Interactions-Polynomial Terms

Andrew Zieffler

April 18, 2016

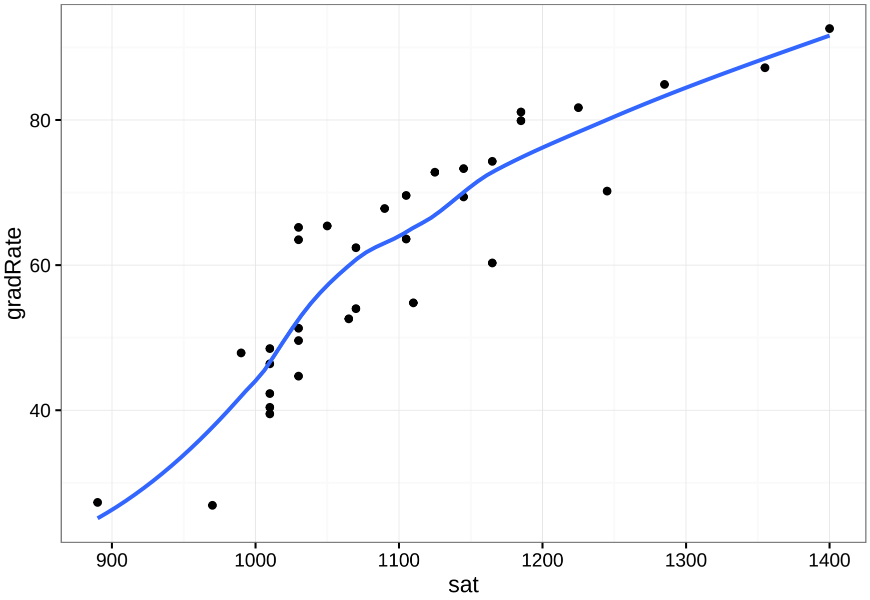
Read in Data and Load Libraries

```
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)
```

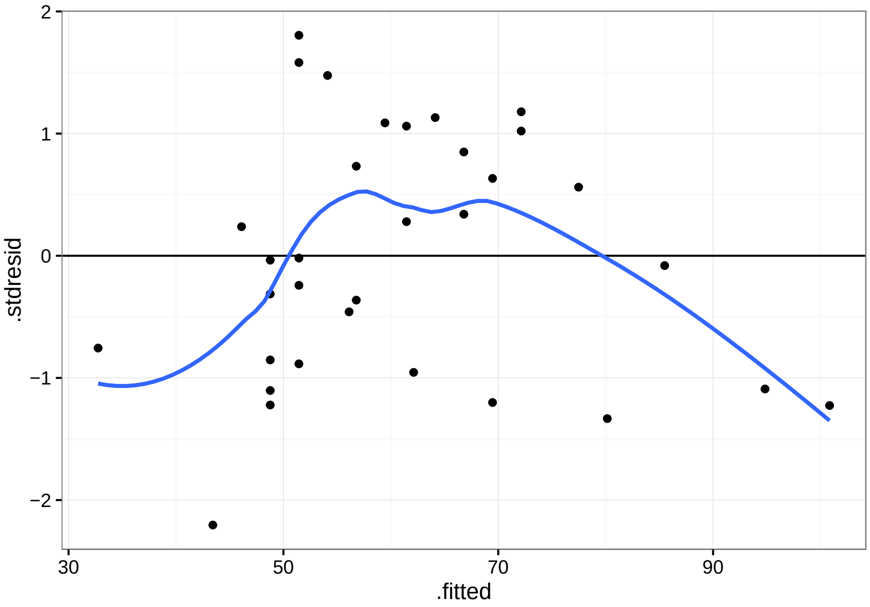
Examine Relationship between Graduation Rate and SAT Scores



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.

```
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.

Polynomials

One way of modeling non-linearity is with the use of polynomials. In regression, a polynomials would include predictors that have powers other than one. For example, x^2 (quadratic term), or x^3 (cubic term). Note that

$$x^2 = x \times x.$$

So the quadratic term, x^2 is a product of x times itself. Recall that products are how we express interactions. Thus the quadratic term of x^2 is really the interaction of x with itself. To model this, we simply (1) create the product term, and (2) include the product term and all constituent main-effects in the regression model.

```
mn$sat_quadratic = mn$sat * mn$sat
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
##   sat_quadratic
## 1       1060900
## 2       1134225
## 3       1311025
## 4       1960000
## 5       1404225
## 6       1311025
```

```
# Fit model
lm.1 = lm(gradRate ~ sat + sat_quadratic, data = mn)
summary(lm.1)
```

```
##
## Call:
## lm(formula = gradRate ~ sat + sat_quadratic, data = mn)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -12.7867  -5.0969   0.3968   5.0011  13.6869
##
## 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 **
## sat_quadratic -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
```

Since this is an interaction model, we start by examining the interaction term. This term is statistically reliable ($p = .008$), suggesting that the quadratic term explains variation above and beyond the linear term. This suggests that we should keep the quadratic term in the model.

Interpretation of a Significant Polynomial Term

How do we interpret the quadratic term? First, we will write out the fitted model.

$$\hat{\text{Graduation Rate}} = -366.3 + 0.63(\text{SAT}) - 0.0002(\text{SAT}^2)$$

From algebra, you may remember that the coefficient in front of the quadratic term (-0.0002) informs us of whether the quadratic is an upward-facing U-shape, or a downward-facing U-shape. Since our term is negative, the U-shape is downward-facing. It also indicates whether the U-shape is skinny or wide. The intercept and linear terms help us locate the U-shape in the coordinate plane (moving it right, left, up, or down from the origin). You could work these out algebraically, but typically, we will just plot the predicted values and interpret from the plot.

Refit the model using the I() function

Before we create the plot, we use a different method of fitting polynomial terms in a regression. Rather than create a new variable in the data set, we insert the polynomial directly into the model using the `I()` function.

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

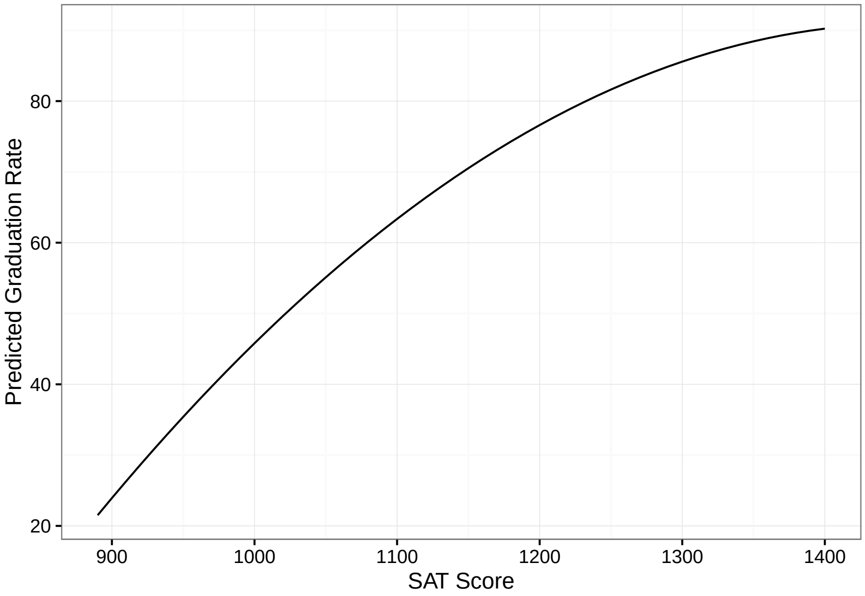
```
##
## Call:
## lm(formula = gradRate ~ sat + I(sat^2), data = mn)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -12.7867  -5.0969   0.3968   5.0011  13.6869
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -3.663e+02  9.862e+01  -3.715 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 `I()` function forces R to create a separate term for the quadratic (which is essentially what we do by adding the product term to the model). Without it, R will do the computation `sat + sat^2` and use those values as a single variable…not what we want. The other advantage for plotting is that we have only used a single predictor, `sat` in specifying the model. Thus we only need to include `sat` in our plot data rather than both `sat` and `sat_quadratic`.

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

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

# Examine data
head(plotData)
```

```
##   sat     yhat
## 1 890 21.50981
## 2 900 23.93244
## 3 910 26.31207
## 4 920 28.64869
## 5 930 30.94230
## 6 940 33.19291
```

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

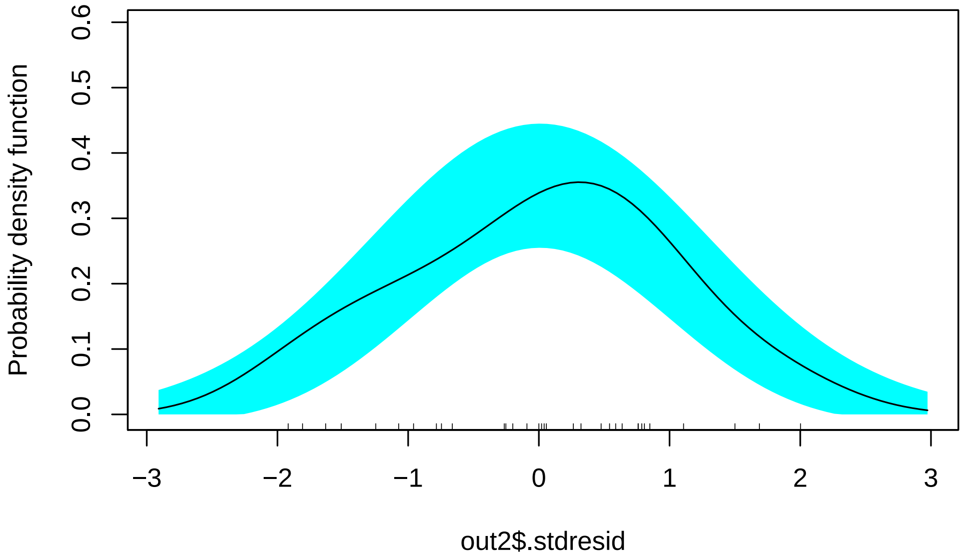


The quadratic relationship is expressed in the predicted values. Aside from plotting them versus SAT scores, there is nothing further we need to do to get the quadratic relationship to appear. The plot, more importantly, helps us interpret the relationship between SAT scores and graduation rates. The effect of SAT on graduation rate depends on SAT score (definition of an interaction). For schools with low SAT scores, the effect of SAT score on graduation rate is positive and fairly high. For schools with high SAT scores, the effect of SAT score on graduation rate remains positive, but it has a smaller effect on graduation rates.

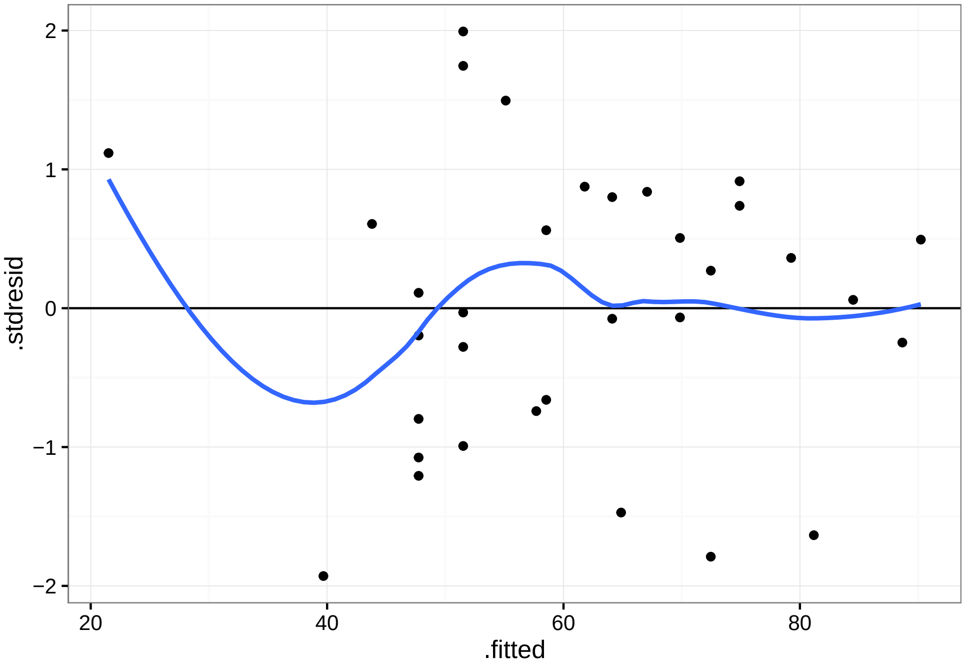
Re-Examining the Residuals for the Quadratic Model

Since we fitted a different model, we should examine the residuals to see whether the assumptions for the model seem satisfied.

```
out2 = fortify(lm.2)

# Check normality
sm.density(out2$.stdresid, model = "normal")
```



```r
# Check other assumptions
ggplot(data = out2, aes(x = .fitted, y = .stdresid)) +
  geom_point() +
  geom_hline(yintercept = 0) +
  geom_smooth(se = FALSE) +
  theme_bw()
```



Based on the plots, the quadratic model seems to meet the assumptions for regression. At the very least, it clearly does so better than the linear model.

Adding Other Predictors

We can also control for other predictors by including them as terms in the model as well. Below we include the `public` predictor.

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

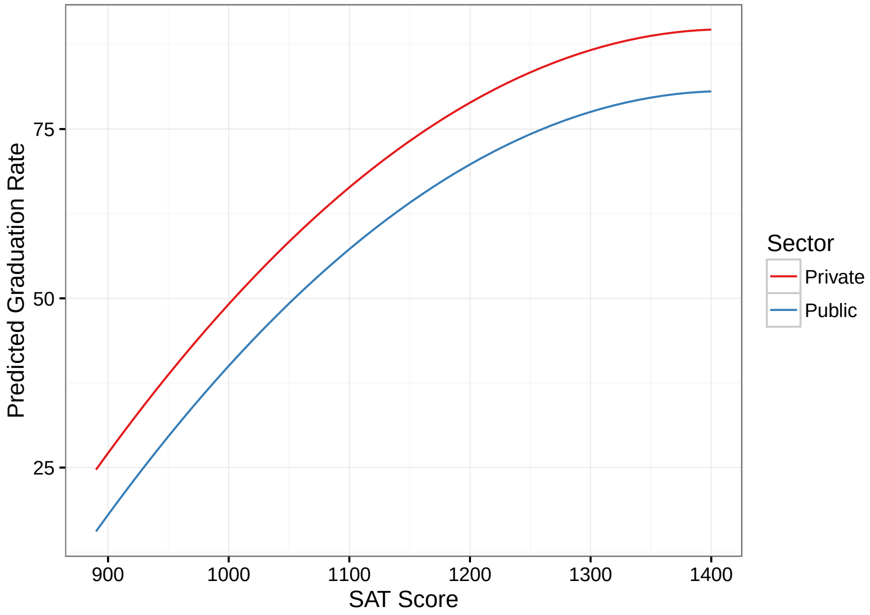
```
##
## Call:
## lm(formula = gradRate ~ sat + I(sat^2) + public, data = mn)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -16.139  -3.207   0.687   2.902  10.388
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -3.842e+02  7.941e+01  -4.838 3.98e-05 ***
## sat          6.704e-01  1.393e-01   4.814 4.25e-05 ***
## I(sat^2)    -2.371e-04  6.059e-05  -3.912 0.000507 ***
## public      -9.125e+00  2.187e+00  -4.171 0.000251 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.644 on 29 degrees of freedom
## Multiple R-squared:  0.897,  Adjusted R-squared:  0.8863
## F-statistic: 84.14 on 3 and 29 DF,  p-value: 2.048e-14
```

Here there is still a quadratic effect of SAT on graduation rates, even after controlling for differences in sector. Similarly, there are differences in sector (public schools have a graduation rate 9.1% lower than private schools, on average), even after controlling for the linear and quadratic effects of SAT. Plot the model to aid interpretation.

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

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

# Coerce public into a factor for better plotting
plotData$public = factor(plotData$public, levels = c(0, 1), labels = c("Private", "Public"))

# Examine data
head(plotData)
```

```
##   sat  public     yhat
## 1 890 Private 24.68509
## 2 900 Private 27.14522
## 3 910 Private 29.55794
## 4 920 Private 31.92324
## 5 930 Private 34.24112
## 6 940 Private 36.51159
```

```r
# Plot
ggplot(data = plotData, aes(x = sat, y = yhat, group = public, color = public)) +
    geom_line() +
  theme_bw() +
  xlab("SAT Score") +
  ylab("Predicted Graduation Rate") +
  scale_color_brewer(name = "Sector", palette = "Set1")
```



The plot shows the quadratic effect of SAT scores on graduation rate; the effect of SAT on graduation rates is positive, but this effect declines for increasingly higher SAT scores, after controlling for sector differences. Private schools have higher graduation rates, on average, than public schools for all levels of SAT score.