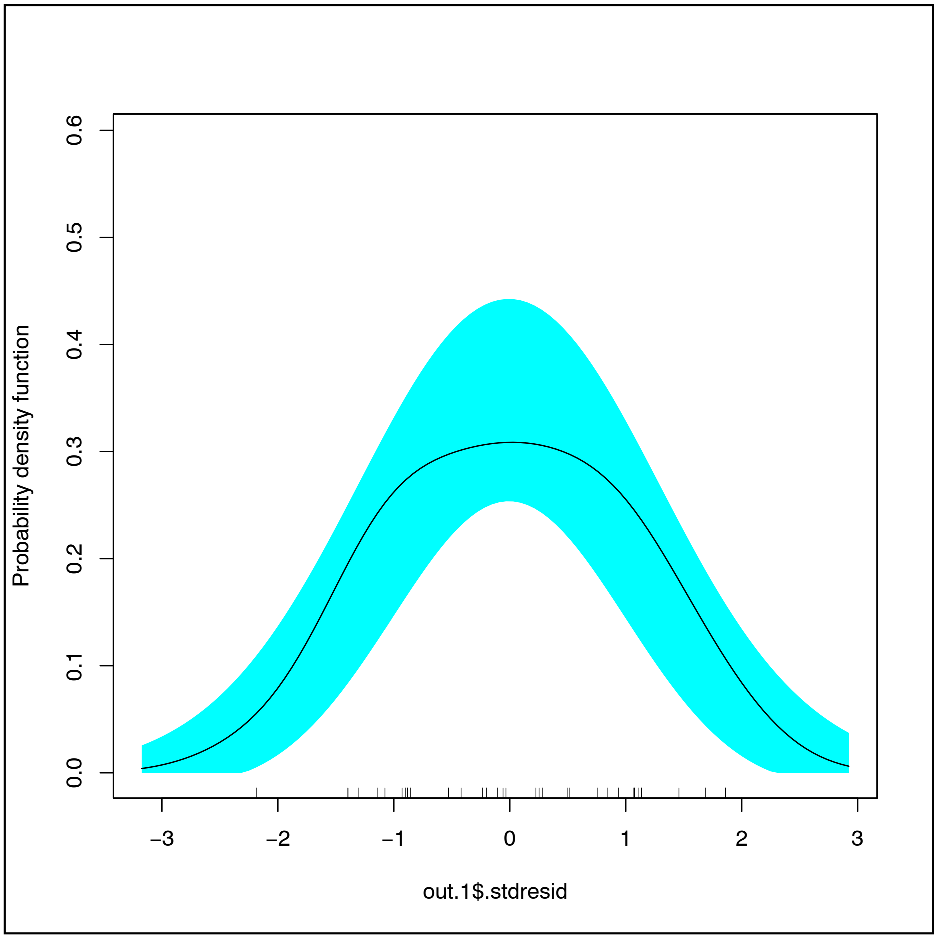
$$\text{L2sat} = \log_2(1600) = 10.64$$

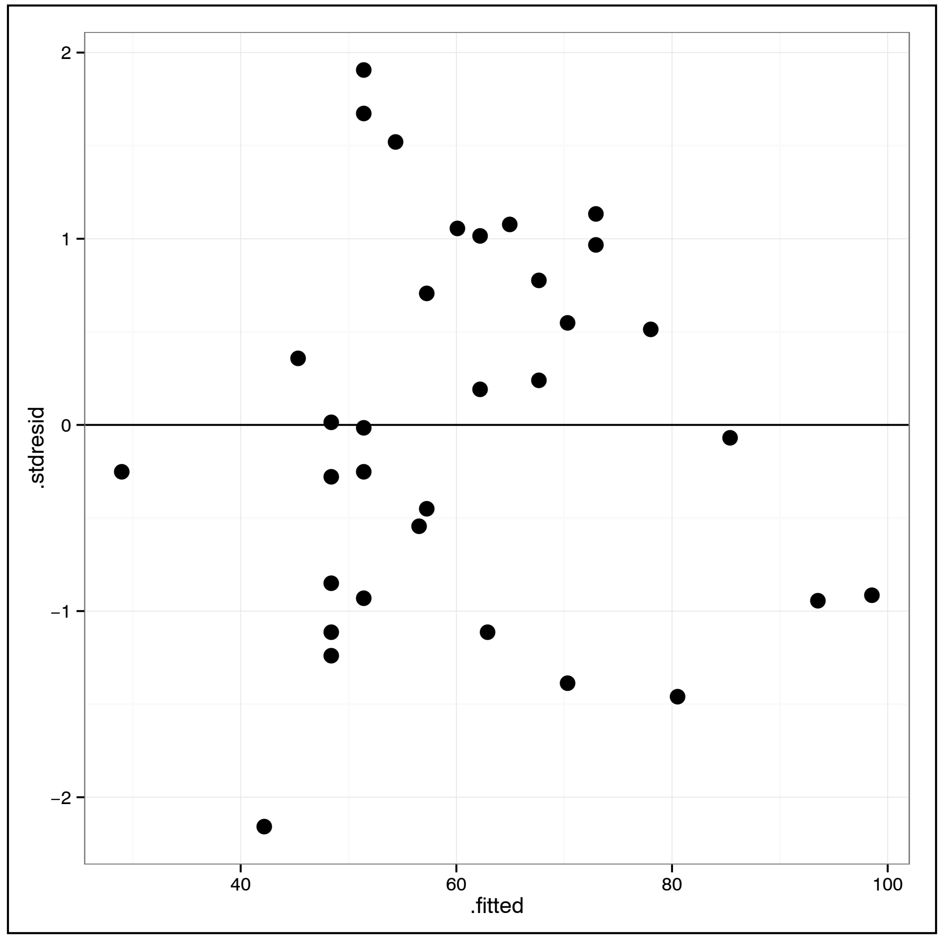
$$\text{grad}\hat{\text{R}}\text{ate} = -1014 + 106(10.64) = 113.84$$

The predicted graduation rates differ by 106.

Be careful…when your variables are measured in percents (i.e., graduation rates), it is easy to say something that is wrong. Here the difference is 106%. But, it would be **wrong** to say that 113.84 is 106% of 7.84!

Check the residuals





Plotting

Set up your predictors to predict from the log model.

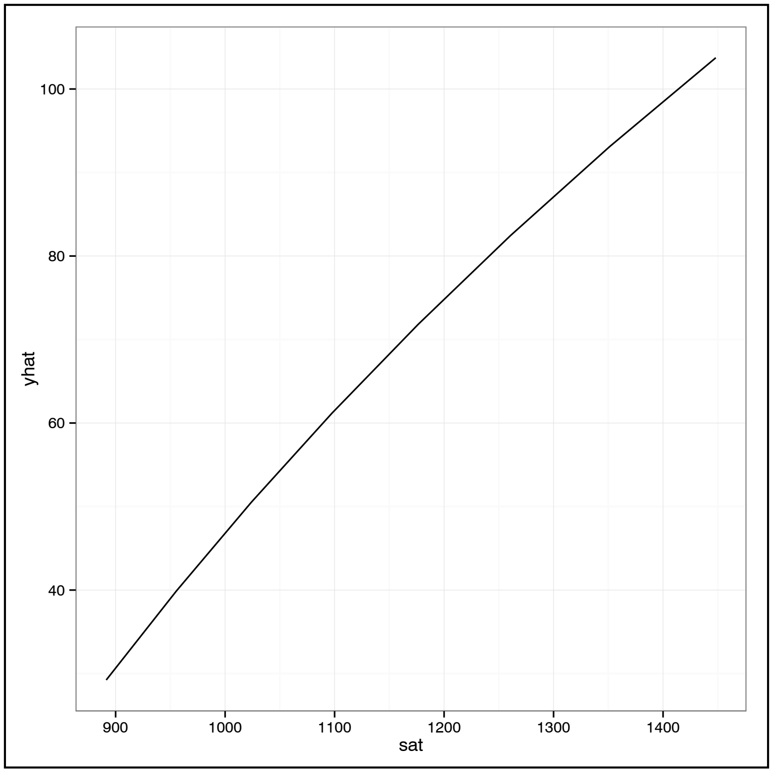
```
> plotData = expand.grid(
    L2sat = seq(from = 9.80, to = 10.5, by = 0.1)
    )

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

> 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
```

Back-transform any variable that you initially log-transformed

```
> plotData$sat = 2 ^ plotData$L2sat

> 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 using the *non-log* predictor and outcome.

```
> ggplot(data = plotData, aes(x = sat, y = yhat)) +
    geom_line() +
    theme_bw()
```



Choosing the Base of the Logarithm

College	L10sat	sat
A	2	
B	3	
C	4	
D	5	

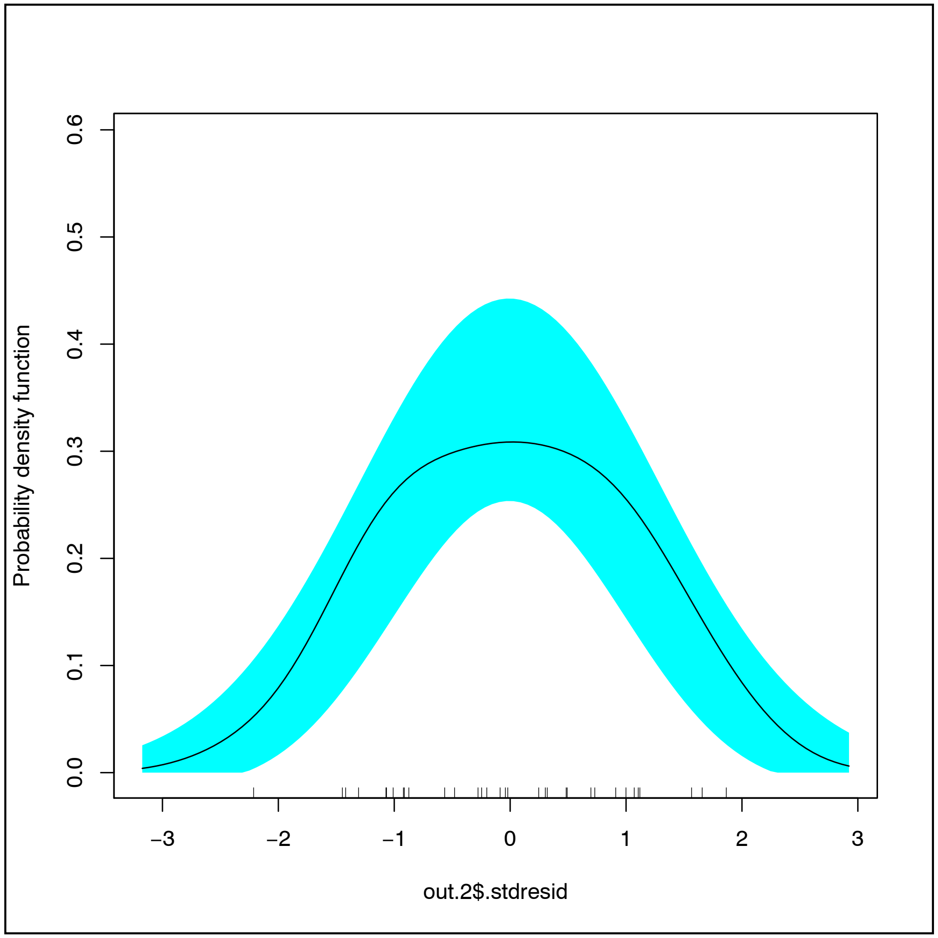
```
> mn$L10sat = log(mn$sat, base = 10)
> lm.2 = lm(gradRate ~ L10sat, data = mn)
> summary(lm.2)


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
```

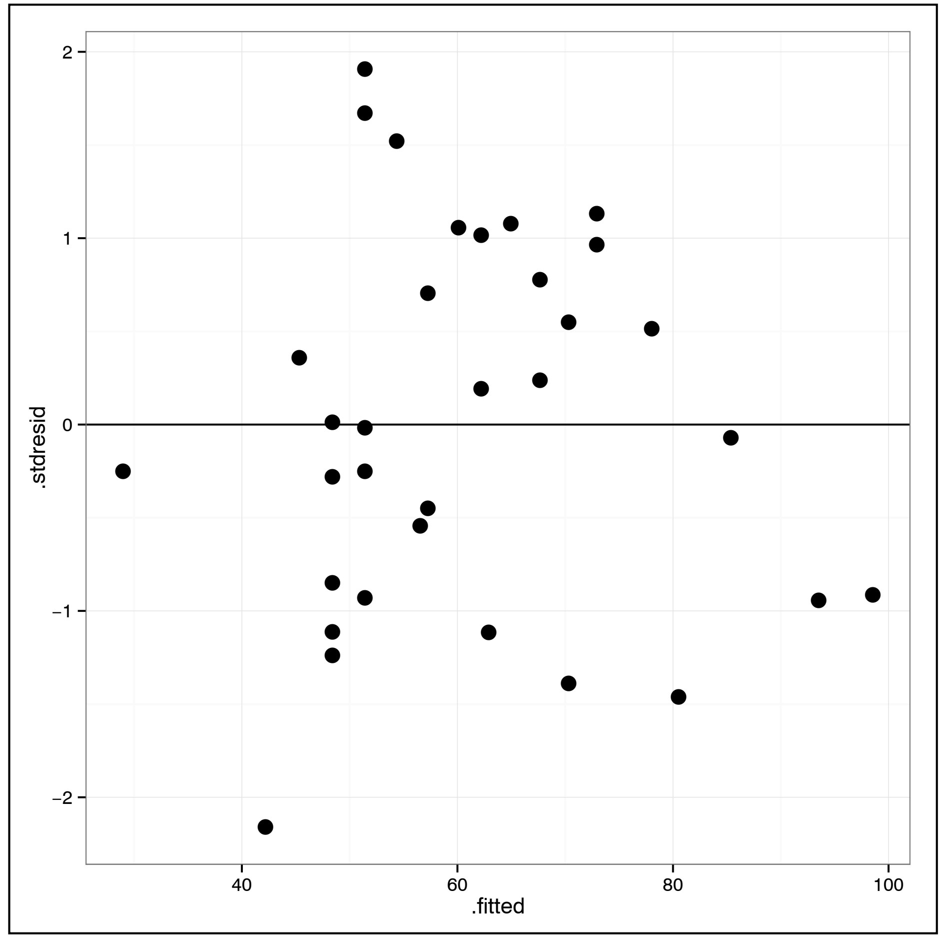
Differences in median SAT scores explain roughly 81% of the variation in graduation rates, $F(1, 31) = 133.3$, $p < .001$. **This is identical to the base-2 choice of logarithm.**

The average graduation rate for all schools with a median SAT score of 1 is predicted to be –1014. **This is identical to the base-2 choice of logarithm.**

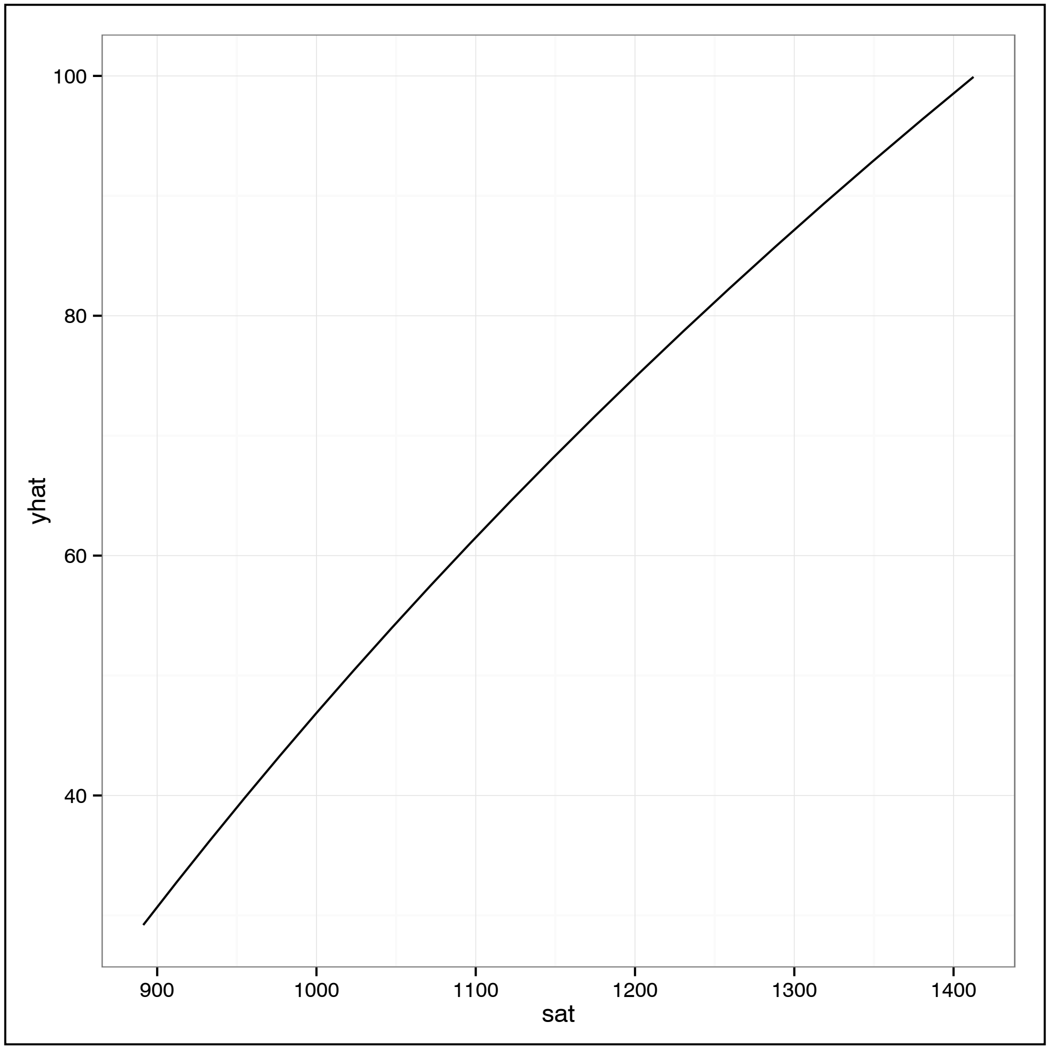
Each **ten-fold difference** in median SAT score is associated with a difference of 353% difference in the predicted graduation rate.

The residuals are identical to the base-2 choice of logarithm….not just the plots, but the size of the residuals. So are the fitted values (y-hats)…so the plots are the same.





Because the fitted values are the same,the plot of the model will also be the same.



Choice of logarithm does not affect the explanation, model fit (residuals) or predictions…at all!

The only thing it affects is the interpretation of the slope (i.e., two-fold difference in SAT scores *vs* ten-fold difference in SAT scores).

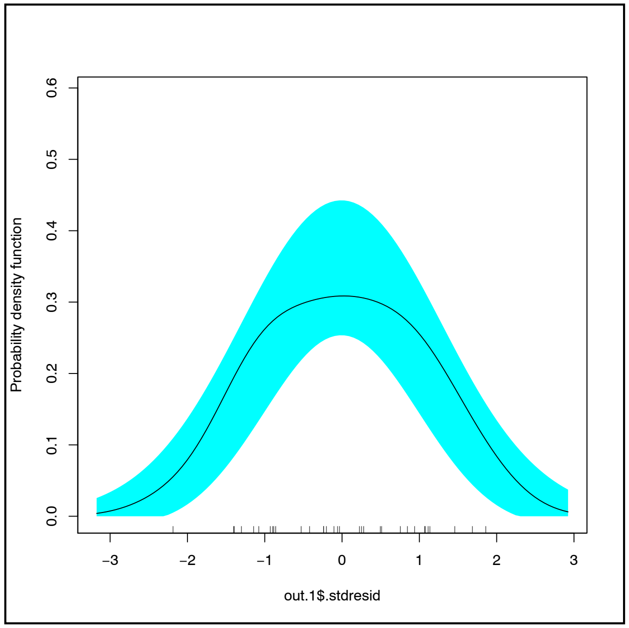
Choose a log base by what kind of differences are realistic…in thinking about SAT scores, ten-fold differences are unrealistic (e.g., 800 —> 8000) ….maybe so are two-fold differences.

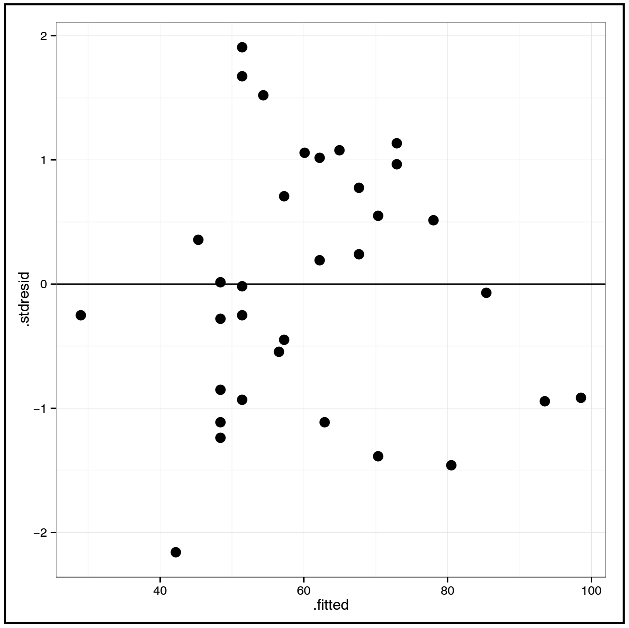
Log-model vs Quadratic model

```
> lm.3 = lm(gradRate ~ sat + lm(sat^2), data = mn)
> summary(lm.3)


Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -3.663e+02  9.862e+01  -3.715 0.000831 ***
sat          6.272e-01  1.727e-01   3.631 0.001040 **
I(sat^2)    -2.150e-04  7.507e-05  -2.864 0.007559 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 7.019 on 30 degrees of freedom
Multiple R-squared:  0.8351, Adjusted R-squared:  0.8241
F-statistic: 75.97 on 2 and 30 DF,  p-value: 1.81e-12
```

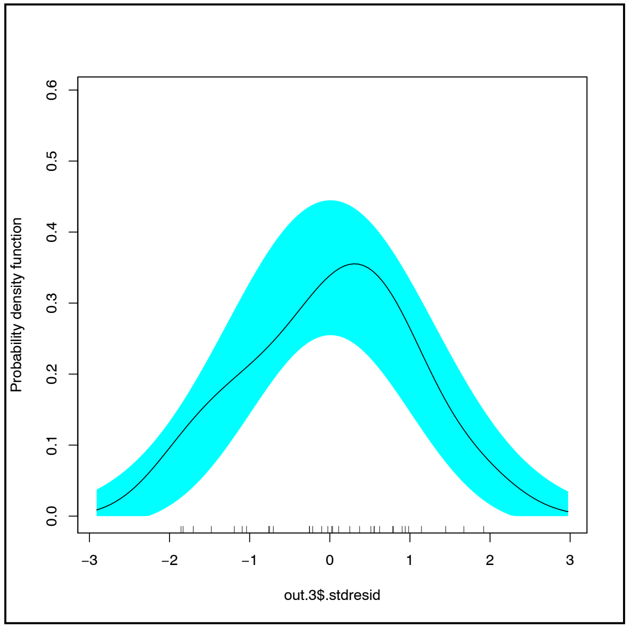
The quadratic model is also statistically reliable and the R^2 value is comparable to the log-model.

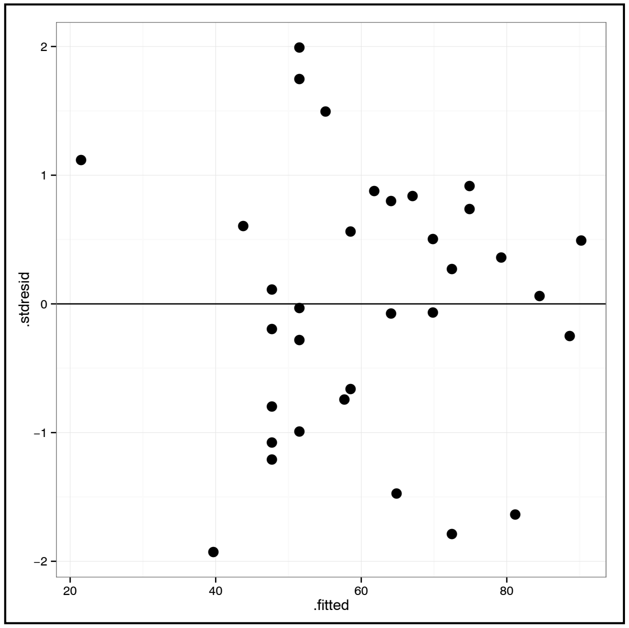
Log-model





Quadratic-model





Occam's Razor of modeling:
If two models fit equally well, choose the simpler model.

In our case the log-transformed model is simpler than the quadratic model because it has one predictor vs the two predictors fro the quadratic model.

Based on the residuals, do they fit equally well?